16 Testing Measures of Association

In Chapters 6 and 7 the two most common measures of association were presented: regression and correlation coefficients. The regression coefficient, or b, measures the change in the criterion variable as a function of a one-unit change in the predictor variable. The correlation coefficient, or r, is a regression coefficient with both variables standardized (expressed as Z scores); that is, both variables have means of zero and variances of one. It is a symmetric measure of association that varies from -1 to $+1$. For both measures of association, a value of zero indicates no linear association. In this chapter, methods are presented to test hypotheses about correlation and regression coefficients.

The use of the correlation coefficient or r does not require a specification about the direction of the causal effect. That is, if a researcher correlates the degree to which a parent uses physical punishment and how aggressive the children are, there is no need to make any assumptions about which of the following causal patterns is true.

1. Physical punishment causes aggression.
2. Aggressive children make parents use physical aggression.
3. Physical punishment and aggression are two signs of a troubled family.

Correlations make no presumption about what is the independent variable and what is the dependent variable. There is then not a single complete model when one tests correlation coefficients because there are three distinct ways in which the correlation could come about. The complete model then presumes some unspecified causal network that brings about association between the variables. The restricted model is that there is no association between the variables.

A regression coefficient, when used in explanation and not in prediction or description, does make a clear statement about a causal ordering. The predictor variable is the independent variable and the criterion is the dependent variable. The complete model is

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

The coefficient in the model is called the *regression coefficient*. Both variables are measured at the interval level of measurement. The complete model for regression is the same as one-way analysis of variance except that the independent variable is measured at the interval level. The restricted model is the same as the complete model, but the independent variable has no effect on the dependent variable.

In this chapter, first tests of correlation coefficients are presented because procedures to do so are relatively simpler than tests of regression coefficients. Then the somewhat more complicated tests of regression coefficients are described. In the last section of the chapter, rules for determining which type of test is most appropriate are presented.

Tests of Correlation Coefficients

In this section, the following tests of correlation coefficients are presented:

1. How to test a single correlation coefficient.
2. How to test whether two correlation coefficients computed from different samples are equal. Correlations computed using different groups of persons are called *independent* correlations.
3. Testing more than two independent correlation coefficients.
4. How to test whether two correlations computed from the same sample are equal to each other.

A Single Correlation Coefficient

Consider two variables: the number of times a person nods his or her head in a conversation and the degree to which the person likes his or her partner in the conversation. The two variables are nods and liking. The two variables are correlated across 30 pairs of persons and the correlation is .45. The correlation indicates that the more one nods during a conversation the more one likes one's partner. However, one might wonder whether the .45 value in the sample might have just occurred by chance. That is, if there were a thousand pairs of persons, would the correlation be zero? Is the .45 value due to sampling error or does it reflect a true positive correlation? A way is needed to evaluate whether a sample correlation coefficient is significantly different from zero.

It turns out that the distribution of r does not closely correspond to any of the major sampling distributions. However, for a population correlation of zero, r divided by the square root of $1 - r^2$ is approximately normally distributed with a mean of zero and a variance of $n - 2$.

The test of the null hypothesis that a correlation coefficient equals zero is

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

In words, the correlation coefficient is divided by the square root of one minus the correlation squared, and then this quantity is multiplied by the square root of the sample size minus two. Under the null hypothesis that the population correlation equals zero, this quantity has a t distribution with $n - 2$ degrees of freedom, where n is the number of pairs of scores. So, one computes $r\sqrt{(n-2)/(1-r^2)}$ and determines whether it equals or exceeds the critical values for t in Appendix D. The degrees of freedom are $n - 2$ and one rounds down to the closest value in the first column in Appendix D.

So for the nods and liking example if $r = .45$ and $n = 30$, then

$$t(28) = \frac{.45\sqrt{28}}{\sqrt{1 - .45^2}} = 2.666$$

which is statistically significant at the .02 level. It would be concluded that the .45 correlation cannot be explained by sampling error and is significantly greater than zero.

If the correlation is significantly different from zero, the correlation can be either negative or positive. If the researcher wishes to allow the null hypothesis to be false in only one direction (e.g., he or she expects the correlation to be positive), then the p value should be cut in half and the test is called a *one-tailed test.* Most researchers agree that a one-tailed test should not ordinarily be done because if the correlation is very large but in the unpredicted direction, it would still be deemed statistically significant.

Assumptions. A correlation coefficient as a measure of association presumes that the relationship is linear. That is, a change in one unit in the X variable results in the same amount of change in Y regardless of the value of X. As explained in Chapter 7, other types of relationships are not adequately captured by a linear measure of association, and some are even totally missed. The reader is referred to Chapter 7 for a more extensive discussion of the linearity question.

The second assumption is that each pair of (X, Y) scores is independent of all other pairs. Such an assumption presumes that person is the sampling unit. That is, each person provides one and only one pair of scores.

Both variables must be normally distributed. More technically, the two variables have a joint normal distribution. Research has indicated that p values are not considerably altered by the violation of the normality assumption. However, the linearity and the independence assumptions cannot be violated with impunity.

Interpretation. The discussion in Chapter 7 concerning the pitfalls in interpreting correlations is relevant. A significant correlation means that the variables are associated. It in no way indicates the direction of causation. Of course, if the researcher believes that the variables are causally related, a correlation is comforting; however, the correlation does not by itself indicate the direction of causation.

Power. The power of a test is the probability of rejecting the null hypothesis when the null hypothesis is false. Tests of correlation have moderate levels of power. For a given value of the population r, a given n, and a given alpha, power can be determined. Table 16.1 gives the power for the correlation coefficient for a small, medium, and large effect sizes. As discussed in Chapter 7, a small correlation is a value of .1, a medium correlation is a value of .3, and a large correlation is .5. The values given in Table 16.1 are for the .05 level of significance. The n in the table is the total sample size or the number of (X, Y) pairs. The entry in the table is the power multiplied by 100. So if a researcher contemplates doing a study with 40 persons and the correlation is expected to be moderate in size, the probability of rejecting the null hypothesis is .48. This means that for every two times that the study is done, the null hypothesis will be rejected about once.

For a given r, alpha, and level of power desired, the n that is needed for that power can be determined. These sample sizes are given in Table 16.2. As an example, consider the n needed to achieve 70% power for a moderate

TABLE 16.1 **Power Table[a] for Correlation Coefficients, $\alpha = .05$**

	Population Correlation		
n	.1	.3	.5
10	6	13	33
20	7	25	64
40	9	48	92
80	14	78	99
100	17	86	99
200	29	99	99

[a]Taken from Cohen (1977).

NOTE: Each entry in the table is the probability of rejecting the null hypothesis times 100 for a given population correlation and sample size.

TABLE 16.2 **Sample Size Required* for Correlation Coefficients, α = .05**

	Population Correlation		
Power	.1	.3	.5
.25	166	20	8
.50	384	42	15
.60	489	53	18
.70	616	66	23
.80	783	84	28
.90	1046	112	37
.95	1308	139	46
.99	1828	194	63

*Taken from Cohen (1977).
NOTE: Each entry represents the sample size needed to achieve a given level of power for a given population correlation.

effect size of .3. The n would have to be 66 to have a 70% chance of being significant.

Example. Manning and Wright (1983) investigated the degree to which learning pain control strategies would reduce the use of painkilling medication during labor and childbirth. For 52 women who were giving birth, Manning and Wright correlated how much time the women devoted to learning a pain control strategy with their use of painkilling drugs during labor. The correlation was found to be −.22, and hence the women who learned the pain control strategy used fewer drugs. The test of that correlation is

$$t(50) = \frac{-.22\sqrt{50}}{\sqrt{1 - (-.22)^2}} = -1.595$$

which is not significantly different from zero at the .05 level. It is concluded that the population correlation may be zero and that the −.22 correlation is within the limits of sampling error given a sample size of 52.

Test of the Difference Between Two Independent Correlations

Assume that the correlation between nods and liking is computed separately for male and female pairs. For 15 male pairs the correlation is .13, and for 15 female pairs the correlation is .68. At issue is whether the correlation is significantly larger for females than it is for males. When the correlations are computed from two different samples (e.g., males and females) the two correlations are said to be *independent*.

To evaluate whether two correlations are significantly different from one another, one might be tempted to test this hypothesis by first testing whether the correlation is statistically significant for males and then testing whether the correlation is statistically significant for females. The t for males is .473, which is not significant, and 3.344 for females, which is significant at the .01 level of significance. The fact that the correlation is significantly greater than zero for females and is not for males does not necessarily mean that the correlation is significantly larger for females than for males. Statistical logic does not follow ordinary logic. If the number x is equal to zero and the number y is greater than zero, then y must be greater than x. This is simple logic. However, it is not necessarily true that if correlation x is not significantly greater than zero and y is significantly larger than zero, y is a significantly larger correlation than x. One must explicitly test whether the two correlations differ and not rely on the significance tests of the correlations calculated individually. The example illustrates this. Although the .68 correlation is statistically significant and the .13 value is not, it will be seen that the difference between the correlations is not statistically significant.

To test whether these correlations are significantly different from one another, the correlations are transformed. Each correlation must be altered by what is called *Fisher's r to z* transformation. This r to z transformation is defined as

$$\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

Actually the transformation is not usually computed by hand or even by calculator, but rather the value is simply looked up in a table. Table 1 in Appendix F presents the Fisher z transformation values for correlation coefficients. To find the Fisher r to z value in Appendix F, locate r in the left column and then determine its z in the right column. If r is negative, follow the same procedure but give the z value a negative sign. Table 2 in Appendix F contains a table for going from z to r. First, round z to two decimal places, and then locate the appropriate value of z in the left column and top row of the table and find the appropriate value of r.

The r to z transformation has little or no effect on small correlations, but for large correlations the Fisher z value is larger than r. Unlike r, the Fisher z has no upper and lower limit. It is important not to confuse this transformation with the Z or standardizing transformation. Fisher's r to z is for correlations and the Z-score transformation is for a sample of data. Also it should not be confused with the standard normal or Z distribution. Fisher's r to z transformation is applied only to correlations.

The effect of this transformation is to make the sampling distribution of the transformed coefficient nearly normally distributed. When the population correlation is not equal to zero, the distribution of r is somewhat skewed. Fisher's z transformation also makes the variance of correlation coefficient approximately the same regardless of the value of the population correlation. For a given population correlation, the distribution of the Fisher's z values for a sample of size n has virtually a normal distribution with a variance of $1/(n - 3)$. Thus, the standard error of a Fisher z transformed correlation is $1/\sqrt{n - 3}$.

If there are two correlations with sample sizes n_1 and n_2, respectively, they are each transformed into Fisher's z values. These Fisher z transformed values are denoted as z_1 and z_2. Under the null hypothesis that the population correlations are equal, the following has approximately a standard normal distribution.

$$Z = \frac{z_1 - z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$$

Therefore, if the above quantity is greater than or equal to 1.96 or less than or equal to −1.96, the two correlations are significantly different at the .05 level of significance. To determine the p value, the value of Z is located in Appendix C. The p value equals twice the quantity of .5 minus the probability. So if Z is 1.96 the p value is two times .5000 − .4750 which equals .05.

For the .68 and .13 correlations for females and males, the Z is only 1.71, which is not significant at the .05 level. Hence sampling error is a plausible explanation for the difference between the two correlations.

Assumptions. The use of Fisher's z to test the difference between correlations requires that the two correlations be independent. One condition required for independence is that the correlations are computed using two different sets of persons. If the same persons are used to compute both correlations, the correlations are called *correlated correlations*. This topic is discussed later in this chapter.

Interpretation and Power. If the Z is statistically significant, then it is concluded that the population correlations differ in the two groups. If the two are not significantly different, the null hypothesis that the correlations are equal is retained. However, the power of the test that compares the correlations between two samples is quite low. For instance, if $n_1 = n_2 = 50$ and population correlations are .10 and .40, which seems like a large difference, there is only a 35% chance of rejecting the null hypothesis. There must be fairly large sample sizes before having a reasonable chance of rejecting the null hypothesis that the two correlations are unequal.

This low power in showing that correlations differ in the two groups has made it very difficult to show that standardized tests, such as the Scholastic Aptitude Test or SAT, have differential validity across the races. Some have argued that standardized tests are less valid for minorities, particularly blacks. The validity of a standardized test is often measured by a correlation coefficient—for example, the correlation between SAT and college grades. Thus, critics of standardized tests have argued that the correlation is lower for blacks than for whites. However, because of low power, the null hypothesis of no difference is very difficult to show to be false, and so very rarely is the null hypothesis rejected.

Example. Wheeler, Reis, and Nezlek (1983) correlated feelings of loneliness with the extent to which persons felt they had opportunities to disclose or discuss things about themselves with others. The correlations were computed separately for 43 men and 53 women. The correlation for the men was –.57 and for women it was –.21. So, persons who were lonely said they had few opportunities to discuss things about themselves. Using Table 1 of Appendix F, the z value for the –.57 correlation is –.6475, and for the –.21 correlation the z value is –.2132. The test that the coefficients differ is

$$Z = \frac{-.6475 - (-.2132)}{\sqrt{\dfrac{1}{43 - 3} + \dfrac{1}{53 - 3}}} = -2.05$$

which, from Appendix C, has a p value of .0404—that is, two times (.5000 – .4798). Because the p value is less than .05, the difference is judged statistically significant at the .05 level. So, loneliness and the absence of self disclosure correlate significantly more highly among men than women.

More than Two Independent Correlations

Suppose the correlation between socioeconomic status and school achievement is computed for students for four schools. The null hypothesis to be tested is that the correlations do not vary across schools. The correlations are first transformed to Fisher z values. The mean of the z values, $\bar{z}$, is computed weighted by $n - 3$. So for the example of four schools,

$$\bar{z} = \frac{z_1(n_1 - 3) + z_2(n_2 - 3) + z_3(n_3 - 3) + z_4(n_4 - 3)}{n_1 - 3 + n_2 - 3 + n_3 - 3 + n_4 - 3}$$

This is the average of the four correlations weighted by sample size. In general to average correlations, the r's are converted to Fisher z values and are multiplied by the sample size less three, these values are summed, and this total is divided by the number of subjects in all groups less three times the number of correlations. This $\bar{z}$ can be converted back into a correlation by using Table 2 of Appendix F to obtain the average of the correlations.

To test whether the average correlation is significantly different from zero, the average z is divided by

$$\sqrt{\frac{1}{n_1 - 3 + n_2 - 3 + n_3 - 3 + n_4 - 3}}$$

which is approximately distributed as Z, the standard normal distribution, given the null hypothesis of a zero correlation.

A researcher might also wish to test whether correlations, as a group, significantly differ from one another. Here the null hypothesis is not whether the average correlation is zero, but that the population correlations are the same in each group. To do so, the following is computed:

$$(n_1 - 3)(z_1 - \bar{z})^2 + (n_2 - 3)(z_2 - \bar{z})^2 + (n_3 - 3)(z_3 - \bar{z})^2 + (n_4 - 3)(z_4 - \bar{z})^2$$

where $\bar{z}$ is the Fisher's z average of the correlations. In general, to evaluate whether correlations computed in k groups are significantly different from one another, one first averages the Fisher z values, weighting by sample size less three. One then computes the deviation of each Fisher z from this average, squares, multiplies by sample size less three and sums. The resulting quantity is approximately distributed as chi square with $k - 1$ degrees of freedom, k being the number of correlations. As described in Chapter 11, chi square (symbolized by χ^2) is a positively skewed distribution with a lower limit of zero and an upper limit of positive infinity. If the chi square value exceeds the values tabled in Appendix G, it is deemed significant at the appropriate level. One rejects the null hypothesis that the correlation is the same for all groups.

Example. Consider five hypothetical correlations between ability in mathematics and in reading in five different countries. The correlations and their Fisher z values from Table 1 of Appendix F are as follows:

Country	r	n	z
France	.65	55	.7753
England	.53	76	.5901
Mexico	.56	44	.6328
Italy	.44	68	.4722
Canada	.74	39	.9505

The average z is

$$\frac{52(.7753) + 73(.5901) + 41(.6328) + 65(.4722) + 36(.9505)}{52 + 73 + 41 + 65 + 36}$$

which equals .6526. Rounding this z value to two decimal places and using Table 2 of Appendix F, it corresponds to an r of .5717. The test that this correlation is zero is tested by

$$Z = \frac{.6526}{\sqrt{\dfrac{1}{52 + 73 + 41 + 65 + 36}}} = 10.66$$

which is statistically significant at the .001 level. So, across the five countries, the correlation between reading and mathematics ability is significantly different from zero at the .001 level of significance.

The test that the population correlations are all equal to each other is

$$52(.7753 - .6526)^2 + 73(.5901 - .6526)^2 + 41(.6328 - .6526)^2 + 65(.4722 - .6526)^2 + 36(.9505 - .6526)^2 = 6.39$$

A χ^2 with four degrees of freedom of 6.39 is not significant at the .05 level of significance. So, the correlation between mathematics and reading skill does not significantly differ from country to country.

Correlated Correlations

The methods in the previous sections have assumed that when two or more correlations are being compared, different sets of subjects are being compared. Often there is one set of persons or one sample, and two correlations are computed from their data, and these two correlations are compared.

Ordinarily when two or more correlations are compared, the same two variables are correlated. But sometimes correlations involving different variables are compared. For instance, one might wish to compare the correlation of mother's education with child's verbal skill to the correlation of father's education with the child's verbal skill. So, the variables involved in the comparison of correlations need not be the same.

The Fisher z transformation cannot be used for comparing these two correlations because the same persons are used. Consider the correlations between X_1 with X_3 and X_2 with X_3. If the correlation between X_1 and X_2 is 1.00, the correlation between X_1 and X_3 must be the same as the correlation between X_2 and X_3. This is a mathematical necessity. If the correlation between X_1 and X_2 is -1.00, the correlation between X_1 and X_3 must be the same as the correlation between X_2 and X_3 but with the opposite sign. Again this is a mathematical necessity. Thus, the size of the correlation between X_1 and X_2 influences how similar r_{13} and r_{23} are. The statistical test must take into account the degree to which X_1 correlates with X_2. This can be done by using a procedure known as the Williams modification of the Hotelling test.

For this test variable X_3 is correlated with two other variables, X_1 and X_2. There are then three correlations:

r_{12}: correlation between X_1 and X_2

r_{13}: correlation between X_1 and X_3

r_{23}: correlation between X_2 and X_3

The test of whether the population correlation between X_1 and X_3 equals the population correlation between X_2 and X_3 is

$$t(n-3) = \frac{(r_{13} - r_{23})\sqrt{(n-1)(1 + r_{12})}}{\sqrt{2K\dfrac{(n-1)}{(n-3)} + \dfrac{(r_{23} + r_{13})^2}{4}(1 - r_{12})^3}}$$

where

$$K = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}$$

The quantity is distributed as t with $n - 3$ degrees of freedom given the null hypothesis that the population correlation between X_1 and X_3 is equal to the population correlation between X_2 and X_3. If the t is statistically significant, it is concluded that the difference between the two correlations cannot be explained by sampling error.

The test just described has one variable (X_3) that is in both correlations. Consider a test of the difference between two correlation coefficients in one sample where none of the variables are the same. There are now four different variables: X_1, X_2, X_3, and X_4. They give rise to six different correlations:

r_{12}: correlation between X_1 and X_2

r_{34}: correlation between X_3 and X_4

r_{13}: correlation between X_1 and X_3

r_{14}: correlation between X_1 and X_4

r_{23}: correlation between X_2 and X_3

r_{24}: correlation between X_2 and X_4

At issue is the test that the population correlations between X_1 and X_2 and between X_3 and X_4 are equal to each other.

First, the correlations that are being compared are transformed into Fisher's z values: z_{12} and z_{34}. The test is

$$Z = \frac{\sqrt{(n - 3)}(z_{12} - z_{34})}{\sqrt{2 - Q(1 - r^2)^2}}$$

where

$$Q = (r_{13} - r_{23}r)(r_{24} - r_{23}r) + (r_{14} - r_{13}r)(r_{23} - r_{13}r) + (r_{13} - r_{14}r)(r_{24} - r_{14}r) + (r_{14} - r_{24}r)(r_{23} - r_{24}r)$$

and

$$r = \frac{r_{12} + r_{34}}{2}$$

This test, called the *Pearson-Filon test,* is approximately distributed as Z, the standard normal distribution (*not* Fisher's z) under the null hypothesis that the population correlation between X_1 and X_2 is equal to the population correlation between X_3 and X_4. This standard normal approximation is quite good if n

is at least 20. This test was developed by Pearson and Filon (1898) and modified by Steiger (1980).

Power. Ordinarily tests of correlations are somewhat more powerful when computed from a single sample than from multiple samples. Nonetheless, the power of the test of the difference between independent correlations is so low that being more powerful still means the test of correlated correlations has relatively low power. It should be noted that in some special cases, the power in the one-sample case can actually be lower than in the two-sample case.

Example. The illustration is taken from Jacobson's (1977) research on the fear of peers among infants. He studied 23 infants and measured their cognitive development and their attachment to a parent when they are presented with a novel stimulus. Research in developmental psychology has shown that for very young children, intelligence and fear are positively associated. The child needs to have the intelligence to realize that a stimulus may be harmful. But as children mature, it is the intelligent children who are less afraid and feel less need to seek a parent for comfort. The older children realize that the novel stimulus is not harmful.

Two different correlations are to be compared. One correlation is between cognitive ability and fear for young infants and the second is between the same two variables for older infants. Jacobson's study confirms the theory. The correlation between cognitive ability and fear is .416 for infants of ten months, which indicates that the smarter children are more afraid. The correlation is –.413 between cognitive ability at ten months and fear at twelve months, which indicates that the smarter children are now less afraid. The correlation between fear over the two-month period is –.343. Fear at ten months is denoted as X_1, fear at twelve months as X_2, and cognitive ability as X_3. So the correlations are

$$\begin{aligned} r_{12} &= -.343 \\ r_{13} &= .416 \\ r_{23} &= -.413 \end{aligned}$$

These are correlated correlations that share one variable in common and that is the cognitive development measure. Using the Williams modification of the Hotelling test, the formula is

$$t(20) = \frac{(-.413 - .416)\sqrt{22[1 + (-.343)]}}{\sqrt{2K\frac{22}{20} + \frac{(-.413 + .416)^2}{4}[1 - (-.343)]^3}}$$

where

$$K = 1 - (-.343)^2 - .416^2 - (-.413)^2 + 2(-.343)(.416)(-.413)$$

The value of K is .657 and the $t(20)$ is –2.622, which from Appendix D is statistically significant at the .02 level. Thus the correlations are significantly different at the .02 level of significance.

Jacobson also measured cognitive ability at twelve months. So, the correlation between cognitive ability and fear for infants at ten and twelve months can be compared. The correlation between ability and fear is .416 at ten months. At twelve months the correlation becomes –.408. These are two correlated correlations that share no variables in common. The variables are denoted as follows:

X_1: ten-month cognitive ability

X_2: ten-month fear

X_3: twelve-month cognitive ability

X_4: twelve-month fear

The six correlations between the four variables are:

$$\begin{aligned} r_{12} &= .416 \\ r_{34} &= -.408 \\ r_{13} &= .556 \\ r_{14} &= -.413 \\ r_{23} &= -.075 \\ r_{24} &= -.343 \end{aligned}$$

The correlations that are to be compared are .416 and –.408. Their Fisher's z values are .443 and –.433. (To increase accuracy, Appendix F is not used and the z's are directly computed.) The average of r_{12} and r_{34} is

$$.004 = \frac{.416 + (-.408)}{2}$$

The value of Q is

$$\begin{aligned} &[.556 - (-.075)(.004)][-.343 - (-.075)(.004)] \\ &+ [-.413 - (.556)(.004)][-.075 - (.556)(.004)] \\ &+ [.556 - (-.413)(.004)][-.343 - (-.413)(.004)] \\ &+ [-.413 - (-.343)(.004)][-.075 - (-.343)(.004)] = -.319 \end{aligned}$$

Now the Pearson-Filon test gives

$$Z = \frac{\sqrt{20}[.443 - (-.433)]}{\sqrt{2 - (-.319)(1.000)}} = 2.57$$

which from Appendix C has a p value of .0102. Therefore the correlations between fear and cognitive ability are significantly different at ten and twelve months.

Test of Regression Coefficients

In this section tests of regression coefficients are presented and the following cases discussed:

1. a single coefficient equal to zero,
2. two independent coefficients equal to each other, and
3. two correlated coefficients equal to each other.

As will be seen, tests of regression coefficients are more computationally complex than tests of correlation coefficients.

Single Regression Coefficient

The regression equation in which the variable X is the predictor and the variable Y is the criterion is

$$Y = a + b_{YX}X + e$$

As an example, consider the number of packs of cigarettes smoked per day as the predictor variable and life expectancy in years as the criterion variable. Research has shown that the coefficient is about –4.0. That is, for every pack of cigarettes smoked per day, one lives on the average four fewer years.

The test that a regression coefficient b_{YX} is not significantly different from zero is

$$t(n-2) = \frac{b_{YX}\sqrt{SS_X}}{s_{Y \cdot X}}$$

where $s_{Y \cdot X}$ is the standard deviation of the errors (see Chapter 6) and equals

$$s_{Y \cdot X} = \sqrt{\frac{SS_Y - b_{YX}^2 SS_X}{n - 2}}$$

The SS_X and SS_Y are the sum of squares for variables X and Y, respectively. In analysis of variance terms, they are the sum of squares total. They equal

$$SS_X = \sum(X - \bar{X})^2$$

and

$$SS_Y = \sum(Y - \bar{Y})^2$$

Their computational formulas are

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

An alternative and simpler way to test b_{YX} is to convert b_{YX} to r_{XY} by the formula

$$r_{XY} = b_{YX}\left(\frac{s_X}{s_Y}\right)$$

The correlation coefficient can be tested for significance. The resulting t value is the same as would be obtained by a direct test of b_{YX}. This fact can be used for determining the power of the test. One converts b to r and uses Table 16.1 to determine the power and Table 16.2 to determine the n necessary to achieve a given level of power.

If X and Y are reversed by having Y predict X, the test of a regression coefficient in which Y predicts X or b_{XY} is

$$t(n\text{–}2) = \frac{b_{XY}\sqrt{SS_Y}}{s_{X \cdot Y}}$$

where

$$s_{X \cdot Y} = \sqrt{\frac{SS_X - b_{XY}{}^2 SS_Y}{n - 2}}$$

The value of t will be the same regardless whether the test is of b_{XY}, b_{YX}, or r_{XY}.

As an example, assume that b_{YX} is 1.5 and SS_X is 33.5 and SS_Y is 140.2 and n is 131. The value of $s_{Y \cdot X}{}^2$ is

$$\frac{140.2 - 1.5^2(33.5)}{129} = .503$$

The test of the slope is

$$t(129) = \frac{1.5\sqrt{33.5}}{\sqrt{.503}} = 12.241$$

which is statistically significant at the .001 level of significance.

Two Independent Regression Coefficients

In this case there is a pair of regression coefficients computed from two different groups of persons. For instance, there are regression coefficients for both males and females of the effect of cigarette smoking on life expectancy. The coefficient for males is –4.32 and for females the value is –3.93. That is, cigarette smoking reduces life expectancy more for males than for females.

The hypothesis of a difference between two regression coefficients is similar to the hypothesis of an interaction in two-way analysis of variance. Both evaluate whether the effect of one variable changes as a function of another. In two-way analysis of variance the two independent variables are measured at the nominal level of measurement. In the regression case one independent variable is at the interval level of measurement and its linear effect on the dependent variable is presumed to change as a function of a dichotomous independent variable.

In the restricted model there is a single slope, and in the complete model there is a slope for each group. The regression coefficient in each sample is computed. These coefficients are denoted as b_1 and b_2 for the two samples. At issue is whether the difference between the two coefficients can be explained by sampling error.

The test statistic is

$$t(n_1+n_2-4) = \frac{b_1 - b_2}{s_{Y \cdot X}\sqrt{\dfrac{1}{SS_{X_1}} + \dfrac{1}{SS_{X_2}}}}$$

where SS_{X_1} and SS_{X_2} are the sum of squares for X for the first and second groups, respectively, and

$$s_{Y \cdot X} = \sqrt{\frac{SS_{Y_1} + SS_{Y_2} - b_1{}^2 SS_{X_1} - b_2{}^2 SS_{X_2}}{n_1 + n_2 - 4}}$$

or, alternatively,

$$s_{Y \cdot X} = \sqrt{\frac{(n_1 - 2)s_{Y \cdot X_1}{}^2 + (n_2 - 2)s_{Y \cdot X_2}{}^2}{n_1 + n_2 - 4}}$$

The formula for $s_{Y \cdot X}$ is the *pooled error standard deviation*. That is, it is a pooling or averaging of the two error variances, each weighted by its degrees of freeodm. This test seems to involve quite a bit of tedious computation. Actually it involves little more than computing and testing two regression coefficients.

As with the difference between two correlation coefficients, the power of the test of the difference between two regression coefficients is quite low. Even if the slopes are quite different, the sample sizes must be quite large before one has a reasonable chance of detecting that the slopes are indeed different.

If the null hypothesis that the regression coefficients are equal is not rejected, the two regression coefficients can be averaged by the following formula:

$$b_p = \frac{b_1 SS_{X_1} + b_2 SS_{X_2}}{SS_{X_1} + SS_{X_2}}$$

The term b_p is called the *pooled regression coefficient*. The pooled coefficient can be viewed as a weighted average where the weights are the sum of squares of the predictor.

The pooled regression coefficient can be tested for significance using the formula

$$t(n_1+n_2-4) = \frac{b_p\sqrt{(SS_{X_1} + SS_{X_2})}}{s_{Y \cdot X}}$$

where $s_{Y \cdot X}$ is the pooled error standard deviation.

The above formula can be generalized to the case in which there is more than one regression coefficient. The generalization, which is found in advanced texts (Winer, 1971), is similar in its computation to a one-way analysis of variance.

As an example consider a researcher who investigates the effect of attitudes about wearing seat belts on behavior for those who heard a series of communications about the importance of wearing seat belts and a group who did not. The criterion is denoted as B for behavior and the predictor as A for attitude. The results are as follows:

	Communication	
	Heard	*Did Not*
b_{BA}	.58	.23
SS_B	68.1	56.3
SS_A	25.0	19.1
n	60	60

The pooled error standard deviation is

$$\sqrt{\frac{68.1 + 56.3 - .58^2(25.0) - .23^2(19.1)}{60 + 60 - 4}} = .996$$

The test that the slopes differ is

$$t(116) = \frac{.58 - .23}{.996\sqrt{\frac{1}{25.0} + \frac{1}{19.1}}} = 1.156$$

This value is not statistically significant at the .05 level. Therefore, the slopes do not significantly differ.

The pooled slope is

$$\frac{.58(25.0) + .23(19.1)}{25.0 + 19.1} = .428$$

The test that the pooled coefficient equals zero is

$$t(116) = \frac{.428\sqrt{(25.0 + 19.1)}}{.996} = 2.854$$

This value is statistically significant at the .001 level. So, attitude toward seat belts significantly predicts behavior across both communication groups.

Two Correlated Regression Coefficients

This case is identical to the previous case except that the regression coefficients are computed from the same sample. For instance, does the number of siblings better predict fourth-grade vocabulary skill than fifth-grade vocabulary skill? If the same persons were measured at fourth and fifth grades, the regression coefficients are computed from the one sample and are correlated.

To test whether these coefficients are the same, the changes in vocabulary skill are computed by subtracting fourth-grade vocabulary scores from fifth-grade scores. This change score would be the criterion, and number of siblings would be the predictor variable in a regression equation. The test of the coefficient from this regression equation would evaluate the difference between regression coefficients. So if X_1 is used to predict X_2 and X_3, to evaluate the difference between b_{21} and b_{31} the regression coefficient $b_{(3-2)1}$ is computed. As with the paired t test described in Chapter 13, difference scores can be used to test hypotheses with paired data.

Choice of Test

Throughout the entire chapter an obvious question arises. Should the test of association be made using a correlation coefficient or regression coefficient? In the case of a single measure of association, the choice of the significance test does not matter. That is, the t value is the same regardless of whether r or b is computed. However, when two or more measures of association are compared, the result from a test of the regression coefficients differs from the result from a test of the correlations. Which measure is to be preferred?

There are three important factors that can guide the decision. First, if there is a clear causal ordering of the two variables, then the regression coefficient is preferred. Because a regression coefficient assumes a causal ordering (the predictor causes the criterion) and a correlation coefficient does not, a regression coefficient is the coefficient of choice when the variables can be causally ordered.

Second, if the variances of the variables are not the same in both groups, the regression coefficient is preferred. As explained in Chapter 7, correlations are affected by variability. Variables with less variability tend to exhibit lower correlations. To evaluate whether two variances are equal the following statistical tests can be employed. For two independent groups, the ratio of sample variances is computed:

$$\frac{s_1^2}{s_2^2}$$

where s_1^2 is greater than s_2^2. If the variances are equal in the population, s_1^2/s_2^2 has an F distribution with $n_1 - 1$ degrees of freedom on the numerator and $n_2 - 1$ degrees of freedom on the denominator. For this test, the p value is doubled because the F ratio is formed by always putting the larger variance on the numerator. If there is a single sample and the purpose is to test whether the variance of X_1 is different from the variance of X_2, begin by computing $X_1 - X_2$ and $X_1 + X_2$. Then correlate the difference, $X_1 - X_2$, with the sum, $X_1 + X_2$, and test whether it is significantly different from zero. The test of this correlation evaluates whether the two variances are equal.

Third, if the unit of measurement changes from group to group, the correlation coefficient is preferred. Thus, for example, if the groups are French and English children and the variables are vocabulary and intelligence, a correlation should be used because different tests would be used in the different countries. However, if males and females within a country were compared, the regression coefficient would be preferred.

Summary

Tests of a regression coefficient or a correlation coefficient are accomplished by a t test with degrees of freedom of sample size less two. Tests of two or more independent correlations is aided by the *Fisher's r to z* transformation. The Fisher's r to z transformation (not to be confused with the Z or standard normal distribution) makes the distribution of the transformed correlation approximately normal. The Fisher's z transformation can be used to pool correlations computed across different samples as well as test whether the correlations are equal.

Sometimes one seeks to compare two correlations that are computed from the same sample. These correlations are called *correlated correlations*. When correlations are themselves correlated, the tests are computationally complicated but straightforward. When the correlated correlations involve three variables, the Williams modification of the *Hotelling test* is used. When the correlated correlations involve four different variables the *Pearson-Filon* test is used.

A test of a single regression coefficient is identical to the t test of a single correlation coefficient. Also, two regression coefficients from different samples can be tested for equality. If they are equal, they can be pooled and the pooled coefficient can be tested to determine whether it is different from zero.

The decision of whether to test either the correlation or the regression coefficient is aided by considerations of causal ordering, equal variances, and unit of measurement.

Problems

1. According to Pulling et al. (1980) the correlation between age and susceptibility to glare is .742 for 148 subjects. Test whether the popula-

tion correlation between the two variables is different from zero. Interpret the result.

2. Convert the following correlations to Fisher z values:

a. −.13	b. .07	c. .91	d. .73
e. .41	f. −.32	g. −.21	h. .53

3. Convert the following Fisher z values to correlation coefficients:

a. −.86	b. −.43	c. .91	d. .06
e. .19	f. .39	g. −.25	h. −1.03

4. Given $b_{YX} = .31$, $n = 44$, $SS_X = 31.93$, and $SS_Y = 22.41$, test whether the population regression coefficient is significantly different from zero.

5. Evaluate whether the population correlations are equal if $r_1 = .23$, $r_2 = .48$, $n_1 = 212$, and $n_2 = 136$.

6. Given $r_{XY} = .39$ and $n = 84$, test whether the population correlation is significantly greater than zero.

7. Variables 1, 2, and 3 are measured on the same set of persons. Test whether the population correlation between variables 1 and 2 is equal to the population correlation between variables 1 and 3 if $r_{12} = .28$ and $r_{13} = .45$. The correlation between X_2 and X_3 is .63 and $n = 175$.

8. Given

$$r_{12} = .43 \qquad r_{14} = .10$$
$$r_{34} = .14 \qquad r_{23} = .16$$
$$r_{13} = .55 \qquad r_{24} = .49$$

and $n = 145$, test whether the difference between r_{13} and r_{23} is statistically significantly different from zero. Test also whether r_{12} is significantly different from r_{34}.

9. Given the following,

	Males	*Females*
b_{YX}	.23	.44
SS_Y	44.5	38.2

Sample	*Correlation*	*n*
1	.69	137
2	.46	108
3	.56	132
4	.66	115

Average them using the Fisher r to z transformation. Test whether the pooled correlation is significantly different from zero. Also test whether the correlations significantly differ from each other.

11. Evaluate whether a correlation of .43 is statistically significant with a sample size of 68.

12. What is the power of the following tests?

	r	n
a.	.1	10
b.	.5	40
c.	.3	100
d.	.1	200

13. For the following cases, how many subjects would be needed to achieve the desired level of power?

	r	*Power*
a.	.1	.50
b.	.5	.25
c.	.3	.90
d.	.5	.50

14. According to Holahan and Moos (1985), the correlation between seeing oneself as easy-going and feeling that one's family is supportive is .21 for 267 men. Test whether the correlation is statistically significant.

15. A sample consisting of 76 females was tested by Schifter and Ajzen (1985). These women's weight loss correlated .41 with the perceived control in losing weight and .25 with intention to lose weight. The correlation between intention and control is .36. Test whether perceived control correlates significantly higher with weight loss than perceived control correlates with intention.

16. According to Neff (1985), the relationships between education (E) and the reporting of depressive symptoms (D) are:

	Whites	*Blacks*
b_{DE}	−.0236	−.0510
s_D	.47	.59
s_E	2.19	2.20
n	658	171

a. Test each of the regression coefficients for statistical significance.
b. Test whether the coefficients are significantly different from each other.
c. Pool the coefficients and test whether the pooled coefficient is different from zero.

17. Given that $n = 148$ and

$$r_{12} = .67 \quad r_{13} = .40 \quad r_{23} = .26$$
$$r_{34} = .21 \quad r_{14} = .53 \quad r_{24} = .19$$

Test whether the correlation between variables one and two is significantly different from the correlation between variables three and four.

18. For the following correlations from three different groups of persons

r	n
.61	96
.23	39
−.15	76

a. Average the correlations using Fisher z.
b. Test whether the average correlation is significantly different from zero.
c. Test whether the correlations differ from each other.