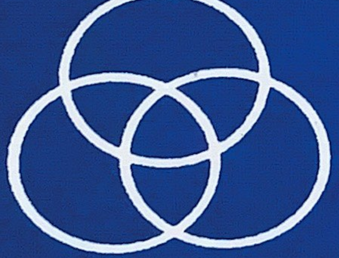
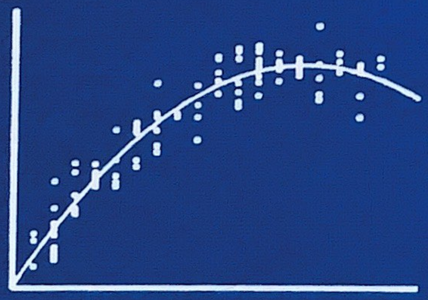
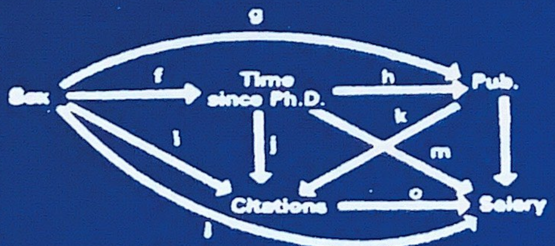
Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences

Third Edition







Jacob Cohen
Patricia Cohen
Stephen G. West
Leona S. Aiken

2

Bivariate Correlation and Regression

One of the most general meanings of the concept of a relationship between a pair of variables is that knowledge with regard to one of the variables carries information about the other. Information about the height of a child in elementary school has implications for the probable age of the child, and information about the occupation of an adult can lead to more accurate guesses about her income level than could be made in the absence of that information.

2.1 TABULAR AND GRAPHIC REPRESENTATIONS OF RELATIONSHIPS

Whenever data have been gathered on two quantitative variables for a set of subjects or other units, the relationship between the variables may be displayed graphically by means of a scatterplot.

For example, suppose we have scores on a vocabulary test and a digit-symbol substitution task for 15 children (see Table 2.1.1). If these data are plotted by representing each child as a point on a graph with vocabulary scores on the horizontal axis and the number of digit symbols on the vertical axis, we would obtain the scatterplot seen in Fig. 2.1.1. The circled dot, for example, represents Child 1, who obtained a score of 5 on the vocabulary test and completed 12 digit-symbol substitutions.

When we inspect this plot, it becomes apparent that the children with higher vocabulary scores tended to complete more digit symbols (d-s) and those low on vocabulary (v) scores were usually low on d-s as well. This can be seen by looking at the average of the d-s scores, M_{d_v}, corresponding to each v score given at the top of the figure. The child receiving the lowest v score, 5, received a d-s score of 12; the children with the next lowest v score, 6, obtained an average d-s score of 14.67, and so onto the highest v scorers, who obtained an average of 19.5 on the d-s test. A parallel tendency for vocabulary scores to increase is observed for increases in d-s scores. The form of this relationship is said to be positive, because high values on one variable tend to go with high values on the other variable and low with low values. It may also be called linear because the tendency for a unit increase in one variable to be accompanied by a constant increase in the other variable is (fairly) constant throughout the scales. That is, if we

TABLE 2.1.1
Illustrative Set of Data on Vocabulary and Digit-Symbol Tests

Child (no.)	Vocabulary	Digit-symbol
1	5	12
2	8	15
3	7	14
4	9	18
5	10	19
6	8	18
7	6	14
8	6	17
9	10	20
10	9	17
11	7	15
12	7	16
13	9	16
14	6	13
15	8	16

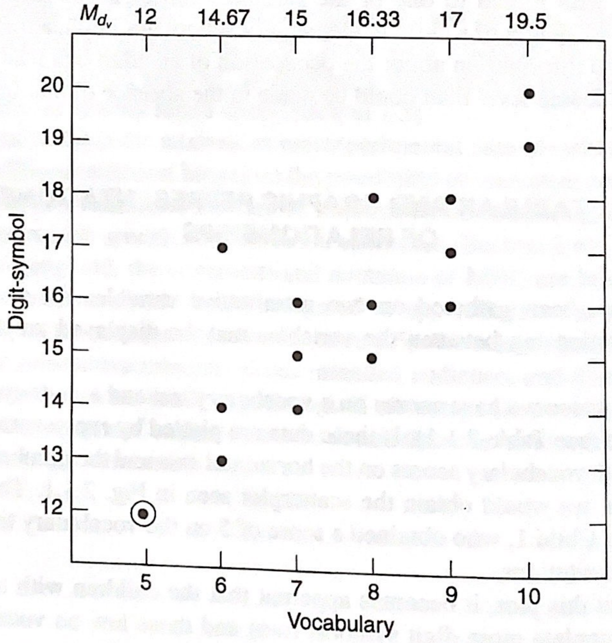


FIGURE 2.1.1 A strong, positive linear relationship.

were to draw the straight line that best fits the average of the d-s values at each v score (from the lower left-hand corner to the upper right-hand corner) we would be describing the trend or shape of the relationship quite well.

Figure 2.1.2 displays a similar scatterplot for age and the number of seconds needed to complete the digit-symbol task. In this case, low scores on age tended to go with high test time in seconds and low test times were more common in older children. This relationship may be

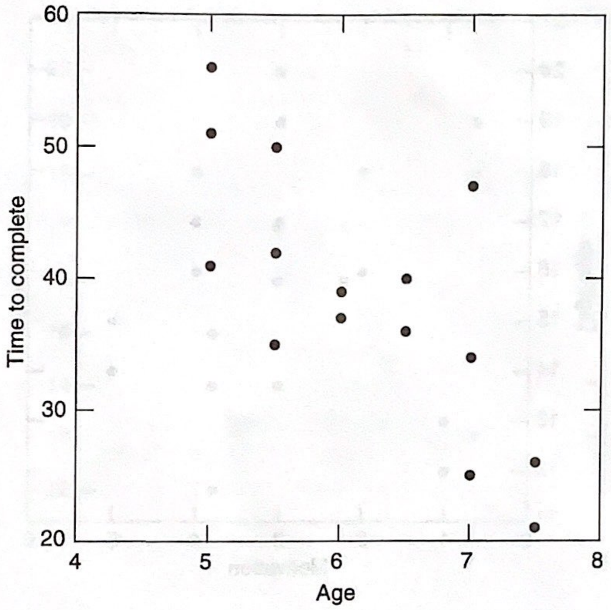


FIGURE 2.1.2 A negative linear relationship.

said to be negative and linear. It should also be clear at this point that whether a relationship between two variables is positive or negative is a direct consequence of the direction in which the two variables have been scored. If, for example, the vocabulary scores from the first example were taken from a 12-item test, and instead of scoring the number correct a count was made of the number wrong, the relationship with d-s scores would be negative. Because such scoring decisions in many cases may be essentially arbitrary, it should be kept in mind that any positive relationship becomes negative when either (but not both) of the variables is reversed, and vice versa. Thus, for example, a negative relationship between age of oldest child and income for a group of 30-year-old mothers implies a positive relationship between age of first becoming a mother and income.[1]

Figure 2.1.3 gives the plot of a measure of motivational level and score on a difficult d-s task. It is apparent that the way motivation was associated with performance score depends on whether the motivational level was at the lower end of its scale or near the upper end. Thus, the relationship between these variables is curvilinear. Finally, Fig. 2.1.4 presents a scatterplot for age and number of substitution errors. This plot demonstrates a general tendency for higher scores on age to go with fewer errors, indicating that there is, in part, a negative linear relationship. However, it also shows that the decrease in errors that goes with a unit increase in age was greater at the lower end of the age scale than it was at the upper end, a finding that indicates that although a straight line provides some kind of fit, clearly it is not optimal.

Thus, scatterplots allow visual inspection of the form of the relationship between two variables. These relationships may be well described by a straight line, indicating a rectilinear (negative or positive) relationship, or they may be better described by a line with one or more curves. Because approximately linear relationships are very common in all sorts of data, we will concentrate on these in the current discussion, and will present methods of analyzing nonlinear relationships in Chapter 6.

[1] Here we follow the convention of naming a variable for the upper end of the scale. Thus, a variable called *income* means that high numbers indicate high income, whereas a variable called *poverty* would mean that high numbers indicate much poverty and therefore low income.

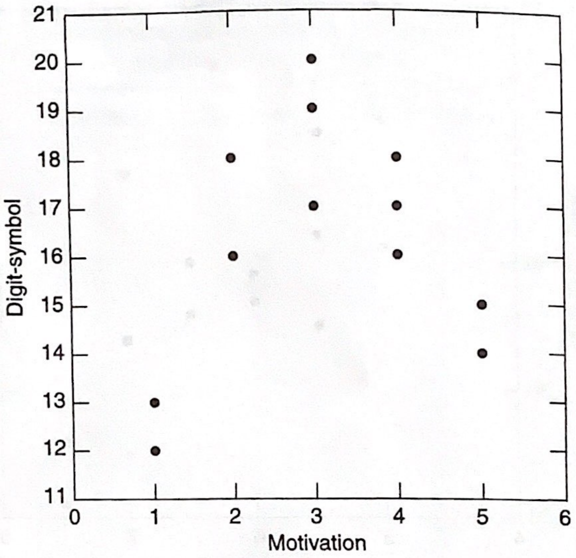


FIGURE 2.1.3 A positive curvilinear relationship.

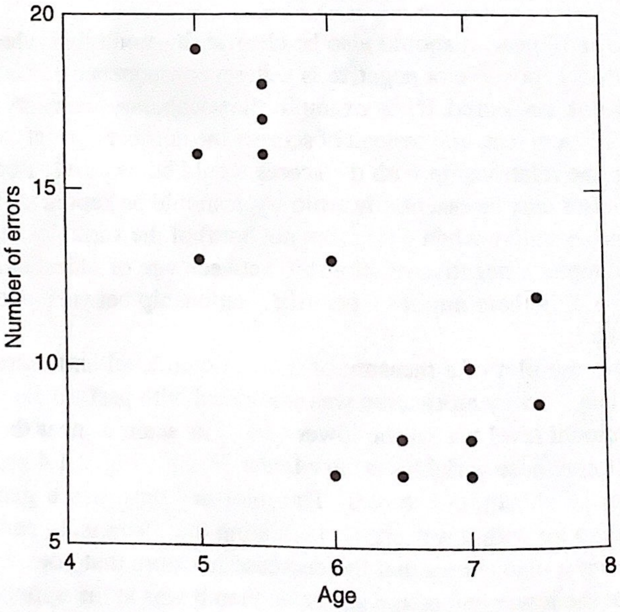


FIGURE 2.1.4 A negative curvilinear relationship.

Now suppose that Fig. 2.1.1 is compared with Fig. 2.1.5. In both cases the relationship between the variables is linear and positive; however, it would appear that vocabulary provided better information with regard to d-s completion than did chronological age. That is, the degree of the relationship with performance seems to be greater for vocabulary than for age because one could make more accurate estimates of d-s scores using information about vocabulary than using age. To compare these two relationships to determine which is greater, we need an index of the degree or strength of the relationship between two variables that will be comparable from one pair of variables to another. Looking at the relationship between v and d-s scores,

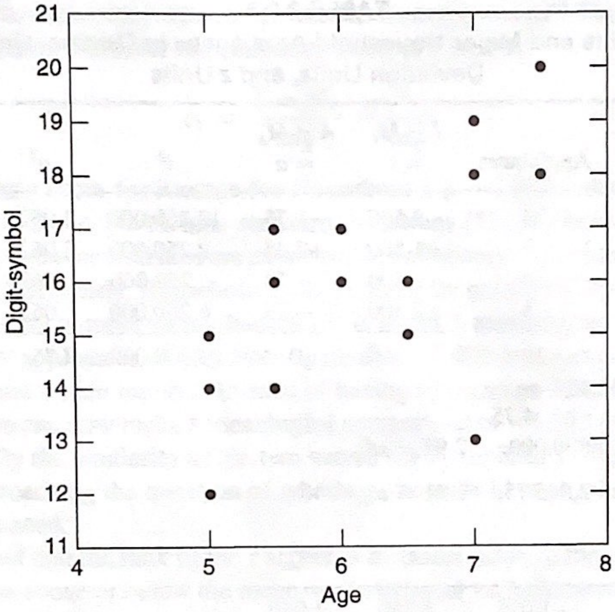


FIGURE 2.1.5 A weak, positive linear relationship.

other questions come to mind: Should this be considered a strong or weak association? On the whole, how great an increase in digit-symbol score is found for a given increase in vocabulary score in this group? If d-s is estimated from v in such a way as to minimize the differences between our estimations and the actual d-s scores, how much error will, nevertheless, be made? If this is a random sample of subjects from a larger population, how much confidence can we have that v and d-s are linearly related in the entire population? These and other questions are answered by correlation and regression methods. In the use and interpretation of these methods the two variables are generally treated as interval scales; that is, constant differences between scale points on each variable are assumed to represent equal "amounts" of the construct being measured. Although for many or even most scales in the behavioral sciences this assumption is not literally true, empirical work (Baker, Hardyck, & Petrinovich, 1966) indicates that small to moderate inequalities in interval size produce little if any distortion in the validity of conclusions based on the analysis. This issue is discussed further in Chapter 6.

2.2 THE INDEX OF LINEAR CORRELATION BETWEEN TWO VARIABLES: THE PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT

2.2.1 Standard Scores: Making Units Comparable

One of the first problems to be solved by an index of the degree of association between two variables is that of measurement unit. Because the two variables are typically expressed in different units, we need some means of converting the scores to comparable measurement units. It can be readily perceived that any index that would change with an arbitrary change in measurement unit—from inches to centimeters or age in months to age in weeks, for example—could hardly be useful as a general description of the strength of the relationship between height and age, one that could be compared with other such indices.

TABLE 2.2.1
Income and Major Household Appliances in Original Units, Deviation Units, and *z* Units

House- hold	Income	Appliances	$I - M_I$ $= i$	$A - M_A$ $= a$	i^2	a^2	Rank I	Rank A
1	24,000	3	−3,500	−1.75	12,250,000	3.0625	1	1
2	29,000	7	+1,500	+2.25	2,250,000	5.0625	3	4
3	27,000	4	−500	−.75	250,000	.5625	2	2
4	30,000	5	+2,500	+.25	6,250,000	.0625	4	3
Sum (Σ)	110,000	19	0	0	21,000,000	8.75		
Mean	27,500	4.75						

$sd_I^2 = \Sigma i^2/(n-1) = 7{,}000{,}000;\quad 2.92 = sd_A^2$

$sd_I = \sqrt{\Sigma i^2/(n-1)} = 2{,}645.75;\quad 1.71 = sd_A$

	$i/_{sd_I} = z_I$	$a/_{sd_A} = z_A$	z_I^2	z_A^2
1	−1.323	−1.025	1.750	1.050
2	+0.567	+1.317	0.321	1.736
3	−0.189	−0.439	0.036	0.193
4	+0.945	+0.146	0.893	0.021
Σ	0	0	3.00	3.00

CH02EX01

To illustrate this problem, suppose information has been gathered on the annual income and the number of major household appliances of four households (Table 2.2.1).[2] In the effort to measure the degree of relationship between income (I) and the number of appliances (A), we will need to cope with the differences in the nature and size of the units in which the two variables are measured. Although Households 1 and 3 are both below the mean on both variables and Households 2 and 4 are above the mean on both (see i and a, scores expressed as deviations from their means, with the means symbolized as M_I and M_A, respectively), we are still at a loss to assess the correspondence between a difference of $3500 from the mean income and a difference of 1.5 appliances from the mean number of appliances. We may attempt to resolve the difference in units by ranking the households on the two variables—1, 3, 2, 4 and 1, 4, 2, 3, respectively—and noting that there seems to be some correspondence between the two ranks. In so doing we have, however, made the differences between Households 1 and 3 ($3000) equal to the difference between Households 2 and 4 ($1000); two ranks in each case.

To make the scores comparable, we clearly need some way of taking the different variability of the two original sets of scores into account. Because the standard deviation (*sd*) is an index of variability of scores, we may measure the discrepancy of each score from its mean (x) relative to the variability of all the scores by dividing by the *sd*:

(2.2.1)

$$sd_X = \sqrt{\frac{\sum x^2}{n-1}},$$

[2] In this example, as in all examples that follow, the number of cases (n) is kept very small in order to facilitate the reader's following of the computations. In almost any serious research, the n must, of course, be very much larger (Section 2.9).

where Σx^2 means "the sum of the squared deviations from the mean."[3] The scores thus created are in standard deviation units and are called *standard* or z scores:

$$z_X = \frac{X - M_X}{sd_X} = \frac{x}{sd_X}. \tag{2.2.2}$$

In Table 2.2.1 the z score for income for Household 1 is -1.323, which indicates that its value ($24,000) falls about 1⅓ income standard deviations ($2646) *below* the income mean ($27,500). Although income statistics are expressed in dollar units, the z score is a pure number; that is, it is unit-free. Similarly, Household 1 has a z score for number of appliances of -1.025, which indicates that its number of appliances (3) is about 1 standard deviation (1.71) below the mean number of appliances (4.75). Note again that -1.025 is not expressed in number of appliances, but is also a pure number. Instead of having to compare $24,000 and 3 appliances for Household 1, we can now make a meaningful comparison of -1.323 (z_I) and -1.025 (z_A), and note incidentally the similarity of the two values for Household 1. This gives us a way of systematically approaching the question of whether a household is as relatively wealthy as it is relatively "applianced."

It should be noted that the rank of the z scores is the same as that of the original scores and that scores that were above or below the mean on the original variable retain this characteristic in their z scores. In addition, we note that the difference between the incomes of Households 2 and 3 ($I_2 - I_3 = \$2000$) is twice as large, and of opposite direction to the difference between Households 2 and 4 ($I_2 - I_4 = -\$1000$). When we look at the z scores for these same households, we find that $z_{I2} - z_{I3} = .567 - (-.189) = .756$ is twice as large and of opposite direction to the difference $z_{I2} - z_{I4} = .567 - .945 = -.378$ (i.e., $.756/-.378 = -2$). Such proportionality of differences or distances between scores,

$$\frac{X_i - X_j}{X_m - X_n} = \frac{z_{X_i} - z_{X_j}}{z_{X_m} - z_{X_n}} \tag{2.2.3}$$

is the essential element in what is meant by retaining the original relationship between the scores. This can be seen more concretely in Fig. 2.2.1, in which we have plotted the pairs of scores. Whether we plot z scores or raw scores, the points in the scatterplot have the same relationship to each other.

The z transformation of scores is one example of a linear transformation. A linear transformation is one in which every score is changed by multiplying or dividing by a constant or adding or subtracting a constant or both. Changes from inches to centimeters, dollars to francs, and Fahrenheit to Celsius degrees are examples of linear transformations. Such transformations will, of course, change the means and *sd*s of the variables upon which they are performed. However, because the *sd* will change by exactly the same factor as the original scores (that is, by the constant by which scores have been multiplied or divided) and because z scores are created by subtracting scores from their mean, all linear transformations of scores will yield the same set of z scores. (If the multiplier is negative, the signs of the z scores will simply be reversed.)

Because the properties of z scores form the foundation necessary for understanding correlation coefficients, they will be briefly reviewed:

[3] As noted earlier, this edition employs the population estimate of *sd* with $n - 1$ in the denominator throughout to conform with computer program output, in contrast to earlier editions, which employed the sample *sd* with n in the denominator in earlier equations in the book and moved to the population estimate when inferences to the population involving standard errors were considered, and thereafter.

Also note that the summation sign, Σ, is used to indicate summation over all n cases here and elsewhere, unless otherwise specified.

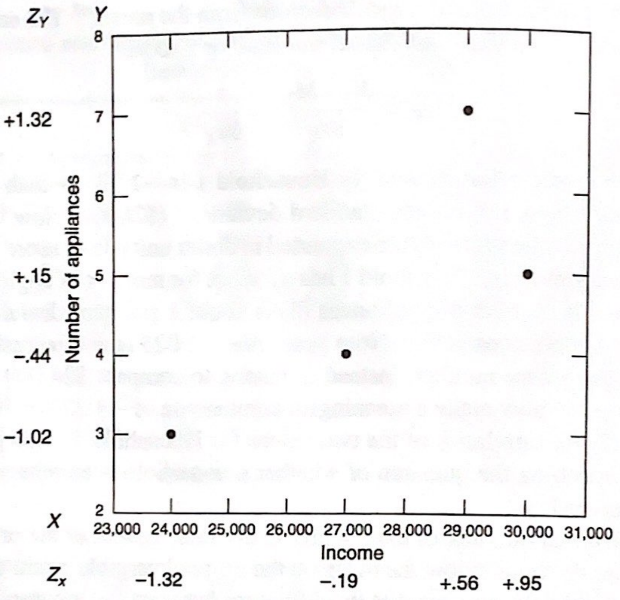


FIGURE 2.2.1 Household income and number of appliances.

1. The sum of a set of z scores (Σz) (and therefore also the mean) equals 0.
2. The variance (sd^2) of the set of z scores equals 1, as does the standard deviation (sd).
3. Neither the shape of the distribution of X nor its absolute correlation with any other variable is affected by transforming it to z (or any other linear transformation).

2.2.2 The Product Moment Correlation as a Function of Differences Between z Scores

We may now define a perfect (positive) relationship between two variables (X and Y) as existing when all z_X and z_Y pairs of scores consist of two exactly equal values. Furthermore, the degree of relationship will be a function of the departure from this "perfect" state, that is, a function of the differences between pairs of z_X and z_Y scores. Because the average difference between paired z_X and z_Y and is necessarily zero (because $M_{z_Y} = M_{z_X} = 0$), the relationship may be indexed by finding the average[4] of the squared discrepancies between z scores, $\Sigma(z_X - z_Y)^2/n$.

For example, suppose that an investigator of academic life obtained the (fictitious) data shown in Table 2.2.2. The subjects were 15 randomly selected members of a large university department, and the data include the time in years that had elapsed since the faculty member's Ph.D. was awarded and the number of publications in professional journals.

Several things should be noted in this table. Deviation scores ($x = X - M_X$ and $y = Y - M_Y$) sum to zero. So do z_X and z_Y. The standard deviations, sd_{z_X} and sd_{z_Y}, are both 1, M_{z_X} and M_{z_Y} are both 0 (all of which are mathematical necessities), and these equalities reflect the equal footing on which we have placed the two variables.

We find that the squared differences (Σsquared) between z scores sums to 9.614, which when divided by the number of paired observations equals .641. How large is this relationship? We have stated that if the two variables were perfectly (positively) related, all z score differences

[4] Because we have employed the sample-based estimate of the population sd, with a divisor of $n - 1$, when z scores have been based on this sd this equation should also use $n - 1$.

TABLE 2.2.2
z Scores, z Score Differences, and z Score Products on Data Example

Case	X Time since Ph.D.	Y No. of publications	$\frac{X_i - M_X}{sd_X} = z_{X_i}$	$\frac{Y_i - M_Y}{sd_Y} = z_{Y_i}$	$z_X - z_Y$	$z_X z_Y$
1	3	18	−1.020	−.140	−.880	.142
2	6	3	−.364	−1.225	.861	.446
3	3	2	−1.020	−1.297	.278	1.322
4	8	17	.073	−.212	.285	−.015
5	9	11	.291	−.646	.938	−.188
6	6	6	−.364	−1.008	.644	.367
7	16	38	1.821	1.307	.514	2.380
8	10	48	.510	2.030	−1.520	1.035
9	2	9	−1.238	−.791	−.447	1.035
10	5	22	−.583	.150	−.732	−.087
11	5	30	−.583	.728	−1.311	−.424
12	6	21	−.364	.077	−.441	−.028
13	7	10	−.146	−.719	.573	.105
14	11	27	.728	.511	.217	.372
15	18	37	2.257	1.235	1.023	2.787
Σ	115	299	0	0	0	
Σ squared	1235	8635	14	14	9.614	
M	7.67	19.93			.641	.613
sd^2	19.55	178.33	1	1		
sd	4.42	13.35	1	1		

would equal zero and necessarily their sum and mean would also be zero. A perfect negative relationship, on the other hand, may be defined as one in which the z scores in each pair are equal in absolute value but opposite in sign. Under the latter circumstances, it is demonstrable that the average of the squared discrepancies times $n/(n-1)$ always equals 4. It can also be proved that under circumstances in which the pairs of z scores are on the average equally likely to be consistent with a negative relationship as with a positive relationship, the average squared difference times $n/(n-1)$ will always equal 2, which is midway between 0 and 4. Under these circumstances, we may say that there is no linear relationship between X and Y.[5]

Although it is clear that this index, ranging from 0 (for a perfect positive linear relationship) through 2 (for no linear relationship) to 4 (for a perfect negative one), does reflect the relationship between the variables in an intuitively meaningful way, it is useful to transform the scale linearly to make its interpretation even more clear. Let us reorient the index so that it runs from −1 for a perfect negative relationship to +1 for a perfect positive relationship. If we divide the sum of the squared discrepancies by $2(n-1)$ and subtract the result from 1, we have

$$r = 1 - \left(\frac{\sum (z_X - z_Y)^2}{2(n-1)} \right), \tag{2.2.4}$$

[5] Note that this equation is slightly different from that in earlier editions. The $n/(n-1)$ term is necessary because the sd used here is the sample estimate of the population sd rather than the sample sd which uses n in the denominator.

which for the data of Table 2.2.2 gives

$$r = r = 1 - \left(\frac{9.614}{28}\right) = .657.$$

r is the product moment correlation coefficient, invented by Karl Pearson in 1895.[6] This coefficient is the standard measure of the linear relationship between two variables and has the following properties:

1. It is a pure number and independent of the units of measurement.

2. Its value varies between zero, when the variables have no linear relationship, and +1.00 or −1.00, when each variable is perfectly estimated by the other. The absolute value thus gives the degree of relationship.

3. Its sign indicates the direction of the relationship. A positive sign indicates a tendency for high values of one variable to occur with high values of the other, and low values to occur with low. A negative sign indicates a tendency for high values of one variable to be associated with low values of the other. Reversing the direction of measurement of one of the variables will produce a coefficient of the same absolute value but of opposite sign. Coefficients of equal value but opposite sign (e.g., +.50 and −.50) thus indicate equally strong linear relationships, but in opposite directions.

2.3 ALTERNATIVE FORMULAS FOR THE PRODUCT MOMENT CORRELATION COEFFICIENT

The formula given in Eq. (2.2.4) for the product moment correlation coefficient as a function of squared differences between paired z scores is only one of a number of mathematically equivalent formulas. Some of the other versions provide additional insight into the nature of r; others facilitate computation. Still other formulas apply to particular kinds of variables, such as variables for which only two values are possible, or variables that consist of rankings.

2.3.1 *r* as the Average Product of *z* Scores

It follows from algebraic manipulation of Eq. (2.2.4) that

$$r_{XY} = \frac{\sum z_X z_Y}{n-1}. \tag{2.3.1}$$

The product moment correlation is therefore seen to be the mean of the products of the paired z scores.[7] In the case of a perfect positive correlation, because $z_X = z_Y$,

$$r_{XY} = \frac{\sum z_X z_Y}{n-1} = \frac{\sum z^2}{n-1} = 1.$$

For the data presented in Table 2.2.1, these products have been computed and $r_{XY} = 9.193/14 = .657$, necessarily as before.

[6] The term *product moment* refers to the fact that the correlation is a function of the product of the *first moments*, of X and Y, respectively. See the next sections.

[7] If we used zs based on the *sample sd* which divides by n, this *average* would also divide by n.

2.3.2 Raw Score Formulas for *r*

Because *z* scores can be readily reconverted to the original units, a formula for the correlation coefficient can be written in raw score terms. There are many mathematically equivalent versions of this formula, of which the following is a convenient one for computation by computer or calculator:

$$r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}. \tag{2.3.2}$$

When the numerator and denominator are divided by n^2, Eq. (2.3.2) becomes an expression of *r* in terms of the means of each variable, of each squared variable, and of the *XY* product:

$$r_{XY} = \frac{M_{XY} - M_X M_Y}{\sqrt{(M_X^2 - M_{X^2})(M_Y^2 - M_{Y^2})}}. \tag{2.3.3}$$

It is useful for hand computation to recognize that the denominator is the product of the variables' standard deviations, thus an alternative equivalent is

$$r_{XY} = \frac{\sum xy/(n-1)}{sd_X sd_Y} \tag{2.3.4}$$

This numerator, based on the product of the *deviation* scores is called the *covariance* and is an index of the tendency for the two variables to *covary* or go together that is expressed in deviations measured in the original units in which *X* and *Y* are measured (e.g., income in *dollars* and *number* of appliances). Thus, we can see that *r* is an expression of the covariance between standardized variables, because if we replace the *deviation* scores with *standardized* scores, Eq. (2.3.4) reduces to Eq. (2.3.1).

It should be noted that *r* inherently is *not* a function of the number of observations and that the $n - 1$ in the various formulas serves only to cancel it out of other terms where it is hidden (for example, in the *sd*). By multiplying Eq. (2.3.4) by $(n - 1)/(n - 1)$ it can be completely canceled out to produce a formula for *r* that does not contain any vestige of *n*:

$$r_{XY} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}. \tag{2.3.5}$$

2.3.3 Point Biserial *r*

When one of the variables to be correlated is a dichotomy (it can take on only two values), the computation of *r* simplifies. There are many dichotomous variables in the behavioral sciences, such as yes or no responses, left- or right-handedness, and the presence or absence of a trait or attribute. For example, although the variable "gender of subject" does not seem to be a quantitative variable, it may be looked upon as the presence or absence of the characteristics of being female (or of being male). As such, we may decide, arbitrarily, to score all females as 1 and all males as 0. Under these circumstances, the *sd* of the gender variable is determined by the proportion of the total *n* in each of the two groups; $sd = \sqrt{PQ}$, where *P* is the proportion in one group and $Q = 1 - P$, the proportion in the other group.[8] Because *r* indicates a relationship between two standardized variables, it does not matter whether we choose 0 and 1 as the two values or any other pair of different values, because any pair will yield the same absolute *z* scores.

[8] Note that here the *sd* is the sample *sd* (divided by *n*) rather than the sample-based estimate of the population σ. As noted earlier, because the *n*s in the equation for *r* cancel, this difference is immaterial here.

TABLE 2.3.1
Correlation Between a Dichotomous and a Scaled Variable

Subject no.	Stimulus condition (X)	Task score (Y)	X_A	X_B	z_Y	z_A	z_B	$z_Y z_A$	$z_Y z_B$
1	NONE	67	0	50	−0.41	−.802	.802	0.329	−0.329
2	NONE	72	0	50	1.63	−.802	.802	−1.307	1.307
3	NONE	70	0	50	0.81	−.802	.802	−0.650	0.650
4	NONE	69	0	50	0.41	−.802	.802	−0.329	0.329
5	STIM	66	1	20	−0.81	1.069	−1.069	−0.866	0.866
6	STIM	64	1	20	−1.63	1.069	−1.069	−1.742	1.742
7	STIM	68	1	20	0	1.069	−1.069	0	0
Sum		476	3	260	0	0	0	−4.565	4.565
Mean		68	.429	37.14	0	0	0		
sd in sample		2.45	.495	14.9		M_Y NONE = 69.5		M_Y STIM = 66.0	

CH02EX02

For example, Table 2.3.1 presents data on the effects of an interfering stimulus on task performance for a group of seven experimental subjects. As can be seen, the absolute value of the correlation remains the same whether we choose (X_A) 0 and 1 as the values to represent the absence or presence of an interfering stimulus or choose (X_B) 50 and 20 as the values to represent the same dichotomy. The sign of r, however, depends on whether the group with the higher mean on the other (Y) variable, in this case the no-stimulus group, has been assigned the higher or lower of the two values. The reader is invited to try other values and observe the constancy of r.

Because the z scores of a dichotomy are a function of the proportion of the total in each of the two groups, the product moment correlation formula simplifies to

$$r_{pb} = \frac{(M_{Y_1} - M_{Y_0})\sqrt{PQ}}{sd_Y}, \tag{2.3.6}$$

where M_{Y_1} and M_{Y_0} are the Y means of the two groups of the dichotomy and the sd_Y is the sample value, which is divided by n rather than $n - 1$. The simplified formula is called the point biserial r to take note of the fact that it involves one variable (X) whose values are all at one of two points and one continuous variable (Y). In the present example,

$$r_{pb} = \frac{(66.0 - 69.5)\sqrt{(.429)(.571)}}{2.45} = -.707. \tag{2.3.7}$$

The point biserial formula for the product moment r displays an interesting and useful property. When the two groups of the dichotomy are of equal size, $p = q = .5$, so $\sqrt{PQ} = .5$. The r_{pb} then equals half the difference between the means of the z scores for Y, and so $2r_{pb}$ equals the difference between the means of the standardized variable.

2.3.4 Phi (ϕ) Coefficient

CH02EX03

When both X and Y are dichotomous, the computation of the product moment correlation is even further simplified. The data may be represented by a fourfold table and the correlation computed directly from the frequencies and marginals. For example, suppose a study investigated the

TABLE 2.3.2
Fourfold Frequencies for Candidate Preference and Homeowning Status

	Candidate U	*Candidate V*	Total
Homeowners	A 19	B 54	$73 = A + B$
Nonhomeowners	C 60	D 52	$112 = C + D$
Total	$79 = A + C$	$106 = B + D$	$185 = n$

difference in preference of homeowners and nonhomeowners for the two candidates in a local election, and the data are as presented in Table 2.3.2. The formula for r here simplifies to the difference between the product of the diagonals of a fourfold table of frequencies divided by the square root of the product of the four marginal sums:

(2.3.8)
$$r_\phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$
$$= \frac{(54)(60) - (19)(52)}{\sqrt{(73)(112)(79)(106)}} = -.272$$

Once again it may be noted that this is a computing alternative to the z score formula, and therefore it does not matter what two values are assigned to the dichotomy because the standard scores, and hence the absolute value of r_ϕ will remain the same. It also follows that unless the division of the group is the same for the two dichotomies ($P_Y = P_X$ or Q_X), their z scores cannot have the same values and r_ϕ cannot equal 1 or -1. A further discussion of this limit is found in Section 2.10.1.

2.3.5 Rank Correlation

CH02EX04

Yet another simplification in the product moment correlation formula occurs when the data being correlated consist of two sets of ranks. Such data indicate only the ordinal position of the subjects on each variable; that is, they are at the ordinal level of measurement. This version of r is called the Spearman rank correlation (r_S). Because the *sd* of a complete set of ranks is a function only of the number of objects being ranked (assuming no ties), some algebraic manipulation yields

(2.3.9)
$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

where d is the difference in the ranks of the pair for an object or individual. In Table 2.3.3 a set of 5 ranks is presented with their deviations and differences. Using one of the general formulas (2.3.4) for r,

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$
$$= \frac{-3}{\sqrt{10}\sqrt{10}} = -.300$$

TABLE 2.3.3
Correlation Between Two Sets of Ranks

I.D.	X	Y	x	x^2	y	y^2	xy	d	d^2
1	4	2	1	1	−1	1	−1	2	4
2	2	1	−1	1	−2	4	2	1	1
3	3	4	0	0	1	1	0	−1	1
4	5	3	2	4	0	0	0	2	4
5	1	5	−2	4	2	4	−4	−4	16
Sum	15	15	0	10	0	10	−3	0	26

TABLE 2.3.4
Product Moment Correlation Coefficients for Special Kinds of Data

Data type	Coefficient
A scaled variable and a dichotomous variable	Point biserial r (r_{pb})
Two dichotomous variables	ϕ or r_ϕ
Two ranked variables	Spearman rank order r (r_S)

The rank order formula (2.3.9) with far less computation yields

$$r_S = 1 - \frac{6(26)}{5(24)}$$
$$= 1 - \frac{156}{120} = -.300,$$

which agrees with the result from Eq. (2.3.4).

We wish to stress the fact that the formulas for r_{pb}, r_ϕ, and r_S are simply computational equivalents of the previously given general formulas for r that result from the mathematical simplicity of dichotomous or rank data (Table 2.3.4). They are of use when computation is done by hand or calculator. They are of no significance when computers are used, because whatever formula for r the computer uses will work when variables are scored 0–1 (or any other two values) or are ranks without ties. It is obviously not worth the trouble to use special programs to produce these special-case versions of r when a formula such as Eq. (2.3.2) will produce them.

2.4 REGRESSION COEFFICIENTS: ESTIMATING *Y* FROM *X*

Thus far we have treated the two variables as if they were of equal status. It is, however, often the case that variables are treated asymmetrically, one being thought of as a dependent variable or criterion and the other as the independent variable or predictor. These labels reflect the reasons why the relationship between two variables may be under investigation. There are two reasons for such investigation; one scientific and one technological. The primary or scientific question looks upon one variable as potentially causally dependent on the other, that is, as in part an effect of or influenced by the other. The second or technological question has for its goal forecasting, as for example, when high school grades are used to predict college grades with no implication that the latter are actually caused by the former. In either case the measure of this effect will, in general, be expressed as the number of units of change in the Y variable per unit change in the X variable.

To return to our academic example of 15 faculty members presented in Table 2.2.2, we wish to obtain an estimate of Y, for which we use the notation $\hat{Y}$, which summarizes the average amount of change in the number of publications for each year since Ph.D. To find this number, we will need some preliminaries. Obviously, if the relationship between publications and years were perfect and positive, we could provide the number of publications corresponding to any given number of years since Ph.D. simply by adjusting for differences in scale of the two variables. Because, when $r_{XY} = 1$, for any individual j, the estimated $\hat{z}_{Y_j}$ simply equals z_{X_j}, then

$$\frac{\hat{Y}_j - M_Y}{sd_Y} = \frac{X_j - M_X}{sd_X},$$

and solving for j's estimated value of Y,

$$\hat{Y}_j = \frac{sd_Y(X_j - M_X)}{sd_X} + M_Y,$$

and because M_X, M_Y, and sd_Y are known, it remains only to specify X_j and then $\hat{Y}_j$ may be computed.

When, however, the relationship is not perfect, we may nevertheless wish to show the estimated $\hat{Y}$ that we would obtain by using the best possible "average" conversion or prediction rule from X in the sense that the computed values will be as close to the actual Y values as is possible with a linear conversion formula. Larger absolute differences between the actual and estimated scores $(Y_j - \hat{Y}_j)$ are indicative of larger errors. The average error $\Sigma(Y - \hat{Y})/N$ will equal zero whenever the overestimation of some scores is balanced by an equal underestimation of other scores. That there be no consistent over- or underestimation is a desirable property, but it may be accomplished by an infinite number of conversion rules. We therefore define *as close as possible* to correspond to the least squares criterion so common in statistical work—we shall choose a conversion rule such that not only are the errors balanced (they sum to zero), but also the sum of the squared discrepancies between the actual Y and estimated $\hat{Y}$ will be minimized, that is, will be as small as the data permit.

It can be proven that the linear conversion rule which is optimal for converting z_X to an estimate of $\hat{z}_Y$ is

(2.4.1) $$\hat{z}_Y = r_{XY} z_X.$$

To convert to raw scores, we substitute for $\hat{z}_Y = (\hat{Y} - M_Y)/sd_Y$ and for $\hat{z}_X = (\hat{X} - M_X)/sd_X$.

Solving for $\hat{Y}$ gives

(2.4.2) $$\hat{Y} = r_{XY} sd_Y \frac{(X - M_X)}{sd_X} + M_Y.$$

It is useful to simplify and separate the elements of this formula in the following way. Let

(2.4.3) $$B_{YX} = r_{XY} \frac{sd_Y}{sd_X},$$

and

(2.4.4) $$B_0 = M_Y - B_{YX} M_X,$$

from which we may write the regression equation for estimating Y from X as

$$\hat{Y} = B_{YX}X + B_0. \tag{2.4.5}$$

Alternatively, we may write this equation in terms of the original Y variable by including an "error" term e, representing the difference between the predicted and observed score for each observation:

$$Y = B_{YX}X + B_0 + e \tag{2.4.6}$$

These equations describe the regression of Y on X. B_{YX} is the regression coefficient for estimating Y from X and represents the rate of change in Y units per unit change in X, the constant by which you multiply each X observation to estimate Y. B_0 is called the regression constant or Y intercept and serves to make appropriate adjustments for differences in size between X and Y units. When the line representing the best linear estimation equation (the Y on X regression equation) is drawn on the scatterplot of the data in the original X and Y units, B_{YX} indicates the slope of the line and B_0 represents the point at which the regression line crosses the Y axis, which is the estimated $\hat{Y}$ when $X = 0$. (Note that B_0 is sometimes represented as A or A_{YX} in publications or computer output.)

For some purposes it is convenient to *center* variables by subtracting the mean value from each score.[9] Following such subtraction the mean value will equal 0. It can be seen by Eq. (2.4.4) that when both the dependent and independent variables have been centered so that both means $= 0$, the $B_0 = 0$. This manipulation also demonstrates that the predicted score on Y for observations at the mean of X must equal the mean of Y. When only the IV is centered, the B_0 will necessarily equal M_Y. For problems in which X does not have a meaningful zero point, centering X may simplify interpretation of the results (Wainer, 2000). The slope B_{YX} is unaffected by centering.

The slope of a regression line is the measure of its steepness, the ratio of how much Y rises (or, when negative, falls) to any given amount of increase along the horizontal X axis. Because the "rise over the run" is a constant for a straight line, our interpretation of it as the number of units of change in Y per unit change in X meets this definition.

Now we can deal with our example of 15 faculty members with a mean of 7.67 and a *sd* of 4.58 years since Ph.D. (Time) and a mean of 19.93 and a *sd* of 13.82 publications (Table 2.2.2). The correlation between time and publications was found to be .657, so

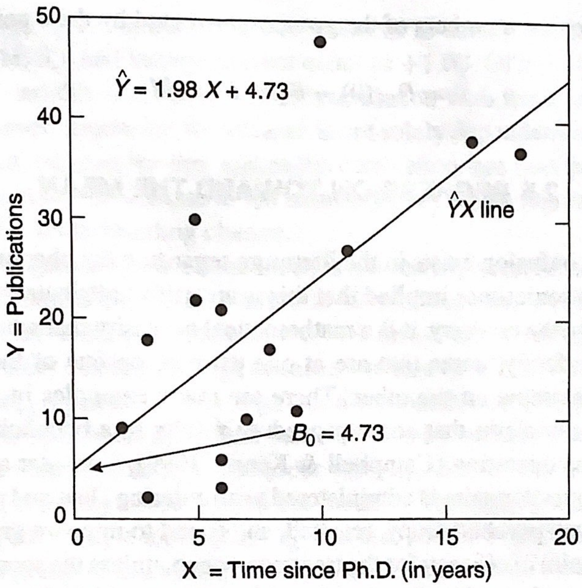
$$B_{YX} = .657(13.82/4.58) = 1.98,$$

$$B_0 = 19.93 - 1.98(7.67) = 4.73.$$

The regression coefficient, B_{YX}, indicates that for each unit of increase in Time (X), we estimate a change of $+1.98$ units (publications) in Y (i.e., about two publications per year), and that using this rule we will minimize our errors (in the least squares sense). The B_0 term gives us a point for starting this estimation—the point for a zero value of X, which is, of course, out of the range for the present set of scores. The equation $\hat{Y}_X = B_{YX}X + B_0$ may be used to determine the predicted value of Y for each value of X, and graphed as the $\hat{Y}X$ line in a scatterplot, as illustrated for these data in Fig. 2.4.1.

We could, of course, estimate X from Y by interchanging X and Y in Eqs. (2.4.3) and (2.2.2). However, the logic of regression analysis dictates that the variables are not of equal status, and estimating an independent or predictor variable from the dependent or criterion variable

[9] As will be seen in Chapters 6, 7, and 9, centering on X can greatly simplify interpretations of equations when relationships are curvilinear or interactive.



X	2	3	4	5	6	7	8	9	10
$\hat{Y}$	8.70	10.68	—	14.64	16.63	18.61	20.59	22.58	24.56
z_X	−1.24	−1.02	—	−0.58	−0.36	−0.15	0.07	0.29	0.51
Mz_Y	−0.84	−0.69	—	−0.40	−0.25	−0.10	0.05	0.20	0.35

X	11	12	13	14	15	16	17	18
$\hat{Y}$	26.54	—	—	—	—	36.46	—	40.42
z_X	0.73	—	—	—	—	1.88	—	2.34
Mz_Y	0.50	—	—	—	—	1.24	—	1.53

FIGURE 2.4.1 Regression of publications on time since Ph.D.

makes no sense. Suffice it to say that were we to do so, the line estimating *X* from *Y* (the *X* on *Y* regression) would not be the same as the line estimating *Y* from *X* (the *Y* on *X* regression). Neither its slope nor its intercept would be the same.

The meaning of the regression coefficient may be seen quite well in the case in which the independent variable is a dichotomy.[10] If we return to the example from Table 2.3.1 where the point biserial $r = -.707$ and calculate

$$B_{YX} = -.707\left(\frac{2.45}{.495}\right) = -3.5,$$

we note that this is exactly the difference between the two group means on *Y*, 66 − 69.5. Calculating the intercept, we get

$$B_0 = 68 - (-3.5)(.428) = 69.5,$$

which is equal to the mean of the group coded 0 (the no-stimulus condition). This must be the case because the best (least squares) estimate of *Y* for each group *is* its own mean, and the

[10] Chapter 8 is devoted to the topic of categorical IVs, for which we provide only a brief introduction here.

regression equation for the members of the group represented by the 0 point of the dichotomy is solved as

$$\hat{Y} = B_{YX}(0) + B_0 = B_0 = M_Y.$$

2.5 REGRESSION TOWARD THE MEAN

A certain amount of confusion exists in the literature regarding the phenomenon of regression toward the mean. It is sometimes implied that this is an artifact attributable to regression as an analytic procedure. On the contrary, it is a mathematical necessity that whenever two variables correlate less than perfectly, cases that are at one extreme on one of the variables will, on the average, be less extreme on the other. There are many examples in the literature where investigators mistakenly claim that some procedure results in a beneficial result when only the regression effect is operating (Campbell & Kenny, 1999). Consider a research project in which a neuroticism questionnaire is administered to an entering class and the students with the poorest scores are given psychotherapy, retested, and found to improve greatly. The "artifact" is the investigator's claim of efficacy for the treatment when, unless the scores remained exactly the same so that the correlation between pretest and posttest was 1.0, they were certain to have scores closer to the mean than previously.

Although the number of cases in a small data set may be too small to show this phenomenon reliably at each data point, examination of the z_X and z_Y values in Fig. 2.4.1 will illustrate the point. The median of time since Ph.D. for the 15 professors is 6 years. If we take the 7 cases above the median, we find that their mean z score is $+.82$, whereas the mean z score for the 5 professors below the median is $-.92$. Now, the mean z score for *number of publications* for the older professors is only .52 and the mean z score for publications for the younger professors is $-.28$. The cases high and low in years since Ph.D. (X) are distinctly less so on publications (Y); that is, they have "regressed" toward the mean. The degree of regression toward the mean in any given case will vary with the way we define *high* and *low*. That is, if we defined high time since Ph.D. as more than 12 years, we would expect an even greater difference between their mean z on time and the mean z on publications. The same principle will hold in the other direction: Those who are extreme on number of publications will be less extreme on years since Ph.D. As can be seen from these or any other bivariate data that are not perfectly linearly related, this is in no sense an artifact, but a necessary corollary of less than perfect correlation.

A further implication of this regression phenomenon is evident when one examines the consequences of selecting extreme cases for study. In the preceding paragraph, we found that those whose Ph.D.s were no more than 5 years old had a mean z score for years since Ph.D. of $-.92$, but a mean z score for number of publication of $-.28$. An investigator might well be tempted to attribute the fact that these new Ph.D.s are so much closer to the mean on number of publications than they are on years since Ph.D. to their motivation to catch up in the well-documented academic rat race. However, recognition that a less than perfect correlation is a necessary and sufficient condition to produce the observed regression toward the mean makes it clear that any specific substantive interpretation is not justified. (There is a delicious irony here: the lower the correlation, the greater the degree of regression toward the mean, and the more to "interpret," spuriously, of course.)

Because regression toward the mean *always* occurs in the presence of an imperfect linear relationship, it is also observed when the variables consist of the same measure taken at two points in time. In this circumstance, unless the correlation is perfect, the extreme cases at Time 1 will be less extreme at Time 2. If the means and *sds* are stable, this inevitably means that low scores improve and high scores deteriorate. Thus, on the average over time, overweight people lose weight, low IQ children become brighter, and rich people become poorer. To ask why these examples of regression to the mean occur is equivalent to asking why correlations between time points for weight, IQ, and income are not equal to $+1.00$. Of course, measurement error is one reason why a variable will show a lower correlation with itself over time, or with any other variables. However, regression to the mean is not solely dependent on measurement error, but on any mechanism whatsoever that makes the correlation less than perfect. Campbell and Kenny (1999) devote an entire volume to the many ways in which regression to the mean can lead to complexities in understanding change.

The necessity for regression toward the mean is not readily accessible to intuition but does respond to a simple demonstration. Expressed in standard scores, the regression equation is simply $\hat{z}_Y = r_{XY}z_X$ (Eq. 2.4.1). Because an r of $+1$ or -1 never occurs in practice, $\hat{z}_Y$ will necessarily be absolutely smaller than z_X, because r is less than 1. Concretely, when $r = .40$, whatever the value of z_X, $\hat{z}_Y$ must be .4 as large (see a comparable set of values below Fig. 2.4.1). Although for a single individual the actual value of z_Y may be larger or smaller than z_X, the expected or average value of the z_Ys that occur with z_X, that is, the value of $\hat{z}_y$, will be .4 of the z_X value (i.e., it is "regressed toward the mean"). The equation holds not only for the expected value of z_Y for a single individual's z_X, but also for the expected value of the mean z_Y for the mean z_X of a group of individuals. Of course, this holds true even when Y is the same variable measured at a later time than X. Unless the correlation over time is perfect, indicating no change, or the population mean and *sd* increase, *on the average*, the fat grow thinner, the dull brighter, the rich poorer, and vice versa.

2.6 THE STANDARD ERROR OF ESTIMATE AND MEASURES OF THE STRENGTH OF ASSOCIATION

In applying the regression equation $\hat{Y} = B_{YX}X + B_0$, we have of course only approximately matched the original Y values. How close is the correspondence between the information provided about Y by X (i.e., $\hat{Y}$), and the actual Y values? Or, to put it differently, to what extent is Y associated with X as opposed to being independent of X? How much do the values of Y, as they vary, coincide with their paired X values, as they vary: equivalently, how big is e in Eq. (2.4.6)?

As we have noted, variability is indexed in statistical work by the *sd* or its square, the variance. Because variances are additive, whereas standard deviations are not, it will be more convenient to work with sd^2_Y. What we wish to do is to partition the variance of Y into a portion associated with X, which will be equal to the variance of the estimated scores, $sd^2_{\hat{Y}}$, and a remainder not associated with X, $sd^2_{Y-\hat{Y}}$, the variance of the discrepancies between the actual and the estimated Y scores (e). (Those readers familiar with ANOVA procedures may find themselves in a familiar framework here.) $sd^2_{\hat{Y}}$ and $sd^2_{Y-\hat{Y}}$ will sum to sd^2_Y, provided that $\hat{Y}$ and $Y - \hat{Y}$ are uncorrelated. Intuitively it seems appropriate that they should be uncorrelated because $\hat{Y}$ is computed from X by the optimal (OLS[11]) rule. Because $\hat{Y} = B_{YX}X +$ (a constant), it is just a linear transformation of X and thus necessarily correlates perfectly with X. Nonzero correlation between $\hat{Y}$ and $Y - \hat{Y}$ would indicate correlation between X (which completely determines $\hat{Y}$) and $Y - \hat{Y}$, and would indicate that our original rule was not optimal. A simple algebraic proof confirms this intuition; therefore:

$$sd^2_Y = sd^2_{\hat{Y}} + sd^2_{Y-\hat{Y}} = sd^2_{\hat{Y}} + sd^2_e, \tag{2.6.1}$$

[11] We introduce the term ordinary least squares (OLS) here, to represent the model that we have described, in which simple weights of predictor variable(s) are used to estimate Y values that collectively minimize the squared discrepancies of the predicted from the observed Ys, so that any other weights would result in larger average discrepancy.

and we have partitioned the variance of Y into a portion determined by X and a residual portion not linearly related to X. If no linear correlation exists between X and Y, the optimal rule has us ignore X because $B_{YX} = 0$, and minimize our errors of estimation by using M_Y as the best guess for every case. Thus we would be choosing that point about which the squared errors are a minimum and $sd^2_{Y-\hat{Y}} = sd^2_Y$. More generally we may see that because (by Eq. 2.4.1) $\hat{z}_Y = r_{XY}z_X$,

$$sd^2_{\hat{z}_Y} = \frac{\sum (r_{XY}z_X)^2}{n-1} = r^2_{XY}\frac{\sum z^2_X}{n-1} = r^2_{XY},$$

and because $sd^2_{z_Y} = 1$, and

(2.6.2)
$$sd^2_{z_Y} = r^2_{XY} + sd^2_{z_Y - \hat{z}_Y},$$

then r^2_{XY} is the proportion of the variance of Y linearly associated with X, and $1 - r^2_{XY}$ is the proportion of the variance of Y *not* linearly associated with X.

It is often helpful to visualize a relationship by representing each variable as a circle.[12] The area enclosed by the circle represents its variance, and because we have standardized each variable to a variance of 1, we will make the two circles of equal size (see Fig. 2.6.1). The degree of linear relationship between the two variables may be represented by the degree of overlap between the circles (the shaded area). Its proportion of either circle's area equals r^2, and $1 - r^2$ equals the area of the nonoverlapping part of either circle. Again, it is useful to note the equality of the variance of the variables once they are standardized: the size of the overlapping and nonoverlapping areas, r^2, and $1 - r^2$, respectively, must be the same for each. If one wishes to think in terms of the variance of the original X and Y, one may define the circles as representing 100% of the variance and the overlap as representing the proportion of each variable's variance associated with the other variable. We can also see that it does not matter in this form of expression whether the correlation is positive or negative because r^2 must be positive.

We will obtain the variance of the residual (nonpredicted) portion when we return to the original units by multiplying by sd^2_Y to obtain

(2.6.3)
$$sd^2_{Y-\hat{Y}} = sd^2_Y(1 - r^2).$$

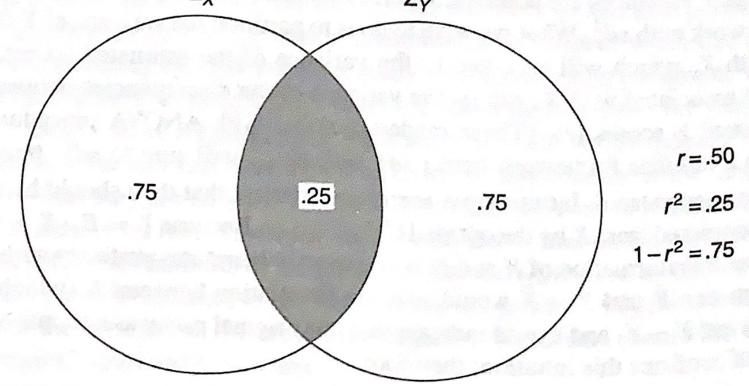


FIGURE 2.6.1 Overlap in variance of correlated variables.

[12] Such figures are called Venn diagrams in mathematical statistics. Here we call them "ballantines," a name taken from a logo for a now-defunct beer company, because we use them illustratively only, and do not wish to imply the mathematical precision that should accompany a Venn diagram.

The standard deviation of the residuals e, that is, of that portion of Y not associated with X is therefore given by

(2.6.4)
$$sd_{Y-\hat{Y}} = sd_Y\sqrt{1-r^2}.$$

For example, when $r = .50$, the proportion of shared variance $= r^2 = .25$, and $.75$ of $sd_{\hat{Y}}^2$ is not linearly related to X. If the portion of Y linearly associated with X is removed by subtracting $B_{YX}X + B_0\ (= \hat{Y})$ from Y, the sd of the residual is reduced compared to the original sd_Y to $sd_{Y-\hat{Y}} = sd_Y\sqrt{.75} = .866\ sd_Y$.

We see that, in this case, although $r = .50$, only 25% of the variance in Y is associated with X, and when the part of Y which is linearly associated with X is removed, the standard deviation of what remains is .866 as large as the original SD_Y.

To make the foregoing more concrete, let us return to our academic example. The regression coefficient B_{YX} was found to be 1.98, the intercept B_0 was 4.73, and r_{XY} was .657. Table 2.6.1 gives the Y, X, and z_Y values and estimated $\hat{Y}$ and $\hat{z}$ from the regression equations (2.4.5) and (2.4.1), which for these values are:

CH02EX05

$$\hat{Y} = 1.98\,X_0 + 4.73 \quad \text{and}$$

$$\hat{z}_Y = .657\,z_X.$$

The $Y - \hat{Y}$ values are the residuals for Y estimated from X or the errors of estimate in the sample. Because $\hat{Y}$ is a linear transformation of X, $r_{Y\hat{Y}}$ must equal $r_{XY}\ (= .657)$. The correlations between $Y - \hat{Y}$ and $\hat{Y}$ must, as we have seen, equal zero. Parallel entries are given for the standardized $\hat{z}_Y$ values where the same relationships hold.

Turning our attention to the variances of the variables, we see that

(2.6.5)
$$\begin{aligned}\frac{sd_{\hat{Y}}^2}{sd_Y^2} &= \frac{sd_{\hat{z}_Y}^2}{1} = r^2\\ &= .657^2 = .4312.\end{aligned}$$

The ratio $sd_{Y-\hat{Y}}/sd_Y = \sqrt{1-r^2} = .754$, which is called the coefficient of alienation, is the part of sd_Y that remains when that part of Y associated with X has been removed. It can also be thought of as the coefficient of *non*correlation, because r is the coefficient of correlation. The standard deviation of the residual scores is given by Eq. (2.6.4) as $sd_{Y-\hat{Y}} = sd_Y\sqrt{1-r^2} = 13.35(.754) = 10.07$, as shown in Table 2.6.1. For the bivariate case, the population *variance error of estimate* or *residual variance* has $df = n - 2$ and is given by

(2.6.6)
$$SE_{Y-\hat{Y}}^2 = \frac{\sum(Y-\hat{Y})^2}{n-2} = \frac{(1-r_{XY}^2)\sum(Y-M_Y)^2}{n-2}.$$

For the two summations, Table 2.6.1 gives in its $\Sigma\sqrt{x^2}$ row, 1521.51 for the $Y - \hat{Y}$ column and 2674.93 for the Y column. Substituting, we get

$$SE_{Y-\hat{Y}}^2 = \frac{1521.51}{15-2} = \frac{(1-.657^2)2674.93}{15-2},$$

and both equations give 117.04. When we take square roots, we obtain the *standard error of estimate*:

(2.6.7)
$$SE_{Y-\hat{Y}} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{(1-r_{XY}^2)\sum(Y-M_Y)^2}{n-2}},$$

which equals 10.82. Here, too, $df = n - 2$.

TABLE 2.6.1
Estimated and Residual Scores for Academic Example

	X Time since Ph.D.	Y No. of publications	$\hat{Y}$	$Y - \hat{Y}$	$\hat{z}_Y$	$z_Y - \hat{z}_Y$	$\hat{Y}_W$	$Y - \hat{Y}_W$	$\hat{Y}_V$	$Y - \hat{Y}_V = e$
	3	18	10.68	7.32	−.67	.53	10.60	7.40	11.07	6.93
	6	3	16.63	−13.63	−.24	−.99	16.60	−13.60	16.77	−13.77
	3	2	10.68	−8.68	−.67	−.63	10.60	−8.60	11.07	−9.07
	8	17	20.59	−3.59	.05	−.26	20.60	−3.60	20.57	−3.57
	9	11	22.58	−11.58	.19	−.84	22.60	−11.60	22.47	−11.47
	6	6	16.63	−10.63	−.24	−.77	16.60	−10.60	16.77	−10.77
	16	38	36.46	1.54	1.20	.11	36.60	1.40	35.77	2.23
	10	48	24.56	23.44	.33	1.70	24.60	23.40	24.37	23.63
	2	9	8.70	0.30	−.81	.02	8.60	.40	9.17	−.17
	5	22	14.65	7.36	−.38	.53	14.60	7.40	14.87	7.13
	5	30	14.65	15.36	−.38	1.11	14.60	15.40	14.87	15.13
	6	21	16.63	4.37	−.24	.32	16.60	4.40	16.77	4.23
	7	10	18.61	−8.61	−.10	−.62	18.60	8.60	18.67	8.67
	11	27	26.54	0.46	.48	.03	26.60	.40	26.27	73
	18	37	40.42	−3.42	1.48	−.25	40.60	3.60	39.57	2.57
M	7.67	19.93	19.93	0	0	0	19.93	0	19.93	0
sd	4.577	13.82	8.77	10.07	.657	.754		10.072	8.40	10.07
sd^2	19.56	178.3	76.98	101.42	.431	.569		101.44	70.60	101.57
$\Sigma\lvert x_i\rvert$				120.29				116.40		120.07
$\Sigma\sqrt{x_i^2}$		2674.93		1521.51						

$$r_{Xz_X} = r_{Yz_Y} = r_{X\hat{Y}} = r_{z_X\hat{z}_Y} = r_{\hat{Y}X} = 1.$$

$$r_{XY} = r_{z_Xz_Y} = r_{Y\hat{Y}} = .657$$

$$r^2_{Y(Y-\hat{Y})} = .5689; r_{(Y-\hat{Y})\hat{Y}} = r_{(Y-\hat{Y})X} = 0.$$

Finally, $\hat{Y}_W$ and $\hat{Y}_V$ in Table 2.6.1 have been computed to demonstrate what happens when any other regression coefficient or weight is used. The values $B_{WX} = 2.0$ and $B_{VX} = 1.9$ were chosen to contrast with $B_{YX} = 1.98$ (the regression constants have been adjusted to keep the estimated values centered on Y). The resulting sd^2 for the sample residuals was larger in each case, 101.44 and 101.57, respectively as compared to 101.42 for the least squares estimate. The reader is invited to try any other value to determine that the squared residuals will in fact always be larger than with 1.98, the computed value of B_{YX}.

Examination of the residuals will reveal another interesting phenomenon. If one determines the *absolute* values of the residuals from the true regression estimates and from the $\hat{Y}_W$, it can be seen that their sum is smaller for both $Y - \hat{Y}_W$ (116.40) and $Y - \hat{Y}_V$ (120.07) than it is for the true regression residuals (120.29). Whenever residuals are not exactly symmetrically distributed about the regression line there exists an absolute residual minimizing weight different from B_{YX}. To reiterate, B_{YX} is the weight that minimizes the squared residuals, not their absolute value. This is a useful reminder that ordinary least squares (OLS), although very useful, is only one way of defining discrepancies from estimation, or error.[13]

[13] Chapter 4 will introduce alternative methods, which are further presented in later chapters.

2.7 SUMMARY OF DEFINITIONS AND INTERPRETATIONS

The product moment r_{XY} is the rate of linear increase in z_Y per unit increase or decrease in z_X (and vice versa) that best fits the data in the sense of minimizing the sum of the squared differences between the estimated and observed scores.

r^2 is the proportion of variance in Y associated with X (and vice versa).

B_{YX} is the regression coefficient of Y on X. Using the original raw units, it is the rate of linear change in Y per unit change in X, again best fitting in the least squares sense.

B_0 is the regression intercept that serves to adjust for differences in means, giving the predicted value of the dependent variable when the independent variable's value is zero.

The coefficient of alienation, $\sqrt{1 - r^2}$, is the proportion of sd_Y remaining when that part of Y associated with X has been subtracted; that is, $sd_{Y-\hat{Y}}/sd_Y$.

The standard error of estimate, $SE_{Y-\hat{Y}}$, is the estimated population standard deviation (σ) of the residuals or errors of estimating Y from X.

2.8 STATISTICAL INFERENCE WITH REGRESSION AND CORRELATION COEFFICIENTS

In most circumstances in which regression and correlation coefficients are determined, the intention of the investigator is to provide valid inferences from the sample data at hand to some larger universe of potential data—from the statistics obtained for a sample to the parameters of the population from which it is drawn. Because random samples from a population cannot be expected to yield sample values that exactly equal the population values, statistical methods have been developed to determine the confidence with which such inferences can be drawn. There are two major methods of statistical inference, estimation using confidence intervals and null hypothesis significance testing. In Section 2.8.1, we consider the formal model assumptions involved. In Section 2.8.2, we describe confidence intervals for B_{YX}, B_0, r_{XY}, for differences between independent sample values of these statistics. In Section 2.8.3, we present the null hypothesis tests for simple regression and correlation statistics. Section 2.8.4 critiques null hypothesis testing and contrasts it with the approach of confidence limits.

2.8.1 Assumptions Underlying Statistical Inference with B_{YX}, B_0, $\hat{Y}_i$, and r_{XY}

It is clear that no assumptions are necessary for the computation of correlation, regression, and other associated coefficients or their interpretation when they are used to describe the available sample data. However, the most useful applications occur when they are statistics calculated on a sample from some population in which we are interested. As in most circumstances in which statistics are used inferentially, the addition of certain assumptions about the characteristics of the population substantially increases the useful inferences that can be drawn. Fortunately, these statistics are *robust*; that is, moderate departure from these assumptions will usually result in little error of inference.

Probably the most generally useful set of assumptions are those that form what has been called the *fixed linear regression model*. This model assumes that the two variables have been distinguished as an independent variable X and a dependent variable Y. Values of X are treated as "fixed" in the analysis of variance sense, that is, as selected by the investigator rather than sampled from some population of X values.[14] Values of Y are assumed to be randomly sampled for each of the selected values of X. The residuals ("errors") from the mean value of Y for each value of X are assumed to be normally distributed in the population, with equal variances across the full range of X values. It should be noted that no assumptions about the shape of the distribution of X and the total distribution of Y per se are necessary, and that, of course, the assumptions are made about the population and not about the sample. This model, extended to multiple regression, is used throughout the book.

2.8.2 Estimation With Confidence Intervals

A *sampling* distribution is a distribution of the values of a sample *statistic* that would occur in repeated random sampling of a given size, n, drawn from what is conceived as an infinite population. Statistical theory makes possible the estimation of the shape and variability of such sampling distributions. We estimate the population value (*parameter*) of the sample statistic we obtained by placing it within a *confidence interval* (*CI*) to provide an estimate of the margin of error (*me*), based on these distributions.

Confidence Interval for B_{YX}

We have seen that B_{YX} is a regression coefficient that gives the slope of the straight line that estimates Y from X. We will see that, depending on the context, it can take on many meanings in data analysis in MRC, including the size of a difference between two means (Section 2.4), the degree of curvature of a regression line (Chapter 6), or the effect of a datum being missing (Chapter 11).

Continuing our academic example, we found in Section 2.4 that for this sample the least squares estimate of $B_{YX} = 1.98$, indicating that for each additional year since Ph.D. we estimate an increase of 1.98 publications, that is, an increase of about two publications. If we were to draw many random samples of that size from the population, we would get *many* values of B_{YX} in the vicinity of $+1.98$. These values constitute the *sampling distribution* of B_{YX} and would be approximately normally distributed. The size of the vicinity is indicated by the standard deviation of this distribution, which is the *standard error* (SE) of B_{YX}:

$$\text{SE}_{B_{YX}} = \frac{sd_Y}{sd_X}\sqrt{\frac{1 - r_{YX}^2}{n-2}} \qquad \text{(2.8.1)}$$

Substituting,

$$\text{SE}_{B_{YX}} = \frac{13.82}{4.58}\sqrt{\frac{1 - .657^2}{15-2}} = .632.$$

Because this is a very small sample, we will need to use the t distribution to determine the multiplier of this SE that will yield estimates of the width of this interval. Like the normal distribution, the t distribution is a symmetrical distribution but with a relatively higher peak in the middle and higher tails. The t model is a family of distributions, each for a different number of *degrees of freedom* (df). As the df increase from 1 toward infinity, the t distribution becomes progressively less peaked and approaches the shape of the normal distribution. Looking in

[14] In the "multilevel" models discussed in Chapters 14 and 15 this assumption is not made for all independent variables.

Appendix Table A, we find that the necessary t at the two-tailed 5% level for 13 df is 2.16. Multiplying .632 by 2.16 gives 1.36, the 95% *margin of error* (*me*). Then, the 95% *confidence limits* (CLs) are given as $1.98 \pm 1.36 = +.62$ as its lower limit and +3.34 as its upper limit. If 1.98 is so much smaller than the population value of B_{YX} that only 2.5% of the possible sample B_{YX} values are smaller still, then the population value is 1.36 publications *above* 1.98, that is, 3.34 (see Fig. 2.8.1), and if 1.98 is so much larger that only 2.5% of the possible sample B_{YX} values are larger still, then the population value is 1.36 publications *below* 1.98, that is, .62 (see Fig. 2.8.2). Thus, the 95% CI is +.62 to +3.34. This CI indicates our 95% certainty that the population value falls between +.62 and +3.34. Note for future reference the fact that the CI for B_{XY} in this sample does *not* include 0 (see Section 2.8.3).

Although the *single most likely* value for the change in number of publications per year since Ph.D. is the sample value 1.98, or about 2 publications per year, we are 95% confident that the true change falls between .62 and 3.34 publications per year since Ph.D. This may be too large an interval to be of much use, as we should have expected when we examined so

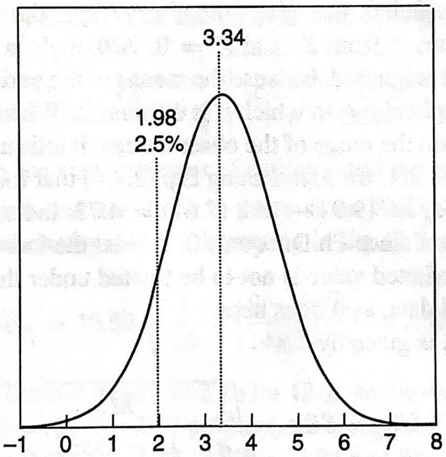


FIGURE 2.8.1 Expected distribution of Bs from samples of 15 subjects when the population $B = 3.34$.

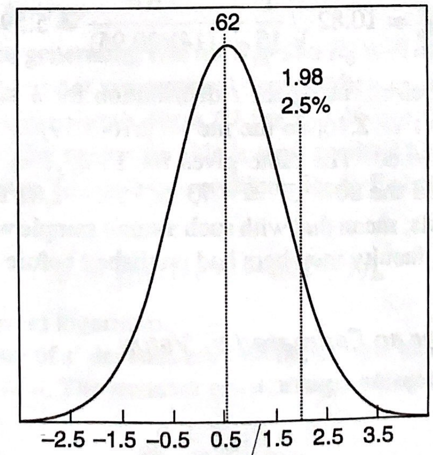


FIGURE 2.8.2 Expected distribution of Bs from samples of 15 subjects when the population $B = 0.62$.

small a sample. Were we to have found the same sample value of 1.98 on a sample as large as 62, the standard error of B_{YX} would go down to .294 (Eq. 2.8.1). When $n = 62$, the *df* for $SE_{B_{YX}}$ is $n - 2 = 60$, so *t* for the 95% $CI = 2.00$ (Appendix Table A). The *me* (margin of error) is now 2.00 (.294) = .588, less than half as large as before, so the 95% *CI* is now 1.98±(.588) = 1.40 to 2.56, from about 1.5 to 2.5 publications per year since Ph.D., distinctly narrower and thus more useful.

Although 95% *CI*s are the most frequently used, other degrees of confidence, greater or smaller, may be preferred. A multiplier of 2.6 will give an approximate 99% *CI*, and 1.3 an approximate 80% interval for all but the smallest samples. Since standard errors are always reported in computer output, and should always be reported in research reports, one can easily approximate a *CI* that includes 68% (about ⅔) of the cases in the sampling distribution by taking the *me* for the sample B_{YX} value to equal its *SE*, so the approximate 68% *CI* is $B_{YX} \pm SE_{B_{YX}}$. The odds are then approximately 2 to 1 that the population B_{YX} value falls between those limits.

Confidence Interval for B_0

B_0 is the regression coefficient that gives the *Y intercept*, the value of $\hat{Y}$ when the $\hat{Y}X$ regression line that estimates *Y* from *X* is at $X = 0$. Although in many behavioral science applications this coefficient is ignored, because the means of the variables are essentially on an arbitrary scale, there are applications in which it is of interest. When zero on the *X* scale has a useful meaning, and is within the range of the observations, it tells us what the expected value of *Y* is for $X = 0$. In Section 2.4, we found using Eq. (2.4.4) that for our running example the intercept $B_0 = M_Y - B_{YX}M_X = 19.93 - 1.98\ (7.67) = 4.73$, indicating a predicted value of 4.73 publications when years since Ph.D. equals 0, that is, the individual has just obtained a Ph.D. Of course, such a predicted value is not to be trusted under the circumstances in which it falls outside the observed data, as it does here.

The standard error of B_0 is given by

(2.8.2)
$$SE_{B_0} = SE_{Y-\hat{Y}}\sqrt{\frac{1}{n} + \frac{M_X^2}{(n-1)\,sd_X^2}}.$$

We found from Eq. (2.6.7) that for this example, $SE_{Y-\hat{Y}} = 10.82$. Substituting from Table 2.6.1 for $n = 15$, $M_X = 7.67$, and $sd^2 = 4.58^2 = 20.95$.

$$\mathrm{SE}_{B_0} = 10.82\sqrt{\frac{1}{15} + \frac{7.67^2}{(14)(20.95)}} = 5.59.$$

We generate *CI*s for B_0 as before, using the *t* distribution for $n - 2 = 13$ df. For the 95% *CI*, Appendix Table A gives $t = 2.16$, so the me $= 2.16(5.59) = 12.07$ and the 95% $CI = 4.73 \pm 12.07 = -7.34$ to 16.80. The table gives for 13 $df, t = 1.35$ for the 80% *CI*, so $me = 1.35(5.59) = 7.55$, and the 80% $CI = 4.73 \pm 7.55$, −2.82 to 12.28. These large *CI*s, with their negative lower limits, mean that with such a small sample we cannot even confidently say whether, on the average, faculty members had published before they got their degrees!

Confidence Interval for an Estimated $\hat{Y}_i$ Value

When we employ the regression equation

(2.4.5)
$$\hat{Y} = B_{YX}X + B_0$$

to estimate a particular $\hat{Y}_i$ from a particular value of X_i, what we find is the *Y* coordinate of the point on the $\hat{Y}X$ regression line for that value of *X*. In the sample data, the *Y* values are scattered above and below the regression line and their distances from the line are the *residuals* or *errors*. The *standard error of estimate* (Eq. 2.6.7) estimates their variability in the population. In our running example estimating number of publications from number of years since Ph.D., we found $SE_{Y-\hat{Y}}$ to equal 10.82. Let's write the regression equation to estimate $\hat{Y}_i$, the number of publications estimated for a specific faculty member with 9 years since Ph.D. The equation for these values was found as $\hat{Y}_i = 1.98X + 4.73$. Substituting $X_i = 9$, we find $\hat{Y}_i = 22.58$.

It is useful to realize that, whatever sampling error was made by using the sample B_{YX} ($= 1.98$) instead of the (unavailable) population regression coefficient, it will have more serious consequences for X values that are more distant from the X mean than for those near it. For the sake of simplicity, let us assume that both X and Y are z scores with means of 0 and standard deviations of 1. Suppose that $B_{YX} = .20$ for our sample, whereas the actual population value is .25. For new cases that come to our attention with $X_i = .1$, we will estimate $\hat{Y}_i$ at .02 when the actual mean value of Y for all $X_i = .1$ is .025, a relatively small error of .005. On the other hand, new values of $X_i = 1.0$ will yield estimated $\hat{Y}_i$ values of .20 when the actual mean value of Y for all $X_i = 1$ is .25, the error (.05) being 10 times as large.

When a newly observed X_i is to be used to estimate $\hat{Y}_i$ we may determine the standard error and thus confidence limits for this $\hat{Y}_i$. The standard error of $\hat{Y}_i$ is given by

(2.8.3)
$$SE_{\hat{Y}_i} = SE_{Y-Y_i}\sqrt{\frac{1}{n} + \frac{(X_i - M_X)^2}{(n-1)sd_X^2}},$$

where SE_{Y-Y_i} (Eq. 2.6.7) is the standard error of estimate and is based on $n - 2df$. We found from the regression equation that for $X_i = 9$ years since Ph.D., we estimate $\hat{Y}_i = 22.58$ publications. We find its standard error by substituting in Eq. (2.8.3),

$$SE_{\hat{Y}_i} = 10.82\sqrt{\frac{1}{15} + \frac{(9 - 7.67)^2}{(14)(20.95)}} = 2.92$$

For the 95% *CI*, Appendix Table A gives $t = 2.16$ for 13 *df*, so the *me* is 2.16 (2.92) = 6.30 and the 95% $CI = 22.58 \pm 6.30 = 16.3$ to 28.9 publications (rounding). For the 80% *CI*, the table gives $t = 1.35$ for 13 *df*, so the $me = 1.35$ (2.92) = 3.94 and the *CI* is $22.58 \pm 3.94 = 18.6$ to 26.5 publications (rounding). These *CI*s are uselessly large because of the large SE_{Y_0}, due mostly in turn to the smallness of the sample.

Confidence Interval for r_{XY}

The approach we used in generating *CI*s for B_{YX} and B_0 will not work for r_{XY} because the sampling distribution for r_{XY} is not symmetrical except when ρ_{YX} (the population r_{XY}) equals 0. That is, the lower and upper limits for a *CI* for r_{XY} do not fall at equal distances from the obtained sample value. The reason for this is that, unlike $SE_{B_{YX}}$, the SE_r varies with ρ_{YX}, which is, of course, unknown. To solve this problem, R. A. Fisher developed the z prime (z') transformation of r:

(2.8.4)
$$z' = \tfrac{1}{2}[\ln(1 + r) - \ln(1 - r)],$$

where ln is the natural (base e) logarithm.

The sampling distribution of z' depends only on the sample size and is nearly normal even for relatively small values of n. The standard error of a sample z' is given by

(2.8.5)
$$SE_{z'} = \frac{1}{\sqrt{n-3}}$$

Appendix Table B gives the r to z' transformation directly, with no need for computation.

To find the *CI* for a sample r, transform the r to z' and, using the $SE_{z'}$ and the appropriate multiplier for the size of the *CI* desired, find the *me* and then the lower and upper limits of the *CI* for z'. Then transform them back to r. For our academic example, we found the r between years since Ph.D. and number of publications to be .657. In Appendix Table B we find the z' transformation to be approximately $z' = .79$. With $n = 15$, we find from (2.8.4) that

$$SE_{z'} = \frac{1}{\sqrt{15-3}} = .289.$$

Then, using the multiplier 1.96 from the *normal distribution* for the 95% limits (Appendix Table C), we find $1.96(.289) = .57$ as the *me* for z', so $.79 \pm .57$ gives the 95% limits for z' as .22 and 1.36. But what we want are the 95% limits for r, so using Appendix Table B we transform these z' values back to r and obtain $r = .22$ (from .22) and .88 (from 1.36). Thus, we can expect with 95% confidence that the population r is included in the approximate *CI* .22 to .88. Note that these limits are not symmetrical about the sample r of .657.

The 95% *CI* for r in this example, .22 to .88, is very wide, as are all the *CI*s for this small sample of $n = 15$.[15] The odds of inclusion here are 95 : 5 (that is, 19 to 1). For narrower and thus less definitive limits, the 80% *CI* gives 80 : 20 (4 to 1) odds of inclusion. To find it, we proceed as before, using the normal curve multiplier for an 80% *CI* of 1.28 (Appendix Table C). We first find the confidence limits for z' by subtracting and adding the $me = 1.28\,(.29) = .38$ to the sample z' of .79, obtaining .41 and 1.17. From Appendix Table B we convert z' to r to find the approximate 80% *CI* for r to be .39 (from .41) to .82 (from 1.17). This is yet another object lesson in precision (or, rather, its lack) with small samples. For most purposes, limits as wide as this would not be of much use.

Confidence Interval for the Difference Between Regression Coefficients: $B_{XY_V} - B_{XY_W}$

Given the many uses to which regression coefficients are put, the size of the difference between a pair of B_{YX} sample values coming from different groups is often a matter of research interest. The *SE* of the difference between two independent B_{YX} values is a function of their standard errors, whose formula we repeat here for convenience:

$$SE_{B_{YX}} = \frac{sd_Y}{sd_X}\sqrt{\frac{1-r_{YX}^2}{n-2}}. \tag{2.8.1}$$

Assume that the sample in Section 2.4 in which we found the regression coefficient describing the relationship between time since Ph.D. and number of publications, 1.98, was drawn from University V and numbered 62 cases. Substituting the sample values found in Section 2.4 in Eq. (2.8.1), we find its standard error to be .294. Now assume that in a random sample of 143 cases from University W, we find $sd_{Y_W} = 13.64$, $sd_{X_W} = 3.45$, and $r_W = .430$. Substituting these values in Eq. (2.4.3), we find $B_W = 1.70$, and in Eq. (2.8.1) we find $SE_{B_W} = .301$. Now, the difference between B_V and B_W is $1.98 - 1.70 = .28$. The standard error of the difference between the two coefficients is

$$SE_{B_V - B_W} = \sqrt{(SE_{B_V})^2 + (SE_{B_W})^2} \tag{2.8.6}$$

Substituting, we find

$$SE_{B_V - B_W} = \sqrt{(.294)^2 + (.301)^2} = .42$$

[15] Indeed, it would be foolish to place any serious faith in the adequacy of the estimate based on such a small sample, which is employed here only for illustrative purposes.

Using the multiplier 2 (a reasonable approximation of 1.96) for the 95% *CI*, we find the *me* for the difference between the *B* values, 2 (.42) = .84, and obtain the approximate 95% *CI* for $B_V - B_W$ as .28 ± .84 = −.56 to +1.12. This means that the confidence limits go from University V's slope being .56 (about ½ of a publication) *smaller* per year since Ph.D. to being 1.12 (about 1) publication larger. Take particular note of the fact that the 95% *CI* includes 0. Thus, we cannot conclude that there is *any* difference between the universities in the number of publications change per year since Ph.D. at this level of confidence.

Equation (2.8.6) gives the standard error of the difference between regression coefficients coming from different populations as the square root of the sum of their squared standard errors. This property is not unique to regression coefficients but holds for *any* statistic—means, standard deviations, and, as we see in the next section, correlation coefficients as well.

Confidence Interval for $r_{XY_V} - r_{XY_W}$

We cannot approach setting confidence limits for differences between *r*s using the *z′* transformation because of the nonlinear relationship between them—equal distances along the *r* scale do not yield equal distances along the *z′* scale (which can be seen in Appendix Table B).

Recent work by Olkin and Finn (1995) has provided relatively simple means for setting confidence intervals for various functions of correlation coefficients. For *large* samples, the difference between r_{YX} in two independent samples, V and W, is normally distributed and is given approximately by

+ or − ?

(2.8.7)
$$SE_{r_V - r_W} = \sqrt{\frac{1 - r_V^2}{n_V} + \frac{1 + r_W^2}{n_W}}.$$

Returning to the example in which we compared the regression coefficients for our running problem, we can estimate confidence intervals for the difference between the correlations of .657 for University V (n_V = 62) and .430 for University W (n_W = 143). Substituting in Eq. (2.8.7),

$$SE_{r_V - r_W} = \sqrt{\frac{1 - .657^2}{62} + \frac{1 - .430^2}{143}} = .122$$

.227

The difference between the *r*s is .657 − .430 = .277. Assuming normality, the 95% *CI* uses 1.96 as the multiplier, so the 95% *me* is 1.96 (.122) = .239. Then the approximate 95% *CI* is .277 ± .239 = +.04 to +.52. We interpret this to mean that we can be 95% confident that the ρ_{YX} of time since Ph.D with number of publications for University V is .04 to .52 *larger* than that for University W. Note here that the confidence interval of the difference between the *r*s of the two universities does not include 0, but the *CI* of the difference between their regression coefficients does. This demonstrates that correlation and regression coefficients are different measures of the degree of linear relationship between two variables. Later, we will argue that regression coefficients are often more stable across populations, in contrast to *r*s that reflect population differences in variability of *X*. In the preceding example, we saw sd_X = 4.58 in the original University V and sd_X = 3.45 in the comparison University W. The smaller *r* in University W is apparently attributable to their faculty's constricted range of years since Ph.D.

2.8.3 Null Hypothesis Significance Tests (NHSTs)

In its most general meaning, a null hypothesis (H_0) is a hypothesis that a population effect size (ES) or other parameter has some value specified by the investigator. The term "null" arises from R. A. Fisher's statistical strategy of formulating a proposition that the research data may be able to *nullify* or reject. By far, the most popular null hypothesis that is tested is the one that posits that a population effect size, such as a correlation coefficient or a difference between means, is *zero*, and the adjective "null" takes on the additional meaning of no relationship or no effect. We prefer to use the term "nil" hypothesis to characterize such propositions for reasons that will become clear later (J. Cohen, 1994).

The Nil Hypothesis Test for B_{YX}

In our running example of the 15 faculty members, we found that the regression coefficient for the number of publications on number of years since Ph.D. was 1.98 ($= B_{YX}$), which means that, on the average in this sample, each additional year since Ph.D. was associated with about two publications. The standard error of the coefficient ($SE_{B_{YX}}$) from Eq. (2.8.1) was .632. Let's perform a *t* test of the nil hypothesis that in the population, each additional year since Ph.D. is associated on the average with *no* additional publications, that is, that there is no linear relationship between years since Ph.D. and publications. We will perform this test at the $p < .05$ ($= \alpha$) significance level. The general form of the *t* test is

$$t = \frac{\text{sample value} - \text{null-hypothetical value}}{\text{standard error}} \quad (2.8.8)$$

which, for regression coefficients, is

$$t = \frac{B_{YX} - H_0}{SE_{B_{YX}}}. \quad (2.8.9)$$

Substituting,

$$t = \frac{1.98 - 0}{.632} = 3.14,$$

which, for $df = n - 2 = 13$ readily meets the $\alpha = .05$ significance criterion of $t = 2.16$ (Appendix Table A). We accordingly reject H_0 and conclude that there is a greater than zero relationship between years since Ph.D. and number of publications in the population. Note, however, that neither the size nor the statistical significance of the *t* value provides information about the *magnitude* of the relationship. Recall, however, that when we first encountered the $SE_{B_{YX}}$ at the beginning of Section 2.8.2, we found the 95% *CI* for B_{YX} to be +.62 to +3.34, which *does* provide a magnitude estimate. Moreover, note that the 95% *CI does not include 0.* After we have determined a *CI* for B_{YX}, a *t* test of the nil hypothesis for B_{YX} is unnecessary—once we have a *CI* that does not include 0, we know that the nil hypothesis can be rejected at that significance level (here, $\alpha = .05$). However, if the only relevant information about a population difference is whether it has some specified value, or whether it exists at all, and there are circumstances when that is the case, then *CI*s are unnecessary and a null hypothesis test is in order.

For example, assume that we wish to test the proposition as a non-nil null hypothesis that the population regression coefficient is 2.5 publications per year since Ph.D.: H_0: population $B_{YX} = 2.5$. We can proceed as before with Eq. (2.8.9) to find $t = (1.98 - 2.5)/.632 = .82$, which is not significant at $\alpha = .05$, and we can conclude that our results are consistent with the possibility that the population value is 2.5. But since the 95% *CI* (+.62 to +3.34) contains the null-hypothetical value of 2.5, we can draw the same conclusion. However, by obtaining the 95% *CI* we have the *range* of B_{YX} values for which the H_0 cannot be rejected at $\alpha = .05$. Not only 2.5 or 0, but *any value* in that range cannot be rejected as a H_0. Therefore, one may think of a *CI* as a range of values within which the H_0 *cannot* be rejected and outside of which H_0 *can* be rejected on the basis of this estimate. The *CI* yields more information than the NHST.

The Null Hypothesis Test for B_0

In the previous section, we found the Y intercept for our running example $B_0 = 4.73$ and, using its standard error (Eq. 2.8.2), found $SE_{B_0} = 5.59$. We can perform a t test for 13 df of the H_0 that the population intercept equals 0 in the usual fashion. Using Eq. (2.8.7) for B_0 and substituting in Eq. (2.8.7), we find

$$t = \frac{4.73 - 0}{5.59} = .85,$$

which fails to meet conventional significance criteria. (In Section 2.8.2 we found 95% and 80% *CI*s, both of which included 0.)

The Null Hypothesis Test for r_{XY}

When ρ_{XY} (the population r_{XY}) $= 0$, the use of the Fisher z' transformation is unnecessary. The t test of the *nil* hypothesis for r_{XY}, H_0: $\rho_{XY} = 0$, is

$$\text{(2.8.10)} \qquad t = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1 - r_{XY}^2}} \quad \text{with} \quad df = n - 2.$$

Returning to our running example, the r_{XY} between years since Ph.D. and publications for the sample of 15 faculty members was .657. Substituting,

$$t = \frac{.657\sqrt{15-2}}{\sqrt{1 - .657^2}} = 3.14.$$

The $\alpha = .05$ significance criterion for t with 13 df is 2.16, readily exceeded by 3.14. We conclude that $\rho_{XY} \neq 0$. (The 95% *CI* was found via the Fisher z' transformation in the previous section to be .22 to .88.)

The Null Hypothesis Test for the Difference Between Two Correlations with Y: $r_{XY_V} - r_{XY_W}$

In Section 2.8.2 we presented a method for setting approximate confidence intervals for differences between independent rs suitable for large samples. For an approximate nil hypothesis test, suitable for samples of any size, we again resort to the Fisher z' transformation. The relevant data for the two universities are

University	N	r_{XY}	z'_{XY}
V	62	.657	.79
W	143	.430	.46

To test the H_0 that the difference between the population correlations: $\rho_V - \rho_W = 0$, we test the equivalent $H_0 : z'_V - z'_W = 0$ by computing the normal curve deviate

$$\text{(2.8.11)} \qquad z = \frac{z'_V - z'_W}{\sqrt{1/(n_V - 3) + 1/(n_W - 3)}}.$$

Substituting,

$$z = \frac{.79 - .46}{\sqrt{1/(62-3) + 1/(143-3)}} = 2.13,$$

which exceeds 1.96, the two-tailed $\alpha = .05$ significance criterion for the normal distribution (see Appendix Table C), and we can conclude that University V's ρ_{XY} is probably larger than University W's. The reason that we can test for z's and conclude about ρs is that there is a one-to-one correspondence between z' and ρ so that when the z's are not equal, the ρs are necessarily also not equal. (The 95% *CI* for the difference between the rs was previously found to be +.04 to +.52.)

2.8.4 Confidence Limits and Null Hypothesis Significance Testing

For more than half a century, NHST has dominated statistical inference in its application in the social, biological, and medical sciences, and for just as long, it has been subject to severe criticism by methodologists including Berkson (1946), Yates (1951), Rozeboom (1960), Meehl (1967), Lykken (1968), and Tukey (1969), among others. More recently, many methodologists, including J. Cohen (1990, 1994) and a committee of the American Psychological Association (Wilkinson of the APA Task Force on Statistical Inference, 1999), among others, have inveighed against the excessive use and abuse of NHST.

We have seen repeatedly that when confidence intervals on statistics or effect sizes are available, they include the information provided by null hypothesis tests. However, there may be a useful role for NHST in cases where the direction of systematic differences is of much more interest than their magnitude and the information provided by confidence intervals may simply be distracting (Harlow, Mulaik, & Steiger, 1997). In addition, as we will see in subsequent chapters, significance tests are useful guides to the decision as to whether certain variables are or are not needed for the explanation of Y. Abelson (1995) notes the usefulness of NHST in making categorical claims that add to the background substantive scientific lore in a field under study.

2.9 PRECISION AND POWER

For research results to be useful, they must be accurate or, at least, their degree of accuracy must be determinable. In the preceding material, we have seen how to estimate regression parameters and test null hypothesis after the sample data have been collected. However, we can plan to determine the degree of precision of the estimation of parameters or of the probability of null hypothesis rejection that we shall be able to achieve.

2.9.1 Precision of Estimation

The *point estimate* of a population parameter such as a population B or ρ is the value of the statistic (B, r) in the sample. The margin of error in estimation is the product of the standard error and its multiplier for the degree of inclusion (95%, 80%) of the confidence interval. The standard error is a function of the sample size, n. We show how to estimate $n*$, the sample size necessary to achieve the desired degree of precision of the statistics covered in Section 2.8.2.

We begin by drawing a trial sample of the data for whose statistics we wish to determine *CI*s. The sample of $n = 15$ cases we have been working with is much too small to use as a trial sample, so let's assume that it had 50 rather than 15 cases so that we can use the same statistics as before: $M_X = 7.67$, $sd_X = 4.58$, $sd_Y = 13.82$, $r_{XY} = .657$, $B_{YX} = 1.98$, $B_0 = 4.73$, and $SE_{Y-\hat{Y}} = 10.82$.

We use the approximate multipliers (t, z) of the standard errors to determine the inclusion of the confidence limits: 99%, 2.6; 95%, 2; 80%, 1.3; and 68%, 1. The standard errors for the regression/correlation statistics of our $n = 50$ sample are as follows:

Estimated B_{YX}

Eq. (2.8.1) $$SE_{B_{YX}} = \frac{13.82}{4.58}\sqrt{\frac{1-.657^2}{50-2}} = .329$$

Estimated intercept

Eq. (2.8.2) $$SE_{B_0} = 10.82\sqrt{\frac{1}{50}+\frac{7.67^2}{(50-1)(20.95)}} = .301$$

Estimated value of $\hat{Y}$ for a case where $X = 9$

Eq. (2.8.3) $$SE_{\hat{Y}_i} = 10.82\sqrt{\frac{1}{50}+\frac{(9-7.67)^2}{(50-1)(20.95)}} = 1.59.$$

Estimated r_{YX}

Eq. (2.8.5) $$SE_{z'} = \frac{1}{\sqrt{50-3}} = .146$$

Estimated difference between B in two populations

Eq. (2.8.6) $$SE_{B_V-B_W} = \sqrt{.329^2+.329^2} = \sqrt{.2165} = .465.$$

Estimated difference between r's in two large samples from different populations

Eq. (2.8.7) $$SE_{r_V-r_W} = \sqrt{\frac{1-.657^2}{50}+\frac{1-.430^2}{50}} = \sqrt{.01136+.01630} = .166.$$

The *SE* is inversely proportional to $\sqrt{n}$ to a sufficient approximation when n is not small. Quadrupling n cuts *SE* approximately in half. To make a standard error x times as large as that for $n = 50$, compute $n* = n/x^2$, where $n*$ is the necessary sample size to attain x times the *SE*. For example, we found $SE_{B_{YX}} = .329$ for our sample of $n = 50$ cases. To make it half (.5) as large, we would need $n* = 50/.5^2 = 200$.

To change a standard error from *SE* to *SE*∗, find $n* = n(SE/SE*)^2$. For example, to change the $SE_{B_{YX}}$ from .329 (for $n = 50$) to $SE* = .20$, we would need $n* = 50\ (.329/.20)^2 = 135$ cases.

For differences between *B*s and *r*s, use their statistics from the trials to determine the desired changes in the *SE*s for the two samples and compute the anticipated *SE* of the difference (Eqs. 2.8.6 and 2.8.7). Adjust the *n*s as necessary.

2.9.2 Power of Null Hypothesis Significance Tests

In Section 2.8.3, we presented methods of appraising sample data in regard to α, the risk of mistakenly rejecting the null hypothesis when it is true, that is, drawing a spuriously positive conclusion (Type I error). We now turn our attention to methods of determining β,[16] the probability of *failing* to reject the null hypothesis when it is false (Type II error), and ways in which it can be controlled in research planning.

[16] We have been using β to represent the standardized regression coefficient. It is used here with a different meaning for consistency with the literature.

Any given test of a null hypothesis is a complex relationship among the following four parameters:

1. The power of the test, the probability of rejecting H_0, defined as $1 - \beta$.
2. The region of rejection of H_0 as determined by the α level and whether the test is one-tailed or two-tailed. As α increases, for example from .01 to .05, power increases.
3. The sample size n. As n increases, power increases.
4. The magnitude of the effect in the population, or the degree of departure from H_0. The larger this is, the greater the power.

These four parameters are so related that when any three of them are fixed, the fourth is completely determined. Thus, when an investigator decides for a given research plan the significance criterion α and n, the power of the test is determined. However, the investigator does not know what this power is without also knowing the magnitude of the effect size (*ES*) in the population, the estimation of which is the whole purpose of the study. The methods presented here focus on the standardized effect size, r in the present case.

There are three general strategies for estimating the size of the standardized population effect a researcher is trying to detect as "statistically significant":

1. To the extent that studies have been carried out by the current investigator or others which are closely similar to the present investigation, the *ES*s found in these studies reflect the magnitude that can be expected. Thus, if a review of the relevant literature reveals rs ranging from .32 to .43, the population *ES* in the current study may be expected to be somewhere in the vicinity of these values. Investigators who wish to be conservative may determine the power to detect a population ρ of .25 or .30.

2. In some research areas an investigator may posit some minimum population effect size that would have either practical or theoretical significance. An investigator may determine that unless $\rho = .05$, the importance of the relationship is insufficient to warrant a change in the policy or operations of the relevant institution. Another investigator may decide that a population correlation of .10 would have a material import for the adequacy of the theory within which the experiment has been designed, and thus would wish to plan the experiment so as to detect such an *ES*. Or a magnitude of B_{YX} that would be substantively important may be determined and other parameters estimated from other sources to translate B_{YX} into ρ.

3. A third strategy in deciding what *ES* values to use in determining the power of a study is to use certain suggested conventional definitions of *small*, *medium*, and *large* effects as population $\rho = .10$, .30, and .50, respectively (J. Cohen, 1988). These conventional *ES*s, derived from the average values in published studies in the social sciences, may be used either by choosing one of these values (for example, the conventional medium *ES* of .30) or by determining power for all three populations. If the latter strategy is chosen, the investigator would then revise the research plan according to an estimation of the relevance of the various *ES*s to the substantive problem. This option should be looked upon as the default option only if the earlier noted strategies are not feasible.

The point of doing a power analysis of a given research plan is that when the power turns out to be insufficient the investigator may decide to revise these plans, or even drop the investigation entirely if such revision is impossible. Obviously, because little or nothing can be done after the investigation is completed, determination of statistical power is of primary value as a preinvestigation procedure. If power is found to be insufficient, the research plan may be revised in ways that will increase it, primarily by increasing n, or increasing the number of levels or variability of the independent variable, or possibly by increasing α. A more complete general discussion of the concepts and strategy of power analysis may be found in J. Cohen (1965, 1988). It is particularly useful to use a computerized program for calculating the statistical power of a proposed research plan, because such a program will provide a graphic depiction of the effect of each of the parameters (ES, n, α) on the resulting power to reject a false null hypothesis.

2.10 FACTORS AFFECTING THE SIZE OF r

2.10.1 The Distributions of X and Y

Because $r = 1.00$ only when each $z_X = z_Y$, it can only occur when the shapes of the frequency distributions for X and Y are exactly the same (or exactly opposite for $r = -1.00$). The greater the departure from distribution similarity, the more severe will the restriction be on the maximum possible r. In addition, as such distribution discrepancy increases, departure from homoscedasticity—equal error for different predicted values—must also necessarily increase. The decrease in the maximum possible value of (positive) r is especially noticeable under circumstances in which the two variables are skewed in opposite directions. One such common circumstance occurs when the two variables being correlated are each dichotomies: With very discrepant proportions, it is not possible to obtain a large positive correlation.

For example, suppose that a group of subjects has been classified into "risk takers" and "safe players" on the basis of behavior in an experiment, resulting in 90 risk takers and 10 safe players. A correlation is computed between this dichotomous variable and self classification as "conservative" versus "liberal" in a political sense, with 60 of the 100 subjects identifying themselves as conservative (Table 2.10.1). Even if all political liberals were also risk takers in the experimental situation, the correlation will be only (by Eq. 2.3.6):

$$r_\phi = \frac{400 - 0}{\sqrt{90 \cdot 10 \cdot 40 \cdot 60}} = .272.$$

It is useful to divide the issue of the distribution of variables into two components, those due to differences in the distribution of the underlying constructs and those due to the scales on which we have happened to measure our variables. Constraints on correlations associated with differences in distribution inherent in the constructs are not artifacts, but have real interpretive meaning. For example, gender and height for American adults are not perfectly correlated, but we need have no concern about an artificial upper limit on r attributable to this distribution difference. If gender completely determined height, there would only be two heights, one for men and one for women, and r would be 1.00.

TABLE 2.10.1
Bivariate Distribution of Experimental and Self-Reported Conservative Tendency

		Experimental		
		Risk takers	Safe players	Total:
Self-report	Liberal	40	0	40
	Conservative	50	10	60
	Total:	90	10	100

Similarly the observed correlation between smoking and lung cancer is about .10 (estimated from figures provided by Doll & Peto, 1981). There is no artifact of distribution here; even though the risk of cancer is about 11 times as high for smokers, the vast majority of both smokers and nonsmokers alike will not contract lung cancer, and the relationship is low because of the nonassociation in these many cases.

Whenever the concept underlying the measure is logically continuous or quantitative[17]—as in the preceding example of risk taking and liberal versus conservative—it is highly desirable to measure the variables on a many-valued scale. One effect of this will be to increase the opportunity for reliable and valid discrimination of individual differences (see Section 2.10.2). To the extent that the measures are similarly distributed, the risk of underestimating the relationship between the conceptual variables will be reduced (see Chapter 4). However, the constraints on r due to unreliability are likely to be much more serious than those due to distribution differences on multivalued scales.

The Biserial r

When the only available measure of some construct X is a dichotomy, d_X, an investigator may wish to know what the correlation would be between the underlying construct and some other quantitative variable, Y. For example, X may be ability to learn algebra, which we measure by d_X, pass–fail. If one can assume that the "underlying" continuous variable X is normally distributed, and that the relationship with Y is linear, an estimate of the correlation between X and Y can be made, even though only d_X and Y are available. This correlation is estimated as

$$r_b = \frac{(M_{Y_P} - M_{Y_Q})PQ}{h(sd_Y)} = r_{pb}\frac{\sqrt{PQ}}{h}, \tag{2.10.1}$$

where M_{Y_P} and M_{Y_Q} are the Y means for the two points of the dichotomy, P and Q $(=1-P)$ are the proportions of the sample at these two points, and h is the ordinate (height) of the standard unit normal curve at the point at which its area is divided into P and Q portions (see Appendix Table C).

For example, we will return to the data presented in Table 2.3.1, where r_{pb} was found to be $-.707$. We now take the dichotomy to represent not the presence or absence of an experimentally determined stimulus but rather gross (1) versus minor (0) naturally occurring interfering stimuli as described by the subjects. This dichotomy is assumed to represent a continuous, normally distributed variable. The biserial r between stimulus and task score will be

$$r_b = \frac{(66 - 69.5)(.428)(.572)}{.392(2.45)} = -.893$$

where .392 is the height of the ordinate at the .428, .572 break, found by linear interpolation in Appendix Table C and $r_{pb} = -.707$.

The biserial r of $-.893$ may be taken to be an estimate of the product moment correlation that would have been obtained had X been a normally distributed continuous measure. It will always be larger than the corresponding point biserial r and, in fact, may even nonsensically exceed 1.0 when the Y variable is not normally distributed. When there is no overlap between the Y scores of the two groups, the r_b will be at least 1.0. It will be approximately 25% larger than the corresponding r_{pb} when the break on X is $.50 - .50$. The ratio of r_b/r_{pb} will increase

[17] *Continuous* implies a variable on which infinitely small distinctions can be made; *quantitative* or *scaled* is more closely aligned to real measurement practice in the behavioral sciences, implying an ordered variable of many, or at least several, possible values. Theoretical constructs may be taken as continuous, but their measures will be quantitative in this sense.

as the break on X is more extreme; for example with a break of .90 – .10, r_b will be about two-thirds larger than r_{pb}.

Confidence limits are best established on r_{pb} or, equivalently, on the difference between the Y means corresponding to the two points of d_X.

Tetrachoric r

As we have seen, when the relationship between two dichotomies is investigated, the restriction on the maximum value of r_ϕ when their breaks are very different can be very severe. Once again, we can make an estimate of what the linear correlation would be if the two variables were continuous and normally distributed. Such an estimate is called the tetrachoric correlation. Because the formula for the tetrachoric correlation involves an infinite series and even a good approximation is a laborious operation, tetrachoric *r*s are obtained by means of computer programs. Tetrachoric r will be larger than the corresponding phi coefficient and the issues governing their interpretation and use are the same as for r_b and r_{pb}.

Caution should be exercised in the use of biserial and tetrachoric correlations, particularly in multivariate analyses. Remember that they are not observed correlations in the data, but rather hypothetical ones depending on the normality of the distributions underlying the dichotomies. Nor will standard errors for the estimated coefficients be the same as those for the product moment coefficients presented here.

2.10.2 The Reliability of the Variables

In most research in the behavioral sciences, the concepts that are of ultimate interest and that form the theoretical foundation for the study are only indirectly and imperfectly measured in practice. Thus, typically, interpretations of the correlations between variables as measured should be carefully distinguished from the relationship between the constructs or conceptual variables found in the theory.

The reliability of a variable (r_{XX}) may be defined as the correlation between the variable as measured and another equivalent measure of the same variable. In standard psychometric theory, the square root of the reliability coefficient $\sqrt{r_{XX}}$ may be interpreted as the correlation between the variable as measured by the instrument or test at hand and the "true" (error-free) score. Because true scores are not themselves observable, a series of techniques has been developed to estimate the correlation between the obtained scores and these (hypothetical) true scores. These techniques may be based on correlations among items, between items and the total score, between other subdivisions of the measuring instrument, or between alternative forms. They yield a reliability coefficient that is an estimate (based on a sample) of the population reliability coefficient.[18] This coefficient may be interpreted as an index of how well the test or measurement procedure measures whatever it is that it measures. This issue should be distinguished from the question of the test's *validity*, that is, the question of whether *what* it measures is what the investigator intends that it measure.

The discrepancy between an obtained reliability coefficient and a perfect reliability of 1.00 is an index of the relative amount of measurement error. Each observed score may be thought of as composed of some true value plus a certain amount of error:

(2.10.2)
$$X = X_t + X_e.$$

[18] Because this is a whole field of study in its own right, no effort will be made here to describe any of its techniques, or even the theory behind the techniques, in any detail. Excellent sources of such information include McDonald (1999) and Nunnally & Bernstein (1993).

These error components are assumed to have a mean of zero and to correlate zero with the true scores and with true or error scores on other measures. Measurement errors may come from a variety of sources, such as errors in sampling the domain of content, errors in recording or coding, errors introduced by grouping or an insufficiently fine system of measurement, errors associated with uncontrolled aspects of the conditions under which the test was given, errors due to short- or long-term fluctuation in individuals' true scores, errors due to the (idiosyncratic) influence of other variables on the individuals' responses, etc.

For the entire set of scores, the reliability coefficient equals the proportion of the observed score variable that is true score variance

$$r_{XX} = \frac{sd^2_{X_t}}{sd^2_X} \tag{2.10.3}$$

Because, as we have stated, error scores are assumed not to correlate with anything, r_{XX} may also be interpreted as that proportion of the measure's variance that is available to correlate with other measures. Therefore, the correlation between the observed scores (X and Y) for any two variables will be numerically smaller than the correlation between their respective unobservable true scores (X_t and Y_t). Specifically,

$$r_{XY} = r_{X_tY_t}\sqrt{r_{XX}r_{YY}}. \tag{2.10.4}$$

Researchers sometimes wish to estimate the correlations between two theoretical constructs from the correlations obtained between the imperfect observed measures of these constructs. To do so, one corrects for attenuation (unreliability) by dividing r_{XY} by the square root of the product of the reliabilities (the maximum possible correlation between the imperfect measures). From Eq. (2.10.4),

$$r_{X_tY_t} = \frac{r_{XY}}{\sqrt{r_{XX}r_{YY}}}. \tag{2.10.5}$$

Thus, if two variables, each with a reliability of .80, were found to correlate .44,

$$r_{X_tY_t} = \frac{.44}{\sqrt{(.80)(.80)}} = .55.$$

Although correlations are subject to attenuation due to unreliability in either or both variables, bivariate regression coefficients are not affected by unreliability in Y. This can be seen from the following, where we consider unreliability only in Y. The regression coefficient expressed as the relationship between the perfectly reliable variables [by Eq. (2.4.3)] is

$$B_{Y_tX_t} = r_{X_tY_t}\left(\frac{sd_{Y_t}}{sd_{X_t}}\right). \tag{2.10.6}$$

By Eq. (2.10.5), when $r_{XX} = 1.0$, $r_{XY} = r_{XY_t}\sqrt{r_{YY}}$. By Eq. (2.10.3),

$$r_{YY} = \frac{sd^2_{Y_t}}{sd^2_{Y_t} + sd^2_{Y_e}} \quad \text{and} \quad sd_Y = \sqrt{sd^2_{Y_t} + sd^2_{Y_e}}$$

so

$$r_{XY_t} = \frac{r_{XY}}{\sqrt{sd^2_{Y_t}/(sd^2_{Y_t} + sd^2_{Y_e})}} \quad \text{and} \quad r_{XY} = r_{XY_t}\frac{sd_{Y_t}}{\sqrt{sd^2_{Y_t} + sd^2_{Y_e}}}.$$

Therefore, using Eq. (2.4.3) where $B_{YX} = r_{XY}(sd_Y/sd_X)$, substituting:

$$B_{YX} = r_{XY_t}\left(\frac{sd_{Y_t}}{\sqrt{sd^2_{Y_t} + sd^2_{Y_e}}}\right)\left(\frac{\sqrt{sd^2_{Y_t} + sd^2_{Y_e}}}{sd_X}\right)$$

and canceling

$$= r_{XY_t}\left(\frac{sd_{Y_t}}{sd_X}\right) = B_{Y_tX}$$

As is generally true for coefficients based on a series of estimates, caution must be used in interpreting attenuation-corrected coefficients, because each of the coefficients used in the equation is subject to sampling error (as well as model assumption failure). Indeed, it is even possible to obtain attenuation-corrected correlations larger than 1.0 when the reliabilities come from different populations than r_{XY}, are underestimated, or when the assumption of uncorrelated error is false. Obviously, because the disattenuated r is hypothetical rather than based on real data, its confidence limits are likely to be very large.[19]

To reiterate, unreliability in variables as classically defined is a sufficient reason for low correlations; it *cannot* cause correlations to be spuriously high. Spuriously high correlations may, of course, be found when sources of *bias* are shared by variables, as can happen when observations are not "blind," when subtle selection factors are operating to determine which cases can and cannot appear in the sample studied, and for yet other reasons.

2.10.3 Restriction of Range

CH02EX06

A problem related to the question of reliability occurs under conditions when the range of one or both variables is restricted by the sampling procedure. For example, suppose that in the data presented in Table 2.2.2 and analyzed in Table 2.6.1 we had restricted ourselves to the study of faculty members who were less extreme with regard to years since Ph.D., occupying the restricted range of 5 to 11 years rather than the full range of 3 to 18 years. If the relationship is well described by a straight line and homoscedastic, we shall find that the variance of the Y scores about the regression line, $sd^2_{Y-\hat{Y}}$, remains about the same. Because when $r \neq 0$, sd^2_Y will be decreased as an incidental result of the reduction of sd^2_X, and because $sd^2_Y = sd^2_{\hat{Y}} + sd^2_{Y-\hat{Y}}$, the proportion of sd^2_Y associated with X, namely, $sd^2_{\hat{Y}}$, will necessarily be smaller, and therefore, r^2 $(= sd^2_{\hat{Y}}/sd^2_Y)$ and r will be smaller. In the current example, r decreases from .657 to .388, and r^2, the proportion of variance, from .432 to .151. (See Table 2.10.2.) When the relationship is completely linear, the regression coefficient, B_{YX}, will remain constant because the decrease in r will be perfectly offset by the increase in the ratio sd_Y/sd_X. It is 2.456 here, compared to 1.983 before. (It increased slightly in this example, but could just as readily have decreased slightly.) The fact that regression coefficients tend to remain constant over changes in the variability of X (providing the relationship is fully linear and the sample size sufficiently large to produce reasonable estimates) is an important property of regression coefficients. It is shown later how this makes them more useful as measures of relationship than correlation coefficients in some analytic contexts (Chapter 5).

[19] Current practice is most likely to test "disattenuated" coefficients via latent variable models (described in Section 12.5.4), although the definition and estimation is somewhat different from the reasoning presented here.

TABLE 2.10.2
Correlation and Regression of Number of Publications on a Restricted Range of Time Since Ph.D.

	Publications	Time since Ph.D.	
	Y	X	
	3	6	
	17	8	
	11	9	
	6	6	$r_{XY} = .388\ (.657)^a$
	48	10	
	22	5	$r^2_{XY} = .150\ (.431)$
	30	5	
	21	6	$sd_{Y-\hat{Y}} = 11.10\ (10.42)$
	10	7	
	27	11	$B_{YX} = 2.456\ (1.983)$
M	19.50	7.30	
sd	12.04	1.31	
sd^2	144.94	1.71	

[a] Parenthetic values are those for the original (i.e., unrestricted) sample.

Suppose that an estimate of the correlation that would be obtained from the full range is desired, when the available data have a curtailed or restricted range for X. If we know the sd_Y of the unrestricted X distribution as well as the sd_{X_C} for the curtailed sample and the correlation between Y and X in the curtailed sample (r_{X_CY}), we may estimate r_{XY} by

$$\tilde{r}_{YX} = \frac{r_{YX_C}(sd_X/sd_{X_C})}{\sqrt{1 + r^2_{YX_C}\left(\left(sd^2_X/sd^2_{X_C}\right) - 1\right)}} \tag{2.10.7}$$

For example, $r = .25$ is obtained on a sample for which $sd_{X_C} = 5$ whereas the sd_X of the population in which the investigator is interested is estimated to be 12. Situations like this occur, for example, when some selection procedure such as an aptitude test has been used to select personnel and those selected are later assessed on a criterion measure. If the finding on the restricted (employed) sample is projected to the whole group originally tested, $\tilde{r}_{XY}$ would be estimated to be

$$\tilde{r}_{XY} = \frac{.25(12/5)}{\sqrt{1 + .25^2[(12/5)^2 - 1]}} = \frac{.60}{\sqrt{1.2975}} = .53$$

It should be emphasized that .53 is an estimate and assumes that the relationship is linear and homoscedastic, which might not be the case. There are no appropriate confidence limits on this estimate.

It is quite possible that restriction of range in either X or Y, or both, may occur as an incidental by-product of the sampling procedure. Therefore, it is important in any study to report the sds of the variables used. Because under conditions of homoscedasticity and linearity regression coefficients are not affected by range restriction, comparisons of different samples using the same variables should usually be done on the regression coefficients rather than on the correlation coefficients when sds differ. Investigators should be aware, however, that the questions answered by these comparisons are not the same. Comparisons of correlations answer the question, Does X account for as much of the variance in Y in group E and in Group F? Comparisons of regression coefficients answer the question, Does a change in X make the same amount of score difference in Y in group E as it does in group F?

Although the previous discussion has been cast in terms of restriction in range, an investigator may be interested in the reverse—the sample in hand has a range of X values that is large relative to the population of interest. This could happen, for example, if the sampling procedure was such as to include disproportionately more high- and low-X cases and fewer middle values. Equation (2.10.7) can be employed to estimate the correlation in the population of interest (whose range in X is less) by reinterpreting the subscript C in the equation to mean changed (including increased) rather than curtailed. Thus, r_{YX_C} and sd_{X_C} are the "too large" values in the sample, sd_Y is the (smaller) sd of the population of interest, and the estimated r in that population will be smaller. Note that the ratio sd_Y/sd_{X_C}, which before was greater than one, is now smaller than one. Because the correlation (the *ES*) will be higher in a sample with a larger *sd*, sampling in order to produce a larger *sd*, as in studies in which the number of "cases" is larger than in a random sample of the general population, is a major strategy for increasing the statistical power of a study.

2.10.4 Part-Whole Correlations

Occasionally we will find that a correlation has been computed between some variable J and another variable W, which is the sum of scores on a set of variables including J. Under these circumstances a positive correlation can be expected between J and W due to the fact that W includes J, even when there is no correlation between J and $W - J$. For example, if k test items of equal *sd* and zero r with each other are added together, each of the items will correlate exactly $1/\sqrt{k}$ with the total score. For the two-item case, therefore, each item would correlate .707 with their sum, W, when neither correlates with the other. On the same assumptions of zero correlation between the variables but with unequal *sd*s, the variables are effectively weighted by their differing sd_i and the correlation of J with W will be equal to $sd_J/\sqrt{\Sigma sd_i^2}$, where *sd*s are summed over the items. Obviously, under these circumstances $r_{J(W-J)} = 0$. In the more common case where the variables or items are correlated, the correlation of J with $W - J$ may be obtained by

$$r_{J(W-J)} = \frac{r_{JW}sd_W - sd_J}{\sqrt{sd_W^2 + sd_J^2 - 2r_{JW}sd_W sd_J}} \qquad (2.10.8)$$

This is not an estimate and may be tested via the usual t test for the significance of r.

Given these often substantial spurious correlations between elements and totals including the elements, it behooves the investigator to determine $r_{J(W-J)}$, or at the very least determine the expected value when the elements are uncorrelated before interpreting r_{JW}. Such a circumstance often occurs when the interest is in the correlation of a single item with a composite that includes that item, as is carried out in psychometric analysis.

Change Scores

It is not necessary that the parts be literally added in order to produce such spurious correlation. If a subscore is subtracted, a spurious negative component in the correlation will also be produced. One common use of such difference scores in the social sciences in the use of post- minus pretreatment (change) scores. If such change scores are correlated with the pre- and posttreatment scores from which they have been obtained, we will typically find that subjects initially low on X will have larger gains than those initially high on X, and that those with the highest final scores will have made greater gains than those with lower final scores. Again, if $sd_{\text{pre}} = sd_{\text{post}}$ and $r_{\text{pre post}} = 0$, the $r_{\text{pre change}} = -.707$ and $r_{\text{post change}} = +.707$. Although in general, we would expect the correlation between pre- and posttreatment scores to be some positive value, it will be limited by their respective reliabilities (Section 2.10.2) as well as by individual differences in true change.

If the post- minus pretreatment variable has been created in order to control for differences in pretreatment scores, the resulting negative correlations between pretreatment and change scores may be taken as a failure to remove all influence of pretreatment scores from posttreatment scores. This reflects the regression to the mean phenomenon discussed in Section 2.5 and the consequent interpretive risks. The optimal methods of handling this and related problems are the subject of a whole literature (Collins & Horn, 1993) and cannot be readily summarized. However, the appropriate analysis, as always, depends on the underlying causal model. (See Chapters 5, 12, and 15 for further discussion of this problem.)

2.10.5 Ratio or Index Variables

Ratio (index or rate) scores are those constructed by dividing one variable by another. When a ratio score is correlated with another variable or with another ratio score, the resulting correlation depends as much on the denominator of the score as it does on the numerator. Because it is usually the investigator's intent to "take the denominator into account" it may not be immediately obvious that the correlations obtained between ratio scores may be spurious—that is, may be a consequence of mathematical necessities that have no valid interpretive use. Ratio correlations depend, in part, upon the correlations between all numerator and denominator terms, so that $r_{(Y/Z)X}$ is a function of r_{YZ} and r_{XZ} as well as of r_{YX}, and $r_{(Y/Z)(X/W)}$ depends on r_{YW} and r_{XZ} as well as on the other four correlations. These correlations also involve the coefficients of variation

(2.10.9)
$$v_X = \frac{sd_X}{M_X}$$

of each of the variables. Although the following formula is only a fair approximation of the correlation between ratio scores (requiring normal distributions and homoscedasticity and dropping all terms involving powers of v greater than v^2), it serves to demonstrate the dependence of correlations between ratios on all vs and on rs between all variable pairs:

(2.10.10)
$$r\,(Y/Z)\,(X/W) = \frac{r_{YX}v_Yv_X - r_{YW}v_Yv_W - r_{XZ}v_Xv_Z - r_{ZW}v_Zv_W}{\sqrt{v_Y^2 + v_Z^2 - 2r_{YZ}v_Yv_Z}\sqrt{v_X^2 + v_W^2 - 2r_{XW}v_Xv_W}}$$

When the two ratios being correlated have a common denominator, the possibility of spurious correlations becomes apparent. Under these circumstances, the approximate formula for the correlation simplifies, because $Z = W$. If all coefficients of variation are equal when all three variables are uncorrelated we will find $r_{(Y/Z)(X/Z)} \approx .50$.

Because the coefficient of variation depends on the value of the mean, it is clear that whenever this value is arbitrary, as it is for many psychological scores, the calculated r is also arbitrary. Thus, ratios should not be correlated unless each variable is measured on a ratio scale, a scale for which a zero value means literally none of the variable (see Chapters 5 and 12). Measures with ratio scale properties are most commonly found in the social sciences in the form of counts or frequencies.

At this point it may be useful to distinguish between rates and other ratio variables. Rates may be defined as variables constructed by dividing the number of instances of some phenomenon by the total number of opportunities for the phenomenon to occur; thus, they are literally proportions. Rates or proportions are frequently used in ecological or epidemiological studies where the units of analysis are aggregates of people or areas such as counties or census tracts. In such studies, the numerator represents the incidence or prevalence of some phenomenon and the denominator represents the population at risk. For example, a delinquency rate may be calculated by dividing the number of delinquent boys ages 14–16 in a county by the total number of boys ages 14–16 in the county. This variable may be correlated across the counties in a region with the proportion of families whose incomes are below the poverty level, another rate. Because, in general, the denominators of these two rates will reflect the populations of the counties, which may vary greatly, they can be expected to be substantially correlated. In other cases the denominators may actually be the same—as, for example, in an investigation of the relationship between delinquency rates and school dropout rates for a given age-gender group. The investigator will typically find that these rates have characteristics that minimize the problem of spurious correlation. In most real data, the coefficients of variation of the numerators will be substantially larger than the coefficients of variation of the denominators, and thus the correlation between rates will be determined substantially by the correlation between the numerators. Even in such data, however, the resulting proportions may not be optimal for the purpose of linear correlation. Chapter 6 discusses some nonlinear transformations of proportions, which may be more appropriate for analysis than the raw proportions or rates themselves.

Experimentally produced rates may be more subject to problems of spurious correlation, especially when there are logically alternative denominators. The investigator should determine that the correlation between the numerator and denominator is very high (and positive), because in general the absence of such a correlation suggests a faulty logic in the study. In the absence of a large correlation, the coefficients of variation of the numerator should be substantially larger than that of the denominator if the problem of spurious correlation is to be minimized.

Other Ratio Scores

When the numerator does not represent some subclass of the denominator class, the risks involved in using ratios are even more serious, because the likelihood of small or zero correlations between numerators and denominators and relatively similar values of v is greater. If the variables do not have true zeros and equal intervals, correlations involving ratios should probably be avoided altogether, and an alternative method for removing the influence of Z from X or Y should be chosen, as presented in Chapters 3 and 12.

The difficulties that may be encountered in correlations involving rates and ratios may be illustrated by the following example. An investigator wishes to determine the relationship between visual scanning and errors on a digit-symbol (d-s) task. All subjects are given 4 minutes to work on the task. Because subjects who complete more d-s substitutions have a greater opportunity to make errors, the experimenter decides, reasonably enough, to determine the error rate by dividing the number of errors by the number of d-s substitutions completed. Table 2.10.3 displays the data for 10 subjects. Contrary to expectation, subjects who completed more d-s tasks did not tend to produce more errors ($r_{ZX} = -.105$), nor did they scan notably more than did low scorers ($r_{ZY} = .023$). Nevertheless, when the two ratio scores are computed, they show a substantial positive correlation (.427) in spite of the fact that the numerators showed slight negative correlation (−.149), nor is there any tendency for scanning and errors to be correlated for any given level of d-s task completion. Thus, because $r_{ZZ} = 1$, the $r_{(X/Z)(Y/Z)}$ may here be seen to be an example of spurious correlation.[20]

[20] An alternative method of taking into account the number completed in considering the relationship between errors and number of scans might be to partial Z (see subsequent chapters).

TABLE 2.10.3
An Example of Spurious Correlation Between Ratios

Subject	No. completed d-s tasks (Z)	No. errors (X)	No. scans (Y)	Error rate (X/Z)	Scan rate (Y/Z)
1	25	5	24	.20	.96
2	29	3	30	.10	1.03
3	30	3	27	.10	.90
4	32	4	30	.12	.94
5	37	3	18	.08	.49
6	41	2	33	.05	.80
7	41	3	27	.07	.66
8	42	5	21	.12	.50
9	43	3	24	.07	.56
10	43	5	33	.12	.77

$r_{ZX} = -.105, \quad r_{ZY} = .106, \quad r_{XY} = -.149$

$r_{(X/Z)(Y/Z)} = .427$

2.10.6 Curvilinear Relationships

When the relationship between the two variables is only moderately well fitted by a straight line, the correlation coefficient that indicates the degree of linear relationship will understate the predictability from one variable to the other. Frequently the relationship, although curvilinear, is monotonic; that is, increases in Z are accompanied by increases (or decreases) in Y, although not at a constant rate. Under these circumstances, some (nonlinear) monotonic transformation of X or Y or both may straighten out the regression line and provide a better indication of the size of the relationship between the two variables (an absolutely larger r). Because there are several alternative ways of detecting and handling curvilinear relationships, the reader is referred to Chapters 4 and 6 for a detailed treatment of the issues.

2.11 SUMMARY

A linear relationship exists between two quantitative variables when there is an overall tendency for increases in the value of one variable to be accompanied by increases in the other variable (a positive relationship), or for increases in the first to be accompanied by decreases in the second (a negative relationship); (Section 2.1). Efforts to index the degree of linear relationship between two variables must cope with the problem of the different units in which variables are measured. Standard (z) scores are a conversion of scores into distances from their own means, in standard deviation units, and they render different scores comparable. The Pearson product moment correlation coefficient, r, is a measure of the degree of relationship between two variables, X and Y, based on the discrepancies of the subjects' paired z scores, $z_X - z_Y$. r varies between -1 and $+1$, which represent perfect negative and perfect positive linear relationships, respectively. When $r = 0$, there is no linear correlation between the variables (Section 2.2).

r can be written as a function of z score products, a function of variances and covariance, or in terms of the original units. Special simplified formulas are available for r when one variable is a dichotomy (point biserial r), when both variables are dichotomies (r_ϕ), or when the data are two sets of complete ranks (Spearman rank order correlation); (Section 2.3).

The regression coefficient, B_{YX}, gives the optimal rule for a linear estimate of Y from X, and is the change in Y units per unit change in X, that is, the slope of the regression line. The intercept, B_0, gives the predicted value of Y for a zero value of X. B_{YX} and B_0 are optimal in the sense that they provide the smallest squared discrepancies between Y and estimated $\hat{Y}$. r is the regression coefficient for the standardized variables. When X is centered, $B_0 = M_Y$ (Section 2.4). Unless $r = 1$, it is a mathematical necessity that the average score for a variable being estimated (e.g., $\hat{Y}$) will be relatively closer to M_Y than the value from which it is being estimated (e.g., X) will be to its mean (M_X) when both are measured in *sd* units (Section 2.5).

When Y is estimated from X the *sd* of the difference between observed scores and the estimated scores (the sample standard error of estimate) can be computed from r and sd_Y. The coefficient of alienation represents the error as a proportion of the original sd_Y. r^2 equals the proportion of the variance (sd^2) of each of the variables that is shared with or can be estimated from the other (Sections 2.6 and 2.7).

The two major methods of statistical inference are estimation and null hypothesis testing. The formal model assumptions are presented (Section 2.8.1), confidence intervals are given for B_{YX}, B_{Y0}, r_{XY}, for differences between independent sample values of these statistics, and for the estimated $\hat{Y}_i$ (Section 2.8.2). Given α, confidence intervals provide the range of values within which the corresponding population values can be expected to fall. In Section 2.8.3, we present the null hypothesis tests for simple regression and correlation statistics. Section 2.8.4 critiques null hypothesis testing and contrasts it with the use of confidence intervals.

The degree of accuracy (precision) in the estimation of parameters is reflected in the statistic's confidence interval. The probability of null hypothesis rejection (statistical power) can be assessed before the research sample is collected (Section 2.9). Methods of finding the sample size to produce a margin of error for a given degree of inclusion in the confidence interval (95%, 80%) are presented (Section 2.9.1) and methods are given for determining the sample size needed for the desired statistical power, that is, the probability of rejecting the null hypothesis (Section 2.9.2).

A number of characteristics of the X and Y variables will affect the size of the correlation between them. Among these are differences in the distribution of the X and Y variables (Section 2.10.1), unreliability in one or both variables (Section 2.10.2), and restriction of the range of one or both variables (Section 2.10.3). When one variable is included as a part of the other variable, the correlation between them will reflect this overlap (Section 2.10.4). Scores obtained by dividing one variable by another will produce spurious correlation with other variables under some conditions (Section 2.10.5). The r between two variables will be an underestimate of the magnitude of their relationship when a curved rather than a straight line best fits the bivariate distribution (Section 2.10.6). Under such circumstances, transformation of one or both variables or multiple representation of one variable will provide a better picture of the relationship between the variables.