C. Tolmie

REFIGURING ANTHROPOLOGY

First Principles Of Probability & Statistics

David Hurst Thomas
American Museum of Natural History

Waveland Press, Inc.
Prospect Heights, Illinois

For information about this book, write or call:

Waveland Press, Inc.
P.O. Box 400
Prospect Heights, Illinois 60070
(312) 634-0081

For permission to use copyrighted material, the author is indebted to the following:

FIG. 2.1. By permission of the Trustees of the British Museum (Natural History).
TABLE 2.3. (p. 24) *From Physical Anthropology: An Introduction* by A. J. Kelso. Reprinted by permission of the publisher, J. B. Lippincott Company. Copyright © 1974. (p. 25) Reproduced by permission of the Society for American Archaeology from *Memoirs of the Society for American Archaeology*, Vol. 11, 1956.
FIG. 3.1. From Hulse, Frederick S. *The Human Species: An Introduction to Physical Anthropology*. Copyright © 1963 by Random House, Inc.
FIG. 3.2. From Dozier, Edward P., *The Pueblo Indians of North America*. Copyright © 1970 by Holt, Rinehart and Winston, Inc. Reproduced by permission of Holt, Rinehart and Winston. [This book reissued 1983 by Waveland Press, Inc.]
FIG. 3.4. Reproduced by permission of the Society for American Archaeology from *American Antiquity*, Vol. 35 (4), 1970.
FIG. 3.5. Reproduced by permission of the American Anthropological Association from the *American Anthropologist*, Vol. 73 (3), 1971.
FIG. 13.14. From *Biometry* by Robert R. Sokal and F. James Rohlf. W. H. Freeman and Company. Copyright © 1969.

Second Printing
The 1976 version of this book was entitled *Figuring Anthropology*.

ISBN 0-88133-223-2

Printed in the United States of America.

14 Correlation Coefficients

● *If experimentation is the Queen of the Sciences, surely statistical methods must be regarded as the Guardian of the Royal Virtue.*—M. Tribus

14.1 CORRELATION IN A SAMPLE

The notion of correlation was briefly introduced in Section 13.3. The population correlation coefficient ρ was defined as the square root of the Coefficient of Determination. But ρ applies only to correlation within a population. When a sample is involved, a new statistic, called r, must be considered. Therefore, r estimates ρ from a sample of variates. Originally derived by biometrist Karl Pearson, the sample correlation coefficient is also commonly called the *Pearson Product-Moment Coefficient.* The correlation coefficient (and its nonparametric equivalents) plays such a critical role in bivariate statistical analysis that correlation must be given especial attention in this chapter.

Consider the scattergram in Fig. 14.1. Coordinate systems can help in assessing the degree of scatter in a swarm of points. The coordinates describe four equal quadrants, labelled *A, B, C,* and *D.* If the swarm of points is randomly distributed about their sample means—in this case, about the origin of the graph—then the number of points should be roughly distributed throughout all the quadrants. But if a linear relationship is present, then the points will be distributed unequally among the four quadrants. A *positive linear relationship* produces a distribution in which quadrants *A* and *C* contain more points than would quadrants *B* and *D.* Similarly, a *negative linear relationship* places an excess of points in quadrants *B* and *D.* So the relative abundance of points among the four quadrants can serve as a rough indicator of linearity; Fig. 14.1 illustrates this simple principle.

Now consider the scattergram in Fig. 14.2, and assume that these datum

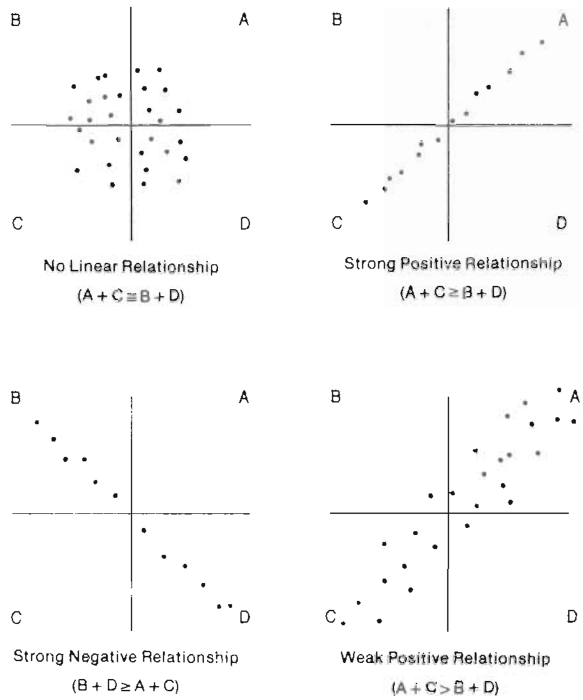


Fig. 14.1

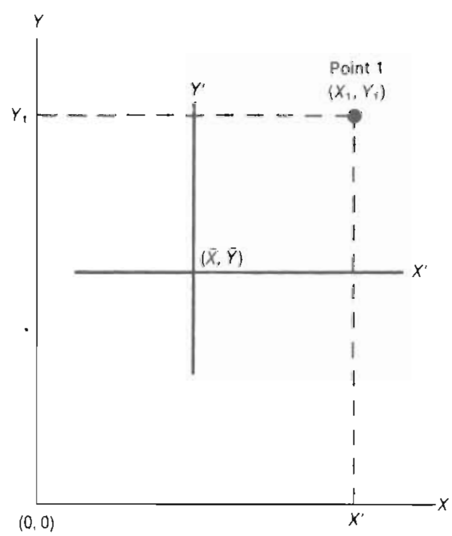


Fig. 14.2

points (X_i, Y_i) represent random samples from a homoscedastic bivariate normal population. If the intention were to describe merely the *form* of the relationship between X and Y, then a *descriptive* line could be fit to these points, using Model II regression (Section 13.8). But let us set aside the question of form for the moment and consider only the *strength* of the linear relationship between X and Y. Following convention, the origin of the $X-Y$ coordinate system is set at (0, 0); any point can be located on Fig. 14.2 merely by describing its vertical and horizontal relationship to the zero origin. This coordinate system can be redefined (as in Fig. 14.3) such that the origin is placed at the two sample means, $\bar{X}$ and $\bar{Y}$. Each of these new axes can be labelled X' and Y' to distinguish them from the original axes X and Y which originated at (0, 0).

Any individual datum point can now be located precisely on the coordinate system by measuring its horizontal and vertical distance from the origin. When the origin was taken as zero, then any point is exactly $(X_i - 0) = X_i$ horizontal units from the origin and $(Y_i - 0) = Y_i$ vertical units above the origin. With the origin redefined as $\bar{X}$ and $\bar{Y}$, the distance to the origin becomes $(X_i - \bar{X})$ horizontal units and $(Y_i - \bar{Y})$ vertical units. The values of the product between these two distances, $(X_i - \bar{X})(Y_i - \bar{Y})$, provides a general distance figure from the point to the graph origin. When a point falls within quadrant A, then the product $(X_i - \bar{X})(Y_i - \bar{Y})$ must be a positive number (Fig. 14.3). Similarly, points in quadrant C must also have a positive product, while the points in B and D must always produce a negative product of deviations.

This reasoning can be generalized from a single point to the entire swarm of points on a scattergram. A strong, positive linear relationship produces a positive value of the *sum of products*, which is $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$, because most of the points must lie within quadrants A and C. The larger this sum, the stronger must be the positive linearity between X and Y. Similarly, a negative linearity produces a larger negative value for $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$. No linear trend is

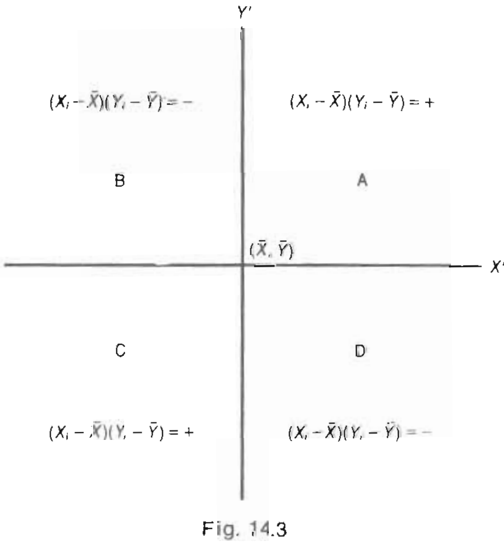


Fig. 14.3

evident whenever the sum of the products of the deviations from $\bar{X}$ and $\bar{Y}$ is zero.

So, clearly, the sum of deviations is a serviceable indicator of the strength of a bivariate linear relationship, but this index is hampered by a couple of limitations. The sum obviously increases as the number of points, n, increases, and therefore the sum of deviations is useful only as long as samples of identical size are to be compared. Moreover, the sum is expressed in the units of analysis—that is, the diverse units of the scales of X and Y—rather than in simple absolute units. The initial difficulty is remedied simply by dividing the sum by n: $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})/n$ expresses the *average* deviation about $\bar{X}$ and $\bar{Y}$, thereby stabilizing the sum against fluctuations in sample size. Each individual deviation can then be divided by the sample standard deviation in order to cancel the actual units of analysis; a similar procedure was used to derive z, the standardized normal deviate. This modified indicator measures the dispersion of a bivariate normal sample about its individual means $\bar{X}$ and $\bar{Y}$,[1] as follows:

$$r = \frac{1}{n}\sum \frac{(X_i - \bar{X})}{S_X} \cdot \frac{(Y_i - \bar{Y})}{S_Y}$$

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})/n}{\sqrt{\Sigma(X_i - \bar{X})^2(Y_i - \bar{Y})^2/(n-1)}} \tag{14.1}$$

The sample correlation coefficient r is simply an alternative algebraic form for estimating ρ, derived in Section 13.3. This form in Equation (14.1) is preferable because it illustrates just how affinity can be computed over a series of bivariate pairs.

The correlation coefficient has a number of desirable properties:

1. A value of $\rho = 0$ (estimated by r) indicates that no linear relationship exists between two variables. They are *linearly unrelated.*
2. The *magnitude of* ρ (estimated by r) denotes the *strength* of the linear relationship. Large absolute values of ρ indicate a close relationship, while smaller absolute values of ρ indicate that X and Y are only weakly related.
3. The *sign of* ρ denotes the *direction* of the relationship.
4. The maximum value of $\rho = +1.00$ indicates a perfect positive correlation (larger X means larger Y) and the maximum negative value of $\rho = -1.00$ indicates a perfect negative correlation (larger X means smaller Y).

The sample correlation coefficient is very closely related to the sample regression constants in a mathematical sense. The correlation coefficient is simply the slope constant b multiplied by the ratio of the sample standard deviations of X and Y:

$$r = b_{Y \cdot X}\frac{S_X}{S_Y} \tag{14.2}$$

Because the regression constant b is expressed in the specific units of analysis ("so many unit changes in Y for every unit change in X"), multiplication by the standard deviations will "standardize" the slope into a dimensionless statement of correlation. But this intimate relationship between r and b should not be

[1]For large samples, the difference between n and $n-1$ becomes negligible, so they simply "cancel" in Expression 14.1.

taken as meaning that they are equivalent expressions, to be interchanged at will. They rely upon some rather different assumptions, and the methods of correlation and regression fulfill rather different needs in social science.

14.2 COMPUTING THE CORRELATION COEFFICIENT

A "computing" method for finding r is given in the following computing formula:

$$r = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\sqrt{(\Sigma X^2 - n\bar{X}^2)(\Sigma Y^2 - n\bar{Y}^2)}} \tag{14.3}$$

Like most of the computing formulas, Expression (14.3) avoids the difficulties of actually determining the individual deviations about the sample mean. Some examples should clarify the meaning and computation of the correlation coefficient.

I employ a number of students in the archaeological lab of the American Museum of Natural History in New York City. The bulk of their duties consists of measuring and classifying archaeological artifacts. One task, for instance, involves taking ten measurements on all of the projectile points processed through the laboratory. Aside from the general grousing I have come to expect from such vacuous duties, a couple of students made a seemingly legitimate protest. "Why do we have to measure *length*, *thickness*, and *weight* for each artifact," they asked, "when we all *know* that the three variables are functions of but a single variable—*size*. Because most of the artifacts are broken, only thickness can be measured with accuracy; both weight and length are generally only estimates from broken artifacts. Why must length and weight be determined when we already know thickness?"

The question set me thinking just how closely thickness, weight, and length were really related in these artifacts. This is an issue of correlation: If length, thickness, and weight are highly correlated, then they are also redundant, and one measurement will serve just as well as three. So I told the students that if they could demonstrate adequate correlation, they could junk the redundant measurements.

They began first with the thickness and weight variates for eight Elko Eared projectile points (Table 14.1). Their first step was to plot the data on a scattergram (Fig. 14.4). The symbols X and Y have been assigned arbitrarily in this case because no predictions are involved. A linear trend appears to be evident, but there is also a good deal of scatter. How correlated are weight and thickness?

All the necessary terms have been computed in Table 14.1, and from Formula (14.3) the coefficient of correlation is found to be

$$r = \frac{155.49 - 8(5.10)(3.76)}{\sqrt{[(212.80 - 8(5.10)^2][(114.61 - 8(3.76)^2]}}$$
$$= +0.78$$

The value of $r = +0.78$ is a rather high value, demonstrating a certain validity of the students' complaints. Weight and thickness are indeed quite redundant, and should produce quite similar results in any typological scheme.

TABLE 14.1 Comparison of weight and thickness measurements for eight Elko Eared projectile points from Reese River, central Nevada (Thomas 1971a)

Thickness, cm X	Weight, grams Y	XY	X^2	Y^2
5.0	3.3	16.50	25.00	10.89
4.6	3.5	16.10	21.16	12.25
4.8	4.0	19.20	23.04	16.00
5.8	3.8	22.04	33.64	14.44
5.3	4.2	22.26	28.09	17.64
4.3	3.5	15.05	18.49	12.25
6.7	4.5	30.15	44.89	20.25
4.3	3.3	14.19	18.49	10.89
40.8	30.1	155.49	212.80	114.61

$\bar{X} = 40.8/8 = 5.10$; $\bar{Y} = 30.1/8 = 3.76$.

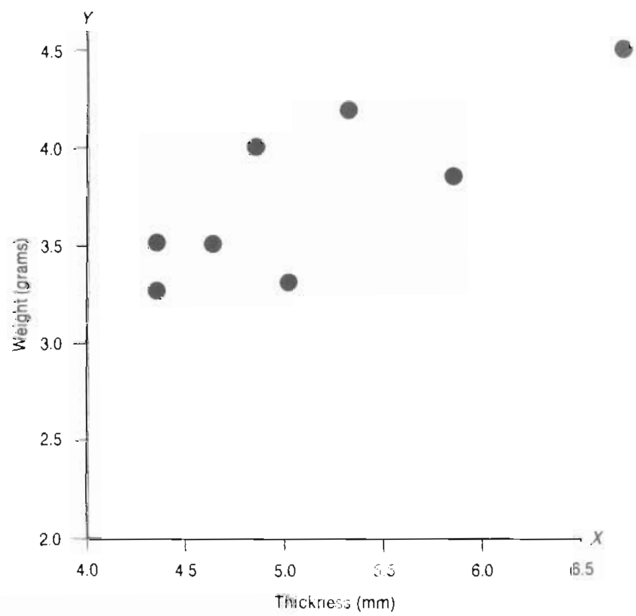


Fig. 14.4 Thickness versus weight of eight Elko Eared points from Reese River, Nevada (data from Thomas 1971a).

Example 14.1

According to population ecologist Paul Ehrlich, "the single most important factor in a country's reproductive rate is the motivation of the people toward the regulation of family size . . . if a couple is determined not to have more than two children, they usually will not, regardless of whether there is a birth control clinic down the street" (Ehrlich and Ehrlich 1972: 318).

Do the following data from seven Latin American cities support Ehrlich's contention that the actual birth rate is correlated with social norms regarding ideal family size?

Desired family sizes of women in seven Latin American cities.

Latin American Cities	Average Number of Children Wanted	1971 Birth Rate of Country
Bogota, Columbia	3.6	44
Buenos Aires, Argentina	2.9	22
Caracas, Venezuela	3.5	41
Mexico City, Mexico	4.2	42
Panama City, Panama	3.5	41
Rio de Janeiro, Brazil	2.7	38
San Jose, Costa Rica	3.6	45

As in all correlation cases, assignment of X and Y is totally arbitrary: The number of desired children is assigned to X and the 1971 birth rate is Y.

X	Y	X^2	Y^2	XY
3.6	44	12.96	1,936	158.40
2.9	22	8.41	484	63.80
3.5	41	12.25	1,681	143.50
4.2	42	17.64	1,764	176.40
3.5	41	12.25	1,681	143.50
2.7	38	7.29	1,444	102.60
3.6	45	12.96	2,025	162.00
24.0	273	83.76	11,015	950.20

$$\bar{X} = \frac{24.0}{7} = 3.43 \qquad \bar{Y} = \frac{273}{7} = 39.00$$

$$r = \frac{950.20 - 7(3.43)(39)}{\sqrt{[(83.76 - 7(3.43)^2][(11,015) - 7(39)^2]}}$$
$$= +0.607$$

These data clearly support Ehrlich's contention that norms are positively related to actual birth rate, at least in Latin America.

evada

14.3 THE MEANING OF CORRELATION AND REGRESSION: USE AND ABUSE

Many practicing anthropologists feel perfectly at home with most statistical techniques, yet they experience a certain trepidation when faced with the issues of correlation and regression. The computations are so closely related arithmetically that one is often tempted to compute both r and the regression equation for all sets of bivariate data, in hopes of covering all the bases. But these two techniques are hardly interchangeable, and their misuse is notoriously common throughout the literature of anthropology. It is difficult to know which statistic is more abused, the chi-square test or the regression-correlation duo. Neither misuse is particularly amusing.

It seems that relatively few problems arise when the regression and correlation statistics are employed in purely descriptive fashion. The population regression coefficients readily describe the *form* of the linear relationship, while the parametric correlation coefficient ρ measures the *degree of dispersion* about this regression axis. But difficulties seem to arise when samples are generated and then inferences extended back to the parent population. Two intersecting criteria must be considered whenever such bivariate samples are to be analyzed: (1) the precise objective of analysis, and (2) the exact nature of the variables which were sampled. Table 14.2 summarizes the following discussion.

Both regression and correlation assume at least interval scales of measurement. When either X or Y fails to qualify as fully interval, then one must turn to one of the nonparametric techniques considered later in this chapter. As long as the values of X are fixed—that is, whenever the levels of the predictor variable are under the control of the experimenter—the least squares approximation of regression should be used to describe the precise relationship between X and Y. Least square regression permits the analyst to predict the probabilistic outcomes of Y, given information about the predetermined levels of X. This agreeable situation occurs most frequently in disciplines such as psychological experimentation, educational testing, and agricultural field studies. The investigator predetermines a value of X and then measures the attendant responses on Y_i. The X variable really has no "distribution" in the strict sense, so

TABLE 14.2 Relationship between correlation and regression (after Sokal and Rohlf 1969: 497).

Nature of Selecting X_i and Y_i	Purpose of Investigation: Determine dependence relationship (*prediction*)	Purpose of Investigation: Establish strength of association (*interdependence*)
X fixed, Y random	Model I regression (least squares)	*Meaningless*, except as measure of goodness of fit between data and line of regression (use r^2)
Both X and Y random	Model II regression (Bartlett's method)	Correlation coefficient

assumptions are not required about the variance of X. X has been selected rather than sampled.

How secure are these predictions based upon Model I regression? To determine the variability of Y about an arbitrary X, the amount of variance "accounted for" by X must be compared with the total variability in Y. There is no variability in X. The Coefficient of Nondetermination was previously defined as

$$k^2 = 1 - \frac{\text{variability in } Y \text{ not accounted for by } X}{\text{total variability in } Y}$$

$$= 1 - \frac{S^2_{X \cdot Y}}{S^2_Y}$$

When $k^2 = 0$, then the variability in Y is completely determined from a knowledge of X.

The accuracy of Model I regressions can alternatively be expressed as the Coefficient of Dispersion, defined earlier as $r^2 = (1 - k^2)$. Whenever $r^2 = 1$, the X variability is said to account for the total variability observed in Y. *But the correlation coefficient r is meaningless in the contexts of Model I regression.* The sample correlation always assumes a bivariate normal distribution, which is clearly never the case when levels of X are fixed. Lacking the bivariate normal distribution, one cannot infer ρ from r. Thus, although the correlation coefficient is computationally related to r^2, r is merely the square root of the Coefficient of Determination—these two statistics are grounded in very different assumptions. The statistics r and r^2 are not interchangeable for Model I regression; r^2 has meaning only when X has been fixed. Furthermore, regression is a predictive technique for guessing the value of Y given X, while the Coefficient of Determination provides an estimate of goodness of fit for these predictions.

A rather different statistical situation exists whenever both X and Y represent random variables. In general, sampling from bivariate normal populations implies an interest more in the strength of relationships than in their form. So bivariate populations are more generally involved with correlation as the basic analytical tool. The correlation approach applies to the sampling situation, whereas the (Model I) regression approach implies a more closely controlled experimental study. Should one actually need to *predict* one random variable from another, then Bartlett's method (Section 13.8) for curve fitting can be used to describe the observed form of articulation between random X and Y. The labels *predictor* and *predicted* are assigned arbitrarily in Model II because there is no structural difference between the distributions of the two variables.

14.4 TESTING r FOR STATISTICAL SIGNIFICANCE

Extreme values of r are relatively easy to interpret. As long as r hovers about zero, then one can feel quite assured that no substantial degree of correlation exists between the X and Y populations. That is, r is almost certainly near zero. Similarly, as r approaches the maximum values of $r = \pm 1.00$, then linear correlation seems a virtual certainty. But because r is merely a statistical estimator of ρ, the sampling errors cannot be ignored. Whenever r assumes an intermediate value between zero and unity, the correlation coefficient should be assessed for *statistically significant deviations by chance.* A number of null hypotheses can be considered, depending upon the precise research objective

14.4.1 Testing against a Specific ρ

The most common statistical test for the significance of the sample correlation coefficient is to determine whether ρ differs from zero. Are the two variables correlated?

$$H_o: \ \rho = 0 \qquad H_o: \ \rho \geq 0 \qquad H_o: \ \rho \leq 0$$
$$H_1: \ \rho \neq 0 \qquad H_1: \ \rho < 0 \qquad H_1: \ \rho > 0$$

This test attempts to determine if the observed deviation of r from zero is sufficiently large to represent a rare sampling event. The test can be phrased in either one- or two-tailed forms.

Observed values of r can be converted to the familiar t-statistic as follows:

$$t = \frac{r - \rho}{\sqrt{(1 - r^2)/(n - 2)}} \tag{14.4}$$

Two degrees of freedom are lost in computing Expression (14.4), so for any r, df $= (n - 2)$.

The significance of r can also be tested by using simple distribution tables. The sampling distribution of r is known to vary, depending both upon ρ and n. For the special case of $H_o: \rho = 0$, the probability values have been compiled in Table A.11 (Appendix). Entering this table with $(n - 2)$ degrees of freedom, one can readily determine the *critical values* of r at the common levels of statistical significance. This table applies only to the two-tailed case, so the sign of r is ignored. Whenever direction has been specified—that is, one predicts either positive or negative correlation—the one-tailed probabilities can be found as simply twice those listed in Table A.11. So the critical value for a directional test with $(n - 2)$ degrees of freedom at $\alpha = 0.05$ is found under $\alpha = 0.10$.

Special circumstances sometimes arise in which one wishes to test whether ρ is equal to some value other than zero, such as $\rho = 0.90$ or $\rho = -0.75$. Such testing requires a conversion of r to z (discussed in Section 14.4.2). The interested reader is referred to discussions in Alder and Roessler (1972 214–215) or Sokal and Rohlf (1969: 519) for the specifics.

Example 14.2

It was determined in Example 14.1 that the correlation between actual birth rate in Latin America and the ideal family size is $r = +0.610$. Does this coefficient indicate that ρ is significantly different from zero?

We must first assume that the seven examples were randomly generated from the Latin American population; if Ehrlich arbitrarily selected the best cases, a significance test is unwarranted. The number of degrees of freedom in this case are equal to $(n - 2) = (7 - 2) = 5$. Table A.11 indicates that the critical value for $\alpha = 0.01$ is $r = 0.8745$. We must conclude that no significant difference exists between the coefficient $r = +0.610$ and $\rho = 0.0$ at the 0.01 level.

Note: Here is a case in which a strict, insensitive dependence upon conventional statistical levels could be misleading. If more cases had been employed in the study, the results almost surely would have been significant. But even granting the small sample size and the lack of significance, one should probably not ignore so large a value as $r = +0.610$, even though it falls short of the tabled value. Ehrlich's example simply has too few cases to demonstrate a correlation which is probabilistically significant.

14.4.2 Confidence Limits of *r*

The statistical confidence limits about an observed correlation coefficient are sometimes more useful than testing for significance against specific values of ρ. Finding the interval about r is complicated by the fact that ρ must be known prior to finding the standard error of r. This unrealistic procedure led Sir Ronald Fisher to derive a second index of correlation, known as Z.[2]

$$Z = \frac{1}{2}\log_e \frac{1+r}{1-r} \tag{14.5}$$

where $\log_e$ is the natural logarithm based upon the constant base $e = 2.718$. As before, this conversion need not be accomplished every time because the values of Table A.12 provide ready access to the conversion of r to Z, so the computations in (14.5) can usually be avoided.

The sampling distribution of Z is known to be approximately normal, with a standard error approximated by

$$\sigma_Z \simeq S_Z = \frac{1}{\sqrt{n-3}} \tag{14.6}$$

How these two quantities are used to determine confidence limits about r is illustrated by a worldwide study of cultural patterns of child rearing made by Barry and Paxson (1971). They reported that the correlation between general indulgence during childhood and the use of carrying devices for infants (such as cradleboards) to be $r = +0.65$ for a sample of 42 societies. What are the 95 percent limits for ρ?

From Table A.12 we find that the observed value of $r = +0.65$ converts to $Z = 0.775$; an identical value of Z is found by using Formula (14.5). The standard error of this Z is found from (14.6):

$$\sigma_Z = \frac{1}{\sqrt{42-3}} = 0.160$$

[2]Once again the statistical terminology conspires against us. Although the *log* conversion of r is denoted by Z, do not confuse this "zee" with z, the symbol used here for the standardized normal deviate. These two "zee's" have nothing in common.

The confidence interval about Z is given by standard methods, using the normal curve:

$$\begin{aligned}\text{confidence interval} &= Z \pm 1.96\sigma_Z \\ &= 0.775 \pm 1.96(0.160) \\ &= 0.775 \pm 0.314\end{aligned}$$

So the 95 percent confidence intervals run from +0.461 to +1.089. But these intervals are still in values of Z and must be converted back to values of r. Again using Table A.12, we find the confidence intervals of r to run between $r = +0.43$ and $r = +0.80$. Clearly, this correlation between carrying devices and the indulgence of infants is positively correlated in the worldwide population. That is, we can be almost certain that ρ lies between $r = +0.43$ and $r = +0.80$. In this case, the confidence interval seems to provide more useful results than would the t-test against $H_o: \rho = 0$.

The question arises: Since the 95 percent confidence interval does not include $\rho = 0$, is this equivalent to rejecting H_o: $\rho = 0$ versus H_1: $\rho \neq 0$ at the 0.05 level, using (14.4)? The answer is yes. This confidence interval tells us that in the long run, we can expect to obtain intervals which will contain the (unknown) ρ about 95 percent of the time. As with all confidence intervals, an entire range of hypotheses has been implicitly tested. All null hypotheses suggesting values of ρ outside the interval +0.43 and +0.80 are implicitly rejected. Null hypotheses with $+0.43 \leq \rho \leq +0.80$ are not rejected.

Note further that the confidence limits are not symmetrical about $r = +0.65$, since the general distribution of ρ produces a diminishing effect upon the confidence limits for the positive values of r. When the correlation coefficient is negative, the lower confidence limits will be closer to r than will the upper limit.

14.4.3 Testing for a Difference between Two ρ

Two independent correlation coefficients can be statistically compared by transforming the r into standardized normal deviates. The raw r must first be converted to Z by using Table A.12. The standard error of the difference between two Z is given by

$$\sigma_{Z_1 - Z_2} = S_{Z_1 - Z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{14.7}$$

The statistical difference between the two Z can be computed as in earlier tests for differences. Another example from population ecology will illustrate the computations.

Paul Ehrlich contends that the families of a DC (developed country) generally come closer to their ideal size than those of the UDC (underdeveloped country). Ehrlich contends that the UDC generally exceed their desired family size, due to socioeconomic, religious, and political factors. Example 14.1 determined that the correlation between number of children wanted and the Latin American birth rate is $r = +0.610$. It is also known that the correlation is $r = +0.818$ for nine European countries. Assuming the European countries to be DC and the Latin American countries to be UDC, is the difference in correlations large enough to support Ehrlich's hypothesis?

Let us term the correlation for Europe as r_1, with $n_1 = 9$. The statistical hypotheses in this one-tailed test are

$$H_0: \rho_1 \leq \rho_2 \qquad H_1: \rho_1 > \rho_2$$

The one-tailed critical value of the standardized normal deviate, z, is found in Table A.3 to be 1.64.

The log conversion values of r are found from Table A.12 to be $Z_1 = 1.157$ and $Z_2 = 0.709$. Let me caution you once again not to confuse the meaning of the standardized normal deviate (z) with the log conversion of the correlation coefficient (Z). The standard error for the difference between the two Z conversions is found from Expression (14.7):

$$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{9-3} + \frac{1}{7-3}} = 0.645$$

The standardized normal deviate for the difference is

$$z = \frac{(Z_1 - Z_2) - 0}{\sigma_{Z_1 - Z_2}} = \frac{1.157 - 0.709}{0.645}$$
$$= 0.695$$

The computed figure for z fails to exceed the critical value of $z = 1.64$, so the null hypothesis cannot be rejected. The census data from Latin America and Europe fail to support Ehrlich's contention that developed countries come closer to their ideal family size than do the underdeveloped countries.

14.5 RANK-ORDER CORRELATION

Measures of statistical correlation always involve *pairs* of observations; each of the pairs represents a bivariate random sample of size n. The correlation coefficient is but one measure of linear correlation, and r is the appropriate measure of affinity between X and Y *only* as long as three criteria are met:

1. Both X and Y are at least interval-scale variables.
2. The distribution of Y and X is bivariate normal.
3. Variables X and Y are related in linear fashion.

Use of r becomes suspect when any of these conditions is not met, and this section discusses two important nonparametric alternatives to the parametric correlation coefficient. Specifically, these nonparametric alternatives apply when conditions (1) and/or (2) are not met; the nonparametric methods still assume a linear relationship. As we will see, the nonparametric methods of correlation are particularly helpful in analyzing cross-cultural samples which are so common in today's ethnology.

14.5.1 Spearman's Rank-Order Correlation Coefficient

Data from the *Ethnographic Atlas* (Murdock 1967) have been used from time to time to illustrate various of the statistical techniques. Another source of easily retrievable ethnographic data is Human Relations Area Files (HRAF) with central headquarters in New Haven, Connecticut; the HRAF records are made available on a subscription basis to over two hundred universities and museums throughout the world.

The physical process of coding requires the analyst to decide whether or not a given trait is present within a society. Sometimes these decisions are quite easy to code:

(Col. 1) *Regional Identity*:
- Africa
- Circum-Mediterranean
- East Eurasian
- Insular Pacific
- North America
- South America

(Col. 39) *Type of Animal Husbandry*:
- bovine animals
- camels
- deer
- equine
- pigs
- other

But many interesting cultural variables are by their nature judgmental in character, such as *degree of anxiety, kind of family organization, intensity of agriculture* or *frequency of warfare*. Variables of this sort require the coder to make rather subjective decisions, decisions which tend to vary between analysts. A common control in coding ethnographic data is to employ multiple judges, who are unaware of the hypothesis and each of whom independently codes the same literature. The scores can then be compared to determine the relative objectivity of the categories.

In one such study, Bacon and others (1965) attempted a rather ambitious study of drinking behavior throughout the world. Ethnographic literature was assembled for a large sample of societies, and then raters independently coded these data into comparable cross-cultural categories. One variable under investigation in this study was *hostility and resentment of males while drinking*, which was categorized into the following divisions (Bacon and others 1965: 340):

A. Little or no expressions of resentment
B. Verbal expression of mild resentment such as slight impoliteness
C. Moderate quarreling
D. Serious quarreling
E. Quarreling frequently accompanied by physical fighting
F. Serious physical combat
G. Physical combat involving frequent injury to other persons

Every society in the sample was then independently scored by two investigators (Barry and Buchwald). When their ratings agreed, then the scale was considered to be relatively objective and hence acceptable. But if the judges disagreed on a number of cases, then they knew that the coding scheme lacked the necessary objectivity and required redefinition.

Suppose that the two judges obtained the following codes for the *hostility and resentment* variable:

	Judges	
Society	1st	2d
Ainu	B	B
Cayapo	D	D
Chukchee	G	G
Cuna	E	E
Ifugao	F	F
Maori	A	C

The two judges agreed in all cases except the New Zealand Maori. Is this single deviation to be expected by chance, or does there appear to be an inordinate amount of disagreement on the *hostility and resentment* scale?

The *Spearman Rank-Order Correlation Coefficient*, designated by r_S, is an index derived to analyze exactly this sort of situation. Originally defined in 1904, this measure is the earliest of the family of nonparametric statistics based upon ordinal ranking.[3] The statistic r_S compares the overall similarity of two ordinal rankings. The two judges' rankings can be considered as rank orderings: Each society is ranked relative to the others in terms of hostility and aggression.

Spearman's Rank-Order Correlation Coefficient is defined as

$$r_s = 1 - \frac{6\Sigma d_i^2}{n^3 - n} \tag{14.8}$$

where d_i is the raw difference between rankings of variate pair i, and n is the total number of such pairings.[4] Like r, Spearman's r_S ranges from +1.0 for a perfect positive correlation to −1.0 for absolute negative correlation.

To compute r_S, the societies must first be placed in rank-order for each scale. Judge A rated the six societies in the following order: (1) Maori, (2) Ainu, (3) Cayapo, (4) Cuna, (5) Ifugao, and (6) Chukchee. Judge B's results are similar, except that the Maori and Ainu are placed in reverse order, with Ainu receiving the rank of 1. The d_i are then found by subtracting rankings of judges. These scales are presented in Table 14.3. The sum of the deviations must always equal zero, a fact which provides a handy check for errors in either adding or subtracting the deviations. The d_i are then squared and summed, providing Σd_i^2. Formula (14.8) can now be applied to the results of Table 14.3:

$$r_S = 1 - \frac{6(2)}{6^3 - 6} = 1 - \frac{12}{210}$$
$$= +0.94$$

[3]Spearman's index is sometimes designated as rho, but the simpler r_S is used here to avoid confusion with the population parameter of the parametric correlation coefficient ρ.

[4]The derivation of r_S can be found in Siegel (1956: 203–204).

	Judge's Rankings			
Society	A	B	d_i	d_i^2
Maori	1	2	−1	1
Ainu	2	1	+1	1
Cayapo	3	3	0	0
Cuna	4	4	0	0
Ifugao	5	5	0	0
Chukchee	6	6	0	0
				$\Sigma d_i^2 = 2$

Spearman's r_S is interpreted in a manner similar to r, so we see that the two ranking scales are indeed quite closely correlated. A second, more complicated example will further illustrate the versatility of r_S.

Villages of northern India generally host representatives from some 5 to 25 endogamous social groupings known as *castes*. Villagers regard each caste as higher (or lower) than another in terms of prestige and esteem. The result is a tightly structured social hierarchy. But a certain amount of disagreement generally exists as to the exact social ranking of particular castes. Stanley Freed, of the American Museum of Natural History, collected an interesting series of data from the small village of Shanti Nagar, northern India (Freed 1963). Freed interviewed a series of randomly selected male informants. Each informant was given a set of movable cards, upon each of which was written the name of a single caste. Informants were requested to arrange the caste cards in their appropriate social ordering. One informant, of the *Brahman* (priest) caste, arranged his cards into the following order: *Brahman* (priest), *Baniya* (merchant), *Jat* (farmer), *Baigari* (beggar), *Mali* (gardener), *Gola Kumhar* (potter), *Lohar* (blacksmith), *Jhinvar* (water carrier), *Maher Kumhar* (potter), *Nai* (barber), *Chamar* (leather worker), *Chuhra* (sweeper). A second informant, a member of the sweeper (*Chuhra*) caste, produced the following social ranking using the same deck of cards: *Brahman*, *Baniya*, *Jat*, *Jhinvar*, *Lohar*, *Mali*, *Bairagi*, *Nai*, *Gola Kumhar*, *Mahar Kumhar*, *Chamar*, *Chuhra*. The two rankings clearly contain many similarities—both place *Brahman* at the highest end of the hierarchy and the *Chuhra* at the bottom, for example, but the order of some intermediate ranks differs. How similar are the two social rankings?

This is obviously a problem in correlation: How closely does the *Brahman*'s ordering correlate with that of the sweeper? The standard correlation coefficient is irrelevant in this context because the caste rankings achieve only ordinal status. But Spearman's coefficient is readily applicable.

The data must first be assigned numerical rank orderings. The *Brahman*'s sequence has been numbered in order from the highest (1) to the lowest caste (12) on Table 14.4. The *Chuhra*'s ordering was then assigned the numbers according to the first sequence. Identical results will result if the *Chuhra*'s ordering is used as the first reference sequence; the *Brahman*'s ordering was arbitrarily selected. In this manner, the two rank orders can be compared simply by subtracting the two columns, with the absolute result tabulated in the difference (d_i) column. The differences are then squared and summed as before, and Σd_i^2 is substituted into Formula (14.8):

$$r_S = 1 - \frac{6(44)}{12^3 - 12} = +0.85$$

The rather high value of 0.85 indicates a very large degree of correspondence between the two orderings. Despite the fact that informants came from the very extremes of the caste spectrum, both agreed on their relative social standing. Freed went on to use r_S to compare the pairwise results for 23 other randomly selected informants to produce a median ranking scale for an entire village (Freed 1963: table 4). In this manner, an objective means of determining caste ranking was devised.

TABLE 14.4

Social Rankings					
Brahman		Chuhra		d_i	d_i^2
1	*Brahman*	1	*Brahman*	0	0
2	*Baniya*	2	*Baniya*	0	0
3	*Jat*	3	*Jat*	0	0
4	*Baigari*	8	*Jhinvar*	−4	16
5	*Mali*	7	*Lohar*	−2	4
6	*Gola Kumhar*	5	*Mali*	1	1
7	*Lohar*	4	*Baigari*	3	9
8	*Jhinvar*	10	*Nai*	−2	4
9	*Mohar Kumhar*	6	*Gola Kumhar*	3	9
10	*Nai*	9	*Mahar Kumhar*	1	1
11	*Chamar*	11	*Chamar*	0	0
12	*Chuhra*	12	*Chuhra*	0	0
				0	44

The case of Indian social castes is particularly useful to illustrate just how r_S functions, since castes are a perfect example of rank orderings which occur in social contexts. Suppose that two informants from Shanti Nagar produced exactly identical orderings as shown in Table 14.5.

Because the informants have agreed, the Σd_i^2 must equal zero, producing the following r_S of unity:

$$r_S = 1 - \frac{6(0)}{12^3 - 12} = 1 - \frac{0}{1710}$$
$$= +1.00$$

The opposite case would be an unlikely situation in which informants produce exactly reverse orderings, as in Table 14.6.

TABLE 14.5.

Social Rankings			
A	B	d_i	d_i^2
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	0
6	6	0	0
7	7	0	0
8	8	0	0
9	9	0	0
10	10	0	0
11	11	0	0
12	12	0	0
		0	0

TABLE 14.6

Social Rankings			
A	B	d_i	d_i^2
1	12	−11	121
2	11	− 9	81
3	10	− 7	49
4	9	− 5	25
5	8	− 3	9
6	7	− 1	1
7	6	1	1
8	5	3	9
9	4	5	25
10	3	7	49
11	2	9	81
12	1	11	121
		0	572

Spearman's coefficient for perfect disagreement is found to be

$$r_s = 1 - \frac{6(572)}{1716} = 1 - \frac{3432}{1716} = 1 - 2.00$$
$$= -1.00$$

Thus, the theoretically maximum value of $\Sigma d_i^2 = 572$ produces a perfectly correlated coefficient of $r_s = -1.00$. Of course no two informants could be expected to disagree in such extreme fashion.

Testing Spearman's r_S for statistical significance. Many applications of correlation require little more than relative measures of covariation (or the lack of it):

1. Do two *Brahman* informants tend to agree on caste ranking more closely than a *Brahman* and a *Chuhra*?
2. Does method A of coding ethnographic data produce more agreement among coders than does method B?

Answers to inquiries such as this can be directly provided by the r_S coefficients.

But cases sometimes arise in which one must generalize beyond the samples at hand to larger populations. The null hypothesis is that the variables lack true association, and only by chance has r_S deviated from zero. Two procedures are presented below which allow us to apply the hypothesis-testing procedures to r_S. But before considering these techniques, be certain to recognize one important sampling stricture. As long as the r_S coefficient is used strictly as *description*, then there are no restrictions on sampling. But if the r_S is interpreted to infer population characteristics from incomplete samples, then sampling must be random. That is, *addition of a test for statistical significance presupposes randomly generated variates from a specific population.*

The sampling distribution of r_S has been compiled in Table A.13.[5] These critical values refer to the one-tailed case in which the direction of association has been clearly specified. Positive values of r_S predict that large X should be paired with large Y, while a negative relationship pairs large X with small Y.

Table A.13 makes quick work of assessing the statistical significance of r_S. Spearman's rank-order coefficient was computed in the earlier cross-cultural drinking study to be $r_S = +0.94$. The critical value of r_S at $\alpha = 0.05$ with $n = 6$ is found from Table A.13 to be $r_S = 0.829$. We are justified in rejecting the null hypothesis in this case because the observed correlation is more extreme than this critical figure. We conclude that the two judges do not differ significantly in their scaling of drinking behavior in these six societies. Similarly, Table A.13 indicates that the correlation between the two Indian informants is significant beyond the 0.01 level. Positive associations were expected in both cases, so each test is one-tailed.

A second method for testing the significance of r_S is available when ten or more pairs are involved. The r_S distribution approaches normality as n increases, and the following approximation holds when $n \geq 10$:

$$t = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \tag{14.9}$$

with $(n-2)$ degrees of freedom.

The caste data in the last section can be used as an example of the normal approximation of r_S:

$$t = \frac{0.85\sqrt{12-2}}{\sqrt{1-0.85^2}} = \frac{2.69}{0.5260}$$
$$= 5.11$$

[5]See Siegel (1956: 210–211) for a derivation of the r_S distribution.

-fectly
ıld be

A one-tailed test is involved, so at the 0.01 level with $(12-2)=10$ degrees of freedom, Table A.4 indicates that the critical value is $t_{0.02} = 2.764$. The observed statistic for r_S far exceeds this critical value, and the result is declared to be statistically significant. The correlation between *Brahman* and *Chuhra* informants is significantly distinct from zero, so the null hypothesis is rejected. Note that this significance test should not be attempted if the informants had not been selected in some random manner.

r_S with tied observations. Spearman's rank-order coefficient compares two ordinal scales to determine the degree of similarity between rankings. The computations assumed an underlying continuous scale for both variates, so ties should not occur between the observations. But in practice, ties are known to occur rather frequently within anthropological scaling, and it becomes necessary to correct the computations of r_S to compensate for tied ranks.

The presence of ties tends to lower the computed value of r_S. A correction factor for tied variates is

$$T = \frac{t^3 - t}{12} \tag{14.10}$$

If three scores were tied at the same rank, for example, then the correction is $T = 3^3 - 3/12 = 2$. This computation is to be performed for all sets of tied variates within the X variable. The quantity ΣT_X is the sum of the corrections on the X variable. Similar computations are performed on the Y ranking, to produce ΣT_Y.[6]

The following formula should be used for r_S whenever ties are present:

$$r_S = \frac{\Sigma X^2 + \Sigma Y^2 - \Sigma d^2}{2\sqrt{\Sigma X^2 \Sigma Y^2}} \tag{14.11}$$

where

$$\Sigma X^2 = \frac{n^3 - n}{12} - \Sigma T_X \qquad \text{and} \qquad \Sigma Y^2 = \frac{n^3 - n}{12} - \Sigma T_Y$$

This correction for ties admittedly complicates the computations somewhat, but many of the correction factors turn out to be redundant; once computed, these terms often recur in the same formula, thereby simplifying the arithmetic. An example should clarify the computation of r_S with ties.

Few students of anthropology would question that economic and political development are functionally related in modern industrial societies. But the cognate notion that this relation holds for *nonindustrial* societies has been the subject of long debate in anthropology. Melvin Ember attempted to test the relationship between political and economic factors in a cross-cultural analysis (Ember 1963). A random sample of 24 societies was drawn from the 565 contemporary and historical cultures in the "World Ethnographic Sample" (see Table 14.7). Each society was then ranked according to its relative economic and political development. Economic specialization is known to be quite closely correlated with the maximum community size, so *economic development* was operationally defined as the "upper limit of community size." Ember defined

[6]Confusion sometimes arises as to just which tied scores are "corrected." We are concerned here only with ties occurring *within ranks* of each variable. Ties between pairs of variates simply reduce to $d_i = 0$.

political authority as "the number of different political officials who participate in all levels of government" (for example, clan chief and head of an extended family). Table 14.7 contains the rankings of these economic and political indicators for the 24 sample societies. Can we say that economic and political development are positively correlated in nonindustrial societies?

TABLE 14.7 Relationship between upper limit of community size and differentiation of political authority (data from Ember 1963: table 5)

Rank Order of Societies	Community Size	Political Authority	d_i	d_i^2
Kaska	1	6	−5	25.00
Caribou Eskimo	2	2.5	−0.5	0.25
Kutubu	3	10	−7	49.00
Xam	5	2.5	2.5	6.25
Naron	5	6	−1	1.00
Mataco	5	6	−1	1.00
Tiwi	8.5	2.5	6	36.00
Ojibwa	8.5	10	−1.5	2.25
Bacairi	8.5	10	−1.5	2.25
Acholi	8.5	17	−8.5	72.25
Guahibo	11.5	2.5	9.0	81.00
Timucua	11.5	15	−3.5	12.25
Ontong Java	13	15	−2	4.00
Chamarro	15	10	5	25.00
Lango	15	10	5	25.00
Samoa	15	18.5	−3.5	12.25
Cuna	17	21	−4	16.00
Omaha	19	20	−1	1.00
Teton	19	13	6	36.00
Didinga	19	18.5	0.5	0.25
Huron	21	15	6	36.00
Tswana	22.5	22	0.5	0.25
Ashanti	22.5	23	−0.5	0.25
Thai	24	24	0	0.00
			0.0	444.50

As before, Spearman's coefficient judges the relationship between these two ordinal scales, but the computations differ somewhat from previous cases because of the tied scores. The sum of squared deviations is found as before; then the correction for ties must be applied. The community size rankings are tied into six sets: The score 8.5 occurs four times, the scores 5.0, 15.0, and 19 each occur three times, and both 11.5 and 22.5 occur twice each.

$$\Sigma X^2 = \frac{n^3 - n}{12} - \Sigma T_x$$

$$= \frac{24^3 - 24}{12} - \left[\left(\frac{4^3 - 4}{12}\right) + 3\left(\frac{3^3 - 3}{12}\right) + 2\left(\frac{2^3 - 2}{12}\right)\right]$$

$$= 1150 - 12 = 1138$$

A similar procedure is followed to correct for ties on the political authority scale.

$$\Sigma Y^2 = \frac{24^3 - 24}{12} - \left[2\left(\frac{3^3-3}{12}\right) + \frac{2^3-2}{12} + \frac{4^3-4}{12} + \frac{5^3-5}{12}\right]$$

$$= 1152 - \left[2\left(\frac{3^3-3}{12}\right) + \frac{2^3-2}{12} + 3\left(\frac{3^3-3}{12}\right) + \frac{4^3-4}{12} + \frac{5^3-5}{12}\right]$$

$$= 1152 - (4.0 + 0.5 + 5.0 + 10.0) = 1150 - 19.5$$

$$= 1130.5$$

These figures can now be substituted into computing Formula (14.11) for when ties are present.

$$r_s = \frac{1138 + 1130.5 - 444.5}{2\sqrt{1138(1130.5)}} = 0.80$$

Using the normal approximation to Spearman's rank-order coefficient, the value of t is found to be

$$t = \frac{0.80\sqrt{22}}{\sqrt{1 - 0.80^2}} = 6.25$$

This figure is highly significant at $df = (n - 2) = 22$. The conclusion is that Ember's random sample strongly supports the hypothesis that economic and political development are positively associated in nonindustrial societies.

Had we failed to correct for the ties, the computed value of r_s would have been

$$r_S = 1 - \frac{6(444.5)}{24^3 - 24} = 1 - \frac{2{,}667}{13{,}800}$$

$$= 0.81$$

Although the difference proves slight in this case, an inordinate number of ties can cause r_S to seriously overestimate the actual correlation if the correction for ties has not been applied.

Example 14.3

Grammatical sex gender is known to correlate with a number of semantic categories which include Freudian sexual symbols, metaphorical extension, and sex role attributes such as beauty and masculinity. But such findings generally relate to a single language, or a few norms over several languages. Robert Munroe and Ruth Munroe have attempted to generalize these findings by examining the underlying relationships between sexual grammar and social structural factors in a cross-cultural study (Munroe and Munroe 1969). A sample of nine languages was selected from those discussed in the *Ethnographic Atlas* and then coded for (1) structural bias toward sex and (2) the prevalence of male-gendered nouns. Male cultural bias was considered to be present under *any one* of the following conditions: patrilocal residence (col. 16 in the *Atlas*), patrilineal kin groups (col. 20 in the *Atlas*), or patri-inheritance (cols. 74 and 76 in the *Atlas*). Each item present is assigned one positive point, and each item rated for "female bias present" receives one negative. Thus, a scale has been devised which ranges from +3 for strong structural bias toward males to −3 for strong structural bias for females; zero indicates the lack of structural sexual bias. The frequency of male and/or female nouns was expressed as "percent of male gender nouns" and words assigned to neuter gender were not recorded. The data for these societies are presented below.

Does this sample support the hypothesis that societies with a tendency toward male bias in social structure also manifest a bias toward male-gendered nouns in their grammar?

Society	Structural Sex Bias raw	Structural Sex Bias rank	Male-Gender Nouns raw, %	Male-Gender Nouns rank	d_i	d_i^2
Lebanese (Arabic)	+3	7.5	64	9	−1.5	2.25
Kanawa (Hausa)	+3	7.5	63	7	0.5	0.25
Nama Hottentots	+3	7.5	63	7	0.5	0.25
Gujarati	+3	7.5	52	5	2.5	6.25
Irish	+2	5	63	7	−2.0	4.00
French Canadians	+1	3	48	4	−1.0	1.00
Byelorussians	+1	3	44	2	1.0	1.00
Greeks	+1	3	35	1	2.0	4.00
Dutch	0	1	45	3	−2.0	4.00
					0.0	23.00

The correction factors for ties are first computed.

$$\Sigma X^2 = \frac{9^3-9}{12} - \left[\left(\frac{4^3-4}{12}\right) + \left(\frac{3^3-3}{12}\right)\right]$$
$$= 60 - (5+2)$$
$$= 53$$

Spearman's coefficient is computed to be

$$r_s = \frac{53 + 54 - 23}{2\sqrt{53(58)}} = +0.79$$

For $n = 9$, this value of r_s is found to be significant at the 0.01 level.

Note that had the correction for ties *not* been applied, the following value would have been obtained:

$$r_s = 1 - \frac{6(23)}{9^3 - 9} = 1 - \left(\frac{138}{720}\right)$$
$$= +0.81$$

This uncorrected value is also significant at beyond the 0.01 level.

14.5.2 Kendall's Tau

Spearman's coefficient of rank-order correlation compares two ordinal rankings in terms of their relative association; r_S is based upon the magnitude of the squared differences between the ranks. The same sets of data can be viewed from a rather different statistical perspective, and a second index of rank-order correlation emerges.

This new method can be illustrated using Freed's data on caste ranking in northern India. Two informants, a *Brahman* and a *Chuhra*, rated the 12 castes into appropriate orders. Although their overall conception of the social order was quite similar, the specific rankings were by no means identical, and *Kendall's tau* provides a new method for assessing this correlation.

The Kendall's tau statistic, τ, can be computed by either of two rather different methods. The first technique requires pair-by-pair enumeration such as completed in Table 14.8. Note that the *Brahman*'s responses (termed the X_i) have been placed into sequence and assigned ranks as before. The Y_i are also ranked, but each caste in Y_i receives the corresponding rank number from the X ordering. Thus, even though the *Chuhra* ranked the *Jhinvar* caste fourth, *Jhinvar* receives rank 8 to correspond with its placement in the *Brahman* ranking. The second step requires that we determine the exact number of *larger ranks* for every Y_i. Beginning with rank Y_1 (*Brahman*), we find there are precisely 11 larger rankings (Y_2 through Y_{12} are all larger). The rank Y_2 (*Baniya*) has ten larger subsequent ranks, and so forth. The fourth ranking, Y_4 (*Jhinvar*) has an assigned rank of 8, so only four subsequent ranks (10, 9, 11, and 12) are greater. After all ranks on Y have been enumerated, the total counts, ΣC_i, are summed to 56.

Kendall's tau can now be found through the enumeration method. First the *numerator* must be determined by the following formula:

$$\text{numerator} = 4\Sigma C_i - n(n-1) \quad (14.12)$$

TABLE 14.8

	X_i		Y_i	Subsequent Ranks Larger than Y_i	Counts, C_i
1	*Brahman*	1	*Brahman*	2,3,8,7,5,4,10,6,9,11,12	11
2	*Baniya*	2	*Baniya*	3,8,7,5,4,19,6,9,11,12	10
3	*Jat*	3	*Jat*	8,7,5,4,10,6,9,11,12	9
4	*Baigari*	8	*Jhinvar*	10 9,11,12	4
5	*Mali*	7	*Lohar*	10 9,11,12	4
6	*Gola Kumhar*	5	*Mali*	10,6,9,11,12	5
7	*Lohar*	4	*Baigari*	10,6,9,11,12	5
8	*Jhinvar*	10	*Nai*	11,12	2
9	*Mahar Kumhar*	6	*Gola Kumhar*	9,11,12	3
10	*Nai*	9	*Mahar Kumhar*	11,12	2
11	*Chamar*	11	*Chamar*	12	1
12	*Chuhra*	12	*Chuhra*		0
					$\Sigma C_i = 56$

where n is the number of variate pairs. In the example,

$$\text{numerator} = 4(56) - 12(11) = 92$$

The entire Kendall's tau statistic is defined to be

$$\tau = \frac{\text{numerator}}{\sqrt{[n(n-1) - \Sigma T_X][n(n-1) - \Sigma T_Y]}} \tag{14.13}$$

where the numerator is computed from Expression (14.12). This formula for tau includes an automatic correction term for ties within rankings. For every set of ties within a ranking, the individual correction term is given by $t(t-1)$. If, for example, five variates were tied at a single rank, then $t(t-1) = 5(4) = 20$. The sum ΣT_X represents the total of the $t(t-1)$ corrections for the X scale, and ΣT_Y sums corrections over the Y ranks. When no ties are present, then $\Sigma T_X = \Sigma T_Y = 0$. Use of this correction will be illustrated in the examples to follow.[7]

The value of tau in the caste stratification example is computed from Formulas (14.12) and (14.13) to be

$$\tau = \frac{92}{\sqrt{[12(11) - 0][12(11) - 0]}} = \frac{92}{132}$$
$$= +0.697$$

Because the maximum of tau is ± 1.00, this measure clearly indicates that the agreement is close between the two informants. We know that the same data produced a value of $r_s = +0.85$ in the last section, so it becomes clear that τ and r_s measure somewhat different conceptions of correlation. Tau has certain intrinsic advantages over r_s as a measure of correlation, but τ unfortunately involves a bit of rather tedious computation.

Some of this computational burden is alleviated by using a *graphical solution* for finding the numerator of tau. The initial step in the graphic method is to list the sets of ranks as before. Then like ranks are connected by straight lines, as in Fig. 14.5. The number of intersections of these lines, termed ΣI, are then counted. If the ranks were ordered in identical fashion, then all connecting lines would be exactly parallel, with no intersections occurring at all. The sum ΣI increases as the rankings become more dissimilar. The two caste rankings produce lines which cross a total of $\Sigma I = 10$ times. This sum can be converted to the numerator of tau by the following expression:

$$\text{numerator} = n(n-1) - 4\Sigma I - \Sigma T_Y \tag{14.14}$$

where ΣI is the number of crossings on the graph and ΣT_Y is the sum of the $t(t-1)$ factors for the second (Y) ranking. For the caste rankings on Table 14.8,

$$\text{numerator} = 12(11) - 4(10) - 0 = 132 - 40 = 92$$

[7]It becomes necessary to confess once again that the statistical notation lacks consistency. Throughout this book, Greek letters have generally denoted parameters and italic letters refer to statistics. But the system has broken down: Kendall's *tau* is clearly a statistic and not a parameter, but the statistic is called τ. The problem is obvious, of course, since t has already been assigned to Student's test. But the distinction between statistic and parameter should be kept firmly in mind by this point, so the devious terminology ought not cause undue conceptual difficulty.

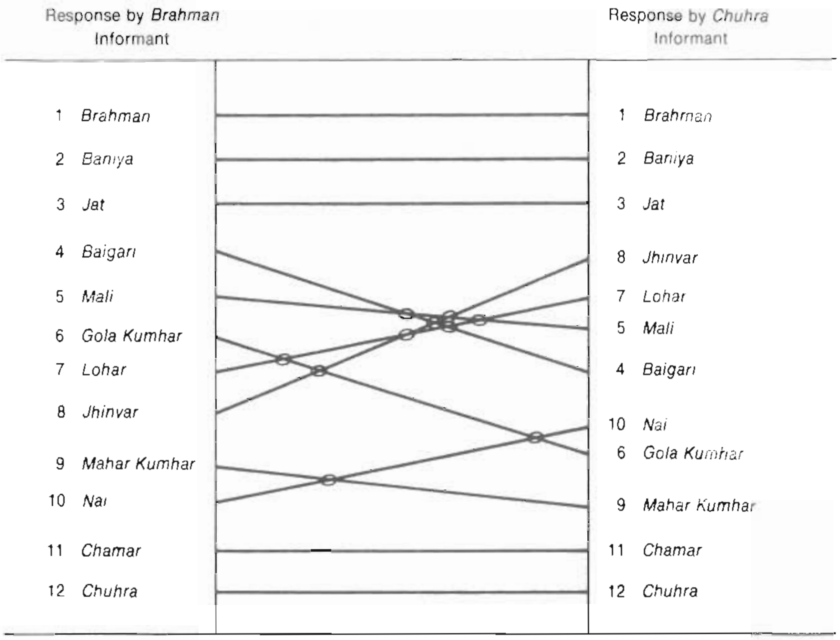


Fig. 14.5 Graphical solution for Kendall's τ.

This same numerator was computed by the enumeration methods, so the resulting values of τ are identical. Whenever ties occur, the lines should be drawn such that lines for tied variates do not cross. The graphical solution generally works better when only a moderate number of ties occurs.

The statistical significance of Kendall's τ can be tested in two ways, assuming that the samples were randomly generated. Whenever $n \leq 10$, the two critical values of the numerator of τ are given as follows (after Sokal and Rohlf 1969: 537):

	Numerator	
n	$\alpha = 0.05$	$\alpha = 0.01$
5	20	—
6	26	30
7	30	38
8	36	44
9	40	52
10	46	58

These critical values are exact only when no ties occur. A table of small n corrected for ties can be found in Burr (1960).

If the sample size exceeds 10, then the distribution of Kendall's tau can be approximated by a normal distribution. The null hypothesis of $H_o: \tau = 0$ can be tested by using a form of the standardized normal deviate:

$$z = \frac{\tau_{\text{observed}} - 0}{\sqrt{2(2n+5)/9n(n-1)}} \qquad (14.15)$$

The probability of such large differences between the observed tau and the null value of $\tau = 0$ can be found as usual from Table A.3.

The normal approximation method indicates that for the caste example,

$$z = \frac{0.697}{\sqrt{2(24+5)/108(11)}} = \frac{0.697}{\sqrt{0.0488}}$$
$$= 3.154$$

Table A.3 indicates a value of only $A = 0.0008$ corresponding to so small a z, so the probability that the obtained τ would deviate from $\tau = 0$ (in either direction) by chance alone is only $p = 0.0016$. The null hypothesis is rejected, and we conclude that the *Brahman* and *Chuhra* indeed rank the castes in a statistically indistinguishable manner.

An exact test of significance for Kendall's tau is also given by Naroll (1974), but the computations are so tedious as to require a computer for any large-scale application.

Example 14.4

Let us test the hypothesis that the more a society depends upon hunting, the more nomadic (that is, the less sedentary) will be that society. A sample of seven societies was randomly selected from the *Ethnographic Atlas*, and these societies were coded for the *dependence upon hunting* variable (col. 8) and the *settlement pattern* variable (col. 30).

Society	Hunting, percent dependence	Settlement Pattern
Copper Eskimo	36–45	Seminomadic communities
Djafun	0–5	Nomadic bands
Fox	36-45	Seminomadic communities
Gros Ventre	76-85	Nomadic bands
Makin	6-15	Complex settlements
Shasta	26-35	Semisedentary communities
Wishram	16-25	Semisedentary communities

Do these data support the hypothesis?

Kendall's tau coefficient is useful in this case, and τ will be computed by both methods. To find τ by enumeration, the variates must first be ranked in descending order of X (hunting), and then the ΣC_i can be computed.

Society	Hunting Rank	Settlement Rank	Subsequent Ranks	Total
Gros Ventre	1	1.5	3.5,3.5,5.5,5.5,7,(1.5)	5.5
Fox	2.5	3.5	(3.5),5.5,5.5,7	3.5
Copper Eskimo	2.5	3.5	5.5,5.5,7	3
Shasta	4	5.5	(5.5),7	1.5
Wishram	5.5	5.5	7	1
Makin	5.5	7		0
Djafun	7	1.5		0
				$\Sigma C_i = 14.5$

Note that the score "0.5" has been added to the C_i if the subsequent ranking is tied with the reference rank; these cases are enclosed in parentheses:

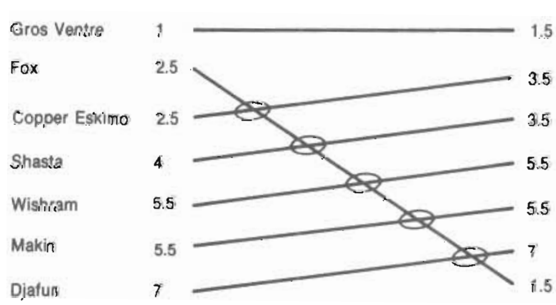
$$\text{numerator} = 4(14.5) - 7(6) = 58 - 42 = 16$$

Kendall's tau is thus

$$\tau = \frac{\text{numerator}}{\sqrt{[42-4][42-6]}} = \frac{16}{\sqrt{38(36)}}$$
$$= 0.433$$

Because n is less than 10, the tabled values indicate that the numerator (16) does not reach the critical value of 30; this sample has insufficient evidence to cause rejection of the null hypothesis. There is no reason to suspect a correlation between hunting and settlement pattern, based upon this limited sample.

The numerator of τ can also be found using the graphic method:



There are a total of five crossings, so the

$$\text{numerator} = 7(6) - 4(5) - 6 = 16$$

This is the same value of the numerator as found earlier, so the same value of tau must result.

Example 14.5

Spearman's r_S was applied earlier to Ember's cross-cultural study of community size and political authority (Table 14.7). Kendall's tau can also determine the relationship between these two variables.

Rank Order of Societies	Community Size	Political Authority	Sum of Larger Subsequent Ranks
Kaska	1	6	18
Caribou Eskimo	2	2.5	20.5
Kutubu	3	10	14
Xam	5	2.5	19
Naron	5	6	16.5
Mataco	5	6	16
Tiwi	8.5	2.5	16.5
Ojibwa	8.5	10	13.5
Bacairi	8.5	10	13
Acholi	8.5	17	7
Guahibo	11.5	2.5	13
Timucua	11.5	15	8
Ontong Java	13	15	7.5
Chamarro	15	10	9.5
Lango	15	10	9
Samoa	15	18.5	5.5
Cuna	17	21	3
Omaha	19	20	3
Teton	19	13	5
Didinga	19	18.5	3
Huron	21	15	3
Tswana	22.5	22	2
Ashanti	22.5	23	1
Thai	24	24	0
			$\Sigma C_i = 226.5$

The following ties exist on the X variable: Rank 5 (tied three times), rank 8.5 (tied four times), rank 11.5 (tied twice), rank 15 (tied three times), rank 19 (tied three times), and rank 22.5 (tied twice). The correction for ties on X is

$$T_X = 3(3-1) + 4(4-1) + 2(2-1) + 3(3-1) + 3(3-1) + 2(2-1)$$
$$= 34$$

Be certain to note here that t represents the *number of ties* in each rank rather than the *value* of the tied ranks. The correction for ties on the Y variable is

$$T_Y = 4(4-1) + 3(3-1) + 5(5-1) + 3(3-1) + 2(2-1)$$
$$= 46$$

The numerator of τ can then be computed:

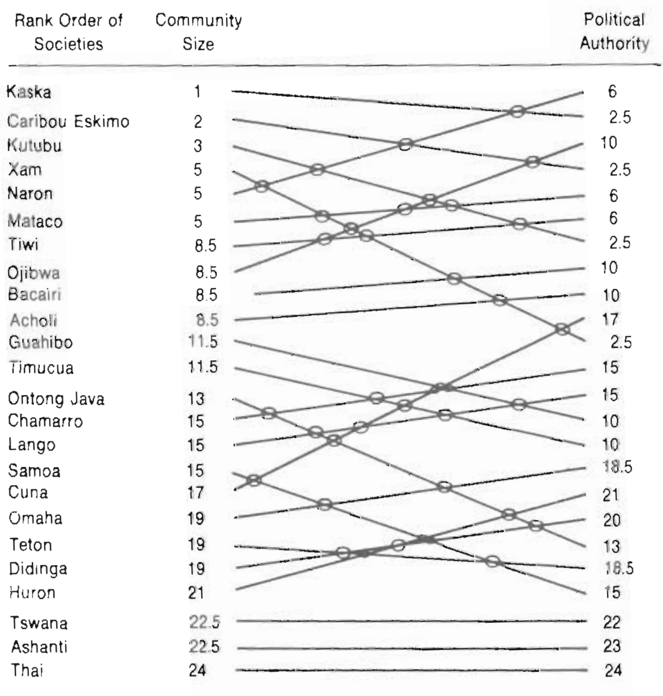
$$\text{numerator} = 4(226.5) - 24(23) = 906 - 552$$
$$= 354$$

Tau is then computed to be

$$\tau = \frac{354}{\sqrt{[24(23) - 34][24(23) - 46]}}$$
$$= 0.69$$

The graphical method can also be used to determine the numerator of tau. In the diagram below, exactly 38 crossings occur. Thus,

$$\text{numerator} = 4(226.5) - 24(23) = 906 - 552$$
$$= 354$$



Thus, the same numerator is found by both methods and of course the values of tau are identical.

Formula (14.15) can be used to test the significance of this value of tau:

$$z = \frac{0.69}{\sqrt{2(2 \cdot 24 + 5)/9 \cdot 24(23)}}$$
$$= 4.72$$

Table A.3 indicates that this result is clearly significant.

14.5.3 Comparison of Spearman's r_s and Kendall's τ

τ and r_S have been computed for the same data, and the values of the two nonparametric correlation coefficients were found to differ. The reason for this is that the two coefficients are based upon rather different underlying models, and hence they do not measure "correlation" in exactly the same manner. But it turns out that r_S and τ utilize exactly the same amount of information, so they have identical *power*. This means that tests of significance based upon either τ or r_S will reject a false H_o at exactly the same level of statistical significance (see Siegel 1956:chapter 9). But several properties of τ have led to a general preference of Kendall's τ over r_S.

Probably the chief advantage of Kendall's statistic is that the distribution of τ approaches the normal more rapidly than does r_S. In fact, the distribution of τ is virtually identical to the normal in samples as small as $n = 9$. So τ seems generally more accurate than r_S in testing for statistical independence between ranked variables, especially when small or moderately sized samples are involved.

The direct and simple interpretation of τ also renders Kendall's tau generally more suitable for use in anthropology. The probability of any value of τ is defined simply in terms of concordant and discordant pairs which, while sometimes tedious to compute, present few conceptual difficulties. But the Spearman's r_S is based upon the squared sums of differences and becomes meaningful only through tortuous analogy with the parametric correlation coefficient r.

Kendall's τ seems also to produce a more meaningful result when a large number of ties are present. As would be expected, Spearman's r_S rather closely follows the Pearson Correlation Coefficient when the underlying distribution is more-or-less continuous, that is, when relatively few ties occur in ranking. But τ is often more accurate when a large number of cases must be classified into a relatively few ordinal classes.

τ has the final advantage over r_s in that tau can be generalized into a *partial correlation coefficient*. This statistic, called $\tau_{XY \cdot Z}$ is particularly useful when observations upon two variables might in fact result from a causal connection with a third related variable. As Siegel (1956) has pointed out, a strong correlation of stature and vocabulary among school children might well be due to an important interrelationship with a third variable, such as age. Kendall's partial correlation is a close relative of Kendall's τ, and these statistics can be helpful in sorting out a number of related variables (see Siegel 1956: 223–229 and Conover 1971:253–255 for a discussion of the techniques of partial correlation).

All of these reasons seem to have convinced anthropologists to rely more upon τ than upon r_S. The main disadvantage of Kendall's tau is, of course, the somewhat tedious computations required whenever n is large. The graphic solution is of some help, especially when few ties are present. But an even greater boon has been the recent availability of computer programs to compute both r_S and τ with little effort (see, for example, the NONPAR CORR program available in the SPSS system devised by Nie, Bent, and Hull 1970).

The most important point to remember when working with rank-order correlation is that whichever coefficient is employed, the resulting value is specific only to that particular coefficient. It is quite improper to compute r_s for one data set, Kendall's τ on a second set, and then compare the values to determine relative degree of intercorrelation. For reasons detailed above, the coefficients measure correlation on different scales, and they are expected to produce different results.

14.5.4 Gamma

Regardless of whether τ or r_s is used, the question of excessive ties within ranks can become a serious problem. Although both τ and r_s can be corrected for ties difficulties often arise when computing the cumbersome corrections. The normal approximation also becomes less valid as the number of ties increases. The correlation procedure can be simplified somewhat by grouping the ordinal variates into a few ranked categories. In effect, the data can thus be reduced into two tight ordinal sequences (A and B) within the standard R × C format (Blalock 1972: 421). For example, for case A:

Variable 2	Variable 1			
	Low	Medium	High	Total
High	0	0	30	30
Medium	0	30	0	30
Low	30	0	0	30
Total	30	30	30	90

A chi-square test for independence within an R × C table is commonly applied to analyze the relationship between two variables such as these, but as Naroll (1970c:163) has correctly pointed out, this practice has a major shortcoming. Compare case A with this second example, case B:

Variable 2	Variable 1			
	Low	Medium	High	Total
High	0	30	0	30
Medium	30	0	0	30
Low	0	0	30	30
Total	30	30	30	

The chi-square test will tell us that the two cases are identical. The variables in case A are arranged into a definite monotonic trend: Low variates predict low variates and high variates predict high variates. Case B lacks this notable trend. Unfortunately, a chi-square test for independence is blind to this difference because the test is not sensitive to changes on an ordinal scale. Chi-square must not be used in such cases.

The gamma coefficient has been designed for precisely those cases in which chi-square falters. Gamma (γ) requires that *both scales be ordinal*, and yet the data are grouped into the conventional R × C format; hence, gamma is particularly effective when so many ties occur that neither τ nor r_S can be used in comparing two ordinal rankings. The value of gamma is given by

$$\gamma = \frac{\text{no. of concordant pairs } \textit{minus} \text{ no. of discordant pairs}}{\text{no. of concordant pairs } \textit{plus} \text{ no. of discordant pairs}}$$

$$= \frac{\Sigma F_i - \Sigma D_i}{\Sigma C_i + \Sigma D_i} \tag{14.16}$$

A pair is termed *concordant* if they are ordered into the proper sequence and *discordant* if the sequencing is reversed. The first method used to find Kendall's tau in fact involved counting the number of concordant pairs (ΣC_i). The following (hypothetical) example shows how the pairs can be enumerated to find the gamma coefficient.

A study has been initiated to test the hypothesis that premarital sexual promiscuity is more prevalent among "primitive" than among "civilized" societies. A cross-cultural survey resulted in a random sample of 232 societies, each of which can be rated on (1) the level of sociopolitical complexity and (2) norms of premarital sexual behavior. Is there a relationship between the two variables?

These data are assembled in Table 14.9. Both scales are ordinal, but because the number of ties is excessive, neither τ nor r_S will serve as a suitable indicator of correlation. The chi-square statistic would adequately handle the format of the data, but χ^2 reduces such data to nominal form, thereby ignoring the important ordinal relationship.

To apply *gamma*, it is first necessary to determine the number of concordant pairs within the sample data. The levels of sociopolitical complexity have been ranked into five ordered categories, running from relatively low to very high levels of integration. Similarly, the norms of premarital sexual behavior have been scaled into three categories from high to low promiscuity. The upper left-hand cell ($\text{cell}_{1,1}$) contains six societies—those at the *state* level of complexity with a high degree of premarital sexual promiscuity. All six societies are "ties" in the earlier sense of τ and r_S. Furthermore, all societies listed on the first row ("state") are also tied with respect to the level of integration with $\text{cell}_{1,1}$. That is, all the societies on the first row are tied with respect to sociocultural

TABLE 14.9

Level of Sociocultural Complexity	Premarital Sexual Promiscuity			
	Weakly Prohibited	Prohibited	Strongly Prohibited	Total
State	6	8	19	33
City	7	15	20	42
Town	18	4	2	24
Village	36	18	19	73
Band	52	6	2	60
Total	119	51	62	232

integration, and all the cells in the first column are tied with regard to "high sexual promiscuity." For such tables, societies are termed *concordant* if they rank lower than those grouped in $cell_{1,1}$. Since row 1 and column 1 contain only societies tied with those in $cell_{1,1}$, none of these societies could possibly either agree (concur) or disagree (demur) with the theoretical ranking. But all the societies *down one row* ("city" or below) and *over one column* (the "moderate" and "low" columns) agree with the rankings of $cell_{1,1}$, and hence are termed *concordant societies*. The total number of these cases is given by the sum of the societies in $cell_{2,2}$ (15 societies), plus the total in $cell_{2,3}$ (four societies) and so forth:

$$15 + 20 + 4 + 2 + 18 + 19 + 6 + 2 = 86 \text{ societies}$$

Thus, 86 societies are concordant with the six societies in $cell_{1,1}$. The total number of *concordant pairs* (as distinct from *concordant societies*) is thus

$$6(86) = 516 \text{ pairs}$$

The concordant pairs must then be computed for the remaining cells in a similar manner. Moving to the second row of column 1 ($cell_{2,1}$), the cell contains 7 societies. Again, societies in the first column and second row represent "ties," so the concordant pairs involve only those groups below and to the right of $cell_{2,1}$:

$$7(4 + 2 + 18 + 19 + 6 + 2) = 7(51) = 357 \text{ pairs}$$

This process of enumeration continues down the first row—note that the last row of the first column will produce no concordant pairs—and then onto the second column, and so forth. The total number of *concordant pairs* for Table 14.9 is thus

$$\begin{aligned}\Sigma C_i &= 6(86) + 7(51) + 18(45) + 36(8) + 8(43) \\ &\quad + 15(23) + 4(21) + 18(2) \\ &= 2780 \text{ pairs}\end{aligned}$$

The total number of *discordant pairs* is found in a manner reverse to that described above. A discordant pair matches a selected reference case with all societies known to rank lower on either scale. So, the procedure begins with the upper right-hand cell ($cell_{1,3}$) and proceeds to enumerate *down and to the left*. The total number of pairs discordant with $cell_{1,3}$ are

$$19(15 + 7 + 4 + 18 + 18 + 36 + 6 + 52) = 19(156) \text{ pairs}$$

The next count is obtained from $cell_{2,3}$:

$$20(4 + 18 + 18 + 36 + 6 + 52) = 20(134) = 2680 \text{ pairs}$$

The total number of *discordant pairs* for the entire table is

$$\begin{aligned}\Sigma D_i &= 19(156) + 20(134) + 2(112) + 19(58) \\ &\quad + 8(113) + 15(106) + 4(88) + 18(52) \\ &= 10{,}752 \text{ pairs}\end{aligned}$$

We now have both quantities necessary to compute gamma from (14.16):

$$\gamma = \frac{2780 - 10{,}752}{2780 + 10{,}752} = -0.589$$

This rather strong negative value of gamma moderately supports the hypothesis that premarital sexual promiscuity is correlated with the less advanced levels of sociopolitical organization. That is, a low level of sociocultural complexity is found to occur with high promiscuity, and high complexity predicts a lower level of promiscuity. This is why the value of gamma is negative. Naroll's (1974) exact test of significance applies to gamma as well as Kendall's tau.

Example 14.6

In Chapter 11, Divale's hypothesis for the evolution of matrilocality was discussed (Example 11.9). In addition to predicting the gross change toward matrilocality upon warfare and migration, Divale also predicted a definite cycle of residence patterns, beginning with patrilocality and evolving into uxorilocality.

Do the three *descent* types appear to correlate with Divale's evolutionary sequence of *residence* types (data from Divale 1974)?

	Descent Types			
Residence	Matrilineal	Ambilineal	Patrilineal	Total
Uxorilocal	1	5	1	7
Matrilocal	52	0	0	52
Matrilocal/ avunculocal	5	0	0	5
Avunculocal	50	0	0	50
Avunculocal/ virilocal	7	2	0	9
Virilocal	26	29	21	76
Patrilocal	4	2	542	548
Total	145	38	564	747

Gamma is an appropriate coefficient with which to assess the relationship between these two ordinal pairs. The number of *concordant pairs* is found as follows:

$5(21+542) = 2815$ pairs
$52(2+29+21+2+542) = 52(596) = 30,992$ pairs
$5(2+29+21+2+542) = 5(596) = 2,980$ pairs
$50(2+29+21+2+542) = 50(596) = 29,800$ pairs
$7(29+21+2+542) = 4,158$ pairs
$26(2+542) = 14,144$ pairs
$2(21+542) = 1,126$ pairs
$29(542) = 15,718$ pairs
$1(2+29+21+542) = 596$ pairs

$$\Sigma C_i = 102,329$$

Similarly, the number of *discordant pairs* is

$1(52+5+50+7+2+26+29+4+2) = 177$ pairs
$21(4+2) = 126$ pairs
$5(52+5+50+7+26+4) = 720$ pairs
$2(26+4) = 60$ pairs
$29(4) = 116$ pairs

$$\Sigma D_i = 1199$$

From Formula (14.16), the gamma coefficient is found to be

$$\gamma = \frac{102{,}329 - 1{,}199}{102{,}329 + 1{,}199} = 0.977$$

This strong value of $\gamma = 0.976$ leaves little doubt of a positive association between residence and descent, in the order hypothesized by Divale (1974).

14.6 CORRELATION ON THE NOMINAL SCALE

Probably the most common measure of statistical correlation—perhaps better termed *association*—between nominal variables is the chi-square statistic, considered in detail in Chapter 11. But one major difficulty with chi-square is that its value depends upon the size of sample, n. To see that this is so, examine the following contingency table:

25	35	60
35	25	60
60	60	120

The value of the χ^2 statistic can be readily computed to be 3.32, with a single degree of freedom. Because this value is not significant at the 0.05 level, the null hypothesis of no association would generally not be rejected.

Let us now modify the frequencies slightly by *doubling* the figures in the above contingency table:

50	70	120
70	35	120
120	120	240

A single degree of freedom remains, but the chi-square statistic is now inflated to $\chi^2 = 6.68$. The results are now found to be significant beyond the 0.01 level, and H_0 would be rejected.

What has happened here? Although the actual numbers of the two tables differ—the second is exactly twice the first—the *relationships of the cells to one another remains constant*. The chi-square statistic is a direction function of sample size, and hence chi-square can never be used as an indicator of the *strength* of a relationship. The two tables above are identical in terms of their percentage relationship, and such proportional similarity is often of interest to the social scientist. If one merely reports a chi-square test as (1) significant or (2) not significant, then the analysis of contingency tables would be totally obfuscated.

A number of coefficients have been proposed from time to time to measure the strength of relationship within a contingency table. No attempt will be made to cover the range of such coefficients. Every statistic is designed to fit specific needs, and only a couple of the most important measures of association will be introduced.

14.6.1 The Phi Coefficient

Look carefully again at the two contingency tables presented above. As the cell frequency increased, the chi-square statistic simultaneously increased and eventually made the grade as "statistically significant," even though the relationship between the two variables was absolutely unchanged. In fact, χ^2 increased not only with n, but also exactly in proportion to n; the second value of χ^2 is about twice the first (within a small rounding error). When n doubles, so does χ^2; when n triples, χ^2 also triples. This same difficulty has been encountered a number of times before, as in Chapter 4, when we needed a measure of sample dispersion. You will remember that the sum of squared deviations about the mean, $\Sigma(X_i - \bar{X})^2$ increased in exact proportion to sample size. This difficulty was solved by dividing the squared deviations by n; the *sample variance*, S^2, was the resultant statistic. Because S^2 is an average of the squared deviations, the statistic is independent of sample size. A similar strategy will enable us to "salvage" the chi-square statistic for assessing the strength of nominal associations. The process of simply dividing by n will free the chi-square statistic from the undesirable inflation due to increasing sample size:

$$\frac{\chi^2}{n} = \phi^2 \qquad (14.17)$$

This new expression is known as *phi squared* (pronounced "fee squared"). Here, ϕ^2 is simply a variant of χ^2, which has been rendered independent of sample size. The most commonly encountered form is simply ϕ ("fee"), the square root of Expression (14.17). Because ϕ applies only to the 2×2 contingency format, the following computational formula is helpful:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \qquad (14.18)$$

where a, b, c, and d represent the various cell frequencies defined earlier for Fisher's Exact Test [Formula (11.10)]. Note that ϕ consists merely of the difference between the diagonals ($ad - bc$), divided by the square root of the product of the row and column totals.

Unlike its cousin the chi-square statistic, the phi coefficient has some definite limits: ϕ ranges only between $+1$ and -1. This range readily follows from the definition of ϕ. The following contingency table indicates the extreme of *absolute positive association*:

50	0	50
0	50	50
50	50	100

The phi coefficient is found to be

$$\phi = \frac{(50)(50) - (0)(0)}{\sqrt{(50)(50)(50)(50)}} = \frac{2500}{2500} = +1.00$$

As expected, a perfect positive association produces a value of $\phi = +1.00$. Such situations will occur whenever every instance of variable 1 occurs with every instance of variable 2, and an absence of variable 1 invariably denotes the absence of variable 2.

The opposite extreme arises whenever variables 1 and 2 *never* co-occur (that is, when cells *a* and *d* are empty), which produces the case of *absolute negative association*:

0	50	50
50	0	50
50	50	100

The phi coefficient for this case is

$$\phi = \frac{(0)(0) - (50)(50)}{\sqrt{(50)(50)(50)(50)}} = \frac{-2500}{2500} = -1.00$$

Perfect negative association will always produce $\phi = -1.00$.

Finally, there is the situation in which variables 1 and 2 have no association at all:

25	25	50
25	25	50
50	50	100

$$\phi = \frac{(25)(25) - (25)(25)}{\sqrt{(50)(50)(50)(50)}} = \frac{0}{2500} = 0.00$$

These simple examples provide an intuitive understanding to the meaning of various values of ϕ.

Example 14.7

O'Nell and Selby (1968) have postulated that Zapotec culture allows men more opportunity for escape than females. Community pressure, for instance, effectively bars most females—especially younger women—from using alcohol as an escape from reality. If this notion is correct, then one expects to find relatively more males than females attending socially sanctioned fiestas.

To test this hypothesis, O'Nell and Selby conducted a census at one *cuelga* ("a drunken fiesta lasting from three to five days, customarily celebrated in honor of a person's Saint's day") in the village of Santo Tomas Mazaltepec, a Zapotec pueblo near Oaxaca, Mexico.

Reported attendance at *Cuelgas* in Santo Tomas Mazaltepec: Differential response by sex.

	Attend	Do Not Attend	Total
Men	26	4	30
Women	12	13	25
Total	38	17	55

Does there appear to be a strong association between sex and attendance at the *cuelga*? Is this difference significant at the 0.01 level?

The first question is one of *strength* of association, so the ϕ coefficient must be computed:

$$\phi = \frac{26(13) - 4(12)}{\sqrt{30(25)(38)(17)}} = \frac{290}{696.1} = 0.42$$

The phi coefficient indicates a positive association of moderate strength in the predicted direction.

The *statistical significance* of this association can be found by using the standard chi-square statistic (corrected for continuity).

O_i	E_i	$\lvert O_i - E_i \rvert$	$\lvert O_i - E_i \rvert - \frac{1}{2}$	$[O_i - E_i - \frac{1}{2}]^2$	$[O_i - E_i - \frac{1}{2}]^2/E_i$
26	20.7	5.3	4.8	23.04	1.11
4	9.3	5.3	4.8	23.04	2.48
12	17.3	5.3	4.8	23.04	1.33
13	7.7	5.3	4.8	23.04	2.99
					$\chi^2 = 7.91$

This value of χ^2 is significant beyond the 0.01 level, but the moderate value of ϕ should sound a note of caution against undue preoccupation with the high level of statistical significance.

Example 14.8

In his study of the southern Yanomamö, Chagnon (1967) suggested that warlike tribes tend to emphasize a cultural norm of *ferocity*—the more ferocious the warrior, the higher is his prestige. A cultural manifestation of this ferocity is a strong tendency for one group to attack the neighboring peoples in an attempt to expand its social territory. The following data have been extracted from Otterbein (1970: 130–149, appendices C and D).

	Territory Constant or Contracting	Territory Expanding
Attacking continual or frequent	Amara	Abipon
	Fox	Aztec
	Ila	Egyptians
	Mossi	Jivaro
	Nandi	Mundurucu
	Papago	Plains Cree
	Saramacca	Sema
	Tibetans	Somali
	Wishram	Thai
		Timbira
		Tiva
Infrequent	Albanians	Japanese
	Amba	
	Ambo	
	Andamanese	
	Copper Eskimo	
	Dorobo	
	Gisu	
	Hawaiians	
	Lau	
	Marshallese	
	Monachi	
	Motilon	
	Mutair	
	Orokaiva	
	Santa Ana	
	Tikopia	
	Tiwi	
	Toda	
	Trumai	

Do these results support the hypothesis that a strong relationship exists between cultural "ferocity" and the tendency for a society to expand its territory? Is this difference statistically significant?

The chi-square statistic for this 2×2 table is as follows:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
9	14	−5	25	1.786
11	6	5	25	4.167
19	14	5	25	1.786
1	6	−5	25	4.167
				$\chi^2 = 11.9060$

This value is significant beyond the 0.01 level. To measure the strength of this association, the ϕ coefficient can be computed from Formula (14.18):

$$\phi = \frac{9(1) - 11(19)}{\sqrt{(20)(20)(28)(12)}} = -0.546$$

Phi could also be computed directly from χ^2 using (14.17):

$$\phi^2 = \frac{\chi^2}{n}$$

$$= \frac{11.906}{40} = 0.298$$

which corresponds closely with the value $\phi^2 = -0.546^2 = 0.2981$ computed from (14.18). Although the χ^2 is statistically significant, the value of $n = 40$ and $\phi = -0.546$ warns that the relationship is not totally overwhelming. In actual research we would also have corrected χ^2 for continuity.

14.6.2 Tau-*b*

The ϕ coefficient is clearly restricted to the 2×2 nominal format, but another related measure, called tau-*b* (or τ_b) expands ϕ to the general $R \times C$ contingency table. Like ϕ, τ_b assumes only a nominal level of measurement. Tau-*b* was initially defined in an important series of articles by Goodman and Kruskal (1954, 1959, and 1963); also see Blalock (1972: 300–302).

To illustrate how τ_b operates, let us return to Example 14.7. Remember that O'Neil and Selby investigated the relationship between sex roles and participation in the Zapotecan *cuelga*. We found earlier that the ϕ coefficient for this contingency table was $\phi = 0.42$. Chi-square indicates this to be significant beyond the 0.01 level.

Now we can examine this same situation using a different probabilistic approach. Suppose that the ethnographer had missed the actual *cuelga*, but was told that 38 of the 55 people in the village attended.

Attend *cuelga* (B_1)	38
Did not attend *cuelga* (B_2)	17
	55 residents of Santo Tomas Mazaltepec

Given only this limited census, how well could an ethnographer guess which individuals attended, and which stayed away? We could, for example, line up the 55 villagers and form them into two groups: One group of 38 who we thought attended and the remaining 17 people who we figured stayed away.

Since we have no outside information, these two groups could be assigned randomly. How many people are incorrectly classified? The probability that any single informant is *incorrectly* placed into group B_1, those who attended the *cuelga*, is clearly $p(B_2) = 17/55$. Because 38 individuals must be independently assigned into group B_1, in the long run, we expect to make only $38 - 11.7 = 26.3$ correct assignments to the group which attended the *cuelga*. Similarly, the error of *incorrectly* assigning informants to the second group is $p(B_1) = 38/55$. Since 17 such assignments are made, we expect that again there will be 17(38/55) = 11.7 errors. (There is only 1 degree of freedom here, so an error in group B_1 automatically implies a corresponding error in group B_2.) Thus, we estimate a total of $11.7 + 11.7 = 23.4$ errors if the informants are randomly assigned to groups B_1 and B_2.

These errors were made by blind chance. What would happen if we were given additional information about the *cuelga*? Suppose somebody told the ethnographer that, of the 38 villagers attending, exactly 26 were males. Now we could reconstruct the 2×2 contingency table considered earlier in Example 14.7

	Men A_1	Women A_2	Total
Attend *cuelga* (B_1)	26	12	38
Did not attend *cuelga* (B_2)	4	13	17
Total	30	25	55

Does a knowledge of the sex ratio at the *cuelga* help in deciding whether or not individual informants attended? Let's see.

We know that 26 of 30 males in Santo Tomas Mazaltepec attended the *cuelga*. Thus, the probability that any particular male *did not attend* is only $p(A_1B_2) = 4/30$. We must now guess at which 26 attended, so we can expect to make about $26(4/30) = 3.5$ errors of assignment. We also expect to make about $4(26/30) = 3.5$ errors when guessing which males did not attend (again, note the single degree of freedom). The probability that a randomly selected woman did not attend the *cuelga* is $p(A_2B_2) = 13/25$. A total of $12(13/25) = 6.2$ errors are likely in deciding which women attended, and $13(12/25) = 6.2$ errors are expected among those who did not attend. Therefore, a total of four kinds of errors exist when we try to reconstruct the cells of the contingency table: $3.5 + 3.5 + 6.2 + 6.2 = 19.4$. These errors are expected when guessing attendance, once sex is known.

So how much does a knowledge of this second variable improve our estimate of who attended the *cuelga*? The proportional diminution of errors is defined as

$$\tau_b = \frac{(\text{no. of errors when } A \text{ is not known}) - (\text{no. of errors when } A \text{ is known})}{\text{no. of errors when } A \text{ is not known}} \tag{14.19}$$

For the example at hand,

$$\tau_b = \frac{23.4 - 19.4}{23.4} = 0.17$$

We can say that a knowledge of the sex distribution saves us about $23.4 - 19.4 = 4.0$ errors in the long run.

So τ_b is a measure of just how well one variable predicts a second. In this example we attempted to predict *cuelga* attendance based upon a knowledge of the sex ratio at the festival. That makes *cuelga* attendance (B) the *dependent* (or predicted) variable and sex ratio (A) the *independent* (or predictor) variable. If A and B are statistically independent, then a knowledge of A should have no effect whatsoever on the outcome of B. Tau-b measures the strength of association of A, given B.

This same positive association was assessed in Example 14.7, using another coefficient, ϕ. The coefficients ϕ and τ_b are rather closely related: $\tau_b = \phi^2$. In this example, $\phi = 0.42$ and $\tau_b = 0.18$. Tau-b is often more meaningful than ϕ because of the intuitively obvious meaning of τ_b.

But τ_b has a second and more important advantage over ϕ. Remember that ϕ can be computed only for 2×2 contingency tables. Tau-b has no such restrictions and is applicable to any R × C table. As an illustration of this, let us once again consider Raymond Firth's kinship and residence cross-tabulation for Tikopia (from Table 11.1).

	Clan			
Village	A_1 Ravena	A_2 Namo	A_3 Faea	Total
B_1: Kafika	31	2	43	76
B_2: Tafua	4	16	46	66
B_3: Taumako	39	6	16	61
B_4: Fanarere	10	3	2	15
Total	84	27	107	$218 = n$

The chi-square statistic computed earlier indicates that the null hypothesis of no association must be rejected. That is, that residence and kinship are not statistically independent on Tikopia. The τ_b coefficient now permits us to assess just *how much association* exists between kinship and residence in this sample.

First consider this question: What does a knowledge of kinship (A) tell us about residence (B)? Kinship is taken here to be the independent variable, and we wish to assess its impact upon residence. If there were no association, the τ_b should be zero, and a knowledge of A would not reduce the errors of assigning informants to residences. τ_b is computed as before. First we find the number of errors resulting in B when A is unknown. A total of 76 Tikopians live in the village of Kafika. The probability of error in randomly assigning an informant to Kafika is $p(\bar{B}_1) = (218 - 76)/218 = 142/218$. In the long run, we can expect to commit about $76(142/218) = 49.5$ errors in assigning informants to Kafika. The other three villages are handled in a similar fashion: For Tafua, we expect $66(152/218) = 46.0$ errors; for Taumako, $61(157/218) = 43.9$ errors; for Fanarere, $15(203/218) = 14.0$ errors. A total of $49.5 + 46.0 + 43.9 + 14.0 = 153.4$ errors are thus estimated for assigning villagers without a knowledge of their kinship. Now we must find what improvement, if any, there is when clan affinities of the villages are known. We know that 31 Tikopians of the Ravena clan live at Kafika, so a total of $31(53/84) = 19.6$ errors is likely. The errors for the other cells are found in a similar manner:

Ravena	Namo	Faea
31(53/84) = 19.6	2(25/27) = 1.9	43(64/107) = 25.7
4(80/84) = 3.8	16(11/27) = 6.5	46(61/107) = 26.2
39(45/84) = 20.9	6(21/27) = 4.7	16(91/107) = 13.6
10(74/84) = 8.8	3(24/27) = 2.7	2(105/107) = 2.0
53.1	15.8	67.5

Total errors in B, when A is known, are $53.1 + 15.8 + 67.5 = 136.4$. The relative improvement in estimating B from a knowledge of clan affiliation is thus

$$\tau_b = \frac{153.4 - 136.4}{153.4} = \frac{17.0}{153.4} = 0.11$$

This is a rather low value, indicating only a weak association between kinship and residence at Tikopia. χ^2 and τ_b tell us rather different things about the same sample. The chi-square statistic was quite large, suggesting that we reject H_0 and conclude that kinship and residence are not independent at Tikopia. But τ_b warns us that this relationship, while significant, is not a very strong one. Even relatively weak associations can prove statistically significant, provided a large enough sample is involved (discussed further in Chapter 16).

Thus, τ_b provides an analog to ϕ, and τ_b is applicable to the general $R \times C$ case. But the conditions for τ_b must be rather carefully defined. Remember the question asked of the Tikopian sample: What does kinship imply about residence? Residence was the dependent variable. Another question remains unanswered by τ_b: What does residence (B) tell us about kinship (A)? Kinship is now put in the dependent position, and this is a rather different situation. A second coefficient must be defined to assess the impact of B on A.

$$\tau_a = \frac{(\text{no. of errors in } A \text{ when } B \text{ is unknown}) - (\text{no. of errors when } B \text{ is known})}{\text{no. of errors when } B \text{ is unknown}} \tag{14.20}$$

The value for the Tikopian sample is $\tau_a = 0.17$, indicating a slightly stronger value than for τ_b. In general $\tau_a \neq \tau_b$.

14.7 CURVILINEAR REGRESSION AND CORRELATION

● *We didn't know we was poor until we went to town.*—R. Cash Hancock

This consideration of correlation has stressed repeatedly that the techniques apply only to situations in which *linear relationships* are suspected between predictor and predicted variables. A host of other mathematical techniques exist which consider nonlinear (that is, *curvilinear*) relationships. Unfortunately, once the linear approximation is known to be invalid, then a bewildering variety of nonlinear regression possibilities jump forth: exponential growth or decay, asymptotic, logistic growth, polynomials of various orders, and so on. The strategy of curvilinear regression consists of sorting through the numerous possibilities at hand in hopes of finding a single curve which best fits the data. Most of these techniques are more advanced than the present scope, and some rather sophisticated analyses of variance designs are often required to determine just which line produces the "best fit." For these reasons, a detailed discussion of curvilinear regression is not attempted here, and the interested reader is referred to Hays (1973: chapter 16) and Sokal and Rohlf (1969: 468–476). Computer programs for fitting curvilinear regression are also available in most comprehensive statistical packages for computers.

There is one technique of curvilinear correlation and regression, however, which deserves mention, for it is not only indeed critical to many anthropological situations, but is also relevant to the current approach. The technique is *logarithmic transformation* and the principle is quite simple: Nonlinear relationships are mathematically converted into linear proportions, and then the standard linear models of correlations and regression can be applied as before. The logarithmic transformation is thus really a device whereby the tedious techniques of curvilinear regression can be avoided. There is a clear analogy between the logarithmic transformation in regression and the transformations considered earlier to approximate the normal distribution.

The simplest logarithmic relationship is the *geometric series*: Growing populations of any species tend to expand geometrically until increase is slowed by extraneous factors.[6] Let us consider the hypothetical situation in which a lifeboat is washed ashore on the deserted island of Malthus. Coincidentally enough, the boat contains two compatible strangers, one male and one female. Upon landing and realizing they have no hope of escape, the strangers decide first to be friends and second to attempt colonization of the island. The population must increase rapidly so that the offspring can survive in this hostile land, so they make an informal pact, agreeing that the rate of growth must be exactly double: two children for every adult. Thus, while the first generation of Malthusians number only two individuals—the original refugees—the second generation will jump to a population of four (two children for each original Malthusian). Once the second generation ceases reproducing, the third generation will number eight people. The fourth generation jumps to sixteen people, and so on. This situation of geometric population increase is graphed in Fig. 14.6. The actual quantitative increase skyrockets as the generations go by, but the basic rate of reproduction remains constant—two offspring for each adult. This characteristic situation cannot be described by the linear regression and correlation techniques considered thus far because the quantitative increase obviously is not linear. But a very simple transformation will allow analysis of this increase *as if it were* linear.

Let us take the common logarithm of the population within each generation and plot these transformed population figures. The population of the first

[6] These extraneous influences constitute the "checks to increase" discussed by Charles Darwin. Whether these influences originate from factors intrinsic within the population or from extrinsic forces, such as food or weather, remains an open issue among population ecologists.

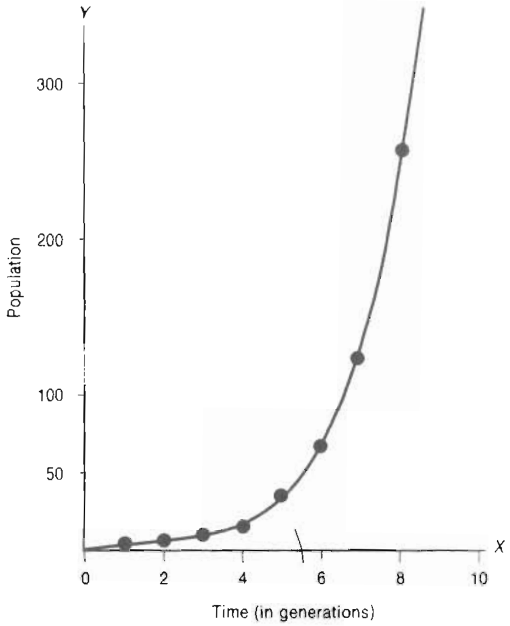


Fig. 14.6

generation is known to be 2, so log(population) is $\log_{10} 2 = 0.301$. The log of the population of generation 2 is log $4 = 0.602$ (exactly *twice* that of the first generation). The log of population in generation 3 is log $8 = 0.903$, exactly *three* times that of generation 1, and so forth. When log(population) is plotted against time (in generations), a perfectly linear relationship results (Fig. 14.7). Bartlett's method of regression determines the best fit to describe this line as

$$\log \hat{Y} = bX = 0.301X$$

Few should be surprised that the slope of this line is log $2 = 0.301$, and that the Y-intercept is through the origin. The correlation coefficient obviously must be $r = +1.00$. Although only ten generations were used in this computation, the equation $Y = 0.301X$ allows projection for the population of any given generation. The projected population for generation 15, for example, is

$$\log \hat{Y} = 0.301(15) = 4.5150$$
$$\hat{Y} = \text{antilog}(4.5150) = 32{,}734$$

Assuming constant conditions, the population of generation 95 is estimated to be 3.935×10^{28} individuals, obviously time for Zero Population Growth.

This simple logarithmic transformation of Y converts a clearly curvilinear relationship into a straight line, and the methods of regression and correlation can be applied to these transformed scores with impunity. An even easier method of analysis is available whenever only a best-fit regression line is necessary.

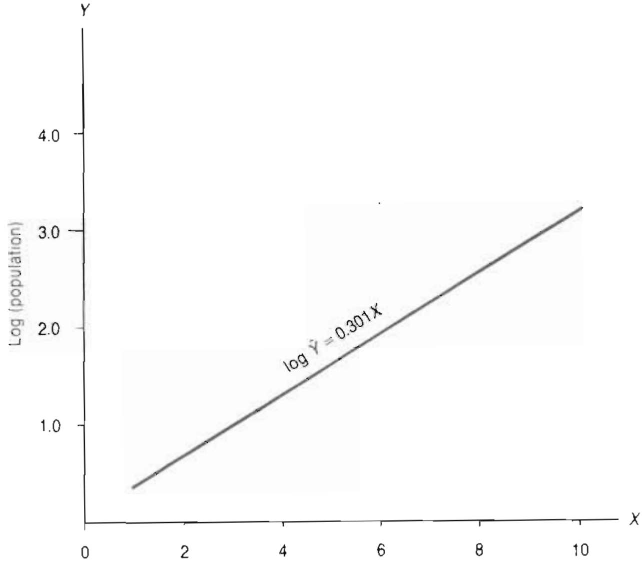


Fig. 14.7

Especially constructed graph paper (*semilog paper*) automatically performs the logarithmic transformation on a set of raw Y_i scores (Fig. 14.8). The X-axis remains on the common arithmetic scale, but the Y-axis is graduated along a logarithmic progression. That is, the distance between 1 and 10 is exactly equal to the distance between 10 and 100, which is exactly equal to that between 100 and 1000. The Y units are thus scaled in exponential fashion, just as one would progress mechanically by using a table of common logarithms. The generational times are plotted on the horizontal axis as before, but the population figures are plotted along the logarithmic divisions of Y. The Malthusian data in Fig. 14.8 has been plotted on *semilog* paper, producing results identical to those of Fig. 14.7, but without the bother of resorting to the logarithmic tables. *Semilog* paper is handy whenever one variable must be transformed and when only an estimate of the regression line is needed. Once the descriptive line of best fit is drawn, the graph can be used in a fashion similar to an actual regression equation, in order to predict interim values of Y from X. When the points tend to scatter—that is, when $\rho \neq \pm 1.00$—the graphic method provides a useful first step in deciding whether or not a semilogarithmic relationship in fact exists. If so, then the individual logs can be found in the tables, and the more rigorous regression and correlation coefficients can be computed. If the points turn out not to be linear, then little effort has been wasted on the preliminary graphics.

A second sort of logarithmic transformation is available when *both* variables must be converted from arithmetic to logarithmic scales. To illustrate the *allometric* or *log–log transformation*, let us examine the relationship which has come to be called *Naroll's constant*. Raoul Naroll correctly realized that the relationship between human population and the floor areas they inhabit would

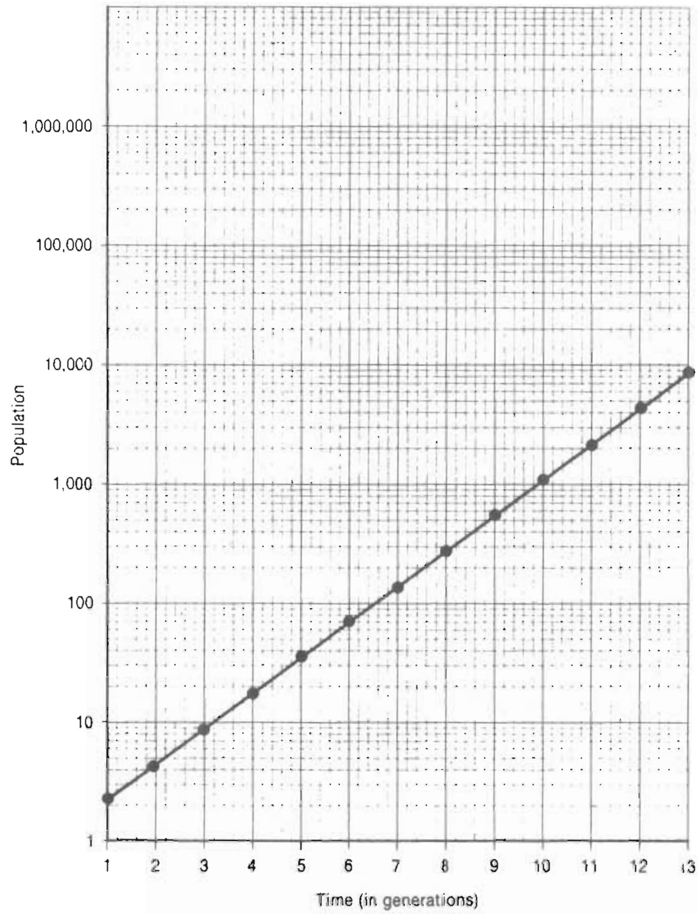


Fig. 14.8

be of interest to prehistoric archaeologists. Specifically, if a systematic relationship could be demonstrated between population and floor area by using a cross-cultural survey technique, then archaeologists would be able to estimate prehistoric population size simply from a knowledge of site size. Naroll (1962a table 1) constructed a cross-cultural sample of 18 societies: 6 each come from North America and Oceania, 3 from South America, 2 from Africa, and 1 from Eurasia. This sample has been reproduced in Table 14.10. Does there appear to be a systematic relationship between the size of a community and the physical area it occupies?

This is clearly a problem amenable to regression analysis, and a scattergram of the data has been constructed in Fig. 14.9. A Model II regression provides the best linear fit to be

$$\hat{Y} = 939.3 + 0.487X$$

TABLE 14.10 Cross-cultural data relating floor area to population size (after Naroll 1962a: table 1).

Society	Largest Settlement (L.S.)	Estimated Population of L.S.	Estimated Floor Area of L.S.
Vanua Levu	Nakaroka	75	412.8
Eyak	Algonik	120	836
Kapauku	Botekubo	181	362
Wintun	?	200	900
Klallam	Port Angeles	200	2,420
Hupa	Tsewenalding?	200	2,490
Haluk	Ifaluk	252	3,024
Ramkokamekra	Ponto	298	6,075
Bella Coola	Bella Coola	400	16,320
Kiwai	Oromosapua	400	1,432.2
Tikopia	Tikopia	1,260	8,570
Cuna	Ustupu	1,800	5,460
Iroquois	?	3,000	13,370
Kazak	?	3,000	63,000
Ila	Kasenga	3,000	47,000
Tonga	Nukualofa	5,000	111,500
Zula	?	15,000	65,612
Inca	Cuzco	200,000	167,220

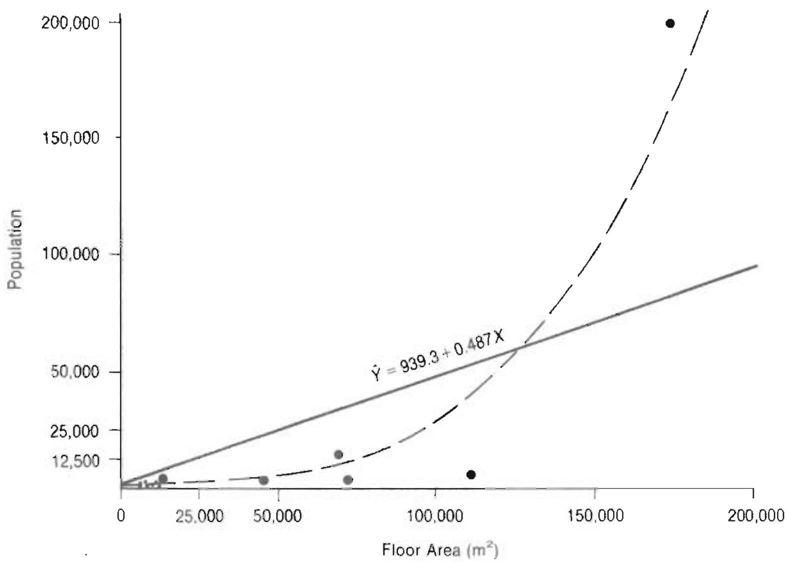


Fig. 14.9

This line has been fitted to the data, but the difficulties with this approach should be obvious. The linear correlation for these data is $r = +0.776$, denoting that only about 60 percent of the variability in Y can be accounted for by X. The scattergram indicates that because the bulk of Naroll's cases have a site size (X) somewhat smaller than about 10,000 m^2, the data points are hopelessly bunched into the lower left of the scattergram while the point representing the Inca of Cuzco lies isolated in the upper extreme. The common linear regression derived above has little relevance to the empirical scatter of points because *this relationship is obviously not linear*. In fact, it has been difficult to represent all the data on a single graph: When the Inca are included, the bulk of the cases becomes blurred. Predictions emanating from a linear description of nonlinear phenomena area are generally found to be spurious and misleading.

But the data on Fig. 14.9 look suspiciously like an *allometric* function, as estimated by the dashed line. If floor area and settlement size are indeed allometric pairs, then a log–log transformation of the scales should provide a more suitable procedure of estimation. Before attempting the transformation, it is a good idea to see how well the allometric function fits the actual data.

Standard log–log paper provides a quick-and-dirty method for determining whether or not an allometric curve sufficiently describes a set of data. To plot data on log–log paper, follow these steps:

1. *Determine the number of "cycles" present in the data.* The number of logarithmic cycles within a data set refers to the number of meaningful decimal places. The population figures on Table 14.10 range between 75 and 200,000 people, encompassing the following decimal digits: tens, hundreds, thousands, ten thousands, and hundred thousands. At least five logarithmic cycles will be required to describe the populations. The floor areas range from 412.8 to 167,220 m^2, so at least four log cycles are involved (hundreds, thousands, ten thousands, and hundred thousands). Thus, the appropriate log–log paper must contain at least four or five logarithmic cycles.
2. *Establish arithmetic scales on the log–log graph.* Following the conventions of regression, the horizontal axis will depict X, so the X-axis is divided into floor areas (expressed in square meters) and Y-axis is divided to represent population. The scalar intervals follow an exponential rather than arithmetic frequency, and therefore care must be taken to avoid errors in transcription.
3. *Plot the data points on the scattergram.* If more than 50 or so points are involved, the data should be grouped into a frequency distribution so that the means of each frequency class can be plotted instead of each individual point.

A perfectly allometric relationship will produce a straight line on log–log paper; the degree of scatter represents the amount of deviation within the sample. If log–log paper is unavailable, then the common (or natural) log table conversions can be plotted on an arithmetic scale, but use of the log–log paper saves the tedium of finding the $2n$ logs.

Once the log–log scattergram is available, one can readily determine whether computing allometric regression is worthwhile. Figure 14.10 indicates that

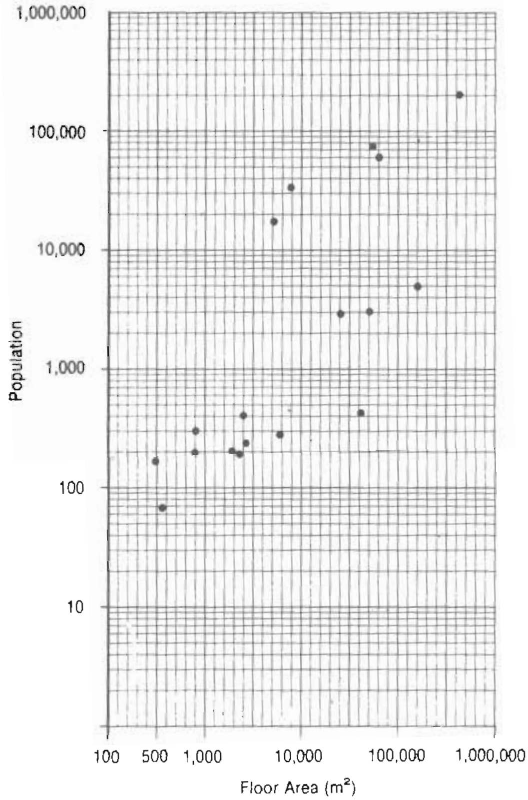


Fig. 14.10

although a good deal of dispersion remains, a log–log transformation does indeed render the points much more linear than merely the raw variates would indicate. Thus, the graph indicates that log–log regression would be a suitable technique to apply to Naroll's data.

The general allometric formula is expressed as

$$\log \hat{Y} = \log a + b \log X \tag{14.21}$$

where a and b are the common coefficients of linear regression. As mentioned earlier, the log–log transformation is not technically a method of curvilinear regression; the variates are merely converted from a curvilinear (geometric) to a linear (arithmetic) scale of measurement and then treated *as if they were linear.*

Table 14.11 shows the computations necessary to find the proper regression line for Naroll's population data. All raw scores are initially converted into common logarithms; then the standard computations are used to find the value of the two regression constants. The Model II method of fitting the regression line has been employed here because the X variable, the *estimated floor area of largest settlement*, is a random variable, subject to errors of sampling. The final

TABLE 14.11 Log–log transformation of Naroll's data on Table 14.10.

Society	X	$\log X$	Y	$\log Y$	
Kapauku	362	2.55871	181	2.25768	
Vanua Levu	412.8	2.61574	75	1.87506	
Eyak	836	2.92221	120	2.07918	$\log \bar{X}_1 = 2.93179$
Wintun	900	2.95424	200	2.30103	$\log \bar{Y}_1 = 2.23601$
Kiwai	1,432.2	3.15600	400	2.60206	
Klallam	2,420	3.38382	200	2.30103	
Hupa	2,490	3.39620	200	2.30103	
Ifaluk	3,024	3.48058	252	2.40140	
Cuna	5,460	3.73719	1,800	3.25527	
Ramkokamekra	6,075	3.78355	298	2.47422	
Tikopia	8,570	3.93298	1,260	3.10037	
Iroquois	13,370	4.12613	3,000	3.47712	
Bella Coola	16,320	4.21272	400	2.60206	
Ila	47,000	4.67210	3,000	3.47712	
Kazak	63,000	4.79934	3,000	3.47712	$\log \bar{X}_3 = 4.79528$
Zulu	65,612	4.81697	15,000	4.17609	$\log \bar{Y}_3 = 3.78873$
Tonga	111,500	5.04727	5,000	3.69897	
Inca	167,220	5.22329	200,000	5.30103	

$$\bar{X} = \frac{68.81904}{18} = 3.82328 \quad \bar{Y} = \frac{53.15784}{18} = 2.95321$$

$$b' = \frac{3.78873 - 2.23601}{4.79528 - 2.93179} = 0.833$$

$$a' = 2.953 - 0.833(3.823) = -0.232$$

$$\log \hat{Y} = 0.833 \log X - 0.232$$

allometric line of regression is given by

$$\log Y = 0.833 \log X - 0.23$$

Sample values from this equation have been computed on Table 14.11, and the line of regression has been fitted to the datum points of Fig. 14.11. Note how much closer the data cluster around the log–log approximation than did the same data about the simple arithmetic regression. The linear correlation coefficient for the allometric plot is $r = +0.878$, accounting for over 77 percent of the variance in Y. Thus, the log–log transformation represents 77% − 60% = 17% improvement over the simple, untransformed method.

So, the mechanics of allometric correlation and regression offer few additional computational challenges. But unexplained so far is just why the log–log transformation works so handily and just what an allometric relationship really signifies. One even hears occasional grumblings that transformations smack of "cooking one's data" in order to obtain more successful results. The fact is that this general suspicion of data transformations is grounded in an interesting ethnocentric fallacy. Many of our mathematically naive colleagues seem to feel

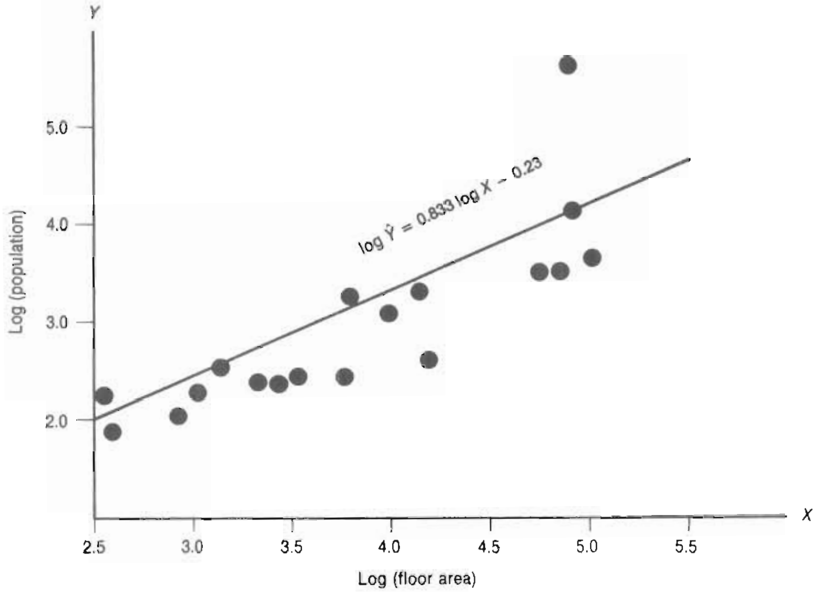


Fig. 14.11

that the common decimal scale is somehow more "natural," that arithmetic scales are somehow better equipped to reflect the characteristics of man and nature. Until recently, Western schooling has inflated the importance of the arithmetic relationship (largely by default) until we are led to look with suspicion at any deviation from the commonplace decimal system. Recent developments such as the "New Math" and talk of the United States converting to metric equivalents has fortunately undermined some of this suspicion, but many social scientists still seem to look askance at nonarithmetic scales. I should think, however, that anthropologists, of all the social scientists, should be most sensitive toward regarding one's own system as somehow "most natural" or "better" than other systems; this is simply an ethnocentric viewpoint.

Specifically with reference to the log–log transformation, there are solid reasons why the allometric relationship provides a more valid measuring scale in some contexts. The allometric formula was originally introduced in 1881 by Snell to express the relationship between brain size and body mass in mammals, and allometric relationships have since been discovered to operate in a number of social and biological circumstances. Specifically, the allometric equation holds that the *ratio* between two variables is roughly constant. For example, adult human weight is known to range roughly between 67 and 89 kg, over a range of about 22 kg. In *Macaca mulatta*, however, the adult weight varies from about 5.7 to 12.0 kg, a range of only 6.3 kg. Thus, it could be stated that the weight of *Homo sapiens* is vastly more variable than that of the *Macaca*. But such a facile pronouncement suffers from the fallacy of using observed range of variability as a point of comparison. Such a measure is clearly unfair. A more suitable method of comparison involves the relative *proportion* of variability, such as the ration between the greatest and the least values. For man, this ratio is about $89/67 = 1.33$, while for the macaque the ration is $12/5.7 = 2.10$. Clearly, the macaque is somewhat more variable than man, and this relationship has been obscured by using the common arithmetic scaling procedure.

The concept of *proportionate range* can best be expressed in logarithmic form. If X_1 is the largest variate and X_2 is the smallest, then the logarithm of the ratio $(X_1/X_2) = \log X_1 - \log X_2$. For man, this ratio was determined previously to be about $X_1/X_2 = 1.328$. But alternatively, one can state that, for *Homo sapiens* $\log X_1 - \log X_2 = \log 1.328 = 0.51587$. Because $\log X_1$ is the logarithm of the heaviest human and $\log X_2$ is the log of the lightest adult, then $\log X_1 - X_2$ is the range of logarithms for modern human mass. That is, because few humans weigh more than 89 kg ($\log 89.0 = 1.94939$) and few weigh less than 67 kg ($\log = 1.82607$), the range of log weights can be determined by subtraction $1.94939 - 1.82607 = 0.12332$, or about 0.12. This basic logarithmic, or *allometric* difference is roughly constant over most human measurements, although sometimes variability may be exceedingly wide or narrow for specific characteristics. The logarithmic range for human *stature*, for example, is known to be only about 0.17. Thus, the ratios are roughly constant, even though the individual arithmetic measurements vary widely. Thus, the logarithmic transformations help isolate the fundamental principles underlying the jumble of observable phenomena.

Allometric relationships have been recognized in a wide variety of biological and social phenomena. Demographers often apply the log–log formula to analyze the growth of urban centers, and allometrics have been used to describe changing word frequencies in languages. The size of the human cranium is known to change allometrically with respect to body height as the individual grows from childhood to maturity. Naroll has recently established a log–log relationship between the number of occupational specialities and the absolute population size. The interested reader is referred to the excellent review article by Naroll and Bertalanffy (1965), in which the principle of allometry is considered throughout the biological and social sciences. Simpson, Roe, and Lewontin (1960: chapter 15) also treat the mathematical aspects of the allometric constant, with specific reference to animal growth.

SUGGESTIONS FOR FURTHER READING

Adkins (1964: chapters 12–14)
Blalock (1972: chapters 17, 18)
Siegel (1956)
Sokal and Rohlf (1969: chapter 15)

EXERCISES

14.1 Return to the variates in Exercise 13.1.

(a) Find the correlation coefficient for these data.

(b) Is this correlation significantly different from zero (at 0.05)?

(c) Find the 95 percent confidence limits for this r.
(d) What percent of the variance in Y is explained by the linear relationship existing between X and Y?

*14.2 Return to the data of Exercise 13.3.
(a) Compute the correlation coefficient.
(b) Is this correlation significantly different from zero (at 0.05)?
(c) Compute the 95 percent confidence limits for r.

14.3 A correlation coefficient of 0.60 is found in a sample of 25 pairs. Is this significantly different from zero at the 0.05 level of significance?

14.4 How large a correlation coefficient is needed from a sample of size 15 to justify the claim that the variables are linearly related (at the 0.05 level)?

14.5 A sample of 95 pairs has a correlation coefficient 0.80. Is this significantly different from $\rho = 0.50$ at the 0.05 level?

14.6 Use Spearman's r_s to test the hypothesis that the upper limit of community size is directly proportional to the degree of economic specialization (data from Ember 1963: table 1).

Relationship between upper limit of community size and number of types of economic specialist.

Rank Order of Societies	Community Size	Economic Specialization
Yagua	1	6.5
Naron	2	1.5
Ulithi	3	11
Hupa	4	11
Ainu	5	1.5
Lesu	6	11
Egedesminde	7	11
Moken	8	11
Ramkokamekra	9	3.5
Bella Coola	10.5	3.5
Kiwai	10.5	14.5
Tikopia	12	22
Ona	14	6.5
Nuer	14	16
Samoa	14	17.5
Flathead	16	6.5
Hopi	17	20
Crow	18	14.5
Cuna	19	17.5
Nama	20	6.5
Dahomey	21	20
Zululand	22	20
Nupe	23	23
Inca	24	24
Aztec	25	25

14.7 Test the hypothesis in Exercise 14.6, using Kendall's tau. Which coefficient seems more appropriate? What are the strengths and weaknesses of each?

14.8 The following data have been extracted from Ember (1963: table 2):

Relationship between relative importance of agriculture and number of types of economic specialist.

	Types of Economic Specialist	
Agriculture	2–5	6 or more
Relatively unimportant	8	1
Relatively important	3	10

(a) Use an appropriate test to determine whether there is a statistically significant relationship between agriculture and the types of economic specialists (at 0.05).

(b) What is the *strength* of this relationship?

14.9 Ember also compared the importance of agriculture with the number of types of political officials (1963: table 6).

	Number of Types of Political Officials	
Agriculture	Fewer than 5	5 or more
Relatively unimportant	8	1
Relatively important	4	11

(a) Is this a statistically significant association (assume one-tailed test at 0.05)?

(b) Which association is stronger, that between agriculture and economic specialization (Exercise 14.8) or the present example?

14.10 In Exercise 11.4, how strong is the association between the sex of applicant and admission to graduate school? Which statistical measure (correlation or significance) seems more appropriate here? Why?

14.11 In Exercise 11.7, how strong is the relationship between Bugandan drinking patterns and age?