CHAPTER 10

DESCRIPTIVE STATISTICS FOR BIVARIATE DISTRIBUTIONS

In the last chapter we discussed statistics that provided summary measures of the properties of univariate distributions. Also, in our introduction to the bivariate realm (Chapter 8) the discussion was confined to an intuitive understanding of statistical relationships through the use of percentages in analyzing contingency tables. In fact, without mentioning it as such, you have already been exposed to a statistical index of association.[1] After computing percentages vertically (in the direction of the independent variable) and comparing (horizontally) across categories of the independent variable, the percentage difference is called epsilon (symbolized ε). Any time ε is greater than zero there is evidence of a relationship between the variables under examination. The larger the ε the stronger the relationship, although the association may not be large, statistically significant, nor practically meaningful. Epsilon values range between 0 (which would indicate a state of statistical independence in which case the variables do not covary) to 100 (which indicates maximum covariation and maximum association between variables). Hence epsilon ranges on a scale from 0 to 100:

range of ε : 0 ——————————— 100

minimum ε maximum ε

Epsilon is a convenient first step for discovering a relationship between variables but its limitations are quickly seen in bivariate tables with larger than 2 x 2 dimensions. In such cases there are several different percentage comparisons that could be made. Secondly, this index of association does not describe the relationships in the entire table. A third restriction occurs when data are not in a contingency table format, in which case ε is not appropriate. Under the latter circumstances, more powerful tools exist for ferreting out the relationship between variables (quanitatively speaking) and its underlying meaning.

Therefore, in this section we will introduce and explicate some conventional statistical procedures for determining the precise numerical value of the degree of correspondence between variables construed in a bivariate table.

Measures of association, commonly called correlation coefficients, are founded upon two types of relationships that provide sufficient evidence that the variables are associated: 1) the principle of the joint occurrence of attributes and 2) the principle of covariation.[2] In this section, the former will be applicable since the data are presented in tabular form and the measurement level is nominal or, at best, ordinal. The term "attribute" is a clue to the nature of the data being scrutinized.

Statisticians have developed, devised, and refined numerous statistical indices of relationships. Since our purpose is to be synoptic rather than exhaustive in treatment, only a few of these statistics will be introduced. At the nominal level there is a family of correlational statistics that are based upon a measure called chi square (symbolized x^2). To develop the rational for the chi square based statistics let us turn to Table 10.1. We know that an association exists between political party preference and attitudes toward the pardoning because epsilon is greater than zero ($\epsilon = 36$). Notice that in a fourfold table the absolute magnitudes of epsilon are identical for both row comparisons. Other than knowing this value falls about a third of the way between the minimum and maximum epsilon value, we do not know the overall association between the two variables. To calculate the numerical relationship one strategy is to construct a model of no association. In other words, if the two variables were not related what would be the expected (or theoretically based) frequencies? The observed empirical frequencies of 225, 270, 602, and 156 have already been obtained. The logic underlying chi-square based association is this: What would we expect the respective cell frequencies to be if party preference and attitude were not correlated?

Table 10.1

ATTITUDES TOWARD PARDONING BY POLITICAL PREFERENCE

	Democrats			Republicans		Totals
	n	%		n	%	
Pardon ("Yea")	225	27	$\epsilon = 36$	270	63	495
No Pardon ("Nay")	602	73	$\epsilon = 36$	156	37	758
	827	100%		426	100%	1,253

It might be tempting to say that if the total sample consists of 1,253, then, if the two variables were not related, about ¼ or 313.25 cases (because there are four cells in this table) of the total should be found in each cell. The pitfall in this line of reasoning is that no account is made of the actual number of Republicans and Democrats actually surveyed, nor the number who said "yea" and "nay". Not only are there almost twice as many Democrats (827 vs. 426) but any observer of the political scene knows that a dramatic decision (first-time one at that) like this is affected by one's party loyalty. To construct a model of no association necessitates taking into account the actual marginal distributions (e.g., number of Republicans and Democrats and number of yea-sayers and nay-sayers) of the data. Not only is there an approximate 2:1 ratio between Democrats and Republicans, but there is almost a 2:1 ratio between nay-sayers and yea-sayers.

With this as a backdrop statisticians think like this: If the total sample is comprised of 758 nay-sayers, then (426/1253) (758) of the nay-sayers (or 258) should be Republicans who chose "no pardon". Similarly, if 827 of the

total sample are Democrats and, again, 758 of the responses are nay, then (758/2153) (827) of the nay-sayers (or 500) should be Democrats. The same reasoning applies to the other two cell frequencies. Specifically, one would expect that 426/1253 times 495 (or 168) of the Republicans to say "pardon" and 827/1253 times 495 (or 327) of the Democrats to judge the decision correct. Notice that the expected frequencies (the root of the model of no association) are generated using the empirically obtained marginal distributions. In fact, a concise formula for generating the expected frequencies (E_f) in a contingency table can be forwarded and reads as follows:

$$E_f = \frac{\text{row marginal total x column marginal total}}{\text{grand total}}$$

or simply:

$$E_f = \frac{\text{(row total) (column total)}}{N}$$

For each of the cells in the contingency tables the E_f's would be:

$n_{11} = (495)\ (827)/1253 = 326.71$

$n_{12} = (495)(426)/1253 = 168.29$

$n_{21} = (758)(827)/1253 = 500.\ 29$

$n_{22} = (758)(426)/1253 = \underline{257.71}$

$\Sigma = 1253.00$

These are the frequencies expected if, in fact, no relationship exists between X and Y. Notice that the sum of E_f's is equal to the total sample size of 1253. Having established the E_f's we may use the following working table (Table 10.2) to compute the chi-square value:

Table 10.2

WORKING TABLE FOR
CONSTRUCTING A MODEL OF NO ASSOCIATION

Cell	O_f	E_f	$O_f - E_f$	$(O_f - E_f)^2$	$(O_f - E_f)^2 / E_f$
n_{11}	225	326.71	-101.71	10344.92	31.66
n_{12}	270	168.29	101.71	10344.92	61.47
n_{21}	602	500.29	101.71	10344.92	20.68
n_{22}	156	257.71	-101.71	10344.92	40.14
	Σ=1253	Σ=1253	Σ=0		$\sum = x^2$ =153.95

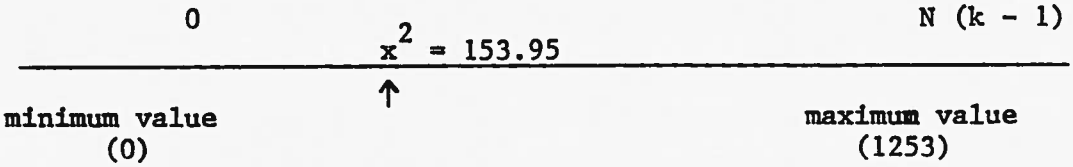
The respective E_f's are subtracted from their counterpart O_f's (producing a value called delta, symbolized "Δ"), the difference is squared, and the squared difference is divided by E_f. When the last column is summed the quantity chi-square is obtained. Conceptually, it represents the discrepancy between observed and expected frequencies, adjusting for expected cell frequencies. With this value an entire "family" of statistics, appropriately called the chi square or delta based statistics, appropriate for computing indices of association can be computed. A common fallacy is to think of the chi-square value itself as an indicator of association. This is not true. It is actually a statistical measure of significance although it is used in the numerator of several different association measures.[3]

Before specifying specific association statistics, several additional words apropos chi square are germane. First, chi square must be a positive number because the $O_f - E_f$ differences are squared. If no association exists (i.e., when O_f and E_f correspond identically) chi square will be zero. Second, its upper limit is a function of N (sample size) and k (number of categories) and is expressed as follows:

$$\text{upper limit of } x^2 = N\,(k - 1)$$

where k = number of rows or columns, whichever is smaller.

Hence the scale of values over which x^2 can range in the present case is charted as follows:



For didactic purposes, we have located the maximum x^2 value for our data along with the computed value.

Before presenting several chisquare based statistics of association, a brief review of the logic behind the x^2 computation is in order. When percentage differences are employed, there are several different comparisons that could be made and they geometrically increase with the r x c configuration of the table. Similarly, in a two by two table there are four delta values. Because of these limitations statisticians desire a single summary measure describing the association in the table as a whole. This aggregate summarization is produced as follows: 1) the delta values are squared, otherwise the algebraic sum would equal zero (see colmumn 4 of 10.2) and this would be a most undesirable condition if further computation is involved; 2) the respective squared differences, Δ^2, are divided by the corresponding cell frequencies. The purpose of such a division is that a particular deviation implies more of an association when the expected frequency is small than when it is large; 3) the ratios are summed over all cells providing the researcher with a single number (which as we said earlier is one reason for computing measures of association in the first place); and 4) because the upper limit of x^2 is a function of N and k (see range of x^2 above) it is necessary to divide by N (which is its maximum value in a fourfold table) to take into account its maximum possible value.

NOMINAL MEASURES OF ASSOCIATION

There are several strains of this lineage of statistics. Again, rather than be exhaustive, a select few indices will be chosen for illustrative purposes. Bear in mind that virtually all measures of association, particularly those not discussed, are simple refinements to make allowances for mathematically undesirable properties of the other coefficients. Specifically, some statistics do not have a maximum value of 1.00 which makes identification of a perfect correlation difficult (the case with the contingency coefficient, C), and some can exceed 1.00 in tables with more than 2 x 2 cellular arrangements (e.g., the phi coefficient, ϕ).

Phi. Although this statistic has a problem (e.g., under certain conditions its maximum value can exceed 1.00) it is a logical step from the computation of the chi square value. The formula[1] for ϕ and ϕ^2 is:[4]

$$\phi = \sqrt{x^2/N} \quad \text{or} \quad \phi^2 = x^2/N. \quad \text{or} \frac{n_{11}n_{12} - n_{12}n_{21}}{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22}}$$

Note that by dividing by N, the maximum value for x^2 in a 2x2 table, the ratio is an expression of the relationship between what we've obtained and what we could obtain. In short, the phi coefficient is an attempt to norm the obtained statistic to some standard. For our data, substituting into the ϕ formula, we have:

$$\phi = \sqrt{\frac{153.95}{1253}} = \sqrt{.1299} = .35$$

The problem with ϕ is that when r x c is greater than two in both, the maximum value can exceed 1.00 since the upper limit of x^2 $\left[N\ (k-1)\right]$ can be larger than N. In such situations it is best to use a different statistic. The maximum value of ϕ^2 is k-1 where k is the smaller of the r or c.

Cramer's V. This statistic may be the best all-purpose nominal measure of association because it overcomes most of the intrinsic deficits of the other statistics. Computationally it is obtained via:

$$V = \sqrt{x^2/N+}$$

Note that it too has x^2 in the numerator (the reason why we consider it a chi square based statistic) and a quantity called t along with N in the denominator. t refers to the smaller of the two quantities: (r - 1) or (c- 1). For the data in Table 10.1, V is computed as follows:

$$V = \sqrt{\frac{153.95}{(1253)\ (1)}} = \sqrt{.1229} = .35$$

Both phi and Cramer's V turn out to be identical. Often in 2 x 2 tables statistics that would yield different values in larger tables are identical in 2 x 2 ones.[5]

Yule's Q. Another statistic of association is Yule's Q and will now be discussed because of its computational simplicity and an interpretation not available with some chi - squared based statistics.[6] It is special case of another measure (gamma) and is restricted to use with 2 x 2 tables as the Q formula makes clear:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

$Q = (AD - BC) / (AD + BC)$

where:

Cell A (or n_{11})	Cell B (or n_{12})
Cell C (or n_{21})	Cell D (or n_{22})

Only the cross products in the diagonal cell frequencies are employed. With respect to the sign value of Q it is important to remember the general rule that the sign of any nominal level coefficient is uninterpretable. For the data in table 10.1, Q is:

$$Q = \frac{(255)(156) - (270)(602)}{(255)(156) + (270)(602)} = \frac{35100 - 162540}{35100 + 162540} = \frac{127440}{197640} = -.64$$

Yule's Q possesses an interpretation known as the proportional reduction in error ("PRE") interpretation. This means that so much variation in the dependent variable can accounted for or explained by the independent variable. Substantively for the Q of .64 we can say that 64% of the "error" in predicting attitudes toward pardon is due to the association of attitudes with political preference.

Lambda. Lambda, also known as Guttman's coefficient of predictability,

has a PRE interpretation but is a different sort of association statistic. Measures of association may be divided into symmetric and asymmetric types. Symmetric statistics like ∅, V, and Q produce identical numerical values regardless of how the two variables are arranged in the contingency table. In other words, it does not matter--for the coefficient's sake--which variable is across the heading or which is down the stub. Asymmetric statistics, like lambda, are influenced by the particular tabular arrangement since the purpose is to predict one variable from a knowledge of the other. Consequently, which variable is the predictee and which variable is the predictor may alter the computed coefficient's value. Lambda is a particularly useful association measure when there exists a clear-cut independent and dependent variable. As one might surmise, the causal variable will be used as the predictor and the effect variable as the predictee.

For the data in Table 10.1 it is clear that the causal variable is political party preference and the attitude toward the pardoning the effect variable. Calling the former X and the latter Y the lambda formula reads:

$$\lambda_{YX} = \frac{\Sigma f_i - \Sigma f_d}{N - \Sigma f_d}$$

where: f_i= largest cell frequency within each category of the independent variable
f_d= largest marginal frequency of the dependent variable totals
N= total number of respondents

The system of double subscripts means that the first notation (Y) to the right of the Greek symbol (λ) is the effect variable and the second notation (X) is the causal variable. Substituting the data into the formula:

$$\lambda = \frac{602 + 270 - 758}{1253 - 758} = \frac{114}{495} = .23$$

This is interpreted to mean that by knowing the empirical relationship between X and Y the magnitude of prediction error can be reduced by 23% (.23 x 100) over what could be achieved by knowing only the marginal totals.

Lambda as a Proportional Reduction in Error (PRE) Measure.

The generic PRE formula reads:

$$\frac{E_1 - E_2}{E_1}$$

where: E_1 = "errors" made by rule number 1
E_2 = "errors" made by rule number 2

Rule 1 (for E_1). By knowing only the dependent variable totals the best prediction ("best" in the sense that it will produce the fewest number of "errors" in the long run) would be the mode of that variable. Since the mode in Table 10.1 is "nay" that would comprise the better predictor.

Rule 2 (for E_2). By having the cross-classification of two variables--attitude toward pardoning and political preference--the best prediction would be that of the mode within categories of the independent variable.

Prediction Errors. For rule 1 we would make 495 errors by predicting "nay". This is so because by predicting nay we would incorrectly predict the response of the 495 subjects who said "yea". For rule 2 we would make 381 errors (225 by predicting "nay" for Democrats and 156 by predicting "yea" for Republicans).

Definition of Measure. The proportional reduction in error achieved by using rule 2 rather than rule 1 is:

$$\frac{E_1 - E_2}{E_1} = \frac{495 - 381}{495} = .23$$

The PRE basis of lambda can be reiterated as follows. If all you knew were the marginal totals ("yea" = 495 and "nay" = 758) of the dependent variable your best single prediction for each response would be "nay". Employing such a procedure would result in 495 "errors". By adding a second variable (political preference) your best single prediction for Democrats would be "nay" while your best single prediction for Republicans would be "yea". Employing this procedure you would make a total of 381 "errors" (225 for Democrats and 156 for Republicans). Therefore, by having additional information, that is, the distribution of responses for both Republicans and Democrats, you would make 114 fewer errors, Hence, the proportion by which you could reduce prediction errors is .23 (495-381÷495).

Some Important Observations on Statistics of Relationship.

It stands to reason that if two variables are independent of each other the index of association should yield a value of zero. Similarly, if two variables are perfectly correlated then the measure of relationship should reflect this fact and produce a value of +1.00 if they are perfectly positively associated and a value of -1.00 if they are perfectly negatively associated. However, the meaning of a perfect relationship can occur in at least two different ways.[7] We will call these two procedures: 1) the stringent model of perfect correlation, and 2) the less stringent model of perfect correlation.

1) The Stringent Model of Perfect Correlation.

Let us consider the simplest case, that of a 2 x 2 contingency table. When all the cell frequencies fall into the diagonals of the table, and by extension, no observations are registered in the other cells, the conditions for a perfect correlation using the stringent model are met: For example, a perfect positive correlation would take on the following appearance:

		Variable X: High	Variable X: Low
Variable Y	High	10	0
	Low	0	10

$$\phi = \frac{100 - 0}{\sqrt{(10)(10)(10)(10)}} = 1.00$$

This situation involves two dichotomous variables, each subclassified into high and low categories. Notice that each value of variable X is associated with only one value of variable Y (e.g., all high variable X values are in the high variable Y category and all low variable X values are in the low variable Y category).

Similary, a perfect negative correlation would take on the following form:

		Variable X	
		High	Low
	High	0	10
Variable Y			
	Low	10	0

$$\phi = \frac{0 - 100}{\sqrt{(10)(10)(10)(10)}} = \frac{-100}{\sqrt{10000}} = \frac{-100}{100} = -1.00$$

Again all high variable X values are in the low variable Y category and all low variable X values are in the high variable Y category.

2) The Less Stringent Model of Perfect Correlation.

Let us also consider the simplest case, that of a 2 x 2 contingency table. If one of the cell frequencies has no observations in it and, corollarily, the other three register some observations, then the conditions for the less stringent model of perfect correlation are met. To illustrate, a perfect positive correlation would take the following appearance:

		Variable X	
		High	Low
	High	10	10
Variable Y			
	Low	0	10

$$Q = \frac{100 - 0}{100 + 0} = \frac{100}{100} = 1.00$$

Notice that only one cell contains no observations.

For a perfect negative correlation the table would take the following form:

		Variable X	
		High	Low
	High	10	10
Variable Y			
	Low	10	0

$$Q = \frac{0 - 100}{0 + 100} = \frac{-100}{100} = -1.00$$

Again only one cell fails to contain any frequencies.

A significant query becomes: When do we use statistics based upon the stringent model of perfect correlation and when do we use statistics based upon the less stringent model. A guideline can be advanced. If the categories of the independent variable are known or thought to influence the dependent variable, then statistics based on the first model (e.g., Ø) are probably most appropriate. Regarding our previous example (Table 10.1) we would expect political preferance--both Republican and Democrat--to affect attitudes toward the pardoning. On the other hand, in certain types of experimental research where, for example, those innoculated with a flu vaccine are expected to have a reaction while those not vaccinated are not expected to have a reaction, statistics based on the second model (e.g., Yule's Q) would probably be best to employ.

Ordinal Measures of Association

Spearman's Rho. When data have been ranked (i.e., conform to ordinal level measurement assumptions) a useful, accurate, and computationally simple measure of association called Spearman's rho (r_s) is appropriate. To compute this coefficient requires one to subtract the difference between two sets of ranks, square and sum these differences, and finally substitute the summed difference into the formula:

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

For the data in Table 10.2 the eponymous rho will be computed. Each of the universities in the "Big Ten" are located down the far left hand column and in the adjacent columns appear the predicted finish and the actual finish. Our job is to determine the association between the two sets of ranks. A third column "D" (difference between ranks) is addended and the difference between a given team's rank is recorded. Notice that the algebraic sum of the D column must be 0 and this should be used as a check on your work. Each of the "D" values is squared and placed in an adjacent column labeled

"D^2". Finally, the "D^2" is summed and the sum substituted into the r_s formula. The denominator "N" refers to the number of ranked cases, teams in this case. Performing this:

$$r_s = 1 - \frac{6(45.5)}{10(10^2-1)} = 1 - \frac{273}{990} = .724$$

Spearman's rho may be interpreted exactly as Pearson's r, that is, in terms of sheer magnitude and in terms of the proportional reduction in error since it is a product moment association coefficient for rank-ordered data.[9] The value of rho will fall between -1.00 (a perfect negative association) and +1.00 (a perfect positive association). A value of zero indicates no association between ranks.

TABLE 10.2

FOOTBALL ACTION PREDICTIONS

Team	Rank X	Rank Y	D	D^2
Michigan	1	2	-1	1
Ohio State	2	1	+1	1
Minnesota	3	4	-1	1
Michigan State	4	3	+1	1
Indiana	5	10	-5	25
Purdue	6	7	-1	1
Illinois	7	5	+2	4
Northwestern	8	8.5	-.5	.25
Wisconsin	9	6	3	9
Iowa	10	8.5	1.5	2.25
			$\Sigma D=0$	$\Sigma D^2=45.5$

Goodman and Kruskal's Gamma. Another useful measure of association for correlating ranks is gamma (G). Suppose we're interested in the correspondence between two variables, job dissatisfaction and automation, for five business enterprises. Each of the organizations and its workers have been studied and ranked on these two dimensions. Table 10.3 contains these hypothetical organizations are placed in their natural order (from 1 to n) for one variable the ranks ranks. Operationally, after the ranks of the second variable are located in of the juxtaposition. Then two new columns are addened: 1) agreements and 2) inversions. To determine the frequency of agreements exclusive attention is paid to the column not arrayed in perfect order. We ask: "How many ranks above it (e.g., General Telephone) are smaller?" Since there are no ranks above General Telephone, a 0 is entered in the agreement column. Then we move to the second organization, Illinois Agricultural Association, and ask the same question. Since there is 1 rank above it that is smaller (e.g., General Telephone has a rank of 1,) we place a one in the agreement column. We proceed in this fashion until the last organization is examined. Since two ranks above Clay Dooley are smaller (e.g. General Telephone and State Farm Insurance Company) we place a 2 in the column adjacent to Clay Dooley. Finally we sum the number of agreements, which is 6 in the present case.

TABLE 10.3

RANK OF FIVE ORGANIZATIONS ON AUTOMATION AND JOB DISSATISFACTION

Organization	Rank on Automation	Rank on Job Dissatisfaction	Agreements	Inversions
General Telephone	1	1	0	0
Illinois Agriculture Association	2	5	1	0
State Farm Insurance Co.	3	2	1	1
Illinois State University	4	4	2	1
Clay Dooley Mfg.	5	3	2	2
			$\Sigma fa=6$	$\Sigma fi=4$

In the final column ("inversions") we again pay exclusive attention to the ranking in imperfect order but this time ask: "How many ranks above it are larger?" Since there are no ranks above General Telephone, we place a 0 at the juncture. Since no ranks are larger than Illinois Agricultural Association we also place a 0 in the inversion column. This process is continued for all organizations until Clay Dooley is reached. It has two ranks above it which are larger and a 2 is placed in the appropriate column. Finally, the sum of inversions is determined, in this case 4. These summed frequencies are substituted into the following formula:

$$G = \frac{\sum f_a - \sum f_i}{\sum f_a + \sum f_i} = \frac{6-4 = 2}{6+4 \quad 10} = .20$$

The correlation between the two sets of ranks is .20, a low positive association. Substantively, this appears to be a modest correlation between automation and job dissatisfaction. Gamma will vary between - 1.00 (a perfect negative association) and +1.00 (a perfect positive association). A value of zero would indicate no association between ranks.

Gamma is also a widely used statistic of association when ordinal level data appear in a contingency table format. Consider the data in Table 10.4.

TABLE 10.4
INCIDENCE OF CHEATING BEHAVIOR BY PRESSURE FOR SUCCESS[10]

		Pressure for Success			
		Low	Moderately Low	Moderately High	High
CHEATING BEHAVIOR	Cheated	3(20%)	6(22%)	5(25%)	13(81%)
	Possibly Cheated	1(7%)	5(19%)	5(25%)	1(6%)
	No Cheating	11(73%)	16(59%)	10(50%)	2(13%)
		15 100%	27 100%	20 100%	16 100%

Computationally, gamma is obtained with the following formula:[11]

$$G = \frac{n_s - n_d}{n_s + n_d}$$

where n_s = the number of same-ordered pairs
n_d = the number of different-ordered pairs

To obtain n_s, the number of same ordered pairs (sometimes called concordant pairs), we first locate the positive diagnol. The positive diagonal is the one in which high pressure and cheated coincide (n_{14}) and low pressure and no cheating coincide (n_{31}). To calculate the number of concordant pairs we multiply the frequency in each cell of the table by all frequencies above and to the right. Thus,

$$11(5+6+5+5+1+13) + 1(6+5+13) + 16(5+5+1+13) + 5(5+13) + 10(1+13) + 5(13) = 11(35) + 1(24) + 16(24) + 5(18) + 10(14) + 5(13)$$
$$= 385+24+384+90+140+65$$
$$= 1088$$

To obtain n_d, the number of different-ordered pairs (sometimes called discordant pairs), we first locate the negative diagonal. The negative diagonal is the one in which low pressure and cheating coincide (n_{11}) and high pressure and no cheating coincide (n_{34}). To calculate the number of discordant pairs we multiply the frequency in each cell of the table by all frequencies above and to the left. Thus,

$$2(5+5+5+6+1+3)+1(5+6+3)+10(5+6+1+3)+5(6+3)+16(1+3)+5(3)=2(25)+1(14)+10(15)+5(9)+16(4)+5(3)$$
$$=50+14+150+45+64+15$$
$$=338$$

Substituting n_s and n_d into the gamma formula we have:

$$\frac{1088-338}{1088+338} \quad \frac{750}{1426} = .52$$

Substantively, the correlation coefficient tells us that a moderate correlation ($\hat{r}$) exists between pressure for success and the incidence of cheating behavior. Cheating is more likely to occur when pressure is high and less likely to occur when pressure is low. This same observation can be inferred from examining the percentages in Table 10.4. By examing the percentages in the table's

most "extreme" cells we notice that 73% of the time low pressure produced no cheating whereas 81% of the time high pressure produced cheating behavior.

Gamma as a Proportional Reduction in Error(PRE) Measure.

The same generic PRE formula appearing earlier can be used for interpretating gamma.

Rule 1(for E_1). Since gamma involves predictions for pairs of cases, we want to predict whether a given pair is same-ordered(i.e., similar) or different-ordered(i.e., dissimilar) in terms of its rankings on two variables. By knowing only the dependent variable totals we are not certain how to predict the orders of pairs. In other words it may be best to assume that concordant and discordant pairs are equal in number. For gamma we predict all pairs to be either same-ordered or different ordered. However, theoretically, we would make errors about 50% of the time.

Rule 2(for E_2). With an additional variable cross-classified with a dependent variable we would predict same-order for all pairs if concordant pairs outnumbered discordant pairs.

Prediction Errors. With information on the order of one variable without knowledge of the order on another variable a random guess of order on the second variable is about the best we could do. However, in using this rationale the number of prediction errors ($\hat{r}$) would amount to one-half (or fifty percent) of the total number of n_s and n_d pairs. Since the total number of concordant and discordant pairs is 1426, and ½ of 1426 is 713, a total of 713 prediction errors would be expected. Thus, for rule 1 there would be 713 errors. With knowledge of a second variable we predict same order because the number of concordant pairs (1088) exceeds the number of discordant pairs(338). For rule 2 the number of prediction errors would be the smaller of the n_s or n_d pairs. For the present n_d is smaller than n_s; hence the number of prediction errors would be 338.

Definition of Measure. The proportional reduction in error achieved by using rule 2 rather than rule 1 is:

$$\frac{E_1-E_2}{E_1} = \frac{713 - 338}{713} = .52$$

Comments on Gamma and Related Statistics. Gamma represents the pair-by-pair comparison procedure (see endnote #1) for determining if two variables are associated. To fully comprehend the nature of gamma it is necessary to think in terms of pairs of cases rather than in terms of individual observations. Furthermore, gamma entails the computation of untied pairs only. In a contingency table of two ordinal level variables there are a variety of tied pairs. For example, there are pairs tied on X (the independent variable), Y (the dependent variable), and on X and Y (both the independent and dependent variables). Moreover, there exists the total number of pairs for the table as a whole. Let us illustrate the computation of these other pair types.

To compute the number of pairs tied on X, t_x, we multiply the frequencies in each column by the sum of all cell frequencies below them. Thus, for the data in Table 10.4:

$$t_x = 3(1+11) + 1(11) + 6(5+16) +5(16) + 5(5+10) +5(10) + 13(1+2) + 1(2)$$
$$= 36 + 11 + 126 + 80 + 75 + 39 + 2$$
$$= 419$$

Similarly, to compute the number of pairs tied on Y, t_y, we multiply the frequencies in each row by the sum of all cell frequencies across from them. Thus, for the data in Table 10.4:

$$t_y = 3(6+5+13) + 6(5+13) + 5(13) + 1(5+5+1) +5(5+1) + 5(1)+ 11(16+10+2) + 16(10+2) + 10(2)$$
$$= 72 + 108 + 65 + 11 + 30 + 5 +308 + 192 + 20$$
$$= 811$$

To compute the pairs tied on both X and Y, t_{xy}, we employ the formula: $n_i(n_i - 1)/2$ for each cell. Thus, for the data in Table 10.4:

$n_{11} = 3(3 - 1)/2 = 3$

$n_{12} = 6(6 - 1)/2 = 15$

$n_{13} = 5(5 - 1)/2 = 10$

$n_{14} = 13(13 - 1)/2 = 78$

$n_{21} = 1(1 - 1)/2 = 0$

$n_{22} = 5(5 - 1)/2 = 10$

$n_{23} = 5(5 - 1)/2 = 10$

$n_{24} = 1(1 - 1)/2 = 0$

$n_{31} = 11(11 - 1)/2 = 55$

$n_{32} = 16(16 - 1)/2 = 120$

$n_{33} = 10(10 - 1)/2 = 45$

$n_{34} = 2(2 - 1)/2 = 0$

$= 347$

Finally, the total number of pairs, T, for the data in Table 10.4 is obtained by $n(n - 1)/2$ or $78(78 - 1)/2 = 3003$.

When the various pair types are listed along with their respective values note that they sum to the total number of pairs in the contingency table:

$$n_s = 1088$$

$$n_d = 388$$

$$t_x = 419$$

$$t_y = 811$$

$$t_{xy} = 347$$

$$T = 3003$$

Various other measures of ordinal association (e.g., tau-a, tau-b, and tau-c) can be calculated from these combinations of pair types.[12] However, they are beyond the scope of this book.

Interval-Ratio Level Measures of Association

When both sets of observations conform to the assumptions for interval/ratio level measurement one of the most sophisticated and well-developed statistics of association, the Pearsonian product-moment correlation coefficient (r) can be computed. Figure 10.1 presents three different sets of bivariate data that will be employed to illustrate the nature of Pearson's r.

This coefficient measures the linear relationship between variables and assumes linearity in addition to the measurement level assumption noted above. To determine if the correspondence between the two data sets is approximately described by a straight line (i.e., a condition known as linearity or rectilinearity), a recommended practice is to construct a scattergram for the data. A scattergram is a graphic technique employed with bivariate data just like the polygon, histogram, and ogive were constructed for univariate data. In such instances the proper construction of this graph enables the analyst to visualize the distribution of scores, in the present case the joint distribution of scores.

To construct a scattergram scores on the Y variable are located along the ordinate and scores of the X variable are placed along the abscissa. Then a dot corresponding to the intersection of each (pair of X-Y coordinates) X and Y score is plotted. The overall configuration of dots in the scattergram permits an intuitive appreciation of the existence, direction, and degree of correlation between the variables being studied. Scattergrams and Pearson r values for each of the three hypothetical distributions have been completed and appear in Figure 10.1.

FIGURE 10.1
BIVARIATE DATA AND SCATTERGRAMS REPRESENTING PERFECT POSITIVE, PERFECT NEGATIVE, AND CURVILINEAR RELATIONSHIPS

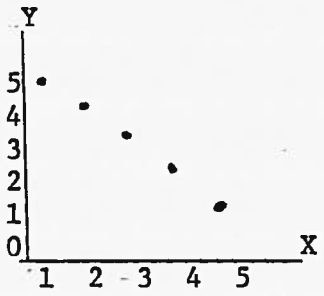
a.

X	Y
1	1
2	2
3	3
4	4
5	5

perfect positive:
r = +1.00

b.

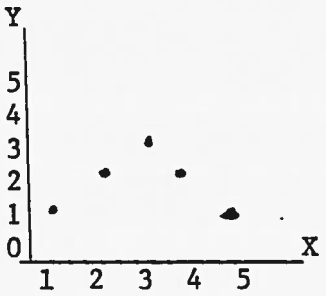
X	Y
1	5
2	4
3	3
4	2
5	1



perfect negative:
r = -1.00

c.

X	Y
1	1
2	2
3	3
4	2
5	1

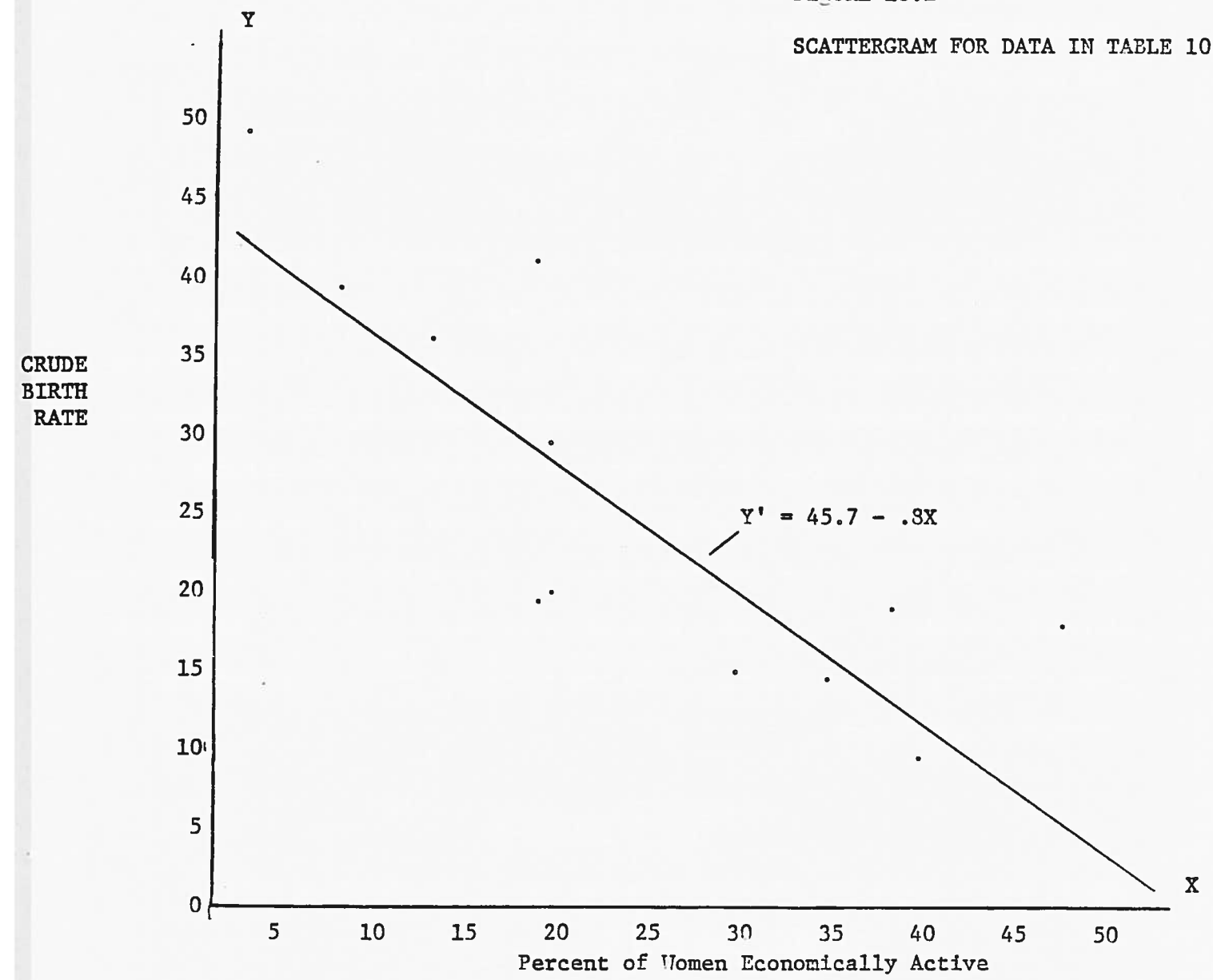


curvilinear
relationship:
r = 0.00

Example 1. The data in columns X and Y in Table 10.5 are the values for twelve nations on the percentage of women aged 14 and above who are economically active and the crude birth rate (defined as the number of births in a given year per 1000 of the population).[13] To determine if the data are linearly related we construct a scattergram, Figure 10.2. The plotted points provide evidence that X and Y are linearly related and, at the same time, indicate a negative slope and association.

FIGURE 10.2

SCATTERGRAM FOR DATA IN TABLE 10.5



To compute the correlation between the two sets of measurements both a conceptual and computational formula will be provided. The former formula enables one to think or conceptualize the mathematical rationale behind r while the latter usually facilitates computations when some electronic device (e.g., hand calculator) is used.[15] The "thinking formula" reads as follows:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N s_X s_Y} = \frac{\Sigma xy}{N s_X s_Y}$$

The numerator should have a familiar cast to it. Heretofore in computing the standard deviation we computed mean deviation values, that is, the extent to which a raw score deviates from the mean of its distribution. With r we first obtain the mean deviates for each score on each variable, multiply the mean deviates together, and finally sum the mean deviate products. The numerator is called the <u>covariation</u> and the <u>covariance</u> when divided by N. The denominator of the formula is nothing more than the number of observations (N) multiplied by the standard deviation of X (s_X) and the standard deviation of Y (s_Y). Performing these operations and substituting (see working format, Table 10.5) into the r formula we have the following correlation coefficient

$$r = \frac{1607.42}{(12)(12.95)(12.08)} = -.856$$

Let us double check the conceptual formula with a computational formula:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\left[N\Sigma X^2 - (\Sigma X)^2\right]\left[N\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

$$r = \frac{12\,(6022) - (287)(319)}{\sqrt{\left[12(8875) - (287)^2\right]\left[12(10{,}231) - (319)^2\right]}} = -.857$$

The two computed r's are, as expected, virtually identical.

TABLE 10.5

BIVARIATE DATA FOR PERCENT OF WOMEN ECONOMICALLY ACTIVE(X) AND CRUDE BIRTH RATE (Y) FOR 12 NATIONS[14]

Nation	X	Y	X^2	Y^2	XY
Algeria	2	48	4	2304	96
Argentina	19	21	361	441	399
Denmark	34	14	1156	196	476
East Germany	40	11	1600	121	440
Guatemala	8	41	64	1681	328
India	12	37	144	1369	444
Ireland	20	22	400	484	440
Jamaica	20	31	400	961	620
Japan	37	19	1369	361	703
Phillipines	19	42	361	1764	798
United States	30	15	900	225	450
USSR	46	18	2116	324	828
Total	287	319	8875	10,231	6022

$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N = 8875 - (257)^2/12 = 2010.92$$

$$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/N = 10,231 - (319)^2/12 = 1750.92$$

$$\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/N = 6022 - (287)(319)/12 = 1607.42$$

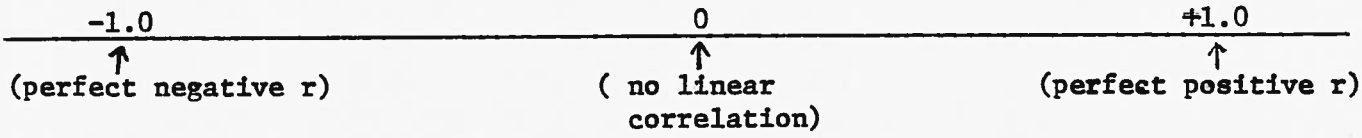
$$s_X = \sqrt{\frac{\Sigma x^2}{N}} = 12.95$$

$$s_Y = \sqrt{\frac{\Sigma y^2}{N}} = 12.08$$

$$\bar{X} = \Sigma X/N = 287/12 = 23.92$$

$$\bar{Y} = \Sigma Y/N = 319/12 = 26.58$$

To interpret r we have recourse to several guidelines. The Pearson product-moment correlation coefficient measures the degree of linear relationship between variables. It is possible for a correlation to exist and r to be nil when the scatterplot is curvilinear. One can imagine the computed r to fall at some point along the correlational scale that runs the gamut from -1.0 through 0 to +1.0. The closer to 1.0, regardless of sign, the more perfect the degree of linear fit. Although perfect correlations are rare as are r's of 0.0, a perfect positive association would be denoted by a sign and value of +1.0 and a perfect negative association by -1.0 with no linear relationship showing a value of r = 0.0.



Another interpretation involves the square of r (r^2) rather than the raw correlation coefficient and is called the proportional reduction in error interpretation. When r is squared the resulting value is known as the coefficient of determination and indicates how much of the variation in Y (dependent variable) is explained by X (independent variable). In short, it provides insight into the explanatory power of the presumed causal variable. When variables are not construed in a causal framework as is the case with height and weight, the former interpretation is probably more salient. When unity is subtracted from r^2 (i.e., $1-r^2$) a concept called the coefficient of non-determination is produced. This latter coefficient indicates how much of the variation in Y is not accounted for by X, or by inference, how much variation in Y is attributable to other factors not included in the analysis (to determine and assess the contribution of other factors takes us into the multivariate realm). Again, this feature is most applicable when cause-effect connections are being explored. For illustrative purposes, the coefficients of determination and non-determination for the data in Table 10.5 are $(-.857)^2=.73$

(or 73% when multiplied by 100) and .27 (or 27% when multiplied by 100), respectively. Note that the two proportions total 1.00 and, similarly, the two percentages total 100.

Simple linear Regression. Whereas the Pearsonian r measures the degree and direction of the correlation between variables, regression analysis enhances understanding the form of the relationship. The concept of regression implies prediction, predicting the values of the dependent variable from a knowledge of the values of the independent variable. Linear implies the two variables can be described, at least roughly, by a straight line as opposed to a curved line in which case curvilinear regression would be appropriate.

Linear Functions. Given the goal of regression, there are several additional formulae which could be used to describe how the dependent variable Y changes as a function of the independent variable X. Here we confine ourselves to the simplest class of such mathematical formulae, those corresponding to linear functions.

> The formula $Y' = a + bX$ expresses the dependent variable Y as a linear function of the independent variable X, with a slope b (beta) and Y-intercept a (alpha). a and b are referred to as regression coefficients and are constants for a given data set.

The algebraic equation for a linear function (straight line) is $Y' = a + bX$ where Y' stands for the value of the dependent variable one is predicting, a is the Y-intercept, the point at which the regression line intercepts the Y axis (also the value of Y when $X = 0$), and b is the regression coefficient representing how much a change in Y is produced by a unit change in X (also called the slope value of the regression equation).

The task becomes one of computing the two regression constants a and b. Several computational formulae are available but to show the correspondence between correlation and regression we will use formulae that enable us to sub-

stitute the values already computed in Table 10.5.[16] Therefore, for determining the Y-intercept and regression coefficient the following formulae will be used:

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{1607.42}{2010.92} = -.80$$

$$b = \bar{Y} - b\bar{X} = 26.58 - (-.80)(23.92) = 45.7$$

The regression equation for these data would be expressed as:

$$Y' = 45.7 - .8X$$

This is plotted on the regression line in Figure 10.2. A word on interpretation is in order. Because the slope (b) is negative, the relationship between the two variables is also (this consistency must always be the case). In general, the larger the percentage of females economically active, the smaller the crude birth rate tends to be. Specifically, b = -.8 indicates that (on the average) an increase of one in the percentage of economically active women corresponds to a decrease of .8 in the crude birth rate. This implies that if one nation has 20% of its females in the labor force and another has 30%, the first nation has 8 more births per 1000 (10 x .8 = 8) population.[17]

How well does the prediction fit the data? Consider Algeria with an actual crude birth rate of 48. The prediction equation is Y' = 45.7 - .8X where X = 2. Hence, 45.7 - 1.6 = 44.1. The prediction error is the differences between actual and predicted values. For Algeria the prediction error is 3.9. The prediction errors are commonly referred to as residuals. A "positive" residual is one in which the prediction is too small, a "negative" residual is a prediction too large. Notice that the algebraic sum of residuals is zero (or nearly so). Table 10.6 contains all residual errors for the twelve nations. The smaller the absolute value of the residuals (or the smaller the sum of the squared residuals) the better the prediction. Graphically, the residual for an observation can be represented by the vertical distance between the actual observation and the regression line. Figure 10.2 displays this notion.

TABLE 10.6

PREDICTIONS ON CRUDE BIRTH RATE AND CORRESPONDING RESIDUALS

Nation	Percent of Economically Active Women X	Crude B/R Y	Predicted Crude Birth Rate $Y'=45.7-.8X$	Residual $Y-Y'$	Residual $(Y-Y')^2$
Algeria	2	48	44.10	3.90	15.21
Argentina	19	21	30.51	-9.51	90.44
Denmark	34	14	18.52	-4.52	20.43
East Germany	40	11	13.73	-2.73	7.45
Guatemala	8	41	39.31	1.69	2.86
India	12	37	36.11	.89	0.79
Ireland	20	22	29.71	-7.71	59.44
Jamaica	20	31	29.71	1.29	1.66
Japan	37	19	16.13	2.87	8.24
Philippines	19	42	30.51	11.49	132.98
United States	30	15	21.72	-6.72	45.16
USSR	46	18	8.93	9.07	82.26
Total				$\Sigma = 0$	$\Sigma = 466.92$

Method of Least Squares. There is a residual for each observation in a data set. The algebraic sum of all residuals (i.e., $\Sigma (Y_i - Y')$) equals zero (see column 5 of Table 10.6). The usual way to summarize the size of the residuals is to calculate the sum of squared prediction errors.[18] This quantity, denoted by SSE, is produced by the formula:

$$SSE = \Sigma(Y_i - Y_i')^2$$

In short, for each score the residual is found, squared, and the SSE is computed by summing all squared residuals. The measure SSE is referred to as the error sum of squares or the residual sum of squares. The better the prediction equation the smaller the residuals and the smaller the summary measure SSE tends to be. The prediction equation here is the one with the smallest value of SSE out of all possible linear prediction equations. The criterion used in choosing the best prediction equation is the one which yields the smallest sum of squared prediction errors. In Table 10.6 (column 6) the residuals are squared and we obtain SSE = 466.92. This value is the smallest value produced by any prediction equation.

Pearson r. The value of r can be interpreted as a "standardized slope." It will vary between -1.0 and +1.0 and does not depend on the units of measurement (e.g., pounds, ounces, grams; yards, feet, inches, etc.). The standardization is accomplished by multiplying the b value by the standard deviation ratio of X and Y. r is related to the slope by the formula:

$$r = \left(\frac{s_X}{s_Y}\right) b$$

In the special case where the standard deviations of X and Y are equal, r = b.

The value of the Pearson r can be obtained using a variety of formulae (see endnote 15). Using the one above:

$$s_X = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{2010.92}{12}} = 12.95$$

$$s_Y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{1750.92}{12}} = 12.08$$

$$r = \left(\frac{12.95}{12.08}\right) (-.80) = -.857$$

The Pearson r as a Proportional Reduction in Error Measure.

The generic PRE formula reads:

$$\frac{E_1 - E_2}{E_1}$$

Rule 1 (for E_1). Suppose we know the distribution of Y values without knowing which Y value corresponds to a specific observation of X. The best predictor would be $\bar{Y}$, the mean of the Y variable because the mean possesses that property around which the squared deviations will be minimal (i.e., $\Sigma (Y_i - \bar{Y})^2$ = minimum).

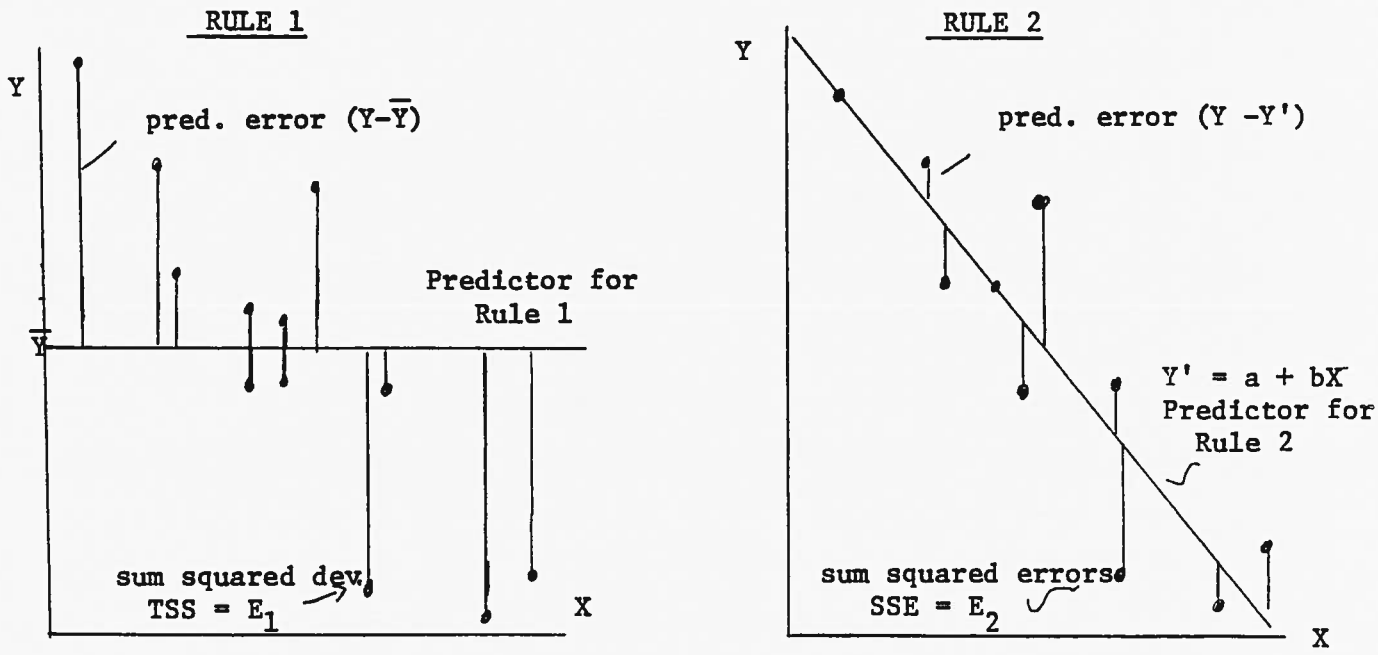
Rule 2 (for E_2). If we knew the relationship between X and Y, then the best predictions for the Y values would be those using the prediction equation: $Y' = a + bX$. For each observation we could substitute the approximate X value to obtain the predicted Y value.

Prediction Errors. For rule 1 we could obtain prediction errors by subtracting each Y value from the mean of Y, square the mean deviates and sum the squared mean deviates. In formula form, $\sum (Y_i - \bar{Y})^2$ = errors by rule number 1 (this is called the total sum of squares, abbreviated "TSS"), or $(48 - 26.58)^2 + (21 - 26.58)^2 \ldots + (18 - 26.58)^2$. Or, more simply, we could obtain the total sum of squares via: $\sum Y^2 - \frac{(\sum Y)^2}{N}$ or $10,231 - \frac{(319)^2}{12} = 1750.92$.

For rule 2 we obtain SSE by subtracting the predicted values (Y'_i) from the actual values (Y_i), square and sum. For the present data this would amount to: $\sum (Y_i - Y')^2$ or $(48 - 44.10)^2 + (21 - 30.51)^2 \ldots (18 - 8.93)^2$. This operation would produce a SSE of 466.92. Graphically, the computation of TSS and SSE is represented in Figure 10.3.

FIGURE 10.3

GRAPHIC REPRESENTATION OF RULE 1 AND E_1, RULE 2 AND E_2



Definition of measure. The proportional reduction in error achieved by using the prediction equation instead of $\overline{Y}$ is called the coefficient of determination and denoted by r^2. The PRE formula for r is:

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{TSS - SSE}{TSS}$$

Substituting the present values into the above formula yields $r^2 = .733$, or

$$\frac{1750.92 - 466.92}{1750.92} = .733$$

We interpret $r^2 = .73$ as follows: Using the prediction equation $Y' = 45.7 - .8X$ the amount of error (as measured by the sum of squared errors) is 73% smaller than when $\overline{Y}$ is used as the predictor. Equivalently, the amount of error using the prediction formula is only 27% as large as the amount of error using $\overline{Y}$ as the predictor (TSS = 1750.92; SSE = 466.92; 466.92/1750.92 = .27).

Summary

Descriptive statistics for bivariate distributions have been discussed in this chapter. To understand the statistics of relationships we began with a consideration of the most intuitively grasped procedure, that of percentaging the table in the direction of the independent variable and comparing across categories of the dependent variable. The resulting value--epsilon or the percentage change index--indicates whether there is an association but is limited insofar as it fails to provide an exact numerical indicator of the magnitude of the correlation. While the epsilon procedure is a convenient first step its limitations become particularly noticeable when researchers deal with tables containing more than 2 x 2 dimensions. To overcome these liabilities indices of association for the entire table are desirable.

A family of correlational statistics based upon chi square were addressed. These measures are premised on the basis of a model of no association, that is, the obtained frequencies are systematically compared with those frequencies expected if there were no relationship between the variables. Both the phi coefficient and Cramer's V were computed and interpreted as chi square or delta-based measures of association. Another nominal level coefficient--Yule's Q--was discussed as was lambda. Since phi and Q are based upon different conceptions of perfect relationships, special attention was paid to two distinct models of perfect relationships:

1) the stringent model, and 2) the less stringent model.

For ordinal level data two representative indices of association--Spearman's rho and Goodman and Kruskal's gamma were presented. Examples entailing computation and interpretation of the respective statistics were included. The elegance of these statistics resides in their proportional reduction in error interpretation. The logic of the PRE interpretation was unfolded, particularly for gamma computed from data in a contingency table.

The Pearson product-moment correlation coefficient (r), another PRE statistic, is ideally suited to use with interval-ratio level data. Since this coefficient measures the linear relationship between variables and assumes linearity, guidelines for assessing this property of data sets were considered. Specifically, a scattergram, a graphic device for bivariate data, permits the analyst to judge the nature of the data. Calculating and interpretating r (using PRE procedures) followed an illustration.

Since simple regression is an extension of correlation, the linear regression equation--Y'= a + bX--was determined and interpreted for a given set of data. The manner in which the regression equation is used for predictive purposes was highlighted.

Finally, since methodologists have derived and refined numerous statistics of relationships it was necessary to be selective rather than exhaustive in treatment. Coefficients of association with the PRE interpretation were given special attention.

Important Concepts Discussed in This Chapter

Epsilon
Principle of the Joint Occurence of Attributes
Principle of Covariation
Model of No Association
Chi Square
Observed Frequencies
Expected Frequencies
Delta
Phi
Cramer's V
Yule's Q
Lambda (coefficient of predictability)
Gamma
Concordant Pairs
Discordant Pairs
Pairs Tied on X
Pairs Tied on Y
Pairs Tied on X and Y
Total Number of Pairs
Pearson's r
Scattergram
Coefficient of Determination
Coefficient of Non-determination
Simple Linear Regression

The Stringent Model of Perfect Correlation

The Less Stringent Model of Perfect Correlation

Spearman's rho

a (Y intercept)

b (slope)

Proportional Reduction in Error

Chapter 10 Endnotes

[1]There are several procedures for determining if two variables are related to one another. Robert Weiss (Statistics in Social Research, N.Y.: Wiley. has suggested five general procedures for establishing relationships between two variables. They are:

1) Departure from independence between two variables. By constructing a model of no association (i.e., determine what the data would look like if no relationship existed)and comparing the empirical distribution with it, one can determine if an association exists.

2) Magnitude of subgroup differences. Assuming a cross-classification of data one can determine if an association exists as well as its magnitude by direct comparisons of subgroup proportions or percentages (e.g., epsilon or the percentage difference value).

3) Summary of pair-by-pair comparisons. Another procedure entails forming all possible comparisons of one member of the sample with another. In each of these comparisons one must decide whether the two factors occurred together or not. When all results of the pair-by-pair comparisons are made the association would be measured by the preponderance of concordant (same-ordered) or discordant (different ordered) pairs.

4) Proportional reduction of probable error. We first determine the number of prediction errors by knowing only the marginal totals of the dependent variable. Then, with another variable, we determine the number of prediction errors also. The more we are able to reduce prediction errors with the added knowledge of the second variable the stronger is the association between the two.

5) Extent to which increments in one variable occur together with increments in another variable. We take as our measure of association the extent to which increases in one variable are accompanied by increases in the other, or decreases in one by decreases in the other.

A Guide for Interpreting Coefficients of Association: Guilford's Table

*Magnitude of Raw Coefficient**	*Degree of Relationship*
Less than or equal to ±.20	Slight, almost negligible
±.21 to ± .40	Low correlation, definite but small
±.41 to ± .70	Moderate correlation, substantial
±.71 to ± .90	High correlation, marked
±.91 to ±1.00	Very high and dependable

*This assumes the coefficient is *statistically significant*!