Chapter 5
Statistical Analysis of Cross-Tabs

D. White and A. Korotayev
2 Jan 2004
Html links are live

Introduction

Descriptive statistics includes collecting, organizing, summarizing and presenting descriptive data. We assume here, as with the Standard Sample, that the collection, data cleaning, and organization of data into variables has been done, and that the student has access to a database through a program such as SPSS, the Statistical Package for the Social Sciences. Further, we will assume that the student has instructions on using the software to create cross-tabulations of variables.

Inferential statistics includes determining relationships using correlation coefficients and testing hypotheses with significance tests. These may be used for evaluating predictions by comparison to the null hypothesis and for the similarity between two predicted outcomes (comparison of a theoretical model to an expected outcome from the model, testing whether two sets of observations give the same result, and so forth). Those will be the primary concern of this Chapter, and other issues of statistical inference will be taken up in Chapter 7 and 8.

Cross tabulation of qualitative data is a basic tool for empirical research. Cross tabulations (cross tabs for short) are also called contingency tables because they are used to test hypotheses about how some variables are contingent upon others, or how increases in one affects increases, decreases or curvilinear changes in others. Problems of causal influences or feedback relationships are difficult to make, of course, without experimental controls or data over time. Contingency analysis, however, is a good place to begin in testing theories or developing hypotheses to be tested with more rigorously collected data. The use of control variables in studying correlations can also be of use in replicating results and identifying more complicated contingencies by which variables interact or influence one another.

Our goal is to help students to obtain and understand the statistics they need in doing empirical research using cross tabulations of variables that are available for analysis of observational samples, notably in the social sciences, using here our examples from the Standard Cross-Cultural Sample.

Section 1 provides practical advice for contingency table analysis in SPSS. A more general introduction to statistical analysis proceeds in the three sections that build one on the other. The student is recommended to study these sections because they provide the basis for statistical reasoning. To deal with your data analysis rather than simply apply correlational and significance tests mechanically without understanding them, it will be invaluable to study and understand the concepts of statistical reasoning in order to reason

from them rather than from ad hoc interpretation of the mechanical procedures in section 1.

Section 2 introduces measurement (nominal, ordinal, interval and ratio scales) and correlation, which are closely connected. Basic methods are presented for getting useful correlations from nominal and ordinal data.

Section 3 takes up those topics in statistics that derive their analytical power from the use of probability theory. We begin with probabilistic inference and the three laws of probability (independent events, sample spaces, and mutually exclusive events). From these we derive expected frequencies and the null hypothesis of statistical independence. We then explain how from a comparison of expected and actual frequencies for cross-tabulations on our data we can derive two useful statistics: the chi-square measure of departure from statistical independence and the phi-square all-purpose correlation coefficient.

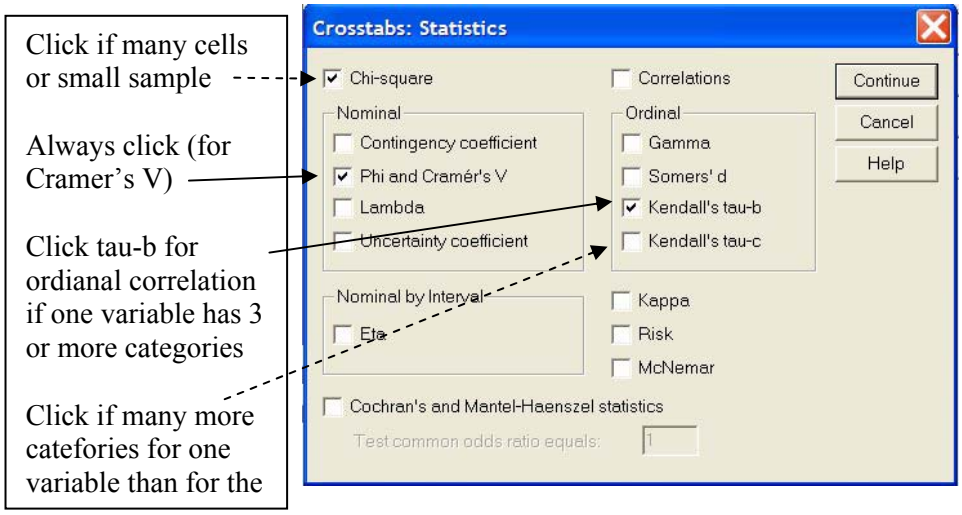
Section 4 unites the two previous sections. Interpreting correlations derived from cross-tables and the testing of hypotheses from them requires the concepts in statistical analysis that derive from probability theory reviewed in the previous section. When strictly independent events having two characteristics that are independently defined are tabulated in a contingency table, the laws of probability can be used to model, from the marginal totals (rows, columns) of the table, what its cell values would be if the variables were statistically independent. The actual cell values of the frequency table can be used to measure the correlation between the variables (with zero correlation corresponding to statistical independence) but they must also be compared to expected values under the null hypothesis of statistical independence. In this section we show how expected frequencies are used to give a significance test or estimate of the probability that the departure of the observed correlation from zero (statistical independence) is simply a matter of chance given the laws of statistical independence. For every correlation, we may and should compute the proper tests of significance.

The conclusion reviews and summarizes the relevance of basic statistics for empirical research on sample data. Some on-line calculators and references are recommended for student use. Some important statistical topics, however, are deferred to Chapters 7 and 8. Chapter 7 takes up one factor hypotheses about reliability of our measures (one factor models of multiple variables) and third factor hypotheses that offer the close approximation to replicating an experiment. Replication is a desideratum of all scientific work. You cannot do without thinking about design effects in how samples are constructed, or proper controls for nonindependence of the sample cases (Galton's Problem of autocorrelation); topics that are reviewed in Chapter 8. For example, when the sample of observations departs from strict independence because of network interactions between them, the correlations between interacting neighbors measured on the same variables can be used to deflate effective sample size in getting accurate significance tests.

Section 1: Practical Advice for Contingency Tables in SPSS

Correlation Coefficients: Phi´ (Cramer's V) and Tau-b

We recommend the use of tau-b for ordinal correlations and for nominal correlations either Cramer's V or phi, the latter if one variable has only two categories. Chi-square is optional and is useful mostly for understanding how phi or adjusted phi (Cramer's V) is calculated.



The advantage of using Cramer's V and tau-b is that when the numbers of categories of the row and column variables are roughly equal, they are measured more or less on the same scale (bounded by -1 and +1, although tau-b cannot reach these values the more the inequality in number of categories). There are three possibilities where one or both are significant.

a. tau-b ~ 0 (non-significant) and Cramer's V is significant: There is no ordinal trend but some nominal correlation.

b. tau-b is weaker than Cramer's V and both are significant: There is some ordinal trend but also some additional nominal correlation.

c. tau-b ~ Cramer's V (they are roughly equal) and both significant: There is only and ordinal trend.

If there is significant nominal correlation in excess of the ordinal, i.e., either alone or in addition to ordinal, then there is some prediction to be gained category by category that is independent of ordinal prediction. Since nominal correlation may in general include ordinal correlation there is no possibility that Cramer's V is significantly weaker than tau-b.

Significance Tests

In SPSS

Significance tests come automatically once your correlation coefficients have been chosen. Since Cramer's V and tau-b are symmetric measures of correlation, they are shown in the SPSS output under the *value* column, and significance is shown in the rightmost *Approximate Significance* column. Here, to show more decimal places for the significant measure, we have right-clicked this table and selected SPSS Pivot Table Object, Edit, clicked the two significance measures in which we are interested, and then Cell Properties, and increased the decimal precision. We did not bother with doing this for tau-c since its significance always equals that of tau-b. Significance for Phi and Cramer's V is also equal also. Whenever Phi and Cramer's V have different values, you must always take Cramer's V as the relevant coefficient. In this case there were 5 row categories and 6 column categories, which is by Phi diverged from Cramer's V and tau-b diverged from tau-c. Note that in this case tau-c is *lower* than tau-b, which has to do with the severity of the normalization. This is the reason we do not recommend tau-c.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.557			.000
	Cramer's V	.278			.000035
Ordinal by Ordinal	Kendall's tau-b	.365	.052	6.300	.00000000030
	Kendall's tau-c	.270	.043	6.300	.000
N of Valid Cases		179			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

When interpreting these results according to rules (a)-(c) for comparing tau-b and Cramer's V, the fact that tau-b is equal or slightly greater (in this case greater) than Cramer's V leads us to the conclusion that the correlation between these two variables is ordinal. The variables correlated here were percentage dependence on hunting (row variable) and gathering (column variable). Although these cannot sum to more than 100% neither variable has very many cases where this percentage exceeds forty and they are otherwise strongly positively correlated with high significance.

With small samples statistical significance may be difficult to assess in SPSS. If one or more cells of the cross-tabulation have an expected count less than 5, the significance calculation for Cramer's V (and phi) will be invalid. This occurs if some of the row and column totals of the table are small. If you are not sure whether this is the case then click the choice for Chi-Square tests, which will tell you how many expected cell counts are less than 5, as in the case below. If there is one or more cells with and an expected frequency less than 5, and neither variable has more than 6 categories, you may use a web-based calculator for Fisher's Exact significant test.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.069[a]	3	.558
N of Valid Cases	11		

a. 8 cells (100.0%) have expected count less than 5. The minimum expected count is .45.

Web-based calculators for Fisher Exact Test of Significance

For 2x2 tables and n $\leq$ 100, use http://faculty.vassar.edu/lowry/fisher.html. If n > 100 and tables up to 2x5, you may use http://home.clara.net/sisa/ord2.htm. Otherwise, use http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html for a table up to 6x6 and no cell value $\geq$ 100. The procedure is simply to enter your table cell values into the appropriate cells in the calculator.

Recoding and k-cotomizing

SPSS allows you to recode variables and to dichotomize, tricotomize or, in general, k-cotomize one or both of your variables. This may have the advantage of reducing the number of categories to fit into a simpler table, or collapsing categories that have similar percentage distributions relative to another variable. You can also do this collapsing of categories by hand, creating new cells by adding up values in the cells to be combined. When this is done, you may use the web-based calculators for Fisher Exact test of significance and the following web-based calculator for Cramer's V and tau-b.

Web-based calculator for Cramer's V and Tau-b

Cramer's V: http://faculty.vassar.edu/lowry/newcs.html.
Tau-b: http://members.aol.com/johnp71/ordinal.html.

Section 2: Measurement and Correlation

Measurement

Nominal, Ordinal, Interval and Ratio Scale Variables

An nominal variable is a classification of events into mutually exclusive and exhaustive classes. An ordinal variable is a classification of events into mutually exclusive and exhaustive classes that are rank-ordered by some criterion. An interval variable is a classification of events by a numeric measurement in which one unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole scale. Operations of addition and multiplication by constants do not change the interval property of the scale. Such transformations are needed, for example, in

converting centigrade to Fahrenheit temperature. Here we multiply ^{o}C by 1.6 and add 32 to get ^{o}F, or subtract 32 from F^{o} and divide by 1.6 to get ^{o}C. Twenty degrees is not twice as hot, however, as ten. That property is the characteristic of a ratio scale such as the Kelvin scale of temperature in which O^{o}K is absolute zero, a true zero point (-459^{o}F, -273^{o}C) and 20^{o}K is twice the heat (measured by motion of molecules) as 10^{o}K.

We consider a few examples of nominal, ordinal, interval and ratio-scale variables, starting with the ratio-scale. These are variables with which you dealt mostly at school. These are such variables as length measured in meters; mass measured in kilograms, power measured in watts, or resistance measured in ohms. An important feature of such variables is that you can use all the mathematical operations with them, and any of the tests employed in applied statistics. For example, if the height of a tree 10 years ago was 4 meters, and now it is 8 meters, we have the grounds to divide 8 by 4, getting two; and to conclude that in 10 years the height of this tree has doubled.

We noted the difference between the ratio and interval scale with the example of Kelvin vs. Fahrenheit or Celsius temperature scales. The difference is not a mere scientific convention. When we say that the growth of temperature from 10^{o}K to 20^{o}K means the doubling of temperature, this has important implications for reality. For example, according to one of the physical laws of gases, the doubling of a gas's temperature should lead to the doubling of its pressure too (given that its volume does not change). If we expect that the doubling of gas's temperature from 10^{o}C to 20^{o}C will lead to the doubling of its pressure, we will be thoroughly disillusioned (the growth of pressure will be just under 4%). A correct prediction must use the Kelvin scale.

In the social sciences you will rarely confront the difference between these two types of variables unless the variables you are using deal with frequency of events (a ratio measure), density or percentages, where true zero points are apparent. In further explanations we will not distinguish between interval and ratio-scale variables, denoting both of them as interval as is customary and dealing with true zero points as something extra. For us the difference between interval and ordinal variables is of much more importance. A typical example of an ordinal variable is the school grading system. Take, for example, the US school grading system: A, B, C, D, E, F. It easy to assign numbers to grades, as in A = 6, B = 5, C = 4, D = 3, E = 2, and F = 1. The number of mathematical operations that apply to these numbers is much more restricted. If John had an F = 1 grade in mathematics in the beginning of the first year, D = 3 at the beginning of the second, and C = 4 at its end, the application of a calculation such as (3 – 1) / (4 – 3) = 3 to conclude that in the last year the growth of John's mathematical knowledge slowed down by a third would not make sense.

Ordinal and interval scales could be used to measure the same parameters. As exemplified in Table 5.0, using the intervally measured population density data we can find that in culture H population density is higher than in culture E by 18 persons per sq. mile (20-2), while in culture E population density is higher than in culture D by 1.5 persons/mi² (2-.5). The first interval is nine times that of the second, a valid interval property. If we wanted to use the true zero point in a density scale, the population density

in culture E is one-tenth that of culture H, and 4 times that of D. None of this we can find using ordinally measured data, and it is not difficult to see that conclusions drawn from the numbers in the light and dark columns might turn out to be rather different. We can divide "4" (population density rank value of culture H) by "2" (population density rank value of culture D), or substract 2 from 4, but the resultant 2 will not tell us anything real about how much higher is the first population density than the second (it is actually higher by 19.5 persons\ mi² mile, or 40 times).

Table 5.0

Culture Identification	Population Density		Average Community Size	
	Measured on Ordinal Scale[1]	Measured on Interval Scale (persons per sq. mile)	Measured on Ordinal Scale[2]	Measured on Interval Scale (persons)
A	1	0.01	1	20
B	1	0.15	1	40
C	2	0.3	2	60
D	2	0.5	2	90
E	3	2	3	110
F	3	4	3	180
G	4	8	4	230
H	4	20	4	350
I	5	30	5	450
J	5	90	5	940
K	6	120	6	1,200
L	6	480	6	4,700
M	7	520	7	5,100
N	7	1,140	7	10,200

The only thing that ordinally measured data let us know about relative population density in cultures D, E, and H, is that in culture H it is higher than in E, and in D it is lower than in E and H. D does not differ much from E, however, in terms of the interval between them. One may even wonder if one can do serious calculations with such ordinal numbers, or if one can extract anything important from such data. As we will see below, however, one can extract even from such data amazingly large amounts of useful information.

[1] 1 = < 1 person per 5 sq. mile
2 = 1 person per 1-5 sq. mile
3 = 1-5 persons per sq. mile
4 = 6-25 persons per sq. mile
5 = 26-100 persons per sq. mile
6 = 101-500 persons per sq. mile
7 = over 500 persons per sq. mile

[2] 1 = < 50
2 = 50-99 persons
3 = 100-199 persons
4 = 200-399 persons
5 = 400-999 persons
6 = 1,000-4,999 persons
7 = > 5,000 persons

Serious calculation is even possible with nominal data, grouped into discrete, unordered categories. As an example (see Korotayev, 2003), we may cite religious (or linguistic) affiliation of cultures, e.g., "Christian," "Islamic," "Buddhist," "Hindu," etc. (or "Indo-European," "Altaic," "Uralic," "Dravidian"). Note that such data will be still normally coded with numbers (e.g. 1 = "Indo-European," 2 = "Altaic," 3 = "Uralic," 4 = "Dravidic"), however, these numbers are totally arbitrary and, unlike numbers used to code ordinal data, do not contain useful information per se. With ordinally measured population density, knowing that a society has population density rank or rank value 3 informs us that its population density is higher than the one of all the societies with rank 2, and lower than in societies with rank 4. Linguistic affiliation is a nominal variable. Hence, if we know that the language of a given culture is Uralic (coded as "3"), this does not let us know that the respective society has a characteristic which is somehow lower than in cultures coded as "4" (Dravidic). Even to such data, however, it is possible to apply some meaningful statistical methods.

A special class of nominal data is constituted by dichotomous variables. We would like to stress that interval and nominal variables can be easily re-coded into dichotomous ones. Such variables have only two values. For example, we can easily re-code the nominal data on religious affiliation, denoting all the Christian cultures as "1 = christianization present," designating all the other cultures as "0 = christianization absent." A classical example of a dichotomous variable that does not usually require recoding is gender. It has only two categories – "female" and "male." Again, if we know that a certain person is not male, we can be reasonably sure that she is female; and conversely if we can be reasonably sure that he is male; hence the dictotomy. The most important feature of dichotomies, however, is that these nominal variables can be considered as ordinal because they can always be ranked (1,2) and they are interval by default because there is only one interval. "Male – female," for example, can be regarded as a scale having just 2 ranks and one interval. We can apply to the analysis of such variables all the methods for ordinal, interval and nominal data. Such variables can be coded using numbers "1" and "0" to denote the presence of an attribute (e.g., "femaleness") and "0" its absence or opposite ("maleness"). Such coding is arbitrary, so that coding "male" as "1." and female as "0" would make equal sense.

Correlation

Functional and Correlational Relationships

Many humanists are certain that no scientific laws could be discovered in the human social world. How can humans possessing free will be subject to laws of science determining their behavior? To our mind, this misunderstanding can be partly accounted for by the fact that the mathematics taught at secondary school are mostly those of functional relationships. The notion of function in mathematics denotes a relationship between variable X and Y such that one and only one value of variable Y corresponds to

any one value of variable X.[3] The mathematics describing functional relationships are extremely important for natural sciences, especially in physics. For example, we can determine the current (amperage I) in an electric circuit from the values of voltage (V) and resistance (R) according to the Ohm's Law (I = V/R), and if we hold R constant, then V is a function of I.

For the human world, however, the relationships of mathematical functions are not so useful. We can even say that the humanitarians' observation about human free will is not entirely without grounds. For example, Ember and Ember (1971) show that the presence of internal warfare strongly predicts patrilocal residence, defined as residence after marriage with the family of the groom. But imagine that you live in such a society with a high level of internal warfare. Does this mean that you personally are bound to choose just patrilocal residence? Of course not. If (for whatever reasons) you decide firmly to bring your husband to live with your parents, and apply enough energy to do this you will successfully "break" the regularity of this pattern for patrilocality. But will this mean that the above mentioned regularity does not exist? No, it does not. In the context of high internal warfare it will be in the interest of parents to retain their sons in their household as a means of defense (and this may come to be supported by tradition and public opinion) rather than let them disperse when they marry. Although you for your special reasons may be able to arrange for your husband to leave his parents and come to live with you in a separate household, we would still have all grounds to expect that in the context of high internal warfare most couples might still live patrilocally. No residence rule is followed by everyone and you would be one of the exceptions. Note the two levels here: one is choice at the individual level, the other a social pattern in which all these various choices might be aggregated. If more people opted as did you and your husband (or similarly, if you are the husband and you went to live separately from your parents and brothers), the social pattern would change accordingly.

Notice too that because it is warfare that tends to predict patrilocality we have taken for granted in this argument that this was a determination of causality. We have implicitly also given the example of choice in this case as a willful exertion to "break" this regularity. The temporal relationship between the two variables might be different if patrilocality develops first, and having groups of related males in different local groups tends to make conflicts escalate into patterns of internal warfare. There is a well established and tested theory for this explanation, called fraternal interest theory, which predicts that patrilocal groups are more likely (a *probability* rather than a determination) to engage in feuding. Some patrilocal societies manage to avoid such conflicts, but others do not. Choice might well be entailed at any stage: whether a majority of people decide to reside patrilocally and, given that, whether they manage to avoid feuding between groups of males, or if not, whether they prevent feuding from escalating into internal warfare. No strict determination is involved if the argument is probabilistic. The first choice (where to reside) may alter the social environment in such a way that a certain outcome becomes more probable (feuding). That result is not determined, but is still only a contingency.

[3] While some function f(x) is always uniquely defined, there may be different values of x that give the same result, e.g., $f(x_1)= f(x_2)$.

We would not call Ember and Ember's finding an anthropological "law," as they do. What they have found is that patrilocal residence tends to be a necessary condition of internal warfare but not a sufficient one, and equivalently, internal warfare tends to be a sufficient condition of patrilocal residence but not a necessary one. Before you claim to have found a law you might want to ascertain the temporal ordering and actual evidence or logic as to cause and effect. Because consequences follow causes, you might want to explore the necessary and sufficient conditions for one or another variables, you might want to see if the relationship replicates under changeable conditions, and so forth.

To repeat, even if according to some stronger socioanthropological law you are expected to behave in a certain way, do not be alarmed – it does not imply that you have no free will because you may well behave in a way contrary to predictions. Or the "law" might not exist because it does not entail a temporal causality that would tend to influence your behavior. But this will not necessarily show that the *regularity* does not exist. The regularity is something crying out to be explained, and a prediction is far from an explanation. We might predict that technology stocks will go up (because of belief in a new economy) and be correct for a time, but the explanation is easy to disconfirm if you give negative evidence a chance to appear (just wait for a business downturn, for example, that affects new technology). Your behavior, when contrary to a prediction, will contribute to whether social regularities are manifested as correlations, and will break the appearance of functional relationships because there are at least two outcomes in a situation that was predicted to have only one. A theoretical prediction, as opposed to a prediction that merely describes and extrapolates a regularity, ought to be based on a more general explanation for which many different tests are possible, and so if false, ought to be fairly easy to disconfirm using a variety of different tests.

What do Correlations Mean?

One might make a stronger argument that functional relationships are not observed in the human world not because of free will but because social scientists normally have no opportunity to use controlled laboratory data to determine functional ones. We mention above, for example, that we can determine which value of amperage (I) corresponds to any value of voltage (V) according to the Ohm's Law ($I \sim V : \text{given } R$), assuming that resistance (R) of the given electric circuit is constant. To hold resistence constant, we need laboratory controls. But in the real world the resistance of a given piece of metal varies at different points because the temperature is never exactly uniform and the higher the metal's temperature is the higher its resistence. So if we measured precisely the amperage in some electric circuit in a car at different times of day, in different seasons and in different phases of the car's work, the resultant data might not look like a functional dependence. If all these different conditions simply created errors of measurement that were random, a correlation between amperage and voltage holding resistance constant would approximate the functional relation between the two variables (a straight line within a cloud of data that are scattered by random errors). But the errors might be biased so as to produce a result that does not show the underlying lawful relation.

If we were to take the analogy of the laboratory seriously, however, we would begin to question whether the observed correlation is one that will replicate under all the possible conditions in which we could imagine testing an hypothesis that the X – Y relationship will change under conditions specified by a third variable Z. These are called *third factor* tests. Correlations that can be replicated under a wide variety of conditions are said to be *robust*. Another type of third factor test is this: we observe that X – Y are uncorrelated but we think we will find that X – Y will correlate robustly under condition Z.

We will review third factor hypotheses and statistical tests in Chapter 7. For present purposes, we have to consider first the type of prediction that corresponds to different correlations so that we can understand what we get when calculating correlation coefficients.

A correlation, also called a measured of association, must be characterized by the *type of prediction* that is tested for, and a measure called the correlation *coefficient* that conveys the direction of the prediction (does X entail more Y or less Y?) and the strength of the prediction. Correlation coefficients are designed or constructed so that zero means no correlation, +1 means a perfect positive correlation, -1 means a perfect negative correlation, and anything else is a less-than-perfect correlation, either negative or positive, depending on its value. The strength of a correlation, applied to data, will always vary between 0 and ±1. Some correlations, have no sign or direction because of how they are computed, like Phi², Cramer's V, lambda and the contingency coefficient, only a strength between 0 and 1.

Not all correlations are of the functional type, so it is useful to distinguish the three major types as functional, order-based, and categorical. They do not always correspond to the levels of measurement (interval, ordinal and nominal) because each of the three types of correlations has distinct subtypes. Different types of correlation may measure weaker or stronger types of predictions, the strongest being a bifunctional prediction of association between two variables X and Y in which each value of X has a unique corresponding value of Y and vice versa. You must always consider which types of correlation are most relevant to the kinds of hypotheses you are testing as well as the measurement level of variables.

Functional Coefficients (interval or ordinal variables or 2 x 2 tables)

Functional coeficients are the strongest general type of correlations but come in three varieties according to their strength, although most of them converge to the same result for cross-tables of two row and two column variables (2 x 2 tables).

1. A *linear* correlation between X and Y assumes that they are measured on a interval scale such that a given interval of change in one variable corresponds to a constant interval change in the other.
Example. Figure 3.1, our scatterplot of Agricultural Reliance and Fixity of Settlement from Chapter 3, shows a dotted black line that would represent a linear relationship such at one unit increase in agricultural contribution to local food supply leads to a constant

unit increase in fixity of settlement. These are not true interval scales, but in this case one unit of ordinal increase in the one variable is associated with one unit of ordinal increase in the other.

2. A *bifunctional* correlation between X and Y tests for the extent to which, given a value of X, there will be only one value of Y, and *vice versa*.
Example. We can see from Figure 3.1, where the red line represents average values of the Y variable at each unit increment of the X variable, that bifunctionality does not describe the general pattern of relationship between these two variabales. This is because between values 1 and 3 on the agricultural contribution scale X axis there is almost no change in the mean on the Y axis, and between values 5 and 6 on the X axis there is something of the same pattern. If we used the red line to guess the best categorical prediction between X and Y there would be several values of X that have the same predicted value of Y.

3. A *simple functional* correlation between X and Y tests for the extent to which, for every value of X, there will be only one value of Y.
Example. Figure 3.3, the scatterplot of Population Density and Juridictional Hierarchy from Chapter 3, has a series of black circles over some pairs of values that show a simple functional prediction, in this case, from Y to X, but we can see that if we try to find a simple functional prediction from the 7 categories of X to the five categories of Y (with 5), several values of X will predict one value of Y, so simple functional prediction fails.[4] The red line in that plot showed the average values on the Y coordinate (Juridictional Hierarchy, scaled from 1 to 5 in equal intervals) for each value of the X coordinate (roughly, the log to base 5 of the population density), two interval variables and the red line is approximately linear. The best-fit straight line to the red line would qualify as a bifunctional prediction if we allowed a continuous measure of each variable.

The linear correlation is the strongest type. It is also a functional relationship because it is by definition bifunctional and a bifunctional correlation is by definition a functional one. The simple functional correlation is the weakest type of prediction among these three. The linear correlation the strongest. The linear correlation is also stronger in its assumptions about measurement because it requires two interval variables. The other two types of functional correlations do not require anything more than *categories* for measuring variation. These weaker types of relationship can be measured with order-based or categorical correlations.

The four the most commonly used functional correlations can be ordered from the strongest to the weakest as follows:

- **Pearson's *r*** is a linear correlation that requires two interval variables and performs best when those variables are normally distributed. R-square (r^2) measures the explained proportion of *variance* in squared differences from a

[4] Here we are taking the domain and range of our functional prediction in the examples of Figure 3.1 and 3.3 as discrete integers. If we chose to take continuous measures with decimal values we could treat the third example as a test of a bifunctional hypothesis without worrying about difference in the number of discrete categories.

prediction, i.e., the average squared differences between variables x_1 and x_2 over each observation i among n observations is $\Sigma_i(x_{1i} - x_{2i})^2/n$.

- **Kendall's tau-b** is a bifunctional correlation between two ordinal or ranked variables based on the number of agreements (P) and inversions (Q) in the rankings. It discounts the number T_o of ties (which are allowed in functional correlations) so that a value of -1 or +1 can only be obtained from square tables. Its square measures the proportion of variance explained in the ordinal rankings. It is called a nonparametric test because it is not affected by whether the data are normally distributed; and it outperforms Pearson's *r* when data are non-normal.

$$\text{tau-b} = (P - Q)/ \text{SQRT}[((P + Q + Y_0)(P + Q + X_0))]$$

- **Stuart's tau-c** is a variant of tau-b for tables with three or more categories in both the row and column variable, where, if m is the number of rows or columns, whichever is smaller, and n is sample size,

$$\text{tau-c} = (P - Q)*[2m/(n^2(m-1))]$$

Although tau c can attain plus or minus 1 even if the two variables do not have the same number of categories, its value is normally less than tau-b. If there are no ties on either variable the two measures are identical. When the numbers of rows and columns are similar, tau-b is recommended because it is more sensitive to ordinal covariation.

- **Spearman's rho**. A nonparametric bifunctional correlation between two ordinal variables. The values of each of the variables are assigned ranks from smallest to largest, and the Pearson correlation coefficient is computed on the ranks. Its square measures the proportion of variance explained for the ranks, but without discounting ties.
- **Lambda**. A functional correlation for ordinal or categorical data in which errors in guessing that each observation of the dependent variable corresponds to its modal category are compared to errors in guessing based on the modal category of the dependent variable knowing the value of the independent variable. It measures the proportionate reduction in error (PRE) for reduction in errors in predicting the dependent given knowledge of the independent variable. Lambda is unaffected by the ordering of the columns of either the independent or dependent variable.

There are some types of relationships for which none of these three types are applicable. For example, the finding that patrilocality (X) predicts internal warfare (Y) is neither a bifunctional nor a functional prediction, because it does not say what happens (to Y) when patrilocality (X) is absent. They require another kind of correlation coefficient based on order or category.

Relational Correlation Coefficients: Order- and Category-Based

Statisticians have designed a number of correlation coefficients for weaker forms of relationship. In contrast to a functional correlation, a relational correlation between X and Y is one in which more than one value of variable Y may correspond to a value of variable X. The first set of coefficients we consider are appropriate for the ordinal level

of measurement and the second set for any level of measurement. One might think that the order-based correlations are stronger than the category-based, but this is not always so. Gamma and order-based correlations, for example, are weaker than most of the category-based coefficients.

For ordinal cross tabulations as well as tabulations of interval variables and scatterplots it is helpful to consider whether the predictions of a dependent variable Y are increasing or decreasing with increases in X. Relationships between variables are *monotonically* increasing or decreasing when these relative changes are consistently in the same direction, and monotonically nonincreasing or nondecreasing when they do not actually reverse course, but may be consistently constant or increasing, or constant or decreasing.

Order-Based Correlation Coefficients (Ordinal variables and 2 x 2 tables)

Ordinal or order-based correlations are weaker than functional ones, but not always stronger than categorical correlations.[5] For ranked variables the ordinal coefficients tau-b and Spearman's rho have been discussed, but these are functional not relational coefficients and they are based on direct prediction from a rank on the first variable to a single rank on the second. A different approach, which is relational, is based on comparing pairs of cases rather than single cases. Freeman (1986) provides an excellent synthesis of ordinal relationships. Among these order-based correlations are those based on summing the number of pairs of cases where the ranks of the two variables are lo-lo/hi-hi (C = concordant for the pair of cases) or lo-hi/hi-lo (D = discordant for the pairs of cases). Some measures, such as Somer's d, also take into account the number of pairs that are tied on the independent variable (Xo) and that are tied on the dependent variable (Yo), or both (T = tied pairs = Xo + Yo). Three of the common statistics may be ranked from strongest to weakest as follows:

- **Somer's symmetric D** is an order-based measure (for two ordinal variables) computed as (C-D)/(C+D+T). Note the similarity to the bifunctional correlation. The way such a correlation is normed by the denominator is interpreted as a proportional reduction of errors (PRE) in making order-based predictions rather than variance explained.
- **Somer's d** is an order-based measure computed as (C-D)/(C+D+Xo) that takes into account the number of pairs not tied *on the independent variable*. It is used for predicting the column categories of a dependent variable (columns) from categories of an independent variable (rows). Note the similarity to the simple functional correlation.

[5] As noted above, 2 x 2 tables may be analyzed by ordinal coefficients because dichotomies may be considered the simplest form of a rank order (hi/ho). Note that recoding of dichotomies will affect the signs of the respective correlation coefficients accordingly. For example, the positive sign of a correlation with gender would mean that the respective characteristic is more typical for females if they were coded as "1"; but would mean it is more typical for males if males were coded as "1." When interpreting the results of statistical analyses of dichotomous variables it is important to take into account how the respective variables were coded. But this would apply to other types of variables as well.

- **Gamma** is a weak order-based measure computed as (C-D)/(C+D).[6]

These relational correlations imitate the functional approach, but not to predict one category from another. Instead, they compare pairs of cases. In a 2 x 2 table for patrilocality and internal warfare, for example, with 40 cases of internal warfare (all patrilocal) and 60 cases with no internal warfare, 20 patrilocal and 40 non patrilocal, Gamma=1.0 indicates that for every pair of societies that differ on both variables, patrilocality always occurs with internal warfare. A gamma of 1 requires at least one zero cell. Why this might be useful information is explained in Appendix 1 at the end of this chapter. Since Somer's d measures the proportion of pairs of cases in which the dependant variable of internal warfare was correctly predicted for pairs that differed on the independent patrilocality variable, Somer's d = .67 does not tell you whether there is a zero cell in the 2 x 2. Somer's symmetric D is more stringent and is 0.5 for this example.[7]

Categorical Correlation Coefficients (Nominal variables and 2 x 2 tables)

Categorical correlations are appropriate for measuring weak forms of relationship such as those between nominal variables. They are also useful, with proper care, in putting together, piece by piece, curvilinear and non-linear relationships among categorical or even ordinal variables, as in the scatterplot analyses of Chapter 3. The following three commonly used categorical correlations are all based on chi-square (χ^2), which is a means of evaluating all sorts of contingent relationships as departures from "chance," which we will discuss after introducing the concepts of probability and statistical independence:

- **Phi²**
- **Adjusted Phi´² (Cramer's V)**
- **Contingency coefficient**

Since departure from chance does not indicate a particular direction or sign for these correlations, the value of each of these coefficients ranges between zero and 1. Zero denotes no correlation between the row and column variables and values close to 1 indicates a strong association between them. They can attain a value of 1 for tables of any dimension. Because we need probability theory as discussed in Section 3 to define

---

[6] For 2 x 2 tables this coefficient is very often denoted as Yule's Q, so designated by G. Udney Yule after Quetelet, a great Belgian statistician. Q is very easy to calculate without a computer, using just a calculator, or even just a pencil and a piece of paper. If we designate cells of 2 x 2 table in the following way:

a	B
c	D

then Q will be calculated using the extremely simple formula: $Q = (ad - bc)/(ad + bc)$.

[7] The relevant 2 x 2 table here (with internal war / no war across the top and patrilocal / not patrilocal as the rows, is table

a=40	b=20
c=0	d=40

For computing the gamma coefficient, C=1600, D=0, Xo=800, and T=1600. Somer's d=.67, and Somer's symmetric d=0.5

statistical independence and to develop the basis for these measures, we do not discuss them until Section 4.

What does the Strength of a Correlation Mean?

We put this question last because it depends on the type of correlation. The square of a functional correlation measures the percentage of covariation between two variables explained by the correlation. If a Pearson's r, tau-b or Spearman's rho equals .71, for example, we can say that 50% of the variance in each of the variables is in common with the other: variance accounted for by the correlation. Because we are not dealing with the analysis of variance in this book, it is easier to explain co-varation if a casual link exists between variables X and Y. So imagine that X is one of the causes of Y. Imagine that the correlation between the variables is – 0.5, or + 0.5. Let us square these figures. In each case we get 0.25: the variation of X explains 25% of variation of Y (or to simplify it, X causes Y by 25%). And what if the correlation coefficient between these variables is + 0.9? – 0.7? + 0.6? – 0.4? + 0.3? – 0.2? + 0.1?

In most statistical textbooks you will read that correlations with coefficients of > |0.7| are strong, the ones with coefficients of |0.5| – |0.7| are medium, and the ones with < |0.5| weak. We would stress that for cross-cultural research these conventions do not make much sense. They developed in the field where the modern applied statistic methods used in social sciences originated – i.e., in Biology. There, especially in laboratory and controlled experiments, these conventions make sense. In Sociocultural Anthropology in general and Cross-Cultural Research in particular, however, most important sociocultural variables seem to be determined by a considerable number of factors. Rarely does it appear to be possible to single out X determining Y by more than 50% (which would correspond to correlation of > 0.7 level). In this context, any X determining Y by 25–50% should be considered a rather strong factor. Factors determining 10–25% of variation in a dependent variable should be rather considered as moderately strong. One even wonders if we have sufficient grounds not to pay attention to factors explaining 6-10% of variance. Some characteristics look like a resultant of *ca.* 10 factors, as is our impression, for example, in studying state formation. If we do not pay attention to correlations of 0.25–0.3 level, or possibly even weaker if significant, will we be able to explain complex sociocultural phenomena?

We propose the following rough typology of strength of correlations that takes into account the types of correlation and sociocultural data for which nominal and ordinal categories are used, bearing in mind that the weaker levels of correlation may be still considered quite strong in cross-cultural research. When we say that a gamma of 1 is only a medium correlation, however, we mean that you only know half the story. To have gamma = 1 in a 2 x 2 table, there must be at least one zero cell, which is important information about the relationship (see Appendix 1) but not a strong functional prediction for cross-cultural research

	Strong Spearman's rho, Tau-b and c, Phi, Pearson's r, Somer's d (symmetric)	**Medium** Predictors: Somer's d	**Weak** Predictors: Gamma
• |1.0|	determinate	Strong	Medium
• > |0.7|	very strong	Medium	Weak
• |0.5| – |0.7|	strong	weak	quite weak*
• |0.3| – |0.5|	medium	quite weak*	very weak
• |0.2| – |0.3|	weak*	very weak	extremely weak
• |0.1| – |0.2|	very weak	extremely weak	Negligible
• < |0.1|	extremely weak	Negligible	Negligible

* in cross-cultural research these may be still considered quite strong, if significant

Section 3: Using the Analytic Power of Statistics

Probabilistic Inference

The most central ideas in statistics involve how to derive statistical judgments from probability theory.

Nominal Variables and the Laws of Probability

The product law of independent events states that their joint probability is the product of their independent probabilities. In a toss of two fair coins, for example, the first and second toss are event classes with equal but independent probabilities of ½, and the chances of two heads in a row is ½ · ½ = ¼. The sample space Ω of outcomes for a single toss is {H,T} and that for two tosses Ω = {HH, HT, TH, TT} has four outcomes when the difference in order between HT and TH is taken into account. Abstractly, the probability of an event E that falls into two classes C_i and C_j in a sample space Ω that occur with independent probabilities $P\Omega(C_i)$ and $P\Omega(C_j)$ is

$$P\Omega(E|E \text{ in } C_i \text{ and } E \text{ in } C_j) = P\Omega(C_i) \cdot P\Omega(C_j). \qquad \text{Law 1. Product Law of independent events}$$

A nominal variable is a classification of events into m mutually exclusive and coexhaustive classes $C_{j=1,m}$, as for example a coin toss having m=2 with j=1 for heads and j=2 for tails. The total probability law for mutually exclusive and coexhaustive classes of any sample space Ω is that probabilities must sum to one:

$$\Sigma(E|E \text{ in } P\Omega(C_j))_{j=1,m} = P(\Omega) = 1. \qquad \text{Law 2. Summation to One Law of sample spaces}$$

The additive law of mutually exclusive classes is that the probability of events in two such classes is the sum of their independent probabilities:

$$P\Omega(E|E \text{ in } C_{j=1} \text{ or } C_{j=2}\text{: mutually exclusive}) = P\Omega(C_{j=1}) + P\Omega(C_{j=2}). \text{ Law 3 Additivity Law of exclusive events}$$

The probability laws apply if the probability of each class is computed as its fraction in the sample space Ω. In so doing we treat a nominal classification as a probabilistic variable. The fraction or relative frequency of events in a class is its frequency divided by N, the size of Ω. We also have to take into account the difference between a set of ordered outcomes, such at HT or TH in two coin tosses, versus the unordered outcome of one head and one tail. To compute this probability as an unordered outcome (HT) = {HT, TH}, we have to use Law 3 to add the probabilities of the outcomes in an ordered sample space Ω that resolve to those in an unordered sample space Ω^*. Thus $P\Omega^*$ (HT) = PΩ{HT,TH} = ¼ + ¼ = ½.

Unlike heads and tails in a fair coin toss, the probability of even or odd in roulette does not sum to one because casino profits derive from a sample space Ω with 37 outcomes, one of which is neither {red, black} nor {even, odd}. Here the probability PΩ(even) = PΩ(odd) = 18/37 < ½ and by law 3 PΩ(even or odd) = 18/37 + 18/37 = 36/37 < 1. Adding probabilities, law 2 is satisfied by P(Ω)= PΩ(even or odd) or PΩ(neither even nor odd) = 36/37 + 1/37 = 1. Although the probability PΩ(odd) = 18/37 < ½ and that of PΩ(red) = 18/37 < ½, the probability of both red and odd is not the direct product of their probabilities on the space Ω of 37 outcomes. It is only in the reduced sample space Ω′ of the 36 outcomes where red and odd are independent that PΩ′(red and odd) = ½ · ½ = ¼ = 8/36 by law 3. For the space Ω of 37 outcomes, law 2 applies so that PΩ({red and odd} or outcome 37) = PΩ({red and odd}) + PΩ(outcome 37) = 8/37 + 1/37 = 9/37. Hence PΩ{red and odd}) = 8/37 < ¼. Probabilities in a coin toss would be the same as those in roulette if the chances of the coin balancing on an edge were 1/37.

Buy a roulette wheel: you see 37 slots

Ad for gambling: You imagine only 36, the "0" slot is hidden

Application of probability theory to empirical examples requires careful assessment of what event classes are independent and in what sample space. Law 3, for example, is especially sensitive to nonindependence of units sampled where there are clustered similarities in the sampling space Ω. These problems are taken up in Chapter 8. As

Kruskal (1988:929) warned: "Do not multiply lightly," meaning: take care in evaluating the assumption of independence in law 3, and make appropriate adjustments. Dow (1993), for example, has examined the adjustments required for the analysis of cross-tabulations for samples with clustered similarities, which we consider in Chapter 8.

The remainder of this chapter shows how to use probability laws, with careful attention as to independence and non-independence, to derive probability distributions for different results of cross-tabulations, and then to associate a probability, called the statistical significance test, with an observed result.

For example, $p < .05$ for an observed distribution means that there is less than one chance in 20 of the observed distribution occurring if the variables were independent. The lower this probability the less likely that the variables are independent and the more likely that the observed deviation in the table from a zero correlation is simply random.

Expected Frequency and the Null Hypothesis

The observed frequencies in a cross tabulation of variables can be compared to the frequencies that would be expected given the laws of probability for independent events.

Expected Probability

By Law 1, under the assumption of independent probabilities for each variable, the expected probability of events in any cell of a cross tabulation of nominal variables is defined as the product of probabilities of this type of event for each variable. For a simple example, we take a fair roulette table with only 36 outcomes corresponding to the numbers 1-36, half of which are assigned to red and half to black for each of the odd and the even number series. Thus, P(red and odd) = ¼ corresponds to the upper left cell of Table 5.1, a cross tabulation of independent probabilities of outcomes for our two variables (color and series).

Table 5.1. Probabilities of outcomes at the roulette table with no slot 37

	RED	BLACK	Total
ODD	¼	¼	½
EVEN	¼	¼	½
Total	½	½	1

Expected Frequencies given Independent Probabilities: the Null Hypothesis

In N=36 throws at this roulette table, the expected frequencies are those in Table 5.2:

Table 5.2. Expected Frequencies at the roulette table with no slot 37

	RED	BLACK	Total
ODD	9	9	18
EVEN	9	9	18
Total	18	18	N=36

Given an empirical cross tabulation, the null hypothesis is that the variables tabulated are statistically independent, that is, there is no correlation between them. The expected, or theoretical, frequencies for Table 5.2 are computed directly from the probabilities of a fair roulette table, which are uniformly ½ for the marginal frequencies of the table (18 of 36), and ½ · ½ or ¼ (9 of 36) for the cell combinations odd/black, red/even, and so forth. In empirical research, however, these theoretical probabilities are rarely known in advance and must be estimated from the actual marginal (row, column) totals of a table. For example, say an investigator surveyed 36 people whether they preferred the color red or the color blue and the outcomes of the survey for differing numbers of males and females is as shown in Table 5.3.

Table 5.3. Survey Results

	RED	BLUE	Total
FEMALE	4	15	19
MALE	7	10	17
Total	11	25	N=36

Overall, blue was preferred to red by more than 2:1 (25/11=2.27). Is there a difference between the preferences of males and females? The expected frequencies $\hat{H}_{r,c}$ under the null hypothesis of no difference can be computed as follows by the formula $\hat{H}_{r,c} = N \cdot P_r \cdot P_c$ where the probabilities of row and column characteristics P_r and P_c are estimated from the relative proportions of the row and column sums, F_r and F_c in the total sample. Hence $P_r = F_r / N$ for row $_{r=1,2}$, and $P_c = F_c / N$ for column $_{c=1,2}$. Thus $P_{r=1}$ (red) = $F_{r=1}$ (red) / N = 11/36 = .305555, as in Table 5.4.

Table 5.4. Survey Results with row and column probabilities calculated

	RED	BLUE	Total F_r (female, male)	P_r
FEMALE	4	15	19	.527777
MALE	7	10	17	.472222
Total F_c (red, blue)	11	25	N=36	1.000000
P_c (red, blue)	.305555	.694444	1.000000	

Expected frequencies for test survey results are given in Table 5.5 by formula $\hat{H}_{r,c} = N \cdot P_r \cdot P_c$. If the survey had turned up results where the cell frequencies were very close to the expected frequencies in Table 5.5, such as cell values 6:13::5:12, the correlation between our two variables would be very close to zero. Try, for example, calculating gamma with the 2 x 2 cell values 6:13::5:12.

Table 5.5. Expected Frequencies from row and column probabilities

	RED	BLUE	Total F_r (female, male)	P_r
FEMALE	5.805	13.195	19	.527777
MALE	5.195	11.805	17	.472222
Total F_c (red, blue)	11	25	N=36	
P_c (red, blue)	.305555	.694444		

Next, we may ask two questions related to the null hypothesis of statistical independence:

(1) How far are the actual survey cell values from the expected values?

(2) What is the likelihood of this difference occurring under the null hypothesis?

The last question is known as a test of statistical significance.

The Chi-square (χ^2) statistic for measuring departure from the Null Hypothesis of Statistical Independence

One way to calculate departure from the null hypothesis of statistical independence is to compute the chi-square statistic made for this purpose. Because larger differences between expected and actual frequencies are more important than smaller ones, the differences are computed, then squared, and finally, divided by the expected frequency. These squared differences for males and females in each response category in the survey can be calculated this way, as in Table 5.6, and the result divided by the expected frequency, as shown in Table 5.7. To get a total chi-square statistic for a cross tabulation, the chi-square values $\chi^2_{r,c}$, for individual r,c cells are summed, so that $\chi^2 = \Sigma\ \chi^2_{r,c}$. In the case of Table 5.7, $\chi^2 = 1.708$. Here is the formula for the chi-square over the i=1,k cells of a crss tabulation:

$$\chi^2 = \Sigma_{i=1,k}\ ((O_i - E_i)^2)/E_i \qquad \text{Equation 1}$$

Expressed in means (μ) and standard deviations (σ), this is equivalent to

$$\chi^2 = \frac{(x_1-\mu)^2}{\sigma^2} + \frac{(x_2-\mu)^2}{\sigma^2} + \cdots + \frac{(x_k-\mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{(x_i-\mu)^2}{\sigma^2} \qquad \text{Equation 2}$$

Table 5.6. Difference between Survey Frequencies and Expected Frequencies

Difference:	RED	BLUE	Squared:	RED	BLUE
FEMALE	1.805	1.805		3.26	3.26
MALE	1.805	1.805		3.26	3.26

Table 5.7. Squared Differences divided by Expected Frequencies (χ^2)

	RED	BLUE	
FEMALE	3.26 / 5.805 = .562	3.26 / 13.195 = .247	$\chi^2 = 1.7$
MALE	3.26 / 5.195 = .623	3.26 / 11.805 = .276	

Chi-square is a standardized measure for answering question (1): How far are the actual cell values in any given table from the expected values? How do we use it to answer question (2): What is the likelihood of this difference occurring under the null hypothesis? In theory, question (2) can be answered directly from probability theory. For simplicity, consider the hypothetical case where, holding the marginals of the survey responses constant, we obtained the most negative possible correlation consistent with the hypothesis that females tend to like blue while males tend more toward liking red, as shown in Table 5.8. By Law 1, the zero cell in this table has a probability P{female,red} = .527777 · .305555 = .1612649035, and by Law 2 and 3, the probability P{not female or

not red} = 1 - P{female,red} = .838735098765. Then by Law 1 once more, the probability that the zero cell will occur is the probability P{not female or not red} raised to the 36^{th} power, which is less than .001. This gets more complicated when there is no zero cell (see Fisher's exact test).

Table 5.8. A hypothetical strongly negative correlation

	RED	BLUE	Total F_r (female, male)	P_r
FEMALE	0	19	19	.527777
MALE	11	6	17	.472222
Total F_c (red, blue)	11	25	N=36	
P_c (red, blue)	.305555	.694444	$p < .001$	$\chi^2 = 25$

Converting Chi-square into a Test of Significance

The chi-square statistic is constructed to allow you to look up an approximate and conservative estimate of the probability that any table with a given number of rows and columns and a calculated chi-square value will occur by chance under the null hypothesis.[8] Later we will give a direct calculation of this probability, Fisher's exact test of significance, so bear in mind that chi-square is only an estimate. For the example in Table 5.3, where $\chi^2 = 1.7$, the corresponding probability is p = .19. This value signifies that, under the null hypothesis, when color preference is statistically independent of gender, similar or more extreme difference in responses than those in Table 5.3 will be observed in about 19% of similar surveys with the same sample size. This lends support to the null hypothesis. To use this procedure to evaluate the null hypothesis, however, you must first calculate the degrees of freedom of your cross tabulation.

Degrees of Freedom

The degrees of freedom, df, of a cross tabulation are the number of cells that must be filled before all cell values are determined by the marginal (row, column) totals. In a 2 x 2 table there is one degree of freedom. If R is the number of rows and C the number of columns,

$$df = (R-1) \cdot (C-1) \qquad \text{Equation 3}$$

Significance Tests: The Chi-square approximation and alternatives[9]

On-line calculators at urls www.stat.sc.edu/webstat/ and home.clara.net/sisa/signif.htm give the correct probability of p = .19 for Table 5.3 given its chi-square value and degrees of freedom df=1 as a 2 x 2 table. To see the chi-square distribution from which these are calculated see http://faculty.vassar.edu/lowry/tabs.html. The distribution has only one

[8] Because the shape of the theoretical distribution relating chi-square values to probabilities changes from a monotonically decreasing function at one degree of freedom towards a more normal or Gaussian curve as degrees of freedom increase, there are lookup tables and calculators to make this conversion.

[9] See http://bmj.bmjjournals.com/collections/statsbk/8.shtml for an introduction to Chi-square, and see http://www.uvm.edu/~golivett/introbio/lab_reports/chi.html for a table of critical values.

tail, "departure from randomness."

The chi-square procedure breaks down, however, when the expected frequency for a cell is less than 5, and the problem grows more severe as the expected value drops closer to zero. This is not a problem in the example above where the expected values for each of the four cells in the 2 x 2 table are all greater than five, so the use of a lookup table (http://home.clara.net/sisa/signif.htm) to convert the chi-square value of 1.71 to a significance level of p = .19 is valid. When a small expected value occurs in the denominator, however, the chi-square cell value becomes overly sensitive to small variations between the actual and expected value of the cell. There are number of adjustments implemented in the SPSS package to correct for this defect, such as Yates' corrected chi-square.

An alternative to chi-square is to compute an exact probability using Fisher's exact test of significance, which we will shortly learn how to compute and use in significance testing. First, however, we exploit a rescaling of chi-square that provides a very sensitive measure of patterns of correlation.

Building on Chi-square: The all-purpose Phi Coefficient

The Phi-square correlation coefficient Φ^2 and adjusted Φ'^2 (Cramer's V)

The correlation coefficient phi (Φ) square is computed as $\Phi^2 = \chi^2/N$, which in the case of Tables 5.3-5.7 is $\Phi^2 = 1.71 / 36 = .0475$, with a significance of p = .19. This is very easy to compute by hand once you know the value of χ^2. For a table with a minimum number m of rows and columns where m > 2, an adjusted phi-square (Φ'^2, pronounced phi-prime square, also known as Cramer's V) is calculated as in Equation 4.

$$\Phi'^2 = \Phi^2/\sqrt{(m-1)} \qquad \text{Equation 4}$$

Φ'^2 is a measure of the variance accounted for in predicting one variable from the other. So long as there are either two rows or two columns in the contingency table, $\Phi'^2 = \Phi^2$, but when this is no longer the case, the ordinary Φ^2 can exceed one and and when this happens phi, the square root of phi, exceeds the ordinary bounds of a correlation coefficient. Phi′ will always be within the bounds of $-1 \leq \Phi' \leq 1$.

For our example in Tables 5.3-5.7, the phi-square (with no adjustment needed) is extremely weak because it is so close to zero: Less than 5% of the variance in color choice is predicted from differences in gender. One would conclude that there is no evidence here that males and females differ in their responses to this survey question.

The significance of phi, including phi′, is estimated by the significance of chi-square or the Fisher exact test, whichever is appropriate.[10] Calculators for Fisher can be found for not only 2 x 2 tables but for tables having up to 2x5 cells (see http://home.clara.net/sisa).

Assigning a Positive or Negative sign to the Phi′ correlation coefficient Φ′

Phi′ is the positive or negative square root of Φ′². Because Φ′² as calculated from χ^2 is always positive, when we compute $\Phi' = \sqrt{(\Phi'^2)}$ we have a choice of assigning a positive or negative sign to Φ′ because square roots of positive numbers may be positive or negative. If inspection of the scatterplot or cross tabulation shows a negative correlation, we assign a value where $-1 \leq \Phi' < 0$. In Table 5.8 the correlation between female and red, (or male and blue) is negative and $\Phi = \Phi' = -\sqrt{(25/36)} = -5/6 \sim .83$.

Evaluating Cross Tabulations of Nominal Variables

Care must taken in giving a sign to Φ′ because Φ′² in general is a measure of predictability between nominal variables. If our respondents fell into three sets, as in Table 5.9, we could not give a sign to Φ′ because there is neither a strictly positive nor a strictly negative correlation with age. Because Φ′² = 1 the relationship between age and color preference is perfectly predictive, but an inspection of Table 5.9 shows a curvilinear relationship between age and color preference.[11]

Table 5.9. A hypothetical perfect correlation, with Φ² = Φ′² =1

	RED	BLUE	Total F_r (yo,mid,el)	P_r
15-25 year olds	0	12	12	.333333
26-40 year olds	12	0	12	.333333
41-60 year olds	0	12	12	.333333
Total F_c (red, blue)	12	24	N=36	
P_c (red, blue)	.333333	.666667		

Evaluating Cross Tabulations of Ordinal Variables

If, however, the sample responses had looked like those in Table 5.10, we could assign the Φ′ correlation between age > 25 and preference for blue a negative sign, Φ′ = -1.[12] The three categories of age in this table constitute an ordinal variable because they are rank ordered by the criterion of age. If in this context the proportions of red::blue in the columns of successive rows and those of younger :: middle :: older in successive columns are monotonically nonincreasing or nondecreasing in every case, phi can be given a sign and interpreted as an ordinal correlation.

[10] As before, the chi-square estimate of significance is inaccurate when expected frequencies fall below five for any cell in the table. This does not deter the use of the phi or phi′ because the correlation is exact but not its p-value calculated by chi-square. The significance of phi′ in this case can be calculated by Fischer's test, which is appropriate for uneven marginals and small sample sizes.

[11] Note in this case χ^2=36 and phi = χ^2/N =1, with no adjustment.

[12] Equivalently, we could say that Φ′ = 1 between age > 25 and preference for red.

Table 5.10. A hypothetical perfect correlation, with Φ′²=Φ²=1 and Φ′=Φ= -1

	RED	BLUE	Total F_r (yo,mid,el)	P_r
15-25 year olds	0	12	12	.333333
26-40 year olds	12	0	12	.333333
41-60 year olds	12	0	12	.333333
Total F_c (red, blue)	12	24	N=36	
P_c (red, blue)	.666667	.333333		

In general, the sign of correlations between dichotomous or ordinal variables are an indication of which diagonal in the cross tabulation has the greatest concentration of scores. For 2 x 2 tables as in Table 5.8, where a, b, c, and d represent the frequencies in the four cells, phi also may be calculated directly from the formula in Equation 5.

$$\phi = \frac{a(d) - b(c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$ Equation 5

Use the values 4:15::7:10 to verify that in Table 5.8 Φ′ = Φ = -.218 and Φ′² = Φ² =.0475.

2 x 2 Cross Tabulations of Ordinal Variables with a sign for Φ

Every cross tabulation involving two dichotomous variables, as in Table 5.3, is by default an ordinal table, because a dichotomy is a binary order. Hence it is always possible to assign a positive or negative correlation coefficient to Φ′ for a 2 x 2 table. In the example above, Φ′ is negative. Note that Φ′² for a 2 x 2 table is equivalent to Pearson's r² (and to botth tau-b square and rho square) and is interpreted as variance accounted for.

Evaluating Cross Tabulations of Categorical Variables

The most precise contingency analysis of a cross tabulation is to calculate a signed fractional Φ′ for every cell in the table. This is done by dividing the chi square value for each cell (obs-exp)²/exp by N√(m-1) to give a fractional Φ′² for each cell. These values sum to the Φ′² for the whole table, which will be $\leq$ 1 (and in the 2x2 case equivalent to Pearson's r^2). Each fractional Φ′² is interpreted as a co-variance predicted for that cell that contributes (by summing) to the co-variance predicted for the whole table. The square root of the Φ′² values for each cell can be given the sign of the (observed-expected) difference for that cell. These signed Φ′ values are comparable across different contingency tables because they are normalized between -1 and +1, with zero indicating no correlation. Curvilinear and complex patterns of association between two categorical variables can be detected by the pattern and strengths of these fractional phi′ values.

A fractional Φ′² value for one cell will always be less than the full 2x2 Φ′² for that cell versus all other values in the full table. A strong fractional Φ′² is a good index of where the statistical significance in the overall table is coming from. Fractional Φ′² values, then, can be a guide to tests of more specific hypotheses about relationships among the

dichotomized categories in the table.[13]

Fisher Exact Significance Test

Significance in a statistical sense is the probability that the observed cell values of a cross tabulation with a given set of marginal totals would occur by chance alone if two nominal variables were statistically independent. For a 2 x 2 cross tabulation, Fisher's exact test is the most direct, accurate and powerful way to calculate this. Significance tests have greater *power* when they can describe more accurately the probability that an observed distribution represents a deviation from the null hypothesis of statistical independence. Fisher's exact is a more powerful significance test than chi-square because it is based on direct application of the laws of probability. Because these laws are combinatorial and they involve long series of multiplication or division, the Fisher test is usually only applied to relatively small samples where greater power is needed. For large samples, results based on chi-square, which are approximations, are relatively good (powerful) and in so the Fisher test may not be needed unless, for example, one of the expected values the cells is less than five. For our example in Table 5.3, with df=1 and $\chi^2 = 1.7$, the chi-square calculators give a probability of p = .19 which is close to but less sensitive than the exact probability of p = .17 from Fisher's exact test. Next we will use this example to show how Fisher's exact test is calculated from the laws of probability.

Fisher Exact Test for 2 x 2 Cross Tabulations

The 2 x 2 cross tabulation in Table 5.3, as we have seen, has $\Phi^2 = \Phi'^2 = .0475$, which corresponds to a Φ of -.218. If we consider the row and column totals of this table as fixed, a probabilistic sampling space for this table may be constructed by considering all possible distributions of the 36 cases that have combinations of {red, blue} by {female, male} that conform to these fixed marginals. Each distribution D_i for can be indexed from i = 0 to k, where k is the minimum of all the row and column sums, in this case 11. In the example, i could be used to index the frequency in the{female, red} cell, which when set determines the frequency in every other cell in the table (recall that df=1 for 2x2 tables). Table 5.8 would correspond to the case where i = 0 and Table 5.3 to the case where i=4. Those two tables along with others indexed for i=0 to 4 are shown in Table 5.11. This is a selective subset of all the possible tables with the same marginal (row, column) totals and includes only those tables that are the same or more extreme than Table 5.3 in their correlation bewteen the two variables.

[13] Because fractional Φ'^2 coefficients are based on an exact formula they do not require expected values ≥ 5 in the cells of the table. If you require an exact significance test for one cell as against all the remaining values in the table you can use Fisher's Exact test for the significance of the full 2x2 table Φ'^2 result. If you are doing rxc tables and looking for dichotomies that have predictive power, it is useful to use fractional Φ' values remembering that their explained fractional variance is Φ'^2.

Table 5.11. Possible tables from the probabilistic sampling space of Table 5.3, but which are more extreme (higher negative correlation) than Table 5.3

i=4	RED	BLUE	Total
FEMALE	4	15	19
MALE	7	10	17
Total	11	25	N=36
	p_4=.12547, 5.3		
i=3	RED	BLUE	Total
FEMALE	3	16	19
MALE	8	9	17
Total	11	25	N=36
	p_3=.03921, 5.11.1		
i=2	RED	BLUE	Total
FEMALE	2	17	19
MALE	9	8	17
Total	11	25	N=36
	p_2=.00692, 5.11.2		
i=1	RED	BLUE	Total
FEMALE	1	18	19
MALE	10	7	17
Total	11	25	N=36
	p_1=.00062, 5.11.3		
i=0	RED	BLUE	Total
FEMALE	0	19	19
MALE	11	6	17
Total	11	25	N=36
	p_0=.00002, 5.11.4		

Additive $p = \Sigma\, p_{i=1,k}\ (p_i)$ = .17224 = Fisher's Exact Test

Given this indexing system for the sampling space from which actual distributions may be drawn with fixed independent probabilities of the variables tabulated, the probability of each specific distribution Di for i = 0 to k can be computed exactly by the permutational formula in Equation 6. Here the product of successive integers from 1 to N (N! or N factorial) defines the size of the total sample space of all possible cell values of 2 x 2 tables with N=36. Subsamples that meet the four marginal constraints for cells F_{rc}! $F_{\sim rc}$! $F_{r\sim c}$! $F_{\sim r\sim c}$! and marginals $F_{r.}$!$F_{\sim r.}$!$F_{.c}$!$F_{.\sim c}$! are defined by the expression (M choose k) that refers to the number of ways there are to draw k elements in a cell given the marginal total of size N.[14] Then the probability P_{rc} of the value in a cell with F_{rc} elements is:

$$P_{rc} = \left(F_{r.} \text{ choose } F_{rc}\right)\left(F_{\sim r.} \text{ choose } F_{\sim rc}\right)\left(F_{.c} \text{ choose } F_{r\sim c}\right)\left(F_{.\sim c} \text{ choose } F_{\sim r\sim c}\right)/N!$$
$$= F_{r.}!F_{\sim r.}!F_{.c}!F_{.\sim c}! \,/\, N!\, F_{rc}!\, F_{\sim rc}!\, F_{r\sim c}!\, F_{\sim r\sim c}! \qquad \text{Equation 6}$$

Starting with the observed table, Table 5.3, and continuing through the tables with more extreme departure from randomness but fixed marginal totals (those in Table 5.3), p-

[14] In Equation 6 r and c designate a row and column of a 2 x 2 tble, ~r and ~c designate the other row and column, so that rc, r~c, ~rc and ~r~c designation the four cells, while r. is a row total and .c is a column total.

values for five distinct tables are computed, indexed on the {red, female} cell frequency i=4 to 0:

P_{rc} = 19!17!11!25!/36!4!15!7!10! = .12547 (Table 5.3)
= 19!17!11!25!/36!3!16!8!9! = .03921 (Table 5.11.1)
= 19!17!11!25!/36!2!17!9!8! = .00692 (Table 5.11.2)
= 19!17!11!25!/36!1!18!10!7! = .00062 (Table 5.11.3)
= 19!17!11!25!/36!0!19!11!6! = .00002 (Table 5.11.4)
These sum to p = .17224

Thus p = .17224 is the exact probability for the occurrence of the Table 5.3 or more extreme distributions with the same marginal totals under the null hypothesis of independence. This means that if two independent variables were generated randomly so as to arrive at their distributions in Table 5.3 (19 females vs. 17 males; 11 red, 25 blue), there is a 17.224% chance that the 'random' table would have the same or stronger correlation than the observed table.

Fisher Exact Test for up to 6 x 6 Cross Tabulations

Fisher's exact test of significance test may be calculated for cross-tabulations with unordered categories by calculating the probability of the observed tables and adding the probabilities of even more unusual tables. Keep in mind that this calculation will not take into account how ordinal categories are rank-ordered, which must be taken into account separately if an ordinal correlation coefficient is used. These calculations can be conveniently done for up to six rows and six columns at the web site calculator at http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html.

Fisher Exact one- and two-tailed Test and Data-Mining Errors

Unlike the chi-square distribution, which has only one tail, the Fisher Exact test is able to calculated a one-tailed test of departure in the same direction as the observed correlation from a two-tailed test of departure either in that direction or equally extreme in the opposite direction. Normally the one-tailed test is used, and is most appropriate where the observed correlation is in the same direction as stipulated by prior hypothesis or theoretical expectation.

The two-tailed test is used to avoid the danger of opportunistic findings. These occur because if you examine k correlations between random, you are likely to find 1/X correlations that test as significant at $p \sim 1/k$. Each of these significance test results, however, should really be multiplied by k, and rejected if $k \cdot p > 1$. Similarly, if you do not know how two variables are related but find in an exploratory test that p is significant using a one-tailed test, you are really overestimating significance, because a random correlation might have reached this level with that probability, but also might have reached a correlation as extreme as this one but in the opposite direction. The two-tailed test is more conservative because it computes the probability at the other tail of opposite correlation, and adds the two probabilities by Law 3 since they are mutually exclusive. The two-tailed test is less likely to treat a random correlation as significant. The one-

tailed test is more powerful in that is less likely to reject a valid correlation due to small sample size. To minimize these kinds of errors in statistical inference you must think about significance in groups of tests where you are data-mining and take care to use one- and two-tailed tests appropriately.

Section 4: Evaluating Correlations and Models

Significance

Significance tests indicate whether it is worthwhile to interpret a contingency table. A significant result means that the cells of a contingency table should be interpreted. A non-significant test means that no effects were discovered and that chance could explain the observed differences in the cells. In this latter case, an interpretation of the cell frequencies is not useful. Standard probability cut-offs considered significant are $p \leq .05$ and $p \leq .01$, which mean there are respectively fewer than five chances in 100 or one in 100 that the observed correlation is due to chance alone.

However, as Rosnow and Rosenthal (1989:1277) put it: "Surely, God loves the .06 nearly as much as the .05" (see also Rosenthal 1991:220; Munroe et al. 2000:17). In general we prefer to call the correlations which are significant within the p=0.05–0.10 range "marginally significant," and we believe that they need special treatment. Our general recommendation is to avoid overly strong statements like “the hypothesis is rejected” and "the hypothesis is supported" on the basis of such results. We believe that such results do not provide sufficient grounds to make either of these statements, but rather suggest that the hypothesis needs to be tested further before any accurate conclusions regarding it could be made.

An Example: The Swindler Test

Statistical significance is used to evaluate a correlation independently of its strength. Because of historical reasons the appearance of this characteristic is extremely misleading. The higher the significance index (*p*), the LOWER the significance. For example, 0.9 denotes a very strong correlation, but a negligibly low significance level; 0.01 denotes a negligibly weak correlation but a relatively high significance level.

What complicates the situation is that while a correlation may be strong and significant, or weak and insignificant it may instead be strong and insignificant or weak and highly significant. It is important to understand that correlation strength and correlation significance are two completely independent parameters. It is also important to know that if you randomly reduce the size of the sample, the correlation will on average stay the same, but its level of significance will decrease. If the original significance is a probability p between 0 and 1, the significance for half the sample will drop to the square root of p. So if you have $p < .01$ (considered highly significant) and you lose half your sample due to missing data, for example, you may expect your significance to drop to $p < .10$ (considered barely significant or by the .05 standard, not significant). On the other

hand, if you have $p < .001$ for a small sample, this is a correlation worth paying attention to.

Let us consider the following example. A swindler has approached you with a suggestion to play a "head-and-tails" game. The rules of the game which he suggests to you are quite peculiar – if both his coin and yours have the same result (both either "heads" or "tails") you pay him one dollar; and if one of them is "heads" while the other is "tails," he pays one dollar to you. It seems difficult to imagine how one could cheat the other with such rules, so you agree to the game.

Suppose in the first eight tosses the coins come up both "heads" three times and both "tails" three times. One time your coin had a "head" and his a "tail." Another your coin was a "tail" and his a "head." As a result, you have paid him $6 and he has paid to you $2. You have lost $4. You might have already started feeling that you are being cheated. There are some cases where victims of such cheating managed to bring swindlers to the court and to win their cases with the help of experts in mathematical statistics. But what do you think – will any court consider the evidence we have at the moment convincing? Will a specialist in mathematical statistics be ready to support your case in court? We are sure that even you have no knowledge in mathematical statistics at all, you must be feeling that the correct answer to both questions above is "No." But why?

Let us first cross-tabulate the results of the first eight tosses. The respective cross tabulations will look as follows (see Table 5.12):

Table 5.12. Results of the first 8 tosses, percentaged by row

		His Coin		Total
		0 (head)	1 (tail)	
Your Coin	0 (head)	3 *75%*	1 *25%*	4 *100%*
	1 (tail)	1 *25%*	3 *75%*	4 *100%*
Total		4 *50%*	4 *50%*	8 *100%*

The correlation between two variables seems to be evident – indeed; if your coin shows "head," his coin is likely to show a "head" too. If his coin shows "tail" so does your coin. We can easily test the strength of this correlation, and it will turn out to be fairly high (phi = rho = r = + .5). But what will be the significance of this correlation? Just p=.243 by Fisher's exact test.[15] This means that there is an approximately 1 chance out of 4 to get the results specified above even if the game was perfectly fair and nobody had tried to cheat the other. This chance is rather high and the respective significance would be totally unacceptable for a court, as this would imply a high probability of wrongly

[15] We used the 1-tailed significant test because we expected to be cheated, and the results turned out to be in the expected or "predicted" direction.

convicting an innocent person.[16] Such significance would not be acceptable either for scientific research, both in natural and social sciences.

Note that although the case cannot be brought to the court, we would still be likely to stop the game because if we generalized the observation of 3 chances in 4 as one of being cheated that would look risky enough not to continue the game. This is similar to hypothesis-testing circumstances in which you would take seriously even insignificant correlations, although they do not prove anything. For research purposes the implication is that if you test your hypothesis and correlation turn out to be in the predicted direction, but insignificant, you cannot consider your hypothesis to be supported, but still we would advise you to consider such results as suggesting that your hypothesis still might be true, but that it should be re-formulated more precisely and re-tested. In any case, we would advise you not to stop your research and to continue it.

Suppose we are police agents and we have decided to find out definitively if our fellow is a swindler or not. So we continue the game. This is analogous to increasing the size of your sample. Suppose that after 1600 tosses we will get the following results:

Table 5.13. Results of 1600 tosses, percentaged by row

		His Coin		Total
		0 (head)	1 (tail)	
Your Coin	0 (head)	500 *62.5%*	300 *37.5%*	800 *100%*
	1 (tail)	300 *37.5%*	500 *62.5%*	800 *100%*
Total		800 *50%*	800 *50%*	1600 *100%*

The correlation in Table 5.13 is visibly weaker than in Table 5.12. Indeed, after 8 tosses you paid $3 per $1 paid to you (that is, you had 3 cases in the predicted direction per 1 case in the opposite direction). After 1000 tosses in the predicted direction mixed with 600 cases in the opposite direction for every $1 paid to you you lost only 1 dollar and 67 cents. The correlation strength in the series of 1600 is indeed twice as weak as in the initial series of 8 (phi = rho = r = + .25 compared with + .5 in the first series). But what has happened with significance? Is it also lower in Table 5.13?

The actual situation is just the other way round. The significance in the final series is higher than in the first. And not just higher, but enormously, incredibly higher. The significance of the correlation in the final series, measured by Fisher's exact test, 1-tailed) is 7.5×10^{-24}. If we write it with ordinary decimals, it will look as follows: .000000000000000000000075. In a fair game, had you started your game immediately after the Big Bang 15 billion years ago, and continued it until now making 1 toss every second, there would be less than 1 chance in a trillion trllions to lose $400 dollars or more in any randomly selected 1600-toss series. Such evidence, of course, can well be

[16] Courts of law might require that this probability be less than one in a million, or a virtual impossibility, given the presumption of innocence until proof is beyond a shadow of a doubt for serious crimes or a preponderances of evidence for less serious offenses.

submitted to the court. This evidence establishes that this game is not based on a fair coin.

To summarise: correlation significance is measured by an index (usually denoted as *p*) which takes values between 0 and 1.0. It can be interpreted as the chance of the observed correlation occuring by chance only. Hence, *p* = .000000001 means that there is 1 chance out of billion that the observed correlation happened by chance so the statistical significance of the respective correlation is VERY HIGH. In contrast *p* = .5 means that there is 1 chance out of 2 (1:2 = .5) that the observed correlation happenned by chance only, so the statistical significance of the respective correlation is VERY LOW.

Sample Size

The swindler example illustrates a basic principle of significance testing: if you are sampling randomly from a population and comparing the result to some norm, the significance of the difference between observed behavior and that norm will increase the larger the sample unless that norm is perfectly calibrated to an actual and exact norm for the population sampled. In tossing a coin, for example, a series of tosses can be considered a random sample of that coin's behavior if each toss can be considered a statistically independent event. A fair coin has a norm of 50% tails and 50% heads. If you sample 2 coin tosses and get HH, the significance of the difference between this observation is p=.25, but if you sampled HHH it would be p=.125 and at HHHH p=.0625.

Note that as we doubling the size of a random sample from N=2 to N=4, the significance is squared and in getting smaller significance is thereby gained. The reason for this relationship between significance and random sample size is easy to derive from probability Law 1: If you have an extreme result of sample 1 with probability p and the same result from independent sample 2 with probability p, then the probability of both extreme results, compounded,[17] is p^2. In general, we have as Equation 7 for random samples,

$$p^f(N)=p(N/f), \text{ so that } p(N)=\sqrt[f]{p(N/f)} \qquad \text{Equation 7}$$

Similarly, if you divide a sample randomly in half, you can expect to take the square root of the original significance and therefore lose significance. In general, for a significance level p(N) at a sample size N, if you sample 1/f (e.g., ½, ⅓, ¼, etc.) then $p(N)= p(N/f)^f= \sqrt[f]{p(N/f)}$.

[17] Of course, this would not work with one sample with two heads and another with one head and one tail, for which the probabilities are ¼ and ½, because there are four possible ways to get one tail in four tosses, namely, HHHT, HHTH, HTHH and HTTT in a sample space of 16 possible outcomes so that probability is ¼ and not ⅛. To get these two probabilities to converge, you have to consider the two sequential tosses {H,H} and {H,T} as equivalent to those of {H,T} and {H,H}, the combined probability of which (by Law 3) is ½. So be careful how you use the Laws of probability

Section 5: Conclusion

Review of Concepts

There are three main parameters of a correlation.

1) The **sign** of the correlation. The correlation may be either positive, or negative. Theoretically, it may be 0, i.e. may have no sign at all, which corresponds to absence of any relationship between two respective variables. The correlation is positive if the growth of value of variable X is accompanied by the growth of value of variable Y. For example, above (in Diagram 3.3) we see that growth of population density tends to be accompanied by the growth of political complexity. Hence, we have all grounds to expect that the correlation between these two variables will turn out to be positive.

Exercise. *Look at Figure 3.1 in Chapter 3 and say what is the sign of correlation between reliance on agriculture and fixity of settlement. Explain why*.

The correlation is negative if the growth of value of variable X is accompanied by the decline of value of variable Y. For example, in Figure 3.2 we see that growth of political complexity tends to be accompanied by the decline of polygyny levels. Hence, we have all grounds to expect that the correlation between these two variables (for complex cultures) will turn out to be negative.

Will it be also negative for simple cultures? Why?

Note that correlations between nominal non-dichotomous variables have no sign.

2) The second parameter of correlation is correlation **strength**. It is measured with various correlation coefficients. Below we will list the most widely used ones:

Comparison of Correlation Coefficients

	Functional	Relational	Order	Category	2 x 2 Table
• Pearson's *r*	Yes				converges
• Spearman's Rho	Yes		Yes		converges
• Kendall's tau-b	Yes		Yes		converges
• Phi, Cramer's V		Yes		Yes	converges
• Somer's symmetric		Yes	Yes		
• Somer's d		Yes	Yes		row → col
• Gamma		Yes	Yes		weak*

* Weakness in this case is not a defect: see Appendix.

Most correlation coefficients take values between – 1.0 and + 1.0: – 1.0 corresponds to perfect negative correlation (i.e., to negative functional relationship); + 1.0 corresponds

to perfect positive correlation (i.e., to positive functional relationship). The exceptions are Cramer's V, phi² and the contingency coefficient which are appropriate for the measurement of correlation between nominal variables, as well as curvilinear and non-linear correlations.

When you are writing up your research results, it does not advisable to state that a correlation that you have found 'proves' your hypothesis. Proof is mostly a matter of logic and involves rigorously deriving and testing all the logical consequences of the theory that generates your hypothesis, including tests with data that have a time dimension. You may have strong evidence, but the evidence you would need to 'prove' an empirical hypothesis that is not already tautological is probably beyond the level of cross-cultural research that draws on a sample where data on each case through time is lacking. It is better to say your evidence support or contradict an hypothesis, or indicates that the hypothesis needs to be reformulated.

3) The third parameter of correlation is **significance**. For Fisher exact tests, one-tailed, given N values randomly distributed with the same frequencies across categories on each variable, this is the probability that an equal or stronger correlation will appear, which is to say, when the variables are statistically independent of one another. Statistical independence constitutes the null hypothesis. The shorthand, then, is to say that the significance value p is the probability of the expected result under the null hypothesis. Other statistical tests assume independence and approximate the Fisher Exact either by measuring deviation from cell frequencies expected from the marginal (row and column) totals of the cross-tabulation and converting those deviations into a chi-square and then a significance value, or by estimating significance from sample size and strength of correlation, but taking the type of correlation into account.

It is crucial to keep in mind that if you have two estimates of the same correlation from samples of different size, the size of the sample will radically affect the significance test but will not bias the estimate of the strength of the correlation. Thus, if you have a strong correlation but small sample size, do not reject the correlation on the basis of significance; the case for the correlation is undecided unless you have a sufficient sample size to reach significance. Thus, working with larger samples is always preferable to working with smaller ones.

There is another problem, however, which is that a very weak correlation may reach significance with a sufficiently large sample. Should we reject such a correlation because it explains very little of the covariation between the variables? The answer is again conditional. If we think that one variable is an outcome that is affected by many variables, each of which contributes a small amount to explaining the outcome (dependent) variable, then we should keep our weak correlation, but find others that contribution additional effects, and eventually try to test whether these various factors, when combined, do have a strong combined correlation to the outcome variable. This takes us into the realm of multivariate analysis that goes well beyond the framework we have established for ourselves here, although we do go into three-factors hypotheses in Chapter 7. Cross-cultural correlations may also be low not because they lack validity or

the concept measured lacks explanatory power, but because the reliability of the measure is low. This is a matter that we also take up in Chapter 7, under the single factor model for measuring reliability. Single-factor models for multiple measures of the same concept can also help to develop combined measures that have higher reliability and help to overcome the problems of consistently weak correlations in a domain of study.

Summary

When strictly independent events having two sets of characteristics that are independently defined are tabulated in a contingency table, the laws of probability can be used to model, from the marginal totals (rows, columns) of the table, what its cell values would be if the variables were statistically independent. The actual cell values of the frequency table can be used to measure the correlation between the variables (with zero correlation corresponding to statistical independence), they can be compared to expected values under the null hypothesis of statistical independence, and they can be used to estimate a significance test of the probability that the departure of the observed correlation from zero (statistical independence) is simply a matter of chance.

Independence of events and independence of definitions are preconditions for statistical tests that one must be careful to satisfy. In Chapter 7 we will take up the case where the definitions of two or more variables are not independent but measure the same thing, so that correlations will be indicators of reliability rather than, for example, causal relationship or influence. In Chapter 8 we will look at how, when the sample of observations departs from strict independence because of observed interactions between them, the correlations between interacting neighbors measured on the same variables can be used to deflate effective sample size in obtaining accurate significance tests.

References

Freeman, Linton C. “Order-based Statistics and Monotonicity: A Family of Ordinal Measures of Association." *Journal of Mathematical Sociology*, 12, 1986, 49-69.

Garson, David. Quantitative Methods in Public Administration, David_Garson@ncsu.edu.

Validity http://www2.chass.ncsu.edu/garson/pa765/validity.htm

Correlation http://www2.chass.ncsu.edu/garson/pa765/correl.htm

Measures of Association
http://www2.chass.ncsu.edu/garson/pa765/association.htm#pairs

Ordinal Association: Gamma, Kendall's tau-b and tau-c, Somers' d
http://www2.chass.ncsu.edu/garson/pa765/assocordinal.htm
http://www2.chass.ncsu.edu/garson/pa765/association.htm#pairs

Reliability http://www2.chass.ncsu.edu/garson/pa765/reliab.htm

Significance http://www2.chass.ncsu.edu/garson/pa765/signif.htm

Probability http://www2.chass.ncsu.edu/garson/pa765/probability.htm

Chi-Square Significance Tests
http://www2.chass.ncsu.edu/garson/pa765/chisq.htm

Appendix 1: Interpreting Gamma Coefficients

At present few scholars use gamma for cross-tab analyses. In SPSS you even cannot order a gamma correlation matrix – such an option simply has not been developed by SPSS designers. You can only calculate gamma through the "Crosstabs" menu. The web site http://home.clara.net/sisa/ord2.htm, however, may be used to compute gamma for tables up to 2 x 5 rows and columns.

Then, why did we advise you to "tick" Gamma when you do your cross tabulations in SPSS? The answer is simple – just because gamma coefficients provide you with extremely useful information, which cannot be adequately substituted with such standard measures of correlation strength as phi, Spearman's rho, or Pearson's *r*.

What is the difference between them?

It is easier to explain this difference for 2 x 2 tables.

For example, let us consider the relationship between general nonsororal polygyny and matrilocality. As was shown by Murdock (1949:206, 216) the general non-sororal polygyny tends to destroy the matrilocality. The test of this hypothesis would yield the following results (see Table 5.14):

Table 5.14:

General Non-Sororal Polygyny * Matrilocality (Standard Cross-Cultural Sample)

Uxori-/Matrilocal Residence	*General Non-Sororal Polygyny*		Total
	0 (absent)	1 (present)	
0 (absent)	**100** *73.5%*	**46** *95.8%*	146
1 (present)	**36** *26.5%*	**2** *4.2%*	38
Total	136 *100.0%*	48 *100.0%*	184

The data on postmarital residence for this table are from Murdock & Wilson, 1972, 1985 [SCCS, 1999, file STDS03.SAV; SCCS, 2002]. The data on non-sororal polygyny are from Murdock, 1985, file STDS09.DAT [SCCS, 1999, file STDS09.SAV; SCCS, 2002].

NOTE: $p = 0.0004$, one tail, by Fisher's exact test
Phi = Rho = $r = -0.24$; $p = 0.001$
Gamma = – 0.78; $p = 0.00001$

The standard measure of correlation strength for 2 x 2 tables is phi (note, however, that for such tables phi = rho = *r*). For our Table 5.14 phi suggests that we are dealing here with a rather weak correlation, whereas gamma suggests that the correlation is unequivocally strong.

Which of these coefficients should we take into account in this case?

Our answer is – "Of course, gamma."

Why?

The standard correlation measures (like rho, or *r*) will achieve their maximum level (i.e., 1.0) if the relationship between variable X and Y is lineal, and if every value of variable X perfectly predicts a certain value of variable Y.

Turning back to Table 5.14 we may say that phi (= rho = *r*) would indicate a strong correlation if not only the presence of general non-sororal polygyny predicted strongly the absence of matrilocal residence, but also if the absence of general nonsororal polygyny predicted as strongly the presence of matrilocal residence. In other words for our table we would have maximum levels of |phi| (= |rho| = |*r*|) if the absence of general nonsororal polygyny were not only a necessary, but also a sufficient condition of matrilocality.

But this was not the hypothesis we tested!

Murdock maintained just that the development of general nonsororal polygyny should tend to destroy matrilocality. But he was not stupid; and he knew ethnographic data sufficiently well not to claim that the disappearance of general nonsororal polygyny would necessarily lead to the development of matrilocal residence. Note that only if the latter claim were made, it would be appropriate in our case to use phi [= rho = *r*] to test the respective hypothesis. And if we are testing just the statement that the development of general nonsororal polygyny should tend to destroy matrilocality, we MUST use gamma, and NOT phi [= rho = *r*].

If we estimate the strength of general nonsororal polygyny as a predictor of nonmatrilocal residence using phi [= rho = *r*] we will have an impression that this is a very weak predictor. And we shall be wrong!

It will be gamma that will allow us to see that we are dealing with a really strong predictor, even though not a functional one. In other words, such measures as rho or *r* are only appropriate if not only the relationship between variables X and Y is linear. In the world populated by human beings one encounters such relationships very rarely. Here, gamma turns out to be much more useful measure of correlation strength than rho and the stronger correlations. But one must also ask with gamma, especially when close to zero:

is this telling me that there is a statistical entailment where, that category X entails Y, but not vice versa?

Let us consider once more 2 x 2 tables, such at Table 5.15.

Table 5.15:

TRAIT X		*TRAIT Y*
	0 (absent)	1 (present)
0 (absent)		
1 (present)		

phi [= rho = *r*] will be an appropriate measure of correlation strength if the hypothesis, *e.g.*, implies not only that the absence of trait X is a sufficient condition of the absence of trait Y, but also that the presence of the trait X is a sufficient condition of the presence of trait Y, *i.e.*, that in the whole cross tabulation we are dealing with sufficient conditions only.

But what to do if our hypothesis is formulated something like what follows: "The state may originate not before the population density reaches level X." Note that, on the one hand, our hypothesis implies that the population density being < X is a sufficient condition of the absence of the state, but, on the other hand, population density being > X is implied to be a necessary, but NOT sufficient condition of state formation. If our hypothesis implies the presence of conditions that are necessary, but not sufficient, the appropriate measure of correlation strength is gamma.

Hence, while preparing to create and analyze your cross tabulations, try to think whether your hypothesis logically implies not only that the absence of trait X is a sufficient condition of the absence of trait Y, but also that the presence of trait X is a sufficient (and not just necessary) condition of trait Y. If you are not sure about the both do not forget to "tick" gamma.

But even if you are sure, still do not forget to do this! You will be surprised how often you will get gamma significantly higher than phi, or rho. You will be surprised how often what you thought to be a sufficient condition turns out to be a necessary condition, but not a sufficient one.