

# Chapter 5

## Statistical Analysis of Cross-Tabs

D. White and A. Korotayev

10 Dec 2003

Html links are live

Some new text added in Blue 30 Oct 2004

### 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. Our discussion proceeds in three sections that build one on the other. Section 1 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 2 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 3 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.

Section 4, excerpted here, provides practical advice for contingency table analysis in SPSS. Although the student can skip directly to this section, reading the other sections is highly recommended because they provide the basis for statistical reasoning. It will often be much more useful to understanding the concepts of statistical reasoning to deal with your data analysis rather than simply apply correlational and significance tests mechanically and without understanding.

### Section 4: Practical Advice for Contingency Tables in SPSS

#### Correlation Coefficients: Phi' (Cramer' s V), Tau-b, and 2x2 Gamma

We recommend the use of tau-b for ordinal correlations and for nominal correlations either Cramer's V or phi, the latter plus Gamma 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.

## Statistical Analysis of Cross-Tabs

Click if many cells or small sample ----->  Chi-square

Always click (for Cramer's V) ----->  Phi and Cramér's V

Click tau-b for ordinal correlation if one variable has 3 or more categories ----->  Kendall's tau-b

Click if many more categories for one variable than for the other ----->  Gamma

Click Gamma if the table is 2x2 (two rows and two cols). Then inspect if one of the four cells is close to zero, but not its opposite, or if both are close to zero. In the first case you have an asymmetric "If X then Y" but not the converse.

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.

If yours is a 2x2 table, then the absolute (unsigned) values of tau-b = Cramer's V, in which case checking whether Cramer's V is greater or equal to tau-b is meaningless. Note that even in a 3x2 table they could differ if the larger than expected cells are staggered opposites rather than ordered in one of the diagonals. In a 2x2 it is crucial to check if one of the cells is close to zero while the opposite on the diagonal is not. You would then have an asymmetric "If X then Y" but not the converse, or "If X then not Y" but not the converse, or "If not X then Y" but not the converse. The converses are: "If Y then X," "If Y then not X" and "If not Y then X." Pay attention here to the logical form of the relationship.

## 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 <sup>a</sup>	Approx. T <sup>b</sup>	Approx. Sig.
Nominal by	Phi	.557			.000
Nominal	Cramer's V	.278			.000035
Ordinal by	Kendall's tau-b	.365	.052	6.300	.00000000030
Ordinal	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.

## Statistical Analysis of Cross-Tabs

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.069 <sup>a</sup>	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](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 4: 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 $\rightarrow$ 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,  $\phi^2$  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 contribute 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

## Chapter 5

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)**

<i>Uxori-/Matrilocal Residence</i>	<i>General Non-Sororal Polygyny</i>		Total
	0 (absent)	1 (present)	
<i>0 (absent)</i>	<b>100</b> 73.5%	<b>46</b> 95.8%	146
<i>1 (present)</i>	<b>36</b> 26.5%	<b>2</b> 4.2%	38
<i>Total</i>	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  
 $\Phi = \text{Rho} = r = -0.24$ ;  $p = 0.001$   
 $\Gamma = -0.78$ ;  $p = 0.00001$

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.

## Chapter 5

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 matrilocality, but also if the absence of general nonsororal polygyny predicted as strongly the presence of matrilocality. 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 matrilocality. 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:*

<i>TRAIT X</i>	<i>TRAIT Y</i>	
	0 (absent)	1 (present)
<i>0 (absent)</i>		
<i>1 (present)</i>		

## Statistical Analysis of Cross-Tabs

$\phi$  [=  $\rho$  =  $r$ ] 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.