GOODMAN AND KRUSKAL'S TAU b: MULTIPLE AND PARTIAL ANALOGS

Louis N. Gray
Washington State University

J. Sherwood Williams
Virginia Commonwealth University

1. Introduction

In situations in which A and B are unordered polytomies, the asymmetrical measure of association proposed by Goodman and Kruskal [3] and labeled τ_b, is a useful and increasingly popular measure [1]. Like their λ_b measure, it has a clear operational interpretation and, as noted by Costner [2], may be given a proportional reduction in error (PRE) interpretation. Its primary advantage over their lambda measures lies in the fact that the aspect of association captured by the tau measures is not as easily obscured by highly skewed marginals. Additionally, since $\tau_b = 0$ is equivalent to independence, this measure is attractive to those who prefer to think of association in relation to statistical independence [3, p. 760].

In many research situations, however, we wish to examine the association between two polytomies when controlling for others, i.e., we wish to know the extent of effect of introduction of additional variables on the association between the initial two. In other research situations, we may be interested in the ability of two, or more, polytomies to predict, in some sense, the distribution of a dependent or criterion set of categories, i.e., multiple association. As greater emphasis in research areas focuses upon multivariate and causal analyses [7], measures of partial and multiple association become of more importance in answering practical research questions; while such extensions exist of lambda measures, it is of value to develop multivariate measures based upon the logic of τ_b, for use in those situations for which λ measures may seem inappropriate. Fortunately, it is possible to develop measures of this sort based upon the arguments for development of multiple and partial measures suggested by Goodman and Kruskal [3, p. 760-762].

2. The Basic Measure

Goodman and Kruskal [3, p. 759] describe τ_b as measure of proportional prediction based upon a method which reconstructs the population on the basis of the marginal distribution of the dependent variable and the conditional distribution of the independent variable. If the B polytomy is to be predicted and contains categories $B_1, B_2, \ldots, B_\beta$, and the A polytomy is used as the predictor and contains categories $A_1, A_2, \ldots, A_\alpha$, then τ_b is the relative decrease in error of correct placement. The prediction rules and probable errors are given by the following rules:

Case 1: Guess B_1 with probability $\rho_{.1}$, B_2 with probability $\rho_{.2}$, etc. The long run proportion of errors in case (1) will be $1-\Sigma_b \rho^2_{.b}$.

Case 2: Guess B_1 with probability $\rho_{a1}/\rho_{a.}$ (The conditional probability of B_1 given A_α), B_2 with probability $\rho_{a2}/\rho_{a.}$, etc. The long run proportion of errors in case (2) will be $1-\Sigma\Sigma \rho_{a.}(\rho_{ab}/\rho_{a.})^2$.

Hence the relative decrease in the proportion of incorrect predictions as we move from case (1) to case (2) is

$$\tau_b = \frac{(1-\Sigma_b \rho_{.b}^2)-(1-\Sigma\Sigma_{ab} \rho_{a.}(\rho_{ab}/\rho_{a.})^2)}{1-\Sigma_b \rho_{.b}^2} \qquad (2.1)$$

and a natural estimator for sample data is

$$t_b = \frac{\sum\sum \frac{f^2_{ab}}{f_{a.}} - \sum \frac{f^2_{.b}}{n}}{n - \sum \frac{f^2_{.b}}{n}} \qquad (2.2)$$

where f_{ab} indicates the observed frequency of the a^{th} category of A and the b^{th} category of B, $f_{a.}$ and $f_{.b}$ indicate marginal frequencies for the A and B polytomies, respectively, and n indicates the number of cases observed.

3. A Partial Coefficient

Goodman and Kruskal [3, p. 761] suggest two basic methods for the development of partial measures of association; we shall make use of the second, in which the measure is based directly on the probabilities of error. For the present exposition we shall consider only the three variable case, for purposes of simplicity, but extension to any number of variables is direct.

The approach we shall adopt utilizes information about the category of a third (control) variable to make proportional predictions about joint category membership. Our prediction and error rules now become:

Case 1: Guess B_1 with probability $\rho_{.1c}/\rho_{..c}$, B_2 with probability $\rho_{.2c}/\rho_{..c}$, etc. The long run proportion of errors in case (1) will be $1-\sum\sum_{bc}\rho_{..c}(\rho_{.bc}/\rho_{..c})^2$. In other words we are now guessing placement of cases within the categories of the C polytomy.

Case 2: Guess B_1 with probability $\rho_{a1c}/\rho_{a.c}$, B^2 with probability $\rho_{a1c}/\rho_{a.c}$, etc. The long run proportion of errors in case (2) will be $1-\sum\sum\sum_{abc}\rho_{a.c}(\rho_{abc}/\rho_{a.c})^2$.

Hence the relative decrease in the proportion of incorrect predictions as we move from case (1) to case (2) for the partial measure is

$$\tau_b(BA|C) = \frac{\sum\sum\sum_{abc}\rho_{a.c}(\rho_{abc}/\rho_{a.c})^2 - \sum\sum_{bc}\rho_{..c}(\rho_{.bc}/\rho_{..c})^2}{1-\sum\sum_{bc}(\rho_{..c}(\rho_{.bc}/\rho_{..c})^2} \qquad (3.1)$$

and a natural estimator for sample data is

$$t_b(BA|C) = \frac{\sum\sum\sum \frac{f^2_{abc}}{f_{a.c}} - \sum\sum \frac{f^2_{.bc}}{f_{..c}}}{n - \sum\sum \frac{f^2_{.bc}}{f_{..c}}} \qquad (3.2)$$

where f_{abc} indicates the joint frequency in the a^{th} category of A, the b^{th} category of B, and the c^{th} category of C, $f_{a.c}$ and $f_{.bc}$ indicate the joint (marginal) frequencies in the ac^{th} and bc^{th} categories, respectively, $f_{..c}$ indicates the marginal frequency in the c^{th} category of C, and n indicates the number of cases observed.

This measure may be given the same interpretation as τ_b with the addition that we allow the C polytomy to "explain" as much of the "variation" in the B polytomy as it can before investigating the relationship of B and A; in this sense it is somewhat analogous to the classical partial correlation coefficient.

4. A Multiple Coefficient

A multiple analogue to τ_b may also be developed, based upon the method suggested by Goodman and Kruskal [3, p. 761-762]. Again utilizing the three variable case for simplicity we see that our rules for prediction and error become:

Case 1: Guess B_1 with probability $\rho_{.1}$, B_2 with probability $\rho_{.2}$, etc. The long run proportion of errors in case (1) will be $1-\sum_b \rho^2_{.b}$.

Case 2: Guess B_1 with probability $\rho_{a1}/\rho_{a.c}$, B_2 with probability $\rho_{a2c}/\rho_{a.c}$, etc. The long run

the proportion of errors in case (2) will be $1-\Sigma\Sigma\Sigma_{a.c}(\rho_{abc}/\rho_{a.c})^2$.

Hence the relative decrease in the proportion of incorrect predictions as we move from case (1) to case (2) for the multiple measure is

$$\tau_b(B|AC)=\frac{\Sigma\Sigma\Sigma_{abc}\rho_{a.c}(\rho_{abc}/\rho_{a.c})^2-\Sigma_b\rho^2_{.b.}}{1-\Sigma_b\rho^2_{.b.}} \quad (4.1)$$

and a natural estimator for sample data is

$$t_b(B|AC)=\frac{\Sigma\Sigma\Sigma_{abc}\frac{f^2_{abc}}{f_{a.c}}-\Sigma_b\frac{f^2_{.b.}}{n}}{n-\Sigma_b\frac{f^2_{.b.}}{n}} \quad (4.2)$$

where all terms are as previously defined.

This measure may be given an interpretation similar to the basic measure, except that we are now using a set of independent variables (polytomies) to predict our dependent variable (polytomy) instead of a single variable; in this sense, it is somewhat analogous to the classical multiple correlation coefficient.

5. An Example

The following hypothetical example is taken from a study by Williams and McGrath (1975). The data has been manipulated so as to inflate the values of the Goodman and Kruskal Tau's. Table 1 is the source from which all of the following calculations are made.

The zero order coefficient between gun ownership and violence proneness is

$$t_b=\frac{[((91)^2+(119)^2)/210]+[((187)^2+(32)^2)/219]+[((38)^2+(155)^2)/193]-[((316)^2+(306)^2)/622]}{622-[((316)^2+(306)^2)/622]}$$

$$=\frac{106.867+164.352+131.964-311.080}{622-311.080}$$

$$=\frac{92.103}{310.920}=.296$$

To introduce the effect of the control variable, residence, the following calculations are necessary for t_b (ba|c)

$$\frac{[((12)^2+(10)^2)/22]+[((17)^2+(12)^2)/29]+[((15)^2+(14)^2)/29]+]((79)^2+(109)^2)/188]+[((170)^2+(20)^2)/190]+[((23)^2+(141)^2)/164]-[((44)^2+(36)^2)/80]-[((272)^2+(270)^2)/542]}{622-[((44^2+(36)^2)/80]-[((272)^2+(270)^2)/542]}$$

$$=\frac{415.595-311.404}{622-311.404}=.336$$

The partial coefficient of .336, when compared with the zero order coefficient .296, indicates that when residence is controlled the relationship between gun ownership and violence proneness becomes stronger.

The multiple coefficient for the Goodman and Kruskal τ_b suggested here has the same first term, in both the numerator and the denominator, as does the suggested partial coefficient. The second term is found by the following calculation:

Table 1. Residence, Violence Proneness, and Gun Ownership (Hypothetical Data).

	Residence (c)								
	Rural Violence Proneness (a)				Urban Violence Proneness (a)				Grand
Own Gun (b)	High	Medium	Low	Total	High	Medium	Low	Total	Total
Yes	12	17	15	44	79	170	23	272	316
No	10	12	14	36	109	20	141	270	306
Total	22	29	29	80	188	190	164	542	622

$$[((316)^2+(306)^2)/622] = 311.080$$

The multiple coefficient is

$$t_b\ (b|ac) = \frac{415.595-311.080}{622-311.080} = .336$$

When both predictor variables are considered together, the proportion of error reduced in predicting gun ownership is .336 as compared with the .296 PRE using only violence proneness to predict gun ownership.

6. Sampling Theory

In their third paper, Goodman and Kruskal [5], investigate asymptotic results for standard errors and variances for several of the measures they had suggested. By generalizing from their results it is possible to develop variance expressions for the multiple and partial measures suggested here. For the basic measure Goodman and Kruskal [5, p. 354] point out that the quantity $\sqrt{n}\ (t_b-\tau_b)$ is asymptotically normally distributed with zero mean and variance

$$4\{\Sigma\Sigma_{ab}[\frac{\rho_{ab}}{\rho_{a.}}(1-\Sigma_b\rho^2_{.b})-\rho_{.b}(1-\Sigma\Sigma_{ab}\frac{\rho^2_{ab}}{\rho_{a.}})]^2\rho_{ab}$$

$$-[\Sigma\Sigma_{ab}\frac{\rho^2_{ab}}{\rho_{a.}}-\Sigma_b\rho^2_{.b}]^2\}/(1-\Sigma_b\rho^2_{.b})^4 \qquad (6.1)$$

so that, if all terms are well-defined, $\sqrt{n}\ (t_b-\tau_b)$ divided by the square root of the sample analogue of (6.1) is asymptotically unit-normal.

The generalization for the partial τ_b measure is much like the generalization for λ_b measures. Essentially we need only add the additional subscripts and sum over the additional variables. Thus, the quantity $\sqrt{n}\{t_b(B,A|C)-\tau_b(B,A|C)\}$ is asymptotically normally distributed with zero mean and variance

$$4\{\Sigma\Sigma\Sigma_{abc}[\frac{\rho_{abc}}{\rho_{a.c}}(1-\Sigma\Sigma_{bc}\rho^2_{.bc})-\rho_{.bc}(1-\Sigma\Sigma\Sigma_{abc}\frac{\rho^2_{abc}}{\rho_{a.c}})]^2\rho_{abc}$$

$$-[\Sigma\Sigma\Sigma_{abc}\frac{\rho^2_{abc}}{\rho_{a.c}}-\Sigma\Sigma_{bc}\rho^2_{.bc}]^2\}/(1-\Sigma\Sigma_{bc}\rho^2_{.bc})^4 \qquad (6.2)$$

Thus, $\sqrt{n}$ multiplied by the difference between the estimated and hypothesized values of $\tau_b\ (B,A|C)$ divided by the sample analogue of (6.2) is asymptotically unit-normal.

The multiple measure is easily handled by the basic result presented in (6.1). If we call the A_aC_c combination by the new name D_d (a composite category) we may simply investigate the association between B and D. Thus, the multiple coefficient is a special case of the bivariate coefficient and a simple re-labeling of the subscripts in (6.1) allows us to utilize the basic results in testing the multiple coefficient.

Unfortunately, these results only apply in the situation in which the following conditions hold [5, p. 349-354]:

(1) There is separate multinomial sampling in the rows (columns);
(2) The row (column) margins are known;
(3) Sampling rates in the several rows (columns) are such that $n_{a.}=n\rho_{a.}$ --that is, the sample sizes in rows are proportional to the known row marginals.

Since it is rare that population row marginals are known in some research situations, these results are of limited value. On the other hand, they are suggestive and allow tests of hypotheses other than $\tau_b=0$, in situations in which the assumptions are felt to be approximately met.

7. Multiple-Partial Relations

It is of interest to note that the relationship between multiple and partial tau measures parallels that of classical correlation analysis. The classical relationship between multiple and partial measures is given by

$$1-R^2_{b.ac}=(1-\rho^2_{bc})(1-\rho^2_{ba.c}) \quad (7.1)$$

where ρ^2_{bc} and $\rho^2_{ba.c}$ here indicate the population correlations for the bivariate relationship and the partial relationship respectively. The analogous relationship for the tau measures is given by

$$1-\tau_b(B|A,C)=(1-\tau_b(B,C))(1-\tau_b(B,A|C)). \quad (7.2)$$

As noted by Goodman and Kruskal [5, p. 333] the same analogous relationship holds for their lambda measures as well, but, similar to their earlier results, we may not express multiple and partial tau measures as simple functions of overall zero-order relationships.

8. Discussion

The family of τ measures developed and elaborated by Goodman and Kruskal [3, 4, 5] would seem to be a particularly useful set of measures for survey or non-experimental data. It is particularly in such research situations that we find ourselves unable to control the marginals of some of the distributions of interest; thus, the use of λ measures may not capture some aspects of association of interest, due to extreme modality of response. The τ measures are particularly useful in this sort of situation since the aspect of association they capture is not so highly effected by marginal distributions; but partial and multiple applications, which may be useful for detailed causal or multivariate analyses [6, 7] have not previously been explicitly developed--it is hoped that the present contribution will serve to rectify this situation.

9. References

[1] Blalock, H. M., Social Statistics, McGraw-Hill, 1960, p. 232-234.

[2] Costner, H. L., "Criteria for Measures of Association," American Sociological Review, 30 (1965), 341-353.

[3] Goodman, L. and Kruskal, W., "Measures of Association for Cross Classification, Journal of the American Statistical Association, 49 (1954), 732-764.

[4] Goodman, L. and Kruskal, W., "Measures of Association for Cross Classification II: Further Discussion and References," Journal of the American Statistical Association, 54 (1959), 123-163.

[5] Goodman, L. and Kruskal, W., "Measures of Association for Cross Classification III: Approximate Sampling Theory," Journal of the American Statistical Association, 58 (1963), 310-364.

[6] Land, K. C. "Principles of Path Analysis," in E. F. Borgatta (Ed.) Sociological Methodology 1969, Jossey-Bass, 1968, p. 3-37.

[7] Simon, H. A. "Spurious Correlation: A Causal Interpretation," Journal of the American Statistical Association, 49 (1954), 467-479.

[8] Williams, J. S., and J. H. McGrath, III, "Social Psychological Dimensions of Gun Ownership," a paper presented to the American Society of Criminology, 1975.