The Statistician (1995)
44, *No.* 2, *pp.* 195–212

The behaviour of nominal and ordinal partial association measures

By MICHAEL OLSZAK and GILBERT RITSCHARD†

University of Geneva, Switzerland

[Received September 1993. Final revision September 1994]

SUMMARY

This paper is concerned with the sampling behaviour of raw and partial measures of association between categorical variables. It summarizes the asymptotic results established for raw measures and extends them in a systematic way for the derived partial associations. The validity of the asymptotic results is then stressed by means of a simulation study. Three proportional reduction in error of prediction measures are considered for nominal variables and three concordance–discordance indices for ordinal variables.

Keywords: Categorical variables; Multiway contingency tables; Nominal and ordinal partial association measures

1. Introduction

In social science research, it is often of interest to summarize the causal dependence between variables with a valued causal graph. The strength of direct and indirect links between metric variables is then summarized with regression coefficients. For links between categorical variables, the task is less obvious, however. Recent developments in inferential categorical data analysis are dominated by parametric model fitting. Nevertheless, no model has been proposed so far which provides single synthesized indices of the direct and indirect links between polytomous variables. Log-linear models are mainly concerned with the detection of association patterns. Like factorial correspondence analysis, they focus on the linkages between categories rather than variables, whereas models like logistic or Poisson regression try to explain the probability of falling in a given category. Parameters of a logistic regression may provide insight on the link between dichotomous variables. Nevertheless, for polytomous variables, we do not obtain synthesized indices, but a whole set of parameters. Likewise, Goodman's row–column association model and its variants provide some kind of association parameters. They apply, however, to ordinal variables only. Furthermore, the number of parameters exceeds the number of two-by-two links when more than two variables are considered. Thus, the parametric modelling approach is, generally, of little help for valuing and testing a causal graph between categorical variables. The old association measures, and their derived partial association indices, remain the most suitable tools for that purpose. This paper deals with the inferential aspects regarding their estimates. It recalls the asymptotic variance of raw measures and extends them in a systematic way for partial measures. It then investigates the behaviour of the estimates through simulation studies.

Several raw association indices have been proposed. Pearson (1948), Tschuprow (1918) and Cramer (1946) have, for instance, proposed measures based on Pearson's X^2. There are also proportional reduction in error (PRE) indices which measure the proportion of the reduction

†*Address for correspondence*: Department of Econometrics, University of Geneva, 102 Boulevard Carl-Vogt, CH-1211 Geneva 4, Switzerland.
E-mail: ritschar@uni2a.unige.ch

 0039–0526/95/44195

in the error of prediction gained by using the information on the first variable for predicting the second. Nominal PRE measures have mainly been introduced by Goodman and Kruskal (1954, 1963, 1972) and Theil (1970). Kendall (1938, 1945), Stuart (1953), Somers (1962) and Wilson (1974) have developed measures that are applicable to ordinal variables. For brevity, we consider only three nominal PRE measures and three ordinal measures. Section 2 recalls their definitions and introduces notation.

Section 3 is devoted to the measure of partial association. A few such measures have been derived from raw indices, e.g. by Goodman and Kruskal (1954), Davis (1967) and Quade (1974). Their approach is systematized here in the form of a general expression that is applicable to any ordinal or nominal PRE measure. We propose also, in Appendix A, computable formulae for the asymptotic variance of the estimates of all partial indices considered.

Section 4 provides an empirical insight into the behaviour of the partial indices and their estimates. It reports simulation results which illustrate the sample behaviour of the estimates. Results are also given which illustrate the effect of a reinforcement of the association on the one hand, and that of a reinforcement of interaction on the other hand. Finally, the consequences of disaggregating a category are examined.

2. Notation and raw association measures

The joint distribution of two categorical variables A and B is characterized by a simple two-way contingency table. Let A be the row variable with a set I of l categories and B the column variables with a set J of c categories. We denote by p_{ij} the joint probability $P(A = i, B = j)$, by p_{i+} the probability of being in row i, i.e. $p_{i+} = P(A = i)$, and by p_{+j} the probability of being in column j, i.e. $p_{+j} = P(B = j)$. These define the theoretical distribution.

For estimations we consider the following sampled quantities: n_{ij}, the number of observations falling into the cell (i, j), $n_{i+} = \Sigma_{j\in J} n_{ij}$, the total of observations in row i, $n_{+j} = \Sigma_{i\in I} n_{ij}$, the total in column j, and n the grand total of observations, i.e. $n = \Sigma_i\Sigma_j n_{ij}$.

Letting A be the dependent variable and B the independent variable, the three main measures are

$$\lambda_{AB} = \frac{\sum_j p_{mj} - p_{m+}}{1 - p_{m+}}, \tag{1}$$

$$\tau_{AB} = \frac{\sum_i \sum_j p_{ij}^2/p_{+j} - \sum_i p_{i+}^2}{1 - \sum_i p_{i+}^2}, \tag{2}$$

$$u_{AB} = \frac{\sum_i \sum_j p_{ij} \log_2\left(\dfrac{p_{i+}p_{+j}}{p_{ij}}\right)}{\sum_i p_{i+} \log_2 p_{i+}} \tag{3}$$

where p_{mj} and p_{m+} are respectively the maximum in column j and the maximum among the row totals.

The first two are due to Goodman and Kruskal (1954). The first, λ, presupposes a deterministic prediction rule, whereas the second, τ, is based on a stochastic prediction rule. The third is Theil's uncertainty coefficient (Theil, 1970) based on Shannon's (1948) measure of entropy.

Sample estimates of the above measures are obtained by replacing the probabilities p_{ij}, p_{i+} and p_{+j} with the sample frequencies n_{ij}/n, n_{i+}/n and n_{+j}/n. Asymptotic standard errors of the estimates are reproduced in Appendix A.

Ordinal measures estimate the difference between the chance of finding a pair of observations with concordant ranking on the two ordinal variables considered and the chance of finding a discordant pair. The measures take their value between -1 and 1.

The most popular measures, among which are Kendall's τ_a (Kendall, 1938) and τ_b (Kendall, 1945), Stuart's (1953) τ_c, Goodman and Kruskal's (1954) γ, Somers's (1962) d and Wilson's (1974) e, differ mainly in the way in which they account for ties. The first, τ_a, can simply be expressed as the difference between the probability π^{c} of a concordant pair and that, π^{d}, of a discordant pair. The others are standardized forms of τ_a which, unlike τ_a, can reach their bounds -1 and 1 for contingency tables. Let π^{t}_A, π^{t}_B and π^{t}_{AB} denote respectively the probabilities of a pair with a tie on the dependent variable A only, on the independent variable B only and on both variables A and B. The main standardized measures read

$$\gamma = \frac{\pi^{\mathrm{c}} - \pi^{\mathrm{d}}}{\pi^{\mathrm{c}} + \pi^{\mathrm{d}}}, \tag{4}$$

$$d_{AB} = \frac{\pi^{\mathrm{c}} - \pi^{\mathrm{d}}}{\pi^{\mathrm{c}} + \pi^{\mathrm{d}} + \pi^{\mathrm{t}}_A}, \tag{5}$$

$$\tau_b = \frac{\pi^{\mathrm{c}} - \pi^{\mathrm{d}}}{\{(\pi^{\mathrm{c}} + \pi^{\mathrm{d}} + \pi^{\mathrm{t}}_A)(\pi^{\mathrm{c}} + \pi^{\mathrm{d}} + \pi^{\mathrm{t}}_B)\}^{1/2}}. \tag{6}$$

Alternative forms can easily be derived, using for instance the equalities

$$\begin{aligned}\pi^{\mathrm{c}} + \pi^{\mathrm{d}} + \pi^{\mathrm{t}}_A &= 1 - \pi^{\mathrm{t}}_B - \pi^{\mathrm{t}}_{AB} \\ &= 1 - \sum_j p^2_{+j}.\end{aligned} \tag{7}$$

These measures refer to different degrees of perfect association. The smaller the denominator, the weaker the degree is. Goodman and Kruskal's γ, for instance, takes account only of the pairs without any tie. Thus, for $\pi^{\mathrm{d}} = 0$, we obtain a perfect positive association even for very small (but non-zero) values of π^{c}. It is also worth mentioning that, for non-square tables with no empty rows or columns, τ_b cannot reach its bounds and Somers's d can reach them only when the dependent variable has at least as many categories as the independent variable.

Sample estimates of the ordinal measures are obtained by replacing the probabilities π^{c}, π^{d}, π^{t}_A, π^{t}_B and π^{t}_{AB} with the sample frequencies C/T, D/T, T^A/T, T^B/T and T^{AB}/T, where $T = n(n-1)/2$ is the total number of pairs, C the number of concordant pairs, D the number of discordant pairs, T^A, T^B and T^{AB} the number of pairs with ties respectively only on A, only on B and on both A and B. Asymptotic variances of these estimates are given in Appendix A.

3. Partial association measures

A few partial association measures have been proposed for categorical variables. Goodman and Kruskal (1954) have discussed partial measures based on their λ_{AB}, and Davis (1967) has proposed a partial coefficient for Goodman and Kruskal's γ. See Quade (1974) for a more general discussion on partial correlation for metric and ordinal data. We propose a unified approach for defining nominal and ordinal partial measures.

We define a partial coefficient as a weighted average of conditional association indices computed for each state of a third variable E. Assume that E has a set K of q categories, and let $\theta_{AB|k}$ denote a generic association index measuring the association between A and B for a given state k of E. The partial measure is then

$$\theta_{AB|E} = \sum_{k \in K} \omega_k \theta_{AB|k}, \tag{8}$$

where the weights ω_k are non-negative and sum to 1.

This definition extends straightforwardly to the case of more than three variables. Indeed, we then must simply consider E as the set of all combinations of the categories of the variables which we want to control for.

The difficulty concerns the choice of the weights. A first approach, which corresponds to the first alternative considered by Goodman and Kruskal (1954), is to take the probabilities p_{++k} of each state k of the conditional variable E. A more satisfactory way is to base the weights on the denominator in the definition of the conditional indices. We thus obtain partial measures which can be interpreted in the same terms as their raw counterparts.

For PRE nominal measures, this denominator is $p(\text{error}|k)$, the probability, when we are in state k of E, of making a prediction error on A in the absence of information on B. We suggest that the conditional indices should be weighted according to the joint probability of being in state k of E and making a prediction error, i.e. $p_{++k}\, p(\text{error}|k)$. The weights are then of the form

$$\omega_k = \frac{p_{++k}\, p(\text{error}|k)}{\sum_{k'} p_{++k'}\, p(\text{error}|k')}. \tag{9}$$

For Goodman and Kruskal's λ and τ, and Theil's u, these weights are respectively

$$\omega_k^{\lambda} = \frac{p_{++k} - p_{m+k}}{1 - \sum_{k'} p_{m+k'}}, \tag{10}$$

$$\omega_k^{\tau} = \frac{p_{++k} - \sum_j p_{i+k}^2/p_{++k}}{1 - \sum_{k'}\sum_i p_{i+k'}^2/p_{++k}}, \tag{11}$$

$$\omega_k^{u} = \frac{\sum_i p_{i+k} \log_2(p_{i+k}/p_{++k})}{\sum_{k'}\sum_i p_{i+k'} \log_2(p_{i+k'}/p_{++k})}. \tag{12}$$

We thus obtain the following partial λ, $\lambda_{AB|E}$, partial (nominal) τ, $\tau_{AB|E}$, and partial uncertainty coefficient $u_{AB|E}$

$$\lambda_{AB|E} = \frac{\sum_k \left(\sum_j p_{mjk} - p_{m+k}\right)}{1 - \sum_k p_{m+k}}, \tag{13}$$

$$\tau_{AB|E} = \frac{\sum_k \left(\sum_i \sum_j p_{ijk}^2/p_{+jk} - \sum_i p_{i+k}^2/p_{++k}\right)}{1 - \sum_k \sum_i p_{i+k}^2/p_{++k}}, \tag{14}$$

$$u_{AB|E} = \frac{\sum_k \sum_i \sum_j p_{ijk} \log_2(p_{i+k}p_{+jk}/p_{ijk}p_{++k})}{\sum_k \sum_i p_{i+k} \log_2(p_{i+k}/p_{++k})}. \tag{15}$$

Note that index (13) is precisely Goodman and Kruskal's second alternative.

A similar approach has been advocated by Quade (1974) for those ordinal association measures which can be expressed in the form

$$\theta = \frac{\pi^c - \pi^d}{\pi^r}, \tag{16}$$

where π^r denotes the probability of obtaining a relevant pair. For Somers's d_{AB}, for example, π^r is given by $\pi^c + \pi^d + \pi^t_A$. Partial ordinal measures are thus obtained by averaging the conditional indices according to the probabilities π^r_k of obtaining a relevant pair tied on the kth state of E. It is worth mentioning that the π^r_k are joint probabilities, i.e. $\pi^r_k = p^2_{++k}\pi^{r|k}$, where p^2_{++k} is the probability of obtaining a pair with both observations in the kth subtable and $\pi^{r|k}$ the conditional probability of a relevant pair among those tied on the state k of E.

Resulting partial measures for γ, τ_b and Somers's d, for example, are then

$$\gamma_{AB|E} = \frac{\sum_k (\pi^c_k - \pi^d_k)}{\sum_k (\pi^c_k + \pi^d_k)}, \tag{17}$$

$$d_{AB|E} = \frac{\sum_k (\pi^c_k - \pi^d_k)}{\sum_k (\pi^c_k + \pi^d_k + \pi^t_{Ak})}, \tag{18}$$

$$\tau^b_{AB|E} = \frac{\sum_k (\pi^c_k - \pi^d_k)}{\sum_k (\pi^c_k + \pi^d_k + \pi^t_{Ak})^{1/2}(\pi^c_k + \pi^d_k + \pi^t_{Bk})^{1/2}} \tag{19}$$

where π^t_{Ak} denotes the probability of obtaining a pair tied on A and the kth state of E but not on B. For computation, we may use the equalities

$$\pi^c_k + \pi^d_k + \pi^t_{Ak} = p^2_{++k} - \sum_j p^2_{+jk}, \tag{20}$$

$$\pi^c_k + \pi^d_k + \pi^t_{Bk} = p^2_{++k} - \sum_i p^2_{i+k}, \tag{21}$$

$$\pi^c_k + \pi^d_k = p^2_{++k} - \sum_i p^2_{i+k} - \sum_j p^2_{+jk} + \sum_j \sum_i p^2_{ijk}. \tag{22}$$

Appendix A provides the asymptotic variances of the estimates. The variance for the estimate of the partial λ can also be found in Goodman and Kruskal (1963). We established the others by means of the δ-rule. The variances found for the ordinal measures can be shown to be alternative expressions of the formula given by Quade (1974). Ours sum over cells, whereas Quade gives an expression where the sums are over the observations and is, therefore, computationally less efficient.

4. Simulation study

The purpose of this section is to provide empirical insight into the behaviour of the measures of partial association. We first examine the behaviour of the sample estimates of the partial indices as well as the estimates of their standard deviation. Section 4.2 presents a comparative study of the evolution of the various partial measures and their asymptotic variances when we smoothly go from full independence to perfect association. Similarly, we examine the effect of interaction by considering a smooth change from no to full interaction. A final simulation illustrates the consequences of disaggregating a category. Except when otherwise indicated, the row variable A is considered to be dependent.

4.1. *Sample behaviour of partial measure estimates*

We present here simulation results for samples drawn from four theoretical distributions for a 3 × 3 × 3 cross-table. Tables 1–4 describe the distributions considered. They provide the marginal distribution of the third variable (p_{++k}). The three cross-tables give the joint distributions ($p_{ij|k}$) of the first two variables conditional on each category of the third variable.

The first distribution D (Table 1) corresponds to near dependence. We did not retain full dependence to allow some discrepancy between the samples drawn. The second distribution M (Table 2) is a case of medium dependence, and the third distribution I (Table 3) a case of near independence. The fourth distribution X (Table 4) exhibits interaction, i.e. strong differences between conditional distributions.

For each distribution, 3000 samples of 300 units were drawn. For our 3 × 3 × 3 tables, the sample size $n = 300$ corresponds to an average cell frequency of approximately 11. Table 5 shows the figures obtained for the partial measures and Table 6 their standard errors.

TABLE 1
Near dependence case D

$p_{++1} = 0.1$			$p_{++2} = 0.3$			$p_{++3} = 0.6$		
0.33	0.01	·	0.33	·	0.01	0.33	·	·
·	0.33	·	·	0.33	·	·	0.33	0.01
·	·	0.33	·	·	0.33	·	·	0.33

TABLE 2
Medium association case M

$p_{++1} = 0.1$			$p_{++2} = 0.3$			$p_{++3} = 0.6$		
0.3	0.05	0.05	0.2	0.05	0.05	0.2	0.05	0.05
0.05	0.2	0.05	0.05	0.3	0.05	0.05	0.2	0.05
0.05	0.05	0.2	0.05	0.05	0.2	0.05	0.05	0.3

TABLE 3
Near independence case I

$p_{++1} = 0.1$			$p_{++2} = 0.3$			$p_{++3} = 0.6$		
0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.11	0.11
0.11	0.11	0.11	0.11	0.12	0.11	0.11	0.11	0.11
0.11	0.11	0.12	0.11	0.11	0.11	0.11	0.11	0.11

TABLE 4
Interaction case X

$p_{++1} = 0.1$			$p_{++2} = 0.3$			$p_{++3} = 0.6$		
0.11	0.11	0.11	0.33	0.01	·	0.2	0.05	0.05
0.11	0.12	0.11	·	0.33	·	0.05	0.3	0.05
0.11	0.11	0.11	·	·	0.33	0.05	0.05	0.2

TABLE 5
Simulation results: estimates and bias†

		θ	$\bar{\hat{\theta}}$	$\bar{\hat{\theta}}-\theta$			θ	$\bar{\hat{\theta}}$	$\bar{\hat{\theta}}-\theta$
$\lambda_{AB\|E}$	D	0.9849	0.9839	−0.0009 (0.00)	$\gamma_{AB\|E}$	D	0.9961	0.9960	−0.0001 (0.11)
	M	0.5000	0.4949	−0.0051 (0.00)		M	0.6875	0.6884	0.0009 (0.19)
	I	0.0000	0.0555	0.0555 (0.00)		I	0.0159	0.0153	−0.0007 (0.33)
	X	0.6010	0.6093	0.0083 (0.00)		X	0.7035	0.7027	−0.0008 (0.24)
$\tau_{AB\|E}$	D	0.9709	0.9716	0.0007 (0.01)	$d_{AB\|E}$	D	0.9862	0.9862	0.0000 (0.42)
	M	0.2992	0.3148	0.0155 (0.00)		M	0.5260	0.5278	0.0017 (0.03)
	I	0.0002	0.0205	0.0203 (0.00)		I	0.0106	0.0102	−0.0004 (0.34)
	X	0.4719	0.4836	0.0118 (0.00)		X	0.5679	0.5681	0.0002 (0.43)
$u_{AB\|E}$	D	0.9589	0.9635	0.0046 (0.00)	$\tau^{b}_{AB\|E}$	D	0.9862	0.9862	−0.0001 (0.32)
	M	0.2517	0.2756	0.0239 (0.00)		M	0.5260	0.5278	0.0018 (0.03)
	I	0.0002	0.0199	0.0197 (0.00)		I	0.0106	0.0102	−0.0004 (0.34)
	X	0.4397	0.4555	0.0158 (0.00)		X	0.5679	0.5681	0.0002 (0.44)

† Values in parentheses are p-values. θ denotes the true value of the partial association index and $\bar{\hat{\theta}}$ the mean estimated value among the m simulations.

Note that the ordinal measures behave much more reliably than the nominal measures. The greatest average bias (difference between the mean estimated value of the partial measure and its true value) is only 0.0018 for the ordinal measures, but about 30 times greater, i.e. 0.0555, for the nominal measures. The estimated nominal partial measures, however, seem to be slightly less scattered. Thus, unsurprisingly, using a t-test, the mean estimated values of

TABLE 6
Simulation results: standard deviations†

		σ_∞	σ_s	$\bar{\hat{\sigma}}_\infty$	$\sigma_s-\sigma_\infty$	$\bar{\hat{\sigma}}_\infty-\sigma_\infty$	$\bar{\hat{\sigma}}_\infty-\sigma_s$
$\lambda_{AB\|E}$	D	0.0088	0.0092	0.0093	0.0004 (0.00)	0.0005 (0.00)	0.0001
	M	0.0481	0.0484	0.0492	0.0003 (0.18)	0.0011 (0.00)	0.0008
	I	0.0367	0.0261	0.0494	−0.0106 (0.00)	0.0127 (0.00)	0.0233
	X	0.0431	0.0411	0.0433	−0.0020 (0.02)	0.0002 (0.00)	0.0022
$\tau_{AB\|E}$	D	0.0162	0.0158	0.0155	−0.0004 (0.04)	−0.0008 (0.00)	−0.0004
	M	0.0434	0.0440	0.0423	0.0007 (0.00)	−0.0011 (0.00)	−0.0017
	I	0.0011	0.0082	0.0111	0.0070 (0.00)	0.0100 (0.00)	0.0030
	X	0.0399	0.0405	0.0396	0.0006 (0.00)	−0.0003 (0.00)	−0.0008
$u_{AB\|E}$	D	0.0184	0.0184	0.0162	−0.0001 (0.02)	−0.0022 (0.00)	−0.0022
	M	0.0370	0.0394	0.0369	0.0024 (0.00)	−0.0001 (0.00)	−0.0025
	I	0.0010	0.0081	0.0105	0.0071 (0.00)	0.0094 (0.00)	0.0024
	X	0.0365	0.0378	0.0363	0.0013 (0.00)	−0.0001 (0.00)	−0.0015
$\gamma_{AB\|E}$	D	0.0042	0.0043	0.0044	0.0001 (0.01)	0.0002 (0.00)	0.0001
	M	0.0566	0.0576	0.0568	0.0011 (0.06)	0.0003 (0.00)	−0.0008
	I	0.0828	0.0832	0.0830	0.0004 (0.35)	0.0002 (0.02)	−0.0002
	X	0.0617	0.0626	0.0619	0.0009 (0.12)	0.0002 (0.09)	−0.0007
$d_{AB\|E}$	D	0.0084	0.0084	0.0085	0.0000 (0.41)	0.0001 (0.01)	0.0001
	M	0.0521	0.0530	0.0520	0.0010 (0.07)	−0.0001 (0.45)	−0.0010
	I	0.0552	0.0557	0.0554	0.0005 (0.29)	0.0002 (0.00)	−0.0003
	X	0.0584	0.0592	0.0583	0.0009 (0.13)	−0.0001 (0.09)	−0.0009
$\tau^{b}_{AB\|E}$	D	0.0084	0.0084	0.0085	0.0001 (0.29)	0.0001 (0.01)	0.0001
	M	0.0520	0.0529	0.0519	0.0010 (0.07)	−0.0001 (0.18)	−0.0010
	I	0.0552	0.0557	0.0554	0.0005 (0.29)	0.0002 (0.00)	−0.0003
	X	0.0583	0.0592	0.0582	0.0009 (0.15)	−0.0001 (0.04)	−0.0009

†Values in parentheses are p-values. σ_∞ denotes the asymptotic standard deviation of $\hat{\theta}$, σ_s the standard deviatio of the m computed $\hat{\theta}$s, and $\bar{\hat{\sigma}}_\infty$ the square root of the mean estimated asymptotic variance.

the nominal partial measures differ significantly at the 5% level from their corresponding true values. The mean bias seems statistically negligible for the ordinal measures.

To evaluate the standard deviation, three values have been computed (Table 6): the theoretical asymptotic standard deviation σ_∞ computed from the true distribution, the standard deviation σ_s of the $m = 3000$ sample estimates $\hat{\theta}$ of the partial association measure and $\bar{\hat{\sigma}}_\infty$, the square root of the mean of the estimated asymptotic variances.

Figs 1 and 2 show that all three measures are very near each other for the ordinal indices and differ slightly, especially in the independence case I, for the nominal indices. The estimated standard error is either very near the sample standard error or overestimates it. The hypothesis $V(\hat{\theta}) = \sigma^2_\infty$ can be tested with a χ^2-test based on σ^2_s, and the hypothesis $E(\hat{\sigma}^2_\infty) = \sigma^2_\infty$ by means of a t-test based on $\bar{\hat{\sigma}}^2_\infty$. The corresponding p-values are given in Table 6. Both hypotheses are almost always rejected for the nominal measures. The second hypothesis is also rejected for ordinal measures in the cases of near dependence and near independence. There are thus obviously some biases. However, the biases remain very small in absolute terms, i.e. less than 0.013 for the standard deviation of the nominal indices, and less than 0.0011 for that of the ordinal indices. Compare these figures, for instance, with the differences between the values of the six indices computed for the same table. The standard deviations (not shown in Table 6) of the 3000 computed $\hat{\sigma}^2_\infty$s remain less than 0.001. Adding, for instance one such standard deviation to a 0.04^2 variance increases its square root by about 0.01. This is virtually the maximal bias observed for nominal cases, but considerably more than the biases for ordinal cases.

Except for the dependence case D, the standard deviations are greater for the ordinal measures. Also, the partial measure based on Goodman and Kruskal's γ has, except for case D, the greatest variance among all six measures. Likewise, the partial λ appears to be the less reliable nominal measure.

To summarize, the estimates of the association measures, as well as those of their asymptotic variances, seem to be slightly biased for the three nominal indices considered. The biases are not statistically significant for the ordinal measures. The figures computed show, nevertheless, that the biases remain small and that the estimates of the asymptotic

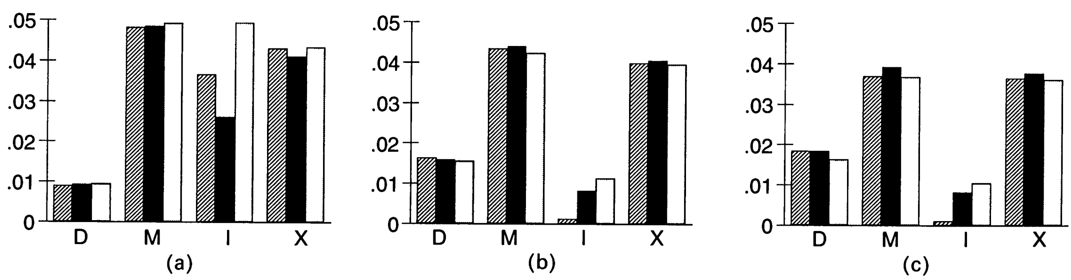

Fig. 1. Standard deviates (σ_∞, σ_s, $\bar{\hat{\sigma}}_\infty$) for the nominal partial measures (a) λ, (b) τ and (c) u

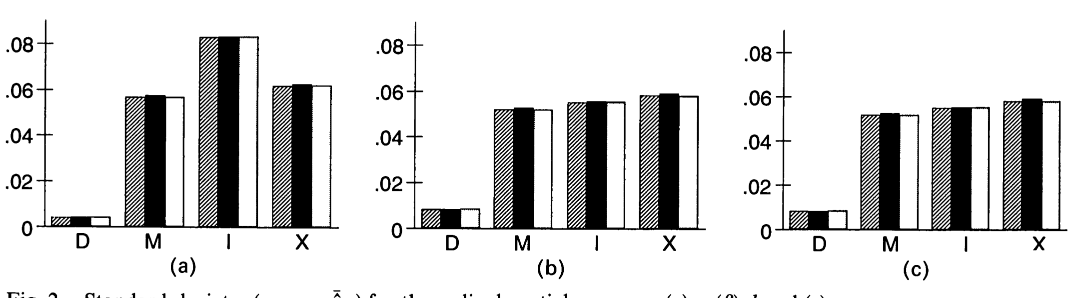

Fig. 2. Standard deviates (σ_∞, σ_s, $\bar{\hat{\sigma}}_\infty$) for the ordinal partial measures (a) γ, (β) d and (c) τ_b

variance provide reasonably good and, hence, reliable approximations of the variances of the estimators.

These conclusions remain true for smaller samples. We ran a similar simulation with 1000 samples of size $n = 100$ corresponding to an average of 3.7 units per cell. The computed biases and standard deviation are, indeed, greater (approximately doubled), but the results remain qualitatively the same. The bias remains less than 0.1 for the nominal measures and less than 0.0025 for the ordinal measures. Likewise, the standard deviation is less than 0.1, except for the partial γ.

4.2. *From independence to perfect association*

We examine now how the partial association indices and their asymptotic standard errors evolve between full independence and complete dependence. We have thus progressively transformed a distribution which exhibits perfect independence into one which exhibits perfect association. The simulation was carried out for $3 \times 3 \times 2$ cross-tables. Table 7 describes the initial distribution and Table 8 the final distribution. The intermediate situations were generated through 200 equally spaced steps. In this simulation, we obtain at each step two identical conditional distributions. The partial association indices are then equal to the raw measures.

The six partial measures as well as the six asymptotic standard deviations were computed at each step assuming a sample size of $n = 100$.

Fig. 3 shows how the measures behave. Note, for instance, how Goodman and Kruskal's τ and the uncertainty coefficient u behave similarly. Note also the similarity between the ordinal measures τ_b and d. As we might have expected, the γ-measure always provides higher values than the other indices. λ remains at 0 as long as the column maxima remain in the same row.

Fig. 4 plots the evolution of the asymptotic standard error of the estimates of the partial measures. Again, the asymptotic errors of Somers's d and of Kendall's τ_b on the one hand and those of Goodman and Kruskal's τ and the uncertainty coefficient u on the other hand behave very similarly. The jumps in the curve of the standard deviation of λ illustrate the discontinuities induced by row changes of the column maxima. Note, however, that the formula used does not take account of the possible row changes of the column maxima. It is therefore not valid in the neighbourhood of points where these changes occur. The most

TABLE 7
Initial distribution: independence

$p_{++1} = 0.5$			$p_{++2} = 0.5$		
1/6	1/6	1/6	1/6	1/6	1/6
1/9	1/9	1/9	1/9	1/9	1/9
1/18	1/18	1/18	1/18	1/18	1/18

TABLE 8
Final distribution: perfect association

$p_{++1} = 0.5$			$p_{++2} = 0.5$		
1/3	0	0	1/3	0	0
0	1/3	0	0	1/3	0
0	0	1/3	0	0	1/3

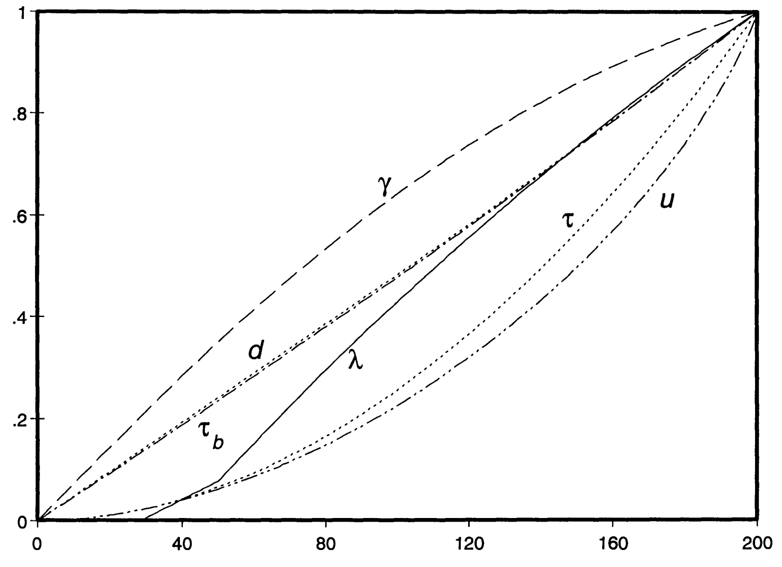

Fig. 3. Evolution of the association measures during a move from independence to dependence

interesting information provided by Fig. 4 is that nominal measures have a smaller variance around independence, whereas the nominal τ and u have greater variance in the case of strong association. The first point follows from the fact that ordinal measures may take negative values, whereas nominal measures are bounded below by 0.

4.3. *Effect of interaction*

The previous study considered situations without interaction. A similar analysis is undertaken here to illustrate the effect of interaction, i.e. of differences between the two conditional cross-tabulated distributions.

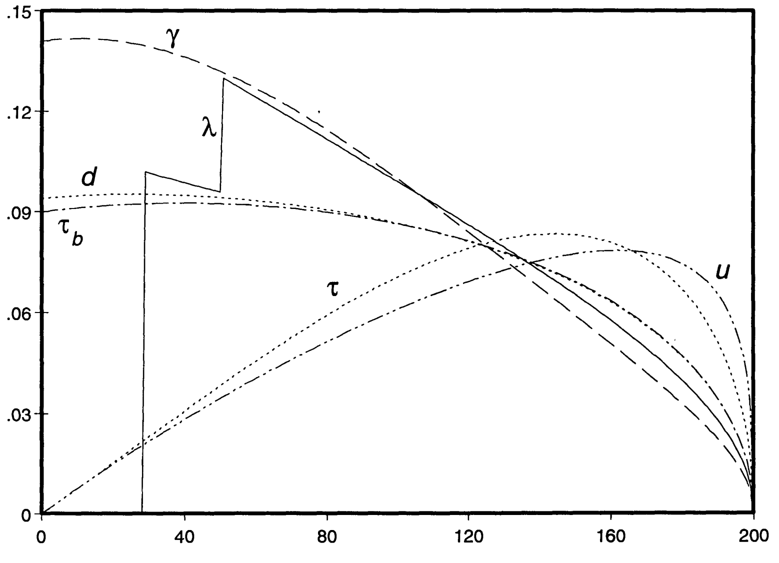

Fig. 4. Evolution of the asymptotic standard errors during a transition from independence to dependence

We have thus progressively transformed a case with two identical conditional distributions towards a case with one distribution exhibiting perfect positive association and the other perfect negative association. The initial case is the same as that used in the previous experiment (Table 7). The final situation is described in Table 9. Again we consider 200 equally spaced steps.

The figures obtained correspond to a sample size of $n = 100$. Since the two categories of the third variable are assumed equiprobable, the partial measures are a simple average of the indices computed for each of the two conditional tables. This implies, here, that the ordinal partial measures are 0 at each step. Likewise, since nominal measures are not affected by permutations of rows, the partial nominal measures are exactly the same as in the previous experiment (Fig. 5).

Let us look at the asymptotic standard errors. Fig. 6 shows that the standard error increases when we reinforce interaction for all three ordinal partial measures. For the nominal partial indices, the effect of interaction is the same here as that of dependence. Remember, however, that, unlike the ordinal partial measures, the values of the nominal measures change along the path generated.

4.4. *Disaggregating categories*

The aggregation level of the categories obviously influences the association measures (Fig. 7). We illustrate this point here by successively disaggregating the second column category of a small initial 2 × 2 table. The theoretical distributions considered are depicted in Table 10.

TABLE 9
Final distribution: strong interaction

$p_{++1} = 0.5$			$p_{++2} = 0.5$		
1/3	0	0	0	0	1/3
0	1/3	0	0	1/3	0
0	0	1/3	1/3	0	0

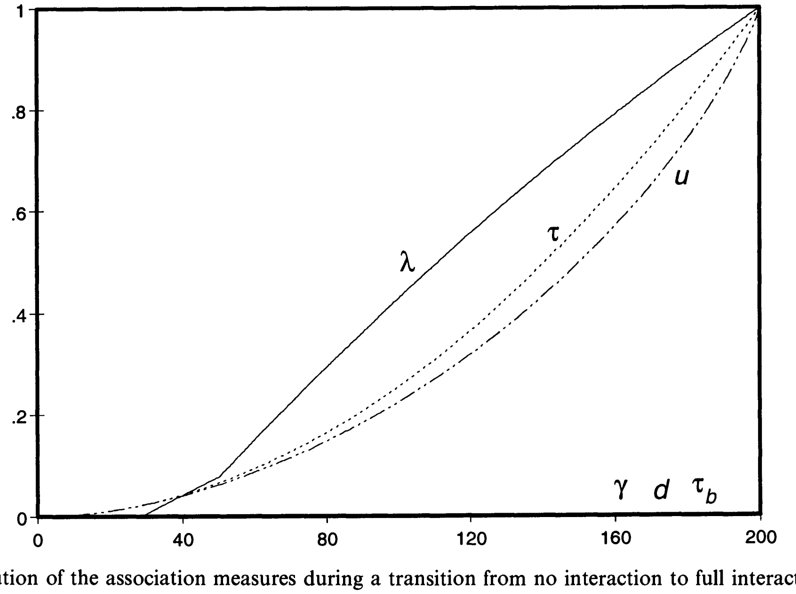

Fig. 5. Evolution of the association measures during a transition from no interaction to full interaction

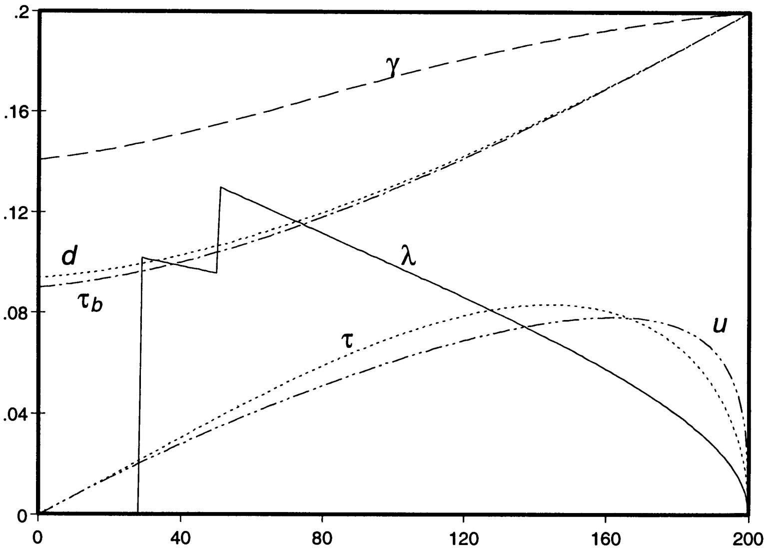

Fig. 6. Evolution of the asymptotic standard errors during a transition from no interaction to full interaction

The consequences of the disaggregation on asymmetrical measures depend on the nature of the variable concerned. Consider first the case where we modify the independent variable. The nominal measures are not affected; nor are their standard errors. Ordinal measures all decrease to some asymptotic value: $\hat{\gamma}$ from 0.88 to 0.72, $\hat{\tau}_b$ from 0.6 to 0.49 and d from 0.6 to 0.4. The standard errors of $\hat{\tau}_b$ and $\hat{d}$ also decrease, whereas the discrepancy of $\hat{\gamma}$ increases (Fig. 8). If the modified variable is now the dependent variable, the nominal measures decrease strongly: $\hat{\lambda}$ from about 0.6 to 0, $\hat{\tau}$ from 0.36 to 0.12 and $\hat{u}$ from 0.28 to 0.08. Their discrepancy also decreases. Among the ordinal measures, $\hat{d}$ remains constant, whereas the symmetric measures $\hat{\gamma}$ and $\hat{\tau}_b$ behave, of course, as in the previous case. Note that the standard error of $\hat{d}$ increases slightly though $\hat{d}$ itself remains constant. We obtain similar, but amplified, evolutions of the measures and their standard errors when we disaggregate symmetrically the two columns of the starting matrix. The only exception concerns $\hat{\lambda}$ which decreases smoothly when the column variable is dependent.

Note that the decreases observed in the association indices are valid for the specific homogeneous disaggregation rule considered, i.e. one that does not add any new information. In practice, however, we usually have some knowledge about the distribution inside categories. Using this information when splitting categories may then lead sometimes to a reinforcement of the association.

It is worth mentioning that Sections 4.2–4.4 are only illustrative. The results rely on the extreme cases chosen and on the arbitrary path followed from one extreme case to the other.

TABLE 10
Column disaggregation

Initial distribution		*After k − 1 steps*			
0.4	0.1	0.4	0.1/k	...	0.1/k
0.1	0.4	0.1	0.4/k	...	0.4/k

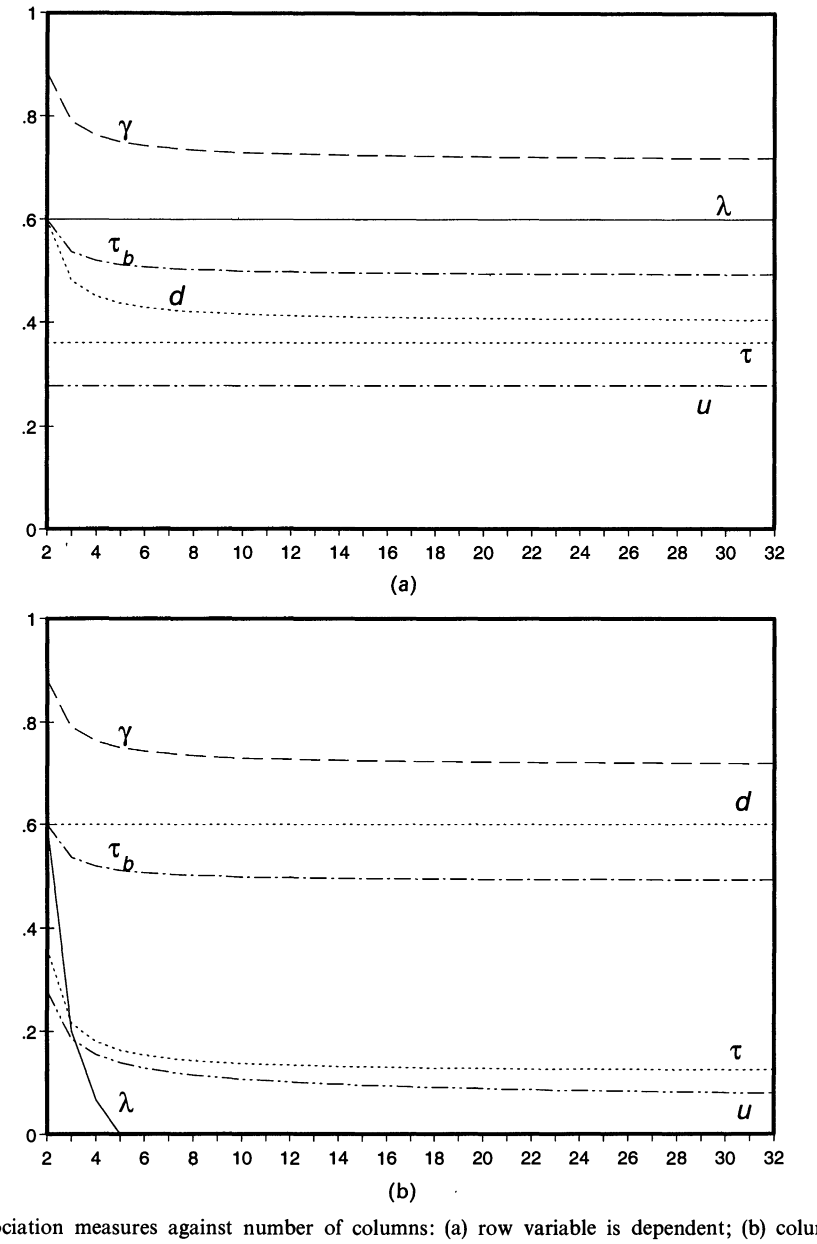

Fig. 7. Association measures against number of columns: (a) row variable is dependent; (b) column variable is dependent

However, they provide an instructive insight into the behaviour of the estimated association measures.

5. Conclusion

The results given in this paper provide insight into the scope and reliability of partial association measures. It has been shown that asymptotic variances can be used reliably for making inference on sample partial measures, even with relatively small samples. Another

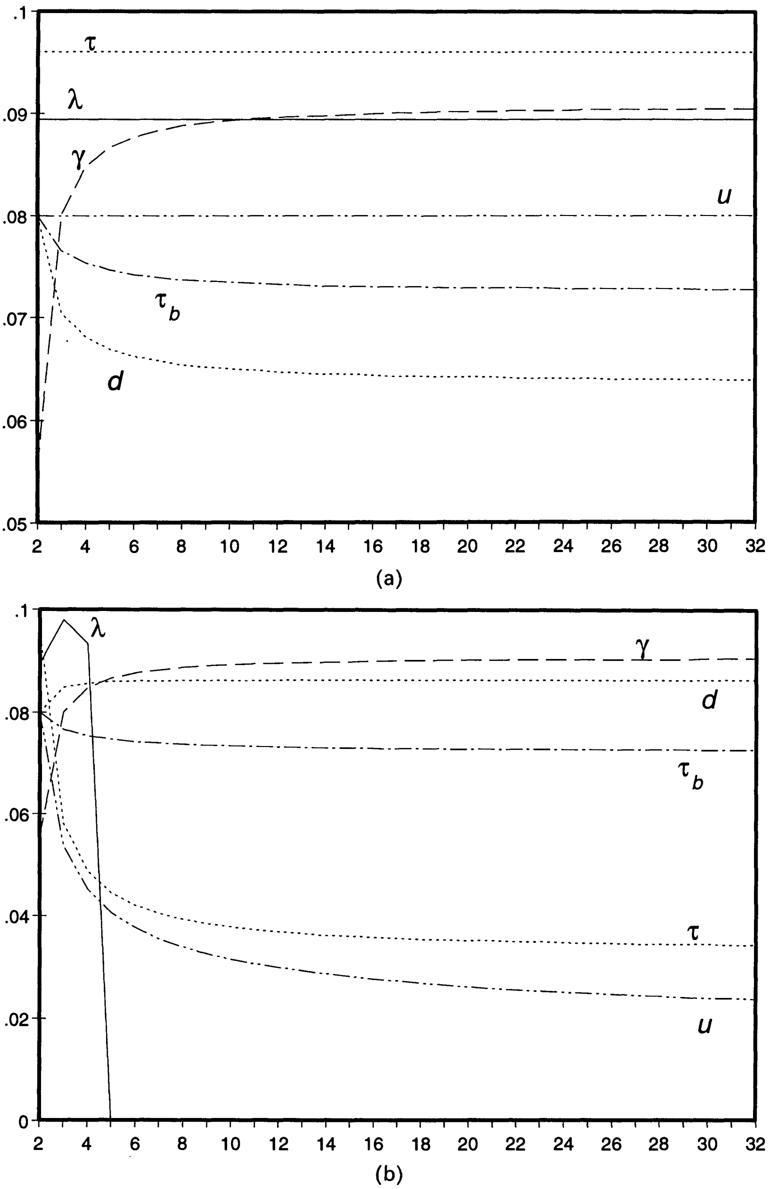

Fig. 8. Asymptotic standard errors against number of columns: (a) row variable is dependent; (b) column variable is dependent

important finding from the simulations carried out is the relatively high discrepancy of the estimates of indices based on the nominal λ and ordinal γ of Goodman and Kruskal. If we want to test the significance of the association, other measures should then be considered.

Partial measures, such as those analysed, synthesize the strength of the direct linkage between two variables which themselves depend on other common factors. It may then be tempting to use them, especially the asymmetrical measures, for evaluating the paths of causal models involving categorical variables (see for example Kellerhals *et al.* (1992)). One of the goals of path analysis, as stated by Wright (1934) and Blalock (1971), however, is to provide a decomposition of raw associations in terms of direct and indirect linkages. This implies some constraints between path coefficients and the raw association. Say, to simplify, that the sum of the direct and indirect linkages between two variables should equal their raw association. Ordinal partial measures comply broadly with such a constraint. Nominal partial measures do not. Indeed, they cannot take negative values, and thus the sum of the direct and indirect linkages may strongly exceed the raw association. Path coefficients measured with partial association indices should then be interpreted with special care.

Appendix A: asymptotic variances

We give here the asymptotic variances of the various raw and partial association measures considered. The formulae can be derived by using the δ-rule.

Broadly, according to the δ-rule, the asymptotic variance of the estimate of a ratio of the form $\theta = v(\ldots, p_{ij}, \ldots)/\delta(\ldots, p_{ij}, \ldots)$ derivable with respect to the p_{ij}s is

$$\sigma^2_\infty(\hat{\theta}) = \frac{1}{n\delta^4}\sum_{i\in I}\sum_{j\in J} p_{ij}(\phi_{ij} - \bar{\phi})^2 \tag{23}$$

with $\phi_{ij} = v(\partial\delta/\partial p_{ij}) - \delta(\partial v/\partial p_{ij})$ and $\bar{\phi} = \Sigma_i \Sigma_j p_{ij}\phi_{ij}$.

The raw and partial λs rely on maxima and are therefore not derivable with respect to the p_{ij}s or p_{ijk}s at points where probability maxima move from one row to another. The δ-rule then applies only outside these points.

A.1. *For estimates of raw nominal measures*

The asymptotic variance of the estimate $\hat{\lambda}$ is

$$\sigma^2_\infty(\hat{\lambda}_{AB}) = \frac{\left(1 - \sum_{j\in J} p_{mj}\right)\left(p_{m+} + \sum_{j\in J} p_{mj} - 2\sum_{j\in J_{m+}} p_{mj}\right)}{n(1 - p_{m+})^3} \tag{24}$$

where J_{m+} is the set of columns whose maximum is in the same row as the marginal maximum p_{m+}. See Goodman and Kruskal (1963).

For the estimate $\hat{\tau}$ of Goodman and Kruskal's τ, we have $\delta = 1 - \Sigma_i p^2_{i+}$ and

$$\phi_{ij} = 2p_{i+}\left(1 - \sum_{i'}\sum_{j'} \frac{p^2_{i'j'}}{p_{+j'}}\right) + \left(1 - \sum_{i'} p^2_{i'+}\right)\left\{2\frac{p_{ij}}{p_{+j}} - \sum_{i'\in I}\left(\frac{p_{i'j}}{p_{+j}}\right)^2\right\}. \tag{25}$$

The asymptotic variance of the estimate $\hat{u}$ of the uncertainty coefficient is given by the expression

$$\sigma^2_\infty(\hat{u}_{AB}) = \frac{\sum_i\sum_j p_{ij}\{H(A)\log_2(p_{ij}/p_{+j}) - H(A|B)\log_2 p_{i+}\}^2}{n\,H(A)^4} \tag{26}$$

where $H(A)$ denotes the entropy of A, i.e. $H(A) = -\Sigma_i p_{i+}\log_2 p_{i+}$, and $H(A|B)$ the entropy of A conditional to B, i.e.

$$H(A|B) = -\sum_j p_{+j}\sum_i (p_{ij}/p_{+j})\log_2(p_{ij}/p_{+j}).$$

See for instance Agresti (1986).

For the estimates $\hat{\gamma}$ and $\hat{d}_{AB}$ of the ordinal measures, we obtain

$$\sigma_\infty^2(\hat{\gamma}) = \frac{16}{n(\pi^c + \pi^d)^4} \sum_i \sum_j p_{ij}(\pi^c \pi_{ij}^d - \pi^d \pi_{ij}^c)^2, \tag{27}$$

$$\sigma_\infty^2(\hat{d}_{AB}) = \frac{4 \sum_i \sum_j p_{ij} \left\{ (\pi^c - \pi^d)(1 - p_{+j}) - \left(1 - \sum_{j'} p_{+j'}^2\right)(\pi_{ij}^c - \pi_{ij}^d) \right\}^2}{n\left(1 - \sum_j p_{+j}^2\right)^4} \tag{28}$$

where π_{ij}^c and π_{ij}^d denote the probabilities of obtaining an observation respectively concordant or discordant with those in cell (i, j), i.e.

$$\pi_{ij}^c = \sum_{i'<i} \sum_{j'<j} p_{i'j'} + \sum_{i'>i} \sum_{j'>j} p_{i'j'}, \tag{29}$$

$$\pi_{ij}^d = \sum_{i'<i} \sum_{j'>j} p_{i'j'} + \sum_{i'>i} \sum_{j'<j} p_{i'j'}. \tag{30}$$

For $\hat{\tau}_b$, the asymptotic variance is given by equation (23) with

$$\delta = \left(1 - \sum_i p_{i+}^2\right)^2 \left(1 - \sum_j p_{+j}^2\right)^2$$

and

$$\phi_{ij} = 2(\pi_{ij}^c - \pi_{ij}^d)\left\{\left(1 - \sum_{i'} p_{i'+}^2\right)\left(1 - \sum_{j'} p_{+j'}^2\right)\right\} + \tau_b \left\{p_{i+}\left(1 - \sum_{j'} p_{+j'}^2\right) + p_{+j}\left(1 - \sum_{i'} p_{i'+}^2\right)\right\}^{1/2}. \tag{31}$$

A.2. *For estimates of partial association*

For partial indices, the δ-rule provides asymptotic variances in the form

$$\sigma_\infty^2(\hat{\theta}_{AB|E}) = \frac{1}{n\delta^4} \sum_k \sum_j \sum_i p_{ijk}(\phi_{ijk} - \bar{\phi})^2 \tag{32}$$

where $\bar{\phi}$ is $\Sigma_k \Sigma_j \Sigma_i p_{ijk} \phi_{ijk}$.

The asymptotic variance of $\hat{\lambda}_{AB|E}$ is given by (see also Goodman and Kruskal (1963), p. 333)

$$\sigma_\infty^2(\hat{\lambda}_{AB|E}) = \frac{\left(1 - \sum_k \sum_j p_{mjk}\right)\left(\sum_k \sum_j p_{mjk} + \sum_k p_{m+k} - 2 \sum_k \sum_{j \in J_{m+k}} p_{mjk}\right)}{n\left(1 - \sum_k p_{m+k}\right)^3} \tag{33}$$

For $\hat{\tau}_{AB|E}$, we have

$$\delta = 1 - \sum_k \sum_i \frac{p_{i+k}^2}{p_{++k}}, \tag{34}$$

$$\phi_{ijk} = \left(-\frac{2p_{i+k}}{p_{++k}} + \frac{1}{p_{++k}^2} \sum_{i'} p_{i'+k}^2\right)\left(1 - \sum_{k'} \sum_{j'} \sum_{i'} \frac{p_{i'j'k'}^2}{p_{+j'k'}}\right) - \delta\left(\frac{1}{p_{+jk}^2} \sum_{i'} p_{i'jk}^2 - \frac{2p_{ijk}}{p_{+jk}}\right). \tag{35}$$

For $\hat{u}_{AB|E}$, we have

$$\delta = \sum_k \sum_i p_{i+k} \log_2\left(\frac{p_{i+k}}{p_{++k}}\right), \tag{36}$$

$$\phi_{ijk} = \log_2\left(\frac{p_{i+k}}{p_{++k}}\right) \sum_{k'} \sum_{i'} \sum_{j'} p_{i'j'k'} \log_2\left(\frac{p_{i'+k'}p_{+j'k'}}{p_{i'j'k'}p_{++k'}}\right) - \delta \log_2\left(\frac{p_{i+k}p_{+jk}}{p_{ijk}p_{++k}}\right). \tag{37}$$

For $\hat{\tau}^{\mathrm{b}}_{AB|E}$, we have

$$\delta = \sum_k \left\{\left(p^2_{++k} - \sum_i p^2_{i+k}\right)\left(p^2_{++k} - \sum_j p^2_{+ik}\right)\right\}^{1/2}, \tag{38}$$

$$\phi_{ijk} = -2\delta(\pi^{\mathrm{c}}_{ij,k} - \pi^{\mathrm{d}}_{ij,k}) + \left\{\sum_{k'} (\pi^{\mathrm{c}}_{k'} - \pi^{\mathrm{d}}_{k'})\right\}$$
$$\times \left\{\frac{\left(p^2_{++k} - \sum_{j'} p^2_{i'+k}\right)(p_{++k} - p_{+jk}) + \left(p^2_{++k} - \sum_{j'} p^2_{+j'k}\right)(p_{++k} - p_{i+k})}{\left(p^2_{++k} - \sum_{i'} p^2_{i'+k}\right)^{1/2}\left(p^2_{++k} - \sum_{j'} p^2_{+j'k}\right)^{1/2}}\right\} \tag{39}$$

where $\pi^{\mathrm{c}}_{ij,k}$ and $\pi^{\mathrm{d}}_{ij,k}$ denote the probabilities of obtaining a case which falls in category k of E and is respectively concordant or discordant according to the first two variables A and B with those in cell (i, j, k), i.e.

$$\pi^{\mathrm{c}}_{ij,k} = \sum_{i'<i} \sum_{j'<j} p_{i'j'k} + \sum_{i'>i} \sum_{j'>j} p_{i'j'k}, \tag{40}$$

$$\pi^{\mathrm{d}}_{ij,k} = \sum_{i'<i} \sum_{j'>j} p_{i'j'k} + \sum_{i'>i} \sum_{j'<j} p_{i'j'k}. \tag{41}$$

For the ordinal indices $\hat{\gamma}_{AB|E}$ and $\hat{d}_{AB|E}$ the expressions are a little simpler. The following is the final form of their asymptotic variances in terms of the probabilities π^{c}_k of a concordant pair and π^{d}_k of a discordant pair among those tied on state k of E:

$$\sigma^2_\infty(\hat{\gamma}_{AB|E}) = \frac{16}{n\left\{\sum_k (\pi^{\mathrm{c}}_k + \pi^{\mathrm{d}}_k)\right\}^4} \sum_k \sum_j \sum_i p_{ijk}\left(\pi^{\mathrm{d}}_{ij,k} \sum_{k'} \pi^{\mathrm{c}}_{k'} - \pi^{\mathrm{c}}_{ij,k} \sum_{k'} \pi^{\mathrm{d}}_{k'}\right)^2, \tag{42}$$

$$\sigma^2_\infty(\hat{d}_{AB|E}) = \frac{4}{n}\left\{\sum_k (\pi^{\mathrm{c}}_k + \pi^{\mathrm{d}}_k + \pi^{\mathrm{t}}_{Ak})\right\}^{-4} \sum_k \sum_j \sum_i p_{ijk}$$
$$\times \left\{(p_{++k} - p_{+jk}) \sum_{k'} (\pi^{\mathrm{c}}_{k'} - \pi^{\mathrm{d}}_{k'}) - (\pi^{\mathrm{c}}_{ij,k} - \pi^{\mathrm{d}}_{ij,k}) \sum_k (\pi^{\mathrm{c}}_{k'} + \pi^{\mathrm{d}}_{k'} + \pi^{\mathrm{t}}_{Ak'})\right\}^2. \tag{43}$$

References

Agresti, A. (1986) Applying R^2-type measures to ordered categorical data. *Technometrics*, **28**, 133–138.

Blalock, H. M. (1971) *Causal Models in the Social Sciences*. Chicago: Aldine-Atherton.

Cramer, H. (1946) *Mathematical Methods for Statistics*. Princeton: Princeton University Press.

Davis, J. A. (1967) A partial coefficient for Goodman and Kruskal's gamma. *J. Am. Statist. Ass.*, **62**, 189–193.

Goodman, L. A. and Kruskal, W. H. (1954) Measures of association for cross classifications. *J. Am. Statist. Ass.*, **37**, 54–115.

———(1963) Measures of association for cross classifications: III, Approximate sampling theory. *J. Am. Statist. Ass.*, **58**, 310–364.

———(1972) Measures of association for cross classifications: IV, Simplification of asymptotic variances. *J. Am. Statist. Ass.*, **67**, 415–421.

Kellerhals, J., Montandon, C. and Ritschard, G. (1992) Social status, types of family interaction and educational styles. *Eur. J. Sociol.*, **23**, 308–325.

Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

———(1945) The treatment of ties in rank problems. *Biometrika*, **33**, 239–251.

Pearson, K. (1948) Mathematical contributions to the theory of evolution: XIII, On the theory of contingency and its relation to association and normal correlation. In *Karl Pearson's Early Papers* (ed. E. S. Pearson). Cambridge: Cambridge University Press.

Quade, D. (1974) Nonparametric partial correlation. In *Measurement in the Social Sciences* (ed. H. M. Blalock, Jr), pp. 369–397. Chicago: Aldine-Atherton.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423, 623–656.

Somers, R. H. (1962) A new asymmetric measure for ordinal variables. *Am. Sociol. Rev.*, **27**, 799–811.

Stuart, A. (1953) The estimation and comparison of strengths of association in contingency tables. *Biometrika*, **40**, 105–110.

Theil, H. (1970) On the estimation of relationships involving qualitative variables. *Am. J. Sociol.*, **76**, 103–154.

Tschuprow, A. A. (1918) On the mathematical expectation of moments of frequency distribution. *Biometrika*, **12**, 140–169.

Wilson, T. P. (1974) Measures of association for bivariate ordinal hypotheses. In *Measurement in the Social Sciences* (ed. H. M. Blalock, Jr), pp. 327–342. Chicago: Aldine-Atherton.

Wright, S. (1934) The method of path coefficients. *Ann. Math. Statist.*, **5**, 161–215.