Simple Measures of Association for the Triple Dichotomy[†]

By Robert H. Somers

The University of Michigan, Ann Arbor, Mich.

Much useful work has been done in recent years in providing ways of measuring association between two attributes where the bivariate distribution is presented in the form of a contingency table. This paper is not intended to introduce any new measures, but instead to apply one of the simplest—the "percentage difference"—to one of the simplest cases that arises when one is faced with the description of relations among not two, but three attributes. The case dealt with here is that of the triple dichotomy.

As an introduction, consider what is meant by the "percentage difference": for a fourfold table such as that shown in Table 3, the percentage difference will be defined as the difference in conditional probabilities.

$$d_{yx} = \Pr(Y|X) - \Pr(Y|\bar{X})$$

$$= \frac{a_.}{A_.} - \frac{b_.}{B_.} \tag{1a}$$

$$= \frac{a_. d_. - b_. c_.}{A_. B_.}. \tag{1b}$$

It should be noted that this measure is not symmetric, meaning that, in general, $d_{yx} \neq d_{xy} = (a_. d_. - b_. c_.)/C_. D_.$, and thus the measure does not satisfy the axioms suggested by Edwards (1963). It is used, nevertheless, on the assumption that some situations may warrant an asymmetric measure, such as the conditions which led Goodman and Kruskal (1954) to propose measures having different interpretations but the same asymmetric characteristics. In addition, the percentage difference has an operational interpretation which may be generalized to ordered contingency tables (see Somers, 1962) which will not be emphasized here. Rather, the coefficient is used here because of its simplicity and widespread use as a measure of "departure from statistical independence", and because it is formally identical to the coefficient of regression of Y on X when 0, 1 scores are assigned to the absence, presence of the attributes, respectively. It is of further interest that $|d_{yx}| \leqslant 1$.

Table 1

First conditional distribution

For Z:

	X	$\bar{X}$	*Totals*
Y	a_1	b_1	C_1
$\bar{Y}$	c_1	d_1	D_1
Totals	A_1	B_1	N_1

[†] Support for this paper derived in part from National Science Foundation Grant NSF-G-17438 administered by the Bureau of Applied Social Research at Columbia University.

TABLE 2

Second conditional distribution

For $\bar{Z}$:

	X	$\bar{X}$	*Totals*
Y	a_2	b_2	C_2
$\bar{Y}$	c_2	d_2	D_2
Totals	A_2	B_2	N_2

TABLE 3

Unconditional distribution

For Z and $\bar{Z}$ combined:

	X	$\bar{X}$	
Y	$a_.$	$b_.$	$C_.$
$\bar{Y}$	$c_.$	$d_.$	$D_.$
	$A_.$	$B_.$	$N_.$

where the "." means summation, e.g. $N_. = N_1+N_2$.

In a similar fashion we may define the conditional percentage differences, for the distributions conditional on the presence or absence of Z, in Tables 1 and 2:

$$d_{yx|z} = \frac{a_1 d_1 - b_1 c_1}{A_1 B_1} \tag{2}$$

$$d_{yx|\bar{z}} = \frac{a_2 d_2 - b_2 c_2}{A_2 B_2}. \tag{3}$$

Following the Yule formula (Kendall and Stuart, 1961, p. 543), for a triple dichotomy, the "excess" of d_{yx} over the "spurious" portion coming from the side relations of X to Z and of Y to Z may be expressed as

$$d_{yx} - d_{yz} d_{zx} = \left(\frac{w_{x|z}}{w_x}\right) d_{yx|z} + \left(\frac{w_{x|\bar{z}}}{w_x}\right) d_{yx|\bar{z}} \tag{4}$$

where

$$w_{x|z} = \frac{A_1 B_1}{N_1}, \quad w_{x|\bar{z}} = \frac{A_2 B_2}{N_2}, \quad \text{and} \quad w_x = \frac{A_. B_.}{N_.}.$$

In contrast to the Yule formula, equation (4) requires equal weights to be applied to the terms d_{yx} and $(d_{yz} d_{zx})$, which leads to the particularly simple result that when there is no "partial association" then

$$d_{yx} = d_{yx} d_{zx}. \tag{5}$$

Since the d's are asymmetric, the denominators have to be chosen appropriately, i.e. the order of the subscripts is important.

The weights assigned by this formula to the coefficients arising from the conditional tables are made more complex than in Yule's formula, however, and it is of interest to try and give them an interpretation. In particular, we may ask, under what conditions will the weights assigned to the conditional coefficients in the right-hand side of equation (4) have the same ratio as the total number of observations in the two conditional distributions, i.e. what condition is required so that

$$\frac{w_{x|z}}{w_{x|\bar{z}}} = \frac{N_1}{N_2}. \tag{6}$$

It can readily be seen that a necessary and sufficient condition is that

$$A_1 B_1/N_1^2 = A_2 B_2/N_2^2;$$

the required condition is that the conditional variances be equal. Thus certain minimal structural conditions must be satisfied in order that $d_{yx} - d_{yz} d_{zx}$ be a "natural" average of the conditional percentage differences. A later remark will reconsider this average of the conditional percentage differences.

In work now in progress, this analysis of a triple dichotomy has been generalized to the analysis of three-way contingency tables where the categories are assumed to have a natural *ordering*, and the generalization of d_{yx} is an asymmetric coefficient of rank correlation (for ties) related to Kendall's τ_b.

In such a situation, it becomes relevant to partition $d_{yx} - d_{yz} d_{zx}$ into two parts:

$$u'_{yx.z} = N_2^2(a_1 d_1 - b_1 c_1) + N_1^2(a_2 d_2 - b_2 c_2) \tag{7a}$$

and

$$u''_{yx.z} = N_1 N_2\{(a_1 d_1 - b_1 c_1) + (a_2 d_2 - b_2 c_2)\} \tag{7b}$$

which sum to

$$u_{yx.z} = A.B.N_1 N_2(d_{yx} - d_{yz} d_{zx}).$$

I shall consider further partitioning shortly, but first we may consider what would be a reasonable coefficient of "partial association" derived from these percentage differences. One possibility would be to use the difference, $d_{yx} - d_{yz} d_{zx}$ itself. In other circumstances it may be useful to consider a norm for it. In particular, suppose one wishes to quantify the partial association in a table where one intuitively reasonable form of "maximum partial association" occurs, illustrated by the distribution in Table 4. In this table, the conditional marginal distributions of X within Z have been retained from Tables 1 and 2, but the off-diagonal cells within the conditional fourfold tables have been decreased to zero. If one had observed a distribution of this form, one would have observed the maximum possible partial association, given the observed "side" association between X and Z. Hence it is reasonable to standardize the value of $d_{yx} - d_{yz} d_{zx}$, and compute the ratio

$$\frac{d_{yx} - d_{yz} d_{zx}}{\max\{(d_{yx} - d_{yz} d_{zx}) \mid X, Z \text{ relation}\}}$$

when one wishes to see how large the partial association is relative to the maximum that could be achieved under this condition. It is easy to see that, from Table 4,

$$\max\{(d_{yx} - d_{yz} d_{zx}) \mid X, Z \text{ relation}\} = 1 - d_{xz} d_{zx}.$$

Hence it seems appropriate to suggest the symbol $d_{yx.z}$ for the ratio

$$d_{yx.z} = \frac{d_{yx} - d_{yz} d_{zx}}{1 - d_{xz} d_{zx}}, \tag{8}$$

since this is, of course, the product-moment formula for computation of the first-order partial regression coefficient from the total regressions.

The application of these measures to one of the "tortuous" hypothetical tabulations given by Edwards is illustrated in Table 5. No particular insight into the character of the distribution is claimed, except to note that in ordinary product–moment analysis it is not unusual for the total coefficient to differ in sign from the partial. Here the marginal relations, d_{yz} and d_{zx}, are of appropriate sign and sufficient magnitude to yield a product (+0·0450) which offsets the negativeness of the conditionals.

TABLE 4

Maximum possible partial association, provided the relation of X to Z is unchanged

For Z:

	X	$\bar{X}$	*Totals*
Y	A_1	0	A_1
$\bar{Y}$	0	B_2	B_2
Totals	A_1	B_2	N_1

For $\bar{Z}$:

	X	$\bar{X}$	*Totals*
Y	A_2	0	A_2
$\bar{Y}$	0	B_2	B_2
Totals	A_2	B_2	N_2

TABLE 5

Application to Edwards's tabulation

Z

	X	$\bar{X}$
Y	3	5
$\bar{y}$	4	6

$\bar{Z}$

	X	$\bar{X}$
Y	9	6
$\bar{y}$	5	3

$$d_{yx|z} = \frac{-2}{(7)(11)} \qquad d_{yx|\bar{z}} = \frac{-3}{(14)(9)}$$

$$\left(\frac{w_{x|z}}{w_x}\right) d_{yx|z} = \frac{(7)(11)(41)}{(21)(20)(18)}\left[\frac{-2}{(7)(11)}\right] = (0{\cdot}4176)(-0{\cdot}0260) \quad = -0{\cdot}011$$

$$\left(\frac{w_{x|\bar{z}}}{w_x}\right) d_{yx|\bar{z}} = \frac{(14)(9)(41)}{(21)(20)(23)}\left[\frac{-3}{(14)(9)}\right] = (0{\cdot}5348)(-0{\cdot}0238) \quad = -0{\cdot}013$$

$$d_{yz}\,d_{zx} = \left[\frac{-86}{(18)(23)}\right]\left[\frac{-91}{(21)(20)}\right] = (-0{\cdot}2077)(-0{\cdot}2167) \quad = +0{\cdot}045$$

$$d_{yx} = \frac{9}{(21)(20)} = 0{\cdot}021 \quad = +0{\cdot}021 \text{ (sum)}$$

$$d_{yx} - d_{yz}\,d_{zx} = -0{\cdot}024$$

$$d_{yx.z} = \frac{d_{yx} - d_{yz}\,d_{zx}}{1 - d_{zx}\,d_{xz}} = \frac{-0{\cdot}0236}{0{\cdot}9524} = -0{\cdot}025$$

It was noted above that, when there are more than two categories on each attribute and these categories have a natural ordering for each attribute, it is useful to partition $d_{yx} - d_{yz}\,d_{zx}$. Interpretation of the coefficient $d_{yx.z}$ introduced above also suggests a partitioning of the denominator, $1 - d_{xz}\,d_{zx}$, into

$$u'_{x.z} = 4A_1 B_1 A_2 B_2 + (A_1 B_2 + A_2 B_1)(A_1 A_2 + B_1 B_2) \tag{9a}$$

and

$$u''_{x.z} = N_1 N_2 (A_1 B_1 + A_2 B_2), \tag{9b}$$

which, in turn, add to

$$u_{x.z} = A_{.} B_{.} N_1 N_2 (1 - d_{xz} d_{zx}).$$

By constructing two separate measures of partial correlation, from the definitions presented in expressions (7) and (9),

$$d'_{yx.z} = \frac{u'_{yx.z}}{u'_{x.z}}, \tag{10a}$$

and

$$d''_{yx.z} = \frac{u''_{yx.z}}{u''_{x.z}}, \tag{10b}$$

it can be shown that the latter is exactly equal to the weighted average of the percentage differences in the conditional distributions:

$$d''_{yx.z} = \frac{1}{v_{.}}(v_1 d_{yx|z} + v_2 d_{yx|\bar{z}}),$$

where

$$v_i = A_i B_i \quad \text{and} \quad v_{.} = v_1 + v_2.$$

The measure designated here as $d_{yx.z}$ is, in turn, the weighted average of $d'_{yx.z}$ and $d''_{yx.z}$, where the denominators, taken relative to their sum, are again used as weights. In the numerical example of Table 5, $d'_{yx.z} = -0{\cdot}0249$, and $d''_{yx.z} = -0{\cdot}0246$; they are very close, although not identical, in value. In contingency tables which cross-classify attributes having more than two categories, where those categories have a natural ordering, and d_{yx} is generalized to a coefficient which measures "monotonic correlation" (Burr, 1960), then the same partitioning into $d'_{yx.z}$ and $d''_{yx.z}$ may be accomplished, with an important difference: while $d''_{yx.z}$ remains a weighted average of the conditional coefficients, the other portion, $d'_{yx.z}$ can no longer be expressed as a function of "within-conditional" distributions, as in expression (7a) above, but instead must be interpreted as arising from "between" the conditional distributions.

Considerable work has been done recently on the analysis of interaction in contingency tables, especially the simplest case of a triple dichotomy. Much of this work stems from the introduction by Bartlett (1935) and is summarized in Lewis (1962). As the fifth edition of Snedecor (1956) notes, the analogy between the double dichotomy and triple dichotomy, upon which Bartlett apparently based his criterion of interaction, which we may call (working with frequencies instead of proportions)

$$I_B = a_1 d_1 b_2 c_2 - a_2 d_2 b_1 c_1,$$

is not a good analogy, and consequently "it must be rare . . . that in a three-way table one would be interested in the hypothesis" that $I_B = 0$ (Snedecor, 1956, p. 231).

In contrast to this view of the triple dichotomy, and the analysis of variance analogies developed in a recent article by Darroch (1962), the present view derives from analogies with the case of multiple rankings, where the rank ordering of the observations is perhaps the most crucial information. When ties are present among all of the rankings, it has been found easier to develop and interpret asymmetric, rather than symmetric, measures of rank correlation and partial rank correlation. This leads to a different view of the triple dichotomy from Bartlett's criterion, which is symmetric in the variables.

From this point of view, it seems natural to consider the triple dichotomy as representing an asymmetric situation in which analysis of variance analogies may be formulated by taking the proportion "positive" on the dependent variable, say, Y, within the levels of the independent variables, X and Z here, as means for those cells. This makes the triple dichotomy a special case of an unbalanced design (i.e. having unequal numbers of observations in each cell) where the interaction is that of first-order interaction between X and Z as they affect Y rather than of second-order interaction between X, Y and Z (a confusion, it seems to me, of "n's and means").

With this view, one may apply the formulas for interaction in an unbalanced design (e.g. Winer, 1962) to the proportions (taken as means) in each of the cells of the fourfold table (reduced from the triple dichotomy) shown in Table 6. The formulas show that, in the notation of Winer (1962, p. 292)

$$SS_{xz(adj)} = (d_{yx|z} - d_{yx|\bar{z}})^2 \left\{ \frac{w_{x|z}\, w_{x|\bar{z}}}{w_{x|z} + w_{x|\bar{z}}} \right\}, \tag{11}$$

the w's being those used in expression (4). It is, of course, not generally true that $I_B = 0$ implies $SS_{xz(adj)} = 0$, for in the triple dichotomy it may be shown that

$$I_B = t_{12}\, d_{yx|z} - t_{21}\, d_{yx|\bar{z}}, \tag{12}$$

where

$$t_{ij} = \tfrac{1}{2} A_i\, B_i (a_j\, d_j + b_j\, c_j).$$

TABLE 6

Conditional probabilities of Y, given X and Z level, from the notation of Tables 1–3

	Z	$\bar{Z}$	*Averages*
X	$\frac{a_1}{A_1}$	$\frac{a_2}{A_2}$	$\frac{a_.}{A_.}$
$\bar{X}$	$\frac{b_1}{B_1}$	$\frac{b_2}{B_2}$	$\frac{b_.}{B_.}$
Averages	$\frac{C_1}{N_1}$	$\frac{C_2}{N_2}$	$\frac{C_.}{N_.}$

ACKNOWLEDGEMENTS

No student of Paul Lazarsfeld can come away from Columbia University without a certain reverence for the mysteries of "the dichotomous cube". Some of Lazarsfeld's recent work, not cited by Lewis, is added to the references below. The referee of this paper has also been helpful in leading me to relevant literature; neither, of course, is responsible for any errors I may have committed.

REFERENCES

BARTLETT, M. S. (1935), "Contingency table interactions", *J. R. statist. Soc., Suppl.*, **2**, 248–252.

BURR, E. J. (1960), "The distribution of Kendall's score S for a pair of tied rankings", *Biometrika*, **47**, 1 and 2, 151.

DARROCH, J. N. (1962), "Interactions in multifactor contingency tables", *J. R. statist. Soc.* B, **24**, 251–263.

EDWARDS, A. W. F. (1963), "The measure of association in a 2×2 table", *J. R. statist. Soc.* A, **126**, 109–114.

GOODMAN, L. A. and KRUSKAL, W. H. (1954), "Measures of association for cross classifications", *J. Amer. statist. Ass.*, **49**, 732–764.

KENDALL, M. G. and STUART, ALAN (1961), *The Advanced Theory of Statistics*, Vol. 2. New York: Hafner.

LAZARSFELD, P. F. (1961), "The algebra of dichotomous systems", in H. Solomon, ed., *Studies in Item Analysis and Prediction*, pp. 111–157. Stanford: University Press.

LEWIS, B. N. (1962), "On the analysis of interaction in multidimensional contingency tables", *J. R. statist. Soc.* A, **125**, 88–117.

SNEDECOR, G. W. (1956), *Statistical Methods.* Ames: Iowa State.

SOMERS, R. H. (1962), "A new asymmetric measure of association for ordinal variables", *Amer. Sociol. Rev.*, **27**, 799–811.

WINER, B. J. (1962), *Statistical Principles in Experimental Design.* New York: McGraw-Hill.