

NONPARAMETRIC PARTIAL CORRELATION

by

Dana Quade

University of North Carolina

Institute of Statistics Mimeo Series No. 526

May 1967

Supported by National Institutes of Health  
Grant No. GM-12868-03

DEPARTMENT OF BIostatISTICS  
UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.

# NONPARAMETRIC PARTIAL CORRELATION<sup>1</sup>

Dana Quade

University of North Carolina

Let there be given a random sample of observations on  $(X, Y, Z)$ . Assume that any pair of observations may be classified as matched or not with respect to  $Z$ , and as concordant, discordant, or tied with respect to  $X$  and  $Y$ . Let  $N_M$  be the total number of matched pairs, of which  $N_C$ ,  $N_D$ , and  $N_T$  are the numbers concordant, discordant, and tied. Then an intuitively reasonable measure of partial correlation is the "index of matched correlation,"  $t^* = (N_C - N_D)/N_M$ . The properties of this index, including its interpretation and its asymptotic sampling distribution, are discussed.

## 1. INTRODUCTION

It is often considered desirable to measure the correlation between two variables, say  $X$  and  $Y$ , when a third variable, say  $Z$ , has been controlled, or its effects have been removed. Such a measure is generally termed a "partial correlation" between  $X$  and  $Y$  given  $Z$ . (In what follows we devote some attention to more exact definition).

The most customary measure is of course the Pearsonian product-moment partial correlation, usually defined as the product-moment correlation between the residuals remaining after performing an ordinary linear least-squares regression of  $X$  on  $Z$  and the corresponding residuals after regressing  $Y$  on  $Z$ . Such a procedure makes no sense unless  $X$ ,  $Y$ , and  $Z$  are all metric variables (though  $Z$  may be multivariate). Its interpretation is difficult in any case, and particularly so if the true regression curves of  $X$  and  $Y$  on  $Z$  are non-linear; and

---

<sup>1</sup> Supported by National Institutes of Health Grant No. GM-12868-03.

assumptions of normality are required in order to proceed beyond mere description to statistical inference.

Another measure of partial correlation is that given by Kendall [8] and based on his measure of rank correlation ( $\tau$ ). This measure permits use of ordinal variables X, Y, and Z (though Z here must be univariate), but since it does not offer more ease of interpretation than the product-moment partial correlation, and since procedures for statistical inference based on it are not available, it appears to have been little used. An interesting but neglected variant, due to Goodman [5], does lead to inference.

When the variables are categorical, the data being expressed as a contingency table, it has long been customary to calculate "partial associations", namely measures of association between X and Y in the subsamples defined by fixing on a single category of Z. In a recent paper, Davis [3] has suggested an overall measure of partial association based on the index G proposed by Goodman and Kruskal [6,7], which is also closely related to Kendall's  $\tau$ . Although Davis did not develop procedures for statistical inference using his measure, this can be done by the methods of the present paper, as we point out at the end of Section 4.

It is our purpose here to develop a general "index of matched correlation" which may be briefly defined, for a population, as "the difference between the conditional probabilities of concordance and discordance of a random pair of observations, given that the pair is matched on Z"; in a sample the index is obtained by substituting in estimates for the required probabilities. Three important properties of this index are:

(i) Ease of interpretation: The proposed index is based on the two simple concepts (defined in Section 2) involved in determining whether a pair of observations chosen at random is matched or not with respect to the controlled

variable  $Z$ , and whether such a pair is concordant or discordant with respect to  $X$  and  $Y$ .

(ii) Wide applicability: The proposed index may be used to control for a completely arbitrary variable, including even a multivariate  $Z$  in which each component separately may be metric, ordinal, or purely nominal; it is necessary only to have a definition of "matching". And the variables  $X$  and  $Y$  need not be more than ordinal or ordered-categorical, although of course they can not be purely nominal since the very word "correlation" implies an ordered (and directed) measure of association.

(iii) Usability for statistical inference: It will be shown, at least for large samples, that the sampling distribution of the proposed index is normal, and how its standard error can be calculated.

Section 2 is devoted to various preliminary questions, mainly of definition; in Section 3 we present our proposed index in some detail; Sections 4 and 5 are devoted to examples and discussion; and a general Theorem, giving the distribution of the ratio of two U-statistics, is relegated to an Appendix.

## 2. CONDITIONAL AND PARTIAL CORRELATION

Any pair of observations  $(X_1, Y_1, Z_1)$  and  $(X_2, Y_2, Z_2)$ , where  $X$  and  $Y$  are at least ordinal or ordered-categorical, may be classified as

concordant (with respect to  $X$  and  $Y$ ) if  $X_1 < X_2, Y_1 < Y_2$  or  $X_1 > X_2, Y_1 > Y_2$ ,

discordant (with respect to  $X$  and  $Y$ ) if  $X_1 < X_2, Y_1 > Y_2$  or  $X_1 > X_2, Y_1 < Y_2$ ,

or tied (with respect to  $X$  and  $Y$ ) if  $X_1 = X_2$  or  $Y_1 = Y_2$  or both;

and let the probabilities of these three events be  $p_C, p_D, p_T$  respectively.

Then as an index of total correlation (or "unconditional," or "marginal") between  $X$  and  $Y$  we may take

$$\tau_{XY} = p_C - p_D,$$

the difference between the probabilities of concordance and discordance.

(When X and Y are clear from the context we may omit indicating them, thus writing  $\tau$  rather than  $\tau_{XY}$ ). It is immediately verifiable that this index always lies in the range  $-1 \leq \tau \leq 1$ , taking the value +1 if  $p_C = 1$  and -1 if  $p_D = 1$ ; and  $\tau = 0$  if  $p_C = p_D$ , for which a sufficient, though not necessary, condition is that X and Y be (marginally) independent.

The index  $\tau_{XY}$  or  $\tau$  is, of course, Kendall's  $\tau_a$  [8]; a more common variant, generally called Kendall's tau, is his rank correlation coefficient

$$\tau_b = \frac{p_C - p_D}{\sqrt{1-p_X} \sqrt{1-p_Y}}$$

where  $p_X$  (or  $p_Y$ ) is the probability that a randomly-chosen pair of observations will have the same value of X (or Y). Another version in common use, particularly for categorical data, is the Goodman-Kruskal index [6]

$$\gamma = \frac{p_C - p_D}{p_C + p_D}.$$

If we reexpress the first index as

$$\tau_a = \frac{p_C - p_D}{p_C + p_D + p_T},$$

Then it becomes clear that the only point of difference among them is the treatment of tied pairs; if  $p_T = 0$  then the three indices coincide. Further discussion on this is given in Section 5; we note in passing, however, that

$$0 \leq |\tau_a| \leq |\tau_b| \leq |\gamma| \leq 1.$$

Let us now define an index of conditional correlation between X and Y given  $Z = z$ , namely

$$\tau_{XY|Z}(z) = P_{C|Z=z} - P_{D|Z=z}$$

where  $P_{C|Z=z}$  (or  $P_{D|Z=z}$ ) is the probability that a randomly-chosen pair of observations  $(X_1, Y_1, Z_1)$  and  $(X_2, Y_2, Z_2)$  will be concordant (or discordant), conditional upon the event that  $Z_1$  and  $Z_2$  are both fixed at the same point  $z$ . To express this in more mathematical notation, let

$$W((x_1, y_1), (x_2, y_2)) = \begin{cases} 1 & \text{if } x_1 < x_2, y_1 < y_2 \text{ or } x_1 > x_2, y_1 > y_2 \\ 0 & \text{if } x_1 = x_2 \text{ or } y_1 = y_2 \text{ or both} \\ -1 & \text{if } x_1 < x_2, y_1 > y_2 \text{ or } x_1 > x_2, y_1 < y_2; \end{cases}$$

then if the marginal joint distribution of X and Y is  $F(x, y)$  we have

$$\tau_{XY} = \iint W((x_1, y_1), (x_2, y_2)) dF(x_1, y_1) dF(x_2, y_2),$$

and if the conditional distribution of X and Y given  $Z = z$  is  $C(x, y|z)$  then

$$\tau_{XY|Z}(z) = \iint W((x_1, y_1), (x_2, y_2)) dC(x_1, y_1|z) dC(x_2, y_2|z).$$

In this last definition the nature of the variable Z is completely unspecified; Z may be categorical and/or multivariate, and in fact need not be random.

Finally, let us define an index of partial correlation between X and Y given Z, written  $\tau_{XY|Z}$ , as a weighted average of the indices of conditional correlation  $\tau_{XY|Z}(z)$  over the values  $z$  of Z. Since we have been considering pairs of observations, we weight the conditional correlation at  $z$  in proportion to the probability that a randomly-chosen pair of observations  $(X_1, Y_1, Z_1)$

and  $(X_2, Y_2, Z_2)$  will be such that  $Z_1 = Z_2 = z$ . Thus if  $Z$  is a discrete random variable, with marginal probability function  $h(z)$ , the probability of observing  $Z_1 = Z_2 = z$  is  $h^2(z)$ , and we define

$$\tau_{XY|Z} = \frac{\sum_z h^2(z) \tau_{XY|Z}(z)}{\sum_z h^2(z)} .$$

On the other hand, if  $Z$  is a continuous random variable, with marginal density function  $h(z)$ , we may define

$$\tau_{XY|Z} = \frac{\int_z h^2(z) \tau_{XY|Z}(z) dz}{\int_z h^2(z) dz} .$$

In both cases  $Z$  is allowed to be multivariate. (We shall not attempt to define an index of partial correlation for the more difficult case in which  $Z$  is neither purely continuous nor purely discontinuous.)

The terminology adopted here is admittedly not standard; but indeed no standard terminology exists. Confusion has arisen from restricting study, as is commonly done, to the multivariate normal distribution, in which the correlations between  $X$  and  $Y$  conditioned on  $Z=z$  are the same for every  $z$ , and hence also the same as any weighted average of them; i.e., in our terminology the conditional correlations and the partial correlation are all equal. (This is usually stated in terms of product-moment correlations, of course, but it holds equally true for the indices of correlation under discussion here.) Such a state of affairs cannot be expected in general, however, and hence our concepts must be defined more precisely. For example, Kendall and Stuart [9, p. 317]

state that "... the correlations between variables when other variables are held constant, i.e., conditionally upon those other variables taking certain fixed values .... are the so-called partial correlations." (In our terminology they would be conditional correlations.) But on the following page, after showing that in the multivariate normal distribution, these (product-moment) correlations obey the familiar "partial correlation formula"

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2} \sqrt{1-\rho_{YZ}^2}} ,$$

Kendall and Stuart use this relationship to define partial correlation for any parent distribution. The more usual definition of partial correlation (see, for example, Cramér [1], Blalock [2], and Yule [13]) is that it is the correlation between the two sets of residuals remaining after separate linear regressions of X on Z and Y on Z; this is equivalent to definition in terms of the formula, but perhaps easier to interpret. (Kendall's partial rank correlation coefficient can be defined in terms of the total rank correlations by means of the same formula, but he gives [8] an entirely different interpretation.)

On the other hand, Yule [13] stated that the partial correlation "... should be regarded, in general, as of the nature of an average correlation ...", and Blalock [1, p. 332] notes that "... the partial correlation coefficient can be interpreted as a weighted average...". This second concept, which we have adopted, is also widely prevalent. For example, Goodman and Kruskal [6] use it to define a measure of partial correlation corresponding to their measure  $\lambda$  based on optimal prediction, and Davis [3]

uses it in developing a partial correlation based on their index  $\gamma$ .

It may be mentioned, however, that different weighting schemes are possible; thus, in Goodman and Kruskal's measure the conditional correlation at  $z$  is weighted in proportion to the probability that a single observation on  $Z$  will take the value  $z$ . Davis also considers this possibility, although he adopts finally the same scheme as ours, which is much simpler when working with indices based on concordant and discordant pairs; we will devote no further consideration to any other weighting scheme.

Consider now the estimation of these various parameters given a random sample of  $n$  observations  $(X_i, Y_i, Z_i)$ ,  $1 \leq i \leq n$ . From these observations it is possible to choose a pair in  $N = n(n-1)/2$  different ways. Of those pairs, let there be  $N_C$  concordant,  $N_D$  discordant, and  $N_T$  tied, with respect to  $X$  and  $Y$ . The obvious estimate of the total correlation  $\tau_{XY}$  is then

$$t_{XY} = \frac{N_C - N_D}{N},$$

the difference between the proportions of concordant and discordant pairs in the sample. The alternative indices may be estimated analogously: for  $\tau_b$  use

$$t_b = \frac{N_C - N_D}{\sqrt{N - N_X} \sqrt{N - N_Y}}$$

where  $N_X$  (or  $N_Y$ ) is the number of pairs tied on  $X$  (or  $Y$ ); and for  $\gamma$  use

$$G = \frac{N_C - N_D}{N_C + N_D}.$$

All of these are, in fact, already standard statistics.

If  $Z$  is discrete, the index  $\tau_{XY|Z}(z)$  of conditional correlation may be estimated by

$$t_{XY|Z}(z) = \frac{N_C(z) - N_D(z)}{N(z)} ,$$

where  $N(z)$  is the number of pairs of observations  $((X_i, Y_i, Z_i), (X_j, Y_j, Z_j))$  such that  $Z_i = Z_j = z$ ,  $N_C(z)$  is the number of these pairs which are concordant with respect to  $X$  and  $Y$ , and  $N_D(z)$  the corresponding number discordant; if it should happen that  $N(z) = 0$  then  $t_{XY|Z}(z)$  is undefined. To estimate the index  $\tau_{XY|Z}$  of partial correlation, simply add numerators and denominators over all possible values of  $Z$ , obtaining

$$t_{XY|Z} = \frac{\sum_z [N_C(z) - N_D(z)]}{\sum_z N(z)} .$$

The denominator of this estimate is the number of pairs of observations  $((X_i, Y_i, Z_i), (X_j, Y_j, Z_j))$  for which  $Z_i = Z_j$ , i.e., which are tied on  $Z$ , and the numerator is the number of these pairs which are concordant with respect to  $X$  and  $Y$ , less the number discordant.

On the other hand, if  $Z$  is continuous, the method of estimation just presented will not serve, because there will be no pairs of observations tied on  $Z$ . But let  $D(z_i, z_j)$  be a suitable measure of distance defined for all pairs  $(z_i, z_j)$  of conceivable values of  $Z$ , and for each  $\epsilon > 0$  define

$$\tau_{XY|Z}(z, \epsilon) = P_{C|Z}(z, \epsilon) - P_{D|Z}(z, \epsilon) ,$$

where  $p_{C|Z}(z, \epsilon)$  (or  $p_{D|Z}(z, \epsilon)$ ) is the probability that a randomly-chosen pair of observations  $(X_1, Y_1, Z_1)$  and  $(X_2, Y_2, Z_2)$  will be concordant (or discordant) with respect to  $X$  and  $Y$ , conditional upon the event that both  $D(Z_1, z) \leq \epsilon$  and  $D(Z_2, z) \leq \epsilon$ . Then a reasonable estimate of  $\tau_{XY|Z}(z, \epsilon)$  is

$$t_{XY|Z}(z, \epsilon) = \frac{N_C(z, \epsilon) - N_D(z, \epsilon)}{N(z, \epsilon)}$$

where  $N(z, \epsilon)$  is the number of pairs of observations  $((X_i, Y_i, Z_i), (X_j, Y_j, Z_j))$  such that both  $D(Z_i, z) \leq \epsilon$  and  $D(Z_j, z) \leq \epsilon$ ,  $N_C(z, \epsilon)$  is the number of these pairs which are concordant with respect to  $X$  and  $Y$ , and  $N_D(z, \epsilon)$  the corresponding number discordant. Now suppose, as seems reasonable in most situations, that  $\tau_{XY|Z}(z, \epsilon)$  is continuous in  $\epsilon$  for  $\epsilon \geq 0$ , so that the index of conditional correlation as previously defined is the limiting case

$$\tau_{XY|Z}(z) = \lim_{\epsilon \rightarrow 0} \tau_{XY|Z}(z, \epsilon) ;$$

then  $t_{XY|Z}(z, \epsilon)$  should be a reasonable estimate of  $\tau_{XY|Z}(z)$  for  $\epsilon$  sufficiently small. Similarly, if  $N(\epsilon)$  is the number of pairs of observations  $((X_i, Y_i, Z_i), (X_j, Y_j, Z_j))$  such that  $D(Z_i, Z_j) \leq \epsilon$ , with  $N_C(\epsilon)$  the number of these pairs which are concordant with respect to  $X$  and  $Y$ , and  $N_D(\epsilon)$  the corresponding number discordant, then

$$t_{XY|Z}(\epsilon) = \frac{N_C(\epsilon) - N_D(\epsilon)}{N(\epsilon)}$$

should be a reasonable estimate of  $\tau_{XY|Z}$  for sufficiently small  $\epsilon$ . The quantity  $\epsilon$  may be referred to as the tolerance.

The "suitable measure"  $D$  need not be the Euclidean distance function, and in fact what is required does not exactly correspond with what mathematicians call a distance. Let  $Z$  be the vector  $\underline{z} = (z^{(1)}, z^{(2)}, \dots, z^{(m)})'$ , where  $m \geq 1$ . Then it appears to be sufficient that  $D(\underline{z}_i, \underline{z}_j)$  be a continuous and nondecreasing function of the absolute differences  $|z_i^{(k)} - z_j^{(k)}|$  for  $1 \leq k \leq m$ , with  $D(\underline{z}_i, \underline{z}_j) = 0$  if and only if  $\underline{z}_i = \underline{z}_j$ . Some convenient measures are:

(maximum component distance)

$$D_1(\underline{z}_i, \underline{z}_j) = \max_k c_k |z_i^{(k)} - z_j^{(k)}|,$$

(city-block distance)

$$D_2(\underline{z}_i, \underline{z}_j) = \sum_k c_k |z_i^{(k)} - z_j^{(k)}|,$$

(weighted Euclidean distance)

$$D_3(\underline{z}_i, \underline{z}_j) = (\underline{z}_i - \underline{z}_j)' Q (\underline{z}_i - \underline{z}_j),$$

(Mahalanobis distance)

$$D_4(\underline{z}_i, \underline{z}_j) = (\underline{z}_i - \underline{z}_j)' S^{-1} (\underline{z}_i - \underline{z}_j),$$

where the  $c$ 's are arbitrary positive constants,  $Q$  is an arbitrary symmetric positive definite matrix, and  $S$  is the sample variance matrix of the  $Z$ 's.

We realize that justification of the various estimates suggested here has so far been entirely intuitive. We postpone discussion of their sampling distributions to the end of Section 3, since all of them may be viewed as special cases of the general index there presented.

### 3. MATCHED CORRELATION

We now return to the general problem set forth in Section 1, to measure the correlation between two variables X and Y when a third variable Z has been controlled. Let there be established some rule by which it can be decided whether any two observations are "matched" or "not matched" with respect to Z: that is, whether they are equal or unequal, within some reasonable tolerance. (We expect that such a rule will generally be based on essentially non-statistical considerations; that is, the definition of what does or does not constitute a "match" will have to be made appropriately for the individual application one has in mind.)

Then our proposed measure, an index of matched correlation, is

$$\tau_{XY|M(Z)}^* = P_{C|M(Z)} - P_{D|M(Z)} ,$$

the difference between the conditional probabilities of concordance and discordance with respect to X and Y of a random pair of observations, given that the pair is matched on Z. (We may write simply  $\tau^*$  when X, Y, Z and M are clear from the context.) This index, as is true of all others considered, is restricted to the range  $-1 \leq \tau^* \leq 1$ :  $\tau^* = 1$  if  $P_{C|M(Z)} = 1$ , that is, if all pairs matched on Z are necessarily concordant with respect to X and Y; and  $\tau^* = -1$  if  $P_{D|M(Z)} = 1$ , that is, if all such pairs are discordant. And  $\tau^* = 0$  if  $P_{C|M(Z)} = P_{D|M(Z)}$ , that is, if pairs matched on Z are equally likely to be concordant or discordant with respect to X and Y; note that for this to be true it is generally not sufficient (or necessary) that X and Y be

conditionally independent given  $Z$ , although it will be at least "nearly sufficient" if the tolerance for matching is small.

Consider estimating the index of matched correlation from a random sample of  $n$  observations  $(X_i, Y_i, Z_i)$ ,  $1 \leq i \leq n$ . Then from among all  $N = n(n-1)/2$  possible pairs of observations count the number which are matched on  $Z$ , say  $N_M$ ; and among these count the numbers concordant and discordant with respect to  $X$  and  $Y$ , say  $N_{CM}$  and  $N_{DM}$ , respectively. Then an intuitively reasonable estimate of  $\tau^*$  is

$$t^* = \frac{N_{CM} - N_{DM}}{N_M} .$$

This sample index also lies in the range  $-1 \leq t^* \leq 1$ , taking the value 1 if the observed matched pairs are all concordant, -1 if they are all discordant, and 0 if there are equally as many concordant as discordant matched pairs. (If it should happen that a sample contained no matched pairs then  $t^*$  would be regarded as undefined.)

Now, for each  $i$ ,  $1 \leq i \leq n$ , let  $M_i$  be the number of observations  $(X_j, Y_j, Z_j)$ ,  $j \neq i$ , which are matched on  $Z$  with the observation  $(X_i, Y_i, Z_i)$ ; and let  $W_i$  be the number of these which are concordant with  $(X_i, Y_i, Z_i)$  with respect to  $X$  and  $Y$ , less the number discordant. Then  $\sum M_i = 2N_M$  (the factor 2 appears because each matched pair is counted twice),  $\sum W_i = 2(N_{CM} - N_{DM})$ , and hence

$$t^* = \frac{\sum W_i}{\sum M_i} .$$

Furthermore,

a convenient formula for the

standard error of  $t^*$  is

$$s^* = \frac{2}{(\sum M_i)^2} \sqrt{\sum M_i^2 (\sum W_i)^2 - 2\sum M_i \sum W_i \sum M_i W_i + (\sum M_i)^2 \sum W_i^2},$$

and the distribution of the ratio

$$\frac{t^* - \tau^*}{s^*}$$

approaches the standard normal as  $n$  tends to infinity. Hence, it is possible to obtain large-sample tests and confidence intervals for  $\tau^*$ , the population index of matched correlation.

The above results may be proven as follows. Establishment of the rule for matching is equivalent to choosing a matching function  $M(z_i, z_j)$ , defined for all conceivable pairs of observations on  $Z$ , such that  $M(Z_i, Z_j) = 1$  if  $Z_i$  and  $Z_j$  are matched, but  $M(Z_i, Z_j) = 0$  if not. To satisfy the mathematical requirements it is sufficient to take any arbitrary function  $M(z_i, z_j)$  which is symmetric in  $z_i$  and  $z_j$  and assumes the values 0 and 1 only; this includes functions not corresponding to any process which could be called "matching" in ordinary language. Then the probability that a randomly-chosen pair of observations  $(X_1, Y_1, Z_1)$  and  $(X_2, Y_2, Z_2)$  will be matched on  $Z$  is

$$p_M = \iint M(z_1, z_2) dG(x_1, y_1, z_1) dG(x_2, y_2, z_2),$$

where  $G(X, Y, Z)$  is the joint distribution function of  $(X, Y, Z)$ , and the U-statistic<sup>1</sup> for estimating  $p_M$  is

$$U_2 = \binom{n}{2}^{-1} \sum_{i < j} M(Z_i, Z_j);$$

---

<sup>1</sup> See the Appendix for definitions. Those readers who are not mathematically inclined may skip the remainder of this paragraph, the next two paragraphs, and the Appendix.

but a little reflection will show that  $U_2$  is precisely  $N_M/N$ . Similarly, the probability that a random pair of observations will be both matched on  $Z$  and concordant with respect to  $X$  and  $Y$ , less the probability that it will be matched but discordant, is

$$P_{CM} - P_{DM} = \iint W((x_1, y_1), (x_2, y_2))M(z_1, z_2) dG(x_1, y_1, z_1) dG(x_2, y_2, z_2) ;$$

the U-statistic for estimating this from the sample is

$$U_1 = \binom{n}{2}^{-1} \sum_{i < j} W((X_i, Y_i), (X_j, Y_j))M(Z_i, Z_j) = \frac{N_{CM} - N_{DM}}{N} .$$

Hence, the proposed sample estimate of the population index of matched correlation is

$$t^* = \frac{N_{CM} - N_{DM}}{N_M} = \frac{U_1}{U_2} ,$$

the ratio of two U-statistics.

The results of the Appendix are now applicable. Identify the general random variable  $X$  used there with  $(X, Y, Z)$ , and let  $\theta_1 = P_{CM} - P_{DM}$ ,  $\theta_2 = P_M$ ,  $m_1 = m_2 = 2$ ,  $\phi_1 = MW$ , and  $\phi_2 = M$ . Note that the components are  $V_1^{(i)} = W_i/(n-1)$  and  $V_2^{(i)} = M_i/(n-1)$ , so that with a little algebra it can be verified that

$$s^2 = \frac{4n^2}{n-1} \cdot \frac{\sum_i M_i^2 (\sum W_i)^2 - 2 \sum M_i \sum W_i \sum M_i W_i + (\sum M_i)^2 \sum W_i^2}{(\sum M_i)^4}$$

Finally, since  $s^* = s\sqrt{n-1}/n$ , and for an asymptotic result  $n$  and  $(n-1)$  may be considered equivalent, the result quoted above follows directly from the Theorem.

The conditions required by the Theorem, namely " $\theta_2 > 0$  and  $0 < \sigma^2 < \infty$ ", may be interpreted as follows. First,  $\sigma^2 = 4(\theta_2^2 \zeta_{11} - 2\theta_1 \theta_2 \zeta_{12} + \theta_1^2 \zeta_{22}) / \theta_2^4$ , where  $\theta_2$  is  $p_M$ . Thus it is necessary to have  $p_M > 0$ , and then  $\sigma^2 < \infty$  since  $\phi_1 = MW$  and  $\phi_2 = W$  are bounded. We shall suppose also that  $\zeta_{11} > 0$ , the interpretation of which is somewhat esoteric; for discussion of a similar condition required in the asymptotic theory of Goodman and Kruskal's  $G$ , see page 364 of their paper [7]. At any rate, for it to fail is, as they say, "... an unlikely state of affairs in most applications". With  $\zeta_{11} > 0$  and  $p_M > 0$ , we have  $\sigma^2 > 0$  if  $\zeta_{22} = 0$ , which will be true only if  $p_M = 1$ , or if  $\zeta_{22} > 0$  and  $\zeta_{12}^2 < \zeta_{11} \zeta_{22}$ , which will then be true if  $-1 < \tau^* < 1$ . Thus for practical purposes we may regard the asymptotic results as valid if  $p_M > 0$  and  $-1 < \tau^* < 1$ .

Goodman and Kruskal [7] have established the upper bound  $2(1-\gamma^2)/(p_C+p_D)$  for the variance of the asymptotic distribution of  $\sqrt{n}(G-\gamma)$ . The corresponding upper bound for the index of matched correlation would be:

$$\text{asymptotic variance of } \sqrt{n}(t^*-\tau^*) = \sigma^2 \leq \frac{2(1-\tau^{*2})}{p_M} .$$

To prove that this bound is valid for the case where  $X$  and  $Y$  are continuous, so that it is impossible for pairs to be tied on  $X$  and  $Y$ , we may use Goodman and Kruskal's argument exactly, if we interpret their subscript "s" as indicating "concordant and matched", "d" as "discordant and matched", and "t" as "not matched". The bound can also be shown to hold if  $\tau^* = 0$  whether or not tied pairs can occur. Unfortunately, I have not been able to show whether it holds in the remaining case (tied pairs possible,  $\tau^* \neq 0$ ), although obviously it must hold at least approximately if ties are unlikely or if  $\tau^*$  is small.

One use for such a bound, as Goodman and Kruskal indicate, is to allow the possibility of "conservative" procedures for situations where one might be uncertain about the estimate  $s^*$  or unwilling to take the trouble to calculate it. For example, a "conservative"  $100(1-\alpha)\%$  large-sample confidence interval for  $\tau^*$  is formed by the set of values  $\tau^*$  which satisfy the quadratic inequality

$$n N_M (t^* - \tau^*)^2 \leq 2N(1-\tau^{*2}) z_{\alpha/2}^2 ,$$

where  $z_{\alpha/2}$  is defined to make the probability be  $\alpha/2$  that a normal deviate will exceed it; the unknown  $p_M$  in the upper bound has been estimated by  $N_M/N$ .

A second use for the bound is that it shows (specifically only for the most extreme case, but the relationship is at least qualitatively true in general) how the variance of the sample index decreases (to 0) as the population index approaches  $\pm 1$  and how it increases as the probability that a random pair will be matched approaches 0. This will be seen more clearly in the examples of Section 4.

Any of the sample indices of conditional and partial correlation suggested in Section 2 may now be treated as a special case of matched correlation by defining the appropriate matching function,  $M$ , in accordance with the following:

For the index:	Define $M(Z_i, Z_j) = 1$ if: (and define $M=0$ otherwise)
----------------	--

---

$t_{XY|Z}(z)$   
(discrete  $Z$ )

$$Z_i = Z_j = z$$

$t_{XY|Z}$   
(discrete  $Z$ )

$$Z_i = Z_j$$

$$t_{XY|Z}(z, \epsilon) \quad \max \{D(Z_i, z), D(Z_j, z)\} \leq \epsilon$$

(continuous Z)

$$t_{XY|Z}(\epsilon) \quad D(Z_i, Z_j) \leq \epsilon$$

(continuous Z)

The fact that  $(t^* - \tau^*)/s^*$  is asymptotically a normal deviate, where  $ns^{*2}$  converges to  $\sigma^2$  in probability, implies that  $t^*$  is a consistent estimator of  $\tau^*$ ; indeed, its asymptotic mean squared error is  $\frac{\sigma^2}{n} + o(\frac{1}{n})$ . Hence a fortiori, if  $Z$  is discrete, then the suggested estimators of the indices of conditional and partial correlation are also consistent, with mean squared errors of order  $\frac{1}{n}$ . However, if  $Z$  is continuous, then the suggested sample index of matched correlation, as an estimator of a population index of conditional or partial correlation, may be biased, even asymptotically, for any fixed tolerance. Presumably consistency could be achieved by reducing the tolerance appropriately as the sample size increases; but the methods of this paper are insufficient to study this point.

#### 4. EXAMPLES

For our first example, which will illustrate the method of computation in some detail, we use the data of Table 1. These fictitious data are part of a set originally invented as illustrative material for a course in descriptive statistics; they are supposed to represent the sex, IQ, height, and final examination results of a class of fourth graders. We consider the correlation between the ordinal variable  $X$  = examination result, recorded as A, B, C, D, or F, and the metric variable  $Y$  = height, recorded in inches. Our "given" variable is a bivariate  $\underline{Z}$  with its first component

Table 1

SEX, IQ, HEIGHT, AND FINAL EXAMINATION RESULTS FOR A CLASS  
OF FOURTH-GRADE CHILDREN (fictitious data)

i	Result of exam	Height (in.)	Sex	IQ	Without matching		Matching on sex only		Matching on IQ <sup>1</sup> only		Matching on sex and IQ <sup>1</sup>	
	X	Y	Z <sub>1</sub>	Z <sub>2</sub>	M	W	M	W	M	W	M	W
1	F	50	M	85	24	19	12	9	4	3	2	2
2	D	58	M	92	24	-12	12	-3	9	-4	5	-1
3	D	54	M	93	24	2	12	5	10	2	6	5
4	A	56	M	96	24	9	12	1	10	2	5	-1
5	C	55	M	100	24	3	12	6	10	2	6	2
6	C	58	M	102	24	-1	12	1	11	0	6	-1
7	B	57	M	103	24	7	12	2	10	3	5	1
8	C	53	M	109	24	3	12	2	10	1	5	1
9	F	54	M	115	24	1	12	4	9	-3	4	-2
10	B	57	M	118	24	7	12	2	8	3	5	2
11	A	49	M	120	24	-21	12	-11	7	-6	4	-4
12	D	52	M	123	24	7	12	6	7	0	4	0
13	B	60	M	128	24	12	12	6	6	1	3	0
14	C	51	F	83	24	0	11	0	4	-2	1	-1
15	B	50	F	86	24	-13	11	-6	5	-2	1	-1
16	C	52	F	98	24	1	11	0	9	-2	3	-2
17	D	57	F	99	24	-9	11	-9	10	-3	3	-1
18	F	53	F	105	24	6	11	0	11	6	5	2
19	C	53	F	106	24	3	11	1	11	1	5	0
20	A	54	F	111	24	2	11	5	10	0	4	2
21	C	55	F	114	24	3	11	-3	9	2	4	0
22	C	51	F	121	24	0	11	0	8	0	3	1
23	C	52	F	131	24	1	11	0	5	3	3	2
24	A	55	F	135	24	7	11	7	3	1	2	2
25	B	54	F	140	24	5	11	5	2	2	2	2

<sup>1</sup> within a tolerance of 10 units.

the nominal variable  $Z_1 = \text{sex}$  and its second component  $Z_2 = \text{IQ}$ .

The sample index of matched correlation between examination result and height with the effects of sex and IQ removed, that is, between X and Y given both  $Z_1$  and  $Z_2$ , is computed using the values of  $M_i$  and  $W_i$  in the last section of the table. The matching function is

$$M(z_i, z_j) = \begin{cases} 1 & \text{if } z_{1i} = z_{1j} \text{ and } |z_{2i} - z_{2j}| \leq 10 \\ 0 & \text{otherwise;} \end{cases}$$

that is, two children are regarded as matched if they are of the same sex and differ in IQ by no more than 10 units. The first child (say) is therefore matched with exactly two others, namely the second and third (for convenience in hand computation the data have been sorted on the variables to be controlled for), hence  $M_i = 2$ ; and he is concordant with both of them--in particular, he is the shortest of the three, and also received the lowest grade -- hence  $W_i = 2$  also. The values of  $M_i$  and  $W_i$  for the other 24 children can be checked similarly, and indeed it would be instructive for the reader to check at least one or two more. We may then compute  $\sum M_i = 96$ , indicating that there are 48 matched pairs of children, and  $\sum W_i = 10$ , indicating that there are 5 more concordant pairs than discordant; hence the index is  $t^* = \sum W_i / \sum M_i = 10/96 = .104$ . (There are actually 22 concordant pairs, 17 discordant, and 9 tied; without modification, however, the computational scheme here presented does not provide these numbers.) Having calculated  $\sum M_i^2 = 422$ ,  $\sum M_i W_i = 50$ , and  $\sum W_i^2 = 90$ , we then find  $s^* = .191$ . Thus the index is smaller than its standard error and therefore not significantly different from zero. If this sample could be regarded as large, we would take  $t^*/s^* = .545$  as a normal deviate in making such a test. We could also produce the 95% (say) confidence interval  $t^* \pm 1.96s^*$ , or  $(-.270, +.469)$ , for the population index  $\tau^*$ . However, with only 25 observations and 48 matched pairs (which are not independent of each other) it is best to be somewhat restrained in making such inferences.

The first section of Table 1 (labeled "without matching") shows the components for the index of total correlation, obtained by defining  $M(z_i, z_j) \equiv 1$ ;

in this case  $M_i = (n-1) = 24$  for all  $i$ , every observation being considered matched with every other. We then have 300 pairs, of which there are 21 more concordant than discordant (actually there are 122 concordant pairs, 101 discordant, and 77 tied) and hence the index takes the value  $t^* = 21/300 = .070$ . Its standard error may be computed according to the formulas given earlier and turns out to be  $s^* = .136$ . Again we conclude that the correlation is not significant.

The other two sections of Table 1 show the components for indices where matching has been performed on only one of the two variables, either sex or IQ; the computations proceed in exactly the same manner. Results are summarized in Table 2. Note that the two indices of conditional correlation given sex are obtainable almost as byproducts of the computation for the index of matched (or, partial) correlation given sex: to obtain the conditional correlation among males, we take  $M_i$  and  $W_i$  the same as for the matched correlation if the  $i$ -th student is male, and take  $M_i = W_i = 0$  if the  $i$ -th student is female; and for the conditional correlation among females we do the reverse.

Table 2

Correlation	Matching function: $M(\underline{Z}_i, \underline{Z}_j) = \text{lif}$	$\Sigma M_i$	$\Sigma W_i$	$\Sigma M_i^2$	$\Sigma M_i W_i$	$\Sigma W_i^2$	$t^*$	$s^*$
Total	(always)	600	42	14400	1008	1726	.070	.136
Matched on sex	$Z_{1i} = Z_{1j}$	288	30	3324	360	600	.104	.165
Conditional on male	$Z_{1i} = Z_{1j} = \text{"male"}$	156	30	1872	360	374	.192	.224
Conditional on female	$Z_{1i} = Z_{1j} = \text{"female"}$	132	0	1452	0	226	.000	.228
Matched on IQ <sup>1</sup>	$ Z_{2i} - Z_{2j}  \leq 10$	198	10	1744	88	178	.051	.133
Matched on sex and IQ <sup>1</sup>	$Z_{1i} = Z_{1j}$ and $ Z_{2i} - Z_{2j}  \leq 10$	96	10	422	50	90	.104	.191

<sup>1</sup> within a tolerance of 10 units

The values of  $\sum M_i$ ,  $\sum W_i$ ,  $\sum M_i^2$ ,  $\sum M_i W_i$ , and  $\sum W_i^2$  for the matched correlation are equal to the sums of the corresponding values for the two conditional correlations. A similar situation will obtain whenever the variable being controlled for is discrete.

The computational scheme is thus extremely simple, in principle. It is extremely tedious in practice, however, although it is admirably suited to an electronic computer.

Now let us consider an example in which the underlying population distribution is known. For  $1 \leq i \leq 50$  let  $C_{1i}$ ,  $C_{2i}$ , and  $C_{3i}$  be the entries in columns 01, 02, and 03 of the table of random normal deviates given by Dixon and Massey [4], and for  $51 \leq i \leq 100$  continue with columns 11, 12, and 13. Now define  $X_i = C_{1i} + C_{3i}$ ,  $Y_i = C_{2i} + C_{3i}$ ,  $Z_i = C_{3i}$ , for  $1 \leq i \leq 100$ . Then simple considerations show that the population total (product-moment) correlation is  $\rho_{XY} = \frac{1}{2}$ , with partial correlation  $\rho_{XY|Z} = 0$ ; the corresponding sample values happen to be  $r_{XY} = .533$ ,  $r_{XY|Z} = -.009$ . In a normal population

$$\tau = \frac{2}{\pi} \sin^{-1} \rho ;$$

hence  $\tau_{XY} = 1/3$ ,  $\tau_{XY|Z} = 0$ . Results of computing the sample index of matched correlation for various values of the tolerance are given in Table 3. Recall that matching with infinite tolerance gives the index of total correlation, and that with matching with zero tolerance (when  $Z$  is continuous, as here, this is of course not possible except as a limiting case) gives the partial correlation. As the tolerance decreases through finite positive values the population index of matched correlation decreases also; but since the relationship for a normal population is rather complicated, we have not computed the exact values.

Table 3

Tolerance for matching on Z <sup>1</sup>	Number of Matched pairs	Population index	Sample index	Standard error
$\epsilon$	$N_M$	$\tau^*$	$t^*$	$s^*$
$\infty$	4950	.333	.363	.053
3.00	4804	--	.343	.052
2.00	4164	--	.261	.053
1.50	3486	--	.179	.055
1.00	2514	--	.086	.060
.75	1942	--	.049	.063
.50	1308	--	.040	.069
.25	674	--	-.003	.078
0	0	.000	--	--

<sup>1</sup> The true standard deviation of Z is  $\sigma_Z = 1.000$ , with  $s_Z = .997$ .

We can see, however, the steady decrease in the sample values. We also see that as the tolerance decreases, and the number of matched pairs correspondingly, the standard error increases; this would be expected, of course, on intuitive grounds, and also from the form of the upper bound given in Section 3; but the increase is not drastic until a very small tolerance has been reached.

We now present three examples, using previously published data, in which we compare the index of matched correlation with other measures. Consider first the example originally presented by Yule [13] and extensively quoted since, in which X = estimated average earnings of agricultural laborers, Y = ratio of the number of paupers receiving "outdoor" relief to the number receiving relief in the workhouse, and Z = percentage of population on relief, for n = 38 rural districts. The product-moment total correlations are  $r_{XY} = -.13$ ,  $r_{XZ} = -.66$ , and  $r_{YZ} = +.60$ , from which the partial correlation

formula gives  $r_{XY|Z} = +.44$ . The results of computing the index of matched correlation, summarized in Table 4, show a similar relationship.

Table 4

Tolerance for matching on Z <sup>1</sup>	Number of matched pairs	Index of matched correlation	Standard error
$\epsilon$	$N_M$	$t^*$	$s^*$
$\infty$	703	-.078	.096
2.00	500	.136	.089
1.50	393	.226	.092
1.00	269	.294	.107
.50	142	.331	.115

<sup>1</sup> The standard error of Z is  $s_Z = 1.29$ .

For our second example we use the data of **Angell** quoted by Blalock [1, p. 300] for  $n = 29$  non-Southern cities of 100,000 or more. Here X is an index of moral integration "derived by combining crime-rate indices with those for welfare effort", Y is an index of heterogeneity "measured in terms of the relative numbers of nonwhites and foreign-born whites in the population", and Z is "a mobility index measuring the relative numbers of persons moving in and out of the city". The product-moment total correlations are  $r_{XY} = -.156$ ,  $r_{XZ} = -.456$ ,  $r_{YZ} = -.513$ , with partial correlation  $r_{XY|Z} = -.511$ . Results for the index of matched correlation are summarized in Table 5, and as in the previous example they agree nicely with those found by the more standard method. In these two examples the index increases in absolute value as the tolerance is reduced, and since a correlation is the more accurately determined the farther it is from zero this has to some extent cancelled out the otherwise - expected increase in standard error.

Table 5

Tolerance for matching on Z <sup>1</sup>	Number of matched pairs	Index of matched correlation	Standard error
$\epsilon$	$N_M$	$t^*$	$s^*$
$\infty$	406	-.138	.100
20	349	-.209	.090
15	286	-.294	.079
10	215	-.349	.085
5	125	-.488	.103
2	47	-.532	.134
1	24	-.583	.165

<sup>1</sup> The standard error of Z is  $s_Z = 9.66$ .

For our last example we use the data of Hajda quoted by Davis [3], which were obtained from a sample survey of Baltimore women. Here X is a dichotomy, taking the values "high" or "low" according as the respondent was above or below 45 years of age; Y is another dichotomy, taking the values "high" or "low" according as she had or had not read a book recently; and Z is her educational attainment, recorded in three categories "college", "high school", and "less than high school". Since it may be instructive to see the full calculations for a problem involving categorical data, we present them in Table 6. Essentially, what has been done is to list each of the 12 possible values of (X,Y,Z), the frequency (F) of occurrence of the value, and the M and W which correspond to any one of the F observations at that value. We then have

$$t^* = \frac{\sum FW}{\sum FM}$$

and

$$s^* = \frac{2}{(\sum FM)^2} \sqrt{\sum FM^2 (\sum FW)^2 - 2 \sum FM \sum FW \sum FMW + (\sum FM)^2 \sum FW^2} .$$

Table 6

Age	Book Reading	Education	Frequency	M	W	C	D	T
X	Y	Z	F					
High	High	College	104	348	46	46	0	342
	Low		36	348	-163	0	163	185
Low	High		163	348	-36	0	36	312
	Low		46	348	104	104	0	244
High	High	High School	159	954	327	327	0	627
	Low		179	954	-290	0	290	664
Low	High		290	954	-179	0	179	775
	Low		327	954	159	159	0	795
High	High	Less than High School	54	545	133	133	0	412
	Low		335	545	-24	0	24	521
Low	High		24	545	-335	0	335	210
	Low		133	545	54	54	0	491

In the example,  $\Sigma FM = 1330092$ ,  $\Sigma FW = -3718$ ,  $\Sigma FM^2 = 1073601726$ ,  $\Sigma FMW = -1531320$ , and  $\Sigma FW^2 = 55729114$ ; hence  $t^* = -.0028$ ,  $s^* = .0112$ . (A number of shortcuts obviously could be taken in the computations for this rather simple example; a general computer program, however, would probably best proceed from the formulas as given.)

The "partial coefficient for Goodman and Kruskal's gamma" proposed by Davis [3], namely  $G_{XY|Z}$ , is for this example the same as our index of partial correlation, or matched correlation with zero tolerance, except that he does not include in the denominator those pairs which are matched on Z but tied with respect to X and Y. The comparison of his measure with ours is then the same as between the Goodman-Kruskal G and the Kendall  $t_a$ , as will be discussed in Section 5. This may be made clearer from the last three columns of Table 6, which show how many of the matched pairs are concordant, discordant,

or tied (labeled "C", "D", and "T", respectively); we have  $W = C-D$ ,  $M = C+D+T$ . If, however, we redefine  $M = C+D$ , thus leaving out the ties, we obtain  $\Sigma FM = 259554$ ,  $\Sigma FW = 3718$  (as before),  $\Sigma FM^2 = \Sigma FW^2 = 55729114$  (this equality holds whenever  $X$  and  $Y$  are both dichotomous, but not in general), and  $\Sigma FMW = -1070650$ . Then Davis' coefficient and its standard error may be calculated by substituting these values in the formulas given above, yielding  $G_{XY|Z} = -3718/259554 = -.0143$ , with standard error .0581. Thus Davis' coefficient and its standard error are both about five times larger than the corresponding values for the index of matched correlation. The validity of this procedure still follows from the argument of Section 3, even though the matching function now depends not only on the  $Z$ 's but also on the  $X$ 's and  $Y$ 's: specifically, we now have  $M((x_1, y_1, z_1), (x_2, y_2, z_2)) = 1$  only if  $z_1 = z_2$  and also  $x_1 \neq x_2$ ,  $y_1 \neq y_2$ , with  $M = 0$  otherwise.

## 5. DISCUSSION

Whether the denominators of sample indices of correlation (total, conditional, partial, or matched) based on counting concordant and discordant pairs should include pairs tied on  $X$  and  $Y$  is probably to some extent a matter of taste. For one thing, omitting the tied pairs generally simplifies computation somewhat, (at least hand computation), by making all the  $M_i$ 's smaller. But then again, the theory is perhaps a little more elegant when the matching function depends only on the  $Z$ 's.

But from the standpoint of interpretation, I prefer to retain the ties, in general, on the grounds that the variables  $X$  and  $Y$ , even if not continuous as recorded, usually represent underlying continua, with ties occurring only because of imprecision in measurement (or grouping later). In that case it

seems to me that the descriptive measure we adopt should involve a certain penalty (in the sense of giving values closer to zero) for such imprecision. Thus, and this is essentially the argument given by (presumably) Karl Pearson [10] in a similar situation, even if all untied pairs are concordant it is best that the index not be allowed to take on the value +1, since this might imply a perfect positive correlation of which we cannot be certain; we have no way of knowing how the tied pairs might split if they could be properly resolved.

A somewhat different argument might run as follows. If X and Y as recorded really do represent imprecise or grouped measurements on an underlying continuous population then it may be a reasonable goal to determine the correlation in that continuous population. Consider, therefore, what proportions of those pairs which we recorded as tied would be concordant and discordant if such ties were properly resolved. This requires guessing at the correlations within subpopulations where the range of X or Y is severely restricted. Surely one would ordinarily expect such subcorrelations to be much smaller, on the whole, than the correlation for unrestricted X and Y. Suppose they are put equal to zero, which is effectively accomplished in the sample by considering one half of the tied pairs as concordant and the other half as discordant. Then the index of total correlation in the adjusted data will be

$$\text{(new) } t_{a, t_b}, \text{ or } G = \frac{(N_C + \frac{1}{2}N_T) - (N_D + \frac{1}{2}N_T)}{N} = \text{(old) } t_a$$

(the denominator for any of the indices is N for the adjusted data, where no ties remain); that is, we will obtain what was  $t_a$  in the original data. This corresponds to what might be called a "pessimistic" view which will generally

result in underestimating the strength of the underlying correlation. Similar remarks hold true for the sample indices of conditional and partial correlation proposed here, since their denominators also include tied pairs. On the other hand, suppose the tied pairs are allocated in the same proportions as the untied ones, which is equivalent to an "optimistic" assumption that the subcorrelations are on the average equal to the overall correlation. Then the index of total correlation in the adjusted data will be

$$(\text{new})t_a, t_b, \text{ or } G = \frac{(N_C + \frac{N_C N_T}{N_C + N_D}) - (N_D + \frac{N_D N_T}{N_C + N_D})}{N} = (\text{old}) G;$$

that is, we will arrive at what was the Goodman-Kruskal index in the original data. Again similar remarks apply for any other index whose denominator omits the tied pairs, as for example Davis' coefficient; all these will generally result in overestimating the strength of the underlying correlation. In most contexts underestimation would be considered preferable to overestimation, and in addition it is likely that the underestimate will be closer to the truth; hence we are led to retain the tied pairs in the denominators of our indices rather than discard them.

The more customary index of total correlation, Kendall's  $t_b$ , is a compromise between the other two and hence might well provide the most precise estimate in general. My personal impression, admittedly based on a rather limited number of examples, is that in most cases  $t_b$  lies about halfway between the other two indices, with the underlying true correlation generally between it and  $t_a$ ; this might be a fruitful area for more thorough investigation. Of course,  $t_b$  is more difficult to interpret in terms of the measurements actually at hand, and certainly much more difficult to work with both theoretically and numerically.

If X and Y definitely do not represent underlying continua, then the heuristic arguments given above would not apply. Goodman and Kruskal [6, p. 736] note that "The desirability of assuming an underlying joint continuum was one of the issues of a heated debate forty [now more than fifty] years ago between Yule on the one hand and K. Pearson and Heron on the other. Yule's position was that very frequently it is misleading and artificial to assume underlying continua; Pearson and Heron argued that almost always such an assumption is both justified and fruitful." I appear to be on the side of Pearson and Heron. Goodman and Kruskal take, in effect, the opposite view when they claim (p. 737) that "... it is in fact desirable that a measure of association reflect the classes as defined for the data"; thus they are not interested in estimating the correlation in any hypothetical underlying continuous population.

A question to which we have no definite answer concerns the distribution of our index in small samples, and indeed the proper definition of "small" in this context. It appears that Monte Carlo sampling will be required to come to a satisfactory conclusion, but the following speculation may be offered. The index  $t^*$  of matched correlation is similar in many respects to the Goodman and Kruskal  $G$ , so the number of matched pairs observed may have an importance for  $t^*$  roughly equivalent to that of the number of untied pairs for  $G$ . Fairly extensive sampling experiments by Rosenthal [11] for 5 x 5 crossclassifications over a wide range of true values of  $\gamma$  showed the distribution of  $(G-\gamma)/s_G$ , where  $s_G^2$  is the maximum likelihood estimator of the variance of  $G$ , to be reasonably close to the standard normal in samples of size  $n=25$  or 50 for  $|\gamma| < .50$ . There was a tendency for  $s_G^2$  to underestimate the variance, and this was particularly true for larger values of  $\gamma$ . The probability of a tie in a 5 x 5 cross-classification cannot be less than .20, and in the representative

examples presented by Rosenthal it varies from about .25 up to more than .40; hence, since the total number of pairs  $N$  is 300 at  $n=25$  and 1225 at  $n=50$ , it appears that her experiments must typically have involved some 200 untied pairs at  $n=25$  and 800 at  $n=50$ . We may then imagine that similar results would be obtained for indices of matched correlation based on numbers of matched pairs in that range. It seems likely that the estimate  $s^*$  of the standard error will also tend to understate the true value: for example,  $s^* = 0$  if all the matched pairs are concordant, all discordant, or all tied, and this is not unlikely in very small samples. Other estimates are, of course, possible; ours, based on the results of Sen [12], was chosen almost entirely on the basis of its simplicity. This might be another suitable area for further investigation.

Another question which may be asked is this: if  $Z$  is continuous, then how does the proposed estimate of the index of conditional or partial correlation depend on the tolerance allowed in matching? The examples of the previous section have thrown some light on this question, the results being much as one would have expected on intuitive grounds. When the tolerance is infinite, matched correlation is equivalent to total correlation. As the tolerance decreases to zero, the population matched correlation approaches the appropriate conditional or partial correlation; but in a sample the number of matched pairs decreases also, leaving less and less data on which to base the estimate, whose variance accordingly increases. Thus the optimal tolerance for estimating a conditional or partial correlation is a compromise -- a large value may give too much bias, and a small value too much variance. As mentioned at the end of Section 3, a way out of this difficulty might be to reduce the tolerance allowed as the sample size increases; for example, one might decide in advance

to base the index on the  $N_M$  most closely matched pairs for some fixed  $N_M$ . This will be somewhat troublesome for calculation, since it requires a sorting of the pairs of observations according to their distances apart with respect to  $Z$ ; but it does have the advantage of providing a non-random denominator for the index. The asymptotic results obtained in this paper do not strictly apply to such a situation, but for large  $N_M$  there will probably be no difference for practical purposes. Of course, it may be that in practice an index of matched correlation in the population will be accepted as the proper object of interest in itself, regardless of whether it equals the somewhat abstract index of conditional or partial correlation; the definition of matching can be arbitrarily established on non-statistical grounds appropriate to the individual application, and the interpretation of the result is then extremely simple.

The preceding paragraphs suggest a possible disadvantage involved in using the index of matched correlation, in that its sampling distribution is not known for smaller samples; but this is true of the competitive measures also, except for the product-moment partial correlation -- and then only if we have normality and linearity of regression. One may also ask whether much efficiency is lost by the nonparametric procedure in the latter situation; I intend to devote a subsequent paper to this point, but will state that the loss does not appear unreasonable when compared with the gain in ease of interpretation and in freedom from restrictive assumptions.

It may be useful to mention, although it is implicit in the very general definition allowed for "matching", that in many situations it will be convenient to match only after first transforming the variable  $Z$ . For example, if  $Z$  is the age of an individual, one might hesitate to designate a match as "within so many years" on the grounds that the same difference in age means

more for young individuals than old ones. This could be handled easily, however by transforming  $Z$  to  $\log Z$ , say, instead of using  $Z$  directly.

In conclusion, we remark that the matching function  $M(z_i, z_j)$  may be generalized, by dropping the requirement that it take on the values 0 and 1 only, and allowing it to vary throughout that range. For example, if  $D(z_i, z_j)$  is a distance function as described in Section 2, then one could take  $M(z_i, z_j) = \exp\{-D(z_i, z_j)\}$ . This would give a matched correlation in which "perfectly" matched pairs have full weight, "near" matches a little less weight, and so on. Computations would, of course, be more difficult, since simple counts of pairs of different types would no longer suffice, but the principles would remain valid.

## 6. REFERENCES

- [1] Blalock, Hubert M., Jr., Social Statistics. McGraw Hill Book Company, New York. 1960.
- [2] Cramér, H., Mathematical Methods of Statistics. Princeton University Press. 1951.
- [3] Davis, James A., "A partial coefficient for Goodman and Kruskal's gamma," Journal of the American Statistical Association, 62 (1967), 189-193.
- [4] Dixon, Wilfrid J., and Massey, Frank J., Jr., Introduction to Statistical Analysis. Second Edition, McGraw-Hill Book Company, New York. 1957.
- [5] Goodman, Leo A., "Partial tests for partial taus," Biometrika 46 (1959), 425-32.
- [6] Goodman, Leo A., and Kruskal, William H., "Measures of association for cross classifications," Journal of the American Statistical Association, 49 (1954), 723-64.
- [7] Goodman, Leo A., and Kruskal, William H., "Measures of association for cross classifications. III: Approximate sampling theory," Journal of the American Statistical Association, 58 (1963), 310-364.

- [8] Kendall, M. G., Rank Correlation Methods. Third Edition, Hafner Publishing Company, New York. 1962.
- [9] Kendall, M. G., and Stuart, Alan, The Advanced Theory of Statistics, Vol. 2, Hafner Publishing Company, New York. 1961.
- [10] [Pearson, Karl ?], "Remarks on Professor Steffensen's measure of contingency, Editorial," Biometrika 26 (1934) 255-60.
- [11] Rosenthal, Irene, "Distribution of the sample version of the measure of association, gamma," Journal of the American Statistical Association, 61 (1966) 440-453.
- [12] Sen, P. K., "On some convergence properties of U-statistics," Calcutta Statistical Association Bulletin, 10 (1960), 1-18.
- [13] Yule, G. Udny, An Introduction to the Theory of Statistics. Charles Griffin and Company, London. 1911.

## APPENDIX: THE RATIO OF TWO U-STATISTICS

Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed random variables (possibly multivariate) with cumulative distribution function  $F(x)$ . Consider two estimable parameters  $\theta_1$  and  $\theta_2$  with degrees  $m_1 \geq 1$  and  $m_2 \geq 1$  and symmetric unbiased estimators  $\phi_1(x_1, x_2, \dots, x_{m_1})$  and  $\phi_2(x_1, x_2, \dots, x_{m_2})$  respectively, so that

$$\theta_1 = \int \dots \int \phi_1(x_1, \dots, x_{m_1}) dF(x_1) \dots dF(x_{m_1})$$

and

$$\theta_2 = \int \dots \int \phi_2(x_1, \dots, x_{m_2}) dF(x_1) \dots dF(x_{m_2}).$$

Then the U-statistics for estimating  $\theta_1$  and  $\theta_2$  are

$$U_1 = \binom{n}{m_1}^{-1} \sum_{(1)} \phi_1(X_{c_1}, \dots, X_{c_{m_1}})$$

and

$$U_2 = \binom{n}{m_2}^{-1} \sum_{(2)} \phi_2(X_{c_1}, \dots, X_{c_{m_2}}),$$

where the summations (1) and (2) extend over  $1 \leq c_1 < \dots < c_{m_1} \leq n$  and  $1 \leq c_1 < \dots < c_{m_2} \leq n$  respectively.

It is by now well-known that if

$$\zeta_{11} = \int [\int \dots \int \phi_1(x_1, x_2, \dots, x_{m_1}) dF(x_2) \dots dF(x_{m_1})]^2 dF(x_1) - \theta_1^2,$$

$$\zeta_{12} = \int [\int \dots \int \phi_1(x_1, x_2, \dots, x_{m_1}) dF(x_2) \dots dF(x_{m_1})]$$

$$[\int \dots \int \phi_2(x_1, x_2, \dots, x_{m_2}) dF(x_2) \dots dF(x_{m_2})] dF(x_1) - \theta_1 \theta_2,$$

and

$$\zeta_{22} = \int [\int \dots \int \phi_2(x_1, x_2, \dots, x_{m_2}) dF(x_2) \dots dF(x_{m_2})]^2 dF(x_1) - \theta_2^2$$

are all convergent, then as  $n$  tends to infinity the asymptotic joint distribution of  $\sqrt{n}(U_1 - \theta_1)$  and  $\sqrt{n}(U_2 - \theta_2)$  is bivariate normal with zero means and variance matrix

$$\begin{bmatrix} m_1^2 \zeta_{11} & m_1 m_2 \zeta_{12} \\ m_1 m_2 \zeta_{12} & m_2^2 \zeta_{22} \end{bmatrix} ;$$

this includes the possibility of a degenerate normal distribution.

Now define also

$$V_1^{(i)} = \binom{n-1}{m_1-1}^{-1} \sum_{(1i)} \phi_1(X_i, X_{c_1}, \dots, X_{c_{m_2-1}})$$

and

$$V_2^{(i)} = \binom{n-1}{m_2-1}^{-1} \sum_{(2i)} \phi_2(X_i, X_{c_1}, \dots, X_{c_{m_2-1}})$$

for  $1 \leq i \leq n$ , where the summations (1i) and (2i) extend over  $1 \leq c_1 < \dots < c_{m_2-1} \leq n$  and  $1 \leq c_1 < \dots < c_{m_2-1} \leq n$  respectively, except that  $c_j = i$  must not occur for any  $j$ ; if  $m_k = 1$  then  $V_k^{(i)} = \phi_k(x_i)$ ,  $k=1,2$ . The  $V$ 's are what Sen [12] calls components of the U-statistics; note that

$$U_1 = \frac{1}{n} \sum_{i=1}^n V_1^{(i)} \quad \text{and} \quad U_2 = \frac{1}{n} \sum_{i=1}^n V_2^{(i)} .$$

Then Sen has shown that

$$s_{11} = \frac{1}{n-1} \sum_{i=1}^n (V_1^{(i)} - U_1)^2$$

and

$$s_{22} = \frac{1}{n-1} \sum_{i=1}^n (V_2^{(i)} - U_2)^2$$

converge in probability to  $\zeta_{11}$  and  $\zeta_{22}$  respectively; and it is a trivial extension of his results to show that

$$s_{12} = \frac{1}{n-1} \sum_{i=1}^n (v_1^{(i)} - U_1)(v_2^{(i)} - U_2)$$

converges in probability to  $\xi_{12}$  also. Finally, define

$$\sigma^2 = \frac{m_1^2 \theta_2^2 \xi_{11} - 2m_1 m_2 \theta_1 \theta_2 \xi_{12} + m_2^2 \theta_1^2 \xi_{22}}{\theta_2^4};$$

then we can obtain the following

Theorem. If  $U_1$  and  $U_2$  are two U-statistics as defined above, and if  $\theta_2 > 0$  and  $0 < \sigma^2 < \infty$ , then the random variable

$$\frac{\sqrt{n}}{s} \left( \frac{U_1}{U_2} - \frac{\theta_1}{\theta_2} \right)$$

converges in distribution to the standard normal, where

$$s^2 = \frac{m_1^2 U_2^2 s_{11} - 2m_1 m_2 U_1 U_2 s_{12} + m_2^2 U_1^2 s_{22}}{U_2^4}$$

Proof. As  $n$  increases to infinity,  $s^2$  converges in probability to  $\sigma^2$ ; hence by the convergence theorem of Cramér [2, p. 254] it follows that the desired asymptotic distribution must be the same as that of

$$\frac{\sqrt{n}}{\sigma} \left( \frac{U_1}{U_2} - \frac{\theta_1}{\theta_2} \right) = \frac{\sqrt{n}}{\sigma U_2 \theta_2} (U_1 \theta_2 - U_2 \theta_1).$$

And  $U_2$  converges in probability to  $\theta_2 > 0$ , so by the same theorem the asymptotic distribution must be the same as that of

$$\frac{\sqrt{n}}{\sigma \theta_2} (U_1 \theta_2 - U_2 \theta_1).$$

But, from the asymptotic joint normal distribution of  $U_1$  and  $U_2$ , it follows that  $\sqrt{n} (U_1\theta_2 - U_2\theta_1)$  is asymptotically normal with mean 0 and standard deviation precisely  $\sigma\theta_2^2$ ; hence the present theorem.

Remark: For any finite  $n$ ,  $U_2$  or  $s$  might vanish, with the ratio of the theorem then undefined. Thus in general the ratio may have no mean and variance.