stichting
mathematisch
centrum

MC

SA

AFDELING MATHEMATISCHE STATISTIEK SW 13/71 DECEMBER

D. QUADE

NONPARAMETRIC PARTIAL CORRELATION

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

NONPARAMETRIC PARTIAL CORRELATION [1]

Dana Quade

University of North Carolina, and Mathematical Center, Amsterdam

1. INTRODUCTION

It is often desired to measure the correlation between two variables, say X and Y, controlled for a third variable, say Z. Any such measure may be called a *partial correlation*, written C(X,Y|Z). Here C indicates *correlation* and | indicates *controlled for*; the varying interpretations of these two concepts form a basis for distinguishing among the indices proposed so far in the literature. In the next three sections I shall briefly review these concepts and the indices derived from them; afterwards I shall present a new index with examples and discussion.

2. CONCEPTS OF CORRELATION

Consider first correlation between X and Y, with Z being ignored at present. In general terms one may say that X and Y are positively correlated if there is a tendency for high values of X to occur together with high values of Y, and low values of X with low values of Y; they are negatively correlated if high values of X tend to occur with low values of Y, and low X with high Y. And correlation *per se* means either positive or negative correlation.

Quantitative indices of correlation C(X,Y) are generally standardized so that

(i) $-1 \leq C(X,Y) \leq 1$ or $0 \leq C(X,Y) \leq 1$,

where the values +1 and -1 are attainable in case of extreme or *perfect* positive or negative correlation; the second set of limits applies to those indices which do not distinguish between the two directions. In addition, correlation indices are ordinarily required to satisfy certain properties of symmetry, such as

(ii) $C(X,Y) = C(Y,X)$ and $C(X,Y) = -C(-X,Y) = -C(X,-Y) = C(-X,-Y)$.

Furthermore, it is considered desirable for them to have some form of invariance, meaning in general terms that if X and Y are separately transformed to new variables $X' = f(X)$ and $Y' = g(Y)$, where f and g are taken from a suitable class of functions, then

(iii) $C(X',Y') = C(X,Y)$.

In particular, *linear invariance* obtains if $C(X',Y') = C(X,Y)$ whenever $f(X) = a_X + b_X X$ and $g(Y) = a_Y + b_Y Y$ with both b_X and $b_Y > 0$. The more restrictive condition of *monotonic invariance*, which is required if the index is to be suitable for ordinal data, obtains if $C(X',Y') = C(X,Y)$ whenever f and g are both monotonic increasing functions.

The first and best-known index is undoubtedly the classical *product-moment correlation* of Pearson, which may be defined by the formula

$$\rho(X,Y) = \frac{\text{(covariance of X and Y)}}{\text{(standard deviation of X).(standard deviation of Y)}}.$$

It is difficult to provide any interpretation of this index unless X and Y

are both metric variables. In that case ρ measures the tendency of the population to be concentrated on a straight line; in fact ρ may well be called, as it has been by some authors, the coefficient of linear correlation. We have perfect positive (negative) correlation if the entire population lies exactly on a straight line of positive (negative) slope.

A rival concept of correlation, which requires no more than ordinal measurement of X and Y, is based on consideration of pairs of observations, for example (X_1,Y_1) and (X_2,Y_2). Such a pair is called *concordant* if $X_1 < X_2$ and $Y_1 < Y_2$ or if $X_1 > X_2$ and $Y_1 > Y_2$; that is, if the observation with the smaller value of X also has the smaller value of Y, and the one with the larger X has the larger Y, or, to put it another way, if the ordering of the pair is the same with respect to both variables. The pair is *discordant* if $X_1 < X_2$ and $Y_1 > Y_2$ or if $X_1 > X_2$ and $Y_1 < Y_2$; that is, if the observation with the smaller X has the larger Y, or if the ordering of the pair given by one variable is opposite to the ordering given by the other. If $X_1 = X_2$ or $Y_1 = Y_2$ or both then the pair is *tied*. Let p_C, p_D, and p_T be the respective probabilities that a randomly-chosen pair is concordant, discordant, or tied; $p_C + p_D + p_T = 1$. Then a possible index of correlation is *Kendall's tau-a* [9], defined as

$$\tau_a(X,Y) = p_C - p_D.$$

This may be interpreted as the difference between the probability that a random pair will be concordant and the probability that it will be discordant; we have perfect positive (negative) correlation if random pairs are concordant (discordant) with certainty, as is the case when the entire

population lies on some monotonically increasing (decreasing) curve. Note however that if ties can occur, as in particular is the case with categorized variables, then τ_a cannot reach the limits +1 and -1. Such an infelicity can be avoided by using a variation on the same theme, namely the *Goodman-Kruskal index* [6]

$$\gamma(X,Y) = \frac{P_C - P_D}{P_C + P_D} = \frac{P_C - P_D}{1 - P_T} .$$

This may be interpreted as the difference between the conditional probability that a random pair will be concordant, and the conditional probability that will be discordant, given that it is not tied; we now have perfect positive (negative) correlation if discordant (concordant) pairs are impossible, whether tied pairs are possible or not - unless tied pairs occur with certainty, in which case γ is undefined. Another well-known variant, *Kendall's tau-b* [9], may be defined as follows. Let p_{TX} be the probability that the random observations (X_1,Y_1) and (X_2,Y_2) will be such that $X_1 = X_2$; that is, the probability that the pair is tied on X, whether or not it is tied on Y. Similarly let p_{TY} be the probability of a tie on Y. Then

$$\tau_b(X,Y) = \frac{P_C - P_D}{\sqrt{1-p_{TX}}\sqrt{1-p_{TY}}} .$$

This index, though often used, has no simple interpretation; its advantages are more theoretical in nature. It is not difficult to see that τ_b always lies between τ_a and γ - usually it is very nearly halfway between them - so that $0 \leq \tau_a \leq \tau_b \leq \gamma \leq 1$ or $-1 \leq \gamma \leq \tau_b \leq \tau_a \leq 0$. Note that the only

difference among these indices, and other variants which appear in the literature but which I shall not discuss here, lies in their treatment of tied pairs: in fact, if $p_T = 0$ they all coincide.

A third basic concept of correlation does not view it as describing a population, but rather operationally as measuring the value of knowing something about one variable when one needs to know something about the other. For example, suppose we will be asked to guess the component Y of an observation (X,Y) taken at random, and that if our guess is Y_1 when the true value is Y_2 then we will suffer some non-negative *loss* $L(Y_1,Y_2)$. Consider two situations: (1) we will be given no further information before we must guess Y; and (2) we will first be told the value of X. Let R_1 and R_2 be the expected loss, or *risk*, in the two situations. Then an index of value of knowing X is the *proportional reduction in risk* of Situation 2 as compared with Situation 1, or

$$PRR(X,Y) = 1 - \frac{R_2}{R_1} .$$

Since clearly R_2 can be no greater than R_1 we have $0 \leq PRR \leq 1$; the value 1 is attained if $R_2 = 0$, that is if knowledge of X reduces the risk to zero, and PRR = 0 if $R_2 = R_1$, that is, if knowledge of X is of no value whatever for the purpose of guessing Y. Note that the direction of the correlation between X and Y is ignored, and indeed it is irrelevant; with this concept an index of association can be constructed even for variables X and Y which have no ordering. As just defined, PRR is not symmetric with respect to X and Y, but this can be corrected as follows. Suppose that we are equally likely to be asked to guess either Y or X, and that in Situation 1

we will be given no further information, but in Situation 2 before we must guess the one variable we will be told the value of the other; then just redefine R_1 and R_2, and hence PRR, in the obvious manner. In this generality the present concept was first formalized by Goodman and Kruskal [6, Section 7]; a somewhat less general version is well-known as *proportional reduction in error* or *PRE*. All the above may be further generalized, of course, by replacing the requirement to "guess Y" by more general situations then this simple prediction.

Many of the indices originally based on other concepts of correlation can also be given PRE interpretations. For example, the well-known PRE interpretation of the product-moment correlation proceeds as follows. Suppose we must guess Y when the loss will be equal to squared error. In Situation 1 the minimum risk is achieved by using the mean of Y, and is then equal to the variance, say $\sigma^2(Y)$; in Situation 2 it is achieved by using the conditional mean of Y given X, and is then equal to the conditional variance, say $\sigma^2(Y|X)$. The proportional reduction in risk is then equal to $1 - \sigma^2(Y|X)/\sigma^2(Y)$; but this is just $\rho^2(X,Y)$ if the conditional mean is a linear function of X. As another example, suppose two observations (X_1,Y_1) and (X_2,Y_2) are to be taken at random, and we will be required to guess whether $Y_1 < Y_2$ or $Y_1 > Y_2$. If we guess correctly we lose 0, and if incorrectly 1, except that if it should happen that $Y_1 = Y_2$ - but we are not permitted to make this our guess - then we lose the amount $\frac{1}{2}$. If we are given no information about X_1 and X_2 our risk is $R_1 = \frac{1}{2}$ no matter what strategy we adopt for guessing the ordering of Y_1 and Y_2; we may as well toss a coin. But if we are told the ordering of X_1 and X_2 then we may adopt the following minimum-risk guessing scheme: for $X_1 = X_2$ toss a coin anyway;

but otherwise guess the ordering of Y_1 and Y_2 so as to make the pair concordant (discordant) if τ_a is positive (negative). Then the risk turns out to be $R_2 = \min(p_C, p_D) + \frac{1}{2}p_T$, and the proportional reduction in risk is then PRR = $|\tau_a|$. For a PRE interpretation of Goodman and Kruskel's gamma see Costner [3], and for τ_b - this one being rather strained - see Wilson [19].

There are of course many other concepts of correlation, yielding for example the familiar Spearman's rho, various form of median and quadrant correlation, and more; but since these do not yet seem to have been used in measuring partial correlation I shall not treat them here. For further discussion see the papers by Goodman and Kruskal [6], [7], and [10].

3. CONCEPTS OF CONTROL

Let us now consider what it means to control for the variable Z. It seems possible to distinguish at least four different concepts in the literature, of which the most basic may be called *holding Z constant*. The usual technique here is to partition the population into strata within each of which Z is indeed constant, at least approximately. Then contingency tables are displayed, or summary parameters - in particular measures of correlation, which may then be called *conditional correlations* - are provided, for each of the strata. Of course, to reduce the variation in Z to a reasonable range often requires so many strata that the mind cannot comprehend them all, and some may occur with such small probability that with any realistic number of observations sampling variation will hide the relationships they should show. One way out, suggested by Rosenberg [14], is standardization: we might call the correlation in the standardized

population the partial correlation. Alternatively, we may define the partial correlation as an average conditional correlation. This approach was formalized by Goodman and Kruskal [6, Section 11], who applied it there to their coefficient λ, an index most appropriate when X and Y are purely nominal.

An important index obtained by holding Z constant is Davis' partial correlation coefficient based on Goodman and Kruskal's gamma [2]. Davis considers the case where X,Y, and Z are all categorized, so that the population might be displayed as a 3-way contingency table. Let p_i be the probability that a random observation will have the i-th value of Z, and let $p_{Ci}(p_{Di}, p_{Ti})$ be the probability that a random pair will be both tied on Z at its i-th value and also concordant (discordant, tied) with respect to X and Y, so that $p_{Ci} + p_{Di} + p_{Ti} = p_i^2$. Davis then defines his index of partial correlation as

$$\gamma(X,Y|Z) = \frac{\sum p_{Ci} - \sum p_{Di}}{\sum p_{Ci} + \sum p_{Di}},$$

where $\sum_i p_{Ci}$ ($\sum_i p_{Di}$) is the total probability that a random pair will be tied on Z and concordant (discordant) with respect to X and Y. Thus γ is the difference between the probability that a random pair tied on Z but not on X or Y will be concordant with respect to X and Y, and the probability that it will be discordant. But if we write γ_i for the conditional Goodman-Kruskal correlation between X and Y at the i-th level of Z, that is

$$\gamma_i = \frac{p_{Ci} - p_{Di}}{p_{Ci} + p_{Di}},$$

then we see that the partial correlation can be re-expressed as

$$\gamma(X,Y|Z) = \frac{\sum(p_{Ci}+p_{Di})\gamma_i}{\sum(p_{Ci}+p_{Di})} :$$

that is, γ is a weighted average of the conditional correlations in which the weight given to the i-th correlation is proportional to the probability that a random pair of observations will be tied on Z at its i-th value but not tied on X or Y.

When considered as a weighted average, Davis' γ may seem to use rather unusual weights. Goodman and Kruskal [6] suggested that it might seem natural to use weights proportional to the probabilities of the various levels of Z, the p_i's. Another reasonable approach might be to use equal weights for all strata. However, Davis' weights are somewhat simpler in this context, and are intuitively reasonable in view of his original definition. Furthermore, as he states, and as was verified empirically in considerable Monte Carlo experimentation by Reynolds [13], the three weighting schemes do not differ appreciably in typical research situations.

The second major concept of control may be called *adjusting for Z*. To do this we proceed as follows. Let f be a suitable function to use in predicting X from Z, in that the *residual* X' = X - f(Z) is concentrated about zero as closely as possible according to some reasonable criterion. Similarly let g be suitable for predicting Y from Z, with Y' = Y - g(Z) the corresponding residual. Then let the index of partial correlation C(X,Y|Z) = C(X',Y'), the total correlation between the residuals. (To correlate what might more properly be called *adjusted values*, obtained perhaps by adding the respective means of X and Y to X' and Y', is of course equi-

valent when the index being used is linearly invariant.) If in particular the criterion of concentration is variance, then f and g are the *regression functions*, the conditional means of X and Y given Z, and if product-moment correlation is used, we obtain the classical product-moment partial correlation coefficient, $\rho(X,Y|Z)$. If furthermore it happens that the regression functions are linear in Z, and that the conditional variances do not depend on Z, then the same result can also be obtained directly from the familiar *partial correlation formula*:

$$\rho(X,Y|Z) = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{1-\rho^2(X,Z)}\ \sqrt{1-\rho^2(Y,Z)}},$$

which is even used as a definition of partial correlation by some authors. The formula can of course be generalized in the well-known manner to allow for multiple and curvilinear regression. In principle it would seem possible to implement the concept of adjustment by using different methods of prediction, different indices of correlation, or both, but I have not yet seen any other partial correlation measures of this type in the literature.

Although there may be no problem in holding constant a categorical Z, Somers [18, p.972] claims that with a continuous Z methods derived from that concept "would be inapplicable, except by approximation, since each subgroup on the control variable would have no more than one observation". It must be admitted that in such a case the theoretical average conditional correlation generally cannot be estimated without some bias. However, it will be shown that this bias can be made negligible in practice, particularly by using the techniques proposed in Section 5. A more important objection is that a summary average may have no useful interpretation without the

assumption, so far implicit, that the conditional correlations being avaraged are not substantially different. For discussion of this point see Ploch [11]. The corresponding difficulty with methods based on the concept of adjustment, of course, is the need for structural assumptions, that is for knowledge of the functional form of the relationships of X and Y to Z. For example, the product-moment partial correlation as found from the simple formula is entirely inappropriate unless both X and Y have linear regressions on Z.

The two basic concepts of control discussed so far are often confused in the literature because of the importance of one special case in which they are entirely equivalent: when X, Y, and Z have a joint normal distribution. In such a population the conditional correlations between X and Y given Z = z are the same for every z, and hence also the same as their average. In addition, the conditions for applying the simple partial correlation formula hold, and the measure of partial correlation so obtained turns out to be identical with the constant conditional correlation. But since such a state of affairs obviously cannot be expected in general it would seem best always to keep clearly in mind what is to be meant by controlling for Z before attempting to choose an appropriate measure of measure of partial correlation.

A third concept of control is employed in constructing Kendall's [9] partial correlation coefficient. Suppose that X, Y, and Z are all at least ordinal, and to simplify matters assume for the moment that ties are impossible. Then any randomly-chosen pair of observations such as (X_1,Y_1,Z_1) and (X_2,Y_2,Z_2) can be classified as to whether it shows X and Y concordant or discordant with Z, the probabilities of these events being arranged as

in the following fourfold table:

		X and Z: Concordant	X and Z: Discordant
Y and Z	Concordant	p_0	p_X
	Discordant	p_Y	p_Z

Specifically, p_0 is the probability of the pair being *non-discordant*, meaning that $(X_1-X_2)(Z_1-Z_2) > 0$ and $(Y_1-Y_2)(Z_1-Z_2) > 0$, which imply that $(X_1-X_2)(Y_1-Y_2) > 0$ also; p_X is the probability that the pair is *X-discordant*, meaning that $(X_1-X_2)(Z_1-Z_2) < 0$ but $(Y_1-Y_2)(Z_1-Z_2) > 0$, and hence $(X_1-X_2)(Y_1-Y_2) < 0$; and similarly p_Y (p_Z) is the probability of the pair being *Y-discordant* (*Z-discordant*). Note that, with respect to X and Y, non-discordant and Z-discordant pairs are concordant, while X- and Y-discordant pairs are discordant; hence $p_C = p_0 + p_Z$, $p_D = p_X + p_Y$. According to Somers [18, p.974], Kendall [9] argues that if the non- and Z-discordant pairs predominate over the X- and Y-discordant pairs the partial correlation is positive, whereas if the X- and Y-discordant pairs predominate it is negative; and "if they are proportionately the same, then the partial is zero, that is, if the fourfold table exhibits statistical independence". This is because the non- and Z-discordant pairs show X and Y "rising and falling together, regardless of the change in the control variable", while the X- and Y-discordant pairs show X and Y "moving consistently in opposite directions, regardless, again, of the behavior of the control variable". Kendall proposed using as an index of partial correlation the well-known ϕ-coeffi-

cient computed from the table: that is

$$\phi(X,Y|Z) = \frac{p_0 p_Z - p_X p_Y}{\sqrt{(p_0+p_X)(p_Y+p_Z)(p_0+p_Y)(p_X+p_Z)}} .$$

He also achieved the surprising result that the same result is obtained if one substitutes the Kendall total correlations into the partial correlation formula! This apparent coincidence was first explained by Somers [17]. Further development by Hawkes [8], in terms of a formal "regression of pairwise differences", suggests that the partial correlation formula is still valid if ties can occur - and can even be extended to more than one control variable - provided the tau-b version is used for the total correlations; however Somers [18] prefers to discard the ties and calculate the ϕ-coefficient from only the untied pairs. A somewhat different line of development is pursued by Goodman [5]. It is difficult to describe his approach except in terms of a sample of observations, say (X_i, Y_i, Z_i) for $i = 1,2,\ldots,n$. Then for any positive integer $k < n$ a fourfold table can be constructed which classifies the $(n-k)$ pairs in which the ranks of the two values of Z differ by exactly k. The index of association in such a table may then be regarded as a partial index of partial association. Note that if k is very small then the table includes only pairs in which Z is approximately constant.

A final concept of controlling for Z extends to partial correlation the proportional reduction in error concept of total correlation. In general terms, suppose we will be asked to make a statement about Y, subject to specified losses in case of error, in one of two situations: (1) we will be given information about Z but not about X; (2) we will be given information

about both Z and X. Then the proportional reduction in expected loss for the second situation as compared with the first may be taken as a measure of partial correlation. Several of the indices of partial correlation presented previously can also be given such PRE interpretations. For example, the product-moment partial correlation can be obtained just as the total correlation, by specifying the loss as the squared error in predicting Y. Also, a PRE interpretation of the Davis coefficient follows from that of Goodman and Kruskal's gamma if the statement we must make is a prediction of the ordering on Y of two random observations having the same value of Z, where in Situation 1 we will be told only the common value of Z but in Situation 2 we will also be told the ordering on X. It should again be noted that the PRE concept cannot produce an index which distinguishes positive from negative correlation.

4. SAMPLE MEASURES AND SAMPLING THEORY

In the preceding I have defined indices of partial correlation strictly in terms of populations. The sample analogues I regard as so obvious that it is not worthwhile to write them down; suffice it to say that population moments are to be replaced by sample moments, that the probability of a pair having any given characteristic is to be replaced by the proportion of sample pairs having that characteristic, and that Greek letters in the notation are to be replaced by the corresponding Latin ones. At any rate, the necessary definitions can be found in the literature already cited.

An interesting little example presented by Somers [18] illustrates beautifully several of the concepts discussed above. Consider the sample

of 6 observations listed in Table 4.1. Holding Z constant immediately produces 3 subgroups of 2 observations each, namely (a and b), (c and d), and (e and f). Within each of these subgroups we see perfect positive correlation, and hence without further ado we put the partial correlation, viewed as average conditional correlation, equal to +1. Attempting to adjust for Z by means of linear regression - although actually Somers presented his example as involving strictly ordinal variables - we calculate total product-moment correlations $r(X,Z) = r(Y,Z) = 0$, and hence from the partial correlation formula $r(X,Y|Z) = r(X,Y) = 1/17$. However, plots of X and Y against Z suggest that linear regression is not appropriate; and on fitting quadratic functions instead we find perfect positive correlation between the residuals, in agreement with the previous result.

A complete listing of the 15 possible pairs of observations is given in Table 4.2. With respect to X and Z there are 4 concordant and 4 discordant pairs, and similarly with respect to Y and Z, so $t_b(X,Z) = t_b(Y,Z) = 0$. Hence if we use the standard formula to produce an index of partial correlation, as suggested by Hawkes, we will have $t_b(X,Y|Z) = t_b(X,Y)$. Now with respect to X and Y we find 5 concordant and 4 discordant pairs, and also 4 tied on X and 4 on Y (including 2 tied on both X and Y); thus the result is $(5-4)/(15-4) = 1/11$. Somers himself - and this is his example - constructs the fourfold table based only on pairs not tied on any variable, obtaining

		X and Z: Concordant	X and Z: Discordant
Y and Z	Concordant	1	2
	Discordant	2	1

and hence a phi-coefficient equal to -1/3. He remarks

Table 4.1

Observation Identification	X	Y	Z
a	1	2	1
b	2	3	1
c	2	1	2
d	3	2	2
e	1	2	3
f	2	3	3

Table 4.2

Pair	Classification with respect to:						
	One variable at a time			Two variables at a time			All three variables
	X	Y	Z	XY	XZ	YZ	XYZ
ab	U	U	T	C	T	T	T
ac	U	U	U	D	C	D	Y-discordant
ad	U	T	U	T	C	T	T
ae	T	T	U	T	T	T	T
af	U	U	U	C	C	C	non-discordant
bc	T	U	U	T	T	D	T
bd	U	U	U	D	C	D	Y-discordant
be	U	U	U	C	D	D	Z-discordant
bf	T	T	U	T	T	T	T
cd	U	U	T	C	T	T	T
ce	U	U	U	D	D	C	X-discordant
cf	T	U	U	T	T	C	T
de	U	T	U	T	D	T	T
df	U	U	U	D	D	C	X-discordant
ef	U	U	T	C	T	T	T

Note: C = Concordant, D = Discordant, T = Tied, U = Untied.

[18, p.976] that this is an "example from which most investigators, using subgroup analysis, would draw an erroneous conclusion". The reader will have to judge for himself. At any rate one may well agree with his further remark that "partial association among ordinal variables is not a simple notion that can be easily summarized in a single statistic".

So far I have considered only the descriptive and operational interpretations of measures of partial correlation, in light of the basic concepts involved; there now follow a few remarks on the sampling theory, for lack of which inference in this area is fraught with difficulties.

A statistic derived from holding Z constant, namely an average conditional correlation, is likely to be approximately normally distributed simply because it is an average. Hence only an estimate of variance is needed, and for this it suffices to have the first two moments of the conditional correlations. For example, Goodman and Kruskal [7] have given suitable estimates of the moments of their index gamma, and thence an approximation to the distribution of Davis' coefficient might be obtained. This line of thought does does not seem to have been pursued, however, and Davis [2] reports himself unable to obtain any sampling theory. (In the next section I shall derive the asymptotic sampling distribution of Davis' coefficient from a different approach.)

The distribution of the product-moment partial correlation coefficient is known exactly if the conditions for applying the simple partial correlation formula hold, and if in addition X and Y are conditionally jointly normal given Z. Although this last assumption is unnecessary for an asymptotic result, it is clear that strong assumptions are still required, and little is known of the effects of departures from them. For the Kendall-

type measures of partial correlation there are a few results in large-sample theory but nothing of any value in practice other than Goodman's [5] asymptotic "partial tests" for his partial indices and even those require fairly strong structural assumptions. All in all, for the indices of partial correlation in the literature the sampling theory is in a most unsatisfactory state.

5. PARTIAL CORRELATION BASED ON MATCHING

Suppose there is given a population of variables (X,Y,Z), where X and Y are at least ordinal, but Z is entirely without restriction - possibly nominal and/or multivariate - and an index of partial correlation between X and Y, controlled for Z, is desired. In this section I shall develop a general index based on the concept of correlation in terms of the concordance and discordance of pairs of observations, and on the concept of control in terms of holding Z constant or, more precisely, in terms of the notion of *matching*. Speaking intuitively, two observations are considered matched if their values of Z are "practically" equal. For what follows, however, it is sufficient if there has been established any specific rule whatever by which it can always be decided whether two observations are matched or not. Then an intuitively reasonable way of measuring the partial correlation, imitating the wording used by Goodman and Kruskal [6] in defining their correlation index gamma, is to find how much more probable it is to get like than unlike orders with respect to X and Y when pairs of observations matched on Z are chosen at random from the population.

More specifically, let MATCH be the event that a randomly-chosen pair

of observations will be matched, and let C (D) be the event that the pair will be concordant (discordant) with respect to X and Y. Assume without further ado that P{MATCH} > 0. Then I propose an *index of matched correlation*

$$\theta(X,Y|Z) = P\{C|MATCH\} - P\{D|MATCH\},$$

the difference between the conditional probabilities of concordance and discordance of a randomly-chosen pair of observations, given that the pair is matched. This index is standardized so that $-1 \leq \theta \leq 1$: $\theta = 1$ if $P\{C|MATCH\} = 1$, that is, if all matched pairs are necessarily concordant; and $\theta = -1$ if $P\{D|MATCH\} = 1$, that is, if all such pairs are discordant. And $\theta = 0$ if $P\{C|MATCH\} = P\{D|MATCH\}$; that is, if matched pairs are equally likely to be concordant or discordant.

Suppose the index θ is to be estimated from a random sample of n observations (X_i,Y_i,Z_i), $i = 1,2,\ldots,n$. Among the $N = n(n-1)/2$ possible pairs of observations let the number which are matched be N_M, and among these let the number concordant (discordant) with respect to X and Y be N_{CM} (N_{DM}). Then the obvious estimate of θ is

$$T(X,Y|Z) = \frac{N_{CM} - N_{DM}}{N_M},$$

the difference between the proportion of matched pairs in the sample which are concordant and the proportion discordant. (If it should happen that a sample contained no matched pairs then T might be arbitrarily set equal to zero.) We have $-1 \leq T \leq 1$ also, with $T = 1$ if the observed matched pairs

are all concordant, T = -1 if they are all discordant, and T = 0 if there are equally as many concordant as discordant matched pairs.

Let us now consider the sampling distribution of the index T. For each i, i = 1,2,...,n, let M_i be the number of observations (X_j,Y_j,Z_j), $j \neq i$, which are matched 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), less the number discordant. Then $\Sigma M_i = 2N_M$ - the factor 2 appears because each matched pair is counted twice - and $\Sigma W_i = 2(N_{CM}-N_{DM})$; hence

$$T(X,Y|Z) = \frac{\Sigma W_i}{\Sigma M_i} .$$

This method of computation leads to a convenient formula for the asymptotic standard error of T, namely

$$S(X,Y|Z) = \frac{2}{(\Sigma M_i)^2} \sqrt{\Sigma W_i^2 (\Sigma M_i)^2 - 2\Sigma W_i \Sigma M_i \Sigma W_i M_i + (\Sigma W_i)^2 \Sigma M_i^2}.$$

Furthermore, the sampling distribution of T is asymptotically normal: that is, for large n the quantity $(T-\theta)/S$ is approximately a standard normal variable. The only assumptions required for this, other than that sampling is random, are:

$$P\{MATCH\} > 0, \quad -1 < \theta < 1, \quad \sigma^2 > 0.$$

The parameter σ^2 is defined in the Appendix, where the proof of these results may be found; as is explained there, the possibility of σ^2 vanishing

seems remote for any real sitation. A corollary of these results is that T is always a consistent estimator of θ.

Thus statistical inference based on the index of matched correlation is possible, at least in large samples. Let Q_α be the critical value for a normal deviate Q, so that $P\{|Q| \geq Q_\alpha\} = \alpha$. Then, for example, a two-sided confidence coefficient $100(1-\alpha)\%$, is

$$T - SQ_\alpha \leq \theta \leq T + SQ_\alpha,$$

and the hypothesis H_0: $\theta = \theta_0$ can be rejected in favor of the alternative H_1: $\theta \neq \theta_0$ if and only if the value θ_0 lies outside this confidence interval. One-sided tests and confidence intervals can also be constructed, in an obvious manner.

As a special case the hypothesis that $\theta = 0$ might be rejected if and only if $|T/S| \geq Q_\alpha$. However, for this null hypothesis an alternative test involving only the W's and not the M's may be preferable: namely, reject if and only if

$$\frac{\sqrt{n}\,\bar{W}}{2\sqrt{\Sigma(W_i-\bar{W})^2}} \geq Q_\alpha$$

where $\bar{W} = \Sigma W_i/n$. (Do not neglect the factor 2 which may give this formula an unfamiliar look).

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:

$$\text{asymptotic variance of } \sqrt{n}(T-\theta) \leq \frac{2(1-\theta^2)}{P\{MATCH\}} .$$

Suppose matched pairs cannot be tied on X or Y, as for example when X and Y are continuous, or when ties are simply excluded in the definition of matching. Then the upper bound may be proved valid using Goodman and Kruskal's argument exactly, but interpreting their subscript "s" as indicating "concordant and matched", "d" as "discordant and matched", and "t" as "not matched". The bound is also easily shown to hold if $\theta = 0$ whether or not matched pairs can be tied. Unfortunately, I have not been able to prove it in the remaining case (matched tied pairs possible, $\theta \neq 0$), although obviously it must hold at least approximately if such ties are unlikely or θ is small.

One use for such a bound, as Goodman and Kruskal indicate, is to allow the possibility of "conservative" inference procedure in situations where use of an asymptotic standard error seems unjustified or its calculation is inconvenient. Then, for example, a "conservative" $100(1-\alpha)\%$ confidence interval for θ is formed by the set of values θ which satisfy the quadratic inequality

$$nN_M(T-\theta)^2 \leq 2N(1-\theta^2)Q_\alpha^2,$$

where the unknown value of P{MATCH} in the bound has been estimated by N_M/N. A second use for the bound is to show, at least qualitatively, how the variance of T decreases to 0 as θ approaches +1 or -1, and increases as the probability of obtaining a match decreases.

The index of matched correlation may be regarded as a somewhat generalized version of partial correlation in the sense of average conditional correlation. Suppose for simplicity that Z is a purely discrete random variable. For each possible value z of Z let E(z) be the event that two randomly-chosen observations (X_1,Y_1,Z_1) and (X_2,Y_2,Z_2) have $Z_1 = Z_2 = z$, that is, are tied on Z at z. Then the conditional correlation between X and Y given Z = z, as measured by Kendall's tau-a, is

$$\tau(X,Y|Z=z) = P\{C|E(z)\} - P\{D|E(z)\}.$$

Let us now construct an average conditional correlation, or partial correlation. With this form of correlation index, it seems reasonable to weight the conditional correlation at z in proportion to the probability of observing a pair tied at z. This yields

$$\tau(X,Y|Z) = \frac{\sum_z P\{E(z)\}\ \tau(X,Y|Z=z)}{\sum_z P\{E(z)\}}.$$

But, using the definition of $\tau(X,Y|Z=z)$ in terms of conditional probabilities,

$$P\{E(z)\}\ \tau(X,Y|Z=z) = P\{C \text{ and } E(z)\} - P\{D \text{ and } E(z)\}.$$

Let TIE be the event that the randomly-chosen pair of observations is tied on Z; that is, TIE is the union of all events E(z). Then the denominator of the last expression for $\tau(X,Y|Z)$ becomes P{TIE}, the numerator is

P{C and TIE} - P{D and TIE}, and hence

$$\tau(X,Y|Z) = P\{C|TIE\} - P\{D|TIE\}.$$

Thus we see that

$$\theta(X,Y|Z) = \tau(X,Y|Z) \qquad \text{if MATCH = TIE,}$$

that is, the index of matched correlation is a true partial correlation, in the sense of average conditional correlation, if two observations are defined as matched when their values of Z are equal.

If the probability function of the discrete random variable Z is h(z), so that $P\{E(z)\} = h^2(z)$, the partial correlation can be written

$$\tau(X,Y|Z) = \frac{\sum h^2(z)\ \tau(X,Y|Z=z)}{\sum h^2(z)}.$$

If instead we have a continuous Z, with density function h(z), then we may write the analogous expression

$$\tau(X,Y|Z) = \frac{\int h^2(z)\ \tau(X,Y|Z=z)\ dz}{\int h^2(z)\ dz},$$

where $\tau(X,Y|Z=z)$ is now the correlation within the conditional distribution of X and Y given Z = z; but expressions involving h^2 do not have such an intuitive interpretation. In addition, a random sample will now have no

tied pairs on which to base a sample estimate of τ. It is here that matching becomes invaluable: instead of demanding that all pairs used in the sample index of partial correlation be exactly tied, we relax the requirement to allow also pairs which are defined as matched though only "practically" tied. It is convenient to introduce here the word *tolerance* as a general term to indicate the maximum discrepancy allowed between two observations before they must be declared unmatched; for example, if MATCH = TIE then the tolerance is zero. Now the sample index T of matched correlation strictly estimates the population index θ, but in any real situation if the tolerance is small then θ will be essentially equivalent to τ. The two population indices will ordinarily not quite be identical in value, however. A particular example which should be noted is the case where X and Y are conditionally independent given Z = z for every z. This is sufficient -- although of course not necessary - to imply that each conditional correlation $\tau(X,Y|Z=z) = 0$, and hence that the partial correlation $\tau(X,Y|Z) = 0$ also; but it does not imply that $\theta(X,Y|Z) = 0$.

It should be clear that what I have been calling the "index" of matched correlation is really a whole family of indices, distinguished from one another by their definitions of matching. And there is indeed no restriction on the definition of matching, beyond the requirement that P{MATCH} > 0. In particular, by proper choice of the definition of matching - though it may then seem a bit unintuitive - one can produce several previously-proposed indices as special cases of the index of matched correlation. For example, suppose we define that "two observations are *always* matched". Then P{MATCH} = 1, and

$$\theta(X,Y|Z) = P\{C\} - P\{D\} = \tau_a(X,Y),$$

the total correlation between X and Y, as measured by Kendall's tau-a. In this case $M_i = n-1$ for $i = 1,2,\ldots,n$, and a little algebra will show that the sample index of matched correlation $T = t_a$, the corresponding sample index of total correlation. Furthermore, if C_i (D_i) is the number of observations concordant (discordant) with the observation (X_i,Y_i,Z_i), then the standard error of t_a is

$$S = \frac{2}{n(n-1)} \sqrt{\Sigma W_i^2 - (\Sigma W_i)^2/n},$$

where $W_i = C_i - D_i$ for $i = 1,2,\ldots,n$. As a second example, suppose we define that "two observations are matched if and only if they are not tied on X or on Y". Then the event MATCH is just the union of the events C and D, and

$$\theta(X,Y|Z) = \frac{P\{C\} - P\{D\}}{P\{C\} + P\{D\}} = \gamma(X,Y|Z),$$

the Goodman-Kruskal index of total correlation. Here $W_i = C_i - D_i$ again, and $M_i = C_i + D_i$, for $i = 1,2,\ldots,n$, and the standard error turns out to be

$$S = \frac{4}{(\Sigma C_i + \Sigma D_i)^2} \sqrt{\Sigma C_i^2 (\Sigma D_i)^2 - 2\Sigma C_i \Sigma D_i \Sigma C_i D_i + (\Sigma C_i)^2 \Sigma D_i^2}\ .$$

This expression may be compared with the maximum likelihood estimator of

asymptotic standard deviation given by Goodman and Kruskal [7]; the two are asymptotically equivalent. And as a final example, define that "two observations are matched if and only if they are tied on Z but not tied on X or on Y". With this definition the index of matched correlation becomes

$$\theta(X,Y|Z) = \frac{P\{C \text{ and } TIE\} - P\{D \text{ and } TIE\}}{P\{C \text{ and } TIE\} + P\{D \text{ and } Tie\}} = \gamma(X,Y|Z),$$

Davis' partial coefficient for Goodman and Kruskal's gamma. If C_i (D_i) is now redefined as the number of observations which are concordant (discordant) with the observation (X_i,Y_i,Z_i) with respect to X and Y and also tied with it on Z, then $W_i = C_i - D_i$ and $M_i = C_i + D_i$ just as for the total coefficient, and the asymptotic standard error of the partial coefficient has the same formula. Thus we see how statistical inference with Davis' coefficient is possible also.

6. EXAMPLES

The first example, which will illustrate the method of computation in some detail, is based on the data of Table 6.1. Let X be the examination result, an ordinal variable recorded as A, B, C, D, or F; and let Y be the metric variable height, recorded in inches. The variable to be controlled for is a bivariate Z of which the first component is the nominal variable sex (Z_1) and the second component is IQ (Z_2).

The sample index of matched correlation between examination result and height controlled for sex and IQ, that is, between X and Y given both Z_1 and Z_2, is obtained using the values of M_i and W_i shown in the last section

of the table. In this computation 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, for instance, 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. One may then compute $\Sigma M_i = 96$, indicating that there are 48 matched pairs of children, and $\Sigma W_i = 10$, indicating that there are 5 more concordant pairs than discordant; hence the index is $T = \Sigma W_i / \Sigma M_i =$ $= 10/96 = .104$. (Of the matched pairs, actually 22 are concordant, 17 discordant, and 9 tied; without modification, however, the computational scheme here presented does not provide these numbers.) Having calculated $\Sigma M_i^2 = 422$, $\Sigma M_i W_i = 50$, and $\Sigma W_i^2 = 90$, one also finds $S = .191$. Thus the index is smaller than its standard error and certainly not significantly different from zero in the statistical sense. If this sample could be regarded as large, one could take $T/S = .545$ as a normal deviate in making such a test, and could also produce the 95% (say) confidence interval $T \pm 1.96S$, or $(-.270, +.469)$, for the population index θ. 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 6.1, labeled "without matching", shows the components for the index of total correlation, which can be obtained by defining that all pairs are matched, so that $M_i = n-1 = 24$ for all i. We then have 300 matched pairs, of which there are 21 more concordant than

Table 6.1

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

	Result of exam.	Height (in.)	Sex	IQ	Without matching		Matching on sex only		Matching on IQ* only		Matching on sex and IQ*	
i	X	Y	Z_1	Z_2	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

* within a tolerance of 10 units.

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 the correlation is not significant.

The other two sections of Table 6.1 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 6.2. Note that the two indices of conditional correlation given sex are obtainable almost as byproducts of the computation for the index of matched (or, in this case, partial) correlation given sex: to obtain the conditional correlation among males, 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 do the reverse. The values of ΣW_i, ΣM_i, ΣW_i^2, ΣM_i^2, and $\Sigma W_i M_i$ for the matched correlation indices are equal to the sums of the corresponding values for the two conditional correlation indices. A similar situation will obtain whenever the variable being controlled for is discrete.

Now let us consider an example in which the underlying population distribution is known. For i = 1,2,...,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 i = 51, 52,...,100 continue with columns 11, 12, and 13. Define $X_i = C_{1i} + C_{3i}$, $Y_i = C_{2i} + C_{3i}$, $Z_i = C_{3i}$, for i = 1,2,...,100. Then simple considerations show that the population total (product-moment) correlation is $\rho(X,Y) = \frac{1}{2}$, with partial correlation $\rho(X,Y|Z) = 0$; the corresponding sample values happen to be r(X,Y) = .533,

Table 6.2

Correlation	Pair is matched if:	ΣM_i	ΣW_i	ΣM_i^2	$\Sigma M_i W_i$	ΣW_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	$Z_{1i}=Z_{1j}$="male"	156	30	1872	360	374	.192	.224
Conditional on female	$Z_{1i}=Z_{1j}$="female"	132	0	1452	0	226	.000	.228
Matched on IQ*	$\lvert Z_{2i}-Z_{2j}\rvert \leq 10$	198	10	1744	88	178	.051	.133
Matched on sex and IQ*	$Z_{1i}=Z_{1j}$ and $\lvert Z_{2i}-Z_{2j}\rvert \leq 10$	96	10	422	50	90	.104	.191

* within a tolerance of 10 units.

Table 6.3

Tolerance for matching on Z* ϵ	Number of Matched pairs N_M	Population index θ	Sample index T	Standard error S
∞	4950	.333	.363	.053
3.00	4804	.311	.343	.052
2.00	4164	.237	.261	.053
1.50	3486	.168	.179	.055
1.00	2514	.090	.086	.060
.75	1942	.055	.049	.063
.50	1308	.026	.040	.069
.25	674	.007	-.003	.078
0	0	.000	--	--

* The true standard deviation of Z is $\sigma_Z = 1.000$, with $s_Z = .997$.

$r(X,Y|Z) = -.009$. In a normal population

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

hence $\tau(X,Y) = 1/3$, $\tau(X,Y|Z) = 0$. If two observations (X_i,Y_i,Z_i) and (X_j,Y_j,Z_j) are defined to be matched if and only if $|Z_i-Z_j| \leq \varepsilon$, then with X, Y, and Z as specified above we have

$$\tau(X,Y|Z) = \tau(\varepsilon) = \frac{1}{3} P\{|Q| \leq \frac{\varepsilon}{\sqrt{2}}\}^2,$$

where Q is a normal (0,1) variable. (The proof of this may be found at the end of the Appendix.) Note that $\tau(\varepsilon)$ decreases steadily from 1/3 to 0 as decreases from $+\infty$ to 0. Results of computing the sample index of matched correlation for decreasing values of the tolerance ε are shown in Table 6.3, and these show a similar steady decrease. Note also 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 5; but the increase is not drastic until a very small tolerance has been reached.

Three examples will now be presented, using previously published data, in which the index of matched correlation may be compared with other measures. Consider first the example originally presented by Yule [20] and extensively quoted since, in which X is the estimated average earnings of agricultural laborers, Y is the ratio of the number of paupers receiving "outdoor" relief to the number receiving relief in the workhouse, and Z is the percentage of population on relief, for n = 38 rural districts. The

product-moment total correlations are $r(X,Y) = -.13$, $r(X,Z) = -.66$, and $r(Y,Z) = +.60$, so that the partial correlation formula gives $r(X,Y|Z) = +.44$. The results of computing the index of matched correlation, summarized in Table 6.4, show a similar relationship.

A second example uses 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 hetorogeneity "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 $r(X,Y) = -.156$, $r(X,Z) = -.456$, $r(Y,Z) = -.513$, with partial correlation $r(X,Y\ Z) = -.511$. Results for the index of matched correlation are summarized in Table 6.5, 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 6.4

Tolerance for matching on Z* ϵ	Number of matched pairs N_M	Index of matched correlation T	Standard error S
∞	703	-.078	.096
2.00	500	.136	.089
1.50	393	.226	.092
1.00	269	.294	.107
.50	142	.331	.115

* The standard error of Z is $s_Z = 1.29$.

Table 6.5

Tolerance for matching on Z* ϵ	Number of matched pairs N_M	Index of matched correlation T	Standard error S
∞	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

* The standard error of Z is $s_Z = 9.66$.

The last example uses the data of Hajda quoted by Davis [2], which were obtained from a sample survey of Baltimore women. Here X is a dichotomy, taking the value "high" ("low") if the respondent was above (below) 45 years of age; Y is another dichotomy, taking the value "high" ("low") if she had (had not) recently read a book; and Z distinguishes three categories of educational attainment, "college", "high school", and "less than high school". Two definitions of matching will be considered: the first, producing a straightforward partial correlation coefficient, declares two observations matched if they are tied on Z; whereas the second, producing Davis' partial coefficient for Goodman and Kruskal's gamma, declares them matched only if they are both tied on Z and also not tied on X or Y. Since it may be instructive to follow the calculations for a problem involving categorical data, Table 6.6 shows them in some detail. There are listed the 12 possible values of (X,Y,Z), and the frequency with which each occurs in the sample, labeled F. Then are shown how many observations are both tied on Z and concordant (discordant, tied) with respect to X and Y with each of the observations at a given value, labeled C (D,T). We have $W = C - D$; for the first definition of matching, $M_1 = C + D + T$, and for the second, $M_2 = C + D$. In either case

$$T = \frac{\Sigma FW}{\Sigma FM}$$

and

$$S = \frac{2}{(\Sigma FM)^2}\sqrt{\Sigma FM^2(\Sigma FW)^2 - 2\Sigma FM\Sigma FW\Sigma FMW + (\Sigma FM)^2\Sigma FW^2}\ .$$

Table 6.6

Age X	Book Reading Y	Education Z	Frequency F	C	D	T	W	M_1	M_2
High	High	College	104	46	0	302	46	348	46
	Low		36	0	163	185	-163	348	163
Low	High		163	0	36	312	-36	348	36
	Low		46	104	0	244	104	348	104
High	High	High School	159	327	0	627	327	954	327
	Low		179	0	290	664	-290	954	290
Low	High		290	0	179	775	-179	954	179
	Low		327	159	0	795	159	954	159
High	High	Less than High School	54	133	0	412	133	545	133
	Low		335	0	24	521	-24	545	24
Low	High		24	0	315	210	-335	545	335
	Low		133	54	0	491	54	545	54

In the second case the equivalent formula for S in terms of C's and D's given at the end of section 5 might also be used; in grouped-data form it is

$$S = \frac{4}{(\Sigma FC+\Sigma FD)^2}\sqrt{\Sigma FC^2(\Sigma FD)^2-2\Sigma FC\Sigma FD\Sigma FCD+(\Sigma FC)^2\Sigma FD^2}\ .$$

For the example, $\Sigma FW = -3718$ and $\Sigma FW^2 = 55729114$. For the first definition of matching, $\Sigma FM = 1330092$, $\Sigma FM^2 = 1073601726$, and $\Sigma FMW = -1531320$, yielding partial correlation $T = -.0028$ with $S = .0112$. For the second definition, $\Sigma FM = 259554$, $\Sigma FM^2 = \Sigma FW^2$ (this equality would hold whenever X and Y are both dichotomous, but not in general), and $\Sigma FMW = -1070650$, yielding Davis' coefficient $T = -.0143$ with $S = .0581$. Note that both Davis' coefficient and its standard error are about five times larger than when ties on X and Y are retained rather than discarded, and the level of significance for testing the null hypothesis of no partial correlation is thus about the same. The alternative test for this null hypothesis would of course be identical for the two definitions of matching, since it depends only on the W's. By the way, it is obvious that a number of shortcuts could have been taken in the calculations for this rather simple example; a general computer program, however, would probably best proceed from the formulas as given.

In this same example the total correlation between age (X) and book reading (Y) is $-.0596$ as measured by t_a, or $-.2412$ as measured by G, and this is significantly different from zero at $\alpha < .01$; thus holding education constant has reduced the correlation by substantially more than 90%, to a clearly insignificant value. On the other hand, using the partial correlation

formula with t_b as suggested by Hawkes, we calculate $t_b(X,Y) = -.1206$, $t_b(X,Z) = -.2394$, and $t_b(Y,Z) = .4139$, and hence $t_b(X,Y|Z) = -.0243$, for a reduction of only 80%. And if we adopt Somers' method, we have the four-fold table

		X and Z	
		Concordant	Discordant
Y and Z	Concordant	68987	180932
	Discordant	15600	27456

from which $\phi = -.0674$; this again illustrates the difference in results which can arise from different concepts of control.

7. DISCUSSION

Since the sampling theory presented above is strictly asymptotic, you may well ask for the distribution of the index of matched correlation in small samples, or at least for the proper definition of "small" in this context. I can give no really satisfactory answer at this stage, but offer the following speculation. The general index T of matched correlation has the same form as its special case, the Goodman and Kruskal index G, in that it is a ratio of which the denominator is the number of sample pairs falling in a specified class and the numerator is the difference between the numbers of pairs in two subclasses of that class. It seem reasonable that the validity of asymptotic methods in finite samples may depend most directly on the total number of pairs observed in the special class, which

for T consists of pairs which are matched, while for G it consists of pairs which are not tied on X or on Y. Fairly extensive sampling experiments by Rosenthal [15] for 5×5 cross-classifications over a wide range of true values of γ showed the distribution of $(G-\gamma)/s$, where s^2 is the maximum likelihood estimator of the asymptotic variance of G, to be reasonably close to the standard normal in samples of n = 25 or 50 for $|\gamma| < .50$. The probability of a tie in a 5×5 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 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. One may then speculate that similar results would be obtained for indices of matched correlation based on numbers of matched pairs in that range. There was a tendency for s^2 to underestimate the variance of G, particularly for larger values of G. Very possibly S^2 tends to underestimate the variance of T also: for instance, S = 0 if the matched pairs are all concordant, all discordant, or all tied, and this is not unlikely in very small samples. Other estimates are, of course, possible: the one I have used, based on the work of Sen [16], was chosen almost entirely on the basis of its simplicity.

You may also ask for guidelines in choosing the definition of matching. Now, in the preceding I have implicitly assumed that such a definition is to be based on substantive considerations, and one might take the attitude that this is not really a statistical question at all. Yet I may still offer some remarks, particularly for the case where ties on Z are rare or nonexistent. (If ties are common then the simple definition that MATCH = TIE should nearly always be satisfactory.)

Suppose the immediate goal is to estimate the partial correlation as defined in Section 5. If the tolerance were infinite, so that all pairs were considered matched, then the matched correlation would be equivalent to the total correlation. As the tolerance decreases to zero, the population matched correlation approaches the partial correlation; but in a sample the number of matched pairs decreases also, leaving a smaller and smaller basis for the estimate, whose variance accordingly increases. Thus the optimal tolerance for estimating a partial correlation is a compromise: a large value may have too much bias, a small value too much variance. Presumably the investigator will first propose a definition of matching based on totally non-statistical substantive grounds. If this definition implies too few matched pairs, say less than 200, a relaxation might be suggested to make the asymptotic theory more tenable. And if the proposed definition implies a very large number of matched pairs, say more than 1000, it might be tightened to reduce possible bias. On the other hand, it might well be in practice that the easily-understood population index of matched correlation would be accepted as the proper object of interest in itself, regardless of whether it equalled the somewhat abstract index of partial correlation; then presumably the statistician should comment only on the sample size and not on the definition of matching itself.

A related question of interest to the mathematical statistician is this: what happens asymptotically as the sample size increases if the definition of matching is simultaneously tightened? Presumably a consistent estimator of the partial correlation could be obtained in this manner, but the theory has not yet been worked out. Similar questions arise if the definition of matching is made relative rather than absolute: for instance,

one might decide to pair each observation with the k others most closely matched to it, or simply to use the K most closely matched pairs out of the total N. Such a decision would confer the advantage of making the number of matched pairs fixed instead of random; but again the theory is not available.

It may be also noted, with respect to this last point, that both practical and theoretical problems are raised in attempting to order pairs of observations according to closeness of matching. In general this requires defining a sort of *distance* function - though it need not have all the properties which mathematicians imply by use of that term - to measure the discrepancy between any two points in the sample space. Where such a function can be defined, however, a generalization of the concept of matching is possible. Specifically, let $D((X_1,Y_1,Z_1),(X_2,Y_2,Z_2))$ be the distance, or discrepancy, between any two observations (X_1,Y_1,Z_1) and (X_2,Y_2,Z_2). Then for example one might give to a pair of observations with discrepancy D the weight $f(d) = 1/D$ or $\exp(-D)$, say. Define

$$M_i = \sum_{j \neq i} f(D((X_i,Y_i,Z_i),(X_j,Y_j,Z_j))$$

for $i = 1,2,\ldots,n$, and W_i similarly as the difference between weighted sums of concordant and discordant pairs, and hence a generalized index $T = \Sigma W_i / \Sigma M_i$. It is not difficult to show that the Theorem of the Appendix applies for such generalized indices also, and thus that the entire asymptotic theory is still valid.

Even if such a generalized index is not contemplated, a distance function may still be extremely useful in practice. For example, suppose we have a vector-valued $\underline{Z} = (Z^{(1)},Z^{(2)},\ldots,Z^{(m)})'$, and we want to balance off

discrepancies due to the various components of Z. This can be accomplished conveniently by declaring two observations matched if and only if the distance between them is no greater than some fixed amount, which of course corresponds to what has been called the tolerance. Then two observations may be called matched if they differ little on the average, though they may differ more on some components if they are particularly close on others. It is often suitable to use distance functions of the form

$$D((X_i,Y_i,\underline{Z}_i),(X_j,Y_j,\underline{Z}_j)) = (\underline{Z}_i-\underline{Z}_j)' \; W(\underline{Z}_i-\underline{Z}_j)$$

where W is a matrix of weights. If $W = V^{-1}$, where V is the sample variance matrix of Z, then D becomes the *Mahalanobis distance* which gives each component of Z equal importance in determining a match.

It may also be useful to point out, by the way, 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.

Let us now move on to the troublesome question of ties. In constructing an index of correlation based on the notion of concordant and discordant pairs, should tied pairs be included or not? The index of matched correlation is defined in sufficient generality that one may have it either way, and the choice is to some extent a matter of taste. For one thing, exclusion of tied pairs often simplifies computations, at least hand computations,

if only by diminishing the numbers involved; and the sample indices which it produces may be more satisfying since they are greater in magnitude. On the other hand, the theory is perhaps more elegant when matching can be defined in terms of the control variables alone.

But suppose we take the point of view that X and Y, even if not continuous as recorded, usually represent underlying continuous variables, with ties occurring only because the measurements are imprecise or because they are grouped afterwards. Then it is a reasonable goal to determine the correlation in the underlying continuous population. (For simplicity, consider only the total correlation at present.) We may begin by asking, how many of the tied pairs we observe would be found concordant, and how many discordant, if they could be properly resolved? This requires guessing the correlations within subpopulations where the range of X and/or Y is restricted. One would ordinarily expect such subcorrelations to be smaller, on the whole, than the correlation for unrestricted X and Y. As an extreme case, set them all equal to zero. This can be effectively accomplished in the sample by adjusting the data so that half the tied pairs are counted as concordant and half discordant. Then

$$N_C' = N_C + \tfrac{1}{2}N_T, \quad N_D' = N_D + \tfrac{1}{2}N_T,$$

where N_C (N_D,N_T) is the number of concordant (discordant, tied) pairs in the data as recorded, and N_C' (N_D') is the number of concordant (discordant) pairs after adjustment. And the adjusted correlation index is

$$t' = \frac{N_C' - N_D'}{N} = \frac{N_C - N_D}{N} = t_a$$

(the denominator for the adjusted data is unequivocally N, since there are now no ties); that is we get the same result as if we had calculated t_a from the original data. As a second extreme case, set the subcorrelations equal to the total correlation. In the sample this amounts to allocating the tied pairs in the same proportions as the untied ones. Then

$$N' = N_C + \frac{N_C}{N_C + N_D} N_T, \quad N'_D = N_D + \frac{N_D}{N_C + N_D} N_T,$$

and the adjusted correlation index is

$$t' = \frac{N'_C - N'_D}{N} = \frac{N_C - N_D}{N_C + N_D} = G;$$

that is, the result is now the same as if we had calculated G from the original data. In general, those measures which include ties may be regarded as conservative, or pessimistic, since they tend to underestimate the strength of any underlying correlation; whereas those which discard ties are optimistic, tending to overestimate its strength. Probably in most contexts underestimation would be preferable to overestimation, thus suggesting that tied pairs be retained.

On the other hand, at least for the total correlation it is possible to compromise, by accepting the index t_b which always lies between the other two. My personal impression, admittedly based on a rather limited number of examples, is that the correlation τ of an underlying continuous population almost always lies between the τ_a and γ of the modified population determined by imposing some grouping on it. Thus t_b may well be a good overall estimator for realistic cases, since t_a tends to declare the correlation somewhat

too weak and G makes it much too strong, although peculiar populations can be invented to favor any of the three indices. It might also be mentioned that considerable numerical work by Reynolds [13] also suggests that G is inferior to t_a and especially t_b, for a somewhat different purpose but perhaps for the same reasons. Of course, t_b is more difficult to interpret in terms of the measurements actually at hand, and it is certainly much more difficult to work with both numerically and theoretically.

Also, t_b is not in general a special case of the index of matched correlation. However, the following suggestion may be made. Consider a modified index

$$T = \frac{N_{CM} - N_{DM}}{N_{CM} + N_{DM} + \frac{1}{2}(N_{XM}+N_{YM})} ,$$

where N_{CM} and N_{DM} are as defined in Section 5, and N_{XM} (N_{YM}) is the number of matched pairs which are tied on X but not Y (Y but not X). Equivalently, in the alternative computational scheme for T, replace M_i by the number of observations which are matched with the observation (X_i,Y_i,Z_i) and are either concordant or discordant with it, plus half the number of matched observations which are tied with it on X but not Y, or on Y but not X, for $i = 1,2,\ldots,n$; leave W_i unchanged. The asymptotic results then hold without further modification. If all pairs are considered matched, this proposal yields the total correlation index t_b if the number of pairs tied on X equals the number tied on Y; otherwise it gives a value between t_b and t_a but ordinarily very close to t_b. Further work will be required, of course, for a complete evaluation of this proposal.

A few remarks may be made with respect to computational matters. It is perhaps a disadvantage that the calculation of an index of matched correlation must always begin from scratch, since there is no formula by which one of these indices can be determined from others previously found. Yet the partial correlation formula is sometimes deceptively easy, since its numerical instability in the presence of highly correlated variables is not always obvious. This is not so with matched correlations, where any instability is always clearly indicated, if not by the asymptotic variance formula, then certainly by a paucity of matched pairs. Of course, any statistic which requires individual consideration of all pairs of observations is tedious to calculate; even on the computer, although a matched correlation program maybe simple and short, the time it requires may be long. This computational problem can be avoided by grouping the data, but unfortunately the resulting ties reduce the precision of the estimate. For large numbers of observations it may be preferable to consider only a sample of the possible pairs; but inference procedures would have to be modified accordingly.

In review, let me summarize the comparison between matched correlation, as an index of the partial correlation between X and Y given Z, and its major competitors. Since Davis' coefficient is not a competitor but is instead a special case of matched correlation, the main rivals would appear to be the product-moment partial correlation and the Kendall-Somers Hawkes measures. Of these the former is inapplicable, or at any rate difficult to interpret, unless X and Y are metric variables; the latter require Z to be at least ordinal. Small sample theory for the product-moment

partial correlation is available, but only under strong assumptions including normality, and even for asymptotic results the form of relationship of X and Y to Z must be known; for the Kendall-type measures sampling theory is practically non-existent. On the other hand, the proposed new index has the following clear advantages:

1) The *applicability* of matched correlation is almost unlimited. It may be used to control for a completely arbitrary variable Z, even a multivariate Z in which each component separately may be metric, ordinal, or purely nominal, provided only that a definition of matching can be supplied. And the variables X and Y need be no more than ordinal, including ordered-categorical.

2) The *interpretation* of matched correlation is based on two very simple concepts: determination as to whether two observations are matched or not, and as to whether they are concordant or discordant with respect to X and Y. The index may then be defined as the probability that a randomly-chosen matched pair will be concordant, less the probability that it will be discordant. (This definition applies to the sample index also, if we think of choosing two observations from the sample, at random and without replacement.)

3) Asymptotic *sampling theory* for matched correlation indices is available, without restrictive assumptions, and hence statistical inference is possible at least in large samples.

REFERENCES

[1] Hubert M. Blalock Jr., *Social Statistics*, McGraw-Hill Book Company, New York. (1960).

[2] James A. Davis, "A partial coefficient for Goodman and Kruskal's gamma", *Journal of the American Statistical Association*, 62 (1967), 189-193.

[3] Herbert L. Costner, "Criteria for measures of association", *American Sociological Review*, 30 (1965), 341-353.

[4] Wilfrid J. Dixon and Frank J. Massey Jr., *Introduction to Statistical Analysis*, Second Edition, McGraw-Hill Book Company, New York. (1957).

[5] Leo A. Goodman, "Partial tests for partial taus", *Biometrika*, 46 (1959), 425-432.

[6] Leo A. Goodman and William H. Kruskal, "Measures of association for cross classifications", *Journal of the American Statistical Association*, 49 (1954), 723-764.

[7] Leo A. Goodman and William H. Kruskal, "Measures of association for cross classifications. III: Approximate sampling theory", *Journal of the American Statistical Association*, 58 (1963), 310-364.

[8] Roland K. Hawkes, "The multivariate analysis of ordinal measures", *American Journal of Sociology*, 76 (1971), 908-926.

[9] M.G. Kendall,,*Rank Correlation Methods*, Third Edition, Hafner Publishing Company, New York. (1962).

[10] William H. Kruskal, "Ordinal measures of association", *Journal of the American Statistical Association*, 53 (1958), 814-861.

[11] Donald R. Ploch, "An interaction test for Goodman and Kruskal's gamma", unpublished paper, Yale University (1969).

[12] M.L. Puri and P.K. Sen, *Nonparametric Methods in Multivariate Analysis*, John Wiley and Sons, Inc., New York. (1971).

[13] H.T. Reynolds, *Making Causal Inferences with Ordinal Data*, Institute for Research in Social Science, University of North Carolina, Chapel Hill. (1971).

[14] Morris Rosenberg, "Test factor standardization as a method of interpretation", *Social Forces*, 41 (1962), 54- .

[15] Irene Rosenthal, "Distribution of the sample version of the measure of association, gamma", *Journal of the American Statistical Association*, 61 (1966), 440-453.

[16] P.K. Sen, "On some convergence properties of U-statistics", *Calcutta Statistical Association Bulletin*, 10 (1960), 1-18.

[17] Robert H. Somers, "The rank analogue of product-moment partial correlation and regression, with application to manifold, ordered contingency tables", *Biometrika*, 46 (1959), 241-246.

[18] Robert H. Somers, "An approach to the multivariate analysis of ordinal data", *American Sociological Review*, 33 (1966), 971-977.

[19] Thomas P. Wilson, "A proportional-reduction-in-error interpretation for Kendall's tau-b", *Social Forces*, 47 (1969), 340-342.

[20] G. Udny Yule, *An Introduction to the Theory of Statistics*, Charles Griffin and Company, London. (1911).

(FOOTNOTE)

[1] Supported by National Institutes of Health Grants No. GM-12868 and GM-38906.

APPENDIX

Consider estimating the value of a parameter ω in the distribution of some random variable Q, possibly multivariate. Define the *degree* of ω as the size of the smallest random sample from which ω can be estimated with no bias and with finite variance. (If no such estimator can be found for a any sample size then we may say that the degree of ω is infinite.) For example, suppose we have a normal population with mean μ and variance σ^2. If the parameter of interest is $\omega = \mu$, then the degree is 1 and the corresponding unbiased estimator, which we call the *kernel*, is $w(Q_1) = Q_1$. If the parameter of interest were $\omega = \sigma^2$, with μ known, then the degree would be 1 again, and the kernel $w(Q_1) = (Q_1-\mu)^2$. But with μ not known, we require at least two observations, that is, the degree is 2. Now there are infinitely many possible estimators: two candiates are $w_1 = (Q_1-Q_2)^2/2$ and $w_2 = Q_1^2 - Q_1Q_2$. In such situations we reject those estimators, such as w_2, which depend on the ordering of the observations, and define the kernel as the symmetric estimator, here w_1. This estimator in general is unique and has minimum variance.

Now suppose that the parameter of interest is ω, of degree k, with kernel $w(Q_1,Q_2,\ldots,Q_k)$, and that we have a random sample of size $n > k$. Calculate w for every subset of k observations out of the n available; take the sum; and divide by the number of such subsets, namely $\binom{n}{k} = n!/k!(n-k)!$, the number of combinations of n things taken k at a time. The resulting average is called the *U-statistic* for estimating ω; in symbols,

$$W = \sum_{C} w(Q_{i_1}, Q_{i_2}, \ldots, Q_{i_k}) / \binom{n}{k}.$$

For example, if ω is the population mean, with degree 1 and kernel $w(Q_1) = Q_1$, then

$$W = \sum_{C} w(Q_{i_1}) / \binom{n}{1} = \sum Q_i / n = \bar{Q},$$

the sample mean. And if ω is the population variance, where the mean is unknown, then

$$W = \sum_{C} \tfrac{1}{2}(Q_{i_1} - Q_{i_2})^2 / \binom{n}{2},$$

where $\binom{n}{2} = n(n-1)/2$. This can be re-expressed as

$$W = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (Q_i - Q_j)^2,$$

and after some algebra it can be shown that

$$W = \frac{1}{n-1} \sum_{i} (Q_i - \bar{Q})^2 = s^2,$$

the usual unbiased estimated of variance.

U-statistics are always unbiased - that is what the "U" stands for - and they have various other nice properties. For a summary of the by now sizeable body of theory which has been worked out for them, see Chapter 3

of Puri and Sen [12]. I shall present here only a few of the most relevant results.

Let a function Z_{WW} be defined as

$$Z_{WW}(Q_1,Q_2,\ldots,Q_{2k-1}) = w(Q_1,Q_2,\ldots,Q_k)\ w(Q_k,Q_{k+1},\ldots,Q_{2k-1}) - \omega^2.$$

Note that the set of k observations on which the first w depends and the set of k on which the second depends have exactly one observation in common. And let the expected value of Z_{WW}, assuming it exists, be ζ_{WW}. Then we have

$$\lim_{n\to\infty} n\ \mathrm{var}(W) = k^2\ \zeta_{WW} \geq 0,$$

or, in words, the asymptotic variance of W is $k^2\ \zeta_{WW}/n$. An estimate of ζ_{WW} can be obtained by the following method, due to Sen [16], which also provides an alternative expression for W. For each i = 1,2,...,n calculate w for only those subsets of k observations out of the n available which include the i-th observation - there are $\binom{n-1}{k-1}$ of these - and let W_i be their sum: in symbols

$$W_i = \sum_{C_i} w(Q_i,Q_{j_1},Q_{j_2},\ldots,Q_{j_{k-1}}).$$

Then it is not difficult to see that

$$W = \sum W_i/n\binom{n-1}{k-1}.$$

But in addition it can be shown that

$$S_{WW} = \frac{1}{n-1}\{\sum W_i^2/\binom{n-1}{k-1}^2 - nW^2\}$$

is a consistent estimator of ζ_{WW}. And furthermore, if $\zeta_{WW} > 0$ then W is asymptotically normally distributed: that is, for large n the quantity $\sqrt{n}(W-w)/k\sqrt{S_{WW}}$ is approximately a standard normal variable.

Thus if ω is the population variance σ^2, so that $w(Q_1,Q_2) = (Q_1-Q_2)^2/2$, and $W = s^2$, then

$$Z_{WW}(Q_1,Q_2,Q_3) = \frac{1}{4}(Q_1-Q_2)^2(Q_2-Q_3)^2 - \sigma^4.$$

The expected value of this can be worked out as

$$\zeta_{WW} = \frac{1}{4}(\mu_4-\sigma^4),$$

where μ_4 is the fourth central moment of Q; if Q is normal then $\mu_4 = 3\sigma^4$. Now, according to the theory of U-statistics, the asymptotic variance of s^2 must be $4\zeta_{WW}/n$, or $(\mu_4-\sigma^4)/n$; and this can be verified, since the exact varience is well-known, namely

$$\text{var}(s^2) = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4).$$

Furthermore, defining W_i as above, some algebraic manipulation shows that

$$W_i = \sum_j \frac{1}{2}(Q_i-Q_j)^2 = \frac{1}{2}\{n(Q_i-\bar{Q})^2 + (n-1)s^2\}$$

and thence that

$$S_{WW} = \frac{1}{4}\left(\frac{n}{n-1}\right)^3 \left\{\frac{\sum(Q_i-\bar{Q})^4}{n} - \left(\frac{n-1}{n}\right)^2 s^4\right\};$$

this clearly is a consistent estimator of ζ_{WW}, though not generally unbiased. Finally, U-statistics theory claims that the quantity $\sqrt{n}(s^2-\sigma^2)/2\sqrt{S_{WW}}$ must have asymptotically the standard normal distribution, provided only that μ_4 exists and is not zero. This can be easily verified at least where Q is normal, since then $(n-1)s^2/\sigma^2$ has exactly a chi-square distribution with (n-1) degrees of freedom; it is well-known that a chi-square variable approaches normality as its degrees of freedom increase without limit.

The results given above have been extended to the joint distribution of two U-statistics, or indeed of any number of them. Thus in particular if μ is a second parameter, of degree ℓ, with kernel $m(Q_1,Q_2,\ldots,Q_\ell)$, then the U-statistic for estimating it is

$$M = \sum_c m(Q_{i_1},Q_{i_2},\ldots,Q_{i_\ell})/\binom{n}{\ell},$$

with asymptotic variance $\ell^2\zeta_{MM}/n$, where ζ_{MM} is the expected value of

$$Z_{MM} = m(Q_1,Q_2,\ldots,Q_\ell)m(Q_\ell,Q_{\ell+1},\ldots,Q_{2\ell-1}) - \mu^2.$$

Let also ζ_{WM} be the expected value of

$$Z_{WM} = w(Q_1,Q_2,\ldots,Q_k)m(Q_k,Q_{k+1},\ldots,Q_{k+\ell-1}) - \omega\mu;$$

then as $n \to \infty$ the asymptotic joint distribution of the quantities

$U_W = \sqrt{n}(W-\omega)$ and $U_M = \sqrt{n}(M-\mu)$ is bivariate normal with means equal to zero, variances equal to $k^2\zeta_{WW}$ and $\ell^2\zeta_{MM}$, and covariance $k\ell\zeta_{WM}$; this includes the possibility of a degenerate normal distribution.

Finally, if

$$M_i = \sum_{C_i} m(Q_i, Q_{j_1}, Q_{j_2}, \ldots, Q_{j_{\ell-1}})$$

for $i = 1,2,\ldots,n$, then

$$S_{MM} = \frac{1}{n-1}\{\sum M_i^2/\binom{n-1}{\ell-1}^2 - nM^2\}$$

and

$$S_{WM} = \frac{1}{n-1}\{\sum W_i M_i/\binom{n-1}{k-1}\binom{n-1}{\ell-1} - nWM\}$$

are consistent estimators of ζ_{MM} and ζ_{WM}, respectively.

Starting from these known results, the following general theorem concerning the ratio of two U-statistics can be obtained.

Theorem. Let ω and μ be parameters of degrees k and ℓ, respectively, and let W and M be the U-statistics for estimating them from a random sample of size n. Then as $n \to \infty$ the random variable $\sqrt{n}(W/M-\omega/\mu)/s$ has asymptotically the normal distribution with mean 0 and variance 1 where, using the notation established above,

$$s^2 = \frac{k^2M^2S_{WW} - 2k\ell MWS_{WM} + \ell^2W^2S_{MM}}{M^4},$$

assuming only that $\mu \neq 0$ and that

$$\sigma^2 = \frac{k^2\mu^2\zeta_{WW} - 2k\ell\mu\omega\zeta_{WM} + \ell^2\omega^2\zeta_{MM}}{\mu^4} > 0.$$

Proof. The quantity σ^2 is defined since by assumption $\mu \neq 0$. Now as $n \to \infty$ we have $W \to \omega$, $M \to \mu$, $S_{WW} \to \zeta_{MM}$, and $S_{WM} \to \zeta_{WM}$, in probability; hence also $s^2 \to \sigma^2$, in probability. But since by assumption $\sigma^2 > 0$, the asymptotic distribution of $\sqrt{n}(W/M-\omega/\mu)/s$ must be the same as that

$$\frac{\sqrt{n}}{\sigma}\left(\frac{W}{M} - \frac{\omega}{\mu}\right) = \frac{\sqrt{n}}{\sigma M\mu}(W\mu-\omega M).$$

And then since $M \to \mu$ the required asymptotic distribution must also be the same as that of $\sqrt{n}(W\mu-\omega M)/\sigma\mu^2$. The desired result is then easily verified from the known asymptotic joint normal distribution of W and M.

Remark. The ratio of the theorem would be undefined should M or s vanish, and this may be possible for any finite n; thus in general the ratio has no mean and variance.

Now let us see how this Theorem applies to the index T of matched correlation. We may identify the variable Q of this Appendix with the multivariate (X,Y,Z). Suppose we have a random sample of observations $Q_i = (X_i, Y_i, Z_i)$, for $i = 1,2,\ldots,n$. Having established some definition of

matching, consider estimating the parameter $\mu = P\{\text{MATCH}\}$. Its degree is 2, and its kernel is

$$m(Q_1,Q_2) = \begin{cases} 1 & \text{if the observations } Q_1 \text{ and } Q_2 \text{ are matched} \\ 0 & \text{otherwise.} \end{cases}$$

We then have

$$M_i = \sum_{j \neq i} m(Q_1,Q_2), \qquad i = 1,2,\ldots,n,$$

which agrees with the definition, given in Section 5, that M_i is the number of observations matched with the i-th observation; and the U-statistic for estimating μ turns out to be $M = N_M/N$. Similarly, the parameter

$$\omega = P\{\text{C and MATCH}\} - P\{\text{D and MATCH}\}$$

is also of degree 2; its kernel is

$$w(Q_1,Q_2) = \begin{cases} 1 & \text{if } Q_1 \text{ and } Q_2 \text{ are concordant and matched} \\ -1 & \text{if } Q_1 \text{ and } Q_2 \text{ are discordant and matched} \\ 0 & \text{if } Q_1 \text{ and } Q_2 \text{ are tied or unmatched,} \end{cases}$$

corresponding to this we have

$$W_i = \sum_{j \neq i} w(Q_1, Q_2), \qquad i = 1,2,\ldots,n,$$

and the U-statistic for estimating ω turns out to be $W = (N_{CM}-N_{DM})/N$. Then finally the sample index is $T = W/M$, and the population index is $\theta = \omega/\mu$.

After substituting $k = \ell = 2$ into their definitions, a little algebra shows that

$$S_{WW} = \frac{1}{n-1} \Sigma\left(\frac{W_i}{n-1} - W\right)^2,$$

$$S_{MM} = \frac{1}{n-1} \Sigma\left(\frac{M_i}{n-1} - M\right)^2,$$

and

$$S_{WM} = \frac{1}{n-1} \Sigma\left(\frac{W_i}{n-1} - W\right)\left(\frac{M_i}{n-1} - M\right),$$

and thence

$$s^2 = \frac{4n^2}{(n-1)(\Sigma M_i)^4} \{\Sigma W_i^2 (\Sigma M_i)^2 - 2\Sigma W_i \Sigma M_i \Sigma W_i M_i + (\Sigma W_i)^2 \Sigma M_i^2\} = \frac{n^2 S^2}{n-1},$$

where S is as defined in Section 5. Now when this result is substituted into the Theorem we find that the quantity $\sqrt{(n-1)/n}(T-\theta)/S$ is asymptotically a normal (0,1) variable, but this agrees with Section 5, since for large n the factor $\sqrt{(n-1)/n}$ can be ignored.

The assumption that

$$\sigma^2 = \frac{4}{\mu^4}\,(\mu^2\zeta_{WW} - 2\mu\omega\zeta_{WM} + \omega^2\zeta_{MM}) > 0$$

may be interpreted as follows. Let $F(Q)$ be the distribution function of $Q = (X,Y,Z)$. Then

$$\zeta_{WW} = \iiint w(Q_1,Q_2)w(Q_2,Q_3)dF(Q_1)dF(Q_2)dF(Q_3) - \omega^2,$$

$$\zeta_{MM} = \iiint m(Q_1,Q_2)m(Q_2,Q_3)dF(Q_1)dF(Q_2)dF(Q_3) - \mu^2,$$

$$\zeta_{WM} = \iiint w(Q_1,Q_2)m(Q_2,Q_3)dF(Q_1)dF(Q_2)dF(Q_3) - \omega\mu,$$

and hence

$$\sigma^2 = \frac{4}{\mu^4} \iiint r(Q_1,Q_2)r(Q_2,Q_3)dF(Q_1)dF(Q_2)dF(Q_3)$$

$$= \frac{4}{\mu^4} \int \{\int r(Q_1,Q_2)dF(Q_2)\}^2\, dF(Q_1)$$

where

$$r(Q_1,Q_2) = \mu w(Q_1,Q_2) - \omega m(Q_1,Q_2).$$

Now clearly $\sigma^2 = 0$ is possible only if $\int r(Q_1,Q_2)dF(Q_2) = 0$ with probability 1 under F; that is, if

$$P(R) = \frac{\int w(R,Q)\, dF(Q)}{\int m(R,Q)\, dF(Q)} = \frac{\omega}{\mu} = \theta.$$

But P(R) is seen to be the probability that if an observation Q matched with the specified point R is drawn at random it will be concordant with R, less the probability that it will be discordant with R: if this probability is totally independent of R, that is of all the components (X,Y,Z) of R, then and only then can $\sigma^2 = 0$. This does occur in the extreme cases where $\theta = +1$ or -1; otherwise, quoting Goodman and Kruskal in a similar context [7, p.364] "we suggest that this is an unlikely state of affairs in most applications". Thus for practical purposes one may regard the asymptotic results for the sample index of matched correlation as valid provided only that the probability of a match is positive and that the population index is neither +1 nor -1.

For the following argument I am indebted to an anonymous referee of an earlier version of this paper. It yields the population value of the index of matched correlation in the second example of Section 6. Let $E(z_1,z_2)$ be the event that $Z_1 = z_1$ and $Z_2 = z_2$ in two random observations (X_i,Y_i,Z_i) $i = 1,2$, and let the conditional joint distribution of $U = X_1-X_2$ and $V = Y_1-Y_2$ given $E(z_1,z_2)$ be $F(u,v|z_1,z_2)$. Let also $G(z)$ be the distribution function of Z. Then

$$P\{C \mid E(z_1,z_2)\} - P\{D \mid E(z_1,z_2)\}$$

$$= P\{UV>0 \mid E(z_1,z_2)\} - P\{UV<0 \mid E(z_1,z_2)\}$$

$$= 1 - 2F(0,\infty \mid z_1,z_2) - 2F(\infty,0 \mid z_1,z_2) + 4F(0,0 \mid z_1,z_2)$$

$$= W(z_1,z_2), \text{ say.}$$

If matching is defined in terms of Z alone, then the index of matched correlation is

$$\theta = \frac{\iint_M W(z_1,z_2)\, dG(z_1)\, dG(z_2)}{\iint_M dG(z_1)\, dG(z_2)}$$

where M is the set of pairs (z_1,z_2) which are considered matched. If M is the region where $|z_1-z_2| \leq \varepsilon$, and if H is the distribution function of the difference between two independent Z's, then the denominator of θ is

$$P\{MATCH\} = H(\varepsilon) - H(-\varepsilon) = 2H(\varepsilon) - 1$$

since H must be symmetric about zero. Now if in addition we have $X_i = A_{1i} + A_{3i}$, $Y_i = A_{2i} + A_{3i}$, $Z_i = A_{3i}$, where the A's are independent and identically distributed, then

$$F(u,v \mid z_1,z_2) = H(u - (z_1-z_2))\, H(v - (z_1-z_2)),$$

and

$$W(z_1,z_2) = [1 - 2H(z_1-z_2)]^2.$$

Thus the numerator of θ is

$$\int_{-\varepsilon}^{\varepsilon} [1 - 2H(\varepsilon)]^2 \, dH(\varepsilon) = \frac{1}{3} [2H(\varepsilon) - 1]^3$$

and finally

$$\theta = \frac{1}{3} [2H(\varepsilon) - 1]^2 = \frac{1}{3} P^2\{MATCH\}.$$

For the case where Z is a normal (0,1) variable,

$$P\{MATCH\} = P\{|Z_1-Z_2| \leq \varepsilon\} = P\{|Z| \leq \frac{\varepsilon}{\sqrt{2}}\}.$$

It may be noted that the first part of this argument gives a general approach to the evaluation of population indices of matched correlation.