SMALL-SAMPLE PROPERTIES OF A FAMILY OF NONPARAMETRIC PARTIAL CORRELATION MEASURES

by

C. David Hardison

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1372

December 1981

SMALL-SAMPLE PROPERTIES OF A FAMILY OF
NONPARAMETRIC PARTIAL CORRELATION MEASURES

by

C. David Hardison

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill

1981

Approved by:

Dana Quade
Advisor

Reader

Bertram H. Kaplan
Reader

C. DAVID HARDISON. SMALL-SAMPLE PROPERTIES OF A FAMILY OF NONPARAMETRIC PARTIAL CORRELATION MEASURES (UNDER THE DIRECTION OF DANA QUADE).

Quade (1974) proposed a family of correlation measures of the form

$$T = \frac{C_R - D_R}{R}$$

where R is the number of relevant pairs and C_R and D_R are the numbers of these which are concordant and discordant, respectively.

We have investigated small-sample properties of T, the sample index of matched correlation; of three methods for obtaining its standard error; and of the quantity $Z=(T-\theta)/\text{s.e.}(T)$, where θ is the population index.

This was done both from a theoretical standpoint and in a Monte Carlo study using trivariate normal data. The study indicated that T and s.e.(T) are biased estimates of θ and the true standard deviation of T, respectively. This bias is usually positive. The distribution of Z is affected by these biases; we believe, however, that the practitioner can safely assume that this quantity attains standard normality for $|\theta| < .5$ when the sample size is at least 50. Furthermore, we presented strong evidence that T is at least normally distributed, although its mean may not equal θ.

Finally, we illustrated the general applicability and flexibility of matched partial correlation on data from one clinic of the Lipid Research Clinics Prevalence Study. We included analogous results from a parametric study of the associations between high-density lipoprotein cholesterol and other plasma lipid and lipoprotein concentrations by Davis _et al_. to provide the reader with a frame of reference.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my friend and advisor, Dana Quade. His guidance and support went beyond the call of duty. Equal thanks are due the other members of my committee, Ed Davis, Bert Kaplan, Gary Koch, and P. K. Sen, for their helpful comments and suggestions throughout this academic odyssey.

I cannot adequately thank my wife, Lucy, for her continuous support, encouragement, and, especially, her understanding. I give special thanks to my father and to my dearly departed mother - it is to her that I dedicate this dissertation.

Finally, I wish to thank Beryl Glover for her excellent first typing of this dissertation and Marla Cunningham and Velma Walker for their superb final typing.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd)

LIST OF TABLES

LIST OF TABLES (Cont'd)

LIST OF FIGURES

CHAPTER I

REVIEW OF THE LITERATURE ON NONPARAMETRIC PARTIAL CORRELATION

1.1 Total Correlation

We shall begin by discussing measures of nonparametric total correlation, of which nonparametric partial correlation measures are generalizations, but first we establish some terminology and notation. Let (X_i, Y_i) and (X_j, Y_j) be a pair of observations on the variables X and Y which are measured at least ordinally. We call the pair concordant if $(X_i-X_j)(Y_i-Y_j) > 0$, discordant if $(X_i-X_j(Y_i-Y_j) < 0$, or tied if $(X_i-X_j)(Y_i-Y_j) = 0$. Suppose we observe a random sample of size n; then, of the $\binom{n}{2}$ pairs in the sample, let C, D, and T denote the numbers which are concordant, discordant, and tied, respectively. We call a tied pair tied on X if $X_i=X_j$, tied on Y if $Y_i=Y_j$, and tied on X and Y if $X_i=X_j$ and $Y_i=Y_j$. Of the T tied pairs, let T_X, T_Y, and T_{XY} denote the numbers which are tied on X, tied on Y, and tied on X and Y, respectively; note that $T=T_X+T_Y-T_{XY}$.

Kendall (1962) described two total correlation measures

$$t_a(X,Y) = \frac{C-D}{C+D+T} \quad \text{and } t_b(X,Y) = \frac{C-D}{\sqrt{C+D+T_X}\sqrt{C+D+T_Y}} \quad .$$

We may think of $t_a(X,Y)$ as the proportion of sample pairs which are concordant less the proportion which are discordant. Although $t_b(X,Y)$ has no simple interpretation, it obviously is always at least as large as $t_a(X,Y)$ in the absolute sense, with equality when there are no tied pairs. A drawback of both $t_a(X,Y)$ and $t_b(X,Y)$ is that neither can attain +1 or −1 if ties occur. But let g(X,Y) be the proportion of sample pairs which are concordant, less the proportion which are discordant, among those which are not tied. This index, which was proposed by Goodman and Kruskal (1954), may be written in our notation as

$$g(X,Y) = \frac{C-D}{C+D} \quad .$$

It is undefined if all pairs are tied, i.e., if $T = \binom{n}{2}$; but on the other hand, perfect positive [g(X,Y) = 1] or negative [g(X,Y) = −1] correlation can obtain whether or not ties occur, thus overcoming the drawback of Kendall's measures. Further examination of these indexes produces the following order relationship among them:

$$0 \leq |t_a(X,Y)| \leq |t_b(X,Y)| \leq |g(X,Y)| \leq 1.$$

Other measures have been proposed which, like those discussed above, differ only in their treatment of ties; in fact, all the measures are identical if no ties occur. For example, Somers (1962) defines the asymmetric measures

$$d(X,Y) = \frac{C-D}{C+D+T_X} \quad \text{and} \quad d(Y,X) = \frac{C-D}{C+D+T_Y} \quad ,$$

which are such that $d(W,Y)d(Y,X) = t_b^2\,(X,Y)$. Wilson (1974) proposes another variant,

$$e(X,Y) = \frac{C-D}{C+D+T-T_{XY}} \quad .$$

Recently Schollenberger and Agresti (1979) proposed refinements to $t_b(X,Y)$ and $g(X,Y)$ for the case when Y is strictly ordinal and X is measured on an interval scale. In particular, they define

$$t_b'(X,Y) = \frac{\Sigma\,(X_j-X_i)\;s(Y_j-Y_i)}{\sqrt{\Sigma(X_j-X_i)^2\;\Sigma s(Y_j-Y_i)^2}} \quad \text{and} \quad g'(X,Y) = \frac{C_w-D_w}{C_w+D_w}$$

where the summation is over all pairs of observations, s is the sign function, i.e.,

$$s(t) = \begin{cases} 1 \text{ if } t>0 \\ 0 \text{ if } t=0 \\ -1 \text{ if } t<0 \end{cases} \quad ,$$

and C_w (D_w) is the weighted number of concordant (discordant) pairs with weight (X_j-X_i). To compare these two measures, note that $C_w-D_w = \Sigma(X_j-X_i)s(Y_j-Y_i)$.

1.2 Partial Correlation

Turning our attention to nonparametric partial correlation, suppose that, in addition to X and Y, we observe a third variable, say Z. There are two main rival approaches to defining the correlation between X and Y "controlled for" or "partialling out" Z.

The first approach was that of Kendall (1942). He proposed the earliest nonparametric measure of partial correlation, which we denote $v(X,Y|Z)$. Using the notation of Quade (1974), and assuming that Z is at least ordinal and that there are no ties, we have

$$v(X,Y|Z) = \frac{N_O N_Z - N_X N_Y}{\sqrt{(N_O+N_X)(N_Y+N_Z)(N_O+N_Y)(N_X+N_Z)}}$$

where

N_O = number of non-discordant pairs, i.e., those for which $(X_1-X_2)(Z_1-Z_2)>0$ and $(Y_1-Y_2)(Z_1-Z_2)>0$,

N_X = number of X-discordant pairs, i.e., those for which $(X_1-X_2)(Z_1-Z_2)<0$ and $(Y_1-Y_2)(Z_1-Z_2)>0$,

N_Y = number of Y-discordant pairs, i.e., those for which $(X_1-X_2)(Z_1-Z_2)>0$ and $(Y_1-Y_2)(Z_1-Z_2)<0$,

and

N_Z = number of Z-discordant pairs, i.e., those for which $(X_1-X_2)(Z_1-Z_2)<0$ and $(Y_1-Y_2)(Z_1-Z_2)<0$.

We note that $v(X,Y|Z)$ can be obtained by substituting the bivariate t_b for each pair of variables into the familiar product-moment form of the partial correlation coefficient, i.e.,

$$v(X,Y\ Z) = \frac{t_b(X,Y)-t_b(X,Z)t_b(Y,Z)}{\sqrt{[1-t_b^2(X,Z)][1-t_b^2(Y,Z)]}} \ .$$

Kendall was surprised by this fact but, as we will see, Hawkes (1971) was not.

Similar to his decomposion of t_b, Somers (1968) defined two asymmetric partial correlation measures related to $v(X,Y|Z)$,

$$d^*(X,Y|Z) = \frac{N_O}{N_O+N_X} - \frac{N_Y}{N_Y+N_Z} \text{ and } d^*(Y,X|Z) = \frac{N_O}{N_O+N_Y} - \frac{N_X}{N_X+N_Z} \ ,$$

where the asterisk indicates that ties must be excluded for the measures to be valid. These are related to $v(X,Y|Z)$ analogously to the total correlation case, namely,

$$d^*(X,Y|Z)\ d^*(Y,X|Z) = v^2(X,Y|Z)$$

Hawkes (1971) overcame the restriction of Kendall's and Somer's partial measures to data without ties by recognizing that regression about the origin of the signed differences of the pairs of observations produces regression coefficients which are in fact correlations and of which the previous partial measures are special cases. More specifically, he solved the normal equations given by

$$\begin{bmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(X,Z) & \mathrm{var}(Z) \end{bmatrix} \begin{bmatrix} d(Y,X|Z) \\ d(Y,Z|X) \end{bmatrix} = \begin{bmatrix} \mathrm{cov}(X,Y) \\ \mathrm{cov}(Z,Y) \end{bmatrix}$$

where

$$\mathrm{var}(X) = \frac{\Sigma s^2(X_j-X_i)}{2N^2},\ \mathrm{cov}(X,Y) = \frac{\Sigma s(X_j-X_i)s(Y_j-Y_i)}{2N^2},\ \text{etc.}$$

The extension to more variables is obvious. We see that Hawkes has constructed a linear model of the sign scores for all pairs of observations; this approach is described in detail by Ploch (1974).

Williams and Singh (1974) suggested obtaining a partial g by substituting the three total g's into a product-moment-like formula; that is, they proposed

$$g^*(X,Y|Z) = \frac{g(X,Y)-g(X,Z)g(Y,Z)}{[1-g^2(X,Z)][1-g^2(Y,Z)]}$$

They claimed that this measure "takes into account the monotonic tendencies of all the bivariate relationships".

The second major approach to nonparametric partial correlation, first suggested by Goodman and Kruskal (1954) for the case when Z can be grouped, is based on weighted averages of total correlation measures computed at each level of Z, of the form

$$c(X,Y|Z) = \frac{\sum_k w_k c_k(X,Y)}{\sum_k w_k}$$

where $c_k(X,Y)$ is some bivariate ordinal correlation computed at the k-th level of Z and $w_k/\Sigma w_k$ is some weight. In this vein, Davis (1967) proposed a partial counterpart of Goodman and Kruskal's g. Borrowing notation again from Quade, let C_k (D_k) be the number of pairs tied on Z at its k-th value that are also concordant (discordant) with respect to X and Y. Then Davis' partial g becomes

$$g(X,Y|Z) = \frac{\sum_k w_k g_k(X,Y)}{\sum_k w_k}$$

where $w_k = C_k + D_k$. The weight can be interpreted as the proportion of pairs tied on the k-th category of Z that are not tied on X or Y. A weighted version of t_b, denoted $\bar{t}_b$, was proposed by Agresti (1977), with

$$w_k = \sqrt{C_k+D_k+T_{XK}}\ \sqrt{C_k+D_kT_{YK}}$$

where T_{XK} and T_{YK} are the obvious restrictions of T_X and T_Y to the k-th value of Z.

The partial g of Davis may be written (and in fact was first written) as

$$g(X,Y|Z) = \frac{\sum_k C_k - \sum_k D_k}{\sum_k C_k + \sum_k D_k} \quad .$$

In this form we see that the denominator is simply the number of pairs tied on Z but not on X or on Y, while the numerator is the number of these that are concordant less the number discordant. This suggested to Quade (1967, 1974) a family of measures of the form

$$T = \frac{C_R - D_R}{R}$$

where R is the number of relevant pairs and C_R and D_R are the number of these which are concordant and discordant, respectively. He defined relevance in terms of any rule that, when applied to a pair of observations, results in the pair being classified as relevant or not. In the case of Davis' partial g, the obvious rule is "a pair is relevant if it is tied on Z but not on X or on Y".

Many of the total correlation measures previously mentioned are also members of the family. For example, T becomes Kendall's t_a if all pairs are relevant, or Goodman and Kruskal's g if all pairs are relevant unless they are tied. Similarly, suppose that pairs are relevant unless they are tied on X; then Somers' d(Y,X) is produced. For partial correlation Quade suggested defining relevance in terms of matching, where "two observations are intuitively considered matched if their values of Z are 'practically' equal": i.e., a pair of observations on the variables X and Y is matched if their values of a third variable Z, possible multivariate, agree with respect to some prespecified rule. The rule might involve a tolerance imposed on the values of Z, or possibly a distance function if Z is multivariate. In any case, a pair is relevant if it is matched on Z.

1.3 Asymptotic Sampling Theory

Assuming full multinomial sampling over an AxB cross-classification of X and Y, Goodman and Kruskal (1963, 1972) employed the "delta method", to be discussed in detail in Chapter 2, to establish the asymptotic normality of $\sqrt{n}(g-\gamma)$, where γ is the population counterpart of g. (This shows *a fortiori* that g converges in probability to γ). Let p_{ij} denote the observed proportion in the (i,j) cell, for i=1,..., A and j=1,...,B. They demonstrated that a consistent estimate of the asymptotic variance

(AV) of $\sqrt{n}(g-\gamma)$ is

$$s_\Delta^2(g) = \frac{16}{(P_C + P_D)^4} \sum_{i,j} p_{ij}(P_C D_{ij} - P_D C_{ij})^2 ,$$

where

$$C_{ij} = \sum_{i'>i} \sum_{j'>j} p_{i'j'} + \sum_{i'<i} \sum_{j'<j} p_{i'j'} ,$$

$$D_{ij} = \sum_{i'>i} \sum_{j'<j} p_{i'j'} + \sum_{i'<i} \sum_{j'>j} p_{i'j'} ,$$

$$P_C = \sum_{i,j} p_{ij} C_{ij} ,$$

and

$$P_D = \sum_{i,j} p_{ij} D_{ij} .$$

Similarly, they showed that $\sqrt{n}[d(Y,X) - \Delta(Y,X)]$ is asymptotically normal with its AV estimated by

$$s_\Delta^2[d(Y,X)] = \frac{4}{(1 - P_{TX})^4} \cdot \left\{ \sum_{i,j} p_{ij}[(P_C - P_D)(1 - P_{i.}) - (1 - P_{TX})(C_{ij} - D_{ij})]^2 \right\}$$

where

$$P_{i.} = \sum_{ij} p_{ij} \text{ and } P_{TX} = \sum_i p_{i.}^2$$

and, of course, $\Delta(Y,X)$ is the population counterpart of $d(Y,X)$. They also proposed an adjustment to $s_\Delta^2(g)$ for use when testing the hypothesis that $\gamma=0$, namely,

$$s_\Delta^2(g)^* = \frac{16}{(P_C + P_D)^2} \sum_{i,j} p_{ij}(D_{ij} - C_{ij})^2 .$$

Although the Goodman-Kruskal adjustment is asymptotically correct, it was pointed out by Brown and Benedetti (1977) that the assumption $\gamma=0$ places a restriction on g which should be taken into account when the delta method is applied. Their resulting variance estimate, say $s_{\Delta}^{2}(g)'$, is smaller than $s_{\Delta}^{2}(g)^{*}$. Hence Brown and Benedetti claimed that Goodman and Kruskal's estimate is conservative, and in particular that

$$s_{\Delta}^{2}(g)' = s_{\Delta}^{2}(g)^{*} - \frac{16(P_C - P_D)^2}{(P_C + P_D)^2} \quad .$$

Grizzle, Starmer, and Koch (1969) described a general approach, discussed in Chapter 2, for the analysis of contingency table data using the delta method. Forthofer and Koch (1973) extended the method and exemplified the computation of several of Goodman and Kruskal's g measures, along with estimates of their asymptotic variances and covariances. This approach was first applied to partial measures by Kritzer (1977), who demonstrated its use to compute Davis' partial g and Hawkes' version of Kendall's partial t_b. When restricted to measures of correlation in contingency tables, the Grizzle-Starmer-Koch method is simply a means of expressing the measures and estimates of their asymptotic variance-covariance structure as a series of matrix operations on the observed cell proportions; the matrices, however, vary considerably from application to application and their construction is often complex. In Chapter 3, we show how this approach can be greatly simplified.

Agresti (1977) was the first to apply this technique to a partial measure; he chose his $\bar{t}_b$. We extend the notation above to an AxBxC cross-classification with each observed cell proportion denoted p_{ijk} and define the following quantities:

$$C_{ijk} = \sum_{i'>i}\sum_{j'>j} p_{i'j'k} + \sum_{i'<i}\sum_{j'<j} p_{i'j'k} \quad ,$$

$$D_{ijk} = \sum_{i'>i}\sum_{j'<j} p_{i'j'k} + \sum_{i'<i}\sum_{j'>j} p_{i'j'k} \quad ,$$

$$C_k = \sum_{i,j} p_{ijk}C_{ijk} \; , \; D_k = \sum_{i,j} p_{ijk}D_{ijk} \quad ,$$

$$P_C = \sum_k C_k \qquad , \; P_D = \sum_k D_k \quad ,$$

$$R_k = [\sum_{i,j} p_{ijk}]^2 \quad , \; R_{ik} = [\sum_j p_{ijk}]^2 \; , \text{ and } R_{jk} = [\sum_i p_{ijk}]^2 \quad .$$

Kendall's t_b at any k then becomes

$$t_{bk} = \frac{C_k - D_k}{w_k} \quad ,$$

where

$$w_k = \sqrt{(R_k - R_{ik})(R_k - R_{jk})} \quad ,$$

and thus

$$\bar{t}_b = \frac{\sum_k w_k t_{bk}}{\sum_k w_k} \quad .$$

Agresti showed that $\sqrt{n}(\bar{t}_b - \bar{\tau}_b)$ is asymptotically normal and that its asymptotic variance can be estimated by

$$s_{\Delta}^{2}(\bar{t}_{b}) = \sum_{i,j,k} p_{ijk} \left\{ \frac{2(C_{ijk}-D_{ijk})}{\sum_{k} w_{k}} - \frac{(P_C-P_D)[(R_k-R_{jk})(p_{..k}-p_{i.k})+(R_k-R_{ik})(p_{..k}-p_{.jk})]}{w_k[\sum_k w_k]^2} \right\}^2$$

where $p_{\cdot\cdot k}$, $p_{i\cdot k}$, and $p_{\cdot jk}$ represent the obvious marginals.

On the other hand, Quade (1967, 1974) recognized that any measure in the family he defined is actually the ratio of two U-statistics. He proceeded to establish the large-sample theory for this family by demonstrating asymptotic normality using results of Heoffding (1948). Inference is made possible by an approximation formula for the asymptotic variance which he provided, based on the method of componenets of U-statistics due to Sen (1960). (As is the case for the delta method, Chapter 2 contains details on U-statistics, components, etc.) In particular, Quade showed that $\sqrt{n}(T-\theta)$ is asymptotically distributed as a normal random variable with AV estimated by

$$s_U^2(T) = \frac{4}{R^4} \sum_h [R(C_{R_h} - D_{R_h}) - (C_R - D_R)R_h]^2\}$$

where R_h is the number of relevant pairs which include observation (X_h,Y_h,Z_h) for $h=1,\ldots,n$ and C_{R_h} (D_{R_h}) is the number of these pairs which are concordant (discordant).

According to Singh (1974) we should use only those ordinal measures of association which have a well-understood sampling theory. Although this criterion seems reasonable, it would exclude all those measures that fit into the linear model framework. This is not intended to imply that no work has been done; the research in this area, however, does not provide any insight into the use of such measures in actual practice and thus is not discussed here.

1.4 Small-sample Distributions

While there are results on the large sample theory for some measures, theoretical studies of their behavior in small samples are nonexistent. A few researchers have empirically investigated the small sample properties of total correlation measures via Monte Carlo simulations; most, however, have restricted their studies to one or two cross-classifications. Goodman and Kruskal (1963) presented sampling results for 3x4 and 2x3 cross-classifications with 100 replications of size 50 samples and 50 replications of size 200 samples, indicating a reasonable agreement between the distribution of $\sqrt{n}(g-\gamma)/\sqrt{AV}$ and the standard normal distribution. They did not examine the behavior of this quantity when AV is estimated. In a similar study, Rosenthal (1966) considered the distribution of $\sqrt{n}(g-\gamma)/\sqrt{AV}$ using both the population AV and its estimate, with γ varying from 0 to 1. For 100 replications of a 5x5 cross-classification with sample sizes of 25 and 50, she found a tendency for heavier tails than the normal over much of the range of γ when the estimates of AV are used. Brown and Benedetti discovered that their correction to the AV when $\gamma=0$ improves the behavior of the distribution of $\sqrt{n}(g-\gamma)/\sqrt{AV}$. Their investigation examined 4x4 and 8x8 cross-classifications with sample sizes of 25, 50, 100, 200 and 25, 50, 100, 200, 400, respectively. Upon examination of the empirical results, they recommended using Student's t distribution with .4n degrees of freedom for an even better approximation when testing $\gamma=0$. Chow et. al. (1974) examined how closely Kendall's t_b approximates the true correlation when the underlying distribution is bivariate normal. They concluded that nonparametric correlations should be interpreted with caution when the true correlation is expected to be small (near zero) or large (near unity). More recently, Gans and Robertson (1981a, 1981b) investigated the behavior of g, t_b, the Pearson product-moment correlation, and Spearman's rank correlation for sample sizes 10 (10) 40 from 2x2 and 2x3 tables. They found that g converged to its asymptotic normal distribution much slower than the others to theirs, and that g is more likely to be significantly biased. For the larger sample sizes, however, the variances and expectations of all four measures were close to their asymptotic values.

An extensive Monte Carlo experiment was conducted by Reynolds (1971, 1974) to ascertain the effects of various w_k in the class of weighted partial correlation measures. In particular, he examined the use of equal weights for all k, weights proportional to the number of observations at each k, and weights proportional to the number of pairs at each k. He did not detect a "best" weighting scheme among the three when applied to t_a, t_b, g, and d(Y,X) for 10 replications of samples of size 600 for many different cross-classifications varying from 3x3x2 to 10x10x10. (If a variable had A categories, he divided the range of the outcomes into A intervals of equal length.) He assumed the "spurious" correlation model arising most often in path analysis. Simply, this model assumes that when X, Y, and Z are trivariate normal, the true correlation between X and Y controlling for Z is zero. Using Reynolds' (1974) notation, the model can be expressed by the recursive system of equations:

$$Z = \xi_1$$
$$X = Z + \xi_2$$
$$Y = Z + \xi_3$$

where ξ_1 is distributed as a normal variable with mean zero and variance 1, i.e., ξ_1 is (N(0,1), ξ_2 is $N(0,\sigma^2_{\xi_2})$, and ξ_3 is $N(0,\sigma^2_{\xi_3})$ with the variances depending on the correlation levels among the variables. For high (.70), medium (.49), and low (.30) correlation between X and Y, he compared how closely the aforementioned measures, and the product moment partial for categorical data, which he denoted $\bar{r}$, estimate the true partial of zero. Reynolds concluded that d(Y,X) weighted by the number of pairs at each level of Z is the most acceptable nonparametric measure and he recommended it over $\bar{r}$ because (i) the differences between d(Y,X) and $\bar{r}$ are small, (ii) $\bar{r}$ produces invalid results when relationships are monotonic but nonlinear (as he empirically illustrated), and (iii) the effect of categorization on the sampling distribution of $\bar{r}$ has not been investigated while a sampling theory is available for d(Y,X) a´la Quade (1974).

On the other hand, Kim (1975) argued against the use of ordinal measures of partial association for underlying continuous variables; but all of his evidence was based on Kendall's partial t_b and very small amounts of artificial data. Smith (1978) attacked Kim on these and other points in favor of the ordinal strategy.

Agresti (1977) examined the stability of his $\bar{t}_b$, Kendall's partial t_b, Renyolds' t_b with weights equal to the number of pairs at each level of Z, and Davis' partial g, for the spurious correlation model in a 2x2xc cross-classification. For c taking on values from 2 through 10, he claimed that $\bar{t}_b$ performs better and is more stable than the other measures. That is, it approximates zero better and is less variabale as c decreases from 10 to 2.

CHAPTER II

METHODS FOR THE CONSTRUCTION OF APPROXIMATE TESTS

2.1 U-statistics

Let $W_1,\ldots,W_n$ be a random sample of size n and let $f(W_1,\ldots,W_m)$ be any unbiased estimate of some parameter θ, where $m \leq n$ is the smallest number of observations needed to estimate θ; m is called the degree of θ. Then the U-statistic for θ is defined by

$$U(W_1,\ldots,W_n) = \frac{1}{\binom{n}{m}} \sum_{C_m} f^*(W_{\gamma_1},\ldots,W_{\gamma_m}),$$

where C_m indicates that the sum is over all $\binom{n}{m}$ combinations of the subscripts. The statistic f* is called the kernel of $U(W_1,\ldots,W_n)$; it is the symmetric form of $f(W_1,\ldots,W_n)$ given by

$$f^*(W_1,\ldots,W_m) = \frac{1}{m!} \sum_{P_m} f(W_{\alpha_1},\ldots,W_{\alpha_m}),$$

where P_m indicates that the sum is over the m! permutations of the subscripts. In the sequel we shall assume, without loss of generality, that $f(W_1,\ldots,W_m)$ is already symmetric, and we shall denote $U(W_1,\ldots,W_n)$ simply by U. Note that U is always an unbiased estimate of θ, hence the choice of letter.

Many of the statistics that we encounter most often are U-statistics or functions thereof. For example, let $W_i = (X_i, Y_i)$ $i=1,\ldots,n$ be a random sample from the bivariate population of X and Y and let

$$f_1(W_1) = X_1 ,$$

$$f_2(W_1) = Y_1 ,$$

$$f_3(W_1,W_2) = \frac{1}{2}(X_2-X_1)^2$$

$$f_4(W_1,W_2) = \frac{1}{2}(Y_2-Y_1)^2 ,$$

and

$$f_5(W_1,W_2) = \frac{1}{2}(X_2-X_1)(Y_2-Y_1) .$$

Then the U-statistics corresponding to $f_1,\ldots,f_5$ are

$$U_1 = \frac{1}{\binom{n}{1}} \sum_{C_1} f_1(W_{\gamma_1}) = \frac{1}{n} \sum_i X_i = \bar{X}$$

$$U_2 = \frac{1}{\binom{n}{1}} \sum_{C_1} f_2(W_{\gamma_1}) = \frac{1}{n} \sum_i X_i = \bar{Y}$$

$$U_3 = \frac{1}{\binom{n}{2}} \sum_{C_2} f_3(W_{\gamma_1},W_{\gamma_2}) = \frac{1}{n(n-1)} \sum_{1\le i\le j\le n} (X_j-X_i)^2$$

$$= \frac{1}{n-1} \sum_i (X_i-\bar{X})^2 = S^2(X) \quad ;$$

similarly,

$$U_4 = \frac{1}{n-1} \sum_i (Y_i-\bar{Y})^2 = S^2(Y)$$

and

$$U_5 = \frac{1}{n-1} \sum_i (X_i-\bar{X})(\bar{Y}_i-Y) = S(X,Y)$$

Thus in addition the sample product-moment correlation coefficient $r(X,Y)$ can be expressed as a function of U-statistics:

$$r(X,Y) = \frac{S(X,Y)}{S(X)X(Y)} = h(U_3,U_4,U_5) \quad .$$

Hoeffding (1948) established a central limit theory for U-statistics and demonstrated many of their nice theoretical properties. Let

$$f_{(k)}(w_1,\ldots,w_k) = E[f(W_1,\ldots,W_k,\ W_{k+1},\ldots,W_m\ W_1=w_1,\ldots,W_K=w_k)]$$

and let

$$\eta_1^{(k)} = \mathrm{Var}[f_{(k)}(W_1,\ldots,W_k)] \text{ and } \eta_1^{(0)}=0$$

where E and Var are the expectation and variance operators, respectively. Then

$$\mathrm{Var}(U) = \frac{1}{\binom{n}{m}} \sum_{k=1}^{m} \binom{m}{k}\binom{n-m}{m-k}\eta_1^{(k)} \quad .$$

Hoeffding showed that if $E[f(W_1,\ldots,W_m)^2]$ is finite, then $\sqrt{n}(U-\theta)$ is asymptotically distributed as a normal random variable with mean zero and variance $m^2\eta_1^{(1)}$. Unfortunately, at least the form of the underlying distribution must be known to calculate $\eta_1^{(1)}$, except for a few very special cases. However, Sen (1960) overcame this problem by introducing his method of components, which produces a consistent estimate of $\eta_1^{(1)}$. He defined the i-th component of U as

$$U_{(i)} = \frac{1}{\binom{n-1}{m-1}} \sum_{C_{m-1}} f(W_i, W_{\delta_1}, \ldots, W_{\delta_{m-1}}) \quad ,$$

where C_{m-1} indicates that the sum is over $\binom{n-1}{m-1}$ combinations of n subscripts containing i. In other words, $U_{(i)}$ denotes the sum of f calculated for all subsets of m observations out of the n which include observation i; the $U_{(i)}$ are related to U by

$$U = \frac{1}{n} \sum_i U_{(i)} \quad .$$

Sen showed that

$$S_U^2 = \frac{1}{n-1} \sum_i (U_{(i)} - U)^2$$

is a consistent estimate of $\eta_1^{(1)}$. Thus, under the condition of Hoeffding's central limit theorem, $\sqrt{n}(U-\theta)/mS_U$ is approximately a standard normal variable for large n.

Hoeffding extended his results to the joint distribution of any number of U-statistics, but we consider the results only for the bivariate case. Let θ_1 and θ_2 be parameters of degree m_1 and m_2 with kernels $f_1(W_1,\ldots,W_{m_1})$ and $f_2(W_1,\ldots,W_{m_2})$ and U-statistics U_1 and U_2, respectively. Then he showed that $\sqrt{n}(U_1-\theta_1)$ and $\sqrt{n}(U_2-\theta_2)$ are asymptotically bivariate normal with means 0, variances $m_1^2\eta_1^{(1)}$ and $m_2^2\eta_2^{(1)}$, and covariance $2m_1m_2\eta_{12}^{(1)}$ where $\eta_2^{(1)}$ is defined analogously to $\eta_1^{(1)}$ and

$$\eta_{12}^{(k)} = \mathrm{Cov}[f_{1\,(K)}\,(W_1,\ldots,W_k),\ f_{2\,(k)}\,(W_1,\ldots,W_k)] \quad .$$

Applying the method of components to obtain a consistent estimate of $\eta_{12}^{(1)}$ yields

$$S_{U_1U_2} = \frac{1}{n-1}\sum_i (U_{1(i)}-U_1)(U_{2(i)}-U_2)$$

where $U_{1(i)}$ and $U_{2(i)}$ denote the i-th components of U_1 and U_2, respectively.

It is easy to prove - - for the details see Quade (1967) - - a general theorem concerning the ratio of two U-statistics. In particular, $\sqrt{n}(U_1/U_2 - \theta_1/\theta_2)/S$ has an asymptotic standard normal distribution, where

$$S^2 = \frac{1}{U_2^4}(m_1^2U_2^2S_{U_1}^2 - 2m_1m_2U_1U_2S_{U_1U_2} + m_2^2U_1^2S_{U_2}^2) \quad .$$

under the assumption that

$$\frac{1}{\theta_2^4}(m_1^2\theta_2^2\eta_1^{(1)} - 2m_1m_2\theta_1\theta_2\eta_{12}^{(1)} + m_2^2\theta_1^2\eta_2^{(1)}) > 0 \quad .$$

2.2 The Jackknife

The jackknife technique is a method of obtaining approximate tests for parameters of interest. Let θ be the parameter of interest, and T some estimator of θ. Further, let $W_1,\ldots,W_N$ be a random sample of size N which can be partitioned into n groups of size k (in practice, generally n=N and k=1), and let T_i be the estimate based on T after the deletion of the i-th group of observations. Then the jackknife estimate of θ and its estimated variance are given by

$$T_J = \frac{1}{n} \sum_i \hat{T}_i \text{ and } S_J^2 = \frac{1}{n-1} \sum_i (\hat{T}_i - T_J)^2$$

where

$$\hat{T}_i = nT - (n-1)T_i \quad .$$

The technique was first introduced by Quenouille (1949) for the case n=2. Later (1956) he showed the usefulness of the jackknife estimate in bias reduction: if T is baised on order 1/N, then T_J reduces the bias to order $1/N^2$. (It is possible to construct second, third, and higher order jackknife estimates to reduce bias of order $1/N^2$, $1/N^3$, etc., but these methods are mostly of theoretical interest and are not discussed here.) Tukey (1958) showed that $\sqrt{n}(T_J-\theta)/S_J$ is approximately distributed as a Student's t random variable with n-1 degrees of freedom; he is responsible for the name "jackknife", given because of the technique's almost universal applicability.

Arvesen (1969) established several important results which link the theories of the jackknife and U-statistics. His Theorems 8 and 9 can be combined to produce the following when the estimate

T of θ is actually a function of U-statistics (for ease of presentation, we restrict the result to functions of two U-statistics although Arvesen's proofs hold in general): Assume that $T=g(U_1,U_2)$ where g has continuous first partial derivatives and U_1 and U_2 are the the U-statistics for estimating some parameters θ_1 and θ_2 such that $\theta=g(\theta_1,\theta_2)$. Define $T_i=g(U_{1i},U_{2i})$ where $U_{\ell i}$ denotes U-statistics ℓ calculated without the i-th group of observations for $\ell=1,2$. Then $\sqrt{n}(T_J-\theta)/S_J$ is asymptotically distributed as a standard normal random variable, where T_J and S_J are functions of the T_i as defined above.

2.3 The Delta Method

The delta method is a general technique for obtaining a central limit theory for a function of dependent random variables. Cramer (1946) presented a somewhat restricted version of the delta method, while Hoeffding and Robbins (1948) proved very general results employing this technique; here, however, we discuss the version given by Goodman and Kruskal (1963) since it is more relevant to our interests.

Assume that W_n and V_n are two sequences of random variables (n=1,2,...) and that ω and ν are constants such that the pairs

$$[\sqrt{n}(W_n-\omega),\ \sqrt{n}(V_n-\nu)]$$

have an asymptotic bivariate normal distribution with zero means, variances σ_W^2 and σ_V^2, and covariance σ_{WV}. Let g(a,b) be a function with continuous first partial derivatives at (ω,ν). Then $\sqrt{n}[g(W_n,V_n) - g(\omega,\nu)]/S_\Delta$ has an asymptotic standard normal distribution, where

$$S_\Delta^2 = \left[\frac{\partial g(W_n,V_n)}{\partial W_n} \bigg|_{W_n=\omega}\right]^2 S_W^2$$

$$+ 2\left[\frac{\partial g(W_n,V_n)}{\partial W_n} \bigg|_{W_n=\omega}\right] \left[\frac{\partial g(W_n,V_n)}{\partial V_n} \bigg|_{V_n=\nu}\right] S_{WV} + \left[\frac{\partial g(W_n,V_n)}{\partial V_n} \bigg|_{V_n=\nu}\right]^2 S_V^2$$

and S_W^2, S_V^2, and S_{WV} are any consistent estimates of σ_W^2, σ_V^2, and σ_{WV}, respectively.

As metioned earlier, Goodman and Kruskal applied this technique to several measures, utilizing well known properties of the multinominal probability distribution, but only for a two-way cross-classification. Extending their approach to multi-way cross-classification is combersome to say the least, unless matrix conventions are used. These conventions have been popularized by their use in the Grizzle, Starmer, and Koch (GSK) method for the analysis of categorical data. Assume full multinomial sampling over the entire multi-way cross-classification. Instead of viewing the data set as a contingency table, think of it as a one-dimensional vector of the observed proportions in the cells. That is, define

$$\underset{\sim 1xr}{p'} = [p_1,\ldots,p_r] \text{ and } \underset{\sim 1xr}{\pi'} = [\pi_1,\ldots,\pi_r]$$

where p_i and π_i are the observed proportion and the expected cell probability, respectively, in cell i and r is the total number of cells. Then it is well known that the variance-covariance matrix of the p_i is given by

$$\underset{\sim rxr}{V} = \frac{1}{n}(D_\pi - \underset{\sim\sim}{\pi\pi'})$$

with estimate

$$\underset{\sim rxr}{S} = \frac{1}{n}(D_p = \underset{\sim\sim}{pp'})$$

where n is the overall sample size and D_X denotes an rxr diagonal matrix with the elements of an rx1 vector $\underset{\sim}{X}$ on the diagonal. Let $G(\underset{\sim}{p})$ denote any scalar function of the p_i's that has continuous first partial derivatives at $\underset{\sim}{\pi}$. Then the delta method yields an asymptotic standard normal variable in the form of $\sqrt{n}[G(p) - G(\underset{\sim}{\pi})]/S_\Delta$ where

$$S_\Delta^2 = \underset{\sim}{c}'\underset{\sim\sim}{Sc}$$

with $\underset{\sim}{c}$ a rx1 vector whose i-th element is given by

$$c_i = \frac{\partial F(\underset{\sim}{p})}{\partial p_i} \quad .$$

A recent application of two of the three methods given above for the construction of approximate tests is given by Hochberg (1981), who demonstrated the use of both U-statistics and the delta method to estimate the non-null variance of the Wilcoxon-Mann-Whitney statistic applied to grouped ordered data. He conjectured that the U-statistic estimator is less biased than the one based on the delta method in many instances.

CHAPTER III

THEORETICAL PROPERTIES OF THREE ESTIMATES

OF THE VARIANCE OF T

3.1 The Jackknife and the Method of Components Estimates of the Variance of a U-statistic

We begin this chapter by examining the relationship between the components and the jackknife pseudovalues of a U-statistic. As before, let U denote a U-statistic of degree m, say, for estimating some parameter θ, $U_{(i)}$ denote the i-th component of U, and U_i denote U computed with the i-th observation removed. Then

$$U = \frac{1}{n} \sum_i U_{(i)}$$

and the jackknife estimate of θ based on U is given by

$$U_J = \frac{1}{n} \sum_i \hat{U}_i$$

where

$$\hat{U}_i = nU - (n-1)U_i \ .$$

Thus, we are interested in comparing the components, $U_{(i)}$, to the jackknife pseudovalues, $\hat{U}_i$. We must point out that Sen (1977) alludes to some of the following relationships in his study of invariance principles relating to jackknifing and their role in sequential analysis.

By definition we have

$$\binom{n-1}{m}U_i + \binom{n-1}{m-1}U_{(i)} = \binom{n}{m}U, \quad i = 1(1)n$$

and, after some algebra, we have

$$U_i = \frac{nU - mU_{(i)}}{n - m} .$$

Now

$$\begin{aligned}
\hat{U}_i &= nU - (n-1)U_i \\
&= nU - \left(\frac{n-1}{m-1}\right)(nU - mU_{(i)}) \\
&= \frac{n(n-m)\ U - n(n-1)\ U + m(n-1)\ U_{(i)}}{n - m} \\
&= \frac{m(n-1)\ U_{(i)} - n(m-1)\ U}{n - m} .
\end{aligned}$$

Hence, it is interesting but not surprising to note that

$$U_J = U .$$

Similarly, we can write

$$\begin{aligned}
U_{(i)} &= \frac{(n-m)\ \hat{U}_i + n(m-1)\ U}{m(n-1)} \\
&= \frac{(n-m)\ \hat{U}_i + n(m-1)\ U_J}{m(n-1)} .
\end{aligned}$$

We next obtain an analogous relationship between S_J^2 and S_U^2:

$$S_J^2 = \frac{1}{n-1} \sum_i (\hat{U}_i - U_J)^2$$

$$= \frac{1}{n-1} \sum_i (\hat{U}_i - U)^2$$

$$= \frac{1}{n-1} \sum_i \left\{\frac{m(n-1)\ U_{(i)} - n(m-1)\ U}{n - m} - U\right\}^2$$

$$= \frac{1}{n-1} \sum_i \left\{\frac{m(n-1)}{n-m} (U_{(i)} - U)\right\}^2$$

$$= \frac{m^2(n-1)^2}{(n-m)^2} \cdot S_U^2 \quad .$$

Recall that $\sqrt{n}(U-\theta)/m\eta_1^{(1)}$ is asymptotically distributed as a standard normal random variable and that S_U^2 is a consistent estiate of $\eta_1^{(1)}$. Thus the components of U-statistics variance estimate is $\frac{m^2}{n} S_U^2 = \hat{\sigma}_U^2$, say. Recall also that $\sqrt{n}(U_J-\theta)/S_J$ is asymptotically distributed as a Student's t random variable with n-1 degrees of freedom, i.e., the asymptotic variance of U_J is $S_J^2/n = \hat{\sigma}_J^2$, say. Then

$$\hat{\sigma}_J^2 = \frac{m^2(n-1)^2}{n(n-m)^2} S_U^2$$

$$= \frac{(n-1)2}{(n-m)^2} \hat{\sigma}_U^2 \quad .$$

Therefore, $\hat{\sigma}_J^2 \geq \hat{\sigma}_U^2$ although they are of the same order of magnitude regardless of m, and are asymptotically equivalent. Efron and Stein (1981) show that the jackknife estimate of the variance of U is always biased upwards, i.e., they show that

$$E(\hat{\sigma}_J^2) \geq \sigma_U^2$$

where σ_U^2 is the true variance of U and E(.) is the expectation operator.

Efron and Stein provide a much more general result which applies not just to U-statistics but also to the index of matched partial correlation T. The conditions are that the statistic must have a finite second moment and that the sample observations must be independent. (Note that symmetry or identically distributed observations are not required.) They show that

$$E(\hat{\sigma}_J^2) \geq \mathrm{Var}(U.)$$

where $U. = \frac{1}{n}\sum_i U_i$.

Thus, if we can assume that the true variance of T, say σ_T^2, is such that

$$\mathrm{Var}^{(n)} = \frac{n-1}{n}\mathrm{Var}^{(n-1)}$$

where

$$\mathrm{Var}^{(k)} = \mathrm{Var}\{T(W_1,\ldots,W_k)\} \quad ,$$

then it follows that

$$E(\hat{\sigma}_J^2) \geq \sigma_T^2$$

with $\hat{\sigma}_J^2$ now the jackknife estimate of the variance of T. We must point out that, although the sample size modification is valid in many instances, there are statistics for which this is not the case. The sample median is one example. However, if we are willing to assume that the sample size modification is valid in the case of matched partial correlation, then the jackknife estimate of the variance of T is always biased upwards.

3.2 The Delta Method Versus the Components of U-statistics for the Ratio of Two U-statistics Both of Degree 2

Consider a random sample of size n obtained by full multinomial sampling over some multi-way cross-classification with the cells indexed by a single subscript i running from 1 through r. Let

π_i = Pr {random observation falls in cell i} ,

p_i = proportion of sample observations in cell i ,

n_i = number of sample observations in cell i ,

where

$$\sum_i \pi_i = \sum_i p_i = 1, \ \sum_i n_i = n, \text{ and } n_i = np_i \ .$$

Denote by $\underset{\sim}{\pi}$, $\underset{\sim}{p}$, and $\underset{\sim}{n}$ the rx1 vectors containing the π_i, p_i, and n_i, respectively, for $i=1,\ldots,r$. Let $\Omega=\{(\omega_{ij})\}$ and $\Lambda= \{(\lambda_{ij})\}$ be two symmetric rxr matrices. Then $\omega=\underset{\sim}{\pi}'\ \Omega\ \underset{\sim}{\pi}$ and $\lambda=\underset{\sim}{\pi}'\ \Lambda\ \underset{\sim}{\pi}$ are two quadratic forms in $\underset{\sim}{\pi}$. The U-statistics to estimate ω and λ are

$$W = \frac{\underset{\sim}{n}'\ \Omega\ \underset{\sim}{n} - \underset{\sim}{\phi}'\ \underset{\sim}{n}}{n^2 - n} \quad \text{where } \underset{\sim}{\phi}'\ (\omega_{11},\omega_{rr})$$

and

$$L = \frac{\underset{\sim}{n}'\ \Lambda\ \underset{\sim}{n} - \underset{\sim}{\psi}'\ \underset{\sim}{n}}{n^2 - n} \quad \text{where } \underset{\sim}{\psi}' = (\lambda_{11},\ldots,\lambda_{rr})$$

respectively.

Now, consider

$$\theta = \frac{\omega}{\lambda} \ .$$

The estimate of θ is

$$T = \frac{W}{L} = \frac{\underset{\sim}{n}'\ \Omega\ \underset{\sim}{n} - \underset{\sim}{\phi}'\ \underset{\sim}{n}}{\underset{\sim}{n}'\ \Lambda\ \underset{\sim}{n} - \underset{\sim}{\psi}'\ \underset{\sim}{n}} = \frac{n^2\ \underset{\sim}{p}'\ \Omega\ \underset{\sim}{p} - n\ \underset{\sim}{\phi}'\ \underset{\sim}{p}}{n^2\ \underset{\sim}{p}'\ \Lambda\ \underset{\sim}{p} - n\ \underset{\sim}{\psi}'\ \underset{\sim}{p}} \ .$$

We have in mind the special case where

$$\omega_{ij} = \begin{cases} 1 & \text{if an observation in cell } i \text{ and one in cell } j \text{ form a concordant relevant pair} \\ -1 & \text{...dicordant relevant pair} \\ 0 & \text{...tied and/or not relevant pair } \{\text{Note: } \omega_{ii}\equiv 0\} \end{cases}$$

and

$$\lambda_{ij} = \begin{cases} 1 \text{ if an observation in cell } i \text{ and one} \\ \quad \text{in cell } j \text{ form a relevant pair} \\ \\ 0 \text{ otherwise} \end{cases}$$

However, the added generality does not add much complexity.

In this setting, the estimate of the asymptotic variance of T given by the components of U-statistics method becomes

$$S_U^2(T) = \frac{4}{(\sum_i n_i L_i)^4} \{(\sum_i n_i L_i)^2 \sum_i n_i W_i^2 - 2(\sum_i n_i L_i)(\sum_i n_i W_i)\sum_i n_i W_i L_i$$

$$+ (\sum_i n_i W_i)^2 \sum_i n_i L_i^2\} ,$$

where W_i and L_i are the components of the U-statistics W and L, respectively;

$$W_i = \frac{1}{n-1} (\sum_j \omega_{ij} n_i - \omega_{ii})$$

$$L_i = \frac{1}{n-1} (\sum_j \lambda_{ij} n_i - \omega_{ii})$$

for i=1(1)r, where each sum is over j=1(1)r.

Now we desire a matrix representation of $S_U^2(T)$. Note the following:

$$(n-1)\sum_i n_i W_i = \sum_{ij}\sum n_i \omega_{ij} n_j - \sum_i \omega_{ii} n_i = \underset{\sim}{n}' \, \Omega \, \underset{\sim}{n} - \underset{\sim}{\phi}' \, \underset{\sim}{n} = n(n-1)W;$$

by analogy $(n-1)\sum_i n_i L_i = n(n-1)L$;

$$(n-1)^2\sum_i n_i W_i^2 = \sum_i n_i(\sum_j \omega_{ij}n_j)^2 - 2(\sum_i n_i\omega_{ii})(\sum_j \omega_{ij}n_j) + \sum_i n_i\omega_{ii}^2$$

$$= \underset{\sim}{n}'\ \Omega\ D_n\Omega\underset{\sim}{n} - 2\underset{\sim}{n}'\ \Omega\ D\phi\underset{\sim}{n} + \underset{\sim}{n}'\ D_\phi\underset{\sim}{\phi}$$

(Recall that D_X is the rxr diagonal matrix formed from the rx1 vector $\underset{\sim}{X}$)

$$= \underset{\sim}{n}'\ \Omega\ D_n\Omega\underset{\sim}{n} - 2\underset{\sim}{n}'\ \Omega\ D_n\underset{\sim}{\phi} + \underset{\sim}{\phi}'\ D_n\underset{\sim}{\phi}$$

(since $D_X\underset{\sim}{Y} = D_Y\underset{\sim}{X}$ for any two rx1 vectors $\underset{\sim}{X}$ and $\underset{\sim}{Y}$)

$$= (\Omega\underset{\sim}{n} - \underset{\sim}{\phi})'D_n(\Omega\underset{\sim}{n} - \underset{\sim}{\phi});$$

by analogy, $(n-1)^2\sum_i n_i L_i^2 = (\Lambda\underset{\sim}{n} - \underset{\sim}{\psi})'D_n(\Lambda\underset{\sim}{n} - \underset{\sim}{\psi})$;

and finally,

$$(n-1)^2\sum_i n_i W_i L_i = \sum_i n_i(\sum_j \omega_{ij}n_j - \omega_{ii})(\sum_k \lambda_{ik}n_k - \lambda_{ii})$$

$$= \underset{\sim}{n}'\ \Omega\ D_n\Lambda\underset{\sim}{n} - \underset{\sim}{n}'\ \Omega\ D_\psi\underset{\sim}{n} - \underset{\sim}{n}'\ \Lambda\ D_\phi\underset{\sim}{n} + \underset{\sim}{n}'D_\phi\underset{\sim}{\psi}$$

$$= \underset{\sim}{n}'\ \Omega\ D_n\Lambda\underset{\sim}{n} - \underset{\sim}{n}'\ \Omega\ D_n\underset{\sim}{\psi} - \underset{\sim}{n}'\ \Lambda\ D_n\underset{\sim}{\phi} + \underset{\sim}{\phi}'D_n\underset{\sim}{\psi}$$

$$= (\Omega\underset{\sim}{n} - \underset{\sim}{\phi})'D_n(\Lambda\underset{\sim}{n} - \underset{\sim}{\psi})\quad .$$

Therefore, the estimate of the variance of T produced by the method of components can be written as

$$S_U^2(T) = \frac{4}{[n(n-1)]^2L^4}\{\ L^2(\Omega \underset{\sim}{n}-\underset{\sim}{\phi})'D_n(\Omega \underset{\sim}{n}-\underset{\sim}{\phi})$$

$$-\ 2WL(\Omega \underset{\sim}{n}-\underset{\sim}{\phi})'D_n(\Lambda \underset{\sim}{n}-\underset{\sim}{\psi}) + W^2(\Lambda \underset{\sim}{n}-\underset{\sim}{\psi})'D_n(\Lambda \underset{\sim}{n}-\underset{\sim}{\psi})\ \} \quad .$$

$$= \frac{4}{[n(n-1)]^2L^4}\{[L(\Omega \underset{\sim}{n}-\underset{\sim}{\phi}) - W(\Lambda \underset{\sim}{n}-\underset{\sim}{\psi})]'D_n[L(\Omega \underset{\sim}{n}-\underset{\sim}{\phi})$$

$$-\ W(\Lambda \underset{\sim}{n}-\underset{\sim}{\psi})\phi]\} \ .$$

Next we apply the delta method to the scalar function $F(\underset{\sim}{p})=T$. We know that we can estimate the variance of T by

$$S_\Delta^2(T) = \frac{1}{n}\underset{\sim}{a}'(D_p - \underset{\sim}{p}\underset{\sim}{p}')\underset{\sim}{a} \text{ where } \underset{\sim}{a} = \frac{\partial T}{\partial \underset{\sim}{p}} \quad .$$

In particular, T=W/L so that

$$\underset{\sim}{a} = \frac{L(\frac{\partial W}{\partial \underset{\sim}{p}} - W(\frac{\partial L}{\partial \underset{\sim}{p}})}{L^2}$$

and

$$\frac{\partial W}{\partial p} = \frac{1}{n(n-1)}(2n^2\Omega\underset{\sim}{p} - n\underset{\sim}{\phi}) = \frac{1}{n-1}(2\Omega \underset{\sim}{n} - \underset{\sim}{\phi})$$

and

$$\frac{\partial L}{\partial \underset{\sim}{p}} = \frac{1}{n(n-1)}(2n^2\Lambda\underset{\sim}{p} - n\underset{\sim}{\psi}) = \frac{1}{n-1}(2\Lambda \underset{\sim}{n} - \underset{\sim}{\psi}) \quad .$$

Hence,

$$S_\Delta^2(T) = \frac{1}{nL^4}\left\{\left[L\left(\frac{\partial W}{\partial \underset{\sim}{p}}\right) - W\left(\frac{\partial L}{\partial \underset{\sim}{p}}\right)\right]'(D_p - \underset{\sim}{p}\underset{\sim}{p}')\left[L\left(\frac{\partial W}{\partial \underset{\sim}{p}}\right) - W\left(\frac{\partial L}{\partial p}\right)\right]\right\}$$

$$= \frac{4}{n^2(n-1)^2L^4}\{L^2(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi})'(D_n - \underset{\sim}{n}\underset{\sim}{p}')(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi})$$

$$- 2WL(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi})'(D_n - \underset{\sim}{n}\underset{\sim}{p}')(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})$$

$$+ W^2(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})'(D_n - \underset{\sim}{n}\underset{\sim}{p}')(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\}$$

$$= \frac{4}{[n(n-1)]^2L^4}\left\{\left[L(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\right]'\right.$$

$$\left.(D_n - \underset{\sim}{n}\underset{\sim}{p}')\left[L(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\right]\right\} .$$

We have obtained matrix representations of both variance approximation methods in order to facilitate their comparison. We are particularly interested in their difference, namely,

$$S_U^2(T) - S_\Delta^2(T)$$

$$= \frac{4}{[n(n-1)]^2L^4}\{L^2\left[(\underset{\sim}{\phi}'D_n\underset{\sim}{\phi} - 2\underset{\sim}{\phi}D_n\Omega\underset{\sim}{n}) - \frac{1}{4}(\underset{\sim}{\phi}'D_n - 4\underset{\sim}{\phi}'D_n\Omega\underset{\sim}{n})\right]$$

$$- 2WL\left[(\underset{\sim}{\phi}'D_n\underset{\sim}{\psi} - \underset{\sim}{\phi}'D_n\Lambda\underset{\sim}{n} - \underset{\sim}{\psi}'D_n\Omega\underset{\sim}{n}) - \frac{1}{4}(\underset{\sim}{\phi}'D_n\underset{\sim}{\psi} - 2\underset{\sim}{\phi}'D_n\Lambda\underset{\sim}{n} - 2\underset{\sim}{\psi}'D_n\Omega\underset{\sim}{n})\right]$$

$$+ W^2\left[(\underset{\sim}{\psi}'D_n\underset{\sim}{\psi} - 2\underset{\sim}{\psi}'D_n\Lambda\underset{\sim}{n}) - \frac{1}{4}(\underset{\sim}{\psi}'D_n\underset{\sim}{\psi} - 4\underset{\sim}{\psi}'D_n\Lambda\underset{\sim}{n})\right]$$

$$+\left[L(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\right]'\underset{\sim}{n}\underset{\sim}{p}'\left[L(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\right]$$

$$= \frac{4}{[n(n-1)]^2 L^4}\{\frac{3}{4}(L\underset{\sim}{\phi}-W\underset{\sim}{\psi})'D_n(L\underset{\sim}{\phi}-W\underset{\sim}{\psi}) - (L\underset{\sim}{\phi}-W\underset{\sim}{\psi})'D_n(L\Omega\underset{\sim}{n}-W\Lambda\underset{\sim}{n})$$

$$+ \left[L(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\right]'\underset{\sim}{n}\underset{\sim}{p}'\left[L(\Omega\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} - \frac{1}{2}\underset{\sim}{\psi})\right]\} \ .$$

There are no obvious simplifications of the above difference. Consider, however, the following alternative approach which assumes a structure in the U-statistics of which the index of matched partial correlation is a special case. Let

$$c = \frac{4}{n^2(n-1)^2L^4}$$

and

$$\underset{\sim}{\gamma} = L(\Omega\underset{\sim}{n}-\underset{\sim}{\phi}) - W(\Lambda\underset{\sim}{n} -\underset{\sim}{\psi}) \quad .$$

Then

$$S_U^2(T) = c\{\underset{\sim}{\gamma}'D_n\underset{\sim}{\gamma}\} \quad .$$

Note that

$$\underset{\sim}{\gamma}'\underset{\sim}{n} = \underset{\sim}{n}'\underset{\sim}{\gamma} = L(\underset{\sim}{n}'\Omega\underset{\sim}{n}-\underset{\sim}{n}'\underset{\sim}{\phi}) - W(\underset{\sim}{n}'\Lambda\underset{\sim}{n} - \underset{\sim}{n}'\underset{\sim}{\psi})$$

$$= L[n(n-1)W] - W[n(n-1)L]$$

$$= 0 \quad .$$

Now,

$$S_\Delta^2(T) = c\{(\underset{\sim}{\gamma}+\frac{L}{2}\underset{\sim}{\phi} - \frac{W}{2}\underset{\sim}{\psi})'(D_n-\underset{\sim}{n}\underset{\sim}{p}')(\underset{\sim}{\gamma}+\frac{L}{2}\underset{\sim}{\phi} - \frac{W}{2}\underset{\sim}{\psi}) \quad .$$

Suppose

$$\underset{\sim}{\phi} = \alpha\underset{\sim}{j} \text{ and } \underset{\sim}{\psi} = \beta\underset{\sim}{j} \text{ where } \underset{\sim}{j} = (1,\ldots,1)'$$

and α and β are some constants (e.g., they are 0 and 1 if T is the index of matched partial correlation). Then

$$S_{\Delta}^{2}(T) = c\{\left[\underset{\sim}{\gamma} + (\frac{L\alpha-W\beta}{2})\underset{\sim}{j}\right]'(D_n-\underset{\sim}{n}\underset{\sim}{p}')\left[\underset{\sim}{\gamma} + (\frac{L\alpha-W\beta}{2})\underset{\sim}{j}\right]\}$$

$$= c\{\left[\underset{\sim}{\gamma} + (\frac{L\alpha-W\beta}{2})\underset{\sim}{j}\right]'D_n\left[\underset{\sim}{\gamma} + (\frac{L\alpha-W\beta}{2})\underset{\sim}{j}\right]\}$$

$$- \frac{c}{n}\{\left[\underset{\sim}{\gamma} + (\frac{L\alpha-W\beta}{2})\underset{\sim}{j}\right]'\underset{\sim}{n}\underset{\sim}{n}'\left[\underset{\sim}{\gamma} + (\frac{L\alpha-W\beta}{2})\underset{\sim}{j}\right]\}$$

$$= c\{\underset{\sim}{\gamma}'D_n\underset{\sim}{\gamma} + (L\alpha-W\beta)\underset{\sim}{\gamma}'D_n\underset{\sim}{j} + \left[\frac{(L\alpha-W\beta)^2}{4}\right]\underset{\sim}{j}'D_n\underset{\sim}{j}\}$$

$$- \frac{c}{n}\{\underset{\sim}{\gamma}'\underset{\sim}{n}\underset{\sim}{n}'\underset{\sim}{\gamma} + (L\alpha-W\beta)\underset{\sim}{\gamma}'\underset{\sim}{n}\underset{\sim}{n}'\underset{\sim}{j} + \left[\frac{(L\alpha-W\beta)^2}{4}\right]\underset{\sim}{j}'\underset{\sim}{n}\underset{\sim}{n}'\underset{\sim}{j}\} \quad .$$

But

$$\underset{\sim}{j}'D_n\underset{\sim}{j} = \underset{\sim}{j}'\underset{\sim}{n} = \underset{\sim}{n}'\underset{\sim}{j} = \underset{\sim}{n}$$

and

$$D_n\underset{\sim}{j} = \underset{\sim}{n}$$

which implies

$$\underset{\sim}{\gamma}'D_n\underset{\sim}{j} = \underset{\sim}{\gamma}'\underset{\sim}{n} = 0 \quad .$$

So

$$S_{\Delta}^{2}(T) = c\{\underset{\sim}{\gamma}'D_n\underset{\sim}{\gamma} + 0 + \left[\frac{(L\alpha-W\beta)^2}{4}\right]\cdot n\}$$

$$- \frac{c}{n}\{0\cdot 0 + 0 + \left[\frac{(L\alpha-W\beta)^2}{4}\right]\cdot n^2\}$$

$$= c\{\underset{\sim}{\gamma}'D_n\underset{\sim}{\gamma}\} + c\{\left[\frac{(L\alpha-W\beta)^2}{4}\right]\cdot n\} - \frac{c}{n}\{\left[\frac{(L\alpha-W\beta)^2}{4}\right]\cdot n^2\}$$

$$= S_U^2(T) + 0$$

or

$$S_{\Delta}^{2}(T) - S_U^2(T) = 0 \quad \text{(independent of } \alpha \text{ and } \beta\text{)}.$$

Therefore, when T is the index of matched partial correlation, i.e., when $\alpha=0$ and $\beta=1$, the components of U-statistics and the delta method estimates of the variance of T are equivalent.

CHAPTER IV

PLANNING A MONTE CARLO STUDY

4.1 Introduction

In this chapter we dicuss several topics which evolved during the planning of a Monte Carlo study to investigate the small-sample properties of the index of matched partial correlation (T) and estimates of its variance. In Sections 2, 3, and 5 we study the true value of T, i.e., θ, in general and in trivariate normal populations. Sections 2 and 3 are particularly intended to establish a setting for the discussion of the Monte Carlo study outlined in Section 4. In Section 6, we explore three numerical methods for approximating θ.

4.2 Population Value of θ

Suppose that we have a pair of random observations (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2) from some trivariate population. Let $U = X_1 - X_2$, $V = Y_1 - Y_2$, and $W = Z_1 - Z_2$. Let also $F(a,b|w)$ denote the conditional joint distribution function of U and V given W=w and let G(w) be the distribution function of W. Then

$$\Pr\{\text{Concordance} \mid W=w\} - \Pr\{\text{Discordance} \mid W=w\}$$
$$= \Pr\{UV>0 \mid W=w\} - \Pr\{UV<0 \mid W=w\}$$
$$= 1 - 2F(0,\infty \mid W) - 2F(\infty,0 \mid w) + 4F(0,0 \mid w).$$

If pairs are considered matched when $|Z_1 - Z_2| < \varepsilon$, then the index of matched partial correlation becomes

$$\theta = \frac{\int_{-\varepsilon}^{\varepsilon}\{1-2F(0,\infty \mid w) - 2F(\infty,0 \mid w) + 4F(0,0 \mid w)\}\, dG(w)}{\int_{-\varepsilon}^{\varepsilon} dG(w)} .$$

Now we impose a joint distribution on X, Y, and Z. In particular, we assume a general trivariate normal distribution in which

$$\begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \\ X_2 \\ Y_2 \\ Z_2 \end{pmatrix} \sim N_6 \left[\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \\ \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}, \begin{matrix} \underset{\sim}{\Sigma} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Sigma} \end{matrix} \right] \qquad \text{where } \underset{\sim}{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix}$$

Applying well-known results from multivariate normal theory, we have that

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, 2\underset{\sim}{\Sigma} \right]$$

and

$$\begin{pmatrix} U \\ V \end{pmatrix} \Bigg| W \sim N_2 \left[\begin{pmatrix} \mu_{U\cdot W} \\ \mu_{V\cdot W} \end{pmatrix}, \begin{pmatrix} \sigma_{U\cdot W}^2 & \sigma_{UV\cdot W} \\ \sigma_{UV\cdot W} & \sigma_{V\cdot W}^2 \end{pmatrix} \right]$$

where

$$\mu_{U\cdot W} = \frac{\sigma_{XZ}}{\sigma_Z^2} W \text{ and } \mu_{V\cdot W} = \frac{\sigma_{YZ}}{\sigma_Z^2} W \quad ,$$

$$\sigma_{U\cdot W}^2 = 2(\sigma_X^2 - \frac{\sigma_{XZ}^2}{\sigma_Z^2}) \;, \text{ and } \sigma_{V\cdot W}^2 = 2(\sigma_Y^2 - \frac{\sigma_{YZ}^2}{\sigma_Z^2}) \quad ,$$

and

$$\sigma_{UV\cdot W} = 2(\sigma_{XY} - \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma_Z^2}) \quad .$$

Now let

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \exp\left[-t^2/2\right]$$

and

$$\psi(t_1,t_2,\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(t_1^2-2\rho t_1 t_2+t_2^2)\right]$$

denote the standard univariate and bivariate normal densities, respectively. Then the corresponding distribution functions become

$$\Phi(a) = \int_{-\infty}^{a} \phi(t)dt$$

and

$$\Psi(a,b,\rho) = \int_{-\infty}^{a} \int_{-\infty}^{b} \psi(t_1,t_2,\rho)dt_2dt_1 \quad .$$

Using a simple transformation, we can write

$$F(a,b\ W) = \psi\left(\frac{a-\mu_{U\cdot W}}{\sigma_{U\cdot W}}, \frac{b-\mu_{V\cdot W}}{\sigma_{V\cdot W}}, \rho_{UV\cdot W}\right)$$

where

$$\rho_{UV\cdot W} = \frac{\sigma_{UV\cdot W}}{\sigma_{U\cdot W}\sigma_{V\cdot W}} = \frac{\rho_{XY}-\rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}} = \rho_{XY\cdot Z} \quad ,$$

i.e., $\rho_{UV\cdot W}$ is the product-moment partial correlation. Note that

$$F(a,\infty|W) = \Psi\left(\frac{a-\mu_{U\cdot W}}{\sigma_{U\cdot W}}, \infty, \rho_{UV\cdot W}\right) = \Phi\left(\frac{a-\mu_{U\cdot W}}{\sigma_{U\cdot W}}\right)$$

and

$$F(\infty,b|W) = \Psi\left(\infty, \frac{b-\mu_{V\cdot W}}{\sigma_{V\cdot W}}, \rho_{UV\cdot W}\right) = \Phi\left(\frac{a-\mu_{V\cdot W}}{\sigma_{V\cdot W}}\right) \quad .$$

Finally,

$$dG(W) = \frac{1}{\sqrt{2}\sigma_Z}\,\Phi\left(\frac{W}{\sqrt{2}\sigma_Z}\right)\,dW \quad .$$

Thus,

$$= \frac{\int_{-\epsilon}^{\epsilon}\left\{1-2\Phi\left(\frac{-\mu_{U\cdot W}}{\sigma_{U\cdot W}}\right)-2\Phi\left(\frac{-\mu_{V\cdot W}}{\sigma_{V\cdot W}}\right)+4\Psi\left(\frac{-\mu_{U\cdot W}}{\sigma_{U\cdot W}}, \frac{-\mu_{V\cdot W}}{\sigma_{V\cdot W}}, \rho_{UV\cdot W}\right)\right\}\frac{\phi\left(\frac{W}{\sqrt{2}\sigma_Z}\right)dW}{\sqrt{2}\sigma_Z}}{\int_{-\epsilon}^{\epsilon}\frac{\phi\left(\frac{W}{\sqrt{2}\sigma_Z}\right)}{\sqrt{2}\,\sigma_Z}\,dW} \quad .$$

Using the substitution $W = \sqrt{2}\, \sigma_Z t_3$, and writing out the various conditional parameters, we get

$$= \frac{\int_{-\gamma}^{\gamma} \{1-2\Phi(ht_3)-2\Phi(kt_3)+4\Psi(ht_3,kt_3,\rho_{xy\cdot z})\}\phi(t_3)\, dt_3}{\int_{-\gamma}^{\gamma} \phi(t_3)\, dt_3}$$

where

$$\gamma = \frac{\varepsilon}{\sqrt{2}\,\sigma_Z}\ , \quad h = \frac{-\rho_{XZ}}{\sqrt{1-\rho_{XZ}^2}}\ , \quad \text{and } k = \frac{-\rho_{YZ}}{\sqrt{1-\rho_{YZ}^2}} \quad .$$

In order to facilitate the evaluation of θ, we express θ in terms of single, double, and triple definite integrals:

$$\theta = \frac{P_1-2P_2-2P_3-4P_4}{P_1}$$

where

$$P_1 = \int_{-\gamma}^{\gamma} \phi(t_3)dT_3 = 2\Phi(\gamma)-1 \quad ,$$

$$P_2 = \int_{-\gamma}^{\gamma} \Phi(ht_3)\phi(t_3)dt_3 = \int_{-\gamma}^{\gamma}\int_{-}^{ht_3} \phi(t_1)\phi(t_3)dt_1dt_3 \quad ,$$

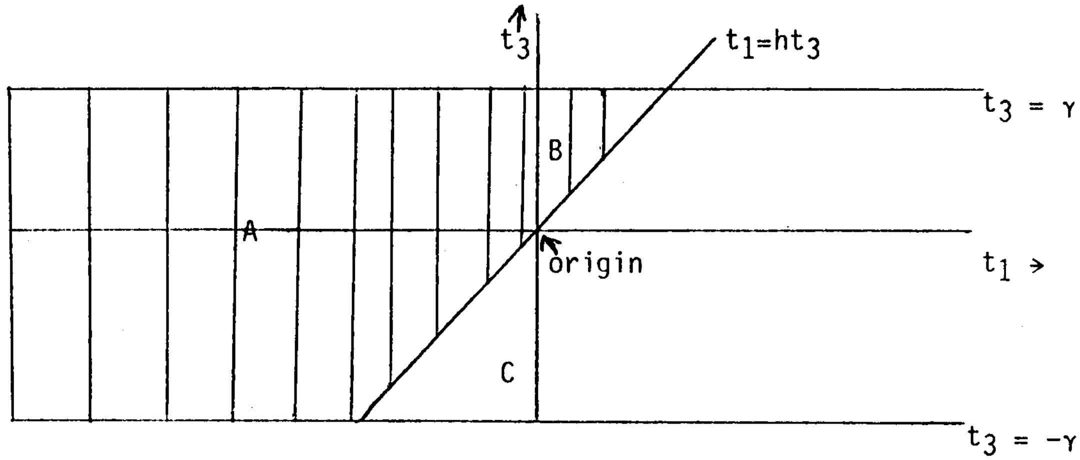
$$P_3 = \int_{-\gamma}^{\gamma} \Phi(kt_3)\phi(t_3)dt_3 = \int_{-\gamma}^{\gamma}\int_{-\infty}^{kt_3} \phi(t_2)\phi(t_3)dt_2dt_3 \quad ,$$

and

$$P_4 = \int_{-\gamma}^{\gamma} \Psi(ht_3,kt_3,\rho_{xy\cdot z})\phi(t_3)dt_3$$

$$= \int_{-\gamma}^{\gamma}\int_{-\infty}^{ht_3}\int_{-\infty}^{kt_3} \psi(t_1,t_2,\rho_{xy\cdot z})\phi(t_3)dt_2dt_1dt_3 \quad .$$

Consider the region of integration for P_2 in the figure below.



Thus we can write

$$P_2 = \iint_{A\cup B} \phi(t_1)\phi(t_3)dt_1dt_3$$

$$= \iint_A \phi(t_1)\phi(t_3)dt_1dt_3 + \iint_B \phi(t_1)\phi(t_3)dt_1dt_3 \quad .$$

Since the integrand is symmetric, i.e., since

$$\phi(t_1)\phi(t_3) = \phi(-t_1)\phi(-t_3) \quad ,$$

the integral over the region B must be the same as the integral over the region C. We now have that

$$P_2 = \iint_{A\cup C} \phi(t_1)\phi(t_3)dt_1dt_3$$

$$= \int_{-\gamma}^{\gamma}\int_{-\infty}^{0} \phi(t_1)\phi(t_3)dt_1dt_3$$

$$= \int_{-\gamma}^{\gamma}\phi(t_3)dt_3 \int_{-\infty}^{0}\phi(t_1)dt_1$$

$$= P_1 \cdot \Phi(0)$$

$$= \tfrac{1}{2}P_1 \quad .$$

By analogy, $P_3 = P_1$ and $\theta = \frac{4P_4}{P_1} - 1$. P_4 appears to admit no further simplification and, therefore, requires a numerical approximation. Three methods are discussed in Section 4.6.

4.3 What should $\underset{\sim}{\Sigma}$ be?

Although the possible values of $\underset{\sim}{\Sigma}$ are limitless, we have chosen four models of association which, we believe, cover a broad spectrum of the possibilities of interest. In the sequel, let the random variables denoted by A_{ji} be independent and identically distributed normal variates with mean zero and variance one.

Case 1: Z independent of X and Y

Let $X_i = (A_{1i} + cA_{2i})/\sqrt{1+c^2}$

$Y_i = (A_{3i} + cA_{2i})/\sqrt{1+c^2}$

$Z_i = A_{4i}$

Note that all X, Y, and Z variables are standardized, which provides convenience without loss of generality. We have then

$$E(X_iY_i) = \mathrm{Cov}(X_i,Y_i) = \rho_{xy} = \frac{c^2}{1+c^2}$$

and since $\rho_{xz} = \rho_{yz} = 0$, $\rho_{xy\cdot z} = \rho_{xy} = \rho.$, say,

$$c = \sqrt{\frac{\rho.}{1-\rho.}} \text{ and}$$

$$\underset{\sim}{\Sigma} = \begin{pmatrix} 1 & \rho. & 0 \\ \rho. & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad |\rho.| \leq 1.$$

Case 2: Common Cause and Intervening Variable Models (Spurious Correlation)

Suppose that Z is a common cause of the variables X and Y. We illustrate this relationship via the following path diagram, a device developed by Wright (1918) in his treatise that laid the foundation of path analysis:

$$\begin{tikzcd} & X \\ Z \arrow[ur] \arrow[dr] & \\ & Y \end{tikzcd}$$

Let

$$X_i = (cA_{1i} + A_{3i}) / \sqrt{1+c^2}$$

$$Y_i = (cA_{2i} + A_{3i}) / \sqrt{1+c^2}$$

$$Z_i = A_{3i}$$

Then

$$E(X_i Y_i) = \text{Cov}(X_i Y_i) = \rho_{xy} = \frac{1}{1+c^2} = \rho^2 ,$$

say. Also,

$$E(X_i Z_i) = \text{Cov}(X_i, Z_i) = \rho_{xz} = \frac{1}{\sqrt{1+c^2}} = \rho$$

and, by design, $\rho_{yz} = \rho$. Thus,

$$c = \frac{\sqrt{1-\rho^2}}{\rho} \text{ and } \underset{\sim}{\Sigma} = \begin{pmatrix} 1 & \rho^2 & \rho \\ \rho^2 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} , \ |\rho| \leq 1 .$$

Of course, $\rho_{XY \cdot Z} = 0$. Note further that the intervening variable model

$$X \longrightarrow Z \longrightarrow Y \quad (\text{or} \quad Y \longrightarrow Z \longrightarrow X)$$

gives the same form of $\underset{\sim}{\Sigma}$.

Case 3: Common Effect Model

The model in Case 3 assumes that the variable Z is a common effect of X and Y, i.e.,

$$\begin{tikzcd} X \arrow[dr] & \\ & Z \\ Y \arrow[ur] & \end{tikzcd} .$$

Let
$$X_i = A_{1i}$$
$$Y_i = A_{2i}$$
$$Z_i = (A_{1i} + A_{2i} + cA_{3i})/\sqrt{2+c^2} .$$

Then

$$E(X_i Z_i) = \text{Cov}(X_i, Z_i) = \rho_{XZ} = \frac{1}{\sqrt{2+c^2}} .$$

This time we let $\rho_{XZ} = \rho$ and $\rho_{YZ} = -\rho$ in order to produce a positive partial correlation, $\rho_{xy \cdot z} = \rho.$, say. Thus

$$\rho = \sqrt{\frac{\rho.}{1+\rho.}} , \quad c = \sqrt{\frac{1-\rho.}{\rho.}} ,$$

and

$$\underset{\sim}{\Sigma} = \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & -\rho \\ \rho & -\rho & 1 \end{pmatrix} , \quad |\rho| \le \frac{1}{\sqrt{2}} .$$

Case 4: Developmental Sequence

Reynolds (1974) considered a causal relationship that assumed Z influenced Y which, in turn, influenced X. The following path diagram illustrates this model, which he called a "developmental sequence":

$$Z \longrightarrow Y \longrightarrow X$$

(Equivalently, one might have $Z \longrightarrow X \longrightarrow Y$.)

Let $X_i = (bA_{1i} + cA_{2i} + A_{3i}) / \sqrt{b^2+c^2+1}$

$Y_i = (cA_{2i} + A_{3i}) / \sqrt{1+c^2}$

$Z_i = A_{3i}$.

Then

$$E(X_i Y_i) = \mathrm{Cov}(X_i, Y_i) = \rho_{xy} = \frac{\sqrt{1+c^2}}{\sqrt{1+b^2+c^2}} \quad .$$

$$E(X_i Z_i) = \mathrm{Cov}(X_i, Z_i) = \rho_{xz} = \frac{1+c^2}{\sqrt{1+b^2+c^2}} \quad .$$

and

$$E(Y_i Z_i) = \mathrm{Cov}(Y_i Z_i) = \rho_{yz} = \frac{1}{\sqrt{1+c^2}} \quad .$$

In order to represent the covariance structure as a function of only one parameter, we set $\rho_{xy} = \rho_{yz} = \rho$ and $\rho_{xy\cdot z} = \rho.$, say.

This yields

$$\rho_{xz} = \rho^2 \ , \ \rho = \frac{\rho.}{\sqrt{1-\rho.^2}} \quad ,$$

$$c = \frac{\sqrt{1-2\rho.^2}}{\rho.} \quad , \qquad b = \frac{c\sqrt{1-2\rho.^2}}{\rho.} \quad ,$$

and

$$\underset{\sim}{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}, \quad |\rho| \leq 1 \quad .$$

This formulation, however, restricts $\rho.$ such that $|\rho.| \leq \frac{1}{\sqrt{2}}$.

In summary, we have presented four models whose covariance structures can each be conveniently expressed as a function of only one parameter, a fact whose usefulness will become apparent in the Monte Carlo studies. Also, formulating the variates as linear combinations of univariate normal deviates suggests a method of generating the Monte Carlo data.

4.4 Design of the Monte Carlo Study

In the preceding section, we have discussed θ in terms of two parameters, the partial correlation, $\rho.$, and the tolerance for a match, ϵ. Thus we are interested in the behavior of θ and its sample estimates along with their variances for various values of $\rho.$ and ϵ. We chose $\epsilon = .1(.1).5$, which allows matches in a range from .2 to 1.0 standard deviation, and $\rho. = 0, .2, .5, .7, .9$, which covers the (positive) range of partial correlations. Of course, we needed to conduct the study for each of the four aforementioned cases. Note, however, that $\rho.$ is identically zero in Case 2. In this instance we decided to vary the total correlations, denoted by ρ, over the same range, i.e., $\rho = 0, .2, .5, .7, .9$. Some computational savings are made by noting that for $\rho. = 0$ and $\rho = 0$, Cases 1 through 4 are identical. These values of the parameters would suggest the design illustrated by Figure 4.1. The '—' for the $\rho. = .9$, CASE = 4 combination signifies an empty cell in the design. Recall from Section 3 that $|\rho.| \leq 1/\sqrt{2} = .707$ for Case 4. If we call each case-by-correlation-by-tolerance combination a scenario, our design comprises 80 scenarios.

Now let n denote the sample size under investigation. Then for each case-by-correlation combination we must generate

FIGURE 4.1

Monte Carlo Design

<u>CASE</u>	0	.2	.5	.7	.9
			ρ.		
1		$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$
3		$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$
4	$\left\{\underset{\sim}{\varepsilon}\right\}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$
			ρ		
2		$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$	$\underset{\sim}{\varepsilon}$

Note: $\underset{\sim}{\varepsilon}$ denotes the five values of ε: (.1, .2, .3, .4, .5).

n random observations with the appropriate multivariate normal distribution. Next we calculate θ, T, T_J, and the variances of T and T_J for each combination. This will, of course, be replicated M times, where M is some number we choose beforehand. We decided to fix M at 100, since Kleijnen (1974) recommends a relatively smaller number of replicates to examine the behavior of a design under different conditions (as opposed to a relatively larger number of replicates when a precise estimate for a particular condition is desired). Thus for each n to be investigated we must generate 16M = 1600 samples of size n; and for each sample we must calculate the sample statistics 5 times, once for each value of ϵ, or 8000 times in all. Obviously, the calculation of each sample statistic 8000 times for each n would require considerable computation time, so we conducted a study to determine the feasibility of the proposed design. The results are described in Appendix A. We concluded that the jackknife estimation process would require too much computation time to make the full design possible; therefore, we eliminated the calculation of T_J and $S_J^2(T_J)$ from the Monte Carlo study.

4.5 Behavior of θ in the Models

Before investigating the behavior of θ in special cases of each of our four models, we first consider its characteristics for both infinite and zero tolerances. Let

$$Q(w) = 1-2F(0,\infty|w) - 2F(\infty,0\ w) + 4F(0,0|w)$$

where, as in section 4.2, $F(U,V|w)$ is the conditional joint distribution function of $U = X_1-X_2$ and $V = Y_1-Y_2$ given $W = Z_1-Z_2 = w$. Then

$$\theta = \theta(\varepsilon) = \frac{\int_{-\varepsilon}^{\varepsilon} Q(w)dG(w)}{\int_{-\varepsilon}^{\varepsilon} dG(w)}$$

where $G(w)$ is the marginal distribution function of W. First, consider "$\theta(\infty)$", where we are writing

$$\theta(\infty) = \lim_{\varepsilon \to \infty} \theta(\varepsilon).$$

For this we return to first principles. With infinite tolerance the denominator of θ is identically 1 and the numerator is

$$\Pr\{UV<0\} - \Pr\{UV>0\}$$

which is the total Kendall correlation between X and Y (ignoring Z). In a trivariate normal population, then,

$$\theta(\infty) = \frac{2}{\pi} \sin^{-1}(\rho_{XY}).$$

Next to find

$$\lim_{\varepsilon \to \infty} \theta(\varepsilon),$$

let

$$Q_U(\varepsilon) = \sup_{-\varepsilon<W<\varepsilon} Q(w)$$

and

$$Q_L(\varepsilon) = \inf_{-\varepsilon<W<\varepsilon} Q(w).$$

Then clearly

$$Q_L(\varepsilon) \leq \theta(\varepsilon) \leq Q_U(\varepsilon).$$

Hence, if $Q(w)$ is continuous in w at $w=0$, it follows that

$$\lim_{\varepsilon \to 0} \theta(\varepsilon) = Q(0).$$

In the trivariate normal,

$$Q(w) = 1-2\Phi(\alpha w) -2\Phi(\beta w) +4\Psi(\alpha w,\beta w,\rho.)$$

where

$$\alpha = \frac{-\rho_{XZ}}{\sigma_Z\sqrt{2}\sqrt{1-\rho^2_{XZ}}} \quad \text{and } \beta = \frac{-\rho_{YZ}}{\sigma_Z\sqrt{2}\sqrt{1-\rho^2_{YZ}}} .$$

Thus,

$$\begin{aligned} Q(o) &= 1-2\Phi(0) -2\Phi(0) + 4\Psi(0,0,\rho.) \\ &= 1 - 1 - 1 + 4\left[\tfrac{1}{4} + \frac{1}{2\pi} \sin^{-1}(\rho.)\right] \\ &= \frac{2}{\pi} \sin^{-1}(\rho.) \end{aligned}$$

which is the Kendall correlation in the conditional distribution of X and Y given any fixed value of Z.

Now we study θ under various conditions in each of the four cases.

Case 1:

Here $\alpha=0$ and $\beta=0$ since $\rho_{XZ}=0$ and $\rho_{YZ}=0$, respectively. Hence,

$$Q(w) = Q(0) \text{ for all } w$$

and

$$\theta(\varepsilon) = Q(0) = \frac{2}{\pi} \sin^{-1}(\rho.) \text{ for all } \varepsilon;$$

thus,

$$\theta(0) = \theta(\infty) = \frac{2}{\pi} \sin^{-1}(\rho.).$$

Therefore, θ is independent of the tolerance and depends only on the true product-moment partial correlation.

Case 2:

In the expression for $\theta(\varepsilon)$, transform from t to $\dfrac{W}{\sigma_Z\sqrt{2}}$

and let

$$\lambda = \frac{\rho}{\sqrt{1-\rho^2}} \text{ and } \gamma = \frac{\varepsilon}{\sigma_Z\sqrt{2}} .$$

Then, since $\rho_{XZ} = \rho_{YZ} = \rho$ in Case 2,

$$\theta = \theta(\gamma) = \frac{\int_{-\gamma}^{\gamma} \{1-4\Phi(-\lambda t) + 4\Psi(-\lambda t,-\lambda t,\rho.)\}\phi(t)dt}{\int_{-\gamma}^{\gamma}\phi(t)dt} .$$

Note that the denominator of θ is simply 2 $(\gamma)-1$. Also in Case 2, $\rho.=0$ so

$$\Psi(-\lambda t,-\lambda t,\rho.) = [\Phi(-\lambda t)]^2.$$

Now suppose that $\rho = \frac{1}{\sqrt{2}}$; then $\lambda = 1$ and the numerator of θ becomes

$$\int_{-\gamma}^{\gamma} \{1-4\Phi(-t) + 4[\Phi(-t)]^2\}\phi(t)dt$$

$$= \int_{-\gamma}^{\gamma}[1-2\Phi(-t)]^2\phi(t)dt$$

$$= [1-2\Phi(-t)]^3 /6\Big|_{-\gamma}^{\gamma}$$

$$= [2\Phi(\gamma)-1]^3 /3$$

Thus

$$\theta = [2\Phi(\gamma) -1]^2 /3$$

when $\rho = 1/\sqrt{2}$ in Case 2.

Furthermore,

$$\theta(0) = 0 ,$$

and

$$\theta(\infty) = \frac{2}{\pi} \sin^{-1}(\rho_{XY}) .$$

Case 3:

The degenerate situation when $\rho. = 1$ appears to be the only instance which admits simplification of θ for Case 3, in which case θ is also 1 (and $\rho = 1/\sqrt{2}$). However,

$$\theta(0) = \frac{2}{\pi} \sin^{-1}(\rho.)$$

and

$$\theta(\infty) = 0 .$$

Case 4:

Unfortunately, no simplification is possible for Case 4 except when the tolerance is zero or infinite. Then we have

$$\theta(0) = \frac{2}{\pi} \sin^{-1}(\rho.)$$

and

$$\theta(\infty) = \frac{2}{\pi} \sin^{-1}\left(\frac{\rho.}{\sqrt{1-\rho.^2}}\right).$$

The "nice" forms of θ found for the above situations provide a means of checking and comparing the approximation methods which we discuss in the next section.

4.6 Three Methods for Approximating P_4

In this section we investigate three numerical methods which can be used to evaluate P_4 and hence, θ, for our scenarios. In particular, we consider the following:

(i) numerical integration of P_4

(ii) Maclaurin series expansion of P_4 in γ

(iii) expansion of the bivariate normal distribution function portion of P_4 in ρ.

(i) Numerical integration of P_4

Stroud (1971) provides an algorithm for numerically evaluating a three-dimensional iterated integral by the 16-point Gauss-Legendre method. We must write P_4 as an iterated integral in order to apply the method:

$$P_4 = \int_{-\gamma}^{\gamma} \int_{-\infty}^{ht_3} \int_{-\infty}^{\frac{kt_3-\rho_{xy.z}t_1}{\sqrt{1-\rho_{xy.z}^2}}} \phi(t_1)\phi(t_2)\phi(t_3)dt_2dt_1dt_3$$

since

$$\int_{-\infty}^{kt_3}\psi(t_1,t_2,\rho)dt_2 = \int_{-\infty}^{\frac{kt_3-\rho t_1}{\sqrt{1-\rho^2}}} \phi(t_1)\phi(t_2)dt_2.$$

P_4 is now in the form required by Stroud's algorithm with the exceptions of the "$-\infty$'s" which are the lower boundaries of the region of integration for the innermost integrals. In this situation some arbitrary, large, negative number must be chosen which will produce a good approximation. We chose -13 since it was the largest negative number which did not cause any underflow conditions during the implementation of this algorithm; the accuracy of the approximation is discussed at the end of this section.

(ii) Maclaurin series expansion of P_4

Before beginning the expansion, notice the following relationships among the univariate and bivariate normal densities and distributions and their derivatives.

a. $\Phi'(ht_3) = h\phi(ht_3)$

b. $\phi'(ht_3) = -h^2t_3\phi(ht_3)$

c. $\Psi(ht_3,kt_3,\rho) = \int_{-\infty}^{ht_3} \int_{-\infty}^{kt_3} (t_1,t_2,\rho)dt_2dt_1$

$$= \int_{-\infty}^{ht_3} \int_{-\infty}^{\frac{kt_3 - \rho t_1}{\sqrt{1-\rho^2}}} \phi(t_1)\phi(t_2)dt_2dt_1$$

$$= \int_{-\infty}^{ht_3} \phi(t_1)\ \Phi\ \left(\frac{kt_3-\rho t_1}{\sqrt{1-\rho^2}}\right)$$

d. $\Psi'(ht_3,kt_3,\rho) = h\phi(ht_3)\Phi\left(\frac{kt_3-\rho ht_3}{\sqrt{1-\rho^2}}\right)$

$$+ \int_{-\infty}^{ht_3} k\phi(t_1)\phi\ \left(\frac{kt_3-\rho t_1}{\sqrt{1-\rho^2}}\right)\ dt_1$$

$$= h\phi(ht_3)\Phi\ \left(\frac{kt_3-\rho ht_3}{\sqrt{1-\rho^2}}\right) + \int_{-\infty}^{ht_3} \frac{k}{\sqrt{2\pi}\sqrt{1-\rho^2}}$$

$$\cdot \exp\left[-\frac{(t_1^2+k^2t_3^2-2k\rho t_3t_1+\rho^2t_1^{\ 2})}{2(1-\rho^2)}\right]\ dt_1$$

$$= h\phi(ht_3)\Phi\ \left(\frac{kt_3-\rho ht_3}{\sqrt{1-\rho^2}}\right)$$

$$+ \int_{-\infty}^{ht_3} \frac{k}{\sqrt{2\pi}} \exp\left[-\frac{(k^2t_3^{\ 2}-k^2t_3^{\ 2}\rho^2)}{2(1-\rho^2)}\right] \frac{\exp\left[-\frac{(t_1-kt_3\rho)^2}{2(1-\rho^2)}\right]}{\sqrt{2\pi}\ \sqrt{1-\rho^2}}\ dt_1$$

$$= h\phi(ht_3)\Phi\ \left(\frac{kt_3-\rho ht_3}{\sqrt{1-\rho^2}}\right) + k\phi(kt_3)\Phi\ \left(\frac{ht_3-kt_3\rho}{\sqrt{1-\rho^2}}\right)$$

Note that this is an example of the differentiation of a definite integral with respect to a parameter. The solution is widely available -- see Gradshteyn and Ryzhik (1980), for example -- and is presented below:

$$\frac{d}{da}\int_{\Psi(a)}^{\zeta(a)} f(x,a)dx = f(\zeta(a),a)\ \frac{d\zeta(a)}{da} - f(\Psi(a),a)\ \frac{d\Psi(a)}{da}$$

$$+ \int_{\Psi(a)}^{\zeta(a)} \frac{df(x,a)}{da}\ dx.$$

Recall that the MacLaurin series for a function f(x) is useful for approximating f for values of x near zero. It is given by

$$f(x) \doteq f(0) + f'(0)x + \frac{f''(0)}{2!}\ x^2 + \ldots + \frac{f^{(n)}(0)}{n!}\ x^{n} + \ldots,$$

where $f^{(n)}(x)$ denotes the nth-derivative of f with respect to x. We need a few more preliminaries before we derive the expansion of P_4 as a function of γ. Let

$$g(\gamma) = \int_0^{\gamma} h(t_3)dt_3;\ g(-\gamma) = \int_0^{-\gamma} h(t_3)dt_3 = -\int_{-\gamma}^{0} h(t_3)dt_3.$$

Thus, if

$$f(\gamma) = \int_{-\gamma}^{\gamma} h(t_3)dt_3,$$

then $f(\gamma) = g(\gamma) - g(\gamma)$. Furthermore, we have that

$$f^{(n)}(0) = \begin{cases} 0, \text{ for n even (including n=0)} \\ 2g^{(n)}(0), \text{ for n odd} \end{cases}$$

Therefore, the Maclaurin series expansion for this particular f is given by

$$f(\gamma) \doteq 2g'(0)\gamma + \frac{g'''(0)\gamma^3}{3} + \ldots + \frac{2g^{(2n+1)}(0)\gamma^{2n+1}}{(2n+1)!} + \ldots$$

In our application,

$$f(\gamma) = P_4$$

and

$$g(\gamma) = \int_0^{\gamma} \Psi(ht_3, kt_3, \rho)\phi(t_3)dt_3$$

(for notational convenience, we have dropped the subscripts from ρ), so

$$g'(\gamma) = \Psi(h\gamma, k\gamma, \rho)\phi(\gamma),$$

$$g''(\gamma) = \Psi'(h\gamma, k\gamma, \rho)\phi(\gamma) - \gamma\Psi(h\gamma, k\gamma, \rho)\phi(\gamma)$$

$$= \Psi'(h\gamma, k\gamma, \rho)\phi(\gamma) - \gamma g'(\gamma),$$

and

$$g'''(\gamma) = \Psi''(h\gamma, k\gamma, \rho)\phi(\gamma) - \gamma\Psi'(h\gamma, k\gamma, \rho)\phi(\gamma) - g'(\gamma) - \gamma g''(\gamma).$$

Now

$$\Psi'(h\gamma, k\gamma, \rho) = h\phi(h\gamma)\Phi\left[\frac{(k-\rho h)\gamma}{\sqrt{1-\rho^2}}\right] + k\phi(k\gamma)\Phi\left[\frac{(h-k\rho)\gamma}{\sqrt{1-\rho^2}}\right].$$

Thus

$$\Psi''(h\gamma, k\gamma, \rho) = \frac{h(k-\rho h)}{\sqrt{1-\rho^2}}\,\phi(h\gamma)\phi\left[\frac{(k-\rho h)\gamma}{\sqrt{1-\rho^2}}\right] - h^3\gamma\phi(h\gamma)\Phi\left[\frac{(k-\rho h)\gamma}{\sqrt{1-\rho^2}}\right]$$

$$+ \frac{k(h-\rho k)}{\sqrt{1-\rho^2}}\,\phi(k\gamma)\,\phi\left[\frac{(h-k\rho)\gamma}{\sqrt{1-\rho^2}}\right] - k^3\gamma\phi(k\gamma)\Phi\left[\frac{(h-k\rho)\gamma}{\sqrt{1-\rho^2}}\right].$$

Hence,

$$g'(0) = \left(\tfrac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}\right)\frac{1}{\sqrt{2\pi}} = \frac{1}{2\pi\sqrt{2\pi}}\left(\frac{\pi}{2} + \sin^{-1}\rho\right)$$

and

$$g'''(0) = \left[\frac{2hk-\rho(h^2+k^2)}{2\pi\sqrt{1-\rho^2}}\right]\frac{1}{\sqrt{2\pi}} - \left(\tfrac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}\right)\frac{1}{\sqrt{2\pi}}$$

$$= \frac{1}{2\pi\sqrt{2\pi}} \left[\frac{2hk-\rho(h^2+k^2)}{\sqrt{1-\rho^2}} - \frac{\pi}{2} - \sin^{-1}\rho \right].$$

Finally,

$$P_4 \doteq \frac{1}{2\pi\sqrt{2\pi}} \{(\pi+2\sin^{-1}\rho)\gamma + 1/3 \left[\frac{2hk-\rho(h^2+k^2)}{\sqrt{1-\rho^2}} - \frac{\pi}{2} - \sin^{-1}\rho \right] \gamma^3\}.$$

(iii) Expansion of the Bivariate Normal Distribution Function Portion of θ in ρ

Once again for notational convenience, we drop the subscript from ρ. Let

$$L(a,b,\rho) = \int_a^\infty \int_b^\infty \psi(t_1,t_2,\rho)dt_1dt_2 .$$

Pearson (1901) obtained the expansion

$$L(a,b,\rho) = \phi(a)\phi(b) + \phi(a)\phi(b) \left[\rho + \frac{\rho^2}{2!} ab + \frac{\rho^3}{3!}(a^2-1)(b^2-1)+\ldots\right]$$

$$= \sum_{i=0}^{\infty} \delta_i(a)\delta_i(b)\rho^i ,$$

where

$$\delta_i(c) = \frac{(-1)^{i-1}}{\sqrt{i!}} \frac{d^{i-1}\phi(c)}{dc} \quad \text{for } i \geq 1$$

and

$$\delta_0(c) = \Phi(c).$$

Note first that

$$\Psi(a,b,\rho) = \Phi(a) + \Phi(b) + L(a,b,\rho) - 1.$$

Next, recall that

$$P_4 = \int_{-\gamma}^{\gamma}\Psi(ht,kt,\rho)\phi(t)dt.$$

Hence,

$$P_4 = \int_{-\gamma}^{\gamma} [\Phi(ht) + \Phi(kt) - 1]\ \phi(t)dt + \int_{-\gamma}^{\gamma} L(ht,kt,\rho)\phi(t)dt$$

$$= P_2 + P_3 - P_1 + \int_{-\gamma}^{\gamma} L(ht,kt,\rho)\phi(t)dt$$

$$= \tfrac{1}{2}P_1 + \tfrac{1}{2}P_1 - P_1 + \int_{-\gamma}^{\gamma} L(ht,kt,\rho)\phi(t)dt$$

$$= \int_{-\gamma}^{\gamma} L(ht,kt,\rho)\phi(t)dt.$$

Thus, if we use Pearson's expansion, P_4 becomes

$$P_4 = \int_{-\gamma}^{\gamma} \{\Phi(ht)\Phi(kt)$$

$$+ \phi(ht)\phi(kt) \left[\rho + \frac{\rho^2}{2!} hkt^2 + \frac{\rho^3}{3!} (h^2t^2-1)(k^2t^2-1) + \ldots\right] \} \phi(t)dt .$$

Unfortunately, no further simplification appears possible.

We recognize the following difficulties with Pearson's expansion for our problem. If we chose to approximate P_4 using this method we would first have to determine an appropriate term in the expansion at which to truncate the infinite series. Then, we would have to use numerical integration on this "approximate" integrand to obtain a value of P_4. Furthermore, Pearson pointed out that this method is only accurate for small ρ, but we have some large values of ρ in our design. Therefore, we believe that Pearson's expansion is not a suitable candidate for a method to approximate P_4.

In Section 4.5 we investigated the behavior of θ in each of our four models for special cases of the covariance structure. The results provide us with some easily obtainable values of θ with which to compare the Maclaurin series expansion of P_4, and hence, θ, with the same using numerical integration. Computer programs were written in FORTRAN to compare the two methods for the special cases when $\rho_{\cdot}=0$, $\rho=1/\sqrt{2}$, and $\varepsilon=.1(.1).5$. These programs are included in Appendix B. We present the outcomes of this comparison in Table 4.1 and we use the following notation: θ is, of course, the true value of the index of matched partial correlation and, for these special cases,

$$\theta = [2\Phi(\gamma)-1]^2/3$$

where $\gamma=\varepsilon/\sqrt{2}$; θ_I denotes the value of θ approximated by numerical integration; θ_M denotes the value of θ approximated by the Maclaurin series expansion; $|\theta-\theta_I|$ and $|\theta-\theta_M|$ represent the obvious absolute differences. We observe that θ_I differs from θ by at most 10^{-5} while θ_M differs from θ by 10^{-5} in the best case. Hence, we conclude that θ_I provides use with a much better approximation of θ in these special cases. Furthermore, we do not

TABLE 4.1

A Comparison of the Maclaurin Series and Numerical Integration Approximation of $\dot{\theta}$ for Case 2, $\rho = 1/\sqrt{2}$, and $\epsilon = .1(.1).5$

ϵ	θ	θ_I	$\theta-\theta_I$	$\lvert\theta_M\rvert$	$\lvert\theta-\theta_M\rvert$
.1	$.106\times10^{-2}$	$.105\times10^{-2}$	$.795\times10^{-5}$	$.105\times10^{-2}$	$.619\times10^{-5}$
.2	$.422\times10^{-2}$	$.421\times10^{-2}$	$.289\times10^{-5}$	$.425\times10^{-2}$	$.292\times10^{-4}$
.3	$.941\times10^{-2}$	$.941\times10^{-2}$	$.105\times10^{-5}$	$.957\times10^{-2}$	$.161\times10^{-3}$
.4	$.165\times10^{-1}$	$.165\times10^{-1}$	$.385\times10^{-6}$	$.170\times10^{-1}$	$.510\times10^{-3}$
.5	$.255\times10^{-1}$	$.255\times10^{-1}$	$.260\times10^{-6}$	$.267\times10^{-1}$	$.123\times10^{-2}$

compare the two methods for any other special cases since numerical integration is the overwhelming method of choice.

Finally, we present the values of θ, approximated by numerical integration, for each scenario in our Monte Carlo design in Table 4.2. This table clearly illustrates the patterns of θ which we expect in our study. In particular, θ increases with $\rho_{\cdot}$ (or ρ in Case 2) for every case; θ also increases with ϵ in Cases 2 and 4, it is constant for all ϵ in Case 1, and it decreases with ϵ in Case 3.

TABLE 4.2

Values for θ For The Monte Carlo Design

Approximated by Numerical Integration *

		Tolerance						
Case	ρ.**	0	.1	.2	.3	.4	.5	∞
	0	0	0	0	0	0	0	0
	.2	.128	.128	.128	.128	.128	.128	.128
1	.5	.333	.333	.333	.333	.333	.333	.333
	.7	.494	.494	.494	.494	.494	.494	.494
	.9	.713	.713	.713	.713	.713	.713	.713
	.99***	.910	.910	.910	.910	.910	.910	.910
	0	0	0	0	0	0	0	0
	.2	0	.000	.000	.000	.001	.001	.128
2	.5	0	.000	.001	.003	.006	.009	.333
	.7	0	.001	.004	.009	.016	.024	.494
	.9	0	.004	.018	.039	.067	.100	.713
	.99***	0	.050	.173	.318	.446	.543	.910
	0	0	0	0	0	0	0	0
	.2	.128	.128	.127	.126	.124	.122	0
3	.5	.333	.332	.330	.325	.319	.311	0
	.7	.494	.492	.487	.478	.466	.452	0
	.9	.713	.709	.697	.677	.652	.621	0
	.99***	.910	.906	.865	.808	.755	.714	0
	0	0	0	0	0	0	0	0
	.2	.128	.128	.128	.128	.128	.128	.131
4	.5	.333	.333	.334	.334	.335	.336	.392
	.7	.494	.500	.518	.546	.579	.615	.873

* Exact values for ρ. or ρ=0, and for tolerance = 0 or ∞.

** ρ in Case 2.

*** ρ. = .99 is discussed in Chapter 5 and is included here for future reference.

CHAPTER V

THE SMALL-SAMPLE PROPERTIES OF T AND THE COMPONENTS OF U-STATISTICS ESTIMATE OF ITS VARIANCE

5.1 Preliminaries

In this chapter we investigate empirically the small-sample properties of T and $S_U^2(T)$. More specifically, we address the following questions:

1. How good is T as an estimator of θ?
2. How good is $S_U(T)/\sqrt{n}$ as an estimator of the standard deviation of T?
3. Is $\sqrt{n}(T-\theta)/S_U(T)$ approximately distributed as a standard normal random variable?

To provide the necessary data, we conducted a Monte Carlo study for samples of size 50, 25, and 10 according to the protocol set forth in the preceding chapter. We took 100 replications of each scenario (case by correlation by tolerance combination) at each sample size. The data are summarized using various descriptive statistics which are focused on the specific questions above. In order to discuss these measures, we introduce some notation and definitions.

Let the matched partial correlation index and its computed standard error obtained for the i-th replication of a scenario for a given sample size be denoted by T_i and S_i, respectively, $i=1(1)100$. Note that T_i and S_i are undefined if there are no matched pairs. Let M be the number of times T_i and S_i are defined; M is usually 100, except for a few scenarios of the size 10 samples. Now define

$$Z_i = \begin{cases} \dfrac{T_i-\theta}{S_i} & \text{if } S_i > 0, \\ +\infty \text{ if } S_i=0 \text{ and } T_i>\theta, \\ -\infty \text{ if } S_i=0 \text{ and } T_i<\theta, \\ \text{undefined if } T_i \text{ is undefined} \end{cases}$$

where θ is the true population value of the index of matched partial correlation for the scenario. (Recall that θ is known exactly for Case 1 but must be approximated by numerical integration otherwise.)

For Question 1 we first consider the mean and standard deviation of the T_i, which we denote $\bar{T}$ and S_T, respectively, where

$$\bar{T} = \frac{1}{M}\Sigma T_i \text{ and } S_T^2 = \frac{1}{M-1}\Sigma\,(T_i-\bar{T})^2$$

with the sum over all replications for which T_i is defined. We estimate the bias in T as an estimator of θ by $b_T = \bar{T}-\theta$. We expect some bias in T except when $\theta=0$, in which case it seems obvious that T should be symmetrically distributed about θ. Also, we compute an estimate of the root mean squared error of T, say S_E, where

$$S_E^2 = \frac{1}{M}\Sigma\,(T_i-\theta)^2.$$

For Question 2, the quality of $S_U(T)/\sqrt{n}$ as an estimator of the standard deviation of T is examined through its estimated bias which we denote b_S. This is defined as $b_S = \bar{S}-S_T$ where

$$\bar{S} = \frac{1}{M}\,\Sigma S_i\ .$$

We expect b_S to be small relative to S_T.

Finally, for Question 3 we are, of course, interested in the standard normality of the Z_i. We pursue this question primarily through the significance or non-significance of the two-sided Kolmogorov-Smirnov goodness-of-fit statistic.

$$D = \sup_{i=1(1)M}\ \sup\{|a_{(i)} - \frac{i-1}{M}\ ,\ \frac{i}{M} - a_{(i)}|\}$$

where $a_{(i)}$ is the i-th order statistic of

$$a_i = \begin{cases} \Phi(Z_i) \text{ if } S_i>0, \\ \dfrac{1+\operatorname{sgn}(T_i-\theta)}{2} \text{ if } S_i = 0, \end{cases} \quad i = 1(1)M.$$

The asymptotic critical values for testing D at the .05 and .01 levels of significance are $1.36/\sqrt{M}$ and $1.63/\sqrt{M}$, respectively.

Secondly, we examine the behavior of the tails of the distribution of the Z_i. If Z_i is standard normal, we expect .1M values of the Z_i to be greater than 1.645 in absolute value, .05M to be greater than 1.960, and .01M to be greater than 2.576. Thus, we count the occurrences of these events in our data.

We also consider the mean $\bar{Z}$ and the standard deviation S_Z of the Z_i:

$$\bar{Z} = \frac{1}{M^*}\Sigma Z_i \quad \text{and} \quad S_Z^2 = \frac{1}{M^*-1}\Sigma(Z_i-\bar{Z})^2,$$

where M* is the number of times that $S_i>0$, and each sum is taken over these M* values of Z_i. We cannot, however, place as much weight on the proximity of these quantities to their asymptotic values, which are 0 and 1, respectively. This is because $S_i=0$ occurs with positive probability, so that, strictly speaking, the population moments of the Z_i do not exist.

5.2 Results for n = 50 (or more)

Tables 5.1-5.4 summarize the data generated for Cases 1-4, respectively, in the Monte Carlo study for samples of size 50. Note that '*' and '**' denote significance of D at the .05 and .01 levels, respectively. This notation is used throughout the sequel.

TABLE 5.1 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 50

CASE=1

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.020	0.151	0.020	0.151	0.188	0.038	5	2	1	0.123	0.117	0.850
0.0	0.2	0.000	100	100	0.003	0.134	0.003	0.133	0.146	0.012	9	4	2	0.065	0.028	0.954
0.0	0.3	0.000	100	100	0.007	0.116	0.007	0.115	0.131	0.016	5	4	1	0.091	0.056	0.910
0.0	0.4	0.000	100	100	0.012	0.108	0.012	0.108	0.123	0.016	7	3	1	0.092	0.098	0.901
0.0	0.5	0.000	100	100	0.012	0.101	0.012	0.101	0.119	0.018	6	3	0	0.094	0.113	0.874
0.2	0.1	0.128	100	100	0.134	0.165	0.006	0.165	0.185	0.019	10	5	1	0.081	0.058	0.945
0.2	0.2	0.128	100	100	0.139	0.130	0.011	0.130	0.147	0.017	10	6	1	0.112	0.110	0.928
0.2	0.3	0.128	100	100	0.145	0.115	0.017	0.116	0.133	0.017	11	5	1	0.125	0.172	0.926
0.2	0.4	0.128	100	100	0.144	0.103	0.015	0.104	0.124	0.021	8	4	0	0.111	0.159	0.872
0.2	0.5	0.128	100	100	0.143	0.097	0.014	0.097	0.119	0.022	6	3	0	0.104	0.154	0.862
0.5	0.1	0.333	100	100	0.347	0.145	0.014	0.145	0.172	0.027	8	4	1	0.127	0.160	0.913
0.5	0.2	0.333	100	100	0.340	0.113	0.007	0.113	0.136	0.023	7	4	0	0.085	0.119	0.881
0.5	0.3	0.333	100	100	0.345	0.100	0.012	0.100	0.121	0.021	9	5	2	0.148*	0.181	0.939
0.5	0.4	0.333	100	100	0.347	0.092	0.013	0.092	0.112	0.020	7	4	2	0.120	0.196	0.929
0.5	0.5	0.333	100	100	0.346	0.091	0.013	0.091	0.107	0.016	8	6	1	0.111	0.185	0.944
0.7	0.1	0.494	100	100	0.499	0.117	0.006	0.117	0.154	0.037	5	4	1	0.117	0.130	0.817
0.7	0.2	0.494	100	100	0.498	0.092	0.005	0.092	0.120	0.028	6	3	1	0.130	0.115	0.840
0.7	0.3	0.494	100	100	0.507	0.084	0.013	0.085	0.104	0.019	7	4	2	0.162*	0.205	0.889
0.7	0.4	0.494	100	100	0.513	0.080	0.019	0.081	0.095	0.015	11	6	3	0.125	0.272	0.932
0.7	0.5	0.494	100	100	0.511	0.082	0.017	0.083	0.091	0.008	9	7	3	0.113	0.262	1.006
0.9	0.1	0.713	100	100	0.721	0.080	0.008	0.080	0.118	0.038	6	4	0	0.179**	0.179	0.761
0.9	0.2	0.713	100	100	0.723	0.060	0.010	0.060	0.089	0.030	3	2	1	0.172**	0.206	0.751
0.9	0.3	0.713	100	100	0.727	0.056	0.015	0.058	0.076	0.020	6	5	1	0.191**	0.306	0.840
0.9	0.4	0.713	100	100	0.730	0.053	0.017	0.055	0.068	0.015	7	5	4	0.214**	0.377	0.857
0.9	0.5	0.713	100	100	0.727	0.051	0.014	0.053	0.064	0.013	9	5	2	0.192**	0.346	0.894

TABLE 5.2 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 50

CASE=2

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.020	0.151	0.020	0.151	0.188	0.038	5	2	1	0.123	0.117	0.850
0.0	0.2	0.000	100	100	0.003	0.134	0.003	0.133	0.146	0.012	9	4	2	0.065	0.028	0.954
0.0	0.3	0.000	100	100	0.007	0.116	0.007	0.115	0.131	0.016	5	4	1	0.091	0.056	0.910
0.0	0.4	0.000	100	100	0.012	0.108	0.012	0.108	0.123	0.016	7	3	1	0.092	0.098	0.901
0.0	0.5	0.000	100	100	0.012	0.101	0.012	0.101	0.119	0.018	6	3	0	0.094	0.113	0.874
0.2	0.1	0.000	100	100	0.019	0.152	0.019	0.152	0.188	0.036	6	2	1	0.108	0.109	0.858
0.2	0.2	0.000	100	100	0.006	0.134	0.005	0.134	0.146	0.012	11	2	2	0.080	0.039	0.960
0.2	0.3	0.000	100	100	0.008	0.115	0.007	0.114	0.131	0.017	7	4	1	0.093	0.061	0.903
0.2	0.4	0.001	100	100	0.014	0.103	0.014	0.104	0.123	0.019	7	1	1	0.087	0.110	0.870
0.2	0.5	0.001	100	100	0.015	0.097	0.014	0.098	0.118	0.021	5	2	0	0.095	0.122	0.847
0.5	0.1	0.000	100	100	0.020	0.149	0.020	0.150	0.189	0.040	4	1	1	0.119	0.116	0.837
0.5	0.2	0.001	100	100	0.009	0.134	0.007	0.134	0.147	0.012	10	3	2	0.079	0.057	0.952
0.5	0.3	0.003	100	100	0.013	0.117	0.010	0.117	0.132	0.015	8	3	1	0.086	0.081	0.922
0.5	0.4	0.006	100	100	0.019	0.107	0.014	0.108	0.122	0.015	6	4	1	0.078	0.116	0.911
0.5	0.5	0.009	100	100	0.021	0.101	0.012	0.101	0.118	0.017	5	2	0	0.095	0.107	0.884
0.7	0.1	0.001	100	100	0.022	0.148	0.021	0.148	0.189	0.041	4	1	1	0.113	0.122	0.833
0.7	0.2	0.004	100	100	0.014	0.131	0.010	0.131	0.147	0.016	8	5	2	0.104	0.074	0.933
0.7	0.3	0.009	100	100	0.020	0.115	0.011	0.115	0.131	0.016	8	4	1	0.078	0.088	0.904
0.7	0.4	0.016	100	100	0.026	0.105	0.010	0.105	0.121	0.017	8	5	1	0.082	0.086	0.890
0.7	0.5	0.024	100	100	0.034	0.100	0.009	0.100	0.116	0.016	9	2	0	0.096	0.079	0.886
0.9	0.1	0.004	100	100	0.025	0.151	0.020	0.152	0.189	0.038	5	2	1	0.099	0.121	0.855
0.9	0.2	0.018	100	100	0.027	0.122	0.009	0.121	0.146	0.025	7	3	0	0.127	0.075	0.859
0.9	0.3	0.039	100	100	0.046	0.107	0.007	0.107	0.129	0.022	6	3	0	0.085	0.061	0.851
0.9	0.4	0.067	100	100	0.071	0.093	0.004	0.093	0.116	0.024	7	3	0	0.104	0.040	0.816
0.9	0.5	0.100	100	100	0.105	0.084	0.004	0.084	0.108	0.024	4	2	0	0.118	0.046	0.792

TABLE 5.3 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 50

CASE=3

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.020	0.151	0.020	0.151	0.188	0.038	5	2	1	0.123	0.117	0.850
0.0	0.2	0.000	100	100	0.003	0.134	0.003	0.133	0.146	0.012	9	4	2	0.065	0.028	0.954
0.0	0.3	0.000	100	100	0.007	0.116	0.007	0.115	0.131	0.016	5	4	1	0.091	0.056	0.910
0.0	0.4	0.000	100	100	0.012	0.108	0.012	0.108	0.123	0.016	7	3	1	0.092	0.098	0.901
0.0	0.5	0.000	100	100	0.012	0.101	0.012	0.101	0.119	0.018	6	3	0	0.094	0.113	0.874
0.2	0.1	0.128	100	100	0.105	0.154	-.023	0.155	0.186	0.032	4	1	1	0.108	-.086	0.855
0.2	0.2	0.127	100	100	0.111	0.121	-.016	0.121	0.146	0.026	5	2	1	0.103	-.082	0.898
0.2	0.3	0.126	100	100	0.110	0.109	-.016	0.110	0.130	0.021	7	3	1	0.121	-.103	0.920
0.2	0.4	0.124	100	100	0.110	0.102	-.015	0.103	0.121	0.019	4	3	2	0.112	-.118	0.884
0.2	0.5	0.122	100	100	0.110	0.095	-.012	0.097	0.116	0.020	5	4	1	0.140*	-.132	0.857
0.5	0.1	0.332	100	100	0.353	0.121	0.021	0.122	0.173	0.052	5	3	0	0.163**	0.169	0.751
0.5	0.2	0.330	100	100	0.350	0.090	0.020	0.091	0.134	0.044	2	0	0	0.185**	0.161	0.698
0.5	0.3	0.325	100	100	0.343	0.088	0.018	0.088	0.120	0.032	3	2	0	0.146*	0.130	0.766
0.5	0.4	0.319	100	100	0.333	0.085	0.014	0.084	0.111	0.025	3	0	0	0.101	0.048	0.781
0.5	0.5	0.311	100	100	0.319	0.083	0.008	0.083	0.105	0.022	4	1	0	0.101	-.078	0.809
0.7	0.1	0.492	100	100	0.515	0.113	0.023	0.115	0.155	0.042	5	3	1	0.185**	0.236	0.758
0.7	0.2	0.487	100	100	0.502	0.106	0.016	0.106	0.119	0.013	11	4	1	0.111	0.196	0.950
0.7	0.3	0.478	100	100	0.493	0.084	0.015	0.083	0.106	0.022	3	1	0	0.110	0.072	0.791
0.7	0.4	0.466	100	100	0.480	0.081	0.014	0.082	0.097	0.015	4	3	0	0.075	-.056	0.853
0.7	0.5	0.452	100	100	0.466	0.073	0.015	0.077	0.091	0.018	5	0	0	0.168**	-.224	0.799
0.9	0.1	0.709	100	100	0.707	0.099	-.002	0.099	0.126	0.027	2	1	0	0.127	0.116	0.738
0.9	0.2	0.697	100	100	0.706	0.074	0.009	0.074	0.092	0.018	6	3	0	0.150*	0.225	0.812
0.9	0.3	0.677	100	100	0.690	0.066	0.013	0.067	0.080	0.014	3	2	1	0.154*	0.273	0.857
0.9	0.4	0.652	100	100	0.665	0.064	0.014	0.065	0.074	0.010	12	6	0	0.154*	0.302	0.918
0.9	0.5	0.621	100	100	0.635	0.060	0.014	0.061	0.070	0.011	8	5	1	0.136*	0.306	0.911

TABLE 5.4 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 50

CASE=4

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.020	0.151	0.020	0.151	0.188	0.038	5	2	1	0.123	0.117	0.850
0.0	0.2	0.000	100	100	0.003	0.134	0.003	0.133	0.146	0.012	9	4	2	0.055	0.028	0.954
0.0	0.3	0.000	100	100	0.007	0.116	0.007	0.115	0.131	0.016	5	4	1	0.091	0.056	0.910
0.0	0.4	0.000	100	100	0.012	0.108	0.012	0.108	0.123	0.016	7	3	1	0.092	0.098	0.901
0.0	0.5	0.000	100	100	0.012	0.101	0.012	0.101	0.119	0.018	6	3	0	0.094	0.113	0.874
0.2	0.1	0.128	100	100	0.146	0.138	0.017	0.138	0.186	0.049	2	2	0	0.133	0.124	0.777
0.2	0.2	0.128	100	100	0.135	0.123	0.007	0.123	0.145	0.022	5	2	1	0.105	0.082	0.880
0.2	0.3	0.128	100	100	0.138	0.108	0.010	0.108	0.130	0.022	6	3	1	0.118	0.114	0.853
0.2	0.4	0.128	100	100	0.143	0.100	0.015	0.100	0.121	0.021	5	4	0	0.116	0.153	0.852
0.2	0.5	0.128	100	100	0.144	0.096	0.015	0.096	0.116	0.021	6	1	0	0.115	0.167	0.856
0.5	0.1	0.333	100	100	0.342	0.122	0.008	0.122	0.174	0.052	4	1	0	0.125	0.107	0.755
0.5	0.2	0.334	100	100	0.341	0.115	0.007	0.114	0.134	0.020	8	3	0	0.130	0.137	0.872
0.5	0.3	0.334	100	100	0.345	0.101	0.011	0.101	0.119	0.018	6	3	0	0.135	0.171	0.855
0.5	0.4	0.335	100	100	0.349	0.092	0.014	0.093	0.110	0.017	9	4	0	0.145*	0.212	0.853
0.5	0.5	0.336	100	100	0.348	0.088	0.012	0.089	0.104	0.016	8	1	0	0.152*	0.217	0.853
0.7	0.1	0.500	100	100	0.509	0.114	0.009	0.115	0.155	0.041	5	4	1	0.140*	0.201	0.799
0.7	0.2	0.518	100	100	0.524	0.094	0.006	0.098	0.115	0.022	5	0	0	0.175**	0.332	0.832
0.7	0.3	0.546	100	100	0.549	0.074	0.003	0.092	0.095	0.021	12	3	0	0.238**	0.661	0.840
0.7	0.4	0.579	100	100	0.581	0.061	0.001	0.106	0.081	0.020	33	20	4	0.476**	1.172	0.889
0.7	0.5	0.615	100	100	0.618	0.053	0.003	0.134	0.070	0.017	57	50	22	0.654**	1.886	0.992

Relevant to Question 1, we note that T appears to have a small positive bias. As stated previously, there should be no bias if $\theta=0$, but our study yields an estimated bias ranging from .003 to .020 (depending on ε) for that situation. Where $\theta>0$, the estimated bias averages about .01, with maximum just over .02; it is always positive except for Case 3, $\rho. = .2$ where the estimated value ranges from $-.012$ to $-.023$, and for one instance of estimated bias $-.002$ for $\rho. = .9$ and $\varepsilon = .1$, also in Case 3. For this sample size (50) the empirical standard error S_T is typically about .10, ranging from just over .05 to just over .15; it is, of course, inversely related to θ (high correlations are more accurately estimated than low ones) and also inversely related to ε (because larger values of ε imply larger numbers of matched pairs), but neither relationship is exceptionally strong within the range of scenarios under study.

We next note that S_U appears to overestimate the true standard deviation. Its estimated bias is positive for every scenario, being typically about .02, with maximum value just over .05. This bias is smaller for larger values of ε, presumably because then there are more matched pairs on which to base S_U; there is no clear pattern of relationship with case or with θ.

Despite the biases in T and S_U, the standardized values $Z_i = (T_i-\theta)/S_i$ appear to be distributed approximately as standard normal random variables for n=50, at least if θ is not too large. Thus in Case 2, where θ never exceeds .1, no value of D is significant at the .05 level; in Cases 1 and 4 some significantly large values of D occur for the larger values of θ; in Case 3 we observe significant D's for large θ, though also for some situations which we discuss later. Overall, the tails of the distribution of the Z_i are too "light"; e.g., in Case 2 we observe 6.68, 2.80, and 0.84 values of Z_i in the tails on the average when we expect 10, 5, and 1 values, respectively. The last three scenarios of Case 4 are the only exceptions, and there we observe just the opposite: extremely heavy tails.

We note that, corresponding to the positive bias in T, $\bar{Z}$ also was greater than zero for every scenario except a few in Case 3, and equaled .159 on average. Furthermore, corresponding to the overestimation of the true standard deviation by $S_U(T)/\sqrt{n}$, S_Z was .869 on average, and less than unity in every scenario except one (S_Z = 1.006 for Case 1, $\rho_{\bullet}$= .7, ε = .5). Thus we considered the proposition that T might be normally distributed, but without specifying the mean and standard deviation. Define

$$Z_i^* = \frac{T_i - \bar{T}}{S_T}$$

and

$$D^* = \sup_{i=1(1)M} \sup\{\,|\Phi(Z^*_{(i)}) - \frac{i-1}{M}|,\,|\frac{i}{M} - \Phi(Z^*_{(i)})|\,\},$$

where $Z^*_{(i)}$ is the i-th order statistic of the Z^*_i, i=1(1)M; then D^* is the Lilliefors (1967) goodness-of-fit statistic for testing the hypothesis of normality, with asymptotic critical values of $.886/\sqrt{M}$ and $1.031/\sqrt{M}$ at the .05 and .01 levels of significance, respectively. Table 5.5 contains the observed values of D^* for n=50. The only significant values of D^* appear for the 3 largest values of θ for both Cases 3 and 4. (One other test statistic is significant at the .05 level, but since we expect 1 out of 20 due to chance we feel that we can dismiss it.) This suggests that the problem in achieving standard normality of the Z_i is due more to the biases than to the shape of the actual distribution.

Note finally that we see a clear pattern of $\bar{Z}$ increasing with θ. Although a similar trend is not evident for S_Z, we notice that the larger $\bar{Z}$'s tend to occur with the relatively smaller S_Z's. In fact the significant D values in Case 3 seem to be explained by this phenomenon. The values of D which are significant in Cases 1 and 4 also appear to be severely affected by this $\bar{Z}$-S_Z relationship. Recall also that the underlying trivariate normal distribution is degenerate for $\rho. = 1/\sqrt{2} = .707$ in Case 4, while the distribution is not degenerate until $\rho.=1$ in the other cases. We may expect some instability near the points of degeneracy.

TABLE 5.5

Lilliefors Goodness-Of-Fit Statistic

For Samples Of Size 50

R	TOL	THETA	M	M_STAR	CASE_1	CASE_2	CASE_3	CASE_4
0.0	0.1	0.000	100	100	0.047	0.047	0.047	0.047
0.0	0.2	0.000	100	100	0.056	0.056	0.056	0.056
0.0	0.3	0.000	100	100	0.046	0.046	0.046	0.046
0.0	0.4	0.000	100	100	0.038	0.038	0.038	0.038
0.0	0.5	0.000	100	100	0.035	0.035	0.035	0.035
0.2	0.1	0.128	100	100	0.064	0.055	0.045	0.061
0.2	0.2	0.128	100	100	0.057	0.047	0.076	0.070
0.2	0.3	0.128	100	100	0.057	0.047	0.057	0.064
0.2	0.4	0.128	100	100	0.060	0.061	0.050	0.044
0.2	0.5	0.128	100	100	0.054	0.030	0.047	0.061
0.5	0.1	0.333	100	100	0.067	0.057	0.089*	0.065
0.5	0.2	0.334	100	100	0.058	0.041	0.054	0.075
0.5	0.3	0.334	100	100	0.063	0.051	0.054	0.078
0.5	0.4	0.335	100	100	0.063	0.048	0.043	0.070
0.5	0.5	0.336	100	100	0.045	0.063	0.061	0.063
0.7	0.1	0.500	100	100	0.044	0.060	0.079	0.066
0.7	0.2	0.518	100	100	0.055	0.061	0.068	0.070
0.7	0.3	0.546	100	100	0.062	0.055	0.071	0.138**
0.7	0.4	0.579	100	100	0.062	0.037	0.077	0.104**
0.7	0.5	0.615	100	100	0.065	0.042	0.071	0.115**
0.9	0.1	0.709	100	100	0.045	0.050	0.106**	–
0.9	0.2	0.697	100	100	0.064	0.059	0.107**	–
0.9	0.3	0.677	100	100	0.062	0.032	0.100*	–
0.9	0.4	0.652	100	100	0.074	0.054	0.064	–
0.9	0.5	0.621	100	100	0.055	0.070	0.055	–

The tendency to have Z_i which are not standard normally distributed when the values of the population parameters are near points of degeneracy suggests increasing the size of the samples and examining the behavior of the Z_i in these instances. Thus, we generated 100 samples of size 100 for Case 4 with $\rho.$ remaining at .7; we increased $\rho.$ to .99 in Cases 1 and 3 and ρ to .99 in Case 2. The tolerance varied as before, with $\varepsilon = .1(.1).5$. The results are in Table 5.6.

We now observe a smaller positive bias in T than before, averaging .009 for Case 2 with maximum of .014 and .003 in Cases 1 and 4 with maximum of .006; for Case 3, however, tolerances of .1, .2, and .5 result in negative estimated bias ranging from -.011 to -.005. Due to the increase in sample size, S_T is much smaller, typically a little more than .04, and ranging from .016 to .090. Similarly, $S_U/\sqrt{n}$ tends to overestimate the true standard deviation by about .004 with maximum .015. One exception is the Case 4, $\rho. = .7$, $\varepsilon = .2$ scenario where $b_S = -.001$. As expected, this bias decreases as ε increases.

There is a definite tendency towards standard normality for Case 4, $\rho. = .7$; only the value of D for ε=.1 is significant at the .05 level. The tails also improve, although they still tend to be lighter than expected. For Cases 1, 2, and 3 the large D's occur for small values of ε with the exception of the significant

TABLE 5.6 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 100

CASE	R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
1	0.99	0.1	0.910	100	100	0.913	0.024	0.003	0.024	0.035	0.010	4	3	1	0.168**	0.197	0.789
1	0.99	0.2	0.910	100	100	0.911	0.021	0.001	0.021	0.026	0.005	6	5	0	0.138*	0.183	0.842
1	0.99	0.3	0.910	100	100	0.912	0.019	0.002	0.019	0.022	0.003	8	4	0	0.138*	0.206	0.892
1	0.99	0.4	0.910	100	100	0.912	0.017	0.002	0.017	0.020	0.003	9	6	1	0.102	0.214	0.933
1	0.99	0.5	0.910	100	100	0.912	0.016	0.002	0.016	0.018	0.002	9	6	2	0.120	0.239	0.958
2	0.99	0.1	0.050	100	100	0.064	0.090	0.014	0.090	0.104	0.015	6	3	2	0.127	0.138	0.901
2	0.99	0.2	0.173	100	100	0.185	0.059	0.012	0.070	0.078	0.009	4	3	0	0.148*	0.161	0.900
2	0.99	0.3	0.318	100	100	0.326	0.054	0.008	0.054	0.063	0.009	7	1	0	0.129	0.142	0.870
2	0.99	0.4	0.446	100	100	0.451	0.047	0.005	0.047	0.053	0.006	10	2	0	0.102	0.128	0.898
2	0.99	0.5	0.543	100	100	0.547	0.041	0.004	0.041	0.045	0.004	9	2	0	0.119	0.146	0.919
3	0.99	0.1	0.906	100	100	0.895	0.027	-.011	0.029	0.038	0.011	3	3	1	0.177**	-.189	0.753
3	0.99	0.2	0.865	100	100	0.857	0.027	-.008	0.028	0.032	0.005	4	2	0	0.131	-.152	0.837
3	0.99	0.3	0.808	100	100	0.811	0.029	0.003	0.029	0.032	0.002	8	3	0	0.089	0.193	0.936
3	0.99	0.4	0.755	100	100	0.761	0.031	0.006	0.031	0.033	0.002	9	2	1	0.174**	0.281	0.947
3	0.99	0.5	0.714	100	100	0.709	0.032	-.005	0.032	0.034	0.003	6	5	0	0.096	-.056	0.909
4	0.70	0.1	0.500	100	100	0.506	0.082	0.006	0.081	0.086	0.005	10	4	0	0.141*	0.148	0.911
4	0.70	0.2	0.518	100	100	0.524	0.067	0.005	0.067	0.066	-.001	12	6	1	0.121	0.163	1.002
4	0.70	0.3	0.546	100	100	0.548	0.054	0.003	0.054	0.056	0.002	7	3	0	0.114	0.111	0.966
4	0.70	0.4	0.579	100	100	0.582	0.046	0.002	0.045	0.047	0.002	8	7	0	0.076	0.112	0.974
4	0.70	0.5	0.615	100	100	0.617	0.040	0.001	0.039	0.041	0.001	8	4	1	0.093	0.100	0.983

D for Case 4, $\rho.=.99$, $\epsilon=.4$. Although the size of the samples has doubled, the large values of D seem to be caused by the biases in our estimate of θ and in our estimate of the true standard deviation of T just as before: the significant D's occur when $\bar{Z}$ is large and S_Z is relatively small. For example, $\bar{Z} = .281$ and $S_Z = .947$ when Case is 3 and $\epsilon = .4$ which results in D=.174, a value that is significant at the .01 level. The tails of the Z_i are also lighter than expected for these three cases.

5.3 Results for n=25 or 10

In Tables 5.7-5.10 we display the results of the Monte Carlo study for samples of size 25. We note an increase in the estimated bias b_T when $\theta=0$, decreasing from .057 when $\epsilon=.1$ to .037 when $\epsilon=.5$. The bias is always positive for Cases 2 and 4, typically about .04 and .03, respectively. On the other hand, when $\theta>0$ in Cases 1 and 3, the bias is negative, averaging a little more than -.002 and -.015, respectively. The empirical standard deviation is still, of course, inversely related to both θ and ϵ; it also increases to slightly less than .2 on the average with maximum .306, which is as expected with the decrease in sample size.

Not surprisingly, $S_U(T)/\sqrt{n}$ overestimates the true standard deviation with b_S increasing slightly, about .002 on average, as compared to n=50. Although the pattern is not as clear as before, there is a tendency for this bias to decrease with an increasing

TABLE 5.7 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 25

CASE=1

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.057	0.304	0.057	0.308	0.338	0.034	11	8	4	0.150*	0.206	1.108
0.0	0.2	0.000	100	100	0.043	0.235	0.043	0.238	0.256	0.020	8	5	2	0.156*	0.162	1.035
0.0	0.3	0.000	100	100	0.037	0.202	0.037	0.205	0.225	0.023	10	6	3	0.160*	0.153	1.008
0.0	0.4	0.000	100	100	0.039	0.190	0.039	0.193	0.206	0.016	8	7	4	0.157*	0.161	1.054
0.0	0.5	0.000	100	100	0.037	0.175	0.037	0.178	0.195	0.020	10	5	2	0.190**	0.181	1.003
0.2	0.1	0.128	100	99	0.136	0.302	0.008	0.301	0.326	0.024	12	6	4	0.078	0.065	1.049
0.2	0.2	0.128	100	100	0.127	0.239	-.002	0.237	0.251	0.012	13	8	3	0.072	0.053	1.087
0.2	0.3	0.128	100	100	0.122	0.201	-.006	0.201	0.219	0.017	12	5	1	0.058	0.013	1.022
0.2	0.4	0.128	100	100	0.131	0.180	0.003	0.179	0.198	0.019	11	7	1	0.073	0.046	0.978
0.2	0.5	0.128	100	100	0.134	0.177	0.005	0.177	0.186	0.008	12	8	2	0.085	0.064	1.048
0.5	0.1	0.333	100	97	0.345	0.285	0.012	0.284	0.298	0.012	11	7	4	0.100	0.120	0.989
0.5	0.2	0.333	100	100	0.340	0.210	0.006	0.209	0.235	0.026	11	7	3	0.102	0.121	0.997
0.5	0.3	0.333	100	100	0.326	0.180	-.007	0.179	0.207	0.027	10	8	0	0.073	0.035	0.943
0.5	0.4	0.333	100	100	0.334	0.164	0.000	0.163	0.187	0.023	8	5	1	0.098	0.074	0.931
0.5	0.5	0.333	100	100	0.327	0.158	-.006	0.157	0.176	0.018	9	6	2	0.089	0.064	1.008
0.7	0.1	0.494	100	98	0.488	0.237	-.006	0.236	0.279	0.042	10	6	4	0.107	0.126	0.931
0.7	0.2	0.494	100	100	0.496	0.183	0.003	0.182	0.218	0.035	10	5	2	0.110	0.173	0.961
0.7	0.3	0.494	100	100	0.490	0.162	-.004	0.161	0.189	0.028	10	4	1	0.087	0.118	0.911
0.7	0.4	0.494	100	100	0.498	0.146	0.004	0.145	0.168	0.022	6	2	1	0.120	0.150	0.914
0.7	0.5	0.494	100	100	0.494	0.138	0.000	0.137	0.156	0.018	6	3	2	0.103	0.157	0.981
0.9	0.1	0.713	100	91	0.697	0.194	-.016	0.194	0.222	0.027	11	10	10	0.098	0.001	0.777
0.9	0.2	0.713	100	100	0.695	0.143	-.018	0.143	0.176	0.033	8	6	3	0.104	0.144	1.072
0.9	0.3	0.713	100	100	0.702	0.120	-.011	0.120	0.150	0.030	8	6	3	0.125	0.144	1.013
0.9	0.4	0.713	100	100	0.708	0.106	-.005	0.106	0.132	0.026	9	5	3	0.127	0.137	0.928
0.9	0.5	0.713	100	100	0.705	0.103	-.008	0.103	0.120	0.017	10	6	1	0.100	0.112	0.934

TABLE 5.8 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 25

CASE=2

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.057	0.304	0.057	0.308	0.338	0.034	11	8	4	0.150*	0.206	1.108
0.0	0.2	0.000	100	100	0.043	0.235	0.043	0.238	0.256	0.020	8	5	2	0.156*	0.162	1.035
0.0	0.3	0.000	100	100	0.037	0.202	0.037	0.205	0.225	0.023	10	6	3	0.160*	0.153	1.008
0.0	0.4	0.000	100	100	0.039	0.190	0.039	0.193	0.206	0.016	8	7	4	0.157*	0.161	1.054
0.0	0.5	0.000	100	100	0.037	0.175	0.037	0.178	0.195	0.020	10	5	2	0.190**	0.181	1.003
0.2	0.1	0.000	100	100	0.058	0.306	0.058	0.310	0.338	0.032	11	8	4	0.140*	0.210	1.114
0.2	0.2	0.000	100	100	0.040	0.233	0.039	0.235	0.255	0.022	9	6	1	0.175**	0.143	1.054
0.2	0.3	0.000	100	100	0.033	0.199	0.032	0.200	0.224	0.025	10	5	2	0.129	0.128	0.996
0.2	0.4	0.001	100	100	0.037	0.188	0.036	0.191	0.204	0.016	9	9	3	0.168**	0.144	1.075
0.2	0.5	0.001	100	100	0.039	0.171	0.038	0.174	0.193	0.023	9	7	2	0.186**	0.175	1.038
0.5	0.1	0.000	100	100	0.054	0.303	0.054	0.306	0.339	0.036	10	7	2	0.129	0.189	1.084
0.5	0.2	0.001	100	100	0.038	0.233	0.037	0.235	0.258	0.024	9	6	3	0.164**	0.154	1.042
0.5	0.3	0.003	100	100	0.039	0.204	0.036	0.206	0.224	0.021	12	6	3	0.146*	0.156	1.006
0.5	0.4	0.006	100	100	0.044	0.191	0.039	0.194	0.206	0.015	10	8	3	0.178**	0.169	1.043
0.5	0.5	0.009	100	100	0.041	0.175	0.032	0.177	0.194	0.019	9	8	3	0.153*	0.159	1.024
0.7	0.1	0.001	100	100	0.046	0.298	0.045	0.300	0.339	0.041	10	5	2	0.118	0.163	1.066
0.7	0.2	0.004	100	100	0.038	0.225	0.034	0.226	0.257	0.032	10	6	3	0.162*	0.140	1.001
0.7	0.3	0.009	100	100	0.051	0.199	0.042	0.202	0.224	0.025	11	7	1	0.149*	0.190	0.974
0.7	0.4	0.016	100	100	0.058	0.187	0.042	0.191	0.205	0.017	11	8	2	0.179**	0.190	1.017
0.7	0.5	0.024	100	100	0.061	0.173	0.037	0.176	0.192	0.019	8	6	1	0.138*	0.185	1.007
0.9	0.1	0.004	100	100	0.053	0.287	0.049	0.290	0.336	0.049	10	8	2	0.124	0.182	1.048
0.9	0.2	0.018	100	100	0.057	0.235	0.039	0.237	0.252	0.017	12	9	3	0.113	0.174	1.095
0.9	0.3	0.039	100	100	0.080	0.210	0.041	0.213	0.218	0.008	13	8	3	0.134	0.223	1.063
0.9	0.4	0.067	100	100	0.107	0.190	0.040	0.193	0.196	0.006	13	8	3	0.154*	0.236	1.095
0.9	0.5	0.100	100	100	0.132	0.172	0.032	0.174	0.180	0.008	14	6	3	0.142*	0.217	1.048

TABLE 5.9 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 25

CASE=3

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.057	0.304	0.057	0.308	0.338	0.034	11	8	4	0.150*	0.205	1.108
0.0	0.2	0.000	100	100	0.043	0.235	0.043	0.238	0.256	0.020	8	5	2	0.156*	0.162	1.035
0.0	0.3	0.000	100	100	0.037	0.202	0.037	0.205	0.225	0.023	10	6	3	0.160*	0.153	1.008
0.0	0.4	0.000	100	100	0.039	0.190	0.039	0.193	0.206	0.016	8	7	4	0.157*	0.161	1.054
0.0	0.5	0.000	100	100	0.037	0.175	0.037	0.178	0.195	0.020	10	5	2	0.190**	0.181	1.003
0.2	0.1	0.128	100	100	0.116	0.294	-.012	0.293	0.335	0.041	9	9	2	0.064	0.021	1.017
0.2	0.2	0.127	100	100	0.120	0.224	-.008	0.223	0.256	0.032	11	5	0	0.048	-.009	0.972
0.2	0.3	0.126	100	100	0.106	0.209	-.020	0.209	0.222	0.013	9	9	6	0.076	-.058	1.095
0.2	0.4	0.124	100	100	0.097	0.195	-.027	0.196	0.200	0.005	8	6	4	0.099	-.108	1.098
0.2	0.5	0.122	100	100	0.091	0.181	-.031	0.184	0.185	0.004	12	9	2	0.114	-.157	1.102
0.5	0.1	0.332	100	99	0.323	0.248	-.009	0.247	0.311	0.063	9	5	2	0.104	0.056	0.916
0.5	0.2	0.330	100	100	0.311	0.191	-.019	0.191	0.234	0.043	4	3	1	0.075	-.018	0.869
0.5	0.3	0.325	100	100	0.314	0.170	-.011	0.171	0.203	0.033	5	2	0	0.060	-.009	0.890
0.5	0.4	0.319	100	100	0.306	0.163	-.013	0.164	0.185	0.022	6	5	0	0.073	-.041	0.933
0.5	0.5	0.311	100	100	0.294	0.154	-.017	0.158	0.173	0.019	10	3	2	0.115	-.120	0.986
0.7	0.1	0.492	100	98	0.461	0.249	-.031	0.250	0.295	0.046	7	5	5	0.089	0.053	0.993
0.7	0.2	0.487	100	100	0.478	0.166	-.009	0.166	0.220	0.054	6	2	0	0.095	0.063	0.844
0.7	0.3	0.478	100	100	0.478	0.152	0.000	0.152	0.188	0.036	4	2	1	0.077	0.055	0.883
0.7	0.4	0.466	100	100	0.462	0.148	-.005	0.151	0.167	0.019	9	4	0	0.093	-.055	0.940
0.7	0.5	0.452	100	100	0.445	0.142	-.006	0.150	0.158	0.015	7	4	1	0.140*	-.164	0.947
0.9	0.1	0.709	100	92	0.695	0.182	-.014	0.181	0.221	0.039	10	9	8	0.116	-.002	0.737
0.9	0.2	0.697	100	100	0.675	0.141	-.022	0.142	0.175	0.034	6	4	3	0.113	0.079	0.924
0.9	0.3	0.677	100	100	0.661	0.118	-.017	0.119	0.149	0.031	6	3	0	0.100	0.034	0.830
0.9	0.4	0.652	100	100	0.637	0.119	-.015	0.119	0.136	0.017	6	2	0	0.069	0.029	0.862
0.9	0.5	0.621	100	100	0.610	0.120	-.011	0.120	0.128	0.008	9	6	0	0.074	0.079	0.957

TABLE 5.10 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 25

CASE=4

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	100	100	0.057	0.304	0.057	0.308	0.338	0.034	11	8	4	0.150*	0.206	1.108
0.0	0.2	0.000	100	100	0.043	0.235	0.043	0.238	0.256	0.020	8	5	2	0.156*	0.162	1.035
0.0	0.3	0.000	100	100	0.037	0.202	0.037	0.205	0.225	0.023	10	6	3	0.160*	0.153	1.008
0.0	0.4	0.000	100	100	0.039	0.190	0.039	0.193	0.206	0.016	8	7	4	0.157*	0.161	1.054
0.0	0.5	0.000	100	100	0.037	0.175	0.037	0.178	0.195	0.020	10	5	2	0.190**	0.181	1.003
0.2	0.1	0.128	100	99	0.180	0.296	0.052	0.299	0.333	0.037	11	6	4	0.143*	0.197	1.037
0.2	0.2	0.128	100	100	0.157	0.228	0.029	0.228	0.253	0.025	11	7	1	0.140*	0.139	1.038
0.2	0.3	0.128	100	100	0.152	0.198	0.024	0.199	0.221	0.023	8	5	2	0.150*	0.141	1.004
0.2	0.4	0.128	100	100	0.160	0.183	0.031	0.185	0.204	0.020	12	6	2	0.178**	0.179	1.016
0.2	0.5	0.128	100	100	0.159	0.171	0.031	0.173	0.192	0.021	10	5	2	0.195**	0.197	0.999
0.5	0.1	0.333	100	99	0.354	0.271	0.021	0.270	0.315	0.044	11	9	2	0.114	0.190	1.021
0.5	0.2	0.334	100	100	0.339	0.211	0.005	0.210	0.245	0.034	12	8	2	0.138*	0.109	0.970
0.5	0.3	0.334	100	100	0.352	0.176	0.017	0.176	0.209	0.033	9	7	1	0.159*	0.159	0.909
0.5	0.4	0.335	100	100	0.361	0.161	0.026	0.162	0.188	0.028	9	5	3	0.153*	0.222	0.926
0.5	0.5	0.336	100	100	0.352	0.151	0.016	0.151	0.177	0.026	11	5	3	0.145*	0.194	0.925
0.7	0.1	0.500	100	97	0.527	0.221	0.027	0.223	0.277	0.056	6	3	3	0.158*	0.180	0.790
0.7	0.2	0.518	100	100	0.540	0.184	0.022	0.189	0.211	0.026	11	6	4	0.190**	0.400	1.017
0.7	0.3	0.546	100	100	0.576	0.154	0.030	0.174	0.167	0.013	17	14	7	0.266**	0.676	1.065
0.7	0.4	0.579	100	100	0.604	0.134	0.024	0.173	0.142	0.007	29	17	8	0.365**	0.992	1.166
0.7	0.5	0.615	100	100	0.629	0.117	0.014	0.179	0.123	0.006	38	30	13	0.459**	1.333	1.259

tolerance; once again there is no clear pattern of relationship with case or with θ.

We note next that the Z_i are not standard normal for θ near zero. The fact that there are no significant values of D for larger values of θ in Cases 1 and 3 leads us to suspect that the Kolmogorov-Smirnov statistic does not have adequate power to detect deviations from standard normality when they do indeed exist for these situations.

Similarly, in Table 5.11, we find more significant values of D* when the size of the samples drops to 25, but they occur in no particular pattern. Here we conclude that any pattern is probably hidden by the lack of power of the Lilliefors statistic to detect deviations from normality under these circumstances.

In our studies of samples of size 10 displayed in Tables 5.12-5.15, we first note that M* is usually about 95 when ϵ=.1, that is, T_i is undefined about 5 percent of the time because there are no matched pairs. M, the number of replications for which $S_i>0$, ranges from 22 to 100 depending on ϵ; it increases with ϵ because the number of matched pairs also increases with ϵ which, in turn, increases the probability of a nonzero standard error. The effect of these relationships on b_T, if any, is not clear; we note, however, that this bias is usually larger than for n=25 and 50, presumably due to the smaller sample size. The bias is negative

TABLE 5.11

Lilliefors Goodness-Of-Fit Statistic

For Samples of Size 25

R	TOL	THETA	M	M_STAR	CASE_1	CASE_2	CASE_3	CASE_4
0.0	0.1	0.000	100	100	0.075	0.075	0.075	0.075
0.0	0.2	0.000	100	100	0.082	0.082	0.082	0.082
0.0	0.3	0.000	100	100	0.088	0.088	0.088	0.088
0.0	0.4	0.000	100	100	0.075	0.075	0.075	0.075
0.0	0.5	0.000	100	100	0.106**	0.106**	0.106**	0.106**
0.2	0.1	0.128	100	99	0.073	0.065	0.083	0.061
0.2	0.2	0.128	100	100	0.080	0.084	0.087	0.081
0.2	0.3	0.128	100	100	0.058	0.066	0.053	0.100*
0.2	0.4	0.128	100	100	0.040	0.092*	0.067	0.099*
0.2	0.5	0.128	100	100	0.066	0.097*	0.042	0.111**
0.5	0.1	0.333	100	99	0.084	0.061	0.052	0.054
0.5	0.2	0.334	100	100	0.055	0.087	0.062	0.087
0.5	0.3	0.334	100	100	0.076	0.081	0.051	0.108**
0.5	0.4	0.335	100	100	0.066	0.094*	0.053	0.072
0.5	0.5	0.336	100	100	0.060	0.078	0.061	0.065
0.7	0.1	0.500	100	97	0.057	0.082	0.076	0.084
0.7	0.2	0.518	100	100	0.062	0.103**	0.073	0.067
0.7	0.3	0.546	100	100	0.082	0.068	0.048	0.095*
0.7	0.4	0.579	100	100	0.124**	0.091*	0.062	0.081
0.7	0.5	0.615	100	100	0.105**	0.051	0.050	0.070
0.9	0.1	0.709	100	92	0.098*	0.074	0.052	–
0.9	0.2	0.697	100	100	0.064	0.048	0.059	–
0.9	0.3	0.677	100	100	0.057	0.055	0.066	–
0.9	0.4	0.652	100	100	0.062	0.087	0.069	–
0.9	0.5	0.621	100	100	0.059	0.068	0.067	–

TABLE 5.12 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 10

CASE=1

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	95	40	-.046	0.788	-.046	0.765	0.295	-.493	55	55	55	0.340**	-.035	0.544
0.0	0.2	0.000	100	76	-.049	0.590	-.049	0.589	0.417	-.173	30	27	26	0.130	-.038	0.935
0.0	0.3	0.000	100	90	-.045	0.480	-.045	0.480	0.408	-.073	19	17	13	0.086	0.008	1.094
0.0	0.4	0.000	100	99	-.025	0.422	-.025	0.421	0.400	-.022	20	13	9	0.078	0.001	1.363
0.0	0.5	0.000	100	99	-.024	0.392	-.024	0.390	0.369	-.022	19	14	7	0.085	-.015	1.324
0.2	0.1	0.128	94	53	0.104	0.694	-.024	0.670	0.405	-.289	41	41	41	0.240**	-.081	0.568
0.2	0.2	0.128	100	79	0.056	0.567	-.072	0.568	0.429	-.138	29	27	23	0.121	-.012	0.990
0.2	0.3	0.128	100	95	0.099	0.423	-.029	0.422	0.442	0.019	19	14	10	0.082	-.033	1.152
0.2	0.4	0.128	100	97	0.109	0.406	-.019	0.405	0.392	-.015	21	19	10	0.077	0.043	1.324
0.2	0.5	0.128	100	100	0.113	0.380	-.015	0.378	0.368	-.012	18	15	8	0.070	0.001	1.379
0.5	0.1	0.333	94	43	0.417	0.645	0.084	0.628	0.321	-.325	51	51	51	0.420**	-.229	0.488
0.5	0.2	0.333	100	72	0.359	0.517	0.026	0.515	0.385	-.132	31	29	28	0.241**	-.106	0.751
0.5	0.3	0.333	100	88	0.361	0.401	0.027	0.400	0.394	-.007	19	16	16	0.137*	-.002	1.000
0.5	0.4	0.333	100	92	0.355	0.380	0.021	0.378	0.353	-.026	18	15	12	0.109	0.072	1.119
0.5	0.5	0.333	100	95	0.354	0.346	0.021	0.345	0.335	-.011	17	14	8	0.104	0.149	1.210
0.7	0.1	0.494	94	39	0.508	0.624	0.015	0.602	0.285	-.339	55	55	55	0.460**	-.239	0.356
0.7	0.2	0.494	100	66	0.508	0.481	0.014	0.479	0.349	-.132	35	34	34	0.310**	-.152	0.695
0.7	0.3	0.494	100	83	0.511	0.354	0.017	0.352	0.366	0.012	20	17	17	0.170**	-.017	0.777
0.7	0.4	0.494	100	87	0.516	0.319	0.023	0.318	0.330	0.010	16	14	13	0.130	0.065	0.813
0.7	0.5	0.494	100	91	0.512	0.301	0.018	0.300	0.308	0.007	17	14	10	0.117	0.118	0.934
0.9	0.1	0.713	94	22	0.715	0.547	0.003	0.527	0.143	-.404	74	73	72	0.660**	-.621	0.464
0.9	0.2	0.713	100	37	0.769	0.395	0.056	0.397	0.169	-.225	65	65	64	0.610**	-.431	0.631
0.9	0.3	0.713	100	55	0.743	0.308	0.030	0.308	0.212	-.096	49	46	45	0.450**	-.359	0.681
0.9	0.4	0.713	100	66	0.735	0.269	0.023	0.269	0.213	-.056	36	35	34	0.340**	-.211	0.672
0.9	0.5	0.713	100	72	0.749	0.235	0.036	0.237	0.196	-.039	30	28	28	0.280**	-.061	0.700

TABLE 5.13 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 10

CASE=2

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	95	40	-.046	0.788	-.046	0.765	0.295	-.493	55	55	55	0.340**	-.035	0.544
0.0	0.2	0.000	100	76	-.049	0.590	-.049	0.589	0.417	-.173	30	27	26	0.130	-.038	0.935
0.0	0.3	0.000	100	90	-.045	0.480	-.045	0.480	0.408	-.073	19	17	13	0.086	0.008	1.094
0.0	0.4	0.000	100	99	-.025	0.422	-.025	0.421	0.400	-.022	20	13	9	0.078	0.001	1.363
0.0	0.5	0.000	100	99	-.024	0.392	-.024	0.390	0.369	-.022	19	14	7	0.085	-.015	1.324
0.2	0.1	0.000	95	40	-.046	0.788	-.046	0.765	0.295	-.493	55	55	55	0.340**	-.035	0.544
0.2	0.2	0.000	100	74	-.031	0.602	-.031	0.600	0.404	-.198	31	29	28	0.149*	-.041	0.920
0.2	0.3	0.000	100	91	-.040	0.476	-.040	0.475	0.410	-.066	20	18	12	0.085	0.059	1.117
0.2	0.4	0.001	100	99	-.023	0.415	-.024	0.413	0.400	-.014	18	12	9	0.069	0.010	1.342
0.2	0.5	0.001	100	99	-.014	0.393	-.015	0.391	0.367	-.026	17	14	9	0.081	0.028	1.359
0.5	0.1	0.000	95	41	-.064	0.780	-.065	0.759	0.302	-.478	54	54	54	0.350**	0.014	0.532
0.5	0.2	0.001	100	75	-.034	0.584	-.035	0.582	0.414	-.170	29	27	26	0.130	-.037	0.815
0.5	0.3	0.003	100	92	-.039	0.453	-.043	0.452	0.419	-.034	19	17	11	0.085	0.075	1.075
0.5	0.4	0.006	100	99	-.012	0.386	-.017	0.384	0.407	0.022	16	11	7	0.053	0.015	1.209
0.5	0.5	0.009	100	99	0.005	0.365	-.003	0.363	0.373	0.008	14	11	7	0.052	0.053	1.220
0.7	0.1	0.001	95	43	-.071	0.765	-.072	0.745	0.323	-.442	52	52	52	0.340**	-.010	0.519
0.7	0.2	0.004	100	74	-.029	0.594	-.033	0.592	0.411	-.184	29	28	27	0.139*	-.059	.811
0.7	0.3	0.009	100	93	-.014	0.435	-.023	0.434	0.428	-.008	15	13	10	0.104	0.070	1.002
0.7	0.4	0.016	100	99	-.005	0.365	-.020	0.364	0.406	0.041	14	9	4	0.099	0.013	1.084
0.7	0.5	0.024	100	100	0.015	0.347	-.010	0.345	0.378	0.031	13	11	4	0.088	0.008	1.212
0.9	0.1	0.004	95	44	-.065	0.758	-.070	0.738	0.333	-.424	51	51	51	0.330**	-.033	0.490
0.9	0.2	0.018	100	77	0.024	0.579	0.007	0.576	0.429	-.149	28	27	25	0.158*	-.051	0.910
0.9	0.3	0.039	100	92	0.028	0.466	-.011	0.463	0.423	-.043	18	12	10	0.092	-.031	1.002
0.9	0.4	0.067	100	99	0.049	0.385	-.018	0.384	0.401	0.015	18	11	6	0.063	-.006	1.176
0.9	0.5	0.100	100	100	0.098	0.323	-.002	0.321	0.383	0.060	10	4	4	0.072	0.036	1.068

TABLE 5.14 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 10

CASE=3

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	95	40	-.046	0.788	-.046	0.765	0.295	-.493	55	55	55	0.340**	-.035	0.544
0.0	0.2	0.000	100	76	-.049	0.590	-.049	0.589	0.417	-.173	30	27	26	0.130	-.038	0.935
0.0	0.3	0.000	100	90	-.045	0.480	-.045	0.480	0.408	-.073	19	17	13	0.086	0.008	1.094
0.0	0.4	0.000	100	99	-.025	0.422	-.025	0.421	0.400	-.022	20	13	9	0.078	0.001	1.363
0.0	0.5	0.000	100	99	-.024	0.392	-.024	0.390	0.369	-.022	19	14	7	0.085	-.015	1.324
0.2	0.1	0.128	97	48	0.222	0.719	0.094	0.711	0.312	-.407	49	49	49	0.330**	0.019	0.647
0.2	0.2	0.127	100	75	0.235	0.552	0.108	0.560	0.416	-.135	31	31	28	0.227**	-.016	1.023
0.2	0.3	0.126	100	93	0.165	0.454	0.039	0.454	0.430	-.024	22	18	12	0.113	0.028	1.182
0.2	0.4	0.124	100	96	0.172	0.427	0.048	0.428	0.389	-.038	23	15	9	0.117	0.088	1.270
0.2	0.5	0.122	100	99	0.146	0.380	0.024	0.379	0.368	-.012	20	13	4	0.099	0.110	1.234
0.5	0.1	0.332	92	48	0.387	0.632	0.054	0.605	0.329	-.304	46	45	45	0.370**	-.336	0.708
0.5	0.2	0.330	99	69	0.303	0.555	-.026	0.550	0.357	-.198	35	32	31	0.230**	-.255	0.755
0.5	0.3	0.325	100	85	0.324	0.434	-.001	0.432	0.364	-.070	22	19	18	0.150*	-.289	1.128
0.5	0.4	0.319	100	90	0.326	0.402	0.007	0.400	0.334	-.068	19	16	12	0.112	-.112	1.111
0.5	0.5	0.311	100	93	0.329	0.374	0.018	0.373	0.321	-.053	20	16	10	0.103	0.008	1.177
0.7	0.1	0.492	94	37	0.476	0.657	-.016	0.634	0.266	-.391	58	58	57	0.470**	-.410	0.470
0.7	0.2	0.487	100	66	0.552	0.421	0.065	0.424	0.341	-.079	37	36	35	0.340**	-.285	0.775
0.7	0.3	0.478	100	80	0.511	0.371	0.033	0.370	0.341	-.030	23	22	21	0.200**	-.139	0.815
0.7	0.4	0.466	100	85	0.485	0.358	0.019	0.357	0.310	-.048	20	17	16	0.150*	-.054	0.931
0.7	0.5	0.452	100	90	0.459	0.341	0.007	0.340	0.301	-.040	18	14	11	0.110	-.008	1.010
0.9	0.1	0.709	91	20	0.805	0.403	0.096	0.393	0.151	-.251	74	71	71	0.700**	-.750	0.481
0.9	0.2	0.697	100	46	0.757	0.306	0.060	0.310	0.242	-.064	54	54	54	0.540**	-.358	0.387
0.9	0.3	0.677	100	64	0.730	0.265	0.053	0.269	0.253	-.012	36	36	36	0.360**	-.131	0.503
0.9	0.4	0.652	100	75	0.689	0.271	0.037	0.272	0.248	-.022	27	26	25	0.250**	-.056	0.730
0.9	0.5	0.621	100	83	0.645	0.272	0.024	0.271	0.251	-.021	20	18	18	0.170**	0.014	0.858

TABLE 5.15 DESCRIPTIVE STATISTICS FOR THE SAMPLES OF SIZE 10

CASE=4

R	TOL	THETA	M	M_STAR	MEAN_T	STD_T	BIAS_N_T	RMSE_T	MEAN_SE	BIAS_N_S	P10	P05	P01	D	MEAN_Z	STD_Z
0.0	0.1	0.000	95	40	-.046	0.788	-.046	0.765	0.295	-.493	55	55	55	0.340**	-.035	0.544
0.0	0.2	0.000	100	76	-.049	0.590	-.049	0.589	0.417	-.173	30	27	26	0.130	-.038	0.935
0.0	0.3	0.000	100	90	-.045	0.480	-.045	0.480	0.408	-.073	19	17	13	0.086	0.008	1.094
0.0	0.4	0.000	100	99	-.025	0.422	-.025	0.421	0.400	-.022	20	13	9	0.078	0.001	1.363
0.0	0.5	0.000	100	99	-.024	0.392	-.024	0.390	0.369	-.022	19	14	7	0.085	-.015	1.324
0.2	0.1	0.128	95	43	0.135	0.760	0.007	0.737	0.318	-.442	52	52	52	0.310**	-.053	0.561
0.2	0.2	0.128	100	76	0.101	0.581	-.028	0.579	0.422	-.158	30	29	24	0.153*	-.018	0.915
0.2	0.3	0.128	100	93	0.078	0.470	-.050	0.471	0.419	-.052	23	14	9	0.086	0.021	1.110
0.2	0.4	0.128	100	100	0.105	0.414	-.023	0.413	0.398	-.016	18	13	7	0.086	0.026	1.357
0.2	0.5	0.128	100	100	0.128	0.387	-.000	0.385	0.370	-.017	18	14	6	0.081	0.077	1.404
0.5	0.1	0.333	95	40	0.364	0.702	0.031	0.681	0.302	-.399	56	55	55	0.420**	-.241	0.464
0.5	0.2	0.334	100	70	0.330	0.572	-.004	0.569	0.358	-.213	33	31	31	0.230**	-.049	0.841
0.5	0.3	0.334	100	92	0.326	0.402	-.008	0.400	0.413	0.011	11	11	8	0.089	0.015	0.894
0.5	0.4	0.335	100	98	0.334	0.355	-.001	0.353	0.371	0.016	11	7	6	0.109	0.074	1.093
0.5	0.5	0.336	100	99	0.337	0.336	0.001	0.335	0.344	0.008	11	8	3	0.115	0.111	1.259
0.7	0.1	0.500	95	34	0.586	0.589	0.086	0.578	0.247	-.342	62	62	61	0.550**	-.407	0.506
0.7	0.2	0.518	100	58	0.601	0.450	0.083	0.460	0.291	-.160	44	43	42	0.410**	-.145	0.747
0.7	0.3	0.546	100	70	0.635	0.353	0.089	0.378	0.280	-.073	35	32	31	0.309**	0.191	0.860
0.7	0.4	0.579	100	83	0.634	0.286	0.054	0.316	0.274	-.012	24	22	19	0.293**	0.401	0.934
0.7	0.5	0.615	100	88	0.658	0.248	0.043	0.294	0.248	0.000	22	18	15	0.353**	0.627	0.984

for $\theta<.024$ in all cases and for $\theta=.128$ in Case 1. S_T, of course, is much larger, due to the decrease in sample size, and ranges from .235 to .788. As a result of the preponderance of zero standard errors, nearly all estimated values of b_S are negative, averaging -.127. This also causes the tails of the distribution of the Z_i to be too heavy - a strong indication that the Z_i are not from a standard normal distribution. Finally, both the Kolmogorov-Smirnov and the Lilliefors goodness-of-fit statistics indicate deviation from standard normality and normality, respectively, in most instances and presumably they would in the others if they had more power.

5.4 Fisher's Z Transformation

In an attempt to improve the distributional properties of T in small samples, we considered the application of Fisher's Z transformation (1921). Fisher examined the convergence to normality of r, the Pearson product-moment correlation, and found that the transformation of r to $1/2\log((1+r)/(1-r))$ approaches normality much faster than r, with an approximate variance $1/(n-3)$ (where n is the sample size) when the population counterpart of r is reasonably small. Fieller et al. (1957) considered applying Fisher's transformation to t_a, Kendall's total correlation coefficient. They found that $1/2\log((1+t_a)/(1-t_a))$ is approximately normally distributed with mean $1/2\log((1+ \ _a)/(1- \ _a))$ and variance

TABLE 5.16

Lilliefors Goodness-Of-Fit Statistic

For Samples of Size 10

R	TOL	THETA	M	M_STAR	CASE_1	CASE_2	CASE_3	CASE_4
0.0	0.1	0.000	95	40	0.227**	0.227**	0.227**	0.227**
0.0	0.2	0.000	100	76	0.095*	0.095*	0.095*	0.095*
0.0	0.3	0.000	100	90	0.097*	0.097*	0.097*	0.097*
0.0	0.4	0.000	100	99	0.097*	0.097*	0.097*	0.097*
0.0	0.5	0.000	100	99	0.075	0.075	0.075	0.075
0.2	0.1	0.128	95	43	0.174**	0.227**	0.190**	0.192**
0.2	0.2	0.128	100	76	0.086	0.102*	0.117**	0.126**
0.2	0.3	0.128	100	93	0.093*	0.083	0.085	0.122**
0.2	0.4	0.128	100	100	0.084	0.075	0.066	0.120**
0.2	0.5	0.128	100	100	0.079	0.066	0.080	0.092*
0.5	0.1	0.333	95	40	0.237**	0.235**	0.204**	0.238**
0.5	0.2	0.334	100	70	0.132**	0.082	0.125**	0.162**
0.5	0.3	0.334	100	92	0.093*	0.105**	0.090*	0.128**
0.5	0.4	0.335	100	98	0.062	0.079	0.087	0.140**
0.5	0.5	0.336	100	99	0.052	0.073	0.059	0.116**
0.7	0.1	0.500	95	34	0.245**	0.228**	0.258**	0.309**
0.7	0.2	0.518	100	58	0.157**	0.088	0.197**	0.222**
0.7	0.3	0.546	100	70	0.088	0.087	0.128**	0.151**
0.7	0.4	0.579	100	83	0.079	0.085	0.076	0.116**
0.7	0.5	0.615	100	88	0.067	0.087	0.082	0.112**
0.9	0.1	0.709	91	20	0.359**	0.221**	0.386**	–
0.9	0.2	0.697	100	46	0.331**	0.094*	0.327**	–
0.9	0.3	0.677	100	64	0.248**	0.094*	0.206**	–
0.9	0.4	0.652	100	75	0.177**	0.067	0.125**	–
0.9	0.5	0.621	100	83	0.143**	0.075	0.101*	–

.437/(n-4). Recall that T and t_a are equivalent when all pairs are considered matched; i.e., $\varepsilon=\infty$. Therefore, we examined the distribution of $\tilde{Z}=1/2\log((1+T)/(1-T))$. We hoped that this would be approximately normally distributed with mean $\xi=1/2\log((1+\theta)/(1-\theta))$.

We considered two possibilities for the variance. First, we tried the simple $\sigma_1=.437/(n-4)$. This, however, was evidently too small; the standardized variable $(\tilde{Z}-\xi)/\sigma_1$ had typically an estimated standard deviation of about 1.3 for n=50, and the problem was considerably worse for n=25 and n=10.

We tried next to improve on the variance approximation for our transformation by defining what we call the effective sample size, denoted by n_E. The effective sample size is that which would have produced the number of matched pairs we observed if all pairs were considered matched. In other words, if the sample size were n_E, then we would have

$$M = \binom{n_E}{2}$$

pairs when all pairs are matched. Solving for n_E given the M matched pairs actually observed, we find that

$$n_E = \frac{1+\sqrt{1+8M}}{2}.$$

We then try $\sigma_2=.437/(n_E-4)$ (note that σ_2 varies from one replication to the next). But this apparently terribly overestimates the variance, causing S_Z usually to be just under .7 for n=50.

Smaller sample sizes were not investigated because of the overwhelmingly unsatisfactory outcome for samples of size 50.

We acknowledge that these guesses at the variance approximations were quite arbitrary, but we had no theoretical knowledge available to suggest any better choices. Therefore, we terminated our search for the appropriate variance approximation (and suggest further research in this area).

The mean of $\tilde{Z}$, by the way, was consistently greater than just as $\bar{T}$ was greater than θ. This suggests that $\tilde{Z}$ is positively biased as an estimator of ξ and thus, if our results for $\bar{T}$ are any indication, there may be problems in achieving standard normality even when a good approximation to the variance exists. We, therefore, concluded that Fisher's Z transformation is not worthwhile in our context.

5.5 Conclusions

After reviewing all the data from our Monte Carlo study, we conclude that the practitioner should expect T to overestimate θ and $S_U(T)/\sqrt{n}$ to overestimate the true standard deviation of T. These biases, however, are not very large for $|\theta|<.5$. Note that the distribution and, hence, characteristics are symmetric with respect to θ. And, in fact, the standardized value $\sqrt{n}(T-\theta)/S_U(T)$ appears to be standard normally distributed for θ in this range when the sample size is at least 50.

Quade (1974) presents the following thumb-rule:

> If your sample contains at least 200 concordant and 200 discordant pairs, then the asymptotic sampling theory should obtain.

Later (personal communication) he decided that this rule is too conservative and suggested only requiring at least 50 pairs of each kind, though admitting that he has no empirical evidence to support this. Table 5.17 displays the number of replications per scenario in which this "50-50 thumb-rule" suggested by Quade is satisfied for n=50; i.e., the cells of the table are counts of the replications in which there are at least 50 concordant and 50 discordant pairs for a given scenario. Also, the scenarios in which the Z_i are significantly different from standard normal random variables as indicated by the value of the Kolmogorov-Smirnov statistic D are marked, analogous to the preceding tables. It appears that the significance or nonsignificance of D is more directly attributable to the value of $\rho.$ and, hence, to θ, than to the thumb-rule, since significant values of D occur for the larger values of $\rho.$, although we do notice that the number of replications in which the thumb-rule is satisfied is usually small when D is significant. Nevertheless, there are certainly exceptions: D is significant when the count is 53 for Case 1, $\rho.=.7$, and $\varepsilon=.3$ but it is not significant for a count of only 1 when ε decreases

TABLE 5.17

The Number of Replications in Which Quade's 50-50 Thumb-Rule is Satisfied

Case	$\rho_{\cdot}$ (ρ in Case 2)	Tolerance .1	.2	.3	.4	.5
	0	0	87	100	100	100
	.2	0	76	100	100	100
1	.5	0	31	91*	99	100
	.7	0	1	53*	89	96
	.9	0**	0**	0**	4**	34**
	0	0	87	100	100	100
	.2	0	89	100	100	100
2	.5	0	88	100	100	100
	.7	0	88	100	100	100
	.9	0	94	100	100	100
	0	0	87	100	100	100
	.2	0	77	100	100	100*
3	.5	0*	27**	91*	100	100
	.7	0**	6	54	95	100**
	.9	0	0*	1*	30*	84
	0	0	87	100	100	100
	.2	0	84	99	100	100
4	.5	0	27	94	99*	100*
	.7	0*	1**	30**	77**	92**

to .2 for the same case and ρ.; the count is always 0 when the tolerance is .1, but for moderate ρ. D is not significant.

We pursued this question one step further by examining the value of D conditional on the thumb-rule being satisfied. Thus, for n=50, we selected only those replications in which there were at least 50 concordant and 50 discordant pairs and then computed D. We noticed a few scenarios where the conditional D was not significant and the unconditional D was significant. Unfortunately, several scenarios which had a small count and a non-significant unconditional D became significant.

In summary, there is no conclusive evidence based on our study which indicates that the 50-50 thumb-rule should be used, nor is there an indication of one which would be more appropriate. We do, however, point out that the design of the present study does not permit a thorough investigation of this thumb-rule.

CHAPTER VI

PARAMETRIC AND NONPARAMETRIC CORRELATION: AN EXAMPLE

Davis *et al.* (1980) present correlations between plasma lipid and lipoprotein fractions taken from the Lipid Research Clinics (LRC) Prevalence Study which was initiated in 1971 to determine the prevalence of different types of dyslipoproteinemias. In particular, they measure the association of high density lipoprotein (HDL) cholesterol with total cholesterol, total tryglycerides, low density lipoprotein (LDL) cholesterol, very low density lipoprotein cholesterol (VLDL) and the sum of LDL and VLDL. Their data consisted of a 15% random sample of the 10 North American LRCs and was restricted to whites older than 5 years of age who fasted at least 12 hours before their blood was drawn, had no missing data for HDL cholesterol, LDL cholesterol, or VLDL cholesterol, and did not have an outlier on any variate. Figure 6.1 is a reproduction of Table 2 from their report on a final sample of 3493 males, 2606 females not taking gonadal hormones, and 712 females taking hormones, stratified by age category. They performed a logarithmic transformation on the values of each measurement before computing the correlations in order to make the distribution (of the correlations) more symmetric. Next

FIGURE 6.1 CORRELATION OF HDL WITH OTHER LIPIDS / DAVIS ET AL. (1980)

CORRELATION BETWEEN LOG HDL CHOLESTEROL AND

CLASS=MALES

AGE_CAT	N	LOG_CHOL	LOG_TRIG	LOG_LDL	LOG_VLDL	LDL_VLDL
5-9	126	0.340	-.230	-.170	0.040	-.170
10-14	282	0.350	-.410	-.040	-.340	-.120
15-19	293	0.140	-.370	-.220	-.240	-.280
20-29	369	0.040	-.420	-.170	-.370	-.280
30-39	774	0.010	-.490	-.080	-.420	-.290
40-49	706	0.150	-.400	-.010	-.340	-.190
50-59	599	0.140	-.550	-.020	-.410	-.220
60-69	230	0.240	-.520	0.100	-.460	-.160
70+	114	0.220	-.440	0.030	-.370	-.090

CLASS=FEMALES NOT TAKING HORMONES

AGE_CAT	N	LOG_CHOL	LOG_TRIG	LOG_LDL	LOG_VLDL	LDL_VLDL
5-9	108	0.250	-.300	-.160	-.230	-.270
10-14	242	0.380	-.210	-.040	-.020	-.070
15-19	272	0.340	-.320	-.050	-.290	-.130
20-29	277	0.300	-.300	-.080	-.220	-.130
30-39	479	0.070	-.320	-.250	-.290	-.320
40-49	508	0.160	-.380	-.190	-.300	-.280
50-59	373	0.040	-.570	-.180	-.430	-.330
60-69	216	0.000	-.650	-.150	-.620	-.370
70+	131	0.110	-.560	-.170	-.420	-.380

CLASS=FEMALES TAKING HORMONES

AGE_CAT	N	LOG_CHOL	LOG_TRIG	LOG_LDL	LOG_VLDL	LDL_VLDL
15-19	10	-.190	-.520	-.510	-.440	-.550
20-29	233	0.110	-.170	-.280	-.200	-.330
30-39	144	0.100	-.160	-.240	-.290	-.300
40-49	132	0.100	-.190	-.300	-.290	-.340
50-59	127	0.140	-.340	-.200	-.380	-.330
60-69	56	0.480	-.430	-.060	-.340	-.260
70+	10	0.300	0.100	-.060	-.170	-.080

they weighted each transformed correlation by the inverse of its variance and averaged these across the clinics after applying Fisher's transformation on the observed correlations to test the hypothesis that the clinics could be combined.

We present parametric and nonparametric correlations between the same lipids and lipoproteins, computed for one clinic of the LRC Prevalence Study whose data were subject to the aforementioned restrictions but with minor exceptions. We did not exclude outliers because specific definitions were not available for their identification. We must point out, however, that unusually large values have negligible effect on our nonparametric measures, which are based on the direction of differences and not on their magnitude as are the parametric correlations. Also, we excluded observations with missing data on height or weight. This information is needed to compute the Quetelet index, a measure of body mass whose value is the ratio of weight in kilograms to the square of height in centimeters. The Quetelet index was included so that we could investigate its role as a possible confounding variable in the relationships of interest. Our resulting data set contains observations on 920 males, 831 females not taking gonadal hormones, and 140 females taking hormones; thus our sample size overall is much smaller than that of Davis _et al_. Its age distribution is quite different, also; since our clinic was one which studied school children and their parents, we have very few individuals aged 20-29, yet more children under 15 than Davis et al. do. Using these data, we computed the following for each stratum, based on the log-transformed variates:

(i) Pearson product-moment correlation,

(ii) Kendall correlation (all pairs considered matched),

(iii) components of U-statistics estimate of the standard error of (ii),

(iv) matched partial correlation based on a tolerance of .1 for the Quetelet index,

(v) components of U-statistics estimate of the standard error of (iv),

(vi) number of matched pairs used in (v).

The tolerance used in (iv) was set at .1 based on our examination of Table 6.1, where we display the sample size by age category distribution of our data along with the mean, standard deviation, minimum, and maximum of the Quetelet index (multiplied by 1000). We decided upon .1 since it is typically between 20 and 25% of one standard deviation of each age category. Also, this should be large enough to produce an adequate number of matched pairs while small enough to distinguish between individuals with significantly different physiques.

Tables 6.2-6.6 contain our sample size and statistics (i)-(vi) along with the corresponding sample size and average Pearson correlation from Figure 6.1 -- one table each for total cholesterol,

TABLE 6.1

The Distribution of the Quetelet Index for Our

Data by Age Category (Units are 1000 x kg/cm^2)

Age Category	N	Mean_Q	STD_Q	MIN_Q	MAX_Q
5-9	296	1.66	0.26	0.81	3.44
10-14	595	1.92	0.34	1.19	3.95
15-19	293	2.24	0.42	1.55	4.34
20-29	62	2.59	0.60	1.72	4.29
30-39	340	2.52	0.44	1.58	4.16
40-49	232	2.57	0.44	1.65	4.08
50-59	66	2.73	0.42	1.95	4.13
60-69	7	2.56	0.22	2.33	2.85

TABLE 6.2 CORRELATION BETWEEN LOG HDL CHOLESTEROL AND

LIPID OR LIPOPROTEIN=LOG TOTAL CHOLESTEROL CLASS=MALES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	126	0.34	151	0.28	0.21	0.06	0.23	0.06	3611
10-14	282	0.35	284	0.31	0.23	0.04	0.27	0.04	7613
15-19	293	0.14	156	0.15	0.09	0.05	0.17	0.06	2017
20-29	369	0.04	23	0.32	0.26	0.14	0.24	0.21	29
30-39	774	0.01	152	-0.02	-0.02	0.06	-0.04	0.07	1869
40-49	706	0.15	107	-0.13	0.04	0.06	0.09	0.07	849
50-59	599	0.14	41	0.28	0.16	0.11	0.05	0.18	97
60-69	230	0.24	6	0.02	0.07	0.24	0.00	1.00	2
70+	114	0.22	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG TOTAL CHOLESTEROL CLASS=FEMALES NOT TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	108	0.25	145	0.36	0.23	0.05	0.29	0.06	3018
10-14	242	0.38	309	0.33	0.22	0.04	0.24	0.04	9443
15-19	272	0.34	131	0.38	0.21	0.05	0.22	0.06	1418
20-29	277	0.30	19	-0.26	-0.17	0.15	-0.18	0.40	17
30-39	479	0.07	122	-0.03	-0.05	0.06	0.06	0.08	1213
40-49	508	0.16	85	0.04	0.03	0.07	0.03	0.08	510
50-59	373	0.04	19	-0.32	-0.19	0.19	-0.37	0.31	19
60-69	216	0.00	1	.	.	.	.	.	0
70+	131	0.11	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG TOTAL CHOLESTEROL CLASS=FEMALES TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
15-19	10	-0.19	2	1.00	1.00	0.00	.	.	0
20-29	233	0.11	6	-0.59	-0.33	0.41	1.00	0.00	1
30-39	144	0.10	20	0.40	0.31	0.14	0.06	0.29	32
40-49	132	0.10	66	0.03	-0.00	0.07	-0.06	0.11	367
50-59	127	0.14	40	-0.34	-0.00	0.12	0.03	0.17	101
60-69	56	0.48	6	0.69	0.47	0.31	0.33	0.54	3
70+	10	0.30	.	.	.	.	.	.	.

TABLE 6.3 CORRELATION BETWEEN LOG HDL CHOLESTEROL AND

LIPID OR LIPOPROTEIN=LOG LDL CHOLESTEROL CLASS=MALES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	126	-0.17	151	-0.03	-0.02	0.06	-0.02	0.07	3611
10-14	282	-0.04	284	-0.00	0.03	0.04	0.08	0.04	7613
15-19	293	-0.22	156	-0.05	-0.05	0.05	0.04	0.07	2017
20-29	369	-0.17	23	0.12	0.09	0.14	0.10	0.23	29
30-39	774	-0.08	152	-0.01	-0.02	0.05	-0.03	0.07	1869
40-49	706	-0.01	107	0.30	0.13	0.07	0.15	0.09	849
50-59	599	-0.02	41	0.38	0.26	0.09	0.20	0.16	97
60-69	230	0.10	6	0.47	0.33	0.41	0.00	1.00	2
70+	114	0.03	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG LDL CHOLESTEROL CLASS=FEMALES NOT TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	108	-0.16	145	0.05	0.02	0.06	0.08	0.06	3018
10-14	242	-0.04	309	-0.00	0.00	0.04	0.03	0.05	9443
15-19	272	-0.05	131	0.02	-0.02	0.06	-0.05	0.07	1418
20-29	277	-0.03	19	-0.27	-0.18	0.15	-0.06	0.39	17
30-39	479	-0.25	122	-0.27	-0.20	0.05	-0.13	0.07	1213
40-49	508	-0.19	85	-0.11	-0.08	0.07	-0.10	0.09	510
50-59	373	-0.18	19	-0.51	-0.25	0.16	-0.58	0.28	19
60-69	216	-0.15	1	.	.	.	.	.	0
70+	131	-0.17	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG LDL CHOLESTEROL CLASS=FEMALES TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
15-19	10	-0.51	2	1.00	1.00	0.00	.	.	0
20-29	233	-0.28	6	-0.70	-0.33	0.41	1.00	0.00	1
30-39	144	-0.24	20	-0.27	-0.19	0.17	-0.41	0.26	32
40-49	132	-0.30	66	-0.22	-0.16	0.07	-0.17	0.10	367
50-59	127	-0.20	40	0.08	-0.08	0.13	-0.11	0.15	101
60-69	56	-0.06	6	0.45	0.47	0.31	0.33	0.54	3
70+	10	-0.06	.	.	.	.	.	.	.

TABLE 6.4 CORRELATION BETWEEN LOG HDL CHOLESTEROL AND

LIPID OR LIPOPROTEIN=LOG TRIGLYCERIDES CLASS=MALES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	126	-0.23	151	-0.28	-0.15	0.05	-0.15	0.06	3611
10-14	282	-0.41	284	-0.48	-0.30	0.04	-0.24	0.04	7613
15-19	293	-0.37	156	-0.51	-0.36	0.04	-0.32	0.06	2017
20-29	369	-0.42	23	-0.11	-0.08	0.12	-0.17	0.26	29
30-39	774	-0.49	152	-0.55	-0.39	0.05	-0.36	0.06	1869
40-49	706	-0.40	107	-0.57	-0.35	0.05	-0.30	0.07	849
50-59	599	-0.55	41	-0.69	-0.49	0.08	-0.43	0.12	97
60-69	230	-0.52	6	-0.52	-0.33	0.41	-1.00	0.00	2
70+	114	-0.44	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG TRIGLYCERIDES CLASS=FEMALES NOT TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	108	-0.30	145	-0.39	-0.24	0.05	-0.26	0.06	3018
10-14	242	-0.21	309	-0.33	-0.19	0.04	-0.17	0.04	9443
15-19	272	-0.32	131	-0.46	-0.30	0.05	-0.29	0.07	1418
20-29	277	-0.30	19	-0.60	-0.31	0.16	-0.18	0.42	17
30-39	479	-0.32	122	-0.56	-0.40	0.05	-0.30	0.07	1213
40-49	508	-0.38	85	-0.45	-0.27	0.07	-0.19	0.09	510
50-59	373	-0.57	19	-0.63	-0.50	0.16	-0.68	0.23	19
60-69	216	-0.65	1	.	.	.	.	.	0
70+	131	-0.56	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG TRIGLYCERIDES CLASS=FEMALES TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
15-19	10	-0.52	2	-1.00	-1.00	0.00	.	.	0
20-29	233	-0.17	6	-0.74	-0.60	0.00	-1.00	0.00	1
30-39	144	-0.16	20	-0.35	-0.23	0.17	-0.09	0.27	32
40-49	132	-0.19	66	-0.22	-0.12	0.08	-0.11	0.10	367
50-59	127	-0.34	40	-0.51	-0.19	0.11	-0.21	0.17	101
60-69	56	-0.43	6	-0.70	-0.33	0.24	-1.00	0.00	3
70+	10	0.10	.	.	.	.	.	.	.

TABLE 6.5 CORRELATION BETWEEN LOG HDL CHOLESTEROL AND

LIPID OR LIPOPROTEIN=LOG VLDL CHOLESTEROL CLASS=MALES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	126	0.04	151	-0.14	-0.10	0.06	-0.08	0.07	3611
10-14	282	-0.34	284	-0.28	-0.19	0.04	-0.15	0.05	7613
15-19	293	-0.24	156	-0.43	-0.32	0.05	-0.31	0.06	2017
20-29	369	-0.37	23	-0.03	-0.09	0.12	0.03	0.24	29
30-39	774	-0.42	152	-0.49	-0.36	0.05	-0.33	0.06	1869
40-49	706	-0.34	107	-0.55	-0.35	0.06	-0.32	0.07	849
50-59	599	-0.41	41	-0.70	-0.44	0.10	-0.48	0.13	97
60-69	230	-0.46	6	-0.49	-0.33	0.41	-1.00	0.00	2
70+	114	-0.37	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG VLDL CHOLESTEROL CLASS=FEMALES NOT TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	108	-0.23	145	-0.30	-0.20	0.05	-0.19	0.06	3018
10-14	242	-0.02	309	-0.14	-0.11	0.04	-0.10	0.04	9443
15-19	272	-0.29	131	-0.38	-0.26	0.05	-0.26	0.06	1418
20-29	277	-0.22	19	-0.49	-0.17	0.22	-0.06	0.39	17
30-39	479	-0.29	122	-0.45	-0.34	0.06	-0.26	0.07	1213
40-49	508	-0.30	85	-0.41	-0.25	0.07	-0.25	0.10	510
50-59	373	-0.43	19	-0.53	-0.40	0.15	-0.58	0.23	19
60-69	216	-0.62	1	.	.	.	.	.	0
70+	131	-0.42	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG VLDL CHOLESTEROL CLASS=FEMALES TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
15-19	10	-0.44	2	-1.00	-1.00	0.00	.	.	0
20-29	233	-0.20	6	-0.77	-0.73	0.15	-1.00	0.00	1
30-39	144	-0.29	20	-0.14	-0.08	0.18	0.09	0.29	32
40-49	132	-0.29	66	-0.37	-0.24	0.07	-0.23	0.11	367
50-59	127	-0.38	40	-0.40	-0.18	0.11	-0.02	0.19	101
60-69	56	-0.34	6	-0.78	-0.40	0.23	-1.00	0.00	3
70+	10	-0.17	.	.	.	.	.	.	.

TABLE 6.6 CORRELATION BETWEEN LOG HDL CHOLESTEROL AND

LIPID OR LIPOPROTEIN=LOG LDL+VLDL CHOLESTEROL CLASS=MALES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	126	-0.17	151	-0.07	-0.04	0.06	-0.03	0.06	3611
10-14	282	-0.12	284	-0.07	-0.02	0.04	0.04	0.04	7613
15-19	293	-0.28	156	-0.19	-0.15	0.05	-0.11	0.06	2017
20-29	369	-0.28	23	0.08	0.00	0.14	0.07	0.25	29
30-39	774	-0.29	152	-0.25	-0.15	0.06	-0.15	0.07	1869
40-49	706	-0.19	107	-0.32	-0.12	0.07	-0.05	0.08	849
50-59	599	-0.22	41	0.04	0.02	0.10	-0.09	0.17	97
60-69	230	-0.15	6	-0.19	0.07	0.24	0.00	1.00	2
70+	114	-0.09	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG LDL+VLDL CHOLESTEROL CLASS=FEMALES NOT TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
5-9	108	-0.27	145	-0.06	-0.05	0.06	0.01	0.06	3018
10-14	242	-0.07	309	-0.07	-0.04	0.04	-0.01	0.05	9443
15-19	272	-0.13	131	-0.12	-0.08	0.06	-0.10	0.07	1418
20-29	277	-0.13	19	-0.43	-0.30	0.15	-0.18	0.40	17
30-39	479	-0.32	122	-0.35	-0.26	0.05	-0.18	0.07	1213
40-49	508	-0.28	85	-0.27	-0.17	0.06	-0.18	0.09	510
50-59	373	-0.33	19	-0.57	-0.29	0.17	-0.37	0.31	19
60-69	216	-0.37	1	.	.	.	.	.	0
70+	131	-0.38	.	.	.	.	.	.	.

LIPID OR LIPOPROTEIN=LOG LDL+VLDL CHOLESTEROL CLASS=FEMALES TAKING HORMONES

AGE_CAT	N	POP_CORR	OUR_N	PEARSON	KENDALL	K_STDERR	MATCHED	M_STDERR	M_PAIRS
15-19	10	-0.55	2	1.00	1.00	0.00	.	.	0
20-29	233	-0.33	6	-0.73	-0.47	0.31	1.00	0.00	1
30-39	144	-0.30	20	-0.40	-0.25	0.18	-0.38	0.28	32
40-49	132	-0.34	66	-0.30	-0.22	0.07	-0.23	0.10	367
50-59	127	-0.33	40	-0.51	-0.21	0.11	-0.21	0.16	101
60-69	56	-0.26	6	0.23	0.07	0.36	-0.33	0.54	3
70+	10	-0.08	.	.	.	.	.	.	.

total triglycerides, LDL cholesterol, VLDL cholesterol, and the sum of LDL and VLDL cholesterol. Before examining the tables, we mention that we expect the Kendall correlation to be about two-thirds the Pearson measure when the sample size is reasonably large and that both the ratios of the Kendall correlation and the matched partial correlation index to their respective standard errors are approximately distributed as standard normal random variables under the hypothesis of no association between the correlates for moderate sample sizes. Furthermore, we point out that the intention of this study is to describe the behavior of the various measures for our data, and to illustrate the use of the matched partial correlation index as an alternative to the Pearson product-moment correlation; we do not intend to discuss the epidemiology of this study or to make inferences about the various relationships. This is a purely methodological example.

First, the Pearson correlations that we observe in our data, except when our sample size is very small, are generally in the same direction as those in the representative sample, though they tend to show a stronger relationship. In order to compare the magnitude of the difference between the two correlations in a given stratum, we proceed as if the Davis et al. correlation were straightforward instead of an average, and test the hypothesis that this measure is not significantly different from the Pearson correlation that we observe in our data, using Fisher's transformation. Thus, we take

$$\frac{1}{\sqrt{2}} \left\{ \frac{\frac{1}{2}\log\left(\frac{1+r_D}{1-r_D}\right)}{\sqrt{n_D - 3}} - \frac{\frac{1}{2}\log\left(\frac{1+r_S}{1-r_S}\right)}{\sqrt{n_S - 3}} \right\}$$

as a standard normal deviate, where r_D and r_S denote the Pearson correlations from Davis et al. and our sample data, respectively, and n_D and n_S are the corresponding sample sizes for a given stratum. This, of course, is a rough test but it provides an objective criterion for deciding which differences are worth commenting on, if any. We conclude that the respective measures do not differ significantly at the .05 level for any of the correlates except for those strata where $n_D \leq 2$, in which case comparisons are not very meaningful anyway.

Next we notice that the Kendall correlation is indeed about two-thirds the Pearson measure when the absolute value of the Pearson measure is in the range .1 to .7 and the sample size is moderate (at least 100). This is not surprising, since small values of the measures are confounded with sampling error, they both approach 1.00 for strong associations, and the two-thirds rule is an asymptotic result.

Finally, a comparison of the Kendall correlation with the measure matched on the Quetelet index does not reveal any significant differences for any of the correlates. That is, those values of the

Kendall correlation which are significantly different from zero are also significant for the matched index and similarly, the strata where the Kendall measure suggests no association also have insignificant values of the matched index. This indicates that the differences in physique among the study participants from our one clinic do not affect the associations between log HDL cholesterol and the other variates in the categories presented.

As a check to see if our matching criterion was stringent enough, we restricted matches to within a Quetelet index of .05 and recomputed the matched partial correlation index for each stratum. This resulted in slightly more than half as many matched pairs per stratum, but the ratio of the indexes to their standard errors remained about the same. Therefore, we do not see any reason to present tables analogous to Tables 6.2-6.6, and we conclude that the tolerance of .1 is adequate for our purposes.

Johnson et al. (1980) present a method for producing an overall matched partial correlation index as a weighted average of conditional measures. Suppose that the data are partitioned into K groups, and assume that no observations are matched between groups. Let T_i denote the conditional matched partial correlation for group i, $i=1(1)K$, based on M_i matched pairs, with conditional standard error S_i. Also, let T denote the matched partial correlation index of the data ignoring the group distinction, with standard error S.

They show that

$$T = \frac{\sum_i M_i T_i}{\sum_i M_i}$$

and

$$S^2 = \frac{4 \sum_i \sum_j (W_{ij} - M_{ij} T)^2}{(\sum_i M_i)^2},$$

where M_{ij} is the number of matched pairs and W_{ij} is the number of these which are concordant, less the number discordant, which include the j-th observation in the i-th group; $j=1(1)n_i$ with n_i the number of observations in group i. This suggests the following alternative to the analysis approach implemented by Davis *et al.*: Run the procedure described in Johnson *et al.* on the combined clinic data, but with specifications which produce age and sex specific Kendall correlations conditional on each clinic for each pair of correlates. Then compare the correlations and their standard errors across clinics separately for each stratum. If the comparisons suggest that the correlations are reasonably similar from clinic to clinic for each stratum, then no further work is necessary since the procedure automatically calculates the weighted measure discussed above along with its standard error for each age-sex combination. Additionally, the procedure could be used to produce matched partial correlation coefficients with matching based on the Quetelet index and/or any other factor(s) that are of interest.

In summary, we do not argue for the choice of matched partial correlation over its parametric alternative - this cannot be done in general. Rather we attempt to illustrate its facility of use and generality of application which accompany a sampling theory free of restrictive assumptions.

CHAPTER VII

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 Summary

We have investigated small-sample properties of T, the sample index of matched correlation; of three methods for obtaining its standard error; and of the quantity $Z = (T - \theta)/\text{s.e.}(T)$, where θ is the population index.

Considering first the standard error, we compared the components of U-statistics, the delta method, and the jackknife. We found that the components and the delta methods produce equivalent estimates of the standard error of T. Using recent results of Efron and Stein (1981), we established that the jackknife estimate is biased upwards under reasonable conditions; its use in practice, however, was considered unfeasible due to its computational complexity. Therefore, the remainder of the thesis was limited to the components method.

Next, we prepared to look at T, s.e.(T), and Z in a Monte Carlo study using trivariate normal data. We considered θ in that situation with four different covariance structures and ranges of both the tolerance and the true product-moment partial correlation. In these settings, we identified several situations in which θ possessed a "nice" closed form; however, we had to approximate θ most of the time. We chose numerical integration over Maclaurin series expansion and Pearson's expansion as the best approximation method for our particular problem.

Then, we conducted an extensive Monte Carlo study of T, s.e.(T), and Z over a broad spectrum of small-sample situations. The study indicated that T and s.e.(T) are biased estimates of θ and the true standard deviation of T, respectively. This bias is usually positive. The distribution of Z is affected by these biases; we believe, however, that the practitioner can safely assume that this quantity attains standard normality for $|\theta|<.5$ when the sample size is at least 50. Furthermore, we presented strong evidence that T is at least normally distributed, although its mean may not equal θ.

Finally, we illustrated the general applicability and flexibility of matched partial correlation on data from one clinic of the Lipid Research Clinics Prevalence Study. We included analogous results from a parametric study of the associations between high-density lipoprotein cholesterol and other plasma lipid and lipoprotein concentrations by Davis et al. (1980) to provide the reader with a frame of reference.

7.2 Suggestions for Further Research

We have found empirically that the components method produces an estimate of the true standard deviation of T that is almost always positively biased; hence, tests involving Z are conservative. One, of course, would be interested in investigating whether or not this property can be established in either of two ways:

(a) from theoretical considerations, as for the jackknife; or

(b) empirically in a wider class of situations; e.g., for other continuous distributions or for discrete distributions.

Needless to say, method (a) is preferable and, if successful, would erase the need for any empirical work. Furthermore, since the estimate based on the components method leads to a conservative test, are there other estimates which might be better, or can adjustments be made to improve the components estimate? We also found that Fisher's transformation of T might help the approach to normality of the statistic; however, we had no theoretical basis on which to produce a suitable approximation to the standard error of the transformed T. Thus, investigations in this area would be of interest.

Our Monte Carlo study suggested a thumb-rule to the practitioner which, unfortunately, involves the unknown parameter, θ. On the other hand, Quade has suggested a 50-50 thumb-rule involving the observed number of concordant and discordant matched pairs. Although our data indicated that his rule was not always upheld, there was no strong evidence to suggest that it was invalid. (We also pointed out that our design did not permit a proper study of Quade's thumb-rule.) Thus, further work which tests Quade's thumb-rule, or any others which do not involve the unknown parameter, would be valuable.

Gans and Robertson (1981a, 1981b) obtain the exact distribution of Goodman and Kruskal's g in 2x2 and 2x3 tables by expanding it in terms of the true cell probabilities. This or some other technique could be used to obtain the exact distribution of T in discrete cases.

Lastly, other Monte Carlo work could, of course, be done using some other distribution than the trivariate normal.

APPENDIX A

FEASIBILITY STUDY

The best software available, to our knowledge, for calculating the index of matched partial correlation is a SAS procedure called MATPAR (Johnson et al. 1980). The current version of MATPAR produces T and $S_U^2(T)$ for a given sample. Recent results establish the equivalence of $S_\Delta^2(T)$ and $S_U^2(T)$, thus no additional computations are required. However, to compute T_J and $S_J^2(T_J)$, we must invoke MATPAR on the Monte Carlo data from each sample after removing the j-th observation for j = 1(1)n. This requires 80n executions of MATPAR for one replication. We decided to run one replication of the proposed design for n = 50. Unfortunately, this took over one hour of CPU time, not to mention input and output time. The jackknife estimation process accounted for 50/51 or 98% of the time.

We can gain some efficiency by modifying MATPAR so that when it checks for a matched pair of observations on one tolerance, it checks for a match on a vector of tolerances. If the vector of tolerances is in ascending order, a match on the k-th tolerance indicates a match on all subsequent tolerances and no further checking is required. Thus, from whatever amount of time MATPAR requires to run on one replication for a sample size of n, we predict that this modification will produce a decrease of about 50%. Therefore, our estimate of the CPU time required to perform the calculations necessary for a single Monte Carlo run of the above design with n = 50 is 100(1 hour)(.5) = 50 hours. If the follow-up analysis shows that a sample of size 50 yields adequate results, then a smaller sample size will be tried. Otherwise, a larger one must be investigated. So if m different sample sizes are examined, 50m hours is an upper bound on the time needed for the former case, while it is a lower bound for the latter.

We also point out that this benchmark was conducted on an IBM 4331, whose CPU performs operations 1.5 to 2.0 times faster than the IBM 370/155 at the UNC Computation Center which provides the "cheapest" computer time for student research. Furthermore, UNC users are billed for input and output time over and above CPU time. Needless to say, the Monte Carlo studies, as originally proposed, could bankrupt the university.

Note that in the discussion of modifying MATPAR to gain computational efficiency, we did not mention changing the existing algorithm to produce jackknife estimates. We do not believe that this would show any improvement in the computing time required. In fact, the addition would increase the current algorithm of order n^2 to one of order n^3 and the already considerable storage requirements would grow n-fold.

APPENDIX B

FORTRAN PROGRAMS TO COMPARE THE MACLAURIN SERIES EXPANSION AND THE NUMERICAL INTEGRATION OF P_4

Main Program:

```
      IMPLICIT REAL*8(A-H,P-Z)
      EXTERNAL PHI,QMULT3,FMACP4
      R=0.D0
      XH=-1.D0
      XK=-1.D0
      SQ2=DSQRT(2.D0)
      DO 10 I=1,5
      G=.1D0*I/SQ2
      P=2.D0*PHI(G)-1.D0
      THETA=P*P/3.D0
      THETAI=4.D0*QMULT3(G,XH,XK,R)/P-1.D0
      THETMP=4.D0*FMACP4(G,XH,XK,R)/P-1.D0
      ADI=DABS(THETA-THETAI)
      ADMP4=DABS(THETA-THETMP)
      PRINT 4,ITHETA,THETAI,ADI,THETMP,ADMP4
    4 FORMAT(2X,I1,2X,5(E20.8,2X),/)
   10 CONTINUE
      STOP
      END
```

Maclaurin Series Expansion of P_4 in γ:

```
      DOUBLE PRECISION FUNCTION FMACP4(E,A,B,R)
      IMPLICIT REAL*8(A-H,P-Z)
      PI=4.D0*DATAN(1.D0)
      FMACP4=(PI/2+DARSIN(R))*E/(PI*DSQRT(2.D0*PI)) +
     *  ((2.D0*A*B-A*A*R-B*B*R)/DSQRT(1.D0-R*R)  -
     *  PI/2-DARSIN(R))/(6.D0*PI*DSQRT(2.D0*PI))*E*E*E
      RETURN
      END
```

Numerical Integration of P_4:

```
      DOUBLE PRECISION FUNCTION QMULT3(BB,XH,XK,R)
C     THIS PROGRAM APPROXIMATES A 3-DIMENTIONAL ITERATED INTEGRAL
C  OF FN(X,Y,Z).  SUBROUTINE GL016 PROVIDES A TABLE OF THE 16-POINT
C  GAUSS-LEGENDRE FORMULA.  THE LIMITS ON THE VARIABLES ARE
C              AA .LE. X .LE. BB
C             -13 .LE  Y .LE. XH*Y
C             -13 .LE. Z .LE. (XK*X-R*Y)/(1-R*R)
C
      IMPLICIT REAL*8(A-H,P-Z)
      DIMENSION X(40,3),A(40,3),MM(3),DX(16),DA(16)
      CALL GL016(DX,DA,-1.D0,1.D0)
      DO 2 I=1,16
      XX=DX(I)
      X(I,1)=XX
      X(I,2)=XX
      X(I,3)=XX
      AB=DA(I)
      A(I,1)=AB
      A(I,2)=AB
    2 A(I,3)=AB
      MM(1)=16
      MM(2)=16
      MM(3)=16
      AA=-BB
      SQXR=DSQRT(1.D0-R*R)
      H1=(BB-AA)/2.D0
      G1=(BB+AA)/2.D0
      Q1=0.D0
      M1=MM(1)
      M2=MM(2)
      M3=MM(3)
      DO 66 I=1,M1
      UI=H1*X(I,1)+G1
      AI=H1*A(I,1)
      D1=XH*UI
      C1=1.3D1
      H2=(D1-C1)/2.D0
      G2=(D1+C1)/2.D0
      Q2=0.D0
```

```
      DO 44 J=1,M2
      VJ=H2*X(J,2)+G2
      AJ=H2*A(J,2)
      D=(XK*UI-R*VJ)/SQXR
      C=-1.3D1
      H=(D-C)/2.DO
      G=(D+C)/2.DO
      Q=0.DO
   DO 22 K=1,M3
   WK=H*X(K,3)+G
22 Q=Q+A(K,3)*DEXP(-.5DO*(UI*UI+VJ*VJ+WK*WK))
44 Q2=Q2+AJ*H*Q
66 Q1=Q1+AI*Q2

   TWOPI=8.DO*DATAN(1.DO)
   QMULT3=Q1/(TWOPI*DSQRT(TWOPI))
   RETURN
   END
```

Subroutine Used by Numerical Integration Subprogram:

```
      SUBROUTINE GL016(X,A,C,D)
      IMPLICIT REAL*8(A-H,P-Z)
      DIMENSION X(16),A(16),XX(8),AA(8)
      DATA XX(1),AA(1),XX(2),AA(2),XX(3),AA(3),XX(4),AA(4),
     * XX(t),AA(5),XX(6),AA(6),XX(7),AA(7),XX(8),AA(8) /
     * .989400934991649932596154173400,
     * .27152459411754094851780572450-1,
     * .944575023073232576077988415500,
     * .62253523938647892862843836990-1,
     * .865631202387831743880467897700,
     * .95158511682492784809925107600-1,
     * .755404408355003033895101194800,
     * .124628971255533872052476282100,
     * .617876244402643748446671764000,
     * .149595988816576732081501730500,
     * .458016777657227386342419442900,
     * .169156519395002538189312079000,
     * .281603550779258913230460501400,
     * .182603415044923588866763667900,
     * .95012509837637440185319335420-1,
     * .189450610455068496285396723200  /
      DMC=.5D0*(D-C)
      DPC=.5D0*(D+C)
      DO 2 I=1,8
      NI=17-I
      X(I)=-DMC*XX(I)+DPC
      X(NI)=DMC*XX(I)+DPC
      A(I)=DMC*AA(I)
    2 A(NI)=DMC*AA(I)
      RETURN
      END
```

Computes the Normal Distribution Function:

```
      DOUBLE PRECISION FUNCTION PHI(T)
      IMPLICIT REAL*8(A-H,P-Z)
      AT=DABS(T)
      U=1.D0/(1.D0+.2316419D0*AT)
      D=DEXP(-.50D0*T*T)/DSQRT(8.D0*DATAN(1.D0))
      PHI=1.D0=D*U*((((1.330274D0*U-1.821256D0)*U+1.781478D0)*U
     *     -.3565638D0)*U+.3193815D0)
      IF(T) 1,2,2
    1 PHI=1.D0-PHI
    2 RETURN
      END
```

REFERENCES

Agresti, Alan (1977). Considerations in measuring partial association for ordinal categorical data. *Journal of the American Statistical Association 72*: 37-45.

Arvesen, James M. (1969). Jackknifing U-statistics. *Annals of Mathematical Statistics 40*: 2076-2100.

Brown, Morton B. and Jacqueline K. Benedetti (1977). Sampling behavior of tests of correlation in two-way contingency tables. *Journal of the American Statistical Association 72*: 309-315.

Chow, Bryant, James E. Miller and Peter D. Dickinson (1974). Extensions of a Monte-Carlo comparison of some properties of two rank correlation coefficients in small samples. *Journal of Statistical Computing and Simulation 3*: 189.

Cramer, Harald (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Davis, C. E., David Gordon, John LaRosa, Peter D. S. Wood and Max Halperin (1980). Correlations of plasma high-density lipoprotein cholesterol levels with other plasma lipid and lipoprotein concentrations. The Lipid Research Clinics Program Prevalence Study. *Circulation 62 (Suppl.IV)*: 24-30.

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

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Annals of Statistics 9*: 586-596.

Fieller, E. C., H. O. Hartley, and E. S. Pearson (1957). Tests for rank correlation coefficients. *Biometrika 49* : 470-481.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron 1*: 1-32.

Forthofer, R. N. and G. G. Koch (1973). An analysis for compounded functions of categorical data. *Biometrics 29*: 143-157.

Gans, Lydia P. and C. A. Robertson (1981a). The behavior of estimated measures of association in small and moderate sample sizes for 2x3 tables. *Communications in Statistics - Theory and Methods A10*: 1673-1686.

Gans, Lydia P. and C. A. Robertson (1981b). Distributions of Goodman and Kruskal's gamma and Spearman's rho in 2x2 tables for small and moderate sample sizes. To appear in the *Journal of the American Statistical Association*.

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

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

Goodman, Leo A. and William H. Kruskal (1972). Measures of association for cross classifications IV: simplification of asymptotic variances. *Journal of the American Statistical Association 67*: 415-421.

Gradshteyn, I. S. and I. M. Ryzhik, *Tables of Integrals, Series and Products*, Alan Jeffrey, ed., Academic Press, London, 1980.

Grizzle, James E., C. Frank Starmer, and Gary G. Koch (1969). Analysis of categorical data by linear models. *Biometrics 25*: 489-504.

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

Hochberg, Yosef (1981). On the variance estimate of a Wilcoxon-Mann Whitney statistic for group ordered data. *Communications in Statistics-Theory and Methods A10*: 1719-1732.

Hoeffding, Wassily (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics 19*: 293-325.

Hoeffding, Wassily and Herbert Robbins (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal 15*: 773-780.

Johnson, Robert E., Dana Quade, and Richard D. Langston (1980). The matched partial correlation procedure: a user's guide to PROC MATPAR. *SAS Technical Report S-120*, SAS Institute Inc., Cary N.C., 14 p.

Kendall, Maurice G. (1942). Partial rank correlation. *Biometrika 32*: 277-283.

Kim, Jae-On (1975). Multivariate analysis of ordinal variables. *American Journal of Sociology 81*: 261-298.

Kleijnen, Jack P. C. (1974). *Statistical Techniques in Simulation* Marcel Dekker Inc., N.Y.

Kritzer, Herbert M. (1977). Analyzing measures of association derived from contingency tables. *Sociological Methods and Research 5*: 387-418.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association 62*: 399-402.

Ploch, Donald R. (1974). Ordinal measures of association and the general linear model. Ch. 12 in *Measurement in the Social Sciences: Theories and Strategies*, H. M. Blalock (ed.). Aldine Publishing Company, Chicago.

Quade, Dana (1967). Nonparametric partial correlation. Institute of Statistics Mimeo Series No. 526, Chapel Hill, N.C.

Quade, Dana (1974). Nonparametric partial correlation. Ch. 13 in *Measurement in the Social Sciences: Theories and Strategies*, H. M. Blalock (ed.). Aldine Publishing Company, Chicago.

Quenouille, Maurice H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society* (B) *11*: 68-84.

Quenouille, Maurice H. (1956). Notes on bias in estimation. *Biometrika 43*: 353-60.

Reynolds, H. T. (1971). Making causal inferences with ordinal data. Working paper 5, Institute for Social Science Research, Chapel Hill, N.C.

Reynolds, H. T. (1974). Ordinal partial correlation and causal inferences. Ch. 14 in *Measurement in the Social Sciences: Theories and Strategies*, H. M. Blalock (ed.). Aldine Publishing Company, Chicago.

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

Schollenberger, John and Alan Agresti (1979). Measuring association and modeling relationships between interval and ordinal variables. *Proceedings of the Social Statistics Section of the American Statistical Association*: 624-626.

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

Sen, P. K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis. *Annals of Statistics 5* : 316-329.

Singh, B. Krishna (1974). Sampling theory and ordinal measures of association: a sociological perspective and a critique. *Proceedings of the Social Statistics Section of the American Statistical Association*: 409-411.

Smith, Robert B. (1978). Nonparametric path analysis: comments on Kim's "Multivariate analysis of ordinal variables". *American Journal of Sociology 84*: 437-448.

Somers, Robert H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review 27*: 799-811.

Somers, Robert H. (1968). An approach to the multivariate analysis of ordinal data. *American Sociological Review 33*: 971-977.

Stroud, A. H. (1971). *Approximate Calculation of Multiple Integrals.* Prentice-Hall Inc., N.J.

Tukey, John W. (1958). Bias and confidence in not-quite-large samples (Abstract). *Annals of Mathematical Statistics 29*: 614.

Williams, J. Sherwood and B. Krishna Singh (1974). Multivariate analysis through a symmetric partial gamma. *Proceedings of the Social Statistics Section of the American Statistical Association*: 435-437.

Wilson, Thomas P. (1974). Measures of association for bivariate ordinal hypotheses. Ch. 11 in *Measurement in the Social Sciences: Theories and Strategies*, H. M. Blalock (ed.). Aldine Publishing Company, Chicago.

Wright, S. (1918). On the nature of size factors. *Genetics 3*: 367-374.