Statistica Sinica **9**(1999), 535-558

A NEW METHOD FOR APPROXIMATING THE ASYMPTOTIC VARIANCE OF SPEARMAN'S RANK CORRELATION

Craig B. Borkowf

National Heart, Lung and Blood Institute

Abstract: Epidemiologists use Spearman's rank correlation, $\hat{\rho}_s$, and the quantile correlation, $\hat{\rho}_q$, to measure the agreement between the bivariate ranks and the bivariate quantile-categories of bivariate continuous data, respectively. In this paper we explore the relationship between the finite and asymptotic means and variances of these statistics. We show that the asymptotic means and variances of $\hat{\rho}_q$ converge to the same limits as those of $\hat{\rho}_s$, as the number of quantile-categories increases. Also, these point estimates have distributions derived from the "empirical bivariate quantile-partitioned" (EBQP) distribution (Borkowf, Gail, Carroll and Gill (1997)), so we can use nonparametric EBQP methods to estimate the finite variances of these statistics from data and to compute the asymptotic variance of $\hat{\rho}_q$ for any underlying bivariate distribution that satisfies certain regularity conditions. These results imply that we can numerically approximate the asymptotic variance of $\hat{\rho}_s$, for which an explicit formula is not available except in special cases, to a degree of accuracy limited only by computing power.

Key words and phrases: Agreement, empirical bivariate quantile-partitioned (EBQP) distribution, epidemiology, nonparametric, quantile correlation, Spearman's rank correlation.

1. Introduction

Spearman's rank correlation, ρ_s, (Spearman (1904), (1906)) has become one of the most commonly used nonparametric statistics in epidemiological studies. We can easily compute point estimates of this statistic, $\hat{\rho}_s(t)$, where t denotes the sample size, from the bivariate ranks of any bivariate data set. Furthermore, for any given underlying bivariate distribution, we can compute the finite and asymptotic means of $\hat{\rho}_s(t)$ by adapting the methods that Moran (1948) developed for the bivariate normal (BVN) distribution with correlation ρ.

By contrast, explicit formulas for the finite and asymptotic variances of $\hat{\rho}_s(t)$ are not available except in special cases, such as independence or bivariate normality. Under independence, $\mathrm{Var}\{\hat{\rho}_s(t)\} = (t-1)^{-1}$ (Pearson (1907)). For the BVN(ρ) distribution, Kendall (1949) extended Moran's methods for means to construct a complicated polynomial approximation of the asymptotic variance of $\hat{\rho}_s(t)$ as a function of even powers of ρ up to order 8. Subsequently, David, Kendall

and Stuart (1951) and Fieller, Hartley and Pearson (1957) extended Kendall's results to construct polynomial approximations of the asymptotic variance of $\hat{\rho}_s(t)$ as a function of even powers of ρ up to orders 10 and 12, respectively. David and Mallows (1961) developed an even more complicated approximation formula for the finite variance of $\hat{\rho}_s(t)$ as a function of both sample size t and correlation ρ, accurate to 5 decimal places for moderate sample sizes and correlations $|\rho| \leq 0.8$.

These previous methods are not quite satisfactory because, even for the BVN distribution, they require dozens of polynomial expansions of complicated expressions, and these expansions need to be reworked to greater powers in order to obtain greater precision. Also, these methods are generally impractical to extend to other bivariate distributions, especially those that lack the symmetry and other well-established mathematical properties of the BVN distribution. By contrast, the methods that we propose in this paper are conceptually simpler and computationally easier to implement. These new methods can be applied to a broad range of bivariate distributions that satisfy certain regularity conditions, and their precision only depends on the available computing power. Furthermore, with due alteration of details, these methods can also be adapted for other measures of agreement calculated from bivariate ranks, such as Kendall's tau.

We can gain further knowledge about the finite and asymptotic means and variances of the rank correlation, $\hat{\rho}_s(t)$, by exploring its relationship to the quantile correlation, $\hat{\rho}_q(d,t)$, where d denotes the number of quantile-categories. To construct the quantile correlation, we choose the number of quantile-categories into which we wish to partition the original bivariate measurements. We then calculate the correlation of these bivariate quantile-categories. For example, if $t = 100$, and we choose to partition the bivariate measurements into quintile-categories, we obtain $d = 5$ quintile-categories of each set of marginal measurements with $m = 20$ observations in each category.

The relationship between $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ may seem obvious, but so far as we know, no one has exploited this relationship to study finite and asymptotic means and variances of these statistics. It also proves to be technically difficult to construct bounds on the absolute differences between their means and variances, even with appropriate regularity conditions. Furthermore, these statistics can behave quite differently for certain bivariate distributions (such as the three squares distribution, discussed below), and these cases have important theoretical implications.

Because we use the bivariate ranks and the empirical bivariate quantile-categories to define $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, respectively, these statistics have distributions derived from the "empirical bivariate quantile-partitioned" (EBQP) distribution (Borkowf, Gail, Carroll and Gill (1997)). The EBQP distribution describes the distribution of two-way contingency tables with categories defined

by the empirical quantiles of the marginal data. The use of the random empirical quantiles as category cutpoints creates tables with fixed marginal totals, unlike multinomial tables which have fixed category cutpoints and thus random marginal totals. Furthermore, the asymptotic distribution of EBQP tables, and hence of statistics calculated from these tables, depends not only on their asymptotic cell proportions, but also on the conditional distributions evaluated at the corresponding population bivariate quantiles. We can use nonparametric EBQP methods to estimate the finite variances of these statistics and to calculate the asymptotic variance of $\hat{\rho}_q(d,t)$ for any underlying bivariate distribution that satisfies certain regularity conditions.

Unfortunately, we cannot use EBQP methods to compute the asymptotic variance of $\hat{\rho}_s(t)$ directly. We can show, however, that the asymptotic variances of $\{\hat{\rho}_q(d,t)\}$ converge to the asymptotic variance of $\hat{\rho}_s(t)$ as the number of quantile-categories increases ($d \to \infty$). Thus, we can numerically approximate the asymptotic variance of $\hat{\rho}_s(t)$ to a degree of accuracy limited only by computing power.

In this paper we define $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ in mathematical notation and show that these statistics can be computed from EBQP tables. We also define special notation for the finite and asymptotic means and variances of these statistics (Section 2). Next, we study the general relationship between the finite and asymptotic means and variances of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, and consider some specific results for certain underlying bivariate distributions, including the BVN distribution (Sections 3 and 4). We present an example from nutritional epidemiology to demonstrate how we can apply these theoretical results to analyze real data (Section 5). Finally, we discuss these results and suggest some further areas of improvement (Section 6).

2. Preliminary Notation and Definitions

2.1. Notation for calculating $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$

Suppose we collect continuous bivariate data (X_k, Y_k), $k = 1, \ldots, t$, independently and identically distributed from some distribution F. Let $F(x,y)$ have marginal distributions $G(x)$ and $H(y)$, and conditional distributions $G(x|y)$ and $H(y|x)$ defined everywhere in an open domain. Furthermore, let $g(x) = G'(x)$ and $h(y) = H'(y)$ exist and be positive everywhere in the open domain. These regularity conditions guarantee that standard EBQP theory will hold for all EBQP tables with finite dimensions.

To define Spearman's rank correlation and the quantile correlation we bring in the following notation. Using the indicator function $I\{\cdot\}$, define the empirical rank functions for the marginal data as $R_x(x) = \sum_{k=1}^{t} I\{X_k \le x\}$ and $R_y(y) = \sum_{k=1}^{t} I\{Y_k \le y\}$. Because $g(x)$ and $h(y)$ exist everywhere, ties occur in the

marginal data with probability zero, and thus we can rank every observation uniquely. Note that the lowest (highest) rank corresponds to 1 (t).

Next, we define the mean (M_R) and variance (V_R) of the marginal ranks, and the covariance (C_R) and squared difference (D_R) of the bivariate ranks as

$$M_R = t^{-1} \sum_{k=1}^{t} R_x(X_k) = t^{-1} \sum_{k=1}^{t} R_y(Y_k), \tag{2.1}$$

$$V_R = t^{-1} \sum_{k=1}^{t} \{R_x(X_k) - M_R\}^2 = t^{-1} \sum_{k=1}^{t} \{R_y(Y_k) - M_R\}^2, \tag{2.2}$$

$$C_R = t^{-1} \sum_{k=1}^{t} \{R_x(X_k) - M_R\}\{R_y(Y_k) - M_R\}, \tag{2.3}$$

$$D_R = t^{-1} \sum_{k=1}^{t} \{R_x(X_k) - R_y(Y_k)\}^2. \tag{2.4}$$

Note that $M_R = \frac{1}{2}(t+1)$ and $V_R = \frac{1}{12}(t^2-1)$ given that no ties occur, while C_R and D_R are random. Furthermore, these terms satisfy the relationship $D_R = 2V_R - 2C_R$. Using this relationship, we can write the point estimate of ρ_s as

$$\hat{\rho}_s(t) = C_R V_R^{-1} = (V_R - \frac{1}{2} D_R) V_R^{-1} = 1 - \frac{1}{2} D_R V_R^{-1}.$$

In turn, using $V_R = \frac{1}{12}(t^2-1)$, we can rewrite this formula as

$$\hat{\rho}_s(t) = 1 - 6 D_R (t^2-1)^{-1}. \tag{2.5}$$

Now, suppose that instead of t distinct marginal ranks, we wish to partition the marginal data into d quantile-categories. Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x, and let $m = \lceil t/d \rceil$. Then, each marginal quantile-category will have either m or $(m-1)$ observations. We define the empirical quantile-category functions as $Q_x(X_k) = \lceil R_x(X_k) d/t \rceil$ and $Q_y(Y_k) = \lceil R_y(Y_k) d/t \rceil$.

Next, we define M_Q, V_Q, C_Q and D_Q in a fashion analogous to equations (2.1) through (2.4). Thus, we can write the point estimate of $\rho_q(d)$ with d quantile-categories as

$$\hat{\rho}_q(d,t) = C_Q V_Q^{-1} = (V_Q - \frac{1}{2} D_Q) V_Q^{-1} = 1 - \frac{1}{2} D_Q V_Q^{-1}.$$

With balanced data (i.e., $m = t/d$ is an integer), $M_Q = \frac{1}{2}(d+1)$ and $V_Q = \frac{1}{12}(d^2-1)$. In this case, we can rewrite this formula as

$$\hat{\rho}_q(d,t) = 1 - 6 D_Q (d^2-1)^{-1}. \tag{2.6}$$

Note that for balanced data, the sample quantile correlation $\hat{\rho}_q(d,t)$ is algebraically equivalent to the weighted kappa statistic $\hat{\kappa}_w(d,t)$ for $d \times d$ tables (Spitzer, Cohen, Fleiss and Endicott (1967)) with quadratic weights $w_{ij} = 1-(i-j)^2/(d-1)^2$. These weights are chosen to make $\hat{\kappa}_w(d,t)$ equivalent to the interaclass correlation with the quantile-categories as outcome scores (Fleiss and Cohen (1973)).

We set some additional notation that will be useful for the proofs in the following sections. Define the remainder functions as $P_x(X_k) = R_x(X_k) - m\{Q_x(X_k)-1\}$ and $P_y(Y_k) = R_y(Y_k)-m\{Q_y(Y_k)-1\}$. We also define M_p, V_p, C_p and D_p in a fashion analogous to equations (2.1) through (2.4) above, and the mixed difference term (D_{QP}) between the quantile-categories and the remainders as

$$D_{QP} = t^{-1}\sum_{k=1}^{t}\{Q_x(X_k) - Q_y(Y_k)\}\{P_x(X_k) - P_y(Y_k)\}. \tag{2.7}$$

Finally, we note that the four difference terms D_R, D_Q, D_p and D_{QP} satisfy the relationship

$$D_R = m^2 D_Q + D_p + 2mD_{QP}. \tag{2.8}$$

2.2. The relationship between $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ and the EBQP distribution

We now demonstrate that the difference terms D_R and D_Q can be calculated from appropriately constructed EBQP tables. Using the indicator function $I\{\cdot\}$, define the empirical cell proportions $\{p_{ij}^R\}$ $(i,j = 1,\ldots,t)$ for the bivariate ranks by

$$P_{ij}^R = t^{-1}\sum_{k=1}^{t} I\{R_x(X_k) = i\}I\{R_y(Y_k) = j\}.$$

These proportions represent the frequency with which the pairs of ranks occur in a particular data set. Similarly, define the empirical cell proportions $\{p_{ij}^Q\}$ $(i,j = 1,\ldots,d)$ for the bivariate quantile-categories by

$$p_{ij}^Q = t^{-1}\sum_{k=1}^{t} I\{Q_x(X_k) = i\}I\{Q_y(Y_k) = j\}.$$

These proportions represent the frequency with which the pairs of quantile-categories occur in a particular data set. We can show algebraically that

$$D_R = \sum_{i=1}^{t}\sum_{j=1}^{t}(i-j)^2 p_{ij}^R,$$

$$D_Q = \sum_{i=1}^{d}\sum_{j=1}^{d}(i-j)^2 p_{ij}^Q.$$

Because the categories of $\{p_{ij}^R\}$ and $\{p_{ij}^Q\}$ are defined by the bivariate ranks and the bivariate quantile-categories, respectively, these empirical proportions have the EBQP distribution (Borkowf, Gail, Carroll and Gill (1997)). In turn, quantities calculated from these proportions, such as $D_R, D_Q, \hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, also have distributions readily derived from the EBQP distribution. With more effort, we can compute these quantities from appropriately constructed EBQP tables when ties occur in the bivariate ranks, a complication that we consider elsewhere (Borkowf, submitted).

2.3. Several underlying bivariate distributions for the original data

In order to study the relationships between $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, we consider several underlying bivariate distributions. Let (X,Y) have the standard BVN distribution with means 0, variances 1, and correlation ρ. Then (e^X, e^Y) has the standard bivariate log-normal (BLN) distribution with shape parameter ρ, $(|X|,|Y|)$ has the standard bivariate half-normal (BHN) distribution with shape parameter, ρ, and (X^2, Y^2) has the standard bivariate chi-squared (BCS) distribution with correlation $\omega = \rho^2$ (Johnson and Kotz (1972)).

We also consider the three squares (TS) distribution, which represents the case in which the data occur in three clusters, as sometimes happens in epidemiological studies (Borkowf, Gail, Carroll and Gill (1997)). The TS distribution has density 3 on each of three squares placed in a 3×3 grid within the unit square (like the light squares in the logo of the International Biometric Society), and 0 elsewhere. The lower left square is $[0, \frac{1}{3}] \times [0, \frac{1}{3}]$, the middle right square is $(\frac{2}{3}, 1] \times (\frac{1}{3}, \frac{2}{3}]$, and the upper center square is $(\frac{1}{3}, \frac{2}{3}] \times (\frac{2}{3}, 1]$. The variates X and Y are each distributed uniformly on [0,1], but X and Y are dependent, with correlation 4/9. Note that the TS distribution does not satisfy the necessary regularity conditions for the standard EBQP theory at the marginal tertiles. This theory still appears to hold, however, if we define the improper conditional distributions $G(x|y) = \frac{1}{2}G(x|y-\varepsilon) + \frac{1}{2}G(x|y+\varepsilon)$ and $H(y|x) = \frac{1}{2}H(y|x-\varepsilon) + \frac{1}{2}H(y|x+\varepsilon)$, with $0 < \varepsilon < d^{-1}$, at the marginal tertiles. We use the TS distribution to demonstrate how pathologically the finite and asymptotic variances $\{\sigma_q^2(d,t)\}$ and $\{\sigma_q^2(d)\}$ can behave.

In addition to the above underlying bivariate distributions, we mention several other distributions that prove to be useful for establishing boundary conditions. First, under independence (denoted H_0), the cell proportions $\{p_{ij}^R\}$ and $\{p_{ij}^Q\}$ have the multivariate hypergeometric (MH) distribution (Plackett (1981)). In turn, quantities calculated from these proportions, including $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, have distributions readily derived from the MH distribution. Second, by definition, under perfect rank agreement (PRA) the set of all possible data points falls on a monotonic increasing line, while under perfect rank disagreement (PRD)

the set of all possible data points falls on a monotonic decreasing line. Thus, under PRA (PRD), $\hat{\rho}_s(t) = \hat{\rho}_q(d,t) = 1(-1)$.

We now consider two important properties of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ that allow us to broaden the results for these statistics beyond a particular distribution. First, these statistics are invariant to monotonic increasing transformations of the marginal data. Thus, results that apply to the standard BVN (BCS) distribution also apply to all BVN and BLN (all BCS and BHN) distributions with the same correlation/shape parameter, regardless of their location and scale parameters. Second, by symmetry, if we keep X and define a new variable $Z = -Y$, then $\hat{\rho}_s^{xz}(t) = -\hat{\rho}_s^{xy}(t)$ and $\hat{\rho}_q^{xz}(d,t) = -\hat{\rho}_q^{xy}(d,t)$, and hence the means of the new statistics will be the negatives of the original means, but the variances will be unchanged. Thus, without loss of generality, we can study distributions with positive correlations, and obtain corresponding results for negative correlations accordingly.

2.4. Notation for the finite and asymptotic means and variances of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$

We now define some notation for the finite and asymptotic means and variances of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$. First, let $\rho_s(t) = E\{\hat{\rho}_s(t)\}$, $\rho_s = \lim_{t\to\infty} \rho_s(t)$, $\rho_q(d,t) = E\{\hat{\rho}_q(d,t)\}$, $\rho_q(d) = \lim_{t\to\infty} \rho_q(d,t)$ with fixed d, and $\rho_q = \lim_{d\to\infty} \rho_q(d)$.

Moran (1948) proved for the BVN(ρ) distribution that $\rho_s(t) = \frac{6}{\pi}(t+1)^{-1}\{(t-2)\arcsin(\frac{1}{2}\rho) + \arcsin(\rho)\}$, and hence, as Pearson (1907) showed, $\rho_s = \frac{6}{\pi}\arcsin(\frac{1}{2}\rho)$.

Next, by adapting the methods that Moran developed for the BVN distribution, we can show for any underlying bivariate distributions $F(x,y)$ with marginal distributions $G(x)$ and $H(y)$,

$$\rho_s(t) = 12(t+1)^{-1}[(t-2)E\{G(X)H(Y)\} + E\{F(X,Y)\} - \frac{1}{4}(t-1)],$$

and hence

$$\rho_s = 12E\{G(X)H(Y)\} - 3 = \mathrm{Corr}\,\{G(X), H(Y)\}.$$

Tables 1(a) and 2(a) show the finite means $\{\rho_s(2)\}$ and $\{\rho_s(1000)\}$ and the asymptotic means $\{\rho_s\}$ for the BVN, BCS, and TS distributions with selected correlations. Note that $\rho_s(2) < \rho_s(t) < \rho_s(t+1) < \rho_s$ for these distributions with positive correlations for all t. Note too that under independence (PRA) (PRD), $\rho_s(t) = \rho_s = \rho_q(d,t) = \rho_q(d) = \rho_q = 0(1)(-1)$.

By contrast an explicit formula for $\rho_q(d,t)$, $d < t$, is generally not available. However, we may use EBQP methods to calculate $\rho_q(d)$. In Section 3 we compute bounds for the absolute difference between $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$. In turn, we use this result to show that $\rho_q(d) \to \rho_s$ as $d \to \infty$, and hence that $\rho_q = \rho_s$. That is, the asymptotic limit of the quantile correlation, which we cannot calculate directly by EBQP methods, is Spearman's rank correlation.

Table 1. Parameters related to the finite and asymptotic means of Spearman's rank correlation, $\hat{\rho}_s(t)$, and the quantile correlation, $\hat{\rho}_q(d,t)$, for underlying BVN distributions.

	Distribution (correlation)				
	BVN 0.0	BVN 0.25	BVN 0.5	BVN 0.75	BVN 0.9
(a) Finite and asymptotic means of Spearman's rank correlation, $\hat{\rho}_s(t)$:					
$\rho_s(2)$	0.0	0.1609	0.3333	0.5399	0.7129
$\rho_s(1000)$	0.0	0.2391	0.4821	0.7336	0.8909
ρ_s	0.0	0.2394	0.4826	0.7341	0.8915
(b) Asymptotic means of the quantile correlation, $\hat{\rho}_q(d,t)$:					
d					
2	0.0	0.1609	0.3333	0.5399	0.7129
3	0.0	0.1997	0.4084	0.6390	0.7984
4	0.0	0.2152	0.4378	0.6774	0.8356
5	0.0	0.2230	0.4525	0.6963	0.8543
6	0.0	0.2275	0.4609	0.7071	0.8649
7	0.0	0.2304	0.4662	0.7139	0.8716
8	0.0	0.2323	0.4697	0.7183	0.8760
9	0.0	0.2336	0.4722	0.7215	0.8791
10	0.0	0.2346	0.4740	0.7238	0.8814
15	0.0	0.2371	0.4786	0.7293	0.8868
20	0.0	0.2380	0.4802	0.7314	0.8888
25	0.0	0.2385	0.4810	0.7324	0.8898
30	0.0	0.2387	0.4815	0.7329	0.8903
35	0.0	0.2389	0.4818	0.7332	0.8906
40	0.0	0.2390	0.4820	0.7334	0.8908

Next, define the variances $\sigma_s^2(t) = \mathrm{Var}\{t^{\frac{1}{2}}\hat{\rho}_s(t)\}$, $\sigma_s^2 = \lim_{t\to\infty}\sigma_s^2(t)$, $\sigma_q^2(d,t) = \mathrm{Var}\{t^{\frac{1}{2}}\hat{\rho}_q(d,t)\}$, $\sigma_q^2(d) = \lim_{t\to\infty}\sigma_q^2(d,t)$ with fixed d, and $\sigma_q^2 = \lim_{d\to\infty}\sigma_q^2(d)$.

Explicit formulas for the finite variance $\sigma_s^2(t)$ and the asymptotic variance σ_s^2 are not available except in special cases, such as independence and bivariate normality. We can theoretically use EBQP methods to estimate $\sigma_s^2(t) = \sigma_q^2(t,t)$ directly, but current computing power limits us to small sample sizes (about $t \leq 20$) (Borkowf, submitted).

Table 2. Parameters related to the finite and asymptotic means of Spearman's rank correlation, $\hat{\rho}_s(t)$, and the quantile correlation, $\hat{\rho}_q(d,t)$, for underlying BCS and TS distributions.

	Distribution (correlation)				
	BCS 0.25	BCS 0.5	BCS 0.75	BCS 0.9	TS 0.44
(a) Finite and asymptotic means of Spearman's rank correlation, $\hat{\rho}_s(t)$:					
$\rho_s(2)$	0.1111	0.2500	0.4444	0.6323	0.2222
$\rho_s(1000)$	0.1660	0.3680	0.6260	0.8263	0.4438
ρ_s	0.1661	0.3683	0.6265	0.8269	0.4444
(b) Asymptotic means of the quantile correlation, $\hat{\rho}_q(d,t)$:					
d					
2	0.1064	0.2539	0.4742	0.6743	0.3333
3	0.1348	0.3116	0.5556	0.7530	0.5000
4	0.1466	0.3342	0.5853	0.7838	0.4167
5	0.1527	0.3455	0.5995	0.7987	0.4267
6	0.1563	0.3520	0.6074	0.8070	0.4571
7	0.1586	0.3560	0.6123	0.8121	0.4354
8	0.1602	0.3587	0.6155	0.8155	0.4375
9	0.1613	0.3606	0.6177	0.8178	0.4500
10	0.1621	0.3620	0.6194	0.8195	0.4400
15	0.1642	0.3654	0.6233	0.8236	0.4464
20	0.1650	0.3666	0.6247	0.8250	0.4433
25	0.1654	0.3672	0.6253	0.8257	0.4437
30	0.1656	0.3676	0.6257	0.8260	0.4449
35	0.1657	0.3678	0.6259	0.8263	0.4441
40	0.1658	0.3679	0.6260	0.8264	0.4442

Table 3(a) shows the approximated finite variances $\{\sigma_s^2(1000)\}$ (David and Mallows (1961)) and asymptotic variances $\{\sigma_s^2\}$ (Fieller, Hartley and Pearson (1957)) for the BVN distribution with selected correlations. Tables 3(a) and 4(a) show the simulated finite variance $\{\bar{\sigma}_s^2(1000)\}$ for the BVN, BCS and TS distributions with selected correlations. For the BVN model, these simulations support the accuracy of the approximated variances for $|\rho| \leq 0.8$, but suggest that the approximated variances for $\rho = 0.9$ are much less accurate. Note that under independence, $\sigma_s^2(t) = \sigma_q^2(d,t) = t(t-1)^{-1}$ and $\sigma_s^2 = \sigma_q^2(d) = \sigma_q^2 = 1$, while under both PRA and PRD, $\sigma_s^2(t) = \sigma_s^2 = \sigma_q^2(d,t) = \sigma_q^2(d) = \sigma_q^2 = 0$.

Table 3. Finite and asymptotic variances of Spearman's rank correlation, $\hat{\rho}_s(t)$, and the quantile correlation, $\hat{\rho}_q(d,t)$, for underlying BVN distributions.

	Distribution (correlation)				
	BVN 0.0	BVN 0.25	BVN 0.5	BVN 0.75	BVN 0.9
(a) Finite and asymptotic variances of Spearman's rank correlation, $\hat{\rho}_s(t)$:*					
$\sigma_s^2(1000)$	0.0010	0.9046	0.6322	0.2537	0.0542
σ_s^2	1.0	0.9035	0.6309	0.2526	0.0559
$\bar{\sigma}_s^2(1000)$	1.0046	0.9085	0.6316	0.2539	0.0559
SD	0.0042	0.0055	0.0028	0.0010	0.0002
(b) Asymptotic means of the quantile correlation, $\hat{\rho}_q(d,t)$:					
d					
2	1.0	0.9741	0.8889	0.7085	0.4918
3	1.0	0.9418	0.7565	0.4141	0.1545
4	1.0	0.9278	0.7065	0.3392	0.0994
5	1.0	0.9204	0.6818	0.3073	0.0809
6	1.0	0.9160	0.6678	0.2904	0.0721
7	1.0	0.9132	0.6590	0.2804	0.0672
8	1.0	0.9113	0.6530	0.2739	0.0642
9	1.0	0.9099	0.6488	0.2695	0.0623
10	1.0	0.9088	0.6457	0.2663	0.0609
15	1.0	0.9062	0.6380	0.2587	0.0577
20	1.0	0.9051	0.6351	0.2561	0.0566
25	1.0	0.9046	0.6337	0.2548	0.0562
30	1.0	0.9043	0.6329	0.2541	0.0559
35	1.0	0.9041	0.6324	0.2537	0.0557
40	1.0	0.9040	0.6320	0.2534	0.0556

*The finite variance $\sigma_s^2(1000)$ (David and Mallows (1961)) and the asymptotic variances σ_s^2 (Fieller, Hartley and Pearson (1957)) are accurate to 5 decimal places for $|\rho| \leq 0.8$, but they appear to be accurate to only 2 and 3 decimal places, respectively, for $\rho = 0.9$. Also, the simulated finite variances $\overline{\sigma}_s^2(1000)$ and their standard deviations (SDs) are the means and standard deviations of 10 repetitions of the variances of 100,000 simulated values $\hat{\rho}_s(1000)$ computed from BVN data sets of size 1000.

Similarly, an explicit formula for the finite variance $\sigma_q^2(d,t)$ is generally not available. We can theoretically use EBQP methods to estimate $\sigma_q^2(d,t)$, but current computing power limits us to moderate numbers of quantile-categories

(about $d \leq 20$) (Borkowf, Gail, Carroll and Gill (1997)). Furthermore, for underlying bivariate distributions that satisfy certain regularity conditions, we can use EBQP theory to calculate $\sigma_q^2(d)$ and hence to approximate σ_q^2 numerically.

Table 4. Finite and asymptotic variances of Spearman's rank correlation, $\hat{\rho}_s(t)$, and the quantile correlation, $\hat{\rho}_q(d,t)$, for underlying BCS and TS distributions.

	Distribution (correlation)				
	BCS 0.25	BCS 0.5	BCS 0.75	BCS 0.9	TS 0.44
(a) Simulated variances of Spearman's rank correlation, $\hat{\rho}_s(t)$:*					
$\overline{\sigma}_s^2(1000)$	0.9850	0.8179	0.4363	0.1286	0.9006
SD	0.0046	0.0050	0.0015	0.0005	0.0044
(b) Asymptotic variances of the quantile correlation, $\hat{\rho}_q(d,t)$:**					
d					
2	0.9967	0.9696	0.8262	0.5707	3.5556
3	0.9924	0.9045	0.5825	0.2195	1.1250
4	0.9912	0.8737	0.5123	0.1690	0.3747
5	0.9904	0.8566	0.4826	0.1516	1.6649
6	0.9897	0.8463	0.4673	0.1435	0.9478
7	0.9872	0.8395	0.4584	0.1390	0.5513
8	0.9887	0.8347	0.4527	0.1362	1.3359
9	0.9884	0.8314	0.4489	0.1344	0.9213
10	0.9881	0.8288	0.4461	0.1331	0.6416
15	0.9871	0.8224	0.4398	0.1302	0.9083
20	0.9866	0.8199	0.4377	0.1292	1.0593
25	0.9863	0.8187	0.4367	0.1287	0.7888
30	0.9862	0.8181	0.4361	0.1285	0.9030
35	0.9861	0.8176	0.4358	0.1283	0.9891
40	0.9860	0.8174	0.4356	0.1282	0.8295

*The simulated finite variances $\overline{\sigma}_s^2(1000)$ and their standard deviations (SDs) are the means and standard deviations of 10 repetitions of the variances of 100,000 simulated values $\hat{\rho}_s(1000)$ computed from BVN data sets of size 1000.

**The asymptotic variances $\sigma_q^2(d)$ for the TS distribution for $d = 10, \ldots, 40$ are: (10) 0.6416, 1.2040, 0.9124, 0.6952, 1.1333, (15) 0.9083, 0.7306, 1.0893, 0.9061, 0.7557, (20) 1.0593, 0.9048, 0.7743, 1.0375, 0.9040, (25) 0.7888, 1.0210, 0.9034, 0.8002, 1.0081, (30) 0.9030, 0.8096, 0.9976, 0.9027, 0.8174, (35) 0.9891, 0.9025, 0.8239, 0.9819, 0.9023, (40) 0.8295.

In Section 4 we compute bounds for the absolute difference between $\sigma_s^2(t)$ and $\sigma_q^2(d,t)$ under some regularity conditions. In turn, we use this result to show that $\sigma_q^2(d) \to \sigma_s^2$ as $d \to \infty$, and hence that $\sigma_q^2 = \sigma_s^2$. That is, we can use EBQP methods to approximate numerically the asymptotic variance of Spearman's rank correlation.

3. The Relationship Between the Point Estimates $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ and Their Means

In this section we present several results concerning the relationship between the point estimates $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ and between their finite and asymptotic means. To simplify the complicated calculations in Sections 3 and 4, we assume that we are dealing with balanced data (i.e., $m = t/d$ is an integer). The finite results in these sections hold approximately for unbalanced data, but the asymptotic results are unchanged.

Theorem 1. *For any distribution F and for all $2 \le d \le t = dm$,*

$$|\hat{\rho}_s(t) - \hat{\rho}_q(d,t)| < 2\{1 + (d+1)(m+2)m^{-1}\}d^{-2}. \tag{3.1}$$

Proof. See Appendix A.

This theorem gives the maximum difference between $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, and in practice the actual difference tends to be much smaller. Using this theorem for the relationship between these point estimates, we now prove the following relationship between their finite and asymptotic means.

Corollary 1. *For any distribution F and for all $2 \le d \le t$,*

(i) $|\rho_s(t) - \rho_q(d,t)| < 2\{1 + (d+1)(m+2)m^{-1}\}d^{-2}$;
(ii) $|\rho_s - \rho_q(d)| \le 2(d+2)d^{-2}$;
(iii) $\rho_s = \rho_q$.

Proof. Result (i) follows immediately from Theorem 1 by taking expectations. Result (ii) follows from (i) as $m \to \infty$ with d constant. Result (iii) follows from (ii) as $d \to \infty$.

Comment. In the special case where $d = 2$, Kruskal (1958) gave the more precise inequality $\frac{3}{16}\{1 + \rho_q(2)\}^3 - 1 \le \rho_s \le 1 - \frac{3}{16}\{1 - \rho_q(2)\}^3$. Kruskal's methods employ the fact that 2×2 EBQP tables have one free cell, and thus these methods cannot be used for $d > 2$.

Corollary 1(i) shows that $\rho_q(d)$ converges to ρ_s rather slowly (order d^{-1}), but for certain underlying bivariate distributions the rate of convergence is much faster (order d^{-2}). Consider the following definition and property.

Definition 1. A conditional distribution $G(x|y)$ is stochastically nondecreasing if $y_1 \leq y_2$ implies that $P(X > x|Y = y_1) = 1 - G(x|y_1) \leq 1 - G(x|y_2) = P(X > x|Y = y_2)$ for all x.

Property 1. Let $F(x, y)$ be a distribution with stochastically nondecreasing conditional distributions $G(x|y)$ and $H(y|x)$.

The BVN, BLN, BHN and BCS distributions with nonnegative correlations satisfy Property 1, but the TS distribution does not. For those distributions that satisfy Property 1, we obtain the following theorem with tighter bounds for the difference between the finite means $\rho_s(t)$ and $\rho_q(d, t)$ than those given in Corollary 1(i).

Theorem 2. *For any distribution F that satisfies Property* 1 *and for all* $2 \leq d \leq t$,

$$-d^{-2} \leq \rho_s(t) - \rho_q(d, t) \leq \{1 + 2(m + 1)m^{-1}\}d^{-2}. \tag{3.2}$$

Proof. See Appendix A.

Corollary 2. *For any distribution F that satisfies Property* 1 *and for all* $2 \leq d$,

$$0 \leq \rho_s - \rho_q(d) \leq 3d^{-2}.$$

Proof. This result follows from Theorem 2 as $m \to \infty$ with d constant, and from the observation that $\hat{\rho}_q(d, t)$ is asymptotically "biased towards the null" as an estimator of ρ_s for these distributions, so $0 \leq \rho_q(d) \leq \rho_s$.

Tables 1(b) and 2(b) show $\{\rho_q(d)\}$ for the standard BVN, BCS and TS distributions for selected quantile-categories. For the BVN and BCS distributions, the $\rho_q(d)$ do indeed appear to converge rapidly to ρ_s (order d^{-2}), as Corollary 2 predicts. Furthermore, by the invariance and symmetry properties of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d, t)$, these results apply to all BVN, BHN, BCS and BLN distributions, regardless of their location, scale, and correlation/shape parameters.

4. The Relationship Between the Finite and Asymptotic Variances of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d, t)$

The bounds for the difference between the finite variances $\sigma_s^2(t)$ and $\sigma_q^2(d, t)$ are more difficult to calculate. While $\sigma_s^2(t)$ achieves its maximum value of $t(t - 1)^{-1}$ under independence, $\sigma_q^2(d, t)$ can have values that exceed $t(t - 1)^{-1}$ for certain pathological distributions, including the TS distribution (and the nicked square distribution; e.g. Borkowf, Gail, Carroll and Gill (1997)). Let $d_c(F)$ denote the smallest dimension such that $\sigma_q^2(d) \leq 1$ for all $d \geq d_c(F)$. For distributions that satisfy Property 1, $d_c(F) = 2$, while for the TS distribution $d_c(F) = 30$.

Theorem 3. *For any distribution F and for all $d_c(F) \leq d \leq t$,*

$$|\sigma_q^2(d,t) - \sigma_s^2(t)| < (2\sqrt{2}d^3 + 4d^2 - 3)d^{-4}t(t-1)^{-1}. \tag{4.1}$$

Proof. See Appendix A.

Once again, this theorem gives the maximum difference between $\sigma_s^2(t)$ and $\sigma_q^2(d,t)$, and in practice the actual difference tends to be much smaller. Using this theorem for the absolute difference between these finite variances, we prove the following relationships between the corresponding asymptotic variances.

Corollary 3. *For any distribution F and for all $d_c(F) \leq d$,*
(i) $|\sigma_q^2(d) - \sigma_s^2| \leq (2\sqrt{2}d^3 + 4d^2 - 3)d^{-4}$;
(ii) $\sigma_q^2 = \sigma_s^2$.

Proof. Result (i) follows from Theorem 3 as $m \to \infty$ with d constant. Result (ii) follows from (i) as $d \to \infty$.

For those distributions that satisfy Property 1, we obtain the following theorem with tighter bounds for the difference between the finite variances $\sigma_s^2(t)$ and $\sigma_q^2(d,t)$ than those given in Theorem 3.

Theorem 4. *For any distribution F that satisfies Property 1 and for all $2 \leq d \leq t$,*

$$\begin{aligned} -(2d^2+1)d^{-4}t(t-1)^{-1} &< \sigma_q^2(d,t) - \sigma_s^2(t) \\ &< \{2(\sqrt{2}+1)d^2 + 2\sqrt{2}d - (2\sqrt{2}+1)\}d^{-4}t(t-1)^{-1}. \end{aligned} \tag{4.2}$$

Proof. See Appendix A.

Corollary 4. *For any distribution F that satisfies Property 1 and for all $2 \leq d$,*

$$0 \leq \sigma_q^2(d) - \sigma_s^2 \leq \{2(\sqrt{2}+1)d^2 + 2\sqrt{2}d(2\sqrt{2}+1)\}d^{-4}.$$

Proof. This result follows from Theorem 4 as $m \to \infty$ with d constant, and from the observation that for these distributions, $0 \leq \sigma_s^2 \leq \sigma_q^2(d)$.

Tables 1(b) and 2(b) show $\{\sigma_q^2(d)\}$ for the standard BVN, BCS and TS distributions for selected quantile-categories. For the BVN and BCS distributions, the $\sigma_q^2(d)$ do indeed appear to converge rapidly to σ_s^2 (order d^{-2}), as Corollary 4 predicts. Furthermore, by the invariance and symmetry properties of $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$, these results apply to all BVN, BHN, BCS and BLN distributions, regardless of their location, scale, and correlation/shape parameters.

By contrast, for the TS distributions, $\sigma_q^2(d)$ converges more slowly to σ_s^2 for the TS distribution and some values of $\sigma_q^2(d)$ even exceed 1! We observe that these values seem to form three subsequences, one each for $d = 3k$, $d = 3k+1$,

and $d = 3k + 2$. The variances for this first subsequence are defined using the improper conditional distributions at the marginal tertiles mentioned above, and they appear to converge most rapidly to the unknown asymptotic variance σ_s^2, which should be close to the simulated variance $\overline{\sigma}_s^2(1000) = 0.9006$. This result reflects the fact that when $d = 3k$, the EBQP estimation methods take account of the unusual features of the TS distribution at the tertiles. At the same time, the second and third subsequences appear to converge more slowly from below and above, respectively. Thus, we can use the subsequence with $d = 3k$ to approximate σ_s^2 most rapidly.

5. An Example from Nutritional Epidemiology

Pietinen, et al. (1988) conducted an extensive study on Finnish men aged 55-69 to test the reproducibility and validity of several methods of measuring the intake of food items and nutrients. In the validation part of the study, the total fat intake (in grams) of 157 men was measured by two methods. First, the subjects kept prospective "food records" to record the foods they consumed on 12 two-day periods during a 6 month interval. Second, the subjects completed retrospective "food use questionnaires" to estimate how much of certain foods they had consumed during the previous year.

Because both sets of marginal data were skewed to the right, we used natural logarithms to transform these data. Let X and Y denote the log-transformed food record and food use questionnaire measurements, respectively. The means and standard deviations of these measurements (in log-grams) are $\overline{x} = 4.6051$, $s_x = 0.2372$, $\overline{y} = 4.5657$, and $s_y = 0.3792$, and the sample correlation is $\hat{\rho} = 0.5598$.

We performed a series of graphical tests (not shown) to study the underlying bivariate distribution of the fat intake data. We created scatter plots of the data on both the original and natural logarithm scales. We also created normal probability plots of the marginal data and linear combinations of the marginal data. Together, these graphical tests suggest that the log-transformed data are consistent with the BVN distribution, and hence the original data are consistent with the BLN distribution.

In Sections 3 and 4 we assumed that ties occur with probability zero. However, in the fat intake data set there were indeed 12 ties in the X variable and 9 ties in the Y variable. We added tiny random errors to the marginal data to break these ties and computed a sample Spearman's rank correlation of $\hat{\rho}_s(157) = 0.5620$ with no ties. By comparison, if we use the formula for ρ_s that assigns tied observation midranks rather than breaking the ties, we compute $\hat{\rho}_s(157) = 0.5622$ (e.g., Borkowf, submitted). We have found through extensive simulations with BVN data that a moderate amount of rounding breaking ties

usually has only a small impact on the point estimates and estimated variances of $\rho_s, \rho_q(d)$, and other measures of agreement calculated from EBQP tables.

Table 5. Point estimates and estimated variances of the quantile correlation for the fat intake data for various table dimensions, and the corresponding asymptotic values under the BVN model with $\hat{\rho} = 0.5598$.*

		empirical results		BVN model	
d	m	$\hat{\rho}_q(d,t)$	$\hat{\sigma}_q^2(d,t)$	$\rho_q(d)$	$\sigma_q^2(d)$
2	79	0.3885	0.8642	0.3783	0.8569
3	53	0.4762	0.7187	0.4610	0.6902
4	40	0.5082	0.6215	0.4932	0.6307
5	32	0.5414	0.7694	0.5092	0.6022
6	27	0.5382	0.6184	0.5184	0.5862
7	23	0.5534	0.5662	0.5241	0.5762
8	20	0.5651	0.5191	0.5280	0.5696
9	18	0.5316	0.5862	0.5307	0.5649
10	16	0.5468	0.6138	0.5326	0.5615
11	15	0.5632	0.5497	0.5341	0.5589
12	14	0.5590	0.5379	0.5353	0.5569
13	13	0.5644	0.5539	0.5362	0.5553
14	12	0.5569	0.5561	0.5369	0.5541
15	11	0.5642	0.5452	0.5375	0.5530
16	10	0.5668	0.5128	0.5380	0.5522
17	10	0.5671	0.5143	0.5384	0.5514
18	9	0.5555	0.5316	0.5388	0.5508
19	9	0.5530	0.5546	0.5391	0.5503
20	8	0.5537	0.5402	0.5393	0.5499
157	1	0.5620	c.c.	c.c.	c.c.
∞	n.a.	n.a.	n.a.	0.5418	0.5454

*Note that $\hat{\rho}_q(157,157) = \hat{\rho}_s(157) = 0.5620$. Also, under the BVN model with $\hat{\rho} = 0.5598$, we compute $\rho_s = 0.5418$ (Moran (1948)) and $\sigma_s^2 = 0.5454$ (Fieller, Hartley and Pearson (1957)). Abbreviations: c.c.=cannot calculate with current computing resources, n.a.=not applicable.

We also computed the sample quantile correlations $\hat{\rho}_q(d,157)$ and their estimated variances $\hat{\sigma}_q^2(d,157)$ for $d = 2,\ldots,20$ (Table 5). We observe that $\hat{\rho}_q(d,157)$ rapidly approaches $\hat{\rho}_s(157) = 0.5620$, and the differences $\hat{\rho}_s(157) - \hat{\rho}_q(d,157)$ fall well within the conservative bounds of Theorem 1. The $\hat{\sigma}_q^2(d,157)$ are more variable, but presumably approach $\hat{\sigma}_s^2(157)$, which we cannot estimate directly because of limited computing power. Under the BVN model with $\rho = 0.5598$, we calculate $\sigma_s^2(157) = 0.5544$ (David and Mallows (1961)), and the differences $\sigma_s^2(157) - \hat{\sigma}_q^2(d,157)$ also fall well within the conservative bounds of Theorem 3. Nonparametrically, we can also estimate the finite variance $\sigma_q^2(20,157)$ from a

20×20 EBQP table by $\hat{\sigma}_q^2(20, 157) = 0.5402$, and then use this variance as an approximation for the desired finite variance $\hat{\sigma}_s^2(157)$.

In turn, we can construct a large sample $(1-\alpha)100\%$ confidence interval for ρ_s of the form $\hat{\rho}_s(t) \pm \Phi^{-1}(1-\alpha/2)\{\hat{\sigma}_q^2(d,t)/t\}^{1/2}$, where Φ denotes the standard normal distribution. For the fat intake data, we compute a 95% confidence interval for ρ_s of (0.4470, 0.6770), which shows that a moderate degree of agreement exists between the ranks given by the two methods of measuring fat intake. Borkowf (submitted) discusses the construction of small sample confidence intervals for ρ_s using the t-distribution and Fisher's z-transformation.

For comparison, we can use the sample correlation $\hat{\rho} = 0.5598$ to calculate $\{\rho_q(d)\}$ and $\{\sigma_q^2(d)\}$ under the BVN model (Table 5). These asymptotic means and variances are comparable with the sample values that we estimated from the data. Furthermore, for the BVN distribution with $\rho = 0.5598$ and $t = 157$, we calculate $\rho_s(157) = 0.5387$ and $\rho_s = 0.5418$ (Moran (1948)), $\sigma_s^2(157) = 0.5544$ (David and Mallows (1961)), and $\sigma_s^2 = 0.5454$ (Fieller, Hartley and Pearson (1957)). We observe that $\rho_q(d)$ and $\sigma_q^2(d)$ appear to converge rapidly to ρ_s and σ_s^2, as we expect.

6. Discussion

In this paper we have shown that Spearman's rank correlation, $\hat{\rho}_s(t)$, and the quantile correlation, $\hat{\rho}_q(d,t)$, have distributions derived from the EBQP distribution. Thus, we can use EBQP methods to estimate the finite variances $\sigma_s^2(t)$ and $\sigma_q^2(d,t)$ and to compute the asymptotic variances $\sigma_q^2(d)$. We have proved that the asymptotic means of the quantile correlation converge to the asymptotic mean of Spearman's rank correlation, i.e., $\rho_q(d) \to \rho_s$ as $d \to \infty$ for all underlying bivariate distributions. We have also proved that the asymptotic variances of the quantile correlation converge to the asymptotic variance of Spearman's rank correlation, i.e., $\sigma_q^2(d) \to \sigma_s^2$ as $d \to \infty$ for all underlying bivariate distributions that satisfy certain regularity conditions. We note that while these means and variances converge slowly in general (order d^{-1}), they converge more rapidly for distributions that satisfy Property 1 (order d^{-2}), including the BVN, BLN, BHN and BCS distributions. Indeed, in epidemiological studies we should perform graphical tests to determine whether the data appear to come from an underlying distribution that satisfies Property 1.

We also note that standard EBQP theory assumes that ties occur with probability zero. Nevertheless, we have found it quite satisfactory in practice to break ties by adding tiny random errors to the original bivariate data when there are only a moderate number of ties in the data. Alternatively, for small sample sizes ($t \leq 20$) we can construct EBQP tables that take these ties into account without breaking them (Borkowf, submitted).

In addition, the results for the finite means and variances that we proved in Sections 3 and 4 hold approximately for unbalanced data, but the asymptotic results are unchanged. In practice, we need to adjust EBQP estimation methods slightly to accommodate unbalanced data, and it is preferable to use balanced data for reasons of esthetics and estimation.

At present, due to limits in computing power, we can only calculate the covariances of $d \times d$ EBQP tables for $d \leq 20$ directly from data using matrix algebra in the GAUSS 3.0 programming language (Aptech Systems Inc. (1992)) on a Pentium 133MHz processor. For a given distribution F, we can calculate the asymptotic covariances of $d \times d$ EBQP tables for $d \leq 40$ by iterative methods, but the calculations become prohibitively slow for larger tables. We expect that as computer technology continues to evolve in the next decade, we will be able to compute the covariances of even larger EBQP tables, and thus compute $\hat{\sigma}_s^2(t)$, $\hat{\sigma}_q^2(d,t)$, and $\sigma_q^2(d)$ for larger values of d and t.

For most epidemiological studies, the observation that $\sigma_q^2(d)$ converge rapidly to σ_s^2 for distributions that satisfy Property 1 justifies the current practice of collapsing large EBQP tables by combining consecutive rows and columns to create smaller tables. For examples, given a sample size of t, we can construct a smaller $d \times d$ EBQP table with $m = \lceil t/d \rceil$ or $(m-1)$ observations in each row and column in order to estimate $\sigma_q^2(d,t)$. We can then use $\hat{\sigma}_q^2(d,t)$ as an approximation for the desired finite variance $\hat{\sigma}_s^2(t)$. We can, of course, compute $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$ in the usual manner and, when these point estimates are close, the estimated variances will also tend to be close, even for small values of d.

The author can provide a computer program (EpiQuant 1.0) to estimate the covariances of EBQP tables from bivariate data and to calculate the asymptotic covariance of EBQP tables for the BVN distribution. In turn, the variances of measures of agreement calculated from such tables can be computed from these covariance matrices by the delta method (Bishop, Fienberg and Holland (1975)). This program is written as a batch file in the GAUSS 3.0 programming language (Aptech Systems Inc. (1992)) and comes with a brief technical note to explain its use (Borkowf (1997)). (Readers may contact the author by e-mail: borkowfc@gwgate.nhlbi.nih.gov.)

Acknowledgements

This research was performed while the author held a National Research Council-National Institutes of Health Research Associateship at the National Heart, Lung and Blood Institute's Office of Biostatistics Research. He wishes to thank Mrs. Anne Hartman and her collaborators (Pietinen et al. (1988)) for the fat intake data and Drs. Nancy Geller, Dean Follmann, and Joanna Shih for helpful comments on this manuscript.

Appendix A. Proofs of Theorems

Proof of Theorem 1. First, we compute bounds for the four difference terms from which we compute $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$. For any distribution F, the difference terms D_R, D_Q and D_p achieve their minimum (maximum) values under PRA (PRD). Thus,

$$0 \le D_R \le \frac{1}{3}(t^2-1),$$
$$0 \le D_Q \le \frac{1}{3}(d^2-1), \tag{A.1}$$
$$0 \le D_p \le \frac{1}{3}(m^2-1). \tag{A.2}$$

By contrast, the bounds for D_{QP} require more effort to compute. First, since the elements of the pairs $\{Q_x(X_k), P_x(X_k)\}$ and $\{Q_y(Y_k), P_y(Y_k)\}$ are always independent,

$$t^{-1}\sum_{k=1}^{t} Q_x(X_k)P_x(X_k) = t^{-1}\sum_{k=1}^{t} Q_y(Y_k)P_y(Y_k) = \frac{1}{4}(d+1)(m+1). \tag{A.3}$$

However, the elements of the pairs $\{Q_x(X_k), P_y(Y_k)\}$ and $\{Q_y(Y_k), P_x(X_k)\}$ usually are not independent. We also find that

$$\frac{1}{6}(d+1)(m+1) < t^{-1}\sum_{k=1}^{t} Q_x(X_k)P_y(Y_k) < \frac{1}{3}(d+1)(m+1), \tag{A.4}$$
$$\frac{1}{6}(d+1)(m+1) < t^{-1}\sum_{k=1}^{t} Q_y(Y_k)P_x(X_k) < \frac{1}{3}(d+1)(m+1). \tag{A.5}$$

We can then use equations (2.7) and (A.3) through (A.5) to show that

$$-\frac{1}{6}(d+1)(m+1) < D_{QP} < \frac{1}{6}(d+1)(m+1). \tag{A.6}$$

The difference term D_{QP} approaches its extrema only in rare cases.

Next, we can use equations (2.5), (2.6) and (2.8) to expand $\Delta_M = \hat{\rho}_s(t) - \hat{\rho}_q(d,t)$ as the decomposition

$$\begin{aligned}\Delta_M &= \{1 - 6D_R(t^2-1)^{-1}\} - \{1 - 6D_Q(d^2-1)^{-1}\} \\ &= -6(m^2D_Q + D_p + 2mD_{QP})(t^2-1)^{-1} + 6D_Q(d^2-1)^{-1} \\ &= 6D_Q\{(d^2-1)^{-1} - m^2(t^2-1)^{-1}\} - 6(D_P + 2mD_{QP})(t^2-1)^{-1} \\ &= T_1 + T_2 + T_3, \end{aligned} \tag{A.7}$$

where $T_1 = 6D_Q\{(d^2-1)^{-1} - m^2(t^2-1)^{-1}\}$, $T_2 = -6D_P(t^2-1)^{-1}$, and $T_3 = -12mD_{QP}(t^2-1)^{-1}$.

Also, we have the inequality (for all $2 \leq d < t$),

$$0 < (d^2-1)^{-1} - m^2(t^2-1)^{-1} < d^{-2}(d^2-1)^{-1}. \tag{A.8}$$

We can then use equations (A.1) and (A.8), (A.2) and (A.6) to compute bounds for the three terms in equation (A.7), respectively. Thus,

$$0 \leq T_1 < 6\{\frac{1}{3}(d^2-1)\}\{d^{-2}(d^2-1)^{-1}\} = 2d^{-2}, \tag{A.9}$$

$$0 \geq T_2 \geq -6\{\frac{1}{3}(m^2-1)\}\{(t^2-1)^{-1} > -2d^{-2}, \tag{A.10}$$

$$|T_3| \leq 12m\{\frac{1}{6}(d+1)(m+1)\}(t^2-1)^{-1} < 2(d+1)d^{-2}\{(m+2)m^{-1}\}. \tag{A.11}$$

Together, equations (A.7) and (A.9) through (A.11) yield statement (3.1), as we wished to prove.

Proof of Theorem 2. First, we can compute even tighter bounds for the four difference terms. Under independence (H_0), the empirical cell proportions $\{p_{ij}^R\}$ and $\{p_{ij}^Q\}$ have the MH distribution, and hence the difference terms D_R, D_Q, D_P and D_{QP} have related distributions. Thus, we calculate $E(D_R|H_0) = \frac{1}{6}(t^2-1)$, $E(D_Q|H_0) = \frac{1}{6}(d^2-1)$, $E(D_P|H_0) = \frac{1}{6}(m^2-1)$, and $E(D_{QP}|H_0) = 0$. Corresponding to equation (2.8), these expectations satisfy the relationship $E(D_R|H_0) = m^2E(D_Q|H_0) + E(D_P|H_0)$.

Next, the difference terms D_R, D_Q and D_P achieve their minimum (maximum) values under PRA (independence) for all $2 \leq d \leq t$. Thus,

$$0 \leq E(D_R) \leq E(D_R|H_0),$$

$$0 \leq E(D_Q) \leq E(D_Q|H_0), \tag{A.12}$$

$$0 \leq E(D_P) \leq E(D_P|H_0). \tag{A.13}$$

Furthermore, we calculate that

$$0 \leq \operatorname{Corr}\{Q_x(X_k), P_y(Y_k)\} \leq \operatorname{Corr}\{R_x(X_k), P_y(Y_k)\} < d^{-1}, \tag{A.14}$$

$$0 \leq \operatorname{Corr}\{Q_y(Y_k), P_x(X_k)\} \leq \operatorname{Corr}\{R_y(Y_k), P_x(X_k)\} < d^{-1}, \tag{A.15}$$

where $\operatorname{Corr}\{R_x(X_k), P_y(Y_k)\}$ and $\operatorname{Corr}\{R_y(Y_k), P_x(X_k)\}$ approach their minimum (maximum) values under independence (PRA). We can also use equations (2.7) and (A.3) to show that

$$E(D_{QP}) = -\operatorname{Cov}\{Q_x(X_k), P_y(Y_k)\} - \operatorname{Cov}\{Q_y(Y_k), P_x(X_k)\}. \tag{A.16}$$

Then, we can use equations (A.14) through (A.16) to show that

$$E(D_{QP}|H_0) \geq E(D_{QP}) > -2d^{-1}(V_QV_P)^{\frac{1}{2}} > -\frac{1}{6}m. \tag{A.17}$$

We can use equations (A.8) and (A.12), (A.13) and (A.17) to compute tighter bounds for the expectations of the three terms in equation (A.7), respectively. Thus,

$$0 \leq E(T_1) < 6\{\frac{1}{6}(d^2-1)\}\{d^{-2}(d^2-1)^{-1}\} = d^{-2}, \tag{A.18}$$

$$0 \geq E(T_2) \geq -6\{\frac{1}{6}(m^2-1)\}(t^2-1)^{-1} > d^{-2}, \tag{A.19}$$

$$0 \leq E(T_3) < -12m(-\frac{1}{6}m)(t^2-1)^{-1} < 2d^{-2}\{(m+1)m^{-1}\}. \tag{A.20}$$

Together, equations (A.7) and (A.18) through (A.20) yield statement (3.2), as we wished to prove.

Proof of Theorem 3. First, we compute bounds for the variances of the four difference terms from which we compute $\hat{\rho}_s(t)$ and $\hat{\rho}_q(d,t)$. Under independence (H_0), the difference terms D_R, D_Q, D_P and D_{QP} have distributions related to the MH distribution. Thus, we calculate

$$\mathrm{Var}\,(D_R|H_0) = \frac{1}{36}(t^2-1)^2(t-1)^{-1}, \quad \mathrm{Var}\,(D_Q|H_0) = \frac{1}{36}(d^2-1)^2(t-1)^{-1},$$
$$\mathrm{Var}\,(D_P|H_0) = \frac{1}{36}(m^2-1)^2(t-1)^{-1}, \quad \text{and}$$
$$\mathrm{Var}\,(D_{QP}|H_0) = \frac{1}{72}(d^2-1)(m^2-1)(t-1)^{-1}.$$

Corresponding to equation (2.8), these variances satisfy the relationship $\mathrm{Var}\,(D_R|H_0) = m^4\mathrm{Var}\,(D_Q|H_0) + \mathrm{Var}\,(D_P|H_0) + 4m^2\mathrm{Var}\,(D_{QP}|H_0)$.

Next, for any distribution F, the variances of the difference terms achieve their minimum (maximum) values under PRA/PRD (independence). Thus,

$$0 \leq \mathrm{Var}\,(D_R) \leq \mathrm{Var}\,(D_R|H_0),$$
$$0 \leq \mathrm{Var}\,(D_Q) \leq \mathrm{Var}\,(D_Q|H_0), \tag{A.21}$$
$$0 \leq \mathrm{Var}\,(D_P) \leq \mathrm{Var}\,(D_P|H_0), \tag{A.22}$$
$$0 \leq \mathrm{Var}\,(D_{QP}) \leq \mathrm{Var}\,(D_{QP}|H_0). \tag{A.23}$$

Furthermore, since $|\,\mathrm{Cov}\,(A,B)| \leq \{\mathrm{Var}\,(A)\mathrm{Var}\,(B)\}^{\frac{1}{2}}$, we can use the above bounds on the variances of the difference terms to compute the bounds on their covariances. Thus,

$$|\,\mathrm{Cov}\,(D_Q, D_P)| \leq \frac{1}{36}(d^2-1)(m^2-1)(t-1)^{-1}, \tag{A.24}$$

$$|\,\mathrm{Cov}\,(D_Q, D_{QP})| < \frac{\sqrt{2}}{72}(d^2-1)dm(t-1)^{-1}, \tag{A.25}$$

$$|\,\mathrm{Cov}\,(D_P, D_{QP})| \leq \frac{\sqrt{2}}{72}(m^2-1)dm(t-1)^{-1}. \tag{A.26}$$

Next, corresponding to equation (2.8), the variances and covariances of the difference terms satisfy the relationship

$$\begin{aligned}\mathrm{Var}\,(D_R) &= m^4\mathrm{Var}\,(D_Q) + \mathrm{Var}\,(D_P) + 4m^2\mathrm{Var}\,(D_{QP}) \\ &\quad +2m^2\mathrm{Cov}\,(D_Q, D_P)+4m^3\mathrm{Cov}\,(D_Q, D_{QP})+4m\,\mathrm{Cov}\,(D_P, D_{QP}). \quad \text{(A.27)}\end{aligned}$$

We can then use equations (2.5), (2.6) and (A.27) to expand $\Delta_V = \sigma_q^2(d,t) - \sigma_s^2(t)$ as the decomposition

$$\begin{aligned}\Delta_V &= 36t\mathrm{Var}\,(D_Q)(d^2-1)^{-2} - 36t\mathrm{Var}\,(D_R)(t^2-1)^{-2} \\ &= T_4 + T_5 + T_6, \quad \text{(A.28)}\end{aligned}$$

where $T_4 = 36t\mathrm{Var}\,(D_Q)\{(d^2-1)^{-2} - m^4(t^2-1)^{-2}\}$,

$$\begin{aligned}T_5 &= -36t\{\mathrm{Var}\,(D_P) + 4m^2\mathrm{Var}\,(D_{QP})\}(t^2-1)^{-2}, \quad \text{and} \\ T_6 &= -36t\{2m^2\,\mathrm{Cov}\,(D_Q, D_P)+4m^3\,\mathrm{Cov}\,(D_Q, D_{QP}) \\ &\quad +4m\,\mathrm{Cov}\,(D_P, D_{PQ})\}(t^2-1)^{-2}.\end{aligned}$$

Also, we have the inequality (for all $2 \le d < t$),

$$0 < (d^2-1)^{-2} - m^4(t^2-1)^{-2} < (2d^2-1)d^{-4}(d^2-1)^{-2}. \quad \text{(A.29)}$$

We can then use equations (A.21) through (A.26) and (A.29) to compute bounds for the three terms in equation (A.28). Thus,

$$\begin{aligned}0 \le T_4 &< 36t\{\frac{1}{36}(d^2-1)^2(t-1)^{-1}\}\{(2d^2-1)d^{-4}(d^2-1)^{-2}\} \\ &< (2d^2-1)d^{-4}t(t-1)^{-1}, \quad \text{(A.30)} \\ 0 \ge T_5 &\ge -36t\{\frac{1}{36}(m^2-1)^2 + \frac{1}{18}m^2(d^2-1)(m^2-1)\}(t^2-1)^{-2}(t-1)^{-1} \\ &> -(2d^2-1)d^{-4}t(t-1)^{-1}, \quad \text{(A.31)} \\ |T_6| &< 36t\{\frac{1}{18}m^2(d^2-1)(m^2-1) + \frac{\sqrt{2}}{18}m^3(d^2-1)dm \\ &\quad +\frac{\sqrt{2}}{18}m(m^2-1)dm\}(t^2-1)^{-2}(t-1)^{-1} \\ &< (2\sqrt{2}d^3 + 2d^2 - 2)d^{-4}t(t-1)^{-1}. \quad \text{(A.32)}\end{aligned}$$

Together, equations (A.28) and (A.30) through (A.32) yield statement (4.1), as we wished to prove.

Proof. of Theorem 4. First, we can use equations (A.14) and (A.15) to show that

$$0 \le \mathrm{Cov}\,(D_Q, D_P) < d^{-2}\{\mathrm{Var}\,(D_Q)\mathrm{Var}\,(D_P)\}^{\frac{1}{2}} < \frac{1}{36}(m^2-1)(t-1)^{-1}, \quad \text{(A.33)}$$

$$0 \geq \mathrm{Cov}\,(D_Q, D_{QP}) > d^{-1}\{\mathrm{Var}\,(D_Q)\mathrm{Var}\,(D_{QP})\}^{\frac{1}{2}} > -\frac{\sqrt{2}}{72}(d^2-1)m(t-1)^{-1}, \quad \text{(A.34)}$$

$$0 \geq \mathrm{Cov}\,(D_P, D_{QP}) > -\{\mathrm{Var}\,(D_P)\mathrm{Var}\,(D_{QP})\}^{\frac{1}{2}} > -\frac{\sqrt{2}}{72}(m^2-1)dm(t-1)^{-1}. \quad \text{(A.35)}$$

Next, we can use equations (A.33) through (A.35) to compute tighter bounds for the third term (T_6) in equation (A.28). Thus,

$$\begin{aligned} &-36t\{\frac{1}{18}m^2(m^2-1)\}(t^2-1)^{-2}(t-1)^{-1} < T_6 \\ &< 36t\{\frac{\sqrt{2}}{18}m^3(d^2-1)m + \frac{\sqrt{2}}{18}m(m^2-1)dm\}(t^2-1)^{-2}(t-1)^{-1}, \end{aligned}$$

which implies that

$$-2d^{-4}t(t-1)^{-1} < T_6 < 2\sqrt{2}(d^2+d-1)d^{-4}t(t-1)^{-1}. \quad \text{(A.36)}$$

Together, equations (A.28), (A.30), (A.31) and (A.36) yield statement (4.2), as we wished to prove.

References

Aptech Systems Inc. (1992). *The GAUSS System Version* 3.0. Aptech Systems Inc., Maple Valley, Washington.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge.

Borkowf, C. B. (1997). *The Empirical and Parametric Bivariate Quantile-Partitioned Distributions*. Dissertation in Statistics, Cornell University.

Borkowf, C. B., Gail, M. H., Carroll, R. J. and Gill, R. D. (1997). Analyzing bivariate continuous data grouped into categories defined by empirical quantiles of marginal distributions. *Biometrics* **53**, 1054-1069.

Borkowf, C. B. A new nonparametric method for estimating the finite sample variance of Spearman's rank correlation. Submitted.

David, F. N. and Mallows, C. L. (1961). The variance of Spearman's rho in normal samples. *Biometrika* **48**, 19-28.

David, S. T., Kendall, M. G. and Stuart, A. (1951). Some questions of distribution in the theory of rank correlation. *Biometrika* **38**, 131-140.

Fieller, E. C., Hartley, H. O. and Pearson, E. S. (1957). Tests for rank correlation coefficients. I. *Biometrika* **44**, 470-481.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33**, 613-619.

Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley, New York.

Kendall, M. G. (1949). Rank and product-moment correlation. *Biometrika* **36**, 177-193.

Kruskal, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.* **53**, 814-861.

Moran, P. A. P. (1948). Rank correlation and product-moment correlation. *Biometrika* **35**, 203-206.

Pearson, K. (1907). *On Further Methods of Determining Correlation.* Drapers' Company Research Memoirs, Biometric Series IV, Mathematical Contributions to the Theory of Evolution, XVI. Dulau, London.

Pietinen, P., Hartman, A. M., Haapa, E., Rasanen, L., Haapakoski, J., Palmgren, J., Albanes, D., Virtamo, J. and Huttunen, J. K. (1988). Reproducibility and validity of dietary assessment instruments. I. A self-administered food use questionnaire with a portion size picture booklet. *Amer. J. Epidemiology* **128**, 655-666.

Plackett, R. L. (1981). *The Analysis of Categorical Data*. 2nd edition. Macmillan, New York.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72-101.

Spearman, C. (1906). 'Footrule' for measuring correlation. *British Journal of Psychology* **2**, 89-108.

Spitzer, R. L., Cohen, J., Fleiss, J. L. and Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry* **17**, 83-87.

National Heart, Lung and Blood Institute, Division of Epidemiology and Clinical Applications, Office Biostatistics Research, Two Rockledge Centre, Room 8100D, 6701 Rockledge Drive, MSC 7938, Bethesda, MD 20892-7938, U.S.A.

E-mail: borkowfc@gwgate.nhlbi.nih.gov

(Received December 1997; accepted July 1998)