PSYCHOMETRIKA-VOL. 42, NO. 3
September 1977

ON THE MEAN AND VARIANCE OF THE TETRACHORIC CORRELATION COEFFICIENT

MORTON B. BROWN

TEL-AVIV UNIVERSITY

JACQUELINE K. BENEDETTI

UNIVERSITY OF CALIFORNIA AT LOS ANGELES

Estimates of the mean and standard deviation of the tetrachoric correlation are compared with their expected values in several 2 × 2 tables. Significant bias in the mean is found when the minimum cell frequency is less than 5. Three formulas for the standard deviation are compared and guidelines given for their use.

Key words: fourfold correlation, 2 × 2 frequency table, bivariate normal.

Introduction

Data from a bivariate normal distribution are often displayed in a fourfold table where each variable is recorded as a dichotomy. Based upon the cell frequencies of the 2 × 2 table, Pearson [1901] proposed the tetrachoric correlation coefficient as an estimate of the correlation of the underlying bivariate normal distribution. The cell probabilities correspond to the probabilities of the four quadrants of the bivariate normal having the same marginals. For example, let the observed frequency table be

$$\begin{array}{cc|c} a & b & a+b \\ c & d & c+d \\ \hline a+c & b+d & N=a+b+c+d \end{array}$$

where N is the total frequency. Let z_1 and z_2 be standard normal deviates corresponding to the marginal probabilities $(a + c)/N$ and $(a + b)/N$; i.e.,

$$\Phi(z_1) = \frac{(a+c)}{N}$$

$$\Phi(z_2) = \frac{(a+b)}{N}$$

This research was performed when the first author was on leave at the University of California at Los Angeles and was supported in part by NIH Special Research Resources Grant RR-3. The second author was also supported by NIH Fellowship 5 F22 GM00328-02.

Requests for reprints should be addressed to Dr. Jacqueline Benedetti at the Department of Biomathematics, AV-111, UCLA School of Medicine, Los Angeles, Ca. 90024.

where $\Phi(z)$ is the *cdf* of the standard normal. Then the tetrachoric correlation r is obtained by solving

$$\int_{-\infty}^{z_2}\int_{-\infty}^{z_1}\phi(x_1, x_2, r)\,dx_1\,dx_2 = \frac{a}{N} \tag{1}$$

for r where $\phi(x_1, x_2, r)$ is the bivariate normal density with means zero, variances one and correlation r. The probabilities of the four quadrants formed by dichotomizing the variables by the lines $x_1 = z_1$ and $x_2 = z_2$ are equal to a/N, b/N, c/N and d/N.

When the marginal probabilities are equal (i.e., $z_1 = z_2 = 0$), the tetrachoric correlation r is obtained by

$$r = -\cos\left(\frac{2\pi a}{N}\right). \tag{2}$$

Otherwise the integral in (1) is replaced by the infinite tetrachoric series expansion in powers of r and a root of (1) must be found numerically. Manually this is a major undertaking and tables were prepared to ease some of the calculations [Everitt, 1910; Pearson, 1931]. Chesire *et al* [1933] prepared computing diagrams to aid in finding r. Recently computer programs have been written to solve the computational problem [e.g., see Froemel, 1971]. However, many programs err by truncating the infinite series too quickly.

Hamdan [1970] showed that the tetrachoric correlation is the maximum likelihood estimate of correlation in the 2 × 2 table. Pearson [1901] also developed a formula for the standard error of the tetrachoric correlation which is tabulated by Hayes [1943] for several values of the correlation. Due to its computational complexity, Pearson [1913] proposed an approximation that could be tabulated in parts. Hamdan [1970] gives an alternate formula based on his maximum likelihood approach.

It is well-known that the product-moment correlation is a biased estimate of the underlying correlation. Similarly, we show numerically that the tetrachoric correlation is also biased, and indicate when bias is severe.

We also compute the population standard deviation and compare it with the formulas of Pearson and Hamdan. We then suggest guidelines for when these formulas can be used.

The Mean and Variance of the Tetrachoric Correlation

Let the underlying probabilities of the cells in a 2 × 2 contingency table be $(p_{11}, p_{12}, p_{21}, p_{22})$.

$$\begin{array}{cc|c} p_{11} & p_{12} & p_{1+} \\ p_{21} & p_{22} & p_{2+} \\ \cline{1-2} p_{+1} & p_{+2} & 1 \end{array}$$

As with Fisher's exact test the marginal frequencies p_{1+}, p_{2+}, p_{+1} and p_{+2} are regarded as fixed a priori. The probability of observing the fourfold table (a, b, c, d) is

$$(3) \qquad f(a, b, c, d) = k^{-1} \frac{p_{11}{}^{a} p_{12}{}^{b} p_{21}{}^{c} p_{22}{}^{d}}{a!\, b!\, c!\, d!}$$

where

$$k = \sum \frac{p_{11}{}^{a} p_{12}{}^{b} p_{21}{}^{c} p_{22}{}^{d}}{a!\, b!\, c!\, d!}$$

subject to the constraints $a + b = Np_{1+}$ and $a + c = Np_{+1}$. Under the hypothesis that $\rho = 0$, (3) reduces to the hypergeometric density used in Fisher's exact test.

Let ρ be the tetrachoric correlation of the fourfold table $(Np_{11}, Np_{12}, Np_{21}, Np_{22})$. Then the expected value of r is

$$(4) \qquad E(r) = \Sigma r f(a, b, c, d)$$

and the expected mean square of r is

$$(5) \qquad \sigma^2(r) = \Sigma (r - \rho)^2 f(a, b, c, d)$$

where the summations are subject to the constraints of (3). Note that $\sigma^2(r)$ is computed about ρ and not about $E(r)$.

Approximations of the Tetrachoric Correlation and Its Standard Error

The tetrachoric correlation is difficult to compute unless the marginal frequencies are equal (when the cosine function (2) can be used). Therefore many statistics have been proposed to approximate r [see Castellan, 1966]. A common approximation that uses the cosine function [Guilford, 1965] is

$$(6) \qquad r_{\cos \cdot pi} = \cos \left\{ \frac{\pi}{1 + \left(\dfrac{ad}{bc} \right)^{1/2}} \right\}$$

which is exact when $r = 0$ or the marginal frequencies are equal (when it simplifies to (2)). We include it in our study to reemphasize the error which can occur when an approximation is used.

Pearson's [1901, 1913] formula for the standard deviation based on asymptotic theory is

$$(7) \qquad s_P = \frac{1}{N^{3/2} \phi(z_1, z_2, r)} \left\{ \frac{(a + d)(b + c)}{4} + (a + c)(b + d)\Phi_2{}^2 + (a + b)(c + d)\Phi_1{}^2 + 2(ad - bc)\Phi_1\Phi_2 - (ab - cd)\Phi_2 - (ac - bd)\Phi_1 \right\}^{1/2}$$

where

$$\Phi_1 = \Phi\left(\frac{z_1 - rz_2}{(1 - r^2)^{1/2}}\right) - 0.5$$

$$\Phi_2 = \Phi\left(\frac{z_2 - rz_1}{(1 - r^2)^{1/2}}\right) - 0.5$$

and

$$\phi(z_1, z_2, r) = \frac{1}{2\pi(1 - r^2)^{1/2}} \exp\left\{-\frac{(z_1^2 - 2rz_1z_2 + z_2^2)}{2(1 - r^2)}\right\}.$$

(Note that Pearson defines probable error, which is $0.67449s_p$.) Due to the complex hand calculations for this formula he gives an approximation to it [Pearson, 1913]:

$$(8) \qquad s_A = \frac{\left\{\dfrac{(a + b)(a + c)(b + d)(c + d)}{N^5}(1 - r^2)\left[1 - \left(\dfrac{\sin^{-1} r}{\pi/2}\right)^2\right]\right\}^{1/2}}{\phi(z_1, z_2, 0)}$$

[see also Guilford and Lyons, 1942].

Based upon maximum likelihood, Hamdan [1970] derives an asymptotic standard error as

$$(9) \qquad s_H = \frac{1}{N\phi(z_1, z_2, r)}\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{-1/2}.$$

Note that when $r = 0$, s_p, s_A and s_H all reduce to

$$(10) \qquad s_0 = \frac{\left\{\dfrac{(a + b)(a + c)(b + d)(c + d)}{N^5}\right\}^{1/2}}{\phi(z_1, z_2, 0)},$$

which is appropriate to test the null hypothesis that the correlation is zero.

Results

The parameters ρ, $E(r)$ and $\sigma(r)$ and the statistics $r_{\cos\text{-}pi}$, s_p, s_H, s_A, s_0 are computed for various tables (Np_{11}, Np_{12}, Np_{21}, Np_{22}). The evaluation of r for each "observed" table (a, b, c, d) in (4) and (5) is by a subroutine that uses a varying number of terms in the tetrachoric series such that the bivariate normal probability of a quadrant is evaluated with an error less than 2×10^{-6}.

When $\rho = 0$, the results are given in Table 1 and when $\rho \neq 0$ they are in Table 2.

Bias of the Tetrachoric Correlation

When $\rho = 0$ and p_{1+} and/or p_{+1} is equal to $\frac{1}{2}$, $E(r) = 0$. Otherwise $E(r)$ is a biased estimate of ρ. The bias is most severe when one of the cells has an

TABLE 1

Tables for which the Tetrachoric Coefficient is Zero

p_{11}	p_{12}	p_{21}	p_{22}	N	min. cell frequency	E(r)	$\sigma(r)$	$s_O=s_P=s_H=s_A$
.25	.25	.25	.25	40	10	0	.2439	.2484
				100	25	0	.1559	.1571
				400	100	0	.0784	.0785
.05	.05	.45	.45	40	2	0	.4144	.3387
				100	5	0	.2206	.2142
				400	20	0	.1077	.1071
.01	.01	.49	.49	100	1	0	.7035	.3624
				400	4	0	.2045	.1812
.01	.09	.09	.81	100	1	-.2481	.5990	.2922
				400	4	-.0257	.1825	.1461
				1000	10	-.0074	.0951	.0924
.002	.018	.098	.882	500	1	-.2851	.6002	.2210
				2500	5	-.0175	.1250	.0989
.0004	.0196	.0196	.9604	2500	1	-.3178	.6067	.1672

TABLE 2

Tables for which the Tetrachoric Coefficient is Not Zero

p_{11}	p_{12}	p_{21}	p_{22}	N	min. cell frequency	r	E(r)	$r_{\cos\cdot pi}$	$\sigma(r)$	s_P	s_H	s_A	s_0
.3333	.1667	.1667	.3333	30	5	.5000	.4968	.5000	.2308	.2342	.2342	.2342	.2868
				90	15		.4989		.1346	.1352	.1352	.1352	.1656
				300	50		.4997		.0739	.0740	.0740	.0740	.0907
.45	.05	.05	.45	40	2	.9511	.9501	.9511	.0482	.0460	.0460	.0460	.2484
				100	5		.9507		.0297	.0291	.0291	.0291	.1571
.08	.02	.42	.48	100	2	.4395	.4742	.5394	.2398	.1919	.1908	.1842	.2142
				400	8		.4432		.0970	.0960	.0954	.0921	.1071
.0967	.0033	.4033	.4967	300	1	.7483	.8132	.9003	.1611	.0890	.0877	.0692	.1237
				600	2		.7747		.1026	.0630	.0620	.0489	.0875
.002	.018	.498	.482	500	1	-.4733	-.6097	-.7138	.3261	.1594	.1585	.1355	.1621
				1000	2		-.5267		.2030	.1127	.1121	.0958	.1146
				1500	3		-.4981		.1371	.0920	.0936	.0915	.0783
.075	.025	.025	.875	40	1	.9358	.9292	.9612	.0968	.0851	.0842	.1038	.4620
				80	2		.9328		.0644	.0602	.0595	.0734	.3267
.03	.07	.07	.83	100	3	.4599	.4220	.5691	.2779	.2175	.2163	.2472	.2922
				400	12		.4535		.1104	.1087	.1082	.1236	.1461
.015	.085	.485	.415	400	6	-.5197	-.5249	-.6378	.0939	.0915	.0908	.0858	.1071

expected frequency near zero. When no frequency is less than 5, the bias is negligible.

This is easily understood by considering the change in r as any observed cell frequency appraoches zero. For example, the table (1, 9, 9, 81) yields $r = 0$, whereas for (0, 10, 10, 80), $r = -1.0$. When the theoretical probabilities are (0.01, 0.09, 0.09, 0.81) and $N = 100$, the probability that $a = 0$ (and $r = 1.0$) is large. Hence the expected value is strongly influenced by a zero cell. When $N = 400$ in the above, the probability that $a = 0$ is greatly reduced, as is the bias.

Therefore caution is necessary in interpreting r especially when any observed frequency is less than 5.

When One Cell is Zero

When one and only one cell is zero, it is obvious that $r = \pm 1.0$ for that table. However, the computation of r is based upon the premise that the observed frequencies represent the underlying frequencies exactly. It is clear that the observed frequencies are integer quantities and therefore cannot be more accurate than $\frac{1}{2}$ unit. Hence the cell probability can be no more accurate than $1/(2N)$.

We therefore replaced the zero observed frequency by $\frac{1}{2}$ and modified the other cells to maintain the same marginal totals. When the expected value $E(r)$ was recomputed using this modification, the bias was greatly reduced. For example when $N = 100$, the table with probabilities (0.01, 0.09, 0.09, 0.81) had a modified expected value 0.0187 instead of -0.2481. Similar large reductions were obtained whenever there was an expected frequency of 1 in the table.

The tetrachoric correlation obtained by modifying the zero cell to $\frac{1}{2}$ is the minimum r which is consistent with the rounding errors inherent in using the observed frequencies. Therefore we recommend this adjustment (similar to Yates correction) whenever a single cell is zero.

When two cells are zero (either diagonal or off-diagonal), one of the two variables is redundant and we would assign a correlation of ± 1.0.

The Error in $r_{\cos\text{-}pi}$

When $ad = bc$ both r and $r_{\cos\text{-}pi}$ are zero. When all the marginals are equal, r (which is obtained by (2)) is equivalent to $r_{\cos\text{-}pi}$. Otherwise $r_{\cos\text{-}pi}$ may greatly overestimate the correlation. This can be seen in Table 2 by comparing ρ with its estimate $r_{\cos.pi}$. Therefore, we recommend that $r_{\cos\text{-}pi}$ *not* be used.

The Behavior of the Standard Errors

From Table 1 we see that s_0 rapidly converges to $\sigma(r)$ when $r = 0$. When the minimum cell frequency is at least 5, the convergence is adequate. Note that $s_0 = s_p = s_H = s_A$ when $r = 0$.

When $r \neq 0$ (Table 2), the standard errors s_p and s_H converge slowly to $\sigma(r)$. Pearson's approximate formula s_A, whose development was based on computational ease, is the most biased. Since Hamdan's standard error s_H is similar to s_p and easiest to calculate of the three formulas, we recommend its use. The convergence of s_p and s_H to $\sigma(r)$ when $\rho \neq 0$ is slower than when $\rho = 0$. Therefore we recommend their use when the minimum cell frequency is at least 10, unless the marginal totals are equal when the minimum cell frequency need only be 5.

The Computation of the Tetrachoric Correlation

Froemel [1971] provides a fine comparison of three routines that are available to compute the tetrachoric coefficient. Others are continually being written. Most routines are based upon fitting a tetrachoric series expansion in r and then finding a root of the expansion. For example, the routine in the IBM *Scientific Subroutine Package* [1970] uses the first six terms. Martinson and Hamdan [1975] use the first eight terms in finding the polychoric estimate of correlation which is the tetrachoric correlation when the table is 2 × 2.

Using a criterion that the tetrachoric series has converged when two successive terms are less than 10^{-6}, we found the following approximate relationship between the number of terms and r.

$\lvert r \rvert$	0.25	0.5	0.75	0.8	0.85	0.9	0.95
approximate number of terms	10	18	36	48	60	86	150

Even if the criterion is less rigid, the number of terms must increase with $|r|$. Therefore any routine using a short fixed length series will not be accurate when $|r|$ is large.

REFERENCES

Castellan, N. J. Jr. On the estimation of the tetrachoric correlation coefficient. *Psychometrika*, 1966, *31*, 67–73.

Chesire, L., Saffir, M., & Thurstone, L. L. *Computing Diagrams for the Tetrachoric Correlation Coefficient.* Chicago: University of Chicago Bookstore, 1933. Reprinted by Psychometric Affiliates, Munster, Indiana, 1968.

Everitt, P. F. Tables of the tetrachoric functions for fourfold correlation tables. *Biometrika*, 1910, *7*, 437–451.

Froemel, E. C. A comparison of computer routines for the calculation of the tetrachoric correlation coefficient. *Psychometrika*, 1971, *36*, 165–174.

Guilford, J. P. *Fundamental Statistics in Psychology and Education* (4th ed.). New York: McGraw-Hill, 1965.

Guilford, J. P. & Lyons, T. C. On determining the reliability and significance of a tetrachoric coefficient of correlation. *Psychometrika*, 1942, *7*, 243–249.

Hamdan, M. A. The equivalence of tetrachoric and maximum likelihood estimates of ρ in 2 × 2 tables. *Biometrika*, 1970, *57*, 212–215.

Hayes, S. P. Tables of the standard error of tetrachoric correlation coefficient. *Psychometrika*, 1943, *8*, 193–203.

IBM. System/360 Scientific Subroutine Package, Version III, 5th ed. White Plains: IBM Corporation, 1970.

Martinson, E. O. & Hamdan, M. A. Algorithm AS87 calculation of the polychoric estimate of correlation in contingency tables. *Applied Statistics*, 1975, *24*, 272–278.

Pearson, K. Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philisophical Transactions of the Royal Society of London*, 1901, *195 A*, 1–47.

Pearson, K. On the probable error of a coefficient of correlation as found from a fourfold table. *Biometrika*, 1913, *9*, 22–27.

Pearson, K. (Ed.). *Tables for Statisticians and Biometricians, Part II*. London: Cambridge University Press, 1931.

Manuscript received 6/25/76
Final version received 12/10/76