PSYCHOMETRIKA—VOL. 31, NO. 1
MARCH, 1966

ON THE ESTIMATION OF THE TETRACHORIC CORRELATION COEFFICIENT*

N. JOHN CASTELLAN, JR.†

UNIVERSITY OF COLORADO

This paper presents briefly the rationale of the tetrachoric correlation coefficient. Pearson's results are outlined and several estimates of the coefficient are given. These estimates are compared with Pearson's expressions to determine the relative accuracy of the various approximations in determining the tetrachoric correlation coefficient.

If we have two continuous and normally distributed random variables X and Y, each with mean 0 and variance 1, then we may write the joint density function as

$$(1) \qquad f(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left[-\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right].$$

Suppose that the (x, y)-plane is intersected by two perpendicular planes which are also perpendicular to the (x, y)-plane. Let these two planes intersect at the point (h, k) in the (x, y)-plane. Then the density function is divided into four quadrants at the point (h, k). (see Fig. 1.)

Suppose we have a set of N observations. Then the frequency of observations in each quadrant, with origin (h, k), will be d, c, a, b, where $N = a + b + c + d$, in quadrants I, II, III, IV, respectively. If we then rotate the y-axis 180° about the x-axis, we have the familiar bivariate frequency table,

	−	+
−	a	b
+	c	d

The problem of tetrachoric correlation is to find h and k given the fre-

*Preparation of this paper was supported in part by Fellowship 1-F1-MH-24, 324-01, from the National Institute of Mental Health; and in part by the Tri-Ethnic Research Project, Grant 3M-9156 from the National Institute of Mental Health to the Institute of Behavioral Science, University of Colorado. This paper comprises Publication Number 57 of the Institute. The author would like to thank D. E. Bailey for his helpful comments and criticisms.

†Now at Indiana University.

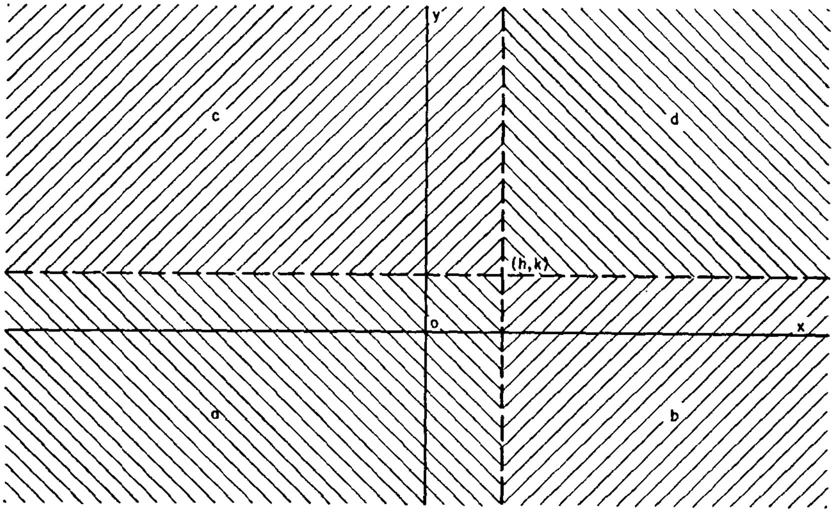

FIGURE 1
Representation of Frequency Distribution

quencies a, b, c, and d. The remainder of this paper outlines several approaches to the determination and approximation of the tetrachoric correlation coefficient.

Pearson's Tetrachoric Correlation

This presentation is based upon Pearson's paper [4]. Proofs are omitted here and the reader is referred to Pearson's paper or Kendall and Stuart [3] for detailed treatment. Let

$$\chi_1 = \sqrt{\frac{\pi}{2}}\frac{(a + c) - (b + d)}{N} = \sqrt{2\pi}\int_0^h \phi(x)\,dx, \tag{2}$$

and

$$\chi_2 = \sqrt{\frac{\pi}{2}}\frac{(a + b) - (c + d)}{N} = \sqrt{2\pi}\int_0^k \phi(y)\,dy, \tag{3}$$

where

$$\phi(w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}w^2\right)$$

is the ordinate of the unit normal curve at w. Now since we know χ_1 and χ_2, we may find an expression for h and k.

$$h = \chi_1 + \frac{1}{3!}\chi_1^3 + \frac{7}{5!}\chi_1^5 + \frac{127}{7!}\chi_1^7 + \cdots, \tag{4}$$

and

$$(5)\qquad k = \chi_2 + \frac{1}{3!}\chi_2^3 + \frac{7}{5!}\chi_2^5 + \frac{127}{7!}\chi_2^7 + \cdots .$$

Now for convenience, let

$$(6)\qquad \epsilon = \frac{ad - bc}{N^2\phi(h)\phi(k)}.$$

Pearson has derived the following expression for ϵ in terms of h, k, and r.

$$(7)\qquad \epsilon = r + \frac{r^2}{2!}hk + \frac{r^3}{3!}(h^2 - 1)(k^2 - 1) + \frac{r^4}{4!}hk(h^2 - 3)(k^2 - 3) + \frac{r^5}{5!}(h^4 - 6h^2 + 3)(k^4 - 6k^2 + 3) + \cdots .$$

Since all of the coefficients of r^n are known in (7) and since we know ϵ from (6), we may determine r. Since $r \leq 1$, and since $n!$ goes to ∞, the terms $r^n/n!$ go to zero and we need only to use the first few terms to determine r.

Pearson has also derived another series expression for ϵ. We can express ϵ as

$$(8)\qquad \epsilon = \exp\left[\tfrac{1}{2}(h^2 + k^2)\right] \int_0^r \frac{1}{\sqrt{1 - r^2}} \exp\left\{-\frac{1}{2}\frac{1}{1 - r^2}(h^2 - 2rkh + k^2)\right\} dr.$$

Now if we let $r = \sin\theta$, we may rewrite (8) as

$$(9)\qquad \epsilon = \exp\left(\tfrac{1}{2}h^2\right) \int_0^\theta \exp\left\{-\tfrac{1}{2}(k\tan\theta - h\sec\theta)^2\right\} d\theta.$$

Further, if we let

$$(10)\qquad \chi = \exp\left[-\tfrac{1}{2}(k\tan\theta - h\sec\theta)^2\right],$$

we may rewrite (9) using Maclaurin's theorem.

$$(11)\qquad \epsilon = \theta + \left[\exp\left(\tfrac{1}{2}h^2\right)\right] \sum_{n=1}^{\infty} \left(\frac{d^n\chi}{d\theta^n}\right)_0 \frac{\theta^{n+1}}{(n + 1)!}.$$

Evaluation of the terms of (11) leads to

$$(12)\qquad \epsilon = \theta + \tfrac{1}{2}hk\theta^2 - (h^2 + k^2 - h^2k^2)\frac{\theta^3}{3!} + hk[h^2k^2 - 3(h^2 + k^2) + 5]\frac{\theta^4}{4!} + \cdots .$$

As $\theta^n/n!$ goes to zero as n gets large, one may use the first few terms of (12) and solve for θ and hence r. As this series converges faster than (7) its use is preferred. A computer program has been written (Castellan and Link [2]) which computes the tetrachoric correlation coefficient using Pearson's series expression (12). This program is fast and accurate, allows for missing data, and constructs the frequency tables. Copies of the program are available from Computer Program Librarian, Institute of Behavioral Science, University of Colorado, Boulder, Colorado, as IBS-007.

Approximations of Tetrachoric Correlation

Because of the difficulty of computing the tetrachoric correlation from (7) or (12), Pearson also gives several other "estimates" of the tetrachoric correlation coefficient. These estimates have the following properties: (i) they are zero when the determinant of the frequency table is zero, i.e., when $ad - bc = 0$; (ii) they are equal to 1 if $b = 0$ or $c = 0$; (iii) they are equal to the tetrachoric correlation coefficient when the response frequencies form median divisions of the distributions, i.e., when $a + b = c + d$ and $a + c = b + d$. These estimates are

$$Q_2 = \frac{ad - bc}{ad + bc}, \tag{13}$$

$$Q_3 = \sin \frac{\pi}{2} \left[\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right] = \cos \pi \left[\frac{1}{1 + \sqrt{\dfrac{ad}{bc}}} \right], \tag{14}$$

$$Q_4 = \sin \frac{\pi}{2} \left\{ \frac{1}{1 + \dfrac{2bc \qquad N}{(ad - bc)\,(b + c)}} \right\} \quad \text{for} \quad ad \geq bc, \tag{15}$$

$$Q_5 = \sin \frac{\pi}{2} \left[\frac{1}{\sqrt{1 + k^2}} \right], \tag{16}$$

where

$$k^2 = \frac{4abcdN^2}{(ad - bc)^2(a + d)(b + c)}.$$

Note that Q_2 is Yule's Q or the Goodman-Kruskal Gamma and does not satisfy condition (iii) above.

Walker and Lev [5] give an estimate of tetrachoric correlation which we shall call Q_6.

$$Q_6 = \sin \frac{\pi}{2} \left[\frac{(a + d) - (b + c)}{N} \right]. \tag{17a}$$

After a few simple manipulations we find that

$$Q_6 = \cos \pi\left(\frac{b + c}{N}\right). \tag{17b}$$

It should be noted that this estimate does yield the tetrachoric correlation exactly when there are median splits; it is *not* 1 if b or c is 0, nor does it equal 0 when $ad - bc = 0$. It does equal 1 when b and c are both equal to 0.

Camp [1] gives a procedure for estimating the tetrachoric correlation coefficient. Let

$$p_1 = \frac{a}{a + c} = \int_{-\infty}^{z_1} \phi(w)\, dw, \tag{18}$$

and

$$p_2 = \frac{d}{b + d} = \int_{-\infty}^{z_2} \phi(w)\, dw, \tag{19}$$

where z_1 and z_2 are abscissa values on the unit-normal curve. Further, let

$$m = \frac{(a + c)(b + d)}{N^2} \frac{z_1 + z_2}{\phi(h)} = p_1 p_2 \frac{z_1 + z_2}{\phi(h)}. \tag{20}$$

Then

$$Q_7 = \frac{m}{\sqrt{1 + \theta m^2}}, \tag{21}$$

where θ is a function of $(a + c)/N$. The relation is tabled below.

$(a + c)/N$	.5	.55	.6	.65	.7	.75	.80	.85	.9
θ	.637	.634	.631	.626	.62	.61	.60	.582	.56

The limitations which Camp gives are roughly that $(a + c)/N < .9$, and that $|r| < .8$. This approximation may be computed easily by using (20) and a table of ordinates and abscissas for the unit-normal curve.

The Standard Error of the Tetrachoric Correlation Coefficient

As the purpose of this paper is to present several different methods for determining the tetrachoric correlation coefficient detailed discussion of its distribution will be omitted here. However, we shall state an approximation of the standard error when testing the null hypothesis that $\rho = 0$. This estimate is appropriate regardless of the determination of r.

$$s_r \doteq \frac{\sqrt{(a + b)(c + d)(a + c)(b + d)}}{\phi(h)\phi(k)N^2\sqrt{N}}. \tag{22}$$

In all cases this sampling error of r is greater than that of a product-moment correlation regardless of splits. The more extreme the splits, the greater the discrepancy in sampling error, and consequently r will compare less favorably with the product-moment correlation.

Comparisons of the Approximations with the Tetrachoric Correlation Coefficient

Pearson selected fifteen tables which covered a "fairly wide range of values." He then calculated Q_2, Q_3, Q_4, and Q_5 and determined the discrepancy between the estimates and the tetrachoric correlation. These differences are summarized in Table 1. In an effort to check Pearson's findings and to evaluate the efficacy Q_6 and Q_7 , an additional sixteen tables were constructed, the estimates were calculated and the mean differences between the Q_i and the tetrachoric coefficients were determined. These data are presented in Table 2 and are also summarized in Table 1.

TABLE 1

Mean Differences (in Per Cent)
Between Tetrachoric Correlation and Various Estimates

	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
Pearson study N = 15	31.2	4.09	3.08	2.87	——	——
This study N = 16	34.15	10.77	8.30	8.17	25.46*	0.81

*This difference was based upon 12 cases. If the other 4 cases are included (cases 11, 12, 14, and 15), the mean difference is 206.91 per cent.

TABLE 2

Data Values and Estimates
Of the Tetrachoric Correlation Coefficient

	a	b	c	d	r	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
1	50	30	10	10	.181	.250	.198	.184	.196	.309	.180
2	30	20	20	30	.309	.385	.309	.309	.309	.309	.307
3	40	10	10	40	.809	.882	.809	.809	.809	.809	.807
4	35	10	25	30	.506	.615	.515	.534	.521	.454	.504
5	400	150	225	225	.364	.455	.369	.365	.368	.383	.367
6	140	20	60	80	.696	.806	.714	.736	.725	.669	.697
7	66	10	14	10	.506	.650	.548	.466	.513	.729	.494
8	60	10	14	10	.484	.622	.521	.448	.491	.695	.473
9	60	5	25	10	.493	.655	.553	.540	.548	.588	.493
10	70	5	20	5	.375	.556	.459	.365	.423	.707	.377
11	45	35	15	5	-.285	-.400	-.322	-.259	-.329	.000	-.288
12	80	10	25	5	.148	.231	.183	.126	.168	.609	.149
13	100	80	10	20	.298	.429	.346	.373	.351	.223	.299
14	100	80	10	10	.069	.111	.087	.084	.087	.156	.069
15	320	100	60	20	.021	.032	.025	.017	.024	.536	.021
16	60	5	12	10	.681	.818	.728	.671	.698	.817	.670

The order of the differences between the estimates was the same in both studies although the magnitude differed between the two studies. The reason for the latter is that Pearson placed some restrictions on his tables (roughly that h and k should be less than 1).

There is little empirical justification for using Q_2 or Q_6 as an estimate. With a mean difference of .81 per cent, Q_7 is clearly the best estimate of the tetrachoric correlation coefficient. In *every one* of the the 16 cases the difference between Q_7 and r was less than ± .01. Consequently, for purposes of manual computation, Q_7 is the best estimate of the tetrachoric correlation coefficient.

REFERENCES

[1] Camp, B. H. *The mathematical part of elementary statistics.* New York: D. C. Heath, 1931. Pp. 300–314.

[2] Castellan, N. J., Jr. and Link, S. W. Tetrachoric correlation program for the IBM 709/7090 (CPA 174). *Behav. Sci.*, 1964, **9**, 292.

[3] Kendall, M. G. and Stuart, A. *Advanced theory of statistics* (Vol. I) (2nd ed.). London: Griffin, 1958.

[4] Pearson, K. I. Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Phil. Trans. roy. Soc. London.* 1901, **195A**, 1–47.

[5] Walker, H. M. and Lev, J., *Statistical·inference.* New York: Holt, Rinehart & Winston, 1953. Pp. 271–275.

Manuscript received January 25, 1965
Revised manuscript received March 25, 1965