PSYCHOMETRIKA—VOL. 36, NO. 2
JUNE, 1971

A COMPARISON OF COMPUTER ROUTINES FOR THE CALCULATION OF THE TETRACHORIC CORRELATION COEFFICIENT*

ERNEST C. FROEMEL

THE UNIVERSITY OF CHICAGO

In calculations of the discriminating-power parameter of the normal ogive model, Bock and Lieberman compared estimates derived from their maximum-likelihood solution with those derived from the heuristic solution. The two sets of estimates were in excellent agreement provided the heuristic solution used accurate tetrachoric correlation coefficients. Three computer methods for the calculation of the tetrachoric correlation were examined for accuracy and speed. The routine by Saunders was identified as an acceptably accurate method for calculating the tetrachoric correlation coefficient.

For a number of years, the only practical method for estimating the discriminating-power parameter of the normal ogive model was a rank-one common factor analysis of the matrix of item tetrachoric correlation coefficients. Recently, however, Bock and Lieberman [1970] obtained a maximum-likelihood solution for estimating this parameter, using the same assumptions but without resorting to tetrachoric correlations. When they compared the two methods, they found that the estimates from them could be in excellent agreement *provided the tetrachoric correlation coefficients used in the former were accurate.* Obtaining such accuracy involved making interpolations in the U. S. Bureau of Standards [1956] tables of bivariate normal distribution. Estimates from the two methods did not agree when the tetrachoric correlations were derived by standard computer routines.

The work reported in the present paper was undertaken with the hope of finding a fast *and* accurate method of calculating tetrachoric correlations. An evaluation of several approaches revealed that a routine developed by Saunders is an acceptably accurate method for computing these correlations.

The Tetrachoric Correlation Coefficient

The tetrachoric correlation is most frequently used when two variables are dichotomized and their relationship presented in a fourfold table. Figure 1

* This research was supported in part by NSF Grant E 1930 to The University of Chicago. The author wishes to thank Dr. David R. Saunders and Dr. Ledyard Tucker for the use of their original materials and Dr. R. Darrell Bock for his many helpful suggestions and his ready counsel throughout the course of this investigation.

gives a schematic representation of such a table. The underlying distribution is assumed to be bivariate normal, an assumption that Carroll [1961] finds more tenable than that underlying ϕ/ϕ_{max}. Each cell represents the proportion of observations which have the four possible point values of the two dichotomous variables. For example, cell a in Figure 1 may represent the proportion of respondents who have correctly answered two test items. From this tabular arrangement, the following quantities are computed:

a. *The marginal proportions:* $p_1 = a + c$ *and* $p_2 = a + b$. The complementary proportions are $q_1 = 1.0 - p_1$ and $q_2 = 1.0 - p_2$.
b. *The normal deviates corresponding to the marginal proportions:* $h = \Phi^{-1}(p_1)$ and $k = \Phi^{-1}(p_2)$. These deviates may be found in tables or may be computed by numerical approximations. In the research reported here, we used a computing routine by Kuki [1966] based on Chebyshev polynomials refined by Newton–Raphson iterations for different sections of the normal curve. The method is quite accurate, even in the tails.

The proportion d may be expressed in terms of h, k, and the unknown correlation ρ as follows:

$$(1)\ d = \Phi(h, k; \rho) = \frac{1}{2\pi\sqrt{1 - p^2}} \int_{-\infty}^{h} \int_{-\infty}^{k} \exp\left[-\frac{(x^2 - 2xy\rho + y^2)}{2\sqrt{1 - \rho^2}}\right] dx\, dy.$$

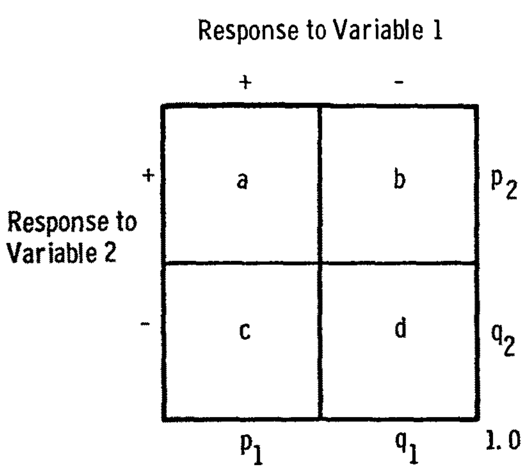

FIGURE 1
Schematic representation of a fourfold table used to generate tetrachoric correlation coefficients.

Three Computer Methods

The next step is to solve for ρ. This paper examines three computer methods which purport to do so. For each routine, the original source program was converted to a subroutine and coded in FORTRAN II. The testing was done on The University of Chicago's IBM 7040/7094 installation. Timing was determined through the 7094 library subroutines, which use the 7040 core clock.

Method 1

Pearson [1901] and Castellan [1966] expand the following series:

$$\frac{ad - bc}{\phi(h)\phi(k)} = \theta + \left[\exp\left(\frac{h^2}{2}\right)\right] \sum_{n=1}^{\infty} \left(\frac{d^n\chi}{d\theta^n}\right)_0 \frac{\theta^{n+1}}{(n+1)!}, \tag{2}$$

where

$$\chi = \exp\left(-\frac{(k \tan \theta - h \sec \theta)^2}{2}\right)$$

and

$$\rho = \sin \theta$$

which yields the following terms used for the evaluation of ρ:

$$\frac{ad - bc}{\phi(h)\phi(k)} = \theta + \frac{hk\theta^2}{2} - \frac{(h^2 + k^2 - h^2k^2)\theta^3}{3!} + \frac{hk(h^2k^2 - 3(h^2 + k^2) + 5)\theta^4}{4!}. \tag{3}$$

Castellan then solves for θ using a Newton-Raphson iterative technique identical to the subroutine POLRT as documented in IBM's *System/360 Scientific Subroutine Package (360A-CM-03X) Version III Programmer's Manual.*

Both Castellan and Pearson prefer this equation to another expansion because of its speed of convergence and of the fact that the terms of the infinite series it uses are often sufficient for accurate computation. However, Castellan and Link [1964] place no restriction on the marginals or cells, whereas Pearson [1901] considers only the condition $h \leq 1$, $k \leq 1$, *i.e.*, splits in the marginals $(p - q)$ no greater than .84 to .16.

The routine tested for this project is an adaptation of Castellan's subroutine TET, which appears under the file number IRS-207 at the Institute of Behavioral Sciences Program Library, University of Colorado, Boulder, Colorado.

Method 2

Method 2 is an adaptation of an unpublished program, PCD, by David R. Saunders, who attributes the idea to Ledyard Tucker (personal communication). In this method, the bivariate normal distribution function is rewritten as follows:

$$(4)\ \Phi(h, k; \rho) = \frac{1}{2\pi} \int_0^\rho \frac{1}{\sqrt{1 - x^2}} \exp\left[\frac{-(h^2 - 2hkx + k^2)}{2(1 - x^2)}\right] dx + \Phi(h)\Phi(k).$$

This form has been discussed by Owen [1956]. It is solved for ρ by substituting the following observed proportions: $\Phi(h, k; \rho) = d$, $\Phi(h) = p_1$, and $\Phi(k) = p_2$. The value of ρ is then found. This value accounts for the area $A_0 = d - p_1p_2$ by successively reducing A_0 according to the iteration:

$$A_{i+1} = A_i - \Delta f(r_{i+1}),$$

where Δ is a small increment, (here $0.0078125 = 2^{-7}$),

$$f(r) = \frac{1}{2\pi\sqrt{1 - r^2}} \exp\left[\frac{-(h^2 - 2hkr + k^2)}{2(1 - r^2)}\right],$$

and $r_{i+1} = r_i + \Delta$ is an approximation to ρ, with $r_0 = \Delta/2$. When A_{i+1} becomes zero or negative, the estimate of ρ is corrected by the interpolation $r_{i+1} = r_{i+1} - (\Delta/2) + A_i/f(r_i)$. The computer routine as coded by Saunders is adjusted to use q if $p > 0.5$ and to use A_0 = minimum cell proportion $- p_1p_2$. This adjustment reduces the accumulation of rounding error.

Different values of Δ were tested in this routine. The value 0.0078125 was finally selected since it was small enough to provide minimal error yet large enough not to prolong iteration.

A method which evaluates the above expression for $\Phi(h, k; \rho)$ using 20-point Gauss–Legendre quadrature and then finds ρ using a Newton–Raphson process was also tested. This method gave exactly the same results as the method described above. However, this more sophisticated technique was slower and, for successful convergence, required that the initial approximation to ρ be very close to the true correlation.

Method 3

This tetrachoric routine may be found in the listing of P-STAT from the Computer Center, Princeton University. It was coded by Roald Buhler using the coding logic of Marilyn Charap [1957] for an algorithm by Ledyard Tucker [1960]. Tucker derives a rational approximation for the tetrachoric correlation as follows:

$$(5)\qquad r_i = \frac{hx \pm \sqrt{h^2x^2 + (x^2 + (x - k)xY_i^2)(Y_i^2 - h^2)}}{x^2 + (x - k)xY_i^2},$$

where

$$x = \frac{\exp\left(-\frac{k^2}{2}\right)}{p_2\sqrt{2\pi}}.$$

The initial value for Y_0 is $\Phi^{-1}(a/p_2)$. Subsequent values of Y_i are based on a system of empirical corrections which are continued until

$$\left|\frac{1}{1 + .8r_i^{12}} - \frac{1}{1 + .8r_{i+1}^{12}}\right| \le 0.0001.$$

Test Data

Data used in testing these routines come from two sources. The first is publications dealing with approximations to ρ containing the true values of ρ [Pearson, 1901; Castellan, 1966]. The second source, which serves more systematic testing needs, is the U. S. Bureau of Standards *Tables of the Bivariate Normal Distribution Function and Related Functions* [1956]. This publication presents the value to six places of:

$$(6) \qquad L(h, k; \rho) = \int_h^\infty \int_k^\infty \frac{1}{2\pi\sqrt{1 - p^2}} \exp\left[-\frac{(x^2 - 2xy\rho + y^2)}{2(1 - p^2)}\right] dx\, dy$$

for certain values of h, k, and ρ. The formula differs from $\Phi(h, k; \rho)$ given previously only in the direction of integration. Practically speaking, this means that the Tables present the value for our a cell rather than for the d cell as shown in Figure 1. With this cell and the marginals, the remaining three cells of a test table can be constructed. These marginals $p_1 = \Phi(h)$ and $p_2 = \Phi(k)$ were taken from the ten-place table of Abramowitz and Stegun [1967].

The data as developed for this study consisted of several sets of tables which hold h and k constant while varying ρ from zero to one. Table 1 shows the values of $L(h, k; \rho)$ for the h, k, and ρ used to develop the data sets used. In some of the data sets, where $.9 \le \rho \le .99$, $L(h, k; \rho)$ cannot be distinguished from $L(h, k; 1.0)$ because one cell of the fourfold table has become zero. In such cases, the largest value of ρ which could produce a distinguishable table was substituted in the tests. This accounts for the holes in Table 1 and for the odd values, such as $\rho = .70$, which was used only for data-set 4. The absolute value of the difference between the tabled and the calculated correlation coefficients is used as a measure of the inaccuracy of each routine.

Results

A discrepancy between results obtained from the Castellan data and those obtained from the Pearson data, both supposedly "representative," suggested the development of two more data sets. As a rough check, data were prepared which represented the tabled values with the closest approximation to the h, k, and ρ by Castellan. This set raised the suspicion that the ρ Castellan presented as the true value was really the ρ calculated by his most accurate method. To verify this, the values of ρ for Castellan's tables were obtained via interpolation from the *Tables of the Bivariate Normal*

TABLE 1

Values of $L(h, k; \rho)$ Used in the Development of Systematic Test Data*

Data Set	1	2	3	4	5	6	7	8	9	10
ρ	h = 0 k = 0	0 1	0 2	0 3	1 1	1 2	1 3	2 2	2 3	3 3
0.00	250000	079328	011375	000675	025171	003609	000214	000518	000031	000002
0.05	257961	084154	012451	000763	028172	004295	000272	000678	000045	000003
0.10	265942	088981	013518	000849	031320	005046	000337	000872	000063	000005
0.20	282047	098633	015596	001010	038069	006742	000490	001370	000115	000011
0.40	315495	117883	019276	001246	053563	010871	000854	002921	000308	000046
0.45	324288	122658	020041	001282	057922	012045	000947	003452	000379	000062
0.50	333333	127398	020724	001309	062514	013266	001037	004053	000460	000082
0.55	342686	132086	021314	001328	067369	014529	001119	004732	000552	000108
0.60	352416	136692	021804	001340	072526	015823	001192	005550	000654	000140
0.70	------	------	------	001349	------	------	------	------	------	------
0.80	397584	153091	022718	------	097637	020860	001344	009825	001131	000372
0.85	------	------	------	------	------	------	001349	------	------	------
0.90	428217	157949	022749	------	115490	022502	------	013361	001319	000610
0.95	449459	158631	------	------	128130	022742	------	016024	001349	000809
0.96	454832	158650	------	------	131352	022748	------	016719	------	000863
0.97	460917	------	------	------	135010	------	------	017514	------	000925
1.00	500000	158655	022750	001350	158655	022275	001350	022750	001350	001350

*Decimal points omitted.

Distribution Function and Related Functions. If the reader should wish to correct Castellan's Table 2 [1966], he may substitute the following interpolated values of r from top to bottom: 0.180, 0.309, 0.809, 0.507, 0.364, 0.699, 0.507, 0.484, 0.496, 0.384, −0.287, 0.149, 0.300, 0.070, 0.020, 0.684. Further discussion of data in this study will be limited to the 167 fourfold tables which include Castellan's tables with these interpolated values for ρ, Pearson's data, and the ten data sets generated for the study. Table 2 summarizes the results for these data by presenting the average error and average time of computation in seconds, averaged over the fourfold tables in each data set. As a further aid to interpretation, Table 3 lists the calculated value of the tetrachoric correlation for selected values of ρ across varying h and k, using the three methods. The value of $\rho = 1.0$ is deleted from Saunders' and Buhler's methods since they assign a value of 1.0 to the correlation if they discover a zero cell, while Castellan's method continues to calculate even in this case.

Discussion

Buhler's routine is the speediest, averaging .0069 seconds per calculation with an average error of .0113. The greatest average errors occur in data-sets 4, 7, 9, and 10, each of which has one or both deviates equal to 3.0. In addition, Table 2 indicates that if h is held constant, the average error increases as k increases. Furthermore, the error generally increases as ρ increases for any given h or k. The information in Tables 2 and 3 suggests that Buhler's routine is a fast one, accurate to two places for moderate ρ when h and k do not exceed 2.0.

The second most rapid routine is Castellan's, averaging .0289 seconds

TABLE 2

Average Error and Average Time for Computation in Seconds Averaged over the Fourfold Tables in Each Data Set*

Data Set	h k	Number of Tables	Castellan		Saunders		Buhler	
			Error	Time	Error	Time	Error	Time
Castellan		16	0017	0333	0007	0365	0020	0073
Pearson		15	0109	0378	0000	0433	0022	0067
1	0 0	15	0000	0089	0000	0433	0058	0067
2	0 1	14	0206	0226	0000	0440	0069	0071
3	0 2	12	2236	0236	0023	0361	0096	0056
4	0 3	11	2673	0227	0007	0318	0396	0061
5	1 1	15	0031	0333	0000	0456	0032	0056
6	1 2	14	1043	0310	0001	0405	0092	0095
7	1 3	12	2568	0361	0005	0375	0266	0042
8	2 2	15	1433	0322	0002	0444	0039	0078
9	2 3	13	1695	0321	0004	0410	0187	0077
10	3 3	15	4011	0322	0014	0467	0194	0078

*Decimal points omitted.

TABLE 3
Calculated Values of ρ for Selected True Values of across Varying h and k Using the 3 Methods*

Method	ρ	h = 0, k = 0	h = 0, k = 1	h = 0, k = 2	h = 0, k = 3	h = 1, k = 1	h = 1, k = 2	h = 1, k = 3	h = 2, k = 2	h = 2, k = 3	h = 3, k = 3
Castellan	0.00	0000	0000	0000	0000	-0000	-0000	-0000	0001	0000	0000
	0.40	4000	3975	9654	9627	3976	2649	9843	2120	5128	2148
	0.60	6000	5919	9515	9623	5982	9656	9849	6517	5158	2157
	0.80	8000	9434	9454	----	8018	9785	9852	6849	5200	2180
	0.90	9000	9636	9452	----	9054	9825	----	7140	5216	2204
	0.95	9500	9664	----	----	9572	9830	----	7369	5219	2244
	0.96	9600	9664	----	----	9673	9830	----	7470	----	2229
	0.97	9700	----	----	----	9773	----	----	7657	----	2235
	1.00	10000	9664	9452	9622	9970	9830	9852	8711	5219	2278
Saunders	0.00	-0000	0000	0000	0000	-0000	-0000	-0002	0001	0012	0087
	0.40	4000	4000	4000	4004	4000	4000	3999	4000	3999	4011
	0.60	6000	6000	6000	6002	6000	6000	6000	6030	6000	6005
	0.80	8000	8000	8006	----	8000	8000	8016	8000	7998	7999
	0.90	9000	9000	8732	----	9000	9001	----	9000	9000	8998
	0.95	9500	9504	----	----	9500	9509	----	9499	9511	9499
	0.96	9600	9602	----	----	9600	9597	----	9600	----	9600
	0.97	9700	----	----	----	9700	----	----	9700	----	9699
Buhler	0.00	-0000	0000	-0000	-0000	-0000	-9000	-0002	0001	0011	0068
	0.40	3982	3997	3980	3836	3979	3981	3867	3982	3916	4141
	0.60	5942	6007	5978	5680	5947	5967	5709	6001	5815	6217
	0.80	7921	8010	7947	----	7990	7920	7444	7851	7545	8140
	0.90	8863	8972	10000	----	8969	8828	----	9041	8288	8787
	0.95	9338	10000	----	----	9419	10000	----	9436	10000	9084
	0.96	9444	10000	----	----	9514	10000	----	9514	----	9099
	0.97	9560	----	----	----	9615	----	----	9594	----	9113

*Decimal points omitted.

for the 167 fourfold tables. However, the average error is .1261, although Table 2 shows that an acceptably small error occurs when h or k is no greater than 1.0. (See data-sets 1, 2, and 5.) When $h = k = 0$, the calculations are accurate to four places for the entire range of ρ. However, when h, k, or both become 1.0, the maximum accuracy is to two places. Hence, Castellan's routine cannot be considered adequate for a general ρ if h or k exceeds 1.0.

For Saunders' routine, the average error is .0005 calculated at an average speed of .0412 seconds, with the greatest errors occurring in data-sets 3 and 10. For data-set 3, the inclusion of a table for $\Phi(0.0, 2.0; 0.9)$ results in a maximum error of .0268. In that case, the maximum cell proportion is 10^{-7}, which could occur in real data only if there were one million cases. Across all other tables, the maximum error is .005 and occurs when $h = k = 3.0$. In addition, Saunders' routine does not appear to be adversely affected by high values of ρ and maintains three-place accuracy for most tables where h or $k \leq 3.0$, and four-place accuracy for most tables where h or k is 2.0 or less. Therefore, it is safe to say that Saunders' routine is accurate for the entire range of values for ρ when h or $k \leq 3.0$.

Summary

This paper presented and evaluated three computer methods for the calculation of the tetrachoric correlation coefficient, ρ. The most accurate for the widest range of data was that of Saunders, which evaluated an integral expression for $\Phi(h, k; \rho)$. Average 7094 time for this routine was .0412 seconds per coefficient. Buhler's method, an empirical approximation, was the quickest at .0069 seconds per coefficient and gave adequate accuracy for moderate ρ when h or k did not exceed 2. Finally, the series expansion of Castellan was found to be accurate only for a restricted range of h or k not exceeding 1. Average time for this routine was .0289 second per coefficient.

When tetrachoric correlations are to be computed for purposes of factor analysis, especially when the cutting points are sometimes extreme or the correlations high, Saunders' routine is to be preferred. Its greater accuracy in this type of application will compensate for the somewhat greater computing time required. A FORTRAN IV subroutine based on Saunders' method may be obtained from the Education Statistics Laboratory, Department of Education, The University of Chicago.

REFERENCES

[1] Abramowitz, M., & Stegun, I. (Eds.), *Handbook of mathematical functions*. Washington, D. C.: U. S. Government Printing Office, 1967. pp. 966–970.

[2] Bock, R. D., & Lieberman, M., Fitting a response model for n dichotomous items. *Psychometrika*, 1970. (In press)

[3] Buhler, R., Tetrachoric program in *P-stat, version 45 source listing*. (Generated from P-stat program tape at The University of Chicago Computation Center, May 17, 1968). pp. 143–148.

[4] Carroll, J. B., The nature of the data or how to choose a correlation coefficient. *Psychometrika*, 1961, 26, 47–372.
[5] Castellan, N., On the estimation of the tetrachoric correlation coefficient. *Psychometrika*, 1966, **31**, 67–73.
[6] Castellan, N., & Link, S., Tetrachoric correlation program for the IBM 709/7090. *Behavioral Science*, 1964, **9**, 292.
[7] Charap, M., Tetrachoric correlation. (An Educational Testing Service program write-up, dated December 1957).
[8] Kuki, H., *Mathematical functions*. (The University of Chicago Computation Center report, 1966). p. 103.
[9] Owen, D., Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 1956, **27**, 1075–1090.
[10] Pearson, K., Mathematical contributions to the theory of evolution. VII: On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society, London*, 1901, **195A**, 1–47.
[11] Tucker, L., *Notes on the approximation of tetrachoric correlation*. (Unpublished notes, dated April 4, 1960).
[12] U. S. Bureau of Standards, *Tables of the bivariate normal distribution function and related functions*. Washington, D. C.: U. S. Government Printing Office, 1956. (Applied Math Series No. 50).

Manuscript received 7/14/70
Revised manuscript received 8/17/70