PSYCHOMETRIKA—VOL. 38, NO. 2
JUNE, 1973

ON THE NUMERICAL APPROXIMATION OF THE BIVARIATE NORMAL (TETRACHORIC) CORRELATION COEFFICIENT

DAVID B. KIRK

EDUCATIONAL TESTING SERVICE

In this paper a rapid and reliable method is found for estimating the value of the Bivariate Normal Correlation Coefficient, ρ, given values of the joint probability and the normal deviates, h and k, or the related areas. This technique finds useful application in the computational approximation of the tetrachoric correlation coefficient, r, when the underlying distributions may be assumed to be normal.

The calculation of the normal bivariate r, or tetrachoric r for the dichotomized case, involves performing an inverse interpolation of the bivariate normal distribution function:

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

since we are effectively given values of L, the standard deviates h and k, and are required to find r.

A major use of r is found in the field of psychology where the data may be measured in or reduced to a two-variable dichotomy. For example, in a testing situation each item may be scored as correct or incorrect, students may be passed or failed, etc. In order to estimate the correlation between these dichotomies, the assumption is made that the underlying traits are continuous and normally distributed or that they were measured in such a way that a normal distribution could be used as a legitimate model. The data may appear in a form similar to the 2 × 2 table shown in Fig. 1.

In order to note the correspondence of the 2 × 2 table with the integral, consider the cell (0, 0), in Figure 1, with a frequency of c or a joint proportion of c/n which corresponds to the value of $L(h, k, r)$. The h and k values are the deviates determined by the areas established by the marginal proportions q_1 and q_2, both of which are less than or equal to .5, of the variables as illustrated in Fig. 2.

Theoretical approaches to the calculation of r have generally relied on an infinite series approach. McNemar [7] gives the first 4 terms of such an infinite series expansion, for which the general form is given by

		Variable 2			
		0 (Wrong)	1 (Right)	Totals	Proportion
Variable 1	1 (Right)	a	b	$a + b$	p_1
	0 (Wrong)	c	d	$c + d$	q_1
	Totals	$a + c$	$b + d$	n	
	Proportion	q_2	p_2		1

FIGURE 1

$$\frac{\frac{c}{n} - q_1 q_2}{z_h z_k} \cong \sum_{j=1}^{\infty} \frac{r^j}{j!} H_{j-1}(h) \cdot H_{j-1}(k) \tag{2}$$

where the H_j are the Tchebycheff-Hermite polynomials (see [6] page 155), and the z_h and z_k are the ordinates on the normal curve as shown in Fig. 2. Experience has shown that this series converges very slowly and a large number of terms are required to calculate values of r close to 1. As an example, for $h = k = 0$, 47 terms and 18 iterations were required to calculate $r = .99$ within an error of .005.

The bivariate normal may be rewritten in the form

$$L(h, k, r) = \frac{1}{2\pi} \int_0^r \frac{1}{\sqrt{1 - x^2}} \exp\left[-\frac{(h^2 - 2hkx + k^2)}{2(1 - x^2)}\right] dx + q_1 q_2 \tag{3}$$

and the value of r approximated by a quadrature technique. In an article by Froemel [2] in which several techniques are compared, a trapezoidal approach to evaluate the above integral written by Saunders is cited as the

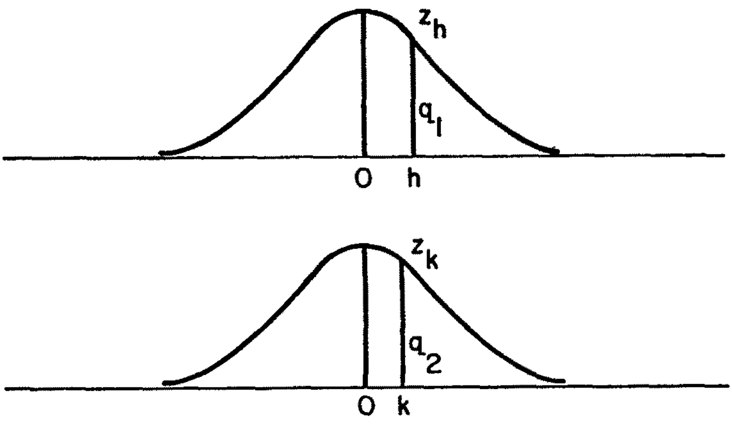

Figure 2

most accurate but slowest. Saunders' technique is to approximate the integral by the sum

$$(4) \qquad \frac{1}{2\pi}\sum_{i=0}^{n} f(x_i)\,\Delta x \quad \text{where} \quad f(x) = \frac{1}{\sqrt{1-x^2}}\exp\left[-\frac{(h^2-2hkx+k^2)}{2(1-x^2)}\right].$$

The value of Δx is fixed at .0078125, $x_0 = \Delta x/2$, $x_{i+1} = x_i + \Delta x$, and successive terms are evaluated until the sum equals $L - q_1q_2$ as determined by a change of sign. The value of $n \cdot \Delta x$ then approximates the value of r. This method gave the representative results at $h = k = 0$ shown in Table 1. However, since the function is smooth over a limited range of integration it seemed reasonable to expect that Gaussian quadrature to evaluate the integral supplemented by a Newton-Raphson iteration to converge on the unknown upper limit, r, might result in an improvement over the trapezoidal approximation.

The derivative becomes:

$$(5) \qquad L'(h, k, r) = \frac{1}{2\pi\sqrt{1-r^2}}\exp\left[-\frac{(h^2-2hkr+k^2)}{2(1-r^2)}\right]$$

and, after a variable transformation, the integral $[L(h, k, r) - q_1q_2]$ becomes:

$$(6) \qquad f(r) = \frac{r}{2\pi}\int_0^1 \frac{1}{\sqrt{1-u^2r^2}}\exp\left[-\frac{(h^2-2hkur+k^2)}{2(1-u^2r^2)}\right]du$$

$$\simeq \frac{r}{2\pi}\sum_{i=0}^{n} w_i g(u_i) \quad \text{where} \quad g(u_i)$$

$$= \frac{1}{\sqrt{1-u_i^2r^2}}\exp\left[-\frac{(h^2-2hku_ir+k^2)}{2(1-u_i^2r^2)}\right]$$

in which the u_i are the roots of the Legendre Polynomials, and the w_i are the associated weights for an $(n + 1)$-point quadrature.

After a starting value is determined, successive values are computed by the Newton-Raphson iteration method:

TABLE 1

Saunders' Algorithm

L	Computed r	True r	Terms Required
.315495	.399999	.40	52
.411699	.849993	.85	109
.428217	.899988	.90	116
.477473	.989857	.99	127

$$r_{i+1} = r_i - \frac{f(r_i) - m}{f'(r_i)} \tag{7}$$

where $m = L(h, k, r) - q_1q_2$. Iteration is continued until

$$|r_{i+1} - r_i| < \epsilon. \tag{8}$$

Using a 5-point quadrature and a convergence criterion of .0001 comparable results for $h = k = 0$ were computed, as shown in Table 2. It is evident that reasonable results are obtained for r with substantially fewer calculations than indicated in Table 1. As a starting estimate for the iteration, the first term of the series expansion was solved for r.

To compute the normal deviates, h and k, Hastings' unmodified approximation [4, p. 192] was first used. It soon became apparent that even an error in the fourth place had a significant effect on the subsequent calculation of r. However, the integral from which the deviates are calculated:

$$\Phi(k) = \frac{1}{\sqrt{2\pi}} \int_0^k \exp\left(-\frac{x^2}{2}\right) dx \tag{9}$$

poses precisely the same computational problem as is faced in the calculation of r, namely, it is necessary to compute and converge on a variable upper limit of a definite integral. Thus, an improvement of the original Hastings' estimates of h and k is readily made, computationally, since the essential tools—Gaussian quadrature and the iteration scheme—are necessarily a part of the program. Table 3 illustrates the magnitude of improvement using both 5- and 8-point quadratures. It is also apparent that these improvements in the h and k values are made rapidly since, at most, two iterations are required.

The use of an 8-point quadrature clearly will increase the accuracy of the computed r as well without a substantial increase of computing time. Certain marginal values of $L(h, k, r)$ which were close to zero or to the areas under the normal curve listed above, converged using an 8-point quadrature but failed when only 5 points were used.

TABLE 2

Gaussian Quadrature-Newton Raphson for $h = k = 0$

L	Computed r	True r	Iterations
.315495	.40003	.40	2
.411699	.85006	.85	4
.428217	.90015	.90	4
.477473	.99492	.99	6

TABLE 3

Different Quadrature Effects on h and k Calculation

Area $.5 - \Phi(h)$	True h	Hastings' Estimate Unmodified	5-Point Quadrature		8-Point Quadrature	
			Error in h	Iterations	Error in h	Iterations
.5	0	$-1.01 \cdot 10^{-7}$	$-.3 \cdot 10^{-13}$	1	$1 \cdot 10^{-22}$	1
.158655254	1	.999968	$4 \cdot 10^{-7}$	2	$3 \cdot 10^{-10}$	2
.022750132	2	2.000438	$2 \cdot 10^{-3}$	2	$1 \cdot 10^{-9}$	2
.001349898	3	3.000314	$2 \cdot 10^{-4}$	2	$1 \cdot 10^{-7}$	2

Since the speed of convergence as well as the actual convergence itself is affected by the starting estimate, the quadratic equation resulting from the inclusion of two terms of the series expansion is solved for r to provide a more accurate starting approximation. This estimate is bounded to prevent its becoming out of range. If the first estimate fails, one final pass for extreme cases is attempted with an arbitrary value. This has resulted in the convergence of all "reasonable" values.

Conclusion and Results

This study has shown that Gaussian Quadrature supplemented by Newton-Raphson iteration provides a rapid method for calculating a reasonable estimate of tetrachoric r. A package of programs may readily be prepared from the basic framework depending on the speed and accuracy requirements of the user. The variable components are:

1. The order of the Gaussian quadrature
2. The technique used to estimate h and k
3. The convergence criteria—one for h and k, the other for r.

A COMPUTER PROGRAM AND SOME EXAMPLES

A representative program using eight-point quadrature is listed as Table 5. For this evaluation, the convergence criteria used were 10^{-5} on h and k, and 10^{-4} on r. The program is written in double precision, FORTRAN IV for an IBM-360. Sufficient comments are included to help a user to adapt it to another configuration, if desired. Only standard Fortran functions are used.

Output for various cases near the singular points are given in Table 4. The comparative values of r are taken from [8]. On a 50 × 50 matrix of "live" data, the time required was .016 sec. per value of r.

TABLE 4

50 Calculated Values of R

P	Q_1	Q_2	R CALC	R(AMS-50)
0.7932800D-01	0.500000000D 00	0.158655254D 00	0.38640D-05	0.0
0.1137500D-01	0.500u00000D 00	0.227500000D-01	0.0	0.0
0.6750000D-03	0.500000000D 00	0.135000000D-02	0.0	0.0
0.8898100D-01	0.158655254D 00	0.500000000D 00	0.10000D 00	0.10
0.1351800D-01	0.227501320D-01	0.500000000D 00	0.99991D-01	0.10
0.8490000D-03	0.134989800D-02	0.500000000D 00	0.99759D-01	0.10
0.5000000D-05	0.134989800D-02	0.134989800D-02	0.10205D 00	0.10
0.3154950D 00	0.500000000D 00	0.500000000D 00	0.40000D 00	0.40
0.3333330D 00	0.500000000D 00	0.500000000D 00	0.50000D 00	0.50
0.1273980D 00	0.158655254D 00	0.500000000D 00	0.50000D 00	0.50
0.1037000D-02	0.158655254D 00	0.134989800D-02	0.50024D 00	0.50
0.4600000D-03	0.227500000D-01	0.135000000D-02	0.49988D 00	0.50
0.1349000D-02	0.134989800D-02	0.500000000D 00	0.70759D 00	0.70
0.3975840D 00	0.500000000D 00	0.500000000D 00	0.80000D 00	0.80
0.1530910D 00	0.158655254D 00	0.500000000D 00	0.80001D 00	0.80
0.4116990D 00	0.500000000D 00	0.500000000D 00	0.85000D 00	0.85
0.1349000D-02	0.158655254D 00	0.134989800D-02	0.84875D 00	0.85
0.4282170D 00	0.500000000D 00	0.500000000D 00	0.90000D 00	0.90
0.1579490D 00	0.158655254D 00	0.500000000D 00	0.89996D 00	0.90
0.2274900D-01	0.227501320D-01	0.500000000D 00	0.87146D 00	0.90
0.4494590D 00	0.500000000D 00	0.500000000D 00	0.95004D 00	0.95
0.1586310D 00	0.500000000D 00	0.158655254D 00	0.94960D 00	0.95
0.1586310D 00	0.500000000D 00	0.158655000D 00	0.94967D 00	0.95
0.1281300D 00	0.158655254D 00	0.158655254D 00	0.95004D 00	0.95
0.2275000D-01	0.227501320D-01	0.500000000D 00	0.89835D 00	0.95
0.2274200D-01	0.227501320D-01	0.158655254D 00	0.95016D 00	0.95
0.1602400D-01	0.227501320D-01	0.227501320D-01	0.95003D 00	0.95
0.1349000D-02	0.227501320D-01	0.134989800D-02	0.95139D 00	0.95
0.8090000D-03	0.134989800D-02	0.134989800D-02	0.95001D 00	0.95
0.1586550D 00	0.158655254D 00	0.500000000D 00	0.95988D 00	0.97
0.2275000D-01	0.158655254D 00	0.227501320D-01	0.95996D 00	0.97
0.4774730D 00	0.500000000D 00	0.500000000D 00	0.99096D 00	0.99
0.1450030D 00	0.158655254D 00	0.158655254D 00	0.99097D 00	0.99
0.1971200D-01	0.227501320D-01	0.227501320D-01	0.99098D 00	0.99
0.1102000D-02	0.134989800D-02	0.134989800D-02	0.99105D 00	0.99
0.2420389D 00	0.500000000D 00	0.500000000D 00	−0.50000D-01	−0.05
0.4600000D-05	0.227501320D-01	0.134989800D-02	−0.19950D 00	−0.20

TABLE 4 (continued)

P	Q1	Q2	R CALC	R(AMS-50)
0.1150267D 00	0.500000000D 00	0.500000000D 00	−0.75000D 00	−0.75
0.9156300D-02	0.158655254D 00	0.500000000D 00	−0.75000D 00	−0.75
0.1179000D-03	0.227501320D-01	0.500000000D 00	−0.75001D 00	−0.75
0.2000000D-06	0.134989800D-02	0.500000000D 00	−0.74998D 00	−0.75
0.7178310D-01	0.500000000D 00	0.500000000D 00	−0.90000D 00	−0.90
0.7048000D-03	0.500000000D 00	0.158655254D 00	−0.90001D 00	−0.90
0.1000000D-06	0.227501320D-01	0.500000000D 00	−0.90174D 00	−0.90
0.5054130D-01	0.500000000D 00	0.500000000D 00	−0.95004D 00	−0.95
0.4516720D-01	0.500000000D 00	0.500000000D 00	−0.96008D 00	−0.96
0.5100000D-05	0.158655254D 00	0.500000000D 00	−0.95711D 00	−0.96
0.3908300D-01	0.500000000D 00	0.500000000D 00	−0.97016D 00	−0.97
0.4000000D-06	0.158655254D 00	0.500000000D 00	−0.95985D 00	−0.97
0.2252670D-01	0.500000000D 00	0.500000000D 00	−0.99096D 00	−0.99
ELAPSED TIME IN MICROSECONDS		0.123654D 07		

The following notation is used in the program:

P—the joint proportion for which both marginal values are $\leq .5$
*FM*1—one marginal proportion, q_1, $\leq .5$
*FM*2—the second marginal proportion, q_2, $\leq .5$
R—the tetrachoric coefficient

An error condition is flagged by an r value greater than 1 being returned. The conditions are:

$R = 2$—Convergence failed; R estimate exceeds 1.
$R = 3$—Convergence not reached after 20 iterations.
$R = 4$—At least one marginal proportion exceeds .5.

Referring to Figure 1, if $(c/n, q_1, q_2)$ or $(b/n, p_1, p_2)$ are used as arguments, the sign of r will be positive for positive correlation and negative for negative correlation. If $(a/n, p_1, q_2)$ or $(d/n, p_2, q_1)$ are used, the sign of r will be opposite the true sign. [7, p. 199].

A test for equal values of P and one or both of the marginal proportions (implying one or more zero cells) should be made before this routine is used, and R set to the appropriate value.

The author is indebted to the reviewers for a number of helpful suggestions.

TABLE 5

```
        REAL FUNCTION TET8*8(P,FM1,FM2,R)
C   THIS FUNCTION EVALUATES THE TETRACHORIC COEFFICIENT,R, GIVEN
C   THE JOINT PROPORTION, P, AND THE CORRESPONDING MARGINAL
C   PROPORTIONS (BOTH<=.5) FM1 AND FM2. THIS MAY BE
C   USED AS AN EXTERNAL FUNCTION OR CALLED AS A SUBROUTINE.
      IMPLICIT REAL*8(A-H,O-Z)
C           FUNCTION EVALUATIONS FOR INTEGRALS
        FN1(W) = DEXP(-XX*W/2.D0)
      FN2(V,W)=DEXP((-H2K2+HK2*V*X)/(2.D0*(1.D0-W*XX)))/DSQRT(1.D0-W*XX)
C           ZEROS OF LEGENDRE POLYNOMIALS AND SQUARED VALUES
      DATA C1,C2,C3,C4,C5,C6,C7,C8/
     1.019855071751232D0,.101666761293187D0,
     2 .237233795041835D0,.408282678752175D0,.591717321247825D0,
     3 .762766204958165D0,.898333238706813D0,.980144928248768D0/
      DATA C1SQ,C2SQ,C3SQ,C4SQ,C5SQ,C6SQ,C7SQ,C8SQ/
     1 .000394223874247D0,.010336130351846D0,.056279873509951D0,
     2 .166694745769052D0,.350129388264702D0,.581812283426281D0,
     3 .807002607765472D0,.960684080371783D0/
C           THE ASSOCIATED WEIGHTS
      DATA W1,W2,W3,W4/.050614268145188D0,.111190517226687D0,
     1 .156853322938943D0,.181341891689181D0/
      DATA RPI,RTPI/.398942280401433D0,2.50662827463100D0/
      EQUIVALENCE (X,OLD)
        R=0.
C           EVALUATION OF H AND K
      J=1
C           CONVERGENCE CRITERION FOR H AND K
      EPS = 1.D-5
      X = FM1
      CON = (.5D0 - FM1)*RTPI
      GO TO 200
  100 J=2
      H = FNEW
     H2 = H*H
      ZH = RPI * DEXP(-H2/2.D0)
      X = FM2
      CON = (.5D0 - FM2)*RTPI
      GO TO 200
  150 FK = FNEW
      FK2 = FK*FK
      H2K2 = H2 + FK2
      HK2 = H*FK*2.D0
      ZK = RPI * DEXP(-FK2/2.D0)
      J=3
```

TABLE 5 (Continued)

```
C           STARTING ESTIMATE FOR R
      A = P-FM1*FM2
      CON = A*6.28318530717959D0
      ZHK = ZH*ZK
      EST = 2.D0*(DSQRT(DABS(HK2*A/ZHK+1.D0))-1.D0)/HK2
        IF(DABS(HK2).LE.1.D-8) EST = A/ZHK
        IF(DABS(EST).GT..80D0)EST = DSIGN(.80D0,A)
C           CONVERGENCE CRITERION FOR R
      EPS = 1.D-4
      GO TO 300
 200  IF(X.GT..5) GO TO 600
C           HASTINGS' APPROXIMATION
 250  E = DSQRT(-2.D0 * DLOG(X))
      E1=((.010328D0*E)+.802853D0)*E + 2.515517D0
      E2 = (((.001308D0*E)+.189269D0)*E+1.432788D0)*E+1.D0
      EST=E-E1/E2
 300  OLD = EST
C           ITERATION LOOP
 350  DO 500 I=1,20
        XX=OLD*OLD
      IF (J.EQ.3) GO TO 400
C           H AND K INTEGRAND EVALUATION FOR GAUSSIAN QUADRATURE
      FNUM=OLD*(W1*(FN1(C1SQ)+FN1(C8SQ))+W2*(FN1(C2SQ)+FN1(C7SQ))
     1+W3*(FN1(C3SQ)+FN1(C6SQ))+W4*(FN1(C4SQ)+FN1(C5SQ)))
      FNEW = OLD - (FNUM - CON)/FN1(1.D0)
      GO TO 450
 400  IF(DABS(OLD).GE.1.D0) GO TO 550
C           R INTEGRAND EVALUATION
      FNUM = OLD*(W1*(FN2(C1,C1SQ)+FN2(C8,C8SQ)) +
     1 W2*(FN2(C2,C2SQ)+FN2(C7,C7SQ))+W3*(FN2(C3,C3SQ)+FN2(C6,C6SQ))
     2 + W4*(FN2(C4,C4SQ) + FN2(C5,C5SQ)))
      FNEW = OLD - (FNUM - CON)/FN2(1.D0, 1.D0)
C           TEST FOR CONVERGENCE
 450  IF(DABS(OLD-FNEW).GT.EPS) GO TO 500
      GO TO (100,150,650),J
 500  OLD = FNEW
      R = 3.
      GO TO 1000
 550  IF(R.EQ.2.) GO TO 1000
      R=2.
      OLD = DSIGN(.97D0,A)
      GO TO 350
 600  R = 4.
      GO TO 1000
 650  R = FNEW
 1000 TET8=R
      RETURN
      END
```

REFERENCES

[1] Chesire, L., Saffir, M. and Thurstone, L. *Computing diagrams for the tetrachoric correlation coefficient.* Chicago: University of Chicago Press, 1933.

[2] Froemel, E. A comparison of computer routines for the calculation of the tetrachoric correlation coefficient, *Psychometrika,* 1971, **36**, pp. 165–173.

[3] *Handbook of Mathematical Functions (AMS-55).* U. S. Department of Commerce, National Bureau of Standards, 1964.

[4] Hastings, C. *Approximations for digital computers.* Princeton: Princeton University Press, 1955.

[5] IBM System 360 Scientific Subroutine Package (360A-CM-03X); Call 360 Time Sharing System, 1970.

[6] Kendall, M. & Stewart, A. *The advanced theory of statistics.* New York: Hefner Publishing Co., 1961.

[7] McNemar, Q. *Psychological statistics.* New York: Wiley, 1955.

[8] Tables of the Bivariate Normal Distribution Function (AMS-50). U. S. Department of Commerce, National Bureau of Standards, 1959.

Manuscript received 6/1/72
Revised manuscript received 10/13/72