

The Pseudo-bivariate Structure of the 2×2 Contingency Table: A Preliminary analysis

E. W. HENRY

THE paper by R. C. Geary [1] on the comparative sensitivity of tests of significance to sample size contains the interesting suggestion that any 2×2 contingency table might be thought of as a sample realisation of a bivariate Normal distribution. Such an approach, if workable, might permit reduced sample size to establish significant correlation between the inherent x and y variates in question. These variates are the pair whose joint probability volume is supposedly estimated over four quadrants of the (x, y) plane by the four entries in the 2×2 table. An important outcome would be the finding of significant correlation in cases where the usual χ^2 test failed to find significant relationship through inadequate size of sample.

The following preliminary investigation shows how a bivariate Normal distribution can always be found to match the relative frequencies in the 2×2 table, hence the designation "Pseudo-Bivariate" above. Part 1 of the essay describes how to interpolate a selected cube, for each corner of which we have exact knowledge of r , the inherent correlation coefficient, by means of published data [2]. Part 2 explains the alternative approach, via the Tetrachoric Functions [3] used with a power series in r . An equation in r is then solved for r by means of Newton's method, as given in [4]. Part 3 shows the trouble-free simplicity of the central case, i.e. when the origin of the supposed Normal bivariate lies at the centre of the 2×2 relative frequencies, one of which is taken to relate to each of four quadrants formed by the x and y axes. Part 4 contains four illustrative numerical examples, each solved by the Taylor interpolation method of Part 1, and also solved by the Tetrachoric functions and Newton's method, as described in Part 2. In Part 5 the critical sample size required by χ^2 is compared with that

required by r , to establish significant relationship. The results of the four numerical examples are used as data to illustrate the marked increase in efficiency (via reduced critical sample size) obtained by means of r . The question is raised as to why χ^2 requires a much larger critical size of sample and an answer tentatively suggested. Some of the implications for 2×2 contingency tables are then examined. I acknowledge a debt to Dr Geary for encouragement and valuable suggestions at several stages of preparing what follows.

This essay avoids discussion of the validity of assuming a Normal bivariate inherent distribution for non-cardinal data. It confines its treatment to methods of finding r for any 2×2 contingency table, on the assumption that r is meaningful for the data in question. The theoretical justification of the procedure of finding r , for non-cardinal data, an aspect commented on by Geary at the end of his paper [1], remains an open question. The views of colleagues on this topic would be welcomed.

The main object of this paper is to study the methodology of deriving estimates of the correlation coefficient r inherent in every 2×2 table. This we do by recourse to many constructed examples illustrating the different kind of situations that can arise.

PART I: *Estimation of r by linear interpolation from the corners of a cube*

We start with the general bivariate Normal distribution, standardised so that

$$(1) \quad \sigma_x = \sigma_y = 1$$

$$(2) \quad \mu_x = \mu_y = 0$$

$$(3) \quad \frac{1}{2\pi\sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2}(x^2 + y^2 - 2rxy)/(1-r^2) \right] dx dy = 1$$

r being the coefficient of correlation between x and y .

$$(4) \quad L(h, k, r) = \frac{1}{2\pi\sqrt{1-r^2}} \int_h^{\infty} \int_k^{\infty} \exp \left[-\frac{1}{2}(x^2 + y^2 - 2rxy)/(1-r^2) \right] dx dy$$

Thus $L(h, k, r)$ is the volume of probability, a fraction of the unit total, above the quadrant between the lines $x = h$, $y = k$ and $x = \infty$, $y = \infty$.

The properties of the distribution are as follows (formulae (5) to (10)):

The marginal distribution of x is given by

$$(5) \quad \int_{-\infty}^h \frac{\exp(-x^2/2) dx}{\sqrt{2\pi}} = P(h)$$

where $P(h)$ is the single-variate standard Normal probability from $-\infty$ to h , and given in tables such as that included in Section 1 of [5].

$$(6) \quad \int_{-\infty}^k \frac{\exp(-x^2/2) dx}{\sqrt{2\pi}} = P(k)$$

having a corresponding meaning for y .

$$(7) \quad L(k, h, r) = L(h, k, r)$$

$$(8) \quad L(-h, k, r) = -L(h, k, -r) - P(k) + 1.0$$

$$(9) \quad L(h, -k, r) = -L(h, k, -r) - P(h) + 1.0$$

$$(10) \quad L(-h, -k, r) = L(h, k, r) + P(h) + P(k) - 1.0$$

$$(11) \quad L(0, 0, r) = 0.25 + \frac{1}{2\pi} \sin^{-1} r.$$

$$(12) \quad L(h, k, 0) = [1 - P(h)][1 - P(k)]$$

The properties (7) to (12) and the notation $L(h, k, r)$ are given in [2, pages VI *et seq.*] Formula (11) will be considered in Section 3 following, as being the central case. Formula (12) shows that for $r = 0$ the bivariate probability for the quadrant is the product of the single-variate probabilities for x and y , over the x -range (h, ∞) and the y -range (k, ∞) , respectively.

The formulae (7) to (10) show that negative values for any or all of h , k and r can be accommodated by means of the L -function for $0 \leq h \leq \infty$; $0 \leq k \leq \infty$; $-1 \leq r \leq 1$. Thus the tabulated values of L for such a range of h , k , and r suffice to evaluate L for any h , k and r . The single-variate Normal probability tabulated values are of course required. It is to be noticed that $L(h_1, k_1, r_1)$ is distinct from $L(h_1, k_1, -r_1)$, i.e. for h and k specified the L -value for a specified r_1 positive is distinct from the L -value for $-r_1$. Evaluation of L for positive and negative h and k values will be illustrated numerically in Part 4, by using formulae (7) and (10) as required.

Estimation of h and k from sample data and selection of a cube containing $L(h, k, r)$

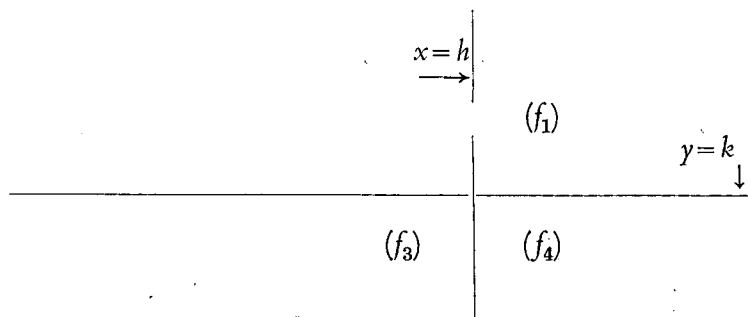
Let us suppose that we are presented with the following typical 2×2 contingency table:

<i>Effect</i>	<i>Not Treated</i>	<i>Treated</i>	<i>Total</i>
Cured	n_2	n_1	$n_1 + n_2$
Not Cured	n_3	n_4	$n_3 + n_4$
Total	$n_2 + n_3$	$n_1 + n_4$	$n_1 + n_2 + n_3 + n_4$

The relative frequencies are as follows:

<i>Effect</i>	<i>Not Treated</i>	<i>Treated</i>	<i>Total</i>
Cured	f_2	f_1	$f_1 + f_2$
Not Cured	f_3	f_4	$f_3 + f_4$
Total	$f_2 + f_3$	$f_1 + f_4$	1.0

The assumed inherent Normal bivariate distribution is now invoked, so as to equate f_1 with $L(h, k, r)$ given by (4). The following sketch may make matters clearer, the f -values quoted in the sketch being the volume of probability related to the quadrant in question:



It is necessary to estimate h and k . By the properties of the bivariate Normal distribution for the marginal distributions of x and y , the estimation of h and k is as follows:

$$(13) \quad P(h) = f_2 + f_3$$

where $P(h)$ is defined in (5) above;

$$(14) \quad P(k) = f_3 + f_4$$

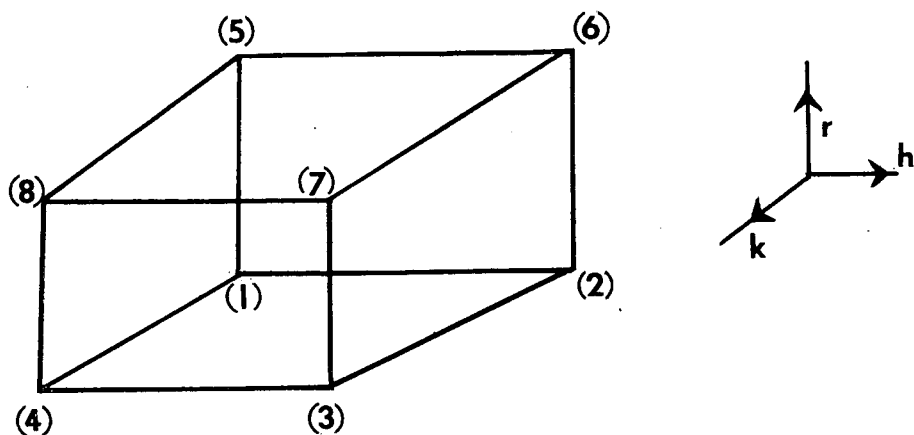
where $P(k)$ is defined in (6) above.

Thus tabulated probabilities for the Normal distribution, such as that in Section 1 of [5], can be used with the f_i values to estimate h and k . The value of L is given by

$$(15) \quad L(h, k, r) = f_1,$$

where the unknown quantity r is to be estimated. Equation (15) may require some manipulation, via equations (8), (9) or (10), to ensure non-negative values of h and k , before proceeding any further.

For the purpose of estimating r it is convenient to consider a three-dimensional space having h , k and r as axes, with $L(h, k, r)$ as a function having a certain value at each point (h, k, r) . L is tabulated in [2], for h and k each in the range $(0, 4)$ at intervals of 0.1 , and for r in the range $(-1, +1)$, at intervals of 0.05 . For ease of exposition we take the r -interval to be 0.1 also, in what follows. In 3-D space the following sketch of a cube describes the situation:



The eight corners numbered (1) to (8) are the corners of a cube having edges of length 0.1 units and parallel to the axes of h , k and r , as indicated. From the tables [2] we can find the value of L at each of eight specified corners, for example L_2 for corner (2) is 0.174428 for $h = 0.3$, $k = 0.5$ and $r = 0.4$ at corner (2). The identification of the corners will be amply illustrated and explained below in Part 4.

In preparing to use the tables [2] we supposedly have first found h and k , as explained above in (13) and (14), via the margins of the 2×2 table and the tabulated single-variate Normal probability distribution, as given for instance in [5]. Generally the point (h, k) thus obtained will be between h_1 and h_2 , k_1 and k_2 , shown with tabulated values of L and r in [2]. Our search now is confined to the L -values for the set of four (h, k) values (h_1, k_1) , (h_1, k_2) , (h_2, k_1) , (h_2, k_2) , these L -values varying with r . We find two such tabulated sets of four L -values, one set having r_1 and the second set r_2 , with $r_2 - r_1 = 0.10$. One chosen set of four L -values must span the L -value given by f_1 from the 2×2 table, i.e. at least one of the set of four must be not less than it and at least one not greater than it. For the second chosen set of four L -values, f_1 may be spanned by them or one of them is nearer to it than any other set of four tabulated L -values for the specified h_1, h_2, k_1, k_2 .

We thus have selected eight tabulated values of $L(h, k, r)$, and the (h, k, r) in question form the eight corners of the cube having edges of length 0.1 units and containing within it the f_1 value of L , the specified h and k and the r , as yet unknown. We are now ready to interpolate for r . We might interpolate from the corners (1), (3), (6), and (8), although we could interpolate from each of the eight, or confine our interpolation to the best corner, as will be explained in Part 4.

We use the general Taylor expansion of a function around a point (h_0, k_0, r_0) , to the first order of the small quantities $\Delta h_0, \Delta k_0, \Delta r_0$, namely

$$(16) \quad L(h, k, r) = L(h_0, k_0, r_0) + \Delta h_0 \left(\frac{\partial L}{\partial h} \right)_0 + \Delta k_0 \left(\frac{\partial L}{\partial k} \right)_0 + \Delta r_0 \left(\frac{\partial L}{\partial r} \right)_0$$

Here $L(h, k, r)$ is the given value f_1 from the 2×2 table. $L(h_0, k_0, r_0)$ is a tabulated value of L at any one corner of the eight chosen corners, say corner (1), Δh_0 and Δk_0 are the changes in h and k , $\left(\frac{\partial L}{\partial h} \right)_0$ etc. are values of the partial derivative at

(h_0, k_0, r_0) . Exact values of $\left(\frac{\partial L}{\partial h} \right)_0$, $\left(\frac{\partial L}{\partial k} \right)_0$ are available but $\left(\frac{\partial L}{\partial r} \right)_0$ must be estimated. Δr_0 is the sole unknown in (16) regarded as an equation.

It is easily shown from (4) that—

$$(17) \quad \frac{\partial L}{\partial h} = -Z(h) [1 - P(h^1)],$$

with $h^1 = (k - rh) / \sqrt{1 - r^2}$

P being given by (4) and Z the normal ordinate, also tabled. There is an analogous formula for $\frac{\partial L}{\partial k}$. We approximate $\left(\frac{\partial L}{\partial r} \right)_0$ by—

$$(18) \quad \left(\frac{\partial L}{\partial r} \right)_0 \doteq [L(h_0, k_0, r_0 + c) - L(h_0, k_0, r_0 - c)] / 2c,$$

c being the smallest interval (actually $\cdot 05$) in the $L(h, k, r)$ tables.

PART 2: *Estimation of r via the tetrachoric functions of h and k , combined with Newton's method of solving an equation for r*

The following method is mainly algebraic and readers may indeed prefer to use the interpolation method given above in Part 1. In the tetrachoric approach a power series in r is developed, the coefficients being constants depending on the specified values of h and k . The power series is then equated with f_1 , explained above as the relative frequency in the top right-hand quadrant of a given 2×2 table. This equation is solved for r , via Newton's method.

The following treatment of the tetrachoric functions is given in the notes for Tables V–VII of [3], with Table VII of [3] having tabulated values of the first 20 tetrachoric functions of h , at intervals of $0\cdot1$ from $h = 0\cdot0$ to $h = 4\cdot0$. Methods of interpolation, for h -values within any interval $0\cdot1$, are shown. The same tetrachoric functions apply to k , each of these functions having only one variate as argument, this being h or k .

The equation for the power series in r is as follows:

$$(19) \quad f_1 = T_0(h) T_0(k) + r T_1(h) T_1(k) + r^2 T_2(h) T_2(k) \\ + \dots + r^s T_s(h) T_s(k)$$

where f_1 , h and k are obtained from the given 2×2 table, as explained above in Part 1 down to and including equation (15), and both h and k non-negative. It may be worth repeating here that equations (8), (9), and (10) may be required, to manipulate (19) in order to ensure non-negative values of h and k , before proceeding any further.

It is now required to define the tetrachoric function $T_s(h)$, which is tabulated in Table VII of [3] for $s = 0$ to $s = 19$ and for $h = 0\cdot0$ ($0\cdot10$) $4\cdot0$, i.e. the tabulated values can cater for the power series in r as far as r^{19} .

$$(20) \quad T_0(h) = \int_h^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

This is the standard Normal single-variate probability integral between h and ∞

$$(21) \quad T_s(h) = \frac{(-1)^{s-1}}{\sqrt{s!}} \frac{d^{s-1}}{dh^{s-1}} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2} \right)$$

defines the tetrachoric function of h for $s = 1, 2, \dots, n$.

Interpolation of $T_s(h)$ and $T_s(k)$

A given 2×2 table will in general yield values of h and k which are not exact multiples of 0.1. For the specified h -value it is therefore necessary to estimate $T_s(h)$ and likewise for k , before solving equation (19) for r .

"If we wish to get the value of $T_s(h)$ correct to the seventh figure, then fourth differences must be used in our interpolation formula; or, if we use central differences, we must take for $\theta (= 1 - \phi)$,

$$(22) \quad T_s(h + \theta) = \theta T_s(h + 0.1) + \phi T_s(h)$$

$$\begin{aligned} & \frac{-\theta\phi}{6} \{ (1 + \theta) \delta^2 T_s(h + 0.1) + (1 + \phi) \delta^2 T_s(h) \} \\ & + \frac{\theta(1 + \theta)\phi(1 + \phi)}{120} \{ (2 + \theta) \delta^4 T_s(h + 0.1) + (2 + \phi) \delta^4 T_s(h) \} \end{aligned}$$

The values of $\delta^2 T_s(h)$ and $\delta^4 T_s(h)$ may be found from the fundamental formulae

$$(23) \quad \begin{cases} \delta^2 T_s(h) = T_s(h + 0.1) + T_s(h - 0.1) - 2T_s(h) \\ \delta^4 T_s(h) = \delta^2 T_s(h + 0.1) + \delta^2 T_s(h - 0.1) - 2\delta^2 T_s(h). \end{cases}$$

(Quotation from [3, pages xlv and xlv].)

Two comments are relevant here. First, the θ quoted above is a fraction of the interval 0.1, thus for example if a value 1.372 is being used for h , then $\theta = 0.72$ and $\phi = 0.28$ for (22) and (23) above, the 0.072 being 72 per cent of the basic unit interval 0.1. Secondly, it is clear that use of powers of r up to and including the n th, in formula (19), will in general involve $2n$ interpolations in order to estimate $T_s(h)$ and $T_s(k)$ for $s = 1, 2, \dots, n$. Thus a considerable amount of preparation is required, before one is ready to solve equation (19) for r .

Solution of the Equation for r , by Newton's Method

As given by Lanczos [4, pages 10 and 11], the general formula for Newton's solution can be stated as follows, for $f(r)$ a function of r :

$$(24) \quad t = \left[-\frac{f'(r_0)}{f(r_0)} + \frac{1}{2} \frac{f''(r_0)}{f'(r_0)} \right]$$

where r_0 is the best available approximation of a root of $f(r)$ fairly near r_0 , t is the increment of r to be added to r_0 , to give a better approximation of the root,

$$f'(r_0) \text{ is } \left(\frac{df}{dr} \right)_{r=r_0} \quad \text{and} \quad f''(r_0) \text{ is } \left(\frac{d^2f}{dr^2} \right)_{r=r_0}$$

For the particular application being investigated here, r_0 is the best available estimate of r which satisfies (19). Equation (19) can be written

$$(25) \quad f_1 - T_0(h) T_0(k) - \sum_{s=1}^n r^s T_s(h) T_s(k) = 0$$

The values of f_1 and the various $T_j(h)$ and $T_j(k)$ for $j = 0, 1, 2, \dots, n$ are given constants.

By trial, two values of r can be found making the left-hand side of (25) negative and positive say r_0 and r_1 . The value sought for r must lie between r_0 and r_1 . We substitute r_0 in the right-hand side of (24) to obtain the increment t_1 and then revise r_0 to become $r_0 + t_1$. Then a second increment t_2 is obtained by using $r_0 + t_1$ instead of r_0 in the right-hand side of (24). Now $(r_0 + t_1 + t_2)$ is used in the right-hand side of (24) to give a further increment t_3 and so on repeatedly, until t_m (for the m th repetition) is satisfactorily small. The value of the estimate of r , the required root of equation (25), is given by

$$(26) \quad \hat{r} = r_0 + t_1 + t_2 + \dots + t_m$$

Two or three iterations generally suffice.

The differentials of f are as follows:

$$(27) \quad \frac{df}{dr} = f'(r) = T_1(h) T_1(k) + 2r T_2(h) T_2(k) + \dots + sr^{s-1} T_s(h) T_s(k)$$

$$(28) \quad \frac{d^2 f}{dr^2} = f''(r) = 2 T_2(h) T_2(k) + 6r T_3(h) T_3(k) + \dots + s(s-1)r^{s-2} T_s(h) T_s(k)$$

It is apparent that a relatively heavy amount of calculation is required, in order to estimate r by the method of tetrachoric functions given above followed by Newton's method of approximating the root of the power series. It is therefore possible that the relatively shorter method of interpolation from the corner of a cube, as explained in Part 1, may be generally preferred and used. Trouble may arise in using (24) if the root being sought happens to be a multiple root. This situation is fortunately not likely for the case of r , the unique single correlation coefficient.

PART 3: *The Central Case, having $h = 0$ and $k = 0$*

Formula (11) above describes the simple case which occurs for $h = k = 0$, i.e.

$$(11) \quad L(0, 0, r) = 0.25 + \frac{1}{2\pi} \sin^{-1} r$$

Thus for f_1 specified, via a given 2×2 table,

$$(29) \quad f_1 = 0.25 + \frac{1}{2\pi} \sin^{-1} r$$

from which

$$(30) \quad r = \sin\{2\pi(f_1 - 0.25)\}$$

Thus for $f_1 > 0.25$, r is positive

for $f_1 = 0.25$, r is zero

for $f_1 < 0.25$, r is negative

for $f_1 = 0.5$, the maximum possible, $r = \sin(\pi/2) = 1.0$ and the whole frequency is divided equally between f_1 and the diagonally opposite f_3 , each being 0.5.

For $f_1 = 0$, the minimum possible, $r = \sin(-\pi/2) = -1.0$ and the whole frequency is divided equally between f_2 and the diagonally opposite f_4 , each being 0.5.

The above properties describe the central case and the estimation of r for a given f_1 is obtained directly via formula (30). For the layout of the relative frequencies in the following sketch

(f_2)	γ	(f_1)
		x
(f_3)		(f_4)

these relative frequencies are as follows:

$$(31) \quad f_1 = f_3 = 0.25 + \frac{1}{2\pi} \sin^{-1} r$$

$$(32) \quad f_2 = f_4 = 0.25 - \frac{1}{2\pi} \sin^{-1} r$$

Their sum is unity, as it should be.

Formula (30) is so easy to use that the conditions necessary for its validation, namely equations (31) and (32), merit a brief examination. Some case can be made for making $(f_2 + f_3)$ equal to $(f_1 + f_4)$, i.e. taking a proportionate equal number of untreated and treated, respectively, in the typical sample data to be analysed for the effects of treatment versus degree of cure. Unless, however, the conditions (31) and (32) for the four relative frequencies are now closely approximated, we are not justified in using (30) to estimate r , because the centre of the bivariate Normal distribution being simulated is in fact not at the division of the unit total between f_1, f_2, f_3 and f_4 .

The conclusion ventured at this point is that there is no harm in looking carefully at the relative frequencies as specified and even possibly adjusting the "Not Treated" pair to make $(f_2 + f_3)$ equal to $(f_1 + f_4)$. Unless at this stage $f_1 \approx f_3$ and $f_2 \approx f_4$ to a high degree of approximate equality, the central case has not been established and formula (30) may not validly be used.

PART 4: Numerical Examples

This part contains four contrived examples, the numerical data being chosen to illustrate positive and negative occurrences of h , k and r . All four examples have r estimated by the interpolation method described in Part 1 above and they are also solved for r by means of the tetrachoric approach explained in Part 2.

Example 1

The following scheme sets out the relative frequencies in a 2×2 contingency table:

$f_2 = 0.2270$	$f_1 = 0.1250$	$f_1 + f_2 = 0.3520$
$f_3 = 0.4715$	$f_4 = 0.1765$	$f_3 + f_4 = 0.6480$
$f_2 + f_3 = 0.6985$	$f_1 + f_4 = 0.3015$	$f_1 + f_2 + f_3 + f_4 = 1.0000$

"

The value 0.6085 for $(f_2 + f_3)$, via (5) above, gives $h = 0.52$.

Likewise 0.6480 for $(f_3 + f_4)$, gives $k = 0.38$, via (6) above.

The value 0.1250 for f_1 gives $L = 0.1250$.

Thus h , k and L are all positive and we are ready to search for r in the tables [2].

The figures in Table I following show the corners of the cube containing the specified L -value, in the right region of h and k and indicating a value of r between 0.1 and 0.2. The serial numbers of the cube's corners are shown in parentheses.

TABLE I: *Basic Values for Example 1.*

Corner	h	k	r	L
(1)	0.5	0.3	0.1	0.131433
(2)	0.6	0.3	0.1	0.117625
(3)	0.6	0.4	0.1	0.106933
(4)	0.5	0.4	0.1	0.119425
(5)	0.5	0.3	0.2	0.145281
(6)	0.6	0.3	0.2	0.130776
(7)	0.6	0.4	0.2	0.119738
(8)	0.5	0.4	0.2	0.132889

Of the eight sets, clearly the actual values of h , k and L (derived from the contingency table) are nearest to those of corner (8), selected as base for detailed working, as follows:

For Corner (8)

$$r_0 = 0.20, f_1 = 0.1250, h_0 = 0.5, k_0 = 0.4,$$

$$\Delta h_0 = 0.02, \Delta k_0 = -0.02, L_0 = L_8 = 0.132889,$$

$$h_0^1 = 0.306186, k_0^1 = 0.428661$$

$$Z(h_0) = 0.352065, 1 - P(h_0^1) = 0.379733,$$

$$Z(k_0) = 0.368270, 1 - P(k_0^1) = 0.334086,$$

$$\partial L / \partial h = -0.133691$$

$$\partial L / \partial k = -0.123034$$

$$\partial L / \partial r \doteq (L_9 - L_4) / (2 \times 0.1) = 0.137015$$

where L_9 is outside the cube* and having

$$h_9 = 0.5, k_9 = 0.4, r_9 = 0.3, L_9 = 0.146828$$

$$\Delta r_0 = \left[f_1 - L_0 - \Delta h_0 \left(\frac{\partial L}{\partial h} \right)_0 - \Delta k_0 \left(\frac{\partial L}{\partial k} \right)_0 \right] / \left(\frac{\partial L}{\partial r} \right)_0$$

which leads to $-0.007676 / 0.137015$, giving

$$\Delta r_0 = -0.05602$$

thus

$$\hat{r} = r_0 + \Delta r_0 = 0.200 - 0.056 = 0.144$$

By similar methods, three further Taylor (cube) interpolations for $\hat{r}$ are as follows:

$$\text{from corner (1)} \quad 0.145$$

$$\text{from corner (3)} \quad 0.149$$

$$\text{from corner (6)} \quad 0.138$$

The average of these four estimates gives $\hat{r} = 0.144$.

It therefore seems likely that $\hat{r} = 0.14$ is correct to two places. Having regard to the fact that our original 2×2 table is invariably deemed a sample for the purpose of inference making, this approximation is quite sufficient.

*It may be pointed out here that an improved estimate of $\partial L / \partial r$ could be obtained from the fully detailed tables [2] by taking $[L(h = 0.5, k = 0.4, r = 0.25) - L(h = 0.5, k = 0.4, r = 0.15)] / 0.1$, that is by using r -intervals of 0.05.

Tetrachoric Interpolation for Example 1

As before we are given by the 2×2 table the values $h = 0.52$, $k = 0.38$, $L = 0.1250$. We will take the power series in r up to and including r^6 and use the interpolation only as far as second differences, omitting the fourth difference terms in (22).

The given h -value is 0.52 so we take $h_0 = 0.5$ and $\theta_h = 0.2$ (which is $1/5$ of the tabular [3] interval of 0.1). Thus $\phi_h = 0.8$. Similarly, we take $k_0 = 0.3$, $\theta_k = 0.8$, $\phi_k = 0.2$. From formula (20) we know that

$$T_0(h_0 + \theta_h) = 1 - P(h) = 1 - f_2 - f_3 = 0.3015,$$

$$T_0(k_0 + \theta_k) = 1 - P(k) = 1 - f_3 - f_4 = 0.3520.$$

Also

$$\frac{-\theta_h \phi_h (1 + \theta_h)}{6} = \frac{-\theta_k \phi_k (1 + \phi_k)}{6} = -0.32$$

$$\frac{-\theta_h \phi_h (1 + \phi_h)}{6} = \frac{-\theta_k \phi_k (1 + \phi_k)}{6} = -0.48$$

Table 2 following shows part of the actual manipulation of the tabular [3] data for estimation of the tetrachoric coefficients of the power series in r , up to r^6 , as given in (19) above. The seven decimal places given in the tables [3] are rounded to six.

TABLE 2: Estimation of tetrachoric coefficients of r for Example 1

Description of entry in row		T_1	T_2	T_3	T_4	T_5	T_6
$T_s(h_0 - 0.1)$	(1)	0.368270	0.104163	-0.126290	-0.085396	0.069442	0.073704
$T_s(h_0)$ where $h_0 = 0.5$	(2)	0.352065	0.124474	-0.107798	-0.098814	0.050217	0.082414
$T_s(h_0 + 0.1)$	(3)	0.333225	0.141375	-0.087065	-0.107742	0.029494	0.085909
$T_s(h_0 + 0.2)$	(4)	0.312254	0.154558	-0.065013	-0.111989	0.008554	0.084230
$\delta^2 T_s(h_0) = (1) + (3) - 2 \times (2)$	(5)	-0.002635	-0.003410	0.002241	0.004490	-0.001498	-0.005215
$\delta^2 T_s(h_0 + 0.1)$	(6)	-0.002131	-0.003718	0.001319	0.004681	-0.000217	-0.005174
$T_s(h_0 + \theta_h) = \theta_h \times (3) + \phi_h \times (2)$ $-0.48 \times (5) - 0.32 \times (6)$	(7)	0.348491	0.128136	-0.103801	-0.100965	0.046152	0.083529
$T_s(k_0 + \theta_k)$	(8)	0.371153	0.099727	-0.129642	-0.082217	0.072994	0.071357
$T_s(h_0 + \theta_h) \cdot T_s(k_0 + \theta_k)$	(9)	0.129343	0.012779	0.013457	0.008301	0.003369	0.005960

Row (9) of Table 2 gives the coefficients of r , apart from r^0 , for which $T_0(h) \cdot T_0(k)$ is 0.106128 , being the product of 0.3015 and 0.3520 . Equation (25) therefore becomes

$$(33) \quad 0.1250 - 0.106128 - 0.129343r - 0.012779r^2 - 0.013457r^3 - 0.008301r^4 \\ - 0.003369r^5 - 0.005960r^6 = 0$$

which can be described as

$$(34) \quad f(r) = 0$$

Then

$$(35) \quad \frac{df}{dr} = \frac{-0.129343 - 0.025558r - 0.040371r^2}{-0.033204r^3 - 0.016845r^4 - 0.035760r^5}$$

and

$$(36) \quad \frac{d^2f}{dr^2} = \frac{-0.025558 - 0.080742r - 0.099612r^2}{-0.067380r^3 - 0.178800r^4}$$

Since $f(0) = 0.018872$ and $f(0.5) = -0.051393$, there is a root between 0.0 and 0.5. We are now ready to use Newton's method to search for this root.

Take $r_0 = 0.25$, thus $f(r_0) = -0.014509$, $f'(r_0) = -0.138875$, $f''(r_0) = -0.053721$

Via equation (24)

$$(37) \quad \frac{1}{t_1} = \frac{-f'(r_0)}{f(r_0)} + \frac{1}{2} \frac{f''(r_0)}{f'(r_0)} = -9.571645 + 0.193419$$

$$t_1 = -0.106630,$$

$$r_0 + t_1 = 0.143337$$

Now setting $r_0 = 0.143337$ we repeat the process, using $f(r_0) = 0.000021$, $f'(r_0) = -0.133943$, $f''(r_0) = 0.039457$, and get

$$\frac{1}{t_2} = 6378.24 + 0.15 ;$$

$$t_2 = 0.000157 ;$$

$$r_0 + t_2 = 0.143527.$$

At this stage we will settle for $\hat{r} = 0.144$ as $f(r_0 + t_2)$ has a value -0.00000104 , small enough. It is worth noticing that for both applications of (37) to give $1/t$, the second term $\frac{1}{2} f''(r_0)/f'(r_0)$ forms an almost negligible part of the total. This result, however, does not necessarily apply in general.

It is consoling to find that the simple interpolation from corner (8) gave $r = 0.144$, agreeing with the relatively precise tetrachoric value 0.144.

Other Worked Examples

We have worked three other examples as follows. The general object was twofold:

- (i) to assess the comparative accuracy of the two methods for estimating r ,
- (ii) to provide additional comparisons of the efficiency of X^2 and r for inferring significance.

The second objective is dealt with in Part 5 below. Table 3 gives the relative frequencies for Examples 2, 3, and 4.

TABLE 3: *Relative frequencies for Examples 2, 3 and 4*

<i>Effect</i>	<i>Untreated</i>	<i>Treated</i>	<i>Total</i>
<i>Example 2</i>			
Cured	0.013000	0.190269	0.203269
Not cured	0.153023	0.643708	0.796731
Total	0.166023	0.833977	1.000000
<i>Example 3</i>			
Cured	0.366371	0.332097	0.698468
Not cured	0.161532	0.140000	0.301532
Total	0.527903	0.472097	1.000000
<i>Example 4</i>			
Cured	0.030471	0.813281	0.843752
Not cured	0.019000	0.137248	0.156248
Total	0.049471	0.950529	1.000000

These examples also involved some negative values of h and k , necessitating changes to the positive values tabled, by using formulae (7) to (10). No complications were encountered. Estimated values of r are as set out in Table 4, with Example 1 results included for completeness, each Taylor estimate being the "nearest corner" estimate, obtained by the cube interpolation method.

TABLE 4: *Estimates of r for Examples 1 to 4*

<i>Example</i>	<i>Via Taylor</i>	<i>Via Tetrachoric</i>
1	0.144	0.144
2	-0.344	-0.344
3	-0.017	-0.017
4	0.353	0.354

There is little doubt that the simpler cube (Taylor) method provides accurate estimates of inherent r in 2×2 tables, the nearest corner being used as base of interpolation.

PART 5: *Comparison of critical sample size for X^2 with that for r*

For any 2×2 table, expressed in relative frequencies we have the following scheme:

Observed			Expected		
f_2	f_1	f_1+f_2	$(f_1+f_2)(f_2+f_3)$	$(f_1+f_2)(f_1+f_4)$	f_1+f_2
f_3	f_4	f_3+f_4	$(f_3+f_4)(f_2+f_3)$	$(f_3+f_4)(f_1+f_4)$	f_3+f_4
f_2+f_3	f_1+f_4	$f_1+f_2+f_3+f_4$ = 1.0	f_2+f_3	f_1+f_4	1.0

The value of X^2 is given by:

$$(38) \quad X^2 = [f_1 - (f_1+f_2)(f_1+f_4)]^2/f_1 + [f_2 - (f_1+f_2)(f_2+f_3)]^2/f_2 \\ + [f_3 - (f_3+f_4)(f_2+f_3)]^2/f_3 + [f_4 - (f_3+f_4)(f_1+f_4)]^2/f_4$$

Following Geary's [1] Examples 2 and 3 we set the critical value of X at 3.841, the 0.95 probability level for one degree of freedom. We equate this with nX^2 , where X^2 is obtained from a given 2×2 table via (38). Thus

$$(39) \quad 3.841 = nX^2$$

The value of n , obtained from (39) for X^2 specified, is the minimum sample size which will establish rejection of the Null Hypothesis, at the 95 per cent significance level. Stated otherwise, the value of n obtained from (39) is the critical size of the sample we are testing, below which size we have not established significant rejection of the hypothesis that the observed X^2 is untypical.

Also following Geary's Examples 2 and 3 [1] we take the 0.95 probability level for the Normal single-variate distribution, 1.96, as equal to $r\sqrt{n}$. For r specified by analysis of a 2×2 table we solve the equation

$$(40) \quad 1.96 = r\sqrt{n}$$

for n , so as to establish the critical size of the sample we are testing for inherent

correlation. Table 5 following sets out the comparison of sample size via X^2 with that via r , for the four examples of Part 4.

TABLE 5: Comparison of critical size of sample via X^2 with that via r for the four examples of Part 4 (critical confidence level at 95 per cent)

Item	Example 1	Example 2	Example 3	Example 4
The value of X^2	0.007413	0.019196	0.00010536	0.020487
$n_1 = 3.841/X^2$	518.2	200.1	36,456	187.5
r	0.144	-0.344	-0.017	0.354
r^2	0.020736	0.118336	0.000289	0.125316
$n_2 = (1.96)^2/r^2$	185.3	32.5	13,293	30.7
n_1/n_2	2.80	6.16	2.74	6.11

The results require little comment. The critical sample size estimated by the X^2 approach is roughly three times to six times as large as that which emerges from the correlation approach. The latter parameter r is some three to six times as efficient as X^2 for establishing significant relationship. This efficiency evidently increases with the magnitude of r . A reasonable question at this point is to what causes may be ascribed the undoubted superiority of r ? Only one cause will be tentatively suggested here: that the assumed inherent bivariate Normal distribution, used to estimate r , takes much fairer account of the four relative frequencies than does X^2 . The truth of this assertion we can partly see, along the following lines. For a given r , the Normal bivariate unit volume can be divided into four parts, by axes through any arbitrary point (h, k) , in an unlimited number of ways. The same value of r is estimable from any such set of four parts, but a different X^2 is obtained for each point (h, k) in general, apart from effects of symmetry. Thus the general relationship between X^2 and r would appear to be arbitrary, apart from the central case for which, via (31) and (32), it can be readily shown that

$$(41) \quad r = \sin\left(\frac{\pi}{2}\sqrt{X^2}\right)$$

with X^2 calculated from the four relative frequencies and having a value of zero for r at zero level and a value of unity for $r = 1.0$. The negative square root can cater for negative values of r .

It is not surprising therefore that in the general non-central situation the lack of a precise relationship between X^2 and r causes a relative loss in efficiency of the former, this loss showing itself in the specification of a much larger size of sample required to establish significant relationship.

There are some implications, for 2×2 contingency tables. All four relative frequencies are of equal importance in establishing the estimate of r and it follows that careful attention must be devoted to the combination of "not treated" and "treated" which justifies the estimation of r . While the same justification is of course required for the data provided to the X^2 test, it is possibly less obvious than in the case of estimating r . It is not the purpose of the present essay to make a detailed examination of what constitutes a meaningful versus a false 2×2 contingency table. It is, however, apparent that for a case permitting valid estimation of the inherent r , a sample of one-third to one-sixth of that required by the X^2 test will establish a significant correlation. This outcome of the comparison of the two approaches to critical sample size, as amply shown by Geary in [1], favours the correlation estimate, for whatever conditions make it meaningful.

In view of the close agreement between the "nearest corner" Taylor interpolation estimate of r and that obtained via the somewhat elaborate tetrachoric approach, for the four numerical examples given above, we firmly recommend the Taylor interpolation method. This will undoubtedly give r correct to two significant figures. We suggest, however, that for the purpose of estimating $(\partial L / \partial r)_0$ via formula (18), the r -intervals of 0.05 on either side of the base corner be used, rather than those of 0.1, applied above for ease of exposition. All that this implies is that having chosen the corner (h_0, k_0, r_0) , we go back to the tables [2] and find $L(h_0, k_0, r_0 + 0.05)$ and $L(h_0, k_0, r_0 - 0.05)$ with $c = 0.05$, before substituting in the right hand side of formula (18) to estimate $\partial L / \partial r$ at r_0 . It is fairly obvious that if r_0 takes the extreme values -1.0 or $+1.0$ we can only use a single r -interval of 0.05 to measure the change in L for h and k constant, this observable change being divided by the r -interval to give $\partial L / \partial r$, instead of the general formula (18).

The assistance of a computer to estimate r can be visualised for either approach. For the cube interpolation the data might be the values of h , k , r and L at each of the eight corners, together with the values of h , k , and L derived from a 2×2 table and (for general formula (18)) the further set of eight L -values required to approximate $\partial L / \partial r$ at each of the corners. The computer would then interpolate for r from each corner and could select the estimate based on the nearest corner, among the computed estimates. To programme the tetrachoric method for a computer would undoubtedly be more difficult, partly because the central difference formulae given above could not be applied for either h_0 or k_0 having a value zero. In this situation forward difference formulae would be required. It thus would seem that the Taylor interpolation method is the more readily adaptable to the computer and a programme to estimate the inherent r might be useful.

REFERENCES

- [1] R. C. Geary, "Sensitivity of Tests of Significance: A Problem Posed". *The Economic and Social Review*, Volume 5, Number 3, Dublin, April 1974.
- [2] US Department of Commerce—"Tables of the Bivariate Normal Distribution Function and Related Functions". National Bureau of Standards, Applied Mathematics, Series 50, June 1959.
- [3] K. Pearson, editor—"Tables for Statisticians and Biometricians, Part II." Issued by the Biometric Laboratory, University College, London, 1931.
- [4] C. Lanczos—"Applied Analysis", Sir Isaac Pitman and Sons, Ltd., London, 1957.
- [5] D. B. Owen—"Handbook of Statistical Tables". Addison-Wesley Publishing Company Inc., Reading, Massachusetts, 1962.
- [6] E. S. Pearson and H. O. Hartley (eds.)—"Biometrika Tables for Statisticians, Volume 1. Cambridge University Press, 1958 edition.