ON THE PROBABLE ERROR OF A COEFFICIENT OF CORRELATION AS FOUND FROM A FOURFOLD TABLE.

By KARL PEARSON, F.R.S.

LET the fourfold table be

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

Then on the assumption that the frequency distribution is normal, we can by aid of Everitt's Tables of the Tetrachoric Functions* rapidly find r. I have shown in a paper published in the *Phil. Trans.* in 1900† that found in this way

Probable error of r

$$= \frac{\cdot 67449}{\sqrt{N}\chi_0}\left\{\frac{(a+d)(c+b)}{4N^2} + \psi_2{}^2\frac{(a+c)(d+b)}{N^2} + \psi_1{}^2\frac{(a+b)(d+c)}{N^2}\right.$$

$$\left. + 2\psi_1\psi_2\frac{ad-bc}{N^2} - \psi_2\frac{ab-cd}{N^2} - \psi_1\frac{ac-bd}{N^2}\right\}^{\frac{1}{2}} \quad \ldots \text{(i)},$$

where
$$\psi_1 = \frac{1}{\sqrt{2\pi}}\int_0^{\beta_1} e^{-\frac{1}{2}z^2}dz, \qquad \psi_2 = \frac{1}{\sqrt{2\pi}}\int_0^{\beta_2} e^{-\frac{1}{2}z^2}dz,$$

$$\beta_1 = \frac{h-rk}{\sqrt{1-r^2}}, \qquad \beta_2 = \frac{k-rh}{\sqrt{1-r^2}},$$

$$\chi_0 = \frac{1}{2\pi}\frac{1}{\sqrt{1-r^2}}e^{-\frac{1}{2}\frac{1}{1-r^2}(h^2+k^2-2rhk)},$$

* *Biometrika*, Vol. VII, p. 436, and Vol. VIII, p. 385.

† *Phil. Trans.* Vol. 195 A, p. 14. Owing to the carelessness of the printers my χ_0 was put as $\sqrt{\chi_0}$ and the last N^2 in the denominator as N_2.

and h and k have their usual meaning defined by the integrals

$$\frac{(a+c)-(b+d)}{2N}=\frac{1}{\sqrt{2\pi}}\int_0^h e^{-\frac{1}{2}z^2}dz=\tfrac{1}{2}\alpha_1, \text{ say};$$

$$\frac{(a+b)-(c+d)}{2N}=\frac{1}{\sqrt{2\pi}}\int_0^k e^{-\frac{1}{2}z^2}dz=\tfrac{1}{2}\alpha_2, \text{ say}.$$

Let $H=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}h^2}$, $K=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}k^2}$ as usual.

Now the formula (i) above for the probable error of r is admittedly laborious in use. I have tried in many ways, while retaining its full accuracy, to throw it into a form involving less laborious calculations; I have not succeeded, however, in achieving any sensible reduction in its complexity, as long as I maintain its complete generality.

Although many hundred fourfold tables have now been published, many of which give such small correlations that their true significance can only be settled by a knowledge of their probable errors, yet I find only 40 to 50 probable errors have so far been determined. This matter seems so regrettable that I have sought for a fairly easy method of determining a closely empirical expression for the probable error of r which is likely to be of service, and can be adapted easily to tables.

I consider first two extreme cases. If h and k are both zero, or the fourfold division at the mean, then $\psi_1=\psi_2=0$*,

Probable error of r

$$=\frac{\cdot 67449\; 2\pi\sqrt{1-r^2}}{\sqrt{N}}\left\{\frac{(a+d)(b+c)}{4N^2}\right\}^{\frac{1}{2}}=\frac{\cdot 67449\sqrt{1-r^2}}{\sqrt{N}}\frac{\pi}{2}\left\{\frac{16ab}{N^2}\right\}^{\frac{1}{2}},$$

since in this case $a=d$, and $b=c$.

But for a division at the mean by Sheppard's Theorem

$$r=\cos\pi\frac{b}{a+b}=\sin\left(\frac{\pi}{2}-\frac{\pi b}{a+b}\right),$$

or

$$(\sin^{-1}r)/\tfrac{1}{2}\pi=(a-b)/(a+b).$$

Hence

$$1-\left(\frac{\sin^{-1}r}{\frac{1}{2}\pi}\right)^2=\frac{4ab}{(a+b)^2}=\frac{16ab}{N^2}.$$

Substituting we have:

$$\text{Probable error of } r=\frac{\cdot 67449}{\sqrt{N}}\frac{\pi}{2}\sqrt{1-r^2}\sqrt{1-\left(\frac{\sin^{-1}r}{90^\circ}\right)^2}\quad\ldots\ldots\text{(ii)},$$

if the angle of the inverse sine be read in degrees.

Again if $r=0$, the probable error of r may be obtained from (i) whatever the values of h and k. For in this case

$$ad-bc=0,\quad \psi_1=\tfrac{1}{2}\alpha_1,\quad \psi_2=\tfrac{1}{2}\alpha_2,\quad \chi_0=HK.$$

* *Phil. Trans.* Vol. 192 A, p. 141 and Vol. 195 A, p. 7.

We have
$$(b+d)/N = \tfrac{1}{2}(1-\alpha_1), \quad (a+c)/N = \tfrac{1}{2}(1+\alpha_1),$$
$$(a+b)/N = \tfrac{1}{2}(1+\alpha_2), \quad (c+d)/N = \tfrac{1}{2}(1-\alpha_2),$$
$$\frac{a+d}{N} = \frac{(a+b)(a+c)}{N^2} + \frac{ad-bc}{N^2} + \frac{(c+d)(b+d)}{N^2} + \frac{ad-bc}{N^2}$$
$$= \tfrac{1}{4}(1+\alpha_2)(1+\alpha_1) + \tfrac{1}{4}(1-\alpha_2)(1-\alpha_1) = \tfrac{1}{2}(1+\alpha_1\alpha_2),$$
since $ad - bc = 0$ in the original population.

Similarly:
$$\frac{c+b}{N} = \tfrac{1}{2}(1-\alpha_1\alpha_2).$$
$$\frac{ab-cd}{N^2} = \frac{a(N-a-c-d)-cd}{N^2}$$
$$= \frac{a}{N} - \frac{(a+c)(a+d)}{N^2}$$
$$= \tfrac{1}{4}(1+\alpha_2)(1+\alpha_1) - \tfrac{1}{4}(1+\alpha_1)(1+\alpha_1\alpha_2)$$
$$= \tfrac{1}{4}\alpha_2(1-\alpha_1^2),$$
and similarly:
$$\frac{ac-bd}{N^2} = \tfrac{1}{4}\alpha_1(1-\alpha_2^2).$$

Hence substituting in (i)

$$\text{Probable error of } r = \frac{\cdot 67449}{\sqrt{N}HK}\{\tfrac{1}{16}(1-\alpha_1^2\alpha_2^2) + \tfrac{1}{16}\alpha_2^2(1-\alpha_1^2) + \tfrac{1}{16}\alpha_1^2(1-\alpha_2^2) - \tfrac{1}{8}\alpha_2^2(1-\alpha_1^2) - \tfrac{1}{8}\alpha_1^2(1-\alpha_2^2)\}^{\frac{1}{2}}$$
$$= \frac{\cdot 67449}{\sqrt{N}HK}\sqrt{\tfrac{1}{16}(1-\alpha_1^2)(1-\alpha_2^2)} \quad \text{.............................(iii).}$$

This can also be put in the form:

$$\text{Probable error of } r = \frac{\cdot 67449}{\sqrt{N}HK}\sqrt{\frac{(a+b)(a+c)(d+b)(d+c)}{N^4}} \quad \text{......(iv).}$$

This is the probable error of r of a fourfold table when the real value of r is zero.

Now as (ii) and (iv) give the reducing factors for the two cases (a) when h and k are both zero but r has any value and (b) when h and k have any values but r is zero, it occurred to me that the combined product of the two would give good results for a considerable range of values of h and k and r. We have to note that (iv) for h and k zero becomes

$$\frac{\cdot 67449}{\sqrt{N}}\frac{\pi}{2}.$$

Hence we take as our formula:

Probable error of r

$$= \frac{\cdot 67449}{\sqrt{N}}\sqrt{1-r^2}\sqrt{1-\left(\frac{\sin^{-1} r}{90^\circ}\right)^2}\frac{\sqrt{\frac{1}{2}(1+\alpha_1)\frac{1}{2}(1-\alpha_1)}}{H}\frac{\sqrt{\frac{1}{2}(1+\alpha_2)\frac{1}{2}(1-\alpha_2)}}{K} \quad \text{...(v).}$$

Now it will be seen that this consists of three parts:

(a) $\sqrt{1-r^2}\sqrt{1-\left(\frac{\sin^{-1} r}{90°}\right)^2}$. This is easy to table for all values of r.

(b) $\frac{\sqrt{\frac{1}{2}(1+\alpha_1)\frac{1}{2}(1-\alpha_1)}}{H}$, and

(c) $\frac{\sqrt{\frac{1}{2}(1+\alpha_2)\frac{1}{2}(1-\alpha_2)}}{K}$.

Both these (b) and (c) can be readily found from a single table rapidly formed from Sheppard's Table of the Probability Integral. The entry to the single table will be $(a+c)/N$ or $(a+b)/N$, i.e. $\frac{1}{2}(1+\alpha)$.

Thus a knowledge of the correlation r and the two division percentages (together with Miss Gibson's Table for $\cdot 67449/\sqrt{N}$), will enable us by the aid of the two new tables to rapidly write down four factors whose product gives the required probable error. I have tested the form (v) against the true probable error as found from (i). In all cases it gave results differing only from the true value at most by about one or two units in the third place of figures—a result amply accurate for all practical purposes.

Illustration I.

211·25	153·75	365
152·75	560·25	713
364	714	1078

The correlation was found to be $\cdot 5557 \pm \cdot 0261$; the probable error from the short formula was ·0265.

Illustration II.

1562	42	1604
383	94	477
1945	136	2081

The correlation was found to be $\cdot 5954 \pm \cdot 0272$; the probable error from the short formula was ·0293.

Illustration III.

455	622	1077
599	1324	1923
1054	1946	3000

The correlation was found to be $\cdot 1811 \pm \cdot 0210$; the probable error from the short formula was ·0199.

Illustration IV.

849	665	1514
205	1281	1486
1054	1946	3000

The correlation was found to be ·6633 ± ·0132; the probable error from the short formula was ·0132.

Illustration V.

1196	223	1419
318	1263	1581
1514	1486	3000

The correlation was found to be ·8464 ± ·0079; the probable error from the short formula was ·0079.

These examples will suffice, I think, to give confidence in the formula and in the tables accompanying this paper. The absence of probable errors from the expressions for fourfold table correlations can no longer be justified on the ground of their great laboriousness.

The following Tables have been calculated by Miss Julia Bell.

Let $\chi_1 = \cdot 67449/\sqrt{N}$. This is given by Miss Gibson's Tables, *Biometrika*, Vol. III. p. 387. Let

$$\chi_r = \sqrt{1-r^2}\sqrt{1-\left(\frac{\sin^{-1} r}{90^\circ}\right)^2},$$

and

$$\chi_a = \frac{1}{H}\sqrt{\tfrac{1}{2}(1+a)\times\tfrac{1}{2}(1-a)}.$$

Then

$$\text{Probable error of } r = \chi_1\cdot\chi_r\cdot\chi_{a_1}\cdot\chi_{a_2}.$$

TABLE I. *Values of χ_r for Values of r.*

r	χ_r	r	χ_r	r	χ_r	r	χ_r	r	χ_r
·00	1·0000	·20	·9717	·40	·8845	·60	·7298	·80	·4843
·01	·9999	·21	·9688	·41	·8785	·61	·7200	·81	·4687
·02	·9997	·22	·9657	·42	·8723	·62	·7099	·82	·4526
·03	·9994	·23	·9625	·43	·8659	·63	·6997	·83	·4362
·04	·9989	·24	·9591	·44	·8594	·64	·6892	·84	·4192
·05	·9982	·25	·9556	·45	·8527	·65	·6785	·85	·4018
·06	·9975	·26	·9520	·46	·8458	·66	·6675	·86	·3838
·07	·9966	·27	·9482	·47	·8388	·67	·6563	·87	·3652
·08	·9955	·28	·9442	·48	·8315	·68	·6448	·88	·3461
·09	·9943	·29	·9401	·49	·8241	·69	·6331	·89	·3262
·10	·9930	·30	·9358	·50	·8165	·70	·6211	·90	·3057
·11	·9915	·31	·9314	·51	·8087	·71	·6088	·91	·2843
·12	·9899	·32	·9268	·52	·8007	·72	·5962	·92	·2620
·13	·9881	·33	·9221	·53	·7926	·73	·5834	·93	·2387
·14	·9862	·34	·9172	·54	·7842	·74	·5702	·94	·2142
·15	·9841	·35	·9122	·55	·7756	·75	·5568	·95	·1882
·16	·9819	·36	·9070	·56	·7669	·76	·5430	·96	·1605
·17	·9796	·37	·9016	·57	·7579	·77	·5288	·97	·1305
·18	·9771	·38	·8961	·58	·7488	·78	·5144	·98	·0972
·19	·9745	·39	·8904	·59	·7394	·79	·4995	·99	·0585
								1·00	·0000

TABLE II.

Values of χ_α for Values of $\frac{1}{2}(1+\alpha)$.

$\frac{1}{2}(1+\alpha)$	χ_α	$\frac{1}{2}(1+\alpha)$	χ_α	$\frac{1}{2}(1+\alpha)$	χ_α	$\frac{1}{2}(1+\alpha)$	χ_α
·50	1·2533	·65	1·2877	·80	1·4288	·95	2·1132
·51	1·2535	·66	1·2928	·81	1·4457	·96	2·2740
·52	1·2539	·67	1·2984	·82	1·4641	·97	2·5071
·53	1·2546	·68	1·3044	·83	1·4844	·98	2·8915
·54	1·2556	·69	1·3109	·84	1·5067	·985	3·2097
·55	1·2569	·70	1·3180	·85	1·5315	·990	3·7333
·56	1·2585	·71	1·3256	·86	1·5590	·991	3·8854
·57	1·2604	·72	1·3338	·87	1·5897	·992	4·0639
·58	1·2626	·73	1·3427	·88	1·6245	·993	4·2784
·59	1·2652	·74	1·3523	·89	1·6640	·994	4·5419
·60	1·2680	·75	1·3626	·90	1·7094	·995	4·8779
·61	1·2712	·76	1·3738	·91	1·7623	·996	5·3278
·62	1·2748	·77	1·3859	·92	1·8249	·997	5·9776
·63	1·2787	·78	1·3990	·93	1·9003	·998	7·0465
·64	1·2830	·79	1·4133	·94	1·9937	·999	9·3870