Vol. LXXV.] [Part.VI.

JOURNAL

OF THE ROYAL STATISTICAL SOCIETY.

MAY, 1912.

On the METHODS of MEASURING ASSOCIATION BETWEEN TWO ATTRIBUTES.

By G. UDNY YULE.

[Read before the Royal Statistical Society, April 23, 1912, The Right Hon. Lord GEORGE HAMILTON, G.C.S.I., President, in the Chair.]

I. *Introduction.*

1. IF in any case of statistical observation we classify the objects or individuals observed into two classes only—*e.g.*, peas into yellow-seeded and green-seeded, or the members of any group of mankind into male and female—the resulting data are of the simplest possible statistical form. If, for each object or individual, we note two characters instead of one, dividing again into two classes only, the data become slightly more complex. We have four classes resulting from the two successive divisions—the class all the members of which possess both characters, the class all the members of which possess the first character but not the second, the class all the members of which possess the second character but not the first, and finally the class all the members of which possess neither of the two characters noted. The data resulting from any such count may, if space is no great consideration, be conveniently represented in the form of a small table such as the following (Macdonell, 10, Table II), which shows the recoveries and deaths amongst vaccinated and unvaccinated patients during the small-pox epidemic at Sheffield in 1887-88. There were 4,703 cases, of which 4,151 were vaccinated

TABLE I.—*Sheffield small-pox epidemic,* 1887-88*: cited from Macdonell* (10).

	Recoveries.	Deaths.	Total.
Vaccinated	3,951	200	4,151
Unvaccinated	278	274	552
Total	4,229	474	4,703

and 4,229 recovered from the disease. Of the vaccinated 3,951 recovered: of the unvaccinated 278.

2. The data given in any such table enable us to test the presence or absence of *association*, as I have termed it, between the two characters or attributes noted. If the two attributes are combined entirely independently, the proportion that possess, say, the first character will be the same, or more or less approximately the same, amongst those which possess and those which do not possess the second. If these two proportions differ, the two attributes are not independent but *associated*: positively associated if the proportion possessing the first character is greater amongst the objects or individuals possessing the second character than amongst those not possessing it, negative in the contrary case. Thus for the Sheffield epidemic we have:—

Proportion of the vaccinated who recovered	0·952
,, unvaccinated who recovered	0·504

There is thus a very marked positive association between vaccination and recovery. The association may, alternatively, be indicated by the proportions derived from the columns instead of the rows of the table:—

Proportion of the recoveries that were vaccinated........	0·934
,, fatal cases ,, ...	0·422

3. A table, such as Table I, may be represented, using the notation employed in my original memoir on association (32) which I have found very convenient, in the general form—

(AB)	$(A\beta)$	(A)
(αB)	$(\alpha\beta)$	(α)
(B)	(β)	N

Here N denotes the whole number of observations, (A) the number of A's, *i.e.*, objects or individuals exhibiting the character A, (α) the number of non-A's, *i.e.*, objects or individuals not exhibiting the character A, whilst (AB) denotes the number exhibiting both characters, and so on. The common test for association employed in the preceding paragraph consists in a comparison of the two proportions.

$$p_1 = (AB)/(B) \qquad p_2 = (A\beta)/(\beta) \qquad (1)$$

or, alternatively of the two proportions

$$p_3 = (AB)/(A) \qquad p_4 = (\alpha B)/(\alpha) \qquad (2)$$

i.e., in a comparison of the proportion of A's amongst the B's with the proportion of *A*'s amongst the non-*B*'s, or in a comparison of the proportion of *B*'s amongst the *A*'s with the proportion of *B*'s amongst the non-*A*'s.

4. Now, suppose that we have tables such as Table I for more than one small-pox epidemic and wish to compare the relative closeness of association between recovery and vaccination in the several cases, to obtain, as it were, some sort of measure of the relative efficacy of vaccination as a factor in recovery. If the numbers of recoveries and deaths of vaccinated and unvaccinated were the same in each case—or at least, in the same proportion to the whole number of observations—no difficulty would arise. If Table II gave figures relating to an actual epidemic no one would, I think, hesitate to say that it showed a markedly closer association between recovery and vaccination than the Sheffield data. So long as the totals of the rows and columns are kept the same, as in Tables I and II, either $p_1 - p_2$ or $p_3 - p_4$ is a fair indication of closeness of association. It is evident from (1) and (2) that if we increase (AB) we increase both $p_1 - p_2$ and $p_3 - p_4$. But matters in practice are not nearly so simple as this, for $(A)/N$ and $(B)/N$ may vary very considerably even in data drawn from similar fields. Table III shows, for example, the association between recovery and vaccination for small-pox patients in the Leicester epidemic, and Table IV for small-pox cases at the Homerton and Fulham hospitals (both cited again from Macdonell). The Leicester data show a relatively very large number of unvaccinated and few deaths, as compared with Sheffield; the Homerton and Fulham data a slightly larger proportion of unvaccinated, but many more deaths. If we

TABLE II.—*Same totals for rows and columns as Table I (imaginary data).*

	Recoveries.	Deaths.	Total.
Vaccinated	4,100	51	4,151
Unvaccinated..............	129	423	552
Total	4,229	474	4,703

TABLE III.—*Leicester small-pox epidemic*, 1892-93 (loc. cit. *Table III*).

	Recoveries.	Deaths.	Total.
Vaccinated............	197	2	199
Unvaccinated....	139	19	158
Total	336	21	357

TABLE IV.—*Small-pox cases at Homerton Hospital*, 1873-84, *and Fulham Hospital*, 1880-85*: doubtful cases excluded* (loc. cit. *Table VI*).

	Recoveries.	Deaths.	Total.
Vaccinated	8,207	692	8,899
Unvaccinated.................	1,424	1,103	2,527
Total	9,631	1,795	11,426

tabulate the proportions of recoveries amongst vaccinated and unvaccinated for Tables I, III and IV, we have :—

District or hospital.	Proportion of recoveries amongst		Difference.
	Vaccinated.	Unvaccinated.	
Sheffield	0·952	0·504	0·448
Leicester.......................................	0·990	0·880	0·110
Homerton and Fulham	0·922	0·564	0·358

What conclusion can we draw from these figures as to the comparative closeness of association between vaccination and recovery in the three cases ? If we go by the differences between the recovery-rates, given in the last column, Leicester must be placed well at the bottom of the list, as the difference is only 0·110 as compared with 0·358 and 0·448 in the two other cases. But is this a fair test ? At Leicester, as already pointed out, the epidemic was apparently a mild one, and even amongst the unvaccinated the recovery-rate was 0·880 : if every one of the vaccinated had recovered, without exception, the difference between the recovery-rates for vaccinated and unvaccinated could not have exceeded 0·120. If we pass this difficulty there is still a further one, namely, that a comparison of the differences $p_1 - p_2$ will no longer give necessarily the same results as a comparison of the differences $p_3 - p_4$—*i.e.*, the tables may not stand in the same order when ranked by the first difference as when ranked by the second. Thus tabulating for the same three tables the proportions of vaccinated amongst recoveries and deaths, we have :—

District or hospital.	Proportion of vaccinated amongst		Difference.
	Recoveries.	Fatal cases.	
Sheffield	0·934	0·422	0·512
Leicester	0·586	0·095	0·491
Homerton and Fulham	0·852	0·386	0·466

If we take the differences between these proportions as the index to degree of association, Leicester, it will be seen, stands above Homerton and Fulham. The comparison of proportion p_3 and p_4 leads therefore to different results from the comparison of p_1 with p_2.

5. But there is another simple quantity that might reasonably be taken as an index of association, viz., the difference between the actual value of the frequency (AB) and the value $(A)(B)/N$ which it would have in the case of independence. This difference,

$$\delta = (AB) - \frac{(A)(B)}{N} \qquad (3)$$

or the ratio δ/N, is the same in magnitude for every compartment of the fourfold table, but positive in sign (if the association is positive) for (AB) and $(\alpha\beta)$, negative for $(A\beta)$ and (αB). Evidently if (A) and (B) are fixed, as well as the whole number of observations, δ increases as (AB) increases, that is as $p_1 - p_2$ or $p_3 - p_4$ increases, so that in this simple case all the suggested indices agree. But again difficulties arise when the $(A)/N$ and $(B)/N$ vary from one case to another. If we work out the values of δ/N for Tables I, III, and IV we obtain the following figures:—

District or hospital.	Value of δ/N.
Sheffield	0·046
Leicester	0·027
Homerton and Fulham	0·062

Thus according to this index Homerton and Fulham should stand at the head of the list, Sheffield next, and Leicester lowest: the three different indices that we have tried have placed the districts in three different orders.

6. These illustrations suffice to show—and it is the sole purpose for which they have been given—that the choice of a measure or index of association is not quite a simple and straightforward matter: that the fundamental quantities which would serve quite well if all tables showed the same ratios for $(A)/N$ and $(B)/N$ give conflicting results when this condition fails to hold, as it invariably fails in practice, and that consequently a useful purpose may be served by an index or "coefficient of association" of somewhat more complex form.

7. Any such coefficient should obviously be zero when the attributes A and B are independent, *i.e.*, when the table takes the form:—

$(A)(B)/N$	$(A)(\beta)/N$	(A)
$(\alpha)(B)/N$	$(\alpha)(\beta)/N$	(α)
(B)	(β)	(N)

It is only for such a table that the association can be said to be zero. Further, it will be convenient if the coefficient can only range between the values ± 1. But the question arises when it should take these limiting values. If we suppose (AB) steadily to increase from the independence-value $(A)(B)/N$, the greatest value it can take is either (A) or (B), whichever is the less. If (A) is the less we reach the limiting case in which $(A\beta)$ is zero, or all A's are B. Similarly, if (B) is the less we will reach the limiting case in which (αB) is zero, or all B's are A. These are the cases of the first two tables below. If, but only if, $(A) = (B)$, and the table is diagonally symmetrical, can we have simultaneously $(\mathrm{A}\beta) = (\alpha B) = 0$, as in the third form of the table. The question is whether it will be more convenient for our coefficient to take the limiting value + 1 in each of these cases, or in the third alone.

(1)			(2)			(3)		
(AB) (αB)	0 $(\alpha\beta)$	(A) (α)	(AB) 0	$(A\beta)$ $(\alpha\beta)$	(A) (α)	(AB) 0	0 $(\alpha\beta)$	(A) (α)
(B)	(β)	(N)	(B)	(β)	N	(B)	(β)	N

Similarly if the association be negative there are again three limiting cases as in the three tables below, and the coefficient may be so formed as to take the value − 1 either in the third case only or in all three. Naturally, if the form chosen is such that the coefficient only becomes + 1 when $(A\beta)$ and (αB) both vanish, it should only take the value − 1 when both (AB) and $(\alpha\beta)$ vanish, but otherwise a coefficient of either type is quite legitimate. As a fact, coefficients of both types have been proposed and used. The only other condition which can be laid down for such a coefficient, from the general standpoint, is that for constant values of $(A)/N$ and $(B)/N$ it should increase as δ, $p_1 - p_2$ or $p_3 - p_4$ increases, decrease as they decrease. Any coefficient which fulfils the conditions stated is a legitimate coefficient of association. The choice between one form and another must depend on such factors

(1)			(2)			(3)		
0 (αB)	$(A\beta)$ $(\alpha\beta)$	(A) (α)	(AB) (αB)	$(A\beta)$ 0	(A) (α)	0 (αB)	$(A\beta)$ 0	(A) (α)
(B)	(β)	N	(B)	(β)	N	(B)	(β)	0

as simplicity of calculation, simplicity of meaning and tractability by algebraic methods. It is the purpose of the present paper to discuss some of the principal forms that have been used, and to add yet one more to their number, a modification of a coefficient which I proposed some twelve years ago.

8. Before proceeding to the special discussions on the several forms, it may be convenient briefly to state here for reference some important relations between δ, the excess of (AB) above its independence value, the frequencies and the chances or proportions p of equations (1) and (2). We have first (Yule 32: the relation is due to Pearson)—

$$\delta = \frac{1}{N}\left\{(AB)(\alpha\beta) - (A\beta)(\alpha B)\right\} \qquad (4)$$

an equation which expresses δ in terms of (AB), $(\alpha\beta)$, &c., the frequencies of the second order, as I term them, without reference to the row and column totals (A) and (B). It is evident that $(AB)(\alpha\beta)$ should be equal to $(A\beta)(\alpha B)$ if the attributes are independent, for either of these products is then equal to $(A)(\alpha)(B)(\beta)/N$. Further, we have, with a little algebra from equations (1), (2), and (3), the relations between δ and the p's.

$$\left.\begin{aligned} p_1 - p_2 &= \frac{N\,.\,\delta}{(B)(\beta)} \\ p_3 - p_4 &= \frac{N\,.\,\delta}{(A)(\alpha)} \end{aligned}\right\} \qquad (5)$$

Finally it may be noted that the greatest possible value of δ is either $(A)(\beta)/N$ or $(\alpha)(B)/N$, according as (A) is less or greater than (B): the greatest possible negative value on the other hand is either $(A)(B)/N$ or $(\alpha)(\beta)/N$ according as (A) is less or greater than (β). The greatest possible negative value is therefore not in general the same as the greatest possible positive value unless

$$(A) = (B) = (\alpha) = (\beta).$$

II. *Coefficients unaffected by selection of either attribute.*

9. The coefficient which I propose to consider first is that given in my memoir of 1900 (32), viz.:—

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \qquad (6)$$

The expression was not derived by any extraneous considerations, but was simply written down as an empirical formula fulfilling the required conditions:—(1) $Q = 0$ when the attributes are independent, for then the numerator, which from (4) is equal to $N\delta$, is zero; (2) when $(A\beta) = 0$ or $(\alpha B) = 0$, or both, $Q = +1$; (3) when $(AB) = 0$ or $(\alpha\beta) = 0$, or both, $Q = -1$. The coefficient is, therefore, one of the first class which takes the limiting value ± 1 in each of the three limiting cases of positive or of negative association (of § 7). It is not at once obvious that it fulfils the fourth essential condition, viz., that it should increase continuously with δ, or decrease continuously with δ, if N and the row and column totals (A) and (B) are kept constant. But if we write (as in 32, p. 273)

$$\kappa = \frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)} = \frac{[(A\beta)_0 - \delta]\,[(\alpha B)_0 - \delta]}{[(AB)_0 + \delta]\,[(\alpha\beta)_0 + \delta]} \qquad (7)$$

where $(AB)_0$ is the value taken by (AB) in the case of independence and so on—values that are constant for constant values of N, (A) and (B)—then

$$Q = \frac{1 - \kappa}{1 + \kappa} \qquad (8)$$

Now $dQ/d\kappa$ is evidently negative, and also $d\kappa/d\delta$: hence $dQ/d\delta$ is positive as required.

10. The coefficient Q (I took the symbol from the initial letter of Quetelet) was, I believe, the first expression to be given as a "coefficient of association" in the sense of the introductory section of this paper. The same formula was again suggested at a later date by Lipps (7, 1905). It has the merit of possessing extreme simplicity of form and consequent rapidity of calculation, but the demerit of not possessing an equal simplicity of interpretation. The numerator, as already pointed out, is N times the difference δ: but to the denominator, so far as I have found, no similar readily understood meaning can be attached. Further, while the coefficient can be interpreted in an extremely interesting sense not given in my original memoir, the work which leads to this interpretation leads also to another coefficient of similar properties, of very little greater complexity of form and much greater simplicity of meaning.

11. The starting point of the train of thought which gave rise to these ideas was a property of Q referred to in my memoir in connection with one of the numerical illustrations (p. 289), but deserving of more stress than is there laid on it. If, namely, we divide through the numerator and denominator of Q by $(B)(\beta)$ in the first place, and by $(A)(\alpha)$ in the second, we have the two expressions—

$$\left.\begin{aligned} Q &= \frac{p_1(1 - p_2) - p_2(1 - p_1)}{p_1(1 - p_2) + p_2(1 - p_1)} \\ &= \frac{p_3(1 - p_4) - p_4(1 - p_3)}{p_3(1 - p_4) + p_4(1 - p_3)} \end{aligned}\right\} \qquad (9)$$

and we have similarly from (7)

$$\kappa = \frac{p_2(1 - p_1)}{p_1(1 - p_2)} = \frac{p_4(1 - p_3)}{p_3(1 - p_4)} \qquad (10)$$

We can therefore express κ, or Q, in terms of p_1 and p_2 alone, or in terms of p_3 and p_4 alone: Q and κ are the same functions of p_1 and p_2 as they are of p_3 and p_4. It was pointed out in § 4 that we could not well use, *e.g.*, the simple difference between p_1 and p_2 or between p_3 and p_4 as an index to association, for the two differences in question would not merely, in general, give different numerical values, but would also give conflicting results. We have now arrived at a function which is the same for p_1 and p_2 as for p_3 and p_4, and does not lead to this apparent contradiction. Thus, using the data of Table III, and the figures already given in § 4 for the p's, we have for Leicester :—

$$\begin{aligned} Q &= \frac{0{\cdot}990 \times 0{\cdot}120 - 0{\cdot}880 \times 0{\cdot}010}{0{\cdot}990 \times 0{\cdot}120 + 0{\cdot}880 \times 0{\cdot}010} \\ &= \frac{0{\cdot}586 \times 0{\cdot}905 - 0{\cdot}095 \times 0{\cdot}414}{0{\cdot}586 \times 0{\cdot}905 + 0{\cdot}095 \times 0{\cdot}414} \\ &= 0{\cdot}862 \end{aligned}$$

12. Now, referring to the general association table in the form given in § 3, it will be seen that p_1 and p_2 are unaltered by multiplying or dividing either or both of the columns of the table by any arbitrary factor ; obviously, we may double or treble the number of B's relatively to the number of β's without thereby affecting $(AB)/(B)$ or $(A\beta)/(\beta)$. The association coefficient Q is, therefore, as it can be expressed in terms of p_1 and p_2 alone, unaltered by this proceeding, *although p_3 and p_4 have been changed* ; the change in p_3 and p_4 consequent upon an alteration in the relative proportion of B's and β's is not such as to affect Q. Regarding an alteration in the proportion of A's or B's as an operation of "selection"—the selection of more or fewer A's or B's—Q is, as I have expressed it in the heading of this section of the paper, a coefficient unaffected by the selection of either attribute.

13. This, as it seems to me, is a most important property, and one of special importance in such cases as those I have chosen for illustrations, in which the proportion of one of the attributes at least—viz., the number of vaccinated—is dependent to a large degree on a purely arbitrary circumstance, the activity of the authorities in enforcing the vaccination laws. Consider a similar case, that of a special hospital experimenting with a new remedy. The remedy would probably be tried at first on a relatively small proportion of the cases, then if successful on more, ultimately on all or nearly all. Suppose further that great care had been taken throughout not to select the cases for treatment on any special grounds, so that they might fairly be regarded as typical—as a random selection from all cases—and that the fatality rates for the treated and non-treated cases respectively had remained constant throughout. It seems to me that it would be at least a great convenience, not to express it in stronger terms, to use a coefficient of association which, in such circumstances, would remain constant, unaffected by the unessential fact of an alteration in the proportion of cases treated. If you told any man of ordinary intelligence that the association between

treatment and recovery was low at the beginning of the experiment, reached a maximum when 50 per cent. of the cases were treated and then fell off again as the proportion of cases treated was further increased, he would, I think, be legitimately puzzled, and would require a good deal of explanation as to what you meant by association. Yet this is the way in which the measure of association that has hitherto been most largely used would behave (Professor Pearson's normal coefficient, cf. below Section IV), as well as the less used product-sum correlation (Section III). The association coefficient Q keeps the same value throughout, quite unaffected by the ratio of cases treated to cases untreated. In all such instances the use of the coefficient Q, or of a coefficient possessing similar properties, seems to me to be clearly indicated. The values of Q in Tables I, III, and IV are 0·90, 0·86, and 0·80 respectively.

14. Two association tables that are not directly comparable owing to the different proportions of A's and B's in the data from which the tables were compiled may be rendered directly comparable by multiplying the frequencies in rows and columns by appropriate factors, and thus the values of Q may be simply illustrated by reducing the original tables to some arbitrarily selected standard form. Thus Tables I, III, and IV are, as we have seen, and as is more clearly evident from the upper section of Table V in which the data are reduced to the proportions per 10,000 observations,* incomparable with one another in any simple way owing to the relatively low percentage of vaccinated at Leicester and the high percentage at Sheffield, combined with the relatively high percentage of recoveries at Leicester and the low percentages at Sheffield and at the Homerton and Fulham hospitals. Let us reduce these percentages to the same values so as to obtain comparability. The only question is what values we shall take as the standard: I have taken first the percentages of the Sheffield data, as the table for Sheffield stands first on the list. To determine the reduction-factors, consider that, if we multiply the top row of, say, the Leicester table by a factor x and the left-hand column by a factor y, the table takes the form—

$xy(AB)$ $y(\alpha B)$	$x(A\beta)$ $(\alpha\beta)$	$xy(AB) + x(A\beta)$ $y(\alpha B) + (\alpha\beta)$
$xy(AB) + y(\alpha B)$	$x(A\beta) + (\alpha\beta)$	—

* This means retaining more figures than the Leicester data (357 cases) are worth. It was convenient, however, for subsequent work to use 10,000 rather than 1,000 as the basis. Using these tables solely as illustrations of general principle, I am not concerned with the possibilities of errors of sampling or of fact.

TABLE V.—*The data of Tables I, III and IV;* (a) *reduced to proportions per* 10,000 *observations;* (b) *reduced to proportions of vaccinated and recoveries in Sheffield data per* 10,000 *observations;* (c) *reduced to* 50 *per cent. of vaccinated and* 50 *per cent. of recoveries per* 10,000 *observations.*

Row and column totals		Sheffield.			Leicester.			Homerton and Fulham.		
		Recoveries.	Deaths.	Total.	Recoveries.	Deaths.	Total.	Recoveries.	Deaths.	Total.
(*a*) Of original data	Vaccinated	8,401	425	*8,826*	5,518	56	*5,574*	7,183	606	*7,789*
	Unvaccinated	591	583	*1,174*	3,894	532	*4,426*	1,246	965	*2,211*
	Total	*8,992*	*1,008*	*10,000*	*9,412*	*588*	*10,000*	*8,429*	*1,571*	*10,000*
(*b*) Reduced to Sheffield form	Vaccinated	8,401	425	*8,826*	8,336	490	*8,826*	8,268	559	*8,827*
	Unvaccinated	591	583	*1,174*	655	519	*1,174*	724	450	*1,174*
	Total	*8,992*	*1,008*	*10,000*	*8,991*	*1,009*	*10,000*	*8,992*	*1,009*	*10,001*
(*c*) Reduced to symmetrical form	Vaccinated	4,076	924	*5,000*	3,929	1,071	*5,000*	3,760	1,240	*5,000*
	Unvaccinated	924	4,076	*5,000*	1,071	3,929	*5,000*	1,240	3,760	*5,000*
	Total	*5,000*	*5,000*	*10,000*	*5,000*	*5,000*	*10,000*	*5,000*	*5,000*	*10,000*

If the ratios of A's to α's and of B's to β's in the standard (Sheffield) table are m and n respectively, we have accordingly two equations for x and y,

$$\left.\begin{aligned} xy(AB) + x(A\beta) &= m\{y(\alpha B) + (\alpha\beta)\} \\ xy(AB) + y(\alpha B) &= n\{x(A\beta) + (\alpha\beta)\} \end{aligned}\right\} \qquad (11)$$

These equations lead to a quadratic, and give for the Leicester table the values $x = 8\cdot98$, $y = 0\cdot17$ as the factors necessary to reduce it to the Sheffield form. For the Homerton and Fulham table the factors found are $x = 1\cdot98$, $y = 1\cdot25$. Multiplications by these factors alters, of course, the total number of observations: dividing through by this new total in each case, we have the tables in the second section of Table V. These tables are directly comparable by eye, and show at once what the coefficient of association Q means when it tells you that the association is highest at Sheffield, next highest at Leicester, and lowest in the Homerton and Fulham cases. Had the percentages of vaccinations and of recoveries been the same at Leicester, and at the Homerton and Fulham hospitals as at Sheffield, of the 8,826 vaccinated 8,336 would have recovered at Leicester, and 8,268 at the London hospitals, as against 8,401 at Sheffield.

15. There is no sufficient reason for selecting Sheffield, however, as the standard form: if Mr. Macdonell, from whose memoir (10) I have cited these tables, had placed them in a different order, I might have taken Leicester or Homerton as the standard instead. There is, in fact, only one form which can be regarded as at all a natural standard form, and that is the symmetrical table in which $(A) = (B) = (\alpha) = (\beta) = N/2$. If we write $m = n = 1$ in equations (11) above, we have as the reduction factors for this case

$$x = \sqrt{\frac{(\alpha B)(\alpha\beta)}{(AB)(A\beta)}}$$

$$y = \sqrt{\frac{(A\beta)(\alpha\beta)}{(AB)(\alpha B)}}$$

or the equivalent symmetrical table is of the form—

$(\alpha\beta)$	$\sqrt{(A\beta)\,(\alpha B)\,(\alpha\beta)/(AB)}$
$\sqrt{(A\beta)\,(\alpha B)\,(\alpha\beta)/(AB)}$	$(\alpha\beta)$

Using the factors given, we obtain the equivalent symmetrical tables for Sheffield, Leicester, and Homerton and Fulham shown in the last section of Table V. These tables (c) are more perspicuous than the last, and give us the following proportions for the recoveries amongst vaccinated and unvaccinated:—

District or hospital.	Proportion of recoveries for the symmetrical table amongst		Difference ω.
	Vaccinated.	Unvaccinated.	
Sheffield	0·815	0·185	0·630
Leicester	0·786	0·214	0·572
Homerton and Fulham	0·752	0·248	0·504

The association-tables being symmetrical, the above table may also be read as giving the proportions of vaccinated amongst recoveries and deaths respectively. These are, of course, not the actual proportions, but the proportions that would have resulted if an omnipotent demon of unpleasant character (no relation of Maxwell's friend) could have visited Sheffield, Leicester, and the two London hospitals, and raised the fatality rate and the proportion of unvaccinated at each to 50 per cent. without otherwise altering the facts.

16. In the symmetrical table p_1 becomes identical with p_3 and p_2 with p_4; further, $p_2 = 1 - p_1$, $p_3 = 1 - p_4$, and conveniently we may denote the proportions in the symmetrical table by

$$\left.\begin{aligned} p_0 &= \frac{(AB)}{(B)} = \frac{(AB)}{(A)} = \frac{2\,(AB)}{N} \\ q_0 &= \frac{(A\beta)}{(\beta)} = \frac{(\alpha B)}{(\alpha)} = \frac{2\,(A\beta)}{N} \end{aligned}\right\} \qquad (12)$$

Let

$$\omega = p_0 - q_0 \qquad (13)$$

Then Q must be expressible in terms of the quantity ω. Writing p_0 and q_0 in terms of the frequencies of the original table we have in fact,

$$\left.\begin{aligned} p_0 &= \frac{1}{1 + \sqrt{\kappa}} \\ q_0 &= \frac{\sqrt{\kappa}}{1 + \sqrt{\kappa}} \end{aligned}\right\} \qquad (14)$$

whence

$$Q = \frac{1 - \kappa}{1 + \kappa} = \frac{2\omega}{1 + \omega^2} \qquad (15)$$

The association coefficient Q for any table is therefore a simple function of the difference between the proportions p_0 and q_0 in the equivalent symmetrical table: to have obtained this significance for it seems to me to be a real gain. The following table shows equivalent values of p_0, q_0, ω and Q. For small values of Q the value of ω is little more than $\frac{1}{2}Q$, but the ratio of ω to Q gradually rises.

p_0.	q_0.	ω.	Q.	p_0.	q_0.	ω.	Q.
0·50	0·50	0·0	0·000	0·80	0·20	0·6	0·882
0·55	0·45	0·1	0·198	0·85	0·15	0·7	0·940
0·60	0·40	0·2	0·385	0·90	0·10	0·8	0·976
0·65	0·35	0·3	0·550	0·95	0·05	0·9	0·994
0·70	0·30	0·4	0·690	1·00	0·00	1·0	1·000
0·75	0·25	0·5	0·800	—	—	—	—

The formula (15) may readily be verified from the values of ω given in § 15. Thus for Sheffield $\omega = 0·630$, and we have therefore

$$Q = \frac{2 \times 0·630}{1 + (0·630)^2} = 0·90$$

agreeing with the result of the direct calculation given at the end of § 13.

17. But the work of the last two paragraphs inevitably suggests the question, why not use ω itself as the coefficient of association, instead of the function Q given by equation (15)? We have from (14) and (15)

$$\omega = \frac{1 - \sqrt{\kappa}}{1 + \sqrt{\kappa}} \qquad (16)$$

Evidently ω, a function of κ which only differs from Q in the fact that κ itself has been replaced by $\sqrt{\kappa}$ (cf. equation 8), possesses just the same *general* properties as Q, *i.e.*, it is zero when the attributes are independent; $+1$ when $(A\beta) = 0$ or $(\alpha B) = 0$ or both; -1 when $(AB) = 0$ or $(\alpha\beta) = 0$ or both; and for the same values of (A), (B), and N it increases continuously as δ increases, decreases as δ decreases.* But, in addition to these essential properties, ω possesses also an extremely simple meaning, viz., that it is equal to the difference between the proportions $(AB)/(B)$ and $(A\beta)/(\beta)$ in the equivalent symmetrical table. For this reason alone I should be inclined to prefer ω to Q for any future work: but its employment is seen from the next section (Section III, § 22) to present an additional advantage, namely, that ω is the coefficient of correlation (the product-sum coefficient) for the equivalent symmetrical table. To this point I recur again (§§ 22—24). In the meantime it is sufficient to point out that ω in unsymmetrical tables always gives numerical values less than Q, as is seen at once

* It may be pointed out that $Q_3 = \sin\frac{\pi}{2}\omega$ is given by Professor Pearson in his memoir (15) as an approximation to his coefficient—the normal coefficient as I term it: but the two possess fundamentally different properties.

from equation (15): for the three cases of Tables I, III and IV, the values have already been given, viz.:—

	Q.	ω.
Sheffield	0·90	0·63
Leicester	0·86	0·57
Homerton and Fulham	0·80	0·50

Q and ω always give, however, results that are consistent with one another, *i.e.*, the same results as regards the order in respect of association in which any series of tables is placed. They are coefficients of precisely the same kind, so that if in two different tables $Q_1 > Q_2$, then $\omega_1 > \omega_2$. To calculate ω, work out first the value of κ from equation (7), take its square root from Barlow's tables, and then ω is given at once by equation (16). Thus for Sheffield

$$\kappa = 0{\cdot}051359$$

$$\sqrt{\kappa} = 0{\cdot}22663$$

$$\omega = \frac{0{\cdot}77337}{1{\cdot}22663} = 0{\cdot}630$$

It is convenient to have a name as well as a symbol for a given coefficient, and I suggest that ω may be termed the *coefficient of colligation.*

18. The standard error of the coefficient Q was given in my memoir on association (32, p. 285): it is

$$\epsilon_q = \frac{1 - Q^2}{2}\sqrt{\frac{1}{(AB)} + \frac{1}{(A\beta)} + \frac{1}{(\alpha B)} + \frac{1}{(\alpha\beta)}} \qquad (17)$$

This expression apparently becomes infinite if one of the frequencies (AB), &c., vanishes: but expressed entirely in terms of the frequencies it takes the form

$$\epsilon_q^2 = \frac{16\,(AB)(A\beta)(\alpha B)(\alpha\beta)}{\{(AB)(\alpha\beta) + (A\beta)(\alpha B)\}^4} \times$$

$$\Big\{(A\beta)(\alpha B)(\alpha\beta) + (AB)(\alpha B)(\alpha\beta) + (AB)(A\beta)(\alpha\beta) + (AB)(A\beta)\,(\alpha B)\Big\}$$

and this vanishes if any one of the frequencies (AB), $(A\beta)$, (αB), $(\alpha\beta)$ becomes zero. Hence the standard error of the coefficient is zero when the coefficient itself is ± 1. The standard error of ω is found by following precisely the same steps, the starting point being the standard error of κ

$$\epsilon_\kappa = \kappa\sqrt{\frac{1}{(AB)} + \frac{1}{(A\beta)} + \frac{1}{(\alpha B)} + \frac{1}{(\alpha\beta)}} \qquad (18)$$

I find,

$$\epsilon_\omega = \frac{1 - \omega^2}{4}\sqrt{\frac{1}{(AB)} + \frac{1}{(A\beta)} + \frac{1}{(\alpha B)} + \frac{1}{(\alpha\beta)}} \qquad (19)$$

If the distribution be completely symmetrical, *i.e.*, if $(A)=(B)=N/2$, the expression admits of an interesting simplification. For in the case of the symmetrical table

$$(AB) = (\alpha\beta) = (1 + \omega)\frac{N}{4}$$

$$(A\beta) = (\alpha B) = (1 - \omega)\frac{N}{4}$$

and hence, for the case of complete symmetry,

$$\epsilon^2_\omega = \frac{1 - \omega^2}{N} \tag{20}$$

For one and the same distribution the ratio of the standard error of ω to that of Q is given by

$$\frac{\epsilon_\omega}{\epsilon_q} = \frac{1 - \omega^2}{2(1 - Q^2)} = \frac{(1 + \omega^2)^2}{2(1 - \omega^2)}$$

This is equal to 0·5 if ω is zero, becomes unity for $\omega = 0\cdot4859$, and rises to infinity as ω approaches unity. It must be remembered, however, that this represents the ratio of two quantities that are indefinitely small. If, instead of the ratio of the standard errors we take the ratio of ϵ_ω/ω to ϵ_q/Q, we have

$$\frac{\epsilon_\omega \, . \, Q}{\epsilon_q \, . \, \omega} = \frac{1 + \omega^2}{1 - \omega^2}$$

This ratio is unity when ω is zero and rises steadily as ω increases. The following table shows as an illustration the standard errors of ω and Q for a symmetrical table of 1,000 observations—the form of table which gives the lowest possible standard error for a given value of ω and a given number of observations:—

Standard errors of ω *and* Q *for symmetrical table of* 1,000 *observations.*

Value of ω.	Standard error of ω.	Corresponding value of Q.	Standard error of Q.
0·0	0·0316	0·0	0·0632
0·1	0·0315	0·198	0·0611
0·2	0·0310	0·385	0·0549
0·3	0·0302	0·550	0·0462
0·4	0·0290	0·690	0·0362
0·5	0·0274	0·800	0·0263
0·6	0·0253	0·882	0·0175
0·7	0·0226	0·940	0·0104
0·8	0·0190	0·976	0·0051
0·9	0·0138	0·994	0·0016

It will be noticed that the standard error of ω falls comparatively slowly as ω increases. For $\omega = 0$ it is 0·032, and only falls to 0·027 when ω rises to 0·5: for the corresponding range of values of Q, the standard error drops from 0·063 to 0·026. Q, although not possessed of so simple and readily interpretable a meaning as ω, is absolutely more stable against fluctuations of sampling than ω over the upper half of the range of ω from 0 to 1, and relatively more stable over the whole range.*

III. *The product sum correlation.*

19. The ordinary formulæ and methods applied to the treatment of correlated variables (excepting such formulæ as are valid only when the correlation is normal) involve no assumptions, explicit or implicit, as to the number of different values of each variable that can occur. If two values alone are possible, the usual many-rowed correlation table reduces to a table with two rows and two columns, such as Tables I—IV. This consideration at once suggests that such tables should be treated by the same familiar methods that are used in the case of variables, regarding the attribute as a variable that is only susceptible of taking one or other of two alternative magnitudes.

20. It is not, however, necessary to suppose that it is the attribute itself which can assume the two magnitudes in question: the variable may be regarded as the number of objects or individuals, in certain classes, which exhibit the attribute A or the attribute B, when the classes contain only a single member. Thus suppose—to use a hard-worked illustration—that we have a bag containing black and white balls, that we draw successively two batches each of ten balls, note the number of black balls in each batch, return them to the bag and draw another pair, and so on. We can then draw up a correlation table between the number of black balls in the first and second batches and calculate the correlation coefficient in the ordinary way. We may then repeat the experiment, reducing the number of balls in each batch from 10 to 9, 8, 7, 6 . . ultimately to 1. The same interpretation continues to hold good for the fourfold correlation table found in this limiting case, *i.e.*, it is a correlation table between "number of black balls in first batch" and "number of black balls in second batch," although these numbers can now only be 0 or 1. Similarly, a record of a single small-pox case must contribute to the general record either 0 or 1 vaccinated patients, either 0 or 1 recoveries. Generally, if we are sorting into A's and α's, B's and β's, the record of one case must show either

* I am glad to take this opportunity of correcting a serious error in the memoir on association (32). In equation (12), p. 286, delete the 2 in the denominator, and in the following tables double the "ratio" and the probable error of Q. Q should, in both cases, be much less stable than the correlation-coefficient for a normal distribution, not more stable.

0 A's or 1 A, 0 B's or 1 B. We may therefore regard our fourfold table as of the form :—

Second variable.	First variable.		Total.
	0	1	
0	$(\alpha\beta)$	$(A\beta)$	(β)
1	(αB)	(AB)	(B)
Total	(α)	(A)	N

21. Treating this precisely like any other correlation table, we have for the means the suffix 1 referring to the A's and the suffix 2 to the B's,

$$\left.\begin{aligned} M_1 &= (A)/N \\ M_2 &= (B)/N \end{aligned}\right\} \qquad (21)$$

For the standard-deviations we have

$$\left.\begin{aligned} \sigma_1{}^2 &= \frac{(A)}{N} - \left\{\frac{(A)}{N}\right\}^2 = \frac{(A)(\alpha)}{N^2} \\ \sigma_2{}^2 &= \frac{(B)}{N} - \left\{\frac{(B)}{N}\right\}^2 = \frac{(B)(\beta)}{N^2} \end{aligned}\right\} \qquad (22)$$

Finally

$$\begin{aligned} \Sigma\,(xy) &= (AB) - N\,.\,M_1\,.\,M_2 \\ &= \delta \end{aligned} \qquad (23)$$

whence

$$r = \frac{N\,.\,\delta}{\sqrt{(A)(\alpha)(B)(\beta)}} \qquad (24)$$

The regressions are given by

$$\left.\begin{aligned} r\frac{\sigma_1}{\sigma_2} &= \frac{N\,.\,\delta}{(B)(\beta)} = p_1 - p_2 \\ r\frac{\sigma_2}{\sigma_1} &= \frac{N\,.\,\delta}{(A)(\alpha)} = p_3 - p_4 \end{aligned}\right\} \qquad (25)$$

the last equalities being written down from equation (5). The expressions check by observing that the mean of the top row is p_2 and the mean of the lower row p_1. Similarly the mean of the left-hand column is p_4, and the mean of the right-hand column p_3. We may therefore write

$$r = \pm\sqrt{(p_1 - p_2)(p_3 - p_4)} \qquad (26)$$

where the minus sign is, as usual, to be taken if both the regressions are negative. Evidently r can only become equal to unity if both $p_1 - p_2$ and $p_3 - p_4$ are unity. But this can only be the case if *both* $(A\beta)$ and (αB), or in the case of negative correlation (AB) and $(\alpha\beta)$, are zero. The correlation coefficient, unlike the association coefficients Q and ω, can therefore only take the values ± 1 in the third cases of § 7.

22. If the table is symmetrical and $(A) = (\alpha) = (B) = (\beta) = N/2$, we have

$$r = p_0 - q_0 = 4\frac{\delta}{N} \tag{27}$$

But $p_0 - q_0$ is the value of ω. We have therefore the important theorem briefly mentioned without proof in §17—*the coefficient of colligation ω for any table is the product-sum correlation r for the equivalent symmetrical table.* These two coefficients r and ω form, accordingly, a natural pair, the first giving the actual correlation in the given table, the second the correlation in a derived table of standard form, thus enabling us to compare the two tables freed from the effects of "selecting" varying proportions of A's and B's.

23. If two tables are "equivalent," *i.e.*, are or may be derived from the same symmetrical table by selecting A's or B's, or both, that table has the lowest correlation for which the ratios of $(A)/(\alpha)$ and $(B)/(\beta)$ are greatest. Thus, suppose that in the general table of § 20 we multiply the lower row by an arbitrary factor x, say to fix our ideas greater than unity: the values of the p's for the rows are unaltered, but for the columns we have

$$p_3 = \frac{p_0 \,.\, x}{q_0 + p_0 \,.\, x}$$

$$p_4 = \frac{q_0 \,.\, x}{p_0 + q_0 \,.\, x}$$

$$p_3 - p_4 = \frac{x}{(q_0 + p_0x)(p_0 + q_0x)}(p_0 - q_0)$$

Hence r has been reduced in the ratio

$$1 : \sqrt{x/(q_0 + p_0x)(p_0 + q_0x)}$$

or

$$r = \omega\sqrt{\frac{x}{(q_0 + p_0x)(p_0 + q_0x)}} \tag{28}$$

Similarly, if we now multiply the right-hand column by an arbitrary factor y, say greater than unity again,

$$p_1 - p_2 = \frac{y}{(q_0 + p_0y)(p_0 + q_0y)}$$

And now

$$r = \omega\sqrt{\frac{xy}{(q_0 + p_0x)(p_0 + q_0x)(q_0 + p_0y)(p_0 + q_0y)}} \tag{29}$$

The value of xy is given by $(AB)/(\alpha\beta)$ and will serve as a rough sort of guide to the magnitude of effect to be expected: the

2 U 2

greater x and the greater y the greater the reduction of r as compared with ω. If x and y are not both greater than unity, but one is greater than unity and the other less $(\alpha B)/(A\beta)$ or its reciprocal, and not $(AB)/(\alpha\beta)$ is the ratio giving the rough indication of the amount of reduction in r to be expected. If, for example, $y = 1/x$ the change is the same as if $y = x$; (AB) and $(\alpha\beta)$ are equal but $(\alpha B)/(A\beta) = x^2$. In practice, therefore, if one wants to get some rough idea as to the probable difference between r and ω—the difference between the actual correlation and the correlation in the equivalent symmetrical table—one should look at both diagonal ratios and take the biggest as the guide.

24. The work may be illustrated again from the tables for Sheffield, Leicester, and the Homerton and Fulham hospitals. I find the following values of r for the three forms of these tables given in Table V. The figures in the last column are, of course, the values of ω from § 17 :—

District or hospital.	Product-sum correlation r for		
	Original form.	Sheffield form.	Symmetrical form.
Sheffield	0·48	0·48	0·63
Leicester....................................	0·23	0·41	0·57
Homerton and Fulham............	0·41	0·34	0 50

The values of $(AB)/(\alpha\beta)$ and $(\alpha B)/(A\beta)$ for the three original tables are as follows :—

	Ratio of	
	Vaccinated-recovered to non-vaccinated non-recovered.	Non-vaccinated recovered to vaccinated non-recovered.
Sheffield	14·4	1·4
Leicester..	10·4	69·5
Homerton and Fulham	7·4	2·1

It will be seen that in the case of Sheffield and the Homerton and Fulham hospitals the first ratio is the guiding one, in the case of Leicester the second. The fact is, that while in the first and third cases the numbers of vaccinated and of recoveries must both be reduced to bring the tables to symmetrical form, in the case of Leicester—paradoxical though it seems—we must largely increase the numbers of the vaccinated, *notwithstanding that they are in a majority already.* The factors found are 2·589 and 0·03725.

Multiplying the numbers of the vaccinated by the first factor, the figures of Table III become

	Recovery.	Death.
Vaccinated................	510·0	5·2
Unvaccinated............	139	19

Now, multiplying the recoveries by 0·03725, we have the symmetrical table

	Recovery.	Death.
Vaccinated................	19·0	5·2
Unvaccinated............	5·2	19

Hence the guiding ratios as to the influence of selection from the symmetrical form are 14·4, 69·5 and 7·4 respectively: corresponding to these ratios we find the Sheffield correlation raised from 0·48 to 0·63 when the table is rendered symmetrical, the Leicester coefficient from 0·23 to 0·57 (more than doubled), and the Homerton coefficient only from 0·41 to 0·50. Looking now at the coefficients for the Sheffield form we see that the value of the Leicester coefficient is raised, as we should expect, while the value of the Homerton coefficient is lowered. As a further illustration, I give in Table VI. three sets of derived tables calculated partly for this and partly for the sake of another comparison (§ 37 below). Thus in set A, the symmetrical table taken as the basis was first altered by multiplying through the row of *B*'s by 19, and then reducing again to a total of 10,000 observations. The table so derived (the second table in the column headed Set A) was further altered by multiplying the column of A's by 11, and thus the third table was obtained. For these three sets we have the following coefficients :—

	Product-sum correlations.		
	First table (symmetrical).	Second table.	Third table.
Set A	0·20	0·09	0·05
Set B	0·50	0·32	0·20
Set C	0·60	0·31	0·20

For sets A and C the multipliers are the same, and the effects of the same sort of magnitude: for set B the multipliers are lower (10 each time) and the effect consequently rather less.

25. This coefficient has been used a good deal of recent years in theoretical work (see the history in §§ 33–5 below), but comparatively

TABLE VI.—*Three sets of tables, the second and third in each set derived from the first by "selection." For set A*, $Q=0·38$, $\omega=0·20$; *for set B*, $Q=0·8$, $\omega=0·5$; *for set C*, $Q=0·88$, $\omega=0·6$.

Multipliers used for derivation.		Set *A*.			Set *B*.			Set *C*.		
		α.	*A*.	Total.	α.	*A*.	Total.	α.	*A*.	Total.
Symmetrical table	β	3,000	2,000	*5,000*	3,750	1,250	*5,000*	4,000	1,000	*5,000*
	B	2,000	3,000	*5,000*	1,250	3,750	*5,000*	1,000	4,000	*5,000*
	Total	5,000	5,000	*10,000*	5,000	5,000	*10,000*	5,000	5,000	*10,000*
Multiply *B* row by 19 in sets *A* and *C*, by 10 in set *B*........	β....	300	200	*500*	682	227	*909*	400	100	*500*
	B	3,800	5,700	*9,500*	2,273	6,818	*9,091*	1,900	7,600	*9,500*
	Total	4,100	5,900	*10,000*	2,955	7,045	*10,000*	2,300	7,700	*10,000*
Multiply *A* column by 11 in sets *A* and *C*, by 10 in set *B*........	β..................................	43	319	*362*	93	310	*403*	46	126	*172*
	B	551	9,087	*9,638*	310	9,288	*9,598*	218	9,609	*9,827*
	Total	*594*	*9,406*	*10,000*	*403*	*9,598*	*10,001*	*264*	*9,735*	*9,9 99*

little, curiously enough, in any practical arithmetical work, hence perhaps the fact that (so far as I am aware) its standard error has not been determined. Taking the coefficient in the form given by equation (24), I have taken the work in the following stages:—(1) The standard errors of the standard deviations $\sqrt{(A)\,(\alpha)}/N$ and $\sqrt{(B)\,(\beta)}/N$: (2) The correlation between errors in the two standard deviations: (3) The standard error of δ: (4) The correlation between errors in δ and errors in a standard deviation: (5) The standard error of r. I have used the following notation: for the correlation table dealt with the constants are r_{12}, σ_1, σ_2: for the errors of sampling ϵ_a denotes the standard error of the frequency (A), ϵ_{ab} of the frequency (AB) and so on, ϵ_δ of δ; $r_{(a)(b)}$ denotes the correlation between errors in (A) and (B), $r_{(a)\,(ab)}$ the correlation between errors in (A) and (AB). The following are known results (Yule 32, Pearson and Filon 14, Pearson and others 19):—

$$\epsilon_a{}^2 = \frac{(A)(\alpha)}{N} \tag{30}$$

and similarly

$$\epsilon_{ab}{}^2 = \frac{(AB)[N - (AB)]}{N} \tag{31}$$

As regards correlation of errors

$$r_{(a)(b)}\ \epsilon_a\ \epsilon_b = \delta \tag{32}$$

$$r_{(ab)(\alpha b)}\ .\ \epsilon_{ab}\ \epsilon_{\alpha b} = -\frac{(AB)(\alpha B)}{N} \tag{33}$$

$$r_{(a)(ab)}\ .\ \epsilon_a\ \epsilon_{ab} = \frac{(\alpha)(AB)}{N} \tag{34}$$

26. To find the standard error of

$$\sigma_1 = \sqrt{(A)(\alpha)}/N$$

we have

$$\frac{d\sigma_1}{\sigma_1} = \tfrac{1}{2}\frac{d\,(A)}{(A)} + \tfrac{1}{2}\frac{d\,(\alpha)}{(\alpha)}$$

$$= -\tfrac{1}{2}\,d\,(A)\frac{(A) - (\alpha)}{(A)(\alpha)}$$

since $d\,(\alpha) = -\,d\,(A)$. Therefore, squaring and summing

$$\frac{\epsilon^2_{\sigma_1}}{\sigma^2_1} = \tfrac{1}{4}\frac{[(A) - (\alpha)]^2}{N\,(A)(\alpha)}$$

$$\epsilon^2_{\sigma_1} = \frac{[(A) - (\alpha)]^2}{4\,N^3} = \frac{0{\cdot}25 - \sigma_1{}^2}{N} \tag{35}$$

This result looks paradoxical, as it is zero if $(A) = (\alpha)$. But it is of course, only true to terms of the order $1/N$, and alterations of order $1/N$ in (A) in a symmetrical table only produce alterations in

σ of the order $1/N^2$. Thus if the values of (A) and (α) are $\frac{N}{2} + a$ and $\frac{N}{2} - a$.

$$\begin{aligned}\sigma_1{}^2 &= \frac{1}{N^2}\left(\frac{N}{2} + a\right)\left(\frac{N}{2} - a\right)\\ &= \frac{1}{4}\left(1 - \frac{4a^2}{N^2}\right)\\ \sigma_1 &= \frac{1}{2}\left(1 - \frac{2a^2}{N^2} - \frac{2a^4}{N^4} - \frac{4a^6}{N^6} - \quad . \quad . \quad .\right)\end{aligned}$$

It may be noted here that the correlation between errors in σ_1 and errors in (A) is -1; this is readily deduced from the differential equation above.

27. From the remark just made it is obvious that the correlation between errors in σ_1 and σ_2 is the same as that between errors in (A) and (B), and is therefore given by (32). Proceeding algebraically

$$\frac{d\sigma_1}{\sigma_1} \cdot \frac{d\sigma_2}{\sigma_2} = \frac{1}{4}\frac{[(A) - (\alpha)][(B) - (\beta)]}{(A)(\alpha)(B)(\beta)} d(A) \cdot d(B)$$

Using (32)

$$r_{\sigma_1\sigma_2}\epsilon_{\sigma_1}\epsilon_{\sigma_2} = \frac{[(A) - (\alpha)][(B) - (\beta)]}{4N^2\sqrt{(A)(\alpha)(B)(\beta)}} \times \delta \qquad (36)$$

Or, dividing out by $\epsilon_{\sigma_1}\epsilon_{\sigma_2}$

$$r_{\sigma_1\sigma_2} = r_{12} \qquad (37)$$

28. To find the standard error of δ proved an unexpectedly lengthy piece of work: starting from the expression

$$\delta = \frac{1}{N}\left\{(AB)(\alpha\beta) - (A\beta)(\alpha B)\right\}$$

Differentiating, squaring, and summing as before, I arrived first at the result

$$\epsilon_\delta^2 = \frac{1}{N^3}\Big\{N(AB)(\alpha\beta)[(AB) + (\alpha\beta)] + N(A\beta)(\alpha B)[(A\beta) + (\alpha B)] \\ - 4[(AB)(\alpha\beta) - (A\beta)(\alpha B)]^2\Big\} \qquad (38)$$

The last portion of the expression in the brackets is $4\,N^2\delta^2$. Getting rid of the second order frequencies by substituting in terms of those of the first order and δ, I find finally,

$$\epsilon_\delta^2 = \frac{1}{N^3}\left\{(A)(\alpha)(B)(\beta) + N\delta[(A) - (\alpha)][(B) - (\beta)] - N^2\delta^2\right\} \qquad (39)$$

Finally, starting from either of the equations.

$$\begin{aligned}\delta &= (AB) - \frac{(A)(B)}{N}\\ &= \frac{1}{N}\left\{(AB)(\alpha) - (\alpha B)(A)\right\}\end{aligned}$$

I arrived at the a third equivalent form of expression.

$$\left.\begin{aligned} \epsilon_\delta^2 = &\frac{(AB)}{N}\left\{N - 2(A) - 2(B) - (AB)\right\} \\ &+ \frac{(A)(B)}{N^2}\left\{(A) + (B) + 6(AB)\right\} \\ &- 4\frac{(A)^2(B)^2}{N^3} \end{aligned}\right\} \qquad (40)$$

If $(A) = (B) = (\alpha) = (\beta) = N/2$ all these expressions reduce to the form

$$\left.\begin{array}{r}\text{For completely}\\ \text{symmetrical table}\end{array}\right\}\epsilon_\delta^2 = N\left(\frac{1}{4} + \frac{\delta}{N}\right)\left(\frac{1}{4} - \frac{\delta}{N}\right) \qquad (41)$$

When δ/N takes its maximum value of $1/4$ this becomes zero as we might expect. If $\delta = 0$, $\epsilon_\delta = \sqrt{N/2}$. The equation (41) checks with the value of ϵ_ω already given in equation (20), remembering that, in a completely symmetrical table $\omega = 4\delta/N$. If δ is zero, it may be noted, the last two terms in (39) vanish, and we have a simple expression for testing the presence of association ${}_2$ near the point of independence, in the general case.

29. To find the correlation between errors in δ and errors in σ_1, differentiating

$$\delta = (AB) - \frac{(A)(B)}{N}$$

we have

$$d\delta = d(AB) - \frac{(A)}{N}\,.\,d(B) - \frac{(B)}{N}d(A)$$

But (see work preceding equation 35)

$$d\sigma_1 = -\frac{1}{2N}\frac{(A)-(\alpha)}{\sqrt{(A)(\alpha)}}d(A)$$

Multiplying these two equations together, summing and simplifying, I find

$$r_{\delta\sigma_1}\,.\,\epsilon_\delta\,.\,\epsilon_{\sigma_1} = \frac{1}{2N^2}\frac{[(A) - (\alpha)]^2}{\sqrt{(A)(\alpha)}}\delta \qquad (42)$$

30. Now proceeding to the final stage: we have

$$\frac{dr}{r} = \frac{d\delta}{\delta} - \frac{d\sigma_1}{\sigma_1} - \frac{d\sigma_2}{\sigma_2}$$

Squaring and summing I find after a good deal of reduction

$$\begin{aligned}\epsilon_r^2 = \frac{1}{N}\Bigg\{1 - r^2 + \left(r + \tfrac{1}{2}r^3\right)&\frac{[(A) - (\alpha)][(B) - (\beta)]}{\sqrt{(A)(\alpha)(B)(\beta)}} \\ &- \tfrac{3}{4}r^2\left[\frac{[(A) - (\alpha)]^2}{(A)(\alpha)} + \frac{[(B) - (\beta)]^2}{(B)(\beta)}\right]\Bigg\} \qquad (43)\end{aligned}$$

If the table be completely symmetrical $(A) = (B) = N/2$, and this reduces to

for completely symmetrical table $$\epsilon_r^2 = \frac{1 - r^2}{N} \qquad (44)$$

thus checking with the expression already found for ω $(= r)$ in this case (equation 20). If the table be diagonally symmetrical only, *i.e.*, if $(A) = (B)$, but these are not 0·5 N

for diagonally symmetrical table $$\epsilon_r^2 = \frac{1}{N}\left\{1 - r^2 + \frac{r[(A) - (\alpha)]^2}{(A)(\alpha)}\left(1 - \tfrac{3}{2}r + \tfrac{1}{2}r^2\right)\right\} \qquad (45)$$

If $r = +1$ this becomes zero, as we might perhaps expect. If $r = -1$ it becomes imaginary. This is as it should be, for only an imaginary table could give $r = -1$ with diagonal symmetry: for a diagonally symmetrical table in which (A) is greater than (α) the greatest negative value that can be taken by r is $-(\alpha)/(A)$. In the general case, if (A) is greater than (α) and also greater than (B), and (B) is greater than (β), the greatest possible positive value of r is $+\sqrt{(\alpha)(B)/(A)(\beta)}$ and the greatest possible negative value $-\sqrt{(\alpha)(\beta)/(A)(B)}$. These results follow from the maximum positive and negative values of δ given at the end of § 8.

The work for obtaining this standard error was rather lengthy, but it has stood every check to which I have subjected it, and is, I hope, correct. Dr. M. Greenwood has been so good as to go through my algebra, and I am glad to take this opportunity of thanking him for his assistance.

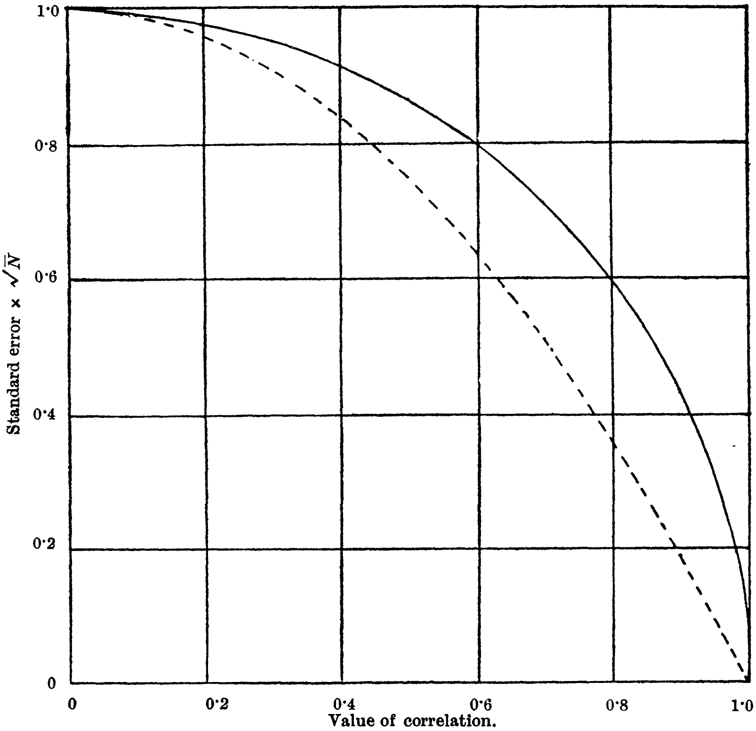
30. For a normal correlation table the standard error of r is (Pearson 14) given by $$\epsilon_r^2 = \frac{(1 - r^2)^2}{N} \qquad (46)$$

Hence if a normal table and a two-rowed table have the same number of observations and the same correlation, the standard error is always least for the former. The ratio is greatest (approaching infinity) for high values of r, as is shown by the following table :—

Value of r.	Standard error × $\sqrt{N}$.	
	Normal table.	Two-rowed table.
0·1	0·99	0·995
0·2	0·96	0·980
0·3	0·91	0·954
0·4	0·84	0·917
0·5	0·75	0·866
0·6	0·64	0·800
0·7	0·51	0·714
0·8	0·36	0·600
0·9	0·19	0·436

The standard error of *r* for the fourfold table drops at first more slowly, afterwards more rapidly, than that for the normal table: fig. 1 shows the course of events more clearly.

FIG. 1.—*Standard error of correlation* × $\sqrt{N}$ *for normal correlation, dotted curve; standard error of correlation* × $\sqrt{N}$ *for symmetrical fourfold table, full curve.*



31. As regards the ratio of the standard error of *r* for a fourfold table to the standard error of the coefficient ω for the same table, I do not find it possible to make any general statement: the ratio may be either greater or less than unity. The values of the standard errors for Tables I, III, and IV, will serve to illustrate the point.

Districtior hospital.	Value of ω.	Standard error.	Value of *r*.	Standard error.
Sheffield	0 ·63	0 ·0169	0 ·48	0 ·0204
Leicester..	0 ·57	0 ·1269	0 ·23	0 ·0407
Homerton and Fulham............ ..,	0 ·50	0 ·0106	0 ·41	0 ·0106

It will be seen that the standard error of *r* is greater than the standard error of ω in the case of Sheffield, that the two standard errors are appreciably equal in the case of Homerton and Fulham, and that the standard error of r is less than one-third of the standard error of ω in the case of Leicester. The ratio of ω to its standard error is greater than the ratio of r to its standard error in the first and last cases, but the converse holds good in the case of Leicester.

32. Taking the three tables under the heading B in Table VI as another illustration, I find the following results:—

Table.	Value of ω	Standard error × $\sqrt{N}$.	Value of r.	Standard error × $\sqrt{N}$.
B 1	0·5	0·866	—	—
B 2	0·5	1·507	0·32	1·040
B 3	0·5	2·467	0·20	2·023

For both the tables B 2. and B 3. the standard error of r is *less* than the standard error of ω. How, then, do we find a *greater* standard error for r in the case of Sheffield above? The answer seems to be that the first effect of moderately increasing the ratio of (A) to (α) is relatively to increase the standard error of r. Thus if the right-hand column and bottom row of table B 1. is multiplied by three, I find $\epsilon_\omega \times \sqrt{N} = 1{\cdot}146$, $r = 0{\cdot}40$, $\epsilon_r \times \sqrt{N} = 1{\cdot}206$. Again, the relatively great excess of ϵ_ω as compared with ϵ_r in the Leicester table is due to the approximate equality of the numbers of vaccinated and unvaccinated. If, say, (A) and (α) are equal, but (B) and (β) are very unequal, the third term in (43) vanishes, but the fourth negative term may remain large, and hence the value of ϵ_r will become relatively small. Thus if we take the table B 2, and multiply the left-hand column by 2·3846, we make $(A) = (\alpha)$. For this table I find $\epsilon_\omega \times \sqrt{N} = 1{\cdot}629$, $r = 0{\cdot}29$, $\epsilon_r \times \sqrt{N} = 0{\cdot}778$, or less than half $\epsilon_\omega \sqrt{N}$. I regret that I have not been able to give any general algebraical investigation as to the influence of altering the relative proportions of A's and B's on the ratio of the standard errors, and these arithmetical illustrations must suffice to show that the matter is not a simple one.

33. It should have been a very obvious matter that the ordinary theory of correlation, once that theory had been freed from any necessary relation to the theory of normal correlation, was applicable in its entirety to the 2 × 2-fold table. The obvious remained unrecognised, however, for a considerable period, so that the history of this coefficient, as they say in the *Arabian Nights*, if it were engraved upon the intellect, would be a lesson to him who would be admonished. The formula was first given, I believe, by myself. It occurs twice in my memoir on association (32, p. 284, p. 287), the first time as the correlation between errors of sampling in (A) and

(B), the second time as the correlation between errors of sampling in the "surplus ratios," as I termed them, viz., $[(A) - (\alpha)]\ /N$ and $[(B) - (\beta)]/N$. But I never recognised the expression as being simply the coefficient of correlation for the fourfold table, nor did I see that it might be used as a coefficient of association. It occurs next, a few months later, in Professor Pearson's memoir "On the correlation of characters not quantitatively measurable" (15, pp. 11–12, pp. 15–16) as the correlation between errors of sampling in (A) and (α) again, and also as the correlation r_{hk} between errors of sampling in the co-ordinates of the means of the two normally distributed variables which the attributes are assumed to represent. Professor Pearson, like myself, did not recognise the expression he obtained as the coefficient of correlation for the fourfold table, but he did point out that it possessed all the essential properties for a coefficient of association, and indicated its principal characteristics as such. We next come to a whole series of researches on the mathematical theory of Mendelian inheritance in a population breeding at random, in which the fourfold table occurs as a possible limiting case, either implied or specifically considered, and by which the use of the coefficient is either implied or expressed. The fundamental memoir is that by Professor Pearson (22, 1904). He discusses the inheritance of the number of pure recessive couplets in a zygote consisting of n couplets, in a population derived from a cross between a pure dominant and a pure recessive, and finds that the correlation between parent and offspring is independent of the magnitude of n and has the value 1/3. If the correlation is independent of n, it must retain the same value in the limit when $n = 1$ and the number of recessive couplets can, therefore, only be 0 or 1. The table in this case is easily found to be of the form

Offspring.	Parent.		Total.
	A.	a.	
A	5	1	6
a	1	1	2
Total	6	2	8

Here $\delta = 0{\cdot}5$, and therefore $r = 4/6 \times 2 = 1/3$ as stated. It was this memoir of Professor Pearson which first directed my attention to the product-sum correlation for a fourfold table. The memoir in question was followed by a short note by myself three years later, the fourfold table being again a possible limiting case (35, 1907). Professor Pearson returned to the same subject in 1909 (27), and in the latter part of this paper specially considered the case of dominance, for a character determined by a single couplet. Assuming that the numbers of dominants, heterozygotes, and recessives in the general population are in the proportions $p^2 : 2pq : q^2$, he finds the correlation between offspring and the nth generation of ancestry to be $q/2^{n-1}(p + 2q)$. If $n = 1, p = q$, we have the previous value 1/3. The following year Dr. Snow (29), in a paper presented to the Royal Society by Professor Pearson, dealt with the correlation between brothers on the same theory. Taking his fourfold Table III (p. 42) and reducing it to the simplest form by putting s infinite and $p = q$, it becomes

Second brother.	First brother.		Total.
	A.	*a.*	
A	41	7	48
a	7	9	16
Total	48	16	64

Here $\delta = 5$, whence $r = 64 \times 5/48 \times 16 = 5/12$, agreeing with the value giving by Mr. Snow $(20s - 3)/3(16s - 1)$ on putting s infinite.

34. In May, 1909, a few weeks later than the second paper on Mendelian inheritance, by Professor Pearson, referred to above, Dr. Franz Boas published a letter in *Science* (1) on "the determination of the coefficient of correlation" (I owe this reference to Dr. Heron (5), in which he directly pointed out the correspondence between attributes and two-valued variables, and deduced the formula for the correlation coefficient. Professor Pearson (28), replying to this letter in the following month, failed to recognise the coefficient or to understand what Dr. Boas was doing, and reproved him accordingly. "I find," he writes, "that Dr. Boas's r is our old friend r_{hk} [formula follows], the correlation in the deviation of the mean of one variable from its mean value with the deviation of the mean of the second variable from its mean value. It is not a true correlation of the first variable with the second variable." After a brief recapitulation of the properties of the coefficient, Professor Pearson continues, "Thus it differs in the simplest cases from the true coefficient of correlation, and often differs considerably . . . . and its use is liable to be misleading, especially if compared with values of the true coefficient found by other processes." By the "true" correlation is apparently meant Professor Pearson's coefficient, dealt with below under the name of the "normal coefficient."

35. In the same year (6, 1909) Johannsen, unaware of Pearson's

use of the coefficient, gave a very full and clear account of it in his "Erblichkeitslehre," an account which has largely influenced my own description above. In the following year Brownlee (2, 1910) directed attention to the very different values given by the product-sum method and by Pearson's normal coefficient when applied to tables of Mendelian inheritance. Finally, in my "Introduction to the theory of Statistics," published last year, I thought it desirable to give a page to the matter (pp. 212—13), first in view of the importance of Professor Pearson's and other memoirs on the theory of Mendelian inheritance in which the coefficient had been used, and secondly, in view of the fact that I had used it myself in the latter portion of the book, in connection with the theory of sampling (cf. pp. 283—4), one of those typical cases which would almost inevitably have led to its use sooner or later. It would have been thought that anyone reasonably acquainted with the theoretical work of the last decade, and especially Professor Pearson and his collaborators, would have found no difficulty in the passage in question. Dr. Heron, however, in an articlे published last August in *Biometrika*, classed the coefficient with "methods which in no circumstances can give correct results" (5, opening paragraph), finds the few lines of simple algebra in my book necessary to arrive at the formula (24) "a process which is nothing short of extraordinary" (5, p. 113), refers again to r_{hk} and cites approvingly Professor Pearson's letter on Boas. At some time in the latter part of last year the mistake seems to have been discovered; in the last issue of *Biometrika*, published at the end of February, the coefficient is recognised by Dr. Snow (30), and the comedy of errors has, it is to be hoped, ended.

IV. *Pearson's normal coefficient.*

36. Professor Pearson's coefficient, which was described and its use illustrated in two memoirs published in 1900 (15, 16), is founded on ideas entirely different from those developed in the preceding sections. "A certain number of characters in living forms," he wrote in the second memoir, "are capable of easy observation, and "thus are in themselves suitable for observation, but they do not "admit of an exact quantitative measurement, or only admit of this "with very great labour. The object of the present paper is to "illustrate a method by which the correlation of such characters "may be effectively dealt with in a considerable number of cases. "The conditions requisite are the following:—(1) The characters "should admit of a quantitative order, although it may be impossible "to give a numerical value to the character in any individual. . . . "(2) We assume that the characters are a function of some "variable, which, if we could determine a quantitative scale, would

"give a distribution obeying—at any rate to a first approximation— "the normal law of frequency." In further explanation and justification of the processes involved, he continues a little further on: "If we take a problem like that of coat colour in horses, it is "by no means difficult to construct an order of intensity of shade. "The variable on which it depends may be the amount of a certain "pigment in the hair, or the relative amounts of two pigments. "Much the same applies to eye-colour. In both cases we may fail "to obtain a true quantitative scale, but we may reasonably argue "that, if we could find the quantity of pigment, we should be able "to form a continuous curve of frequency. We make the assumption "that this curve—to at any rate a first approximation—is a normal "curve. Now if we take any line parallel to the axis of frequency "and dividing the curve, we divide the total frequency into two "classes, which, so long as there is a quantitative order of tint or "colour, will have their relative frequency unchanged, however we, "in our ignorance of the fundamental variable, distort its scale. . . . "Our problem thus reduces to the following one: Given two classes "of one variable, and two classes of a second variable correlated "with it, deduce the value of the correlation."

37. The mathematical problem: *Given that a fourfold table such as Tables I–IV represents a grouping of normally distributed frequency, required to find the correlation,* is solved in the first memoir (15) in a piece of magnificent mathematical work. The solution gives the correlation as the root of an equation

$$\frac{\delta}{N} = \tau_1 . \tau_1' . r + \tau_2 . \tau_2' . r^2 + \tau_3 . \tau_3' . r^3 + \ldots\ldots \qquad (47)$$

where τ_1, τ_2, etc., are certain functions (by no means simple functions) of $(A)/N$, and τ_1', τ_2', etc., are the same functions of $(B)/N$. These functions have comparatively recently been tabulated (Everitt 4). There is only one root to the equation between the limits ± 1, and if $\delta = 0$ this root is zero. If either $(A\beta)$ or (αB) is zero, $r = 1$: this is not obvious from the form of the equation, but it is clear from the fact that a distribution of frequency along a straight line (the limit to the normal distribution when $r = 1$) can pass through three of the four quadrants. The coefficient resembles in this one respect the coefficients Q and ω, not the product sum correlation. The number of terms that it is necessary to retain on the right-hand side of (47) depends on the degree of accuracy desired and the rapidity of convergence of the series. In favourable cases, and if two-figure accuracy is sufficient, four terms may suffice: in unfavourable cases twelve may be hardly enough. It will be seen that the calculation is not a simple matter: even with the aid of Mr. Everitt's tables (4) it occupies a good deal of time, and would be practically impossible without the aid of an arithmetic machine. Neither are there any simple and obvious properties which render the general nature of the coefficient and its relations to the class-frequencies readily comprehensible; it bears no simple and straightforward relation to δ or to the differences p_1—p_2 and p_3—p_4. For a given set of "equivalent" tables with the same values of Q and ω, the value of the coefficient, the "normal "coefficient," as I propose to term it, as it does not at present bear any special name, is greatest for the symmetrical table. In this respect it resembles the product-sum correlation, but the decrease with increasing ratio of $(A)/(\alpha)$ is not so rapid as that of the latter, as is illustrated by the following values for the distributions of sets A and C in Table VI:—

Values of the normal coefficient for Table VI.

	Set A.	Set C.
Symmetrical distribution	0·31	0·81
Second distribution	0·23	0·68
Third distribution	0·19	0·58

The formula for the probable error is complex, and I must refer the reader to the original memoir for it.

38. Here I am concerned rather with the assumptions and their applicability, and the significance of the resulting coefficient. The first remark to be made is one that must occur to almost any reader of the memoir. If the attributes A and α (B and β) represent naturally discrete classes, the fundamental assumption is, according to the precise form in which we take it, either inapplicable or unnecessary—and incapable (as a rule) of verification. For discontinuous attributes—attributes proper, as we might term them—the true correlation is that given by formula (24) or (26): we are dealing with a variable, in fact, which can only take two values as distinct from a variable exhibiting a normal or any other continuous distribution. Tables I, III and IV, as it seems to me, represent precisely such a case. Those who are unvaccinated are all equally non-vaccinated (assuming of course correct observation): no one individual is more unvaccinated or less unvaccinated than another. The fact that the operation of vaccination may have been more or less successfully performed in the case of those who *have* been vaccinated does not lessen this discontinuity; we are considering simply the performance as against the non-performance of the operation of vaccination. Similarly, all those who have died of small-pox are all equally dead: no one of them is more dead or

less dead than another, and the dead are quite distinct from the survivors. Formula (24) gives therefore the correlation between *performance of vaccination* and *death*. From this standpoint Professor Pearson's assumptions are quite inapplicable and do not lead to the true correlation between the attributes. But this is not, apparently, the standpoint taken by Professor Pearson himself. He assumes rather that "the characters are *a function of* some "variable which . . . would give a distribution obeying . . . "the normal law of frequency" (my italics) and he takes as one example of his method (15, page 43) some data as to vaccination (presence or absence of cicatrix) and small-pox (recovery or death). The columns for "recoveries" and "deaths" bear the general heading "Strength to resist small-pox when incurred," and the rows for "cicatrix present" and "cicatrix absent" the general heading "Degree of effective vaccination." In the next example (*ibid.*, page 44) the data refer to recovery or death from diphtheria with and without the administration of antitoxin: in this case no general headings corresponding to those given for the vaccination table are attempted. In neither of these cases am I able to agree that the data can be legitimately regarded as groupings of continuous variables, but apart altogether from any question as to its legitimacy, the assumption is unnecessary and unverifiable. It is unnecessary, because equation (24) or (26) gives the correlation between performance of vaccination and recovery from small-pox, or in the second case between administration of antitoxin and recovery from diphtheria. It is unverifiable because we have only the four frequencies given us, and these afford no guide as to whether the correlation is normal or non-normal. At the best the normal coefficient can only be said to give us in cases like these a hypothetical correlation between supposititious variables. The introduction of needless and unverifiable hypotheses does not appear to me a desirable proceeding in scientific work.

39. Let us pass, then, to the case where the attributes do not represent naturally discrete classes and the hypothesis is consequently less unreasonable, and in the first place let us define the hypothesis itself more closely. It does not appear to me that the statement of the hypothesis given by Professor Pearson in his second illustrative memoir, and cited above, is quite sufficient. The assumption is not really one concerning the forms of distribution of the individual variables supposed to underlie the attributes with which we are dealing, but the form of their correlation, *i.e.*, the distribution of frequency in two dimensions. The distributions of the individual variables might be normal, but the form of the correlation not even remotely normal. If the distributions of the individual variables

were originally skew, the methods of measurement might be so altered as to strain these skew distributions to the normal form (as by Professor Edgeworth's process of "translation"); but this strain may not make the distribution of frequency for the correlation table even remotely normal. The assumption on which alone the normal coefficient may become identical with the correlation is, as it seems to me, that the form of the distribution for the correlation table is either (1) normal as it stands or (2) such that it may be made normal by strains (non-homogeneous strains) parallel to each axis of measurement in succession. If, for example, we take a normal surface, divide it up by k planes parallel to the x-axis and l planes parallel to the y-axis, we may compress or stretch the distances between these planes in any, quite irregular, way, but the normal coefficient will remain the same for every pair of dividing planes, and will give the correlation coefficient for the unstrained normal form of the surface. We assume, then, not necessarily that the correlation surface is normal, but that it is "strained normal," as we may term it, and it is conceivable that "strained normal" may cover markedly skew correlation tables. The question whether it does so or not is one that can only be decided by trial, but on its validity the whole value of the normal coefficient depends.

40. The importance of the assumption is recognised by Professor Pearson in his first memoir (15), and also the probability of its failure. One section of that memoir (§ 5, pp. 14–18) is devoted to finding "a physical meaning for the series in r," but the "physical "meaning" found only amounts to showing that the quantity on the left of the equation, differently expressed in his memoir, is δ/N. As every coefficient of association must necessarily be a function of δ, and equation (47) shows that the normal coefficient is a singularly complicated function, the "physical meaning" is somewhat obscure. Professor Pearson, having arrived at this result, proceeds to take various simpler functions of δ/N which vanish with the normal coefficient and also attain the value unity under the same conditions, and considers five such functions altogether, including my coefficient of association Q. After comparing the values given by such other functions with those given by the normal coefficient in certain actual cases, he finds that one of them (Q_5) gives the closest approximation, and that it may be taken "as a good measure of the degree of "independent variation." "The reader may ask," he continues, "Why is it needful to seek for such a measure? Why cannot we "always use the correlation as determined by the method of this "paper? The answer is twofold. We want first to save the labour "of calculating r where the data are comparatively poor, and so "reaching a fairly approximate result rapidly. But labour-saving

"is never a wholly satisfactory excuse for adopting an inferior "method. The second and chief reason for seeking such a "coefficient as Q* lies in the fact that all our reasoning in this "paper is based upon the normality of the frequency. We "require to free ourselves from this assumption if possible, for "the difficulty, as is exemplified in Illustration V below, is to "find material which actually obeys within the probable errors any "such law." It seems to me that a hostile critic could hardly have penned a more condemnatory passage. If it is difficult to find material obeying the law assumed, what is the special advantage of the normal coefficient ? Its sole claim to value lies in its alleged equivalence to the correlation between the supposed variables represented by the attributes: if it is not equivalent it becomes merely an empirical coefficient of association, with no special significance, and its laborious calculation is a waste of labour. Professor Pearson's Illustration V, to which he refers in the passage cited, consists of an examination of the values of the normal coefficient for different divisions of a correlation table between stature of father and son (Galton's data), from which he concludes that the differences observed between the values of the normal coefficient for different divisions of the table "are sensibly "larger than the probable error of the differences, even in some "cases double; hence it is not the method but the assumption of "normal correlation for such distributions which is at fault. As we "shall hardly get a better variable than stature to hypothesise "normality for, we see the weakness of the position which assumes "without qualification the generality of the Gaussian law of "frequency." The result ought surely to have served as a warning that the normal coefficient could not be safely regarded as giving the correlation coefficient, or as at all equivalent thereto. As a fact it seems to have been forgotten in the course of a few weeks at most. In the very next memoir (16), on the inheritance of coat-colour in horses and eye-colour in man, it is used as equivalent to the correlation-coefficient, without any attempt, apparently, to test the validity of the fundamental assumption as to the form of the correlation. As the classification used was not merely twofold but manifold, the application of such a test (in the same way that it was applied to the stature table) would have presented no difficulty. I take the tables of this memoir accordingly as my first illustration of the impossibility of regarding the normal coefficient as in any way an approximation to the correlation, *i.e.*, the product sum correlation

* As I understand the passage, this does not mean my coefficient of association designated by that symbol, but any empirical coefficient such as Q_5.

that would be found if the frequency distributions for the individual variables were strained into normal form.

41. *Eye-colour in man.*—The classification of eye-colour in the tables of this memoir (16) is 8-fold, the classes being:—1. Light blue; 2. Blue, dark blue; 3. Blue-green, grey; 4. Dark grey, hazel; 5. Light brown; 6. Brown; 7. Dark brown; 8. Very dark brown, black. Professor Pearson worked out the normal coefficient for one division only, as I judge from the results and from an illustration in the earlier memoir (15, p. 39), the division being taken between classes 3 and 4; the matter seems to have been thought of so little importance that, so far as I can find, it is not stated in the second memoir and certainly no stress is laid on it. To test the validity of the assumption that the correlation table can be regarded as "strained normal" in form, I first worked out the normal coefficient for *four* divisions in five of the tables. I grouped classes 3 and 4 together as a central class, and took the division at each corner of this class in turn. Columns A and B below give the values for the symmetrical divisions, and columns C and D for the cross divisions. It will be seen that even the most optimistic observer could not describe the resulting values of the normal coefficient as constant. In Table I they range from 0·58 to 0·28; in Table II from 0·52 to 0·12; in Table III from 0·49 to 0·27, and so on.

Normal coefficients for four divisions of eye-colour tables.

Table of memoir: and relationship.	Divisions (A and C at light end).				Pearson's value in memoir.
	Symmetrical.		Unsymmetrical.		
	A	B	C	D	
I. Father and son	0·50	0·58	0·28	0·39	0·5503
II. Father and daughter	0·35	0·52	0·12	0·29	0·4370
III. Mother and son	0·46	0·49	0·41	0·27	0·4817
IV. Mother and daughter ..	0·40	0·46	0·38	0·25	0·5096
V*b*. Brother and brother	0·50	0·39	0·24	0·24	0·5169

The number of observations is 1,000 for Tables I–IV, 1,500 (made 3,000 by symmetry) for Table V*b*, and consequently the differences are large compared with the probable errors—Professor Pearson gives the probable error of his value of the coefficient for Table III as 0·028—and further, they are obviously systematic. In every table the lowest of the coefficients for the symmetrical divisions is greater than the greatest of the coefficients for the cross divisions. With these results before me, I thought it worth while examining two of the tables in greater detail, and chose for this purpose Tables I and V*b*; the latter being symmetrical, the work for it is somewhat lightened. The frequencies of tints 1, 5, and 8 being relatively small, I did not calculate the value of the coefficient for every possible division, which would have given me 49 coefficients to reckon, I pooled tints 1 with 2, 5 with 6, and 7 with 8, leaving a table with five arrays or 16 possible values for the normal coefficient. The results are given below:—

Normal coefficients for 16 *different divisions of eye-colour Table I of memoir* (16)*: father and son.*

Son : division between.	Father: division between eye-colours.			
	2 and 3	3 and 4	4 and 5	6 and 7
2 and 3	0·50	0·38	0·39	0·32
3 and 4	0·39	0·55	0·42	0·49
4 and 5	0·28	0·47	0·58	0·52
6 and 7	0·27	0·38	0·46	0·51

Normal coefficients for 16 *different divisions of eye-colour Table Vb of memoir* (16)*: brother-brother* (*symmetrical*).

Second brother : division between.	First brother: division between eye-colours.			
	2 and 3	3 and 4	4 and 5	6 and 7
2 and 3	0·50	0·38	0·24	0·24
3 and 4	0·38	0·52	0·35	0·28
4 and 5	0·24	0·35	0·39	0·30
6 and 7	0·24	0·28	0·30	0·31

It will be seen that the greatest values of the coefficient are given by the symmetrical divisions (along the diagonal) and the lowest values by the most unsymmetrical. The tables are not even remotely "strained normal" in form, and the values of the coefficient given by Professor Pearson are the greatest or nearly the greatest values attained. It is clear that they must be considerably in excess of any possible value for the correlation, and it is a matter of some interest to estimate how great such excess may be. Some years ago (34, 1906) I suggested that the mean of four divisions round a central class might give a good approximation to the correlation, though the symmetrical divisions gave values that were clearly too high. But in this I was in error. From the above results it appears that the mean of the normal coefficients for a set of four divisions will itself be appreciably dependent on the precise positions in which the axes of divisions are fixed, and from an illustration below (§ 49) that the correlation may possibly be given by the *lowest of all* the values that the normal coefficient takes. You cannot estimate the correlation with any confidence even by averaging normal coefficients.

42. Professor Pearson has now adopted the method of contingency in lieu of the method of the normal coefficient for use on such tables, and apparently regards this as giving "correlations" and the values as supporting those given by the normal coefficient. But the coefficient of contingency itself only gives an approximation to the correlation coefficient when the correlation is normal, and suffers therefore from precisely the same defect as the normal coefficient. It is quite true that the coefficient of contingency and the normal coefficients for symmetrical divisions give values that often differ comparatively little from each other for tables of the present type: this only shows that the coefficient of contingency, regarded as an approximation to the correlation, is also untrustworthy. To enforce my point, let me give the results of an experiment on the brother-brother table for eye-colour (Table V*b* of Professor Pearson's memoir). Using the fivefold classification of this table as described above I calculated, for the same row and column totals, (1) the frequencies for a normal distribution with $r = 0{\cdot}3$, (2) the frequencies for a distribution with the same value of r, given by the rules $(A_m\, B_n) = 0{\cdot}7\, \frac{(A_m)\,(B_n)}{N}$ (for non-diagonal compartments) $(A_n\, B_n) = 0{\cdot}7\, \frac{(A_n)\,(B_n)}{N} + 0{\cdot}3\,(A_n)$: where it must be remembered that $(A_n) = (B_n)$ for all values of n as the table is symmetrical. This distribution, therefore, consists of a number of independent pairs with perfectly correlated pairs superposed on them down the diagonal. The coefficients of mean square contingency for the original and the two constructed tables were as follows:—

Original table..	$C = 0{\cdot}44$
Normal distribution..	$C = 0{\cdot}26$
Compound distribution..	$C = 0{\cdot}51$

It will be seen that the coefficient of contingency for the compound distribution is nearly double that for the normal distribution, though both have the same coefficient of correlation. For a compound distribution of this kind, the contingency is in fact a measure of the number of arrays quite as much as a measure of the correlation. The correlation is constant for all symmetrical groupings—*i.e.*, whatever the number of arrays—for it is given simply by $n_1 / (n_1 + n_2)$, where n_1 is the number of perfectly correlated pairs and n_2 the number of independent pairs. But the coefficient of mean square contingency for such a table is, t being the number of rows or columns,

$$C = \sqrt{\frac{r^2(t-1)}{1+r^2(t-1)}} \qquad (45)$$

$$= \sqrt{\frac{r^2}{r^2 + \frac{1}{t-1}}}$$

We have therefore only to make t sufficiently large to get C as near unity as we please, whatever the value of r. The course of affairs is illustrated more clearly by the following short table:—

Number of arrays t.	Value of contingency coefficient when r is		
	0·1	0·3	0·5
2	0·099	0·287	0·447
4	0·173	0·461	0·655
6	0·218	0·557	0·745
8	0·256	0·622	0·798
10	0·287	0·669	0·832

It will be seen that the coefficient of contingency may well give its closest approximation to r for a *fourfold* table, not for a six or eight-rowed table as is generally assumed. The fact that C rises as the number of arrays is increased may not be due merely to the influence of errors of sampling or the occurrence of single units in the table as is often stated, but may be an essential property of the distribution, quite independent of the number of observations. The coefficient of contingency is, therefore, no more a trustworthy guide to the value of the correlation between the supposed variables than is the normal coefficient. We must seek some other guide.

43. From several trials—more than are here given—I have come to the tentative conclusion that the best guide to the correlation that would be found for given data, if the grouping were other than that which in fact it is, is the correlation for the existing grouping, provided that you are given at least some five or six arrays. Take your given table, that is to say, treat it exactly as if it were a distribution grouped by uniform equal intervals, and work out the correlation in the ordinary straightforward way. For a moderately large number of arrays this will not differ greatly from the coefficient that would be given by the same data if they were actually grouped by equal intervals or other intervals, and with a larger number of arrays, and the result can be partially checked by evaluating the correlation in the same way for coarser

groupings than those given. The result is an empirical one, and the method I know sounds rash, but it makes no assumption of normality, and I am fairly convinced that it is on the whole the best available. Professor Pearson, it will be remembered, has shewn how curiously stable the coefficient of correlation is to certain small changes of grouping (26); it is more stable than I myself at one time thought for even changes of considerable magnitude. The following are the results for Eye-colour, Table V*b* of Professor Pearson's memoir, with the correlations for the normal distribution of $r = 0{\cdot}3$ worked out as a control, and the correlations taken for tables of 2, 3, 4, 5, and 8 arrays to see the nature of the influence of number of arrays. It is not necessary to work out the correlations similarly for the compound distribution as a further control, for the correlation, as already pointed out, is constant for all symmetrical groupings of such a table. The groupings actually taken for the eight eye-colour classes were:—

Five classes	1, 2 : 3 : 4 : 5, 6 : 7, 8.
Four classes	1, 2 : 3 : 4 : 5, 6, 7, 8.
Three classes	1, 2 : 3 : 4, 5, 6, 7, 8.
Two classes	1, 2, 3 : 4, 5, 6, 7, 8.

Number of arrays.	Value of the correlation coefficient.	
	Actual table.	Normal distribution with same totals for rows and columns.
2	0·34	0·19
3	0·36	0·24
4	0·33	0·25
5	0·30	0·26
8	0·28	0·26
Infinite.	?	0·30

The result emphasises the entire non-normality of the eye-colour table. For the normal distribution the correlation gradually increases towards the known true value as the number of arrays is increased: with five or eight arrays, notwithstanding the extreme irregularity of the grouping, we have the same moderately good approximation to the correlation as is given by the coefficient of contingency in this case. For the eye-colour table the correlation *decreases* as the number of arrays is increased. In the case of the normal distribution we would have no hesitation in estimating the the true value of the correlation for a finer and more uniform classification as something slightly greater than 0·26. Can we have any hesitation in similarly estimating the correlation for the eye-colour table, if we were in a position to adopt a finer and more uniform grouping (without assuming that we will compel that grouping to give us a normal distribution) as something slightly *less* than 0·28 ? It does not seem to me that there can be much doubt about the matter: the correlation between eye-colour of brothers in Galton's data, for a fairly fine classification, may be taken as slightly less than 0·28, not 0·5169—or anything approaching it—as given by Professor Pearson. I have carried out the same test, but not in the same detail, on Tables I–IV of the memoir, for parental inheritance of eye-colour. The results are:—

Table and relationship.	Product-sum correlations for—	
	Actual table.	Threefold grouping.
Table I. Fa. & S.	0·39	0·37
„ II. Fa. & D.	0·31	0·25
„ III. M. & S.	0·32	0·34
„ IV. M. & D.	0·33	0·33

In only one of these four tables does the correlation for the threefold grouping come out larger than that for the original table, as in the case of Table V*b*; but in two of the other cases the difference is very small, and only in the one case (Table II) is the decrease of correlation with decreasing number of rows anything like that of the normal distribution in the last illustration. I do not think, again, that we can be very far out in estimating the correlations that would be found for the parental relationships, with a finer and more regular system of classification, as in the neighbourhood of those given by the left-hand column, instead of in the neighbourhood of those given by Professor Pearson—0·55, 0·44, 0·48, 0·51. The average estimated correlation is in round numbers something like 1/3, not 1/2. There is no real basis for the statement that still continues to be made, to the effect that the correlation for inheritance of eye-colour is approximately 0·5: that value is due solely to the use of a quite inapplicable method of estimating the correlation.

44. *Coat-colour of horses.*—I pass to the tables for coat-colour of horses in the same memoir, but will deal with these more briefly. I have taken only the four tables for parental relationship and worked out the normal coefficients for the four divisions round a central class (Bay) in the same way as before. *A* and *B* give the results for the symmetrical divisions, *C* and *D* for the unsymmetrical divisions, as before for the eye-colour tables. The values given in column *B* correspond to those given by Professor Pearson. It will be seen that his values are again the greatest or nearly the greatest occurring, and that the general run of the coefficients is similar to

that of the coefficients for the eye-colour tables. Tbe coefficients for symmetrical divisions are greatest, for the cross-divisions least. The values of the normal coefficient given by Professor Pearson are again obviously in excess of any possible value for the correlation, and we may again work out the product-sum correlations, as for the eye-colour tables, to test the magnitude of the error. The original tables in this case have 16 arrays: there are 6 principal coat-colours—black, brown, bay, chestnut, roan, and grey—and Professor Pearson has also noted two intermediates between each pair of colours, so that the scale runs: black, black-brown, brown-black, brown, and so on. These 16-fold tables I condensed to 11-fold by pooling each pair of intermediates together—there are relatively very few—and worked out the product-sum correlations for the 11-fold tables and also for the 3-fold tables obtained by taking bay as the central class. Bay, it may be noted, covers nearly 50 per cent. of all the colours. The results are as follows :—

Normal coefficients for four divisions of tables for coat-colour of horses.

Table of memoir and relationship.	Divisions (A and C at black end).				Pearson's value in memoir.
	Symmetrical.		Unsymmetrical.		
	A.	B.	C.	D.	
Table I. Sire and colt	0·34	0·49	0·30	0·23	0·4913
„ II. Sire and filly	0·38	0·54	0·27	0·27	0·5422
„ III. Dam and colt	0·52	0·48	0·35	0·32	0·4862
„ IV. Dam and filly	0·36	0·57	0·28	0·35	0·5668

Table and relationship.	Product-sum correlations for—	
	Elevenfold table.	Threefold table.
I. Sire and colt	0·28	0·29
II. Sire and filly	0·35	0·32
III. Dam and colt	0·37	0·36
IV. Dam and filly	0·34	0·34

It will be seen that the product-sum correlations also closely resemble those obtained for the corresponding eye-colour tables. The true correlation for the inheritance of coat-colour in horses is, I conclude, likewise of the order 1/3 not 1/2. Yet the values of the normal coefficient in this memoir are termed "Coefficients of correlation" without qualification (*cf.* Tables II and VII of the memoir and numerous passages in the text), they are brought without warning into the same tables as product-sum correlations calculated in the ordinary way for measured characters (Tables V and IX)—as in later memoirs also—and they are compared with the theoretical values obtained for the true correlation coefficient on various hypotheses and divers conclusions are drawn from the values.

45. The values given in all the memoirs on the inheritance of similar unmeasured characters are wholly untrustworthy, and tend to be much too high as compared with true correlations. This comment applies, for example, to Professor Pearson's Huxley Lecture (23), to his work with Miss Barrington and Miss Lee on Coat-Colour of Greyhounds (24), and his work with Miss Barrington on Inheritance of Coat-Colour in Cattle (25). Of the "correlations" (normal coefficients) in the Huxley Lecture I have already written (34). They are not calculated on any uniform plan. Some are calculated on the basis of a single division only, some on the basis of two symmetrical divisions, others on the basis of four divisions round a central class. It will have been seen from the above how largely a coefficient may be reduced by taking the mean of four divisions round a central class instead of the value given by a single symmetrical division, and obviously coefficients calculated in such different ways are incomparable with one another. No warning is given to the reader as to the effect of this choice of diverse methods. "Theoretically," he is told (23, p. 149), "the "fourfold divisions ought to be made everywhere possible, and the "weighted mean taken of the results to smooth out irregularities, "but the labour is too great for practical purposes, and we must "content ourselves with a few simple divisions." Labour does not seem, however, to have stood in the way when some adjustment of a coefficient was thought desirable. The value given in the Huxley Lecture for correlation between eye-colour of brothers is 0·54, a value agreeing very closely with that given as the result of Galton's data in the earlier memoir, viz., 0·52. But the latter is calculated from a single division between blue eyes and hazel, the former as the mean of four divisions round a central class. If Professor Pearson had taken in the Huxley Lecture a division corresponding as nearly as possible with that taken in the earlier memoir he would have found, I think, a "correlation" of 0·64, not 0·54. If he adopted the process of taking the mean of four normal coefficients in this case because he thought it gave a closer approximation to the true product-sum correlation, it is a little difficult to see why he did not adopt it uniformly in every case, and warn his readers that when a single division only was possible the result was certainly too high. Since the date of the Huxley Lecture at least Professor Pearson must have realised the non-normality of these inheritance

tables, for by taking the mean of four values of the normal coefficient round a central class, instead of one or two symmetrical divisions only, he reduced the "correlation" for eye-colours of brothers, for example, from 0·64 to 0·54, that for hair-colour from 0·71 to 0·62. In what sense can the normal coefficient be regarded as an approximation to the correlation when it varies so largely for different divisions of the same table as either to require, or to permit, adjustments of this character? If it had been put forward merely as an empirical coefficient of association no one would have expected it to give the same value for all divisions, and its limitations would have been recognised.

46. But, perhaps it may be said, these inheritance tables are exceptional and peculiar in form. They are certainly very markedly non-normal in character, and the distribution of frequency is of a very curious type. Let us take, then, a table that has nothing to do with inheritance, a simple contingency table for eye-colour and hair colour. As it is desirable in such a test to keep errors of sampling as nearly negligible as possible I took the table in Livi's *Antropometria Militare* (Parte I, p. 62, 1896), the table being based on 298,860 observations. The results of the test are shown below in a small table giving the value of the normal coefficient for every possible division of the distribution. The coefficients in the first two rows do not differ greatly from each other merely because the red-haired formed under six per thousand of the population examined, and consequently it makes very little difference whether they come above or below the line of division. But it makes a great deal of difference whether we take the line above or below brown hair, above or below brown eyes. The largest value found for the normal coefficient is 0·64, the lowest 0·32 or just half the highest. Again it cannot be said that the coefficient is anything like constant for different divisions of the table.

Normal coefficients for eye and hair colour (*Livi*, 8).

Division taken for hair-colour between	Division taken for eye-colour between		
	Blue and grey.	Grey and brown.	Brown and black.
Blond and red...........................	0·53	0·52	0·49
Red and brown	0·52	0·52	0·45
Brown and black........................	0·36	0·32	0·64

47. *Age of husband and of wife.*—It might be argued further that, although we must give up the use of the coefficient (as an approximation to the correlation) for cases of heredity, and possibly

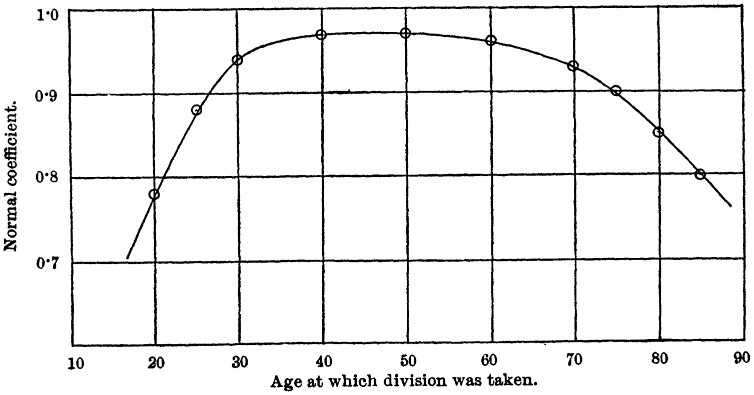
also for all cases of characters in which the differences are really qualitative (as between brown and blue), it might nevertheless hold good for cases in which the differences were merely scalar in kind, *i.e.*, in which we had to deal with purely quantitative variables. We must therefore appeal to empirical experience with ordinary correlation tables. Can we, if the correlation departs at all from normality, as is so often the case in practice, still rely upon the coefficient giving reasonably close approximations to the coefficient of correlation, *i.e.*, can we regard such tables as being to a first approximation "strained normal" in form? My next tests related accordingly to correlation tables exhibiting moderately skew distributions, in which the series of concentric ellipses that form the contour lines of the normal correlation-surface is replaced by a series of non-concentric ovoids. As before, it is desirable for such tests to choose distributions based on large numbers of observations so that the probable errors are small even when the divisions are taken towards the extremities of the distribution. I chose therefore for the first test the correlation table between age of husband and age of wife for the English Census of 1901 (Summary Tables, page 182, a condensed form of the table in 36, page 159). The table is based on 5,317,520 pairs, and errors of sampling are therefore extremely small. From the standpoint of the calculator, however, the table presents the disadvantage that the correlation is high, viz., 0·91, and the approximation to the value of the normal coefficient correspondingly slow, eight to ten or twelve terms of the equation being necessary to give a value fairly trustworthy in the second place of decimals: the supplementary tables since published to facilitate the work in such cases (Everitt 4) were not available for my work. The following were the values found,

Normal coefficients for different divisions of correlation-table between age of husband and age of wife (Census of England and Wales, 1901).

Division taken at age	Value of *h* and *k*.		Normal Coefficient.
	Husband.	Wife.	
20	−3 ·20	−2 ·62	0 ·78
25	−1 ·69	−1 ·39	0 ·88
30	−0 ·93	−0 ·73	0 ·94
40	−0 ·10	+0 ·05	0 ·97
50	+0 ·59	+0 ·74	0 ·97
60	+1 ·18	+1 ·36	0 ·96
70	+1 ·84	+2 ·06	0 ·93
75	+2 ·22	+2 ·46	0 ·90
80	+2 ·65	+2 ·92	0 ·85
85	+3 ·16	+3 ·47	0 ·80

no cross-divisions for one age of the husband and another of the wife being taken in this case: the values of h and k, the co-ordinates of the point of division from the mean of the normal distribution, in terms of the standard-deviations of the normal distribution, are also shown. It will be seen that over the centre of the range the value of the normal coefficient is in excess of the true value of the correlation, while it falls rapidly as the division is moved towards the lower ages (the modal end of the distribution) and more slowly towards the higher ages. The annexed diagram, fig. 2, shows the value of the normal coefficient as a function of the age at which the division was taken: it can hardly be said to be suggestive of any approximate constancy of the coefficient. The lowest value observed is 0·78, the greatest 0·97, a difference of nearly 0·2.

FIG. 2.—*Values of the normal coefficient for different divisions of the correlation table for age of husband and wife.*

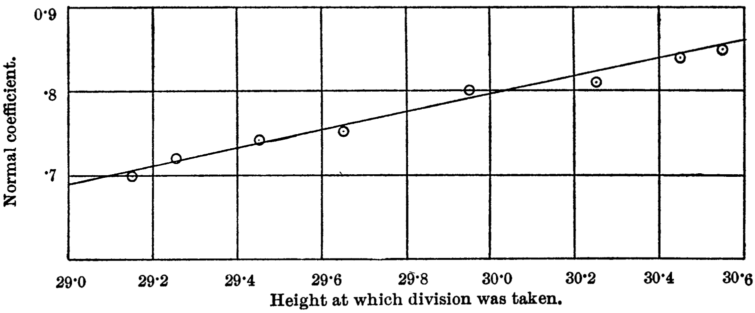


48. *Barometer heights.*—The next test case I took was the table for correlation of barometer heights at Laudale and Southampton given in the well-known memoir by Professor Pearson and Miss Lee (13). The number of observations is 2,922, and the correlation 0·76. The following were the values of the normal coefficient found for divisions taken at the same barometer height in both cases. It will be seen that the normal coefficient is highest at the end of the distribution corresponding to high barometers (the modal end of the distribution), and thence drops steadily. The course of the coefficient as a function of the barometer height at which the division was taken is nearly linear (fig. 3) and quite unlike that of the previous illustration. The range considered is not nearly so great (in terms of h and k) as in the last case, but the difference between the greatest and least values of the normal coefficient is 0·15.

Normal coefficients for different divisions of correlation table between heights of barometer at Laudale and Southampton (Pearson and Lee).

Division taken at height.	Values of *h* and *k*.		Normal coefficient.
	Laudale.	Southampton.	
30·55	+2·10	+1·86	0·85
30·45	+1·75	+1·49	0·84
30·25	+1·01	+0·85	0·81
29·95	+0·18	−0·18	0·80
29·65	−0·54	−1·06	0·75
29·45	−1·01	−1·56	0·74
29·25	−1·43	−2·05	0·72
29·15	−1·62	−2·25	0·70

FIG. 3.—*Values of the normal coefficient for different divisions of a correlation table between barometer heights at two stations.*



49. For my third trial I took a theoretical table, viz., the table for the Mendelian inheritance of number of recessive couplets in a zygote of four couplets (Pearson's theory, 22). The correlation is 1/3, and the annexed table shows the results of the calculation. The mode of the distribution lies near the upper left-hand corner, so in this case the normal coefficient for symmetrical divisions is *lowest* near the modal end, not highest, as in the last illustration.

Normal coefficients for table of Mendelian inheritance for number of recessive couplets on Pearson's theory (22): four couplets.

Offspring: division taken between	Parent: division taken between			
	0 and 1.	1 and 2.	2 and 3.	3 and 4.
0 and 1	0·39	0·37	0·35	0·33
1 and 2	0·37	0·40	0·39	0·37
2 and 3	0·35	0·39	0·41	0·40
3 and 4	0·33	0·37	0·40	0·43

It is a curious chance that the results of these three trials should have differed so markedly from each other; it would seem that we cannot even be sure, from mere inspection, in what kind of way the normal coefficient will vary for different divisions of a skew distribution. Further, the present case is particularly interesting because it would appear from it that the lowest value of the normal coefficient, for four divisions round a central class, may be the nearest to the product-sum correlation. Dr. Brownlee, two years ago (2), directed attention to the much higher values that were given by the normal coefficient for symmetrical divisions, than by the product-sum correlation, for tables of this kind.

50. *Lengths of Ivy Leaves.*—Finally, I took the table for correlation between length of one ivy leaf on a plant and length of a second, given in the memoir on "homotyposis" by Professor Pearson and others (17), but I only calculated in this case the values of the normal coefficient for four divisions round a central class. The divisions were taken at 6·95 and 14·95 eighths of inches, corresponding to values of h and k $-1·19$ and $+1·25$ respectively. For the first symmetrical division, near the modal end of the table, the normal coefficient is 0·70; for the second 0·55; the values for the cross-divisions I made 0·55 also. The product-sum correlation is given by Professor Pearson as 0·5618. Hence the normal coefficient near the modal end of the table is much too high, and the cross-divisions, or a division somewhere towards the upper end of the table, give much closer results. But if we had only been given one fourfold table we should, of course, have been quite ignorant whether our division was near the modal end or the non-modal end; or whether the table was like this and the table for barometer heights, or like the Mendelian inheritance table in form, and could not have had any confidence in the one figure.

51. *Partial Correlations Formed from the Normal Coefficient.*—It is evident enough from the preceding examples that the untrustworthiness of the normal coefficient renders it an exceedingly dangerous proceeding to form partial correlations, as has been proposed, from the normal coefficients. No reliance could be placed on partial coefficients so determined. Further, when we have only a twofold classification, it is quite a simple and straightforward matter to deal with the partial associations directly, *i.e.*, to form the tables showing the association between A and B in the universe of C's and the universe of γ's, and work out whatever coefficient of association we please for such tables. Unfortunately, however, the assumption on which alone the normal coefficient has any special value necessarily ceases to hold, strictly at least, for such partial association tables. If the association tables between

A and *B*, *A* and *C*, *B* and *C*, for the whole series of observations, represent groupings of normally distributed frequency, the table showing the association between *A* and *B* for the universe of *C*'s represents only a grouping of the frequency in one truncated extremity of the normal surface for three variables. Any complete investigation as to the relations between the normal coefficient and the correlation in such partial tables would have to be carried out by the general formulæ given by Professor Pearson in his memoir (15), and would be most laborious. I have only attempted a very limited investigation for the simplest case, and this seems to indicate, as far as it goes, that for such truncated portions of the triple-normal distribution the divergence between the normal coefficient and the true correlation, while quite appreciable, may not be so large as in many of the preceding illustrations.

52. Suppose the attributes *A*, *B*, and *C* to be determined by values of X_1, X_2, and X_3 in excess of assigned values, as on Professor Pearson's hypothesis, and let the division for *C* be taken at the median, *i.e.*, let $(C) = (\gamma) = N/2$. Let the correlations between the three variables be as usual r_{12}, r_{13}, r_{23}. Then I find for the correlation between X_1 and X_2, when X_3 is in excess or in defect of the median,

$$\rho_{12\cdot 3} = \frac{r_{12} - \frac{2}{\pi} r_{13} r_{23}}{\left(1 - \frac{2}{\pi} r_{13}^2\right)^{\frac{1}{2}} \left(1 - \frac{2}{\pi} r_{23}^2\right)^{\frac{1}{2}}},$$

a formula differing only from that for the partial correlation by the introduction of the factor $2/\pi$. If $r_{12} = r_{13} = r_{23} = 0{\cdot}5$, the value of $\rho_{12\cdot 3}$ is 0·4054. If the divisions determining *A* and *B* are taken at the medians of X_1 and X_2, the value of the normal coefficient is 0·40, in very close agreement with the correlation. The whole distribution differs in fact, in this particular case, very little from the normal distribution. I calculated out the whole table, and find the normal coefficient (for symmetrical divisions) ranging only from 0·39 to 0·43, when the division is carried from *h* and *k* = − 1·77 to *h* and *k* = + 1·75.* Apparently the divergence may be greater than this, however, in other cases. Thus, if r_{12} is 0·4, r_{13} and r_{23} 0·6, I find $\rho_{12\cdot 3} = 0{\cdot}2404$, and the normal coefficient for median division of the original variables 0·19. At present I have not been able to carry the matter further. The normal coefficient, it is evident, cannot be absolutely trusted in such a case, but as stated above, its untrustworthiness seems to be much less marked than in many other cases.

* These values of *h* and *k* refer to the truncated distribution: where I speak of the median of X_1, etc., I mean the median for the entire series, not the truncated half-distribution,

53. The work suggests also a test which, it seems to me, may be useful as shewing that some association tables cannot be regarded as representing a grouping of normally distributed frequency. For it is evident that the true correlation in a truncated portion of a triple-normal distribution must lie between the total correlation and the partial correlation; the truncated portion is neither the single array of X_1's and X_2's corresponding to values of X_3 grouped closely round a single type, nor is it the totality of such arrays, but a grouping of several of them. Hence, if the normal coefficients for the partial association tables lie very widely outside the limits given by this rule, we may be fairly certain that the correlation is not normal. An illustration will show more clearly what I mean. The following table, based on some of Dr. Warner's data used for illustrations in my memoir on association (32), show the normal coefficients between certain defects observed in school children:—*A* stands for development defects, *B* for nerve signs, *C* for low nutrition. The "total" normal coefficients are 0·54, 0·59, and 0·51. From these are calculated the partial coefficients, in the next column, in the ordinary way used for partial coefficients of correlation. The normal coefficients for the actual partial association tables, on the other hand, have the values given in the last two columns.

Defects.	Normal coefficients.		Actual normal coefficients for the partial universe of—	
	Total.	Partial.	Defectives.	Undefectives.
A and *B*........	0·54	0·34	−0·23	+0·56
A and *C*........	0·59	0·43	+0·19	+0·68
B and *C*........	0·51	0·28	+0·02	+0·59

It will be seen that the normal coefficients for the universes of defectives are in each case far less than the calculated partial coefficients, while the normal coefficients for undefective universes are even greater than the total coefficients. It does not seem to me possible to regard the frequency in such a case as a grouping of normally distributed frequency.

54. This section of my work has been a very long one, and may now be summarised. The normal coefficient has derived its repute solely from the belief that it gave the true correlation between the variables which the classification was supposed to represent; it is usually, indeed, termed "the correlation" without qualification, and the method spoken of as "Professor Pearson's fourfold-table method of determining the correlation," or some equivalent phrase. It is true that the author of the method gave several warnings as to

its untrustworthiness in his original memoir, but these seem to have been almost immediately forgotten, even by himself; he and his collaborators have repeatedly given the values as "correlations" in tables under the same headings as true product-sum correlations, and written of the values as if they were comparable with product-sum correlations. But there is no justification for the belief that the normal coefficient is in general identical with the correlation. For discrete attributes it gives the wrong value altogether. In other cases, where the attributes are supposed to represent classified values of variables, there is no justification for speaking of the method as a method of "determining" the correlation; at the best, it is an untrustworthy method of estimating it. The values of the normal coefficient not only may mislead if regarded as the correlation, but actually have misled; the values given for numerous cases of heredity are much higher than the true correlations. In view, in fact, of the results of this investigation, it is difficult to see any special use for the coefficient. It may possibly, in an occasional case, be of some service to calculate it as some indication of the value the correlation *may* take in certain purely hypothetical circumstances. It may again be of service as a short method (saving cost and time in tabulation rather than in calculation) of obtaining an approximate value for the correlation of measured characters, *provided that the applicability of the method is in the first place adequately tested*, as in a memoir by Macdonell (9). But such cases are likely to be rare. The coefficient has attained its vogue by masquerading as the correlation; as an empirical coefficient of association, the laboriousness of its method of calculation and the lack of clear meaning would have been quite sufficient to debar it from general use. In conclusion, I may perhaps add that if the attributes A and B are, in fact, continuous variables, and nothing is known about the distribution of frequency in the correlation table which the fourfold table then represents (as must be the case if only the fourfold table is given), the search for a coefficient which shall give a trustworthy approximation to the coefficient of correlation between the variables is, in my opinion, a search for the non-existent; there cannot, from the nature of the case, be any such universally applicable approximation (*cf.* the passages cited below from 32).

V. *General Synopsis, and Comments on a Paper by Dr. Heron.*

55. The work of which a description has been given in the preceding sections has, in fact, left little doubt in my own mind as to the coefficients which are of most practical service. These are the coefficient of colligation ω and the product-sum correlation r

The two form, as pointed out before, a natural pair. The latter, if the attributes are discrete and equivalent to two-valued variables, gives the correlation between these variables; in any case it may be regarded as giving the correlation between the presence of A and the presence of B. The coefficient of colligation gives the value that the correlation would take if both tables were reduced to the same form, *i.e.*, the same totals of rows and columns, by successive selection of each attribute *as an attribute* in turn. It is necessary, of course, to insert the limitation that the attribute shall be selected as such, *i.e.*, that the selection shall be the same for all sub-classes of A's and B's; if either attribute is, in fact, a variable and the selection is different for different sub-classes of A's or B's, the statement will not hold. My original coefficient of association Q possesses the same general properties as the colligation ω, and its use as such is perfectly valid, but it has not the same convenience of direct comparability with the correlation r. For investigations on small-pox and vaccination, such as those of Brownlee (3), Macdonell (10, 11) and Turner (31), the use of Q or ω would, in my opinion, have been more illuminating as well as simpler than the use of the normal coefficient; investigations such as those of Maynard (12) on anti-typhoid inoculation are in the same case. So much for my own views. Let me now, in conclusion, turn to a recent article expressing quite other convictions.

56. That any coefficient which fulfils the conditions laid down in § 7 is a valid coefficient of association in the general sense of the term, is a proposition which I should have thought it unnecessary to state. But it is apparently regarded as by no means self-evident (or, in fact, as untrue) in a paper published some months since by Dr. Heron (5), criticising two passages in my *Introduction to the Theory of Statistics* which deal with the association coefficient Q and the product-sum correlation respectively (pp. 38–9 and 212–3). The latter coefficient, as I have already pointed out, Dr. Heron failed to recognise (§ 35). The association coefficient he condemns because it differs from the normal coefficient, *i.e.*, it gives different values for different divisions of the normal correlation table, and incidentally also because it differs from the product-sum correlation. Speaking of the association and correlation coefficients Dr. Heron complains that I have "failed to apply the most obvious test of their validity, *i.e.*, to compare the results obtained by the two methods when applied to the same data." I was well aware they gave different values; I point out in the *Introduction* (p. 213) that the two coefficients possess essentially different properties, and I am unable to agree that Dr. Heron's "test of validity" is, in fact, any test. Let us apply the suggested "test" to a parallel

instance. The geometric mean and the arithmetic mean often give largely different values. The geometric mean is therefore invalid because it differs from the arithmetic mean, and the arithmetic mean, I suppose, is invalid because it differs from the geometric mean. The median is often, indeed usually, different from both, and therefore it is also invalid—and so on for the mode, the harmonic mean, and every other conceivable form of average. It is needless to insist on the character of such an argument. All forms of average are measures of analogous properties but do not give the same values. The various coefficients that have been suggested for measuring association and correlation differ in precisely the same way. As regards the difference between the association coefficient Q and the normal coefficient, I devoted two or three pages to the relation between association and correlation in my memoir (32). I gave a table showing corresponding values of Q and the coefficient of correlation for a normal distribution, when Q was determined by division of the table at the medians. I then proceeded:—"In the general case the value of Q is necessarily a function of the position of the origin, or of the arbitrary axes which are chosen for dividing A from α and B from β. The evaluation of Q for any pair of axes in the case of normal correlation depends on that of certain definite integrals which have not yet been tabulated. To get some idea of the general character of the dependence, I have calculated the value of Q for every possible pair of axes in the annexed (observed) frequency table. . . . An inspection of the table will show that Q is a minimum for axes near the mean of the whole table, and a maximum for origins near the limits." Dr. Heron devotes the greater part of his article to calculations and curves showing the variation of Q for different divisions of the normal distribution, but there does not seem to me to be any conclusion therein not essentially covered by the work of my memoir. I continue a little lower than the passage cited (*ibid.*, p. 276): "It does not seem possible to obtain for Q a function that shall not vary with the position of the axes in the general case, so long, at all events, as we adhere to certain conditions of symmetry for the function Q that seem to me almost necessary. It may perhaps be possible for a strictly normal frequency distribution." And again (p. 278): "The whole subject of the connection between correlation and association demands further investigation, as it bristles with difficulties and possibilities of fallacy. In some practical cases there seems no doubt that the signs of Q and r would be different, and, indeed, the physical meaning attached to their interpretation." The warning against confounding Q with the coefficient of correlation for a normal or any other distribution seems sufficiently emphatic, but Dr. Heron includes Q, as well as the product-sum correlation for the fourfold table, in a list of coefficients of association (*loc. cit.*, p. 120) which he appears to regard as approximations for "finding the actual value of the coefficient of correlation from a fourfold table"—approximations to the normal coefficient appears to be what is really meant. It would be as legitimate to draw up a list of the various forms of average and head it "Approximations to the Arithmetic Mean," or a list of measures of dispersion and head it "Approximations to the Standard Deviation."

57. If the fact that a coefficient of association varied for different divisions of a correlation table implied invalidity, it is evident from the work of the preceding section that the normal coefficient must be itself invalid. As soon as we leave the narrow field within which normal or "strained normal" correlation holds good, the normal coefficient fluctuates as we change the axes of division quite as largely as any other coefficient. Taking, for example, the symmetrical table for eye-colours of brothers from Professor Pearson's memoir (16)—the table that I used as one illustration in § 41 above—I find the following values for the normal coefficient ψ, the association coefficient Q, the coefficient of colligation ω, and the coefficient of correlation (the product-sum correlation) r.

Summary table showing values of the normal coefficient ψ, *the association coefficient* Q, *the colligation coefficient* ω, *and the product-sum (fourfold) correlation* r *for different divisions of the eye Colour Table Vb of Professor Pearson's memoir* (16).

Division taken for second brother between eye colours.	Coefficient.	Division taken for first brother between eye colours.			
		2 and 3.	3 and 4.	4 and 5.	6 and 7.
2 and 3	ψ	0·50	0·38	0·24	0·24
	Q	0·61	0·49	0·34	0·37
	ω	0·35	0·26	0·18	0·19
	r	0·32	0·22	0·13	0·11
3 and 4	ψ	0·38	0·52	0·35	0·28
	Q	0·49	0·62	0·46	0·40
	ω	0·26	0·35	0·24	0·21
	r	0·22	0·34	0·21	0·15
4 and 5	ψ	0·24	0·35	0·39	0·30
	Q	0·34	0·46	0·51	0·43
	ω	0·18	0·24	0·27	0·23
	r	0·13	0·21	0·23	0·16
6 and 7	ψ	0·24	0·28	0·30	0·31
	Q	0·37	0·40	0·43	0·47
	ω	0·19	0·21	0·23	0·25
	r	0·11	0·15	0·16	0·16

From this table I find the following values of the means of the given coefficients, with their ranges, the standard-deviations and and coefficients of variation (standard-deviation as a percentage of the mean):—

Coefficient.	Mean.	Range.	Standard deviation.	Coefficient of variation.
Normal coefficient	0·33	0·28	0·084	25·5
Association coefficient	0·45	0·28	0·081	18·0
Colligation coefficient	0·24	0·17	0·050	20·9
Correlation coefficient	0·19	0·23	0·066	35·0

If we judge them by the actual range of the values observed for different divisions of this table, the normal coefficient and the association coefficient are the two worst, the colligation is the best coefficient. If we judge by the standard-deviations, the normal coefficient is the worst, the colligation again the best. If we judge them by the coefficients of variation, the product-sum correlation is the worst, the normal coefficient second worst, the association coefficient best. The normal coefficient presents no advantage from the standpoint of variation.

58. But such variation is no test of validity. The view taken in my memoir, that steadiness of an association coefficient for all divisions of a correlation table is impossible, was, I think, correct. Dr. Heron, in fact, while making many numerical comparisons between the normal and other coefficients, does not apply any real test of the value or the validity of the several coefficients with which he deals. If he did not consider it necessary to confute my view, that any function fulfilling the requisite conditions (§ 7) is a valid coefficient of association, he ought to have explained why he differed from Professor Pearson, who, in his memoir (15), took precisely the same view, only regarding it as "a great advantage" (*loc. cit.*, p. 15) or "such great gain that it more than counterbalances the somewhat greater labour of calculation" (*ibid.*, p. 17) to subject the coefficient to the further condition of approximating to the value that the coefficient of correlation would take if the fourfold table represented a grouping of normally distributed frequency. Professor Pearson raised no objection then, and so far as I know has raised no objection since, to my coefficient Q; indeed, he referred to "the extreme elegance and simplicity of Mr. Yule's coefficient of association" (*ibid.*, p. 17). Dr. Heron makes no reference to this view. Again, when dealing with the product-sum coefficient, which he failed to recognise except as Professor Pearson's r_{hk}, he omitted to tell his readers that Professor Pearson (though also failing to recognise it as the product-sum coefficient)

did actually suggest its use as a coefficient of association (*ibid.*, pp. 12, 15–6).

59. Dr. Heron also makes sundry comments on certain conclusions of my illustrative work in the association memoir (*loc. cit.*, pp. 118–119), which it is of interest to discuss, as they serve to exemplify his arguments and my own. "Thus in dealing with Dr. Warner's data in the *Phil. Trans.* paper already cited, Mr. Yule gravely discusses the 'apparent law that associations were on the whole higher where populations were healthier or less defective' . . . . Had Mr. Yule used his 'theoretical value of r'* instead of the coefficient of association, it would have been suggested that the associations were on the whole *lower* where populations were healthier or less defective. Neither coefficient enables any light at all to be thrown on the question at issue." And again (*ibid.*, p. 119): "In the same way Mr. Yule's statements that 'the differences exhibited by the sexes as regards association are so marked that they can hardly have failed to have struck the reader of the foregoing tables,' and that, 'in an immense majority of cases, the associations are greater for females than for males,' must be rejected. The apparent difference arises from the fact that, 'besides being more highly associated, women are also in general less defective.'"

60. Of course, when I speak of "association" in my memoir, I mean the degree of association as measured by my association coefficient and not the association as measured by some coefficient then uninvented, like the normal coefficient; if the use of some such other coefficient should lead to other results it would in no way invalidate, though it might help to elucidate, some of the conclusions to which I came. But amongst those conclusions will not be found any conclusion to the effect that associations are, on the whole, higher amongst the more healthy populations; nor would I have expected such a law to hold, as I see no reason to regard these defects as normal or "strained-normal" variables, and would expect Q to be almost if not quite free from the influence of varying proportions of defectives. Dr. Heron, I fear, must have failed to read to the end my grave discussion of the "apparent law that associations were, on the whole, higher where populations were healthier." What I wrote was (32, p. 301): "These two tables *suggested to me at first sight* an apparent law . . . ." I then

* Dr. Heron means the product-sum coefficient. He fails to understand the following sentence in my book (p. 212): "In some cases, however, a theoretical value is obtainable for the coefficient [of correlation], which holds good even for the limiting case when there are only two rows and two columns." My reference is to cases like that of Pearson's memoir on Mendelian inheritance (22), referred to above (§ 33).

proceeded to test the "apparent law" suggested by the two tables in question on much more extensive data, and concluded (p. 304): "Thus I do not think we can accept the hypothesis without wider evidence; I have mentioned it as it occurred to me, and would probably occur to others, as covering certain of the facts presented." I also refer later (p. 311) to "my first impression (unjustified, I think) that associations were, on the whole, higher in the healthier groups." Now if differences of healthiness do not, in general, coincide with differences in the degree of association, it seems difficult to regard a difference of healthiness in one particular case as an explanation of the difference in degree of association in that case. I did not therefore feel entirely satisfied with Dr. Heron's theory as to the origin of the difference of association between the sexes, the more so as he gave no facts in support of his statement, and I accordingly worked out some coefficients to test it. I took the data used for Table I (p. 300) of my memoir, contrasting the coefficients for boys and girls in Dr. Warner's two investigations, and worked out the values of the normal coefficients and the product-sum correlations, and also added for comparison the values of ω. The results were as follows:—

Table giving the values of the association coefficient Q, colligation coefficient ω, correlation coefficient r, and normal coefficient ψ, for defects in school children, in Warner's earlier (1881—91) *and later* (1892—4) *investigation. A, development defects; B, nerve signs; C, low nutrition; D, mental dullness.*

Defects.	*Q.*		*ω.*		*r.*		*ψ.*	
	Boys.	Girls.	Boys.	Girls.	Boys.	Girls.	Boys.	Girls.
A, B	0·90	0·90	0·62	0·63	0·50	0·46	0·78	0·76
A, C	0·90	0·95	0·63	0·73	0·34	0·46	0·71	0·82
A, D	0·89	0·93	0·61	0·68	0·44	0·47	0·74	0·80
B, C	0·86	0·91	0·57	0·65	0·29	0·39	0·64	0·74
B, D	0·89	0·93	0·61	0·67	0·43	0·47	0·74	0·79
C, D	0·79	0·86	0·49	0·57	0·22	0·29	0·53	0·63
A, B	0·75	0·78	0·45	0·48	0·28	0·27	0·54	0·56
A, C	0·85	0·92	0·55	0·65	0·25	0·35	0·59	0·72
A, D	0·85	0·90	0·55	0·63	0·35	0·41	0·65	0·73
B, C	0·78	0·81	0·48	0·51	0·20	0·23	0·51	0·55
B, D	0·90	0·91	0·62	0·64	0·44	0·43	0·75	0·75
C, D	0·82	0·84	0·52	0·54	0·22	0·24	0·55	0·57

Allowing for the third place of decimals, where the coefficients are equal to the second place, the association coefficients and colligation coefficients are greater for girls than for boys in all

12 cases; the normal coefficients are greater for girls than for boys in 11 cases out of the 12; the correlations are greater for girls than for boys in 9 cases out of the 12. Both associations or colligations and correlations are, on the whole, greater for girls. Dr. Heron's statement to the contrary finds no support in these data.

61. Further, the present table afforded one of the comparisons that showed higher associations for the more defective, and not the less defective, universe, when the sex was the same. The proportions of defectives were, on the whole, less in the second series than in the first, but 10 of the 12 association coefficients were greater in the first series than in the second. The statement that is true of the association and colligation coefficients is also true for normal coefficients and correlations. The general results in this case are, therefore, much on a par with what we often find, *e.g.*, in using different forms of average for calculating index-numbers; the actual values for the averages are altered, but the general conclusions as to differences or changes remain the same.

62. Dr. Heron also objects to my conclusion that association decreases with age. His objection appears to be that the product-sum correlation does not decrease so markedly or regularly with age in one of my cases that he examined (see table below), and that no evidence has been given that the normal coefficient decreases. I never said that the correlation decreased or that the normal coefficient decreased. I said that the association coefficient decreased. The correlation, as it seems to me, does decrease also in this particular example, though its movement is more irregular than that of the association coefficient. It drops sharply from the first age-group to the second, fluctuates irregularly through middle

Association between blindness and mental defect for males (Census 1891): *Q, association coefficient (cited from* 32); *ω, colligation; r, correlation (cited from Heron* 5); *ψ, normal coefficient.*

Age.	Numbers per 100,000 in the given age class.		Q.	ω.	r.	ψ.
	Blind.	Mentally defective.				
5—	26	85	+0·92	+0·66	+0·011	+0·32
15—	44	229	+0·75	+0·45	+0·006	+0·20
25—	56	246	+0·61	+0·33	+0·005	+0·15
35—	93	572	+0·57	+0·31	+0·006	+0·15
45—	147	687	+0·46	+0·24	+0·005	+0·12
55—	247	753	+0·41	+0·22	+0·006	+0·11
65—	435	769	+0·20	+0·10	+0·003	+0·05
75—	1,051	679	−0·13	−0·06	−0·003	−0·03

life, and drops again sharply in old age. That these small figures represent perfectly appreciable intensities of the product-sum correlation r is evident when one remembers their meaning. For the second age-group, for example, the correlation is 0·0060, and this may be derived as follows: Proportion of mentally deranged amongst the blind, 16·047 per thousand; amongst the non-blind, 2·289 per thousand; difference, 13·758 per thousand, or 0·013758. Proportion of blind amongst the mentally deranged, 3·052 per thousand; amongst the non-mentally deranged, 0·430 per thousand; difference, 2·622 per thousand, or 0·002622. The correlation is the square root of the product $0·013758 \times 0·002622$, or 0·0060. Its standard error, I find from equation (43), is 0·0016, or the coefficient is 3·75 times its standard error. If we turn to the normal coefficients which Dr. Heron favours, we find that they decrease with nearly, but not quite, as much regularity as the association or colligation coefficients. But which of the coefficients possesses the clearest meaning? Consider for a moment what the assumption of normality of distribution would imply in any case where there is an increase of, say, the blind from one age-group to the next. This must imply either (1) a fall in the mean of the assumed variable character—goodness of sight, I suppose—if the standard-deviation is constant or falling, or (2) an increase of the standard-deviation if the mean is constant or rising. If the first occurs, then there must be some people in the later age-group who are *much more blind* than any people in the first, and fewer people of first-class sight; if the second, there must still be some people in the later group much blinder than any in the earlier, and there will also be some of much better sight. On the assumption that lies at the base of the normal coefficient, you cannot, in fact, effect a change in the numerical proportion of A's without changing them qualitatively at the same time. The assumption seems to me absurd, to be equivalent in this case to saying that there are certain people entirely deprived of sight in the first age-group, and certain others more than entirely deprived of sight in the second. The normal coefficient is accordingly inapplicable, and its precise values of no special significance. The groups, blind and seeing, mentally deranged and non-deranged, if not absolutely discrete, are very largely distinct from each other, either the blindness or the mental derangement having originated from some definite congenital defect, injury, or disease. That being so, the coefficients most applicable are the coefficients of correlation and of association or colligation. Of these, the colligation and association are the best, as they are largely, if not entirely, free from the influence of the increasing proportions of blind and of mentally deranged. This view is borne out, I think, by the fact that the movements of Q and ω are in a slight degree the most regular. My use of the association coefficient was, I venture to think, perfectly correct.

63. But this particular case is not a good test-case as the proportions of blind and of mentally deranged increase continuously throughout life. A better case is afforded by the data for school children in different standards, one of the cases examined by me in the memoir on association (pp. 306–7, 311). Of 24 total associations, 22 shewed a decrease on passing from one group of standards to the next higher, *although the percentages of defects in a number of cases decrease.* I give in the table below the values of Q, ω, r, and ψ for boys and girls, for a couple of cases. Q and ω decrease uniformly as we pass up through the standards, as also does the normal coefficient ψ. Only the movements of r are irregular. If we examine these irregularities, we see that the increase of r corresponds in the two cases in which it occurs to an increase in the percentage of both or one of the defects. The associations appear quite unaffected by increasing or decreasing percentages of defectives. If the reader goes into the matter more closely he will find that Q and ω show no regularity of connection between the changes in percentages of defects and the changes in the magnitude of the coefficient. Only r and the normal coefficient ψ do this; for both these coefficients the decrease in the coefficient is greatest when the

Change of association with standard in school; Q, association coefficient, cited from (32)*; ω, colligation; r, correlation; ψ, normal coefficient.*

Standards.		Percentage with defects.		Q.	ω.	r.	ψ.
		A.	*B.*				
Boys	Infants	7·9	6·3	0·79	0·49	0·27	0·56
	St. I–III...........	9·9	13·5	0·74	0·44	0·29	0·55
	St. IV–Ex. VII	7·4	8·3	0·72	0·43	0·24	0·50
Girls	Infants	7·8	4·2	0·85	0·56	0·29	0·62
	St. I–III...........	7·3	10·3	0·79	0·49	0·30	0·58
	St. IV–Ex. VII	4·2	8·8	0·75	0·45	0·20	0·49
		C.	*D.*				
Boys	Infants	3·7	6·7	0·87	0·58	0·29	0·63
	St. I–III...........	3·1	9·7	0·79	0·49	0·21	0·52
	St. IV–Ex. VII	1·5	5·4	0·75	0·45	0·12	0·41
Girls	Infants	3·9	5·3	0·88	0·59	0·30	0·64
	St. I–III...........	3·4	8·3	0·83	0·54	0·25	0·58
	St. IV–Ex. VII	2·0	4·5	0·75	0·45	0·13	0·42

percentages of both defects decrease, least when they increase. The facts support the conclusion of the last paragraph and my general view, that it is the association in my original sense of the term, or the colligation, which is least affected by changes in the proportions of defectives; the correlation tends to increase with increasing proportions of defectives, as I should, on the whole, expect, and so to a lesser degree does the normal coefficient. There is in any case no justification for the view that the normal coefficient is independent of the proportion of *A*'s and *B*'s; the view is only correct (1) if the association corresponds to "strained-normal" correlation, and this it rarely does, and (2) if the changes in proportion are effected by a change of mean and not by a change of standard-deviation, *e.g.*, by selection. It may be noted in conclusion that for this table 5 of the 6 associations, 5 of the 6 colligations, 5 of the 6 correlations, and 5 of the 6 normal coefficients are greater for girls than for boys.

64. I should like to have continued this paper for a little longer and to have given a few more illustrations, but it has already extended to more than an ordinary length, and I must bring it to a close. I have for some years been desirous of resurveying the problems of association, in which I have always felt a special interest since the publication of my memoir on that subject some twelve years ago. Professor Pearson's memoir, which immediately followed mine, was a remarkable piece of mathematical work, for which, as mathematical work, I have a great admiration, but the value of the method therein suggested depends entirely on the empirical truth of the assumptions made, on which alone the "normal coefficient," as I have termed it, becomes identical with the correlation. Those assumptions were never adequately tested, even in cases in which such a test was called for and easy of application, and the few tests which I applied some years since in connection with other work (*cf.* 34) sufficed to show that they were, to say the least, of exceedingly doubtful validity. Later on, the product-sum correlation for the fourfold table had been introduced in a somewhat curious way for a special field of work, when its possible applications were of a much wider character, and this it was necessary to emphasise. The situation has therefore been a particularly difficult one for a writer of a text-book, and it was for this reason that I did not treat the subject of measures of association at any greater length in my *Introduction to the Theory of Statistics.* I felt myself justified, after consideration, in giving a brief statement concerning the association coefficient in order, as stated, "to direct the attention of the student to the possibility of forming such a measure of association" (p. 38); I likewise found it not merely desirable but

necessary, as already stated above (§ 35), to give the product-sum correlation. The normal coefficient is quite unsuitable for treatment in an elementary course, but I gave full references, more than once, to the memoir concerned. I have slightly, but only slightly, expanded the passages in question in the second edition; the fuller treatment must, I am afraid, wait for a while, as my second edition went to the printers before the present paper. In the meantime I hope that this discussion may do something to place the measurement of association on a less hypothetical and more rational basis.

References.

1. Boas, F. Letter on "The determination of the coefficient of correlation," *Science*, vol. 29, p. 823, May 21, 1909.
2. Brownlee, J. "The significance of the correlation coefficient when applied to Mendelian distributions," *Proc. Roy. Soc. Edin.*, vol. xxx, p. 473, 1910.
3. Brownlee, J. "Statistical studies in immunity: small-pox and vaccination," *Biometrika*, vol. iv, 1905, p. 313.
4. Everitt, P. F. "Tables of the tetrachoric functions for fourfold correlation tables," *Biometrika*, vii, 1910, p. 437, and Supplementary Tables, *ibid.*, vol. viii, 1912, p. 385.
5. Heron, D. "The danger of certain formulæ suggested as substitutes for the correlation coefficient," *Biometrika*, vol. viii, p. 109, 1911.
6. Johannsen, W. *Elemente der exakten Erblichkeitslehre*, Fischer, Jena, 1909.
7. Lipps, G. F. "Die Bestimmung der Abhängigkeit zwischen den Merkmalen eines Gegenstandes," *Berichte d. Math.-Phys. Kl. d. kgl. Sächsischen Ges. d. Wissenschaften*, February, 1905.
8. Livi, R. "*Antropometria militare*," Parte I, Rome, 1896.
9. Macdonell, W. R. "On criminal anthropometry and the identification of criminals," *Biometrika*, vol. i, 1902, p. 177.
10. Macdonell, W. R. "On the influence of previous vaccination in cases of small-pox," *Biometrika*, vol. i, 1902, p. 375.
11. Macdonell, W. R. "A further study of statistics relating to vaccination and small-pox," *Biometrika*, vol. ii, 1903, p. 135.
12. Maynard, G. D. "A statistical study of anti-typhoid inoculation," *Biometrika*, vol. vi, 1909, p. 366.
13. Pearson, K., and Alice Lee. "On the distribution of frequency (variation and correlation) of the barometric height at divers stations," *Phil. Trans.*, A, vol. 190, 1897, p. 423.
14. Pearson, K., and L. N. G. Filon. "On the probable errors of frequency constants," *Phil. Trans.*, A, vol. 191, 1898, p. 229.
15. Pearson, Karl. "On the correlation of characters not quantitatively measurable," *Phil. Trans. Roy. Soc.*, A, vol. 195, 1900, p. 1.
16. Pearson, Karl, and Alice Lee. "On the inheritance of characters not capable of exact quantitative measurement," *Phil. Trans. Roy. Soc.*, vol. 195, A, 1900, p. 79.

17. Pearson, Karl, and others. "On the principle of homotyposis and its relation to heredity, etc.," *Phil. Trans.*, A, vol. 197, 1901, p. 285.

18. Pearson, Karl. "The law of ancestral heredity" (with appendices on inheritance in racehorses, by N. Blanchard and A. Lee), *Biometrika*, vol. ii, p. 211, 1903.

19. Pearson, Karl, and others (editorial). "On the probable errors of frequency constants," .*Biometrika*, vol. ii, 1903, p. 273.

20. Pearson, Karl, with Alice Lee. "On the laws of inheritance in man," *Biometrika*, vol. ii, p. 357, 1903.

21. Pearson, Karl. "On the theory of contingency and its relation to association and to normal correlation," *Drapers' Co.'s Research Memoirs*, Biometric Series, vol. i, Dulau and Co., 1904.

22. Pearson, Karl. "On a generalised theory of alternative inheritance, with special reference to Mendel's laws," *Phil. Trans. Roy. Soc.*, A, vol. 203, p. 53, 1904.

23. Pearson, Karl. "On the inheritance of the mental and moral characters in man and the comparison with the inheritance of the physical characters" (Huxley Lecture), *Journ. Anthrop. Inst.*, vol. xxxiii, 1903, and *Biometrika*, vol. iii, p. 131, 1904. [References to the latter.]

24. Pearson, Karl, Amy Barrington, and Alice Lee. "On inheritance of coat-colour in the greyhound," *Biometrika*, vol. iii, p. 245, 1904.

25. Pearson, Karl, and Amy Barrington. "On the inheritance of coat-colour in cattle," *Biometrika*, vol. iv, 1906, p. 427.

26. Pearson, Karl. "On certain points connected with scale order in the case of the correlation of two characters which for some arrangement give a linear regression line," *Biometrika*, vol. v, p. 176, 1906

27. Pearson, Karl. "On the ancestral gametic correlations of a Mendelian population mating at random," *Proc. Roy. Soc.*, B, vol. 81, p. 225, 1909.

28. Pearson, Karl (reply to Boas, 1). Letter on "The determination of the coefficient of correlation," *Science*, vol. 30, p. 23, July 2nd, 1909.

29. Snow, E. C. "On the determination of the chief correlations between collaterals in the case of a simple Mendelian population mating at random," *Proc. Roy. Soc.*, B, vol. 83, p. 37, 1910.

30. Snow, E. C. "The application of the correlation coefficient to Mendelian distributions," *Biometrika*, vol. viii, p. 420, 1912.

31. Turner, F. M. "On the correlation between vaccination and small-pox in the London epidemic, 1901–02," *Biometrika*, vol. iv, p. 483, 1906.

32. Yule, G. U. "On the association of attributes in statistics," *Phil. Trans. Roy. Soc.*, A, vol. 194, 1900, p. 257.

33. Yule, G. U. "On a property which holds good for all groupings of a normal distribution of frequency for two variables, with applications to the study of contingency tables for the inheritance of unmeasured qualities," *Proc. Roy. Soc.*, A, vol. 77, p. 324, 1906.

34. Yule, G. U. "On the influence of bias and of personal equation in statistics of ill-defined qualities," *Journ. Anthrop. Inst.*, vol. 36, p. 325, 1906 (Abstract in *Proc. Roy. Soc.*, A, vol. 77, p. 337, 1906).

35. Yule, G. U. "On the theory of inheritance of quantitative compound characters on the basis of Mendel's laws," *Report of the Conference on Genetics* (*Royal Horticultural Soc.*), 1907.

36. Yule, G. U. *An introduction to the theory of statistics*, Griffin and Co., London, 1911.

DISCUSSION *on* MR. YULE'S PAPER.

PROFESSOR EDGEWORTH, in proposing a vote of thanks to the reader of the Paper said that it afforded new guidance to statisticians in the investigation of causal connection between phenomena. He was prepared to find that Mr. Yule's *Colligation* would prove to be *the* method for treating such problems. The very term "colligation" had a felicity unintended by the author. It was the term employed by Whewell to denote the achievement of the scientist when he read a new idea into the facts; as Kepler had connected the observations of the earth's position by the appropriate conception of an ellipse, so Mr. Yule had read a new idea into the dry categories of formal logic. He had extracted a fresh lesson from the "hard-worked illustration" afforded by bags and balls. Experiments with bags and balls and the kindred problems about games of chance had not without good reason fascinated Pascal and subsequent mathematicians. There was a mysterious connection between those apparently trivial phenomena and the deep first principles of Induction. Mr. Yule's use of such illustration was specially adapted to *discrete* data which he had properly distinguished from continuous statistics of frequency. A trial which resulted in the presentation of one or other of two attributes, say white and black, was appropriately compared to an observation which admitted of only two values, a frequency-locus consisting of two points, say $x=0$, $x=1$. Now let them complicate the illustration by introducing a second logical division, and supposing some of the balls in the bag to be *marked* and the rest *plain*. Each observation might now present any one of *four* values: say $x=0$, $y=0$ (corresponding to a ball that was at once white and plain); $x=1$, $y=0$ (black and plain); $x=0$, $y=1$ (white and marked; $x=1$, $y=1$ (black and marked). The records of extractions from such a medley would present a (discontinuous) frequency-surface; for which they might calculate (the averages and) the "product sum." But he demurred, not now for the first time [*Journal*, vol. lxxi (1908) p. 511], to Mr. Yule's doctrine that the method of the product sum "had been freed from any necessary relation to the theory of normal correlation." He (Professor Edgeworth) thought that the propriety of the product-sum in the case of the discontinuous surface (consisting of four points) just mentioned and other abnormal frequency-surfaces consisted in this that the coefficient so determined led directly to the *correlation-coefficient* for a certain *normal* surface; namely, the surface which was formed by the *averages* of large batches taken from the supposed medley. If they extracted (in random fashion) a hundred or a thousand balls from the imagined bag (with due replacement after extraction, unless indeed the bag was supposed to be indefinitely large) and marked a point in the plane of x,y such that the abscissa represented the proportion of white balls (to the total number) in the batch, and the ordinate the proportion of marked balls; and they similarly proceeded with another and another equal batch; the frequency of the observations thus constituted would be approximately normal. This theory had nothing to do with the assumptions of normality in natural phenomena which might be unjustified, as Mr. Yule had insisted. *This* normality was manufactured by the statistician averaging independent items, not furnished ready made by Nature. The appearance of the normal surface with a genuine coefficient of correlation was a great fact; but one which proved less serviceable than might have been hoped for the present purpose; since the coefficient varied with variations in the data which might be regarded as accidental—such as a difference in the proportion of the vaccinated to the total number of persons under observation. To remedy this imperfection, Mr. Yule had proposed an ingenious method. Whereas the averages, or proportions in the long run of black and of marked balls respectively might be any fractions between o and I corresponding to any point within the little square that they had marked out; Mr. Yule in effect so operated upon the balls extracted from the given bag as to construct a new medley, an ideal normal surface of which the centre coincided with the centre of our little square. From given statistics as to vaccination and recovery he constructed an ideal hospital, in which the numbers of the vaccinated and unvaccinated were equal, and likewise the number of deaths from all causes and of recovery were equal. How should they regard this beautiful construction? As a conception or a fact, a colligation merely or an induction, a useful convention or a profound truth? Before attempting to answer these questions, he (Professor Edgeworth) required more time for reflection. He would, in conclusion, advert very briefly to the large part of the Paper which dealt with errors—errors of coefficients, and on the part of the persons who calculated them. With regard to the probable or "standard" errors of frequency-constants, it must always be remembered that, as he had pointed out before now (*Journal*, vol. lxxi, 1908), the computation was based on the hypothesis of *independent* observations. But concrete observations were surely often far from being perfectly independent. There was a correlation between adjacent observations, like consecutive bad strokes at golf. But though the determination of the *absolute* values of the errors might be precarious, the *relative* values of the errors incident to different methods might well be important. As to the controversial part of the Paper he was not prepared to take sides. He admired the work of Mr. Yule without condemning that of others.

Mr. SANGER, in seconding the vote of thanks, said this Paper included all the qualities which a Paper on theoretical statistics was expected to contain; it dealt with a very important subject in a most interesting way, and contained quite new and original matter of a valuable kind—namely, the coefficient of colligation—together with a due proportion of controversial matter. The greater part, in fact, was controversial and was devoted to attacks on rival statisticians of equal eminence. That was always expected, and he was glad to find it there. The importance of Mr. Yule's work on association, which might be said to culminate in this Paper, had not always been sufficiently recognised. To his mind the work began with his memoir, which was only read by experts, and it was made known to the small class of people who had the great advantage of attending Mr. Yule's Newmarch lectures, and was eventually made known to a larger public by his book published last year, which he ventured to prophesy would be recognised as the standard work on the elements of statistical theory. That work on association was of fundamental importance. Its importance consisted in this, that all statisticians before Mr. Yule had this passion for the normal curve which Professor Edgeworth had displayed. He thought that passion required explanation, but it could be explained. The whole difficulty arose in this way: that the theory of statistics was developed from the theory of errors of observation which involved the normal curve, and the difficulty was that mathematicians wished to be able to deal with things which were not discrete. Speaking as a mathematician, he had always felt this prejudice very strongly, and liked to have a formula which gave rise to interesting mathematics. It was Mr Yule who first shook his faith by pointing out that it was very important, when dealing with discrete characters, to get rid of this assumption of normality. When one came to a perfectly simple case like that considered—whether people were vaccinated or unvaccinated, and either died or did not die—or ordinary simple cases where one wished to get some measure of correlation or association, the first way to proceed was to make no assumption and see what was the most reasonable measure one had. That was the fundamental business of statisticians. They collected a mass of figures and the first thing was to reduce them to some kind of order; the next thing was to ascertain whether there was some kind of causal connection between them. They were not interested simply in the abstract figures, but they wanted to know was there or was there not some causal relation between vaccination and the death-rate from small-pox. To find whether the two sets of figures were connected they wanted to form some kind of coefficient. Certainly Mr. Yule's original Q possessed all the properties one wanted, and his new coefficient ω possessed the same properties and *primâ facie* had the same merits. It had the slight demerit, to his mind, that it was rather more difficult to calculate, but that could be got over, because one could construct a table which gave the corresponding values of Q and ω, and then there was no further difficulty. It had the additional merit that it corresponded to another very well-known thing, the product-sum coefficient, on the assumption that the table of four-fold division was equally divided. Professor Edgeworth was rather pained at the process by which the tables were reduced, but, if he understood Mr. Yule correctly, the coefficient ω was not got by any such process. He pointed out that it happened to be the same as the product-sum when these processes were used. He thought it was very important to get this question settled, and should have been rather glad if the Paper had only consisted of the first three sections, so that attention should be centred on that point. He saw no point in the normal coefficient when used for the four-fold table. As Mr. Yule pointed out, when dealing with discrete quantities, when the thing was simply either black or white, it was obviously inapplicable, and when dealing with quantities not discrete one had to test first of all whether it was applicable or not. He had apparently tested a large number of cases, and found in every one that it was not, though there might be cases in which it was. But it all came back to the mathematician's prejudice again; he expected nature to range itself according to a normal curve, but, unfortunately, facts did not arrange themselves so; they were finding that out time after time; consequently they must abandon completely the use of the normal coefficient, not only for discrete quantities, but even where there was no good reason for supposing that the correlation was on the whole normal. One additional reason why he welcomed the Paper was that the rise of Mendelian biology had made a great difference. There they were always dealing with things which were discrete, whereas according to all Galtonian laws they always dealt with things which were thought to be continuous. At present there was this difficulty, that mathematicians had a prejudice in favour of more elegant mathematics, and the Mendelians had not yet learnt algebra; but that day would come, and then Mr. Yule's work would be the work for Mendelians. With regard to the last section of the Paper—the controversy with Dr. Heron—there was this to be said. It was very unfortunate that there were no fixed standardised terms. If one talked of a correlation coefficient for a four-fold table the normal coefficient was one which naturally came into one's head. Dr. Heron had been accustomed to work a great deal with that, and, naturally, not reading Mr. Yule very carefully, he misinterpreted him. Possibly the adjective Mr. Yule used misled him. At any rate, it seemed to him it was a kind of slip they were all liable to fall into, and he did not really attribute very much blame to him.

Mr. R. H. HOOKER said that if there was one thing more than another that had been impressed upon his students by Mr. Yule, it was the fundamental difference in the methods required in the treatment of attributes and variables. The whole treatment of the former had been entirely ignored by statisticians until Mr. Yule some twelve years ago wrote his classical paper on the association of attributes; and no one had, he thought, done anything worth mentioning on the subject since. They could almost feel grateful to the article in *Biometrika* which had led Mr. Yule to thrash this question out thoroughly anew and to examine the coefficients which might be used to measure the association of attributes. As this subject was so absolutely and entirely Mr. Yule's he would like to suggest that it would be preferable not to introduce any new term such as colligation—although that word was perfectly legitimate and appropriate—but to call his new coefficient ω the coefficient of association simply. The old form seemed destined to drop out of use; it had scarcely been used; and he thought that the author of the coefficient might with perfect propriety substitute a corrected form under the same name without creating confusion. The idea of association was now getting very well known; it was simpler (and a simpler term always conveyed a more definite meaning) than colligation. Such a course would hasten the disuse of the less satisfactory form of the coefficient, and save confusion among the less mathematical statisticians who already had a rather unnecessary number of coefficients to measure the same thing.

Dr. GREENWOOD said he should like to associate himself with what had been said as to the extreme value of this piece of work, and if he raised one or two questions it was really with the object of eliciting some further information from Mr. Yule. To begin with, he should like to take a point perhaps rather of dialectics, and ask if Mr. Yule had been quite fair to Professor Pearson's point of view regarding the question of continuity. He was not at all clear that one could dismiss, for instance, Professor Pearson's argument with respect to vaccination cases with the jibe about people being either dead or not dead. It seemed to him what was expressed in the passage cited from Professor Pearson was, that when one had two classes, the deaths of persons who had had small-pox and the recoveries, one might regard those as being susceptible of reduction to a continuous form; of course, if one knew nothing else, Mr. Yule's other criticism that the hypothesis was unverifiable was just. But there was some other evidence. Taking, for instance, the statistics published by Dr. Brownlee in one of his papers (*Biometrika*, vol. v, p. 433), a multiple classification was adopted. Dr. Brownlee classified the cases, in accordance with the form of cutaneous eruption, as confluent, very much pocked, sparse eruption, and very sparse eruption, and he also had an extreme series of hæmorrhagic cases, which might be regarded as an extreme form of confluent. If one looked at the figures, although the distribution was not a normal curve, it suggested a part of one of those "cocked-hat" curves with which they were so familiar in biological statistics, which might be not badly represented by a normal curve. He did not put that forward as being anything like conclusive, but he did submit that one could not sum up this question of vaccination and small-pox as being obviously a case of discrete compartments; he did not think that was at all clear. In the same way, if one studied the other two classes (vaccinated and unvaccinated), it was not an easy thing to say whether a person had been vaccinated or not. It was no doubt an easy thing on paper, but even as to the actual facts of vaccination the criterion was not a very good one. One had all gradations of efficiency of vaccination, and that seemed to be a question which wanted further consideration, the possibility that here also a continuous distribution might arise. Turning to another point, which he hoped some person more competent in mathematics would refer to, he must confess to having some doubt as to the value of the standard error of a constant as a measure of its distribution in every case. Suppose one had data given like those of small-pox and vaccination. One had a sample of M people, p_1 of whom had died of a certain disease, and all those M people had been treated in some particular way. One had another sample of N people, p_2 of whom had died of a certain disease; the question was with regard to that second sample, supposed not to be immunised, whether the death-rate was higher than in the first, significantly. That question could be treated by one of those four-fold tables, and measured by one of the resultant coefficients, but how would it have been looked at before anything had been learned about association or correlation coefficients at all? Assume that nothing whatever was known about such functions, how would that problem have been looked at, say at the end of the eighteenth century? A case like that would have been considered on the basis of Bayes' theorem. Returning to the stock case of a bag of balls, one had taken out from an unlimited bag a limited number of balls of which a certain proportion were black. One had taken from another part of the bag another sample with a different proportion of black balls amongst them; and the problem was: Was that second sample consistent with the first? Was it possible that the "climate" of the other part of the bag had had some effect on the colour of the balls? That was a problem for treatment by the method of Bayes' theorem as modified by Condorcet and La Place. To take an example: suppose one had 100 persons who had been vaccinated, and none of them died of small-pox, and suppose one had another lot of 50 persons, not having been vaccinated, of whom two died of small-pox. From such a table it would be found that three of the coefficients discussed by Mr. Yule became unity, while the fourth, the product moment coefficient, was ·164 with a standard error of about ·06. Using this standard error as an argument and *assuming that random deviations were distributed normally*, that would mean that one had over 300 chances to one against the variables not being correlated; but if one looked at the problem from another point of view and considered what was the likelihood if a sample of 100 gave no successes that a second sample of 50 would give two or more successes, this proved to be (using Bayes' theorem) about one in 10. Looking at the question in one way it would appear that the probability of correlation was extremely high; from another standpoint, the results were not at all inconsistent with the two samples being homogeneous and vaccination without influence upon the chance of dying (the illustration is, of course, imaginary). It seemed to him that the explanation of the apparent paradox was that when one had very skew tables then, precisely as in the case of the so-called probable errors of simple percentages, the fluctuation in the resulting coefficient could not be tested by anything approaching a normal curve. When one had found the coefficient and determined its standard deviation, the latter was no real measure of the reliability of the former. There was perhaps a danger in the use of these coefficients, especially by untrained persons, in these extreme cases. There was nothing whatever in the publications dealing with the different coefficients to prevent the average man concluding that when one had one of the four squares blank one got a coefficient of association equal to unity without any probable error. Of course, common sense showed the average man that this was ridiculous, but he was not told what he had to do in those circumstances, and what value the resultant coefficient possessed as a measure of association.

Dr. E. C. SNOW expressed his opinion of the service Mr. Yule had rendered in writing the Paper, but desired to make a few criticisms of his new coefficient of colligation. The author had referred (§ 33) to a paper of his (Dr. Snow's) and this presented an opportunity of comparing the value of r, which he agreed was the correct coefficient to employ in such a case where perfectly discrete variates were being used, with the corresponding values of Q and ω. Mr. Yule had given the value of r for the table referred to as ·42; the values of Q and ω for the same table were ·77 and ·46. But the given table was only a particular case of a far more general one. If in the general table $4p$ was put equal q the values of r, Q and ω were ·47, ·78 and ·48 respectively. In these two cases, then, ω compared favourably with r. But if p was put equal $4q$ the values were ·33, ·91 and ·64. Thus ω might differ very considerably from r. In the limit it might be as much as three times as large. At the other extreme they could compare the values of Q and ω with those of r found for a case of continuous variation from tables of many rows and columns by reducing the latter in various ways to four-fold tables. Mr. Yule had worked out the normal coefficient and compared it with the true value for a number of cases from Professor Pearson's memoirs, but had not brought into comparison his own Q and ω. Taking first the case of inheritance of stature (15, p. 40), a case in which the frequency distributions were as nearly normal as was ever likely to be attained—the *true* value of the correlation was ·52, and Professor Pearson had found that the normal coefficients for six distinct divisions of the table were ·59, ·56, ·55, ·53, ·52 and ·55, these not differing much from ·52. He (Dr. Snow) found that the corresponding values of Q were ·70, ·66, ·69, ·64, ·65 and ·71, all of them much too high. The values of ω were ·41, ·38, ·40, ·36, ·37 and ·42, these being much too low. Mr. Yule had worked out the normal coefficient for four reductions of a very large table from another of Professor Pearson's memoirs (§ 50). A glance at the original table showed that the distributions were far from normal, and the true value of the correlation found from the unreduced table was ·56. Mr. Yule gave the normal coefficients for the four reductions of the table as ·70, ·55, ·55 and ·55. He (Dr. Snow) found the values of Q to be ·85, ·75, ·92 and ·75, and of ω, ·56, ·45, ·66 and ·45. ω then might differ considerably from the true correlation, and on either side of it. With reference to the Mendelian tables referred to in the Paper, Mr. Yule quoted with apparent approval the use of the normal coefficient by Dr. Brownlee, but he (Dr. Snow) thought that the use of that coefficient in the simple Mendelian case in which the presence or absence of a discrete entity was being dealt with, could not be legitimately defended. He pointed out, too, a simple and very useful trigonometrical relation between Q and ω which Mr. Yule had not mentioned. If in equation (15) Q was put equal sin θ, then ω was at once seen to be tan $\frac{\theta}{2}$. This enabled the one coefficient to be found very quickly from tables when the other was known. He could not agree with Mr. Yule that in future work the two coefficients to be used were r and ω, and he thought that not nearly enough examples of the use of ω had been given. He was of the opinion that in the case of discrete variates r only should be employed, and in the case of continuous variation he saw no evidence that ω was going to improve upon the normal coefficient as an approximation to the correlation.

Mr. YULE, in reply, said that he should like some time to consider the remarks of Professor Edgeworth. With regard to the concluding remarks of Mr. Sanger, it was quite possible that Dr. Heron had some reason for misunderstanding the paragraph in his book concerning the product-sum coefficient for a four-fold table if he read this paragraph without any of the context, but such a proceeding was hardly legitimate. The suggestion put forward by Mr. Hooker, that he should transfer the name "coefficient of association" to the new "coefficient of colligation" had been actually considered when he was writing the Paper, but he had thought it better on the whole to use a new term. With regard to Dr. Greenwood's comments, he quite recognised that something might be said in defence of regarding death from small-pox, in the same way as Professor Pearson, as a *function* of some variable, but it seemed to him that, in adopting the assumption that they were dealing with a variable, they were making an unnecessary hypothesis and were going outside their data. Further, it seemed to him that when Professor Pearson headed the columns for "recoveries" and "deaths" in his association table for small-pox and vaccination with the words "Strength to resist small-pox when incurred" these words implied assumptions that were not altogether justified. Did death or recovery really indicate "strength to resist small-pox when incurred"? It seemed to him that the result was a function of several things—*e.g.*, the strength of the attack as well as the properties of the patient. Again, was the second heading "degree of effective vaccination" a justifiable substitution for "cicatrix present" and "cicatrix absent"? He believed that the great bulk of the cases included under "cicatrix absent" were cases in which vaccination had never been performed and that the two classes were largely distinct. If the cases classed under "cicatrix present" were dealt with alone, he agreed that they might be regarded as giving a frequency distribution of "degree of effective vaccination," but he did not think that this was the important point if the two classes were for the most part distinct. When they passed from, say, Sheffield to a place like Leicester, where there was a strong anti-vaccinationist movement, they largely reduced the proportion of the people who were vaccinated, but there was no reason to suppose that this altered the form of the frequency distribution for "degree of effective vaccination" amongst the vaccinated or the "cicatrix present" cases. Professor Pearson's assumption, which lay at the basis of the normal coefficient, was that in altering the proportion they altered the distribution, that they could not alter the ratio of the number of *A's* to *non-A's* without at the same time altering the qualitative composition of the group of *A's*. Dr. Greenwood also raised an interesting point as to the use of Bayes' theorem for measuring association. He was rather inclined to think that the use, as a measure of association, of the improbability that a certain difference could have arisen as a fluctuation of sampling, confused two distinct things. They should always keep separate, as it seemed to him, (1) the magnitude of the association, (2) the probability that it might or might not have arisen as a fluctuation of sampling. If they had a very small difference between the death-rates for, say, vaccinated and unvaccinated, a coefficient of association should indicate that such difference was small; but they had only to make the number of observations big enough to give any desired degree of improbability that such death-rates could arise in random samples from the same population. He quite agreed, however, that it was very necessary to bear in mind the non-normality of the distribution of errors. Turning to Dr. Snow's remarks, he would point out in the first place that r was necessarily less than ω, as repeatedly indicated in the Paper, and might be indefinitely less; ω was not an approximation to r. He quite realised that the coefficients of association and of colligation for different divisions of the same table in many cases fluctuated more largely than the normal coefficient, but he thought he had made it clear that he did not attach any importance to such fluctuation. He had considered one particular case in his reply to Dr. Heron because that writer seemed to attach great importance to it, and he himself wished to emphasise that the normal coefficient, in the case of certain tables for which it had been largely used, fluctuated as much as other coefficients. It was quite possible that in many cases the normal coefficient might be the best approximation they could get to the value the coefficient of correlation would take if the four-fold table were replaced by a manifold table, but the fact remained that it was quite an untrustworthy approximation. As indicated in the Paper, he did not think it was possible to arrive at any trustworthy approximation; Dr. Snow's concluding remarks were based on some misapprehension of his reasons for recommending ω. He was glad to know that Dr. Snow thought that the product-sum correlation for the four-fold table would prove useful, but he himself thought that ω would prove useful too—indeed, in such cases as small-pox and vaccination more useful than r itself—but he admitted that this remained to be tested. As regards Dr. Snow's recent comments in *Biometrika* on the use of the normal coefficient for Mendelian tables in Dr. Brownlee's paper, he really thought that those comments were a much stronger condemnation of Professor Pearson's than of Dr. Brownlee's work. Professor Pearson had repeatedly used the normal coefficient for inheritance tables, that were in all probability representations of Mendelian inheritance, as if it were an approximation to the product-sum correlation. Dr. Brownlee, as he read his paper, used the normal coefficient simply as empirical coefficient of association, and clearly distinguished it from the product-sum correlation. He by no means agreed accordingly with Dr. Snow in his condemnation of Dr. Brownlee's paper. That paper seemed to him valuable because it was there first pointed out that the normal coefficient did not give the product-sum correlation for Mendelian tables.

The following candidates were elected Fellows of the Society:—

J. M. Ramsay.	**W. F. Sheppard.**