Biometrika Trust

A Short Method of Calculating the Coefficient of Correlation in the Case of Integral Variates

Author(s): J. Arthur Harris

Source: *Biometrika*, Vol. 7, No. 1/2 (Jul. – Oct., 1909), pp. 214–218

Published by: [Biometrika Trust](#)

Stable URL: <http://www.jstor.org/stable/2345370>

Accessed: 17/06/2014 23:46

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Biometrika Trust is collaborating with JSTOR to digitize, preserve and extend access to *Biometrika*.

<http://www.jstor.org>

	Below 1·10	1·10 1·22	1·22 1·34	1·34 1·46	1·46 1·58	1·58 1·70	1·70 1·82	1·82 1·94	1·94 2·06	2·06 2·18	2·18 2·30	2·30 2·42	2·42 2·54	2·54 2·66	2·66 2·78	Above 2·78	
Actual ...	—	4	8	7	14	12	12	5	7	5	2	1	2	0	1	—	
Calculated {	No Corre- lation	1·01	2·42	5·28	8·86	11·69	13·07	12·18	9·97	6·90	4·27	2·36	1·18	—	—	·92	—
	Corre- lation ·01	1·85	3·27	6·02	8·92	11·01	11·71	10·84	8·95	6·64	4·52	2·82	1·64	·90	—	·85	—

These give $P = .46$ and $P = .86$ respectively, the first being a good deal helped by the convention that the tail should not be carried beyond the point at which a single unit may be expected, and the second much less so.

As the empirical curve fitted from the actual moments has a P of .92, the second curve may be considered fairly good depending as it does on a guess following on calculation. On the other hand a P of .46 with so few cases as 80 is not particularly good, and as Prof. Pearson has pointed out to me the graph distinctly gives an idea of greater skewness than is represented by the no correlation curve. I do not however wish to contend that the circumstances attending the production of the sample actually conformed to the arbitrary conditions which I found it necessary to assume in order to simplify the analysis. But seeing that the fit is good and that with such a small sample even the divergent B_1 is not altogether impossible, I think it likely that there was some sort of correlation, though probably not that particular kind which has been assumed in this note.

Conclusions:

- (1) That the approach to normality of the distribution of means of samples drawn from a non-Gaussian population is delayed by the existence of correlation between the individuals composing the samples.
- (2) That on certain arbitrary assumptions the constants of the new distribution can be found given the constants of the old one and r according to formulae given above.
- (3) That using the above formulae and choosing a likely looking value of r , a curve can be drawn to represent the sample in Drs Greenwood and White's paper with fair likelihood.

II. A Short Method of Calculating the Coefficient of Correlation in the Case of Integral Variates.

By J. ARTHUR HARRIS, PH.D., Cold Spring Harbor, Long Island, U.S.A.

For symmetrical correlation tables in which both variates have the same mean and standard deviation, Professor Pearson has suggested* that his difference method "may possibly be of good service," but warns the reader:—

"At the same time too much reliance must not be placed upon the difference method, not only because it assumes normality of distribution but because it involves a somewhat rough method of approximation in the case of the diagonal cell."

* *Drapers' Company Research Memoirs, Biometric Series*, iv. pp. 4—9, 1907.

In applying this difference method to series of material already worked out by the product moment method I have found considerable variation in the weight to be given to the diagonal cell, but many more series should be examined before suggesting any other value than the one-sixth proposed by Professor Pearson. While engaged in this work some modifications of method which, I think, will lighten considerably the calculation of r in some cases came to my attention.

It is well known that the formula

$$r_{xy} = (\sigma_x^2 + \sigma_y^2 - \sigma_v^2) / (2\sigma_x\sigma_y) \dots\dots\dots (i),$$

where σ_x and σ_y are the standard deviations of the two characters and σ_v is the standard deviation of their difference, gives r with the same accuracy as the product moment method irrespective of the nature of the distributions.

In symmetrical tables, in which each individual is used once as a first and once as a second member of a pair, $\sigma_x = \sigma_y$ and the above formula may be written

$$r_{xx} = \frac{2\sigma_x^2 - \sigma_v^2}{2\sigma_x^2} = 1 - \frac{1}{2} \frac{\sigma_v^2}{\sigma_x^2} \dots\dots\dots (ii).$$

From this formula the correlation coefficient may be calculated with great ease and rapidity as follows :

First, determine the positive differences between the first and second member of the pairs from the table, as suggested by Professor Pearson* and as illustrated in the example given below.

After the totals of the columns have been found they are multiplied by the squares of the differences as given at the heads of the several columns. Twice the sum of these products divided by N gives σ_v^2 , for the origin is at 0, the plus and minus deviations are equally great and the first moment necessarily 0.

The standard deviation of the character will usually have been obtained for other purposes but if it is not wanted no roots need be extracted to obtain r by this method. The multiplication of the totals of the several columns by the squares of the differences which they represent requires only a little more work than their multiplication by the numbers themselves, and if a Brunsviga or Comptometer be used all of the arithmetical work can be done in a few minutes, even when the tables are rather large. Thus the whole work can be completed almost if not quite as quickly as that for Pearson's formula (vii)† and with the same accuracy as the product moment method.

Since all may not have access to Pearson's memoir referred to, I give the following illustration of the arithmetic of the method.

Illustration I. Calculation of correlation from symmetrical table. Dr Fernando De Helguero's symmetrical table for number of flowers per inflorescence in *Cicorium Intybus*‡ serves as an example. Beginning at the head of each column in the correlation table we copy down the frequencies to and including the diagonal cell—where the difference between the two variates is 0—under appropriate headings in a series of difference columns. The work can be done systematically and rapidly by copying the first number of the first column in the 0 column. In beginning the copying of each succeeding column from the correlation table a new column of differences, one higher than the last, will be begun and be represented by a single entry for that column while the remaining entries will be placed successively one place to the left. The whole process will be clear at once from a comparison of the work of the illustration with the table of data. The numbers in clarendon type show the order in which the entries from the 12th column were written down.

* *loc. cit.* p. 7.

† *loc. cit.* p. 6.

‡ *Biometrika*, Vol. v. p. 188.

The same value is obtained by the product moment method, the difference between this and the $r=.613$ obtained by De Helguero being due to the slip in his calculation of the standard deviation. It appears from the above that the calculation of r is a very simple process.

TABLE I.
Inflorescences of Cicorium.

	8	9	10	11	12	13	14	15	16	17	Totals
8	—	1	1	1	2	—	—	—	—	—	5
9	1	2	2	15	10	1	2	—	—	—	33
10	1	2	8	19	11	4	1	—	—	—	46
11	1	15	19	134	90	18	1	—	—	—	278
12	2	10	11	90	114	97	9	4	—	—	337
13	—	1	4	18	97	122	53	2	—	—	297
14	—	2	1	1	9	53	96	10	—	—	172
15	—	—	—	—	4	2	10	10	3	—	29
16	—	—	—	—	—	—	—	3	2	1	6
17	—	—	—	—	—	—	—	—	1	—	1
Totals	5	33	46	278	337	297	172	29	6	1	1204

ν = Difference of x and y .

0	1	2	3	4	5	6	7	8	9	
0	1	1	1	2	0	0	0	0	0	
2	2	15	10	1	2	0	0	0	—	
8	19	11	4	1	0	0	0	—	0	
134	90	18	1	0	0	0	—	0		
114	97	9	4	0	0	—	0			$S(\nu^2)$
122	53	2	0	0	—	0				276
96	10	0	0	—	2					224
10	3	0	—	4						180
2	1	—	20							64
0	—	56								50
—	276									—
488										794

$N=1204,$

$\sigma_x=1.3352*, \sigma_x^2=1.7828,$

$\sigma_{\nu^2}=\frac{2 \times 794}{N}=1.318937,$

$r=\left(1-\frac{1}{2} \frac{\sigma_{\nu^2}}{\sigma_x^2}\right)=\left(1-\frac{1}{2} \frac{1.318937}{1.7828}\right)=.630.$

In working with a symmetrical table we copy down only the diagonal cell and the positive differences. But it would be very easy to continue right down the columns: after passing the diagonal cell the differences would take the negative sign and we could calculate the standard deviation from all the differences instead of from a part of them. In the symmetrical table this would be merely a waste of time, but where the tables are not symmetrical (and the two variates have different means and standard deviations) this process can be used to advantage.

* The value 1.3536 given by De Helguero was apparently obtained by taking the root of the second rough moment, i.e. without first subtracting the square of the first rough moment.

In fact I believe that in the case of integral variates the calculation with formula (i) will be considerably more rapid than the conventional product moment method.

A further saving of time can be effected by doing away with the necessity for copying the entries from the correlation table into difference columns. This can be made clear by an illustration.

Illustration II. Calculation of correlation from non-symmetrical tables. Pearl's table for his Series V of *Ceratophyllum**, showing correlation between the position on the branch and the number of leaves per whorl, is given below with the differences between the two variates indicated above each entry as small indices. These can be entered on a table—preferably in red

TABLE II.
Whorls of Leaves in Ceratophyllum.

		Position of Whorl on Branch.										Totals
		1	2	3	4	5	6	7	8	9	10	
Leaves per Whorl.	4	2 ³	2 ²	— ¹	— ⁰	— ⁻¹	— ⁻²	—	—	—	—	4
	5	17 ⁴	5 ³	1 ²	2 ¹	— ⁰	1 ⁻¹	1 ⁻²	—	—	—	27
	6	16 ⁵	6 ⁴	10 ³	1 ²	1 ¹	1 ⁰	— ⁻¹	1 ⁻²	—	—	36
	7	40 ⁶	29 ⁵	18 ⁴	14 ³	9 ²	1 ¹	2 ⁰	4 ⁻¹	1 ⁻²	—	118
	8	19 ⁷	28 ⁶	25 ⁵	17 ⁴	8 ³	7 ²	3 ¹	— ⁰	3 ⁻¹	1 ⁻²	111
	9	5 ⁸	20 ⁷	19 ⁶	26 ⁵	22 ⁴	17 ³	13 ²	6 ¹	4 ⁰	5 ⁻¹	137
	10	7 ⁹	5 ⁸	6 ⁷	4 ⁶	6 ⁵	10 ⁴	8 ³	14 ²	4 ¹	3 ⁰	67
	11	1 ¹⁰	— ⁹	— ⁸	— ⁷	— ⁶	— ⁵	— ⁴	— ³	— ²	— ¹	1
Totals		107	95	79	64	46	37	27	25	12	9	501

ink or pencil to avoid confusion—much more quickly than the products of the deviations of the two variables from their assumed origins, and with much less chance of error, especially in the case of large tables. The work can be most rapidly done by taking first the diagonal cell and then the diagonal rows lying on either side of it. All the entries of any diagonal row will have the same index, but it should be put down in every case to avoid confusion. All that now remains to be done is to sum these diagonal columns and calculate the second moment. For this r can be at once calculated from formula (i) above.

From our illustrative table we have for the sums of the diagonal rows :—

ν	-2	-1	0	1	2	3	4	5	6	7	8	9	10	Total
Frequency ...	4	13	10	17	47	64	90	102	91	45	10	7	1	501

Note that a check for the accuracy of the work is furnished by the identity of the total with the N of the population. In the product moment method there is no convenient check for the accuracy of the multiplications leading to the $S(xy)$. In the present case we find

$$\begin{aligned} \text{S. D. Position} &= 2.3733, \\ \text{S. D. Leaves} &= 1.4003, \\ \text{S. D.}^2 \text{ Difference} &= 4.4380, \\ r &= .475, \end{aligned}$$

agreeing with the value given by Pearl.

* Pub. 58, *Carnegie Institution of Washington*, p. 60, Table 31, 1907.

Obviously this method of obtaining the totals of the difference columns can be used for the symmetrical tables as well, where only the diagonal cell and half of the remaining diagonal rows need be considered, thus doing away with the work required by the method of copying the frequencies from the correlation table into difference columns as suggested by Pearson.

Unfortunately these difference methods are limited in their applicability. When integral variates are under consideration, they give the same result as the product moment method when Sheppard's Correction is not applied in the calculation of the standard deviations. But difficulties arise when other than integral variates are considered. The divergence of the coefficients of correlation calculated by different methods is shown by Wright, Lee and Pearson*. But where the limits of their applicability are borne in mind, I think, these difference methods deserve more consideration in practical work than they have hitherto received.

III. Note on Variation in *Adoxa*.

By J. ARTHUR HARRIS, PH.D., Cold Spring Harbor, L.I., U.S.A.

In examining Whitehead's paper† on the *Moschatel*, certain points came to my attention which seem to merit more detailed consideration than he has given them. What is really needed is a more extensive quantitative investigation than is possible on the data hitherto published. Possibly the indication of certain points of interest not discussed by Whitehead may induce some one who has the opportunity of collecting fresh material to treat the problems in the detailed way which they deserve.

Two questions occur to one examining Whitehead's second table :

(a) Is there any relationship between the number of flowers which an inflorescence produces and the characteristics of these flowers?

(b) What is the similarity of the flowers of an inflorescence?

Both of these, I take it, are of considerable interest to morphologists.

Problem (a). The correlation between the number of flowers per inflorescence and the number of the divisions of the corolla.

In our almost complete ignorance of factors influencing the number of parts of the flower, it seems important to investigate every possible interrelationship between the characteristics of the flower and the characteristics of the individual which produces it. Such investigations may not be expected to yield more than a small part of the information which we desire, but before experimental methods are extensively applied, it seems logical to determine whether there is any dependence of the number of floral parts developed, or any dependence of any characteristics of these floral parts, upon the vegetative organs of the individuals which produce them.

In the present case only one characteristic other than those belonging to the flower itself is available for comparison, i.e. number of flowers per inflorescence. This character seems, however, a particularly interesting one for study. In the *Moschatel* there is one form of inflorescence which is conspicuously the modal one. This type is a very neat, compact, morphological structure and any departure from it appears at once to be "abnormal." The flowers, too, have a normal form, from which a considerable number of deviations are to be found. Are abnormalities in the number of lobes of the corolla associated with abnormalities in the number of flowers borne on the inflorescence? The data for determining the coefficients of correlation are extracted from Whitehead's Table II and presented in the correlation tables

* *Biometrika*, Vol. v. p. 410.

† *Biometrika*, Vol. II, pp. 108—113, 1902.