SPECIAL ARTICLES

ON THE COEFFICIENT OF CORRELATION AS A MEASURE OF RELATIONSHIP

THE theory of correlation deals with the relationship between variable quantities in the case where that relationship lies somewhere between functional dependence and complete independence. In the case of normal correlation for two variables a certain quantity r, which is zero for complete independence and ± 1 for functional dependence, plays an important rôle. The formula for r, in terms of n observed pairs of values of two variables x and y, is

$$(1) \qquad r = \frac{\sum_{i=1}^{i=n} (x_i - x_0)(y_i - y_0)}{\sqrt{\sum_{i=1}^{i=n} (x_i - x_0)^2 \cdot \sum_{i=1}^{i=n} (y_i - y_0)^2}},$$

where x_0 is the mean of the x-values and y_0 the mean of the y-values.[1] This formula has also been given an interpretation for the case of skew correlation[2] which makes r an important quantity in many instances of such correlation.

The quantity r is usually termed the coefficient of correlation and is said to measure the amount of correlation between the variables x and y. This latter statement is too vague as it stands for scientific procedure, so it is desirable to state more precisely what is meant by it. In the case of normal correlation r has been shown to have the following significance:[3] if we take the mean of all the y's corresponding to a given value of x, then the deviation of this mean from the mean of *all* the y's, divided by the standard deviation of the y's, is equal to r times the deviation of the given x-value from the mean of all the x's, divided by the standard deviation of the x's. Thus r may be said to measure the tendency of a given deviation from the mean in one of the variables to be associated with an average deviation from the mean of corresponding magnitude in the other variable.

It is clear that the value of r throws much light on the relationship between two variable quantities in the case of normal correlation. It is not apparent, however, that it gives us in every instance the information we are most interested in obtaining, and it will be shown in what follows, that in certain cases of interest in the applications of the theory of correlation it will not necessarily give it.

The formula (1) is well adapted to the computation of r from observed values of x and y. For our purposes, however, we need a formula which exhibits r as a function of the underlying variable quantities that determine x and y and the relationship between them. We shall now proceed to obtain such a formula on the basis of assumptions similar to those that Pearson used in his derivation of (1).[4]

Let

$$(2) \qquad \begin{aligned} x &= f_1(\epsilon_1, \epsilon_2, \cdots, \epsilon_m), \\ y &= f_2(\epsilon_1, \epsilon_2, \cdots, \epsilon_m), \end{aligned}$$

where the ϵ's are independent variables that follow a Gaussian distribution, and the f's are analytic functions. If we expand the right-hand members of (2) about the mean values of the ϵ's and neglect higher powers than the first,[5] we have

$$(3) \qquad \begin{aligned} x - x_0 &= a_{11}\eta_1 + a_{12}\eta_2 + \cdots + a_{1m}\eta_m, \\ y - y_0 &= a_{21}\eta_1 + a_{22}\eta_2 + \cdots + a_{2m}\eta_m, \end{aligned}$$

where the η's are deviations of the ϵ's from their mean values and x_0 and y_0 are mean values of x and y, respectively.

Since the ϵ's are independent variables following a Gaussian distribution, we have

[1] *Cf.* Pearson, "Regression, Heredity and Panmixia," *Philosophical Transactions of the Royal Society*, 187 *A* (1896); also Bravais, "Analyse mathématique sur les probabilités des erreurs de situation d'un point," Académie des Sciences: Mémoires présentés par divers savants, Ser. 2, Vol. 9 (1846).

[2] *Cf.* Yule, "On the Significance of Bravais's Formulæ for Regression, etc., in the Case of Skew Correlation," *Proceedings of the Royal Society*, Vol. 60 (1897).

[3] *Cf.* Pearson, *l. c.*

[4] *L. c.*

[5] Pearson assumes that the variations of the ϵ's from their mean values are small in comparison with those values, in order to justify the dropping of higher powers. It is more general to assume merely that for the range of values of the ϵ's considered, the f's are sufficiently good approximations to linear functions to warrant the neglect of higher powers.

$$\sum_{v=1}^{v=n} \eta_i^{(v)} \eta_j^{(v)} = 0 \qquad (i \neq j),$$

where

$$(\eta_i', \eta_j'), (\eta_i'', \eta_j''), \cdots, (\eta_i^{(n)}, \eta_j^{(n)})$$

are n pairs of values of η_i and η_j. Hence, substituting in (1) the values of $(x - x_0)$ and $(y - y_0)$ given by (3), we obtain

$$r = \frac{\sum_{i=1}^{i=m} a_{1i} a_{2i} s_i^2}{\sqrt{\sum_{i=1}^{i=m} a_{1i}^2 s_i^2 \cdot \sum_{i=1}^{i=m} a_{2i}^2 s_i^2}}, \tag{4}$$

where the s's are the standard deviations of the ϵ's. The formula (4) for r is well adapted to the discussion of the connection between the value of r and the relationship between x and y. We shall use it first to show that under certain conditions r will not furnish a satisfactory measure of the particular form of relationship in which we are interested.

Consider, for example, the use of correlation in educational investigations. A value for r is computed from the performances of a group of persons in two fields of mental activity, such as two school subjects, and the closeness of relationship between the two fields or subjects is discussed on the basis of this value. It is clear that the value of r is a good measure of the tendency of the members of the group having a given deviation from the mean ability of the whole group in one field, to have an average deviation of corresponding magnitude from the mean ability of the whole group in the other field. It is certainly useful to be able to measure such a tendency, but there is something else which it is more useful from the educational standpoint to be able to measure. Suppose the average ability of the whole group in one field is increased a certain amount by training in that field, and this in turn causes a certain increase in the average ability of the whole group in the other field. The ratio of this latter increase to the former, when each is measured in terms of the standard deviation of the group in the corresponding field, is a very important quantity in educational investigations; it is vital for example in the discussion of such questions as disciplinary values.

We will now proceed to show that under certain conditions this ratio may be much greater than r. Since ability in any complicated field of mental activity like a school subject may be regarded as a function of a great many elementary abilities, the abilities x and y in two subjects may be represented as in equation (2). If we expand about the mean values of the ϵ's at any given time and neglect higher powers than the first,[6] we get equations of the type (3).

Since ability in each of the two subjects will in general depend on certain elementary abilities not involved in the other, we shall consider a case where certain of the a's in the first equation in (3) are zero and certain of the a's in the second equation are zero. Let us suppose then that

$$\begin{aligned} a_{11} = a_{12} = \cdots = a_{1p} &= 0, \\ a_{2,\, m-p+1} = a_{2,\, m-p+2} = \cdots = a_{2m} &= 0, \end{aligned} \tag{5}$$

and let us suppose further that

$$\begin{cases} a_{j,\, p+1} = a_{j,\, p+2} = \cdots = a_{j,\, m-p} = a > 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad (j = 1, 2), \\ a_{1,\, m-p+1} = a_{1,\, m-p+2} = \cdots = a_{1m} = 100a, \\ a_{21} = a_{22} = \cdots = a_{2p} = 100a, \\ s_1 = s_2 = \cdots = s_p = s, \\ m = 902p. \end{cases} \tag{6}$$

If by training in one subject the average ability of a group of persons in that subject is increased a certain amount, it is reasonable to suppose that this increase has been uniformly distributed in the way of corresponding increases in each of the elementary abilities involved in that subject. Since from (6) the standard deviations of the elementary abilities are all equal in the present case, a uniform distribution of the increase would imply an equal increase in each elementary ability. We will assume then that after training in the subject, the mean value of each ϵ of which x is a function is increased by a quantity δ. Since the η's occurring on the right-hand side of the first equation in (3) are deviations from the original means of the ϵ's, the mean value of each of them will now be δ instead of zero.

[6] In the present instance we neglect higher powers on the assumption that ability in the given subject is approximately a weighted mean of the elementary abilities on which it depends.

Hence, in view of (5) and (6), the mean value of x will be

$$x_0' = x_0 + 1{,}000\ p\ a\ \delta. \tag{7}$$

Similarly the mean value of y after the increase in the average value of the ϵ's involved in x, the ϵ's involved in y but not in x remaining constant, will be

$$y_0' = y_0 + 900\ p\ a\ \delta. \tag{8}$$

Therefore, since the standard deviations of x and y, s_x and s_y, are equal,

$$\frac{y_0' - y_0}{s_y} \Big/ \frac{x_0' - x_0}{s_x} = 0.9. \tag{9}$$

It is apparent from (9) that in this instance a certain increase in the average ability x will be accompanied by an increase almost as great in the average ability y.

If r is to be considered in all cases a reliable measure of the closeness of relationship between two fields of mental activity, it ought to be approximately equal to the ratio in (9). Let us see what its value actually is. Making use of equations (4), (5) and (6), we get

$$r = \frac{900pa^2}{\sqrt{(900pa^2+10{,}000pa^2)(10{,}000pa^2+900pa^2)}} = 0.08 \text{ approximately.} \tag{10}$$

We have dealt here with a special case, but it is easy to see from the above discussion that in many other cases we would have discrepancies of the same sort. Hence it is apparent that it is not safe to assume off-hand that r is always the best measure of the relationship between two fields of mental activity. It may be a very poor measure of the form of relationship in which we are interested.[7]

The question naturally arises, under what conditions will r be a good approximation to the value of the ratio in (9)? It is the purpose of the rest of this paper to obtain certain sufficient conditions that this will be the case. It is very easy to see that if all the a's of equation (3) which are not zero are equal to each other in absolute value, and furthermore if the standard deviations of the ϵ's are all equal to each other, r will be exactly equal to the ratio in (9). This leads one to suspect that if these conditions are fulfilled to a sufficient degree of approximation, r will not differ very much from this ratio.

In discussing the general case there are really two ratios of the type (9) to be considered, according as the training has been in the field corresponding to x or in the field corresponding to y. In the special case discussed above these two ratios were identical, so we only considered one of them. Under the hypotheses we shall make in what follows, the discussion for one ratio is practically the same as the discussion for the other, so here too we shall only consider one of them.

We will investigate first the case where all the a's on the right-hand side of the equations in (3) are positive or zero. It is apparent that there is no loss of generality in supposing that the a's which are zero in the first equation are the a's of the first p terms and the a's which are zero in the second equation are the a's of the last q terms. In particular p, or q, or both of them, might be zero.

Since the standard deviations of the ϵ's involved in x are no longer necessarily equal to each other, a uniform distribution over these ϵ's of an increase in x would result in an increase in each ϵ proportional to its standard deviation. Let us suppose then that after training in the field corresponding to x the mean value of each ϵ_v involved in x has been increased by an amount $s_v\delta$. Representing as before by x_0' the mean value of x after the increase in the ϵ's we have

$$x_0' - x_0 = \sum_{v=p+1}^{v=m} a_{1v}s_v\delta. \tag{11}$$

Similarly, if y_0' represents the mean value of y after the increase in the ϵ's, we have

$$y_0' - y_0 = \sum_{v=p+1}^{v=m-q} a_{2v}s_v\delta. \tag{12}$$

[7] We have restricted ourselves in the foregoing discussion to the case of relationship between different fields of mental activity. The mathematical part of the discussion, however, will undoubtedly have a bearing on many applications of the theory of correlation. If for any two variables x and y, the a's of equation (3) satisfy the conditions of our special case, the ratio of the common factors involved in the variation of x and y to all the factors, will, for each variable, be 0.9. Hence r, which is given by (10), will not be a good measure of the closeness of relationship between the two variable quantities.

Hence we have

$$R=\frac{y_0'-y_0}{s_y}\Big/\frac{x_0'-x_0}{s_x}$$

$$=\frac{\sum_{v=p+1}^{v=m-q}a_{2v}s_v}{\sum_{v=p+1}^{v=m}a_{1v}s_v}\cdot\frac{\sqrt{\sum_{v=p+1}^{v=m}a_{1v}^2s_v^2}}{\sqrt{\sum_{v=1}^{v=m-q}a_{2v}^2s_v^2}}. \tag{13}$$

Let us now suppose that two positive quantities a and s, and a positive quantity $\rho<1$, exist, such that

$$\begin{aligned}&a(1-\rho)\leqq a_{1i}\leqq a(1+\rho)\quad(i=p+1,p+2,\cdots,m),\\ &a(1-\rho)\leqq a_{2i}\leqq a(1+\rho)\quad(i=1,2,\cdots,m-q),\\ &s(1-\rho)\leqq s_i\leqq s(1+\rho)\quad(i=1,2,\cdots,m).\end{aligned} \tag{14}$$

It follows readily from (13) and (14) that

$$\left(\frac{1-\rho}{1+\rho}\right)^4\frac{m-p-q}{\sqrt{(m-p)(m-q)}}<R<\left(\frac{1+\rho}{1-\rho}\right)^4\frac{m-p-q}{\sqrt{(m-p)(m-q)}}. \tag{15}$$

Similarly from (4) and (14) we have

$$\left(\frac{1-\rho}{1+\rho}\right)^4\frac{m-p-q}{\sqrt{(m-p)(m-q)}}<r<\left(\frac{1+\rho}{1-\rho}\right)^4\frac{m-p-q}{\sqrt{(m-p)(m-q)}}. \tag{16}$$

We might obtain still narrower limits for the values of R and r than those given in (15) and (16). It is apparent from the limits obtained, however, that if ρ is sufficiently small, r will furnish a good approximation to the value of R.

We will now consider the case where some of the a's on the right-hand side of the equations in (3) are negative. Let us suppose that the first λ of the $(m-p-q)$ η's that appear in both equations have coefficients of the same sign in the two equations, and that the remainder, μ in number, have coefficients of opposite signs. Obviously, an increase in x that is uniformly distributed with regard to the ϵ's involved in x, will be accompanied by a decrease in those ϵ's for which the corresponding η's have negative coefficients in the first equation in (3); also an increase in an η having a negative coefficient in the second equation will cause a corresponding decrease in the value of y. Hence we have for the ratio in (9)

$$R=\frac{\frac{y_0'-y_0}{s_y}}{\frac{x_0'-x_0}{s_x}}$$

$$=\frac{\sum_{v=p+1}^{v=p+\lambda}|a_{2v}|s_v-\sum_{v=p+\lambda+1}^{v=m-q}|a_{2v}|s_v}{\sum_{v=p+1}^{v=m}|a_{1v}|s_v}\cdot\frac{\sqrt{\sum_{v=p+1}^{v=m}a_{1v}^2s_v^2}}{\sqrt{\sum_{v=1}^{v=m-q}a_{2v}^2s_v}}. \tag{17}$$

Let us now suppose that the a's and the s's satisfy equations of the type (14), *i. e.*, equations obtained by replacing the a's in those equations by their absolute values. Then it is easy to see that if $\lambda>\mu$, and ρ is sufficiently small,

$$\begin{aligned}&\frac{\lambda-\mu}{\sqrt{(m-p)(m-q)}}\cdot\frac{(1+\rho^2)(1-\rho)^2}{(1+\rho)^4}\\ &\quad-\frac{2(\lambda+\mu)}{\sqrt{(m-p)(m-q)}}\cdot\frac{\rho(1-\rho)^2}{(1+\rho)^4}<R\\ &<\frac{\lambda-\mu}{\sqrt{(m-p)(m-q)}}\cdot\frac{(1+\rho^2)(1+\rho)^2}{(1-\rho)^4}\\ &\quad+\frac{2(\lambda+\mu)}{\sqrt{(m-p)(m-q)}}\cdot\frac{\rho(1+\rho^2)}{(1-\rho)^4}.\end{aligned} \tag{18}$$

Furthermore, in view of (4), we have for r

$$\begin{aligned}&\frac{\lambda-\mu}{\sqrt{(m-p)(m-q)}}\cdot\frac{1+6\rho^2+\rho^4}{(1+\rho)^4}\\ &\quad-\frac{4(\lambda+\mu)}{\sqrt{(m-p)(m-q)}}\cdot\frac{\rho(1+\rho^2)}{(1+\rho)^4}<r\\ &<\frac{\lambda-\mu}{\sqrt{(m-p)(m-q)}}\cdot\frac{1+6\rho^2+\rho^4}{(1-\rho)^4}\\ &\quad+\frac{4(\lambda+\mu)}{\sqrt{(m-p)(m-q)}}\cdot\frac{\rho(1+\rho^2)}{(1-\rho)^4}.\end{aligned} \tag{19}$$

The corresponding inequalities for the cases where $\lambda\leqq\mu$ are easily obtained. It follows from (18) and (19) or the corresponding inequalities, that r will be a good approximation to R if ρ is sufficiently small.

The case where all the a's on the right-hand side of (3) that are not zero, are negative, does not seem to have any great interest in connection with the applications discussed in this paper. In any event the treatment of that case presents no new difficulties, so we shall not consider it here.

This paper makes no pretense of being an exhaustive treatment of the subject under consideration. Its main object has been to point out as briefly as possible the danger of assuming that the coefficient of correlation is necessarily a satisfactory measure of all forms of relationship between two variable quantities, and at the same time to suggest a method of attack for determining in what way a particular relationship depends on the value of this coefficient.

CHARLES N. MOORE
UNIVERSITY OF CINCINNATI

AN ABERRANT ECOLOGICAL FORM OF UNIO COMPLANATUS DILLWYN

THE variety of *Unio complanatus* Dillw. which is here described was found at Songo Pond, about three miles south of Bethel, Me. The specimens from which it is described were collected in August, 1913. The pond is a headwater of the Crooked River, one of the larger tributaries of the Presumpscot. It lies in a glacial scoop in alluvial sand, and is fed by springs mainly. A small brook a mile long enters it also. The country rock is a granitic gneiss of the eastern range of Montalban gneisses, and the intrusive granites scattered here and there are of the same mineralogy. There is no limy rock in any form within many miles, a fact which will account for the peculiar structure of the shell. The specimens were picked up on a very gently sloping beach of round-grained sand, along the western shore of the pond, and in about two feet of water. The pond is about a mile and a quarter long, from north to south, and averages a quarter of a mile in width.

So far as I can determine, the soft parts of the animal are in every way normal for the species. The aberrancy occurs in the valves, and is in structure and in shape.

The largest of my specimens, and the largest I have seen in the course of eight summers' picking, measures two and three quarters by one and a half inches over all. The greatest thickness, from umbo to umbo, is three quarters of an inch. The following features are normal: hinge size and place, umbo size, place and shape, lateral and pseudocardinal teeth size and shape, scars, pallial line, and sculpture. Epidermis is of normal color, but thicker than usual, and overlaps the edge of the hard part of the shell up to 3/32 of an inch, being most extended at the siphonal region and along the anterior part of the ventral edge in many specimens.

The shape of the shell is almost identical with that of *Anodonta marginata* Say, being roughly rhomboidal. It does not resemble the specimens of *Unio complanatus* from other regions in the American Museum at New York, in this respect. From the posterior end of the hinge, the dorsal edge slopes ventrally, straight, at an angle between 35 and 40 degrees from the line of the hinge. This portion of the edge is nearly straight and about as long as the hinge. It rounds off into the small semicircle of the posterior end. In mature specimens there may be a slight flattening of the posterior end at the point where the mantle forms a pair of siphons by its folding and coherence, but this is not constant and I find it only in the largest specimens. The ventral edge is not a uniform curve, but approaches more or less to three straight lines, equal in length, each making an angle of about ten degrees with the line continuing the edge beyond it. The anterior end has the usual graceful elliptical outline, forming a large curve from hinge to ventral edge.

There are no rays visible on any of my specimens.

The most peculiar feature of the shell is the exceedingly small amount of mineral matter in it. When fresh the shells are horny and somewhat flexible, not unlike two layers of parchment pasted together, in texture. Alcoholic material and fresh are alike easily cut with a small shears, and there is no cracking. The thin nacreous layer breaks into small angular chunks, which adhere to the epidermis. I found only the faintest traces of a prismatic layer, in the largest specimens. Smaller ones fail entirely to show it. In my largest specimens there is at the umbo a larger amount of mineral matter, but even here it is hardly more in amount than at the margin in the normal shell of this species. The epidermis seems to me to be nearly twice as thick as in the normal type. In many specimens I found grains of