"On the Significance of Bravais' Formulæ for Regression, &c., in the case of Skew Correlation."

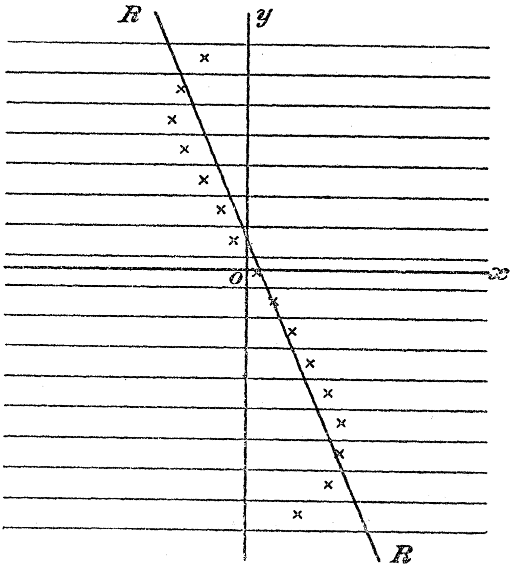
By G. UDNY YULE. Communicated by Professor KARL PEARSON, F.R.S.

Received December 14, 1896,—Read February 18, 1897.

The only theory of correlation at present available for practical use is based on the normal law of frequency, but, unfortunately, this law is not valid in a great many cases which are both common and important. It does not hold good, to take examples from biology, for statistics of fertility in man, for measurements on flowers, or for weight measurements even on adults. In economic statistics, on the other hand, normal distributions appear to be highly exceptional: variation of wages, prices, valuations, pauperism, and so forth, are always skew. In cases like these we have at present no means of measuring the correlation by one or more "correlation coefficients" such as are afforded by the normal theory.

It seems worth while noting, under these circumstances, that in ordinary practice statisticians never concern themselves with the form of the correlation, normal or otherwise, but yet obtain results of interest—though always lacking in numerical exactness and frequently in certainty. Suppose the case to be one in which two variables are varying together in time, curves are drawn exhibiting the history of the two. If these two curves appear, generally speaking, to rise and fall together, the variables are held to be correlated. If on the other hand it is not a case of variation with time, the associated pairs may be tabulated in order according to the magnitude of one variable, and then it may be seen whether the entries of the other variable also occur in order. Both methods are of course very rough, and will only indicate very close correlation, but they contain, it seems to me, the point of prime importance at all events with regard to economic statistics. In all the classical examples of statistical correlation (*e.g.*, marriage-rate and imports, corn prices and vagrancy, out-relief and wages) we are only primarily concerned with the question is a large x usually associated with a large y (or small y); the further question as to the form of this association and the relative frequency of different pairs of the variables is, at any rate on a first investigation, of comparatively secondary importance.

Let Ox, Oy be the axes of a *three dimensional* frequency-surface drawn through the mean O of the surface parallel to the axes of measurement, and let the points marked ⓧ be the means of successive x-arrays, lying on some curve that may be called the curve of regression of x on y. Now let a line, RR, be fitted to this curve,



subjecting the distances of the means from the line to some minimal condition. If the slope of RR is positive we may say that large values of x are on the whole associated with large values of y, if it is negative large values of x are associated with small values of y. Further, if the slope of RR to the vertical be given we shall have a measure of a rough practical kind of the shift of the mean of an x-array when its type y is altered. The equation to RR consequently gives a concise and definite answer to two most important statistical questions. It is also evident that if the means of the arrays actually lie in a straight line (as in normal correlation), the equation to RR must be the equation to the line of regression.

Let n be the number of observations in any x-array, and let d be the horizontal distance of the mean of this array from the line RR. I propose to subject the line to the condition that the sum of all quantities like nd^2 shall be a minimum, *i.e.*, I shall use the condition of least squares. I do this solely for convenience of analysis; I do not claim for the method adopted any peculiar advantage as regards the probability of its results. It would, in fact, be absurd to do so, for I am postulating at the very outset that the curve of regression is only exceptionally a straight line; there can consequently be no meaning in seeking for the *most probable* straight line to represent the regression.

Let x, y be a pair of associated deviations, let σ be the standard deviation of any array about its own mean, and let

$$\mathrm{X} = a+b\mathrm{Y}$$

be the equation to RR. Then for any one array

$$\mathrm{S}\{x-(a+by)\}^2 = \mathrm{S}\{x-(a+b\mathrm{Y})\}^2 = n\sigma^2+nd^2.$$

Hence, extending the meaning of S to summation over the whole surface

$$\mathrm{S}(nd^2) = \mathrm{S}\{x-(a+by)\}^2-\mathrm{S}n\sigma^2.$$

But in this expression $\mathrm{S}(n\sigma^2)$ is independent of a and b, it is, in fact, a characteristic of the surface. Therefore, making $\mathrm{S}(nd^2)$ a minimum is equivalent to making

$$\mathrm{S}\{x-(a+by)\}^2$$

a minimum. That is to say, we may regard our method in another light. We may say that we form *a single-valued* relation

$$x = a+by$$

between a pair of associated deviations, such that the sum of the squares of our errors in estimating any one x from its y by the relation is a minimum. This single-valued relation, which we may call the characteristic relation, is simply the equation to the line of regression RR. There will be two such equations to be formed corresponding to the two lines of regression.

The idea of the method may at once be extended to the case of correlation between several variables x_1, x_2, x_3, &c. Let n be the number of observations in an array of x_1's associated with fixed values X_2, X_3, X_4, &c., of the remaining variables, let σ_1 be the standard deviation of this array, and let d be the difference of its mean from the value given by a regression equation

$$\mathrm{X}_1 = a_{12}\mathrm{X}_2+a_{13}\mathrm{X}_3+a_{14}\mathrm{X}_4+ \ldots\ldots$$

Then, as before, we shall determine the coefficients a_{12}, a_{13} a_{14}, &c., so as to make $\mathrm{S}nd^2$ a minimum. But this is again equivalent to making

$$\mathrm{S}\{x_1-(a_{12}x_2+a_{13}x_3+a_{14}x_4+\ldots.)\}^2$$

a minimum for

$$\mathrm{S}\{x_1-(a_{12}x_2+a_{13}x_3+a_{14}x_4+\ldots.)\}^2 = \mathrm{S}(n\sigma_1^2)+\mathrm{S}(nd^2).$$

Hence, we may say that we solve for a single-valued relation

$$x_1 = a_{12}x_2+a_{13}x_3+a_{14}x_4+\ldots.$$

between our variables; the relation being such that the sum of the squares of the errors made in estimating x_1 from its associated values x_2, x_3, &c., is the least possible. In the case of normal correlation this "characteristic relation" must become the "equation of regression" which gives the means of any x_1-array, as only in this way can $\mathrm{S}nd^2$ be made a minimum, *i.e.*, zero.

It might be said that it would be more natural to form a "characteristic relation" between the absolute values of the variables and not their deviations from the mean. This may, however, be most conveniently done by working with the *mean* as origin until the characteristic is obtained, and then transferring the equation to zero as origin. It would be much more laborious and would only lead to the same result if zero were used *ab initio* as origin.

We may now proceed to the discussion of the special cases of two, three, or more variables. The actual formulæ obtained are not, it will be found, novel in themselves, but throw an unexpected light on the meaning of the expressions previously given by Bravais* for the case of normal correlation.

(1) *Case of Two Variables.*—Since x and y represent deviations from their respective means, we have, using S to denote summation over the whole surface,

$$\mathrm{S}(x) = \mathrm{S}(y) = 0.$$

The characteristic or regression equations which we have to find are of the form

$$\left.\begin{aligned} x &= a_1 + b_1 y \\ y &= a_2 + b_2 x \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1).$$

Taking the equation for x first, the normal equations for a_1 and b_1 are

$$\left.\begin{aligned} \mathrm{S}(x) &= \mathrm{N}a_1 + b_1\mathrm{S}(y) \\ \mathrm{S}(xy) &= a_1\mathrm{S}(y) + b_1\mathrm{S}(y^2) \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots \quad (2),$$

N being the total number of correlated pairs. From the first of these equations we have at once

$$a_1 = 0.$$

From the second

$$b_1 = \frac{\mathrm{S}(xy)}{\mathrm{S}(y^2)}.$$

To simplify our notation let us write

$$\mathrm{S}(x^2) = \mathrm{N}\sigma_1{}^2. \qquad \mathrm{S}y^2 = \mathrm{N}\sigma_2{}^2.$$

$$\mathrm{S}(xy) = \mathrm{N}r\sigma_1\sigma_2.$$

σ_1 and σ_2 are then the two standard-deviations or errors of mean

* "Mémoires par divers Savants," 1846, p. 255, and Professor Pearson's paper on "Regression, Heredity, &c." 'Phil. Trans.,' A, vol. 187 (1896), p. 261 *et seq.*

square. r is Bravais' value of the coefficient of correlation. Re-writing b_1 in terms of these symbols, we have

$$b_1 = r\,\frac{\sigma_1}{\sigma_2} \qquad (3).$$

Similarly, $a_2 = 0, \qquad b_2 = r\,\frac{\sigma_2}{\sigma_1} \qquad (4).$

But the expressions on the right of (3) and (4) are the values obtained by Bravais *on the assumption of normal correlation* for the regression of x on y, and the regression of y on x. That is to say, the Bravais values for the regressions are simply those values of b_1 and b_2, which make

$$\mathrm{S}(x-b_1y)^2 \text{ and } \mathrm{S}(x-b_2y)^2$$

respectively minima, *whatever be the form of the correlation between the two variables.* Again, whatever the form of the correlation, if the regression be really linear, the equations to the lines of regression are those given above (as we pointed out in the introduction). This theorem admits of a very simple and direct geometrical proof.

Let n be the number of correlated pairs in any one array taken parallel to the axis of x, and let θ be the angle that the line of regression makes with the axis of y. Then, for a single array,

$$\mathrm{S}(xy) = y\mathrm{S}(x) = ny^2 \tan\theta,$$

or extending the significance of S to summation over the whole surface,

$$\mathrm{S}(xy) = \mathrm{N} \tan \theta\sigma_2^2,$$

that is,

$$\tan\theta = r\,\frac{\sigma_1}{\sigma_2}.$$

In any case, then, where the regression appears to be linear, Bravais' formulæ may be used at once without troubling to investigate the normality of the distribution. The exponential character of the surface appears to have nothing whatever to do with the result.

To return, again, to the most general case, we see that both coefficients of regression must have the same sign, namely, the sign of r. Hence, either regression will serve to indicate whether there is correlation or no, for there is no reason, *à priori*, why the values of b_1 and b_2, as determined above, should be positive rather than negative. But, nevertheless, the regressions are not convenient measures of correlation, for, on comparing two similar cases, we may find, say,

$$b_1 > b'_1, \qquad b_2 < b'_2,$$

where $b_1 b_2$, $b'_1 b'_2$ are the regressions in the two cases. To which distribution are we, in such a case, to attribute the greater correlation? Bravais' coefficient solves the difficulty, we may say, in one way, by taking the geometrical mean of the two regressions as the measure of correlation. It will still remain valid for non-normal correlation. But there are other and less arbitrary interpretations even in the general case.

Suppose that instead of measuring x and y in arbitrary units we measure each in terms of its own standard deviation. Then let us write

$$\frac{x}{\sigma_1} = \rho \frac{y}{\sigma_2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5),$$

and solve for ρ by the method of least squares. We have omitted a constant on the right-hand side, since it would vanish as before. We have, at once,

$$\rho = \frac{\mathrm{S}(xy)}{\mathrm{S}(y^2)} \frac{\sigma_2}{\sigma_1} = r \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6),$$

That is to say, if we measure x and y each in terms of its own standard deviation, r becomes at once the regression of x on y, and the regression of y on x. The regressions being, in fact, the fundamental physical quantities, r is a coefficient of correlation because it is a coefficient of regression.*

Again, let us form the sums of the squares of residuals in equations (1) and (5). Inserting the values of b_1, b_2, and ρ, we have—

$$\left.\begin{aligned} \mathrm{S}(x-b_1 y)^2 &= \mathrm{N}\sigma_1^2(1-r^2) \\ \mathrm{S}(y-b_2 x)^2 &= \mathrm{N}\sigma_2^2(1-r^2) \\ \mathrm{S}\left(\frac{x}{\sigma_1}-\rho\frac{y}{\sigma_2}\right)^2 = \mathrm{S}\left(\frac{y}{\sigma_2}-\rho\frac{x}{\sigma_1}\right)^2 &= \mathrm{N}(1-r^2) \end{aligned}\right\} \ldots\ldots\ldots \quad (7).$$

Any one of these quantities, being the sum of a series of squares, must be positive. Hence r cannot be greater than unity. If r be equal to unity, or if the correlation be perfect, all the above three sums become zero. But

$$\mathrm{S}\left(\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2}\right)^2$$

can only vanish if

$$\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2} = 0$$

in every case, or if the relation hold good,

* That the regression becomes the coefficient of correlation when each deviation is measured in terms of its standard-deviation in the case of normal correlation has been pointed out by Mr. Francis Galton. *Vide* Pearson 'Phil. Trans.,' A, vol. 187, p. 307, note.

$$\frac{x_1}{y_1} = \frac{x_2}{y_2} = \frac{x_3}{y_3} = \ldots\ldots = \pm\frac{\sigma_1}{\sigma_2} \ldots\ldots\ldots\ldots \quad (8),$$

the sign of the last term depending on the sign of r. Hence the statement that two variables are "perfectly correlated" implies that relation (8) holds good, or that all pairs of deviations bear the same ratio to one another. It follows that in correlation, where the means of arrays are not collinear, or the deviation of the mean of the array is not a linear function of the deviation of the type, r can never be unity, though we know from experience that it can approach pretty closely to that value. If the regression be very far from linear, some caution must evidently be used in employing r to compare two different distributions.

In the case of normal correlation, $\sigma_1\sqrt{1-r^2}$ is the standard deviation of any array of the x variables, corresponding to a single type of y's. $\sigma_2\sqrt{1-r^2}$ is similarly the standard deviation of any array of the y variables, corresponding to a single type of x's. In the general case, the first expression may be interpreted as the mean standard deviation of the x-arrays from the line of regression, and the second expression as the mean standard deviation of the y-arrays from the line of regression. Otherwise we may regard

$$\sigma_1\sqrt{1-r^2}$$

as the standard error made in estimating x from the relation

$$x = b_1 y,$$

and

$$\sigma_2\sqrt{1-r^2}$$

as the standard error made in estimating y from the relation

$$y = b_2 x,$$

these interpretations being independent of the form of the correlation.

(2.) *Case of Three Variables.*

Let the three correlated variables be X_1, X_2, X_3, and let x_1, x_2, x_3 denote deviations of these variables from their respective means. Let us write, for brevity,

$$S(x_1^2) = N\sigma_1^2, \qquad S(x_2^2) = N\sigma_2^2$$

$$S(x_3^2) = N\sigma_3^2$$

$$S(x_1x_2) = Nr_{12}\sigma_1\sigma_2$$

$$S(x_2x_3) = Nr_{23}\sigma_2\sigma_3$$

$$S(x_3x_1) = Nr_{31}\sigma_3\sigma_1.$$

Our characteristic or regression-equation will now be of the form

$$x_1 = b_{12}x_2 + b_{13}x_3 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (9),$$

b_{12} and b_{13} being the unknowns to be determined from the observations by the method of least squares. I have omitted a constant term on the right-hand side, since its least-square value would be zero as before. The two normal equations are now—

$$\mathrm{S}(x_1x_2) = b_{12}\mathrm{S}(x_2^2) + b_{13}\mathrm{S}(x_2x_3)$$

$$\mathrm{S}(x_1x_3) = b_{12}\mathrm{S}(x_2x_3) + b_{13}\mathrm{S}(x_3^2),$$

or replacing the sums by the symbols defined above, and simplifying—

$$\left.\begin{array}{l} r_{12}\sigma_1 = b_{12}\sigma_2 + b_{13}r_{23}\sigma_3 \\ r_{13}\sigma_1 = b_{12}r_{23}\sigma_2 + b_{13}\sigma_3 \end{array}\right\} \quad \ldots\ldots\ldots\ldots \quad (10),$$

whence

$$\left.\begin{array}{l} b_{12} = \dfrac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\,\dfrac{\sigma_1}{\sigma_2} \\ b_{13} = \dfrac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\,\dfrac{\sigma_1}{\sigma_3} \end{array}\right\} \quad \ldots\ldots\ldots\ldots \quad (11).$$

That is, the characteristic relation between x_1 and x_2x_3 is—

$$x_1 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\,\frac{\sigma_1}{\sigma_2}\,x_2 + \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\,\frac{\sigma_1}{\sigma_3}\,x_3 \quad \ldots\ldots \quad (12).$$

Now Bravais showed that *if the correlation were normal*, and we selected a group or array of X_1's with regard to special values h_2 and h_3 of x_2 and x_3, then h_1 being the deviation of the mean of the selected X_1's from the X_1-mean of the whole material,

$$h_1 = b_{12}h_2 + b_{13}h_3,$$

where b_{12} and b_{13} have the values given in (11). But evidently the relation is of much greater generality; it holds good so long as h_1 is a linear function of h_2 and h_3, *whatever be the law of frequency.*

Further, the values of b_{12} and b_{13} above determined, are, under any circumstances, such that

$$\mathrm{S}v^2 = \mathrm{S}[x_1 - (b_{12}x_2 + b_{13}x_3)]^2,$$

is a minimum. If we insert in this expression the values of b_{12} and b_{13} from (11), we have, after some reduction,

$$\mathrm{S}(v^2) = \mathrm{N}\sigma_1^2\left\{1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}\right\}$$

$$= \mathrm{N}\sigma_1^2\{1 - \mathrm{R}_1^2\} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (13),$$

say. In normal correlation $\sigma_1\sqrt{1-R_1^2}$ is the standard deviation of an X_1-array, corresponding to any given types of X_2 and X_3. In general correlation it may be regarded as the mean standard deviation of the X_1-arrays from the plane

$$x_1 = b_{12}x_2 + b_{13}x_3,$$

or as the standard error made in estimating x_1 from x_2 and x_3 by relation (12).

The quantity R is of some interest, as it exactly takes the place of r in the residual expressions (7). R_1 may, in fact, be regarded as a coefficient of correlation between x_1 and (x_2x_3); it can only be unity if the linear relation (9) or (12) hold good in every case.

The quantities b_{12}, b_{13}, &c. (the others may be written down by symmetry), may be termed the net regressions of x_1 on x_2, x_1 on x_3, &c. If we write 2 for 1 and 1 for 2 in the value of b_{12}, we have

$$b_{21} = \frac{r_{12}-r_{13}r_{23}}{1-r_{23}^2}\,\frac{\sigma_2}{\sigma_1},$$

b_{21} being the the net regression of x_2 on x_1. In normal correlation, b_{12} and b_{21} are the regressions for any group of X_1's or X_2's associated with a fixed type of X_3's. Hence, in this case (normal correlation), the coefficient of correlation for such a group is the geometrical mean of the two regressions, or

$$\rho_{12} = \frac{r_{12}-r_{13}r_{23},}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

a quantity that may be called the net coefficient of correlation between x_1 and x_2.* The similar net coefficients between x_1 and x_3, x_2 and x_3, may be written down by interchanging the suffixes.

In normal correlation ρ_{12} is quite strictly the coefficient of correlation for *any* sub-group of X_1's and X_2's, whatever the associated type of X_3's. In generalised correlation this will not be so, and ρ_{12} can only retain an average significance.

The method does not appear to be capable of investigating changes in the net coefficient as we pass from one type to another, but it may be noted that whatever the form of the correlation, ρ_{12} retains three of the chief properties of the ordinary coefficients: (1) it can only be

* My quantities, b_{12}, b_{13}, &c., were termed by Professor Pearson ("Regression &c.," 'Phil. Trans.,' A, vol. 187 (1896), p. 287), "Coefficients of double regression," and quantities like $b_{12}\frac{\sigma_2}{\sigma_1}$, $b_{13}\frac{\sigma_3}{\sigma_1}$, &c., "coefficients of double correlation." My quantities ρ he did not use. Having named the ρ's "net correlation," it seemed most natural to rename the b's "net regressions," as the b's and ρ's are corresponding quantities.

Some of my results given above were quoted by Professor Pearson in his paper (*loc. cit.*, notes on pp. 268 and 287).

zero if both net regressions are zero; (2) it is a symmetrical function of the variables; (3) it cannot be greater than unity; for, by (13),

$$r_{12}^2+r_{13}^2-2r_{12}r_{23}r_{31} < 1-r_{23}^2;$$

or adding $r_{13}^2r_{23}^2$ to both sides, and transferring r_{13}^2 to the right-hand side

$$(r_{12}-r_{13}r_{23})^2 < (1-r_{13}^2)(1-r_{23}^2).$$

If any two coefficients, say $r_{12}r_{13}$, be supposed known, the inequality we have used above will give us limits for the value of the third. Throwing it into the form

$$(r_{23}-r_{12}r_{13})^2 < 1+r_{12}^2r_{13}^2-r_{12}^2-r_{13}^2,$$

we have r_{23} must lie between the limits

$$r_{12}r_{13} \pm \sqrt{r_{12}^2r_{13}^2-r_{12}^2-r_{13}^2+1}.$$

The values of these limits for some special cases are collected in the following table :—

Values of r_{12} and r_{13}.	Limits of r_{23}.
$r_{12} = r_{13} = 0$	0
$r_{12} = r_{13} = \pm 1$	$+1$
$r_{12} = +1,\ r_{13} = -1$	-1
$r_{12} = 0,\ r_{13} = \pm 1$	0
$r_{12} = 0,\ r_{13} = \pm r$	$\pm\sqrt{1-r^2}$
$r_{12} = r_{13} = \pm r$	1 and $2r^2-1$
$r_{12} = +r,\ r_{13} = -r$	$2r^2-1$ and -1
$r_{12} = r_{13} = \pm\sqrt{0{\cdot}5} = 0{\cdot}707$	0 and 1
$r_{12} = +\sqrt{0{\cdot}5}\ \ r_{12} = -\sqrt{0{\cdot}5}$	0 „ -1

One is rather pronc to argue that if A be correlated with B, and B with C, A will be correlated with C. Evidently this is not necessary. A may be positively correlated with B, and B positively correlated with C, but yet A may, in general, be negatively correlated with C. Only, if the coefficients (AB) and (BC) are both numerically greater than 0·707, can one even ascribe the correct sign to the (AC) correlation.

It is evident that one would, in general, expect to make a smaller standard error in estimating x_1 from the two associated variables x_2 and x_3, than in estimating it from one only, say x_2. But it seems desirable to prove this specifically, and to investigate under what conditions it will hold good. The necessary condition is—

$$\frac{r_{12}^2+r_{13}^2-2r_{12}r_{23}r_{13}}{1-r_{23}^2} > r_{12}^2,$$

that is,

$$r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23} > r_{12}^2-r_{12}^2r_{13}^2,$$

or

$$(r_{13}-r_{12}r_{23})^2 > 0.$$

But $(r_{13}-r_{12}r_{23})$ is the numerator of ρ_{13}, the net coefficient of correlation between x_1 and x_3. Hence the standard error in the second case will be always less than in the first, so long as ρ_{13} is not zero. The condition is somewhat interesting.

To take an arithmetical example, suppose one had in some actual case

$$r_{12} = +0\cdot 8$$

$$r_{23} = +0\cdot 5 \qquad r_{13} = +0\cdot 4.$$

One might very naturally imagine that the introduction of the third variable with a fairly high correlation coefficient (0·4) would considerably lessen the standard deviation of the x_1-array; but this is not so, for

$$\rho_{13} = \frac{0\cdot 4-(0\cdot 5\times 0\cdot 8)}{\sqrt{0\cdot 75\times 0\cdot 36}} = 0,$$

so the third variable would be of no assistance.

III. *Case of Four Variables.*

This case is, perhaps, of sufficient practical importance to warrant our developing the results at length as in the last.

If x_1, x_2, x_3, x_4, be the associated deviations of the four variables from their respective means, the characteristic equation will be of the form

$$x_1 = b_{12}x_2+b_{13}x_3+b_{14}x_4 \ldots\ldots\ldots\ldots\ldots \quad (14).$$

The normal equations for the b's are, in our previous notation,

$$\left.\begin{aligned} r_{12}\sigma_1 &= b_{12}\sigma_2+b_{13}r_{23}\sigma_3+b_{14}r_{24}\sigma_4 \\ r_{13}\sigma_1 &= b_{12}r_{23}\sigma_2+b_{13}\sigma_3+b_{14}r_{34}\sigma_4 \\ r_{14}\sigma_1 &= b_{12}r_{24}\sigma_2+b_{13}r_{34}\sigma_3+b_{14}\sigma_4 \end{aligned}\right\}$$

Hence

$$b_{12} = \frac{\begin{vmatrix} r_{12} & r_{23} & r_{24} \\ r_{13} & 1 & r_{31} \\ r_{14} & r_{34} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} & r_{24} \\ r_{23} & 1 & r_{34} \\ r_{24} & r_{34} & 1 \end{vmatrix}} \frac{\sigma_1}{\sigma_2} \ldots\ldots\ldots\ldots\ldots \quad (15),$$

and so on for the others, b_{12}, b_{13}, &c., we may call the net regressions of x_1 on x_2, x_1 on x_3, &c., as before. By parity of notation, we have

$$b_{21} = \frac{\begin{vmatrix} r_{12} & r_{23} & r_{24} \\ r_{13} & 1 & r_{34} \\ r_{14} & r_{34} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{13} & r_{14} \\ r_{13} & 1 & r_{34} \\ r_{14} & r_{34} & 1 \end{vmatrix}} \frac{\sigma_2}{\sigma_1}$$

and we may again call

$$\rho_{12} = \sqrt{b_{12}b_{21}},$$

the net coefficient of correlation between x_1 and x_2. Expanding the determinants, we have, in fact,

$$\rho_{12} = \frac{r_{12}(1-r_{34}^2) + r_{13}(r_{34}r_{24}-r_{23}) + r_{14}(r_{23}r_{34}-r_{24})}{\sqrt{[(1-r_{34}^2) + r_{23}(r_{34}r_{24}-r_{23}) + r_{24}(r_{23}r_{34}-r_{24})][(1-r_{34}^2) + r_{13}(r_{34}r_{14}-r_{13}) + r_{14}(r_{13}r_{34}-r_{14})]}} \quad \ldots\ldots\ldots \quad (16).$$

There are six such net coefficients, ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, ρ_{34}. The above values of the regressions are again those usually obtained on the assumption of normal correlation.* The net correlation ρ_{12} becomes, on that assumption, the coefficient of correlation for any group of the x_1 x_2 variables associated with fixed types of x_3 and x_4. If we write

$$u = x_1 - (b_{12}x_2 + b_{13}x_3 + b_{14}x_4),$$

we have, after some rather lengthy reduction,

$$\frac{1}{\mathrm{N}}\mathrm{S}(u^2) = \sigma_1^2(1-\mathrm{R}_1^2),$$

where

$$\mathrm{R}_1^2 = \frac{\left\{ \begin{array}{c} r_{12}^2 + r_{13}^2 + r_{14}^2 - r_{12}^2 r_{34}^2 - r_{23}^2 r_{14}^2 - r_{13}^2 r_{24}^2 \\ -2(r_{13}r_{14}r_{34} + r_{12}r_{14}r_{24} + r_{12}r_{13}r_{23}) + 2(r_{12}r_{14}r_{23}r_{34} + r_{13}r_{14}r_{23}r_{24} + r_{12}r_{13}r_{24}r_{34}) \end{array} \right\}}{1 - r_{23}^2 - r_{34}^2 - r_{24}^2 + 2r_{23}r_{34}r_{24}}.$$

In normal correlation, $\sigma_1\sqrt{1-\mathrm{R}_1^2}$ is the standard deviation of all x_1-arrays associated with fixed types of x_2, x_3, and x_4. In general correlation, it is most easily interpreted as the standard error made in estimating x_1, by equation (14), from its associated values of x_2, x_3, and x_4.

As in the case of three variables, the quantity R may be considered as a coefficient of correlation. It can range between ± 1, and can only become unity if the linear relation (14) hold good in each individual instance.

We showed at the end of the last section that the standard error made in estimating x_1 from the relation

$$x_1 = b_{12}x_2 + b_{13}x_3$$

* Professor Pearson, "Regression, Heredity, and Panmixia." 'Phil. Trans.,' A, vol. 187 (1896), p. 294.

was always less than the standard error when only x_2 was taken into account, unless

$$\rho_{13} = 0.$$

We may now prove the similar theorem that when we use three variables, x_2, x_3, x_4, on which to base the estimate, the standard error will be again decreased, unless

$$\rho_{14} = 0.$$

The condition that $S(u^2)$, in our present case, shall be less than $S(r^2)$ in the last, is, in fact,

$$\left\{\begin{array}{l} r_{12}^2+r_{13}^2+r_{14}^2-r_{12}^2r_{34}^2-r_{23}^2r_{14}^2-r_{13}^2r_{24}^2 \\ -2(r_{13}r_{14}r_{34}+r_{12}r_{13}r_{23}+r_{12}r_{14}r_{24}) \\ +2(r_{12}r_{14}r_{23}r_{34}+r_{14}r_{13}r_{24}r_{23}+r_{12}r_{13}r_{24}r_{34}) \end{array}\right\}(1-r_{23}^2)$$

$$> (r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23})(1-r_{23}^2-r_{24}^2-r_{34}^2+2r_{23}r_{24}r_{34}).$$

This may be finally reduced to—

$$(r_{14}-r_{13}r_{34}-r_{12}r_{24}-r_{14}r_{23}^2+r_{13}r_{23}r_{24}+r_{12}r_{23}r_{34})^2 > 0,$$

that is

$$\rho_{14}^2 > 0.$$

The treatment of the general case of n variables, so far as regards obtaining the regressions, is obvious, and it is unnecessary to give it at length.

We can now see that the use of normal regression formulæ is quite legitimate in all cases, so long as the necessary limitations of interpretation are recognised. Bravais' r always remains a coefficient of correlation. These results I must plead as justification for my use of normal formulæ in two cases* where the correlation was markedly non-normal.

"Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation which may arise when Indices are used in the Measurement of Organs." By KARL PEARSON, F.R.S., University College, London. Received December 29, 1896,—Read February 18, 1897.

(1) If the ratio of two absolute measurements on the same or different organs be taken it is convenient to term this ratio an *index*.

If $u = f_1(x, y)$ and $v = f_2(z, y)$ be two functions of the three variables x, y, z, and these variables be selected at random so that there exists no correlation between x,y, y,z, or z,x, there will still be found to

* 'Economic Journal,' Dec., 1895, and Dec., 1896, "On the Correlation of Total Pauperism with Proportion of Out-relief."