A SYMMETRIC COEFFICIENT OF CORRELATION FOR SEVERAL VARIABLES*

BY DUNHAM JACKSON

Let $(x_1, x_2, \cdots, x_n)$ and $(y_1, y_2, \cdots, y_n)$ be two sets of n real numbers each, neither set consisting entirely of zeros. In order that the ensuing use of technical terms may be in accordance with established practice, let it be supposed that

$$\sum x_k = \sum y_k = 0,$$

though the mathematical relations to be considered are in themselves not dependent on this hypothesis. The coefficient of correlation between the x's and the y's, defined by the formula

$$r = \frac{\sum x_k y_k}{\sqrt{\sum x_k^2 \sum y_k^2}},$$

may be regarded as a measure of the degree of resemblance of the two given sets of numbers, and may be interpreted as the cosine of the angle $POQ = \theta$, if O, P, Q represent respectively the origin, the point with coordinates $(x_1, x_2, \cdots, x_n)$, and the point $(y_1, y_2, \cdots, y_n)$, in space of n dimensions.†

Let a third set of real numbers $(z_1, z_2, \cdots, z_n)$ be given, not all zero (and subject, let us say, to the condition $\sum z_k = 0$). There may sometimes be occasion to consider a measure of the degree of resemblance of the three sets (x_k), (y_k), and (z_k), not an asymmetric measure of the dependence of a specified set on the other two, as in the case of a coefficient of double correlation, but a formula

* Presented to the Society, April 19, 1924.

† For an elementary exposition of this idea, and for bibliographical references, see a paper by the author entitled *The trigonometry of correlation*, AMERICAN MATHEMATICAL MONTHLY, vol. 31 (1924), pp. 275-280.

treating all three sets alike.* Such a formula will be discussed in the following pages. Some theoretical advantages belonging to it will appear in the course of the discussion; its practical utility must naturally await the test of experience. If it has already been treated elsewhere, as is not unlikely, the writer hopes that the present exposition may nevertheless serve a useful purpose in giving it wider publicity.

Without loss of generality, it may be assumed that the given sets of numbers are normalized, so that

$$\sum x_k^2 = \sum y_k^2 = \sum z_k^2 = 1.$$

Then the points P and Q, and the point R with coordinates $(z_1, z_2, \cdots, z_n)$, are on the surface of a sphere of unit radius about the origin. The discussion is concerned with the degree of propinquity of these three points. As the correlation of the first two sets of numbers is measured by means of the angle POQ, there would be a certain analogy in making use, at the next stage, of the solid angle $O—PQR$, or, in other words, of passing from the length of the arc PQ to the area of the spherical triangle PQR as a quantitative indication. But the area of the triangle is zero whenever the three points lie on the same great circle, whether the points themselves are close together or widely separated, so that it seems necessary to look further for a satisfactory definition.

Let S be the middle point of the chord PQ. The value of the coefficient of correlation r can of course be expressed in terms of the angle $\Phi = \frac{1}{2}\theta = POS$, the relation being

$$r = \cos\theta = \cos 2\Phi = 2\cos^2\Phi - 1.$$

The extreme limiting values of $\cos\Phi$, for all possible positions of P and Q, are not -1 and 1, as in the case of $\cos\theta$, but 0 and 1. As $\cos\Phi = OS$ is the perpendi-

* This problem was suggested to the writer by a study of the distribution of freshman grades, in which Professor W. H. Bussey was engaged.

cular distance of the chord PQ from O, an analogous quantity in some respects would be the distance from O to the plane of P, Q, and R. But a knowledge of this quantity would scarcely give the kind of information that is desired, since the distance is zero whenever the three points are on the same great circle, and may be either zero or arbitrarily near unity for three points arbitrarily close to each other, becoming wholly indeterminate between these limits in the one case for which an unmistakable characterization is most essential, the case of actual coincidence of P, Q, and R.

It seems more instructive to regard OS as the distance from O to the *center of gravity* of the points P and Q, regarded as material particles of equal weight. Let T be defined as the center of gravity of P, Q, and R, and let $\bar{r}_1 = OT$. This $\bar{r}_1$ is 1 if and only if the three points are coincident; it is 0 if and only if they are situated on a great circle at angular distances of 120° from each other; and it has a value intermediate between 0 and 1 in all other cases. A formal expression for it is readily calculated. The coordinates $(w_1, w_2, \cdots, w_n)$ of T are given by

$$w_k = \frac{1}{3}(x_k + y_k + z_k),$$

so that $\bar{r}_1 = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$

$$= \frac{1}{3}\sqrt{\Sigma x_k^2 + \Sigma y_k^2 + \Sigma z_k^2 + 2\Sigma x_k y_k + 2\Sigma x_k z_k + 2\Sigma y_k z_k}.$$

In consequence of the hypothesis that the x's, y's, and z's are normalized, the sums of squares under the last radical are each equal to 1, and the product sums are the respective coefficients of correlation r_{12}, r_{13}, r_{23} of the given sets of numbers, taken two sets at a time. That is,

$$\bar{r}_1 = \frac{1}{3}\sqrt{3 + 2r_{12} + 2r_{13} + 2r_{23}}.$$

Taken on its own merits, independently of any geometric interpretation, this formula defines a number which is mani-

festly subject to the inequalities $0 \leqq \bar{r}_1 \leqq 1$, since the radical is of course to be taken as positive or zero, and none of the numbers r_{12}, r_{13}, r_{23} can exceed 1; the quantity under the radical sign is equal to 9 if $r_{12} = r_{13} = r_{23} = 1$, is 0 if* $r_{12} = r_{13} = r_{23} = -\frac{1}{2}$, and can never be negative,† since it is equal to $9\sum w_k^2$.

A measure somewhat more closely analogous to the ordinary coefficient of correlation, and slightly easier to compute, the measure which it is the main purpose of this paper to suggest, is related to $\bar{r}_1$ as $\cos\theta$ is related to $\cos\varphi$, and *is defined by the formula*

$$\bar{r} = 2\bar{r}_1^2 - 1 = \frac{4}{9}(r_{12} + r_{13} + r_{23}) - \frac{1}{3}.$$

From its dependence on $\bar{r}_1$, it ranges between the extreme values -1 and $+1$, the latter value signifying perfect correlation of all three of the given sets of quantities. It is simply a linear function of the "average coefficient of correlation" of the given sets, adjusted so as to have -1 and $+1$ for its extremes; this adjustment, however, appears from the above treatment not as a mere piece of algebraic formalism, but as a natural corollary of simple geometric relations.

When there are m variables instead of three, that is, m given (normalized) sets of n quantities each, the notation is to be changed by letting $(w_1, \cdots, w_n)$ stand for the co-

---

* These relations, together with the conditions $\Sigma x_k = \Sigma y_k = \Sigma z_k = 0$, are satisfied if, for example, the x's, y's and z's are respectively

$$\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, \cdots, 0\right), \quad \left(-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, \cdots, 0\right),$$
$$\left(0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, \cdots, 0\right),$$

all the entries after the first three being zero in each parenthesis.

† In particular, such a combination as $r_{12} = r_{13} = r_{23} = -1$ is impossible; it is readily seen that if $r_{12} = r_{13} = -1$, the third coefficient r_{23} must necessarily be $+1$.

ordinates of the center of gravity of the m corresponding points of the unit sphere, so that

$$w_k = \frac{1}{m}(x_k + y_k + z_k + \cdots), \qquad (k = 1, 2, \cdots, n);$$

$$\bar{r}_1 = \sqrt{\Sigma w_k^2} = \frac{1}{m}\sqrt{m + 2\Sigma r_{ij}},$$

$$(i = 1, 2, \cdots, m, j = i+1, i+2, \cdots, m);$$

$$\bar{r} = 2\bar{r}_1^2 - 1 = \frac{4}{m^2}\Sigma r_{ij} - \frac{m-2}{m}.$$

The quantities r_{ij} are once more the coefficients of correlation of two sets at a time of the given m sets of numbers, supposed subject to the conditions

$$\Sigma x_k = \Sigma y_k = \Sigma z_k = \cdots = 0.$$

It is readily seen, both algebraically and quasi-geometrically, that $\bar{r}$ varies from the value -1 at one extreme to the value $+1$, denoting perfect correlation, at the other.

It is well known that the center of gravity of a system of particles corresponds to the solution of a simple problem in least squares. If the particles are of equal weight, as is assumed to be the case here, the center of gravity is the point so situated that the sum of the squares of the distances of the various particles from it is a minimum. Something analogous to the idea of least squares can be made apparent in the present connection under a slightly different form.

Since $\cos\theta$ is approximately equal to $1 - \frac{1}{2}\theta^2$, when θ is small, the minimizing of a sum of squares may be looked upon as related to the problem of making a sum of cosines a maximum. Let $(\mu_1, \mu_2, \cdots, \mu_n)$ be the direction cosines of an arbitrary* line OM, that is, an arbitrary set of n numbers subject to the condition $\mu_1^2 + \mu_2^2 + \cdots + \mu_n^2 = 1$. If the point M is taken at unit distance from the origin, the numbers μ_k will be its coordinates. Let the angles POM, $QOM, \cdots$, be denoted by $\varphi_1, \varphi_2, \cdots$, and let the problem be proposed of determining the μ's so that

* It is not assumed that $\Sigma\mu_k = 0$.

$\cos\varphi_1 + \cos\varphi_2 + \cdots + \cos\varphi_m$ shall be a maximum. Since

$$\cos\varphi_1 = x_1\mu_1 + x_2\mu_2 + \cdots + x_n\mu_n,$$

etc., the quantity s to be made a maximum is

$$\begin{aligned} s &= (x_1 + y_1 + \cdots)\mu_1 \\ &\quad + (x_2 + y_2 + \cdots)\mu_2 + \cdots + (x_n + y_n + \cdots)\mu_n \\ &= m(w_1\mu_1 + w_2\mu_2 + \cdots + w_n\mu_n), \end{aligned}$$

or, if λ_k denotes the kth direction cosine of the line from O to the center of gravity* T,

$$\lambda_k = \frac{w_k}{\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}} = \frac{w_k}{\bar{r}_1},$$

then

$$s = m\bar{r}_1(\lambda_1\mu_1 + \lambda_2\mu_2 + \cdots + \lambda_n\mu_n).$$

The last parenthesis is the cosine of the angle TOM, and can not exceed 1, being equal to 1 when $\mu_k = \lambda_k$, that is, when the line OM is drawn through T. So OT is determined as the line for which the sum of cosines is greatest. The maximum value of the sum is seen to be $m\bar{r}_1$.

As $\cos\varphi_i$ is equal to

$$1 - 2\sin^2\left(\frac{1}{2}\varphi_i\right),$$

it can also be said that the sum of the squares of the sines of the half angles has been made a minimum, and hence that there has been literally a solution of a problem in least squares. Let L denote the point with coordinates $(\lambda_1, \lambda_2, \cdots, \lambda_n)$, where the line OT pierces the unit sphere, and let $\varphi_1, \varphi_2, \cdots$, now stand for the angles $POL, QOL, \cdots$. Then $2\sin(\frac{1}{2}\varphi_1)$ is the length of the chord LP, etc., and L is determined on the surface of the sphere as the point for which the sum of the squares of the rectilinear distances from $P, Q, \cdots$, is a minimum. If these distances from $P, Q, \cdots$, to L are denoted by $d_1, d_2, \cdots$,

* The problem becomes indeterminate in the limiting case in which T falls at the origin.

$$\sum_{i=1}^{m} d_i^2 = 4\sum_{i=1}^{m} \sin^2 \frac{\varphi_i}{2} = 2\sum_{i=1}^{m} (1-\cos\varphi_i) = 2(m-m\bar{r}_1),$$

and $\bar{r}_1$ has the value

$$\bar{r}_1 = 1-\frac{1}{2m}\sum_{i=1}^{m} d_i^2.$$

But this reckoning does not seem to lead to any particularly simple geometric interpretation for $\bar{r}$. It may turn out, in some circumstances at least, that $\bar{r}_1$ is the more significant measure after all.

THE UNIVERSITY OF MINNESOTA

METHODS FOR FINDING FACTORS OF LARGE INTEGERS*

BY H. S. VANDIVER

1. *Introduction.* We shall examine, in this paper, the problem of finding factors of integers beyond the range of Lehmer's factor tables, by methods shorter than that of dividing the integer by all the primes less than its square root.

Three methods will be proposed here. The first two depend on the representation of the integer as a definite quadratic form, and the third on the representation as an indefinite quadratic form. As I hope to devote another paper to the development of the last two methods, only outlines and a few examples will be given in connection with them.

The theory of quadratic forms has been applied in several different ways to the problem.†

In particular, Seelhoff‡ gave an expeditious method with the use of tables, which, however, is limited in application,

* Presented to the Society, September 7, 1923, under the title *A method of finding factors of integers of the form* $8n+1$. The author was enabled to carry out this investigation through a grant from the Heckscher Foundation for the Advancement of Research.

† Dickson, *History of the Theory of Numbers*, vol. 1, pp. 361–66.

‡ AMERICAN JOURNAL, vol. 7, p. 264; vol. 8, p. 26.