THE GEOMETRY OF FREQUENCY FUNCTIONS*

BY DUNHAM JACKSON

1. *Introduction.* The Pearson coefficient of correlation, calculated for a finite number of observations, has a geometric interpretation which is simple and almost immediate.† The same thing may be said of the corresponding expression formed for a pair of functions of a continuous variable.‡ When the distribution of the observed quantities is thought of as given by a frequency function, the geometric interpretation of the correlation coefficient is not so obvious. It is the purpose of this paper to show one form that such an interpretation may take.§ The geometric configurations are exactly the same as in the other cases mentioned; the difference is in the manner of setting up the association. This is accomplished by defining an appropriate correspondence between an arbitrary point of a plane, or of space, and an arbitrary linear combination of the variables subjected to measurement.

There will be no assumption that the distributions involved are "normal", in the sense of the Gaussian law. There will be incidental reference to frequency functions having properties that correspond to those of normal orthogonal sets of functions, as the terms are used in the theory of the development of arbitrary functions in series; but

* Presented to the Society, October 25, 1924.

† Cf., e. g., D. Jackson, *The trigonometry of correlation,* AMERICAN MATHEMATICAL MONTHLY, vol. 31 (1924), pp. 275–280; also the paper cited in the next footnote.

‡ Cf., e. g., D. Jackson, *The elementary geometry of function space*; recently submitted to the AMERICAN MATHEMATICAL MONTHLY.

§ For another form, cf. James McMahon, *Hyperspherical goniometry; and its application to correlation theory for n variables,* BIOMETRIKA, vol. 15 (1923), pp. 173–208. The fundamental idea of attaching a geometric meaning to the correlation coefficient appears to be due to Pearson himself.

the word "normal" will not be used in this connection, because of the possibility of misunderstanding. As the paper is concerned primarily with the establishment of certain formal relations, questions of convergence will be avoided for the most part by assuming that the functions considered are different from zero only over a finite range; the concluding paragraph will deal briefly with distributions that "tail off" to zero at infinity.

The order of presentation in the main part of the paper is closely parallel to that followed in the author's article on *The elementary geometry of function space,* to which reference has been made in a previous footnote. There is enough difference in detail, however, to call for an independent treatment at some length.

2. *Frequency Functions in Two Variables.* To begin with the case of two variables, let $\varphi(x, y)$ be defined and continuous as a function of its arguments over a finite region R of the plane; let it be everywhere positive or zero, and not identically zero. There is no essential loss of generality, and there is some gain of simplicity, in assuming that

$$\int\int_R \varphi(x, y)\, dx\, dy = 1.$$

For preliminary consideration, let it be supposed further that φ satisfies the conditions*

$$(1) \qquad \int xy\varphi = 0, \qquad \int x^2\varphi = \int y^2\varphi = 1,$$

where $\int xy\varphi$ is an abbreviated notation for the double integral

$$\int\int_R xy\varphi(x, y)\, dx\, dy,$$

and where the other integrals are to be interpreted similarly.

Let a and b be any two real numbers. Then we have $\int(ax+by)^2\varphi = a^2+b^2$, because of (1); if (a, b) are taken

* A method of constructing an infinite variety of functions φ satisfying the conditions imposed will appear presently.

as the coordinates of a point, the value of the integral is the square of the distance of the point from the origin. If P_1 and P_2 are two points with coordinates (a_1, b_1) and (a_2, b_2), and O the origin, the distance P_1P_2 is the square root of the quantity

$$\int [(a_2 - a_1)x + (b_2 - b_1)y]^2 \varphi = (a_2 - a_1)^2 + (b_2 - b_1)^2,$$

and the cosine of the angle P_1OP_2 is

$$\frac{a_1a_2 + b_1b_2}{\sqrt{a_1^2 + b_1^2}\,\sqrt{a_2^2 + b_2^2}} = \frac{\int (a_1x + b_1y)(a_2x + b_2y)\varphi}{\sqrt{\int (a_1x + b_1y)^2\varphi}\,\sqrt{\int (a_2x + b_2y)^2\varphi}}.$$

Thus the fundamental geometric measures can be expressed, somewhat indirectly and artificially to be sure, in terms of integrals involving the function φ.

In transition, let φ be a function subject to the same hypotheses as before, except that the last two of the conditions (1) are not imposed; and let*

$$\sqrt{\int x^2\varphi} = \sigma, \qquad \sqrt{\int y^2\varphi} = \tau.$$

If new variables ξ, η are introduced by means of the relations $\xi = x/\sigma$, $\eta = y/\tau$, so that, incidentally, $d\xi = dx/\sigma$, $d\eta = dy/\tau$, and if the function $\sigma\tau\varphi(x, y)$ is designated, with regard to its dependence on the new variables, by $\Phi(\xi, \eta)$, it is found that

$$(2)\quad \begin{cases} \int\int \Phi(\xi, \eta)\, d\xi\, d\eta = 1, & \int\int \xi\eta\, \Phi(\xi, \eta)\, d\xi\, d\eta = 0, \\ \int\int \xi^2 \Phi(\xi, \eta)\, d\xi\, d\eta = \int\int \eta^2 \Phi(\xi, \eta)\, d\xi\, d\eta = 1, \end{cases}$$

* Since $\int\varphi = 1$, the quantities σ and τ are the standard deviations of x and y, when φ is interpreted as a frequency function, if $\int x\varphi = \int y\varphi = 0$; these last conditions have no bearing on the formal work in hand, apart from its statistical interpretation, but they may be thought of as included among the conditions imposed on φ, if it appears that the notation is likely to cause confusion otherwise.

the integrals being extended over the region of definition of Φ. A simple substitution has replaced φ by a function satisfying all the original hypotheses.

Now let φ, while retaining the other properties that have been assigned to it, be relieved of the restrictions (1) altogether. Let

$$\sqrt{\int x^2\varphi} = \sigma, \quad \int xy\varphi = p, \quad y' = y - (p/\sigma^2)x, \quad \sqrt{\int y'^2\varphi} = \tau'.$$

It is found by substitution that

$$\int xy'\varphi = \int xy\varphi - \frac{p}{\sigma^2}\int x^2\varphi = 0.$$

As a matter of notation, let $x' = x$, $\sigma' = \sigma$, $\varphi(x,y) = \bar{\varphi}(x',y')$. Since the functional determinant of x' and y' with respect to x and y is 1, three of the above relations may be written in the form

$$\int\int x'^2\bar{\varphi}(x',y')\,dx'\,dy' = \sigma'^2, \quad \int\int y'^2\bar{\varphi}(x',y')\,dx'\,dy' = \tau'^2,$$

$$\int\int x'y'\bar{\varphi}(x',y')\,dx'\,dy' = 0,$$

the range of integration now being that over which $\bar{\varphi}(x',y')$ is defined. The present $\bar{\varphi}$ then corresponds to the function φ in the preceding paragraph. To obtain a Φ satisfying (2), it is sufficient to let

$$\xi = x'/\sigma', \quad \eta = y'/\tau', \quad \sigma'\tau'\bar{\varphi}(x',y') = \Phi(\xi,\eta).$$

The original variables x, y are expressed in terms of ξ, η by the equations

$$x = \sigma'\xi, \quad y = \frac{p}{\sigma'}\xi + \tau'\eta, \tag{3}$$

which are of the general form

$$x = a_1\xi + b_1\eta, \quad y = a_2\xi + b_2\eta. \tag{4}$$

The particular determination of the coefficients specified in (3) is only one of an infinite variety of determinations which will serve essentially the same purpose. For if ξ

and η are subjected to a further transformation defined by setting

$$\xi' = \alpha_1\xi + \beta_1\eta, \qquad \eta' = \alpha_2\xi + \beta_2\eta,$$

where

$$\alpha_1^2 + \beta_1^2 = \alpha_2^2 + \beta_2^2 = 1, \qquad \alpha_1\alpha_2 + \beta_1\beta_2 = 0,$$

and if $\Phi(\xi, \eta) = \overline{\Phi}(\xi' \eta')$, equations of the form (2) are satisfied in terms of the new variables ξ' and η', because of the equations (2) themselves, in terms of ξ and η, and the fact that the functional determinant* of the transformation is ± 1. There are accordingly infinitely many transformations of the form (4) which yield a Φ satisfying (2), if $\Phi(\xi, \eta)$ is equal to $\varphi(x, y)$ multiplied by the absolute value of the determinant of the coefficients. Any transformation (4) fulfilling this requirement may be taken as a basis for the work that is to follow, and it will be understood that the symbols a_1, b_1, a_2, b_2 refer to the coefficients in such a transformation.

With this understanding, it is seen at once that

$$\int x^2\varphi = \int\int (a_1\xi + b_1\eta)^2 \Phi(\xi, \eta)\, d\xi d\eta = a_1^2 + b_1^2,$$

$$\int y^2\varphi = a_2^2 + b_2^2,$$

$$\int xy\varphi = \int\int (a_1\xi + b_1\eta)(a_2\xi + b_2\eta)\, \Phi(\xi, \eta)\, d\xi d\eta = a_1a_2 + b_1b_2;$$

when the variables of integration are not expressly indicated, it is intended always that the integration shall be performed with regard to x and y. *The equations* (4) *serve, when the function* φ *is given, to associate* x *and* y *with two points* P_1 *and* P_2 *in the* (ξ, η)*-plane, having the coordinates* (a_1, b_1) *and* (a_2, b_2) *respectively. Because of the degree of arbitrariness retained by the coefficients in* (4), *the association may*

* To repeat the familiar proof,

$$\begin{vmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{vmatrix}^2 = \begin{vmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{vmatrix} \begin{vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{vmatrix} = \begin{vmatrix} \alpha_1^2 + \beta_1^2 & \alpha_1\alpha_2 + \beta_1\beta_2 \\ \alpha_1\alpha_2 + \beta_1\beta_2 & \alpha_2^2 + \beta_2^2 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1.$$

be carried out in an infinite variety of ways, but in all cases the expressions

$$(5) \qquad \sqrt{\int x^2\varphi}, \quad \sqrt{\int y^2\varphi}, \quad \frac{\int xy\varphi}{\sqrt{\left(\int x^2\varphi\right)\left(\int y^2\varphi\right)}},$$

which are themselves independent of the coefficients, represent respectively the distance OP_1, the distance OP_2, and the cosine of the angle P_1OP_2. The last expression, in case $\int x\varphi = \int y\varphi = 0$, is the coefficient of correlation between two variables x and y distributed according to the frequency function φ, and a geometric meaning has thus been assigned to the correlation coefficient.

The correspondence that has been established renders it possible to prove analytical theorems involving a frequency function by the mere interpretation of geometric facts. Consider, for example, the problem of determining a coefficient λ so as to make the value of the integral $\int(y-\lambda x)^2\varphi$ a minimum. Since this integral is equal to

$$\int\int [(a_2-\lambda a_1)\xi+(b_2-\lambda b_1)\eta]^2\,\Phi(\xi,\eta)\,d\xi\,d\eta,$$

it is the square of the distance between the points $(\lambda a_1, \lambda b_1)$ and (a_2, b_2). The points (a_1, b_1) and (a_2, b_2) have already been denoted by P_1 and P_2; let Q stand for the point $(\lambda a_1, \lambda b_1)$, and θ for the angle P_1OP_2. Then Q is on the line OP_1, the distances OQ and OP_1 being in the ratio of λ to 1. In order that the distance QP_2 may be a minimum, Q must be the foot of the perpendicular from P_2 on OP_1. This means that $\overline{OQ} = \overline{OP_2}\cos\theta$, and

$$\lambda = \frac{\overline{OQ}}{\overline{OP_1}} = \frac{\overline{OP_2}}{\overline{OP_1}}\cos\theta = \frac{\sqrt{\int y^2\varphi}}{\sqrt{\int x^2\varphi}} \cdot \frac{\int xy\varphi}{\sqrt{\left(\int x^2\varphi\right)\left(\int y^2\varphi\right)}} = \frac{\int xy\varphi}{\int x^2\varphi}.$$

If $\int x\varphi = \int y\varphi = 0$, the value of $\int \varphi$ being 1 throughout, the quantities

$$\sigma = \sqrt{\int x^2\varphi}, \qquad \tau = \sqrt{\int y^2\varphi}$$

are the *standard deviations* of x and y respectively, while, as already mentioned, $\cos\theta = r$ is the *coefficient of correlation* of x and y. With this notation, it is seen that $\lambda = (\tau/\sigma)r$. The minimum distance QP_2 is equal to

$$\overline{OP_2} \sin\theta = \tau\sqrt{1-r^2}.$$

The familiar determination of the coefficient of regression of y on x and of the root-mean-square deviation from the line of regression will be recognized at once.

3. *Frequency Functions in Three Variables.* The geometric treatment can be extended to three dimensions without difficulty. To begin, as before, with a preliminary inspection of a special case, let $\varphi(x, y, z)$ be defined throughout a finite region of space, continuous, everywhere positive or zero, not identically zero, and so constituted that

$$(6) \qquad \begin{aligned} \int xy\varphi = \int xz\varphi = \int yz\varphi = 0, \\ \int \varphi = \int x^2\varphi = \int y^2\varphi = \int z^2\varphi = 1, \end{aligned}$$

the integrals being triple integrals with regard to x, y, and z, extended over the region of definition of φ. If P_1 and P_2 are two points in space, with coordinates (a_1, b_1, c_1) and (a_2, b_2, c_2), O being the origin, it is found directly that the squares of the distances OP_1, OP_2, P_1P_2 are represented by the integrals

$$\int (a_1x + b_1y + c_1z)^2\varphi, \qquad \int (a_2x + b_2y + c_2z)^2\varphi,$$

$$\int [(a_2 - a_1)x + (b_2 - b_1)y + (c_2 - c_1)z]^2\varphi.$$

A corresponding evaluation of the quotient

$$\frac{\int (a_1x+b_1y+c_1z)\,(a_2x+b_2y+c_2z)\,\varphi}{\sqrt{\int (a_1x+b_1y+c_1z)^2\varphi}\ \sqrt{\int (a_2x+b_2y+c_2z)^2\varphi}}$$

gives the familiar formula of solid analytic geometry for the cosine of the angle P_1OP_2.

If the last three equations in (6) are replaced by

$$\sqrt{\int x^2\varphi} = \sigma, \qquad \sqrt{\int y^2\varphi} = \tau, \qquad \sqrt{\int z^2\varphi} = \omega,$$

the unit values of the integrals may be restored by setting

$$\xi = x/\sigma, \quad \eta = y/\tau, \quad \zeta = z/\omega, \quad \sigma\tau\omega\varphi(x, y, z) = \Phi(\xi, \eta, \zeta).$$

If φ is not subject to (6) at all, except for the condition $\int\varphi = 1$, a reduction may be performed as follows. Let

$$x' = x, \qquad \sqrt{\int x'^2\varphi} = \sigma', \qquad \int x'y\varphi = p'_{12},$$

$$y' = y - (p'_{12}/\sigma'^2)x', \quad \sqrt{\int y'^2\varphi} = \tau', \quad \int x'z\varphi = p'_{13}, \quad \int y'z\varphi = p'_{23},$$

$$z' = z - (p'_{13}/\sigma'^2)x' - (p'_{23}/\tau'^2)y', \quad \sqrt{\int z'^2\varphi} = \omega',$$

$$\varphi(x, y, z) = \bar{\varphi}(x', y', z').$$

It is readily verified that

$$\int\int\int x'y'\bar{\varphi}(x', y', z')dx'dy'dz'$$

$$= \int\int\int x'z'\bar{\varphi}(x', y', z')dx'dy'dz'$$

$$= \int\int\int y'z'\bar{\varphi}(x', y', z')dx'dy'dz' = 0.$$

It remains to set

$$\xi = x'/\sigma', \qquad \eta = y'/\tau', \qquad \zeta = z'/\omega',$$

$$\sigma'\tau'\omega'\bar{\varphi}(x', y', z') = \Phi(\xi, \eta, \zeta),$$

and the desired reduction is accomplished; that is, Φ has the properties that

$$
\begin{aligned}
&\int\int\int \Phi d\xi d\eta d\zeta = 1, \\
&\int\int\int \xi\eta\Phi d\xi d\eta d\zeta = \int\int\int \xi\zeta\Phi d\xi d\eta d\zeta \\
&\qquad = \int\int\int \eta\zeta\Phi d\xi d\eta d\zeta = 0, \qquad (7) \\
&\int\int\int \xi^2\Phi d\xi d\eta d\zeta = \int\int\int \eta^2\Phi d\xi d\eta d\zeta \\
&\qquad = \int\int\int \zeta^2\Phi d\xi d\eta d\zeta = 1.
\end{aligned}
$$

Let it be understood henceforth that x, y, z are related to a new set of variables ξ, η, ζ, by equations of the form

$$
\begin{aligned}
x &= a_1\xi + b_1\eta + c_1\zeta, \\
y &= a_2\xi + b_2\eta + c_2\zeta, \\
z &= a_3\xi + b_3\eta + c_3\zeta,
\end{aligned}
$$

so that the conditions (7) are satisfied, when $\Phi(\xi, \eta, \zeta)$ is equal to $\varphi(x, y, z)$ multiplied by the absolute value of the determinant of the coefficients. The coefficients may be determined by the calculation of the preceding paragraph, or in any one of an infinite variety of other ways, the passage from one determination to another amounting to a rotation of coordinate axes, with or without a reflection in one of the coordinate planes. Let the point with coordinates (a_i, b_i, c_i) be denoted by P_i, for $i = 1, 2, 3$, and the origin, as usual, by O. The expressions (5) once more represent the distances OP_1 and OP_2 and the cosine of the angle P_1OP_2, while the corresponding expressions obtained by permuting the variables are to be similarly interpreted.

A foundation is thus laid for geometric reasoning on a more extensive scale. If U and V are any two linear combinations of x, y, and z,

$$U = l_1x + m_1y + n_1z, \qquad V = l_2x + m_2y + n_2z,$$

they can be expressed in the form

$$U = A_1\xi + B_1\eta + C_1\zeta, \qquad V = A_2\xi + B_2\eta + C_2\zeta,$$

and then $\int U^2\varphi$ is the square of the distance of the point (A_1, B_1, C_1) from the origin, $(\int UV\varphi)/\sqrt{(\int U^2\varphi)(\int V^2\varphi)}$ is the cosine of the angle subtended at the origin by the points (A_1, B_1, C_1) and $A_2, B_2, C_2)$, and so on. Any linear combination of x, y, and z corresponds to a definite point in three-dimensional space.

Suppose λ and μ are determined so as to minimize the integrals $\int(y-\lambda x)^2\varphi$ and $\int(z-\mu x)^2\varphi$. The points Q_2 and Q_3 corresponding to λx and μx are on the line OP_1, where this line is met by the perpendiculars from P_2 and P_3 respectively. Let $u = y-\lambda x$, $v = z-\mu x$; the coordinates of the points corresponding to u and v are the components of the vectors Q_2P_2 and Q_3P_3. Consequently the expression $(\int uv\varphi)/\sqrt{(\int u^2\varphi)(\int v^2\varphi)}$ is equal to the cosine of the dihedral angle between the planes P_1OP_2 and P_1OP_3. If $\int x\varphi = \int y\varphi = \int z\varphi = 0$, it is at the same time the coefficient of partial correlation between y and z. The geometric figure is exactly the same as in the case of a finite number of observations, and there is no need of repeating the steps by which the coefficients of partial correlation and of double correlation are calculated in terms of ordinary correlation coefficients.* It may be verified that the determinants

$$\begin{vmatrix} \int x^2\varphi & \int xy\varphi \\ \int yx\varphi & \int y^2\varphi \end{vmatrix}, \qquad \begin{vmatrix} \int x^2\varphi & \int xy\varphi & \int xz\varphi \\ \int yx\varphi & \int y^2\varphi & \int yz\varphi \\ \int zx\varphi & \int zy\varphi & \int z^2\varphi \end{vmatrix}$$

represent respectively the square of the area of the

* Cf. the papers on *The trigonometry of correlation* and *The elementary geometry of function space,* previously cited.

parallelogram determined by OP_1 and OP_2, and the square of the volume of the parallelepiped determined by OP_1, OP_2, and OP_3. The values of the determinants, which have the character of Gramian determinants, are therefore always positive.*

4. *Functions Defined over an Infinite Range.* As was said in the introduction, it is not the purpose of this paper to enter into an elaborate discussion of the questions of convergence that arise if the frequency functions considered are supposed defined and different from zero over an infinite range, in a highly arbitrary manner. There is no difficulty, however, in arriving at a formulation general enough to cover the most important statistical applications. Suppose, for example (to concentrate attention on the case of three variables), that

$$\varphi(x,y,z) < Ke^{-|x|-|y|-|z|}$$

for all values of x, y, and z, K being a constant. Then all the integrals that appear in the course of the discussion are absolutely convergent, so that the integral over a finite region approaches a definite limit as the boundary of the region recedes to infinity in any way whatever. To justify the transformations of variable in the infinite integrals, it is sufficient to carry out the transformations over corresponding finite regions, a sphere, for example, in one set of variables, and an ellipsoid in the other set, and then to allow these regions to expand to infinity. Furthermore, though this is not essential to the argument, all the triple integrals can be evaluated as iterated integrals.

THE UNIVERSITY OF MINNESOTA

* The fact that the two-rowed determinant is positive is equivalent to the fact that the value of a coefficient of correlation is always between -1 and $+1$. It is to be noticed that the extreme values $+1$ and -1 can not occur, under the hypotheses on which this paper has been based; a frequency function corresponding to perfect positive or negative correlation is not continuous as a function of the two or more variables involved.