CORRELATION AND GROUP THEORY*

BY D. J. STRUIK

1. *Introduction.* The following considerations, that were first presented at the meeting of the American Mathematical Society at Amherst in September 1928, might seem rather trivial. Indeed, their essential elements are well known to all statisticians. I have, however, never seen an explicit statement of these principles and it might therefore be useful to give a short presentation.†

2. *The Groups of Correlation.* Problems on correlation may be divided into three different groups.

A. Problems in which comparison is made between quantities that can not be expressed in the same units. For instance the marriage rate and the foreign trade of a country. Here the marriage rate may be expressed in number of marriages per thousand of population and the foreign trade in dollars. Change in unit in both cases may be entirely independent. The marriage rate may be expressed in percentage, or per million of population, the foreign trade in thousands of dollars, or in pounds sterling. If the two variables be denoted by x and y, I may just as well introduce variables x', y' defined by the equations

$$x' = \lambda x, \quad y' = \mu y, \tag{1}$$

where λ and μ are arbitrary independent constants.

B. Problems in which comparison is made between quantities that can be expressed in the same units. For instance heights of fathers and heights of sons, age of husband and age of wife. Here the only reasonable change in the variables x and y is the same change of scale

$$x' = \lambda x, \quad y' = \lambda y, \qquad (\lambda \text{ constant}).$$

C. To this type B, belong also problems in which the x and y

* Presented to the Society, September 6, 1928.

† A reference has been made by N. Wiener, *Harmonic analysis and quantum theory*, Journal of the Franklin Institute, vol. 207 (1929), pp. 525–534; particularly p. 531.

are mere coordinates, as the study of bullet holes in a target or the measurement of stars on photographic plates. As in the previous problem, the coordinates x and y are generally taken as rectangular coordinates, and they then allow the additional transformation (a rotation of angle α)

$$x' = x\cos\alpha + y\sin\alpha,$$
$$y' = -x\sin\alpha + y\cos\alpha.$$

3. *The First Group.* Let there be N points in the diagram, each with coordinates (x_n, y_n). Write

$$\bar{x} = \frac{1}{N}\sum x_n, \qquad \bar{y} = \frac{1}{N}\sum y_n,$$
$$\xi_n = x_n - \bar{x}, \qquad \eta_n = y_n - \bar{y},$$
$$N\sigma_x^2 = \sum\xi_n^2, \qquad N\sigma_{xy} = \sum\xi_n\eta_n, \quad N\sigma_y^2 = \sum\eta_n^2,$$

and let σ_x and σ_y be the positive root of σ_x^2 and σ_y^2. The elementary theory of correlation is the theory of invariants of the matrix

$$\begin{Vmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{Vmatrix}$$

under the given groups of transformations.

Take first the given affine transformation (1). Here

$$\bar{x}' = \lambda\bar{x}, \quad \bar{y}' = \mu\bar{y},$$
$$\xi_n' = \lambda\xi_n, \quad \eta_n' = \mu\eta_n,$$
$$\sigma_x' = \lambda\sigma_x, \quad \sigma_{xy}' = \lambda\mu\sigma_{xy}, \quad \sigma_y' = \mu\sigma_y.$$

The point $(\bar{x}, \bar{y})$ has therefore a geometric meaning: the *mean* of the distribution. The simplest rational invariant, is

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}, \quad r'^2 = r^2.$$

The number

$$r = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

is also invariant and may be positive or negative depending on σ_{xy}. It is the *correlation coefficient.* It can easily be shown that $r^2 \leqq 1$.

Invariant lines in the figure are

(a) those connecting the mean with the points (x_n, y_n);
(b) the lines through the mean parallel to the x and y axes;
(c) the lines

$$\eta = \frac{\sigma_{xy}}{\sigma_x^2}\xi = r\frac{\sigma_y}{\sigma_x}\xi,$$

$$\eta = \frac{\sigma_y^2}{\sigma_{xy}}\xi = r\frac{\sigma_x}{\sigma_y}\xi.$$

These lines are *regression lines*. They are generally defined as the lines through the mean for which the sum of the squares of the distances of the points (x, y) to the line, measured in direction parallel to the x and y axes, is a minimum. This is also an affine definition, invariant under transformations (A).*

4. *Central Meaning of r. Every function of σ_x, σ_y and σ_{xy} invariant under the transformations* (1) *is a function of r.* Indeed, such a function satisfies the condition

$$f(\lambda\sigma_x, \mu\sigma_y, \lambda\mu\sigma_{xy}) = f(\sigma_x, \sigma_y, \sigma_{xy}).$$

By a change of variables of functional determinant $\neq 0$

$$\sigma_x' = \sigma_x, \qquad \sigma_y' = \sigma_y, \qquad r = \frac{\sigma_{xy}}{\sigma_x\sigma_y}.$$

This takes the form

$$\phi(\lambda\sigma_x, \mu\sigma_y, r) = \phi(\sigma_x, \sigma_y, r),$$

so that ϕ as a function of σ_x and σ_y is independent of those variables, which proves the theorem. Now r (or $1/r$) is the simplest of those functions; in the practice of statistics it is to be preferred to $Cr($ or $C/r)$, C a constant $\neq 0$, because r runs from -1 to $+1$. A line through the mean of invariant character, depending only on σ_x, σ_y, σ_{xy}, must have the equation

$$\eta = \alpha\xi,$$

* The equation (in our geometric representation) of one regression line can be found in Laplace, *Théorie Analytique des Probabilités*, p. 318 and p. 326 of his *Oeuvres*, vol. 7 (ed. 1886). Laplace gets it, however, from considerations on errors of observation; he does not take two equivalent sets of variables as in the theory of correlation.

where α is a function of $\sigma_x, \sigma_y, \sigma_{xy}$ which is transformed as follows:

$$\alpha' = \frac{\mu}{\lambda}\alpha.$$

Therefore

$$\frac{\alpha'\sigma_x'}{\sigma_y'} = \frac{\alpha\sigma_x}{\sigma_y} = \text{a pure function of } r = \phi(r),$$

so that

$$\alpha = \frac{\sigma_y}{\sigma_x}\phi(r).$$

For

$$\phi(r) = r \quad \text{and} \quad \phi(r) = 1/r$$

we find the regression lines. Their equation is the simplest of those in which σ_{xy} enters.

5. *The Correlation Ellipses.* The regression lines and the lines parallel to the x and y axes form an involution. We arrange these lines in such a way that $\xi = 0$ is conjugate to

$$\eta = \frac{\sigma_y}{\sigma_x}\frac{1}{r}\xi$$

and $\eta = 0$ is conjugate to

$$\frac{\sigma_y}{\sigma_x} r\xi.$$

Then the equation of the involution is

$$\frac{\xi\xi_1}{\sigma_x^2} - \frac{r(\xi\eta_1 + \xi_1\eta)}{\sigma_{xy}} + \frac{\eta\eta_1}{\sigma_y^2} = 0.$$

The double lines of this involution are imaginary and asymptotes of the ellipses

$$\frac{\xi^2}{\sigma_x^2} - \frac{2r\xi\eta}{\sigma_x\sigma_y} + \frac{\eta^2}{\sigma_y^2} = \text{const.},$$

or

$$\sigma_y^2\xi^2 - 2\sigma_{xy}\xi\eta + \sigma_x^2\eta^2 = \text{const.}$$

In such a way the *correlation ellipses* are obtained. Their equation in line coordinates u, v is

$$\sigma_x^2 u^2 + 2\sigma_{xy} uv + \sigma_y^2 v^2 = \text{const.}$$

The regression lines are the lines conjugate to the axis directions with respect to the correlation ellipses.

The arrangement of the four lines in pairs is arbitrary. But the two other combinations would lead either to a set of conic sections without σ_{xy} in their equation or to a set with equation obtainable from the equation of the correlation ellipses by changing r into $1/r$. This last set has real asymptotes and their equation is

$$\frac{\xi^2}{\sigma_y^2} - \frac{2\xi\eta}{\sigma_x\sigma_y r} + \frac{\eta^2}{\sigma_x^2} = \text{const.}$$

The conic sections form a set of hyperbolas. The condition $r = 0$ leads here to asymptotes parallel to the x and y axes. The equation of the correlation ellipses in line coordinates shows, however, that the first curves are the natural ones. (See §8.)

6. *The Second Group.* In problems of type B, all invariants of problem A remain invariants. There are, however, new ones. To these belong, in particular, the symmetry axes of the correlation ellipses. This is geometrically obvious, as our only transformations are similarity transformations that keep the x and y axes in their place. It can easily be verified that the angle θ which the axis of symmetry make with the x axis is determined by the equation

$$\tan 2\theta = \frac{\dfrac{1}{\sigma_x^2} - \dfrac{1}{\sigma_y^2}}{\dfrac{2r}{\sigma_x\sigma_y}},$$

and this expression is unaltered by the transformation

$$x' = \lambda x, \quad y' = \lambda y.$$

If the x and y axes are orthogonal, the symmetry axes of the correlation ellipses are also defined by the property that the

sum of the squares of the orthogonal distances of the points (x_n, y_n) to this line is a minimum. In fact, let

$$y = mx + n$$

be the line satisfying this property; then

$$\phi(m, n) = \sum_{k=1}^{N} \frac{(y_k - mx_k - n)^2}{m^2 + 1}$$

must have a minimum value. This gives

$$\frac{\partial \phi}{\partial m} = 0, \quad \frac{\partial \phi}{\partial n} = 0,$$

or

$$\sum (m^2 + 1)(y_k - mx_k - n)x_k + (y_k - mx_k - n)^2 m = 0,$$
$$\sum (m^2 + 1)(y_k - mx_k - n) = 0.$$

These equations give

$$\bar{y} - m\bar{x} - n = 0,$$

which shows that the line must pass through the mean and

$$\frac{m}{1 - m^2} = \frac{\sigma_{xy}}{\sigma_x^2 - \sigma_y^2},$$

which, by means of the substitution

$$\tan 2\theta = \frac{2m}{1 - m^2},$$

passes into the previous equation for $\tan 2\theta$.*

If the x and y axes are not perpendicular to each other, the symmetry axes do not lose their invariant character. They are, however, not the lines corresponding to the symmetry axes in the case that x and y are plotted orthogonally. The symmetry axes pass into lines that can easily be determined by an affine transformation from the figure in orthogonal axes, if the direction and scale of the old and new x axes coincide, and the direction and scale of the new y axis are given.

* See, for example, K. Lundmark and W. J. Luyten, *On the determination of the colour-equivalent of a star, etc.*, Monthly Notices of the Royal Astronomical Society, vol. 82 (1922), pp. 495–509; particularly p. 505.

7. *The Third Group.* In problems of type C the invariant lines and numbers are those of type B that remain invariant under a rotation of the axes. That means that in such a diagram the symmetry axes, depending only on the position of the points (x_n, y_n) with respect to the mean, keep an invariant meaning.* But regression lines and coefficient of correlation have no meaning in this case, as they depend on the choice of x and y axes. The quantity

$$\sigma_{xy}^2 - \sigma_x^2 \sigma_y^2$$

is an invariant under rotation, but not under a change of scale. The equation has therefore an invariant meaning (perfect correlation).

8. *Partial Correlation.* In problems with more variables we have analogous properties.† In type A, the variables x_1, x_2, x_3, $\cdots$, x_n are affected in this way through a change of scale

$$x_1' = \lambda_1 x_1, \quad x_2' = \lambda_2 x_2, \quad x_3' = \lambda_3 x_3, \cdots, \quad x_n' = \lambda_n x_n,$$

where $\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_n$ are constants.

The object of the elementary correlation theory is here the matrix

$$\begin{Vmatrix} \sigma_{11} & \sigma_{12} \cdots \sigma_{1n} \\ \sigma_{12} & \sigma_{22} \cdots \sigma_{2n} \\ \cdots & \cdots \\ \cdots & \cdots \\ \sigma_{1n} & \sigma_{2n} \cdots \sigma_{nn} \end{Vmatrix},$$

where

$$\sigma_{jk} = \frac{1}{N} \sum (x_j - \bar{x}_j)(x_k - \bar{x}_k),$$

$$\bar{x}_l = \frac{1}{N} \sum x_l, \qquad \text{(sum on all points)}.$$

The elements of the correlation matrix are transformed by the formula

* See, for example, J. L. Coolidge, *An Introduction to Mathematical Probability*, Oxford, 1925, p. 144.

† See, for example, Yule, *An Introduction to the Theory of Statistics*, 6th ed., 1922.

$$\sigma'_{jk} = \lambda_j \lambda_k \sigma_{jk}.$$

An invariant expression is the form

$$\sigma_{11}{}^2 u_1{}^2 + 2\sigma_{12}u_1u_2 + \cdots + \sigma_{nn}u_n{}^2 = \sum \sigma_{jk}u_ju_k, \text{ (sum on all } j, k),$$

where u_k are hyperplane coordinates, so that

$$x_1u_1 + x_2u_2 + \cdots + x_nu_n = 1$$

represents the equation of a hyperplane in the $x_1, x_2, \cdots, x_n$ space. The hypersurfaces

$$\sigma_{11}u_1{}^2 + 2\sigma_{12}u_1u_2 + \cdots + \sigma_{nn}u_n{}^2 = \text{const.}$$

represent the *correlation quadrics* in hyperplane coordinates. In point coordinates, taken from the mean, $\xi_1, \xi_2, \cdots, \xi_n$, they have the equation

$$\begin{vmatrix} 0 & \xi_1 & \xi_2 & \cdots & \xi_n \\ \xi_1 & \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \xi_2 & \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \xi_n & \cdots & \cdots & \cdots & \sigma_{nn} \end{vmatrix} = \text{const.}$$

The rank of the matrix determines the shape of the correlation quadric.

The *regression hyperplanes** are the hyperplanes conjugate to the coordinate axes with respect to these quadric surfaces. They are n in number, and their equations are obtained by replacing one row of σ's in the determinant of the σ's and equating the result to zero, for instance,

$$\begin{vmatrix} \xi_1 & \xi_2 & \cdots & \xi_n \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1n} & \cdots & \cdots & \sigma_{nn} \end{vmatrix} = 0.$$

* See Laplace, *Théorie Analytique des Probabilités*, ed. 1886, p. 327, etc. Laplace, however, only deals with errors of observations, not with equivalent sets of variables.

If we call Σ_{ik} the minor of σ_{ik} in the determinant of the σ's this equation takes the form

$$\Sigma_{11}\xi_1 + \Sigma_{12}\xi_2 + \cdots + \Sigma_{1n}\xi_n = 0.$$

The *partial correlation coefficients* are of the form

$$r_{jk} = \frac{\Sigma_{jk}}{(\Sigma_{jj})^{1/2}(\Sigma_{kk})^{1/2}}.$$

The transformation formulas for the Σ's are of the form

$$\Sigma'_{jk} = \frac{\lambda_j\lambda_k}{\lambda_1^2\lambda_2^2 \cdots \lambda_n^2}\Sigma_{jk},$$

from which the invariant character of r_{jk} immediately follows. The sum of the squares of the distances from the points to the regression planes is in the case of the plane mentioned above

$$\sum \frac{(\xi_1 - \Sigma_{12}\xi_2 - \Sigma_{13}\xi_3 - \cdots - \Sigma_{1n}\xi_n)^2}{\Sigma_{11}^2} = \frac{\begin{vmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{1n} & \cdot & \cdots & \sigma_{nn} \end{vmatrix}}{\begin{vmatrix} \sigma_{22} & \cdots & \sigma_{2n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \sigma_{2n} & \cdots & \sigma_{nn} \end{vmatrix}}.$$

All functions of the σ_{jk} invariant under the given affine transformation are functions of the r_{jk}.

Indeed there are $n(n+1)/2$ quantities σ_{jk}, and $n(n-1)/2$ independent quantities r_{jk}. Since we have

$$\frac{n(n+1)}{2} - \frac{n(n-1)}{2} = n,$$

we can introduce into every function of the quantities σ_{jk} the r_{jk} as new independent variables, leaving the quantities $\sigma_{jj}(j=1, \cdots, n)$ unchanged. In the new function the quantities σ_{jj} change independently, so that invariance of that function means independence of the quantities σ_{jj}, and therefore only dependence on the quantities r_{jk}.

As *dual* partial correlation coefficient we might take the invariant numbers

$$\rho_{jk} = \frac{\sigma_{jk}}{(\sigma_{jj})^{1/2}(\sigma_{kk})^{1/2}}.$$

All invariant functions of the quantities σ_{jk} are also functions of the quantities ρ_{jk} only. But the quantities ρ_{jk} have not the simple relation to the regression hyperplanes.

In problems of type B in more variables the symmetry axes of the correlation quadric come into consideration.

In problems of type C the regression planes and the correlation coefficients lose their sense, but not the symmetry axes. Here the theory becomes the well known theory of quadratic matrices under orthogonal substitutions with the unessential modification that similarity transformations are also permitted.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

NOTE ON THE EXISTENCE OF A POSITIVE FUNCTION ORTHOGONAL TO A GIVEN SET OF FUNCTIONS*

BY N. H. McCOY†

Let the finite set of functions:

$$\{f_j(x)\}: \qquad f_1(x), f_2(x), \cdots, f_m(x)$$

be continuous and linearly independent on the closed interval X, $(a \leqq x \leqq b)$. With reference to this set of functions, L. L. Dines‡ has shown the equivalence of the following properties:

(A) *Every linear combination of the functions changes sign on* X.

(B) *There exists a positive continuous function orthogonal to each function of the set on* X.

A sufficient condition for the set $\{f_j(x)\}$ to have properties (A) and (B) has also been given by Dines.§ It is in a form

---

* Presented to the Society, September 11, 1930.

† National Research Fellow.

‡ *A theorem on orthogonal functions with an application to integral inequalities*, Transactions of this Society, vol. 30 (1928), pp. 425–438.

§ *On completely signed sets of functions*, Annals of Mathematics, vol. 28 (1926), pp. 393–395.