ON THE GENERALIZED DISTANCE IN STATISTICS.

By P. C. MAHALANOBIS.

(*Read January 4, 1936.*)

1. A normal (Gauss-Laplacian) statistical population in P-variates is usually described by a P-dimensional frequency distribution :—

$$df = \text{const.} \times e^{-\frac{1}{2\alpha}\left[A_{11}(x_1-\alpha_1)^2 + A_{22}(x_2-\alpha_2)^2 + \ldots\ldots + 2A_{12}(x_1-\alpha_1)(x_2-\alpha_2) + \ldots\ldots\right]} . dx_1 . dx_2 \,.\,. dx_P \qquad (1{\cdot}0)$$

where

$\alpha_1, \alpha_2 \ldots . \alpha_P$=the population (mean) values of the P-variates $x_1, x_2 \ldots . x_P$ .. .. (1·1)

$\alpha_{ii}=\sigma_i^2$, are the respective variances .. .. (1·2)

$\alpha_{ij}=\sigma_i . \sigma_j . \rho_{ij}$, where ρ_{ij}=the coefficient of correlation between the ith and jth variates .. .. (1·3)

α is the determinant $\mid \alpha_{ij} \mid$ defined more fully in (2·2), and A_{ij} is the minor of α_{ij} in this determinant.

A P-variate normal population is thus completely specified by the set of $P(P+1)/2$ parameters[*] :—

$$(\alpha_1, \alpha_2 \ldots . \alpha_P;\ \alpha_{11}, \alpha_{22} \ldots \alpha_{PP};\ \alpha_{12}, \alpha_{13} \ldots . \alpha_{P-1,\,P}) \quad .. \qquad (1{\cdot}4)$$

A second P-variate normal population can be specified in the same way by the parameters :—

$$(\alpha'_1, \alpha'_2 \ldots . \alpha'_P;\ \alpha'_{11}, \alpha'_{22} \ldots . \alpha'_{PP};\ \alpha'_{12}, \alpha'_{13} \ldots . \alpha'_{P-1,\,P}) \qquad (1{\cdot}5)$$

It is convenient to use the notation (α_i, α_{ij}) and $(\alpha'_i, \alpha'_{ij})$ to represent respectively the two sets of parameters given by (1·4) and (1·5). The α_i's or α'_i's are the mean values, while we may speak of (α_{ij}) or (α'_{ij}) as the respective dispersions.

2. It will be convenient at this stage to introduce the idea of a statistical field, such that at each point in this field there is a specified set of parameters (α_i, α_{ij}) which define a particular population. In other words, each point in a statistical field is the centre of a density-cluster belonging to a particular normal population completely specified by the value of (α_i, α_{ij}) at that particular point of the field.

[*] It may be noted here that such a population can be conveniently represented by a density-cluster in a P-dimensional space in which the position of the cluster is determined by the parameters $(\alpha_1, \alpha_2 \ldots \alpha_i \ldots \alpha_P)$, the size of the equi-frequential surfaces by $(\alpha_{11}, \alpha_{22} \ldots \alpha_{ii} \ldots \alpha_{PP})$, and the orientation of these surfaces by

$$(\alpha_{12}, \alpha_{13} \ldots \alpha_{ij} \ldots \alpha_{P-1,\,P}).$$

[Published April 15th, 1936.

I shall first consider a field in which the dispersion is same at all points. That is

$$\alpha_{ij} = \alpha'_{ij} \qquad .. \qquad .. \qquad .. \qquad .. \qquad (2·0)$$

for all values of i, j, and for all points in the field. In this case two populations can only differ in their mean values.

It has been shown elsewhere * that the distance between the two statistical populations can be conveniently measured by a Δ^2-statistics defined by

$$P \,.\, \Delta^2 = \underset{i=1}{\overset{i=P}{S}} \frac{(\alpha_i - \alpha'_i)^2}{\alpha_{ii}} \qquad .. \qquad .. \qquad .. \qquad (2·1)$$

for P independent variates.

This formula can be easily generalized for P correlated variates. Let us define the fundamental dispersion matrix :—

$$\alpha = \left| \begin{array}{ccccc} \alpha_{11}\,, & \alpha_{12}\,, & . & . & . \quad \alpha_{1P} \\ a_{21}\,, & a_{22}\,, & . & . & . \quad a_{2P} \\ . \quad . & . \quad . & . & . & . \quad . \\ \alpha_{P1}\,, & \alpha_{P2}\,, & . & . & . \quad \alpha_{PP} \end{array} \right| \qquad .. \qquad .. \qquad (2·2)$$

Let $\alpha^{ij} = \dfrac{\text{Co-factor of } \alpha_{ij} \text{ in } \alpha}{\text{Determinant } \alpha}$ .. .. (2·3)

The generalized Δ^2=statistics for P correlated variates can be now written in the form

$$P \,.\, \Delta^2 = \underset{i,j=1}{\overset{i,j=P}{S}} \left[\alpha^{ij} \,.\, (\alpha_i - \alpha'_i) \,.\, (\alpha_j - \alpha'_j)\right] \qquad .. \qquad .. \qquad (2·4)$$

Introducing the summation convention that a set of duplicated suffixes (say μ) will imply a summation for all values of $\mu = 1, 2, \ldots . P$, we get

$$P \,.\, \Delta^2 = \alpha^{\mu\nu} \,.\, (\alpha_\mu - \alpha'_\mu) \,.\, (\alpha_\nu - \alpha'_\nu) \qquad .. \qquad .. \qquad (2·41)$$

If we write

$$d\alpha_\mu \equiv (\alpha_\mu - \alpha'_\mu), \quad d\alpha_\nu \equiv (\alpha_\nu - \alpha'_\nu) \qquad .. \qquad .. \qquad (2.42)$$

we get

$$P \,.\, \Delta^2 = \alpha^{\mu\nu} \,.\, d\alpha_\mu \,.\, d\alpha_\nu \qquad .. \qquad .. \qquad .. \qquad (2·5)$$

It is slightly more convenient to change the notation a little, and define

$$\alpha^{ij} = \sigma_i \,.\, \sigma_j \,.\, \rho_{ij} \qquad .. \qquad .. \qquad .. \qquad .. \qquad (2·61)$$

$$\alpha_{ij} = \frac{\text{Co-factor of } \alpha^{ij} \text{ in } \alpha}{\text{Determinant } \alpha} \qquad .. \qquad .. \qquad .. \qquad (2·62)$$

and

$$(d\alpha)^i = (\alpha_i - \alpha'_i), \quad (d\alpha)_j = (\alpha_j - \alpha'_j) \qquad .. \qquad .. \qquad (2·63)$$

* *Journ. Asiat. Soc. Bengal,* vol. XXVI, pp. 541–588, (1930).

Equation (2·5) can then be written in the form

$$P \, . \, \Delta^2 = \alpha_{\mu\nu} \, . \, (d\alpha)^\mu \, . \, (d\alpha)^\nu \quad \ldots \quad \ldots \quad \ldots \quad (2{\cdot}7)$$

Comparing with the formula for ds^2

$$ds^2 = g_{\mu\nu} \, . \, (dx)^\mu \, . \, (dx)^\nu \quad \ldots \quad \ldots \quad \ldots \quad (2{\cdot}71)$$

we notice that $P \, . \, \Delta^2$ in statistics is the exact analogue of ds^2 in the restricted theory of relativity.

This merely implies that a consistent geometrical representation is possible in both cases. It is possible, however, to use this formal equivalence to establish an exact correspondence between results in the two subjects.

3. We see therefore that a statistical field in which the dispersion is same everywhere (values of $\alpha_{\mu\nu}$'s same at all points of the field and independent of mean values) corresponds to the physical field in the restricted theory of relativity ($g_{\mu\nu}$'s same everywhere and independent of co-ordinate values). In fact $\alpha_{\mu\nu}$'s play the same part in statistics as $g_{\mu\nu}$'s in the theory of relativity, and all the results involving ds^2 can be formally obtained from the results for a statistical field in which the dispersion is constant by putting $P=4$; x_1, x_2, x_3 as the three space co-ordinates, and $x_4 = ct\sqrt{-1}$, where t is the time co-ordinate, c is the velocity of light, and $P_{44} = -1$.

The possibility of transformation to Galilean co-ordinates in physics is now seen to be a special case of a more general transformation from a set of correlated statistical variates to a set of independent statistical variates which is always possible when the dispersion is constant.

4. The expression for the statistical distance

$$P \, . \, \Delta^2 = \alpha_{\mu\nu} \, . \, (d\alpha)^\mu \, . \, (d\alpha)^\nu \quad \ldots \quad \ldots \quad \ldots \quad (2{\cdot}7)$$

is given in terms of the population parameters $\alpha_{\mu\nu}$ and $(d\alpha)^\mu$, and is therefore not subject to sampling fluctuations. In other words equation (2·7) is a functional and not a statistical equation.

In actual practice we are, however, obliged to work with values calculated from finite samples. It is therefore necessary to convert the functional equation (2·7) into a statistical equation by replacing the population parameters $\alpha_{\mu\nu}$ and $(\alpha)^\mu$, $(\alpha')^\mu$ by sample statistics $a_{\mu\nu}$ and $(a)^\mu$, $(a')^\mu$. Weakening the equation (2·7) by such substitution we thus get

$$P \, . \, D_1^2 = a_{\mu\nu} \, . \, (da)^\mu \, . \, (da)^\nu \quad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}0)$$

where D_1^2 is the sample value of the Δ^2-statistic.

As the values of $a_{\mu\nu}$, or $(da)^\mu$ will in general fluctuate for samples drawn from the same two populations, the value of D_1^2 will also be subject to sampling fluctuations. We may consider the problem in two stages.

It is often possible to calculate the values of $a_{\mu\nu}$ by pooling together the variances and coefficients of correlation in a large number of samples on the assumption that the corresponding population values of $a_{\mu\nu}$ are identical. In such cases it is often possible to neglect the sampling fluctuations in $a_{\mu\nu}$ in comparison with the sampling fluctuations in the mean values $(a)^{\mu}$. We may then treat $a_{\mu\nu}$'s as constants and equal to $\alpha_{\mu\nu}$'s, and write :—

$$P \,.\, D_1^2 = \alpha_{\mu\nu} \,.\, (da)^{\mu} \,.\, (da)^{\nu} \quad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}1)$$

in which $(da)^{\mu}$ or $(a)^{\mu}$ and $(a)^{\nu}$ are considered to be subject to sampling fluctuations.

I had shown * that the mathematical expectation of D_1^2 was

$$E(D_1^2) = \Delta^2 + \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \Delta^2 + \frac{2}{\bar{n}} \quad \ldots \quad \ldots \quad (4{\cdot}11)$$

where n_1, n_2 are the respective sizes of the two samples.

We may now define the sample value of the D^2-statistics in the following form :—

$$P \,.\, D^2 = P \,.\, D_1^2 \; - \; \frac{2 \,.\, P}{\bar{n}} \quad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}12)$$

or

$$D^2 = D_1^2 - \frac{2}{\bar{n}} = \frac{1}{P} \cdot [\alpha_{\mu\nu} \,.\, (da)^{\mu} \,.\, (da)^{\nu}] - \frac{2}{\bar{n}} \quad \ldots \quad \ldots \quad (4{\cdot}2)$$

with

$$E(D^2) = \Delta^2 \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}21)$$

that is the mathematical expectation of D^2 is the population value Δ^2.

The moment-coefficients of D^2 were also obtained by me † :—

$$\mu_2(D^2) = \frac{8}{P \,.\, \bar{n}} \left[\Delta^2 + \frac{1}{\bar{n}}\right] \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}22)$$

$$\mu_3(D^2) = \frac{32}{P \,.\, \bar{n}^2} \left[3\Delta^2 + \frac{1}{\bar{n}}\right] \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}23)$$

$$\mu_4(D^2) = \frac{192}{P \,.\, \bar{n}^2} \left[\left(\Delta^2 + \frac{2}{\bar{n}}\right)^2 + \frac{4}{P \,.\, \bar{n}} \left(2\Delta^2 + \frac{1}{\bar{n}}\right)\right] \quad \ldots \quad (4{\cdot}24)$$

* *Journ. Asiat. Soc. Bengal,* vol. XXVI, p. 557, (1930).

† *Journ. Asiat. Soc. Bengal,* vol. XXVI, p. 559, (1930).

The exact distribution of D_1 was found some time ago by Raj Chandra Bose,* and can be written in the following form :—

Let $$\lambda^2=\tfrac{1}{2}\bar{n}\,.\,P\,.\,\Delta^2,\quad L^2=\tfrac{1}{2}\bar{n}\,.\,P\,.\,D_1^2 \qquad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}31)$$

Then the distribution of L is given by

$$f(L)\,.\,dL=\left(\frac{L}{\lambda}\right)^{\frac{P}{2}-1}.\,L\,.\,e^{-\frac{1}{2}(L^2+\lambda^2)}\,.\,I_{\frac{P}{2}-1}(L\lambda)\,.\,dL \qquad \ldots \quad (4{\cdot}32)$$

where I_z is the Bessel function of pure imaginary argument. It was also shown that the moment-coefficients previously given by me [equations (4·21) –(4·24)] were exact and remained valid for correlated variates. Five per cent. and one per cent. values of L for various values of λ have also been calculated by Raj Chandra Bose and Samarendra Nath Ray in conjunction with me and will be shortly published.†

Going a stage further we may consider both the dispersion $(a_{\mu\nu})$ as well as the mean values $(a)^{\mu}$ to be subject to sampling fluctuations, and relax the functional equation (2·7) completely into the statistical equation :—

$$P\,.\,D_2^2=a_{\mu\nu}\,.\,(da)^{\mu}\,.\,(da)^{\nu} \qquad \ldots \quad \ldots \quad \ldots \quad (4{\cdot}4)$$

The exact distribution of D_2^2 is not known, but Raj Chandra Bose has recently succeeded in obtaining the exact moment-coefficients (for the case of P independent variates) which are quoted in an appendix.

5. We have been considering so far the case of a statistical field of populations for which the dispersions ($\alpha_{\mu\nu}$'s) are identical and independent of the mean values $(\alpha)^{\mu}$'s. We may now relax this condition. That is we shall now assume that $\alpha_{\mu\nu}$'s are functions of $(\alpha)^{\mu}$'s. For our purpose it is, however, not necessary that $\alpha_{\mu\nu}$ should be a mathematical function of $(\alpha)^{\mu}$. It is clearly sufficient that $\alpha_{\mu\nu}$ and $(\alpha)^{\mu}$, that is the dispersions and mean values should be statistically connected. This point is one of considerable importance. We assume that given $(\alpha)^{\mu}$, that is the mean values, we have knowledge of the distribution of $\alpha_{\mu\nu}$'s. This we may call a generalized statistical field in which the values of $\alpha_{\mu\nu}$'s will vary from point to point in the field. In such a field we may still continue to use

$$P\,.\,D_1^2=a_{\mu\nu}\,.\,(da)^{\mu}\,.\,(da)^{\nu} \qquad \ldots \quad \ldots \quad \ldots \quad (5{\cdot}0)$$

as the expression for the line element. Here as also in the case where $\alpha_{\mu\nu}$

* *Science and Culture,* vol. I, p. 205, (1935).

† *Science and Culture,* vol. 1, (December, 1935). *Proc. Ind. Sc. Congress,* 1936.

are functionally connected with $(\alpha^{\mu}, \alpha^{\nu})$ the expression will be used only as a differential element the integral of which will depend on the path of integration. Equation (5·0) may be therefore considered to be a perfectly general expression* for the distance between two normal statistical populations. Before it can be used in practice it is, however, necessary to determine its exact distribution (or at least its moment-coefficients) which is now under investigation.

APPENDIX.

The moment-coefficients of the expression defined in (4·4) in the text are given below.

Let

$$D_i^2 = \frac{(a_i - a_i')^2}{(n_1\, a_{ii} + n_2\, a_{ii}')/(n_1 + n_2 - 1)}, \quad \Delta_i^2 = \frac{(\alpha_i - \alpha_i')^2}{\alpha_{ii}} \quad \ldots \quad (6{\cdot}0)$$

$$D_0^2 = \frac{1}{P} \mathop{S}_{i=1}^{i=P} (D_i^2), \qquad \Delta^2 = \frac{1}{P} \mathop{S}_{i=1}^{i=P} (\Delta_i^2) \quad \ldots \quad (6{\cdot}1)$$

the summation extending over all the variates. It is also convenient to put

$$n = n_1 + n_2 - 1 \quad \ldots \quad \ldots \quad \ldots \quad (6{\cdot}2)$$

We then have

$$\mu_1(D_0^2) = \frac{n}{n-3}\left(\Delta^2 + \frac{2}{n}\right) \quad \ldots \quad \ldots \quad \ldots \quad (6{\cdot}3)$$

To compensate for the bias introduced by the finite size of the sample we now introduce a new statistics defined by

$$D^2 = \frac{n-3}{n} \,.\, D_0^2 - \frac{2}{n} \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (6{\cdot}4)$$

We then have the mathematical expectation of D^2

$$E(D^2) = \mu_1\,(D^2) = \Delta^2 \quad \ldots \quad \ldots \quad \ldots \quad (6{\cdot}5)$$

* I may just mention here that there is another alternative method of approaching the problem. A P-variate normal population can be represented by a density cluster in P-dimensional space, and can always be transformed to a set of independent variates. Consider two such clusters or populations. They may differ in their mean values or the position of the clusters. They may also differ in the size of the equi-frequential surfaces. Finally they may differ in the set of independent variates in terms of which they can be specified. In general two such non-identical normal populations can be completely superposed by a translation (equalizing the difference in mean values), a squeeze (equalizing differences in variances), and a rotation (equalizing differences in coefficients of correlation). Various quantitative forms are possible, some of which are under investigation at present.

The moment-coefficients of D^2 are given by the formulæ :—

$$\mu_2(D^2)=\frac{2}{n-5}\left[\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^4}{P^2}\right)+\frac{4(n-2)}{\bar{n}\,.\,P}\Big\}\,\Delta^2+\frac{2}{\bar{n}}\Big\{\right] \quad \cdots \quad \cdots \quad \cdots \quad (6{\cdot}6)$$

$$\mu_3(D^2)=\frac{16}{(n-5)(n-7)}\left[\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^6}{P^3}\right)+\frac{6(n-2)}{\bar{n}\,.\,P}\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^4}{P^2}\right)\right.$$

$$\left.+\frac{2(n-1)(n-2)}{\bar{n}^2\,.\,P^2}\left\{3\,\Delta^2+\frac{2}{\bar{n}}\right\}\right] \quad \cdots \quad (6{\cdot}7)$$

$$\mu_4(D^2)=\frac{12(n+9)}{(n-5)(n-7)(n-9)}\left[\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^8}{P^4}\right)\right.$$

$$+\frac{8(n-2)}{\bar{n}\,.\,P}\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^6}{P^3}\right)+\frac{12}{(n-5)^2}\left[\overset{i=P}{\underset{i=1}{S}}\left(\frac{2\,\Delta_i^4\,\Delta_j^4}{P^4}\right)\right.$$

$$\left.+\frac{8(n-2)}{\bar{n}\,.\,P}\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^4\,\Delta_j^2}{P^3}\right)+\frac{16(n-2)^2}{\bar{n}^2\,.\,P^2}\left(\Delta^2+\frac{2}{\bar{n}}\right)\right]$$

$$+\frac{384(n-2)}{(n-5)^2(n-7)(n-9)}\left[\frac{10n^2-69n+81}{\bar{n}^2\,.\,P^2}\overset{i=P}{\underset{i=1}{S}}\left(\frac{\Delta_i^4}{P^2}\right)\right.$$

$$\left.+\frac{2(n^3-6n^2+2n+9)}{\bar{n}^3\,.\,P^3}\left(2\,\Delta^2+\frac{1}{\bar{n}}\right)\right] \quad \cdots \quad \cdots \quad \cdots \quad (6{\cdot}8)$$

Statistical Laboratory,
Presidency College, Calcutta.

CALCUTTA :—Published by the National Institute of Sciences of India, 1, Park Street,
and Printed by P. Knight, Baptist Mission Press, 41A, Lower Circular Road.