Note on Selection from a Multivariate Normal Population

By A. C. Aitken.

(*Received* 14th *November*, 1934. *Read* 7*th December*, 1934.)

The problem of statistical "selection" is concerned with the alteration induced in a frequency distribution in several variables by an alteration of the parameters in a subsection of the distribution. It may be illustrated by a simple trivariate case, as follows:

From a population characterised by variables x, y, z, correlated and normally distributed, with means 0, 0, 0, variances σ_1^2, σ_2^2, σ_3^2 and product variances $r_{12}\sigma_1\sigma_2$, $r_{13}\sigma_1\sigma_3$, $r_{23}\sigma_2\sigma_3$, a sub-population is extracted by selection in x alone, in such a way that after selection x is still normally distributed, but with mean h and variance s^2. It is required to determine the new values, in the selected population, of the means and variances of y and z, and of the product variances.

This type of problem was first considered (1) by K. Pearson, and was solved by him for the general case of n normally correlated variables of which p variables are selected, that is, are subjected to an alteration of their p means, p variances and $\frac{1}{2}p(p-1)$ product variances, the remaining $n-p$ variables undergoing an induced alteration in respect of their corresponding parameters. It was shown (2) by H. E. Soper that such problems, which include as special cases problems of partial correlation and regression, could be treated conveniently by the use of moment generating functions. This method, which is the same in principle as the Heaviside operational calculus and the use of Fourier, Hankel and other integral transforms in pure analysis, proceeds by setting up a duality between a frequency function and its exponential transform, defined by, *e.g.*

$$G(\alpha, \beta, \gamma) = \iiint_{-\infty}^{\infty} \phi(x, y, z)\, e^{\alpha x+\beta y+\gamma z}\, dx\, dy\, dz,$$

where $\phi(x, y, z)\, dx\, dy\, dz$ is the frequency differential. For example in the case of three normally correlated variables, with variances standardized to unity, we have the very convenient function

$$G(\alpha, \beta, \gamma) = \exp\{\tfrac{1}{2}(\alpha^2+\beta^2+\gamma^2+2r_{12}\,\alpha\beta+2r_{13}\,\alpha\gamma+2r_{23}\,\beta\gamma)\}.$$

https://doi.org/10.1017/S0013091500008063 Published online by Cambridge University Press

Since α interprets x, Soper proceeds by gathering all the terms in α into a squared expression, thus,

$$\exp\{\tfrac{1}{2}(\alpha + r_{12}\beta + r_{13}\gamma)^2 + \tfrac{1}{2}\beta^2(1-r_{12}^2) + \beta\gamma(r_{23}-r_{12}r_{13}) + \tfrac{1}{2}\gamma^2\ (1-r_{13}^2)\}.$$

He then replaces the exponential factor involving α by its equivalent integral involving frequency in x alone, thus

$$G(\alpha, \beta, \gamma) = \left\{(2\pi)^{-\frac{1}{2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, e^{(\alpha+r_{12}\beta+r_{13}\gamma)x}dx\right\} \exp\{\tfrac{1}{2}\beta^2(1-r_{12}^2) + \ldots.\}.$$

Evidently the effect of selection in x, producing a new mean h and a new variance s^2, will be to alter the frequency differential from

$$(2\pi)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}x^2)\,dx \quad \text{to} \quad (2\pi)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(x-h)^2/s^2\}\,dx/s.$$

Making this substitution and integrating back, Soper obtains at once the new moment generating function

$$\exp\{h(\alpha + r_{12}\beta + r_{13}\gamma) + \tfrac{1}{2}[s^2\alpha^2 + (1 - r_{12}^2\,\overline{1 - s^2})\beta^2 + \ldots.] \\ + r_{12}s^2\alpha\beta + r_{13}s^2\alpha\gamma + (r_{23} - r_{12}r_{13}\,\overline{1 - s^2})\beta\gamma\}.$$

By mere inspection the coefficients of α, β, γ in the exponent give the new means as h, $r_{12}h$, $r_{13}h$; the coefficients of $\alpha^2/2!$, $\beta^2/2!$, $\gamma^2/2!$ give the new variances as

$$s^2,\ 1 - r_{12}^2(1 - s^2),\ 1 - r_{13}^2(1 - s^2),$$

and the coefficients of $\alpha\beta$, $\alpha\gamma$, $\beta\gamma$ give the new product variances as

$$r_{12}s^2,\ r_{13}s^2,\ r_{23} - r_{12}r_{13}(1 - s^2).$$

These results furnish the complete solution of this trivariate problem; and the case $s = 0$ gives the solution of the corresponding problem in partial correlation.

General Solution of the Multivariate Problem.

By a slight simplification of the above approach, and the free use of the notation and methods of matrices, a perspicuous formulation of the solution of Pearson's general problem may be obtained.

The n variables may be viewed synoptically[1] as a vector x,

[1] Vectors x and α are matrices of one column of elements, x' and α' are matrices of one row obtained from x and α by "transposition." Quadratic forms are written $x'Ax$ or $\alpha'A\alpha$, where A is symmetric, *i.e.* $A' = A$.

https://doi.org/10.1017/S0013091500008063 Published online by Cambridge University Press

interpreted in the moment generating function by a vector α. It may be assumed without loss of generality that the vector of means is the null vector, and that the n variances have been standardized to have unit value. The moment generating function before selection is therefore, in a suitable notation,

$$G(\alpha) = (2\pi)^{-\frac{1}{2}n}\,|R|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}x'R^{-1}x\right) e^{\alpha' x}\,dx = \exp\left(\tfrac{1}{2}\alpha' R\alpha\right), \qquad (1)$$

where R is the matrix $[r_{ij}]$ of total correlation coefficients, and where the integral is n-ple, dx standing for $dx_1\,dx_2\,\ldots.\,dx_n$.

Lemma. Since

$$\int_{-\infty}^{\infty} \phi(x-h)\,e^{\alpha' x}\,dx = e^{\alpha' h}\int_{-\infty}^{\infty} \phi(x)\,e^{\alpha' x}\,dx, \qquad (2)$$

a vector change of origin from the vector 0 to the vector h induces a factor $\exp(\alpha' h)$ in the moment generating function.

In preparation for selection according to the first p of the n variables, the matrix R is partitioned thus,

$$R = \left[\begin{array}{c:c} R_{pp} & R_{pq} \\ \hdashline R_{qp} & R_{qq} \end{array}\right], \qquad (3)$$

that is, into submatrices of orders $p\times p$, $p\times q$, $q\times p$, $q\times q$, where $p+q=n$. The vector α is correspondingly partitioned, its first p elements being regarded as a vector which without confusion can still be denoted by α, provided the remaining q elements are distinguished as a vector β. The quadratic form $\frac{1}{2}\alpha' R\alpha$ thus appears in partitioned shape as

$$\begin{aligned} &\tfrac{1}{2}\left(\alpha' R_{pp}\,\alpha + 2\alpha' R_{pq}\,\beta + \beta' R_{qq}\,\beta\right) \\ &= \tfrac{1}{2}\{(\alpha + R_{pp}^{-1}\,R_{pq}\,\beta)' R_{pp}\,(\alpha + R_{pp}^{-1}\,R_{pq}\,\beta) + \beta'(R_{qq} - R_{qp}\,R_{pp}^{-1}\,R_{pq})\,\beta\}, \qquad (4) \end{aligned}$$

the vector α being segregated in a first quadratic term.

The exponential of this first quadratic term may be replaced by a multiple integral, involving a frequency differential (with matrix R_{pp}^{-1} in the integrand; *cf.* (1) above) in the first p variables only, multiplied by $\exp\{(\alpha + R_{pp}^{-1}\,R_{pq}\,\beta)' x\}$.

Now let selection operate, (i) by altering the vector of means of these first p variables from 0 to h. By the Lemma noted above in (2), the moment generating function acquires a factor

$$\exp\{(\alpha + R_{pp}^{-1}\,R_{pq}\,\beta)'\,h\}.$$

https://doi.org/10.1017/S0013091500008063 Published online by Cambridge University Press

This yields the first general result, that all the new means are given as coefficients of the α_i and β_i in the elements of the vector

$$(\alpha + R_{pp}^{-1} R_{pq}\beta)' h. \tag{5}$$

Next, (ii) let the variances and product variances of the first p variables be altered. The effect of this is that R_{pp}^{-1}, the matrix in the integrand of frequency for the first p variables, is replaced by some other symmetric matrix, say V_{pp}^{-1}. On integrating back (*cf.* again (1) above) the *reciprocal* matrix V_{pp} must therefore replace R_{pp} as matrix of the first quadratic term in (4). Hence the whole quadratic form (4) is altered by selection to

$$\tfrac{1}{2}\{(\alpha + R_{pp}^{-1} R_{pq}\beta)' V_{pp}(\alpha + R_{pp}^{-1} R_{pq}\beta) + \beta'(R_{qq} - R_{qp}R_{pp}^{-1}R_{pq})\beta\}$$
$$= \tfrac{1}{2}\{\alpha' V_{pp}\alpha + 2\alpha' V_{pp}R_{pp}^{-1}R_{pq}\beta + \beta'(R_{qq} - R_{qp}R_{pp}^{-1}R_{pq} + R_{qp}R_{pp}^{-1}V_{pp}R_{pp}^{-1}R_{pq})\beta\}. \tag{6}$$

The results of Pearson and Soper are contained in the two statements (5) and (6). The second may be expressed thus: the effect of selection is to change

$$\left[\begin{array}{c:c} R_{pp} & R_{pq} \\ \hdashline R_{qp} & R_{qq} \end{array}\right] \text{ into } \left[\begin{array}{c:c} V_{pp} & V_{pp}R_{pp}^{-1}R_{pq} \\ \hdashline R_{qp}R_{pp}^{-1}V_{pp} & R_{qq} - R_{qp}(R_{pp}^{-1} - R_{pp}^{-1}V_{pp}R_{pp}^{-1})R_{pq} \end{array}\right]. \tag{7}$$

Example. To deduce the formulae for selection in one of three variables.

$$\text{Here}\qquad R = \left[\begin{array}{c:cc} 1 & r_{12} & r_{13} \\ \hdashline r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{array}\right], \qquad V_{pp} = [s_1^2].$$

By the general result, the matrix after selection becomes

$$\left[\begin{array}{c:l} [s_1^2] & [s_1^2][1][r_{12}\quad r_{13}] \\ \hdashline \text{etc.} & \begin{bmatrix} 1 & r_{23} \\ r_{23} & 1 \end{bmatrix} - \begin{bmatrix} r_{12} \\ r_{13} \end{bmatrix}[1][r_{12}\quad r_{13}] + \begin{bmatrix} r_{12} \\ r_{13} \end{bmatrix}[s_1^2][r_{12}\quad r_{13}] \end{array}\right].$$

$$= \begin{bmatrix} s_1^2 & r_{12}s_1^2 & r_{13}s_1^2 \\ r_{12}s_1^2 & 1 - r_{12}^2(1 - s_1^2) & r_{23} - r_{12}r_{13}(1 - s_1^2) \\ r_{13}s_1^2 & r_{23} - r_{12}r_{13}(1 - s_1^2) & 1 - r_{13}^2(1 - s_1^2) \end{bmatrix}$$

https://doi.org/10.1017/S0013091500008063 Published online by Cambridge University Press

and here all the results of Soper's example are visible. As for the means, we have

$$(a + R_{pp}^{-1} R_{pq} \beta)' h \equiv (a_1 + r_{12} a_2 + r_{13} a_3) h,$$

which gives them as h, $r_{12} h$, $r_{13} h$, again as found earlier.

REFERENCES.

1. K. Pearson, *Phil. Trans. Roy. Soc.*, London, 200 A (1902), 1-66.
2. H. E. Soper, *Frequency Arrays* (Cambridge, 1922), 20-21.

https://doi.org/10.1017/S0013091500008063 Published online by Cambridge University Press