IV.—A Note on Karl Pearson's Selection Formulæ.

By **D. N. Lawley**, B.A., Moray House, University of Edinburgh. *Communicated by* Prof. GODFREY H. THOMSON, D.C.L., D.Sc., Ph.D.

(MS. received May 5, 1943. Read July 5, 1943)

IT is a well-known fact that Karl Pearson's formulæ expressing the effect of selection on the means, variances and covariances of a multivariate population hold when the variates are such as to be normally distributed both before and after selection. It is not, however, generally known that the formulæ are true under much more general conditions, and in view of a recent controversy it has been thought desirable to establish precisely what these conditions are. In dealing with the problem we shall adopt the shortened vector and matrix notation introduced by Aitken (1). This notation is reproduced below with but slight modifications.

We suppose the differential of frequency for the n-variate population under consideration to be given by

$$df = \phi(x, y)dxdy, \tag{1}$$

where x is used to denote the p variables which are to undergo selection, y the $n-p$ unselected variables, and $\phi(x, y)$ denotes the frequency function

$$\phi(x_1, x_2, \ldots x_p, y_1, y_2, \ldots y_{n-p});$$

dx and dy are also used to indicate the products of differentials dx_i and dy_j, and all variates are measured with respect to their means as origin.

The moment generating function of the distribution is defined to be

$$\begin{aligned} M(t, u) &= \iint \phi(x, y) e^{i(t'x + u'y)} dxdy \\ &= 1 - [t' : u']V\{t : u\}/2! + \ldots, \end{aligned} \tag{2}$$

where t', u' are row vectors of p and $n-p$ elements respectively, x and y are corresponding column vectors, and V is the variance matrix, having the n variances for its diagonal elements and the $\frac{1}{2}n(n-1)$ covariances for its non-diagonal elements. The integration is over the whole range of the x and the y.

Let the differential of frequency for the x alone, variation in the y being ignored, be $\phi(x)dx$. It may be obtained by integrating the differential in (1) over the whole range of the y. Let also the distribution of the y for a fixed set of values of the x be given by the differential $f(x, y)dy$. Then we shall have

$$\phi(x, y) = \phi(x) . f(x, y). \tag{3}$$

It will be convenient to partition the variance matrix V according to selected and unselected variates. The variance quadric of (2) may then be expressed in the form

$$\begin{aligned} &[t' : u'] \left[\begin{array}{c|c} V_{pp} & V_{p, n-p} \\ \hline V_{n-p, p} & V_{n-p, n-p} \end{array}\right] \left[\begin{array}{c} t \\ \hline u \end{array}\right] \\ &= t'V_{pp}t + 2t'V_{p, n-p}u + u'V_{n-p, n-p}u \\ &= (t + \beta u)'V_{pp}(t + \beta u) + u'(V_{n-p, n-p} - \beta' V_{pp}\beta)u, \end{aligned} \tag{4}$$

where

$$\beta = V_{pp}^{-1} V_{p, n-p}.$$

It may be noted that the elements of the matrix β are the linear regression coefficients of the y on the x.

Using (3) we see that

$$M(t, u) = \iint \{f(x, y)e^{iu'(y-\beta'x)}\}\phi(x)e^{i(t+\beta u)'x}dxdy$$

$$= \int g(u, x)\phi(x)e^{i(t+\beta u)'x}dx,$$

where

$$g(u, x) = \int f(x, y)e^{iu'(y-\beta'x)}dy.$$

Now suppose that $f(x, y)$ can be expressed as a function of the set of values $y - \beta'x$. This implies that the regression of the y on the x is linear and also that the form of distribution of the y for a fixed set of values of the x remains constant (except of course for changes of mean). Then $g(u, x)$ does not involve the x and becomes a function of the u alone, say $g(u)$. Hence we shall have

$$M(t, u) = g(u)\int \phi(x)e^{i(t+\beta u)'x}dx.$$

If we observe that

$$\int \phi(x)e^{i(t+\beta u)'x}dx = 1 - (t + \beta u)'V_{pp}(t + \beta u)/2! + \ldots,$$

then in view of (4) it is clear that

$$g(u) = 1 - u'(V_{n-p,\,n-p} - \beta'V_{pp}\beta)u/2! + \ldots \tag{5}$$

Now the effect of selection is to substitute for $\phi(x)dx$ another differential $\psi(x - h)dx$, with a new vector of means h and a new variance matrix W_{pp} in the selected variates. The moment generating function after selection is therefore

$$g(u)\int \psi(x - h)e^{i(t+\beta u)'x}dx$$

$$= g(u)e^{i(t+\beta u)'h}\{1 - (t + \beta u)'W_{pp}(t + \beta u)/2! + \ldots\}$$

$$= 1 + i(t + \beta u)'h - [u'(V_{n-p,\,n-p} - \beta'V_{pp}\beta)u + (t + \beta u)'W_{pp}(t + \beta u)]/2! + \ldots$$

$$= 1 + i(t + \beta u)'h - [t'W_{pp}t + 2t'(W_{pp}\beta)u + u'\{V_{n-p,\,n-p} - \beta'(V_{pp} - W_{pp})\beta\}u]/2! + \ldots$$

Remembering that $\beta = V_{pp}^{-1}V_{p,\,n-p}$, we thus see that the means of the variates y in the modified population are given by the $n - p$ elements of the vector $V_{n-p,\,p}V_{pp}^{-1}h$, while the new variance matrix for the whole n variates is

$$\left[\begin{array}{c|c} W_{pp} & W_{pp}V_{pp}^{-1}V_{p,\,n-p} \\ \hline V_{n-p,\,p}V_{pp}^{-1}W_{pp} & V_{n-p,\,n-p} - V_{n-p,\,p}(V_{pp}^{-1} - V_{pp}^{-1}W_{pp}V_{pp}^{-1})V_{p,\,n-p} \end{array}\right].$$

These results are the selection formulæ when expressed in matrix notation.

The foregoing argument shows that the restriction imposed above on the form of $f(x, y)$ is *sufficient* for the selection formulæ to hold. A less rigorous restriction is, however, permissible. For it is only necessary that in the expansion of $g(u, x)$ (previously defined) in powers of the u the linear and quadratic terms should be independent of the x; only these terms have any influence on the means and the variance matrix. Expanding $g(u, x)$ we obtain

$$g(u, x) = 1 + \int \{iu'(y - \beta'x) - u'(y - \beta'x)(y - \beta'x)'u/2! + \ldots\}f(x, y)dy;$$

so that for the linear and quadratic terms in the u to be independent of the x we must have

$$\int (y - \beta'x)_j f(x, y)dy = 0, \tag{6}$$

$$\int (y - \beta'x)_j(y - \beta'x)_k f(x, y)dy = c_{jk}, \tag{7}$$

where $(y-\beta'x)_j$ represents any element of the vector $y-\beta'x$ and $\{c_{jk}\}$ are a set of constants independent of the x. From (5) it will be seen that c_{jk} is the typical element in the matrix.

$$(V_{n-p,\,n-p}-\beta'V_{pp}\beta)=(V_{n-p,\,n-p}-V_{n-p,\,p}V_{pp}^{-1}V_{p,\,n-p}).$$

Equations (6) express the condition that the regression of the y on the x should be linear, while equations (7) express the condition that the variances and covariances of the distribution of the y for a fixed set of values of the x should be constant, *i.e.* independent of the x. These are the two necessary and sufficient conditions for the selection formulæ to hold.

Since the conditions refer to relationships existing between the y and the x and take no account of the form of distribution of the x alone, it is clear that the selection formulæ, if once applicable, may again be applied when a second selection is performed on the already selected population. If the second selection is such as to restore W_{pp} to V_{pp} then it may be verified that this selection will at the same time restore the *whole* variance matrix V. A reciprocity thus exists between the parent and the selected population. It will also be seen that the operations $V_{pp}\rightarrow W_{pp}$ and $W_{pp}\rightarrow U_{pp}$, when performed in succession, are equivalent to the single operation $V_{pp}\rightarrow U_{pp}$; in other words, the operation of selection is transitive. These facts have already been noted by Aitken.

REFERENCE TO LITERATURE

(1) AITKEN, A. C., 1934. "Note on Selection from a Multivariate Normal Population," *Proc. Edin. Math. Soc.*, IV, 106–110.

(*Issued separately November* 17, 1943.)