OPTIMUM STRATIFICATION WITH TWO CHARACTERS[1]

By S. P. Ghosh[2]

University of California, Berkeley

1. Introduction. An attempt will be made here to extend Dalenius' (1950) theory for univariate stratification to more than one variate. Here we shall be concerned with the most general theory for stratification with bivariates. The generalized variance of the sample means will be taken as a measure of precision for the bivariate characters. We shall call a system of stratification the *optimum stratification*, if it minimizes the generalized variance of the sample means.

In this paper we shall confine ourselves only to rectilinear stratification, i.e., stratification by lines parallel to the axes; and within this class, we shall look for the stratification points for which the optimum stratification is achieved. We shall assume that the number of strata are predetermined and the sampling allocation shall be taken to be proportional allocation. To start with we shall assume that the stratification variables are identical with the variables under analysis, but afterward we shall dispense with this restriction.

2. Mathematical formulation. We shall make the following assumptions.

(i) The bivariates (x, y) have a continuous probability density function $f(x, y)$ in the finite range $x_0 \leqq x \leqq x_k$, $y_0 \leqq y \leqq y_l$.

(ii) The population is infinite.

Though these assumptions will not, in general, be satisfied, yet in practice they will be approximately satisfied. Our problem is to divide this population, defined over the rectangle $[x_0 , x_k] \times [y_0 , y_l]$, into $k \times l$ strata by dividing x at points $x_1 , x_2 , \cdots , x_{k-1}$ and y at the points $y_1 , y_2 , \cdots , y_{l-1}$ such that the generalized variance of the means of the characters of a stratified sample with proportional allocation is minimum with respect to choice of these cut-off points. As our system of stratification of x's and y's will be cutting across each other, i.e., a two-way rectilinear stratification, our solution to the problem would be optimum stratification among rectilinear stratifications. We shall need a set of notation which we shall define below:

$$p_{ij} = \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} f(x, y)\, dx\, dy = \text{proportion in the } (i, j)\text{th cell},$$

$$f_{x_i j} = \int_{y_{j-1}}^{y_j} f(x_i , y)\, dy = \text{marginal density of the } j\text{th cell of } y \text{ evaluated at the point } x_i ,$$

Received March 14, 1962; revised March 12, 1963.

[1] This paper was prepared with the partial support of the Office of Ordnance Research, U. S. A., Grant (DA-ARO(D)-31-124-G183).

[2] Now at Thomas J. Watson Research Center, Yorktown Heights, New York.

$f_{iy_j} = \int_{x_{i-1}}^{x_i} f(x, y_j)\,dx$ = marginal density of the ith cell of x evaluated at the point y_j.

Hence it follows that

$$\mu_{xij} = \frac{1}{p_{ij}} \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} xf(x, y)\,dx\,dy = \text{mean of } x \text{ in the } (i, j)\text{th cell},$$

$$\sigma_{xij}^2 = \frac{1}{p_{ij}} \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} (x - \mu_{xij})^2 f(x, y)\,dx\,dy = \text{variance of } x \text{ in the } (i, j)\text{th cell}.$$

Similarly, we can define μ_{yij} and σ_{yij}^2,

$\mu_{xiy_j} = \frac{1}{f_{iy_j}} \int_{x_{i-1}}^{x_i} xf(x, y_j)\,dx$ = conditional mean of x in the ith cell of x evaluated at the point y_j,

$\mu_{yx_ij} = \frac{1}{f_{x_ij}} \int_{y_{j-1}}^{y_j} yf(x_i, y)\,dy$ = conditional mean of y in the jth cell of y evaluated at the point x_i,

$\sigma_{xyij} = \frac{1}{p_{ij}} \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} (x - \mu_{xij})(y - \mu_{yij}) f(x, y)\,dx\,dy$ = convariance of x and y in the (i, j)th cell.

Suppose we draw a stratified sample of size n, with n_{ij} observations from the (i, j)th cell of this population, and calculate the following sample statistics. Let

(x_{ijk}, y_{ijk}) = kth observation from the (i, j)th cell in the sample, $k = 1, 2, \cdots, n_{ij}$

$$\bar{x}_{ij} = (1/n_{ij}) \sum_{k=1}^{n_{ij}} x_{ijk} = \text{sample mean of } x \text{ in the } (i, j)\text{th cell},$$

$$\bar{x} = \sum_i \sum_j p_{ij} \bar{x}_{ij} = \text{mean of } x \text{ for the stratified sample}.$$

Similarly, we can define $\bar{y}_{ij}$ and $\bar{y}$.

The sampling is independent among the strata. Hence

$$\sigma_x^2 = V(\bar{x}) = \sum_i \sum_j p_{ij}^2 \sigma_{xij}^2 / n_{ij}$$

$$\sigma_y^2 = V(\bar{y}) = \sum_i \sum_j p_{ij}^2 \sigma_{yij}^2 / n_{ij}$$

$$\sigma_{xy} = \operatorname{Cov}(\bar{x}, \bar{y}) = \sum_i \sum_j p_{ij}^2 \sigma_{xyij} / n_{ij}.$$

Hence the generalized variance (G) of the means of a stratified sample will be given by the determinant

$$G = \begin{vmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{vmatrix}.$$

3. Solution for proportional allocation. For proportional allocation with fixed total sample size, the allocations are given by $n_{ij} = p_{ij}n$. Hence we have $\sigma_x^2 = \sum_i \sum_j p_{ij}\sigma_{xij}^2/n$, $\sigma_y^2 = \sum_i \sum_j p_{ij}\sigma_{yij}^2/n$, $\sigma_{xy} = \sum_i \sum_j p_{ij}\sigma_{xyij}/n$.

THEOREM. *The optimum stratification points* (x_i, y_j), $i = 1, 2, \cdots, k-1$; $j = 1, 2, \cdots, l-1$, *among the class of rectilinear stratification, of a bivariate* p.d.f. $f(x, y)$ *with proportional allocation is given by*

$$x_i = D_3(x_i, y) + D_4(x_i, y) \qquad \text{for } i = 1, 2, \cdots, k-1, \tag{3.1}$$

where

$$D_3(x_i, y) = \begin{vmatrix} \dfrac{\sum_j f_{xij}(\mu_{x\,i+1\,j}^2 - \mu_{xij}^2)}{2\sum_j f_{xij}(\mu_{x\,i+1\,j} - \mu_{xij})} & \dfrac{\sigma_{xy}}{\sigma_y^2} \\ \dfrac{\sum_j f_{xij}\{(\mu_{yx_ij} - \mu_{y\,i+1\,j})(x_i - \mu_{x\,i+1\,j}) - (\mu_{yx_ij} - \mu_{yij})(x_i - \mu_{xij})\}}{2\sum_j f_{xij}(\mu_{x\,i+1\,j} - \mu_{xij})} & 1 \end{vmatrix}$$

$$D_4(x_i, y) = \begin{vmatrix} \dfrac{\sigma_x^2}{\sigma_y^2} & \dfrac{\sum_j f_{xij}\{(\mu_{yx_ij} - \mu_{y\,i+1\,j})(x_i - \mu_{x\,i+1\,j}) - (\mu_{yx_ij} - \mu_{yij})(x_i - \mu_{xij})\}}{2\sum_j f_{xij}(\mu_{x\,i+1\,j} - \mu_{xij})} \\ \dfrac{\sigma_{xy}}{\sigma_y^2} & \dfrac{\sum_j f_{xij}(\mu_{y\,i+1\,j} - \mu_{yij})(\mu_{y\,i+1\,j} + \mu_{yij} - 2\mu_{yx_ij})}{2\sum_j f_{xij}(\mu_{x\,i+1\,j} - \mu_{xij})} \end{vmatrix}$$

and

$$y_j = D_7(x, y_j) + D_8(x, y_j) \qquad \text{for } j = 1, 2, \cdots, l-1, \tag{3.2}$$

where

$$D_7(x, y_j) = \begin{vmatrix} \dfrac{\sum_i f_{iy_j}(\mu_{xi\,j+1} - \mu_{xij})(\mu_{xi\,j+1} + \mu_{xij} - 2\mu_{xiy_j})}{2\sum f_{iy_j}(\mu_{yi\,j+1} - \mu_{yij})} & \dfrac{\sigma_{xy}}{\sigma_x^2} \\ \dfrac{\sum_i f_{iy_j}\{(\mu_{xiy_j} - \mu_{xi\,j+1})(y_j - \mu_{yi\,j+1}) - (\mu_{xiy_j} - \mu_{xij})(y_j - \mu_{yij})\}}{2\sum_i f_{iy_j}(\mu_{yi\,j+1} - \mu_{yij})} & \dfrac{\sigma_y^2}{\sigma_x^2} \end{vmatrix}$$

$$D_8(x, y_j) = \begin{vmatrix} 1 & \dfrac{\sum_i f_{iy_j}\{\mu_{xiy_j} - \mu_{xi\,j+1})(y_j - \mu_{yi\,j+1}) - (\mu_{xiy_j} - \mu_{xij})(y_j - \mu_{yij})\}}{2\sum_i f_{iy_j}(\mu_{yi\,j+1} - \mu_{yij})} \\ \dfrac{\sigma_{xy}}{\sigma_x^2} & \dfrac{\sum_i f_{iy_j}(\mu_{yi\,j+1}^2 - \mu_{yij}^2)}{2\sum_i f_{iy_j}(\mu_{yi\,j+1} - \mu_{yij})} \end{vmatrix}.$$

The details of the proof are given in the appendix.

REMARK 1. As in the univariate case, the algebraic determination of the optimum stratification points is a very difficult task, however, they can be evaluated numerically by applying the method of bivariate iteration to Equations (3.1) and (3.2) with Dalenius' solution (1950) for univariate as the starting point for each variable.

COROLLARY 1. *Dalenius' solution for univariate stratification with proportional allocation follows as a special case.*

PROOF. In order to pass from the bivariate situation to the univariate case, we shall consider the following type of limiting situation. We shall suppose that the variable y degenerates on the x-axis, i.e., $\mu_{yij} = \mu_{y\,i+1\,j} = \mu_{yx_ij} = 0$ and f_{x_ij}, μ_{xij}, $\mu_{x\,i+1\,j}$ are independent of j. Hence $D_4(x_i, y) = 0$ and (3.1) becomes $x_i = D_3(x_i, y) = \frac{1}{2}(\mu_{x\,i+1} + \mu_{xi})$.

COROLLARY 2. *The optimum double dichotomy point for a symmetric bivariate distribution is the center of gravity of the distribution.*

PROOF. Without loss of generality we can take the center of gravity of the distribution to be (0, 0). The marginal distributions will also be symmetric about zero, hence by application of Dalenius' method to each variable separately it follows that we may take the starting point of our iteration method as (0, 0). Hence from symmetry it follows that $\mu_{x\,i+1\,j+1} = -\mu_{xij}$, $\mu_{x\,i+1\,j} = -\mu_{xi\,j+1}$, $f_{x_ij} = f_{x_ij+1}$ and $\mu_{yx_ij} = -\mu_{yx_ij+1}$ at $x_i = 0$, which implies $D_3(x_i, y) = D_4(x_i, y) = 0$. Hence from (3.1) it follows that $x_i = 0$. Similarly, from symmetry it follows that $\mu_{y\,i+1\,j+1} = -\mu_{yij}$, $\mu_{y\,i+1\,j} = -\mu_{yi\,j+1}$, $f_{iy_j} = f_{i+1\,y_j}$ and $\mu_{xy_ji} = -\mu_{xy_ji+1}$ at $y_j = 0$ which implies $D_7(x, y_j) = D_8(x, y_j) = 0$. Hence from (3.2) it follows that $y_j = 0$. Hence the corollary is proved.

REMARK 2. Thus for well-known bivariate distributions as bivariate normal, bivariate rectangular distribution, etc., the optimum double dichotomy point is the center of gravity.

4. Optimum stratification when stratification is carried on a specific stratification variable. As before, we shall assume the variables under analysis are (x, y) and the stratification variables shall be denoted by (u, v) and they are related in the following manner,

$$x = \phi(u, v) + \epsilon$$

$$y = \psi(u, v) + \eta$$

where ϕ and ψ are such that $\phi^{-1}(x)$ and $\psi^{-1}(y)$ are defined. ϵ and η are random variables with $E(\epsilon) = E(\eta) = 0$ and $\rho(\epsilon, \eta) = \rho(\epsilon, \phi) = \rho(\epsilon, \psi) = \rho(\eta, \psi) = \rho(\eta, \phi) = 0$ but $\rho(\phi, \psi) \neq 0$. The variances of ϕ, ψ, ϵ, and η shall be denoted by σ_ϕ^2, σ_ψ^2, σ_ϵ^2, and σ_η^2, respectively. σ_ϵ^2 and σ_η^2 are assumed to be known. In general $\phi^{-1}(x)$ and $\psi^{-1}(y)$ will be multiple-valued functions, the only restriction that will be imposed is that for a fixed value $x = x_0$, $y = y_0$; $\phi^{-1}(x_0)$, and $\psi^{-1}(y_0)$ intersect one another only at a single point on the (u, v)-plane. In a very special case, when ϕ is taken to be a function of only u and ψ a function of only v, then an important invariant property is preserved, namely, a system of rectilinear

stratification in the (ϕ, ψ)-plane will remain a rectilinear stratification in the (u, v)-plane.

Now we can go through the same type of algebra as in Section 3 and the cut-off points will be given by the same type of equations as (3.1) and (3.2) with the following changes: $\phi(u_i, v_j)$ will be obtained from (3.1) and $\psi(u_i, v_j)$ from (3.2) replacing x by ϕ, y by ψ, σ_x^2 by $\sigma_\phi^2 + \sigma_\epsilon^2$, and σ_y^2 by $\sigma_\psi^2 + \sigma_\eta^2$. Here also the method of iteration has to be applied to find the values of $\phi(u_i, v_j)$ and $\psi(u_i, v_j)$. In order to find the exact cut-off points we have to solve the simultaneous equations,

$$\phi(u_i, v_j) = a_0, \qquad \psi(u_i, v_j) = b_0 \tag{4.1}$$

where a_0 and b_0 are the right-hand side of (3.1) and (3.2), respectively, with the substitutions mentioned above and $\phi = \phi^{(m)}$ and $\psi = \psi^{(m)}$ where $\phi^{(m)}$ and $\psi^{(m)}$ are the mth iterated value of ϕ and ψ, respectively, and m is the number of steps at which the iteration process stops.

By our assumptions, these equations (4.1) have only one pair of solutions, which gives us our cut-off points.

Special case (I). *When ϕ and ψ are linear functions. Suppose* $\phi(u, v) = \alpha_1 u + \alpha_2 v$, $\psi(u, v) = \beta_1 u + \beta_2 v$. Hence in order that the condition of unique solution be satisfied, we should have $\alpha_1\beta_2 - \alpha_2\beta_1 \neq 0$. Thus the solutions will be $u_i = (a_0\beta_2 - b_0\alpha_2)/(\alpha_1\beta_2 - \alpha_2\beta_1)$ and $v_j = (\alpha_1 b_0 - \beta_1 a_0)/(\alpha_1\beta_2 - \alpha_2\beta_1)$.

Special case (II). *When $\phi = \phi(u)$, i.e., a function of u only and $\psi = \psi(v)$, i.e., a function of v only.* In this case the condition for uniqueness of the solution reduces to the condition that ϕ and ψ are one-to-one functions of u and v, respectively. Hence the solutions are $u_i = \phi^{-1}(a_0)$ and $v_j = \psi^{-1}(b_0)$.

5. Appendix. We shall present here the details of finding the optimum stratification points. We have to minimize G with respect to the points x_i's ($i = 1, 2, \cdots, k-1$) and y_j's($j = 1, 2, \cdots, l-1$). Now

$$\frac{\partial G}{\partial x_i} = \begin{vmatrix} \sigma_x^{2\prime} & \sigma_{xy} \\ \sigma'_{xy} & \sigma_y^2 \end{vmatrix} + \begin{vmatrix} \sigma_x^2 & \sigma'_{xy} \\ \sigma_{xy} & \sigma_y^{2\prime} \end{vmatrix} \tag{5.0}$$

where the primes refer to derivatives with respect to x_i. Let us consider

$$(\partial/\partial x_i)\Big(\sum_i \sum_j p_{ij}\sigma^2_{xij}\Big) = (\partial/\partial x_i)\Big(\sum_j p_{ij}\sigma^2_{xij} + \sum_j p_{i+1\ j}\sigma^2_{x\ i+1\ j}\Big). \tag{5.1}$$

The other terms are independent of x_i and hence vanish. Now

$$\frac{\partial p_{ij}}{\partial x_i} = \int_{y_{j-1}}^{y_j} f(x_i, y)\, dy = f_{x_i j},$$

$$\frac{\partial p_{i+1\ j}}{\partial x_i} = -\int_{y_{j-1}}^{y_j} f(x_i, y)\, dy = -f_{x_i j},$$

and

$$\sigma_{xij}^2 = \frac{1}{p_{ij}} \int_{x_i-1}^{x_i} \int_{y_j-1}^{y_j} x^2 f(x, y)\, dx\, dy - \mu_{xij}^2,$$

$$\sigma_{x\,i+1\,j}^2 = \frac{1}{p_{i+1\,j}} \int_{x_i}^{x_{i+1}} \int_{y_j-1}^{y_j} x^2 f(x, y)\, dx\, dy - \mu_{x\,i+1\,j}^2.$$

Differentiating with respect to x_i and simplifying, we get,

$$p_{ij}(\partial\sigma_{xij}^2/\partial x_i) = f_{x_i j}\{(x_i - \mu_{xij})^2 - \sigma_{xij}^2\}$$

$$p_{i+1\,j}(\partial\sigma_{x\,i+1\,j}^2/\partial x_i) = -f_{x_i j}\{(x_i - \mu_{x\,i+1\,j})^2 - \sigma_{x\,i+1\,j}^2\}.$$

Therefore (5.1) becomes

$$\begin{aligned}(\partial/\partial x_i)\ (\sum_i \sum_j p_{ij}\sigma_{xij}^2) &= \sum_j \sigma_{xij}^2(\partial p_{ij}/\partial x_i) + \sum_j p_{ij}(\partial\sigma_{xij}^2/\partial x_i) \\ &\quad + \sum_j \sigma_{x\,i+1\,j}^2(\partial p_{i+1\,j}/\partial x_i) + \sum_j p_{i+1\,j}(\partial\sigma_{x\,i+1\,j}^2/\partial x_i) \\ &= \sum_j f_{x_i j}(x_i - \mu_{xij})^2 - \sum_j f_{x_i j}(x_i - \mu_{x\,i+1\,j})^2 \\ &= 2\sum_j f_{x_i j}(\mu_{x\,i+1\,j} - \mu_{xij})\{x_i - \tfrac{1}{2}(\mu_{x\,i+1\,j} + \mu_{xij})\}.\end{aligned}$$

Hence

$$(5.2)\quad \sigma_x^{2\prime} = (2x_i/n) \sum_j f_{x_i j}(\mu_{x\,i+1\,j} - \mu_{xij}) + (1/n) \sum_j f_{x_i j}(\mu_{xij}^2 - \mu_{x\,i+1\,j}^2).$$

Using exactly the same technique we get

$$p_{ij} \frac{\partial\sigma_{yij}^2}{\partial x_i} = \int_{y_j-1}^{y_j} \{(y - \mu_{yij})^2 - \sigma_{yij}^2\} f(x_i, y)\, dy,$$

$$p_{i+1\,j} \frac{\partial\sigma_{y\,i+1\,j}^2}{\partial x_i} = -\int_{y_j-1}^{y_j} \{(y - \mu_{y\,i+1\,j})^2 - \sigma_{y\,i+1\,j}^2\} f(x_i, y)\, dy.$$

Hence

$$\begin{aligned}\frac{\partial}{\partial x_i}(\sum_i \sum_j p_{ij}\,\sigma_{yij}^2) &= \sum_j \int_{y_j-1}^{y_j} \{(y - \mu_{yij})^2 - (y - \mu_{y\,i+1\,j})^2\} f(x_i, y)\, dy \\ &= \sum_j (\mu_{yij} - \mu_{y\,i+1\,j}) \int_{y_j-1}^{y_j} (\mu_{yij} + \mu_{y\,i+1\,j} - 2y) f(x_i, y)\, dy.\end{aligned}$$

Using the definitions of $f_{x_i j}$ and $\mu_{yx_i j}$ finally we get

$$(5.3)\quad \sigma_y^{2\prime} = (2/n) \sum_j f_{x_i j}(\mu_{y\,i+1\,j} - \mu_{yij})\{\mu_{yx_i j} - \tfrac{1}{2}(\mu_{yij} + \mu_{y\,i+1\,j})\}.$$

Proceeding exactly similarly and on simplifying we get

$$\begin{aligned}\sigma_{xy}' = (1/n) \sum_j f_{x_i j}\{&(\mu_{yx_i j} - \mu_{y_i j})(x_i - \mu_{xij}) \\ &- (\mu_{yx_i j} - \mu_{y\,i+1\,j})(x_i - \mu_{x\,i+1\,j})\}.\end{aligned} \tag{5.4}$$

Substituting (5.2), (5.3), and (5.4) in (5.0) and making use of the well-known property of a determinant

$$\begin{vmatrix} a_1 + b_1 & c_1 \\ a_2 + b_2 & c_2 \end{vmatrix} = \begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix} + \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}$$

we get

$$\begin{aligned} \frac{\partial G}{\partial x_i} &= \frac{1}{n} \begin{vmatrix} 2x_i \sum_j f_{x_i j}(\mu_{x\, i+1\, j} - \mu_{x i j}) & \sigma_{xy} \\ 0 & \sigma_y^2 \end{vmatrix} \\ &+ \frac{1}{n} \begin{vmatrix} \sum_j f_{x_i j}(\mu_{x i j}^2 - \mu_{x\, i+1\, j}^2) & \sigma_{xy} \\ \sum_j f_{x_i j}\{(\mu_{y x_i j} - \mu_{y i j})(x_i - \mu_{x i j}) - (\mu_{y x_i j} - \mu_{y\, i+1\, j})(x_i - \mu_{x\, i+1\, j})\} & \sigma_y^2 \end{vmatrix} \\ &+ \frac{1}{n} \begin{vmatrix} \sigma_x^2 & \sum_j f_{x_i j}\{(\mu_{y x_i j} - \mu_{y i j})(x_i - \mu_{x i j}) - (\mu_{y x_i j} - \mu_{y\, i+1\, j})(x_i - \mu_{x\, i+1\, j})\} \\ \sigma_{xy} & 2\sum_j f_{x_i j}(\mu_{y\, i+1\, j} - \mu_{y i j})\{\mu_{y x_i j} - \tfrac{1}{2}(\mu_{y i j} + \mu_{y\, i+1\, j})\} \end{vmatrix} \\ &= (2x_i/n) \sum_j f_{x_i j}(\mu_{x\, i+1\, j} - \mu_{x i j})\sigma_y^2 + (D_1/n) + (D_2/n) \quad \text{(say)}, \end{aligned} \tag{5.5}$$

where D_1 and D_2 are the second and third determinants without the factor $1/n$.

Equating (5.5) to zero and dividing both sides by the coefficient of x_i we finally get

$$x_i = D_3(x_i, y) + D_4(x_i, y) \tag{5.6}$$

where $D_3(x_i, y)$ and $D_4(x_i, y)$ have been defined in Section 3.

As G is symmetric in x and y, hence $\partial G/\partial y_j$ can be written directly and on simplifying we get (3.2)

6. Acknowledgement. The author wishes to thank Professor J. L. Hodges, Jr. for his constant guidance and help throughout the entire work. The author would like to thank Professor Dalenius for suggesting the problem and making some valuable comments.

REFERENCES

[1] DALENIUS, T. (1950). The problem of optimum stratification. *Skand. Aktuarietidskr.* **33** 203–213.

[2] DALENIUS, T. (1962). Recent advances in sample survey theory and methods. *Ann. Math. Statist.* **33** 325–349.