PSYCHOMETRIKA—VOL. 17, NO. 3
SEPTEMBER, 1952

NOTES ON A PROBLEM OF MULTIPLE CLASSIFICATION*

FREDERIC M. LORD

EDUCATIONAL TESTING SERVICE

A solution is developed in implicit form for the problem of assigning N men to n jobs, the proportion of men to be assigned to each job being specified in advance.

Suppose that it is desired to assign N men to n jobs, the proportion of men to be assigned to each job being specified in advance. It is desired to maximize the average weighted productivity of the men, the productivity of each man being weighted according to the importance of the job to which he is assigned. It is assumed that the productivity of each man for each job is known in advance and can be used as a basis for assignment. If x_{ia} is the productivity of man a for job i, we can indicate the productivity of all men assigned to job i by $\sum^* x_{ia}$. The quantity to be maximized is then

$$Q = w_1\sum{}^*x_{1a} + w_2\sum{}^*x_{2a} + \cdots + w_n\sum{}^*x_{na}\,, \tag{1}$$

where w_i is the weight assigned to job i.

A solution to an almost identical problem, assuming all weights to be unity, has been given by Hubert E. Brogden (1); these and related problems have more recently been treated by Thorndike (2); Votaw (3) has quite recently made important contributions toward a rapidly converging successive approximation method for the practical solution of such problems.† The present paper is primarily concerned with developing in analytic form an implicit solution that is effectively the same as Brogden's; it is not immediately concerned with the problem of obtaining a practical solution by successive approximations.

The Two-Dimensional Case

Consider first the case when $n = 2$. The contribution of any individual to Q will be w_1x_{1a} if he is assigned to job 1, w_2x_{2a} if he is assigned to job 2. The differential effect of the job assignment on Q is in this case $w_1x_{1a} - w_2x_{2a}$. All individuals for whom the quantity $w_1x_{1a} - w_2x_{2a}$ has a given value are interchangeable with each other in their effect upon Q, and consequently are

*The author wishes to thank Dr. Hubert Brogden and Dr. Paul Horst for their helpful discussion and criticism.

†Dr. Paul S. Dwyer has recently developed important practical methods of solution, as yet unpublished.

interchangeable with each other for purposes of job assignment. Furthermore, if certain individuals characterized by some specified value of $w_1x_{1a} - w_2x_{2a}$ are properly assigned to job 1, then all individuals with higher values must also be assigned to job 1. The same relation obtains between $w_2x_{2a} - w_1x_{1a}$ and assignment to job 2.

Suppose a scatter diagram is plotted with x_1 and x_2 as axes, each individual being represented by a point corresponding to his values of x_1 and x_2. The reasoning just given shows that the optimum assignment of people to jobs corresponds to the division of the scatter diagram into two regions by the line $w_1x_1 = w_2x_2 + k$, where k is a constant to be determined so that the required proportions of individuals fall in the two regions created.

We thus have the result for $n = 2$ that the optimum assignment corresponds to a region bounded by a straight line with a slope of w_2/w_1. This conclusion holds irrespective of the shape of the bivariate frequency distribution represented by the scatter diagram. The intercept of the line must be determined so that the proportions of cases cut off by the line are equal to the predetermined proportions of people to be assigned to the two jobs.

The Three-Dimensional Case

Let us next consider the case where $n = 3$. Using x_1, x_2, and x_3 as axes, a three-dimensional scatter plot may be prepared representing the data. The region of space containing the individuals to be assigned to job 1 will necessarily include all positive values on the x_1 axis above a certain point. Consider the surface bounding this region (region 1) from the region containing the individuals to be assigned to job 2 (region 2).

Suppose that the desired regions have already been set up. There will be of necessity certain individuals who could equally well be assigned to jobs 1 or 2, but who should not be assigned to job 3. Such individuals lie on the boundary surface between jobs 1 and 2. For such people, the x_3 score can have no effect on the decision as to whether they should be assigned to job 1 or to job 2. Consequently, the boundary between region 1 and region 2 is parallel to the x_3 axis. By the same reasoning used for the case where $n = 2$, it follows that this boundary must be represented by the equation $w_1x_1 = w_2x_2 + k_{12}$ where k_{12} is a constant to be determined so as to obtain the proper proportion of people in each region. Similarly the boundaries between regions 1 and 3 and between regions 2 and 3 are respectively $w_1x_1 = w_3x_3 + k_{13}$ and $w_2x_2 = w_3x_3 + k_{23}$. In the present case these three equations define 3 planes, each of which is functionally independent of one of the variables and consequently parallel to the corresponding coordinate axis.

Let us see if any further conditions should be imposed on these planes. Let p_i be the proportion of cases to be assigned to job i. Since $\sum^n p_i = 1$, only $n - 1 = 2$ of the values of p_i can be determined arbitrarily. Since the values of the k's must be adjusted so that each region contains the proper

proportion of cases, not more than one restriction can be imposed on the three k's. It will now be shown that one restriction must be imposed.

Consider the augmented matrix representing the three planes:

$$\left\| \begin{matrix} w_1 & -w_2 & 0 & k_{12} \\ w_1 & 0 & -w_3 & k_{13} \\ 0 & w_2 & -w_3 & k_{23} \end{matrix} \right\| .$$

If this matrix has a rank of 3, the three planes will intersect in only one point, and hence will define $2^n = 8$ regions. Since we require only three regions, we would have to select some combination of these eight regions that would form the three regions required. Furthermore, symmetry requires that each of the three regions be defined with respect to the three coordinate axes in some way that is symmetrical, as far as the axes are concerned, with the definitions for the other two regions—e.g., if one boundary is a half-plane, all boundaries must be half-planes; if two half-planes intersect each other in a line, all half-planes must intersect all others in a line, etc. Since it is manifestly impossible to form three regions from the eight available so as to meet this condition, it follows that the rank of the matrix must be less than 3.

This same conclusion may be reached via an alternative argument. There must be, at least theoretically, certain individuals who could be assigned to any one of the three jobs without changing the value of Q—such individuals will lie at the mutual intersection of the three boundary planes. Moreover, such individuals may vary in at least one dimension of our coordinate space, since their contribution to Q (irrespective of the job to which they are assigned) may vary in one dimension, i.e., may be large, small, or intermediate in value. These cases therefore cannot all be located at a single point in our coordinate space, but must be distributed at least over a line. Since all these cases lie at a place where all three regions are in simultaneous contact, it follows that the planes bounding the regions must all have at least one line in common. This requires that the rank of the matrix given must be less than 3.

Since the w's are predetermined, and since one and only one condition can be imposed on the k's, it follows that the rank of the matrix is exactly 2, and hence that the planes have only a straight line in common.

In summary, for the case when $n = 3$, the boundaries of the three regions are the three planes represented by the matrix given, with the additional condition that the k's must be so chosen as to give the matrix a rank of 2. Each plane is parallel to one coordinate axis and intersects the other two planes in a straight line. The slopes of the planes are determined by the pre-assigned weights and are independent of the distribution of cases in the scatter plot. The intercepts of the planes must be determined so that the planes cut off the desired proportion of the cases.

Since three planes intersecting in the same straight line define six regions,

we must form our three regions from these six. If we denote the plane that separates region i from region j by b_{ij}, we find, for example, that there are at least two separate regions lying between b_{12} and b_{13}. We must choose as region 1 the one of these that includes the end of the x_1 coordinate axis representing large positive values of x_1. This is equivalent to saying that the boundaries of the three regions are three half-planes, each terminating in the same straight line that is its common intersection with the other two half-planes.

The n-Dimensional Case

These conclusions may readily be extended to any number of dimensions. In n dimensions, each of the n regions will be in contact with each of the other regions. The boundary between two such regions will be the $(n-1)$-space represented by the equation $w_i x_i = w_j x_j + k_{ij}$. This boundary is parallel to all coordinate axes except x_i and x_j. There will be $\frac{1}{2}n(n-1)$ such boundaries, corresponding to the $\frac{1}{2}n(n-1)$ possible pairs of regions. All the boundaries will have as their mutual intersection a single straight line. The rank of the augmented matrix representing the boundaries will be $n-1$.

Since all $(n-1)$-spaces acting as boundaries pass through the same straight line, it is possible to make a translation such that they will all pass through the origin. The equations for the boundaries can thus be written $w_i(x_i - b_i) = w_j(x_j - b_j)$ or $w_i x_i = w_j x_j + a_i - a_j$, where the a's (or b's) are constants replacing the k's and determining the intercepts of the $(n-1)$-spaces. In all the preceding formulations k_{ij} may simply be replaced by $a_i - a_j$. With this formulation it is seen that we need determine only the n unknown constants $a_1, a_2, \cdots, a_n$ instead of the $\binom{n}{2}$ values of $k_{11}, k_{12}, \cdots k_{22}, k_{23}, \cdots k_{n-1,n}$.

For $n = 4$, for example, the boundaries are represented by

$$\left\| \begin{array}{cccc} w_1 & -w_2 & 0 & 0 & a_1 - a_2 \\ w_1 & 0 & -w_3 & 0 & a_1 - a_3 \\ w_1 & 0 & 0 & -w_4 & a_1 - a_4 \\ 0 & w_2 & -w_3 & 0 & a_2 - a_3 \\ 0 & w_2 & 0 & -w_4 & a_2 - a_4 \\ 0 & 0 & w_3 & -w_4 & a_3 - a_4 \end{array} \right\|,$$

where the a's are restricted by the fact that the rank of the matrix must be 3.

Formulation in Terms of Definite Integrals

If the form of the multivariate scatter plot can be specified analytically by the continuous frequency function $f(x_1, x_2, \cdots, x_n)$, the proportion of cases in region i can be expressed as a multiple integral, as follows:

$$p_i = \int_{-\infty}^{\infty} \int_{-\infty}^{(w_i x_i - a_i + a_n)/w_n} \int_{-\infty}^{(w_i x_i - a_i + a_{n-1})/w_{n-1}} \cdots \int_{-\infty}^{(w_i x_i - a_i + a_{i+1})/w_{i+1}} \int_{-\infty}^{(w_i x_i - a_i + a_{i-1})/w_{i-1}} \cdots \int_{-\infty}^{(w_i x_i - a_i + a_1)/w_1} f(x_1 , x_2 , \cdots x_n)\, dx_1\, dx_2 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_n\, dx_i . \tag{2}$$

If $n = 4$, the simultaneous equations to be solved for the a's are

$$p_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{(w_1 x_1 - a_1 + a_4)/w_4} \int_{-\infty}^{(w_1 x_1 - a_1 + a_3)/w_3} \int_{-\infty}^{(w_1 x_1 - a_1 + a_2)/w_2} \cdot f(x_1 , x_2 , x_3 , x_4)\, dx_2\, dx_3\, dx_4\, dx_1 ,$$

$$p_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{(w_2 x_2 - a_2 + a_1)/w_1} \int_{-\infty}^{(w_2 x_2 - a_2 + a_4)/w_4} \int_{-\infty}^{(w_2 x_2 - a_2 + a_3)/w_3} \cdot f(x_1 , x_2 , x_3 , x_4)\, dx_3\, dx_4\, dx_1\, dx_2 ,$$

$$p_3 = \int_{-\infty}^{\infty} \int_{-\infty}^{(w_3 x_3 - a_3 + a_2)/w_2} \int_{-\infty}^{(w_3 x_3 - a_3 + a_1)/w_1} \int_{-\infty}^{(w_3 x_3 - a_3 + a_4)/w_4} \cdot f(x_1 , x_2 , x_3 , x_4)\, dx_4\, dx_1\, dx_2\, dx_3 , \tag{3}$$

$$p_4 = \int_{-\infty}^{\infty} \int_{-\infty}^{(w_4 x_4 - a_4 + a_3)/w_3} \int_{-\infty}^{(w_4 x_4 - a_4 + a_2)/w_2} \int_{-\infty}^{(w_4 x_4 - a_4 + a_1)/w_1} \cdot f(x_1 , x_2 , x_3 , x_4)\, dx_1\, dx_2\, dx_3\, dx_4 .$$

The reasoning by which these integrals are written down may be illustrated by the first multiple integral given for the case when $n = 4$. We wish to find the frequency in region 1, which is known to be bounded by the three hypersurfaces $w_2x_2 = w_1x_1 - a_1 + a_2$, $w_3x_3 = w_1x_1 - a_1 + a_3$, and $w_4x_4 = w_1x_1 - a_1 + a_4$. Suppose we wish to integrate first with respect to x_2 . The last two of the hypersurfaces listed are parallel to the x_2 axis, so they cannot serve as limits of integration and may be left out of consideration. Since the hypersurfaces can provide only one limit of integration, the other limit must be either $+\infty$ or $-\infty$, but all points for which $x_2 = +\infty$ are included in region 2. Hence the limits of integration with respect to x_2 are $-\infty$ and $(w_1x_1 - a_1 + a_2)/w_2$. Similar reasoning will give us the limits of integration with respect to x_3 and x_4 . When these three integrations have been accomplished, there remains only one variable in our integrand—all the frequencies have been summed and expressed as a function of a single variable, x_1 . Since, in the general case, there will be some finite frequency corresponding to every possible value of x_1 , we must integrate from $-\infty$ to $+\infty$ with respect to this variable. The other multiple integrals given may be written down by similar lines of reasoning.

Since when $n = 4$ the a's must be so determined that the matrix of the boundary surfaces has a rank of 3, and since $\sum p_i = 1$, any three of the integral equations given is sufficient to determine the a's. In practice, the solution of these equations will be extremely difficult in many cases. If $f(x_1, x_2, \cdots, x_n)$ is taken as the normal multivariate function, the necessary expressions at present available for the multiple integrals are too cumbersome to be of practical use. Unless simpler expressions can be found, the iterative methods developed by Brogden and Votaw will probably be found to be the most satisfactory method for handling actual numerical problems.

The Case where Some Individuals Are to be Rejected

We may consider briefly the additional case where p_0 of the individuals

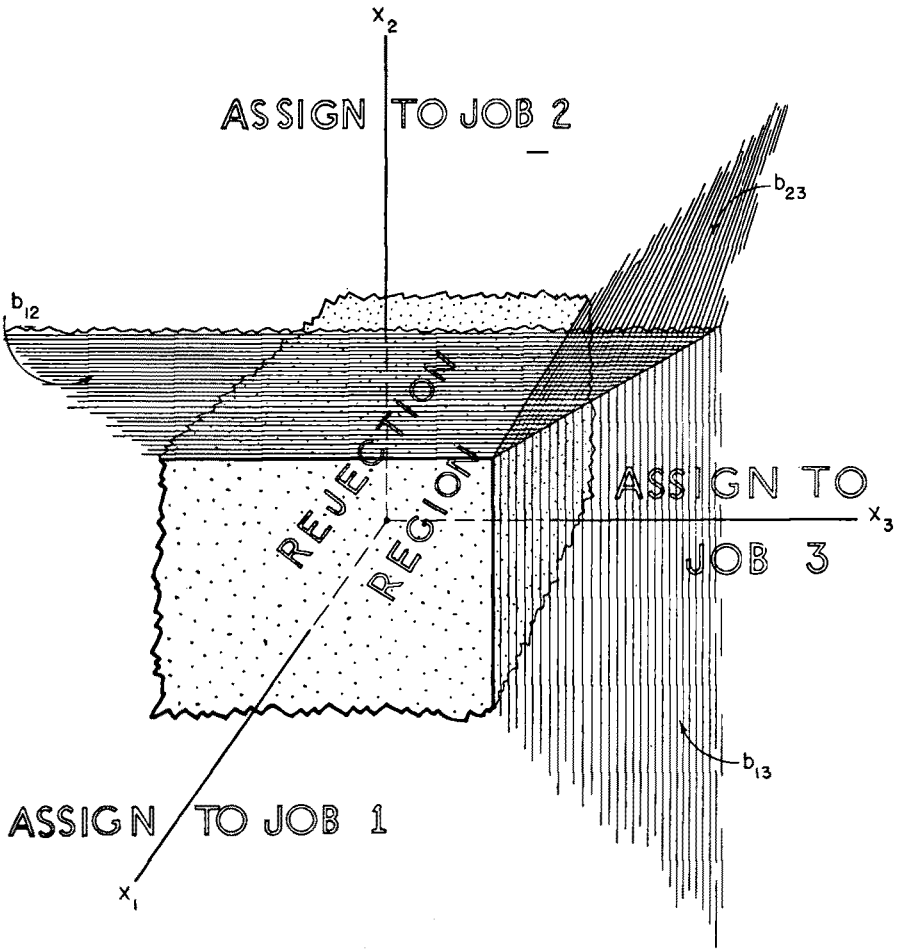

FIGURE 1
Illustrative Assignment Regions and Rejection Region for a Three-job Problem

are not to be assigned to any job, but are to be rejected entirely. Now if individuals in region i are to be rejected, they must be rejected solely on the basis of x_i, irrespective of their score on other variables. Otherwise some people who were included in region i would have lower values of x_i than those of some people who were rejected. Consequently, some hyperplane of the form $x_i = c_i$, where c_i is an unknown constant, must be the boundary (b_{i0}) between region i and the rejection region.

Next consider an individual who lies exactly on the boundary between region i and region j—this individual may equally well be assigned to either job i or job j. If this individual has a low value of x_i, he may simultaneously lie exactly on the boundary between region i and the rejection region, in which case he also may equally well be assigned to job i or rejected entirely. It follows that this individual, and all similar individuals, may equally well be rejected or assigned to job j and that consequently they all lie exactly on the boundary between region j and the rejection region. We have thus proved that the intersection of b_{i0} with b_{ij} coincides with the intersection of b_{j0} with b_{ij}. Since all boundaries between jobs intersect in a single straight line, the boundaries of the rejection region must intersect each other in a point lying on this line.

The rejection region is thus a rectangular region each of whose sides is parallel to $n - 1$ of the coordinate axes. It obviously includes all large negative values of all the variables. (In practice negative values of x_i may not occur, but we need not exclude the possibility of their occurrence in our theoretical discussion.) The intersection of any two sides of this region coincides with their intersection with one of the boundaries between jobs. This may be visualized for the three-dimensional case as follows (see Figure 1): The rejection region is a rectangular solid containing all large negative values of all variables. Only the three upper faces of this rectangular solid can be visualized, since the other three faces are at $-\infty$. Each face is perpendicular to one of the coordinate axes. An oblique line having a positive slope with respect to all axes extends in a positive direction from the corner of the rectangular solid. The boundaries separating the three jobs are three oblique planes each of which joins this line with one of the edges of the rectangular solid and extends to ∞ in the remaining directions.

The analytic equations corresponding to this situation may be readily written down. For the general case the integrals will be the same as given before in equation 2, with the exception that the lower limit of the first integral sign will be c_i instead of $-\infty$. In addition to the n equations of the type given, there will be an $(n + 1)$-th equation, as follows:

$$p_0 = \int_{-\infty}^{c_1} \int_{-\infty}^{c_2} \cdots \int_{-\infty}^{c_n} f(x_1, x_2, \cdots, x_n)\, dx_n \cdots dx_2\, dx_1 . \qquad (4)$$

In addition to the restrictions previously imposed on the a's, we now must impose $n - 1$ restrictions on the c's corresponding to the fact that all

the rejection-region boundaries intersect in a single point through which all the other boundaries pass. These restrictions may be written down as follows:

$$w_i c_i - w_j c_j = a_i - a_j \qquad (i, j = 1, \cdots, n). \tag{5}$$

Only $n - 1$ of the equations in (5) are independent.

REFERENCES

1. Brogden, H. E. An approach to the problem of differential prediction. *Psychometrika*, 1946, 11, 139-154.
2. Thorndike, R. L. The problem of classification of personnel. *Psychometrika*, 1950, 15, 215-236.
3. Votaw, D. F., Jr. Methods of solving some personnel-classification problems. *Psychometrika*, 1952, 17, 255-266.

Manuscript received 10/22/51

Revised manuscript received 1/14/52