MANAGEMENT SCIENCE
Vol. 22, No. 6, February, 1976
Printed in U.S.A.

Notes[*]
I

A MATHEMATICAL PROGRAMMING FORMULATION OF ESTIMATION PROBLEMS RELATED TO CONTINGENCY TABLES[†]

EDWARD L. MELNICK[‡] AND URI YECHIALI[§]

The adjustment of cross-tabulated sampled data to fit known marginal totals is a problem that has been the subject of some articles in the statistical literature (e.g., Deming and Stephan [2], Feinberg [3], and Kullback [7]). This problem occurs, for example, in sampling situations where a complete count of certain characteristics for an individual is obtained and cross-tabulations of these characteristics based upon sampled information are required. A second type of problem occurs when forecasting multivariate models such as input-output models of an economic situation, where the individual sectors are to be adjusted after estimating the more stable marginal totals.

Solutions to the adjustment problem have been proposed which not only satisfy the marginal totals but also retain the relationships of the observed data as measured by a statistical criterion. Since closed form solutions to these problems are not obtainable, iterative convergent procedures have been developed from which approximate solutions are obtained. Presented here is a reformulation of the problem as an integer transportation problem with a convex separable cost function. This representation demonstrates the relationships between some statistical optimization problems and mathematical programming procedures. By appropriately defining the cost function, any of the proposed statistical criteria can be represented and estimates can be obtained which maximize (minimize) the criterion and satisfy the marginal constraints. Thus, instead of considering separate problems, a general model is formulated which encompasses all previously related problems and estimates are obtained from a general algorithm which satisfies *exactly* the statistical criterion in a *finite* number of iterations.

The ensuing discussion will be in terms of an $r \times c$ table, although the ideas can be applied to any multidimensional table. In this note, $\{n_{ij} \mid i = 1, \ldots, r$ and $j = 1, \ldots, c\}$ are the observed frequency counts, $\{n_{i.}\}$ and $\{n_{.j}\}$ their row and column sums, respectively, $\{N_{i.}\}$ and $\{N_{.j}\}$ are the known marginal totals and $\{N_{ij}\}$ are the unknown population frequency counts that are to be estimated. Obviously, $N_{ij} \geqslant n_{ij}$ so that $x_{ij} = N_{ij} - n_{ij} \geqslant 0$ for all i and j. Previously the problem was redefined in terms of relative frequencies. In this framework, the objective becomes the estimation of population probabilities $\{P_{ij} = N_{ij}/N,\ N = \sum_i \sum_j N_{ij}\}$ in a contingency table when the row and column marginal probabilities, $\{P_{i.} = N_{i.}/N\}$ and $\{P_{.j} = N_{.j}/N\}$, are known, and at the same time there is a sample of cell frequencies $\{n_{ij}\}$ where $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$.

In terms of a mathematical programming model, the adjusted problem is stated as

[*] All Notes are refereed.

[†] Processed by Professor Arthur M. Geoffrion, Departmental Editor for Integer and Large-Scale Programming; received April 27, 1973, revised March 14, 1974, May 9, 1975. This paper has been with the authors 13 months for revision.

[‡] New York University.

[§] Tel-Aviv University.

follows: Given a set of observed frequency counts and known marginal frequency totals, find a set of nonnegative integer-variables $\{x_{ij} = N_{ij} - n_{ij}\}$ that minimize (maximize) a separable convex (concave) objective function

$$\sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij}(x_{ij}) \tag{1}$$

and satisfy the transportation-type constraints

$$\sum_{j=1}^{c} x_{ij} = a_i, \qquad i = 1, \ldots, r,$$

$$\sum_{i=1}^{r} x_{ij} = b_j \qquad j = 1, \ldots, c, \tag{2}$$

where

$$a_i = N_{i\cdot} - n_{i\cdot}, \quad b_j = N_{\cdot j} - n_{\cdot j}. \tag{3}$$

The estimated population cell frequencies $\{N_{ij}\}$ are obtained from the computed $\{x_{ij}\}$ terms. Furthermore, the bivariate probability estimates can be computed as $P_{ij}^* = N_{ij}/N$.

Deming and Stephan [2] first considered the adjustment problem. Their estimates were obtained by a least squares method which minimized the sum of the weighted squares of the residuals. This was obtained by minimizing Neyman's modified chi-square

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - nP_{ij})^2}{n_{ij}} \tag{4}$$

subject to the restrictions imposed by the known marginal density functions (i.e., $\sum_j P_{ij} = N_{i\cdot}/N$, $i = 1, \ldots, r$; $\sum_i P_{ij} = N_{\cdot j}/N$, $j = 1, \ldots, c$). This criterion can be represented in the form of (1) by recognizing that the minimization of (4) is equivalent to minimizing

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{x_{ij}^2}{n_{ij}}\,. \tag{5}$$

Estimates which maximize the likelihood function have also been proposed for this problem. These estimates are the values which maximize

$$\frac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} P_{ij}^{n_{ij}} \tag{6}$$

subject to the marginal density constraints. Equation (6) can be written in the form of (1) by recognizing that the estimates which maximize it also maximize

$$\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log(x_{ij} + n_{ij}). \tag{7}$$

Ireland and Kullback [6] estimate the joint probability function by minimizing the discrimination information number between the derived density function and the sampled density function. This is obtained by generating estimates which satisfy the marginal constraints and minimize the discrimination information number defined

$$\sum_{i=1}^{r} \sum_{j=1}^{c} P_{ij} \log(P_{ij}/\pi_{ij}) \tag{8}$$

where $\pi_{ij} = n_{ij}/n$. The form of (1) for this problem is

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}+n_{ij})\log[(x_{ij}+n_{ij})/n_{ij}] \tag{9}$$

since it is minimized by the same estimates minimizing (8).

The solutions to these problems require integers for $\{x_{ij} \mid i = 1, \ldots, r$ and $j = 1, \ldots, c\}$. To obtain this goal proceed as follows: Let $z_{ij} = \min(a_i, b_j)$ for all i, j and linearize each convex (concave) function $f_{ij}(x_{ij})$ over the set of all integers from 0 to z_{ij}. Thus obtain a linear programming problem with $Z = \sum_i^r\sum_j^c z_{ij}$ variables (Hadley [5]). Specifically, the problem is to find nonnegative integer variables x_{ijk} so as to optimize $\{\sum_{i=1}^r\sum_{j=1}^c\sum_{k=1}^{z_{ij}}\alpha_{ijk}x_{ijk}\}$ subject to

$$\sum_{j=1}^{c}\sum_{k=1}^{z_{ij}} x_{ijk} = a_i, \qquad i = 1, \ldots, r,$$

$$\sum_{i=1}^{r}\sum_{k=1}^{z_{ij}} x_{ijk} = b_j, \qquad j = 1, \ldots, c,$$

$$0 \leqslant x_{ijk} \leqslant 1, \quad \text{all } i, j, k \quad \text{where } \alpha_{ijk} = f_{ij}(k) - f_{ij}(k-1). \tag{10}$$

One can show that an optimal solution to this problem (which is not a classical transportation problem any more) is necessarily integral. Thus, optimal integer solutions to problems (5), (7) and (9) can now be obtained using any simplex-based algorithm. A more efficient way might be to use the Graves-Thrall method [4] for capacitated transportation problems with convex separable costs. Another attractive alternative would be to use one of the transportation or network algorithms designed for convex costs. Better than all of these would probably be a slightly modified specialization of Dantzig-Wolfe decomposition [1], [5] for separable programming.

References

1. Dantzig, G. B., *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.
2. Deming, W. E. and Stephan, F. F., "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," *Annals of Mathematical Statistics*, Vol. 11 (December 1940), pp. 427–444.
3. Feinberg, S. E., "An Iterative Procedure for Estimation in Contingency Tables," *Annals of Mathematical Statistics*, Vol. 41 (June 1970), pp. 907–917.
4. Graves, G. and Thrall, R., "Capacitated Transportation Problem with Convex Polygonal Costs," Rand RM-4941-PR, April 1966.
5. Hadley, G., *Nonlinear and Dynamic Programming*, Addison-Wesley Publishing Co., Reading, Mass., 1964.
6. Ireland, C. T. and Kullback, S., "Contingency Tables with Given Marginals," *Biometrika*, Vol. 55 (March 1968), pp. 179–188.
7. Kullback S., "Loglinear Models in Contingency Table Analysis," *The American Statistician*, Vol. 28 (November 1974), pp. 115–122.