NEW ALGORITHMS FOR MULTIDIMENSIONAL SAMPLE ALLOCATION

Shulamith T. Gross and Martin R. Frankel
Baruch College of The City University of New York

1. The multipurpose versus the multidimensional stratification problem.

Some statistical problems, especially those that depend on sophisticated (and current) mathematical techniques for their solution, bear reconsideration years after they are first introduced and solved. Whereas the original solution(s) may have been cumbersome or in other ways inaccessible, newer methods may offer solutions that are simpler and that may possibly make the methods more widely available. Such is the case with the problem of optimal multipurpose stratified survey design.

As is well known among survey research practitioners, the problem of multipurpose stratification goes back to Yates (1960). Cochran (1977) gave a full account of the early solutions proposed for this problem. One of the ways the problem was posed by Yates was the following: A survey is planned in which the practitioner is interested in estimating the mean of several characteristics $X_1,X_2,\ldots,X_J$. The population is divided into H strata, and the cost of sampling is c_h in stratum h, for $h=1,\ldots,H$. The objective is to find the allocation $n=(n_1,n_2,\ldots,n_H)$ that yields minimum total sample costs $\Sigma_h n_h c_h$ subject to restrictions on the variances of the J characteristics. Given some positive constant bounds $V_1,\ldots,V_J$, the variances of the J sample means are bounded:

$$VAR(\bar{x}_j)=\Sigma_h W_h^2 S_{hj}^2 (1/n_h - 1/N_h) <= V_j \quad \text{for } j=1,\ldots,J. \qquad (1)$$

In (1) S_{hj}^2 denotes the population variance of the j'th characteristic in the h'th stratum, and W_h denotes the stratum weight N_h/N where N_h and N denote the stratum size and the population size respectively.

As was pointed out by Huddleston, Claypool and Hocking (1970), this formulation covers the problem of multidimensional sample allocation as well as the problem of multipurpose design. In the former problem, which is the topic of our present investigation, the objective is to estimate the mean of a single characteristic X in the whole population, as well as in various strata defined by one of the categorical variables $C_1,\ldots,C_k$ at a time. The variance constraints are placed on marginal strata characterized by a single class of one of the categorical variables $C_1,\ldots,C_k$. A variance constraint is also placed on the variance of the mean estimate in the whole population, and the optimal stratification sought is the one that minimizes the total sampling costs subject to the marginal strata variance constraints.

In the multidimensional case each stratum can be identified by the tuplet $i=(i_1,\ldots,i_k)$ identifying the class of each one of the classificatory variables $C_1,\ldots,C_k$ in that stratum. If we assume that the classificatory variables $C_1,\ldots,C_k$ have $r_1,\ldots,r_k$ categories respectively, and if $a_1,\ldots,a_k$ constraints are placed on each of the k dimensions respectively, with bounds $V_1,\ldots,V_k$, then the restrictions may written as

$$VAR(\bar{x}_{im}) = \Sigma_{i1,\ldots,i(m-1),i(m+1),\ldots,ik} W_{i1,\ldots,ik}^2 S_{i1,\ldots,ik}^2 \cdot (1/n_{i1,\ldots,ik} - 1/N_{i1,\ldots,ik}) <= V_{im} \quad \text{for } 1<=i_m<=a_m \text{ and } m=1,\ldots k. \qquad (2)$$

The "embedding" of the multidimensional allocation problem in the multipurpose allocation problem is done by identifying each stratum variance constraint in (2) with a new characteristic and thus a new constraint in (1). In the process of identification, the variance constraint on the marginal stratum i_m ($1<=m<=k$ and $1<=i_m<=a_m$) is associated with (population) strata variances that vanish in all strata with index different from i_m on the m'th dimension. Thus any algorithm that solves the multipurpose problem also solves the multidimensional allocation problem minimizing total survey costs. The reverse is not true. In fact, the multidimensional problem, with total survey cost as objective function, is a special case of the former problem, with a very particular structure which we exploit in our approach to the problem. In the remainder of this paper we shall treat the multidimensional allocation problem only, and discuss solutions that are particularly suited to the specific structure of the multidimensional problem.

In section 2 we describe the problem and the solutions in more detail, and in section 3 we demonstrate their use in data arising from a survey of physicians in the US. Final remarks, including a discussion of the variable cost case, are offered in section 4.

2. Algorithmic solutions for multidimensional stratification.

In the multidimensional allocation problem with a single characteristic, the objective need not necessarily be the minimization of total costs. In many practical problems, there is little variation in sampling cost among strata and the total sample size n is thus fixed and completely determined by budgetary constraints. The allocation problem that emerges in that case is then one of finding an allocation scheme for the fixed sample size n to the cross-classified strata that will satisfy the margin constraints (2) without sacrificing the precision of the mean estimate in the total population more than is absolutely necessary. Depending on the interpretation of this goal,

several different solutions to the problem emerge. In the remainder of this paper we assume that sampling costs are equal in all cell strata, and that the total sample size n is fixed in advance. We shall briefly discuss the case of unequal costs in the final section.

The most obvious interpretation of the problem is the determination of the allocation scheme $n=(n_{i1,\ldots,ik}: 1<=i_m<=a_m$ for $1<=m<=k)$ minimizes the variance of the mean estimate in the total population

$$VAR(\bar{x})=\Sigma_{i1,\ldots,ik} W_{i1,\ldots,ik}{}^2 S_{i1,\ldots,ik}{}^2 (1/n_{i1,\ldots,ik} - 1/N_{i1,\ldots,ik})$$

among all allocations that satisfy the constraints in (2) in addition to the "fixed sample size constraint"

$$\Sigma_{i1,\ldots,ik} n_{i1,\ldots ik}=n. \qquad (3)$$

Alternatively we could interpret the goal as minimizing the "deviation" between the desired allocation n and the optimal Neyman allocation $\underline{n}^N$(with constant costs) among all feasible allocations that satisfy (2) and (3). The "deviation" can be interpreted as the Kullback-Lieber distance between the two allocations for example. In the sequel we refer to the resulting algorithm as the minimum information algorithm.

Regardless of the specification of the objective function, the problem of allocating a fixed sample size n subject to constraints formulated by the practitioner need not have a feasible solution. It is one of the salutary effects of our choice of objective functions that are convex in $\underline{n}$, and constraints that are convex in $\underline{n}$, that if a feasible solution exists, an optimal solution exists as well, and then it will be found by our algorithms. If a solution is not found, the output that is provided by the algorithm will aid in determining by how much the total sample size needs to be increased in order for a feasible solution to exist. This will be accomplished by isolating the margins (usually one margin, when the dimensions of the table are not too large) in which the constraints are not satisfied, and then increasing the sample size accordingly.

Before we proceed to describe the results obtained using the minimum information and minimum variance algorithms, we shall describe an additional algorithm, that is not optimal in any known sense, but which enjoys three very useful properties: it is fast, elementary, and very easy to implement. In large examples we ran, it yielded allocations that were very similar to those determined by the "optimal" algorithms, and failed to find a solution when a feasible one existed only for severe constraints. It cannot replace the other algorithms because it is not guaranteed to converge when a solution exists, but it can be regarded as a quick, simple and accessible alternative, that will at times require a larger sample size than is optimally necessary.

The algorithm is a modified version of the iterative proportional fitting (IPF) algorithm. It is in fact the discovery that this simple algorithm provides reasonable and sensible solutions to our allocation problem that led to the present investigation. The method consists of simple cycling through the k dimensions and checking each stratum for compliance with the variance bound requirement. If the constraint is satisfied in stratum $i_m=b$ for instance, the algorithm proceeds to the next stratum; otherwise it multiplies the sample sizes $n_{i1,..,im,..,ik}$ by a positive constant exceeding 1 if $i_m=b$, and by a positive constant smaller than 1 if i_m does not equal b, while preserving the total sample size of n. The algorithm cycles through the margins in order, until the solution stabilizes, or the constraints are all satisfied.

One problem that can be raised regarding the present approach to the allocation problem, regardless of which algorithm is selected, is that whereas the variances of the estimates in the various marginal strata of interest are constrained, the variance of the estimate in the total population is only minimized and thus left to the vagaries of the game of optimization. In practice this is not a serious problem. As was already pointed out by Cochran (op cit, page 116), the optimum Neyman variance can be described as flat, i.e., fairly substantial deviations from the optimal allocation will result in only slight changes in the variance. This "flatness" provides sufficient "space" for finding an allocation that will also satisfy the margin constraints, provided of course the latter are not too severe.

As the discussion progresses, it will become evident that the methods we use to obtain the minimum variance and the minimum information solutions, namely those of convex programming, are but modern versions of the methods used by earlier research workers in this field (e.g., Hartley and Hocking (1963); Kokan and Kahn (1967); Huddleston et al (op.cit.)). Fortunately, these modern methods are much more easily implemented on small or large scale computers, thereby making the techniques much more widely available than they appear to have been thus far. We describe the mathematical details of the algorithm elsewhere (Gross and Frankel 1985). Here we shall demonstrate their use via two examples drawn from a recent survey of physicians in the US.

3. A physicians survey example.

In designing a sample of physicians for the purpose of producing an estimate of malpractice insurance costs, a stratified sample allocation that would yield sufficiently accurate cost estimates for all physicians, for various medical specialties, for different geographic regions and for physician groups defined by degree of urbanization was sought. Thus for the purpose of sample selection physicians were assigned

to strata on the basis of their specialty, their geographic region, and whether or not they lived in a Metropolitan Statistical Area (MSA).

A total of 331,174 doctors were cross-classified by region: North-east (1), North-west (2), South (3) and West (4), by degree of urbanization: non-MSA* (1) and MSA (2), and by Specialty. Seventeen medical specialties were identifed, resulting in a 4x 2x17 table of frequencies in the population. For the sake of simplicity it was assumed that the standard deviation of the cost of malpractice insurance was equal to 1 unit in all 136 strata. This assumption was not necessary for determing the "best" allocation scheme. Budgetary considerations dictated a stratified random sample of n=1000 doctors.

Table 1. Population proportions of doctors by region, degree of urbanization and specialty.

Dimension	stratum	marginal proportion	marginal variance
region	1	.2524	.0040
	2	.2228	.0045
	3	.3026	.0033
	4	.2221	.0045
urbanization	1	.1575	.0063
	2	.8425	.0012
specialty	1	.0776	.0129
	2	.0817	.0122
	3	.1214	.0082
	4	.0267	.0374
	5	.0691	.0144
	6	.0715	.0139
	7	.0663	.0150
	8	.0399	.0250
	9	.0361	.0277
	10	.0211	.0472
	11	.0694	.0144
	12	.0500	.0120
	13	.0725	.0138
	14	.0460	.0217
	15	.0271	.0368
	16	.0538	.0185
	17	.0699	.0143

In Table 1 the population proportions in the marginal strata by geographic region, degree or urbanization and specialty are shown, along with the corresponding variance of the malpractice insurance estimate based on proportional allocation. The variance of the estimate of malpractice insurance in the total population for proportional allocation (which is also the Neyman allocation in this example) is $9.968\ 10^{-4}$.

Because it was assumed that variances were equal in all cell strata, the variances shown on the right hand column in Table 1 are actually the minimum or Neyman variances (for equal sampling costs) for the mean estimates in the corresponding margins. It is apparent then that the variance for the mean estimate of malpractice cost in rural areas is over five times as large as the corresponding variance in urban areas. Thus a uniform bound (of .004) was placed on both marginal strata by urbanization. Also since the four regions were of equal importance, a uniform variance bound of .0045 was placed on all the four strata in the region dimension. Among the medical specialties, it is apparent that rare specialties tend to have larger variance for their mean malpractice insurance estimator. In the examples below, the bounds in the first and second dimension were left unchanged. In the third (specialty) dimension, the bound was successively lowered from .028, to .025 and then .022. The process was stopped at that point, since one of the algorithms (the IPF) did not converge. The remaining minimization algorithms did converge.

In the first two cases, as well as many other combinations of uniform constraints on the marginal variances, for which all three algorithms converged, we found that IPF arrived at an allocation that was hardly distinguishable from that determined by the variance minimization method: both the marginal allocations and the marginal variances obtained were very similar. We also found the allocation determined by the information-distance minimization algorithm very similar to that found by the variance minimization procedure in all the examples we tested. We therefore present only two cases, one in which all three methods converged, and one in which the IPF algorithm diverged, i.e. was unable to come up with an allocation that would satisfy all the marginal variance constraints. In both cases the sample size was 100 and the population variance in all strata was assumed to be 1. The parameters used in the two examples are summarized below.

	case 1	case 2
variance bound on regional strata:	.0045	.0045
variance bound on urbanity strata:	.0040	.0040
variance bound on specialization strata:	.0280	.0220

In Table 2 we present the summary of the allocations determined by the three algorithms. For each case the marginal proportions, the marginal variances for the given allocation and the total variance are given. Note that the marginal proportions and variances for the modified IPF algorithm in the second case are somewhat arbitrary. The algorithm in fact could not reach an allocation that satisfies the marginal variance constraints on all three dimensions simultaneously; it did return however to the same allocation after completing a full cycle through the three dimentions. The allocation shown in this case is simply the one it returned to after adjusting the margins in dimension 3 (specialty). This explains why the variance constrainst are met in that dimension but not in the first dimension (geographic region).

* Metropolitan Statistical Area

Table 2. Strata marginal allocation proportions and variances for three algorithms.

			ALGORITHM					
			IPF		VAR-MIN		INFO-MIN	
case	Dimension	stratum	proportion	variance	proportion	variance	proportion	variance
1	region	1	.2320	.0045	.2322	.0045	.2322	.0045
		2	.2376	.0045	.2346	.0045	.2342	.0045
		3	.2968	.0037	.3025	.0036	.3034	.0036
		4	.2335	.0045	.2306	.0045	.2302	.0045
	urbanization	1	.2535	.0040	.2496	.0040	.2492	.0040
		2	.7465	.0014	.7504	.0014	.7508	.0014
	specialty	1	.0821	.0132	.0825	.0131	.0826	.0131
		2	.0878	.0124	.0881	.0123	.0881	.0123
		3	.1137	.0092	.1142	.0092	.1143	.0092
		4	.0368	.0280	.0360	.0280	.0358	.0280
		5	.0641	.0163	.0644	.0162	.0645	.0162
		6	.0644	.0160	.0647	.0159	.0648	.0159
		7	.0657	.0162	.0660	.0161	.0661	.0162
		8	.0378	.0277	.0380	.0276	.0380	.0276
		9	.0374	.0280	.0371	.0280	.0370	.0280
		10	.0376	.0280	.0359	.0280	.0357	.0280
		11	.0651	.0161	.0655	.0160	.0655	.0160
		12	.0452	.0228	.0455	.0227	.0455	.0228
		13	.0653	.0158	.0656	.0157	.0656	.0157
		14	.0425	.0244	.0427	.0243	.0427	.0243
		15	.0374	.0280	.0362	.0280	.0360	.0280
		16	.0511	.0206	.0513	.0205	.0514	.0205
		17	.0660	.0159	.0663	.0158	.0664	.0158
total variance				.001090		.001063		.001063
2	region	1	.2418	.0046	.2400	.0045	.2400	.0045
		2	.2509	.0046	.2400	.0045	.2843	.0045
		3	.2588	.0045	.2822	.0043	.2843	.0043
		4	.2485	.0045	.2364	.0045	.2357	.0045
	urbanization	1	.2707	.0040	.2515	.0040	.2505	.0040
		2	.7293	.0015	.7485	.0014	.7495	.0014
	speciality	1	.0752	.0151	.0774	.0154	.0776	.0145
		2	.0809	.0142	.0827	.0137	.0829	.0137
		3	.1021	.0106	.1057	.0102	.1059	.0102
		4	.0482	.0220	.0455	.0220	.0454	.0220
		5	.0570	.0189	.0592	.0181	.0593	.0182
		6	.0569	.0187	.0592	.0179	.0592	.0178
		7	.0593	.0182	.0613	.0180	.0615	.0180
		8	.0496	.0220	.0468	.0220	.0465	.0220
		9	.0495	.0220	.0463	.0220	.0461	.0220
		10	.0498	.0220	.0454	.0220	.0453	.0220
		11	.0578	.0188	.0601	.0180	.0615	.0180
		12	.0485	.0220	.0473	.0220	.0472	.0220
		13	.0582	.0182	.0604	.0175	.0605	.0175
		14	.0489	.0220	.0471	.0220	.0469	.0220
		15	.0494	.0220	.0456	.0220	.0455	.0220
		16	.0496	.0220	.0489	.0220	.0489	.0220
		17	.0590	.0184	.0611	.0177	.0612	.0177
total variance				.001156		.001120		.001120

The results displayed in Table 2 for case 1 are typical of the behaviour we observed in other example data. When all three algorithms converge, the allocations to the marginal strata are very similar. In small allocation tables we encountered some divergence among the allocations produced by the three methods in the internal cells of the allocation table, but little differences in the marginal total allocations. In most cases it is safe to assume that the practitioner is far more concerned with latter rather than the former allocations. If such is indeed the case, then in most cases the results obtained simply and inexpensively via the modified IPF algorithm will satisfy the needs of the practitioner. In unusual circumstances, such as the one displayed in case 2, the constraints are so severe that feasible solutions barely exist. In such cases the optimal algorithms will detect a solution, whereas the modified IPF will fail to do so. A quick purusal through the results displayed in Table 2 for the IPF algorithm indicates the practical solution to the problem. Instead of resorting to one of the optimal algorithms, it is possible to simply increase the total sample size so as to ensure that the constraints are satisfied in the region dimension. Such an increased sample size, when submitted again to the algorithm will yield a satisfactory, if not optimal, solution to the allocation problem.

4. Final remarks.

In this paper we offered two optimal and one practical solution to the problem of allocating a sample of fixed total size to cross-classified strata. Although we recognized the general multidimensional stratified allocation problem as a special case of the multipurpose allocation problem we chose, on practical grounds, to reformulate the problem as an allocation problem for a fixed total sample size, and offer specialized solutions which capitalize on the special structure of the multidimensional problem. We described the nature of the proposed algorithms and discussed their behaviour in two real data examples. We recommended, again on practical grounds, the use of the simplest and most easily implementable of the three algorithms, the modified IPF algothrim. Despite the fact that it is apparently not optimal in any formal way, and despite the fact that it may not always yield a solution when one exists, it can be used in a very effective way in practical applications, as we demonstrated via our example 2.

The case of unequal sampling costs can be treated in a manner similar to the one we have just described. The optimal algorithms are modified in an obvious manner and retain most of their structure. The IPF algorithm does lose some of its simplicity, since the determination of the renormalization constants in each step requires the solution of three (quadratic) equations in three variables. The remarks made in the text about the relative behaviour of the IPF and the minimum variance and minimum information algorithms appear to continue to hold. More experience with the programs we have written to implement the algorithms in the unequal cost case is required to make a final determination regarding their relative efficiency.

REFERENCES

Cochran,W.G.(1977).Sampling Techniques. Third Edition. Wiley.New York.

Gross,S.T., and Frankel,M.R.(1985). On Multidimensional Sample Allocation and Post-Stratification Using Frequency Table Adjustment Procedures. Submitted.

Hartley,H.O., and Hocking,R.R. (1963) Convex Programming by Tangential Approximation. Management Science, 9, 600-612.

Huddleston,H.F.,Claypool,P.L.,and Hocking, R.R.(1970). Optimal Sample Allocation to Strata Using Convex Programming. Appl. Stat.19, 273-278.

Kokan,A.R., and Kahn,S. (1967) Optimum Allocation in Multivariate Surveys: An Analytical Solution.JRSS,29,1,115-125

Yates,F.(1960). Sampling Methods for Consensuses and Surveys. Charles Griffin and Co.,London. Third Edition.