AN APPLICATION OF MULTI-PURPOSE SURVEY SAMPLING

William E. Winkler, Energy Information Administration
William E. Winkler, 7705 Heritage Drive, Annandale, VA 22003

1. INTRODUCTION

Much of the sampling literature deals with methods that try to minimize the variance of one variable under fixed bounds on sample size. Most surveys, however, need to assure that variances of two or more (possibly uncorrelated) variables are somehow minimized under fixed bounds on sample size (Neyman, 1934).

1.1. Minimization Measure

This paper will present a strategy for sampling two or more variables that places bounds on sample size and attempts to minimize the variance of two or more variables. The minimization measure is total variation which is defined as the square root of the sum of squares of the coefficients of variation (cv) of the variables.

1.2. The Model

Our approach to multi-purpose sampling comprises three parts. In the first part, we allocate marginal (univariate) samples to multi-purpose cells. The allocation basically uses classical iterative proportional fitting (IPF) in which the initial matrix is the population array induced by the univariate stratifications. The allocation, thus, preserves margins and structural zeros and has an interaction pattern that is consistent with the original population array (Bishop, Fienberg, and Holland, 1977, pp. 178-182). The fitted array will generally have fractional parts.

In the second part, we find a set of nonnegative integer matrices having convex sum equal to the fitted array. Each of the nonnegative integer arrays satisfies the same marginal restraints as the fitted array. If the integer arrays are sampled with probability proportional to their coefficient in the convex sum and sampling within cells is simple random, then we have a probabilistic structure.

In the third part, we obtain estimators of the population parameters and their variances. The final section is the summary.

2. GENERALIZED ITERATIVE PROPORTIONAL FITTING

In this section, we provide a method of systematically allocating two or more univariate samples to multi-purpose population cells. The method assures that the deviations of proportions in sample cells from proportions in population cells are minimized on the average (except for rounding). The systematic fit is provided by the Kullback-Liebler measure which controls both classical iterative proportional fitting (IPF) and Dykstra's generalized iterative fitting procedure (GIFP).

Example 2.1 is used to illustrate how classical iterative fitting can yield samples fitted to multi-way cells that have some sample allocations exceeding population counts. When this occurs, Dykstra's GIFP (1985a,b, 1987) can be used because it allows convex contraints. Consequently, in addition to imposing marginal constraints that are linear, we can bound the within-cell sample allocations by the corresponding population counts. In those cases in which only linear marginal constraints need be satisfied, Dykstra's GIFP provides answers identical to those provided by classical IPF.

Example 2.1. Breakdown of Classical IPF in Sampling Problem

If we perform ordinary IPF with initial matrix N (Table 1) and fixed marginal constraints, we obtain matrix A (Table 2).

We notice that the entries in cell (1,1) exceed available population values of 2 units. If we apply Dykstra's GIFP with initial matrix N and cell (1,1) constrained to be less than or equal two, we obtain matrix B (Table 3).

3. PROBABILISTIC MECHANISM FOR CELL COUNTS

In this section, we determine a probabilistic structure and a random nonnegative integer matrix having expected value equal the fitted matrix of section 2.

3.1. Convex Sum of Nonnegative Integer Matrices

We assume that we have a set of margin counts m_j, $j = 1, 2, \ldots, s$, determined by two or more univariate sampling strategies. That is, each m_j is the sample count associated with a univariate sample. We let N_i, $i \in I$, be the population counts determined by two or more univariate stratifications. With N_i, $i \in I$, as initial matrix, let Dykstra's GIFP converge to g_i, $i \in I$. The array g_i, $i \in I$, necessarily satisfies

$$\sum_{ij \in I_j} g_{ij} = m_j,\ j = 1, \ldots, s, \tag{3.1}$$

$$\text{and } g_i \le N_i,\ i \in I,$$

where I_j is the subset of I associated with

marginal constraint m_j, $j = 1, \ldots, s$.

We wish to find a sequence of arrays M_{ik}, $i \in I$, $k = 1, \ldots, t$, having nonnegative integer entries and margins m_j, $j = 1, \ldots, s$, and positive constants p_k, $k = 1, \ldots, t$, such that

$$\sum_{k=1}^{t} p_k = 1 \text{ and } \sum_{k=1}^{t} p_k M_{ik} = g_i,\ i \in I. \tag{3.2}$$

Representation (3.2) yields a probabilistic structure that yields a random nonnegative integer matrix having expected value g_i, $i \in I$. If such arrays can be found, we merely select an array M_{il}, $i \in I$, with probability proportional to size p_l and then choose a simple random sample of size M_{il} in cell i for all $i \in I$.

The proof of (3.2) is in the appendix. The proof consists of an algorithm that on successive steps yields integer LP problems. The number of steps is bounded by the number of cells in the original array. If each of the successive integer LP problems has a solution, then the algorithm necessarily converges and yields (3.2).

3.2. Example of Convex Sum

Example 3.1. Two Dimensional Iterative Integer LP Procedure

The results from applying the algorithm to the fitted 5x5 matrix of Example 2.1 (Table 3) are presented in Table 4.

The first matrix M_0 is the fitted matrix derived using Dykstra's GIFP. The next sixteen matrices M_k, $k = 1, \ldots, 16$, are the nonnegative integer matrices obtained by the iterative integer LP procedure. The final matrix M_{17} is the convex sum (with the entries in the column headed by p_k used as the coefficients) of the integer matrices M_k, $k = 1, \ldots, 16$.

4. ESTIMATION METHODOLOGY

This section contains the formulas for an unbiased estimator of the population total and an unbiased estimator of its variance. The estimators are based on using the probabilistic structure defined by the sampling strategies of sections 2 and 3.

4.1. Parameter and Variance Estimation

We define some terms needed for the theorem. Let Y_i be the population total of any quantitative variable in cell i and let $\hat{y}_i$ be any unbiased estimator of Y_i. Let $\tilde{M}_i$ be the random nonnegative integer matrix that takes value M_{ij} with probability proportional to size p_j (see formula 3.2).

The first stage of sampling consists of selecting a matrix to determine sample allocations, and the second stage consists of simple random sampling within cells according to the matrix allocation obtained in the first stage. Let $E^{(1)}$ denote expectation with respect to the first stage of sampling and let $E^{(2)}$ denote expectation with respect to the second stage. We note that $E^{(1)}(\tilde{M}_i) = g_i$ and the second stage samples are independent. For each cell i, let $\sigma_i{}^2$ be the variance of the population mean and let $\hat{\sigma}_i{}^2$ be an estimator such that of

$$E^{(2)}(\hat{\sigma}_i{}^2) = \sigma_i{}^2.$$

Although any unbiased estimator of Y_i can be used, in the empirical example we will use

$$\hat{y}_i = (N_i/\tilde{M}_i)\ \Sigma_j\ y_{ij}\ .\ 1_{ij},$$

where y_{ij} is the quantitative value assigned to unit ij in cell i and 1_{ij} is the indicator that unit ij is sampled.

THEOREM. Let g_i, $i \varepsilon I$, be the array obtained using Dykstra's GIFP. Let p_j, $j = 1, \ldots, t$, be nonnegative constants and M_{ij}, $i \varepsilon I$, $j = 1, \ldots, t$, be nonnegative integer matrices such that

$$\sum_{j=1}^{t} p_j = 1 \text{ and } \sum_{j=1}^{t} p_j\ M_{ij} = g_i,\ i \varepsilon I.$$

Let $\tilde{M}_i$, $i \varepsilon I$, be the random matrix of sample allocations. An unbiased estimator of the population total is

$$\hat{T} = \sum_{i \varepsilon I} (\tilde{M}_i/g_i)\ .\ \hat{y}_i. \qquad (4.1)$$

An unbiased estimator of its variance is

$$\hat{v} = \sum_{i \varepsilon I} (\tilde{M}_i/g_i - 1)^2 . \hat{y}_i{}^2 +$$

$$\sum_{i \neq j\, \varepsilon I} (\tilde{M}_i/g_i - 1).(\tilde{M}_j/g_j - 1).\hat{y}_i.\hat{y}_j \qquad (4.2)$$

$$+ \sum_{i \varepsilon I} (2.\tilde{M}_i/g_i - 1).N_i.(N_i - \tilde{M}_i).\hat{\sigma}_i{}^2/\tilde{M}_i.$$

The proofs of (4.1) and (4.2) are in the appendix.

4.2. Example of Variance Estimation

Example 4.1 highlights the multi-purpose sampling techniques of this paper. For fixed size samples, it shows that multi-purpose sampling can yield lower cvs for two variables than judiciously applied univariate techniques.

Example 4.1. Two-Purpose Variance Estimation

Table 5 contains a summary of the main variance results associated with the data base. The second and third columns are cvs. The last column is total variation. The first set of four cvs is for optimal univariate designs in which the stratifying variable agrees with one of the variables being estimated. Diagonal elements are low (0.012 and 0.009).

Off-diagonal elements are dramatically higher (0.334 and 0.407) because the stratifying variables are not highly correlated with the variables for which the cvs are computed. Regression using the two variables yields an R-square value less than 0.2. If, however, we apply standard contingency table techniques (Bishop, Fienberg, and Holland, 1975) to the underlying population matrix N (Table 1), we reject independence at the 95 percent level of confidence.

The final set of numbers are the multi-purpose cvs: 0.096 and 0.039. They are also dramatically higher than the diagonal entries in the first matrix and approximately twice as high the diagonal entries in the second matrix. They are substantially lower than the highest of the off-diagonal entries for the respective variables (0.334 and 0.407).

If we were to use the stratification given by the first row of the first matrix we have total variation 0.334 while multi-purpose sampling yields total variation 0.112. Total variation is defined as square root of the sum of squares of the cv columns in a fixed row (i.e., corresponding to a fixed stratification and estimation methodology). Ignoring the finite population correction (fpc), we would have to increase the sample size by a factor greater than 6 to equal the cvs given for the multi-purpose case.

4.3. Empirical Verification

Five hundred independent samples were drawn to evaluate the empirical performance of the multi-purpose estimator of section 4.2. The empirical biases associated with the two variables were +0.008 and 0.000 and the empirical cvs were 0.103 and 0.044, respectively. If the true parameter is given by T and the independent estimates using (4.1) are given by θ_k, k = 1, 2, ... , 500, then the empirical bias is given by

$$\text{e.b.} = \bar{\theta} - T,$$

where $\bar{\theta} = (1/500) \sum_{k=1}^{500} \theta_k$, and the empirical variance is given by

$$\text{e.v.} = (1/499) \sum_{k=1}^{500} (\theta_k - \bar{\theta})^2 .$$

4.4. Three Dimensional Variance Example

The 3-dimensional example involves 4x4x2 arrays for which the iterative LP procedure converges. The data base has similar characteristics to the data base of example 4.2. Total variation ranges from 0.466 (the best in the case of standard univariate stratification techniques) to 0.115 (multi-purpose) (Table 6). Ignoring the fpc, we would have to increase the sample size by a factor greater than 16 to equal the cvs given for the multi-purpose case.

Based on 200 replications, the empirical biases of the multi-purpose estimators of the three variables are -0.004, 0.000, and -0.004, respectively. The empirical cvs were 0.070, 0.056, and 0.73, respectively.

5. SUMMARY

The results of this paper show that we can exert moderate control over the cvs of two or more variables while sample size is fixed. The multi-purpose sampling techniques give a new methodology for analyzing the relationships of two or more variables. The methods are computationally intensive in that they require:

1. Dykstra's Generalized Iterative Fitting Procedure and
2. An Iterative Integer LP Procedure.

REFERENCES

Bishop, Y., Fienberg, S., and Holland, P. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge, MA.

Dykstra, R. (1985a), "An Interative Procedure for Obtaining I-Projections onto the Intersection of Convex Sets," Ann. Prob., 13 975-984.

Dykstra, R. (1985b), "Computational Aspects of I-Projections," J. Statist. Comput. Simul., 21 265-274.

Dykstra, R. and Wollan, P. (1987), "Algorithm for Iterative Fitting," Applied Statistics, to appear.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," J. R. Statist. Soc. B., 97, 558-606.

Winkler, W. E. (1986), "Multi-Purpose Survey Sampling," 1986 ASA Proceedings of the Section on Survey Research Methods, 618-623.

Winkler, W. E. (1987a), "On Dykstra's Iterative Fitting Procedure," Energy Information Administration Technical Report.

Winkler, W. E. (1987b), "Strata Boundary Determination," Energy Information Administration Technical Report.

Winkler, W. E. (1987c), "On Multi-Purpose Survey Sampling," Energy Information Administration Technical Report.

Winkler, W. E. (1987d), "Appendix to An Application of Multi-Purpose Survey Sampling," Energy Information Administration Technical Report.

Table 1. Population Counts N Induced by Two Univariate Stratifications.

Var 1	Variable 2 Strata					
Strata	1	2	3	4	5	Total
1	2	7	4	1	11	25
2	3	5	7	17	31	63
3	0	10	16	47	85	158
4	2	3	10	78	257	350
5	3	5	29	67	551	655
	10	30	66	210	935	1251

Table 2. Fitted Sample Matrix A Obtained by Classical IPF, Marginal Totals are Fixed

Var 1	Variable 2 Strata					
Strata	1	2	3	4	5	Total
1	2.172	2.393	0.997	0.097	0.341	6.
2	2.098	1.101	1.124	1.059	0.618	6.
3	0.0	1.640	1.914	2.182	1.263	7.
4	0.820	0.387	0.941	2.848	3.004	8.
5	0.911	0.478	2.023	1.814	4.774	10.
	6.	6.	7.	8.	10.	37.

Table 3. Fitted Sample Matrix B Obtained by Dykstra's GIFP, Marginal Totals are Fixed

Var 1	Variable 2 Strata					
Strata	1	2	3	4	5	Total
1	2.000	2.483	1.052	0.103	0.362	6.
2	2.182	1.061	1.101	1.046	0.610	6.
3	0.0	1.614	1.914	2.200	1.272	7.
4	0.860	0.377	0.930	2.840	2.993	8.
5	0.958	0.466	2.003	1.811	4.763	10.
	6.	6.	7.	8.	10.	37.

Table 4. Example of Iterative Integer LP Procedure

		Cell of Matrix M_k							
k	p_k	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(2,1)	(2,2)	(2,3)
0	0.000	2.000	2.483	1.052	0.103	0.362	2.182	1.061	1.101
1	0.140	2.000	3.000	1.000	0.000	0.000	3.000	1.000	1.000
2	0.042	2.000	3.000	1.000	0.000	0.000	3.000	1.000	1.000
3	0.007	2.000	2.000	1.000	0.000	1.000	2.000	2.000	1.000
4	0.054	2.000	3.000	1.000	0.000	0.000	2.000	2.000	1.000
5	0.029	2.000	3.000	1.000	0.000	0.000	2.000	1.000	2.000
6	0.114	2.000	3.000	1.000	0.000	0.000	2.000	1.000	1.000
7	0.104	2.000	3.000	1.000	0.000	0.000	2.000	1.000	1.000
8	0.017	2.000	2.000	2.000	0.000	0.000	2.000	1.000	1.000
9	0.035	2.000	2.000	2.000	0.000	0.000	2.000	1.000	1.000
10	0.003	2.000	2.000	1.000	0.000	1.000	2.000	1.000	1.000
11	0.031	2.000	2.000	1.000	0.000	1.000	2.000	1.000	2.000
12	0.043	2.000	2.000	1.000	0.000	1.000	2.000	1.000	1.000
13	0.103	2.000	2.000	1.000	1.000	0.000	2.000	1.000	1.000
14	0.040	2.000	2.000	1.000	0.000	1.000	2.000	1.000	2.000
15	0.237	2.000	2.000	1.000	0.000	1.000	2.000	1.000	1.000
16	0.001	2.000	2.000	1.000	0.000	1.000	2.000	1.000	2.000
17	1.000	2.000	2.483	1.052	0.103	0.362	2.182	1.061	1.101

Table 4. Example of Iterative Integer LP Procedure (cont.)

Cell of Matrix M_k

k	(2,4)	(2,5)	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(4,1)	(4,2)
0	1.046	0.610	0.000	1.614	1.914	2.200	1.272	0.860	0.377
1	1.000	0.000	0.000	1.000	2.000	2.000	2.000	0.000	1.000
2	1.000	0.000	0.000	1.000	2.000	2.000	2.000	1.000	0.000
3	1.000	0.000	0.000	1.000	2.000	2.000	2.000	1.000	1.000
4	1.000	0.000	0.000	1.000	2.000	2.000	2.000	1.000	0.000
5	1.000	0.000	0.000	1.000	1.000	3.000	2.000	1.000	1.000
6	1.000	1.000	0.000	1.000	2.000	3.000	1.000	1.000	1.000
7	1.000	1.000	0.000	2.000	2.000	2.000	1.000	1.000	0.000
8	1.000	1.000	0.000	2.000	1.000	3.000	1.000	1.000	1.000
9	1.000	1.000	0.000	2.000	2.000	2.000	1.000	1.000	1.000
10	2.000	0.000	0.000	2.000	2.000	2.000	1.000	1.000	1.000
11	1.000	0.000	0.000	2.000	2.000	2.000	1.000	1.000	1.000
12	2.000	0.000	0.000	2.000	2.000	2.000	1.000	1.000	0.000
13	1.000	1.000	0.000	2.000	2.000	2.000	1.000	1.000	0.000
14	1.000	0.000	0.000	2.000	1.000	3.000	1.000	1.000	0.000
15	1.000	1.000	0.000	2.000	2.000	2.000	1.000	1.000	0.000
16	1.000	0.000	0.000	2.000	2.000	2.000	1.000	1.000	1.000
17	1.046	0.610	0.000	1.614	1.914	2.200	1.272	0.860	0.377

Cell of Matrix M_k

k	(4,3)	(4,4)	(4,5)	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)
0	0.930	2.840	2.993	0.958	0.466	2.003	1.811	4.763
1	1.000	3.000	3.000	1.000	0.000	2.000	2.000	5.000
2	1.000	3.000	3.000	0.000	1.000	2.000	2.000	5.000
3	1.000	3.000	2.000	1.000	0.000	2.000	2.000	5.000
4	1.000	3.000	3.000	1.000	0.000	2.000	2.000	5.000
5	1.000	2.000	3.000	1.000	0.000	2.000	2.000	5.000
6	1.000	2.000	3.000	1.000	0.000	2.000	2.000	5.000
7	1.000	3.000	3.000	1.000	0.000	2.000	2.000	5.000
8	1.000	2.000	3.000	1.000	0.000	2.000	2.000	5.000
9	0.000	3.000	3.000	1.000	0.000	2.000	2.000	5.000
10	0.000	3.000	3.000	1.000	0.000	3.000	1.000	5.000
11	0.000	3.000	3.000	1.000	0.000	2.000	2.000	5.000
12	1.000	3.000	3.000	1.000	1.000	2.000	1.000	5.000
13	1.000	3.000	3.000	1.000	1.000	2.000	1.000	5.000
14	1.000	3.000	3.000	1.000	1.000	2.000	1.000	5.000
15	1.000	3.000	3.000	1.000	1.000	2.000	2.000	4.000
16	0.000	3.000	3.000	1.000	1.000	2.000	2.000	5.000
17	0.930	2.840	2.993	0.958	0.466	2.003	1.811	4.763

Table 5. Comparison of Variances For 2-Way, Incomplete Case, Sample Size is 13 Certainty and 37 Noncertainty

	CVs		Total Variation
	Var 1	Var 2	1/
Optimal Univariate			
Stratifying Var 1	.012	.334	.334
Stratifying Var 2	.407	.009	.407
Multi-Purpose	.104	.041	.112

[1/] Square root of the sum of squares of two CV columns.

Table 6. Comparison of Variances For 3-Way, Incomplete Case, Sample Size is 15 Certainty and 50 Noncertainty

	CVs			Total Variation
	Var 1	Var 2	Var 3	1/
Optimal Univariate				
Stratifying Var 1	.001	.490	.372	.615
Stratifying Var 2	.306	.001	.351	.466
Stratifying Var 3	.604	.884	.001	1.071
Multi-Purpose	.071	.055	.071	.115

[1/] Square root of the sum of squares of three CV columns.