Conference on Applied Statistics in Agriculture

1993 - 5th Annual Conference Proceedings

A COMPARISON OF ALGORITHMS FOR SELECTING AN OPTIMUM SAMPLE FROM H STRATA USING k VARIABLES

M. Williams

H. T. Schreuder

Recommended Citation

Williams, M. and Schreuder, H. T. (1993). "A COMPARISON OF ALGORITHMS FOR SELECTING AN OPTIMUM SAMPLE FROM H STRATA USING k VARIABLES," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1383

A COMPARISON OF ALGORITHMS FOR SELECTING AN OPTIMUM SAMPLE FROM H STRATA USING k VARIABLES

M. Williams
H. T. Schreuder
Multiresource Inventory Techniques
Rocky Mountain Forest and Range Experiment Station
USDA Forest Service
Fort Collins, Colorado 80526-2098, U.S.A.

Abstract

In stratified sampling with k different variables and H strata it is often of interest to minimize the survey cost with respect to variance restrictions on each of the k variables. This problem has previously been solved using compromise solutions or using a linear approximation to this nonlinear problem. In this paper a nonlinear optimization routine is tested on this problem. The formulation of the problem in its original form proved problematic. For the test cases run, the transformation $t_h = 1/n_h$, where n_h is the number of samples in stratum h, performed best when k and H are less than 7. As the number of strata and variables increase, the transformation $t_h = n_h^2$ performs better. In addition, simple modifications to the routine used can improve the convergence.

1. Introduction

In survey sampling we often want a fixed sample of n units to be selected from H strata for k variables. Approximate solutions and compromise solutions exist for the allocation of the n units to the H strata. In survey sampling a linear approximation has been advocated (Hartley and Hocking 1963, Huddleston et al. 1970) and little has been published in the statistical literature since then. In this paper, we present and test a nearly exact nonlinear solution to this problem. Such solutions can be important in deciding what number of strata can be optimal for sampling to meet variance and cost constraints. This is an important topic in an annual forest inventory (AFIS) (NC FIA 1992), for example.

2. Review of Literature

Assume that the number of strata are fixed and that a set of plots is to be measured in several strata each year. As noted in Cochran (1977, p. 119), the best allocation for one parameter will in general not be the best allocation for others, so some compromise solution is needed. Chatterjee (1967) (see Cochran 1977, p. 121) suggests choosing the strata sampling sizes n_h (h=1, ..., H, H = number of strata) that minimize the average of the proportional increase in variance, i.e.,

$$n_h = n \sqrt{\sum_j n'^2_{jh}} \Big/ \sum_h \sqrt{\sum_j n'^2_{jh}} \qquad (1)$$

where n'_{jh} is the optimum sample size in stratum h for variable j (j=1, ..., k, k = number of variables).

If the optimum allocations for individual variables are very different so that no compromise is obvious then two problems may result. Cochran (1977, p. 121) discusses these two problems, which were first suggested by Yates (1960). Yates' first problem applies to surveys in which the loss due to an error of a given size in an estimate can be measured in terms of some measure of utility. If there are k variables and quadratic loss functions, then the total loss L_o, expressed as a linear function of the estimated population variances, is

$$L_o = \sum_{j=1}^{k} a_j V(\hat{Y}_j) = \sum_{j=1}^{k} a_j \sum_{h=1}^{H} N_h^2 S_{jh}^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \qquad (2)$$

where $\hat{Y}$ is the estimated population total, a_j is the linear weight assigned to variable j, N_h is the number of units in stratum h, and S^2_{jh} is the variance of the j-th variate in stratum h. The linear cost function is given by

$$C = C_o + \sum_{h=1}^{H} c_h n_h, \qquad (3)$$

where the c_h are the cost coefficients. We obtain

$$n_h \doteq [n N_h A_h / \sqrt{c_h}] \Big/ \sum_{h=1}^{H} N_h A_h / \sqrt{c_h} \qquad (4)$$

where

$$A_h = \sqrt{\sum_{j=1}^{k} a_j S_{jh}^2} \qquad (5)$$

with required total sample size

$$n = \frac{1}{L_o} [\sum_{h=1}^{H} N_h A_h / \sqrt{c_h}] \sum_{h=1}^{H} N_h A_h \sqrt{c_h} \qquad (6)$$

(Cochran 1977).

In the second problem described by Yates (1960), the desired variance for each parameter estimate is specified, giving

$$V(\hat{Y}_j) = \sum_{h=1}^{H} N_h^2 S_{jh}^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \le V_j \quad j = 1, 2, \ldots, k. \qquad (7)$$

The cost C, given in eq. (3), is minimized subject to the constraints V_j and $2 \le n_h \le N_h$. This is a problem in nonlinear optimization. Booth and Sedransk (1969) give an approximate solution to Yates' second problem using eq. (2)

and (3) by selecting the a_j inversely proportional to the V_j and solving the easier first problem specified by Yates. They note that this method will meet the constraint

$$L_0=\sum_{j=1}^{k} a_j V(\hat{Y}_j)\ , \tag{8}$$

but that the individual variance constraints will not necessarily be met.

Algorithms to solve Yates' second problem were first given by Huddleston et al. (1970). The solution for this problem is based on a linear approximation of the nonlinear components of the problem and solving the approximate problem using the simplex method. This method is referred to as a convex programming algorithm and is given by Hartley (1959) and Hartley and Hocking (1963). Allocations obtained by convex programming are often an improvement over the compromise solutions listed earlier.

In survey sampling we have not found additional published approaches to solve this basic nonlinear problem. However, considerable progress has been made on nonlinear constrained optimization over the last 10 years. For example, IMSL (1989) includes successive quadratic programming algorithms, NCONG and NCONF, based on a subroutine developed by Schittkowski (1986). This iterative routine may produce infeasible points during the solution process, but this is not a concern for the problems we wish to solve because infeasible solutions do not cause any numerical problems such as taking the square root of a negative number.

Algorithms exist for the solution of Yates' second problem, as given in eq. (7), and should only need to be modified slightly to fit our problem. The algorithm discussed by Huddleston et al. (1970) requires additional reformulations of the original problem and provides only approximate solutions. We will test the nonlinear technique proposed by Schittkowski (1986).

Huddleston et al. (1970) illustrates the performance of the convex programming technique on a data set with seven variables of interest and 15 strata. We have used this data set for our testing and comparison purposes (Table 1).

3. Methods

The initial effort entailed verifying that the nonlinear optimization technique proposed by Schittkowski performed properly on Yates' second problem, which is given by:

$$\min\ C_0+\sum_{h=1}^{H} c_h n_h$$

$$\text{such that}\quad V(\hat{Y}_j)=\sum_{h=1}^{H} N_h^2\, s_{jh}^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\le V_j \quad j=1,\,2,\,\ldots,\,k. \tag{9}$$

Some initial difficulties in getting the routine to converge to a solution led us to reformulate the problem using two transformations. The transformations, $1/t_h = n_h$, gave a nonlinear objective function and linear constraints:

$$\min\ C_0+\sum_{h=1}^{H} \frac{c_h}{t_h} \tag{10}$$

$$\text{such that}\quad \sum_{h=1}^{H} N_h^2\, s_{jh}^2\left(t_h-\frac{1}{N_h}\right)\le V_j \quad j=1,\,2,\,\ldots,\,k$$

The transformation $t_h^2 = n_h$ gave the nonlinear objective and constraint functions:

$$\min\ C_0+\sum_{h=1}^{H} c_h t_h^2 \tag{11}$$

$$\text{such that}\quad \sum_{h=1}^{H} N_h^2\, s_{jh}^2\left(\frac{1}{t_h^2}-\frac{1}{N_h}\right)\le V_j \quad j=1,\,2,\,\ldots,\,k.$$

Two comparisons for Yates' second problem were made. The first was to compare the results of the nonlinear solution with the approximate solution given by the convex programming method. Since the nonlinear solution is nearly exact, the transformation used to reach the solution is not of interest in the first comparison. In the second comparison, the robustness and speed of convergence for the three formulations of the nonlinear optimization problem were compared. This was performed by randomly generating four sets of initial guesses of the strata sample sizes (n_h). Then the number of failed convergence attempts, the average speed of convergence for each method, and the number of correct solutions were compared.

4. Results

Table 2 lists the solutions generated by the nonlinear optimization routine and the convex programming solution given by Huddleston et al. (1970). Since the solution for the nonlinear routines are identical to within roundoff error, the results are given for only one of the three transformations. The nonlinear solution has a larger overall sample size, but all constraints have been satisfied with a variance in stratum of only two less than the prescribed maximum (table 2). The convex programming solution given by Huddleston et al. (1970) violates the variance restrictions for variables three and six and gives a variance that is less than the variance restriction for the other five strata. In general the difference in sample size is not large. However, for four of the 15 strata the difference between the optimum sample size and the approximate sample size given by the convex programming solution exceeds 10%. This is because the convex programming solution uses a linear approximation of the actual problem.

As mentioned earlier there were some problems with the convergence of the nonlinear optimization technique. To test the robustness of the original problem and the two additional formulations, the data set was reduced in size by dropping a number of variables and strata from consideration and using four sets of randomly chosen starting values. The random starting values were generated using a uniform (0,400) distribution.

Table 2 lists the results of the comparison between the three nonlinear optimization formulations. The performance of the nonlinear optimization routine on the original formulation was unsatisfactory. The nonlinear routine converged to the correct solution only three times out of 20. When a routine did converge to a solution, the amount of time required to find the solution was at least an order of magnitude larger than for the other two routines. For the first transformation, $t_h = 1/n_h$, convergence to the optimum solution occurred in 17 out of 20 test cases. For the three smallest test cases, this transformation converged on every attempt and had the fastest convergence times and the smallest number of function and gradient evaluations. The transformation $t_h = n_h^2$ produced some of the most puzzling results. Convergence occurred for 15 of 20 test cases. For the three smallest test cases this transformation converged much more slowly than the transformation $t_h = 1/n_h$. For the two largest test cases, the transformation $t_h = n_h^2$ converged quickest and actually required fewer function and gradient evaluations than it did for the three smallest test cases.

5. Discussion

The results in tables 2 and 4 indicate two things. First, nonlinear techniques can be used to find exact solutions to Yates' second multivariate allocation problem. Secondly, while nonlinear techniques may be less likely to converge due to numerical problems, transforming the original problem can greatly improve the efficiency of these techniques.

It is not apparent as to why the transformations $t_h = n_h^2$ and $t_h = 1/n_h$ would prove to be more likely to converge than the original problem. The only conditions for using the algorithm given by Schittkowski (1985) are:

a) The first derivatives of the problem functions exist and are continuous on the domain of interest (continuously differentiable).

b) The algorithm is best suited for problems with less than 100 variables.

Both of these conditions are satisfied for the problem in its original formulation. We can gain some insight by examining how the problem is formulated by the IMSL routine.

For a given constrained nonlinear optimization problem of the form

$$\min f(\underline{n}) \qquad \textit{such that} \quad g_j(\underline{n}) \geq 0 \quad j=1,2..k, \tag{12}$$

where $\underline{n} = (n_1\ n_2\ \ldots\ n_h)$, subproblems using a quadratic approximation of the objective function and linearized constraints are formulated and iteratively solved. The quadratic approximation and linearized constraints to be solved are of the form

$$\min \frac{1}{2} d^T B d + \nabla f(\underline{n})^T d \qquad \textit{such that} \quad \nabla g_j(\underline{n})^T d + g_j(\underline{n}), \tag{13}$$

where

$$\frac{1}{2} d^T B d + \nabla f(\underline{n})^T d \approx \frac{1}{2} d^T\, \nabla^2 f(\underline{n})\, d + \nabla f(\underline{n})^T d. \tag{14}$$

The matrix B in (13) and (14) is a positive definite approximation to the Hessian of f($\underline{n}$) and is computed using the BFGS update technique suggested by Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970). The convergence problems for the original formulation are caused by the form of the Hessian and the approximation B. A comparison of the approximation given in (14) follows:

For the transformation $t_h^2 = n_h$

$$d^T \nabla^2 f(\underline{n})\, d + \nabla f(\underline{n})^T d = d^T \begin{vmatrix} 2c_1 & & & 0 \\ & 2c_2 & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & 2c_h \end{vmatrix} d + (2c_1 t_1 \ldots 2c_h t_h)\, d \tag{15}$$

for the transformation $t_h = n_h$

$$d^T \nabla^2 f(\underline{n})\, d + \nabla f(\underline{n})^T d = d^T \begin{vmatrix} 0 & & & 0 \\ & 0 & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & 0 \end{vmatrix} d + (2c_1 \ldots 2c_h)\, d \tag{16}$$

and for the transformation $t_h = 1/n_h$

$$d^T\nabla^2 f(\underline{n})\,d + \nabla f(\underline{n})^T d = d^T \begin{vmatrix} 4\frac{c_1}{t_1^3} & & & 0 \\ & 4\frac{c_2}{t_2^3} & & \\ & & \cdot & \\ & & \cdot & \\ 0 & & & 4\frac{c_h}{t_t^3} \end{vmatrix} d - (2\frac{c_1}{t_1^2}\,..2\frac{c_h}{t_h^2})\,d. \tag{17}$$

Clearly, the forms given by the transformations $t_h = n_h^2$ and $t_h = 1/n_h$ can easily be approximated by the matrix B, since both formulations are always positive definite for $t_h > 0$. The Hessian in the formulation for $t_h = n_h$ is not positive definite. This would imply that B is always going to be a poor approximation to the Hessian. The Schittkowski routine is not appropriate for the problem in its originally proposed form, even though nothing in the documentation would indicate this. An additional item worth noting is that for the transformation $t_h = n_h^2$ the approximating matrix B could be replaced by the true Hessian. This could provide improved speed and likelihood of convergence, but it would require modifying the source code for the Schittkowski routine.

6. Conclusions

The nonlinear technique discussed provides exact optimal solutions to the problem posed, provided the appropriate transformations are used. The transformation $t_h = 1/n_h$ produces the best results when the number of variables to be sampled and the number of strata are less than seven. For larger problems the transformation $t_h = n_h^2$ produces the best results and the convergence of this routine might be improved by replacing the Hessian approximation with the actual Hessian.

References

Booth, G., and Sedransk, J. 1969. Planning some two-factor comparative surveys. Jour. Amer. Stat. Assoc. 64:560-573.

Broyden, C. G. 1970. The convergence of a class of double rank minimization algorithms, parts I and II. J. Inst. Maths. Applns. 6:76-90 and 222-231.

Chatterjee, S. 1967. A note on optimum stratification. Skand. Akt. 50:40-44.

Cochran, W. G. 1977. Sampling techniques. 3rd ed. J. Wiley, NY.

Fletcher, R. 1970. A new approach to variable metric algorithms. Computer J., 13:317-322.

Goldfarb, D. 1970. A family of variable metric methods derived by variational means. Maths. Comp. 24:23-26.

Hahn, J. T., Czaplewski, R. L., Hansen, M. H., and Chen, C. -M. 1992. Building a prototype system to develop and test AFIS procedures. Unpublished document. USDA Forest Service, North Central Forest Experiment Station, St. Paul, Minnesota.

Hansen, M. H., Madow, W. G., and Tepping, B. J. 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Amer. Stat. Assoc. 78:776-807 (includes discussion).

Hartley, H. O. 1959. Nonlinear programming by the simplex method. Econometrika. 29:223-237.

Hartley, H. O. and Hocking, R. R. 1963. Convex programming by tangential approximation. Manage. Sci. 9:600-612.

Huddleston, H. F., Claypool, P. L., and Hocking, R. R. 1970. Optimum sample allocation to strata using convex programming. Appl. Stat. 19:273-278.

IMSL. 1989. Math/library. Fortran Subroutines for Mathematical Applications. Softcover edition. Version 1.1. IMSL Corp., Houston, Texas.

NC FIA. 1992. Problem analysis for Annual Forest Inventory Systems (AFIS). Unpublished Document. USDA Forest Service. North Central Forest Experiment Station, St. Paul, Minn.

Schittkowski, K. 1986. NLPQL: A Fortran subroutine solving constrained nonlinear programming problems (edited by C. L. Mouma). Ann. Operations Research. 5:485-500.

Schreuder, H. T., Gregoire, T. G., and Wood, G. B. 1993. Sampling methods for multiresource forest inventory. J. Wiley and Sons, NY 446p.

Shanno, D. F. 1970. Conditioning of quasi-Newton methods for function minimization. Maths. Comp. 14: 149-160.

Yates, F. 1960. Sampling methods for censuses and surveys. 3rd ed. Charles Griffin and Co., London, 3rd. ed.

Table 1. Data for 15 strata and seven variables given in Huddleston et al. (1970).

Stratum	Cost ($)	Number of population sampling units	Standard deviations, S_{hj}, of the characteristics						
h	c_h	N_h	S_{h1}	S_{h2}	S_{h3}	S_{h4}	S_{h5}	S_{h6}	S_{h7}
1	21.2	4714	1.0	311	27	161	551	30	350
2	20.6	5718	4.5	70	1	208	27	9	331
3	18.3	4686	1.2	135	80	126	152	13	65
4	20.8	6134	1.2	265	266	86	115	99	50
5	23.6	9912	8.6	116	78	35	79	55	24
6	19.7	28044	2.0	74	65	44	34	49	45
7	16.2	24642	2.2	75	1	5	1	5	2
8	20.5	11328	4.5	98	1	13	1	19	8
9	35.3	1144	5.1	844	2015	1386	1	4	372
10	23.7	4948	2.5	321	2507	81	30	91	123
11	18.6	14932	1.7	98	22	43	1	64	86
12	18.1	1378	1.4	88	3	293	22	7	488
13	16.3	7016	4.9	190	6	88	2	27	71
14	27.0	2808	3.0	491	7	85	25	61	203
15	23.4	3038	1.7	67	1	148	1	1	444
$V^{1/2}(\hat{Y}_j)$	(1000)		12.3	709.6	648.9	281.9	218.8	160.3	376.0
Estimated population total (1000)			246	11,827	6,489	5,639	3,646	2,004	6,267
Specified coeff. of variation (%) of population total			5	6	10	5	6	8	6

Table 2. Comparisons of the results generated by the convex programming method and the nonlinear optimization.

	Method of Allocation		
Strata	Huddleston	Nonlin. Opt.	Difference
1	232	232	0
2	103	96	+7
3	87	75	+8*
4	117	104	+13*
5	150	155	-5
6	207	200	+7
7	106	104	+2
8	92	88	+4
9	87	87	0
10	369	443	-74*
11	98	110	-12*
12	33	32	+1
13	83	77	+6
14	27	27	0
15	49	49	0
	---	---	
Total	1840	1881	
Cost	39752	40901	

* indicates > 10% error

Achieved Variance in Each Stratum

Variable constraint		Huddleston	Nonlin. Opt.	Variance
$V(Y_1)$	=	150228775.-	151290000.	151290000.
$V(Y_2)$	=	241790768949.-	245465943804.-	503532160000.
$V(Y_3)$	=	487307626783.+	421071210000.	421071210000.
$V(Y_4)$	=	77370898107.-	79467610000.	79467610000.
$V(Y_5)$	=	46431722592.-	47873440000.	47873440000.
$V(Y_6)$	=	26162049632.+	25696090000.	25696090000.
$V(Y_7)$	=	139612535883.-	141376000000.	141376000000.

- : Achieved variance less than constrained variance
+ : Achieved variance greater than constrained variance

Table 3. Comparison of convergence properties for the three proposed nonlinear optimization problems.

Test problem	Number of successful attempts	Number of unsuccessful attempts	Average run time	Average # of function evaluations	Average # of gradient evaluations
Strata = 2 Variables = 2					
$n_h = t_h$	0	4	*	*	*
$n_h = 1/t_h$	4	0	0.045	3.75	3.50
$n_h = t_h^2$	4	0	0.805	150.50	71.00
Strata = 7 Variables = 4					
$n_h = t_h$	1	3	127.250	7287.00	1981.00
$n_h = 1/t_h$	4	0	1.612	33.75	21.25
$n_h = t_h^2$	4	0	3.170	119.25	63.00
Strata = 4 Variables = 7					
$n_h = t_h$	1	3	62.810	7165.00	1920.00
$n_h = 1/t_h$	4	0	0.665	18.00	13.00
$n_h = t_h^2$	2	2	1.545	107.00	55.00
Strata = 10 Variables = 7					
$n_h = t_h$	1	3	246.090	6331.00	1885.00
$n_h = 1/t_h$	1	3	6.900	53.00	34.00
$n_h = t_h^2$	3	1	1.125	15.00	14.33
Strata = 15 Variables = 7					
$n_h = t_h$	0	4	*	*	*
$n_h = 1/t_h$	4	0	16.410	57.25	32.25
$n_h = t_h^2$	2	2	4.500	22.50	19.50

*Not applicable.