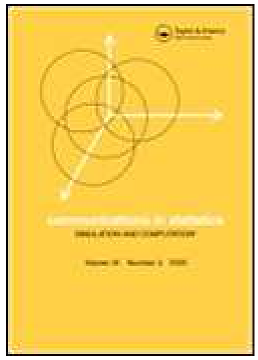
Allocation constraints in stratification

Marcin Kozak, Pawel Jankowski



Allocation constraints in stratification

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2007-0207.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	13-Jun-2008
Complete List of Authors:	Kozak, Marcin; Warsaw University of Life Sciences, Experimental Design and Bioinformatics Jankowski, Paweł; Warsaw University of Life Sciences, Biometry
Keywords:	constraints, optimization, optimum stratification, sample allocation
Abstract:	When a finite population is to be stratified, one of constraints in stratification is that sample sizes from strata may not be greater than the corresponding strata sizes and may not be smaller than two. There are several ways of treating this allocation constraint, each providing an alternative approach to stratification. In the paper it is shown that a choice of the approach has a bearing on stratification efficiency. Unfortunately, no particular approach out of the four compared is shown to be the most efficient for each population studied. In addition, the approaches are applied to stratify a real population.

Allocation constraints in stratification

Marcin Kozak[1]*, Paweł Jankowski[2]

[1] Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences, Nowoursynowska 159, 02-787, Warsaw, Poland,

[2] Department of Biometry, Warsaw University of Life Sciences, Nowoursynowska 159, 02-787, Warsaw, Poland,

*Corresponding author, Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences

Nowoursynowska 159, 02-787, Warsaw, Poland, e-mail: **m.kozak@omega.sggw.waw.pl**

Running head: Allocation constraints in stratification

Allocation constraints in stratification

Abstract

When a finite population is to be stratified, one of constraints in stratification is that sample sizes from strata may not be greater than the corresponding strata sizes and may not be smaller than two. There are several ways of treating this allocation constraint, each providing an alternative approach to stratification. In the paper it is shown that a choice of the approach has a bearing on stratification efficiency. Unfortunately, no particular approach out of the four compared is shown to be the most efficient for each population studied. In addition, the approaches are applied to stratify a real population.

Keywords: constraints, optimization, optimum stratification, sample allocation

Mathematics Subject Classification: Primary 62D05, Secondary 62P20, 62P25.

1. Introduction

In this paper we consider an optimization approach to stratification, which has been recently proved to be superior to approximate stratification procedures (see Kozak and Verma (2006) and the citations therein). Suppose we aim to stratify a finite population U based on an auxiliary (stratification) variable X. Let the aim of stratification be minimizing the variance of an estimator of the population total of a study variable Y subject to fixed sample size n. At the design stage of a survey it is usually assumed that a survey variable (Y) and stratification variable (X) be the same and that there be no non-responses. This is, of course, never the case in practice, yet such an approach is common and is not thought of as controversial. Furthermore, let us consider the common practical situation in which the survey and stratification variables are positively skewed; then, the most efficient approach is to construct a so-called take-all stratum, from which all the elements are taken to the sample (e.g., Hidiroglou, 1986; Lavallée and Hidiroglou, 1988).

The objective function to be minimized in the problem in question is the variance $Var(\hat{t}_X)$ of an estimator $\hat{t}_X$ of the population total of *X*. Assuming that the *L*th stratum is the take-all one, the variance, under the take-all stratum approach, takes the form

$$Var(\hat{t}_X) = \sum_{h=1}^{L-1} S_h^2 W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \tag{1}$$

$$\hat{t}_X = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{k=1}^{n_h} X_{hk} \,;\; t_X = \sum_{h=1}^{L} \sum_{k=1}^{N_h} X_{hk} \,;\; W_h = N_h \Big/ \sum_{h=1}^{L} N_h \,;$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} \left(X_{hk} - \overline{X}_h \right)^2 \text{ for } h = 1, \ldots, L-1; \;\; \overline{X}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} X_{hk}$$

where $\hat{t}_X$ is the unbiased estimator of the population total t_X of *X*; S_h^2 is the population variance of the variable *X* restricted to the *h*th stratum; n_h is the sample size from the *h*th stratum of size N_h; X_{hk} is the value of *X* for the *k*th population element of the *h*th stratum; and $\overline{X}_h$ is the population mean of *X* restricted to the stratum *h*. In equation (1), we have considered the most classical unbiased estimation of the population total under stratified sampling; see, e.g., Särndal et al. (1992).

A vector of stratification points, say $\mathbf{a} = (a_1, \ldots, a_{L-1})^T$, which explicitly defines the subdivision of the population *U* into strata, is a vector of parameters to be searched for (e.g., Lednicki and Wieczorkowski, 2003). It is to be noted that the variance (1) does not directly involve the parameters sought. An objective function $f(\mathbf{a})$ can be written in a general form as

$$f(\mathbf{a}) = Var(\hat{t}_X) \tag{2}$$

where $Var(\hat{t}_X)$ is given in equation (1). The constraints for the function (2) are as follows:

$$2 \le n_h \le N_h, \; h = 1,..., L-1 \tag{3}$$

$$N_h \ge 2; \; h = 1,..., L \tag{4}$$

$$N_L + \sum_{h=1}^{L-1} n_h = n \tag{5}$$

Fulfilling the constraints (3) and (4) is required to ensure that the variance (1) can be evaluated. (Note that the constraint $N_L \ge 2$ is not required here, but it is reasonable to use it in order to obtain a take-all stratum comprising at least two elements.) Sample sizes n_h from strata are usually determined through the Neyman optimum sample allocation, which aims at minimizing the variance (1); after adjusting the formula for the take-all stratum approach, the sample sizes are given by

$$n_h = (n - N_L)\frac{W_h S_h}{\sum_{h=1}^{L-1} W_h S_h}, \; h = 1,..., L-1; \; n_L = N_L \tag{6}$$

There are two possible ways of treating the constraints (3): (i) one does not accept the solution in which any n_h provided by the formula (6) does not fulfil the constraints (3), and changes the stratification points (such an approach was applied, for instance, by Lednicki and Wieczorkowski (2003)); and (ii) one does not accept the allocation and searches for the optimum allocation using numerical optimization. The first approach is obviously easier to implement and provides less time-consuming computation. However, intuition makes us suppose that this approach may give rise to rejecting solutions that either are optimal or may be connections between stratification points considered in a particular step and the optimum points (or a path leading to the optimum points).

The option (i) of treating the constraints can be applied in two manners. First, one can reject points not fulfilling the constraints (3). Second, one can apply the following procedure to adjust the sample sizes for the constraints: determine n_h through the allocation (6) and apply the following formula (for $h = 1,..., L-1$)

$$n_h = 2 \text{ if } n_h < 2;\ n_h = N_h \text{ if } n_h > N_h \tag{7}$$

This procedure makes us accept solutions that would be rejected by the first manner, in which way a set of possible solutions is widened.

In this paper we compare the following approaches to treating the allocation constraints in stratification:

(A) Approach based on not accepting a solution (stratification) in which any n_h provided by the formula (6) does not fulfil the constraints (3).

(B) Approach based on applying the allocation (6) with the adjustment (7).

(C) Approach based on applying numerical optimization to allocate the sample: if n_h's provided by the formula (6) do not fulfil the constraints (3), solve the following problem to find the allocation. Given a vector of stratification points **a**, find such n_h's, $h = 1,..., L-1$, that minimize the objective function (1), i.e., $f(n_1,..., n_{L-1}) = Var(\hat{t}_X)$, under the constraints (3) and (5).

Kozak (2004b) showed that results of stratification determined by numerical optimization depend on stratification points that are used as initial parameters in the optimization. Therefore, we will consider an additional approach as follows:

(D) Approach based on applying approach C with strata boundaries provided by approach B taken as initial parameters in optimization.

Hence, a question to answer is, does the choice of an approach of treating the allocation constraints in stratification have an influence on stratification efficiency? The aim of the paper is to answer this question through a simulation study.

2. Design of experiment

The following aspects of a population and a stratification variable were considered in the experiment: (i) population size N, viz., $N = \{1000, 2000, 5000, 10000, 15000\}$; (ii) number L of strata to be constructed, viz., $L = \{3, 5, 7, 9\}$; (c) sample fraction $f = n / N$ (n being the assumed sample size), viz., $f = \{0.1, 0.2\}$; and (d) parameter σ of the distribution of the stratification variable (see below), viz., $\sigma = \{0.4, 0.6, 0.8\}$. For the sake of convenience, below the population quantities N, L, f, and σ will be referred to as factors.

Stratification variables were generated based on the following formula:

$$X = [\exp(Z)],$$

where Z is the realization of an $N(10, \sigma^2)$ variable (a normal random variable with mean 10 and standard deviation σ) and the function $[\,\cdot\,]$ stands for rounding to integers (to simulate the most often practical situation). As a result of such generation, the variables were positively skewed; the greater the σ value, the greater the skewness was.

For each combination of $N \times L \times f \times \sigma$, 100 independent populations (stratification variables) were generated; thus there were 12000 populations altogether. For an ith population, the four stratification approaches of study were applied and the coefficient of variation $cv_{ki}(\hat{t}_X^i)$ (k referring to the kth approach to stratification, $k = A, B, C, D$) of the estimator $\hat{t}_X^i$ was evaluated using the formula

$$cv_{ki}\left(\hat{t}_X^i\right)=\left(t_X^i\right)^{-1}\sqrt{Var_k\left(\hat{t}_X^i\right)} \tag{8}$$

where $Var_k\left(\hat{t}_X^i\right)$ is the variance (1) of the estimator $\hat{t}_X^i$ under the *k*th approach to stratification, and t_X^i is the total of *X* in the *i*th population.

As Lednicki and Wieczorkowski (2003) did, to perform stratification we have applied *optim* function, which implements the algorithm of the simplex method of Nelder and Mead (1965), available in R language and environment (R Development Core Team 2007). Following Kozak's (2004a) results on efficiency of approximate stratification points used as initial parameters in optimization, stratification points determined by Mahalanobis's (1952) procedure were taken as the initial vector of parameters in optimization. However, because this procedure does not take account of the take-all stratum approach, the point defining the last stratum was changed in such a way that this stratum comprised five population elements with the largest values of the stratification variable. (Note that the take-all stratum contained five elements at the initial stage of stratification, but later the number of its elements was not limited to five.) Whenever these stratification points did not fulfil the constraints (3) and/or (4), we used the points provided by (i) the Ekman (1959) procedure; (ii) then, whenever Ekman's points failed, the Dalenius and Hodges (1959) procedure were applied; and finally, (iii) whenever Dalenius and Hodges' points failed, we used the Gunning and Horgan (2004) procedure. Each such procedure was applied with the above-mentioned adjustment for the take-all stratum approach. However, whenever all these procedures failed to fulfil the constraints (3), the initial strata were constructed based on the following procedure. First, the five-element take-all stratum had been constructed, and then the remaining part of the population was subdivided into $L-1$ strata of equal sizes (or nearly equal, if $(N-5)/(L-1)$ was not an integer).

In the approaches C and D, R's function *optim* was applied to determine the optimum sample allocation; as the initial parameters (sample sizes from strata) in the optimization, the $(L-1)$-vector of twos was taken.

All the computation was performed in R (R Development Core Team, 2007) using self-implemented functions, which can be obtained from the corresponding author upon request.

3. Results

Results of the simulation study are presented in Tables 1-5, for $N = 1000$, 2000, 5000, 10000 and 15000, respectively. For a particular population we calculated ranks for the values of the coefficient of variation (cv), given by Eq. (8), of the estimator obtained under the four approaches. Then we determined the number of times in which (i) a particular approach was the best (cv obtained under approach A, C, or D had rank 1; cv obtained under approach B had rank 1.5 provided that cv obtained under approach D had rank 1.5, too); (ii) a particular approach was the worst (cv obtained under approach A or C had rank 4; cv obtained under approach B had rank 3.5, provided that cv obtained under approach D had rank 3.5, or rank 4; cv obtained under approach D had rank 3.5 provided that cv obtained under approach B had rank 3.5, too). Through this analysis of ranks approach B was recognized as the best in a situation when approach D had not improved its results and approaches A and C had been worse (we would not recognize approach D as the best then, since its application did not provide any gain in efficiency in comparison to its initial parameters that had been provided by approach B). On the other hand, approach D was recognized as the worst when it had not improved the results of approach B, of which the results of approaches A and C had been better (in such a situation approach B was also recognized as the worst).

[Table 1]

[Table 2]

[Table 3]

[Table 4]

[Table 5]

In addition, Table 6 contains mean ranks determined for the approaches under a particular level of the population quantities studied; from this table it follows that in general approach D appears to be the best (in the sense that it provides the most efficient stratification points); approaches A, B and C seem to provide similar results (inferior to those of approach D), even though a slight tendency of approach C to be better than approaches A and B has been detected. Such a result has been obtained for all the levels of the factors studied except for $L=3$, in which case all the approaches provided similar mean ranks though approach D appeared to be slightly better than the other approaches. For all the other factor levels, cv obtained under approach D had the smallest mean rank. Usually cv obtained under approach C had a little smaller mean rank than cvs obtained under approaches A and B, but for some combinations cvs obtained under approaches A, B and C had similar mean ranks. Differences between the values of cv obtained by different approaches were sometimes meaningful (e.g., it sometimes happened that $cv_{worst} > 1.5cv_{best}$, the first cv referring to the worst and the second to the best approach).

[Table 6]

Several interesting situations occurred. Under σ = 0.4 and 0.6, the larger *L*, the more often approach D was the best. However, this situation did not occur under σ = 0.8. In general, under *L* = 5, 7 and 9, approaches A and B, and sometimes C, were quite often the worst. In general, approach B was seldom the best; neither was approach C, although under σ = 0.8, *f* = 0.2, and *L* = 7 and 9 it was usually the best. The hypothesis that approach C and/or D may always be the best, which was mentioned in Introduction, has not been proven correct by our experiment. Nonetheless, in most situations approach D often appeared the best; in addition, it seldom was the worst. That approach D was usually better than approach C is easy to explain—it resulted from more efficient initial strata points used in the former than those used in the latter. Nonetheless, let us recall the combinations σ = 0.8, *f* = 0.2, and *L* = 7 and 9, in which, without any reasonable and explicable reason, approach D was usually worse than approach C.

Approach A has one important drawback that must be mentioned here. There were many populations, especially under high *N*, σ, and *f* values, for which the optimization was unable to perform this stratification based on all the initial values mentioned, for which reason this approach failed to stratify such populations. This situation did not occur for any other approach.

We have not studied the approach in which the results of approach A would be taken as initial parameters to perform optimization in approach C. The main reason was that such an approach would fail in a situation in which approach A did not succeed to provide stratification points fulfilling the constraints (3) and (4).

Based on the results obtained we are not able to choose the best approach explicitly. In practice the best way is to apply all the approaches (even with points provided by the approach A taken as initial parameters if only they are feasible) and to choose the best one. However, if for any reason it is unlikely to be done, approach D should be chosen, as the best and the least risky in our experiment. Of course, a simulation study as it is, it does consider a somewhat limited range of possible populations and stratification approaches (e.g., a population size, variability in a stratification variable, number of strata to be constructed, and the like). It is possible that considering other population and stratification attributes, we could obtain different results. Nonetheless, this would not change our conclusions that we cannot point out any particular approach as the best one, and that the choice of an approach does count even though still no explicit recommendation can be given.

The most possible reason why approaches C and D were not the best for all the populations is that Nelder and Mead's optimization might lead to the local minimum of a function optimized. Moreover, initial stratification points have a bearing on the optimization results (Kozak, 2004b). We have applied several approximate stratification procedures to provide initial stratification points, but there is no certainty that those stratification points are really the best ones. Further efforts should be focused on determining a procedure that would provide more efficient stratification than Nelder and Mead's optimization approach does. Kozak's (2004a) random search algorithm is a very promising one, as claimed by Baillargeon and Rivest (2007), but we need to remember that as a global optimization method this procedure provides random results. Hence the best option is to apply it several times (maybe with various starting points) to ensure that the results obtained are indeed globally and not locally optimum. Worth noting is that Baillargeon and Rivest (2007) implemented a non-random version of the algorithm, which is of course free of the problem of random results.

Deleted: The main aim of this paper was to show that a choice of treating the constraints in stratification does matter; this has been proven indeed. What is more, we have shown that this choice is very important, since it may cause results of stratification under various treating the constraints be very different. In the next section we will present the application of the four approaches for a real population.

4. Example

Here we apply the four allocations for a real data set SHS available in the package *stratification* (Baillargeon and Rivest, 2007) of R (R Development Core Team, 2007). The set contains data for 16057 units from the 2001 Survey of Household Spending (SHS) Statistics Canada; for stratification we will use one variable, namely "household income before taxes". The results are given in Table 7.

Apparently approaches A and D were the most and approach C the least efficient for this particular data set. Interestingly, all the approaches provided the same take-all stratum (even though the boundaries for the take-all strata they provided differed, all of them comprised the same eight population units). Note that due to rounding of sample sizes from strata there were some inconsistencies in overall sample sizes as in none of the cases it equalled 1500 (it was either 1499 or 1501), but for so large a sample size these two elements did not make any real difference.

[Table 7]

5. Conclusion

This paper aimed to show that the way of treating the constraint (3) does matter. This has been proven indeed: The choice may cause results of stratification under various treating the constraints be very different in terms of the precision of estimation of a parameter studied. Unfortunately, of four such ways considered in this paper, none was ultimately the best.

From the results we have concluded that in practical applications the best way is to apply all the approaches and to choose the best one for the particular population. If this is impossible, approach D should be applied as most often the best and the least risky in our experiment.

References

Baillargeon, S. and L.P. Rivest (2007). stratification: Stratification of Survey Populations. R package version 1.0.

Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratification. Journal of the American Statistical Association 54:88—101.

Ekman, G., (1959), An Approximation Useful in Univariate Stratification. Annals of Mathematical Statistics 30:219—229.

Gunning, P. and Horgan, J.M. (2004). A Simple Algorithm for Stratifying Skewed Populations. Survey Methodology 30:159—166.

Hidiroglou, M. (1986). The Construction of a Self-Representing Stratum of Large Units in Survey Design. The American Statistician 40:27—31.

Kozak, M. (2004a). Optimal Stratification Using Random Search Method in Agricultural Surveys. Statistics in Transition 6 (5):797—806.

Kozak, M. (2004b). Optimal Stratification with the Auxiliary Variable. Wiadomosci Statystyczne (Statistical News) 8:29—34 (in Polish).

Kozak, M. and Verma, M.R. (2006). Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency. Survey Methodology 32(2):157—163.

Lavallée, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Population. Survey Methodology 14:3—43.

Lednicki, B. and Wieczorkowski, R. (2003). Optimal Stratification and Sample Allocation between Subpopulations and Strata. Statistics in Transition 6:287—306.

Mahalanobis, P.C. (1952). Some Aspects of the Design of Sample Surveys. Sankhya 12:1—7.

Nelder, J.A. and Mead, R. (1965). A Simplex Method for Function Minimization. Computer Journal 7:308—313.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; URL http://www.R-project.org.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Table 1. Summary of rank-ordering of efficiency of stratification approaches A, B, C, and D, for *N*=1000 and various combinations of σ, *f*, *L*.

σ	*f*	*L*	A[a] best	A[a] worst	A[b] failed	B[c] best	B[d] worst	C[a] best	C[a] worst	D[e] best	D[f] worst
0.4	0.1	3	0	0	0	0	0	0	0	7	0
0.4	0.1	5	12	17	0	5	22	13	38	25	3
0.4	0.1	7	20	28	0	9	24	20	42	48	3
0.4	0.1	9	22	19	0	1	24	8	56	69	0
0.4	0.2	3	2	3	0	0	0	0	0	7	0
0.4	0.2	5	18	33	0	3	6	2	6	46	0
0.4	0.2	7	17	39	0	2	20	11	24	66	0
0.4	0.2	9	18	43	3	1	31	12	24	69	0
0.6	0.1	3	1	0	0	0	0	0	1	12	0
0.6	0.1	5	18	21	1	12	16	17	43	33	4
0.6	0.1	7	19	21	1	6	24	19	50	54	3
0.6	0.1	9	14	16	0	2	15	8	68	76	0
0.6	0.2	3	2	6	0	1	0	0	0	9	0
0.6	0.2	5	16	62	0	2	4	0	4	72	0
0.6	0.2	7	17	39	0	0	30	14	22	68	1
0.6	0.2	9	28	37	2	0	31	23	32	49	0
0.8	0.1	3	0	2	0	0	0	1	0	18	0
0.8	0.1	5	12	32	0	10	28	29	25	40	6
0.8	0.1	7	16	21	0	3	16	13	59	66	2
0.8	0.1	9	5	27	0	3	11	8	62	84	0
0.8	0.2	3	2	9	0	0	0	0	0	6	0
0.8	0.2	5	9	58	1	3	3	2	9	73	1
0.8	0.2	7	23	34	7	0	39	19	22	57	1
0.8	0.2	9	13	66	2	1	22	63	12	23	0

[a] "A best", "C best", "A worst", and "C worst" indicate number of times in which a particular approach (A or C) had rank 1 (best) or rank 4 (worst); [b] "A failed" indicates number of time in which numerical problems occurred under approach A, so the solution was not be found; [c] "B best" indicates number of times in which this approach had rank 1.5 provided that approach D had rank 1.5, too; [d] "B worst" indicates

number of times in which this approach had rank 4 or 3.5 provided that approach D also had rank 3.5; [e] "D best" indicates number of times in which approach D had rank 1; [f] "D worst" indicates number of times in which this approach had rank 3.5 provided that approach B had rank 1, too.

Table 2. Summary of rank-ordering of efficiency of stratification approaches A, B, C, and D, for N=2000 and various combinations of σ, f, L.

σ	f	L	A[a] best	A[a] worst	A[b] failed	B[c] best	B[d] worst	C[a] best	C[a] worst	D[e] best	D[f] worst
0.4	0.1	3	0	0	0	0	0	0	0	4	0
0.4	0.1	5	19	43	0	1	6	0	2	37	1
0.4	0.1	7	25	35	0	1	31	10	15	62	0
0.4	0.1	9	35	29	0	0	42	11	26	54	0
0.4	0.2	3	2	4	0	0	0	0	0	3	0
0.4	0.2	5	13	42	0	0	4	0	2	62	0
0.4	0.2	7	22	47	0	0	17	6	14	72	0
0.4	0.2	9	16	53	0	0	19	4	20	80	0
0.6	0.1	3	0	1	0	0	0	0	0	5	0
0.6	0.1	5	18	57	0	3	10	1	2	58	0
0.6	0.1	7	31	20	0	0	37	8	33	59	0
0.6	0.1	9	30	19	1	0	40	14	41	56	0
0.6	0.2	3	1	4	0	0	0	0	0	12	0
0.6	0.2	5	10	59	0	0	1	1	3	75	0
0.6	0.2	7	17	39	0	1	18	1	19	80	1
0.6	0.2	9	23	48	1	0	33	28	12	49	0
0.8	0.1	3	0	3	0	0	0	0	0	18	0
0.8	0.1	5	17	54	0	3	7	3	5	62	0
0.8	0.1	7	28	21	0	1	39	20	38	51	1
0.8	0.1	9	31	18	0	1	33	17	49	51	1
0.8	0.2	3	1	10	0	0	0	0	0	19	0
0.8	0.2	5	13	58	4	1	3	2	7	68	1
0.8	0.2	7	10	60	7	2	23	45	7	43	1
0.8	0.2	9	3	82	46	0	18	85	0	12	0

[a] "A best", "C best", "A worst", and "C worst" indicate number of times in which a particular approach (A or C) had rank 1 (best) or rank 4 (worst); [b] "A failed" indicates number of time in which numerical problems occurred under approach A, so the solution was not be found; [c] "B best" indicates number of times in which this approach had rank 1.5 provided that approach D had rank 1.5, too; [d] "B worst" indicates

number of times in which this approach had rank 4 or 3.5 provided that approach D also had rank 3.5; [e] "D best" indicates number of times in which approach D had rank 1; [f] "D worst" indicates number of times in which this approach had rank 3.5 provided that approach B had rank 1, too.

Table 3. Summary of rank-ordering of efficiency of stratification approaches A, B, C, and D, for N=5000 and various combinations of σ, f, L.

σ	f	L	A[a] best	A[a] worst	A[b] failed	B[c] best	B[d] worst	C[a] best	C[a] worst	D[e] best	D[f] worst
0.4	0.1	3	0	0	0	0	0	0	0	7	0
0.4	0.1	5	22	35	0	1	3	1	2	41	0
0.4	0.1	7	12	60	0	1	10	2	8	85	0
0.4	0.1	9	8	57	0	0	20	2	12	90	0
0.4	0.2	3	0	0	0	0	0	0	0	8	0
0.4	0.2	5	3	39	0	1	0	2	2	66	0
0.4	0.2	7	17	48	0	0	15	6	14	76	0
0.4	0.2	9	14	49	0	0	18	6	20	80	0
0.6	0.1	3	1	1	0	0	0	0	0	11	0
0.6	0.1	5	10	41	0	1	8	2	6	65	0
0.6	0.1	7	13	47	0	1	20	5	18	81	0
0.6	0.1	9	26	32	0	0	26	2	26	72	0
0.6	0.2	3	0	0	0	0	0	0	0	22	0
0.6	0.2	5	7	47	0	1	3	1	2	73	0
0.6	0.2	7	7	42	0	0	25	9	18	84	0
0.6	0.2	9	4	86	2	0	6	51	8	45	0
0.8	0.1	3	0	6	0	0	0	0	0	11	0
0.8	0.1	5	11	44	0	4	8	1	7	69	0
0.8	0.1	7	22	61	3	0	12	8	10	69	0
0.8	0.1	9	32	17	0	0	37	5	39	63	0
0.8	0.2	3	0	5	0	0	0	0	0	18	0
0.8	0.2	5	14	35	0	1	3	3	7	40	1
0.8	0.2	7	2	76	3	0	23	79	0	19	1
0.8	0.2	9	0	100	85	0	0	96	0	4	0

[a] "A best", "C best", "A worst", and "C worst" indicate number of times in which a particular approach (A or C) had rank 1 (best) or rank 4 (worst); [b] "A failed" indicates number of time in which numerical problems occurred under approach A, so the solution was not be found; [c] "B best" indicates number of times in which this approach had rank 1.5 provided that approach D had rank 1.5, too; [d] "B worst" indicates

number of times in which this approach had rank 4 or 3.5 provided that approach D also had rank 3.5; [e] "D best" indicates number of times in which approach D had rank 1; [f] "D worst" indicates number of times in which this approach had rank 3.5 provided that approach B had rank 1, too.

Table 4. Summary of rank-ordering of efficiency of stratification approaches A, B, C, and D, for N=10000 and various combinations of σ, f, L.

σ	f	L	A[a] best	A[a] worst	A[b] failed	B[c] best	B[d] worst	C[a] best	C[a] worst	D[e] best	D[f] worst
0.4	0.1	3	0	1	0	0	0	0	0	12	0
0.4	0.1	5	15	34	0	1	1	3	2	46	0
0.4	0.1	7	18	49	0	0	12	6	10	75	0
0.4	0.1	9	10	54	0	0	13	3	17	87	0
0.4	0.2	3	0	0	0	0	0	0	0	11	0
0.4	0.2	5	6	31	0	0	1	0	1	63	0
0.4	0.2	7	9	59	0	0	5	3	12	88	0
0.4	0.2	9	8	49	0	0	21	6	13	86	0
0.6	0.1	3	0	1	0	0	0	0	0	18	0
0.6	0.1	5	7	32	0	0	6	2	1	67	0
0.6	0.1	7	12	44	0	0	20	8	21	80	0
0.6	0.1	9	9	43	0	0	27	10	18	81	0
0.6	0.2	3	0	0	0	0	0	0	0	29	0
0.6	0.2	5	11	38	0	1	0	2	4	65	0
0.6	0.2	7	15	54	0	0	14	10	12	74	1
0.6	0.2	9	0	99	0	0	1	77	0	23	0
0.8	0.1	3	0	3	0	0	0	0	0	17	0
0.8	0.1	5	13	35	0	0	3	0	3	56	3
0.8	0.1	7	12	62	3	0	13	14	11	72	0
0.8	0.1	9	26	34	0	0	29	7	27	67	0
0.8	0.2	3	0	0	0	0	0	0	0	23	0
0.8	0.2	5	8	26	0	0	1	2	1	55	0
0.8	0.2	7	4	71	1	0	27	89	1	7	1
0.8	0.2	9	0	100	94	0	0	98	0	2	0

[a] "A best", "C best", "A worst", and "C worst" indicate number of times in which a particular approach (A or C) had rank 1 (best) or rank 4 (worst); [b] "A failed" indicates number of time in which numerical problems occurred under approach A, so the solution was not be found; [c] "B best" indicates number of times in which this approach had rank 1.5 provided that approach D had rank 1.5, too; [d] "B worst" indicates

number of times in which this approach had rank 4 or 3.5 provided that approach D also had rank 3.5; [e] "D best" indicates number of times in which approach D had rank 1; [f] "D worst" indicates number of times in which this approach had rank 3.5 provided that approach B had rank 1, too.

Table 5. Summary of rank-ordering of efficiency of stratification approaches A, B, C, and D, for N=15000 and various combinations of σ, f, L.

σ	f	L	A[a] best	A[a] worst	A[b] failed	B[c] best	B[d] worst	C[a] best	C[a] worst	D[e] best	D[f] worst
0.4	0.1	3	0	0	0	0	0	0	0	10	0
0.4	0.1	5	8	33	0	0	2	0	0	66	1
0.4	0.1	7	6	67	0	0	11	3	4	89	0
0.4	0.1	9	7	57	0	0	10	5	17	88	0
0.4	0.2	3	0	0	0	0	0	0	0	14	0
0.4	0.2	5	1	30	0	0	1	0	0	75	0
0.4	0.2	7	10	47	0	0	8	10	13	80	0
0.4	0.2	9	10	52	0	0	6	3	14	87	0
0.6	0.1	3	1	0	0	0	0	0	0	25	0
0.6	0.1	5	14	36	0	1	3	1	3	69	1
0.6	0.1	7	12	35	0	0	24	7	21	81	0
0.6	0.1	9	12	41	0	0	20	1	21	87	0
0.6	0.2	3	0	0	0	0	0	0	0	33	0
0.6	0.2	5	6	21	0	1	0	0	0	77	0
0.6	0.2	7	10	45	0	1	19	21	15	67	0
0.6	0.2	9	0	100	0	0	0	86	0	14	0
0.8	0.1	3	0	7	0	0	0	0	0	21	0
0.8	0.1	5	7	20	0	1	2	3	2	58	0
0.8	0.1	7	21	48	3	2	11	7	9	66	0
0.8	0.1	9	25	38	2	0	31	13	21	62	0
0.8	0.2	3	0	2	0	0	0	0	0	27	0
0.8	0.2	5	7	21	0	0	7	5	4	53	1
0.8	0.2	7	2	82	0	0	17	93	1	5	0
0.8	0.2	9	0	100	97	0	0	98	0	2	0

[a] "A best", "C best", "A worst", and "C worst" indicate number of times in which a particular approach (A or C) had rank 1 (best) or rank 4 (worst); [b] "A failed" indicates number of time in which numerical problems occurred under approach A, so the solution was not be found; [c] "B best" indicates number of times in which this approach had rank 1.5 provided that approach D had rank 1.5, too; [d] "B worst" indicates

number of times in which this approach had rank 4 or 3.5 provided that approach D also had rank 3.5; [e] "D best" indicates number of times in which approach D had rank 1; [f] "D worst" indicates number of times in which this approach had rank 3.5 provided that approach B had rank 1, too.

Table 6. Mean rank for four approaches compared, under different levels of the population quantities studied.

	Items	A	B	C	D
N=1000	2400	2.73	2.79	2.77	1.71
N=2000	2400	2.74	2.89	2.67	1.70
N=5000	2400	2.93	2.84	2.6	1.64
N=10 000	2400	2.98	2.84	2.54	1.64
N=15 000	2400	2.99	2.85	2.55	1.61
σ=0.4	3200	2.81	2.81	2.77	1.61
σ=0.6	3200	2.85	2.86	2.68	1.61
σ=0.8	3200	2.95	2.86	2.43	1.76
L=3	2400	2.58	2.59	2.59	2.25
L=5	2400	2.92	2.81	2.78	1.49
L=7	2400	2.94	2.99	2.65	1.42
L=9	2400	3.05	2.98	2.49	1.48
f=0.1	4800	2.73	2.84	2.81	1.63
f=0.2	4800	3.02	2.85	2.45	1.69
Totally	96 000	2.87	2.84	2.63	1.66

Table 7. Stratification of household income before taxes for 16 057 units from the 2001 Survey of Household Spending (SHS) Statistics Canada (source: Baillargeon and Rivest, 2007) for seven strata and sample size n = 1500. (Sums of stratum sample sizes are different from 1500 due to rounding to integers.)

Stratum	Approach A			Approach B			Approach C			Approach D		
	k[1]	N_h	n_h	k	N_h	n_h	k	N_h	n_h	k	N_h	n_h
1	0	3717	200	0	3717	178	0	2755	103	0	3717	202
2	21707.55	3351	168	21862.38	2011	51	17076.72	2750	87	21017.47	3351	170
3	37177.37	3094	178	30978.86	4434	311	29363.74	2358	76	37365.61	2930	160
4	55383.14	2722	195	55897.71	2548	149	41727.19	3297	182	54053.91	3570	363
5	78944.03	2079	228	76158.94	1626	103	62832.64	2595	183	86728.29	1886	281
6	112557.23	1086	524	99599.61	1713	701	89082.52	2294	860	138399.45	595	315
7	488693.26	8	8	496513.15	8	8	480472.03	8	8	496177.07	8	8
CV	0.00374			0.00417			0.00436			0.00376		

[1] k is a stratum boundary; a particular hth stratum comprises units of which the stratification variable's values are within the interval $<k_h, k_{h+1})$