Available online at www.sciencedirect.com

SCIENCE DIRECT®

Mathematical and Computer Modelling 43 (2006) 530–535

MATHEMATICAL AND COMPUTER MODELLING

www.elsevier.com/locate/mcm

Stratification and sample size of data sources for agricultural mathematical programming models

Lucinio Judez[a], Carolina Chaya[a], José María de Miguel[a], Rafael Bru[b,*]

[a] *Departamento de Estadística y Métodos de Gestión en Agricultura, ETSIA, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain*

[b] *Institut de Matemàtica Multidisciplinar, ETSEA, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain*

Received 1 July 2005; accepted 6 July 2005

Abstract

A comparison is made between the variance of the estimator of the total of a variable obtained from both a simple and a stratified random sampling, in which the sample sizes of some strata are equal to the stratum population size.

It is shown that in this case, the advantage of the stratified sample could depend on the sample size. The paper presents inequalities that determine, as a function of the sample size, when the variance of the estimator obtained with simple sampling is lower than the variance obtained with the stratified sampling. The results give insight in order to prevent overstratification.

Keywords: Stratified sample; Sample size; Overstratification; Geographical stratification

1. Introduction

The classical results on stratified sampling [1, Ch. 5] provide expressions showing that, independently of the sample size, in most cases, the variance of the estimator of the total of a variable is lower than the one obtained using simple random sampling. In [1, p. 99], it is stated that simple random sampling could perform better than stratified sampling in some cases, pointing out that this could be an academic curiosity rather than something likely to happen in practice. However in [2], the study of the effects of geographical stratification in a Farm Accountancy Data Network (FADN) of the Navarra Autonomous Community of Spain, provides some instances where, depending on the sample size, the precision of the estimator of the total of a variable with a simple random sample is not improved by geographical stratification. This is a real case happening in practice. The authors found that those instances could appear when the sample size in some strata is identical to the size of the strata in the population. This condition is not taken into account to establish the classical results mentioned above [1].

The aim of this work is to study the values of the sample size, n, for which the stratified sampling is better than the simple random sampling and vice versa. This information could help in the decision making in applied

* Corresponding author.

E-mail addresses: lucinio.judez@upm.es (L. Judez), carolina.chaya@upm.es (C. Chaya), josemaria.demiguel@upm.es (J.M. de Miguel), rbru@imm.upv.es (R. Bru).

0895-7177/$ - see front matter
doi:10.1016/j.mcm.2005.07.006

problems involved with sampling. In particular, and given that stratification could lead, sometimes in practice, to higher costs, the findings presented here could be useful in order to decide the degree of stratification of an FADN, that provides information on the level of farm incomes and is used to analyse the effects of policy options [3]. Moreover, mathematical programming models using FADN data to analyse these effects show an exceptional development in the last few years [4–9] but little work has been done on the study of the data source which could improve the results of these models. The values of the sample size are given as a function of the size (N_h) and the standard deviation (S_h) of the strata h in the population, as well as, of the size (N) and the standard deviation (S) of the whole population. A necessary and sufficient condition is given in order to get a better performance of the simple random sampling as compared to the stratified sampling. That condition is also discussed in a particular case. The theoretical results are illustrated with some examples.

2. Simple random sampling versus stratified sampling

Consider a sampling plan to estimate the total of a variable, M, in a given population. That sampling plan is applied to a population of size N either with or without stratification; that is, dividing the population into subgroups called strata and taking some units from each strata or taking units of the whole population (see [1] for details). Since we are dealing with the estimation of a given characteristic of the population from a sample, the variance of the estimator of the characteristic should be considered. Let $V(\widehat{M}_{sp})$ and $V(\widehat{M}_{st})$ be the variance of the estimator of the total of the variable in the case of a simple random sample and in the case of a stratified sample, respectively, see Eqs. (3) and (1). In both cases, n denotes the number of units in the sample.

Consider that the population is stratified in L strata and assume that all the units of the first L_1 strata are included in the sample and these strata will be grouped in the subset T_1.

Then, the set of all strata T will be the union of

$$T = T_1 \bigcup T_2$$

where T_2 contains the remaining strata. Thus, $\mathrm{card}(T) = L$, $\mathrm{card}(T_1) = L_1$ and $\mathrm{card}(T_2) = L_2$.

In each stratum, N_h and n_h denote the number of units of the population and of the sample, respectively. $N_h = n_h$, for all $h = 1, 2, \ldots, L_1$, that is for all strata of T_1. Then, we can write

$$N = N_1' + N_2' \quad \text{and} \quad n = n_1' + n_2'$$

where

$$N_1' = \sum_{h=1}^{L_1} N_h \quad \text{and} \quad n_1' = \sum_{h=1}^{L_1} n_h$$

and

$$N_2' = \sum_{h=L_1+1}^{L} N_h \quad \text{and} \quad n_2' = \sum_{h=L_1+1}^{L} n_h.$$

From the property of the strata in T_1 note that $N_1' = n_1'$.

Then, the variance of the estimator with the stratified sample is [1, p. 93]

$$\begin{aligned} V(\widehat{M}_{st}) &= \sum_{h=1}^{L} N_h\,(N_h - n_h)\,\frac{S_h^2}{n_h} \\ &= \sum_{h=L_l+1}^{L} N_h\,(N_h - n_h)\,\frac{S_h^2}{n_h}, \end{aligned} \tag{1}$$

where S_h^2 is the variance of the variable in the strata h of the population. It is well-known that the variance of the estimator, $V(\widehat{M}_{st})$, is minimized using Neyman allocation [1, p. 98], that is, when

$$n_h = \frac{N_h S_h}{\sum\limits_{h=L_l+1}^{L} N_h S_h} n_2', \quad h = L_1 + 1, L_1 + 2, \ldots, L$$

and in this case $V(\widehat{M}_{st})$ becomes

$$V(\widehat{M}_{st}) = \frac{1}{n - N_1'} \left(\sum_{h=L_1+1}^{L} N_h S_h \right)^2 - \sum_{h=L_1+1}^{L} N_h S_h^2. \tag{2}$$

We recall that the variance of the estimator [1, p. 24] in the case of a simple random sample is

$$V(\widehat{M}_{sp}) = N(N-n)\frac{S^2}{n}, \tag{3}$$

where S^2 denotes the variance of our variable in the whole population, i.e.,

$$S^2 = \frac{1}{N-1}\left[\sum_{h\in T}(N_h - 1)S_h^2 + \sum_{h\in T} N_h (m_h - m)^2\right]$$

in which m and m_h are the mean in the population and in the stratum h of the variable, respectively.

With some algebraic manipulations, from Eqs. (2) and (3), one deduces that the inequality

$$V(\widehat{M}_{sp}) > V(\widehat{M}_{st}) \tag{4}$$

holds if and only if

$$An^2 + Bn + C > 0, \tag{5}$$

where

$$A = \sum_{h=L_1+1}^{L} N_h S_h^2 - NS^2$$

$$B = N^2 S^2 + N_1' \left(NS^2 - \sum_{h=L_1+1}^{L} N_h S_h^2 \right) - \left(\sum_{h=L_1+1}^{L} N_h S_h \right)^2$$

$$C = -N_1' N^2 S^2.$$

Expression (5) is obtained by developing the inequality (4) replacing $V(\widehat{M}_{sp})$ by its expression (3) and $V(\widehat{M}_{st})$ by its expression (2). The conclusion of the above discussion may be written as the following result.

Theorem 1. *Given a population of N units divided into L strata and a stratified sample of size n $(0 < n < N)$ obtained in such a way that, on the one hand, each stratum contains at least one unit and that, on the other hand, we have $n_h = N_h$ in L_1 strata (the first L_1, for instance), that is for all $h = 1, 2, \ldots, L_1$. Then, $V(\widehat{M}_{sp}) > V(\widehat{M}_{st})$ if and only if, n satisfies the inequality $An^2 + Bn + C > 0$, where the coefficients A, B and C are given above.*

We illustrate this result with the following example.

Example 1. Let a population of $N = 83$ units divided into $L = 6$ strata, having the characteristics given in Table 1. The distribution of the stratified sample is given in Table 2.

The mean is $m = 22.08$ and the variance $S^2 = 46.913$. In this case $T_1 = \{1, 2, 3, 4\}$, $T_2 = \{5, 6\}$, $L_1 = 4$ and $L_2 = 2$.

The computation of the coefficients of the quadratic inequality gives $A = -578.45$, $B = 75548.50$ and $C = -1.93909$. The roots of the corresponding quadratic equation are $n = r_1 = 35.10$ and $n = r_2 = 95.51$.

Then, for sample sizes lower or equal than 35, $V(\widehat{M}_{sp}) < V(\widehat{M}_{st})$, and for sample sizes greater than or equal to 36 the inequality reverses.

For the computation of values of Table 2 we recall that the sample size of the strata 1, 2, 3 and 4 are the number of units of the corresponding strata in the population, that is, $n_h = N_h$, for all $h = 1, 2, 3, 4$.

Table 1
Population strata characteristics of Example 1

	Strata					
	1	2	3	4	5	6
N_h	1	1	1	3	52	25
S_h	0	0	0	3.838	5.926	7.718
m_h	26	24	19	10.67	21.48	24.60

Table 2
Details of the allocation as a function of the size of the sample Example 1

n	Stratified sampling with Neyman allocation in the strata subset T_2							Simple random sampling	$V(\widehat{M}_{sp})/V(\widehat{M}_{st}) \times 100$
	n_1	n_2	n_3	n_4	n_5	n_6	$V(\widehat{M}_{st})$	$V(\widehat{M}_{sp})$	
8	1.00	1.00	1.00	3.00	1.23	0.77	122 236.3	36 503.9	29.86
12	1.00	1.00	1.00	3.00	3.69	2.31	38 535.2	16 305.1	74.81
20	1.00	1.00	1.00	3.00	8.61	5.39	14 620.7	12 265.3	83.89
25	1.00	1.00	1.00	3.00	11.68	7.32	9 900.7	9 033.5	91.24
30	1.00	1.00	1.00	3.00	14.76	9.24	7 147.3	6 879.0	96.24
33	1.00	1.00	1.00	3.00	16.60	10.40	5 984.8	5 899.6	98.58
34	1.00	1.00	1.00	3.00	17.22	10.78	5 658.7	5 611.6	99.17
35	1.00	1.00	1.00	3.00	17.83	11.17	5 343.4	5 340.0	99.94
36	1.00	1.00	1.00	3.00	18.45	11.55	5 054.8	5 083.5	100.57
37	1.00	1.00	1.00	3.00	19.06	11.94	4 784.8	4 840.9	101.17
38	1.00	1.00	1.00	3.00	19.68	12.32	4 531.7	4 611.0	101.75
40	1.00	1.00	1.00	3.00	20.91	13.09	4 070.1	4 185.8	102.84
50	1.00	1.00	1.00	3.00	27.06	16.94	2 391.6	2 569.9	107.45

The expression (5) giving the values of n for which stratified samples are more precise than simple random samples simplifies considerably when $L_2 = 1$, that is, when $n_h = N_h$ in all strata but one. In this case,

$$V(\widehat{M}_{st}) = N_2'(N_2' - n_2')\frac{S_2'^2}{n_2'} = N_2'(N - n)\frac{S_2'^2}{n_2'}, \tag{6}$$

where $S_2'^2$ denotes the variance of our variable in the population belonging to the strata subset T_2.

On the other hand, to see the possible inequalities between $V(\widehat{M}_{st})$ and $V(\widehat{M}_{sp})$ we will make use of the following parameter

$$H = \frac{N_1' N S^2}{N S^2 - N_2' S_2'^2}. \tag{7}$$

The result for this case can be stated by the following theorem.

Theorem 2. *Given a population of N units divided into L strata and a stratified sample of size n obtained in such a way that we have $n_h = N_h$ in $L - 1$ strata (the first $L - 1$, for instance), that is for all $h = 1, 2, \ldots, L - 1$. Then, we have two cases,*
(i) *Parameter (7) H is positive. Then, the variance of the unbiased estimator of the total of the variable, M, of the population $V(\widehat{M}_{st})$ is lower than the variance $V(\widehat{M}_{sp})$ obtained with a simple random sample of the same size n, if and only if, $n > H$.*
(ii) *Parameter (7) H is negative. In this case, $V(\widehat{M}_{sp}) < V(\widehat{M}_{st})$, for any sample size n.*

Proof. Dividing expressions (3) and (6) and using (7) we have, for $n < N$

$$\frac{V(\widehat{M}_{sp})}{V(\widehat{M}_{st})} = \frac{Nn_2' S^2}{N_2' n S_2'^2} = \frac{N(n-n_1')S^2}{N_2' n S_2'^2} > 1$$

if and only if $n > H$.

(i) Suppose $H > 0$. Then $V(\widehat{M}_{sp}) > V(\widehat{M}_{st})$, if and only if $n > H$.

(ii) Suppose $H < 0$. Then, the same inequality holds for $n > H$, but in this case $H < 0$, so, there is no value of n for which the precision of the stratified sample is greater than that of the simple sample. □

In addition, note that the above coefficient can be written as

$$\frac{V(\widehat{M}_{sp})}{V(\widehat{M}_{st})} = \frac{NS^2}{N_2' S_2'^2}\left[1 - \frac{n_1'}{n}\right].$$

Then, if we are in case (i) of the theorem where $V(\widehat{M}_{sp}) < V(\widehat{M}_{st})$, for $n < H$, as the size of the sample n increases, both variances will be closer, obtaining equality when $n = H$, and when $n > H$ the inequality will reverse becoming $V(\widehat{M}_{sp}) > V(\widehat{M}_{st})$.

Also notice that the maximum value of the ratio of variances $V(\widehat{M}_{sp})/V(\widehat{M}_{st})$ is: $\frac{S^2}{S_2'^2}\left[\frac{1-1/N_2'}{1-1/N}\right]$, which is given when $n = N - 1$. This effect is illustrated in the next example.

Example 2. Consider a population of $N = 10$ units divided into $L = 5$ strata, each stratum having the characteristics given in Table 3. Details of the allocation of the sample on the strata are given in Table 4.

Table 3
Population strata characteristics of Example 2

	Strata				
	1	2	3	4	5
N_h	1	1	1	3	52
S_h	0	0	0	3.838	5.926
m_h	26	24	19	10.67	21.48

This population has $S^2 = 38.435$ and mean $m = 21$. The subset $T_1 = \{1, 2, 3, 4\}$ and $T_2 = \{5\}$. On the other hand $N = 58$, $N_2' = 52$, $N_1' = 6$ and $S_2' = 35.117$, so the value of the parameter H is 33.18. The values of $V(\widehat{M}_{sp})$ and $V(\widehat{M}_{st})$ as a function of the sample size can be found in Table 4.

Note that for values of sample size $n > H$, that is, for sample sizes greater or equal to 34, the variance of the stratified sample is lower than that of the simple random sample. Note also that the maximum value of the ratio of variances is 1.0922, which is given when $n = 57$.

3. Conclusion

When the population is stratified into L strata and it is aimed to estimate the total (or of the mean) of a variable using unbiased estimators, the sample is required to contain at least one unit of each of the L strata. This implies that in cases of overstratification, that is, when stratification leads some of the strata of the population to contain only one unit, the sample size for these strata is also of one unit. This is not the only case where $n_h = N_h$. This equality may be achieved in the strata that satisfy the inequality $\frac{N_h S_h}{\sum N_h S_h} n \geq N_h$, when the Neyman method is used for the allocation of the sample, as well as in the case of including a priori all the population units of a stratum in the sample.

In these conditions, even when using the optimal allocation of the sample in the remaining strata, that is, in the strata where $n_h \neq N_h$, the variance of the estimator of the total of a variable from the stratified population may be

Table 4
Details of the allocation as a function of the size of the sample Example 2

n	Stratified sampling						Simple random sampling	$V(\widehat{M}_{sp})/V(\widehat{M}_{st}) \times 100$
	n_1	n_2	n_3	n_4	n_5	$V(\widehat{M}_{st})$	$V(\widehat{M}_{sp})$	
8	1.00	1.00	1.00	3.00	2.00	45 652.7	13 932.8	30.52
12	1.00	1.00	1.00	3.00	6.00	14 000.2	8 545.4	61.04
16	1.00	1.00	1.00	3.00	10.00	7 669.7	5 851.8	76.30
20	1.00	1.00	1.00	3.00	14.00	4 956.6	4 235.6	85.45
25	1.00	1.00	1.00	3.00	19.00	3 171.7	2 942.6	92.78
30	1.00	1.00	1.00	3.00	24.00	2 130.5	2 080.6	97.66
33	1.00	1.00	1.00	3.00	27.00	1 690.8	1 688.8	99.88
34	1.00	1.00	1.00	3.00	28.00	1 565.2	1 573.6	100.53
35	1.00	1.00	1.00	3.00	29.00	1 448.3	1 464.9	101.15
36	1.00	1.00	1.00	3.00	30.00	1 339.2	1 362.3	101.73
37	1.00	1.00	1.00	3.00	31.00	1 237.0	1 265.2	102.28
38	1.00	1.00	1.00	3.00	32.00	1 141.3	1 173.3	102.80
40	1.00	1.00	1.00	3.00	34.00	966.8	1 003.2	103.77
50	1.00	1.00	1.00	3.00	44.00	332.0	356.7	107.43
55	1.00	1.00	1.00	3.00	49.00	111.8	121.6	108.76
56	1.00	1.00	1.00	3.00	50.00	73.0	79.6	109.01
57	1.00	1.00	1.00	3.00	51.00	35.8	39.1	109.22

lower than the one obtained from the unstratified population by simple random sampling only for sample sizes greater than a determined value.

In the case of using a geographical stratification criterion, the precision of the estimates gained by stratification is in general unimportant [1, p. 102]. In consequence for this case, when we have some strata whose $n_h = N_h$, the sample size for which the stratified sampling generates estimators more precise than the simple sampling may be high.

Acknowledgements

The work of the first three authors was supported by the Commission of the EU (GENEDEC project FP6-502184) and that of the last author was supported by Spanish grant DGI (FEDER) number AGL2004–03263, Grupos03/062 and Universitat Politècnica de València under its research incentive program.

References

[1] W.D. Cochran, Sampling Techniques, John Wiley, New York, 1977.
[2] L. Judez, C. Chaya, Effects of geographical stratification in a farm accountancy data network on the accuracy of the estimates, Journal of Agricultural Economics 50 (3) (1999) 388–399.
[3] Commission of the European Communities, Farm Accountancy Data Network: an A–Z of methodology, Luxembourg: Office for Official Publications of the European Communities, 1989.
[4] F. Arfini, Mathematical programming models employed in the analysis of the Common Agricultural Policy, Working Paper no. 9. Istituto Nazionale di Economia Agraria, 2001.
[5] A. Barkaoui, J.P. Butault, J.M. Rousselle, Positive Mathematical Progamming and Agricultural Supply within EU under Agenda 2000, in: T. Heckelei, H.P. Witzke, W. Henrishmeyer, Agricultural Sector Modelling and Policy Information Systems (Proceedings of the 65th EAAE Seminar, 29–31 March, 2000 at Bonn University), Vauk Verlag, Kiel: 200, 2001.
[6] S. De Cara, P.A. Jayet, Emissions of greenhouse gases from agriculture: the heterogeneity of abatement costs in France, European Review of Agricultural Economics 27 (3) (2000) 281–303.
[7] T. Heckelei, W. Britz, Models based on Positive Mathematical Programming: State of the Art and Further Extensions 89th EAAE Seminar on “Modelling agricultural policies: state of the art and new challenges”, Parma, 3–5 February, 2005.
[8] S. Huettel, B. Kuepker, W. Kleinhanss, F. Offermann, Assessing the 2003 CAP Reform Impacts on German Agriculture using the Farm Group Model FARMIS, in: 89th EAAE Seminar on “Modelling agricultural policies: state of the art and new challenges”, Parma, 3–5 February, 2005.
[9] L. Judez, C. Chaya, S. Martínez, A. González, Effects of the measures envisaged in “Agenda 2000” on arable crop producers and beef and veal producers: an application of Positive Mathematical Programming to representative farms of a Spanish region, Agricultural Systems 67 (2001) 121–138.