Designing household samples in Brazil using the 2010 census enumeration area frame

Sâmela Batista Arantes [1,3] , Pedro Luis do Nascimento Silva[1,2]
[1]Brazilian Institute of Geography and Statistics, BRAZIL
[2]National School of Statistical Science, BRAZIL
[3]Corresponding author: Sâmela Batista Arantes. E-mail: samela.arantes@ibge.gov.br

Abstract

Complex sample designs are used in most of the household sample surveys conducted by the Brazilian Institute of Geography and Statistics. One of the key steps of sample design comprises the sample size determination, possibly for several variables of interest. This paper presents an approach for calculating sample sizes using design effects (DEFF) and intraclass correlation coefficients estimated for four different sampling strategies using data from the census 2010 enumeration area frame. All four sampling strategies comprise two stage cluster sampling, where in the first stage census tracts are the primary sampling units (PSUs) and households are selected in the second stage. The results indicate that PPS designs for sampling the PSUs are generally more efficient (i.e. lead to smaller sample sizes) because they take advantage of the correlation between target variables and the size of the PSUs. Pareto PPS sampling design (without replacement sampling of PSUs) is more efficient than with replacement PPS sampling of PSUs.

Key Words: design effect, Household sample survey, Pareto sampling, sample size

1 Introduction

The main motivation for this work is the development of tools for the calculation of sample sizes for complex household sample surveys conducted by the Brazilian Institute of Geography and Statistics (IBGE) and others using the census frames provided by IBGE. Designing and sampling for household surveys in Brazil relies heavily on the area frames prepared for and updated by the latest Population Census carried out in the country, since there are no address or household registers of sufficient quality to sample directly. The Census provides population parameters and updated census tracts (enumeration areas). Census tracts are used as primary sampling units (PSUs) in most household surveys carried out by IBGE and in many other surveys conducted by private and non-governmental organizations.

Design effects (DEFFs, Kish, 1965) are useful tools to assist with the calculation of sample sizes for complex sample designs. Some DEFFs are obtained in terms of the intraclass correlation coefficients which measure the homogeneity in the clusters or PSUs.

2 Sampling strategies considered

Cluster sampling is useful to minimize the cost of reaching the sampled elementary units, when compared to unclustered sample designs. For two-stage cluster sampling the elementary units are arranged in groups, and the sample is obtained in two stages: first groups are sampled and then elementary units are selected within the groups selected in the first stage. Four two-stage cluster sampling strategies were considered:

- Strategy 1 (S1) comprises selecting census tracts (stage 1) and households (stage 2) by Simple Random Sampling Without Replacement (SRSWOR) combined with the Horvitz-Thompson estimator;
- Strategy 2 (S2) comprises selecting census tracts and households by SRSWOR

combined with a ratio estimator;

- Strategy 3 (S3) comprises selection of census tracts with probability proportional to size with replacement (PPSR) and SRSWOR selection of households, combined with the natural estimator;
- Strategy 4 (S4) comprises selection of census tracts with Pareto PPS sampling (Rosén, 1997) and SRSWOR selection of households, combined with the Horvitz-Thompson.

The sampling strategies S1, S2 and S3 were studied by Silva and Moura (1990). Here we also considered S4 because this is the method of selection adopted for the selection of the newly designed master sample of the Integrated Household Sample Survey called SIPD implemented by IBGE in 2012 (Freitas et al, 2007).

Some common features of the sampling strategies used by a number of IBGE's household sample surveys are geographical stratification, two stage clustering with census tracts as the PSUs, PPS sampling for selection of the PSUs, and calibration adjustment of survey weights to estimated population totals. These features are presently found in the Monthly Labour Force Survey called PME, the Integrated Household Sample Survey (SIPD) and the Family Budget Survey called POF. Hence the strategies S3 and S4 closely resemble strategies which are in common use at IBGE, with the exception of the weight calibration mentioned here.

Let $\mathrm{U} = \{1,2,\dots,\mathrm{i},\dots,\mathrm{M}\}$ represent the population of PSUs, where M is number of PSUs in the population. Let $U_i=\{1,2,...,j,...,N_i\}$ denote the set of population units in PSU I, where $\mathrm{N_i}$ is the number of units in PSU i, with $\mathrm{N} = \sum_{\mathrm{i}\in\mathrm{U}} \mathrm{N_i}$ denoting the total number of units in the population. Denote by $\mathrm{y_{ij}}$ the value of the variable of interest for unit of the PSU i. Let $\mathrm{Y_i} = \sum_{\mathrm{j}\in 1}^{\mathrm{N_i}} \mathrm{y_{ij}}$ be the total of the variable y for PSU i and $\mathrm{Y} = \sum_{\mathrm{i}\in\mathrm{U}} \mathrm{Y_i}$ be the overall total for y in the population. Denote by $\bar{\mathrm{Y}}_\mathrm{i} = \mathrm{Y_i}/\mathrm{N_i}$ the average of y in PSU i, $\bar{\mathrm{Y}}_\mathrm{C} = \mathrm{Y}/\mathrm{M}$ be the average cluster total, and $\bar{\bar{\mathrm{Y}}} = \mathrm{Y}/\mathrm{N}$ the population average per unit.

Let $\mathrm{s} = \{\mathrm{i_1},\mathrm{i_2},\dots,\mathrm{i_m}\}, \mathrm{s} \subset \mathrm{U}$, be the set of PSUs selected from the population U. Let m be the size of the first stage sample, and $\mathrm{n_i}$ denote the number of units selected in PSU i; with $\mathrm{n} = \sum_{\mathrm{i}\in\mathrm{s}} \mathrm{n_i}$ denoting the overall sample size.

Let $\pi_\mathrm{i} = \Pr(\mathrm{i} \in \mathrm{s})$ be the probability of inclusion of PSU i in the sample s. Let $\mathrm{a_i} = \{\mathrm{j_1},\mathrm{j_2},\dots,\mathrm{j_{n_i}}\}, \mathrm{a_i} \subset \mathrm{U_i}$ denote the set of units sampled in PSU i, and let $\mathrm{A} = \bigcup_{\mathrm{i}\in\mathrm{s}} \mathrm{a_i}$ represent the overall sample of units.

The estimators of totals and their variances under strategies S1 to S4 are described in Arantes (2012). For strategies S1 to S3, estimators were available in Silva and Moura (1990). For S4 we developed the Horvitz-Thompson estimator of the total and its variance considering properties described in Cochran (1977), as:

$$\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{S4}}) = \frac{\mathrm{M}}{\mathrm{M}-1}\left\{\sum_{\mathrm{i}\in\mathrm{U}} \frac{\mathrm{Y_i^2}}{\pi_\mathrm{i}}(1-\pi_\mathrm{i}) - \frac{[\sum_{\mathrm{i}\in\mathrm{U}} \mathrm{Y_i}(1-\pi_\mathrm{i})]^2}{\mathrm{m}-\sum_{\mathrm{i}\in\mathrm{U}} \pi_\mathrm{i}^2}\right\} + \sum_{\mathrm{i}\in\mathrm{U}} \frac{1}{\pi_\mathrm{i}} \mathrm{N_i^2} \frac{1}{\mathrm{n_i}} - \frac{1}{\mathrm{N_i}} \mathrm{S_i^2} \qquad (1)$$

where $S_i^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(y_{ij} - \overline{Y}_i)^2$.

4 Design effects and intraclass correlation coefficients

The design effect (DEFF) is a statistical measure of the efficiency of a complex sample design compared to a design using simple random sampling (SRS). It is defined as the ratio of the variance of an estimator under the complex sampling design and the variance of an estimator under SRS, assuming equal sample sizes for both numerator and denominator. According to Silva and Moura (1990), the intraclass correlation coefficient is a “measure of the degree of homogeneity between the units belonging to certain groups of the population”. Under certain two-stage cluster sampling strategies, it can be shown (see Silva and Moura, 1990) that:

$$\mathrm{DEFF}(\hat{\mathrm{Y}}_{\mathrm{CS2}}) = \frac{\mathrm{V}_{\mathrm{CS2}}(\hat{\mathrm{Y}})}{\mathrm{V}_{\mathrm{SRS}}(\hat{\mathrm{Y}})} \approx 1 + (\bar{\mathrm{n}} - 1)\rho \qquad (2)$$

CS2 represents a two-stage cluster sampling design and SRS represents simple random sampling, respectively, $\bar{\mathrm{n}}$ is the average number of units selected per PSU and ρ is a properly defined intraclass correlation coefficient.

4.1 Design effect for the Strategy S1

For the sampling strategy S1, the design effect can be approximated by a function of the intraclass correlation coefficient as follows:

$$\mathrm{DEFF}(\hat{\mathrm{Y}}_{\mathrm{S1}}) = \frac{\mathrm{V}_{\mathrm{S1}}(\hat{\mathrm{Y}}_{\mathrm{S1}})}{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{SRS}})} \approx \frac{\sigma_{\mathrm{a}}^2}{\sigma_{\mathrm{b}}^2}[1 + \rho_{\mathrm{e}}(\bar{\mathrm{n}} - 1)] \qquad (3)$$

where $\sigma_{\mathrm{a}}^2 = \frac{\sigma_{\mathrm{e}}^2}{\bar{\mathrm{N}}^2} + \frac{\bar{\mathrm{N}}-1}{\bar{\mathrm{N}}}\frac{\mathrm{M}}{\mathrm{N}}\sigma_{\mathrm{d}}^2$, $\sigma_{\mathrm{b}}^2 = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}\in\mathrm{U}}\mathrm{N}_{\mathrm{i}}\sigma_{\mathrm{i}}^2 + \frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}\in\mathrm{U}}\mathrm{N}_{\mathrm{i}}(\bar{\mathrm{Y}}_{\mathrm{i}} - \bar{\bar{\mathrm{Y}}})^2$, $\rho_{\mathrm{e}} = 1 - \frac{\bar{\mathrm{N}}\sigma_{\mathrm{d}}^2}{\sigma_{\mathrm{e}}^2 + (\bar{\mathrm{N}}-1)\sigma_{\mathrm{d}}^2}$ with $\sigma_{\mathrm{e}}^2 = \frac{1}{\mathrm{M}}\sum_{i\in U}(Y_i - \bar{Y}_C)^2$ and $\sigma_{\mathrm{d}}^2 = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}\in\mathrm{U}}\frac{\mathrm{N}_{\mathrm{i}}^2}{\mathrm{N}_{\mathrm{i}}-1}\sigma_{\mathrm{i}}^2$, $\bar{N}$ is the average PSU size and $\sigma_{\mathrm{i}}^2 = \frac{1}{\mathrm{N}_i}\sum_{j\in U_i}(y_{ij} - \bar{Y}_i)^2$.

Note that the intraclass correlation coefficient ρ_{e} is a function of the variance between and within the clusters, and if the variance within the clusters is null this coefficient takes the value 1.

4.2 Design effect for the Strategy S2

For the sampling strategy S2, where the estimator of total is the ratio estimator we have the DEFF is approximated by the following expression:

$$\mathrm{DEFF}(\hat{\mathrm{Y}}_{\mathrm{S2}}) = \frac{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{S2}})}{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{SRS}})} \approx \frac{\sigma_{\mathrm{c}}^2}{\sigma_{\mathrm{b}}^2}[1 + \rho_{\mathrm{f}}(\bar{\mathrm{n}} - 1)] \qquad (4)$$

where $\sigma_c^2 = \frac{1}{\bar{\mathrm{N}}^2}\frac{1}{\mathrm{M}}\sum_{\mathrm{i}\in\mathrm{U}}(\mathrm{Y}_{\mathrm{i}} - \mathrm{N}_{\mathrm{i}}\bar{\bar{\mathrm{Y}}})^2 + \frac{\bar{\mathrm{N}}-1}{\bar{\mathrm{N}}}\frac{1}{\mathrm{N}}\sum_{\mathrm{i}\in\mathrm{U}}\frac{\mathrm{N}_{\mathrm{i}}^2}{\mathrm{N}_{\mathrm{i}}-1}\sigma_{\mathrm{i}}^2$, $\rho_{\mathrm{f}} = 1 - \frac{\bar{\mathrm{N}}\sigma_{\mathrm{d}}^2}{\sigma_{\mathrm{e},2}^2 + (\bar{\mathrm{N}}-1)\sigma_{\mathrm{d}}^2}$ with $\sigma_{\mathrm{e},2}^2 = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}\in\mathrm{U}}(\mathrm{Y}_{\mathrm{i}} - \mathrm{N}_{\mathrm{i}}\bar{\bar{\mathrm{Y}}})^2$.

4.3 Design effect for the Strategy S3

For strategy S3 we have the simpler expression of the DEFF as a function of the intraclass correlation coefficient:

$$\mathrm{DEFF}\left(\hat{\mathrm{Y}}_{\mathrm{S3}}\right) = \frac{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{S3}})}{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{SRS}})} = 1 + (\bar{\mathrm{n}} - 1)\rho_{\mathrm{c}} \qquad (5)$$

where $\rho_{\mathrm{c}} = 1 - \frac{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}\in\mathrm{U}}\mathrm{N}_{\mathrm{i}}\sigma_{\mathrm{i}}^2\left(1+\frac{1}{\mathrm{N}_{\mathrm{i}}-1}\right)}{\sigma^2}$ where $\sigma_y^2 = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}\in\mathrm{U}}\sum_{j\in U_i}\left(y_{ij} - \bar{\bar{Y}}\right)^2$.

4.4 Design effect for the Strategy S4

The design effect for the strategy S4 is obtained by ratio of variances:

$$\mathrm{DEFF}(\hat{\mathrm{Y}}_{\mathrm{S4}}) = \frac{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{S4}})}{\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{SRS}})} \qquad (6)$$

where $\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{SRS}}) = \mathrm{N}^2\left(\frac{1}{\mathrm{n}} - \frac{1}{\mathrm{N}}\right)\mathrm{S}_{\mathrm{y}}^2$, $\mathrm{S}_{\mathrm{y}}^2 = \frac{N}{N-1}\sigma_y^2$ and $\mathrm{V}(\hat{\mathrm{Y}}_{\mathrm{S4}})$ is given in (1).

5 Design effects and intraclass correlation coefficients for 2010 Population Census

The data used in this study refer to the 2010 Population Census microdata for the State of Minas Gerais in Brazil. The 2010 Population Census used a two-questionnaire data collection approach, similar to the one adopted in the US and Canada. Households sampled within each census tract were interviewed using a questionnaire containing a number of items such as housing characteristics, demographics, education, religion, occupation, income, etc. Non-sampled households were interviewed using a shorter questionnaire, containing only a few questions on the household and its members.

Assuming that the sample of households within each census tract was selected by SRSWOR, the Census sample provides unbiased estimates $\bar{y}_i = \frac{1}{n_i} \sum_{j \in a_i} y_{ij}$ and $s_i^2 = \frac{1}{n_i - 1} \sum_{j \in a_i} (y_{ij} - \bar{y}_i)^2$ for the mean $\bar{Y}_i$ and variance S_i^2 within each census tract, respectively. These estimates may be used to obtain consistent estimates for the various design effects and intraclass correlation coefficients described above (see Silva and Moura, 1990, and Arantes, 2012, for details). For characteristics in the short form the intraclass correlation and design effects can be calculated using the expressions provided.

Design effects and intraclass correlation coefficients for the four sampling strategies for Minas Gerais are provided in table 1 for short form variables, and in table 2 for sample or long form variables, in both cases considering $\bar{n}$=14. This choice for the second stage sample size was adopted because this is the number of households selected per census tract for the IBGE integrated household sample survey.

In both tables, we observed the following trends for the intraclass correlation coefficients $\rho_c < \rho_f < \rho_e$ and for the design effects $\mathrm{DEFF}_{S4} < \mathrm{DEFF}_{S3} < \mathrm{DEFF}_{S2} \ll \mathrm{DEFF}_{S1}$, where this last symbol << means “much smaller than”.

Table 1: Intraclass correlation coefficients and design effects for some 2010 Census private permanent household variables for Minas Gerais, by sampling strategy

Variable	Description	S1		S2		S3		S4
		ρe	DEFF_{S1}	ρf	DEFF_{S2}	ρc	DEFF_{S3}	DEFF_{S4}
U2	Number of illiterate residents aged 10 or older	0,13	2,66	0,12	2,63	0,12	2,52	2,46
U9	Number of female heads of household	0,21	4,48	0,08	2,05	0,07	1,94	1,90
U17	Number of households with electricity	0,98	509,81	0,09	2,05	0,13	2,63	2,57
U18	Number of single person households	0,05	1,73	0,02	1,28	0,02	1,23	1,21
U20	Number of residents in households with per capita income below half the minimum wage	0,24	4,52	0,19	3,55	0,16	3,14	3,05

Table 2: Intraclass correlation coefficients and design effects for some 2010 Census private permanent household sample variables for Minas Gerais, by sampling strategy

Variable	Description	S1		S2		S3		S4
		ρe	DEFF$_{S1}$	ρf	DEFF$_{S2}$	ρc	DEFF$_{S3}$	DEFF$_{S4}$
A5	Number of residents unnatural UF in permanent household	0,32	6,14	0,22	3,96	0,18	3,38	3,27
A10	Number of live births to women aged 15 or more in DPP	0,20	4,20	0,06	1,78	0,05	1,69	1,67
A12	Total monthly income of	0,20	3,93	0,16	3,21	0,14	2,82	2,74
A13	Permanent households	0,84	65,33	0,11	2,36	0,11	2,45	2,40

6 Sample size calculations

With two-stage cluster sampling designs for household surveys, possibly using unequal probabilities of selection, calculation of sample sizes required to attain a specified accuracy for an estimator may be achieved by the following procedure (Silva & Moura, 1990; Silva, 2002).

Step 1) Choose a parameter to be estimated (total, average, ratio); in this paper we consider only the case of population totals.

Step 2) Define the maximum margin of error for estimating the parameter of interest.

Step 3) Calculate the sample size $\mathrm{n}_{\mathrm{SRSWOR}}$ required to attain such precision assuming SRSWOR of households.

Step 4) Define the sample take in each selected PSU, that is, establish ñ.

Step 5) Considering the value of ñ defined in 4) estimate the design effect for the two stage design using some previous Census or survey.

Step 6) Calculate the total sample size required for the two stage design that yields the same accuracy as

$$\mathrm{n}_{\mathrm{CS2}} \approx \mathrm{n}_{\mathrm{SRSWOR}} \times \mathrm{DEFF}_{\mathrm{CS2}} \tag{7}$$

Step 7) Calculate the PSU sample size using

$$\mathrm{m}_{\mathrm{CS2}} = \mathrm{n}_{\mathrm{CS2}}/\tilde{\mathrm{n}}. \tag{8}$$

7 Concluding Remarks

Our results show that there are large differences between the sampling strategies considered. SRSWOR sampling of PSUs coupled with the Horvitz-Thompson estimator can be disastrous for variables which are highly clustered, such as the indicator that households have electricity. The strategies that sample PSUs with probability proportional to size (PPS) showed the best performance (smaller DEFFs) for most variables. Sampling strategy S4 which considers Pareto PPS (without replacement) sampling of census tracts yielded smaller DEFFs than all the other strategies considered.

These results provide support for users who wish to work with some of the sampling strategies considered. Arantes (2012) developed R language scripts that calculate all the above intraclass correlations, design effects and sample sizes using data from the Brazilian 2010 census microdata. These may be easily adapted to other data sources, and thus help those planning household sample surveys where a two-stage cluster sampling design is adopted.

8 References

ARANTES, B. S. (2012). Planejamento de pesquisas domiciliares no Brasil utilizando a malha setorial do Censo Demográfico 2010. Rio de Janeiro: Escola Nacional de Ciências Estatísticas, MSc. dissertation.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley.

FREITAS, M. P. S. de, et. al. (2007). *Amostra Mestra para o Sistema Integrado de Pesquisas Domiciliares*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Texto para discussão número 23.

HAGGARD, E.A. (1958) Intraclass correlation and the analysis of variance. New York, Dryden Press. 171p..

IBGE (2011). B*ase de informações do Censo Demográfico 2010*: resultados da Sinopse por setor censitário. Rio de Janeiro: IBGE.

KISH, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

ROSÉN, B. (1997). *On sampling with probability proportional to size*. Journal of Statistical Planning an Inference, nº 62, 159-191.

SILVA, P. L. N., MOURA, F. A. S. (1990). *Efeito de conglomeração da malha setorial do Censo Demográfico 1980*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Texto para discussão número 32.