

Sampling Design for the National Health Survey, Brazil 2013

doi: 10.5123/S1679-49742015000200003

Paulo Roberto Borges de Souza-Júnior

Fundação Instituto Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Rio de Janeiro-RJ, Brasil

Marcos Paulo Soares de Freitas

Instituto Brasileiro de Geografia e Estatística, Coordenação de Métodos e Qualidade, Rio de Janeiro-RJ, Brasil

Giuseppe de Abreu Antonaci

Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas, Rio de Janeiro-RJ, Brasil

Célia Landmann Szwarcwald

Fundação Instituto Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Rio de Janeiro-RJ, Brasil

Abstract

This paper describes the sample design used in the Brazilian 2013 National Health Survey. The target population was comprised of people resident in permanent private households throughout the country. The survey was household-based with stratified sampling and three clustering stages: census tracts form the primary sampling units; households are the units of second stage; and adults (aged 18 years or older) define the third-stage units. The sample size considered the desired level of precision for the estimates of some indicators at different levels of disaggregation and different population groups. The final weighting was a product of inverse selection probabilities at each stage of the sampling plan, including non-response correction procedures and adjustment calibrations for the known population totals. Since this is a complex sample, appropriate procedures must be used during data processing.

Key words: Population Surveys; Cluster Sampling; Sample Size.

Correspondence:

Paulo Roberto Borges de Souza-Júnior – Fundação Instituto Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde/Ministério da Saúde, Av. Brasil, nº 4365, Pavilhão Haity Moussatché, Manguinhos, Rio de Janeiro-RJ, Brasil. CEP: 21040-360. E-mail: paulo.borges@icict.fiocruz.br.

Introduction

The National Health Survey (PNS), conducted between 2013 and 2014, was developed in partnership between the Ministry of Health's Health Surveillance Secretariat (SVS/MS), the Oswaldo Cruz Foundation (Fiocruz) and the Brazilian Institute of Geography and Statistics (IBGE). PNS is a nationwide population-based survey, and its main objective is (i) to produce national-level data about the health status and lifestyles of the Brazilian population, (ii) and also data about health care, regarding access, use of health services, preventive actions, continuity of care, and health care funding.¹

PNS, as part of the IBGE Integrated Household Surveys System (SIPD), used the sample infrastructure built for this system.² This sample planning stage of the research was also conducted by IBGE, in partnership with Fiocruz.

This article describes the sampling plan used in PNS, including the target population, the selection stages, the calculation of sample size and the definition of expansion factors.

Target Population

The target population was comprised of people living in permanent private households (PPH) located in the survey's geographical coverage area. The PPHs are those that were built to be used exclusively for housing and are intended to serve as a home to one or more people.³

The 'geographic coverage' area of the survey was defined as the entire national territory, divided into the census tracts of the 2010 Geographic Operating Base. Areas with special features and sparsely populated were excluded, these being classified by IBGE on the basis of tracts, such as indigenous villages, barracks, military bases, lodgings, camp sites, boats, penitentiaries, penal colonies, jails, nursing homes, orphanages, convents and hospitals. Census tracts located in indigenous lands were also excluded.

Questionnaire

The PNS questionnaire was stratified into three parts, as follows:

Part 1 – Household

Questions about household information and about visits to the household made by the Family Health team and endemic disease health workers.

This first part was answered by the head of the household or person who had this information at the time of the interview.

Part 2 – Household residents

Questions related to the general characteristics of all household residents, including education level, work, income, disabilities, health insurance coverage, use of health services, health of the elderly, mammography coverage and characteristics of children under 2 years old.

The second part was answered by all household members and in the case of a resident being absent or incapable of answering, the head of the household could answer the questionnaire (or part of it) on behalf of the absent or incapable resident, in the same way as is done in the Brazilian National Household Survey (PNAD). If this happened it was recorded on the questionnaire by making a note of which household resident replied, using an identification number.

Part 3 – Individual

Questions asked of a randomly selected resident, aged 18 years or older, regarding other work and social support characteristics, self-perception of health status, accidents and violence, lifestyles, chronic diseases, women's health, prenatal, dental and medical care.

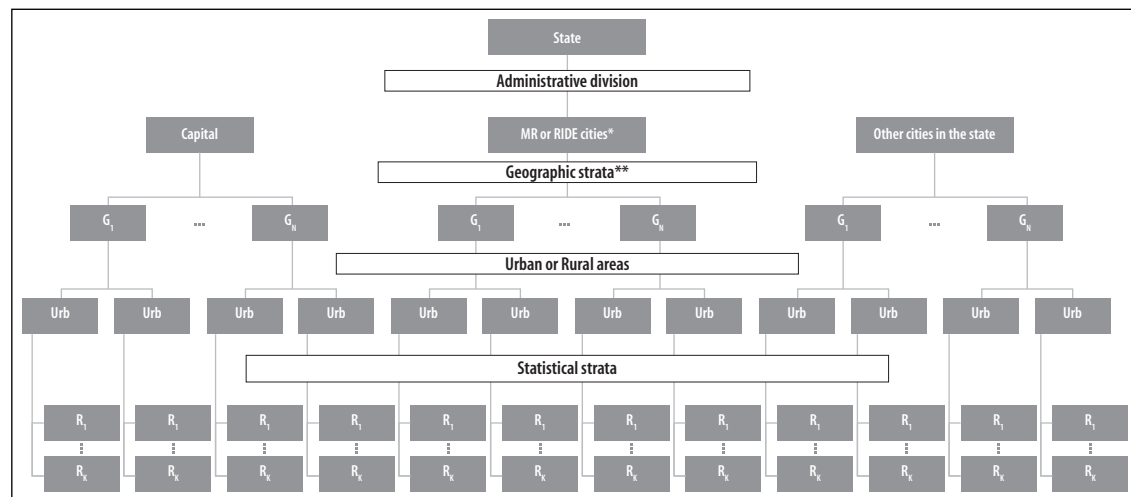
In this part, an adult (≥ 18 years old) was randomly selected among eligible household residents. This part could only be answered by the selected resident and other household residents could not answer on their behalf.

Sampling Plan

The PNS sample was a subsample of the Master Sample of the Integrated Household Surveys System (SIPD). Before describing the PNS sampling plan, we will give a brief description of the Master Sample. For more information about the Master Sample or about SIPD, see Freitas *et al*⁴ and Freitas and Antonaci.⁵

Master Sample

The Master Sample is a group of units of areas selected for use by various studies. These units are considered to be primary sampling units (PSUs) when planning



* MR: Metropolitan Region; RIDE: Integrated Economic Development Region.

** Subdivides the capitals and other big cities into more strata, such as district, sub-district and neighborhoods, for example.

Figure adapted from Freitas et al.

Figure 1 – Master Sample stratification of the Integrated Household Surveys System, used by the National Health Survey, Brazil, 2013

the samples of each of the studies that use the Master Sample, including the National Health Survey (PNS). The PSUs are census tracts or groups of census tracts (when there are few households).

For the purposes of selecting the Master Sample, the PSUs were stratified according to four different criteria (Figure 1):

1. Administrative

This stratifies PSUs by state and, within each state, subdivides PSUs into the state capital city, its Metropolitan Region (MR) or Integrated Economic Development Region (RIDE) (in those states that have these administrative divisions), and the rest of municipalities in the state.

2. Geographic

This subdivides the state capital cities and other big cities into additional strata, such as district, sub-district and neighborhoods, for example.

3. Area situation

This subdivides the geographical strata into urban and rural areas.

4. Statistics

This subdivides the strata above (urban and rural) into homogeneous strata by total household income and total permanent private households (PPH), in order to improve the accuracy of the estimates.

The selection of PSUs within each stratum was carried

out by probability proportional to size sampling (PPS), whereby the number of PPHs was used to measure the size of the PSU.

PNS sampling design

PNS is a household survey and the sample design applied was cluster sampling in three selection stages, with PSU stratification. As part of SIPD, in the first stage PSU selection was obtained by simple random sampling (SRS) among those previously selected for the Master Sample, maintaining the stratification of PSUs used in the Master Sample, as described above.

In the second stage, a fixed number of PPHs in each PSU selected in the first stage was selected by SRS. Household selection was done by using the updated National Address List for Statistical Purposes (CNEFE). It is important to highlight that, for the information contained in the first and second part of the survey, *i.e.*, the parts concerning household characteristics and the set of all household residents, the PNS sampling plan had only these two selection stages.

In the third stage, within each household in the sample, a resident aged 18 or older was selected, also by SRS, to answer the individual survey (part 3). The selection was made from a list of eligible residents compiled during the interview.

Calculating the sample size

The sample size was defined based on the desirable level of accuracy, with 95% confidence intervals (95%CI) to estimate some indicators (or parameters) of interest at different levels of geographical breakdown and specific population groups. To make it possible to set the sample size in these specific population groups, it was necessary to evaluate the proportion of PPHs that had people in these groups, based on 2010 Census data. The desirable level of accuracy for each indicator was based on the extent of the 95% confidence interval expected. From there, we calculated the standard error and the coefficient of variation (CV) desirable for each indicator, and CV was used as a measure of precision in the calculation of sample size.

The initial sample size calculations are based on simple random sampling, it being necessary to consider the sample design effect (DEFF), which indicates how much the cluster sampling (CS), which is the case of PNS, is less effective than SRS. Because of this, the initial sample sizes were increased in order to achieve the same accuracy. The DEFFs used in PNS sample size calculations were estimated based on data from the 2008 Brazilian National Household Survey (PNAD, 2008).

The indicators used for determining the sample size, its respective population groups, the expected values, desirable coefficients of variation, the estimated DEFFs and sizes initially calculated for the household sample, are shown in Table 1.

The formulas used for sample size calculation in each domain are presented below:

Table 1 – Population group, proportion of households with residents in the population group, expected value, coefficient of variation, estimated effect of the sample design, and initial sample size according to the indicator of interest – National Health Survey, Brazil, 2013

Indicator	Population Group	Proportion of households with people in the population group - Census 2010	Expected value of the indicator (%)	Desirable CV (%)	Estimated DEFF	Size of the sample of households
1. Prevalence of diabetes among adults aged 35 or more	Adults 35 years and more selected	60.0	8.0	12.8	1.4	1,641
2. Prevalence of hypertension among adults aged 35 or more	Adults 35 years and more selected	60.0	30.0	8.5	1.6	856
3. Prevalence of depression among adults aged 35 or more	Adults 35 years and more selected	60.0	8.0	12.8	1.9	1,425
4. Hospitalization rate in the previous year	All household residents	100.0	7.0	14.6	2.9	546
5. Use of health services in the previous 2 weeks	All household residents	100.0	14.0	7.3	4.6	1,602
6. Private health insurance coverage	All household residents	100.0	26.0	5.9	10.4	1,447
7. Coverage of preventive screening of cervical cancer among women aged 25-59	Women aged 25-59 years selected	40.0	80.0	3.2	2.2	1,490
8. Coverage of mammography among women aged 50 or more	Women aged 50-69 years selected	10.0	65.0	5.5	1.8	1,671
9. Prevalence of smoking	Selected adults	100.0	17.0	9.0	1.8	1,084
10. Prevalence of overweight and obesity	Selected adults	100.0	12.0	12.8	1.8	811
11. Prevalence of sedentary lifestyle	All household residents	100.0	15.0	10.2	4.1	672
12. Alcohol abuse	Selected adults	100.0	7.0	14.6	1.8	1,125
13. Proportion of who suffered violence with injury	Selected adults	100.0	2.0	25.5	2.8	1,205
14. Proportion of elderly with limitations (functionality issues)	All household elderly	40.0	15.0	13.6	2.1	1,786

CV: Coefficient of variation; DEFF: Effect of Sample Design

Table 2 – Selected sample size according to areas of interest. National Health Survey, 2013.

Area (Region and State)	Number of PSUs selected in the sample	Estimated number of households with interview conducted	Number of households selected in the sample
North	1,161	13,578	17,553
Rondônia	129	1,806	2,322
Acre	180	1,800	2,340
Amazonas	245	2,594	3,365
Roraima	124	1,736	2,232
Pará	246	2,652	3,438
Amapá	107	1,498	1,926
Tocantins	130	1,492	1,930
Northeast	1,916	19,160	24,908
Maranhão	181	1,810	2,353
Piauí	180	1,800	2,340
Ceará	290	2,900	3,770
Rio Grande do Norte	179	1,790	2,327
Paraíba	182	1,820	2,366
Pernambuco	279	2,790	3,627
Alagoas	180	1,800	2,340
Sergipe	179	1,790	2,327
Bahia	266	2,660	3,458
Southeast	1,451	14,510	18,863
Minas Gerais	366	3,660	4,758
Espírito Santo	181	1,810	2,353
Rio de Janeiro	365	3,650	4,745
São Paulo	539	5,390	7,007
South	767	7,670	9,971
Paraná	287	2,870	3,731
Santa Catarina	186	1,860	2,418
Rio Grande do Sul	294	2,940	3,822
Central-West	774	7,740	10,062
Mato Grosso do Sul	179	1,790	2,327
Mato Grosso	162	1,620	2,106
Goiás	253	2,530	3,289
Distrito Federal	180	1,800	2,340
Brazil	6,069	62,658	81,357

*Including, approximately, 23% of estimated losses

PSU: Primary Sample Unit

$$n_{AAS} = \frac{N}{N-1} * P * Q * \frac{I}{CV^2 * P^2 + \frac{P * Q}{N-1}}$$

$$n_{AC} = n_{AAS} * EPA$$

Where,

n_{AAS} is the sample size of people under simple random sampling;

N is the total number of people living in PPHs in the domain;

P is the proportion of interest;

$Q = 1 - P$;

CV is the coefficient of variation of the estimated desirable ratio;

n_{AC} is the sample size in cluster sampling;

EPA is the sample design effect (DEFF).

The data used in the calculations were obtained from the 2010 Census and PNAD 2008.

Initially, we calculated the minimum sample size required to estimate the indicators shown in Table 1, with the desirable coefficient of variation (CV) for each of the disseminating domains, that is, areas where the search is able to generate estimates with sufficient and representative accuracy of the population. The domains are: Brazil, country regions, states, metropolitan region (MR), state capital and the rest of the municipalities in each state. Table 1 shows also the minimum sample

Table 3 – Number of households in the sample and response rates, according to states. National Health Survey, 2013.

Federative Units	Total of occupied households*		Total of households occupied with resident selected for the interview					
	Total (C1)	With interview conducted	Without interview conducted (C3)	Household response rates (%) (C2/C1)	Total (C1s)	With interview conducted (C2s)	Without interview conducted (C3s)	Response rate of the selected residents (%) (C2s/C1s)
Rondônia	1,961	1,849	112	94.3	1,959	1,694	265	86.5
Acre	2,035	1,892	143	93.0	2,033	1,814	219	89.2
Amazonas	2,946	2,795	151	94.4	2,945	2,586	359	87.3
Roraima	1,835	1,749	86	94.5	1,828	1,591	237	86.3
Pará	2,882	2,438	444	84.1	2,879	2,004	875	69.2
Amapá	1,637	1,522	115	93.0	1,637	1,332	305	81.4
Tocantins	1,738	1,601	137	92.1	1,738	1,515	223	87.2
Maranhão	2,072	1,882	190	90.8	2,071	1,774	297	85.7
Piauí	1,957	1,859	98	95.0	1,957	1,804	153	92.2
Ceará	3,224	2,793	431	86.6	3,220	2,560	660	79.5
Rio Grande do Norte	1,987	1,807	180	90.9	1,987	1,691	296	85.1
Paraíba	2,000	1,960	40	98.0	1,998	1,943	55	97.2
Pernambuco	3,043	2,719	324	89.4	3,043	2,591	452	85.1
Alagoas	1,998	1,901	97	95.1	1,995	1,748	247	87.6
Sergipe	1,955	1,734	221	88.7	1,955	1,553	402	79.4
Bahia	2,880	2,776	104	96.0	2,878	2,641	237	91.4
Minas Gerais	4,110	3,932	178	95.7	4,110	3,779	331	91.9
Espírito Santo	2,031	1,894	137	93.3	2,030	1,724	306	84.9
Rio de Janeiro	4,145	3,801	344	91.7	4,145	3,486	659	84.1
São Paulo	6,209	5,623	586	90.6	6,208	5,305	903	85.5
Paraná	3,257	3,122	135	95.9	3,254	3,012	242	92.6
Santa Catarina	1,975	1,721	254	87.1	1,975	1,623	352	82.2
Rio Grande do Sul	3,314	2,996	318	90.4	3,314	2,913	401	87.9
Mato Grosso do Sul	2,016	1,914	102	94.9	2,013	1,809	204	89.9
Mato Grosso	1,827	1,621	206	88.7	1,825	1,476	349	80.9
Goiás	2,831	2,548	283	90.0	2,828	2,423	405	85.7
Distrito Federal	2,129	1,899	230	89.2	2,129	1,811	318	85.1
Brazil	69,994	64,348	5,646	91.9	69,954	60,202	9,752	86.0

*Excluding vacant households.

sizes initially computed for each indicator.

As the MR and the capitals are dissemination domains, the states that do not have MRs have two dissemination domains, namely the state capital and rest of the municipalities in the state. This means that the minimum sample size in these states required to estimate each indicator is about twice that of the size shown in Table 1. In the case of states that have MRs, the required size would be a little more than twice the size, since the MR is also a dissemination domain.

Based on the results found and what was presented above, some adjustments in the minimum sample sizes and in the geographic levels initially thought of as domains were necessary.

After several reviews, we decided to determine the minimum sample size of household per state as being

1800 and, as at least two dissemination domains are expected per state, we determined a minimum sample size of 900 households per domain.

The number of households selected in each PSU was set at 10, however in some domains there were insufficient PSUs in the Master Sample to complete the PNS sample. In these domains, it was necessary to select 14 households per PSU. That is, the number of households per PSU was of 10 or 14, depending on the dissemination domain. The sample sizes by state are shown in Table 2.

The number of households selected for the survey was approximately 23% higher than the minimum sample described above in order to take into account total loss, which includes non-response rate and the rate of misclassification of units in the records used for selection. This

percentage was estimated based on other IBGE surveys, such as the Special Survey on Tobacco (PETaB), which was incorporated into PNAD 2008, although that survey only interviewed one person per household.

Sample Expansion

Because PNS has a complex sampling design and unequal selection probabilities, data analysis requires the definition of the expansion factors or sample weighting for households and all their residents, as well as for the resident selected to answer the third part of the questionnaire. The final weighting is a product of the inverse selection probabilities at each stage of the sampling plan, plus the non-response correction processes and calibration adjustments to the known population totals.

Weighting of primary sampling units (PSUs)

PSU weightings were calculated as the product between the inverse of the PSU selection probability for the Master Sample and the inverse of the PSU selection probability for PNS, as shown below:

$$W_{hi} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}}$$

where,

h is the index of the stratum;

i is the PSU index;

W_{hi} is the basis weight of PSU i of stratum h in PNS;

m_h is the number of PSUs selected in stratum h for Master Sample;

N_{hi} is the number of occupied permanent private households, occupied but without interviews having been conducted (equivalent to closed households) and vacant in PSU i of stratum h [updated data from CNEFE (National Address List for Statistical Purposes) at the time of selection of PSUs for the Master Sample];

N_b is the number of occupied permanent private households, occupied but without interviews having been conducted (equivalent to closed households) and vacant in stratum h [updated data from CNEFE (National Address List for Statistical Purposes) at the time of selection of PSUs for the Master Sample];

m_h^{PNS} is the number of PSUs selected in stratum h for use by PNS.

Weighting households and all their residents

The weightings for households and all their residents were calculated by the product of the corresponding PSU weight and the inverse of the household selection probability within the PSU. Weightings were adjusted for non-response correction and to calibrate the estimates with population totals known through other sources. These weightings are used in the analysis of the answers to the first two parts of the questionnaire which provide information on the characteristics of the households and all their residents.

Households were selected with equal probability in each PSU, so the weighting of the household within the PSU is given by:

$$W_{j|hi} = \frac{N_{hi}^*}{n_{hi}}$$

where,

h is the index of the stratum;

i is the PSU index;

j is the index of the household;

$W_{j|hi}$ is the household selection weighting j in the PSU i of stratum h ;

N_{bi}^* is the number of occupied and closed permanent private households in the PSU i of stratum h [data from the last update of CNEFE at the time of households selection];

n_{bi} is the number of selected households in PSU i of stratum h .

And the basic weighting of the household is obtained by the expression:

$$W_{hij} = W_{hi} * W_{j|hi} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}}$$

After determining the basic weighting, an adjustment was made to compensate interviews lost owing to non-response, *i.e.*, interviews not conducted due to respondent refusal, no contact with the resident or some other reason for loss in occupied households.

The loss adjustment factor was calculated by the ratio between the number of selected and occupied households (households with residents) in the PSU and the number of selected and occupied households with interviews conducted in the PSU. The formula used is shown as follows:

$$W_{hij}^* = W_{hij} * \frac{n_{hi}^*}{n_{hi}}$$

where,

n_{hi}^* is the number of selected and occupied households (households with residents) in the PSU i of stratum b ; and

n_{hi}^{**} is the number of selected and occupied households (households with residents) with interview conducted in the PSU i of stratum b .

After correcting the weightings for non-response, another adjustment was made to the household weightings, referred to as calibration, or post-stratification. The main goal of the calibration of weightings is to get estimates that are consistent with the population projections released by IBGE^{7,8}, so that, in estimating the total population of certain geographical levels, the estimate obtained coincides with the population estimates produced by the IBGE Board of Surveys' Coordination of Population and Social Indicators (COPIS).

Calibration was done using the estimated population as at July 27, 2013, by means of the following expression:

$$W_{hij}^{**} = W_{hij}^* * \frac{P_d^{tri}}{\hat{P}_d^{tri}}$$

where,

P_d^{tri} is the population estimate produced by COPIS for geographical level d on July 27, 2013; and

\hat{P}_d^{tri} is the population estimate obtained with the survey data for geographical level d .

Therefore, the final household weight, with correction for losses and calibrated population totals, is given by:

$$W_{hij}^{**} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * \frac{n_{hi}^*}{n_{hi}^{**}} * \frac{P_d^{tri}}{\hat{P}_d^{tri}}$$

Weighting of selected resident

The selection of the resident who answered to the individual interview was done by simple random sampling. The weighting of the selected resident was therefore calculated by the product of the household weighting multiplied by the number of eligible residents at the household (equivalent to the inverse of the probability of selection). Thus, the basic weighting is given by:

$$W_{hijk} = W_{hij} * O_{hij} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * O_{hij}$$

And the weighting of the selected resident, including allowing for household non-response, is given by:

$$W'_{hijk} = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * \frac{n_{hi}^*}{n_{hi}^{**}} * O_{hij}$$

where,

k is the index of the selected resident;

O_{bij} is the number of residents aged 18 or older in the household j in PSU i of stratum b .

As there was interview loss after resident selection, there was also a need for non-response correction at this stage. Initially, this adjustment was done in a way equivalent to the one already described within the PSU, considering households with resident interviews conducted. After consideration of the estimates of people by sex and characteristics of non-respondents residents, we decided to make the adjustment by sex, because it was found that the loss was higher among men than among women. Weightings, by sex, were as follows:

$$W_{hijk}^M = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * \frac{n_{hi}^*}{n_{hi}^{**}} * O_{hij} * \frac{\sum_j^{n_{hijk}^{**}} W'_{hijk} * \alpha_{hijk}^M}{\sum_i^{n_{hijk}^{***}} W'_{hijk} * \alpha_{hijk}^M}$$

$$W_{hijk}^F = \frac{1}{m_h} * \frac{N_h}{N_{hi}} * \frac{m_h}{m_h^{PNS}} * \frac{N_{hi}^*}{n_{hi}} * \frac{n_{hi}^*}{n_{hi}^{**}} * O_{hij} * \frac{\sum_j^{n_{hijk}^{**}} W'_{hijk} * \alpha_{hijk}^F}{\sum_i^{n_{hijk}^{***}} W'_{hijk} * \alpha_{hijk}^F}$$

where,

n_{hij}^{**} is the number of households selected with interview conducted in the PSU i of stratum b ;

n_{hi}^{***} is the number of residents selected with interview conducted in the PSU i of stratum b ;

α_{hijk}^M indicates whether the resident selected in the household j of PSU i of stratum b is male;

α_{hijk}^F indicates whether the selected resident in the household j of PSU i of stratum b is female.

As household residents are sampled randomly, it is natural that due to this random selection, the population totals obtained with the expansion factors of the selected resident are not exactly the same as population totals obtained with the household expansion factors.

However, residents of the households as a whole form a much larger sample than just the selected residents and thus a more accurate estimate is obtained of these population totals. In order for the estimates to be equal, we chose to calibrate the selected resident weighting so

that the population totals by sex and age groups would correspond to the total obtained with the household weighting. The four age groups used were 18-24 years, 25-39 years, 40-59 years and over 60 years.

The selected resident weighting was calibrated using the following formula, analogous to the weighting of the selected female resident:

$$W_{hijk}^{M*} = W_{hijk}^M * \frac{\hat{P}_{d,M,c}^M}{\hat{P}_{d,M,c}^S}$$

where,

$\hat{P}_{d,M,c}^M$ is the population estimate obtained with the data of the residents of the survey households for geographical level d male sex and age group c ;

$\hat{P}_{d,M,c}^S$ is the population estimate obtained with the data of the selected residents of the survey for geographical level d male sex and age group c .

Non-response rates for households and for the selected resident are presented in Table 3.

Final Considerations

A certain amount of care needs to be taken when processing the data, since they are derived from a complex sample that generates estimates from data collected for households and all their residents (as per the sample selection stage) and for residents aged 18 years or older, who are randomly selected (third stage), including a weighting calibration process in both stages.

All samples originating from cluster sample research must be analyzed using statistical programs with algorithms for complex data analysis, which are able to correct the effect that the conglomeration of primary sampling units has on the estimates, *i.e.*, the sampling plan design effect (DEFF). Most statistical packages have modules or libraries able to correct DEFF, including the Survey library of the R package, the SPSS Complex Sample module, SUDAAN, among others.

The calibration, or post-stratification process, also needs special care at the time of analysis. Generally, when there is no weighting calibration, it is sufficient to input the variables that define the strata, the PSUs,

the type of design and the weightings corrected for non-response into the statistical program at the time of analysis. However, when the weightings go through a process of post-stratification, we must also input the population projections and the categories used for population total adjustments in the weighting calibration process. This information is provided by IBGE along with the database, however, not all statistical packages that perform analysis of complex sample data have this option.

In the case of PNS, there is a set of information related to the household and all its residents, including weightings corrected for non-response, both with and without calibration, the population projections and the categories of population projections. The same is true for the selected resident. The important thing is knowing when to use each one.

The final weightings with calibration adjustment should be used when it is not possible to incorporate the process of post-stratification into the analysis (informing the population projections and the categories of population projections) as the one-off estimates (average, totals, proportions) generated with this weighting are correct.

For professionals with little experience in analyzing data from complex samples, the database and necessary information about the variables that define the sampling plan and that have to be used for the data analysis in the statistical packages for complex data analysis are available at the PNS website (<http://www.pns.icict.fiocruz.br/>).

Authors' Contributions

Souza-Junior PRB, Freitas MPS, Antonaci GA participated in the conception and design of the study, writing, critical review of the contents and approval of the final version of the manuscript.

Szwarcwald CL participated in the conception and design of the study, critical review of the contents and approval of the final version of the manuscript.

All authors approved the final version of the manuscript and declare that they are responsible for all aspects of work, ensuring its accuracy and integrity.

References

1. Szwarcwald CL, Malta DC, Pereira CA, Vieira MLFP, Conde WL, Souza Júnior PRB, et al. National Health Survey in Brazil: design and methodology of application. *Cien Saude Coletiva*. 2014 Feb;19(2):333-42.
2. Instituto Brasileiro de Geografia e Estatística. Diretoria de Pesquisas. Coordenação de Trabalho e Rendimento. Sistema Integrado de Pesquisas Domiciliares – SIPD [Internet]. Rio de Janeiro: IBGE; 2007 [citado 2015 fev 2]. (Texto para discussão; 24). Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/sipd/texto_discussao_24.pdf
3. Instituto Brasileiro de Geografia e Estatística. Coordenação de Comunicação Social. Guia do Censo 2010 para jornalistas. Rio de Janeiro: IBGE; 2010. 40 p.
4. Freitas MPS, Lila ME, Azevedo RV, Antonaci GA. Amostra mestra para o Sistema Integrado de Pesquisas Domiciliares [Internet]. Rio de Janeiro: IBGE; 2007 [citado 2015 fev 2]. (Texto para discussão; 23). Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/sipd/texto_discussao_23.pdf
5. Freitas MPS, Antonaci GA. Sistema Integrado de Pesquisas Domiciliares: amostra mestra 2010 e amostra da PNAD contínua. Rio de Janeiro: IBGE; 2014 [citado 2015 mar 28]. (Texto para discussão; 50). Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv86747.pdf>
6. Cochran WG. Sampling techniques. 3th. New York: John Wiley & Sons; 1977.
7. Vasconcellos MTL, Silva PLN, Szwarcwald CL. Aspectos de amostragem da Pesquisa Mundial de Saúde no Brasil. *Cad Saude Publica*. 2005;21 supl 1:589-99.
8. Ruiz CMM, Silva PLN. Explorando alternativas para a calibração dos pesos amostrais da Pesquisa Nacional por Amostra de Domicílios. In: 6º Congreso de la Asociación Latinoamericana de Población [Internet]; 2014 ago 12-15; Lima. Peru: Asociación Latinoamericana de Población; 2014 [citado 2015 fev 2]. Disponível em: http://www.alapop.org/Congreso2014/DOCSFINAIS_PDF/ALAP_2014_FINAL948.pdf.

Received on 21/02/2015
Approved on 02/04/2015