Sampling design for the *World Health Survey* in Brazil

Aspectos da amostragem da *Pesquisa Mundial de Saúde* no Brasil

Mauricio Teixeira Leite de Vasconcellos [1]
Pedro Luis do Nascimento Silva [1]
Célia Landmann Szwarcwald [2]

[1] *Escola Nacional de Ciências Estatísticas, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brasil.*
[2] *Centro de Informação Científica e Tecnológica, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.*

Correspondence
M. T. L. Vasconcellos
Escola Nacional de Ciências Estatísticas, Instituto Brasileiro de Geografia e Estatística.
Rua André Cavalcanti 106, Rio de Janeiro, RJ 20231-050, Brasil.
mtlv@ibge.gov.br

Abstract

This paper describes the sample design used in the Brazilian application of the World Health Survey. *The sample was selected in three stages. First, the census tracts were allocated in six strata defined by their urban/rural situation and population groups of the municipalities (counties). The tracts were selected using probabilities proportional to the respective number of households. In the second stage, households were selected with equiprobability using an inverse sample design to ensure 20 households interviewed per tract. In the last stage, one adult (18 years or older) per household was selected with equiprobability to answer the majority of the questionnaire. Sample weights were based on the inverse of the inclusion probabilities in the sample. To reduce bias in regional estimates, a household weighting calibration procedure was used to reduce sample bias in relation to income, sex, and age group.*

Survey Methods; Selection Bias; Sampling Studies

Introduction

The *World Health Survey* (WHS) in Brazil was part of a World Health Organization (WHO) project aimed at collecting information on populations' health status and health system performance in member countries, by means of household surveys. In Brazil, the WHS was conducted in 2003, coordinated by the Oswaldo Cruz Foundation with support from the Brazilian Institute of Geography and Statistics (IBGE), in the sample design.

Due to the lack of a centralized and reliable household registry, Brazil's national household surveys use the IBGE Geographic Operational Base for stratification and selection of areas. In the selected areas, updated household registries (or lists) are prepared to use in household selection. Thus, the household surveys use sampling designs with two or more selection stages, combining stratification of primary sampling units (municipalities or census tracts) and selection probabilities proportional to some measure of size, in addition to equiprobable selection of the final sampling units (households) to correct the probabilities of including households in the sample. The selected household, in turn, is normally treated as a closed group in which all the residents are survey targets, even when there are different questions according to individual characteristics, such as employment for members ten years or older or fertility

for women 15 years or older [1,2,3]. To estimate the data, the design's natural expansion factors were initially calculated, corresponding to the inverse inclusion probability in the various selection stages, which are subsequently calibrated to eliminate minor selection biases and generate sampling factors or weights that are used for tabulating and disseminating the data.

In this sense, the WHS sample design was based on traditional household survey designs, with three selection stages: census tracts, households, and adult residents (18 years or older). As a function of the objectives and budget limitations, some adaptations were made, including: (1) the use of inverse sampling as a way of avoiding an increase in the sample size to compensate for non-responses and (2) utilization of data collection instruments from the 2000 Population Census (*Censo Demográfico 2000*)[4] to reduce the costs of listing households in the selected tracts.

This article describes the sample design used in the WHS, considering the census tract stratification and selection; selection of households and adults interviewed; expansion of the sample of households and adults interviewed; calibration of the sampling weights; and data collection strategies used to obtain a representative household sample for Brazil given the available resources. The sample selection and expansion used Statistical Analysis System (SAS), version 8, and the calibration used the Generalized Estimation System (GES), version 4.2, both licensed for use by the IBGE.

Survey population, geographic operational base, and tract stratification

The survey population [5] corresponded to the entire set of permanent private households in Brazil, except for those located in the rural areas of the Northern macro-region and special census tracts (military barracks and bases, lodgings, camps, ships/vessels, prisons, nursing homes, orphanages, convents/monasteries, and hospitals). According to this definition, the sample population included 207,513 tracts (96.2% of the 215,811 census tracts from 2000). According to the 2000 Population Census, of the 45,053,286 permanent private households existing in Brazil, 44,005,362 (97.7%) were covered by the sample population. In terms of the resident Brazilian population for 2000 (169,799,170 inhabitants), 99.7% resided in permanent private households (169,252,872 inhabitants), and exclusions of tracts reduced the survey population to 165,669,837 inhabitants, or 97.6% of the resident population.

The IBGE geographic operational base represents Brazil's territorial division in terms of so-called Federal units (States and the Federal District, or Brasília), municipalities, districts, and other legal subdivisions (sub-districts, administrative regions, neighborhoods, city limits), including the census tracts as the operational subdivision. The tracts are subdivisions that follow the legal boundaries (city limits and other subdivisions of Brazil's territorial division) and completely cover the national territory. Each tract has physical boundaries that are identifiable in the field, chosen so as to contain a certain number of households, which together with their geographic extension allow conducting the census by a single census-taker in the tract. The tracts have their boundaries depicted on individual maps (for the urban tracts) or grouped maps (for rural tracts). The geographic operational base is reviewed on an ongoing basis due to changes in the legislation (creation of new municipalities by reapportionment, alterations in city limits) and the population growth itself, but the data associated with the census tracts are only updated when the censuses and population counts are conducted.

In the case of the WHS, three products of the geographic operational base were used, all pertaining to 2000: (1) the tract-aggregate file with data from the universe of the 2000 Population Census [6]; (2) the tract maps; and (3) the registry with the description of the tract boundaries. The latter two were used during the data collection and were only provided for the tracts selected for the sample, while the first was used for the sample stratification, selection, and expansion.

To ensure that the sample represents the urban and rural areas of the small, medium, and large municipalities, which have important differences in the size and type of health services, the tracts were divided into six strata based on the combination of the tract's situation (urban or rural) and the municipality's population range (< 50,000 inhabitants; 50,000-399,999; and 400,000 +). Within each stratum, the tracts were ordered (before their selection) according to head-of-household's mean income, and each stratum's sample was selected in three stages: (1) census tracts; (2) private households; and (3) adult residents.

Census tract selection

In the first stage, the census tracts were selected systematically with the probability proportional to a measurement of size, defined according to the number of permanent private households in each tract. In order to reduce the variance in the expansion factor, the measurement of the tract size was limited to the interval [50; 2,100], i.e., tracts with 50 or fewer permanent private households were assigned to size 50; those with 50 to 2,100 were assigned to size equal to the real number of permanent private households; and those with more than 2,100 permanent private households were assigned to size 2,100. Note that the prior ordering of tracts meant implicit income stratification in each stratum, which guaranteed the representation of all the socioeconomic levels in the stratum [7].

Budget restrictions set the sample size at 5 thousand households and forced the sample to be more clustered than desirable, with the selection of 20 households per tract, resulting in a sample of 250 tracts. Allocation of the tracts' sample size across the strata was proportional to each stratum's population. Adjustments had to be made to ensure a minimum of five selected tracts per stratum (Table 1).

Selection of tracts and households to be visited was done in the office, prior to the data collection, while selection of the adults was done in the field based on the collection support material provided.

Household selection and inverse sampling

In the second stage, for each sample tract, equiprobability was used to select 60 private households in order to obtain 20 interviews, following the inverse sampling method. This method, originally proposed by Haldane [8] to estimate frequencies and proportions, can be defined as a technique to verify how many units need to be observed in order to obtain a prefixed number of successes, or in this case interviews performed. Application of this method in the WHS consisted of sequentially visiting the previously selected households, recording the occurrences (interview or non-interview by type) until reaching the planned number of 20 interviews per tract.

This method is called inverse sampling because rather than defining the sample size (or number of households to visit to attempt the interview), it defines the number of successes (or interviews performed), considering (for the sample expansion) the number of households actually visited. The method's main advantages are: (1) dispensing with correction of the calculated sample size to compensate for the expected non-response rate, which would be difficult to adjust at the tract level; (2) avoiding the use of over-sampling when the minimum number of interviews needed is not reached (that is, when compensation for non-response is insufficient); (3) dispensing with non-response correction during the sample expansion process; and (4) including a sample screening procedure, which is less costly than the screening technique described by Kalton & Anderson [9].

Considering that the cost-containment measures to keep the WHS data collection costs within the budget limitations included not conducting household listing in the selected tracts, by using the 2000 Population Census Listing (*Folha de Coleta do Censo Demográfico 2000*), inverse sampling for household selection was done simply by selecting (with equal probability between one and the member of

Table 1

Allocation of tract sample size among strata.

Sample strata	Stratum population	Sample size	
		Calculated	Adjusted
Urban areas and population < 50,000	37,485,034	56.6	56
Rural areas and population < 50,000	21,825,293	32.9	32
Urban areas and population from 50,000-399,999	49,089,645	74.1	74
Rural areas and population from 50,000-399,999	4,972,309	7.5	7
Urban areas and population > 400,000	51,228,275	77.3	76
Rural areas and population > 400,000	1,069,281	1.6	5
Total	165,669,837	250.0	250

permanent private households in the tract) 60 random numbers from the household order in the 2000 Population Census Listing. Since the latter sequentially numbers the permanent private households and provides the address for each household, once having selected the ordered numbers for the households to be visited, the list of household addresses to be visited was prepared, following the selection sequence.

This strategy, while reducing the data collection costs, did not produce any probability of selecting households created after the 2000 Population Census, thus limiting the inference to the universe of permanent private households included in the 2000 Census.

Selection of the adult to be interviewed

In the third stage, an adult (18 years or older) was selected in each household interviewed, giving equal probability to all the adult residents in the household. The method proposed by the WHO for this selection consisted of using the so-called Kish numbers, which are equiprobable numbers selected from the sets of natural numbers with 2, 3, 4, 5, or 6 elements, beginning with 1. However, since households with more than six adults are not infrequent in Brazil, it was decided to extend this procedure to households with up to 12 adult residents (the questionnaire provided a space for recording up to 13 residents). Thus, 11 sets of natural numbers were defined, beginning with {1, 2} and ending with {1, 2,..., 12}; and in each set, 5 thousand equiprobable selections were made of only one of the elements in the set. The results of these selections were placed in tables included in the Information Sheet on Tracts, Households, and Persons to Interview. The interviewer thereby determined the order number of the selected adult in the table's box, from the Information Sheet on Tracts, Households, and Persons to Interview, identified by the total number of adults in the household (row) and the number of the household interviewed (column).

The probabilistic selection scheme in the WHS sample design can be represented by the expressions in Figure 1. Inclusion probabilities are calculated on the basis of the number of favorable cases divided by the possible cases, except in the case of household selection, where the estimator proposed by Haldane [8] also considers the ratio between the number of households visited in the target population and the predicted number of successes in the sample, both with less one degree of freedom.

Data collection support material

In addition to the maps and the 2000 Population Census Listing for the selected tracts, the field team received the Information Sheet on Tracts, Households, and Persons to Interview and the instructions on how to identify and canvass the census tracts, based on the 2000 Population Census manuals. The Information Sheet has three main sections: (1) identification data for the selected tract, with the tract codes and the codes and names of different subdivisions in Brazil's territorial division, description of the initial and final point in the tract's boundaries, and indication of tracts contained within the boundaries that should be excluded (generally special tracts with barracks, nursing homes, etc.); (2) list of the ordered numbers of permanent private households to be interviewed, to prepare the list of addresses based on the 2000 Population Census Listing; and (3) a table for selecting the adult to be interviewed and the rotation in the vignettes to be used.

Since each tract was surveyed by two interviewers and each interviewer was responsible for conducting interviews in ten households, each tract had two Information Sheets on Tracts, Households, and Persons to Interview, each one with thirty households to interview (in order of selection) and instructions for the selection of ten adults and ten vignette rotations. The vignettes, aimed at intercultural calibration, were not the same for all the adults interviewed, and were divided into four groups, referred to by the WHO as rotations and identified by the letters A, B, C, and D. Selection in this case was done by the WHO, and on the Information Sheet on Tracts, Households, and Persons to Interview, the 5 thousand rotations furnished by the WHO were recorded, following the sequential enumeration of the households interviewed, defined by the WHO identification of the tract (001 to 250) and of the households in each tract (01 to 20).

Results of interviews

During the data collection it was necessary to replace three selected tracts. Two were due to problems of access: one was located 30 hours by boat from Belém (at the mouth of the Amazon River), thus posing a cost that was incompatible with the survey budget, while the other was experiencing problems with the drug traffic. The third tract was replaced since it was a summer resort area, where most of the dwellings

Figure 1

Probabilistic scheme of the WHS sample.

Where h is the stratum index, i the selected tract index, j the selected household index, and k the selected adult index, the probability of including any given adult is equal to the product of the probabilities of including tract i, represented by $P(S_{hi})$; household j, represented by $P(D_{hij} / S_{hi})$; and adult k, represented by $P(A_{hijk} / D_{hij} \cap S_{hi})$. These probabilities are expressed as follows:

(1) $$P(S_{hi}) = \frac{n_h \times M_{hi}}{M_h} ;$$

(2) $$P(D_{hij} / S_{hi}) = \frac{d_{hi} - 1}{n_{hi} - 1} \times \frac{n_{hi}}{M^*_{hi}} ;$$

(3) $$P(A_{hijk} / D_{hij} \cap S_{hi}) = \frac{1}{M^*_{hij}} ,$$

where:

n_h is the sample size of tracts in stratum h, indicated in Table 1;

M_{hi} is the size associated with tract i in stratum h, defined as the number of permanent private households in the tract limited to the range [50, 2100], that is, $M_{hi} = \max \{ 50; \min [M^*_{hi} ; 2100] \}$;

M_h is the sum of the sizes of all the tracts in stratum h, that is, $M_h = \sum_{i=1}^{N_h} M_{hi}$, where N_h is the number of tracts in the population for stratum h;

n_{hi} is the size of the actual household sample (or of adults, since only one adult is selected per household) in tract i of stratum h, which was set at 20 households per tract, that is, the number of interviews conducted in households in the target population;

d_{hi} is the number of households in the target population (consisting of the permanent private households with at least one resident adult eligible for the interview) visited in tract i of stratum h in order to obtain 20 interviews in the tract;

M^*_{hi} is the number of private households in tract i of stratum h; and

M^*_{hij} is the number of adults eligible for selection in household j of tract i in stratum h.

Thus, the probability of including any given adult, without considering non-response, as shown by , is given by the following (4):

(4) $$P(A_{hijk}) = \frac{n_h \times M_{hi}}{M_h} \times \frac{d_{hi} - 1}{n_{hi} - 1} \times \frac{n_{hi}}{M^*_{hi}} \times \frac{1}{M^*_{hij}} .$$

were only used occasionally (and were thus not selection targets in order not to increase the probability of selecting their owners). The replacement was done by selecting the next tract in the selection registry order (same stratum and equivalent mean income).

In all 250 sample tracts the 20 planned interviews were obtained (Table 2), and it was necessary to visit an average of 34.4 households. The mean number of households visited includes 20 permanent private households interviewed (58.1%); 8.5 refusals by the household as a whole or by the selected adult (24.7%); 3.3 permanent private households that were vacant or had no adults (9.6%); and 2.6 non-existent households or dwelling units that were no longer permanent private households (7.6%). This last group is probably due to fact that this was an old households list (from the 2000 Population Census Listing). As shown in Table 2, the refusal rate increased according to the mean income quintile of the heads-of-households in the tract, as did the mean number of vacant households, while the number of non-existent households tended to drop as the tract's mean income increased, except for the wealthiest quintile.

The mean number of households visited per tract indicates that the sample size in a tra-

Table 2

Mean number of households according to interview results by sample strata and mean income quintiles of heads-of-households in the tract.

Sample strata and income quintiles	Mean number of households				
	Visited	Interviewed	Refusals	Vacant	Non-existent
Urban areas and population < 50,000	29.2	20	5.0	2.3	1.9
Rural areas and population < 50,000	27.1	20	2.8	2.1	2.3
Urban areas and population from 50,000-399,999	35.5	20	9.3	3.9	2.4
Rural areas and population from 50,000-399,999	27.3	20	0.4	2.3	4.6
Urban areas and population > 400,000	40.9	20	3.7	4.0	3.2
Rural areas and population > 400,000	32.2	20	7.0	3.0	2.2
Income quintiles					
1st quintile	30.5	20	3.0	2.8	4.7
2nd quintile	30.7	20	5.8	2.8	2.1
3rd quintile	33.8	20	9.6	2.5	1.7
4th quintile	36.4	20	11.0	3.8	1.6
5th quintile	40.4	20	13.1	4.4	2.9
Total	34.4	20	8.5	3.3	2.6

ditional design should be 35 households to compensate for non-responses and to obtain 20 interviews performed per tract. However, this solution would have been insufficient for 71 tracts in which 40 or more visits were needed (and where the solution would have been over-sampling) and would have been a waste of resources in 102 tracts where fewer than 30 households were visited.

Sample expansion

Natural expansion factors in the design are defined as the inverse probability of a household's inclusion (the product of expressions 1 and 2, in Figure 1) and that of an adult (expression 4, Figure 1).

However, the use of the 2000 Population Census Listing meant the selection of households that were non-existent at the moment of the survey and of some for which it was not possible to determine whether they belonged to the sample population (closed residences). Thus, the probability of including a household, conditioned on the tract's selection (expression 2, Figure 1), had to be subdivided into three probabilities: (1) that of being visited; (2) that of belonging to the sample population or of being eligible; and (3) that of being one of the first 20 eligible households to agree to the interview, which correspond respectively to the three ratios shown in expression 5 of Figure 2. With this correction the design's natural expansion factors (to be applied to the data for the selected adults and the interviewed households) are given respectively by expressions 7 and 8, Figure 2.

The factors calculated by expressions (7) and (8) are fractions and consequently so are the estimates. It was thus decided that all the estimates of the number of persons and households would be used in calculating the totals and percentages without rounding-off, and that the rounding-off would be done at the end for each single estimate. Although this decision can generate tables with different totals from the sum of its portions, it guarantees consistency between identical estimates presented in different tables.

Calibration of expansion factors and estimation of regional data

In official statistics it is common to calibrate the expansion factors of household survey samples. According to Silva [10], the most common justification is to maintain the coherency of population data already published. In addition, up to a point, calibration allows correcting of selection biases and making estimates coherent with population totals obtained from other sources.

In the case of the WHS, since the sample was not selected to furnish data for the macro-regions of Brazil, there was a concentration of the wealthiest tracts in the South and Southeast macro-regions and of the poorest tracts in

Figure 2

Natural expansion factors in the design.

Where h is the stratum index, i the selected tract index, j the selected household index, and k the selected adult index, the corrected probability of including household j in the sample, as represented by $P^c (D_{hij} / S_{hi})$ is the following:

(5) $$P^c (D_{hij} / S_{hi}) = \frac{v_{hi}}{M^*_{hi}} \times \frac{d_{hi} - 1}{v_{hi} - 1} \times \frac{n_{hi}}{d_{hi}} ,$$

where:

n_{hi} is the size of the actual household sample (or of adults, since only one adult is selected per household) in tract i of stratum h, which was set at 20 households per tract, that is, the number of interviews conducted in households in the target population;

d_{hi} is the number of households in the target population (consisting of the permanent private households with at least one resident adult eligible for the interview) visited in tract i of stratum h to in order to obtain 20 interviews in the tract;

M^*_{hi} is the number of private households in tract i of stratum h; and

v_{hi} is the number of households visited in tract i of stratum h.

Thus, the probability of including any given adult, represented by $P^c (A_{hijk})$, is given by the expression (6):

(6) $$P^c (A_{hijk}) = \frac{n_h \times M_{hi}}{M_h} \times \frac{v_{hi}}{M^*_{hi}} \times \frac{d_{hi} - 1}{v_{hi} - 1} \times \frac{n_{hi}}{d_{hi}} \times \frac{1}{M^*_{hij}} .$$

The weight (or expansion factor) to be applied to this adult, represented by W_{hijk} , corresponds to the inverse probability given in (6), as indicated by expression (7):

(7) $$W_{hijk} = \frac{M_h}{n_h \times M_{hi}} \times \frac{M^*_{hi} \times (v_{hi} - 1) \times d_{hi}}{v_{hi} (d_{hi} - 1) \times n_{hi}} \times M^*_{hij}.$$

The weight (or expansion factor) to be applied to the household data, represented by W_{hij} , is given by the expression (8):

(8) $$W_{hij} = \frac{M_h}{n_h \times M_{hi}} \times \frac{M^*_{hi} \times (v_{hi} - 1) \times d_{hi}}{v_{hi} (d_{hi} - 1) \times n_{hi}} .$$

the other regions, which would introduce a bias in regional estimates by not representing the range of income variation in each region. Given this situation, calibration of the sampling weights in the WHS became the only means to correct this situation and generate expansion factors that allow breaking down the national data in groups of macro-regions.

The basic idea of calibration is to estimate factors (called calibration factors) that multiply the design's natural sampling weights (obtained from the inverse probabilities of inclusion in the sample) and thereby furnish the calibrated sampling weights, which have the property of minimizing the difference between the estimated and known population totals for a set of auxiliary calibration variables. This application used the technique known as integrated household weighting, which uses regression to determine the values of the calibration factors so as to simultaneously minimize the differences between the estimated and known totals of households and persons for a set of defined post-strata, so that the household's calibrated expansion factor is the same for all its members.

According to Silva [10], some principles and statistics are normally analyzed to determine the best choice of calibration post-strata and decide between the sets of calibrated factors, without losing sight of the need for a sample size in each post-stratum that allows supposing that the bias in the ratio estimators is negligible [5]. Thus, calibration should seek positive calibration factors to avoid negative or null calibrated weights, which are mathematically possible but which destroy the principle that the expansion factor indicates the number of population units represented by the sample unit, besides generating absurd values in the estimation of some domains. In addition, the mean

of calibration factors should not be significantly different from one, given that its difference from one should represent the necessary addition or subtraction for the total natural expansion factor to reach the population total, which in principle should be small. This aspect leads to another important point, related to large extreme calibration factor values, which influence their mean (leading to their deviating from the desired value of one) and indicate few units in the sample for one or more post-strata, making the ratio estimators' bias non-negligible.

Another relevant point relates to the choice between different sets of calibrated expansion factors in different post-strata and for different auxiliary variables. Silva [10] proposes seven measures to aid the choice and identify problems in the different sets of post-strata, of which we highlight five: (1) the mean absolute values for the differences between the population totals (of the auxiliary variables used) and their estimates (obtained with the natural expansion factor); (2) the mean coefficient of variation of the estimated auxiliary variables' totals, using the Horvitz-Thompson estimator [5]; (3) the proportion of the calibration values that are greater than the maximum limit established in the calibration algorithm; (4) the proportion of those lower than the minimum limit; and (5) the coefficient of variation in the calibration factors.

In the application of the WHS, different sets of post-strata were tested. Some, like those including the sample selection stratum in their definition, were discarded because they were too detailed for the sample size, generating negative or null calibration factors (note that the samples of rural strata in medium and large municipalities had respectively seven and five census tracts, or 140 and 100 households, in the sample, as shown in Table 1), not to mention the fact that some post-strata had population units not represented in the sample. Others, containing only the total population by sample selection stratum (rather than the total for each combination of the other post-stratification variables), after processing so as not to generate negative factors, were abandoned since they led to statistics with higher values than those obtained with the post-stratification used.

The post-strata used for calibration were defined by the combination of four variables: (1) three groups of macro-regions (Northeast; Southeast; and the rest of the country); (2) the mean income quintiles for heads-of-households in the tract for each group of macro-regions; (3) resident's sex; and (4) nine age groups (0-9; 10-17; 18-19; five ten-year groups from 20 to 69 years; and the 70-and-over group), producing a total of 270 post-strata.

Since the WHS had an additional selection stage (that of an adult resident), a two-stage calibration technique was adopted. In the first stage, the calibrated expansion factors for the households were determined by the technique described above and considering the 270 post-strata. In the second stage, the household's calibrated weight was multiplied by the total number of adults in the household, generating provisional weights for the selected adults which were calibrated by the ratio between the total obtained in the 2000 Population Census and the estimate they produced for the defined post-strata. This ratio corresponds to a calibration factor that varies by post-stratum and that was the target of an analysis similar to that described previously.

The estimates obtained by using the household natural expansion factor indicate that there is a trend to underestimate the number of households in the poorest quintile of the Brazilian population in all the groups of macro-regions except for the Northeast (Table 3). In the fourth income quintile an opposite trend was observed, with an overestimation of the number of households in the national total and in all the groups of macro-regions except for the Southeast.

In relation to the estimates for the total population and in terms of income quintiles, the household natural expansion factor leads to estimates with the same trends that were observed for the number of households, as indicated by the relative errors presented in the upper part of Table 4. For the population estimates by age bracket, one observes a trend to underestimate the younger age groups (up to 39 years) and overestimate the older age groups and especially the elderly, with a variation in the case of the Northeast macro-region, where all the age groups were overestimated (data not shown). This trend towards underestimation resulted from the household composition design, which provided space for recording up to six men and seven women from the oldest to the youngest in each sex. However, the data on household composition are not the most important in the survey, which focuses on information for the selected adult. In relation to the population by sex, the relative errors shown in Table 4 indicate similar trends to those observed for the total population, except in relation to 18-19-year-old males. The most relevant point is the slight underestimation of the total male population as a result of the compensation effect between the strong overestimation

Table 3

Number of households in the population and estimated number of households obtained by natural and calibrated household weights, according to regions and income quintiles.

Macro-regions and mean income quintiles in tract	Population data	Natural household weight		Calibrated household weight	
		Estimate	Relative error (%)*	Estimate	Relative error (%)*
Brazil	44,005,362	44,514,246	1.2	44,005,362	0.0
1st income quintile	8,805,572	7,978,516	-9.4	8,805,572	0.0
2nd income quintile	8,799,675	9,170,014	4.2	8,799,675	0.0
3rd income quintile	8,801,614	9,119,175	3.6	8,801,614	0.0
4th income quintile	8,801,022	9,357,769	6.3	8,801,022	0.0
5th income quintile	8,797,479	8,888,772	1.0	8,797,479	0.0
Macro-regions: North, Central West, and South	12,390,489	12,254,368	-1.1	12,390,489	0.0
1st income quintile	2,479,617	2,125,424	-14.3	2,479,617	0.0
2nd income quintile	2,477,840	2,599,268	4.9	2,477,840	0.0
3rd income quintile	2,478,446	2,063,377	-16.7	2,478,446	0.0
4th income quintile	2,478,043	2,896,629	16.9	2,478,043	0.0
5th income quintile	2,476,543	2,569,670	3.8	2,476,543	0.0
Macro-region: Northeast	11,395,951	13,214,573	16.0	11,395,951	0.0
1st income quintile	2,281,320	2,691,767	18.0	2,281,320	0.0
2nd income quintile	2,278,195	3,095,107	35.9	2,278,195	0.0
3rd income quintile	2,278,966	2,274,674	-0.2	2,278,966	0.0
4th income quintile	2,279,699	2,861,139	25.5	2,279,699	0.0
5th income quintile	2,277,771	2,291,886	0.6	2,277,771	0.0
Macro-region: Southeast	20,218,922	19,045,305	-5.8	20,218,922	0.0
1st income quintile	4,044,635	3,161,326	-21.8	4,044,635	0.0
2nd income quintile	4,043,640	3,475,638	-14.0	4,043,640	0.0
3rd income quintile	4,044,202	4,781,124	18.2	4,044,202	0.0
4th income quintile	4,043,280	3,600,001	-11.0	4,043,280	0.0
5th income quintile	4,043,165	4,027,216	-0.4	4,043,165	0.0

* Relative error (%) = (Estimate – population data) x 100 / population data.

of men in the Northeast and underestimation observed in the other macro-regions (data not shown). The calibrated household factor completely corrects all these trends, since optimum estimates were obtained for the calibration factors in the calibration regressions, done with up to five hundred iterations, using 0.001 and 3 as the lower and upper limits for the calibration factors, respectively.

Since the expansion factor for the selected adult is more important for the WHS, based on the reasons already discussed, the lower part of Table 4 shows a comparison between the relative errors of the estimates produced by the factors for the selected adult, both for the natural sample design and the calibrated value. In these estimates one observes an overestimation of the total adult population (18 years or over) and to a greater extent of the adult female population. Distribution by income quintiles shows a similar pattern to that described for the total population, with an inversion of the tendency in the wealthiest fifth. Age distribution shows the same pattern as the total population, with an inversion of the tendency in the 18-to-19-year bracket for the total population.

These results reflect tendencies that are familiar to whoever conducts household surveys, no matter how rigorous the data collection protocol: the more frequent presence of women and the elderly in the home as compared to men and working-age individuals. Thus, calibration of the expansion factors emerges as the best technical alternative to deal with such typical collection and selection biases.

Table 4

Relative errors of estimated total population and by sex, obtained by natural and calibrated weights, according to age groups and mean income quintiles in the tract.

Age groups and mean income quintiles in the tract	Total population		Male population		Female population	
	Natural weight	Calibrated weight	Natural weight	Calibrated weight	Natural weight	Calibrated weight
Relative errors (%)* of the estimates obtained by the household weight						
Brazil	0.5	0.0	-0.9	0.0	1.9	0.0
0 to 9 years	-6.8	0.0	-6.8	0.0	-6.8	0.0
10 to 17 years	-6.3	0.0	-6.3	0.0	-6.4	0.0
18 to 19 years	-1.9	0.0	2.8	0.0	-6.7	0.0
20 to 29 years	-0.3	0.0	-2.1	0.0	1.5	0.0
30 to 39 years	-2.2	0.0	-4.1	0.0	-0.4	0.0
40 to 49 years	6.2	0.0	4.0	0.0	8.2	0.0
50 to 59 years	18.2	0.0	12.5	0.0	23.4	0.0
60 to 69 years	18.0	0.0	11.0	0.0	24.0	0.0
70 years or older	10.9	0.0	15.6	0.0	7.4	0.0
1st income quintile	-11.4	0.0	-12.3	0.0	-10.5	0.0
2nd income quintile	2.7	0.0	2.2	0.0	3.3	0.0
3rd income quintile	2.9	0.0	-0.8	0.0	6.4	0.0
4th income quintile	9.6	0.0	7.3	0.0	11.8	0.0
5th income quintile	-0.2	0.0	0.4	0.0	-0.9	0.0
Relative errors (%)* of estimates obtained by the selected adult weight						
Brazil	4.2	0.0	-0.9	0.0	9.0	0.0
18 to 19 years	1.5	0.0	-3.1	0.0	6.1	0.0
20 to 29 years	-8.1	0.0	-12.8	0.0	-3.6	0.0
30 to 39 years	-1.7	0.0	-8.0	0.0	4.3	0.0
40 to 49 years	6.8	0.0	-2.8	0.0	15.7	0.0
50 to 59 years	28.8	0.0	14.9	0.0	41.6	0.0
60 to 69 years	24.0	0.0	35.9	0.0	13.9	0.0
70 years or older	7.3	0.0	18.8	0.0	-1.4	0.0
1st income quintile	-7.4	0.0	-16.3	0.0	1.2	0.0
2nd income quintile	8.8	0.0	0.2	0.0	17.2	0.0
3rd income quintile	6.4	0.0	4.7	0.0	8.0	0.0
4th income quintile	9.8	0.0	4.8	0.0	14.4	0.0
5th income quintile	3.2	0.0	1.9	0.0	4.2	0.0

* Relative error (%) = (Estimate – population data) x 100 / population data.

Resumo

Este artigo descreve o desenho da amostra da Pesquisa Mundial de Saúde *no Brasil. A amostra foi selecionada em três estágios. No primeiro, os setores censitários foram divididos em seis estratos, definidos pela situação e porte populacional dos municípios, e selecionados com probabilidade proporcional ao seu número de domicílios. No segundo estágio, os domicílios foram selecionados com eqüiprobabilidade, seguindo um esquema de amostragem inversa, para assegurar vinte entrevistas realizadas por setor. No último estágio foi selecionado com eqüiprobabilidade um adulto (18 anos ou mais) por domicílio para responder aos principais quesitos do questionário. A expansão da amostra foi feita com base nas probabilidades de seleção e, para permitir a obtenção de estimativas regionalizadas, os fatores de expansão foram calibrados para assegurar coerência com os totais populacionais por grupos de macrorregiões, quintos de renda, sexo e grupos etários, por meio de estimadores de regressão.*

Métodos de Levantamento; Viés de Seleção; Amostragem

Collaborators

M. T. L. Vasconcellos participated in the definition and selection of the WHS sample and participated in drafting the article. P. L. N. Silva participated in the definition of the sample design and drafting of the article. C. L. Szwarcwald participated in the definition of the sample design and drafting of the article.

References

1. Instituto Brasileiro de Geografia e Estatística. Pesquisa Nacional por Amostra de Domicílio: síntese de indicadores 2003. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2004.
2. Instituto Brasileiro de Geografia e Estatística. Pesquisa Mensal de Emprego. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2002. (Série Relatórios Metodológicos 23).
3. Bianchini ZM, Vieira M. Aspectos de amostragem da Pesquisa de Orçamentos Familiares 1995-1996. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 1998. (Textos para Discussão 93).
4. Instituto Brasileiro de Geografia e Estatística. Metodologia do Censo Demográfico 2000. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2003. (Série Relatórios Metodológicos 25).
5. Cochran WG. Sampling techniques. 3rd Ed. New York: John Wiley & Sons; 1977.
6. Instituto Brasileiro de Geografia e Estatística. Censo Demográfico 2000: agregado por setores censitários dos resultados do universo. 2a Ed. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2003.
7. Madow WG. On the theory of systematic sampling, II. Annals of Mathematical Statistics 1949; 20:333-54.
8. Haldane JBS. On a method of estimating frequencies. Biometrika 1945; 33:222-5.
9. Kalton G, Anderson DW. Sampling rare populations. J R Stat Soc [Ser A] 1986; 149 (Pt 1):65-82.
10. Silva PLN. Calibration estimation: when and why, how much and how. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2004. (Textos para Discussão da Diretoria de Pesquisas 14).

Submitted on 04/May/2005
Final version resubmitted on 13/Oct/2005
Approved on 20/Oct/2005