First reproducible nationwide survey on substance use in Brazil: survey design and weighting

Pedro L. d. N. Silva[1], Mauricio T. L. d. Vasconcellos[1], Francisco I. Bastos[2], Raquel B. De Boni[2], Neilane Bertoni[3], Carolina Coutinho[2], Jurema Mota[2]

[1]IBGE/ENCE, Rua André Cavalcanti, 106, Rio de Janeiro-RJ, 20231-050, Brazil
[2]FIOCRUZ, Avenida Brasil, 4365, Rio de Janeiro-RJ, 21040-360, Brazil
[3]Instituto Nacional de Câncer, Rua Marquês de Pombal, 125, Rio de Janeiro-RJ, 20230-240, Brazil

Abstract

The paper describes the study design adopted in the III Brazilian Household Survey on Substance Use (BHSU-3). This is the first truly reproducible nationwide survey to investigate substance use and related issues carried out in Brazil in 2015. The study adopted a stratified multi-stage probability sampling design and interviewed 16,273 individuals in their households, using face-to-face paper and pencil interviewing. The study population included all residents of private households aged 12-65 at the time of the survey. Calibration weighting was used to compensate for differential age-sex nonresponse. R statistical software was used throughout weighting and tabulation, thus enabling fully reproducible results by analysts having access to the survey microdata. Innovative combination of proven survey methods and statistical leadership of study design team were crucial elements for successful planning, realization and conclusion of the study.

Key Words: probability sampling; survey design; weighting methods; reproducibility.

1. Introduction

This paper broadly describes the methods used to conduct the III Brazilian Household Survey on Substance Use (BHSU-3). A substantial effort was made to ensure transparency and reproducibility of this observational study, as emphasised by (von Elm et al., 2007), and in strict adherence to the UN Fundamental Principles of Official Statistics (UNSD, 1994). Space restrictions prevent full disclosure of all details here, but all details are available in Bastos et al. (2017).

The BHSU-3 was conducted on demand for the National Secretariat for Policies on Drugs (SENAD) as specified in a public bid published on 11 February 2014. One of the core requirements of the call was that the survey should adhere to sampling and collection protocols like those adopted by the Brazilian National Statistics Institute (IBGE). This requirement aimed to ensure that results could be generalized properly to the national population and support reproducibility. The motivation behind this requirement came from the fact that previous surveys on this topic covered only the 108 largest municipalities in Brazil, and thus failed to provide proper estimates for the whole country, since it is expected that drug use patterns vary by size/nature of municipality.

The BHSU-3 aimed to "estimate epidemiologic parameters related to substance use by people of both sexes in the age range 12-65 years old, in all of the national territory, including rural areas". BHSU-3 specific objectives included:

a) Provide direct estimates of prevalence and use pattern (life, year, month), problematic use (heavy, frequent), and incidence over last year of the use of alcohol, tobacco, marijuana, hashish, skank, solvents, cocaine, crack, hallucinogens, Ketamine, Ayahuasca 'tea', ecstasy, steroids, anabolic, anxiolytics (benzodiazepines), sedatives/barbiturates, opioid analgesics, anticholinergics, heroin, amphetamines (anorexigenics), LSD, other synthetic drugs;
b) Provide direct estimates of multiple use of drugs;
c) Provide direct estimates of the number of people who are dependent on alcohol, tobacco and other drugs;
d) Assess perceptions regarding facility to obtain drugs, presence of drug trafficking or people under influence of alcohol and other drugs in their neighborhoods, and risk related to experimental and regular use of alcohol, tobacco and other drugs;
e) Estimate the number of people who have submitted to treatment for use of alcohol, tobacco and other drugs;
f) Estimate age at first use for specific drugs;
g) Estimate prevalence of 'binge drinking' episodes;
h) Estimate prevalence of adverse consequences from abuse of alcohol, tobacco and other drugs.

The **study population** was defined as all residents of private or collective households aged 12-65 years on the date of the survey. The following exclusions applied: residents of indigenous tribe villages, foreigners residing in Brazil, Brazilians who do not speak Portuguese, people with intellectual deficiency or other limitations that prevented responding to the survey, people in institutions such as prisons, hospitals, clinics, shelters, etc.

Data collection for the BHSU-3 was conducted between May and December 2015. Data was collected from 16,273 individuals interviewed in their households, using face-to-face and paper questionnaires. The study was approved by the Ethics Review Board of the Escola Politécnica de Saúde Joaquim Venâncio – FIOCRUZ (CAAE # 35283814.4.0000.5241). Consent was obtained from all selected individuals 18 years old or more by signing of an informed consent form. For those under 18 years of age (legally 'minors' according to Brazilian law), the informed consent form was signed either by a parent or guardian, while the individual signed an assent form.

2. Sample Design and Implementation

The study's geographic domains of interest were defined by the contracting agency as: national (1); urban and rural areas (2); macro-regions (5); set of capital cities of the 26 Brazilian states plus the Federal District (1); set of nine Metropolitan Regions (1); sets of large, medium and small size municipalities (3); and the set of municipalities located on the borders with neighboring countries (1). This complex set of overlapping domains of interest determined the need for a rather complex stratification strategy, described later in this section.

BHSU-3 used a stratified multi-stage probability sampling design – see for example Cochran (1977). Within the defined strata, municipalities were sampled in the first stage. Census enumeration areas (CEAs) were sampled in the second stage. Households were sampled within each sampled CEA. Finally, one eligible (12 to 65 years old) resident was sampled at random within each selected and participating household.

To facilitate comparison with the previous editions of the study (BHSU-1 and BHSU-2) – CEBRID (2002, 2006), each state capital and large city (≥ 200,000 inhabitants in 2010) was included in the sample with certainty, hence in fact turning into a selection stratum. Therefore, in these large municipalities, the design only had up to three stages of selection (CEA, household and resident). All the other municipalities were stratified into the five Brazilian macro-regions (North, Northeast, Southeast, South and Center-west). Within each macro-region, municipalities were further stratified in three groups: (1) border municipalities (with part of their area within 200 km from the terrestrial borders of Brazil); (2) municipalities within metropolitan regions; and (3) other municipalities. Within these three groups, municipalities were further stratified by size, according their population, as small (≤ 11,000 inhabitants), medium (11,000 to 200,000 inhabitants) or large (≥ 200,000 inhabitants). Thus, in total, the population was stratified into 138 strata.

In each stratum that was not a municipality (i.e. for the municipalities not included with certainty in the sample), the municipalities were the primary sampling units (PSUs) and were sampled with probability proportional to size (PPS), considering their population in 2010 as the size measure. In every selected municipality (certainty or sampled), CEAs were first sorted by their average household monthly income, and then sampled with systematic PPS (size was the number of private households).

The households were sampled using equal probability inverse sampling (e.g. see Vasconcellos et al., 2005). This method comprised sampling households at random sequentially for contact from an updated list of residential type addresses in the selected CEA. Instead of using a fixed size sample, it samples potentially eligible households until a stopping rule is reached. Sampling of new addresses to contact stops after having reached 10 complete interviews per selected CEA or having reached 50 contacted households irrespective of the number of complete interviews achieved per CEA. In the last stage, one eligible resident was selected with equal probability among the eligible residents in each participating household.

The total sample size for the study was calculated to estimate a minimum proportion (prevalence) of 2% with a relative error of 30%, confidence level of 95%. An average design effect of 1.5 was anticipated and considered when determining the sample size. Power allocation (with power = 3/4) was used to distribute the total sample size among the strata, using population as the size measure. After the allocation, the sample size reached 16,400 residents (or households) spread in 1,640 CEAs and in 351 municipalities.

When data collection concluded, a total of 16,273 eligible residents provided complete interviews, corresponding to an effective sample size that reached 99.2% of the required sample size. Table 1 provides some statistics on the survey data collection effort and outcome. It shows that the application of the sequential inverse sampling procedure within the selected CEAs resulted in selecting a total of 27,906 addresses which were screened to identify eligible households and residents. Of these, 4,036 could not be contacted after exhausting the attempts of the contact protocol (out of which, 24 addresses were not found). A total of 3,180 were vacant or used only as occasional/non-permanent residence (e.g. rental flats, beach houses, etc.), another 1,052 did not have eligible residents, and 5 had residents with contagious diseases during the interview period. Out of the remaining 19,633 eligible households, 3,057 refused to participate, 271 selected individuals refused to participate, and 32 interviews were started but not successfully completed. Therefore, a non-response rate of 17.1% [100*(3,057+271+32) / 19,633)] out of the confirmed eligible households was observed.

Table 1 – Selected summaries regarding data collection effort and outcome

Outcome of approach	Frequency	Percentage
Total	27,906	100.00
1 - Complete interview	16,273	58.31
2 - Interrupted interview	32	0.11
3 - Household refusal	3,057	10.95
4 - Refusal of selected resident	271	0.97
5 - Contagious disease in household	5	0.02
6 - Vacant or occasional use dwelling	3,180	11.40
7 - Ineligible dwelling	1,052	3.77
8 - Address not found	24	0.09
9 - Non-contact after 4 attempts	4,012	14.38

Table 2 provides some summary statistics on the workforce which was used to conduct the study. Therefore, each interviewer carried out an average of 57 interviews, and was responsible for data collection on 6 CEAs. Supervisors were on average overseeing the work of 7 interviewers.

Table 2 – Selected summaries regarding data collection workforce

Role	Females	Males	Total
Regional coordinator	9	18	27
Supervisor	15	28	43
Interviewer	150	135	285
Total	**174**	**181**	**355**

3. Survey Processing

Because data collection was carried out using paper questionnaires, these were then scanned to create digital records. Scanning was carried out using Fujitsu FI-7160 scanners, driven by a software called KaptureAll® which stores images as .TIFF files. The software also performs data recognition by interpreting manuscript characters written in cursive form, printed characters, bar codes and optical marks. It then processes the recognised characters and creates digital records corresponding to each scanned questionnaire (each printed questionnaire was 25 pages long). Overall close to 425 thousand pages were scanned, providing the data for the 16,273 complete survey questionnaires.

The digital records were then edited and imputed using CSPRO software from US Census Bureau. The edit and imputation rules are provided in full as part of the survey report, such that, in principle, anyone wishing to repeat the operation could do so, if provided with the original or ‘dirty’ version of the survey data, which is maintained as an integral part of the survey outputs for future reference. The ‘clean data’ (edited and imputed) was used for survey weighting and tabulation.

4. Survey Weighting

Four core strategies were used to deal with non-response. First, the sequential inverse sampling procedure was used to screen for eligible households, implying that any contact attempts would be recorded and that the final sample size in each selected CEA would match the required sample size per CEA (10 complete interviews). Second, the contact protocol ensured that households would only be declared ‘lost for interview’ after a substantial contact effort had been made (i.e., at least four visits on different days of the week and times of day to each selected and eligible address). Third, processes used for interviewer selection, training and supervision were designed to ensure that interviewer contacts with selected households were effective and lead to small refusal rates. Finally, calibration (Deville & Särndal, 1992) was applied to the basic sampling design weights to compensate for observed differential non-response.

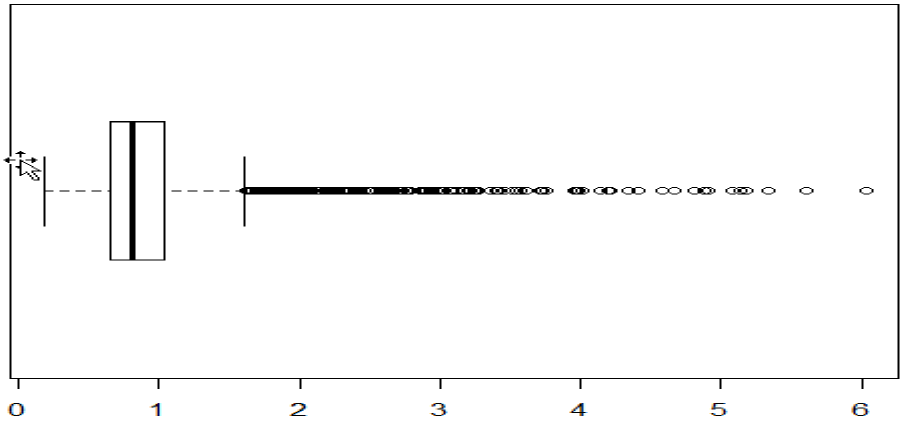
The basic design weights were calculated as reciprocals of each person’s sample inclusion probability. The details of these calculations is available from Bastos et al. (2017). These basic design weights were then adjusted by raking on marginal distributions to compensate for differential non-response by sex (2), age groups (18), macro-regions (5) and size of household (6). Figure 1 provides the detailed classifications used for the weight calibration. Population totals for the calibration were obtained from the Brazilian Continuous National Household Sample Survey for the third quarter of 2015.

Figure 1 – Classes of the variables used for calibration weighting

Sex	**Age groups**
Male	12 years
Female	13 years
	14 years
Macro-region	15 years
North	16 years
Northeast	17 years
Southeast	18 years
South	19 years
Center-West	20 to 24 years
	25 to 29 years
Household size	30 to 34 years
1 resident	35 to 39 years
2 residents	40 to 44 years
3 residents	45 to 49 years
4 residents	50 to 54 years
5 residents	55 to 59 years
6 or more residents	60 to 64 years
	65 years

The weight calibration factors (ratios between the final calibrated weights and the corresponding basic design weights) have their distribution plotted in Figure 2. This shows that a large part of the sample had their weights reduced (there was substantial differential nonresponse by sex, hence responding women had to be down-weighted) while a smaller portion had weights increased. Overall, the smallest weight adjustment factor was 0.18 and the largest was 6.03, which is a reasonable range of values for household surveys.

Figure 2 – Boxplot showing distribution of weight calibration factors



Survey weighting and analysis was carried out using tools from the *tidyverse* (Wickham, 2017), *survey* (Lumley, 2010 and 2018) and *srvyr* (Ellis & Lumley, 2018) packages of the R statistical software. The scripts developed and used for such operations is also maintained as an integral part of the survey outputs. In addition, a file was prepared for dissemination of the microdata in the form of an R 'calibration survey design object' to facilitate analysis. This 'object' which contains all the microdata as well as the relevant information about the anonymized survey strata and primary sampling units, as well as about the survey weights and calibration process used to produce them.

Given this 'object', survey analysis is very easy and safe: data users can easily replicate all or parts of the survey output and perform other analyses of interest simply by specifying the target analysis, and not having to worry about specifying details of the survey design used to obtain the data or about the weighting process. To enable analysis using other software, users would need to extract the survey microdata from this R object, and then use tools available in R to export the file to the desired format. In our own experience, usage of the *survey* and *srvyr* packages made analysis very fast and easy to implement, including for medical doctor members of the survey team.

5. Selected Survey Results

To illustrate what kinds of survey results were produced, Table 3 provides point (number and percentage) and 95% confidence interval estimates (for the percentages) for consumption of three substances: alcohol, marijuana and cocaine. These estimates refer to questions asking respondents about their previous use of the substance at some point in their lives (lifetime), during the last year (12 months) and during the last 30 days (30 days).

A book containing all the details of the survey design, implementation and results was prepared and delivered to contracting agency. This book also contained in an Appendix all relevant pieces of software code, ready for dissemination. The survey microdata was delivered together with the book, ready to be disseminated.

Table 3 – Number (in thousands) and percentage of users of selected substances – Brazil

Substance	Number (thousands)	%	95%CI for %
Alcohol			
Lifetime	101,615	66.4	64.8 - 68.0
12 months	65,943	43.1	41.8 - 44.4
30 days	46,036	30.1	28.9 - 31.3
Marijuana / Skank			
Lifetime	11,772	7.7	7.1 - 8.3
12 months	3,865	2.5	2.1 - 2.9
30 days	2,223	1.5	1.1 - 1.8
Cocaine			
Lifetime	4,683	3.1	2.7 - 3.4
12 months	1,340	0.9	0.7 - 1.1
30 days	461	0.3	0.2 - 0.4

Unfortunately, neither the book with the survey results and methodology nor the survey microdata were released for publication to date. The survey team is free to disseminate only partial results as traditional academic outputs (e.g. conference and journal papers, etc.) but not the complete survey report with its accompanying microdata set. A hard lesson was learned: in the future, involvement in carrying out a similar survey under contract for a government agency would need introduction of explicit contract conditions to avoid what we consider inappropriate withholding of the survey results.

Previous surveys carried out in Brazil on the same topic lack national coverage, proper documentation about methods, access to microdata and to scripts or code used in survey processing and analysis. Therefore, the survey team has made considerable efforts to ensure that this survey would be reproducible and have proper national coverage of the relevant target population, but politics interfered in its timely dissemination.

References

Bastos, F.I.P.M., Vasconcellos, M.T.L.D., De Boni, R.B., Reis, N.B.D. & Coutinho, C.F.D.S. (2017). III Levantamento Nacional Sobre O Uso De Drogas Pela População Brasileira. Rio de Janeiro: Fundação Osvaldo Cruz (FIOCRUZ) & Secretaria Nacional de Políticas sobre Drogas (SENAD).

CEBRID (Centro Brasileiro de Informações sobre Drogas Psicotrópicas). I Levantamento Domiciliar sobre o Uso de Drogas Psicotrópicas no Brasil: Estudo Envolvendo as 107 Maiores Cidades do País 2001. CEBRID, UNIFESP, 2002.

CEBRID (Centro Brasileiro de Informações sobre Drogas Psicotrópicas). II Levantamento Domiciliar sobre o Uso de Drogas Psicotrópicas no Brasil: Estudo Envolvendo as 108 Maiores Cidades do País 2001. CEBRID, UNIFESP, 2006.

Cochran WG. Sampling techniques. 3rd Ed. New York: John Wiley & Sons; 1977.

Deville JC, Särndal CE. Calibration estimators in survey sampling. Journal of the American Statistical Association 1992; 87(418): 376–382.

Ellis, G.F., Lumley T. (2018). Package ‘srvyr’. Documentation available from: https://cran.r-project.org/web/packages/srvyr/srvyr.pdf

Lumley, T. Complex Surveys: A Guide to Analysis Using R. Wiley Series in Survey Methodology. Hoboken: John Wiley & Sons, 2010.

Lumley, T. (2018). Package ‘survey’. Documentation available from: https://cran.r-project.org/web/packages/survey/survey.pdf

United Nations Statistics Division. Fundamental Principles of Official Statistics. 1994, https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

Vasconcellos MTL, Silva PLN, Szwarcwald CL. Sampling design for the World Health Survey in Brazil. Cadernos de Saúde Pública 2005; 21(S): S89-S99.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, et al. (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. PLoS Med 4(10): e296

Wickham, H. (2017). Package ‘tidyverse’. Documentation available from: https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf