WEIGHTING A NON-PROBABILITY WEB SAMPLE SURVEY CARRIED OUT DURING THE PANDEMIC

Marcelo Trindade Pitta[1]; Pedro Luis do Nascimento Silva[2]

[1] Brazilian Network Information Center (NIC.br)
[2] National School of Statistical Sciences

Abstract:

This paper presents the approach used to weight a non-probability web-panel sample survey carried out by the Brazilian Network Information Center (NIC.br) during the COVID-19 pandemic targeting Internet users aged 16 years and over. The weighting approach relied on combining data from a probability sample survey recently conducted by face-to-face interviewing to develop weights for units observed in the web-panel survey. The weighting attempted to mitigate selection and response biases affecting the observed panel sample.

Keywords: Survey; non-probability samples; Web panel; weighting.

1. Introduction:

Research on the use of non-probability samples for producing reliable statistics is on the rise. Several approaches have been proposed to overcome the difficulties arising from such samples, such as selection or response bias. At the same time, official and public statistics producers face increasing challenges arising from the demand for timelier and more disaggregated data, while coping with decreasing resources, increasing non-response rates, and more recently, challenges for collecting data via face-to-face surveys.

The COVID-19 pandemic has added urgency to the need for alternatives that enable such producers to meet the demands for timely, relevant, and accurate information, while facing the additional constraints on survey-taking imposed by social-distancing and other health protection protocols.

The Brazilian Network Information Center (NIC.br) is a non-profit organization responsible for the planning, evaluation, and monitoring the use of information and communication technologies (ICT) in Brazil. NIC.br has conducted yearly face-to-face national probability household sample surveys on the use of ICT (ICTHS) in Brazil since 2011.

In 2020, soon after social distancing restrictions took effect, NIC.br moved to experiment with alternative data collection modes to enable continuing to inform on ICT usage in Brazil during the pandemic. The ‘COVID-19 ICT Web Panel’ (Panel from now on) was a survey targeting Internet users aged 16 years and over carried out between March and May 2020 via a self-administered web-based questionnaire supplemented by telephone interviewing. The Panel aimed to collect information on Internet use during the pandemic. Internet users are individuals who used the network in the three months preceding the interview, following the definition of the International Telecommunication Union (ITU, 2014).

The frame used for selecting the sample for the Panel was obtained from a panel sample of individuals (recruited via the internet) maintained by a market research company that provides data collection services on demand. The frame had approximately 95 thousand panelists aged 16 and over.

The sampling plan used to select the initial sample of panel members was quota sampling, with quotas established considering sex, age, education, macro-region, and social class. Recruitment of the selected sample members for the Panel survey was carried out by the private marketing research firm. The selected units were approached and invited to participate in the Panel survey, but only about 14% responded after the standard follow-up measures. Both the recruitment strategy and subsequent low response implied that the sample available for analysis displayed non-negligible representation bias compared to the study's target population.

In addition to the web-based interviewing, a sample of complementary interviews conducted by telephone was added aiming to reach sub-populations that were more scarcely represented on the web-based sample. The register used to sample for this additional telephone sample was a list of numbers used in other marketing polls by the same marketing firm.

2. Weighting Methodology:

Surveys using quota sampling to select respondents are classified as non-probabilistic. Since sample selection probabilities for responding units are not known, standard unbiased estimators cannot be used to estimate from the available samples. Hence such strategies do not allow the calculation of sampling errors and may lead to selection biases. On the other hand, such surveys often have faster collection periods and are less costly to implement. Numerous recent studies proposed methods to attribute weighting structures to non-probability samples aiming to reduce estimation biases. Some of these studies use a traditional probability sample survey or a census as a reference for calculating the proposed sampling weights, and consider standard approaches for estimating sampling errors, confidence intervals, etc. – see for example Elliott and Valliant (2017) and Valliant (2019).

For weighting the COVID-19 ICT Web Panel, the reference probability sample survey used was the 2019 ICT Household Survey (CGI.br, 2020). The weighting process comprised three steps:

1. Estimation of the total contingent of Internet users aged 16 and over in Brazil who are represented by the respondents of the COVID-19 ICT Web Panel on the reference date; and
2. Estimation of the pseudo-inclusion-probabilities of these respondents, so that their reciprocals could be used as pseudo-weights for the Panel.
3. Calibration of the pseudo-weights by marginal totals of eligible internet users classified by sex, age group, schooling, social class and computer use.

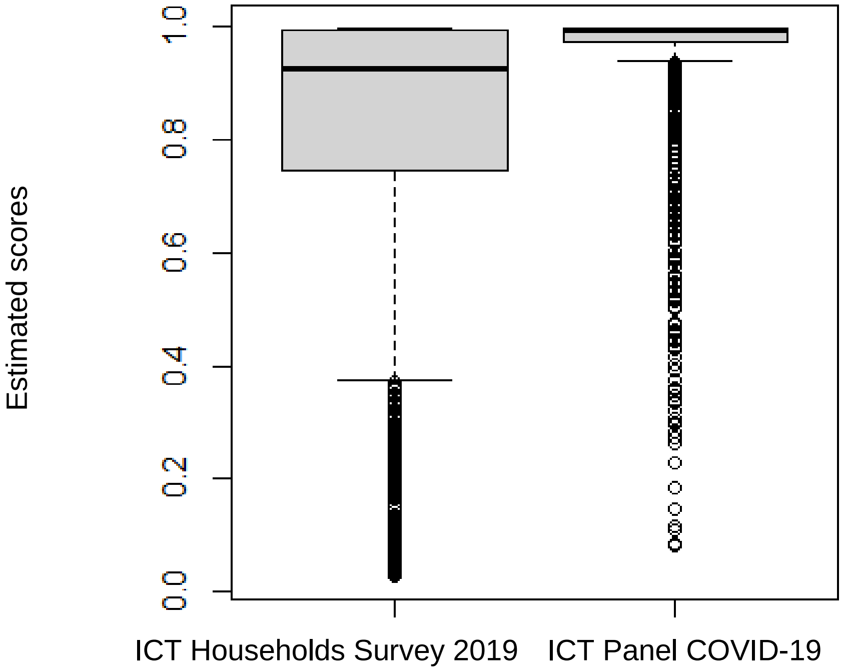
Step 1 - Estimating the contingent of Internet users 16 and over who are represented on the COVID-19 ICT Web Panel

Initial analysis of the Panel respondents showed that the available sample was not representative of the full range of eligible Internet users, with scarce representation of the lower socioeconomic stratum. This lead to an effort to identify what part of the Internet user population we could aim to provide inference for using the Panel respondents. The process of determining this 'survey population' comprised four stages:

a. Update of population totals for the 2019 ICT Household survey used as reference to reflect the population totals for the first quarter of 2020, by recalibrating the survey's sampling weights using population projections provided by IBGE.

b. Fit a logistic regression model with the variable "Indicator for Internet user" as the response and a set of socioeconomic variables observed on 2019 ICT Household survey as predictor variables. All predictors used were also collected on the COVID-19 ICT Web Panel. This model was used to estimate the propensity of an individual being an eligible Internet user. The final model included the following variables: sex, age group, schooling, social

class, and computer use indicator. The fitted model yielded an R^2 of 43.1% and a percentage of correctly classified individuals of 83%.
c. Use the fitted logistic regression model to estimate individual propensity scores for respondents of the COVID-19 ICT Web Panel.
d. Determine a cut-off point above which the propensity score for being an Internet user was sufficiently high, such that the propensity score distribution on the COVID-19 ICT Web Panel was not too thinly represented.
This cut-off point was defined after comparing plots (Figure 1) of the propensity scores from the reference sample and the Panel. Four cut-off points were considered, before selecting the 2/3 value (1/2; 2/3; 3/4; 4/5). This choice was guided by calculating the pseudo-weights for the Panel after each of the cut-off points was applied, then computing the final calibrated weights. The best-performing calibration corresponded to the cut-off point of 2/3 on the propensity scores, meaning that the part of the internet user population 'represented' by the Panel respondents is that of Internet users 16 years and over with a propensity score for being an internet user greater than or equal to 2/3 according to the fitted logistic regression model.
Figure 1: Propensity scores for Internet Users



Source: Cetic.br/NIC.br, ICT Household Survey 2019 (2020) and COVID-19 ICT Web Panel.

Step 2 - Estimation of inclusion pseudo-inclusion-probabilities for determining the pseudo-weights for respondents of the COVID-19 ICT Web Panel
This process consists of first estimating pseudo-inclusion-probabilities for the respondents of the COVID-19 ICT Web Panel (non-probability sample) using the 2019 ICT Household probability sample survey as the reference sample. Then we use their reciprocal as pseudo-weights for the Panel respondents, just as we would do in a traditional probability sample.
To estimate the pseudo-inclusion-probabilities, we pooled together the data from the two samples (reference and Panel). In the pooled dataset, we created a 'Panel membership indicator' variable which was used as response for a logistic regression model. The Panel respondents were allocated to a new 'stratum', separated from the strata used to select the reference probability sample, and received provisional sampling weights equal to one for the purposes of fitting the logistic regression model. We also assumed that there was no

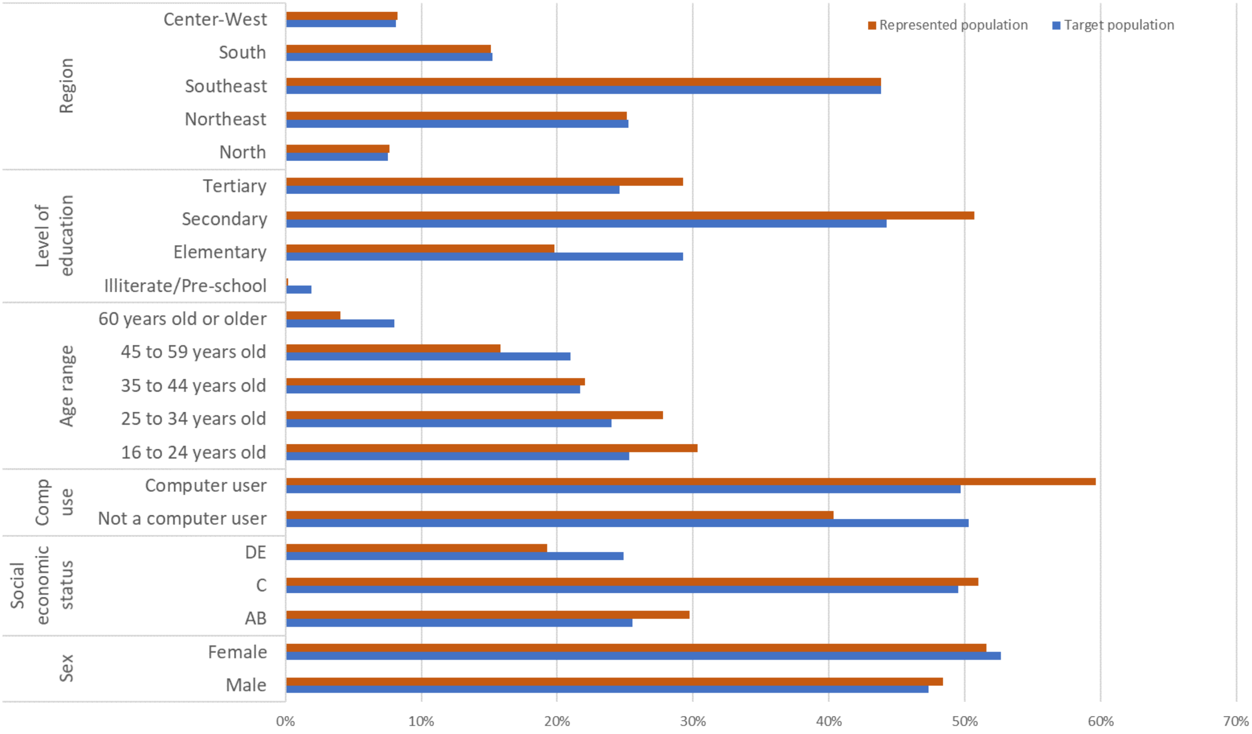
clustering of the Panel respondents, and therefore assigned each one a different PSU identifier.
We then fitted the logistic regression model taking into account the reference sample probability design that included stratification, clustering and differential weighting. The parsimonious model fitted used the following predictor variables common to the two surveys: stratum, social class, indicator for computer use, schooling and number of household members. The fitted model provided estimated propensities that we used as the pseudo-inclusion-probabilities for the respondents of the COVID-19 ICT Web Panel. The reciprocals of these pseudo- inclusion-probabilities were the initial pseudo-weights allocated to each respondent of the Panel.
Step 3 – Weight calibration
These initial pseudo-weights were calibrated for estimated marginal totals of the stratification variable, sex, age group, schooling, social class and computer use using raking calibration. Calibration totals were estimated from the reference probability sample, after having their weights updated to reflect the total population for the first quarter of 2020.

Following Valliant (2019), variance estimation used bootstrap methods. This approach enables considering the uncertainty due to model fitting steps carried out for subsamples taken from the main reference sample.

3. Results:

The approach used resulted in a weighted panel database that represents part of the desired target population. The cut-off of 2/3 defined above, restricted the represented population to 83% of the initial estimated target population – 101M of 121M. The represented population differs from the target population in some characteristics: it is composed of more educated, less vulnerable, younger and with more frequent access to computers (Figure 2).

Figure 2: Comparison of target population and represented population



Source: Cetic.br/NIC.br, ICT Household Survey 2019 (2020) and COVID-19 ICT Web Panel.

The web-based data collection revealed – for this panel frame – that the propensity of reaching the more vulnerable population is limited but, despite the limitations, it was possible to produce statistics about the dynamics of Internet use in the context of the pandemic for a substantial part of the Internet user's population aged 16 and over.

4. Discussion and Conclusion:

The use of innovative methods of data collection has been growing in the last few years. These new developments are the response from the official statistics and academic researchers for the increasing need for more disaggregated timely statistics in conjunction to the use of new data sources – not always collected / compiled according with statistics production concepts.

This article showed one application of these new methodologies that gave results that were limited to part of the intended target population. The methodology used was able to identify some coverage issues but no to correct for all of them, even with the use of a larger traditional probability sample survey as reference for the weighting. As an advantage, we can argue that the Panel was a lot cheaper and faster to run than a traditional probability face-to-face survey. On the other hand, the weighting methodology is complex and relies heavily on models – and good models are not always available. Dissemination of the results and methodology explanation are also points of attention: the target user audience often sets aside details about the part of the target population represented by the Panel, and many are unable to understand the implications of the modeling and variance estimation approaches used.

Considering the development of new methods, much is still to be done in the field of weighting non-probability surveys given the emergence of new data sources that can be used for statistics production. However, at present it is clear that these cannot be used without support from traditional probability sample surveys or censuses as references for benchmarking or bias correction.

References:

1. Dever, Jill A. 2018. "Combining Probability and Nonprobability Samples to Form Efficient Hybrid Estimates: An Evaluation of the Commom Support Assumption." In 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference, 15.
2. Elliott, Michael R. 2009. "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights." Survey Practice 2 (6): 1–7. https://doi.org/10.29115/sp-2009-0025.
3. Elliott, Michael R., and Richard Valliant. 2017. "Inference for Nonprobability Samples." Statistical Science 32 (2): 249–64. https://doi.org/10.1214/16-STS598.
4. Little, Roderick J. A., and Donald B. Rubin. 2002. Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics.
5. Valliant, Richard. 2019. "Comparing Alternatives for Estimation from Nonprobability Samples." Journal of Survey Statistics and Methodology, no. March. https://doi.org/10.1093/jssam/smz003.
6. Valliant, Richard, and Jill A. Dever. 2011. Estimating Propensity Adjustments for Volunteer Web Surveys. Sociological Methods and Research. Vol. 40. https://doi.org/10.1177/0049124110392533.