

<1>Methodological Report ICT Panel

<2>Introduction

The Brazilian Internet Steering Committee (CGI.br), through the Regional Center for Studies on the Development of the Information Society (Cetic.br), a department of the Brazilian Network Information Center (NIC.br), presents the methodology of the ICT Panel survey with Brazilian Internet users.

The COVID-19 pandemic has substantially affected the work of national statistical institutes and other data producers around the world, particularly among Latin American countries. In response, the Cetic.br|NIC.br innovated in the production of indicators using web panel surveys.

This experience proved to be very relevant, and from 2021 on this panel became a new form of investigation used by Cetic.br|NIC.br in the production of statistics about information and communication technologies (ICT) for Internet users. This new survey tool will allow more regular monitoring of indicators on the current scenario and the evaluation and collection of information on new topics and aspects of Internet use in Brazil.

<2>Survey objectives

The ICT Panel aims to collect information on topics associated with the use of technologies by Internet users across Brazil.

<2>Target population

The target population of the survey is made up of Internet users 16 years old or older in Brazil. Internet users are considered to be individuals who have used the Internet in the three months prior to the interview, according to the methodological recommendation of the International Telecommunication Union (ITU, 2014).

<2>Unit of analysis and reference

Individual Internet users 16 years old or older.

<2>Areas of interest for analysis and dissemination

For the units of analysis and reference, the results are reported by domains defined based on the variables and levels described below:

- **sex:** Corresponds to the division into male or female;

- **level of education:** Corresponds to the division into Elementary Education, Secondary Education, and Tertiary Education¹;
- **age group:** Corresponds to the division into the ranges of 16 to 24 years old, 25 to 34 years old, 35 to 44 years old, 45 to 59 years old, or 60 years old and older;
- **region:** Corresponds to the regional division of Brazil, according to criteria of the Brazilian Institute of Geography and Statistics (IBGE), into the macro-regions North, Northeast, Southeast, South, and Center-West;
- **social class:** Corresponds to the division into AB, C or DE, according to the Criterion of Economic Classification Brazil (CCEB), of the Brazilian Association of Research Companies (Abep).

<2>Data collection instruments

<3>Information on collection instruments

Data was collected through a structured questionnaire, with closed-ended questions and predefined answers (single or multiple answers) and, in some cases, open-ended questions analyzed using text analysis methodologies. The questionnaire was self-administered without interviewer mediation.

<3>Themes

The survey investigated topics associated with activities carried out on the Internet and the devices used to go online, based on the reference indicators validated by the ICT Households survey, carried out by CGI.br, and usage indicators related to relevant topics at the time of the survey.

<2>Sampling plan

<3>Survey frame and sources of information

For the sample design of the ICT Panel, a panel of respondents previously recruited by a market research company was used as a primary source, which included panelists 16 years old or older. Panel participants were recruited through a series of channels and methods, including probabilistic research, careful selection of recruitment partners and partnerships with communication and media outlets, continuous evaluation of the response rate of panelists, focus on recruitment actions for specific audiences according to customer needs, and a recruitment process in accordance with the highest market standards. In addition, it is important to mention that panel participants received incentives to respond to the surveys.

¹ The levels of education presented refer to the aggregation of the levels of education declared that are equal to or lower than the categories presented, that is, Elementary Education includes no formal education, completion of preschool, and complete or incomplete Primary Education. The same applies to Secondary and Tertiary Education.

<3>Sample size determination

The sample was sized according to information needs, available resources, and the time frame in which the survey needed information for analysis. The total number of interviews of each wave of the ICT Panel is presented in the "Data Collection Report".

<3>Methods for obtaining the sample

A quota sampling plan was used to obtain the sample of respondents. The quotas were established considering sex, age group, education, macro-region, and social class, and were used to indicate the individuals to be approached for collection through the Web. The sample allocation according to the established criteria was disproportionate to the information contained in the survey frame, given the need to meet the demand for information for all domains of interest. The sample resulting from this collection effort is hereinafter referred to as the ICT Panel.

<2>Field data collection

<2>Data collection method

Data was collected using computer-assisted web interviewing (CAWI), through a programmed and self-administered online questionnaire.

<2>Data processing

<3>Weighting procedures

Sample surveys that use quotas to select respondents are classified as non-probabilistic. Typically, such strategies do not allow the calculation of sampling errors and may carry some selection biases, as the selection probabilities of each unit are not known. Non-probability approaches are common in opinion, voting intention, product evaluation, and customer satisfaction polls. Such surveys generally have shorter collection periods and lower budgets, and do not follow the usual rigor of probability sampling methods to obtain samples.

Recently, the growing demand for more frequent and disaggregated information, in addition to the emergence of new sources of information (Big Data), has promoted numerous studies that try to assign weight structures that allow mitigation of the biases of databases collected by non-traditional methods. In general, such studies use a sample survey or the traditional census as a reference for calculating weights for non-probability sample observations, which then serve as a basis for obtaining estimates of accuracy, confidence intervals, etc. As examples of studies along these lines, Elliott and Valliant (2017) and Valliant (2019) can be cited.

For the ICT Panel, the last ICT Households survey (a probability survey), whose data has been publicly available, was used as a primary reference. Additionally, the results of ICT Households were updated based on the population of the Continuous National Household Sample Survey (Continuous PNAD), from IBGE, referring to the last quarter released. The process of weighting ICT panel respondents was divided into two stages:

1. Estimation of the total contingent of Internet users aged 16 years old or older in Brazil at the reference date of the survey who are represented by the respondents of the ICT Panel.
2. Estimation of pseudo-probabilities of selection of these respondents for ICT Panel weighting.

<4>Step 1 - Estimation of the contingent of Internet users represented in the ICT Panel

The ICT Households survey (last available result), based on a traditional probabilistic approach, allows estimation of the total number of Brazilians 10 years old or older who are Internet users². The ICT panel, on the other hand, includes respondents 16 years old or older who are Internet users, according to internationally adopted parameters (ITU, 2014). In order for the two samples to be comparable, the results of the ICT Households survey referring to the same age group were filtered.

Because the construction of the set of respondents of the ICT Panel is not done in a probabilistic way, it is not possible to consider it a priori as representative of the overall population of Internet users 16 years old and older. To estimate the contingent of the population that is represented by the respondents of the panel, an estimation procedure based on propensity scores was adopted. In this methodology, the propensity scores for being an Internet user were initially calculated according to socioeconomic variables based on the last available ICT Households survey. Then, this same model was used to estimate propensity scores for ICT Panel respondents.

By comparing the distribution of propensity scores for the ICT panel with the one verified in the last ICT Households survey, it was possible to determine which part of the population (or all of it) of Internet users 16 years old or older from the last ICT Households survey could be considered represented by the ICT Panel respondents. This is equivalent to estimating the panel's coverage error relative to the target population initially considered for the survey.

From this comparison, a cut-off point was established that determined, on the basis of the last ICT Households survey, the set of investigated units whose propensity scores seemed well represented by the ICT Panel respondents.

This procedure aimed to determine the population represented by the ICT Panel and consider, for the purpose of comparing results, that this same population was among Internet users in the last ICT Households survey.

The process of determining this population followed four steps:

- I. Population totals update of the latest ICT Households survey using the totals for the last quarter released by the Continuous PNAD carried out by IBGE.
- II. Logistic regression model adjustment with "Internet user" as the response variable and a set of socioeconomic factors common to the ICT Households survey and the ICT Panel as explanatory variables.

² More details at the Cetic.br/NIC.br website.

http://cetic.br/media/microdados/256/ticdom_2019_relatorio_metodologico_v1.0.pdf

This model was then used to estimate the propensity scores for being an Internet user for the respondents of the latest ICT Households survey.

- III. Estimation of propensity scores for ICT panel respondents based on the model adjusted with data from the last ICT Households survey.
- IV. Determination of the cut-off point that separates in the samples of both the last ICT Households and the ICT Panel the portion of the population to be represented.

Step I. Updating population totals from the latest ICT Households survey for the quarter most recently released by the Continuous PNAD

The objective of this step was to update the population estimates for the population 10 years old or older from the last ICT Households survey, based on data released by IBGE in the last Continuous PNAD. The calculations updated the total population 10 years old or older from the estimates reported in the Continuous PNAD microdata. Then, following the same percentage distribution of the calibrators used in the last ICT Households survey, the weights of the survey were updated again according to the new totals of the marginal distributions of the variables considered in the calibration.

Step II. Adjustment of the logistic regression model for the variable "Internet user" among respondents 16 years old or older in ICT Households

This step sought to make a quality estimate of the probability of an individual being an Internet user based on the socioeconomic variables observed in the last ICT Households survey that were also available in the ICT Panel. Several models were tested to obtain a parsimonious model that gave good results in estimating Internet users, according to Formula 1.

Formula 1

$$\log \left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \right) = \alpha + \beta X_i$$

Y_i is an indicator variable, taking the value 1 if individual i is an Internet user, and the value 0 otherwise.

X_i is a vector with the values of the explanatory variables (gender, age group, education, etc.) of individual i

$P(Y_i = 1)$ represents the probability that an individual is an Internet user

α and β are parameters of the model, to be estimated

The estimates for $P(Y_i = 1)$ are provided by the expression:

$$\hat{P}(Y_i = 1) = \frac{\exp(\hat{\alpha} + \hat{\beta}\mathbf{X}_i)}{1 + \exp(\hat{\alpha} + \hat{\beta}\mathbf{X}_i)}$$

These are the so-called propensity scores considered in the methodology, and $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the parameters obtained based on the adjusted model.

The adjusted model used as options for independent variables (\mathbf{X}) only information that was present in both sources: ICT Households and ICT Panel. The final model is presented in the “Data Collection Report.”

Step III. Estimation of propensity scores for ICT Panel respondents

Based on the model adjusted with the data of the last ICT Households survey, the propensity scores for the set of respondents of the ICT Panel were estimated. Then, the propensity score distributions of the ICT Households sample were compared with the scores of the ICT Panel sample for Internet users. The results are presented in the “Data Collection Report.”

Step IV. Determination of the common support population of ICT Households and ICT Panel

If the distributions of the scores obtained in both surveys were different, we sought to identify a section of the sample of Internet users of the ICT Households survey that was more similar to the set of respondents of the ICT Panel. The choice of this cut-out took into account the observation of the score distributions and the variability in weights that were assigned to the respondents of the panel, for a set of possible cut-outs of propensity scores to be an Internet user. This assessment was made by estimating the weights of the ICT Panel respondents according to alternative situations:

- I. Selection of all respondents from both surveys, without cut-out; and
- II. Selection of respondents from both surveys who had propensity scores greater than or equal to a specific fraction.

The fraction was chosen to use "chunks" of the pool of respondents that were comparable from both surveys, which determined a common support population for them.

For each fraction option (common support population determinant), pseudo-weights were estimated for the ICT Panel respondents³, and the cut-outs considered were evaluated according to the variability in weights. The researchers opted for the cut-out in which the resulting weights had the smallest amplitude in the distribution of absolute values, and the calibration factors (ratio between calibrated weights and the basic weights) had an average closer to 1. This is desirable because, in this situation, the calibrated weights were closer to the weights initially established by the pseudo-weight estimation methodology. The results of this stage are presented in the “Data Collection Report.”

³ The methodology for estimating the pseudo-weights is presented in the next section.

<4>Stage 2 - Estimation of pseudo-probabilities of inclusion to determine the weights of the respondents of the ICT Panel

The pseudo-weight estimation process consisted of estimating the pseudo-probabilities of inclusion of the respondents of the ICT Panel (non-probability sample) in the last ICT Households survey (probability sample), as well as using their reciprocals as weights, as in a traditional probability sampling survey. With this, the probability of an individual being selected and responding to the ICT Households survey was estimated based on independent variables (**X**) related to the profile of the respondents, considering that, given these variables (**X**), the probabilities of inclusion were independent from the variables of interest of the survey.

To estimate pseudo-probabilities, data from both samples (probability and non-probability) were stacked into a single database, and inclusion probabilities were estimated using a logistic regression model that took into account the sampling plan of the reference probability survey.

For this study, different possibilities were considered according to the population cut-outs established in the previous section. Such cut-outs aimed to identify the common support population of the two studies by evaluating the weights obtained, as suggested by Valliant (2019).

The pseudo-probability estimation process used the following steps:

- I. Union of cases in the same database (stacking), ensuring the presence of common independent variables (**X**), collected according to the same criteria and concepts. In this basis, an indicator variable **Z** was created, whose value was 1 for ICT Panel respondents (non-probability sample) and 0 for ICT Households respondents (probability sample).
- II. Creation of a column of weights in this file, which considered the weights coming from the probabilistic sample (for its cases) and a weight equal to 1 for the cases of the non-probability sample.
- III. Fitting a logistic regression model having the variable **Z** as a response, taking into account the sample design of the ICT Households survey, to estimate the probabilities of including the ICT Panel respondents in the probability sample.

In the model adjustment, the ICT Panel sample was considered as a separate stratum, and each respondent in that sample was considered to be a distinct primary sampling unit (PSU). This procedure was necessary to designate the structure variables of the sampling plan for the stacked data file of the two surveys.

The most parsimonious model considering the independent variables (**X**) available and common to the two databases is presented in the "Data Collection Report" of the ICT Panel survey. Based on this model, the pseudo-probabilities of inclusion of the ICT Panel respondents in the last ICT Households survey were estimated. The reciprocals of these pseudo-probabilities were the initial weights allocated to each respondent of the ICT Panel.

These initial weights were calibrated for the total estimated marginals of the ICT Households survey variables. The weights thus calibrated were considered for the estimation of all indicators of results of interest and associated precision measures.

<3>Estimation of variance

The estimation process assigned each respondent of the ICT Panel a weight that treated them as a research participant with a sample plan equal to that of the last ICT Households survey, but with a smaller total sample size. In this way, it was possible to estimate variances and margins of error. According to Valliant (2019), there are two possibilities for variance estimation: estimation considering the sample as simple random with replacement, and estimation based on the replication method.

The second method (estimation based on the replication method) has the advantage of considering the estimation of the model and the pseudo-probabilities of inclusion of subsamples taken from the main sample. This allows to include in the estimation of variance the variability associated with the estimation of this model, and this is why this method was chosen. The steps of the procedure were as follows:

- I. From the common (stacked) base used to estimate the pseudo-probabilities model, 200 bootstrap samples were selected using the *as.svrepdesign* function of the survey package of the program R, considering the sample plan.
- II. For each of these 200 replicas, the model was adjusted to estimate pseudo-probabilities of inclusion and corresponding pseudo-weights.
- III. The pseudo-weights of each replica were calibrated and stored for variance estimation.

The variance of estimates of indicators of interest was estimated using Formula 2.

Formula 2

$$\hat{V}(\hat{y}) = \frac{R-1}{R} \sum_{r=1}^R (\hat{y}_r - \hat{y})^2,$$

\hat{y} is the estimate of indicator y obtained using the ICT Panel COVID-19 sample (with 2,511 respondents);

\hat{y}_r is the estimate of indicator y in replica r ;

$R = 200$ is the total of bootstrap replicas formed.

<2>Data dissemination

The results of the ICT Panel are presented according to the classification variables described in the item “Domains of interest for analysis and dissemination”. In some results, rounding caused the sum of partial categories to be different from 100% for single-answer questions. The sum of frequencies in multiple-answer questions usually exceeds 100%. It is worth mentioning that, in the tables of results, hyphens (-) are used to indicate that no respondents chose that item. Furthermore, since the results are presented without decimal places, cells with zero value mean that there was an answer to the item, but it was explicitly greater than zero and lower than one.

The results of this survey are published online and made available on the Cetic.br/NIC.br website (<http://www.cetic.br>). The tables of proportions, totals, and margins of error for each indicator are available for download on the website. For comparison with previous editions of the ICT Households survey, the survey tables are provided considering the same cut-out used in the ICT Panel, when necessary, separating the common support population in the ICT Households survey.

<2>References

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249–64.

International Telecommunication Union. (2014). *Manual for measuring ICT access and use by households and individuals 2014*. http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ITCMEAS-2014-PDF-E.pdf

Valliant, R. (2019). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 231–263.