Rescuing non-probability samples: an experience with model-based inference from a web-panel survey

Marcelo Pitta & Pedro Silva

23 February 2022
About CETIC.br and context

The impact of COVID pandemic

Brief review of approaches for non-prob. samples

The COVID-19 web panel of internet users

Weighting and estimation

Some results and discussion
CETIC.BR AND CONTEXT
CETIC.BR
16 Years Producing ICT Data for Policymaking and Research

Multistakeholder model for Internet governance in Brazil

UNESCO 36th General Conference approves Cetic.br as a Category 2 Center

Source: CGI.br/NIC.br (2021)
## CETIC.BR
### Core Set of Surveys

<table>
<thead>
<tr>
<th>Survey</th>
<th>Collection mode</th>
<th>Periodicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT Households</td>
<td>F2F</td>
<td>Annual</td>
</tr>
<tr>
<td>ICT Kids Online</td>
<td>F2F</td>
<td>Annual</td>
</tr>
<tr>
<td>ICT in Education</td>
<td>F2F and CATI</td>
<td>Annual</td>
</tr>
<tr>
<td>ICT in Health</td>
<td>CATI</td>
<td>Annual</td>
</tr>
<tr>
<td>ICT Enterprises</td>
<td>CATI</td>
<td>Biannual</td>
</tr>
<tr>
<td>ICT Nonprofit Organizations</td>
<td>CATI</td>
<td>Irregular</td>
</tr>
<tr>
<td>ICT Providers</td>
<td>CATI</td>
<td>Biannual</td>
</tr>
<tr>
<td>ICT Electronic Government</td>
<td>CATI</td>
<td>Biannual</td>
</tr>
<tr>
<td>ICT Public Access Centers</td>
<td>CATI</td>
<td>Irregular</td>
</tr>
<tr>
<td>ICT in Culture</td>
<td>CATI</td>
<td>Biannual</td>
</tr>
<tr>
<td>ICT Panel</td>
<td>CATI</td>
<td>Twice a year</td>
</tr>
</tbody>
</table>
PROBABILITY SAMPLING

- Standard approach for surveys producing public and official statistics
- Every sampled unit has its inclusion probability calculated and used to derive its sampling weight
- Standard methods available for unbiased estimation and assessment of precision / accuracy
- Production process often more costly and time consuming
Demand for more timely and disaggregated indicators

Demand for new indicators & topics

While at the same time

Reduction of the resources available for supporting the traditional statistics production process

Increasing survey non-response, irrespective of collection mode
NON-PROBABILITY SAMPLING

- Standard approach for opinion polls and market research
- Unit inclusion probabilities unknown
- Often enable cheaper and faster production process
- Coverage and selection bias issues likely
ICT HOUSEHOLDS

- Traditional **face-to-face** household sample survey, carried out yearly
- Since 2012, uses standard probability sampling
- Stratified multi-stage sampling, with 36 geographical strata
- Sample of **33,000** households, from 2,200 census enumeration areas
- In 2019, achieved **23,490** complete interviews (71%)

2. [https://www.amstat.org/asa/files/pdfs/ASASoreOrderForm.pdf?hkey=a420f2db-cdfdc-b4f1-5fa4c9364a1f](https://www.amstat.org/asa/files/pdfs/ASASoreOrderForm.pdf?hkey=a420f2db-cdfdc-b4f1-5fa4c9364a1f)
IMPACT OF THE COVID PANDEMIC
IMPACT OF THE PANDEMIC

Traditional face-to-face interviewing impossible for most surveys during 2020 and early 2021 \(\rightarrow\) affected ICT Households survey the most

Telephone interviewing with enterprises, schools and health facilities (other target populations of Cetic.br surveys) became more difficult \(\rightarrow\) increased non-response rates

Themes of study – access, use and appropriation of Internet technologies – are correlated to the new ‘ways of life’ and to the possibility of becoming respondents to Cetic.br surveys (of any kind)
Cetic.br developed a contingency plan to collect and publish ICT statistics based on alternative data collection methods:

- **Web panel** survey with Internet users
- **Telephone interviewing** for the ICT Households survey (2020 edition)

  » New methodological approaches
  » Data collection via CATI and WEB
WEB PANEL OF INTERNET USERS
Nonprobability Sample

ICT COVID-19 Panel (Web panel survey)

Target population
Internet users aged 16+ in Brazil

Target domains for estimation
Sex (2), Education (3), Region (5), Age group (5) and Socioeconomic status (4) – not cross-classified

Frame
Web panel of individuals obtained from market research companies, complemented by telephone lists (to reach population with lower SES/education)

Sampling design
Quota sample based on region, sex, age group, SES and education
WEB PANEL OF INTERNET USERS

The sample

- ~91 k panel members contacted
- ~2.5 k panel members responded
- Evidence of selection bias: respondents included larger portions of users who are more educated and of higher socioeconomic status
- Crude unweighted estimates likely to be severely biased
- Fieldwork period: June 23rd to July 8th 2020
HOW TO ESTIMATE FROM CETIC.BR WEB PANEL
ESTIMATION FROM NON-PROBABILITY SAMPLES

Beaumont (2022)

- Design-based approaches
- Model-based approaches
  - Calibration
  - Statistical matching
  - Inverse probability weighting
  - Small area estimation through the Fay-Herriot model
Not feasible in the case at hand:
- Key survey variables not available in contemporary probability sample
- Overlap between probability and panel samples is very small and likely null
ESTIMATION FOR CETIC.BR WEB PANEL

Model-based approaches
- Calibration
- Statistical matching
- **Inverse probability weighting (IPW)**
- Small area estimation through the Fay-Herriot model

**IPW** adopted based on simplicity and timeliness for implementation
- Goal was to estimate for a wide range of target indicators from the panel
- No modelling required for specific target response / survey variables
ESTIMATION FOR CETIC.BR WEB PANEL

Inverse probability weighting (IPW)

- According to Beaumont (2022) a core advantage is:
  "Simplify the modelling effort when there are many variables of interest (only one participation indicator to model)"

- Requires modeling the relationship between an indicator of inclusion in the panel (non-probability) sample and covariates \( x \) available for both the panel and a reference probability sample.
Inverse probability weighting (IPW)
Common support assumptions – Valliant & Dever (2011)

Reference probability sample should:

› Be carried out for the same reference period;
› Collect the \( x \) variables in the same way (overlapping questions);
› Cover the same target population;
› Have no intersection of respondents with the panel survey.
Reference probability sample survey: ICT Households 2019

**Target population**
Permanent private households and their residents aged 10+ in Brazil

**Total sample size**
~ 23,500 interviews (households and individuals – one per household)

**Sampling frame**
IBGE 2010 census enumeration areas database

**Sample design**
Stratified multi-stage sampling of households and residents

**Fieldwork period**
October 2019 till March 2020
Inverse probability weighting (IPW)

Common support assumptions

› Be carried out for the same reference period (not contemplated);

› Collect the $x$ variables in the same way (overlapping questions) (contemplated);

› Cover the same target population (domain in the reference sample);

› Have no intersection of respondents with the panel survey (highly unlikely that a respondent took part in the two surveys).
Reference period support

• Updated the weights for ICT Households 2019 eligible respondents using population estimates from IBGE’s quarterly labour force survey

Target population support

Selected ICT Households 2019 respondents who would be eligible for web panel membership (internet users aged 16+, n=19,231)
METHODS
METHODS

IPW Approach Implemented in 4 Steps

Estimate pseudo-inclusion probabilities $p_i$ for the non-probability sample units via logistic regression model.

Use reciprocals ($1/p_i$) as basic pseudo-weights, considering subpopulations defined by propensity scores from model for indicator of Internet use.

Calibrate pseudo-weights for estimated marginal totals for domains (sex, education, age group, SES and computer user status).

Estimate variances using bootstrap.
RESULTS
Data Collection Statistics

<table>
<thead>
<tr>
<th>Collection mode</th>
<th>Contacts attempted</th>
<th>Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web</td>
<td>18,734</td>
<td>2,539</td>
</tr>
<tr>
<td>Telephone</td>
<td>72,794</td>
<td>88</td>
</tr>
<tr>
<td>Total</td>
<td>91,528</td>
<td>2,627</td>
</tr>
</tbody>
</table>

- ~63k telephone contact attempts were not answered
- CATI collection mode tried to reach more vulnerable population
- Signs of misrepresentation of low-income internet users
MODELS

Model for the indicator of participation in the panel survey

\[
\log \left( \frac{P(Z_i = 1|x_i)}{1 - P(Z_i = 1|x_i)} \right) = \log \left( \frac{p_i}{1 - p_i} \right) = \alpha + \beta x_i
\]

\(Z_i\) indicator variable for unit \(i\) responding to panel sample:
- =1 for panel (non-probability) sample,
- =0 for reference probability sample

\(x_i\) covariates for unit \(i\) (collected in both surveys)
\(\alpha\) and \(\beta\) model parameters

Covariates: sex, age group, SES, education, location and indicator of computer use
Estimates using pseudo-weights appeared biased towards “more connected” Internet users (considering pandemic reached Brazil in March)

<table>
<thead>
<tr>
<th>Internet users 16 years and older (%)</th>
<th>ICT Households</th>
<th>Web Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devices used in the past three months</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>Desktop computer</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>Notebook</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Tablet</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>Cellphone/Smartphone</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Game</td>
<td>36</td>
<td>59</td>
</tr>
<tr>
<td>TV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of connection used to access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet through cellphone/smartphone</td>
<td>80</td>
<td>76</td>
</tr>
<tr>
<td>3g/4g</td>
<td>90</td>
<td>94</td>
</tr>
</tbody>
</table>
RESULTS

- Various alternative calibration adjustments of pseudo-weights were tried, none seemed to correct the apparent bias.

- Calibration using raking with different combinations of variables, even cross-classifications, were not able to correct the bias.

- Solution adopted:

  Fit a model to predict Internet use propensity and identify the part of the internet user population represented by the panel respondents, if such a model exists.
MODELS
Modelling the Propensity of Being an Internet User

- Fitted with data from the reference probability sample (ICT Households 2019)
- Used covariates that were also collected in the panel sample
- Covariates: sex, age group, education, SES and computer use
Modelling the Propensity of Being an Internet User

Internet user propensity model statistics:

- $R^2 = 0.431$;
- Proportion of correctly classified individuals = 83%.
RESULTS

Modelling the Propensity of Being an Internet User

Comparison of internet user propensities across surveys
RESULTS
Determining the ‘representativity’ of the panel survey

- Different propensity distributions in the reference and panel samples suggest the need to identify a ‘user population’ in common

- Pseudo-weights calculated for four different cut-offs:
  - Select all eligible respondents from both surveys;
  - Select eligible respondents from both surveys with propensity scores larger than a threshold – $\frac{4}{5}$, $\frac{3}{4}$ and $\frac{2}{3}$ cut-off values considered.
RESULTS
Defining the ‘common user population’ covered by the panel survey

Calibration factors to pseudo-weights used to choose the threshold

<table>
<thead>
<tr>
<th>Web panel weight calibration factor statistics</th>
<th>Min.</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole sample</td>
<td>0.120</td>
<td>0.283</td>
<td>0.369</td>
<td>0.581</td>
<td>0.630</td>
<td>5.840</td>
</tr>
<tr>
<td>Sample with $q_i \geq 2/3$</td>
<td>0.078</td>
<td>0.484</td>
<td>0.637</td>
<td>0.988</td>
<td>1.038</td>
<td>12.518</td>
</tr>
<tr>
<td>Sample with $q_i \geq 3/4$</td>
<td>0.063</td>
<td>0.623</td>
<td>0.767</td>
<td>0.918</td>
<td>0.930</td>
<td>4.380</td>
</tr>
<tr>
<td>Sample with $q_i \geq 4/5$</td>
<td>0.087</td>
<td>0.447</td>
<td>0.554</td>
<td>0.729</td>
<td>0.791</td>
<td>5.604</td>
</tr>
</tbody>
</table>
RESULTS
Defining the ‘common user population’ covered by the panel survey

Common user population (~101 M) while target population (~ 121 M)

More educated, younger, higher SES and with more computer users

2,511 panel interviews used (from the initial sample of 2,627)
SUMMARY OF THE PROCEDURE

- Population projections first quarter 2020
- ICT Households Survey 2019
- Updated ICT Households Survey population weights
- ICT Web Panel Survey
- Modeling of ‘Internet users’
- Propensity scores
  - Whole sample
  - \( q_i \geq 2/3 \)
  - \( q_i \geq 3/4 \)
  - \( q_i \geq 4/5 \)
- Estimation of pseudo-weights + calibration => evaluation of respondent’s representativeness
- ICT Web Panel with ‘survey weights’

Common support reference sample
SOME RESULTS
## Comparing Common and Full Population

<table>
<thead>
<tr>
<th>Sex</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age range</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 to 24 years old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 to 34 years old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 to 44 years old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45 to 59 years old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 years old or older</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center-West</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Target population
- Represented population
COMPARING COMMON AND FULL POPULATION

Social class
- AB
- C
- DE

Computer use
- Not a computer user
- Computer user

Level of education
- Illiterate/Pre-school
- Elementary
- Secondary
- Tertiary

Target population
Represented population
Some highlights

<table>
<thead>
<tr>
<th>Activity</th>
<th>ICT Households 2018*</th>
<th>ICT Households 2019*</th>
<th>ICT Panel 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Look up for information on health or healthcare services</td>
<td>54</td>
<td>55</td>
<td>72</td>
</tr>
<tr>
<td>Used e-government services</td>
<td>33</td>
<td>36</td>
<td>54</td>
</tr>
<tr>
<td>Look up for financial information, making payments and other financial transactions</td>
<td>36</td>
<td>42</td>
<td>71</td>
</tr>
</tbody>
</table>

*ICT Households refers to the percentage of households with ICT access.
*ICT Panel refers to the percentage of people in the panel with ICT access.

SOME INDICATORS OBTAINED FROM PANEL
Some highlights

- Completing school activities or research: 45, 43, 46
- Taking distance learning courses: 13, 16
- Studying on the Internet on one's own: 42, 45, 55
- Completing work activities: 40, 41, 49

- ICT Households 2018* - 45, 43, 46
- ICT Households 2019* - 13, 16, 42
- ICT Panel 2020 - 40, 41, 49
DISCUSSION
EVALUATING PROS & CONS OF WEB PANEL

Pros

- **Feasibility during pandemic:** data collected without face-to-face interviews
- **Speed:** from planning to publication, took less than two months
- **Cost:** data collection much cheaper than that of traditional face-to-face survey
**Cons**

- **Coverage:** web panel is not representative of the full target population
- **Coverage:** issues remain, even after using a probability survey as reference for weighting
- **Robustness:** approach is model-dependent, and good models may not always be available
- **Complexity:** for developing & explaining methodology
Thank you all!

www.cetic.br
marcelopitta@nic.br
www.ence.ibge.gov.br
pedronsilva@gmail.com

Access the survey in English/Portuguese:
https://cetic.br/en/publicacao/painel-tic-covid-19/
REFERENCES


REFERENCES
