COVID-19 High-Frequency Phone Survey (HFPS) in Latin America

Technical Note on Sampling Design, Weighting, and Estimation*

The COVID-19 High-Frequency Phone Survey (HFPS) 2020 was conducted in 13 Latin American countries: Argentina, Bolivia, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Paraguay, and Peru. It followed a panel sample over three waves of data collection in 12 countries and over four waves in Ecuador.[1] All waves spanned from May to August 2020 and each wave's collection period lasted about ten days on average. The survey was administered to one adult per household. Each respondent was presented with both individual and household-level questions.

All national samples were based on a dual frame of cell and landline phones, and selected as a one-stage probability sample, with geographic stratification of landline numbers. The samples were generated through a Random Digit Dialing (RDD) process covering all cell and landline telephone numbers active at the time of the sample selection.

Survey estimates represent households with a landline or at least one cell phone and individuals of 18 years of age or above who have an active cell phone number or a landline at home.

1. Sampling design

The RDD methodology generates virtually all *possible* telephone numbers in the country under the national telephone numbering plan and then draws a random sample of numbers. This method guarantees full coverage of the population with a phone.[2]

First, in each country, a large first-phase sample was selected in each frame of numbers, with an allocation ranging from 0 percent landlines and 100 percent cell phones to 20 percent landlines and 80 percent cell phones (landline and cell telephone numbers are distinguished by their prefixes). Landline numbers were included with a small share of the total sample in nearly all countries for two reasons: to cover the landline-only households and individuals, who have a low prevalence in most Latin American countries; and to achieve more accurate sex and age sample distributions.[3]

---

* This note was prepared by Ramiro Flores Cruz, partner at Sistemas Integrales and World Bank consultant on survey methodology and sampling, with the financial support from the Latin American and Caribbean Regional Vice Presidency.

[1] Ecuador HFPS had a sample design different to the other HFPS countries since it was based on respondents to the 2019 Human Mobility and Host Community Survey (EPEC by its acronym in Spanish), which collected phone numbers in the field. For more details about EPEC's sample design see *Muñoz, Juan; Muñoz, Jose; Olivieri, Sergio. 2020. Big Data for Sampling Design: The Venezuelan Migration Crisis in Ecuador. Policy Research Working Paper; No. 9329. World Bank, Washington, DC.* https://openknowledge.worldbank.org/handle/10986/34175

[2] Given that the HFPS used a sampling frame of telephone numbers, results represent the population with at least one active phone and exclude the population with no phone.

[3] Survey methodology literature and experience show that cell phone survey respondents are more likely to be male and younger than landline phone respondents due to both cell phone ownership patterns and differential response rates, with females and seniors less likely to answer a call from an unknown number. The underrepresentation of females and seniors in a 100 percent cell phone sample can be compensated via nonresponse weighting adjustment

The landline frame in each country was geographically stratified by department, province, or state, and the sample of landlines was selected with proportionate allocation across these strata. Geographic stratification of cell phones was only done in Argentina, Bolivia, and Mexico.[4] It is important to note that the HFPS sample design allows for obtaining precise estimates at the country level only. Some subnational estimates may have large sampling errors.[5]

The first-phase samples of landline and cell phone numbers were then screened through an automated process to identify the active numbers. The active numbers were then cross-checked with business registries (based on yellow page directories and websites) to identify and remove business numbers not eligible for this survey.

A smaller second-phase sample[6] was then selected from the *active residential* numbers identified in the first-phase sample and was delivered to each country operations team to be contacted by the interviewers.[7]

HFPS sample sizes

The HFPS was conducted in three waves. Table 1 shows the wave 1 final sample size per country and the allocation between both frames.[8] In the first wave, when a cell phone was called, the call answerer was interviewed as long as he or she was 18 years of age or above. When a landline number was called, the interviewer asked to talk to any household member 18 years of age or older. In both cell phone and landline calls, the respondent was asked individual and household questions. Landlines are 10 to 15 percent of the sample in eight countries, 20 percent in two, and 0 percent in three (Table 1).

Wave 1 respondents were recontacted to be interviewed in the second and third waves. The questionnaires across waves included different questions but kept core questions considered key to longitudinal analysis.

---

and calibration. That said, the more unbalanced the sample, the larger the weighting adjustments needed; hence, standard errors in the final survey estimates are larger. The inclusion of landline telephone numbers improves the sex and age representation in the sample. As such, the weighting adjustments and the standard errors of the survey estimates will both be smaller.

[4] Geographic stratification of cell phones was only feasible in these three countries because only in them cell phone number prefixes can be linked to the district (department, province, or state) in which they were issued.

[5] Annex 1 shows how to compute sampling errors for different estimates using Stata.

[6] Note that the selection of phone numbers involves two sampling *phases*, and not two sampling *stages*. The HFPS involves only one sampling stage.

[7] Furthermore, the second-phase sample was delivered in batches to the country teams during fieldwork. Delivering large lists of numbers could have facilitated the "misuse" of the sample by easily replacing non-answering numbers, raising nonresponse rates and potentially increasing nonresponse biases.

[8] The HFPS samples are element samples (i.e., they have one sampling stage), so the design effects are about 1 and the effective sample sizes are similar to the nominal sizes. In contrast, multi-stage cluster samples typically have design effects larger than 1 and the effective sample sizes are smaller than the nominal sizes, generating larger standard errors.

Table 1. Sample size and allocation to cell phones and landlines in HFPS Wave 1

Country	Sample size	Cell phones	Landlines
Argentina	1,000	85%	15%
Bolivia	1,000	100%	0%
Chile	1,000	80%	20%
Colombia	1,000	85%	15%
Costa Rica	800	90%	10%
Dominican Rep	800	85%	15%
Ecuador	1,200	85%	15%
El Salvador	800	90%	10%
Guatemala	800	90%	10%
Honduras	800	100%	0%
Mexico	2,000	80%	20%
Paraguay	800	100%	0%
Peru	1,000	90%	10%

2. Weighting

The HFPS has two sample units: households and individuals. Sampling weights were computed for each unit and should be used according to the estimate of interest. The weighting process involves four steps:

1. Calculation of the inclusion probabilities of landline and cell phone numbers.
2. Computation of design weights for households and individuals.
3. Nonresponse weighting adjustment.
4. Calibration of individual and household weights, using external data from official sources (adjusted for the national phone coverage).

In the second and third HFPS waves, household and individual weights were further adjusted for attrition nonresponse from Wave 1 to 2 and from Wave 1 to 3.

Step 1: Inclusion probabilities of landline and cell phone numbers

A first-phase sample was selected in each of the two frames (cell phone number and landline number frames) with simple random selection without replacement from the entire frame or within geographic strata. The selected numbers were then screened and classified into active and inactive.

The first-phase inclusion probabilities of cell phone and landline numbers are[9]

[9] Inclusion probabilities of cell phones do not show a stratum index since most cell phone samples were not stratified for the reasons stated above. Only cell phone samples for Argentina, Bolivia, and Mexico were stratified.

$$\pi^{C}_{(1)i} = \frac{n^{C}_{(1)}}{N^{C}_{(1)}} = \frac{n^{C}_{(1)A} + n^{C}_{(1)IN}}{N^{C}_{(1)}}$$

$$\pi^{L}_{(1)hi} = \frac{n^{L}_{(1)h}}{N^{L}_{(1)h}} = \frac{n^{L}_{(1)hA} + n^{L}_{(1)hI}}{N^{L}_{(1)h}}$$

where

$\pi^{C}_{(1)i}$ is the first-phase inclusion probability of the i-th cell phone number;

$n^{C}_{(1)}$ is the size of the first-phase sample of cell phones, composed of $n^{C}_{(1)A}$ active cell phones and $n^{C}_{(1)IN}$ inactive cell phones;

$N^{C}_{(1)}$ is the cell phone frame size, the total number of all possible cell phones according to the national numbering plan;

$\pi^{L}_{(1)hi}$ is the first-phase inclusion probability of the i-th landline number in stratum h;

$n^{L}_{(1)h}$ is the size of the first-phase sample of landlines in stratum h, composed of $n^{L}_{(1)hA}$ active landlines and $n^{l}_{(1)hIN}$ inactive landlines; and

$N^{L}_{(1)h}$ is the landline frame size in stratum h, the total number of all possible landline numbers according to the national numbering plan.

Next, two second-phase samples were selected systematically out of the first-phase samples of active cell and active landline telephone numbers. The second-phase inclusion probabilities of cell phones and landlines are

$$\pi^{C}_{(2)i|(1)i} = \frac{n^{C}_{(2)A}}{n^{C}_{(1)A}}$$

$$\pi^{L}_{(2)hi|(1)hi} = \frac{n^{L}_{(2)hA}}{n^{L}_{(1)hA}}$$

where

$\pi^{C}_{(2)i|(1)i}$ is the second-phase inclusion probability of the i-th active cell phone number conditional on being selected in the first phase;

$n^{C}_{(2)A}$ is the size of the second-phase sample of active cell phones;

$\pi^{L}_{(2)hi|(1)hi}$ is the second-phase inclusion probability of the i-th active landline number in stratum h conditional on being selected in the first phase; and

$n^{L}_{(2)hA}$ is the size of the second-phase sample of active landlines in stratum h.

The unconditional inclusion probabilities of the second-phase active cell phones and landlines are

$$\pi_i^C = \pi_{(1)i}^C \, \pi_{(2)i|(1)i}^C = \frac{n_{(1)A}^C + n_{(1)IN}^C}{N_{(1)}^C} \frac{n_{(2)A}^C}{n_{(1)A}^C} = \frac{n_{(1)A}^C + n_{(1)IN}^C}{n_{(1)A}^C} \frac{n_{(2)A}^C}{N_{(1)}^C} = \frac{n_{(2)A}^C}{\widehat{RA}_{(1)}^C N_{(1)}^C} = \frac{n_{(2)A}^C}{\hat{A}_{(1)}^C}$$

$$\pi_{hi}^L = \pi_{(1)hi}^L \, \pi_{(2)hi|(1)hi}^L = \frac{n_{(1)hA}^L + n_{(1)hI}^L}{N_{(1)h}^L} \frac{n_{(2)hA}^L}{n_{(1)hA}^L} = \frac{n_{(1)hA}^L + n_{(1)hIN}^L}{n_{(1)hA}^L} \frac{n_{(2)hA}^L}{N_{(1)h}^L} =$$

$$= \frac{n_{(2)hA}^L}{\widehat{RA}_{(1)h}^L \, N_{(1)h}^L} = \frac{n_{(2)hA}^L}{\hat{A}_{(1)h}^L}$$

where $\widehat{RA}_{(1)}$ is the rate of active phones estimated in the first phase.[10] Hence, the unconditional inclusion probabilities of the second-phase active numbers π_i^C and π_{hi}^L can be expressed as the ratio between the active numbers selected in the second phase and an estimate of the total active numbers in the frame $\hat{A}_{(1)}$.

Step 2: Design weights for households and individuals

The selection probabilities of households and individuals aged 18 years and older are based on the inclusion probabilities of the cell phones and landlines through which they can be reached. Therefore, the computation of household and individual weights should account for multiple chances of selection and for the overlapping between the cell phone and landline frames. This multiplicity weighting adjusts estimates to eliminate the over-representation of households and individuals in the sample that can be reached through more telephone numbers than other households and individuals. It thus eliminates the chance for multiplicity sampling bias.

Multiplicity adjustment

There is multiplicity probability when a household has a larger selection probability because it can be selected through different sample elements (telephone numbers). Households with more than one cell phone or more than one landline number are over-represented in sample designs like this. As a result, their selection probabilities need to be adjusted to account for this increased chance of selection. The multiplicity-adjusted *household* selection probabilities in each frame are computed as

$$\pi_{mj}^C = m_{cj} \pi_i^C$$

$$\pi_{mhj}^L = m_{lj} \pi_{hi}^L$$

[10] $\widehat{RA}_{(1)}$ estimates are highly precise due to the very large size of the first-phase samples.

where

π^{C}_{mj} is selection probability of the j-th household when contacted through a cell phone, adjusted for multiplicity of working cell phones in the household;

m_{cj} is the number of working cell phones in the j-th household;

π^{L}_{mhj} is the selection probability of the j-th household in stratum h when contacted through a landline, adjusted for multiplicity of working landlines in the household; and

m_{lj} is the number of working landlines in the j-th household.

Therefore, if a household has m_c cell phones, its chance of being selected through a cell phone is m_c higher than a household where there is only one cell phone. The same applies to landlines, in which case the multiplicity factor is m_l. Since the number of cell phones and landlines in a household is unknown at the time of the sample design, it needs to be asked during the interview in the questionnaire.

The probability of an *individual* being selected through a cell phone equals the inclusion probability of his or her cell phone number. On the other hand, the probability of an individual being selected through a landline equals the selection probability of his or her household, conditional on the number of working landlines in the household, over the number of individuals aged 18 years and older in the household.

$$\pi^{C}_{k} = \pi^{C}_{i}$$

$$\pi^{L}_{hjk} = \pi^{L}_{mhj} \,/\, \sum_{j} k$$

where

π^{C}_{k} is the selection probability of the k-th individual when contacted through a cell phone;

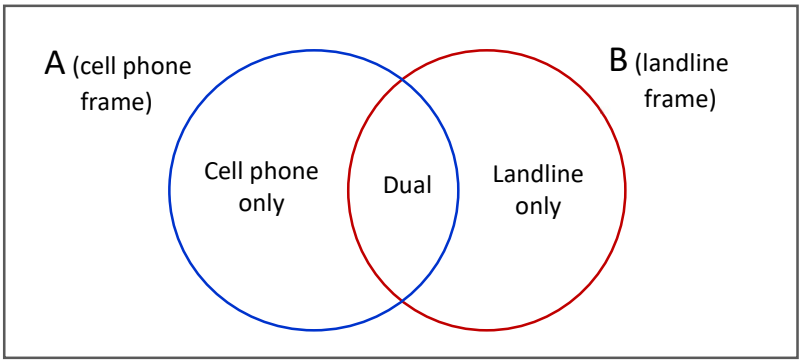
π^{L}_{hjk} is the selection probability of the k-th individual in stratum h when contacted through a landline in the j-th household; and

$\sum_{j} k$ is the number of eligible persons 18 years of age or older in the j-th household.

Overlapping sampling frames

Households and individuals with both cell and landline telephones (dual cases) have a higher probability of being selected than those with only cell phones or only landlines. The following diagram displays the overlapping pattern of the cell phone and landline sampling frames.

Figure 1. Partially overlapping frames



In order to adjust the selection probabilities for multiplicity, it is essential to collect relevant information during the interview. It is necessary to know the domain ownership of the sample households and individuals, plus the number of cell phones and landlines in the sample households. For this purpose, the HFPS questionnaire included the following three questions:

1. *How many working cell phones in total are owned by the persons in your household, including you?*
2. *Is there any working landline in your household?*
3. *How many working landlines are there in your household currently?*

By knowing the domain of ownership, the selection probability for each sample unit can be calculated based on the following probability property

$$P\,(A \cup B) = P\,(A) + P\,(B) - P\,(A \cap B)$$

where $P\,(A \cap B) = P\,(A) \times P\,(B)$, given that A and B are independent

- *In general*, in a dual-frame telephone sample design

$$\pi = \begin{cases} \pi^C & \text{if the sample unit is cell phone only} \\ \pi^L & \text{if the sample unit is landline only} \\ \pi^C + \pi^L - \pi^C \pi^L & \text{if the sample unit is dual} \end{cases}$$

where π^C y π^L are the selection probabilities of the sample units (households or individuals) in domains cell phone only and landline only.

- *In the specific HFPS setting* (with overlapping frames and multiplicity)

Selection probabilities for households are

$$\pi_j = \begin{cases} m_{cj}\pi_i^C & \text{if the household is cell phone only} \\ m_{lj}\pi_{hi}^L & \text{if the household is landline only} \\ m_{cj}\pi_i^C + m_{lj}\pi_{hi}^L - m_{cj}\pi_i^C\, m_{lj}\pi_{hi}^L & \text{if the household has both cell phone and landline} \end{cases}$$

Selection probabilities for individuals are

$$\pi_k = \begin{cases} \pi_i^C \text{ if the individual is cell phone only} \\ m_{lj}\pi_{hi}^L / \sum_j k \text{ if the individual is landline only} \\ \pi_i^C + m_{lj}\pi_{hi}^L / \sum_j k - \pi_i^C \, m_{lj}\pi_{hi}^L / \sum_j k \text{ if the individual has both cell phone and landline telephones} \end{cases}$$

Household and individual design weights, w_{0j} and w_{0k} respectively, are the inverse of the above selection probabilities

$$w_{0j} = \pi_j^{-1}$$

$$w_{0k} = \pi_k^{-1}$$

Step 3: Nonresponse adjustment

When a phone number is called, it is not always possible to carry out an interview. Nonresponse occurs because of a number of constraints. Most common are that nobody answers the call (no contact), the respondent is unwilling to cooperate (refusal), or language barriers exist.

Four main strategies were implemented to minimize nonresponse:

a. The survey management team sent SMS text messages to the sample cell phone numbers before calling to inform that a survey firm would reach out and persuade the phone holder to answer.
b. In most countries, the sample was released to the country operations teams over successive replicates to keep nonresponse monitored and under the central management team's control.
c. Stringent calling protocols were put in place and monitored to ensure a minimum number of attempts on different days and times (5 to 10 attempts depending on the country).
d. The survey offered monetary and non-monetary incentives in most countries to those who cooperated (e.g., gift cards and phone credit).
e. In some countries, the most experienced interviewers recontacted the numbers classified as a "Refusal" to convert them into a "Complete interview".

These actions enabled response rates that were higher than similar RDD sample surveys. Wave 1 response rates and recontact rates in waves 2 and 3 varied across countries. The highest response levels were in Bolivia and Ecuador, while the lowest were in Argentina and Mexico.

The survey attempted to recontact all respondents from wave 1 in waves 2 and 3. Table 2 displays wave 1 response rates and recontact rates for wave 1 to wave 2 and wave 1 to wave 3.

Table 2. HFPS 2020 response and recontact rates by country and wave

Country	Wave 1 response rate	Recontact rate W1/W3	Recontact rate W1/W3
Argentina	14,6%	70,3%	63,7%
Bolivia	34,4%	62,3%	66,1%
Chile	21,3%	62,2%	68,4%
Colombia	26,6%	73,0%	63,8%
Costa Rica	25,7%	79,4%	82,1%
Dominican Rep.	28,4%	83,4%	82,7%
Ecuador	44,6%	83,7%	69,5%
El Salvador	28,7%	77,7%	75,1%
Guatemala	29,2%	77,5%	78,9%
Honduras	20,7%	68,2%	64,6%
Mexico	14,4%	59,0%	54,6%
Paraguay	18,4%	68,0%	63,9%
Peru	28,5%	84,1%	82,1%

The design weights of responding households and individuals were adjusted to compensate for nonresponse and thus reduce potential nonresponse bias on the survey estimates. A class-based nonresponse adjustment was used. Classes were formed by crossing all categories of auxiliary variables that were known to be correlated with the likelihood of responding and were available for both respondents and nonrespondents. Given that the survey used RDD sampling, the information in the sampling frame was limited. The only variables available for respondents and non-respondents were the type of phone number (landline or cell phone) and the corresponding geographic region (known for landlines in all countries and for cell phones only in Argentina, Bolivia and Mexico).

The weighting class nonresponse adjustment is based on the inverse of the weighted response rate estimate in each class. This is the ratio of the sum of the design weights of all units (respondents and nonrespondents) in class c to the sum of the design weights of respondents in that class.

$$a_{jc} = \frac{\sum_{j\in c,R} w_{0j} + \sum_{j\in c,NR} w_{0j}}{\sum_{j\in c,R} w_{0j}} \quad ; \quad a_{kc} = \frac{\sum_{k\in c,R} w_{0k} + \sum_{k\in c,NR} w_{0k}}{\sum_{k\in c,R} w_{0k}}$$

where a_{jc} is the nonresponse adjustment factor that should be applied to responding households in class c, and a_{kc} is the nonresponse adjustment factor for responding individuals in that class. R and NR indicate the responding and nonresponding units, respectively.

Thus, the nonresponse adjusted weights for responding households and individuals are

$$w'_j = w_{0j}\, a_{jc} \quad ; \quad w'_k = w_{0k}\, a_{kc}$$

Step 4: Calibration of individual and household weights

Finally, the weights for the responding households and individuals were calibrated to reflect the total population with phone by sex, age, and region available from external national official sources. This last adjustment has two objectives:

- To further reduce potential nonresponse biases that were not addressed by the nonresponse adjustment in Step 3, by using auxiliary variables from external sources. This can be achieved as long as the calibration auxiliaries are correlated with nonresponse and the study variables.
- To improve the precision of estimators (i.e., reduce the sampling variances), as long as the auxiliaries are correlated with the study variables of interest.[11]

Calibration works by minimizing a measure of the distance between the input weights (nonresponse adjusted weights in this case) and the calibrated weights, under the constraint that the sum of the calibrated weights equals the sum of the totals of the auxiliaries from the external source. Unlike the nonresponse adjustment, weights calibration requires auxiliary variables only for respondents.

Among the existing calibration techniques, the HFPS applied the raking method, using the logit distance function. This method was most suitable given that all available auxiliary variables (region, sex, and age groups) were categorical, the region variable had many categories in most countries, and the overall samples were rather small.

The final weights for responding households and individuals can then be expressed as

$$w_j = w'_j\, g_j = w_{0j}\, a_{jc}\, g_j$$

$$w_k = w'_k\, g_k = w_{0k}\, a_{kc}\, g_k$$

where

w_{0j} is the design weight for the j-th household;

a_{jc} is the nonresponse adjustment factor for households in class c;

g_j is the calibration factor for the j-th household;

w_{0k} is the design weight for the k-th individual;

a_{kc} is the nonresponse adjustment factor for individuals in class c; and

g_k is the calibration factor for the k-th individual.

[11] This objective was not addressed in this survey since it would have entailed computing a large set of replicate weights (with bootstrap or jackknife replication methods), which could be confusing for the final user.

Table 3 shows the data sources used for calibrating the weights in each country. Population totals by sex, age, and region taken from these sources were further adjusted for telephone coverage, using the national phone coverage rates published by the International Telecommunication Union (ITU) from the United Nations.

Table 3. Data sources for the auxiliary data used for weight calibration

Country	Data source used for weight calibration
Argentina	Instituto Nacional de Estadística y Censos. Proyecciones Elaboradas en base al Censo Nacional de Población, Hogares y Viviendas 2010.
Bolivia	Instituto Nacional de Estadística. Proyecciones de Población. 2020.
Chile	Instituto Nacional de Estadística. Estimaciones y Proyecciones de la Población de Chile 1992-2050.
Colombia	Departamento Administrativo Nacional de Estadística. Proyecciones de Población Nacional para el Periodo 2018-2070.
Costa Rica	Centro Centroamericano de Población. Proyecciones Distritales de Población de Costa Rica 2000-2050.
Domincan Rep.	Oficina Nacional de Estadistica. Población Estimada y Proyectada para el Período 1950 – 2100.
Ecuador	World Bank. Ecuador Sociodemographic and Labor Force Survey for Oopulation in Human Mobility - EPEC (2019).
El Salvador	Centro Centroamericano de Población. Proyecciones de Población de El Salvador. 2000-2050.
Guatemala	Instituto Nacional de Estadística. Proyecciones Nacionales 1950-2050.
Honduras	Instituto Nacional de Estadística. Proyecciones de Población 2013-2015.
Mexico	Consejo Nacional de Población. Proyecciones de la Población de México y de las Entidades Federativas, 2016-2050.
Paraguay	Dirección General de Estadística, Encuestas y Censos. Proyección de la población nacional por sexo y edad, 2000-2025. Revisión 2015.
Peru	Instituto Nacional de Estadística e Informática. Estimaciones y Proyecciones de Población. Boletín Especial Nº 21 y 22.

3. Estimation and Sampling Errors

When analyzing the data, it is essential to compute and assess the precision of the survey estimates, i.e., the magnitude of their sampling error. Sampling errors can be expressed through the sampling variances, standard errors, coefficients of variation,[12] and confidence intervals, although all these may also include part of the non-sampling errors.

When estimating sampling errors for means, proportions, ratios, and linear and nonlinear regression parameters, HFPS sample design features and weighting need to be accounted for. If these are not considered, standard statistical software will treat the sample as a simple random sample, which would lead to biased estimates of sampling variances.

[12] The standard error is the square root of the sampling variance. The coefficient of variation is a relative measure of the standard error and is calculated as the ratio between the standard error and the point estimate (it is usually expressed in percentage terms). As a rule of thumb, estimates with coefficients of variation of 1 percent or lower are considered to have a very high level of precision. Coefficients of variation between 1 and 3 percent are generally classified as very good, from 3 to 5 percent as good, from 5 to 10 percent as acceptable, and from 10 to 15 percent as large. Above 15 percent is classified as too large and the corresponding estimate is considered unreliable.

The two most common approaches for estimating sampling errors for complex sample data are: (1) the Taylor Series Linearization (TSL) of the estimator and the corresponding approximation to its variance, or (2) the use of resampling variance estimation techniques, such as balanced repeated replication (BRR), jackknife repeated replication (JRR), and bootstrap. Stata and other statistical software packages use the TSL method as the default for estimating sampling errors.

Annex 1 indicates the Stata script that should be used to account for the HFPS sample design and weighting when computing an estimate based on cross-sectional data (i.e., based on one wave only).

As mentioned, the HFPS has a panel design and the survey attempted to recontact all respondents from wave 1 in waves 2 and 3. Thanks to the overlapping of sample units over the survey waves, panel surveys allow more precise estimates of the change or difference for an indicator between successive waves to be obtained. Sequential cross-sectional surveys, where each wave's sample includes different households and individuals, can also track changes over time. In this case, however, change estimates are less precise (i.e., have a larger sampling error) than with a panel survey.

Thus, the HFPS panel should be able to determine more precisely whether a decrease or increase in a given indicator over time is statistically significant. It should ideally be able to detect small changes between two waves.

Under these conditions, the design-based variance of the change estimate $\hat{\Delta} = \theta_2 - \theta_1$ for the indicator of interest θ is given by

$$var(\hat{\Delta}) = var(\hat{\theta}_1) + var(\hat{\theta}_2) - 2\, corr(\hat{\theta}_1, \hat{\theta}_2) \sqrt{var(\hat{\theta}_1)\, var(\hat{\theta}_2)}$$

where

$\hat{\theta}_1$ is the cross-section estimate of the indicator of interest θ in wave 1;

$\hat{\theta}_2$ is the cross-section estimate of the indicator of interest θ in wave 2;

$\hat{\Delta}$ is the estimate of net change of θ between waves 1 and 2; and

$corr(\hat{\theta}_1, \hat{\theta}_2)$ is the correlation between the two wave indicators.

The above expression shows how the sampling variance of the change estimate (for indicator θ) is reduced. The precision of the change estimate is thus increased due to the existing correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$. Since respondents in a panel are the same in waves 1 and 2, then the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ is expected to be non-zero. The larger the correlation, the more precise the change estimate.[13]

Annex 2 includes the Stata code for testing the change of an indicator between any two HFPS waves, accounting for both the sample design features and the panel overlap. The test output shows the change point estimate, plus the corresponding standard error, t-score, p-value, and 95% confidence interval.

[13] The magnitude of $corr(\hat{\theta}_1, \hat{\theta}_2)$ depends on each particular indicator of interest θ.

Reference literature

Heeringa, S., West, B., and P. Berglund. (2017). Applied Survey Data Analysis (Second Edition). New York, Taylor & Francis Group.

Lohr, S. and J. RAO. (2006). Estimation in Multiple-Frame Surveys, Journal of the American Statistical Association, 101, 1019−1030.

Lohr, S. (2011). Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames, Survey Methodology, 37, 197−213. Statistics Canada.

Skinner, C. and J. Rao. (1996). Estimation in Dual-Frame Surveys with Complex Designs, Journal of the American Statistical Association, 91, 349−356.

Thompson, S. (2012). Chapter 15: Network Sampling and Link-Tracing Designs, in Sampling. New York, Wiley.

Valliant, R., Dever J., and F. Kreuter. (2016). Practical Tools for Designing and Weighting Sample Surveys. New York, Springer.

Annex 1

Stata Code for Weighted Estimates and Sampling Error Computation
Cross-sectional data

This annex provides a set of examples of the Stata syntax for computing estimates and their corresponding sampling errors (measured by standard errors, confidence intervals, and coefficients of variation), accounting for the HFPS sample design and weighting. For more details, data users are referred to the online Stata manual for the svy command (http://www.stata.com/manuals15/svy.pdf).

To specify the sample design features in any of the HFPS datasets, use command:

```
svyset [pweight=w_hh_w1]
*Use weight w_hh_w1 for household-level estimates in wave 1 (w_hh_w2 for wave 2 and w_hh_w3 for wave 3)
*Use weight w_ind_w1 for individual-level estimates in wave 1 (w_ind_w2 for wave 2 and w_ind_w3 for wave 3)
```

Numeric variables (means):

To estimate the mean age of the population 18+, use command:

```
svy: mean q03_07
estat cv
```

To estimate the mean age of the population 18+ by gender, use command:

```
svy: mean q03_07, over(q03_03)
estat cv
```

To estimate the mean age of the population 18+ who did not work in the week prior to the interview, use command:

```
svy, subpop (if q07_01==2): mean q03_07
estat cv
```

Categorical variables (proportions):

To estimate the frequency distribution of persons 18+ according to their level of concern that a family member could fall seriously ill because of COVID-19, use command:

```
svy: tab q10_01, se ci cv
```

To estimate the frequency distribution of persons 18+ on whether they worked in the week prior to the interview by status in employment, use command:

```
svy: tab q07_01 q07_05, col se ci cv
```

To estimate the frequency distribution of households on whether they received money from the government or NGOs, among households where a member lost his or her job since the beginning of the quarantine, use command:

```
svy, subpop (if q07_20==1): tab q11_04, se ci cv
```

Linear regression:

To estimate the regression coefficients of a continuous variable y on two continuous variables x_1 and x_2, use command:

```
svy: regress y x1 x2
```

To estimate the regression coefficients of a continuous variable y on two continuous variables x_1 and x_2 and two categorical variables x_3 and x_4, use command:

```
svy: regress y x1 x2 i.x3 i.x4
```

Annex 2

Stata Code for Testing Changes between HFPS Waves
Panel data

This annex provides the Stata code for testing the change of an indicator between any two HFPS waves, accounting for both the sample design and the panel overlap.

The following example is based on the first two waves of the Colombia HFPS for testing the change in the proportion of persons 18+ who worked for at least one hour in the week before the HFPS interview.

The variable of interest is originally named q07_01 in Wave 1 and p07_01 in Wave 2. Rename it as d07_01 in both waves, so it has the same name in both datasets.

Rename the weights variables w_ind_w1 in Wave 1 and w_ind_w2 in Wave 2 as w_ind.

In both datasets, keep variables caso_se, w_ind, estrato, ola, and the variable to be tested d07_01.

Save both data sets as new files with new names.

```
use HFPS_COL_W1_2020.dta, clear
rename q07_01 d07_01
rename w_ind_w1 w_ind
keep caso_se w_ind estrato ola d07_01
save HFPS_COL_W1_2020_prime.dta, replace

use HFPS_COL_W2_2020.dta, clear
rename p07_01 d07_01
rename w_ind_w2 w_ind
keep caso_se w_ind estrato ola d07_01
save HFPS_COL_W2_2020_prime.dta, replace
```

"Stack" the two resulting datasets, combining them into a single dataset.

```
use HFPS_COL_W1_2020_prime.dta, clear
set more off
append using HFPS_COL_W2_2020_prime.dta, force
```

Test the change in d07_01 for the full population 18+:

```
replace d07_01 = 0 if d07_01 == 2
svyset caso_se [pweight=w_ind]
svy: mean d07_01, over (ola)
lincom [d07_01]2-[d07_01]1
```

Test the change in d07_01 among women 18+:

```
replace d07_01 = 0 if d07_01 == 2
svyset caso_se [pweight=w_ind]
svy, subpop(if q03_03==2): mean d07_01, over (ola)
lincom [d07_01]2-[d07_01]1
```