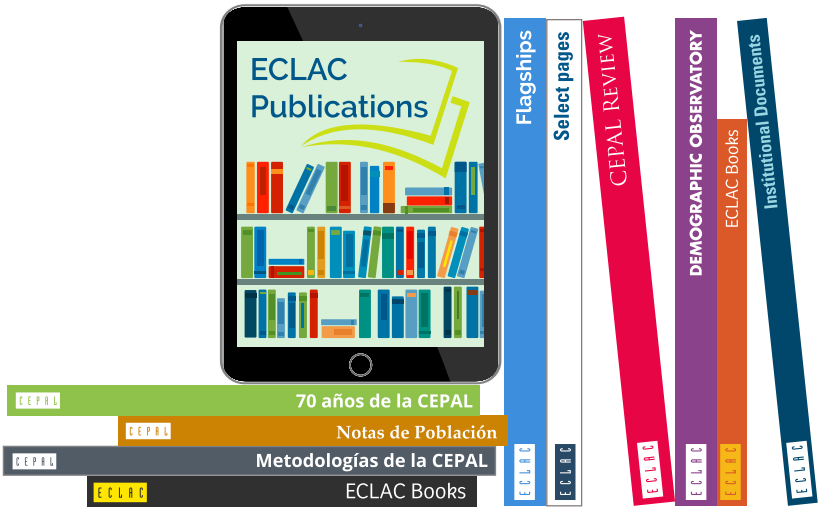
Methodological recommendations on the measurement of the quality of household survey figures

UNITED NATIONS

ECLAC

Statistical Conference of the Americas of ECLAC

Thank you for your interest in this ECLAC publication



Please register if you would like to receive information on our editorial products and activities. When you register, you may specify your particular areas of interest and you will gain access to our products in other formats.

Register

www.cepal.org/en/publications

www.instagram.com/publicacionesdelacepal

www.facebook.com/publicacionesdelacepal

www.issuu.com/publicacionescepal/stacks

www.cepal.org/es/publicaciones/apps

UNITED NATIONS

ECLAC

Methodological recommendations on the measurement of the quality of household survey figures

UNITED NATIONS

ECLAC

Statistical Conference of the Americas of ECLAC

This document was prepared by the working group for the development of methodological recommendations for measuring the quality of household survey figures, established by the Statistical Conference of the Americas of the Economic Commission for Latin America and the Caribbean for the 2022–2023 biennium. The group was coordinated by the National Institute of Statistics of Chile, and the ECLAC Statistics Division served as its technical secretariat. The following countries and agencies made up the group's membership: the National Institute of Statistics and Censuses of Argentina; the Brazilian Institute of Geography and Statistics; the National Administrative Department of Statistics of Colombia; the National Institute of Statistics and Censuses of Costa Rica; the National Office of Statistics and Information of Cuba; the National Office of Statistics of the Dominican Republic; the National Institute of Statistics and Censuses of Ecuador; the National Office of Statistics and Census of El Salvador; the National Institute of Statistics of Guatemala; the National Institute of Statistics and Geography of Mexico; the National Institute for Development Information of Nicaragua; the National Institute of Statistics of Paraguay; the National Institute of Statistics and Informatics of Peru; the National Institute of Statistics of the Plurinational State of Bolivia; and the National Institute of Statistics of Uruguay.

The following individuals participated in the production of this document:

Chile (National Institute of Statistics): Marly Olivares Lathulerie, Miguel Guerrero Herrera, Jonathan Paredes Baez, Iván Tourón Romero, Denisse Lopez Arenas, Felipe Molina Jaque

Brazil (Brazilian Institute of Geography and Statistics): Marcus Morais Fernandes

Colombia (National Administrative Department of Statistics): María Margarita Echeverry Vásquez

Mexico (National Institute of Statistics and Geography): Eric Rodríguez Herrera, José Elías Rodríguez Muñoz

Peru (National Institute of Statistics and Informatics): Nancy Hidalgo

Plurinational State of Bolivia (National Institute of Statistics): Rosemary Villanueva Calle

United Nations publication
LC/CEA.12/11
Distribution: L

Printed at United Nations, Santiago
S.2301135[E]

This publication should be cited as: Working group for the development of methodological recommendations for measuring the quality of household survey figures of the Statistical Conference of the Americas, *Methodological recommendations on the measurement of the quality of household survey figures* (LC/CEA.12/11), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC), 2024.

Contents

Introduction 5

Chapter I
Background on the use of quality standards by national statistical offices 7

Chapter II
Conceptual framework 13

A. Sample surveys and their characteristics 13
1. Sampling frame 14
2. Sampling types 15
3. Issues to be considered in a sampling scheme 19
B. Complex sampling designs and their characteristics 20
C. Population parameters and their estimators 21
D. Confidence interval 22
E. Estimation of sampling errors 23
1. Variance and its estimators 24
2. Measurements of dispersion associated with variance 26
3. Tabulations or disaggregations 29

Chapter III
Elements for assessing the statistical quality of an estimate 31

A. Design effect 31
B. Minimum sample size to apply accuracy measurements 32
1. Sample size 32
2. Effective sample size 33
3. Degrees of freedom 33
C. Types of estimators 34
D. Accuracy measurements used for publication and dissemination 37
1. Standard error 38
2. Coefficient of variation 38
3. Logarithmic coefficient of variation 39
E. Unweighted case count 40

Chapter IV
Quality requirements applicable to estimates 41

Chapter V
Concluding observations 45

Bibliography 47

Tables

Table II.1 Main parameters, estimators and estimated variance in simple random sampling 24
Table II.2 Estimated number of people by sex, according to socioeconomic level 29

Figures

Figure I.1 Frequency of use of quality metrics 10
Figure III.1 0.95 percentile of Student's t-distribution and percentage variation according to degrees of freedom 34
Figure III.2 Maximum coefficient of variation and standard error admitted according to the estimate of *P* (first criterion) 35
Figure III.3 Maximum coefficient of variation and standard error admitted according to the estimate of *P* (second criterion) 36
Figure III.4 Behaviour of the standard error and coefficient of variation according to the estimate of *P* in a simple random sample where n=60 37
Figure III.5 Relationship between sample size and the precision of an indicator using the logit transform 39

Diagrams

Diagram I.1 Study domains 9
Diagram I.2 Other domains of interest 10
Diagram II.1 Processes and subprocesses of GSBPM 5.1: business segment 14
Diagram III.1 Precision of an estimator determined by the concentration or dispersion of the estimated values 37
Diagram IV.1 Flow chart for evaluation of estimates 41
Diagram IV.2 Flow chart for evaluation of tabulations 43

Introduction

Countries require increasingly disaggregated socioeconomic information in order to implement, monitor and evaluate public policies. National statistical offices meet most of that need through various statistical operations and household surveys that cover a wide range of topics, including economic, labour, social, educational, health and time-use matters. National statistical offices are responsible for ensuring that the indicator estimates obtained through household surveys are as robust as possible.

Guidelines are therefore essential to orient national statistical offices and other official producers of statistics and user groups in general and to provide them with reference frameworks on how to assess the statistical quality of the estimates obtained through household sample surveys.[1]

Although numerous household surveys with different objectives exist both in Latin America and the Caribbean and across the world, they share similarities in terms of the standardization of their concepts, methodologies and processes and of the challenges faced. It is therefore possible —and highly desirable— to develop cross-cutting methods and guidelines that can be applied to the entire region.

Accordingly, the Statistical Conference of the Americas of the Economic Commission for Latin America and the Caribbean created a working group for the 2022–2023 biennium tasked with developing recommendations for harmonizing analyses of the quality of the figures yielded by household surveys and for quantifying sampling errors by applying methods that are accessible to all user groups and can be used for all types of estimators. In keeping with that general objective, four specific goals were defined:

(i) Systematize the state of the art in the methods currently used to publish and disseminate household sample survey figures.

(ii) Establish a standardized regional procedure, in line with international recommendations and good practices, for assessing the quality and accuracy of the estimates yielded by processing and analysing household sample surveys. The procedure should be adaptable to the specific reality of each country and each survey.

(iii) Examine the particular characteristics of database anonymization processes and their impact on how household survey users measure sampling errors.

(iv) Define a set of methods suitable for estimating sampling errors in anonymized household survey databases that use replicated weights.

This document begins with a diagnosis of the practices used by the region's national statistical offices to measure the quality of estimates, particularly those yielded by household sample surveys. A conceptual framework related to the scope of the study is provided below, with a particular focus on replicated weight techniques for estimating sampling errors. Based on the review of the literature carried out, the various quality criteria that can be used to determine the statistical quality of estimates are presented in a logical order, and a work flow for applying the examined quality criteria is proposed to serve as the standard reference for assessing the statistical quality of national statistical office estimates.

[1] Probabilistic surveys targeting households are a particular kind of household sample surveys that offer a practical way to obtain up-to-date figures on social conditions and trends, households' socioeconomic behaviour and access to basic services and the impact of social programmes. They are particularly useful for creating household databases and they open up multiple possibilities for establishing interrelationships and conducting analyses using existing data processing systems.

Chapter I

Background on the use of quality standards by national statistical offices

One of the principles of the Code of Good Practice in Statistics for Latin America and the Caribbean (ECLAC, 2011) is a commitment to quality, according to which national statistical offices within national statistical systems must work and cooperate in accordance with international rules, principles and standards. To uphold that principle, the region has undertaken a series of initiatives to further develop methods for assuring the quality of statistical output.

The *Guide for the implementation of a quality assurance framework for statistical processes and outputs*, a joint effort by the national statistical offices of Colombia and Mexico with advice from the technical secretariat of the ECLAC Statistics Division, was published in 2022. Its purpose is to guide Latin American and Caribbean countries in the adoption and application of the *United Nations National Quality Assurance Frameworks Manual for Official Statistics*, which organizes quality principles and their associated requirements into four levels (ECLAC, 2022):

(i) Level A, which deals with the management of statistical systems and covers the fundamental principles of coordinating the national statistical system and managing relations and standards.

(ii) Level B, which addresses the management of the institutional environment based on the principles of professional independence, objectivity, transparency, confidentiality, commitment to quality and adequacy of resources.

(iii) Level C, which deals with the management of statistical processes, which are to be governed by the principles of methodological soundness, cost-effectiveness, appropriate procedures and respondent burden management.

(iv) Level D, which focuses on the management of statistical outputs and covers the fundamental principles of relevance, accuracy and reliability, timeliness and punctuality, accessibility and clarity, coherence and comparability, and metadata management.

The present document aims to further develop levels C and D in particular, to help ensure that the principles of methodological soundness, appropriate procedures, accuracy and reliability are applied in the estimates obtained from probabilistic household surveys.

In line with the above, and in pursuit of its established goals, the working group began by identifying the processes used by the region's national statistical offices to measure the quality of estimates and, specifically, of those obtained through household sample surveys. To that end, it contacted 11 of the region's countries to obtain detailed information on different issues related to measuring the statistical quality of estimates, such as the principles and guidelines followed to assess the quality of parameter estimates, the metrics used, decision thresholds, estimate classifications and anonymization methods.

The guidelines for applying quality criteria and determining the usability of the parameter estimates of interest include those defined by the national statistical offices themselves and those based on international standards. The national statistical offices that have their own guidelines and related documentation include those of Argentina,[2] the Plurinational State of Bolivia (ANDA, 2020), Chile (INE, 2020a), El Salvador, Uruguay (INE, 2021) and Brazil. In the case of Brazil, however, the most recent version is still being drafted, and neither El Salvador nor the Plurinational State of Bolivia have any official documents making their standards public.

[2] See technical notes 1 to 4 of the National Institute of Statistics and Censuses (INDEC, n.d.).

The Central Data and Microdata Catalogue of the National Data Archive of the Plurinational State of Bolivia's National Statistics Institute establishes the disclosure policy, which contains a recommendation on the statistical quality of the estimates made by users. The section indicates the criteria to be used for this purpose and stresses the importance of taking due account of the coefficient of variation and the number of observations from which responses were obtained. It recommends the use of statistical programs that enable the calculation of sampling errors by applying the sampling design used in the survey, and the inclusion of the value of the coefficient of variation (CV) when estimates are made. It also establishes a standard for assessing the sampling errors of the main indicators and at more disaggregated geographic levels. This standard is based on CV and sampling units at the disaggregated level, and it classifies estimates into five categories: optimal estimation, reliable estimation with very good accuracy, estimation with sufficient accuracy, estimation with fair accuracy and non-significant estimation (ANDA, 2020).[3]

In 2020, the National Institute of Statistics of Chile published two documents: "Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares" and "Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares". The former document describes in detail the methodological analysis that the Institute carries out to assess the quality metrics of the parameter estimates covered by the standard, as well as the conceptual framework, the use of the quality metrics and the criteria used in the institute's surveys and in other national statistical offices. In this document, the Institute suggests that CV should not be used to determine the accuracy of proportion or ratio estimators, recommending instead a quadratic function to examine the behaviour of the standard error. It also uses a series of flow charts to present proposed guidelines for assisting user groups in making decisions related to information analysis and publication (INE, 2020b). The second document is a compendium of the first and concisely sets out the concepts and criteria used in estimates. It also includes the flow chart for the application of the criteria to be used in determining the quality of parameter estimates and of the corresponding tabulations (INE, 2020a). In line with these documents, in 2022 the Institute published an R-language package called *calidad* (quality) (CRAN, 2023). This package enables the accuracy of different indicators derived from household surveys to be assessed, on the basis of the standards established by the institution and by ECLAC, which also allows other institutions in the region to use it.

In turn, the Brazilian Institute of Geography and Statistics (IBGE) has a guide for the disclosure of sampling errors in the probability sample surveys it conducts (IBGE, 2021a). The guide presents ways to disclose sampling errors associated with parameter estimates obtained through sample surveys conducted in accordance with the provisions of the second edition of the IBGE code of good statistical practice, which was published in 2021 (IBGE, 2021b). Guidelines are also provided on how to publish information on the quality of those estimates —specifically, on their accuracy— so that user groups can duly assess and make use of the figures produced by IBGE. The guide contains a conceptual framework for terms associated with the statistical quality of estimates, and it provides guidelines on how to publish data on estimate accuracy. IBGE is currently preparing a new document with recommendations for all its surveys that will address the statistical quality of estimates of the parameters of interest.

In June 2022, under ECLAC guidance and in accordance with the organization's most recent recommendations (Gutiérrez and others, 2020), the Department of Statistics and Censuses of El Salvador published the document "Recomendaciones sobre criterios de supresión para investigaciones por muestreo probabilístico", which provides simple and clear guidelines to orient different user groups in assessing the quality of estimates of parameters of interest (DIGESTYC, 2022).

In 2021, the National Institute of Statistics of Uruguay adopted the Technical Standard for Statistical Operation Quality Certification INE-CCOE: 2021-01 (INE, 2021), which addresses the quality of statistical output based on the dimensions proposed by the Statistical Office of the European Communities (Eurostat), namely: relevance, accuracy and precision, timeliness and punctuality, accessibility and transparency, comparability and consistency, and institutional capacity. The Generic Statistical Business Process Model (GSBPM) of the Economic Commission for Europe was used as a reference framework in preparing the document. The technical standard establishes a set of indicators for each dimension of quality, and those indicators are measured by means of standards. The degree of compliance with the standards determines the level of development of each indicator; in other words, how

[3] The Plurinational State of Bolivia is not the only country to use this metadata file system based on the Data Documentation Initiative; it is also quite commonly used by other institutions and national statistical offices. See ANDA (2023).

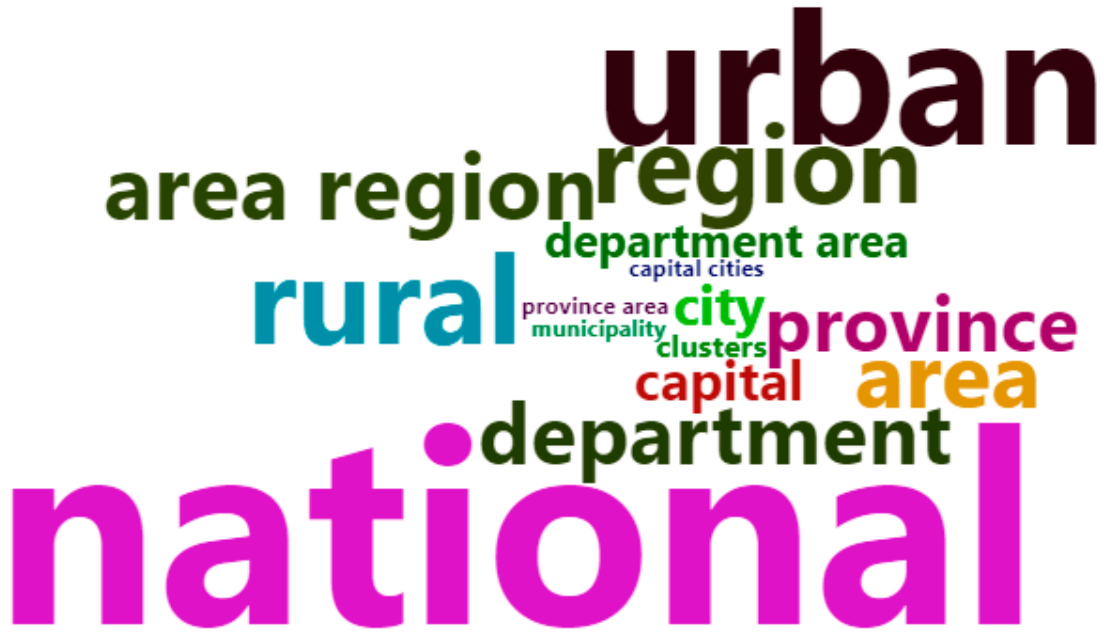
close the statistical operation being evaluated is to the ideal quality level. The level of development is measured on a scale from 1 to 4, where 1 is the lowest expected level of quality and 4 is the highest. For the quality of parameter estimates, the standard requires the calculation of sampling errors related to the main variables. Among the metrics mentioned are standard error, coefficient of variation, confidence intervals, mean squared error and design errors. For estimate quality, the availability of sampling error measurements is evaluated according to whether the data are available or not for the main variables or for the remaining variables, and a score is assigned according to the accuracy of the parameter estimates based on CV values. If the accuracy of the variables is low, an explanation should be provided (INE, 2021).

The region's national statistical offices that to date have not published or shared with the community documents on quality standards related to parameter estimates state that they make use of guidelines formulated by Eurostat, Statistics Canada, the United States Census Bureau and ECLAC.

Another aspect that was considered in the regional consultation on estimate quality was the disaggregation of the information: i.e. the study domains guaranteed in the design and all the other domains of interest for which estimates are produced. This is based on the Sustainable Development Goals, since national statistical offices must ensure that the entire population, in its vast diversity, is represented. In practical terms, national statistical offices must work to ensure that their official statistics remain robust when disaggregated by relevant population characteristics such as income, gender, age, race, ethnicity, migration status, disability and geographic location.

As can be seen on diagram I.1, the study domains —and, consequently, those in which the representativeness of the estimates is guaranteed— generally follow the political and administrative division of the countries at different geographic scales. Most commonly, estimates are made at the national level, by area type (urban and rural) and by regions (and their various counterparts, such as departments, municipalities, States and so on).

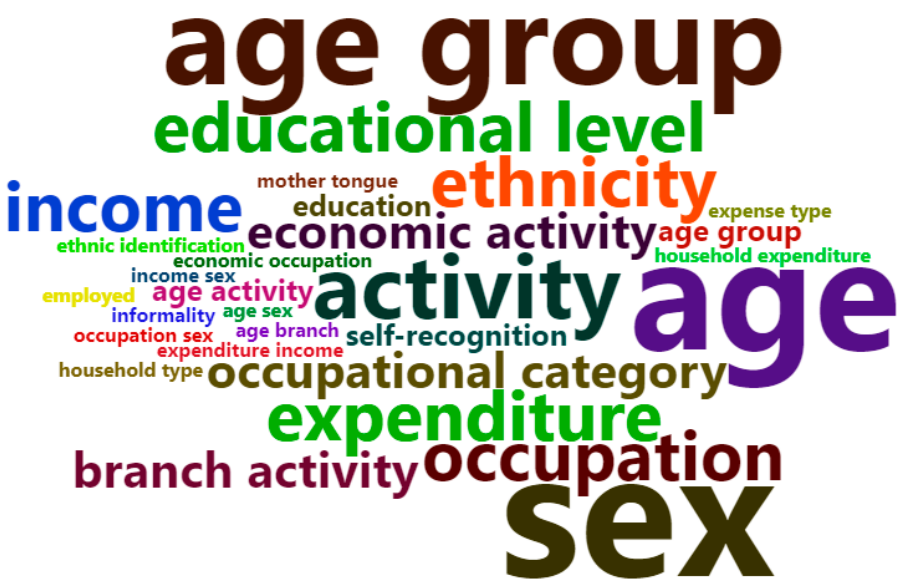
Diagram I.1
Study domains



Source: Economic Commission for Latin America and the Caribbean (ECLAC) and National Institute of Statistics of Chile (INE).
Note: Covers 30 statistical operations.

Diagram I.2, in contrast, shows the other domains of interest: in other words, those where estimate robustness is in most cases not guaranteed. These domains are the most likely to lack statistical quality, with gender, age group, income level, activity, expenditure and others being among the most common.

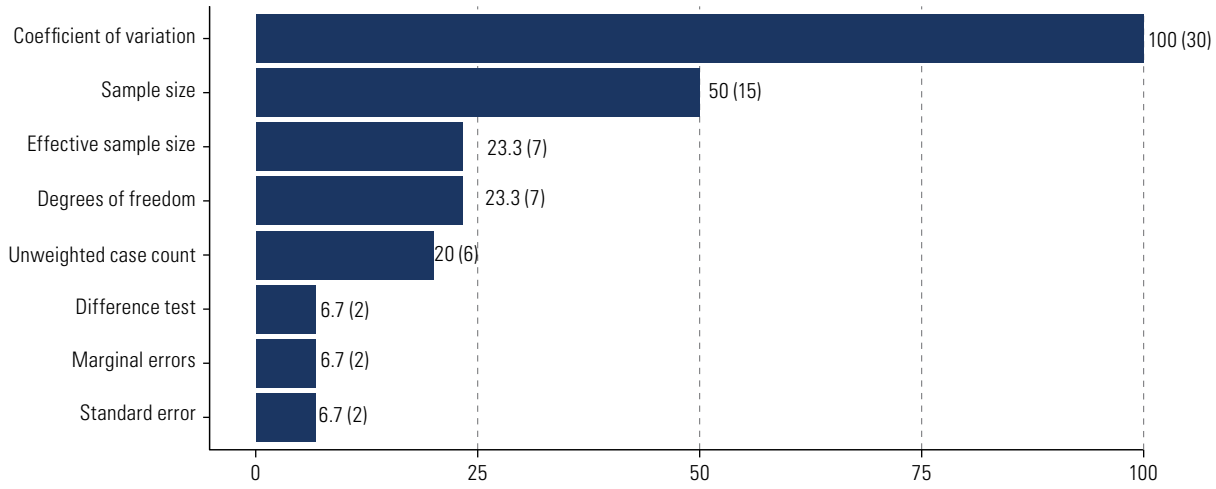
Diagram I.2
Other domains of interest



Source: Economic Commission for Latin America and the Caribbean (ECLAC) and National Institute of Statistics of Chile (INE).
Note: Covers 30 statistical operations.

According to the regional consultation, and based on the 30 statistical operations reported by the universe of 11 institutions, two metrics are frequently used to assess estimate quality: CV (which is used in all the statistical operations in question) and the sample size.[4] Other elements commonly used, but to a lesser extent, are effective sample size, degrees of freedom and unweighted case counts (see figure I.1).[5] The differences observed among the criteria used are generally determined by the type of estimate (total, mean, ratio or percentile).

Figure I.1
Frequency of use of quality metrics
(Percentages)



Source: Economic Commission for Latin America and the Caribbean (ECLAC) and National Institute of Statistics of Chile (INE).
Note: The graph shows the answers obtained to a multiple-choice question based on 30 statistical operations. The numbers in parentheses indicate the number of statistical operations in which the metric is used.

4 Sample size refers to the number of units taken into account in calculating the estimate of the parameter in question.
5 Effective sample size is the ratio between the number of units taken into account in calculating the estimate and the design effect.

The national statistical offices' responses regarding the thresholds set to evaluate each estimate varied from one case to another. The value most commonly used for CV was 15%, followed less frequently by 20% and 30%. In some cases, CV is evaluated using intervals that enable the estimates to be assigned different levels of reliability, accuracy or quality.

Three thresholds were identified with regard to the sample size evaluation: two relating to elements (20 and 60 units), and another relating to the number of clusters in the sample obtained (at least two clusters).[6] For the degrees of freedom, nine degrees was seen to be the minimum value necessary to guarantee estimate quality. When unweighted case counts are used, the estimate is not published if there are fewer than 25 cases and, if there are between 25 and 49 cases, the figure is presented in parentheses to provide a warning about the quality of the estimate.

As noted in the previous paragraphs, a classification of the estimates is also established according to how the quality criteria are evaluated. In some cases, four categories are used for the classification: for example, very good, good, acceptable and referential. Other cases use three categories (reliable, partially reliable and unreliable) or two (reliable or unreliable). Not all the national statistical offices reported that they use classifications of this kind.

It should be noted that the consultation was conducted after the coronavirus disease (COVID-19) pandemic, when response rates for household surveys were significantly affected. The countries were asked whether they had to adjust the quality criteria during this period. In general, the measure adopted by most of the national statistical offices was to discard estimates corresponding to the most disaggregated levels on account of the smaller sample sizes obtained.

Another important aspect of the working group's objectives was for all users, both external and internal to the national statistical offices, to have access to the information contained in the microdatabases published by the national statistical offices, to enable them to assess the estimate quality of the parameters of interest to them.[7] For this to occur requires clear and precise guidelines that enable full respect for the sample designs with which the surveys were conceived and that, at the same time, are simple for users to follow.

In keeping with the principle of statistical confidentiality, the common practice among national statistical offices is for published microdatabases to be anonymized, which in certain cases may entail the suppression of variables needed to estimate the sampling error. At the same time, some estimators require the sampling error to be calculated by means of calculations that are very complex even for specialized personnel. For those reasons, this consultation also inquired about the methods used to estimate sampling errors and to anonymize the data, in order to reveal whether all the microdata users were indeed in a position to estimate sampling errors. It was found that in 96.7% (29) of the statistical operations reported by the national statistical offices, the microdatabases were published. It was also found that 90% of those operations (26 out of 29) used data masking and anonymization methods: for example, the elimination of variables containing personal or sensitive data, such as household identifiers. Less frequently, the region's national statistical offices opted to group together variable categories containing sensitive information, such as nationality, age or occupation; other options included limiting the geographic disaggregation and truncating values for variables such as earnings.

Among the methods used to estimate sampling errors, the most notable are the last cluster method, the variance of the calibrated estimator, and resampling techniques such as the bootstrap or jackknife methods. It was thus found that, contrary to what was believed, several of the region's national statistical offices had already implemented resampling methods, which, according to some experts, represent the best option when sampling errors are to be estimated, since they can be applied to any type of estimator. Moreover, if the replicated weights are made available to individuals, users and researchers, it is feasible for any user with a basic statistical knowledge

6 Assuming two-stage sampling in which the secondary units are dwellings, the clusters in the sample obtained are all those in which at least one dwelling was obtained. Similarly, an obtained household is one that answered, at the least, the basic identification questions to record the household's members and another minimum set of questions addressing the main variables, according to each survey's criteria. Two clusters is the minimum value needed to calculate variance, regardless of the study domain.

7 In the context of household surveys, a microdatabase is a database at the level of the information units in the study: that is, either households or individuals. It also involves a tabular-matrix structure of rows representing individuals and columns representing the various characteristics or variables recorded by means of a questionnaire and associated with each subject.

to estimate the sampling error of the parameter estimates in which they are interested, even if anonymization techniques are applied to the databases. This offers the possibility of evaluating the quality of the information and of making use of it responsibly.

In terms of technological advances, major steps with the automation of calculation processes have also been taken in the region. The National Institute of Statistics of Chile has developed an RStudio package to assess the quality of parameter estimates (INE, 2020a; CRAN, 2023).[8] This package enables estimates obtained from sample surveys to be evaluated by applying its own standard as well as the one proposed by ECLAC. One of the benefits of the package is that it allows the thresholds for each evaluated metric to be modified according to the selected standard. At the same time, the National Institute of Statistics and Censuses of Argentina has designed an application called CEMRepBoot to apply resampling techniques to its surveys. This application enables sampling errors associated with population parameter estimates to be calculated, and it provides users with recommendations on estimates, calculations and the statistical use of survey data (MTESS/OISS/SRT, 2018).

8 See [online] http://www.rstudio.com/.

Chapter II

Conceptual framework

In designing a study, due account must be taken of the research goals, the specific questions to be answered, the available resources and any practical constraints. Choosing the right sampling approach is crucial to ensure the validity and quality of the results and the efficiency of the research process. When working with household surveys, information from all individuals is of equal relevance, and so probabilistic surveys are used to draw inferences about the total population.[9]

The approaches outlined in the above paragraphs are important in research and data analysis, and they complement each other to provide a more complete and detailed picture in different study scenarios. In practice, both probability and non-probability sampling have their advantages and disadvantages, and the choice of the appropriate method will depend on the research context, study objectives and practical constraints. By combining these approaches in a mixed sampling design, a wide variety of research questions can be effectively addressed and a deeper and more complete understanding of the phenomenon under study can be obtained. Statistical offices generally use probability sampling, with which they can achieve population representativeness and eliminate selection bias.

The purpose of this chapter is to introduce the concepts that underpin and generate the statistical quality of an estimated figure reached through a probability sample survey. It covers all stages from the design of the survey to the analysis of the results, and the quality analysis will focus on the latter.

A. Sample surveys and their characteristics

Sample surveys should be designed so that quality statistical information can be obtained efficiently.[10] Accordingly, this section provides illustrative examples of the different characteristics that a sample design should adopt and offers bibliographic references that can be consulted to learn more about the technical requirements of the described designs. It also highlights the importance of ensuring the statistical quality of the parameter estimates of interest as of the moment when the survey is designed.

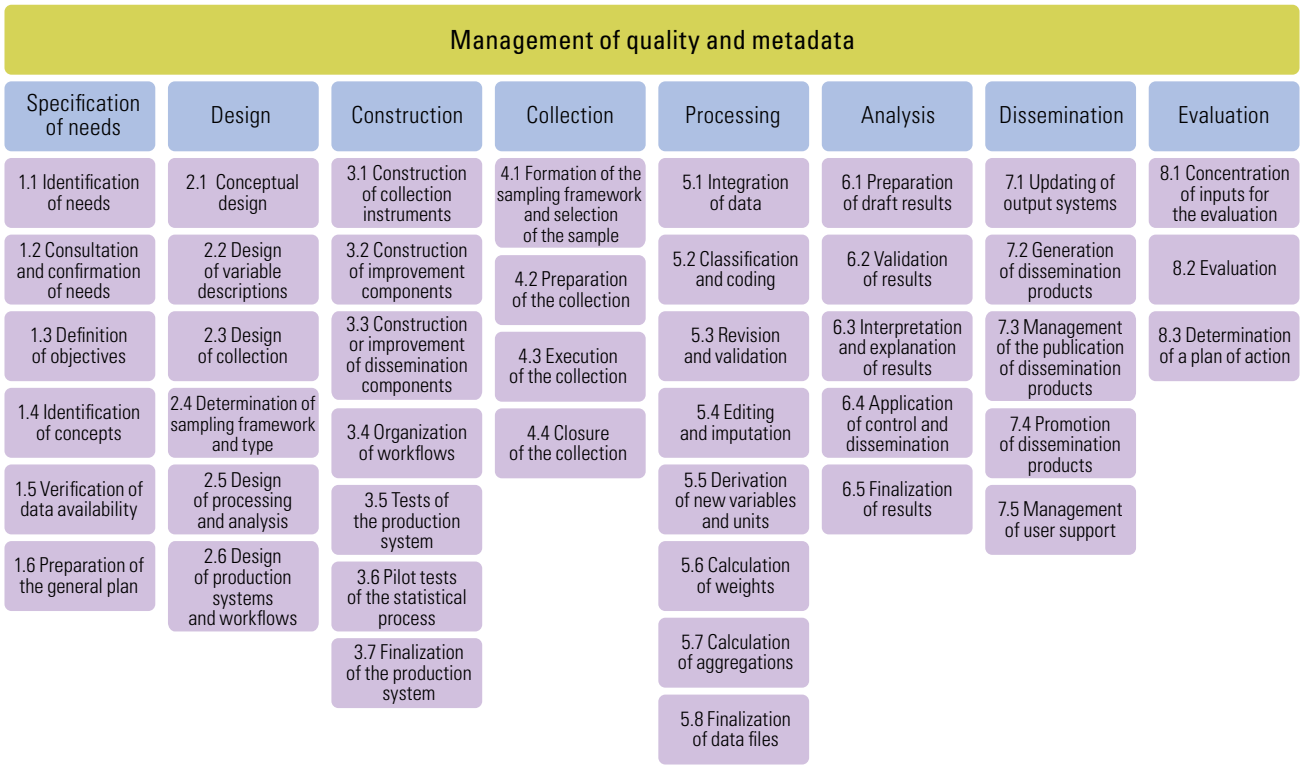
Sample surveys can be conducted under GSBPM proposed by the Economic Commission for Europe.[11] GSBPM is a conceptual framework that describes the different stages and activities involved in the production of official statistics and other statistical products. This generic statistical process model is a useful tool to help statistical offices understand, design, manage and improve the processes through which statistics are produced. GSBPM 5.1 (business segment) consists of eight processes, each of which is made up of a series of subprocesses (see diagram II.1). The processes focus on planning and managing the business aspects of statistical production, and they provide detailed guidance on identifying needs, designing statistical programmes, preparing resources and establishing frameworks for coordination. GSBPM has been adapted by the national statistical institutes of various Latin American countries, including Chile, Colombia, Mexico and Uruguay.

9 Non-probability sampling is used in business surveys to ensure that the economic units most involved in the sector's activities are recorded in all surveys.

10 Sample surveys are those in which random selection methods are used to extract the units to be part of the sample from a list containing all the sample units, which is generally called the sampling frame. Thus, the selection of each sampling unit has a known non-zero probability of being included in the sample and, for this reason, this method is also called the probabilistic selection method.

11 Since November 2013, the Modernisation Committee on Standards, which reports to High-Level Group for the Modernisation of Statistical Production and Services, is responsible for GSBPM. The original version of the main document is available on the Economic Commission for Europe web site (see [online] https://unece.org/statistics/modernstats/gsbpm).

Diagram II.1
Processes and subprocesses of GSBPM 5.1: business segment



Source: Economic Commission for Europe (ECE), "Generic Statistical Business Process Model (GSBPM) version 5.1", Geneva, 2019.

In determining the design of a sample in accordance with subprocess 2.4 of diagram II.1, the following should be considered: constructing the sampling frame, choosing the sampling type, determining the sample size for probability sampling, defining the estimators for probability sampling, identifying the sources of total sampling error and, finally, documenting the sample design, including details on each of the steps. The remainder of this section will provide information on the first two steps: the sampling frame and the type of sampling.

1. Sampling frame

A sampling frame is a listing of all the units in the target population, and one or more samples can be extracted from such a frame. The frame contains information on the main characteristics of the units that make up the target population, together with data through which they can be geolocated, located or contacted. It can be constructed using various sources of information, such as census data, official tax or commercial records, cartographic information and geospatial information.

In constructing a sampling frame, care should be taken to ensure that each of the units in the study universe is included only once, thereby guaranteeing that the sample is representative of the total population. In the case of periodic surveys, mechanisms for updating the sampling frame —to enable the creation of new units and the elimination of those that no longer belong to the study universe— must be established. Sampling frames are generally of two types: area frames and list frames.

Area frames are usually made up of a hierarchy of geographical units: that is, the frame units at one level can be subdivided to form the units at the next level. All the elements included in the frame constitute the frame population. Discrepancies between the survey population and the frame population are known as coverage errors (Statistics Canada, 2003, p. 17).[12]

List frames, in contrast, are a sampling technique used in statistics and surveys to identify a population and to select a representative sample from it. They consist of a complete and detailed list of the population elements to be studied or surveyed (Kish, 1965). List frames are essential for obtaining a representative sample, as they provide a complete enumeration of the population elements to be used in selecting the sample. Population elements can be individuals, companies, households, objects and so on. The list can also include relevant information about each item, such as names, locations, size and demographic characteristics.

2. Sampling types

Depending on the information sought and the characteristics of the study universe, a sampling type must be chosen that ensures that the data collected in the sample are representative of the total population and that collection is carried out in a way that ensures that economic and technical resources are used efficiently.

Sampling can be probabilistic or non-probabilistic. Probability sampling is a better option for collecting information from the entire study population, especially when household surveys are used, and it is suitable for obtaining information that represents an entire population. In probability sampling, each population element has a known, non-zero probability of being selected for the sample. This type of sampling yields accurate estimates of population parameters and guarantees the statistical validity of results. It also allows sampling errors to be measured and confidence intervals around the estimates to be defined, which provides a measurement of the accuracy of the results.

Non-probability sampling, in contrast, is very useful in business surveys, where the participation of the observation units in the different economic variables determines the importance of constantly monitoring some and only intermittently monitoring others: the units that contribute the most to the variables of interest can be included in a non-probabilistic way, while probability sampling can be used for those that contribute less. This type of sampling is useful when a more specific approach is required, targeting certain groups or sectors, as the focus can be placed on specific, more important observation units in certain contexts. In general, researchers establish certain criteria to select some population elements without taking into account the selection probability. In contrast to probability sampling, this method can lead to biases in the results and prevent conclusions from being generalizable to the population as a whole. For this reason, probability sampling is usually preferred over non-probability sampling in designing indicators intended for the definition of public policies. However, the non-probability approach can be more efficient in terms of time and resources, and it can be suitable for exploratory or qualitative studies, or to reveal detailed information on specific cases.

The following paragraphs provide more detailed information on each sampling type.

(a) Probability sampling

Probability sampling is useful for estimating indicators in the population under study. Since each unit is selected randomly, this type of sampling offers several advantages: for example, it eliminates selection bias (Särndal, Swensson and Wretman, 1992), and it enables the use of the units' inclusion probabilities to make reliable estimates and to calculate the associated standard error, which in turn allows inferences to be drawn about the study population (Statistics Canada, 2003, p. 91). Some of the different forms this type of sampling can take are described below.

[12] Coverage error is the difference between the target population and the sample used in a survey or research study. Such errors can occur in two ways: (i) overcoverage, which happens when the sample includes people who are not part of the target population, and (ii) undercoverage, which happens when some people in the target population are not included in the sample.

(i) Simple random sampling

A number is assigned to each of the individuals in the study population, the size of which is N, and those numbers serve as unique identifiers. As many subjects as necessary are chosen to complete the sample size n, with $n \leq N$, by means of a random selection mechanism through which n numbers between 1 and N are drawn. Each number has the same probability of being selected. This procedure is of little or no practical use when the study population is very large; it is also very costly and entails an arduous operational process. It does, however, offer a reference point that can be used to compare other sampling methods. Two types of selection exist: with replacement and without replacement. The entropy or uncertainty of a simple random sample is high: in other words, predicting the type of sample that will be obtained is very difficult. The concept of entropy is useful in estimating the variance of sample estimators. When a sample design has a high level of entropy, an approximation of the second-order probabilities in terms of the first-order inclusion probabilities can be obtained, which is necessary to estimate the variance (Tillé and Haziza, 2010).

(ii) Systematic sampling

All the study population elements are numbered in ascending order (from 1 to N) but, instead of drawing n random numbers, only one is drawn, between 1 and k, where $k=N/n$ and N is the size of the study population. Starting from that random number, the remaining numbers are chosen at regular intervals until the sample is complete. When the study population is in a random order with respect to the variable of interest, this method is the equivalent of simple random sampling but it makes drawing the sample easier. In addition, when the order of the study population tends to present gradual changes in the variable of interest, this method produces lower estimated variances than simple random sampling. This is because the sample is more dispersed over the study population: i.e. the sample is more representative. This type of sampling can be used, for example, to survey opinions on the quality of a company's service, with customers systematically selected from a database.

(iii) Stratified sampling

The study population is divided into strata: in other words, into homogeneous and mutually exclusive classes, such as by age, sex or other categories. The sample is then distributed according to different allocation methods (equal in all strata, proportional to the size of each stratum, Neyman distribution and optimal distribution). This ensures that all strata of interest are represented in the sample. In each sampling stratum, the samples are selected independently, generally using the same selection method except in cases when special strata are created: for example, forced inclusion sampling, in which all elements are selected with probability 1. This type of case generally occurs in surveys of companies or industries in which a very few large companies command considerable shares of sales or worker numbers. In household surveys, which are the focus of this paper, the strata most frequently correspond to geographic areas, identified according to each country's administrative and political division: regions, departments, provinces and so on.

(iv) Single-stage and multi-stage cluster sampling

This is a type of sampling in which the study population is divided into subpopulations, called "clusters", whose elements share a certain characteristic, quality or attribute. An example of these clusters would be groups of dwellings located on two or more city blocks in a defined geographic area. Cluster sampling is used when a complete sampling frame of the target population is not available or would be very expensive to produce, and when data collection costs must be kept down. This occurs mainly when the target population is widely dispersed across an area.

There are two types of cluster sampling: single-stage and multi-stage. The former involves randomly selecting a certain number of clusters (the number necessary to reach the established sample size) and then taking all the elements of each cluster chosen as part of the sample. When the clusters are geographic areas, this type of sampling is generally called area sampling. Single-stage cluster sampling is less accurate than simple random or stratified sampling due to the inherent homogeneity of the sampling units in the selected clusters, since the units are physically close and usually share similar characteristics: in other words, the selection of two or more units from the same cluster may result in redundant information, and this increases the sampling errors of the target indicators.

Multi-stage cluster sampling occurs when the analysis is taken further and, within each cluster, new units are randomly selected and this process is repeated successively until the final units are reached. In certain scenarios sample selection can be a complex process. In some cases it may be difficult to locate each observation unit, leading to a selection by stages. In addition, when the population is too large, it is inefficient to select a sample from it directly, and so it is preferable to form clusters by grouping observation units that share similar characteristics. This grouping reduces the complexity of the selection process and ensures a more representative sample of the target population. Sample selection is a critical aspect of any study, since an improper sample can affect the validity and generalizability of the results obtained. When the study population has been divided into clusters but those clusters range greatly in size, new clusters can be formed within the existing ones, thereby increasing the number of selection stages to two or more. In this situation, one alternative is to combine the clusters with the strata. The clusters are stratified, a certain number are selected from each stratum and, from the groups chosen, the analysis units that will make up the sample are taken.

Two-stage sampling is the most common type of multi-stage sampling: in the first stage, the clusters or areas are randomly selected and, in the second, the last or most basic units of the population as a whole are selected, without the need to select any other type of intermediate unit.

Selection by stages results in lower costs but also leads to an increase in the variance of the estimates. In addition, the calculation of an unbiased variance estimator can become overly complicated, since it involves estimating several variance components associated with each stage. For that reason, simplified variance estimators with a certain level of bias are sometimes used.

(v) Two-phase sampling

Two-phase sampling is somewhat different from multi-stage sampling. Although two-phase sampling also involves taking two samples, these are drawn from the same frame and the selected units have the same structure in each phase. In a two-phase sample, information is collected from a large sample of units, and then more detailed information is collected to obtain a subsample. Two-phase sampling is also known as double sampling, although it may also involve three or more phases. However, as with multi-stage sampling, the design and calculation of sample estimates become more complicated as the number of phases increases.

Double sampling uses the following phases:

- In the first phase, a large sample of s_a units is selected using a simple or stratified sampling design. Auxiliary information is collected —easily and inexpensively— for those units.
- Using the auxiliary information obtained in the first phase, a second phase sample s is selected from s_a, the sample design of which depends on the first sample. The target indicators are recorded in relation to the second sample's units.
- The creation of a frame with useful and reliable information on sample s_a is the basis for the success of double sampling. This type of sampling has been used to make estimates in surveys with high levels of non-responses, or in surveys where it is not easy to identify the target population in the sampling frame.

Examples of population studies where multi-phase sampling is useful include mobile populations (e.g. people who migrate for work or travel frequently), hidden or stigmatized populations (e.g. substance abusers, illegal drug users or sex workers) and low-density populations (e.g. rural or isolated communities).

(vi) Probability proportional to size sampling

This type of sampling design is a particular case of designs with unequal selection probabilities. Probability proportional to size (PPS) sampling is a method in which auxiliary information is used and different selection probabilities are obtained. If the size of the study population units varies and those different sizes are known, they can be used in selecting the sample to improve statistical efficiency. Although it does not eliminate selection bias, PPS sampling can increase the accuracy of the estimates if the size measurements are correlated to the target indicators: in other words, the variance of the estimates will be lower than in an equal probability sampling design (Lavrakas, 2008). PPS sampling, in which the selection probabilities are proportional to unit size, is often

used in household surveys, where the number of dwellings is used as a measurement of population size. Several variants of PPS sampling exist: with replacement, without replacement, Poisson and systematic. With the last of these, large PPS samples can be obtained, but since most of the second-order inclusion probabilities are zero, it does not yield unbiased estimates of the variance.

(vii) Repeated measures sampling

The studies in which this type of sampling is used include, notably, longitudinal surveys in which a probabilistic sample of observation units is tracked over a period of time and several measurements are obtained. The purpose of longitudinal surveys is to collect and analyse data on the growth, change or trends of one or more target indicators over time. This type of sampling generally enables changes in a characteristic of the study population to be measured with greater accuracy than when a series of independent samples is selected. An intermediate design between successively selected independent samples and longitudinal samples is rotating panel sampling, in which a fraction of the sample is replaced each time the survey is conducted. The main purpose of rotating panel samples is to obtain estimates at aggregate levels, and the rotation scheme is designed to control a sample over time, to ensure that cross-sectional estimates are unbiased and to reduce costs and sampling variance.

(b) Non-probability sampling

In non-probability sampling, sample size calculation and selection are based on subjective judgments and criteria; the selection probability of the units in the study population is therefore unknown and accuracy with respect to predefined confidence levels cannot be determined. Nevertheless, this type of sampling —which is also called deterministic— offers a viable alternative when probability sampling would be too costly, when a sampling frame cannot be obtained or when certainty exists that the information collected by means of this method will be sufficiently useful for the purposes of the research.

Several different forms of non-probability sampling exist, and these are discussed below.

(i) Convenience or accidental sampling

This method involves collecting data about the study subjects that are most accessible. It is a quick and low-cost sampling tool, but it has shortcomings as regards representativeness. It is useful as part of the exploratory study process for obtaining guidance in defining the research, but it is not useful for depicting the structures or behaviour of a study population.

(ii) Quota sampling

This type of sampling uses data from subsets or certain strata of the study population —such as sex, age, religion and so on— to select the members that are deemed typical for the purposes of the research. Quota sampling takes its name from the practice of allocating certain proportions of the sample to certain strata of the study population. In the selection process, the person conducting the research decides on the subjects to whom the questionnaire will be administered in accordance with certain previously established general criteria. In practice, those criteria are insufficient to prevent subjective elements from intervening and, for that reason, quota sampling is not used to obtain statistics when a good level of reliability is required. However, solid results have been obtained in practice by combining multi-stage probability sampling with quota sampling in the last sampling units to improve the representativeness of age and sex groups.

(iii) Chain or snowball sampling

Chain or snowball sampling is based on the assumption that the members of a small population know each other. Snowball sampling requires, first, the identification of a small group of people with the characteristics that define the population of interest. Each of those individuals is then asked to identify others with the same characteristics for inclusion in the sample. The process is repeated with those new individuals, who must in turn identify others with the attribute sought. The procedure progresses successively until the desired sample size is reached. Snowball sampling enables a fairly large sample to be drawn from a specific population, but strong

assumptions are needed for the results obtained from such a sample to be generalizable to the study population. Although the identification of members of a small population that would be difficult to locate with other methods is possible with snowball sampling, the sample produced cannot be equated with the results of simple random sampling. Members of the study population with many connections are more likely to be included in the sample than people with few connections, and it is likely that isolated persons cannot be contacted.

(iv) Purposive or judgment sampling

The main characteristic of this sampling method is that both the size of the sample and the selection of its elements are subject to the researcher's judgment; therefore, sufficient knowledge and experience on the subject is required. In this type of sampling, the validity of the results depends on the researcher's knowledge of the phenomenon under study and the availability of statistical data that establish that the selected sample is useful and representative for understanding and analysing patterns, trends or salient characteristics in the target population. The analysis of behaviours carried out through this type of sampling is not limited to specific human actions; it also alludes to patterns, trends or characteristics of the study population. The behaviours can vary according to the research topic, and they can include elements such as consumption habits, attitudes towards a topic, responses to stimuli, cultural practices and so on.

3. Issues to be considered in a sampling scheme

As noted in the previous section, both probability and non-probability sampling are techniques used for research purposes. Important differences exist between the two techniques, however, and choosing between one and the other depends on the goal of the research and the population's characteristics. Probability sampling must be used when the research requires accurate and unbiased estimates of the population, which are essential for the proper planning of public policies because they enable the quality of the results obtained and the confidence in them to be assessed. Non-probability sampling is appropriate when the goal is to understand a particular phenomenon for which random samples cannot be selected.

The choice of a sampling scheme must made in accordance with several principles. Tillé and Wilhelm (2017) identify three: randomization, overrepresentation and restriction. According to the first of these, the sampling design should be as random as possible: i.e. sample selection should be random and unbiased. According to the principle of overrepresentation, units with higher levels of uncertainty should be oversampled, given that the sample should collect as much information as possible from the study population, and that implies a tendency to create sample designs with different selection probabilities. In turn, the principle of restriction involves selecting only those samples that have a set of desirable characteristics; in other words, it entails conducting a verification process (balanced sampling) to avoid samples with empty domains, strata or categories.

Once the sampling scheme has been chosen, various issues concerning the sample must be resolved, in particular its size and the selection procedure.

The first issue to consider is that total sampling error refers to all sources of bias and estimate variance that could affect the accuracy of the sample data and that arise in the sample design and in the collection, processing and analysis of the survey data. According to Groves and others (2004, p. 48), those sources can be classified into two dimensions:

(i) Representation: relating to the target study population to be described in the survey and answering the following question: who is included in the sample?

(ii) Measurement: describing the data to be obtained about the observation unit and answering the following question: what is the survey topic?

Costs, reporting burdens, professionalism, ethics and constraints affect both representation and measurement.

According to Groves and others (2004, p. 48), the representativeness of a sample is affected by four sources of error:

(i) Coverage error: imperfections in the sampling frame lead to incomplete representativeness of the sample with respect to the target population, resulting in poor accuracy of sample estimates and coverage bias.

(ii) Sampling error: occurs when a sample is measured instead of the entire study population. It comprises a fixed error (the bias) and a variable error (the variance). Sampling bias is the systematic failure to observe sample elements that have different values for the target indicator. Sampling variance refers to the variability of the estimates on account of the possible samples obtained, which are affected by the sample size and by clustering, stratification and weighting.

(iii) Non-response error: total or unit non-response occurs when the selected informant cannot be found, when informants refuse to be interviewed or when there are barriers to communication. Non-response bias arises when unit non-response correlates to one or more target indicators.

(iv) Post-survey adjustment error: after data collection has concluded in a survey using probability sampling, adjustments are made to account for selection bias, coverage errors and non-response errors. Due account must be taken of the differences between adjustments made because of the sample design (e.g. adjusting the respondent's selection probability by household size) and those made to eliminate differences between the results of the sample estimate and the available official statistics.

B. Complex sampling designs and their characteristics

Survey sample designs frequently propose different selection and estimation strategies and ultimately choose the one that satisfies the research needs with efficiency, accuracy and cost-effectiveness. In general, selection strategies other than the direct selection of elements are defined, and estimators are chosen to improve the accuracy of the estimates; as a result, the selected sample becomes complex rather than simple.

Complex samples are characterized by the fact that their selection processes entail a number of aspects that differentiate them from direct sampling. Those aspects include the type of frame used, the selection stages or phases, the use of clusters and sample strata, the types of analysis units and so on.

Most household surveys use designs other than simple random sampling. As noted by Heeringa, West and Berglund (2010), stratification is generally used to increase the statistical and administrative efficiency of the sample. Population elements are also grouped into clusters to reduce travel costs and improve interviewing efficiency, resulting in multi-stage sampling designs. Another feature of complex sample designs is that because disproportionate sampling of some population elements is used to increase the sample size in subpopulations of particular interest, weighting must be used in the descriptive estimates of population statistics.

Household surveys usually have an area frame in which people are reached indirectly by defining several stages in the sample. In addition, a particular stratification can be defined in each stage, and the subsamples of these surveys are often used for further research (phases). Sometimes indicators are estimated for dwellings, sometimes for households and sometimes for individuals, which implies a wide variety of analysis units.

All of these characteristics of the sampling schemes typically used in household surveys have an impact on estimator accuracy, and those effects are discussed below.

In addition, in surveys of this type, parameters can be estimated using regression or calibration estimators. In many cases, household surveys are adjusted using those methods. These selection and estimation processes have a direct impact on the estimators, especially those of variance, because of the difficulty of using a mathematical formula to determine all the aspects inherent to those processes.

C. Population parameters and their estimators

As noted by Kish (1965), from the viewpoint of sampling theory, a population parameter is a numerical expression that summarizes the values of one or more characteristics of the N basic units that make up a complete statistical population; thus, it is a summarized measurement of a quality of the distribution of the variable or variables in the defined population. Kish (1965) adds that, depending on the N population values taken by characteristic y in the k-th unit of analysis, the total, the mean, the size of a domain and the proportion are among the population parameters θ that are most commonly estimated. Putting those concepts into practice yields the following formulas:

Population total:
$$Y = f(y_1, ..., y_N) = \sum_{k=1}^{N} y_k \quad (1)$$

Population mean:
$$\mu = \overline{Y} = f(y_1, ..., y_N) = N^{-1} \sum_{k=1}^{N} y_k \quad (2)$$

Domain size:
$$N_d = f(Z_{1d}, ..., Z_{Nd}) = \sum_{k=1}^{N} Z_{kd} \quad (3)$$

Population proportion:
$$P = f(Z_{1d}, ..., Z_{Nd}) = \frac{N_d}{N} = N^{-1} \sum_{k=1}^{N} Z_{kd} \quad (4)$$

Population ratio:
$$R = f(y_{11}, ..., y_{1N}, y_{21}, ..., y_{2N}) = \frac{Y_1}{Y_2} = \frac{\sum_{k=1}^{N} y_{1k}}{\sum_{k=1}^{N} y_{2k}} \quad (5)$$

In the above formulas, Z_{kd} is a dichotomous variable that has a value of 1 when individual k is in domain d and 0 when he or she is not. Other population parameters of a descriptive nature also exist: the ratio between two population totals, the population median, population percentiles, population size and so on. Population analytical indicators, such as the correlation coefficient between variables and the regression coefficients, also exist and could be of interest.

Moreover, Mirás (1985) suggests that, in statistical phenomena, the estimate in finite populations can essentially be approached in the following terms: having defined a characteristic y in population $U=\{u1,u2,...,uk,...,uN\}$, which takes the numerical value y_k in unit u_k, in order to estimate the numerical value θ in function of the N values of y_k, a random sample s is selected and, using the information obtained from measuring the units of that sample, a numeric value of $\hat{\theta}(s)$ is attributed to θ. Therefore, an estimator is a function $\hat{\theta}$ in which each sample s is associated with the numerical value $\hat{\theta}(s)$, which is therefore a particular estimate of θ given the sample s.

Mirás (1985) stresses that the design of an estimator depends on its probability distribution and that this depends on the sampling procedure used; therefore, the formulation of estimators, in the design as a whole, should not be an operation independent of the sampling procedure used.

Martínez (2019) argues that since an estimator $\hat{\theta}$ is a random one-dimensional variable, a good estimator should possess the following characteristics:

- Unbiased. An estimator $\hat{\theta}$ of population parameter θ is said to be unbiased when its bias is zero; in other words, when the expected value of the estimator $\hat{\theta}$ and the value of the parameter θ are equal: $B(\hat{\theta})=E(\hat{\theta})-\theta=0$.
- Consistent. An estimator $\hat{\theta}$ is said to be consistent when, as the sample size increases, it converges in probability towards the population parameter θ being estimated.

Pérez (2005) holds that since the function $\hat{\theta}$ is an estimator of population parameter θ, the values obtained for each sample s of the sample space are called point estimates. Consequently, the problem arises of evaluating the form of the optimal point estimators for the population parameters of interest, which usually involves using unbiased linear estimators of the form $\hat{\theta} = \sum_{k=1}^{n} w_k y_{k'}$ because such estimators usually have the best properties; the w_k values represent the sampling weights or expansion factors.

Specifically, Pérez (2005) argues that when sampling without replacement is used, the optimal choice is the Horvitz-Thompson estimator, $\hat{\theta}_{HT} = \sum_{k=1}^{n} y_k/\pi_{k'}$ where π_k represents the probability of unit u_k being included. The formulas obtained by putting these concepts into practice are provided below.

For sampling without replacement:

Estimated total:
$$\widehat{Y}_{HT} = \sum_{k=1}^{n} y_k/\pi_k \quad (6)$$

Estimated mean:
$$\widehat{\overline{Y}}_{HT} = N^{-1} \sum_{k=1}^{n} y_k/\pi_k \quad (7)$$

Estimated domain size:
$$\widehat{N}_{dHT} = \sum_{k=1}^{n} Z_{kd}/\pi_k \quad (8)$$

Estimated proportion:
$$\widehat{P}_{HT} = N^{-1} \sum_{k=1}^{n} Z_{kd}/\pi_k \quad (9)$$

Estimated ratio:
$$\widehat{R}_{HT} = \sum_{k=1}^{n} y_{1k}/\pi_k \Big/ \sum_{k=1}^{n} y_{2k}/\pi_k \quad (10)$$

D. Confidence interval

The confidence interval is a range or series of values derived from the statistics. It is related to the sample and the sampling design, which also delimits the range in which the population characteristic to be estimated will tend to be found.

The use of confidence intervals requires choosing a confidence level: for example, 90%, 95% or 99%. Since this level varies according to the survey goals and the requirements for estimate accuracy, it is important that the level of confidence used be specified (United Nations, 2009).

If inferences are sought regarding parameter of interest θ within a defined population, a 95% confidence interval is commonly used. To calculate this interval, the critical value $t_{0,975,\,df}$ is used, which corresponds to the percentile of the Student's t-distribution with df degrees of freedom. This value is used to calculate the interval's margin of error in order to ensure that it covers the true value of the parameter with a 95% level of confidence.

Note that when working with large samples, the critical value $t_{0,975,\,df}$ converges to the same confidence level as the percentile of the standard normal distribution denoted by $Z_{1-\alpha/2}$, which is equal to 1.96 when the confidence level is 95%. Therefore, if the sample is large enough, this critical value can be used to calculate the 95% confidence interval instead of the critical value of the Student's t- distribution (INE, 2020b).

The confidence value over the defined population is given by the equation:

$$\left(\hat{\theta} - t_{0,975,df} * se(\hat{\theta}), \hat{\theta} + t_{0,975,df} * se(\hat{\theta})\right) \quad (11)$$

$$LimInf = \hat{\theta} - t_{0,975,df} * se(\hat{\theta}) \quad (12)$$

$$LimSup = \hat{\theta} + t_{0,975,df} * se(\hat{\theta}) \quad (13)$$

where:

θ is the population parameter;

$\hat{\theta}$ is a sampling estimator corresponding to the parameter of interest θ;

$t_{0,975,\ df}$ is the 97.5th percentile of the Student's t-distribution with df degrees of freedom, and

$se(\hat{\theta})$ is the standard error of the estimate, where $se(\hat{\theta}) = \sqrt{v(\hat{\theta})}$.

The length of the confidence intervals influences perceptions of how accurate the estimates are.

The confidence intervals corresponding to the proportions must be contained in the interval (0,1). In some cases, the standard error is large for estimates of proportions close to 0 or 1, and hence the interval limits are less than 0 or greater than 1. In those particular cases, consideration must be given to the possibility of estimating the confidence interval with a variant under which those restrictions can be taken into account. One solution to the problem is to consider a transformation of the estimator. Thus, if $\widehat{P}$ is an estimate of a proportion, the logit transformation of the proportion is defined as (Gutiérrez and others, 2020):

$$\hat{L} = log\left(\frac{\widehat{P}}{1 - \widehat{P}}\right) = logit(\widehat{P}) \tag{14}$$

And the first-order Taylor approximation with respect to $\hat{L}$ is as follows:

$$\hat{L} \cong L(p) + \frac{\partial \hat{L}}{\partial \widehat{P}} / {}_{\widehat{P}=p}\left(\widehat{P} - p\right) = L(p) + \left(\frac{-1}{p(1-p)}\right)\left(\widehat{P} - p\right) \tag{15}$$

Then, the variance of $\hat{L}$ can be written as follows:

$$Var(\hat{L}) = AVar(\hat{L}) = \frac{Var(\widehat{P})}{p^2(1-p)^2} \tag{16}$$

In this way, a $100(1-\alpha)$% confidence interval for L can be defined as follows:

$$\left(\hat{L} - t_{0,975,\ df}\sqrt{Var(\hat{L})}, \hat{L} + t_{0,975,\ df}\sqrt{Var(\hat{L})}\right) = (\hat{L}_1, \hat{L}_2) \tag{17}$$

Finally, the following formula is obtained:

$$\widehat{P} = logit^{-1}(\hat{L}) = \frac{exp(\hat{L})}{1 + exp(\hat{L})} \tag{18}$$

Therefore, the confidence interval corresponding to $\widehat{P}$ is given by the formula:

$$\left(logit(\hat{L}_1), logit(\hat{L}_2)\right) = \left(\frac{exp(\hat{L}_1)}{1 + exp((\hat{L}_1)}, \frac{exp(\hat{L}_2)}{1 + exp((\hat{L}_2)}\right) \subseteq (0,1) \tag{19}$$

E. Estimation of sampling errors

Since variance is a fundamental indicator in determining the accuracy of estimates as one of the factors that determine statistical quality, this section will explore the concept of variance and its estimators, as well as the estimators of other dispersion measurements that are also used to evaluate statistical quality.

1. Variance and its estimators

Currently, the objectives of household sample surveys go far beyond providing, from a descriptive point of view, a series of tabulations that offer point estimates of the parameters of interest. It is increasingly common for them to include analytical objectives aimed at fitting models and testing hypotheses that account for the strength and relationship of variables in the population at different levels. Published results must therefore include appropriate measurements of the precision or accuracy of the estimates derived from the survey data.[13]

Variance is one of the key measurements of accuracy in sample surveys, and it is an indicator of the variability introduced by choosing a sample rather than making a list of the entire population. In addition, calculating variance provides other ways of measuring sampling error, such as the standard error, the coefficient of variation, the design effect and so on. These measurements are characterized by having an algebraic relationship with each other, and so the expression of all the others can be deduced from any given one of them.

(a) Exact formulas

When they can be used, exact methods are the best way to estimate variance. However, most household surveys use sampling designs that are more complex than simple random or stratified sampling, which are the kinds of sampling in which exact methods can be applied.

Exact methods depend on the sample design, the estimate of interest and the weighting procedures used. Table II.1 shows the variance of the most frequently used estimators in simple random sampling, where n is the size of the sample drawn from a population of size N.

Table II. 1
Main parameters, estimators and estimated variance in simple random sampling

Parameter[a]	Estimator	Estimated variance
Population mean $(\bar{y})$	$\hat{\bar{y}} = \sum_{i=1}^{n} y_i / n$	$v(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right) \cdot \frac{S^2}{n}$; with $S^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\bar{y}}\right)^2}{n-1}$
Quantitative total (τ_y)	$\hat{\tau}_y = N \cdot \hat{\bar{y}}$	$v(\hat{\tau}) = N^2 \cdot v(\hat{\bar{y}})$
Proportion of population corresponding to a category d (P_d)	$\hat{P}_d = \sum_{i=1}^{n} I_i / n$ $I_i = \begin{cases} 1, \text{if } i \in \text{class } d \\ 0, \text{if otherwise} \end{cases}$	$v(\hat{P}_d) = \left(1 - \frac{n}{N}\right) \cdot \frac{\hat{P}_d \cdot (1 - \hat{P}_d)}{n-1}$
Ratio between variable and variable x $r_{y/x} = \frac{\tau_y}{\tau_x} = \frac{\bar{y}}{\bar{x}}$	$\hat{r}_{y/x} = \frac{\hat{\tau}_y}{\hat{\tau}_x} = \frac{\bar{y}}{\bar{x}}$	$v(\hat{r}_{y/x}) \doteq \left(1 - \frac{n}{N}\right) \cdot \left(S_y^2 - 2 \cdot \hat{r}_{y/x} \cdot cov(x;y) + \hat{r}_{y/x}^2 \cdot S_x^2\right)$

Source: W. Cochran, *Técnicas de muestreo*, Mexico City, Compañía Editorial Continental (CECSA), 1980.

[a] A parameter is a real function calculated by taking all the population elements into consideration in its calculation. All known descriptive statistics represent parameters if they are obtained using all the population elements. Some examples of those statistics are means, proportions, totals, ratios, variances, percentiles and others derived from algebraic operations performed among them, such as the coefficient of variation, which is the standard deviation divided by the mean. When real functions are applied with only the elements of the selected sample taken into consideration in order to estimate the parameter in question, they are called estimators.

[13] Measurements of statistical accuracy indicate the proximity between the estimator from the survey and the parameter's true value. They are also used as an estimate quality metric.

(b) Final cluster method

The final cluster method for estimating variance is used for estimates based on a sample drawn from a complex sample design, usually multi-stage, clustered and stratified (Hansen, Hurwitz and Madow, 1953, pp. 257–259).

According to this method, the final cluster is obtained by considering the primary sampling unit (PSU) as the last information unit containing information from all the sampling units of the subsequent stages that are nested in it.

Thus, the sampling design resembles single-stage stratified cluster sampling: i.e. a stratified sample of fully enumerated final (last) clusters. This final cluster approach yields a good approximation of estimated variance (Gutiérrez, 2022) and, for that reason, its use is widespread among national statistical offices. This implies that variance is calculated by the intervariance or inter-cluster variance using only inter-PSU totals, without requiring the variance components at each selection stage to be calculated.

The final cluster method and simplified variance estimate in complex sampling are used to estimate population size using the same expansion factors. Unlike other methods that will be discussed later, this approach enables the population size to be established and the finite population correction to be applied to reduce the variance, which determines the populations. This means that when the final cluster method is used, greater accuracy in population estimates is achieved by keeping the populations fixed and adjusting the variance through the finite population correction.

(c) Taylor linearization approximations

Many of the descriptive statistics to be estimated or the estimates of interest are not simple linear functions of the observed values; hence, sampling variance cannot be expressed by a closed-form formula, as is the case with the variance of the sample mean in simple random or stratified sampling.

The linearization method is widely used because it can be applied to almost any sample design and to any statistic that can be linearized or expressed as a linear function of common statistics, such as means or totals, whose coefficients are drawn from partial derivatives necessary to achieve the Taylor series expansion.

Linearization is generally used to estimate ratios or quotients of two variables that can be expressed in an infinite Taylor series centred on the expected (estimated) value of the numerator and the expected (estimated) value of the denominator.

The non-linear estimator is then subjected to an algebraic approximation that retains only the first terms of the infinite Taylor series. This produces an algebraic expression that is a linear function of the sample data, which means that the non-linear estimator ratio has been linearized and that the estimated variance of the linearized function (including the relevant covariance terms) can now be obtained directly.

Once linearized, the variance of the non-linear estimate can be approximated by using the exact methods described above.[14]

(d) Replicated weights

This method involves taking subsamples or replicates of the total sample. First, the variable of interest Y is estimated using the total sample according to the probabilistic design, whatever it is. Replicates are then created or subsamples are selected from the total sample, such that each reflects the sampling plan, adjustments and weighting procedures of the full sample and enables the same estimate as derived from the full sample to be arrived at.

There are several techniques for applying this method, and some software programs already include replicate selection modules. Among those modules, the most commonly used techniques are randomized groups, balanced repeated replication (McCarthy, 1969; Judkins, 1990), replication using the jackknife method (Krewski and Rao, 1981) and the bootstrap and sub-bootstrap techniques (Rao and Wu, 1984).

[14] For detailed technical information on the linearization process, illustrated with examples, see Cochran (1977), Lohr (1999), Heeringa, West and Berglund (2010) and Valliant, Dever and Kreuter (2018).

Suppose, for example, that K replicates are created from a given sample, that an estimate $\hat{\theta}_1, \hat{\theta}_{2,\ldots} \hat{\theta}_k$ of parameter $\hat{\theta}$ corresponds to each of these replicates and that, additionally, the estimate based on the full sample is $\hat{\theta}_0$. In this case, the estimated variance based on replication is obtained by means of the following formula: $V(\hat{\theta}) = \frac{1}{C}\sum_{r=1}^{K}(\hat{\theta}_r - \hat{\theta}_0)^2$, where c is a constant that depends on the estimation method.

These variance estimation methods are old and, previously, few computer programs included or addressed them. However, because of technological progress and advancements with computer technology in particular, most statistical programs nowadays offer special modules for processing data obtained from complex samples, where a complex sample is understood as one in which weighting, stratification and clustering variables are included to estimate the main statistical data associated with a variable of interest and also to estimate their variances.

2. Measurements of dispersion associated with variance

Given the parameter estimate and the variance estimate as a measurement of dispersion by itself, other associated measurements can be obtained, such as the mean squared error *(MSE)*, the standard error *(SE)*, the coefficient of variation *(CV)*, the absolute error (E_A), the relative error (E_R) and the confidence interval *(CI)*. Similarly, depending on the design type, another measurement of dispersion that relates the variance obtained from a complex design to that obtained from simple random sampling can be obtained. This measurement is called the design effect *(Deff)*.

(a) Mean squared error

The mean squared error *(MSE)* or mean squared deviation (MSD) of any estimator $\hat{\theta}$ has two components: one directly related to the variability itself $(\sigma^2[\hat{\theta}])$ and the other registering the difference between the estimator's expected value and the true value of the parameter $(E[\hat{\theta}] - \theta)$.[15] That difference, called the bias of the estimator, is almost impossible to obtain directly because in practice the value of parameter θ is unknown (this bias can be approximated by means of simulations). *MSE* is expressed as follows:

$$MSE(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = \sigma^2[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2 \quad (20)$$

where

$$\sigma^2[\hat{\theta}] = Varianza; Sesgo(\hat{\theta}) = E[\hat{\theta}] - \theta \quad (21)$$

The estimator is said to be unbiased if $Bias(\hat{\theta}) = 0 \Leftrightarrow E[\hat{\theta}] = \theta$

In sample surveys, unbiased or asymptotically unbiased estimators are used, so that the bias is negligible or vanishes with larger sample sizes; therefore, accuracy measurements focus on the variance or its square root —which is called the standard error— or on a transformation of it.

Noted that *MSE* is a function of the same parameter θ, which is usually unknown and is estimated by means of $\hat{\theta}$. By calculating $E[\hat{\theta}]$—i.e. the average of the estimates based on all possible samples of a certain size— the value of the parameter is arrived at, provided the estimator is unbiased. In practice, if the estimator is unbiased, a value of the unknown parameter that is approximately close can be arrived at on the basis of a single sample.

When the estimator is unbiased, *MSE* matches the variance of the estimator, and an unbiased estimator is also said to be more efficient than another unbiased one when its variance is smaller. In general, *MSE* is used to compare two or more estimators, whether unbiased or not, and the one with the lowest value is deemed the most accurate or efficient. If it is a biased estimator, it should be consistent: i.e. as the sample size increases, both the bias and the variance should decrease and converge (asymptotically) to zero.

[15] References to MSE in this document actually mean empirical MSE, which assumes that the population parameter to be estimated is known and has a given or fixed value that could correspond to an administrative record or an exogenous estimate that, in general, is obtained by means of another estimator. That exogenous estimator is assumed to be the population estimator, and so an empirical bias is calculated as the difference between the value obtained with an estimator and the assumed population value.

(b) Standard error

This indicator provides information on how dispersed the estimator is.[16] This metric is easier to interpret than variance, since it uses the same measurement scale as the estimate:

$$SE(\hat{\theta}) = \sqrt{\sigma^2(\hat{\theta})} = \sigma(\hat{\theta}) \tag{22}$$

(c) Coefficient of variation

The coefficient of variation (CV) is a dimensionless measurement —in other words, a pure number that has no units— with which CV of different distributions of variables expressed in different units can be compared. It is calculated as the ratio between the standard error and the estimated parameter:

$$CV(\hat{\theta}) = \sqrt{\sigma^2(\hat{\theta})} \Big/ \hat{\theta} \tag{23}$$

It can also be expressed as a percentage:

$$100 \cdot CV(\hat{\theta})\% = 100 \cdot \sqrt{\sigma^2(\hat{\theta})} \Big/ \hat{\theta}\ \% \tag{24}$$

(d) Absolute error and confidence interval

The absolute error associated with an estimate can be interpreted as the maximum difference expected to be found between the parameter's estimator and its (true) value, with a probability of $100(1-\alpha)\%$, where $(1-\alpha)$ is the confidence level used or the probability that parameter θ is contained in the constructed interval.[17]

The probability that the distance between the parameter estimated in the sample and the parameter's true value equals, at most, the defined absolute error $E_A(\hat{\theta})$ is expressed as:

$$P(|\hat{\theta} - \theta| \leq E_A(\hat{\theta})) = 1 - \alpha \tag{25}$$

Developing equation (25) and resolving the parentheses for parameter θ yields the following equation, which has the form of a confidence interval or interval estimate:

$$P[\hat{\theta} - E_A(\hat{\theta}) \leq \theta \leq \hat{\theta} + E_A(\hat{\theta})] = 1 - \alpha \Leftrightarrow \theta \in \left[\hat{\theta} - E_A(\hat{\theta}); \hat{\theta} + E_A(\hat{\theta})\right]_{1-\alpha} \tag{26}$$

Using the law of large numbers and the central limit theorem, if the sample size is large enough, the probability distribution of the estimator converges to the normal distribution. In practice, however, estimator $\hat{\theta}$ is distributed as a Student's t (the population variance is unknown and is estimated by the same sample), the mean or expectation is equal to the value of parameter θ and the variance equals $\sigma^2(\hat{\theta})$. Resolving equation (26) yields the following equation for the absolute error:

$$E_A(\hat{\theta}) = t^{v}_{1-\alpha/2} \cdot \sqrt{\sigma^2(\hat{\theta})} = t^{v}_{1-\alpha/2} \cdot EE(\hat{\theta}) \tag{27}$$

[16] An indicator or measurement is a descriptive statistic or statistical graph: i.e. a function of the values taken by the variable of interest in the sample, such as the minimum, maximum, average, variance and so on. An indicator is also a function of two other statistics: for example, the coefficient of variation (standard error divided by the mean) or the unemployment rate (total unemployed divided by total workforce).

[17] In probabilistic terms, in the context of repeated sampling —i.e. when 100% of the intervals are constructed with samples of the same size and same confidence level— the probability that the interval contains the population parameter is approximated as the percentage of intervals that will contain the parameter.

where $t^{v}_{1-\alpha/2}$ is the percentile of the Student's t distribution with v degrees of freedom which, in the case of large samples (more than 30 sampling units), converges, at the same confidence level, to the percentile of the standard normal distribution denoted by $Z_{1-\alpha/2}$, which is 1.96 when the confidence level is 95%.

(e) Relative error

The relative error associated with an estimate can be interpreted as the maximum percentage variation that can be expected between the estimated parameter and the true population parameter with a probability of $(1-\alpha)$. Relative error is defined as the ratio between the absolute error and the estimate:

$$E_R(\hat{\theta}) = t^{v}_{1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2(\hat{\theta})} \Big/ \hat{\theta} = t^{v}_{1-\alpha/2} \cdot \frac{SE(\hat{\theta})}{\hat{\theta}} = t^{v}_{1-\alpha/2} \cdot CV(\hat{\theta}) = \frac{E_A(\hat{\theta})}{\hat{\theta}} \quad (28)$$

Formula (28) shows the relationship between the different measurements of accuracy defined above.

(f) Design effect

Kish (1965) defines the design effect $(Deff)$ as the ratio of the variance of an estimate in a sample obtained through a complex design $\sigma^2_c(\hat{\theta})$ to the variance of the same estimate in a sample obtained by means of simple random sampling $\sigma^2_{SRS}(\hat{\theta})$ (United Nations, 2009). He also defines it as the factor by which the variance of an estimate based on a simple random sample of the same size must be multiplied to take account of the complexities of the sample design arising from factors such as stratification, clustering or weighting. The design effect therefore is expressed as follows:

$$Deff(\hat{\theta}) = \frac{\sigma^2_c(\hat{\theta})}{\sigma^2_{SRS}(\hat{\theta})} \iff \sigma^2_c(\hat{\theta}) = \sigma^2_{SRS}(\hat{\theta}) \cdot Deff(\hat{\theta}) \quad (29)$$

The design effect depends on the estimator and the sample design itself: in other words, given a complex sample design, different design effects are obtained depending on the parameter of interest $\hat{\theta}$ that is to be estimated (Chambers and Skinner, 2003; United Nations, 2009). In addition, it must be noted that the sample is not observed in the same way if the designs are different. For example, in a complex two-stage cluster design, the sample is grouped into the selected clusters; in a simple randomized design, however, the same sample's distribution is very different. The only way for the expected variance to be the same in both designs is for the phenomenon to behave similarly in each cluster, or for each cluster to be a faithful or representative copy of the population.

At the same time, if n is calculated using the simple random sampling formula, then $n_c = n \cdot Deff$ is the sample size needed in a complex sample design to achieve the same variance as yielded by simple random sampling. This gives rise to the concept of effective sample size, which is expressed as: $n_{ef} = \frac{n_c}{Deff}$. Thus, if $Deff$ is equal to 2 and the sample size in a complex design is 60, the effective sample size (in a simple random design) is 30.

Finally, it should be noted that the design effect depends on each particular estimator rather than being a global measurement of a survey; as a result, in the same survey, the design effect can vary according to the estimator and the level at which it is evaluated (Chambers and Skinner, 2003; United Nations, 2009).[18]

[18] For more information on the design effect in household surveys, see Gutiérrez (2022).

3. Tabulations or disaggregations

A tabulation or statistical table is a matrix configuration of rows and columns divided into cells containing aggregated information on descriptive statistics from a statistical operation or survey.

Note that cells containing some descriptive statistic —such as an unweighted count, total, average, median, standard error, coefficient of variation or other summary metrics— are themselves a tabulation that, together with other cells containing the same summarized metrics, make up the statistical tables that are published. Sometimes the descriptive statistics of the most relevance to the publication's purpose are all placed together: for example, the weighted and unweighted count, the estimate of the parameter of interest and the dispersion measurement. In general, however, they are presented separately.

For illustrative purposes only, table II.2 shows the total number of people by sex, broken down by socioeconomic level. The descriptive statistics —estimates, standard errors and coefficients of variation, respectively— are presented separately.

Table II.2
Estimated number of people by sex, according to socioeconomic level

Variable	Estimates			Standard errors			Coefficients of variation		
Socioeconomic level	Sex			Sex			Sex		
	Total	Women	Men	Total	Women	Men	Total	Women	Men
Total	t_{00}	t_{01}	t_{02}	se_{00}	se_{01}	se_{02}	cv_{00}	cv_{01}	cv_{02}
Low	t_{10}	t_{11}	t_{12}	se_{10}	se_{11}	se_{12}	cv_{10}	cv_{11}	cv_{12}
Medium	t_{20}	t_{21}	t_{22}	se_{20}	se_{21}	se_{22}	cv_{20}	cv_{21}	cv_{22}
High	t_{30}	t_{31}	t_{32}	se_{30}	se_{31}	se_{32}	cv_{30}	cv_{31}	cv_{32}

Source: Economic Commission for Latin America and the Caribbean (ECLAC).

Although each of the descriptive statistics separately constitutes a tabulation, it is the tabulation of the estimates that is analysed to determine whether it can be published without reliability warning restrictions (reliable, partially reliable or unreliable). In this tabulation, the standard error is analysed, with a range of between 0 and 1, when it is a ratio estimator (or, as a particular case thereof, a proportion estimator). For all the other types of estimators, the coefficient of variation is analysed.

The variables that produce a tabulation are categorical or categorized variables (in the case of numerical variables), and each category represents a subpopulation or portion of the total population. Thus, dealing solely with the statistical table of estimates, there are four tabulations: the total of totals (t_{00}), the marginal totals by socioeconomic level (t_{10}, t_{20} and t_{30}) and by sex (t_{01} and t_{02}), and the cells of the main tabulation containing the interaction between socioeconomic level and sex (t_{11}, t_{12}, t_{21}, t_{22}, t_{31}, t_{32}).

First, if the accuracy of the total of totals t_{00} presents problems, the entire statistical table should not be analysed or published.

Second, the marginal totals of the socioeconomic level and sex variables are evaluated separately. Since the sex variable cannot be collapsed because the gender analysis would be rendered meaningless, if one of the categories (female or male) lacked appropriate accuracy according to the standard, only the statistically reliable category can be analysed.

As for socioeconomic level, the main study variable, each of its categories must be statistically reliable because, if that were not the case, at least one of the sex categories would be unreliable and that would diminish the amount of comparative gender analysis that could be performed. In such a case it is advisable, whenever possible and when it makes sense according to the study topic, to collapse adjacent categories until the new collapsed category is reliable.

Third, once the categories of the socioeconomic level variable have been collapsed, its interactions with the sex variable are evaluated in the main tabulation. At that point the standard is applied to determine whether the cells in the main tabulation are reliable, partially reliable or unreliable, and then the tabulation is assessed to determine whether or not it is publishable. The point is that before the standard is applied to the statistical table, efforts must be made to ensure that the socioeconomic level variable has been adequately collapsed. This is because it guarantees that when the evaluation begins in accordance with the standard flow chart, the tabulation is ready and no further adjustments to the categories will be necessary.

Chapter III

Elements for assessing the statistical quality of an estimate

Principle 6 of the Regional Code of Good Practice in Statistics for Latin America and the Caribbean establishes a commitment to quality, stating that "entities that produce statistics within the national statistical system must work and cooperate in accordance with rules, principles and standards" (ECLAC, 2011, p. 9). All the region's national statistical offices base each stage of the statistical production process on that principle. This document focuses on quality guidelines associated with the results production stage.

In evaluating the quality of estimates, it is useful to consider a series of factors or criteria to help determine which indicators should be applied in each case. This is because the suitability of each criterion may vary depending on factors such as the estimator type, the estimated value and the estimation levels.

The quality criteria presented below are based on the main regional references for the formulation of guidelines on the statistical quality of estimates: the documents "Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional", published by ECLAC (Gutiérrez and others, 2020), and the "Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares", published by the National Institute of Statistics of Chile (INE, 2020a).

A. Design effect

According to Lumley (2010), in large studies the design effects *(Deff)* are usually greater than 1.0, which means that complex designs require larger sample sizes than simple randomized designs. This can be seen by defining *Deff* as a ratio between the variance of a complex sample estimator and the variance from a simple random sample. In theory, the variance in simple random sampling should be smaller than in complex random sampling. In practice, however, *Deff* may be less than 1 or much higher than expected given the sample design or the observations at different estimation levels.

In surveys with complex sample designs, an estimate of *Deff* of less than 1 may be due to the fact that the variance calculation algorithms are based on approximations in which the final cluster method is generally applied, the use of which may result in an underestimation of the variance that is lower than would be obtained with a simple randomized design. Given that situation, and under the assumption that the sampling error of a multi-stage sample is greater than that of a simple random sample, the use of a design effect bounded by 1 in the calculations can be justified.

In addition, if researchers believe that the estimated design effect significantly exceeds the expected levels, they may decide to use upper bounds in the calculation, under the assumption that the high values observed are due to particular design factors and the characteristics of the collected sample, since the inference is drawn from a particular sample out of a large number of possible ones.

In particular, Gutiérrez (2022) recommends placing a warning flag on all figures with a design effect of less than 1.

The situations described above constitute a first warning about estimate quality, so it is the responsibility of the researcher to investigate the factors that could have given rise to that result.

B. Minimum sample size to apply accuracy measurements

Among the factors that affect the magnitude of the sampling variance are the heterogeneity of the study variable, the sample size and the sampling design. As noted by Gutiérrez and others (2020), sample size indirectly affects the width of the confidence interval through the standard error, which generally decreases as sample size increases.

Once the sample's characteristics have been specified, its size (n) must be determined, taking into account that it must be sufficiently representative of the population and must ensure, given a confidence level of $(1-\alpha)$, that the sampling error does not exceed a maximum admissible value.

Similarly, the European Commission (2013) indicates that there are two main strategies for establishing accuracy requirements: first, determining the accuracy thresholds that the main target indicators of the survey must meet, and, second, establishing the minimum effective sample size.

1. Sample size

As explained in INE (2020b), sample size is given by the total number of analysis units considered to obtain the estimates. It is therefore important for the sample to be of sufficient size for estimates of the study population to be made and, if appropriate, to be broken down by sex for the variables of greatest interest.

Both Gutiérrez and others (2020) and INE (2020a) identify sample size as an aspect to be examined in determining an estimate's statistical quality. Gutiérrez and others (2020) note that an adequate sample size ensures that the distribution of the estimators converges to the theoretical distribution from which the percentiles are calculated for determining the confidence interval. Moreover, determining the sample size is of vital importance in identifying the degree of accuracy of the statistical operation. However, there are practical situations in which the minimum sample size is not determined on the basis of criteria of accuracy. Budgetary constraints commonly limit the total number of interviews that can be conducted, so estimates have to be calculated in the absence of a degree of accuracy given by the survey's sample design. In such cases it is even more imperative that sample size be considered as a criterion of statistical quality.

Gutiérrez and others (2020) refer to the suggestions of Barnett-Walker and others (2003), who hold that in order for an estimate to be obtained, there should be at least 100 units of analysis. However, INE (2020a) maintains that 60 units are the minimum necessary for an estimate to be evaluated under the second quality criterion. If an estimate is based on fewer than 60 units, it is directly classified as unreliable and its use is not recommended.

With regard to the mechanism for controlling non-response, the European Commission (2013) notes that, for European Union Statistics on Income and Living Conditions, Commission Regulation No. 1982/2003 provides that the accuracy requirements for the publication of data must be expressed in terms of the number of sample observations on which the statistic is based and the level of partial non-response (in addition to total non-response). It thus stipulates:

- The Commission shall not publish an estimate if it is based on fewer than 20 sample observations, or if non-response for the item concerned exceeds 50%.
- The Commission shall publish the data with a flag if the estimate is based on 20 to 49 sample observations, or if non-response for the item concerned exceeds 20% and is 50% or below.
- The Commission shall publish the data in the normal way when they are based on 50 or more sample observations and item non-response does not exceed 20%.

The recommendations of the Department of Statistics and Censuses (DIGESTYC, 2022) state that El Salvador's Multipurpose Household Survey is to use a minimum of 14 degrees of freedom, which implies at least 15 PSUs. In addition, taking into consideration the 12 units that are selected in the second selection stage gives a minimum of 180 households. Assuming a response rate of 85%, the survey sample is expected to have a minimum of 153 items. For other studies, the suggestion is for the sample to contain at least 100 elements.

2. Effective sample size

The effective sample size is the minimum size a sample must have so that, when simple random sampling without replacement is applied, the same level of accuracy as with a complex sample design is obtained. From this it follows that the minimum effective sample size for simple random sampling must be adjusted for the design effect *Deff*: in other words, for the variation that the sampling design's components —such as stratification or clustering— cause in the sampling efficiency (European Commission, 2013).

As noted by Gutiérrez and others (2020), in household surveys, the design involves selecting a set of households within the same PSU and repeating that selection strategy systematically across the country. It can therefore be concluded that if the variable of interest has a high intraclass correlation, then the reality of individuals and households within the same PSU will have a high level of homogeneity, to the extent that the interpretation could be that the information is repeated and that individuals or households within the same PSU do not provide information in a differentiated manner. Because of the effects of the complex sampling design, therefore, the number of persons contributing to the inference of the indicator is neither the number of persons nor the number of households in the sample, but rather its effective size, n_{eff}, which enables the effects of agglomeration to be deflated.

In line with the above, the effective sample size is calculated by (Kalton, Brick and Lê, 2005):

$$n_{eff} = \frac{n}{Deff} \tag{30}$$

For reference values for evaluating the effective sample size, Gutiérrez and others (2020) uses the thresholds considered in Hornik and others (2002) and Klein and others (2002), where a minimum of 30 units is set.

3. Degrees of freedom

In 1915, Sir Ronald Fisher, the founder of modern statistics, introduced the concept of degrees of freedom in connection with sets of observations. These degrees are given by the number of values that can be arbitrarily assigned before the rest of the variables —once the free ones have been set— automatically take on a value in order to compensate for and equalize a previously known result. According toWalker (1940), in geometric terms, the number of observations is the dimensionality of the original space and each relationship represents a section through that space restricting the sample point to a space of one dimension less. Imposing a relationship upon the observations is equivalent to estimating a parameter from them.

The rationale for using degrees of freedom in assessing statistical quality is that these degrees offer an important element for calculating absolute error and determining the width of confidence intervals according to the associated probability distribution and also according to their relationship to sample size.

As can be seen in equation (31), the width of the confidence interval of the estimator of the parameter of interest $(\hat{\theta})$ depends on the sampling error in terms of the standard error *(se)* and the 0.975 percentile of the Student's t-distribution with the corresponding degrees of freedom *(df)*.

$$(\hat{\theta} - t_{0,975,df} * se(\hat{\theta}), \hat{\theta} + t_{0,975,df} * se(\hat{\theta})) \tag{31}$$

As noted in INE (2020b), in surveys that use a complex sample design, the calculation of the degrees of freedom to estimate the variance is not trivial. Currently, statistical packages for analysing data obtained from complex samples use the fixed degrees of freedom rule, which entails obtaining the degrees of freedom by subtracting the total number of strata from the total number of PSUs.

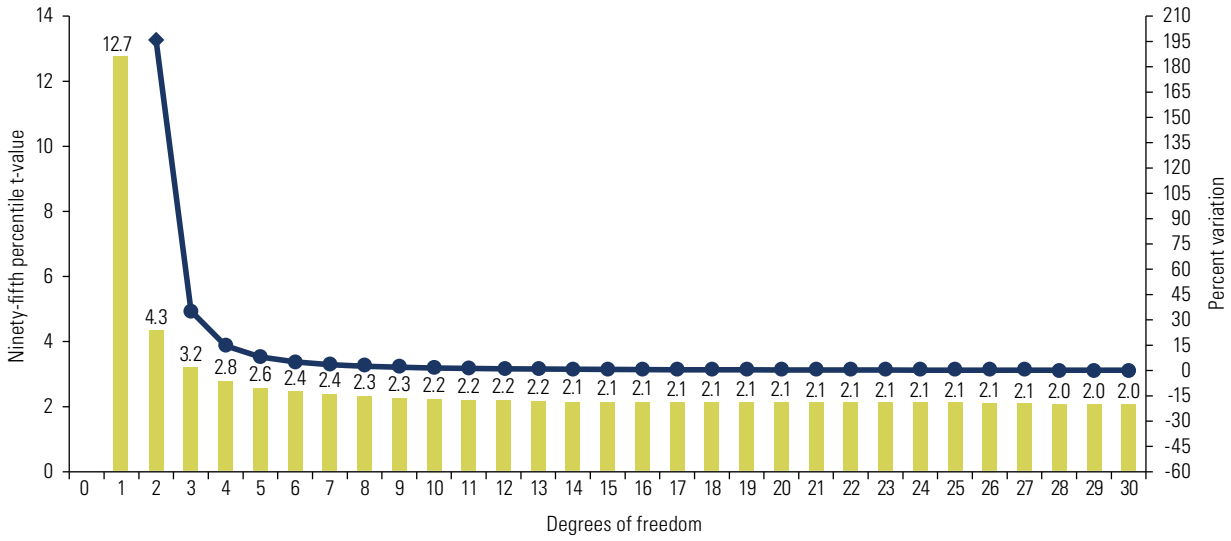
$$df=number\ of\ PSUs-number\ of\ strata \tag{32}$$

However, evidence exists that this methodology leads to an overestimation of the degrees of freedom and, consequently, yields narrower confidence intervals. Alternatively, Korn and Graubard (1999) suggest calculating variable degrees of freedom. With this method, the degrees of freedom are obtained by subtracting the number of sampled PSUs containing the subpopulation and the total number of strata with subpopulation samples from the total number of sampled clusters in the subpopulation. This method is more convenient, especially when the characteristic of interest is not very prevalent.

As for evaluating degrees of freedom as an indicator of quality, Gutiérrez and others (2020) refer to Parker and others (2017), who hold that if the degrees of freedom induced by the subpopulation are fewer than eight, the figure should be omitted.

Similarly, INE (2020b) analyses the behaviour of the ninety-fifth percentile of Student's t-distribution (the percentile associated with a 95% confidence level), since it is the one most commonly used to calculate confidence intervals. Figure III.1highlights the pronounced behaviour of the distribution when the degrees of freedom are few. Thus, it can be seen that when there are 1, 2 and 3 degrees of freedom, the associated values t of the percentile are 12.7, 4.3 and 3.2, respectively, which is equivalent to a percentage variation of between 195.3% and 35.2%. As the number of degrees of freedom increases, the percentage variation decreases and, when there are between eight and nine degrees of freedom, the variation stabilizes.

Figure III.1
0.95 percentile of Student's t-distribution and percentage variation according to degrees of freedom



Source: National Institute of Statistics of Chile (INE), "Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares", *Documento de Trabajo*, Santiago, No. 13, 2020.

Note that if degrees of freedom are used as a criterion for omission, designers and planners of surveys that use two-stage sampling designs in each country should bear in mind that for a figure to be published in a domain of interest, it must have a sufficient number of PSUs (at least 11) so that the rule does not limit its dissemination (Gutiérrez and others, 2020).

C. Types of estimators

As indicated by the European Commission (2013), estimator type plays a fundamental role in defining the criteria for constructing standards to evaluate the statistical quality of estimates, since the accuracy measurements must be adapted to the type of estimator.

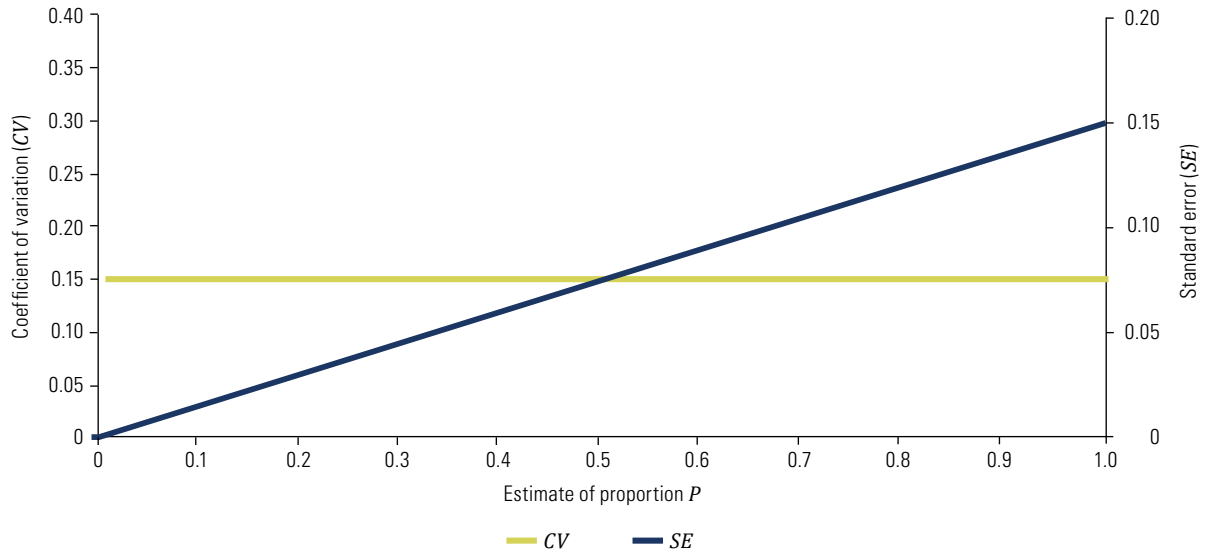
Various analyses have shown that when the estimator is a proportion or a ratio, the coefficient of variation has shortcomings as a method for evaluating the accuracy of the estimate. According to the European Commission (2013), this is because the value of the percentage or proportion has a strong impact on the value of the coefficient of variation, especially when the percentage or proportion is low, and because the coefficients of variation for the percentages or proportions of any characteristic are not symmetrical. Thus, the European Commission (2015) states that in household surveys, results are often presented as proportions or percentages, and that confidence intervals are a better option for presenting the random sampling error associated with the estimate.

In INE (2020a), for example, the accuracy of the estimates is evaluated by considering two classes of estimators: (i) those representing proportions and ratios whose value is between 0 and 1, and (ii) the remaining indicators. In line with this, INE (2020b) analyses the behaviour of the standard error given the coefficient of variation for different levels of the proportion or ratio, applying two initial criteria:

(i) According to the first criterion, the estimate of the proportion is of good quality if the associated coefficient of variation (CV($\hat{p}$)) is less than or equal to 0.15 (15%).

By setting an upper bound for the coefficient of variation, and since the standard error *(se)* is a linear function of it ($se(\hat{p}) = CV(\hat{p}) \cdot \hat{p}$), higher standard errors are permissible as the proportion increases. When prevalence is low, however, very low standard errors are required, which does not necessarily mean that the confidence interval generated is not useful for decision-making (see figure III.2). For example, estimates of P that are close to 0.05 can have a reasonably wide confidence interval when the coefficient of variation is greater than 0.15, and so the establishment of a second criterion that involves setting upper bounds for the coefficient of variationaccording to the values taken by p (the estimate of P) has been proposed.

Figure III.2
Maximum coefficient of variation and standard error admitted according to the estimate of P (first criterion)

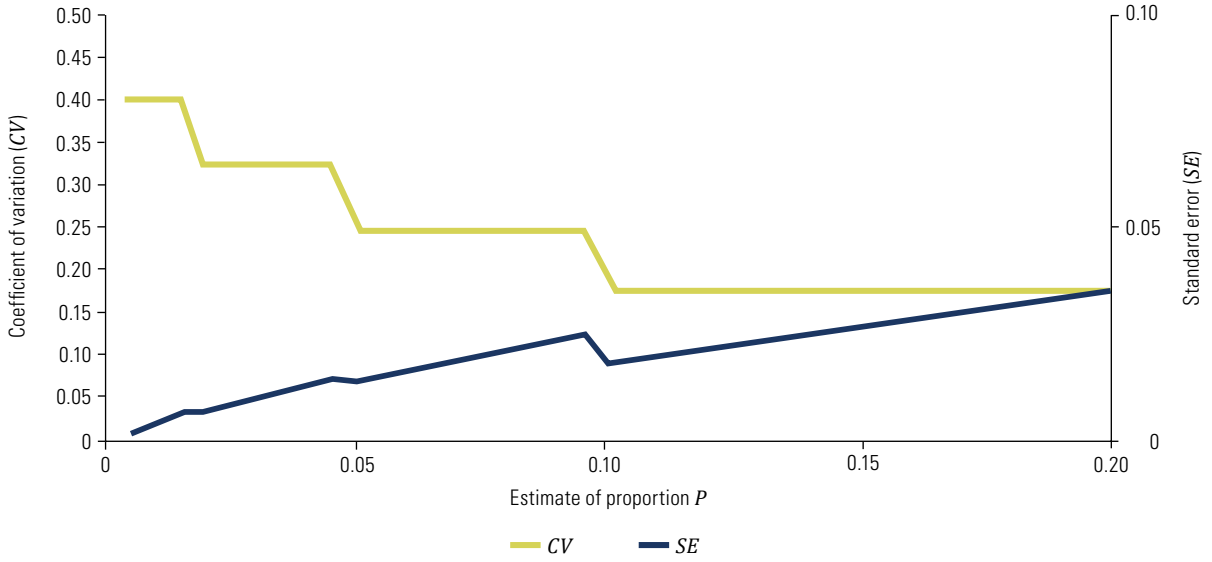


Source: National Institute of Statistics of Chile (INE), "Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares", *Documento de Trabajo*, Santiago, No. 13, 2020.

(ii) According to the second criterion, the estimate of the proportion is of good quality if the associated coefficient of variation (CV($\hat{p}$)) coincides with that shown on figure III.3.

When the bounds are defined according to the values of p, the standard error follows a rising linear trend with slight decreases around the values of p where the coefficient of variation bounds change, i.e. at values 0.02, 0.05 and 0.10. This means that in some cases, a lower value of p is required to have a higher standard error, which breaks the natural linear trend of the standard error, which increases as the values of p rise from 0 to 0.50. This is unreasonable.

Figure III.3
Maximum coefficient of variation and standard error admitted according to the estimate of P (second criterion)



Source: National Institute of Statistics of Chile (INE), "Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares", *Documento de Trabajo*, Santiago, No. 13, 2020.

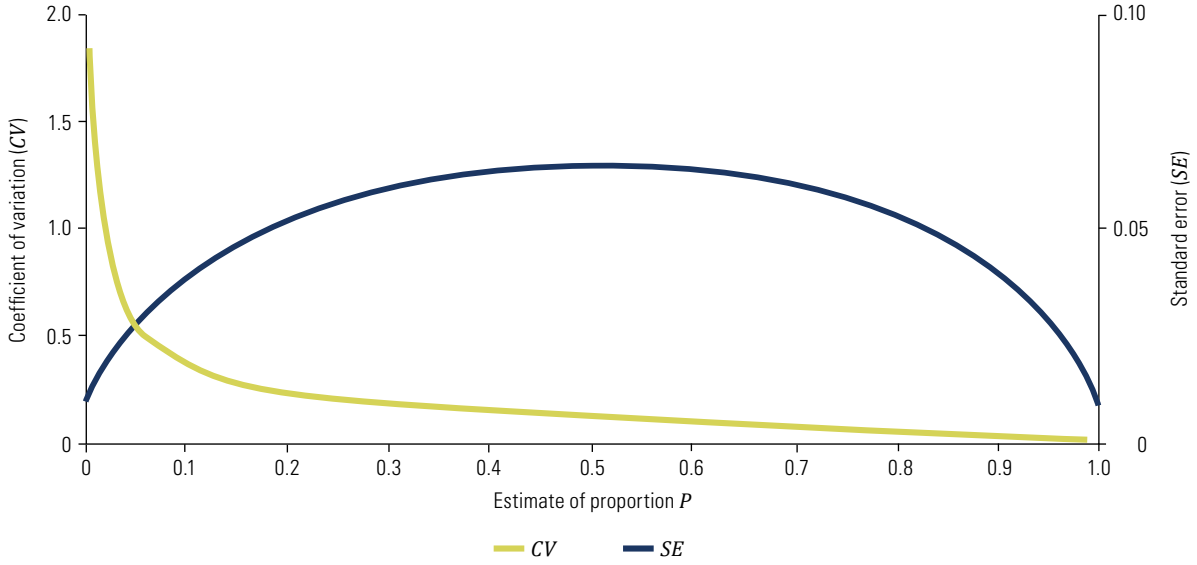
In light of the above findings, the two initial criteria were discarded and it was decided that when prevalence is low, the quality of the estimate should be measured by the standard error and not by the coefficient of variation.

According to INE (2020b), another issue to be considered in assessing accuracy is the dichotomy of the phenomena measured by proportion estimators. Behind an estimate of the proportion of victimized persons, for example, lies the estimate of non-victimized persons: both estimates are associated with the same standard error but not with the same CV, which will be significantly higher for the phenomenon with lower prevalence. Figure III.4 shows the behaviour of the standard error and coefficient of variation for various levels of p with an assumed sample size of 60 units. Note that, for p = 0.05, the standard error is 0.028 and is equal to a coefficient of variation of 0.567; the complement $1\text{-}p=0.95$, however, has the same standard error but a CV of 0.030. It would therefore be illogical to conclude that the estimate of p is not of acceptable quality, but that the estimate of $1\text{-}p$ is.

Based on the above, INE (2020b) holds that the quality of estimates of proportions and ratios with a value between 0 and 1 as a function of dispersion should be evaluated by simultaneously considering the coefficient of variation, the absolute error and the dichotomy of the study phenomena, as well as the way in which the specified thresholds affect the required sample size. The following section discusses the accuracy indicators used in INE (2020a) according to the estimator type.

Gutiérrez and others (2020) state that the coefficient of variation poses a paradox when it is used to evaluate estimators of proportions, since for measurements of the same phenomenon, the coefficients of variation of the estimate of the parameter of interest and of the complement of the proportion are contradictory. Consequently, they offer the alternative of the logarithmic coefficient of variation ($CVLog$), which is discussed in more detail in the section below.

Figure III.4
Behaviour of the standard error and coefficient of variation according to the estimate of P in a simple random sample where n=60



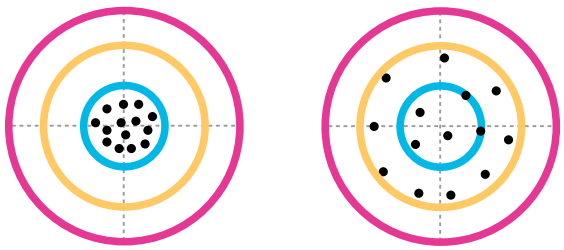
Source: National Institute of Statistics of Chile (INE), "Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares", *Documento de Trabajo*, Santiago, No. 13, 2020.

D. Accuracy measurements used for publication and dissemination

After validating the quality criteria associated with the design effect, sample size and degrees of freedom, the accuracy of the estimate must be evaluated in light of the type of estimator associated with it.

The accuracy of an estimator reports the concentration or dispersion of the estimated values around the value to be estimated (sampling error), and an estimator can be said to be accurate when, after the sampling operation is repeated several times, the magnitude of the deviations is small (see diagram III.1).

Diagram III.1
Precision of an estimator determined by the concentration or dispersion of the estimated values



Source: Economic Commission for Latin America and the Caribbean (ECLAC).

The various metrics used in the quality standards reviewed to assess the accuracy of the estimates are described in the following paragraphs.

1. Standard error

The precision of an estimator is measured mainly through the sampling variance, defined as the sum of the squared deviations from the average value of the estimate obtained using all possible samples. Since it is a squared measurement, its square root is calculated to simplify its interpretation; this is called the standard error *(se)* and is one of the measurements of sampling error.[19]

Due to the shortcomings of the coefficient of variationfor estimating the precision of estimates consisting of proportions or ratios with values of between 0 and 1, INE (2020b) proposes a new method to measure the quality of those estimates in terms of their precision. That method is based on defining standard error thresholds (as a function of p) and using those thresholds to rank estimates according to their reliability for different levels of p.

To determine the best alternative, three functions were evaluated: one linear, one logarithmic and one quadratic. The conclusion was that the quadratic function was the best for being more parsimonious (an important aspect to consider when implementing it) and for overcoming the shortcomings of the coefficient of variation and the other proposed alternatives.[20] Additionally, the quadratic function requires that the sample size be larger when the prevalence is extreme, which undoubtedly creates a margin of safety when analysing infrequent prevalences (INE, 2020b).

This quadratic function is defined as:

$$Máximoerror(ee)tolerable = \begin{cases} \frac{\sqrt[3]{p^2}}{9}; 0 < p \leq 0.5 \\ \frac{\sqrt[3]{(1-p)^2}}{9}; 0.50 < p < 1 \end{cases} \quad (33)$$

2. Coefficient of variation

The coefficient of variation (CV) of an estimate is the ratio of its standard error to the average value of the estimate itself.[21] Thus, CV provides a measurement of the sampling error relative to the characteristic being measured (United Nations, 2009).

This indicator is the measurement most commonly used by national statistical offices for the publication of official figures. Its use is cross-cutting because, by definition, it is relative in nature, which frees the user from the unit of measurement induced by the variable of interest (Gutiérrez and others, 2020). Because of this advantage, comparisons can be made between variables or indicators of different types.

In INE (2020a), except in the case of ratio or proportion estimators, CV is used to analyse the accuracy of the estimates obtained, and different thresholds that refer to the estimate's level of reliability are evaluated: when CV is less than or equal to 15%, the estimate is considered reliable; when it is greater than 15% but less than or equal to 30%, it is said to be partially reliable; and when it exceeds 30%, it is deemed unreliable.

An examination of the criteria used by different national statistical offices reveals that divergences in terms of assessment thresholds are common, and that they are observed both between surveys and between the offices of different countries. To cite a few examples, European Commission (2015) stipulates that if CV is less than or equal to 10%, the estimate is sufficiently accurate and can therefore be published without restrictions; if CV is greater than 10% but less than or equal to 30%, the estimate is less accurate and is therefore marked with the letter M; and, if CV exceeds 30%, the estimate is not accurate enough to be published and is therefore replaced by the letter N. In contrast, for the Argentine Permanent Household Survey (EPH), that country's National Institute of Statistics and Census states that if CV is less than or equal to 16%, the estimate is reliable; if it is greater than 16% but less than or equal to 25%, it is partially reliable; and when it exceeds 25%, it is unreliable.

19 The formula for standard error $SE(\hat{\theta})$ is shown in section (b).
20 For more information, see INE (2020b).
21 The formula for the coefficient of variation $CV(\hat{\theta})$ is shown in section (c).

3. Logarithmic coefficient of variation

Gutiérrez and others (2020) highlight the shortcomings of CV in the particular case of proportions. For example, when the estimate of the parameter of interest is very close to zero, regardless of how small its variance is, the coefficient of variation will be very large and will not reflect the quality of the sampling strategy. However, CV of the complement of the proportion $(1 - P)$ will be very small and reliable, which is a contradiction. As an alternative, Barnett-Walker and others (2003) propose using a logarithmic transformation on the proportion and applying what is called the logarithmic coefficient of variation ($CVLog$) to measure the accuracy of the estimated proportions.

If $P \leq 0.5$, $\hat{L} = -\log(\hat{P})$ is defined, where the first order Taylor approximation is:

$$\hat{L} \cong L + \frac{\partial \hat{L}}{\partial \hat{P}} \Big/_{\hat{P}=P} (\hat{P} - P) = L + \left(\frac{-1}{P}\right)(\hat{P} - P) \tag{34}$$

Then, the variance of $\hat{L}$ will be $Var(\hat{L}) \cong AV(\hat{L}) = \frac{Var(\hat{P})}{P^2}$ and, consequently, the standard error of the transformation will equal the coefficient of variation of the proportion, where:

$$SE(\hat{L}) = \sqrt{AVar(\hat{L})} = \frac{\sqrt{Var(\hat{P})}}{\hat{P}} = CV(\hat{P}) \tag{35}$$

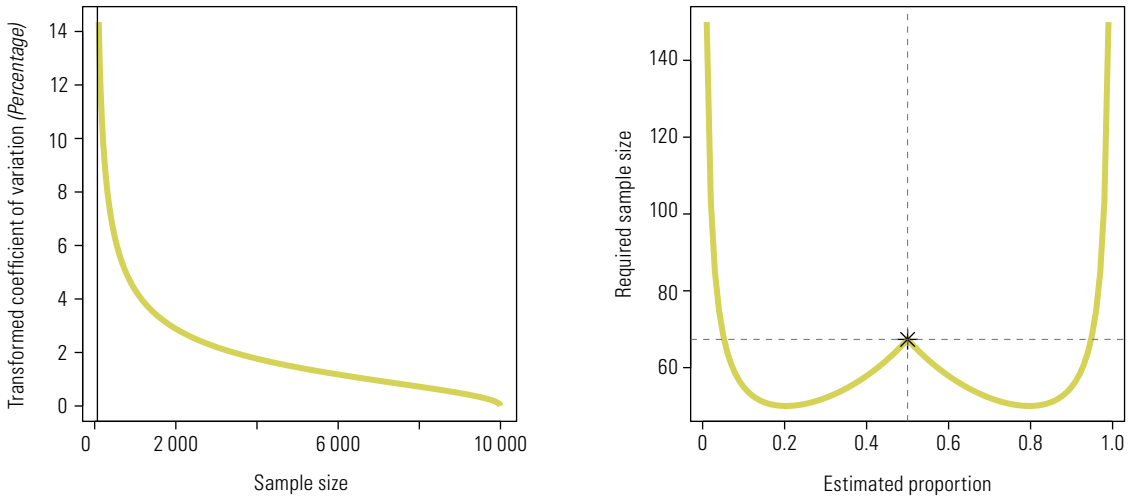
Thus, a smoothing measurement can be defined as the coefficient of variation associated with the transformation:

$$CV(\hat{L}) = S\frac{E(\hat{L})}{\hat{L}} = \frac{CV(\hat{P})}{\hat{L}} \tag{36}$$

Similarly, to maintain symmetry and maintain generality, when $P>0.5$ an adjustment is made by defining $\hat{L} = -\log(1 - \hat{P})$ Consequently, when the proportions lie within the central interval, specifically between 0.2 and 0.8, the coefficients of variation for the estimates, represented as P^P^ and L^L^, show a remarkable similarity. The underlying reason is that L^L^ tends to approach values close to one in this spectrum of proportions. As a result of this behaviour, the coefficient of variation of L^L^, or CV(L^)CV(L^), has a close correspondence with the coefficient of variation of P^P^, denoted as CV(P^)CV(P^).

Figure III.5 shows that, as with the original coefficient of variation, the sample size will increase as greater precision is required in the estimate; unlike the original coefficient of variation, however, the sample size will be identical for phenomena that entail symmetrical proportions.

Figure III.5
Relationship between sample size and the precision of an indicator using the logit transform



Source: A. Gutiérrez and others, "Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional", *Statistical Studies series*, No. 101 (LC/TS.2020/52), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC), 2020.

Finally, according to the value of $\widehat{P}$, the formula for evaluating the logarithmic coefficient of variation ($CVLog$) would be the following:

$$CVlog = \begin{cases} \dfrac{CV(p)}{-log(p)}, & p < 0.5 \\ \dfrac{CV(p)}{-log(1-p)}, & p \geq 0.5 \end{cases} \tag{37}$$

In addition, Gutiérrez and others (2020) suggest using the decision threshold identified by Barnett-Walker and others (2003), under which the figure is omitted if $CVLog$ exceeds 17.5%.

E. Unweighted case count

The unweighted case count refers to the number of elements in the sample that exhibit the characteristic of interest. According to Gutiérrez and others (2020), this indicator has a direct effect on determining the precision of the estimator of interest and is particularly important in the case of ratio and proportion estimators. The formula is defined as:

$$n = \sum_{s} \delta_k^y \tag{38}$$

where δ_k^y is an indicator variable for each individual k in the sample s, which takes the value 1 if the individual is affected by the phenomenon that the variable of interest y represents. Note that by definition this is a random quantity and that it can also be calculated in the sample of a specific population subgroup U_g as follows:

$$n_y^g = \sum_{s} z_{gk} \delta_k^y = \sum_{s_g} \delta_k^y \tag{39}$$

If the incidence of the phenomenon is very low (when the proportion P is close to zero), both the original coefficient of variation and its logarithmic transformation will have high values, since:

$$\lim_{n_y \to 0} CV(\hat{\theta}) = \lim_{n_y \to 0} CV(\hat{L}) = \infty \tag{40}$$

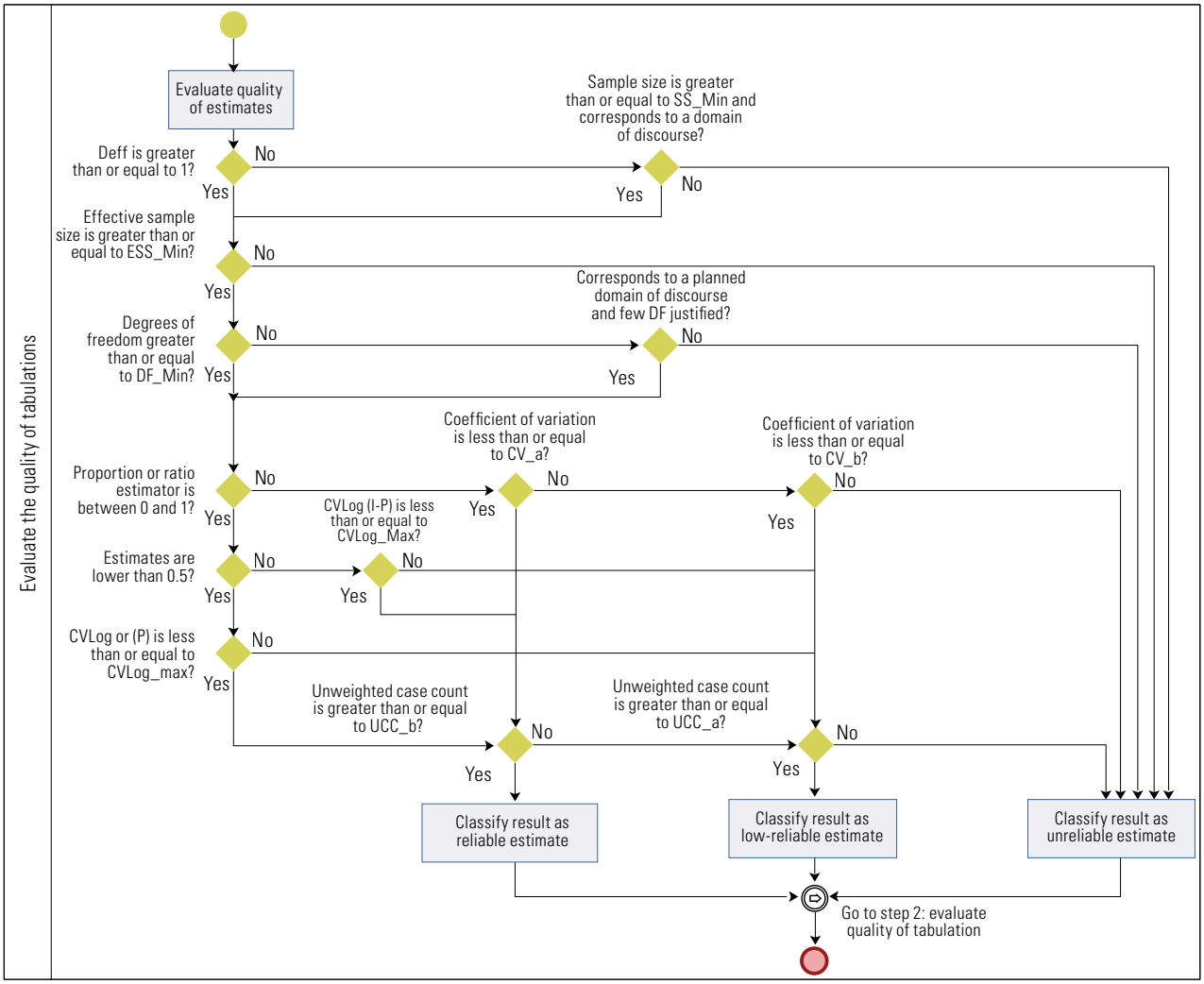
Chapter IV

Quality requirements applicable to estimates

This section aims to formulate a harmonized quality standard that proposes decision thresholds for each quality criterion, together with guidelines to orient users on how the estimates obtained can be used.

Diagram IV.1 provides a flow chart indicating the path to be followed in assessing the quality of the estimates obtained through household sample surveys as proposed by the working group in the harmonized proposal.

Diagram IV.1
Flow chart for evaluation of estimates



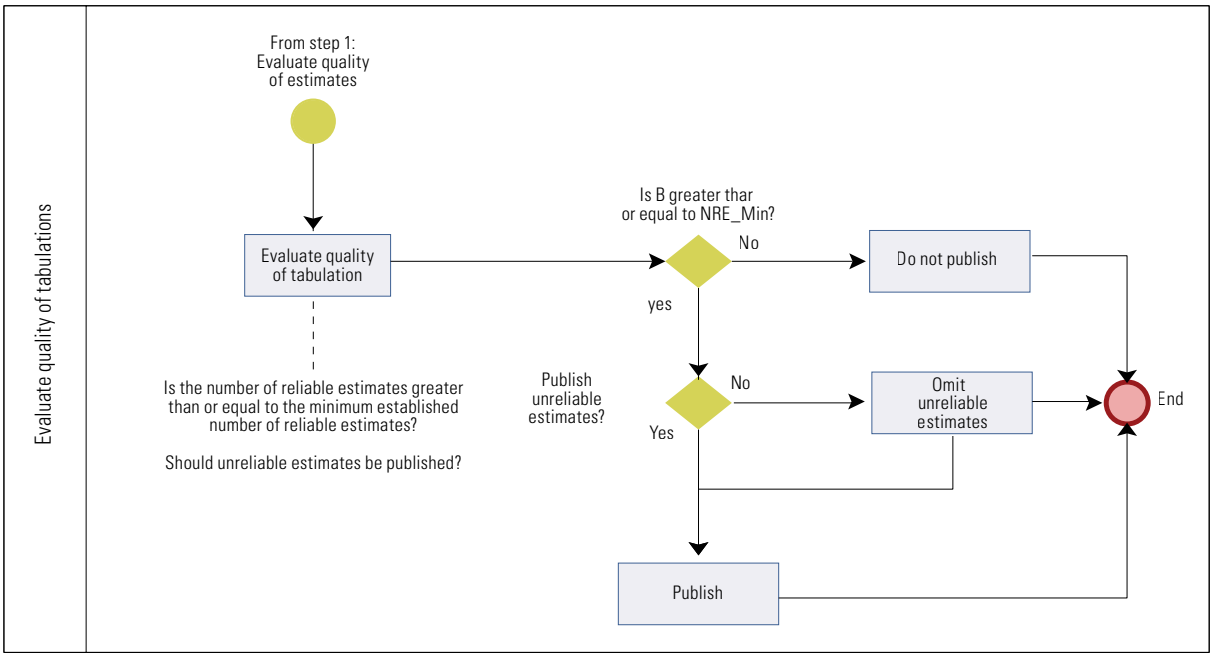
Source: Economic Commission for Latin America and the Caribbean (ECLAC).

A guide to the flow chart shown in diagram IV.1 is presented below, indicating the acceptance ranges for each criterion and the arguments on which they are based:

- Design effect (*Deff*): *Deff* must be calculated for the variable of interest at the disaggregated level where the estimate is found. An acceptance threshold equal to or greater than 1 is suggested. While a *Deff* of less than 1 is uncommon, when this occurs, it is recommended to review the *Deff* estimation mechanism in the software used and to analyse the homogeneity of the phenomenon in the study domain. If *Deff* is less than 1, the sample size is not greater than the threshold established in the following paragraph and the number does not correspond to a study domain, so the result should be classified as an unreliable estimate. When *Deff* is greater than 1, the flow proceeds with the effective sample size as described below.
- Sample size: this element is crucial in determining the publication route of statistical data, as it is the basis for theoretical advances in statistical inference for surveys. Both the precision of confidence intervals and the distribution of estimators are contingent on the adequate size of both the subpopulation and the corresponding sample. On that point, Barnett-Walker and others (2003) suggest that any estimate based on a sample size smaller than 100 units should be eliminated or flagged as not reliable.
- Effective sample size: an acceptance threshold of between 60 and 100 elements is suggested, on the basis of the standard acceptance thresholds of the National Statistics Institute of Chile and ECLAC. This threshold guarantees convergence in the distribution of the estimators, which makes inferences about the population possible. If this threshold is not reached, the estimate should be classified as unreliable; if it is reached, it is suggested to move on to the evaluation of the degrees of freedom.
- Degrees of freedom: as simulations show that the convergence in ratio estimators can be empirically demonstrated from 13 degrees of freedom onwards, and considering the recommendation of 8 degrees made by Parker and others (2017), the minimum values suggested for the degrees of freedom for the estimates are between 9 and 13. If there are insufficient degrees of freedom in the estimate, it is recommended to classify the estimate as unreliable, except when it corresponds to study domains where few degrees of freedom can be justified. In order to calculate the degrees of freedom, the variable degrees of freedom must be calculated: i.e. the number of primary sampling units with observations in the subpopulation minus the number of strata with observations in the subpopulation. Following the evaluation of this element, the next step is the coefficient of variation or the logarithmic coefficient of variation.
- Coefficient of variation: this element is used to evaluate estimates that do not correspond to proportions or ratios with values of between 0 and 1. In this regard, most national statistics offices consider estimates with a coefficient of variation of more than 20% to be "unpublishable" data (Molina, 2022). An acceptance value for reliability of between 15 and 20 (identified on the flow chart as coefficient of variation "a") is recommended. A threshold between 25 and 30 is recommended for estimates to be classified as partially reliable, and when the coefficient of variation is greater than 30 (identified on the flow chart as coefficient of variation "b"), classification of the estimate as unreliable is recommended.
- Logarithmic coefficient of variation: this element is used to evaluate estimates that correspond to proportions or ratios with values of between 0 and 1. As proposed by Barnett-Walker and others (2003), a threshold of 17.5% is recommended for an estimate to be considered reliable when the effective sample size is 60; if the figure exceeds this value, it is recommended that it be considered partially reliable or unreliable in accordance with the thresholds established in the point below this one. This acceptance threshold can be modified on the basis of the value of the logarithmic function of the coefficient of variation, measured as p = 0.5, depending on the values of the effective sample size.
- Unweighted case count: the unweighted case count corresponds to the number of units within the sample affected by the phenomenon of interest. If the unweighted case count is less than 30 (identified on the flow chart as unweighted case count "a"), it is recommended to classify the estimate as unreliable. If the count is greater than 30 but less than or equal to 40 (or 50, depending on the institution), a partially reliable classification is recommended. When the unweighted case count is greater than 50 (identified on the flow chart as unweighted case count "b"), it is recommended that the estimate be classified as reliable.

Diagram IV.2 provides a flow chart for evaluating tabulation quality, which is the next step after the evaluation of the estimate quality.

Diagram IV.2
Flow chart for evaluation of tabulations



Source: Economic Commission for Latin America and the Caribbean (ECLAC).

The flow chart in diagram IV.2 recommends not publishing the tabulation if the number of reliable estimates is lower than a minimum established threshold. To determine that threshold, it is suggested that the number of reliable estimates be greater than or equal to 50% plus one of the total number of estimates shown in the tabulation: i.e. A/2 + 1, where A is the number of estimates in the tabulation. If the decision to publish the tabulation is made, an assessment will be necessary of whether the unreliable estimates should be included in the tabulations or omitted.

The proposed workflows for estimates and tabulations provide the region's national statistical offices with a template for safeguarding the quality of their statistical products. By taking into consideration the thresholds and elements linked to estimate reliability, the actions needed in the design, collection, processing and data analysis stages can be defined and applied. In keeping with Gutiérrez and others (2020), in estimates where there is evidence of coverage difficulties, it is recommended that bias analysis be carried out and that expansion factors be adjusted to mitigate the possible effects.

Chapter V

Concluding observations

In the context of process-based statistical production, the frame of reference for which is the Generic Statistical Business Process Model of the Economic Commission for Europe, the guidelines contained in this document are aimed at the results analysis phase, specifically at the stages involving preparation, finalization of results and application of quality protocols. The results analysis phase is dedicated to the creation of the content of the reports to be published for each statistical operation, such as a survey, and it ensures that the statistics and information to be included meet quality criteria and are available for society's use, analysis and interpretation.

To ensure survey quality, however, quality standards must be observed throughout the production process, from design to publication. Survey quality depends largely on the design, which must be properly structured and in line with the objectives sought. The variables of interest, target population and sampling frame must be clearly defined.

In addition, particular care in the data collection and processing phases is crucial. Robust procedures are essential at these stages: adequate training of field personnel, quality control mechanisms —including the monitoring of indicators— over the course of the data collection process, and rigorous procedures for verifying and cleaning the data collected.

In the analysis process, the reliability of the estimates —which depends on the quality of the data collected and processed— is another vital aspect that cannot be overlooked. Detailed and rigorous data analyses are therefore essential to ensure estimate reliability, and this document identifies specific elements that should be measured and evaluated.

Continuous review and improvement are fundamental aspects of this process, and each version of a survey must evaluate whether the results meet the objectives set and, if necessary, a road map of improvements to be considered in the next production cycle must be established.

Finally, in addition to these recommendations, it must be remembered that transparency and effective communication are essential to ensure survey quality. Transparency and communication involve sharing survey results in a timely manner and in an easy-to-understand format and entail willingness to answer questions and concerns from participants and other stakeholders.

Bibliography

ANDA (National Data Archive) (2023), "Catálogo de Metadatos y Microdatos" [online database] https://anda.ine.gob.bo/index.php/home.

___(2020), "Política de difusión" [online] https://anda.ine.gob.bo/index.php/politicas-difusion.

Barnett-Walker, K. and others (2003), *2001 National Household Survey on Drug Abuse: Statistical Inference Report*, Rockville, Substance Abuse and Mental Health Services Administration (SAMHSA).

Chambers, R. and C. Skinner (eds.) (2003), *Analysis of Survey Data*, Hoboken, Wiley.

Cochran, W. (1980), *Técnicas de muestreo*, Mexico City, Compañía Editorial Continental (CECSA).

___(1977), *Sampling Techniques*, Hoboken, Wiley.

CRAN (Comprehensive R Archive Network) (2023), "Calidad: assesses the quality of estimates made by complex sample designs", Vienna [online] https://cran.r-project.org/web/packages/calidad/index.html.

DIGESTYC (Department of Statistics and Censuses) (2022), "Recomendaciones sobre criterios de supresión para investigaciones por muestreo probabilístico", San Salvador.

ECLAC (Economic Commission for Latin America and the Caribbean) (2022), *Guide for the implementation of a quality assurance framework for statistical processes and outputs* (LC/CEA.11/19), Santiago.

___(2011), *Code of Good Practice in Statistics for Latin America and the Caribbean*, Santiago.

European Commission (2015), *ESS Handbook for Quality Reports*, Brussels.

___(2013), *Handbook on Precision Requirements and Variance Estimation for ESS Households Surveys: 2013 Edition*, Brussels.

Groves, R. and others (2004), *Survey Methodology*, Hoboken, Wiley.

Gutiérrez, A. (2022), *Diseño y análisis estadístico en las encuestas de hogares de América Latina* (LC/PUB.2023/14-P), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC).

Gutiérrez, A. and others (2020), "Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional", *Statistical Studies series*, No. 101 (LC/TS.2020/52), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC).

Hansen, M., W. Hurwitz and W. Madow (1953), *Sample Survey Methods and Theory*, New York, Wiley.

Heeringa, S., B. West and P. Berglund (2010), *Applied Survey Data Analysis*, Boca Ratón, CRC Press.

Hornik, R. and others (2002), *Evaluation of the National Youth Anti-Drug Media Campaign: Fifth Semi-Annual Report of Findings*, Rockville, Westat.

IBGE (Brazilian Institute of Geography and Statistics) (2021a), *Guia para Divulgação de Erros Amostrais nas Pesquisas por Amostragem Probabilística Realizadas pelo IBGE*, Rio de Janeiro.

___(2021b), *Código de Boas Práticas das Estatísticas do IBGE*, Rio de Janeiro.

INDEC (National Institute of Statistics and Censuses of Argentina) (n.d.), "Notas técnicas" [online] https://www.indec.gob.ar/indec/web/Institucional-Indec-Metodologias-2.

INE (National Institute of Statistics of Chile) (2020a), *Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares*, Santiago.

___(2020b), "Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares", *Documento de Trabajo*, No. 13, Santiago.

INE (National Institute of Statistics of Uruguay) (2021), *Norma Técnica de Certificación de la Calidad de Operaciones Estadísticas INE-CCOE: 2021-01*, Montevideo.

Judkins, D. (1990), "Fay's method for variance estimation", *Journal of Official Statistics*, vol. 6, No. 3, Stockholm, Statistics Sweden.

Kalton, G., J. Brick and T. Lê (2005), "Estimating components of design effects for use in sample design", *Household Sample Surveys in Developing and Transition Countries*, series F, No. 96 (ST/ESA/STAT/SER.F/96), New York, United Nations.

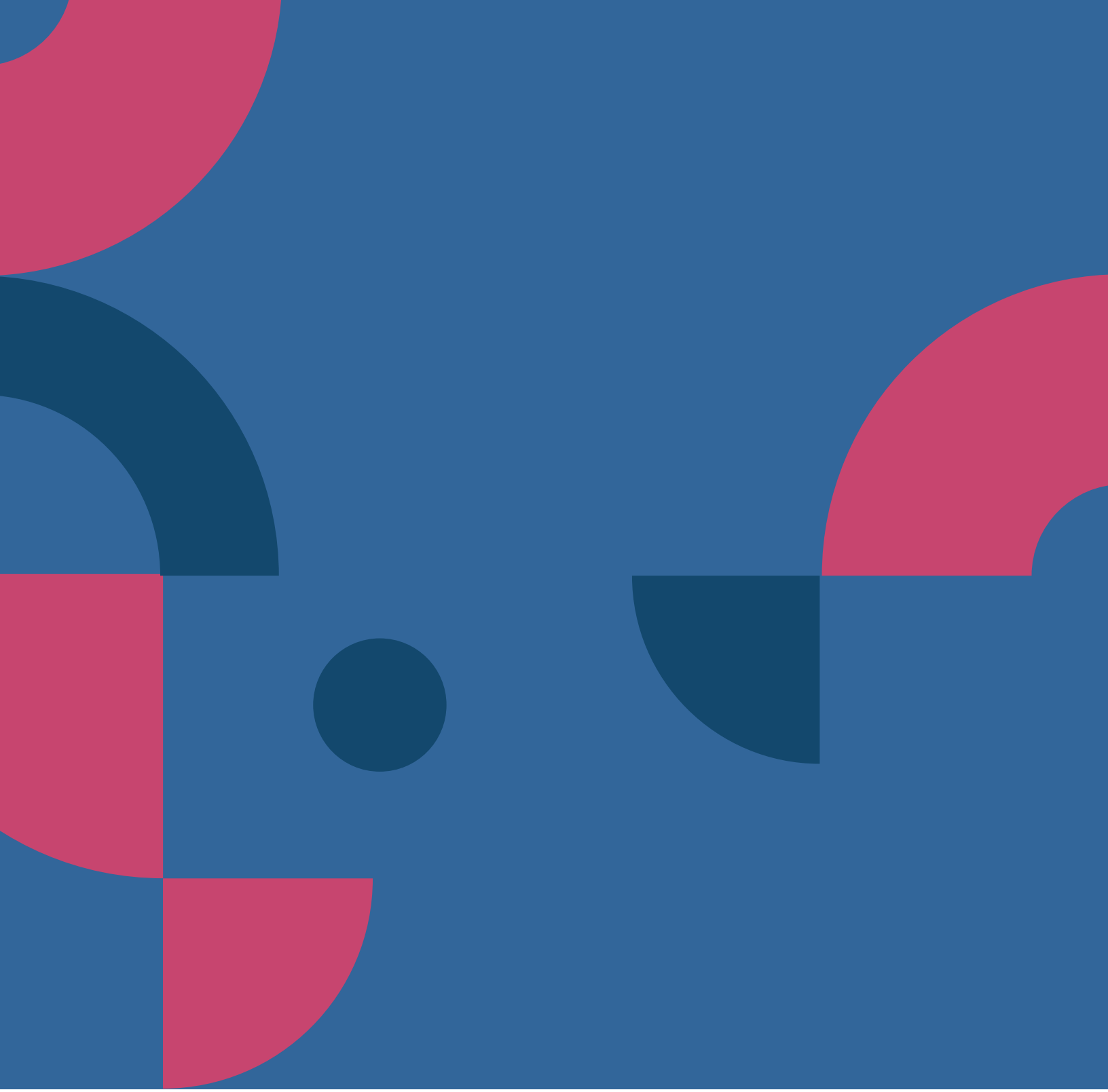
Kish, L. (1965), *Survey Sampling*, New York, Wiley.

Klein, R. and others (2002), "Healthy People 2010 criteria for data suppression", *Healthy People 2010 Statistical Notes*, No. 24, Washington, D.C., American Psychological Association (APA).

Korn, E. and B. Graubard (1999), *Analysis of Health Surveys*, Hoboken, Wiley.

Krewski, D. and J. Rao (1981), "Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods", *The Annals of Statistics*, vol. 9, No. 5, Durham, Institute of Mathematical Statistics (IMS).

Lavrakas, P. (2008), "Probability proportional to size (PPS) sampling", *Encyclopedia of Survey Research Methods*, Thousand Oaks, SAGE Publications.
Lohr, S. (1999), *Sampling: Design and Analysis*, Duxbury Press.
Lumley, T. (2010), *Complex Surveys: A Guide to Analysis Using R*, Hoboken, Wiley.
McCarthy, P. (1969), "Pseudo-replication: half samples", *Review of the International Statistical Institute*, vol. 37, No. 3, The Hague, International Statistical Institute (ISI).
Martínez, C. (2019), *Estadística y muestreo*, Bogotá, ECOE Ediciones.
Mirás, J. (1985), *Elementos de muestreo para poblaciones finitas*, Madrid, National Institute of Statistics.
Molina, I. (2022), "Disaggregating data in household surveys: using small area estimation methodologies", *Statistical Studies series*, No. 97 (LC/TS.2018/82), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC).
MTESS/OISS/SRT (Ministry of Labour, Employment and Social Security/Ibero-American Social Security Organization/Office of the Superintendent of Occupational Risks) (2018), *Encuesta Nacional a Trabajadores sobre Condiciones de Empleo, Trabajo, Salud y Seguridad (ECETSS) 2018: guía de referencia para estimaciones de errores de muestreo por medio de bootstrap*, Buenos Aires.
Parker, J. and others (2017), "National Center for Health Statistics data presentation standards for proportions: data evaluation and methods research", *Vital and Health Statistics*, vol. 2, No. 175, Hyattsville, National Center for Health Statistics (NCHS).
Pennell, B. and others (2017), "A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts", *Total Survey Error in Practice*, P. Biemer and others (eds.), Hoboken, Wiley.
Pérez, C. (2005), *Muestreo estadístico: conceptos y problemas resueltos*, Madrid, Prentice Hall.
Rao, J. and C. Wu (1984), "Bootstrap inference for sample surveys", *Proceedings of the American Statistical Association Survey Research Methods Section*, Alexandria, American Statistical Association (ASA).
Särndal, C., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Berlin, Springer.
Statistics Canada (2003), *Statistics Canada Quality Guidelines Fourth Edition – October 2003*, Ottawa.
Tillé, Y. and D. Haziza (2010), "An interesting property of the entropy of some sampling designs", *Survey Methodology*, vol. 36, No. 2, Ottawa, Statistics Canada.
Tillé, Y. and M. Wilhelm (2017), "Probability sampling designs: principles for choice of design and balancing", *Statistical Science*, vol. 32, No. 2, Durham, Institute of Mathematical Statistics (IMS).
United Nations (2009), "Designing household survey samples: practical guidelines", *Studies in Methods*, series F, No. 98 (ST/ESA/STAT/SER.F/98), New York.
Valliant, R., J. Dever and F. Kreuter (2018), *Practical Tools for Designing and Weighting Survey Samples*, Berlin, Springer.
Walker, H. (1940), "Degrees of freedom", *Journal of Educational Psychology*, vol. 31, No. 4, Washington, D.C., American Psychological Association (APA).



Economic Commission for Latin America and the Caribbean (ECLAC)
Comisión Económica para América Latina y el Caribe (CEPAL)
www.cepal.org/en

LC/CEA.12/11