Data Quality, Sampling and Weighting, Confidentiality and Random Rounding

Data Quality

General

The 1996 Census was a large and complex undertaking and, while considerable effort was taken to ensure high standards throughout all collection and processing operations, the resulting estimates are inevitably subject to a certain degree of error. Users of census data should be aware such error exists, and have some appreciation of its main components, so that they can assess the usefulness of census data for their purposes and the risks involved in basing conclusions or decisions on these data.

Errors can arise at virtually every stage of the census process, from the preparation of materials through the listing of dwellings, data collection and processing. Some errors occur more or less at random, and when the individual responses are aggregated for a sufficiently large group, such errors tend to cancel out. For errors of this nature, the larger the group, the more accurate the corresponding estimate. It is for this reason that users are advised to be cautious when using small estimates. There are some errors, however, which might occur more systematically, and which result in “biased” estimates. Because the bias from such errors is persistent no matter how large the group for which responses are aggregated, and because bias is particularly difficult to measure, systematic errors are a more serious problem for most data users than the random errors referred to previously.

For census data in general, the principal types of error are as follows:

- **coverage errors**, which occur when dwellings and/or individuals are missed, incorrectly included or double counted;
- **non-response errors**, which result when responses cannot be obtained from a small number of households and/or individuals, because of extended absence or some other reason;
- **response errors**, which occur when the respondent, or sometimes the Census Representative, misunderstands a census question, and records an incorrect response;
- **processing errors**, which can occur at various steps including **coding**, when “write-in” responses are transformed into numerical codes; **data capture**, when responses are transferred from the census questionnaire to computer tapes by key-entry operators; and **imputation,** when a “valid”, but not necessarily correct, response is inserted into a record by the computer to replace missing or “invalid” data (“valid” and “invalid” referring to whether or not the response is consistent with other information on the record);
- **sampling errors**, which apply only to the supplementary questions on the “long form” asked of a one-fifth sample of households, and which arise from the fact that the results for these questions, when weighted up to represent the whole population, inevitably differ somewhat from the results which would have been obtained if these questions had been asked of all households.

The above types of error each have both random and systematic components. Usually, however, the systematic component of sampling error is very small in relation to its random component. For the other non-sampling errors, both random and systematic components may be significant.

Coverage Errors

Coverage errors affect the accuracy of the census counts, that is the sizes of the various census universes: population, families, households and dwellings. While steps have been taken to correct certain identifiable errors, the final counts are still subject to some degree of error resulting from persons or dwellings being missed, incorrectly included in the census or double counted.

Missed dwellings or persons result in **undercoverage**. Dwellings can be missed because of the misunderstanding of enumeration area (EA) boundaries, or because they are not apparent (e.g., unmarked dwellings) or appear uninhabitable. Persons can be missed when their dwelling is missed or is classified as vacant, or when individual household members are omitted from the questionnaire because the respondent misinterprets the instructions on whom to include. Some individuals may be missed because they have no usual residence and did not spend census night in any dwelling.

Dwellings or persons that are incorrectly included or double counted result in **overcoverage**. Overcoverage of dwellings can occur when structures unfit for habitation are listed as dwellings, or when units which do not meet the census definition of a dwelling are listed separately instead of being treated as part of a larger dwelling. Double counting of dwellings also can occur because of ambiguity over EA boundaries. Persons can be double counted because their dwelling is double counted or because the guidelines on whom to include on the questionnaire have been misunderstood. Occasionally, someone who is not in the census population universe, such as a foreign resident or a fictitious person, may, incorrectly, be enumerated in the census. On average, overcoverage is less likely to occur than undercoverage and, as a result, counts of dwellings and persons are likely to be slightly underestimated.

In 1996, three studies are used to measure coverage error. In the Vacancy Check, a sample of dwellings listed as vacant was revisited to verify that they were vacant on Census Day. Adjustments have been made to the final census counts for households and persons missed because their dwelling was incorrectly classified as vacant. Despite these adjustments, the final counts are still subject to some undercoverage. Undercoverage tends to be higher for certain segments of the population, such as young male adults and recent immigrants. The Reverse Record Check study is used to measure the residual undercoverage for Canada, and each province and territory. The Overcoverage Study is designed to investigate overcoverage errors. The results of the Reverse Record Check and the Overcoverage Study, when taken together, furnish an estimate of net undercoverage.

Other Non-sampling Errors

While coverage errors affect the number of units in the various census universes, other errors affect the characteristics of those units.

Sometimes, it is not possible to obtain a complete response from a household, even though the dwelling was identified as occupied and a questionnaire was dropped off. The household members may have been away throughout the census period or, in rare instances, the householder may have refused to complete the form. More frequently, the questionnaire is returned but information is missing for some questions or individuals. Considerable effort is devoted to ensure as complete a response as possible. Census representatives edit the questionnaires and follow up on missing information. Their work is then checked by both a supervisor and a quality control technician. Despite this, at the end of the collection stage, a small number of responses is still missing. Although missing entries are eliminated during processing by replacing a missing value by the corresponding entry for a “similar” record, there remain some potential **non-response errors**. This is particularly serious if the non-respondents differ in some respects from the respondents, since this procedure will result in **non-response bias**.

Even when a response is obtained, it may not be entirely accurate. The respondent may have misinterpreted the question or may have guessed the answer, especially when answering on behalf of another, possibly absent, household member. Such errors are referred to as **response errors**. While response errors usually arise from inaccurate information provided by respondents, they can also result from mistakes by the Census Representative when completing certain parts of the questionnaire, such as structural type of dwelling, or when calling back to obtain a missing response.

Some of the questions on the census document require a written response. During processing, these “write-in” entries are given a numeric code. **Coding errors** can occur when the written response is ambiguous, incomplete, difficult to read or when the code list is extensive (e.g., major field of study and place of work). A formal Quality Control (QC) operation is used to detect, rectify and reduce coding errors. Within each work unit, a sample of responses is independently coded a second time. The resolution of discrepancies between the first and second codings determines whether recoding of the work unit is necessary. Except for the Industry and Occupation variables, much of the census coding is now automated, partly in an effort to reduce the extent of coding errors.

The information on the questionnaires is key-entered onto a computer file. Two procedures are used to control the number of **data capture errors**. First, certain edits (such as range checks) are performed as the data are keyed. Second, a sample from each batch of documents is rekeyed and compared with the original entries. Unsatisfactory work is identified and corrected and the remainder of the batch is rekeyed as needed.

Once captured, the data are edited where they undergo a series of computer checks to identify missing or inconsistent responses. These are replaced during the imputation stage of processing where either a response consistent with the other respondent’s data is inferred or a response from a similar donor is substituted. Imputation ensures a complete database where the data correspond to the census counts and facilitate multivariate analyses. Although imputation may introduce errors, the methods used have been rigorously tested to minimize systematic **imputation errors**.

Various studies are being carried out to evaluate the quality of the responses obtained in the 1996 Census. For each question, response rates and edit failure rates have been calculated. These can be useful in identifying the potential for non-response errors and other type of errors. Also, tabulations from the 1996 Census have been or will be compared with corresponding estimates from previous censuses, from sample surveys (such as the Labour Force Survey) and from various administrative records (such as birth registrations and municipal assessment records). Such comparisons can indicate potential quality problems or at least discrepancies between the sources.

In addition to these aggregate-level comparisons, there are some micro-match studies in progress, in which census responses are compared with another source of information at the individual record level. For certain “stable” characteristics (such as age, sex, mother tongue and place of birth), the responses obtained in the 1996 Census, for a sample of individuals, are being compared with those for the same individuals in the 1991 Census.

Sampling Errors

Estimates obtained by weighting up responses collected on a sample basis are subject to error due to the fact that the distribution of characteristics within the sample will not usually be identical to the distribution of characteristics within the population from which the sample has been selected.

The potential error introduced by sampling will vary according to the relative scarcity of the characteristics in the population. For large cell values, the potential error due to sampling, as a proportion of the cell value, will be relatively small. For small cell values, this potential error, as a proportion of the cell value, will be relatively large.

The potential error due to sampling is usually expressed in terms of the so-called “standard error”. This is the square root of the average, taken over all possible samples of the same size and design, of the squared deviation of the sample estimate from the value for the total population.

The following table provides approximate measures of the standard error due to sampling. These measures are intended as a general guide only.

Table: Approximate Standard Error Due to Sampling for 1996 Census Sample Data

Cell Value	Approximate Standard Error
50 or less	15
100	20
200	30
500	45
1,000	65
2,000	90
5,000	140
10,000	200
20,000	280
50,000	450
100,000	630
500,000	1,400

Users wishing to determine the approximate error due to sampling for any given cell of data, based upon the 20% sample, should choose the standard error value corresponding to the cell value that is closest to the value of the given cell in the census tabulation. When using the obtained standard error value, in general the user can be reasonably certain that, for the enumerated population, the true value (discounting all forms of error other than sampling) lies within plus or minus three times the standard error (e.g., for a cell value of 1,000, the range would be 1,000 ± (3 x 65) or 1,000 ± 195).

The standard errors given in the table above will not apply to population or universe (persons, households, dwellings or families) totals or subtotals for the geographic area under consideration (see Sampling and Weighting).

The effect of the particular sample design and weighting procedure used in the 1996 Census will vary, however, from one characteristic to another and from one geographic area to another. The standard error values in the table may, therefore, understate or overstate the error due to sampling.

Sampling and Weighting

The 1996 Census data were collected either from 100% of the population or on a sample basis (i.e. from a random sample of one in five households) with the data weighted up to provide estimates for the entire population. Some of the information in this report was collected on a 20% sample basis and weighted up to compensate for sampling. All table headings are noted accordingly. Note that, on Indian reserves and in remote areas, all data were collected on a 100% basis.

For any given geographic area, the weighted population, household, dwelling or family total or subtotal may differ from that shown in reports containing data collected on a 100% basis. Such variation (in addition to the effect of random rounding) will be due to sampling.

Confidentiality and Random Rounding

The figures shown in the tables have been subjected to a confidentiality procedure known as “**random rounding**” to prevent the possibility of associating statistical data with any identifiable individual. Under this method, all figures, including totals and subtotals are randomly rounded. For 100% data, all counts are rounded to a multiple of “5”. This means that all 100% data will end in either “0” or “5”. For the 20% sample data, all counts greater than “10” are rounded to a multiple of “5”. Counts less than “10” are rounded to either the value “0” or “10”. While providing strong protection against disclosure, this technique does not add significant error to the census data. The user should be aware that totals and margins are rounded independently of the cell data so that some difference between these and the sum of rounded cell data may exist. Also, minor differences can be expected in corresponding totals and cell values among various census tabulations. Similarly, percentages, which are calculated on rounded figures, do not necessarily add up to 100%. Percentage distributions and rates for the most part are based on rounded data, while percentage changes and averages are based on unrounded data. It should also be noted that small cell counts may suffer a significant distortion as a result of random rounding. Individual data cells containing small numbers may lose their precision as a result.

Users should be aware of possible data distortions when they are aggregating these rounded data. Imprecisions as a result of rounding tend to cancel each other out when data cells are re-aggregated. However, users can minimize these distortions by using, whenever possible, the appropriate subtotals when aggregating.

For those requiring maximum precision, the option exists to use custom tabulations. With custom products, aggregation is done using individual census database records. Random rounding occurs only after the data cells have been aggregated, thus minimizing any distortion.

In addition to random rounding, other methods, such as **area suppression** and the suppression of income statistics, have been adopted to further protect the confidentiality of individual responses.

Area suppression is the deletion of all characteristic data for geographic areas with populations below a specified size. The extent to which data are suppressed depends upon the following factors:

- If the data are tabulated from the 100% database, the data are suppressed if the total area population in the area is less than 40.
- If the data are tabulated from the 20% sample database, the data are suppressed if the total non-institutional population in the area from either the 100% or 20% databases is less than 40.

There are some exceptions to these rules:

- Income distributions and related statistics are suppressed if the non-institutional population in the area from either the 100% or 20% databases is less than 250.

For place of work, suppression is required where the labour force working in an area is less than 40. For place of work tables containing both residence and work locations, both standard suppression rules and location of work suppression rules are applied.

In all cases, suppressed data are included in the appropriate higher aggregate subtotals or totals. The suppression technique is being implemented for all products involving subprovincial data (i.e. Profile series, Basic Summary Tabulations, semi-custom and custom data products) collected on a 100% or 20% sample basis.

With cell suppression, the minimum acceptable value for a cell is specified. All cell values below the designated cut-off are deleted and replaced by a dash. However, the suppressed data are included in the appropriate higher aggregate subtotals and totals.

As part of the income statistics suppression, the statistics of income components within cells where population is less than 10 persons are suppressed. The suppression is based on the unrounded number of persons; therefore, it is possible to see, within the cell, a total number of 10 persons for which all statistics on income are suppressed.

For further information on the quality of census data, contact the Social Survey Methods Division at Statistics Canada, Ottawa, Ontario, Canada K1A 0T6, or by dialing (613) 951-6934.