Microdata User Guide

For the Public Access Microdata File

Canadian Legal Problems Survey (CLPS)

2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to:

Statistics Canada
Client Services
Centre for Social Data Integration and Development
Telephone: 613-951-3321 or call toll-free 1-800-461-9050
E-mail: csdid-info-cidds@statcan.gc.ca

Accessing the microdata and ordering information

The 2021 Canadian Legal Problems Survey (CLPS) produces two types of microdata files: master file and public use microdata file (PUMF).

Master file

The master file contains all variables and all in-scope records from the survey collected during a collection period. The file is accessible at Statistics Canada for internal use and in Statistics Canada's Research Data Centres (RDC), and are also subject to custom tabulation requests.

Research Data Centre

The RDC Program enables researchers to use the survey data in the master files in a secure environment in several universities across Canada. Researchers must submit research proposals that, once approved, give them access to the RDC. For more information, please consult the following web page: https://www.statcan.gc.ca/eng/microdata/data-centres

Custom tabulations

Another way to access the master file data is to offer all users the option of requesting the staff in Statistics Canada Centre for Social Data Integration and Development Client Services to prepare custom tabulations. This cost-recovery product allows users who do not possess knowledge of tabulation software products to get custom results. The results are screened for confidentiality and reliability concerns before release. For more information, please contact Client Services.

Public use microdata file

The public use microdata file (PUMF) is developed from the master file using a technique that balances the need to ensure respondent confidentiality with the need to produce the most useful data possible. The PUMF must meet stringent security and confidentiality standards required by the *Statistics Act* before they are released for public access. To ensure that these standards have been achieved, each PUMF goes through a formal review and approval process by an executive committee of Statistics Canada. Variables most likely to lead to identification of an individual are deleted from the data file or are collapsed to broader categories.

To obtain a copy of the PUMF, contact Infostats or consult the Statistics Canada website.

The Data Liberation Initiative

The Data Liberation Initiative (DLI) Program enables students and researchers to use the public use microdata files in several universities across Canada. For more information, please consult the following web page: https://www.statcan.gc.ca/en/microdata/dli

Table of Contents

1.0 Survey description

The Canadian Legal Problems Survey (CLPS) was conducted by Statistics Canada in 2021 on behalf of the Department of Justice Canada. The purpose of the Canadian Legal Problems Survey (CLPS) was to identify the kinds of serious problems people face, how they attempt to resolve them, and how these experiences may impact their lives. The information collected will be used to better understand the various methods people use to resolve problems - not just formal systems such as courts and tribunals, but also informal channels such as self-help strategies.

This user guide has been produced to facilitate the manipulation of the Public Use Microdata File (PUMF) of the survey.

Any question about the data set or its use should be directed to:

Statistics Canada
Client Services
Centre for Social Data Integration and Development
Telephone: 613-951-3321 or call toll-free 1-800-461-9050
E-mail: csdid-info-cidds@statcan.gc.ca

Department of Justice Canada
Research and Statistics Division
284 Wellington Street, 6th floor, Ottawa, ON K1A 0H8
E-mail: rsd.drs@justice.gc.ca

2.0 Concepts and definition

2.1 CLPS concepts and definitions

The concepts and definitions used in this Canadian Legal Problems Survey (CLPS) are for the most part commonly used. However, the definitions of the following terms used in the questionnaire could be different from the ones used in other contexts.

Harassment is any improper conduct by an individual that is directed at and offensive to another individual and that the individual knew or should reasonably to have known would cause offence or harm.

Discrimination means treating someone differently or unfairly because of a personal characteristic or distinction, which, whether intentional or not, has an effect that imposes disadvantages not imposed on others or that withholds or limits access that is given to others.

2.2 Content development

The questionnaire content was developed in close collaboration with Justice Canada (DOJ). The analysts from the Canadian Centre for Justice and Community Safety Statistics (CCJCSS) were involved in the development of the survey, as they are the subject-matter experts for Justice themes within Statistics Canada.

The survey questionnaires from some previous surveys conducted by DOJ or the Canadian Forum on Civil Justice were consulted when building the initial version of the CLPS questionnaire. However, the CLPS findings shouldn’t be compared to those previous surveys in part as there are significant differences between the questionnaires.

Internal and external stakeholders identified by DOJ were also consulted throughout the content development to ensure that the survey questions would address their data needs as much as possible.

Throughout its development, the survey questionnaire was tested 3 times by the experts of StatCan's Questionnaire Design Resource Center (QDRC). Two rounds of cognitive interviews for content feasibility were completed in 2019. The first test took place in February and March 2019 where 35 one-on-one cognitive interviews took place in four cities: Halifax, Winnipeg, Ottawa, Montreal. Interviewees included a mix of languages (French or English), a mix of types of legal problems having been encountered in the past three years and as much as possible, a mix of age, gender, education, employment and income.

The second round of qualitative testing test took place in Ottawa in September 2019, where 19 interviews were conducted in both official languages. Participants were administered the questionnaire, including new questions, on a face-to-face basis (similar to a phone interview).

Based on the results of the qualitative testing, updates were made to the questionnaire and the content was finalized in collaboration with DOJ. A PDF version of the questionnaire was then designed collaboratively with internal partners at StatCan to mimic the screens of the electronic application. The PDF version of the questionnaire and in some cases an electronic section of the questionnaire underwent a third round of qualitative testing by the QDRC in March 2020. The testing was conducted in Montreal, Ottawa and Whitehorse for a total of 15 interviews in English and 12 in French.

Any changes required after qualitative testing were made to the PDF screen designs and communicated to the DOJ. An electronic questionnaire application was then developed and thoroughly tested at StatCan.

The survey consists of the following modules:

- Sociodemographic information (Part 1)
- Problems identification
- Covid-19 pandemic (Not available on the PUMF)
- Problem specific questions for the ones faced in the past 3 years. This includes modules: Consumer purchases and services, Employer and work, Debt and money owed to you, Interaction with police, Family and relationships, Child custody and parental responsibilities, Harassment, Discrimination
- Connections between problems (1 variable only on the PUMF)
- Most serious problem identification
- Assistance with the most serious problem
- Scope and status of the most serious problem
- Legal assistance for the most serious problem
- Costs associated with the most serious problem
- Socioeconomic impacts
- Health and social problems
- Sociodemographic information (Part 2)

Please see the CLPS questionnaire for more detailed information.

3.0 Survey methodology

3.1 Target and survey population

The survey target population includes individuals 18 years of age or older living in one of Canada's 10 provinces, with the exception of individuals living in an institution, in a collective dwelling or on an Indian reserve.

A sample of 42,400 people was randomly selected from the survey frame. The sample consists of a representative sample of 29,972 people from the general population as well as an oversample of 12,428 Indigenous Peoples.

3.2 Sample design

The survey frame used for the CLPS was a person-based list frame, constructed using the 2016 Long-form Census and other administrative files. The sample was made of individuals randomly selected from the frame. Those specific individuals were invited to participate in the survey.

The CLPS frame was stratified by province, and by Indigenous / non-Indigenous status. The Indigenous strata were sub-stratified by Indigenous identity (First Nations, Métis and Inuit) in order to improve the quality of estimates by Indigenous identity.

The sampling frame contained up to five telephone numbers (landline or cellular) to allow for telephone follow-up with a respondent.

3.3 Sample size

The sample initially allocated a 30,000 unit main sample for the non-Indigenous population and a 12,400 oversample for the Indigenous population, where both the main sample and the oversample were to be stratified by province using a Kish allocation. However, due to the relatively small number of Inuit people and Prince Edward Island residents, some strata were combined, resulting in a final allocation of 29,972 for the main sample and 12,428 for the Indigenous oversample.

4.0 Collection instrument and data collection

The data collection procedure for the Canadian Legal Problems Survey (CLPS) was an online electronic questionnaire (EQ) with telephone follow up for non-respondents. The electronic questionnaire data collection started on February 1, 2021 and ended on August 20, 2021.

An invitation letter accompanied by a brochure on the survey was sent to the selected person's personal mailing address to invite them to participate in the survey. Each letter provided a secure access code (SAC) for the respondent to access the electronic questionnaire and complete-it online.

In the survey invitation letter, brochure and reminders, participants were informed that the survey was voluntary and that their information would remain strictly confidential.

Four reminder letters, including a list of help resources, were also sent to try to increase the response rate.

Starting on March 15, 2021, computer-assisted telephone interviews (CATI) were also made by trained interviewers from regional offices to try to contact the selected persons and have them respond to the survey. When required, the interviewers took note of any information or other phone numbers given to try to reach the selected person. Proxy interviews were not permitted for this survey due to the sensitive topic.

5.0 Data processing

Processing transforms survey responses obtained during collection into a form that is suitable for tabulation and data analysis. It includes all data handling activities – automated and manual – after collection and prior to estimation.

5.1 Data capture

5.1.1 Electronic questionnaire (EQ)

For the electronic questionnaire, responses to survey questions were entered directly by the respondents. The electronic questionnaire reduces processing time and costs associated with data entry, transcription errors and data transmission. The responses were securely transferred to Statistics Canada in Ottawa, through industry standard encryption protocols, firewalls and encryption layers.

Some editing was done directly at the time the electronic questionnaire was completed. Where the information was outside an acceptable range (too large or small) of expected values, or inconsistent with the previous entries, the respondent was prompted, through message screens, to verify the information. However, the respondents had the option of bypassing the edits, and of skipping questions if they did not know the answer or refused to answer. Therefore, the data were subjected to further edit processes after they were submitted to head office. When the electronic data were received it was converted to readable text files.

5.1.2 Computer-assisted telephone interviews (CATI)

Responses to survey questions are captured directly by the interviewer at the time of the interview using an electronic questionnaire application. The computerized questionnaire reduces processing time and costs associated with data entry, transcription errors and data transmission. The responses provided by respondents were secure through industry standard encryption protocols, firewalls and encryption layers.

Some editing is done directly at the time of the interview. Where the information entered is out of range (too large or small) of expected values, or inconsistent with the previous entries, the interviewer is prompted, through message screens on the computer, to clarify the information with the respondent. However, for some questions interviewers have the option of bypassing the edits, and of skipping questions if the respondent does not know the answer or refuses to answer. Therefore, the response data are subjected to further editing processes once they arrive in head office.

5.2 Editing

Editing can occur at several points throughout the survey process and ranges from simple preliminary checks performed by the collection application to more complex automated verifications performed at the processing stage after the data have been captured. In general, edit rules are determined by what is logically or validly possible, based upon:

- expert knowledge of the subject matter;
- other related surveys or data;
- structure of the questionnaire and its questions; and
- statistical theory.

There are three main categories of edits: validity, consistency and distribution edits. Validity edits verify the syntax of responses and include such things as checking that the data lie within an allowed range of values. For example, a range edit might be placed on the reported age of a respondent to ensure that it lies between 0 and 121 years.

5.2.1 Recoding

For all records where there is a lack of values (no response provided), a non-response or “not-stated” code (9, 99, 999, etc.) is assigned to the item.

At the recode step, a value of 7 is assigned to any “Don’t know” responses. The Mark All type questions are modified to create dichotomous responses (Yes=1 or No=2) for each answer category.

All text responses (Other – Specify, for example) were removed from the data file and put aside in a separate folder for further manipulations, such as recoding.

5.2.2 Flow edits

The flow of the questionnaire is integrated within the application. For example, in the Problem identification section, we ask respondents to identified the types of problems they had to face in the past 3 years. Depending on their answers, the flow of the questions will be different.

Additional question were presented to respondents who selected specific types of problems; the others skipped those questions.

The flow edits are meant to check if the respondent answered questions which did not apply to them (and should therefore not have been answered). In this case a computer edit automatically eliminated superfluous data by following the flow of the questionnaire implied by answers to previous, and in some cases, subsequent questions.

The skip patterns triggered by some answers provided to some questions, for example a type of serious problem, are considered as flow conditions.

The responses skipped because of skip patterns are modified from Not Stated (« 9 », « 99 », « 999 », etc.) to Valid Skip (« 6 », « 96 », « 996 », etc.).

5.2.3 Consistency edits

Consistency edits verify that relationships between questions are respected. Consistency edits can be based on logical, legal, accounting or structural relationships between questions or parts of a question. The relationship between date of birth and marital status is one example where an edit might be: “a person less than 15 years of age cannot have any marital status other than single - never married.”
Once the “Not Stated” and “Valid Skips” were completed in processing, consistency edits were applied to the data, including the following:

1. Number of persons in household 18 of age or older (PHH_Q02)

   If the number of persons age 18 years or older (PHH_Q02) is greater than the number of persons in the household (PHH_Q01), then set PHH_Q02 to Not Stated.

   Also, if the number of persons age 18 years or older (PHH_Q02) is 0, then set PHH_Q02 to Not Stated.

2. Approximate total cost to deal with problem (CST_Q20)

   If at least one type of cost associated with the problem is reported (CST_Q10) and a total cost of $0 is reported (CST_Q20), then set CST_Q20 to Not Stated.

3. Received reimbursements/settlements for problem (CST_Q30)

   If a reimbursement or settlement is reported (CST_Q30 = 1) and a total reimbursement of $0 is reported (CST_Q40), then set CST_Q30 = 2 (“No”).

5.3 Coding open-ended questions

There were 28 open-ended questions on the CLPS that were recoded as described below. Some of those variables were either used to create other variables or dropped from the PUMF during the risk analysis for confidentiality.

1) Gender of respondent (GDR_S10): The “Or please specify” write in category was recoded back to male, female, gender diverse or Not Stated. For the PUMF, the gender diverse responses were randomly recoded as Male or Female.

2) Sexual orientation (SOR_S01): The “Or please specify” category was recoded. On the PUMF, the sexual orientation is presented as a dichotomous variable: heterosexual and Another sexual orientation.

3) Population group (PG_S05): The other-specify category was recoded to existing categories used to derive the VISMFLP variable presented on the PUMF. It is a dichotomous variable: Visible Minority and Not a visible Minority.

4) Disputes or problems (PRI_S05): The other-specify category was either recoded back into one of the existing categories, left as Other or changed to Not Stated. Corresponding changes were made to related variables, including PRI_Q10 and SERPROBP, for internal consistency.

5) For the following variables, the other-specify category were either recoded back into one of the existing categories or left as Other:

   - Large purchase/service (CON_S10)
   - Employer/job (EMP_S10)

- Debt/money owed to you (DEB_S10)
- Breakdown of family (FAM_S10)
- Child custody (CHL_S10)
- Harassment (DSH_S10)
- Harassment based on (DSH_S20)
- Nature of harassment (DSH_S30)
- Discrimination (DSH_S40)
- Discrimination based on (DSH_S50)
- Action to resolve problem (AST_S10)
- Type of lawyer contacted (AST_S30)
- Reason no lawyer contacted (AST_S40)
- Reason no action taken (AST_S50)
- Help for only part of problem (LGA_S20)
- No legal help (LGA_S30)
- Wish you had to solve problem (STA_S40)
- Costs associated with problem (CST_S10)
- Financial challenges (CST_S50)
- Live after losing your housing (SOC_S30)
- Problems experienced because of serious problem (HLT_S30)

5.4 Creation of derived variables

In order to facilitate data analysis, a number of data items on the microdata file were derived by combining items on the questionnaire. In these cases, two or more variables were used to create a new variable.

5.5 Imputation

Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data. This is done by changing some of the responses and all of the missing values on the record being edited to ensure that a plausible, internally consistent record is created.

For a small number of cases, the province of residence was not stated in the submitted questionnaire. In that situation, it was imputed directly with the province from the sample file using SAS.

5.6 Disclosure control

Statistics Canada is prohibited by law from releasing any information it collects that could identify any person, business, or organization, unless consent has been given by the respondent or as permitted by the Statistics Act. Various confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.

The micro data file will not contain any personal identifiers. Individual responses and results for very small groups will not be published or shared with any stakeholders.

For aggregate or tabular data, confidentiality will be maintained by both cell collapsing and suppression of data where necessary.

6.0 Data quality

This section covers the various factors that impact the quality of data collected in the survey. Survey errors come from a variety of different sources. They can be classified into two main categories: non-sampling errors and sampling errors.

6.1 Non-sampling errors

Non-sampling errors can be defined as errors arising during the course of virtually all survey activities, apart from sampling. They are present in both sample surveys and censuses (unlike sampling error, which is only present in sample surveys). Non-sampling errors arise primarily from the following sources: non-response, coverage, measurement and processing.

6.1.1 Non-response

Non-response is both a source of non-sampling error and sampling error. Non-response results from a failure to collect complete information from all units in the selected sample. Non-response is a source of non-sampling error in the sense that non-respondents often have different characteristics from respondents, which can result in biased survey estimates if non-response bias is not fully eliminated through weighting adjustments. The lower the response rate, the higher the risk of bias. Non-response is also a source of sampling error; this is discussed further in Section 6.2. The overall response rate for CLPS was 50.7%.

6.1.2 Coverage errors

Coverage errors consist of omissions, erroneous inclusions, duplications and misclassifications of units in the survey frame. Coverage errors may cause a bias in the estimates and the effect can vary for different sub-groups of the population.

CLPS data is collected from people aged 18 years and over living in private dwellings within the 10 provinces, excluding individuals living in a collective dwelling, an institution or on an Indian reserve. These exclusions represent approximately 2% of the Canadian population aged 18 and over living in the 10 provinces. It is often not possible to accurately quantify coverage errors. Potential sources of overcoverage errors include respondents who have moved to institutions, collective dwellings, or Indian reserves but whose contact information does not reflect that move.

A significant proportion of the units in the Indigenous oversample did not indicate on CLPS that they currently identified as Indigenous. Conversely, fewer than 1% of units in the main sample indicated that they now identify as Indigenous. The potential bias introduced by this change in Indigenous identity was addressed in the weighting process.

6.1.3 Measurement errors

Measurement errors (or sometime referred to as response errors) occur when the response provided differs from the real value; such errors may be attributable to the respondent, the interviewer, the questionnaire, the collection method or the respondent's record-keeping system. Such errors may be random or they may result in a systematic bias if they are not random.

It is very costly to accurately measure the level of response error and very few surveys conduct a post-survey evaluation. However, interviewer feedback and observation reports usually provide clues as to which questions may be problematic (poorly worded question, inadequate interviewer training, poor translation, technical jargon, no help text available, etc.).

Several measures are taken at Statistics Canada to prevent and reduce the level of response error. These measures include questionnaire review and testing using cognitive methods, the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and content, and continuous monitoring of collection processes.

6.1.4 Processing errors

Processing error is the error associated with activities conducted once survey responses have been received. It includes all data handling activities after collection and prior to estimation. Like all other errors, they can be random in nature, and inflate the variance of the survey's estimates, or systematic, and introduce bias. It is difficult to obtain direct measures of processing errors and their impact on data quality especially since they are mixed in with other types of errors (non-response, measurement and coverage).

6.2 Sampling errors

Sampling error is defined as the error that results from estimating a population characteristic by measuring a portion of the population rather than the entire population. For probability sample surveys, methods exist to calculate sampling error. These methods derive directly from the sample design and method of estimation used by the survey.

The most commonly used measure to quantify sampling error is sampling variance. Sampling variance measures the extent to which the estimate of a characteristic from different possible samples of the same size and the same design differ from one another. For sample designs that use probability sampling, the magnitude of an estimate's sampling variance can be estimated.

Factors affecting the magnitude of the sampling variance include:

1. The variability of the characteristic of interest in the population: the more variable the characteristic in the population, the larger the sampling variance.
2. The size of the population: in general, the size of the population only has an impact on the sampling variance for small to moderate sized populations.
3. The response rate: the sampling variance increases as the sample size decreases. Since non-respondents effectively decrease the size of the sample, non-response increases the sampling variance.
4. The sample design and method of estimation: some sample designs are more efficient than others in the sense that, for the same sample size and method of estimation, one design can lead to smaller sampling variance than another.

The standard error of an estimator is the square root of its sampling variance. This measure is easier to interpret since it provides an indication of sampling error using the same scale as the estimate whereas the variance is based on squared differences.

The coefficient of variation (CV) of an estimate is a relative measure of the sampling error. It is defined as the estimate of the standard error divided by the estimate itself, usually expressed as a percentage (10% instead of 0.1). It is very useful for measuring and comparing the sampling error of quantitative variables with large positive values. However, it is not recommended for estimates such as proportions, estimates of change or differences, and variables that can have negative values.

It is considered a best practice at Statistics Canada to report the sampling error of an estimate through its 95% confidence interval. The 95% confidence interval of an estimate means that if the survey were repeated over and over again, then 95% of the time (or 19 times out of 20), the confidence interval would cover the true population value.

7.0 Weighting

The principle behind estimation in a probability sample is that each unit selected in the sample represents, besides itself, other units that were not selected in the sample. For example, if a simple random sample of size 100 is selected from a population of size 5,000, then each unit in the sample represents 50 units in the population. The number of units represented by a unit in the sample is called the survey weight of the sampled unit.

The following section provides the details of the method used to calculate sampling weights for the CLPS.

The weighting for CLPS consisted of several steps:

1) Calculation of design weights
2) Removal of out-of-scope units
3) Non-response adjustment
4) Treatment of influential weights
5) Calibration

Each of these steps is described in more detail in the following four subsections.

7.1 Design weights

Each unit in the sample was assigned a basic weight, $\boldsymbol{W_{1,i}}$, equal to the inverse of its probability of selection within each province (where i represents the province).

$$W_{1,i} = \left(\frac{Number\ of\ eligible\ units\ on\ the\ frame}{Number\ of\ sampled\ units}\right)$$

This weight takes into account the likelihood that a person was selected for the 2016 Long-form Census. There were 42,400 sampled units with assigned weights.

7.2 Removal of out-of-scope units

Out-of-scope units, such as individuals not living in the provinces or those younger than 18 years of age were excluded from further weighting adjustments. At this stage, there were 317 units identified as out-of-scope. The following step was applied:

If out-of-scope unit,

$$W_{2,i} = 0$$

Otherwise,

$$W_{2,i} = W_{1,i}$$

7.3 Non-response adjustment

The 42,083 remaining units were separated into two groups: 21,170 respondents and 20,913 non-respondents. Because so few units were identified as being out of scope in the previous step, all non-respondents were assumed to be in scope. A combination of auxiliary data available from the frame, as well as collection paradata was used to model response propensity using logistic regression. Only variables which were predictive of both response propensity and the key variables of interest were used in the logistic regression model. Response homogeneous classes were then constructed using the predicted response propensity. Non-response adjustment factors were computed within each weighting class. The 20,913 non-respondents were then removed from further adjustments. The non-response adjusted weights, denoted by $W_{3,i}$, were calculated for the 21,170 respondents using the following formula:

$$W_{3,i} = W_{2,i} * \left(\frac{\sum W_2 \; for \; respondents + \sum W_2 for \; non - respondents}{\sum W_2 \; for \; respondents} \right)$$

For the 20,913 non-respondents, $W_{3,i}$ was set to 0.

7.4 Treatment of influential weights

Units whose Indigenous identity on CLPS differed from the 2016 Long-form Census and units whose province on CLPS was different from the sample file were considered stratum jumpers. For some of these units, weights were significantly bigger than those of the non-stratum jumpers in their final stratum. For example, an individual whose province on the census was Ontario but who had since moved to Prince Edward Island would have a much larger weights than individuals who were in PEI for both the 2016 census and CLPS. Units whose weights were more than three standard deviations from the mean of their stratum had their weights reduced to be equal to the maximum weight of the non-stratum jumpers in their final stratum. There were 41 units whose weights were adjusted in this manner.

7.5 Adjustment to known totals

An adjustment was made to the weights in order to make population estimates consistent with external population counts for persons 18 years and older. This is known as post-stratification. The following external control totals were used:

1) Population totals for each province*sex*age group.
2) Population totals by Indigenous identity.

The population totals for province*sex*age group were obtained from demography estimates, and the population totals by Indigenous identity were estimated using the 2016 Long-form Census. The person weights obtained after this step represent the final person-level weight that is available on the microdata files. The sum of the final weights for the 21,170 records included on the final file represent the estimate of the CLPS target population.

8.0 Guidelines for tabulation, analysis and release

This chapter of the documentation outlines the recommended guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey microdata files. With the aid of these guidelines, users of microdata should be able to produce the same figures as those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

8.1 Rounding guidelines

Users are urged to adhere to the following rounding guidelines when producing estimates and statistical tables computed from these microdata files:

a) Estimates in the main body of a statistical table are to be rounded using the normal rounding technique. In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one.

b) Marginal sub-totals and totals in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves using normal rounding. Averages, rates, percentages, proportions and ratios are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves using normal rounding. Sums and differences are to be derived from their corresponding unrounded components and then are to be rounded themselves using normal rounding.

c) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used resulting in estimates to be published or otherwise released which differ from corresponding estimates published by Statistics Canada, users are urged to note the reason for such differences in the publication or release document(s).

d) Under no circumstances are unrounded estimates to be published or otherwise released by users. Unrounded estimates imply greater precision than actually exists.

8.2 Sample weighting guidelines for tabulation

The CLPS uses a complex sample design and estimation method, and the survey weights are therefore not equal for all the sampled units. When producing estimates and statistical tables, users **must** apply the proper survey weights. If proper weights are not used, the estimates derived from the microdata files cannot be considered to be representative of the survey population, and will not correspond to those produced by Statistics Canada.

8.3 Release guidelines for quality

Before releasing and/or publishing any estimates, analysts should consider the quality level of the estimate. While data quality is affected by both sampling and non-sampling errors, this section covers quality in terms of sampling error. It is considered a best practice at Statistics Canada to report the sampling error of an estimate through its 95% confidence interval (CI). The confidence interval should be released with the estimate, in the same table as the estimate. In addition to the confidence intervals, estimates are categorized into one of three release categories:

Category A
The estimate and confidence interval can be released with no warning. Data users should use the 95% confidence interval to assess whether the quality of the estimate is sufficient. Note that the ‘A’ is not a quality indicator; it should not be released.

Category E
The estimate and confidence interval should be flagged with the letter E (or some similar identifier) and accompanied by a quality warning to use the estimate with caution. Data users should use the 95% confidence interval to assess whether the quality of the estimate is sufficient.

Category F
The estimate and confidence interval are not recommended for release. They are deemed of such poor quality, that they are not fit for any use; they contain a very high level of instability, making them unreliable and potentially misleading. If analysts insist on releasing estimates of poor quality, even after being advised of their accuracy, the estimates should be accompanied by a disclaimer. Analysts should acknowledge the warnings given and undertake not to disseminate, present or report the estimates, directly or indirectly, without this disclaimer. The estimates should be flagged with the letter F (or some similar identifier) and the following warning should accompany the estimates and confidence intervals: “Please be warned that these estimates and confidence intervals [flagged with the letter F] do not meet Statistics Canada’s quality standards. Conclusions based on these data will be unreliable, and may be invalid.”

The rules for assigning an estimate to a release category depends on the type of estimate.

Release Rules for Estimated Proportions and Estimated Counts
Estimated proportions and estimated counts are computed from binary variables. Estimated counts are estimates of the total number of persons/households with a characteristic of interest;

in other words, they are the weighted sum of a binary variable (e.g., estimated number of immigrants). Estimated proportions are estimates of the proportion of persons/households with a characteristic of interest (e.g., estimated proportion of immigrants in the general population). Estimated counts and proportions can also be computed from categorical variables: that is, estimates of the number or proportion of persons/household who belong to a category.

The release rules for estimated proportions and estimated counts are based on sample size. Tables 1, 2, and 3 provide the release rules for the CLPS. The rules in Table 2 are used whenever the domain of interest is at the province-level or below (except estimates for Indigenous peoples) ; in other words, all the respondents that contribute to the estimate belong to the same province. The rules in Table 3 are used whenever estimates are produced at the level of Indigenous identity. Otherwise the rules in Table 1 are used.

Table 1: General rules for proportions and counts

Sample Size (n)	**Release Category**	**Action**
n > 130	A*	Release with no warning; users should use CI as quality indicator
65 ≤ n ≤ 130	E	Release with quality warning; users should use CI as quality indicator
n < 65	F	Suppress the estimate and its CI for quality reasons

Table 2: Rules for proportions and counts for estimates at the province-level or below

Sample Size (n)	**Release Category**	**Action**
n > 75	A*	Release with no warning; users should use CI as quality indicator
37 ≤ n ≤ 75	E	Release with quality warning; users should use CI as quality indicator
n < 37	F	Suppress the estimate and its CI for quality reasons

Table 3: Rules for proportions and counts for estimates at the level of Indigenous identity

Sample Size (n)	**Release Category**	**Action**
n > 90	A*	Release with no warning; users should use CI as quality indicator
45 ≤ n ≤ 90	E	Release with quality warning; users should use CI as quality indicator
n < 45	F	Suppress the estimate and its CI for quality reasons

* Note that ‘A’ is not a quality indicator; it should not be released with the estimate. The 95% confidence interval is the quality indicator.

For estimated proportions, *n* is defined as the unweighted count of the number of respondents in the denominator (not the numerator) of the proportion. For estimated counts, *n* is defined as the unweighted count of the number of respondents with nonzero values that contribute to the estimate.

Release Rules for Means and Totals of Quantitative Variables

The release rules for the estimated means and totals of quantitative variables or amounts are based on the sample size and on the CV of the estimate. Tables 4, 5, and 6 provide the release rules for the CLPS. The rules in Table 5 are used whenever the domain of interest is at the province-level or below (except estimates for Indigenous peoples); in other words, all the respondents that contribute to the estimate belong to the same province. The rules in Table 6 are used whenever estimates are produced at the level of Indigenous identity. Otherwise the rules in Table 4 are used.

Table 4: General rules for means and totals

Sample Size (n)	Release Category	Action
n>130 and CV≤25%	A*	Release with no warning; users should use CI as quality indicator
Otherwise	E	Release with quality warning; users should use CI as quality indicator
n<65 or CV>50%	F	Suppress the estimate and its CI for quality reasons

Table 5: Rules for means and totals for estimates at the province-level or below

Sample Size (n)	Release Category	Action
n>75 and CV≤25%	A*	Release with no warning; users should use CI as quality indicator
Otherwise	E	Release with quality warning; users should use CI as quality indicator
n<37 or CV>50%	F	Suppress the estimate and its CI for quality reasons

Table 6: Rules for means and totals for estimates at the level of Indigenous identity

Sample Size (n)	Release Category	Action
n>90 and CV≤25%	A*	Release with no warning; users should use CI as quality indicator
Otherwise	E	Release with quality warning; users should use CI as quality indicator
n<45 or CV>50%	F	Suppress the estimate and its CI for quality reasons

* Note that ‘A’ is not a quality indicator; it should not be released with the estimate. The 95% confidence interval is the quality indicator.

For estimated means, *n* is defined as the unweighted count of the number of respondents that contribute to the estimate including values of zero. For estimated totals, *n* is defined as the unweighted count of the number respondents with nonzero values that contribute to the estimate.

Release Rules for Differences

In order to assign a release category for an estimated difference between two estimates, the analyst must first determine the release category of each of the two estimates using the rules described above. Next, the release category of the estimated difference or the estimate of change is assigned the lower release category of the two estimates; this can be specified as follows:

- If one or both estimates are category F estimates, then assign the estimated difference to category F and suppress it
- Otherwise, if one or both estimates are category E estimates, then assign the estimated difference to category E
- If both estimates are category A estimates, then assign the estimated difference to category A

Additional Rules Regarding Confidence intervals

The above release rules should suppress most estimates and confidence intervals of poor quality. There are also two conditions that indicate that a confidence interval is of poor quality. An estimate and its confidence interval should be assigned to release category F if either of the following two conditions are true:

- The lower bound of the 95% confidence interval is equal to the upper bound of the interval; in other words, the confidence interval is of length zero. (Exceptions are if the estimate corresponds to a calibration control total.)
- The lower bound or upper bound of the 95% confidence interval is not a plausible value for the estimate. For example, the lower bound for an estimated proportion is negative.

8.4 Guidelines for Statistical Analysis, Variance Estimation and Constructing Confidence intervals

In order to measure the sampling error of estimates, variance estimates need to be calculated and confidence intervals need to be constructed. The CLPS uses a complex sample design and estimation method, which means that there is no simple formula for calculating variance estimates. The survey therefore uses a resampling method called the bootstrap. One thousand sets of bootstrap weights were generated. Essentially, the variance is estimated by calculating the value of the estimate of interest using each set of bootstrap weights and then measuring the variability between the 1,000 bootstrap estimates.

Statistical packages for statistical analysis and variance estimation

It is necessary to use bootstrap weights to compute correct estimates of the variance for this survey. A number of statistical software programs or packages have been developed that are specifically designed for analyses of data from complex survey designs and that can compute variance estimates using replicate weights such as bootstrap weights. These include for example SUDAAN, WesVar, Stata and newer versions of SAS.

Other standard and/or older statistical analysis software packages including, SPSS, versions of SAS prior to version 9.2, do not have an integrated procedure to calculate variance estimates from bootstrap weights when using data based on a complex survey design. These packages should not be used to calculate variance estimates, to construct confidence intervals nor to conduct statistical tests (significance tests, regression analysis, etc.).

SAS version 9.2 and above can calculate variances from bootstrap weights, as well as other types of replicate weights such as Jackknife and Balanced Repeated Replication (BRR) weights. There are also a number of procedures, such as regression, logistic regression for instance, that accommodate replicate weights. Confidence intervals for medians using replicate weights are only available in SAS version 9.3 and above.

It should be noted that software packages that do not explicitly support bootstrap weights but do support the BRR method, can be used with bootstrap weights. While the bootstrap and BRR methods differ in the way in which the replicate weights are built, once the replicate weights are produced, the two methods use a similar formula to compute variance estimates. For more information on the relationship between the bootstrap and the BRR method, please refer to Phillips (2004).

Multiplicative factor

The method used to create the bootstrap weights for CLPS included a step where the bootstrap weights were transformed to remove negative weights. The transformed bootstrap weights require that the variance estimates for CLPS be multiplied by a factor of 4. **It is extremely important to apply this multiplicative factor; omission of the factor would lead to erroneous results and conclusions.**

Statistical software packages that support the BRR with Fay's adjustment can produce correct variance estimates without the need of an extra multiplication step. The multiplicative factor can be specified by using the Fay parameter: for some software packages (e.g., SUDAAN), use a Fay parameter of C, where C is the variance multiplicative factor. For other software packages (SAS, in particular), use a Fay factor k, where $k = 1 - \sqrt{\frac{1}{C}}$. For CLPS, $C = 4$ and $k = 0.5$.

Confidence intervals

The most commonly used method of constructing 95% confidence intervals is the Wald interval, which is of the form $\hat{y} \pm 1.96\sqrt{\hat{\text{var}}(\hat{y})}$ for an estimate $\hat{y}$ with estimated variance $\hat{\text{var}}(\hat{y})$. Wald intervals are based on the assumption that the sampling distribution of $\hat{y}$ is approximately normal. For proportions, the normality assumption is known to break down for small sample sizes and for proportions near zero or one. Three alternative methods of constructing confidence intervals are therefore recommended for proportions: the modified Wilson interval, the modified Clopper-Pearson interval and the logit interval (see Korn and Graubard, 1998; Liu and Kott, 2009). There are options in SAS and SUDAAN to produce confidence intervals using these alternative methods.

The examples below show how alternative methods of constructing confidence intervals are specified for proportions in SAS and SUDAAN, and they include the necessary Fay factor described above.

1. SAS, modified Wilson confidence intervals:

```
PROC SURVEYFREQ
DATA=…. VARMETHOD=BRR (Fay = 0.5);
WEIGHT WTPP;
REPWEIGHTS WRPP1-WRPP1000;
TABLES …. / CL (TYPE=WILSON  ADJUST=NO TRUNCATE=YES)
```

2. SUDAAN, modified Clopper-Pearson confidence intervals:

```
PROC CROSSTAB
DATA=…. DESIGN=BRR SMCONF=50;
WEIGHT WTPP;
REPWGT WRPP1-WRPP1000 / ADJFAY=4;
TABLES ...;
```

Rescaling the weights

As mentioned, it is recommended that users use analysis procedures designed for the analysis of data from complex survey designs, which can use weights to produce estimates and can use bootstrap weights to produce variance estimates. Analysis procedures not designed for the sample survey framework may allow weights to be used (without bootstrap weights). However, these procedures may differ in their definition for the weight, and produce correct estimates but meaningless variance estimates. For analyses such as linear regression, logistic regression and analysis of variance, rescaling the weights can make the variance estimates calculated by the standard packages more reasonable. The weights for the domain of interest should be rescaled so that the average weight is one (1); this can be accomplished by dividing each weight by the overall average weight before the analysis is conducted. The rescaling makes the variance estimates more reasonable, but they only take into account the unequal probabilities of selection - they do not take into account the stratification and clustering of the sample's design. This approach should therefore only be used as a last resort when no procedures that can use bootstrap weights are available; users are warned that the results are approximate.

References

Korn, E.L., and Graubard, B.I. (1998). “Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data”. *Survey Methodology*, 24, 193-201.

Liu, Y.K. and Kott, P.S. (2009). “Evaluating Alternative One-Sided Coverage Intervals for a Proportion”. *Journal of Official Statistics*, Vol. 25, No. 4, 569-588.

Phillips, O. (2004). “Using bootstrap weights with WesVar and SUDAAN” (Catalogue no. 12-002-X20040027032) in The Research Data Centres Information and Technical Bulletin, Chronological index, Fall 2004, vol.1 no. 2 Statistics Canada, Catalogue no. 12-002-XIE.