# Abridged Microdata User Guide

# Survey on Early Learning and Child Care Arrangements

Public Use Microdata File (PUMF)

2022

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
Canadian Centre for Education Statistics
Telephone: 613-951-3321 or call toll-free 1-800-461-9050
E-mail: statcan.education-education.statcan@statcan.gc.ca

**Accessing and ordering information**
The Survey on Early Learning and Child Care Arrangements (SELCCA) produces two types of microdata files: master file and public use microdata file (PUMF). For more information, please consult the following web page:
https://www.statcan.gc.ca/eng/microdata/data-centres

**Master files**
The master files contain all variables and all records from the collected survey data and administrative data. These files are accessible at a cost at Statistics Canada for internal use and in Statistics Canada's Research Data Centres (RDC), real time remote access (RTRA), and virtual data lab (vDL). Master files 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.

**Real Time Remote Access**
The RTRA system is an on-line remote access facility allowing users to run SAS programs, in real-time, against micro-data sets located in a central and secure location. Researchers submit SAS programs to extract results in the form of frequency tables. Confidentiality of the micro data is automated in the RTRA system. Researchers must complete an application form.

**Virtual Data Lab**
The vDL is a cloud-based microdata access centre that allows approved researchers to remotely access data housed by Statistics Canada using a secure IT connection and a protected Cloud environment. Several organizations have already signed memorandums of understanding with Statistics Canada that have allow approved researchers to transition into the vDL environment.

**Custom tabulations**
Another way to access the master files is to offer all users the option of having staff in the Centre for Social Data Integration and Development (CSDID) Client Services 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 the CSDID Client Services at csdid-info-cidds@statcan.gc.ca.

**Public use microdata files**
The public use microdata files (PUMF) are developed from the master files 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, please contact the Infostats team at infostats@statcan.gc.ca.

**Data Liberation Initiative**

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

# Table of Contents

1.0 Introduction ..... 5
2.0 Estimation ..... 5
2.1 Weighting Guidelines ..... 5
2.2 Rounding Guidelines ..... 5
3.0 Data Quality ..... 5
3.1 Non-sampling Errors ..... 6
3.2 Sampling Error ..... 6
3.3 Release Guidelines for Quality ..... 7
3.4 Guidelines for Statistical Analysis, Variance Estimation, and Constructing Confidence Intervals 9

# 1.0 Introduction

This documentation provides basic information on producing estimates for the Survey on Early Learning and Child Care Arrangements (SELCCA) Public Use Microdata File (PUMF) and on reporting the quality of the estimates. For general information about the survey, refer to the online information (IMDB).

# 2.0 Estimation

## 2.1 Weighting Guidelines

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 SELCCA 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.

## 2.2 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.

# 3.0 Data Quality

This section discusses survey error (non-sampling error and sampling error), the rules that should be used to decide whether an estimate can be released for quality reasons, and computing variance estimates and confidence intervals.

## 3.1 Non-sampling Errors

Survey errors come from a variety of different sources. They can be classified into two main categories: non-sampling errors and sampling errors. Non-sampling errors may arise in nearly all survey activities. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered in the system, and errors may be introduced in the processing and tabulation of the data. These are all examples of non-sampling errors. Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates.

Non-response is both a source of non-sampling error and sampling error. Non-response result 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 3.2. A response rate of 58.0% was obtained for the SELCCA.

## 3.2 Sampling Error

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. This is an issue when the response rate is lower than the anticipated response rate used to inflate the initial sample size.
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 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. 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 the confidence interval would cover the true population value 95% of the time (or 19 times out of 20).

## 3.3 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 and 2 provide the release rules for the SELCCA PUMF. The rules in Table 1 are used for estimates at the level of Canada or at the level of Canada by age group. The rules in Table 2 are used whenever the domain of interest is at the province-level or below; in other words, all the respondents that contribute to the estimate belong to the same province. This includes the level of province by age group.

Table 1: General rules for proportions and counts

| Sample Size (n) | Release Category | Action |
|---|---|---|
| n ≥ 115 | A* | Release with no warning; users should use CI as quality indicator |
| 60 ≤ n < 115 | E | Release with quality warning; users should use CI as quality indicator |
| n < 60 | 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 ≥ 80 | A* | Release with no warning; users should use CI as quality indicator |
| 40 ≤ n < 80 | E | Release with quality warning; users should use CI as quality indicator |
| n < 40 | 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 3 and 4 provide the release rules for the SELCCA PUMF. The rules in Table 3 are used for estimates at the level of Canada or at the level of Canada by age group. The rules in Table 4 are used whenever the domain of interest is at the province-level or below; in other words, all the respondents that contribute to the estimate belong to the same province. This includes the level of province by age group.

Table 3: General rules for means and totals

| Sample Size (n) | Release Category | Action |
|---|---|---|
| n≥115 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<60 or CV>50% | F | Suppress the estimate and its CI for quality reasons |

Table 4: Rules for means and totals for estimates at the province-level or below

| Sample Size (n) | Release Category | Action |
|---|---|---|
| n≥80 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<40 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.

## 3.4 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 SELCCA 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).

**Confidence intervals**

The most commonly used method of constructing 95% confidence intervals is the Wald interval, which is of the form $\hat{y} \pm 1.96\sqrt{\text{v}\hat{\text{a}}\text{r}(\hat{y})}$ for an estimate $\hat{y}$ with estimated variance $\text{v}\hat{\text{a}}\text{r}(\hat{y})$. 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.

1. SAS, modified Wilson confidence intervals:

```
PROC SURVEYFREQ
DATA=…. VARMETHOD=BRR;
WEIGHT WTMP;
REPWEIGHTS WRMP1-WRMP1000;
TABLES …. / CL (TYPE=WILSON  ADJUST=NO TRUNCATE=YES)
```

2. SUDAAN, modified Clopper-Pearson confidence intervals:

```
PROC CROSSTAB
DATA=…. DESIGN=BRR SMCONF=50;
WEIGHT WTMP;
REPWGT WRMP1-WRMP1000;
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 of the domain 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.