# Guide For Cross-Sectional Public-Use Microdata File

# Survey of Labour and Income Dynamics (SLID)

# Reference Year 2003

## Table of Contents

# 1. INTRODUCTION

The cross-sectional public-use microdata file for the Survey of Labour and Income Dynamics (SLID) is a collection of income, labour and family variables on persons in Canada and their families. SLID is an annual household survey covering the population of the 10 Canadian provinces with the exception of Indian reserves, residents of institutions and military barracks.

The Survey of Labour and Income Dynamics began collecting data for reference year 1993. Initially, SLID was designed to be, first and foremost, a longitudinal survey, with primary focus on labour and income and the relationships between them and family composition. Then, the decision was made to extend the objectives of SLID to be the primary source of cross-sectional household income data.

For many years, the Survey of Consumer Finances had provided public-use microdata files (PUMFs) to meet the needs of cross-sectional household income data users. SCF PUMFs were released up to and including reference year 1997. For the purpose of standard publications, Statistics Canada has made the transition from SCF to SLID between 1995 and 1996. Therefore, SLID cross-sectional PUMFs are being made available beginning with reference year 1996. The SLID files have been designed to be analogous to those produced for the SCF. The type of income data collected by SLID was identical to that of the former household income survey SCF (Survey of Consumer Finances), with the distinction that SLID respondents had the choice of a traditional income interview and granting permission to Statistics Canada to use their T1 income tax data.

To find more information on comparability between SLID and SCF data please consult the two following documents: *Bridging two surveys - An integrated series of income data from SCF and SLID - 1989-1997* and *Comparison of income estimates from the survey of consumer finances and the survey of labour and income dynamics* (see also section Related products and services)

*How to cite SLID in publications*

For publication of any information based on the SLID microdata files on CD-ROM (75M0010XCB), the following form of accreditation is recommended:

"This analysis is based on Statistics Canada's Survey of Labour and Income Dynamics Public Use Microdata, which contains anonymized data collected in the Survey of Labour and Income Dynamics. All computations on these microdata were prepared by (Name of user). The responsibility for the use and interpretation of these data is entirely that of the author(s)".

## 2. FILE STRUCTURES

Although often referred to as one file, the SLID cross-sectional PUMF is four separate flat files: key, person, economic family and census family. To a large extent, the file structure used for SCF PUMFs has been maintained.

On the person file, there is one record per person in the sample aged 16 and over. Job characteristics such as industry, wage rates and work schedule are included on the person file and relate to the person's main job during the reference year (the job at which the most hours were worked during the year). Although SLID collects data on all jobs held during the year by each person under 70 years old, the characteristics of all other jobs are not included on the SLID PUMFs.

The person file does contain identifiers that allow a researcher to group persons into households, economic families and census families.

In 2003, there was no change to PUMF files structure from the previous reference year.

The sizes of the 2003 public-use files are:

| Files | Number of Records | Number of Variables | Record length |
|---|---|---|---|
| Person file | 57,233 | 130 | 491 |
| Economic family file | 29,846 | 67 | 371 |
| Census family file | 33,461 | 64 | 368 |
| Key file | 71,418 | 14 | 45 |

## 3. USING THE RECORD LAYOUTS, DATA DICTIONARY AND UNIVARIATE DISTRIBUTIONS

Additional information files are provided to assist users of the SLID public-use microdata files. For each of the four data files (key, person, economic family and census family), record layout, data dictionary and univariate distributions are provided. These information files are organized by content themes and in some cases sub-themes.

The following describes the structure of the additional information files:

A. The columns of the record layout file

- *Variable name*. This is the variable name assigned for the microdata file.
- *Type*. Indicates whether the variable is numeric (in the sense that it can logically be used in mathematical operations) or character.
- Sequence. Indicates the order of variable appearance
- Length. Indicates both the number of spaces including the decimal point if there are decimal places and the number of decimal places, if any. For example, a variable which can have values of zero (00.0) to 99.9 would have a format expressed as: 4.1. A variable which can have values of zero (00) to 99 would have a format expressed as: 2.
- *Start position*. This shows the location of the variable on the public use file.
- *Number of categories*. Shows the number of categories in the value set for the variable in question. Applies only to "character" variables. Numeric variables have ranges, which are specified in the data dictionary.
- *Long variable name*. A standardized name, with a maximum of 26 characters, which can be used to quickly identify variables, to label tables, and so on. Although still rather cryptic, it is considerably more revealing than the variable name. However, this longer name obviously excludes a lot of important information contained in the variable description shown in the data dictionary. In short, analysts are warned against making assumptions about the variable definition based on the long variable name.

B. Data Dictionary

The data dictionary presents the complete information about each survey variable on each of the four files. For each variable in the record layout the following information is shown: the variable name, the description or definition, code lists with descriptions or alternatively the range of values that the variable can take on, the variable type, its length (or format), and the population to which the variable pertains, i.e. for whom it is applicable.

C. Univariate Distributions

These distributions are provided to allow users of the public use microdata files to verify totals that they produce. These distributions relate to the public-use files and not to the internal database; the distributions will be similar but not identical due to confidentiality processing procedures used to produce the public-use files.

For character variables, the weighted and unweighted frequencies for each code, including reserved codes (see below), are provided. For numeric variables, the values are broken into several ranges and weighted and unweighted frequencies are provided for each range. The minimum value, the maximum value and the weighted mean (excluding reserved codes) are also provided.

**Missing Values and Reserved Codes**

There are a few types of missing values on the public use file. SLID has adopted standard codes which have a particular meaning. It is important to account for reserved codes in any analysis, particularly with numeric variables. If your calculation of means or aggregates seems too high, check to ensure that you have excluded reserved codes from the calculation. With only a few exceptions, the reserved codes are the highest four values permitted according to the length of the variable. A brief explanation of reserved codes is provided below.

If the coverage of a variable does not extend to a certain population sub-group, then there are no valid values for that sub-group and the values (reserved codes) that do appear are in the form 9, 99, 9.9 and so on, which indicates that the variable is not applicable. The coverage of each variable on the file is referred to in the data dictionary as the “population”.

For certain records, no valid value is available, although the value is applicable. Possibly, the respondent did not provide the information or it failed an edit in processing and the value was not imputed. Such missing values appear with a reserved code such as 7, 97, 9.7 and so on depending on the format. For certain variables, the number of missing values has been reduced through imputation. Missing values for the income variables have been entirely imputed, but most other variables may have missing values.

Finally, a few values may have been coded as 8, 98, 9.8, etc. These represent refusals to particular items in the interview. The approach for dealing with missing values of this last kind depends on the type of analysis being carried out and the extent of missing data. Although the end solution may be to exclude the records with missing values from the analysis a review should first be carried out to assess the impact of missing values on the overall representativeness of the data. Is it possible that a bias results from the missing data – for example, are the (other) characteristics of the people with missing values different from those of the observed part of the sample? It may be necessary to take into account the possible impact in some way. In all cases, analysts should note exclusions of records with missing values in their published results.

## 4. GUIDELINES FOR APPLYING WEIGHTS

The microdata on the public use file are unweighted. It is the responsibility of data users to apply the appropriate weights in any estimates they wish to produce. If proper weights are not used, the results derived from the microdata cannot be considered to be representative of the survey population, and will not correspond to those that would be produced by Statistics Canada. The weights are provided as variables under "Sample control". On the SLID PUMF, the weight variable is named ICSWT26.

## 5. GUIDELINES FOR RELEASE (DATA QUALITY AND ROUNDING)

Microdata users should apply the rules for assessing data quality, below, to all estimates they produce, and retain only those that satisfy the release criteria. Estimates that do not satisfy the release criteria are not reliable.

**Introduction**

The guidelines for release and publication make use of the concept of sampling variability to determine whether the estimates obtained from the microdata are reliable. Sampling variability is the error in the estimates caused by the fact that we survey a sample rather than the entire population. The concept of standard error and the related concept of coefficient of variation and confidence interval provide an indication of the magnitude of the sampling variability.

The standard error and coefficient of variation do not measure any systematic biases in the survey data which might affect the estimate. Rather, they are based on the assumption that the sampling errors follow a normal probability distribution.

Subject to this assumption, it is possible to estimate the extent to which different samples that have the same design and the same number of observations would give different results. This indicates the margin of error that is likely to be included in the estimates derived from our single sample.

For a more complete description of the measures of sampling variability, see A. Satin and W. Shastry, Survey Sampling: A Non-Mathematical Guide, Statistics Canada, Catalogue 12-602E.

## Minimum sizes of estimates for release

Suppression rules, or data reliability cut-offs, are currently established based on the sample size that underlies the estimate. In general, a sample size of 25 observations is required for the estimate to be published. Depending on the type of estimate, this rule can vary slightly. These rules help protect the confidentiality of survey respondents and ensure the reliability of estimates.

### Suppression rules

| ESTIMATE | SUPPRESS IF: |
|---|---|
| **Percentage, Distribution, Proportion/Shares:** | |
| • % under the low-income cutoff (LICO)<br>• Income distribution<br>• Proportion of families with income=0 | Denominator* sample size < 25<br>or<br>Denominator* sample size < 100 and numerator sample size < 5 |
| **Ratios:** | |
| • female/male earnings | Numerator sample size < 25<br>or<br>Denominator sample size < 25 |
| **Quintiles (shares, means and upper income limits)** | |
| • shares of income by quintile<br>• average income by quintile<br>• upper income limits | sample in all quintiles/5 < 25<br>or<br>upper income limit for upper income quintile or total of quintiles |
| **Other estimates** | |
| • Counts<br>• Mean<br>• Medians<br>• Gini coefficients | sample < 25 |

*The denominator sample size refers to the sample size of the total estimate from which the distribution, percentage, proportion or share is derived.

## Estimates of provincial aggregates and means

When producing estimates for provincial aggregates and means it should be noted that for a small number of records, province of residence has been suppressed. This will result in a small bias in provincial estimates.

**Rounding guidelines**

Once it has been determined whether the results obtained are reliable, the level of rounding indicates the level of precision that the data can actually support. The following guidelines for rounding should be used:

- All estimates should be rounded so there are no more than three significant digits.
- Estimates of population sub-groups should be rounded to at least the nearest hundred units.
- Rates and percentages should be rounded to at least one decimal point.

Note that all calculations are to be derived from their unrounded components, and then 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. For example, in normal rounding to the nearest 100, the estimate 49,448 would be rounded down to 49,400 and an estimate of 49,252 would be rounded up to 49,300. The figure 1.78% would be rounded to 1.8%.

**Hypothesis tests provided by statistical software packages**

Microdata users should be aware that the results of hypothesis tests (such as the p values accompanying t statistics or Pearson statistics) that are provided automatically by most standard statistical software packages are incorrect for data provided by surveys with a complex survey design, such as SLID. Such packages calculate these test results under the assumption of simple random sampling. That is, they do not take into account the special sample design features of SLID such as stratification, clustering, and unequal selection probabilities. While many of the standard packages can account for the unequal selection probabilities in the production of estimates by allowing the use of weights, these packages do not properly take the sample design into account when producing variance estimates that form part of most test statistics.

To perform hypothesis tests, a two-step method can be employed with the standard statistical software to form the test statistics. First, estimate the characteristics of interest (total or mean) using the weights provided on the microdata file. Second, obtain approximate variance estimates of these characteristics by rerunning the same software procedure as that used for producing the characteristic estimates but using a scaled weight that consists of the original weight divided by the average of the original weights of all the observations being used in your computations. The standard error can be derived by using the estimate and the rough estimate of the variance. These quantities (estimate, variance, standard error) can then be combined to form test statistics. It must be noted that this method provides only rough approximations to the variance.

It should be noted that users of the SLID PUMF cannot readily obtain better design-based variance estimates through the use of statistical software specifically designed for survey data.

This is because the design information required by these software packages is not currently available on the SLID data file due to confidentiality considerations. However, better variance estimates can be produced by Statistics Canada on a cost-recovery basis.

## 6. CONFIDENTIALITY OF THE PUBLIC-USE MICRODATA

The production of a public-use microdata file includes many safeguards to prevent the identification of any one person. Longitudinal surveys are faced with an extra challenge when it comes to ensuring confidentiality, because data are collected for the same people for several years. For this reason, Statistics Canada plans to release only cross-sectional files from SLID. The number of topics covered in SLID also contributes to the amount of processing required specifically to ensure confidentiality. Confidentiality of the public-use file is ensured mainly by reducing information, i.e. deleting whole variables or suppressing or collapsing some of their detail.

SLID uses a number of techniques to ensure confidentiality:

- The SLID public-use file is comprised of a sample of the households randomly selected from the full SLID sample.

- All the variables which would permit direct identification of individuals are, of course, deleted from the file. This includes name, telephone number, and other data used for collection purposes;

- Collapsing categories. This is applied to categorical (i.e. qualitative) variables.

- Top and bottom coding. Very high and very low values usually are rare or unique in the population. Such extreme values are replaced with the value of an upper or lower limit.

- Rounding. Some variables, particularly monetary values, are rounded.

- Suppression and modification of characteristics was done while preserving integrity of the file for the purpose of producing precise and accurate statistics.

- Imputed records and variables on the file are not identified as such.

- Addition of "noise" (perturbation). Numeric values may have been raised or reduced by unequal amounts and proportions in a random-like fashion (addition of "noise"), while maintaining data integrity for the purpose of producing precise and accurate statistics.

## 7. SLID CONTENT, NOTES and DEFINITION, METHODOLOGY

See the appropriate section in Survey of Labour and Income Dynamics (SLID) - A survey overview

## 8. RELATED PRODUCTS AND SERVICES

See the appropriate section in Survey of Labour and Income Dynamics (SLID) - A survey overview

## 9. QUESTIONS AND COMMENTS

If you have any questions or comments about the data in this CD-ROM product, you can contact the Income Statistics Division.

Telephone: 1-888-297-7355 or 613-951-7355
Facsimile Number: 613-951-3012
Internet: income@statcan.ca

Income Statistics Division
Statistics Canada
Ottawa, Ontario
K1A 0T6