Public-Use Microdata File Guide

The Canadian Perspectives Survey Series 4 – Information Sources Consulted During the Pandemic

The following document is a complement to the Canadian Perspectives Survey Series User Guide for CPSS4 – Information Sources Consulted During the Pandemic. It presents the differences between the master file and the Public-Use Microdata Files (PUMF) as well as a short guide to proper analysis, release and interpretation of the data provided in the PUMF. For more details regarding the CPSS including the target population, sample design or weighting design, please refer to the user guide *CPSS4 Analytical Guide*.

1. Documentation

The CPSS4 – Information Sources Consulted During the Pandemic questionnaire serves as reference for the PUMF data file. The questionnaire will display additional questions that are not represented in the PUMF due to suppressions and other restriction methods that have been applied to the data.

2. Conversion of the Master File to Public-Use Microdata File

The approach for creating a PUMF is intended to balance the requirements for maintaining respondent confidentiality by minimizing disclosure risks, while providing the most useful data to users. Some differences between the CPSS4 – Information Sources Consulted During the Pandemic master file and PUMF include the removal of province from the file and the removal of some labour related variables (see Appendix A). Response categories have also been collapsed for other variables (see Appendix B). Also, bootstrap weights (used in the calculation of variances) are not provided with the PUMF. Users should apply the factors provided in Table 1 of Section 3. These factors will provide approximations for the precision of estimates. **Where differences are noted between values from the master file and those from the PUMF, values from the master file are considered the authoritative source.**

3. Guidelines for statistical analysis

Survey Weights: The CPSS4 is based upon a complex sample design, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents. The sample design was not self-weighted. When producing simple estimates including the production of ordinary statistical tables, users **must** apply the proper survey weights. If proper weights are not used, the estimates derived from the PUMF cannot be considered to be representative of the survey population.

Variance estimation for the PUMF: Variances produced by statistical packages often rely on simple formula that do not take into account the complexity of the sample design. These formula will underestimate the true variances of estimates in the CPSS4. The survey therefore used a resampling method called the bootstrap. For confidentiality reasons, the bootstrap weights are not provided with the PUMF, but were used to create adjustment factors that can be applied to estimates of variance, standard errors or coefficient of variations as calculated by a statistical package. The adjustment factors found in Table 1 should be multiplied by the variance/standard error/coefficient of variation as produced by the statistical software to account for the survey's complex design.

Table 1. Adjustment Factors to be applied to Measure of Precision

Age Category	Adjustment Factor (also called the Design Effect) for a **Variance**	Adjustment Factor for a **Standard Error**
More than one age group combined	3.5	1.8708
15-24	2.5	1.5811
25-34	2.5	1.5811
35-44	2.5	1.5811
45-54	2.5	1.5811
55-64	2.5	1.5811
65-74	3.0	1.7321
75+	3.0	1.7321

The following example was drawn from the CPSS4. For example, let's say we want to know the proportion of people who said their mental health was excellent or very good. The first step would be to create a subset of the PUMF file keeping only respondents who have answered the MH_30 question and create a binary variable:

$$\text{mental\_health} = \begin{cases} 1 & \text{if MH\_30} \in \{1,2\} \text{ (i.e. excellent and very good mental health)} \\ 0 & \text{if MH\_30} \in \{3,4,5\} \text{ (i.e. good, fair and poor mental health)} \end{cases}$$

For this example, assume the name of this subset is pumf_health. To calculate the average proportion of people having a very good or excellent mental health, for all ages together, a proc means using the weights can be performed in SAS.

SAS code example:

```
proc means data=pumf_health mean stderr;
      var mental_health;
      weight PERS_WGT;
run;
```

We find that 0.5534 of the population have a very good or excellent mental health (this could also be expressed as 55% have a very good or excellent mental health).

The standard error (S.E.) obtained by the procedure for this estimate is 0.0076601. However, this does not account for the survey's design and is an underestimate of the true standard error.

To estimate an appropriate standard error, the adjustment factor for standard error from Table 1, needs to be applied:

$$\text{S. E.} = \underbrace{0.0076601}_{\text{from procedure above}} \cdot \underbrace{1.8708}_{\text{from Table 1}}$$

$$= 0.0143$$

Once the S.E. is calculated, other measures such as the coefficient of variation (CV) and the confidence interval (CI) can be calculated.

The coefficient of variation is calculated as:

$$\text{CV} = \frac{\text{S.E.}}{\text{mean}} = \frac{0.0143}{0.5534} = 0.02584$$

And to get the 95% confidence interval:

$$\begin{aligned}\text{C.I.} &= [p - 1.96(\text{S.E.}), p + 1.96(\text{S.E.})] \\ &= [0.5534 - 1.96(0.0143), 0.5534 + 1.96(0.0143)] \\ &= [0.5254, 0.5814]\end{aligned}$$

This confidence interval could also be expressed as [52.5%, 58.1%]

Rounding guidelines: Users are urged to adhere to the following guidelines regarding the rounding of such estimates:

a) Estimates in the main body of a statistical table are to be rounded to the nearest hundred units 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, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last digits are between 50 and 99 they are changed to 00 and the preceding digit is incremented by 1.

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 to the nearest 100 units using normal rounding.

c) Averages, rates and percentages are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves to one decimal using normal rounding. In normal rounding to a single digit, if the final 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 increased by 1. Proportions and ratios are to be computed from unrounded components and then are to be rounded themselves to three decimals using normal rounding.

d) Sums and differences of aggregates are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units (or the nearest one decimal). Sums and differences of percentages (or ratios) are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest one decimal (or three decimals) using normal rounding.

e) 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).

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

Quality Guidelines for Estimates: Before releasing and/or publishing any estimates, users should consider the quality level of the estimate. This section covers quality in terms of sampling error. There are different ways of measuring and reporting sampling error. It is considered a best practice at Statistics Canada to report the sampling error of an estimate through its 95% confidence interval. 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 quality categories:

> **Category A:** Estimates can be released with no warning. Data users should use the 95% confidence interval to decide whether the quality of the estimate is sufficient.
>
> **Category E – Marginal Quality:** Estimates and confidence intervals are deemed of marginal quality. Estimates and confidence intervals should be flagged with the letter E (or some similar identifier) and be accompanied by a warning to use the estimate with caution.
>
> **Category F – Poor Quality:** Estimates and confidence intervals are deemed of poor quality, and are not recommended for release. The estimates contain a very high level of instability, making them unreliable and potentially misleading. If users insist on releasing estimates of poor quality, even after being advised of their accuracy, the estimates should be accompanied by a disclaimer. The user should acknowledge the warnings given and undertake not to disseminate, present or report the estimates,

directly or indirectly, without this disclaimer. They 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."

Table 2 Sample size requirements for Category of Quality

Type of Statistic	Category F	Category E	Category A
Proportion	$n < 90$	$90 \leq n \leq 180$	$n > 180$
Weighted count	$m < 90$	$90 \leq m \leq 180$	$m > 180$

Notation used in Table 2:

n: Domain sample size. For proportions, this is the unweighted count of the number of respondents included in the denominator of the proportion.

m: Unweighted count of the number of respondents with nonzero values that contribute to the estimate.

5. Additional notes for analysis

Labour Variables

PEMPSTC

- The concepts of employed and absent from work in the CPSS-4 are not equivalent to Labour Force Survey (LFS) concepts, and direct comparisons should not be made. In the CPSS-4, "employed" (PEMPSTC response categories 1 through 3) includes all persons who worked or were absent for any reason, including temporary lay-offs. Those who have been temporarily laid off are not treated as absent or employed in the LFS.
- Given the rapidly evolving situation at the time of the survey (e.g. public health directives, government orders), the reference week of the employment status questions (July 12 to July 18, 2020) should be clearly stated when reporting results. Appropriate context should be provided about the nature and extent of restrictions in place at that time.

Appendix A

The following is the list of variables removed from the file in the creation of the PUMF**.**

Variables removed

Variable	Description
Province	Province of residence
Region	Region of residence
CSizeMiz	Community Size and Metropolitan Influence Zones.
LM_05	During that week, did you work at a job or business?
LM_10	During that week, did you have a job or business from which you were absent?
LM_15	What was the main reason you were absent from work that week?
TOTHHSIZ	Total household size

Appendix B

Recoded or Capped Variables

Variable Description	Recoded Label	
Age of respondent (original: Curr_Age) (recoded: AgeGrp)	1 2 3 4 5 6 7	15 to 24 years old 25 to 34 years old 35 to 44 years old 45 to 54 years old 55 to 64 years old 65 to 74 years old 75 years old and over
Household size (original: HHLDSIZE) (recoded: HHLDSIZC)	1 2 3 4 5	1 2 3 4 5 and more
Education level (original: EDUC_LVL) (recoded: PEDUC_LC)	1 2 3 4 5 6 7	Less than high school diploma or its equivalent High school diploma or a high school equivalency certificate Trade certificate or diploma College/Cegep/other non-university certificate or diploma University certificate or diploma below the bachelor's level Bachelor's degree University certificate/diploma/degree above the BA level
Type of dwelling (original: DWELCODE) (recoded: DWELCODC)	1 2 3 4	Single detached house Low-rise apartment less than 5 stories High-rise apartment 5 or more stories Other
Marital status (original: MARSTAT) (recoded: MARSTATC)	1 2 3 4	Married Living Common Law Widowed/Separated/Divorced Single/never married
Immigration status (original: IMMIGRNT) (recoded: IMMIGRNC)	1 2	Born in Canada Landed and not a landed immigrant
Employment status (original: EMPST) (recoded: PEMPSTC)	1 2 3 4 9	Employed and at work at least part of the reference week Employed but absent work for reasons not related to COVID-19 Employed but absent from work due to COVID-19 Not employed Not stated
Main source of information to find out about COVID-19	1 2 3	News outlets Federal health agency Provincial or territorial health agency

(original: BH_05) (recoded: BH_05C)	4 Municipal health agency 5 Federal daily announcements 6 Provincial daily announcements 7 Social media 8 Family, friends or colleagues 9 Health professionals 10 Place of employment 11 Other 12 I do not look for information about this 99 Not stated
Reason accuracy not validated – Did not know how to check/too difficult (original: FC_20C, FC_20E) (recoded: FC_20CE)	Reason accuracy not validated: 1 Reason accuracy not validated – Did not know how to check 2 Reason accuracy not validated – Knew how to check 6 Valid skip 9 Not stated