Public-Use Microdata File Guide

The Canadian Perspectives Survey Series 6
Substance Use and Stigma During the Pandemic

The following document is a complement to the Canadian Perspectives Survey Series User Guide for CPSS6 – Substance Use and Stigma 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 *CPSS6 Analytical Guide*.

1. Documentation

The CPSS6 – Substance Use and Stigma 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 CPSS6 – Substance Use and Stigma 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), or variables can be combined (see Appendix C). 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 CPSS6 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 CPSS6. 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 for a **Variance** (also called the Design Effect)	Adjustment Factor for a **Standard Error**
More than one age group combined	3.5	1.8708
15-24	2.5	1.5811
25-34	3.0	1.7321
35-44	2.5	1.5811
45-54	2.5	1.5811
55-64	3.0	1.7321
65-74	4.0	2.0000
75+	2.5	1.5811

The following example was drawn from the CPSS6. For example, let's say we want to know the proportion of people who do not have not downloaded the COVID Alert app yet. The first step would be to create a subset of the PUMF file keeping only respondents who have answered the PCOV_05A question and create a binary variable:

$$\text{no\_COV\_app} = \begin{cases} 1 & \text{if PCOV\_05A} \in \{2,3,4\} \quad \text{(i.e. has not downloaded the COVID Alert app)} \\ 0 & \text{if PCOV\_05A} \in \{1\} \quad \text{(i.e. has downloaded the COVID Alert app)} \end{cases}$$

For this example, assume the name of this subset is pumf_COVapp. To calculate the average proportion of people who have not downloaded the app, for all ages together, a proc means using the weights can be performed in SAS.

SAS code example:

```
proc means data=pumf_COVapp mean stderr;
      var no_COV_app;
      weight PERS_WGT;
run;
```

We find that 0.6655 (or 66.55%) of the population has not downloaded the COVID Alert app.

The standard error (S.E.) obtained by the procedure for this estimate is 0.0075590. 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:

$$\text{S. E.} = \underbrace{0.0075590}_{\text{from procedure above}} \cdot \underbrace{1.8708}_{\text{from Table 1}}$$

$$= 0.0141$$

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:

$$\text{CV} = \frac{\text{S.E.}}{\text{mean}} = \frac{0.0141}{0.6655} = 0.0212$$

And to get the 95% confidence interval:

$$\text{C.I.} = [p - 1.96(\text{S.E.}) \, , p + 1.96(\text{S.E.})]$$

$$= [0.6655 - 1.96(0.0141), \; 0.6655 + 1.96(0.0141)]$$

$$= [0.6378 \, , 0.6932]$$

This confidence interval could also be expressed as [63.78%, 69.32%]

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.

Appendix A – Removed Variables

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.
TOTHHSIZ	Total household size
LM_05	Work at a job or business
LM_10	Absent from job or business
LM_15	Main reason absent that week
GDR_05	Sex at birth
GDR_10	Gender
HS_05	Hospitalized overnight or longer for problems with emotions, mental health, alcohol, drugs

Appendix B – Recoded or Capped Variables

Variable Description	Recoded Label	
Age of respondent (original: Curr_Age) (recoded: AGEGRP)	1 15 to 24 years old 2 25 to 34 years old 3 35 to 44 years old 4 45 to 54 years old 5 55 to 64 years old 6 65 to 74 years old 7 75 years old and over	
Household size (original: HHLDSIZE) (recoded: PHHLDSZC)	1 1 2 2 3 3 4 4 5 5 and more	
Education level (original: EDUC_LVL) (recoded: PEDUC_LC)	1 Less than high school diploma or its equivalent 2 High school diploma or a high school equivalency certificate 3 Trade certificate or diploma 4 College/Cegep/other non-university certificate or diploma 5 University certificate or diploma below the bachelor's level 6 Bachelor's degree 7 University certificate/diploma/degree above the BA level	
Type of dwelling (original: DWELCODE) (recoded: DWELCODC)	1 Single detached house 2 Low-rise apartment less than 5 stories 3 High-rise apartment 5 or more stories 4 Other	
Marital status (original: MARSTAT) (recoded: PMARSTC)	1 Married 2 Living Common Law 3 Widowed/Separated/Divorced 4 Single/never married	
Immigration status (original: IMMIGRNT) (recoded: IMMIGRNC)	1 Born in Canada 2 Landed and not a landed immigrant	
Employment status (original: EMPST) (recoded: PEMPSTC)	1 Employed and at work at least part of the reference week 2 Employed but absent work for reasons not related to COVID-19 3 Employed but absent from work due to COVID-19 4 Not employed 9 Not stated	Increased or about th Decreased Not applicable Not stated
Past 30 days: Number of drinks (original: ALC_10) (recoded: ALC_10C)	1 1 to 4 2 5 to 9 3 10 to 19 4 20 or more	

	96	Valid Skip
	99	Not Stated
Past 30 days: Frequency of 5 or more drinks on one occasion (original: ALC_15) (recoded: ALC_15C)	1	4 times a week or more
	2	2 or 3 times a week
	3	Once a week
	4	2 to 3 times in the past 30 days
	5	Once in the past 30 days
	6	Not in the past 30 days
	96	Valid Skip
	99	Not Stated
Embarrassing to tell friends/family about my A/D use (original: ST_10A) (recoded: ST_10AC)	1	Agree or Strongly Agree
	2	Disagree
	3	Strongly Disagree
	4	Not Applicable
	6	Valid Skip
	9	Not Stated
Embarrassing to seek help/treatment for my A/D use (original: ST_10B) (recoded: ST_10BC)	1	Agree or Strongly Agree
	2	Disagree
	3	Strongly Disagree
	4	Not Applicable
	6	Valid Skip
	9	Not Stated
Difficulties in obtaining treatment for my A/D use (original: ST_10C) (recoded: ST_10CC)	1	Agree or Strongly Agree
	2	Disagree or Strongly Disagree
	4	Not Applicable
	6	Valid Skip
	9	Not Stated
Discriminated against by health professionals because of A/D use (original: ST_10D) (recoded: ST_10DC)	1	Agree or Strongly Agree
	2	Disagree
	3	Strongly Disagree
	4	Not Applicable
	6	Valid Skip
	9	Not Stated
Often feel alone because of A/D use (original: ST_10E) (recoded: ST_10EC)	1	Agree or Strongly Agree
	2	Disagree
	3	Strongly Disagree
	4	Not Applicable
	6	Valid Skip
	9	Not Stated
Scared how people will react if they find out about my A/D use (original: ST_10F)	1	Agree or Strongly Agree
	2	Disagree
	3	Strongly Disagree

(recoded: ST_10FC)	4 6 9	Not Applicable Valid Skip Not Stated
People have insulted me because of my A/D use (original: ST_10G) (recoded: ST_10GC)	1 2 3 4 6 9	Agree or Strongly Agree Disagree Strongly Disagree Not Applicable Valid Skip Not Stated
People have avoided me because of my A/D use (original: ST_10H) (recoded: ST_10HC)	1 2 3 4 6 9	Agree or Strongly Agree Disagree Strongly Disagree Not Applicable Valid Skip Not Stated
Need to hide my problems with A/D from my friends/family (original: ST_10I) (recoded: ST_10IC)	1 2 3 4 6 9	Agree or Strongly Agree Disagree Strongly Disagree Not Applicable Valid Skip Not Stated

Appendix C – Combined Variables

The values of all combined variables on the master file and on the PUMF are:

Value	Definition
1	Yes
2	No
6	Valid Skip
9	Not Stated

The combined variables on the PUMF are:

Description	Variables on the Master File	Combined Variable on the PUMF
Reasons for decreased use of cannabis	**CAN_25A** (Cost), **CAN_25G** (Other reason), **CAN_25H** (No particular reason)	**CAN25AGH**
	CAN_25B (Personal responsibilities), **CAN_25F** (Lack of privacy to use cannabis at home)	**CAN_25BF**
	CAN_25C (Personal choice), **CAN_25D** (Improved health or pain)	**CAN_25CD**
Reasons for increased use of opioids	**OPI_15B** (Boredom), **OPI_15C** (Loneliness), **OPI_15F** (Other reason), **OPI_15G** (No particular reason)	**OP15BCFG**
	OPI_15A (Stress), **OPI_15D** (Convenience)	**OPI_15AD**
Reasons for decreased use of opioids	**OPI_20C** (Personal choice), **OPI_20E** (Inconvenience), **OPI_20G** (Other reason)	**OPI20CEG**
Reasons for increased use of non-prescribed drugs	**NPD_15F** (Other reason), **NPD_15G** (No particular reason)	**NPD_15FG**
Reasons for decreased use of non-prescribed drugs	**NPD_20A** (Cost), **NPD_20B** (Personal responsibilities), **NPD_20C** (Personal choice), **NPD_20D** (Improved health or pain)	**NP20ABCD**
	NPD_20E (Decreased opportunities for socialization), **NPD_20F** (Inconvenience), **NPD_20G** (Concerned about quality of drugs), **NPD_20I** (No particular reason)	**NP20EFGI**
Drugs used for managing pain – if pain was managed by prescription drugs	**PM_05BC** (Low-dose codeine), **PM_05BG** (Anticonvulsants or anti-seizure meds), **PM_05BI** (Other prescription drugs)	**PM05BCGI**
Drugs used for managing pain – if pain was managed by non-prescription drugs	**PM_05CC** (Low-dose codeine), **PM_05CD** (Opioids), **PM_05CF** (Illegal drugs), **PM_05CH** (Other non-prescription drugs)	**PM5CCDFH**

Drugs used for managing pain – if pain was managed by both prescription and non-prescription drugs	**PM_05DC** (Low-dose codeine), **PM_05DF** (Illegal drugs), **PM_05DH** (Anticonvulsants or anti-seizure meds), **PM_05DJ** (Other presc. and non-presc. drugs)	**PM5DCFHJ**