Microdata User Guide

2019 Canadian Alcohol and Drugs Survey

June to December 2019

Statistics Canada Statistique Canada

Canada

Contents

1.0 Introduction

The Canadian Alcohol and Drugs Survey (CADS) was conducted by Statistics Canada from June to December 2019 with the cooperation and support of Health Canada. This manual has been produced to facilitate the manipulation of the microdata files (master file and PUMF - Public Use Microdata File) of the survey results.

Any questions about the data sets or their 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
Fax: 613-951-4527
E-mail: statcan.csdidclientservice-ciddsservicealaclientele.statcan@canada.ca

Health Canada

Controlled Substances Directorate
Controlled Substances and Cannabis Branch
Ottawa, ON K1A 0K9
E-mail: hc.odss-bssd.sc@canada.ca

2.0 Background

From 1999 until 2012, data on tobacco use were collected annually as part of the Canadian Tobacco Use Monitoring Survey (CTUMS). In 2013, 2015 and 2017, Statistics Canada conducted the Canadian Tobacco, Alcohol and Drug Survey (CTADS). This survey collected data not only on tobacco, but also on alcohol and drugs. For the first time in 2019, the Canadian Alcohol and Drug Survey (CADS) was conducted, collecting data primarily on alcohol and drugs. In 2019, the Canadian Tobacco and Nicotine Survey (CTNS) was conducted to collect data on tobacco and nicotine.

3.0 Objectives

The main objective of this survey is to collect information on Canadians' use of alcohol and drugs. Health Canada and other organizations will use the information to monitor changes in alcohol and drug use.

The other objectives of the Canadian Alcohol and Drugs Survey (CADS) are the following: to measure frequency of alcohol use, to measure frequency of cannabis use, to measure frequency of use of other drugs and to measure potential harmful effects of using alcohol, cannabis and other drugs.

The CADS is the only Statistics Canada survey that meets Health Canada's need for continuous and detailed information on drug consumption and drinking prevalence by province, sex or age group, for age groups 15 to 19, 20 to 24 and 25 and over.

4.0 Survey Methodology

The Canadian Alcohol and Drugs Survey (CADS) was administered between June 10 and December 31, 2019 by electronic questionnaire, and in the case of non-response, by follow-up over the phone.

4.1 Target population and sampling frame

The target population for the CADS was all persons 15 years of age and over living in Canada with the exception of the following persons:

1) residents of the Yukon, Northwest Territories and Nunavut;
2) full-time residents of institutions; and
3) residents of Native reserves

The survey used the Dwelling Universe File (DUF), a file produced at Statistics Canada, as the sampling frame. This was done in order to produce quality estimates at the provincial level and for the different age groups (at the Canada level), and to facilitate an initial contact by mail for the invitation to complete the questionnaire electronically. This sampling frame allows for up to three telephone numbers to allow for telephone follow-up with a household, including landline and cellular telephone numbers. A sample cleaning process to eliminate telephone numbers that were not in service or unknown was conducted prior to sending the sample to the collection team.

Since the survey was conducted using a sample of addresses, almost all households could be contacted by mail. Dwellings that were identified as vacant at the time the sampling frame was created were excluded. Dwellings that had neither a mailing address nor an associated telephone number were also excluded from the sample frame, as they could not be contacted by any of the survey collection modes. However, the survey estimates were weighted to include persons living in these dwellings.

4.2 Sample Design and Allocation

The sample design for the 2019 CADS was a stratified two-phase random sample. The provinces formed the strata. In the first phase, households were selected randomly, and in the second phase, one person was selected from within the household using the age-order selection method. The selection algorithm was based on the number of eligible members in the household and the ordered age of each member.

A letter was sent to the selected household and a household member was selected, via the instructions provided in the letter, to complete the electronic questionnaire. The selected person was invited to complete the questionnaire by accessing it online and entering a secure access code (SAC) provided in the letter.

Age-order selection was also used for CATI respondents (computer-assisted telephone interviews). Selection was done with the interviewer. The instructions in the letter, as well as the selection made with the interviewer, were consistent for the same sampled household to ensure that the same person was selected to participate in the survey for a given household, regardless of the collection mode used to complete the questionnaire.

Kish allocation was the method used to allocate the sample in order to meet quality targets for various domains of interest. The initial sample size was determined by assuming an overall response rate of 50% and a design effect of 1.5. It was determined that a sample size of 22,000 households was required to produce quality estimates at the provincial level. The sample allocation can be found in Section *7.0 Data Quality.*

4.3 Weighting

The principle behind estimation from a probability sample is that each person in the sample "represents", besides himself or herself, several other persons not in the sample. For example, in a 2% simple random sample of the population, each person in the sample represents 50 persons in the population.

The weighting phase is a step which calculates this number (or weight) for each record. This weight, which appears on the microdata files, must be used to derive meaningful estimates from the survey data. For example, if the number of people in Canada who drink alcohol daily is to be estimated, it is done by selecting the records referring to those individuals in the sample with that characteristic (ALC_Q15 = 1) and summing the weights associated to those records. Details of the method used to calculate these weights are presented in *Chapter 10.0 - Weighting*.

5.0 Data Collection

5.1 Questionnaire Design

The questionnaire for CADS 2019 uses many questions from the Canadian Tobacco, Alcohol and Drugs Survey (CTADS) 2017 questionnaire and its previous versions. The questionnaire also contains a high percentage of new questions. The sections sex, gender and age (AGS), maternal experiences with cannabis and alcohol (MEX), Spice (SPI), Kratom (KRT), Mephedrone (MEP), BZP or TFMPP (BZP), injectable drug use (IDU), overdose (OD) and treatment (TT) are new. Many of the already existing sections from CTADS were also modified by changing the question order, adding additional questions, separating questions into more than one question etc.

CADS 2019 was the first time this survey used a self-completed electronic questionnaire, as opposed to only using computer-assisted telephone interviews (CATI).

Specifications defining valid limits and ensuring consistency across questions have been incorporated into the electronic questionnaire application to the extent possible. Additional consistency edits were done during the data processing phase.

5.2 Data Collection and Editing

The data collection was conducted from June through December 2019. Collection was divided into 2 waves. The first wave took place from June 10 to September 22 and the second wave from September 23 to December 31. For each wave, an introductory letter was sent, followed by up to 4 reminder letters for non-responding households.

Valid skip patterns and validation messages appeared throughout the electronic questionnaire. Checks built into the application ensured consistency of responses, identified and corrected outliers and determined who was asked certain questions. As a result, by the end of the collection process, the data were already fairly "clean".

All nonresponding cases after the 4 reminder letters were distributed to two Statistics Canada regional offices for telephone follow-up (CATI). The workload and interviewers in each office were overseen by a project manager. The automatic scheduler used in the CATI system ensured that cases were randomly assigned to interviewers and that calls were made at different times of day on different days of the week to maximize the probability of contact.

6.0 Data Processing

In the past, in the case of the Canadian Tobacco, Alcohol and Drug Survey (CTADS), the main outputs were two "clean" microdata master files, one for the household level information and one for the person level information, as well as an equivalent set of public use microdata files (PUMF). Now, with CADS, only one master file and PUMF are created. This chapter presents a brief summary of the processing steps involved in producing these files.

6.1 Data capture

As the data was collected using an electronic questionnaire and a telephone follow-up (CATI-computer-assisted telephone interviewing), there was no need for a separate data capture system.

6.2 Editing

Some inconsistencies in the responses provided by respondents in the survey have been corrected. At times, respondents provided an information and later in the questionnaire they mentioned the opposite. In collaboration with the survey client, a series of specifications were written to address these inconsistencies.

6.3 Creation of Derived Variables

A number of variables included in the microdata files were calculated by combining data from several questions or within questions to facilitate data analysis. Examples of derived variables include alcohol use status and cannabis use status. The rural/urban characteristic of the community where the respondent lives (DVURBAN) was derived from the postal code.

6.4 Suppression of Confidential Information for the PUMF

The Public Use Microdata File (PUMF) for the CADS differs from the survey “master” files held by Statistics Canada as a result of actions taken to protect the anonymity of individual survey respondents. Since the PUMF is publically available free-of-charge to a wide range of users, it is essential that additional steps are taken to ensure that the respondent data released in a PUMF is safe.

These additional steps include limiting the amount of family and household information on the PUMF, aggregating codes and capping certain variables, or suppressing or perturbing responses for certain respondents. Users requiring access to information excluded from the PUMF may either access the master file through the Research Data Centre Program at Statistics Canada or they may purchase custom tabulations. Estimates generated from custom tabulations will be released to the user, subject to meeting the guidelines for analysis and release outlined in *Chapter 9.0* of this document.

All variables on the master file were looked at in terms of risk of residual disclosure. As a result of the analysis and the remedial actions taken, the PUMF contains 10,293 respondent records and 387 variables. More information can be found in the PUMF data dictionary.

7.0 Data Quality

7.1 Response Rate

For the Canadian Alcohol and Drugs Survey (CADS), the overall response rate is calculated as follows.

The Response Rate is the proportion of records of selected persons, within the scope of the survey, with valid data. Since only one person is selected per household, the household response rate is the same as the person response rate. Therefore, only the term "response rate" is being used. The proportion of responding units is adjusted by a factor estimating the proportion of in-scope households. This adjustment factor is obtained by dividing the number of households in the sampling frame by the number of households according to population projections.

$$\frac{number\ of\ people\ with\ valid\ data}{number\ of\ selected\ households} * adjustment\ factor$$

A **person respondent** (with valid data) has the following characteristics:

- The person completed the age-order selection
- The person answered questions on age, household size and household composition
- The selected person answered the two key questions on alcohol consumption.

Table 1: Response Rate by Province

Province	Total Persons Selected	Total Persons Responding	Adjustment factor for estimating the number of in-scope households	Response Rate (%)
Newfoundland and Labrador	1,935	744	1.28	49.2
Prince Edward Island	1,915	827	1.24	53.3
Nova Scotia	1,923	891	1.17	54.4
New Brunswick	1,910	903	1.17	55.2
Quebec	2,786	1,432	1.12	57.6
Ontario	3,488	1,638	1.05	49.4
Manitoba	1,940	954	1.09	53.5
Saskatchewan	1,938	905	1.11	51.9
Alberta	2,096	983	1.07	50.4
British Columbia	2,185	1,016	1.09	50.8
Total	**22,116**	**10,293**	**1.09**	**50.7**

Note: The adjustment factors in the table above are rounded and the response rates presented can therefore not be reproduced

7.2 Survey Errors

The estimates derived from this survey are based on a sample of households. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used in the survey. The difference between the estimates obtained from the sample and those resulting from a complete count taken under similar conditions is called the sampling error of the estimate.

Errors which are not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered in the electronic questionnaire application 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. Considerable time and effort were made to reduce non-sampling errors in the survey. Quality assurance measures were implemented at each step of the data collection and processing cycle to monitor the quality of the data. These measures included testing of the electronic questionnaire, extensive training of interviewers with respect to the survey procedures and computer-assisted telephone interviewing (CATI) application, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions and testing the application to ensure that range checks, edits and question flow were all programmed correctly.

7.3 Total Non-response

Total non-response can be a major source of non-sampling error in many surveys, depending on the degree to which respondents and non-respondents differ with respect to the characteristics of interest. Total non-response occurred because the interviewer was either unable to contact the respondent or the respondent refused to participate in the survey. Total non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond.

7.4 Partial Non-response

In most cases, partial non-response to the survey occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information. Partial non-response is indicated by codes on the microdata file (i.e. Non-response.)

7.5 Coverage

As mentioned in *Section 4.1 (Target population and sampling frame)*, some households in Canada do not have valid addresses, home telephone numbers, or cellular phone numbers on the survey frame. Individuals living in these households may have unique characteristics which will not be reflected in the survey estimates. Users should be cautious when analyzing subgroups of the population which have characteristics that may be correlated with not having a valid address or either a cellular or landline telephone.

7.6 Measurement of Sampling Error

Since it is unavoidable that estimates from a sample survey are subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error. This section of the document outlines the measures of sampling error which Statistics Canada commonly uses and which it urges users producing estimates from its microdata files to also use.

The basis for measuring the potential size of sampling errors is the standard error of the estimates derived from survey results.

However, because of the wide variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed in terms of the estimate to which it relates. One of the resulting measures often used is called the coefficient of variation (CV) of an estimate; it is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate. This measure of quality has been used in previous iterations of alcohol, drug and tobacco surveys. That being said, since very small proportions are sometimes measured in the CADS, it is preferable to express the quality of the estimates by presenting their confidence intervals. Indeed, a small proportion will have a high CV by construction. On the other hand, its complementary proportion (1-p) will have a small CV per construction. A more appropriate measure of quality in this case is to observe the confidence interval. This also leaves it up to users to determine whether the estimate presented is accurate enough for their needs.

We recommend using the modified Wilson interval, the modified Clopper-Pearson interval or the logit interval confidence intervals for binomial proportions (1/0; yes/no; etc.).

Most statistical software, such as SAS or SUDAAN, can produce these types of intervals. For more information on how they can be computed, please refer to Appendix A.

8.0 Guidelines for Tabulation, Analysis and Release

This chapter of the documentation outlines the 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 Statistics Canada's microdata master file 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

In order that estimates for publication or other release derived from the microdata master file correspond to those produced by Statistics Canada, 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 two 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 hundred units using normal rounding.

c) Averages, proportions, 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.

d) Sums and differences of aggregates (or ratios) are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest hundred units (or the nearest one decimal) 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.

8.2 Sample Weighting Guidelines for Tabulation

The sample design used for the Canadian Alcohol and Drug Survey (CADS) was not self-weighting. When producing simple estimates including the production of ordinary statistical tables, users must apply the survey weight.

If 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 using the master microdata file.

Users should also note that some software packages may not allow the generation of estimates that exactly match those available from Statistics Canada, because of their treatment of the weight field.

8.3 Definitions of Types of Estimates: Categorical and Quantitative

Before discussing how the CADS data can be tabulated and analyzed, it is useful to describe the two main types of point estimates of population characteristics which can be generated from the microdata files for the CADS.

8.3.1 Categorical Estimates

Categorical estimates are estimates of the number, or percentage of the surveyed population possessing certain characteristics or falling into some defined category. The number of people who have ever had a drink or the proportion of stimulant users using prescribed stimulants are examples of such estimates. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Examples of Categorical Questions:

Q: Have you **ever** had a drink?
R: Yes / No

Q: During the past 12 months, were all the stimulants you have used **prescribed to you**?
R: Yes, they all were prescribed/ Some were prescribed and others were not/ No, none were prescribed

8.3.2 Quantitative Estimates

Quantitative estimates are estimates of totals or of means, medians and other measures of central tendency of quantities based upon some or all of the members of the surveyed population. They also specifically involve estimates of the form $\hat{X}/\hat{Y}$ where $\hat{X}$ is an estimate of surveyed population quantity total and $\hat{Y}$ is an estimate of the number of persons in the surveyed population contributing to that total quantity.

An example of a quantitative estimate is the average number of drinks consumed in the past 7 days, per person. The numerator $(\hat{X})$ is an estimate of the total number of drinks consumed in the past 7 days, and its denominator $(\hat{Y})$ is the number of persons who reported consumption of at least one drink in the past 7 days.

Examples of Quantitative Questions:

Q: During the past 7 days, how many drinks did you have each day? (questions for each of the past 7 days)
R: |_|_| drinks

Q: How old were you when you first tried amphetamines or methamphetamine?
R: |_|_|_| years old

8.3.3 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata files by summing the final weights of all records possessing the characteristic(s) of interest. Proportions and ratios of the form $\hat{X} / \hat{Y}$ are obtained by:

a) summing the final weights of records having the characteristic of interest for the numerator $\left(\hat{X}\right)$,

b) summing the final weights of all records for the denominator $\left(\hat{Y}\right)$, then

c) dividing estimate a) by estimate b) $\left(\hat{X} / \hat{Y}\right)$.

8.3.4 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata files by multiplying the value of the variable of interest by the final weight for each record, then summing this quantity over all records of interest. For example, to obtain an estimate of the total number of drinks consumed in the past 7 days, multiply the value reported in question ALC_Q70 (number of drinks consumed each day) by the final weight for the record, then sum this value over all records with ALC_Q70 < 96 (all respondents who reported a value in this field).

To obtain a weighted average of the form $\hat{X} / \hat{Y}$, the numerator $\left(\hat{X}\right)$ is calculated as for a quantitative estimate and the denominator $\left(\hat{Y}\right)$ is calculated as for a categorical estimate. For example, to estimate the average number of drinks consumed in the past 7 days,

a) estimate the total number of drinks consumed in the past 7 days $\left(\hat{X}\right)$ as described above,

b) estimate the number of people $\left(\hat{Y}\right)$ in this category by summing the final weights of all records with ALC_70 < 96, then

c) divide estimate a) by estimate b) $\left(\hat{X} / \hat{Y}\right)$.

8.4 Guidelines for Statistical Analysis

The CADS is based upon a complex sample design, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used. Survey weights must be used when computing survey estimates and doing analyses.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures may differ from that which is appropriate in a sample survey framework without use of the bootstrap weights. In many cases the estimates produced by the packages are correct, but if the variances are not based on the bootstrap weights then the variances calculated are poor.

For complex analysis techniques (for example linear regression, logistic regression and analysis of variance), a method exists which can make the variances calculated by the standard packages more meaningful (if not using the bootstrap weights), by incorporating the unequal probabilities of selection. The method rescales the weights so that there is an average weight of 1.

For example, suppose that an analysis of all male respondents is required. The steps to rescale the weights are as follows:

1. select all respondents from the file who reported SEX = male;
2. calculate the AVERAGE weight for these records by summing the original person weights from the microdata file for these records and then dividing by the number of respondents who reported SEX = male;

3. for each of these respondents, calculate a RESCALED weight equal to the original person weight divided by the AVERAGE weight;
4. perform the analysis for these respondents using the RESCALED weight.

However, because the stratification of the sample's design is still not taken into account, the variance estimates calculated in this way are likely to be under-estimated.

Wherever possible, users should use the bootstrap weights in analyses in order to correctly estimate the variances. If using a statistical package that allows analysis with the bootstrap weights, the user should apply the bootstrap weights and not re-scale. For more details on the use of bootstrap weights in calculating the sampling error used in CVs, variances, and confidence intervals, please see *Appendix A.*

The parameters for cycle 2019 of CADS are:

For the MASTER file:

- Data file: CADS2019ECAD.txt
- Bootstrap weight file: CADS2019ECAD_BSW.txt
- Identification variable: MASTERID
- Survey Weight: WEIGHT
- Number of bootstrap replicates (B): 1000
- Replicate Weights: wrmp0001 to wrmp1000

For the PUMF:

- Data file: CADS2019ECAD_P.txt
- Bootstrap weight file: CADS2019ECAD_P_BSW.txt
- Identification variable: PUMFID
- Survey Weight: WEIGHTP
- Number of bootstrap replicates (B): 1000
- Replicate Weights: wrpp0001 to wrpp1000

8.5 Release Guidelines

8.5.1 Release Guidelines based on Quality

Before releasing and/or publishing any estimates from the CADS, users should consider the quality level of the estimate. Data quality is affected by both sampling and non-sampling errors as discussed in *Chapter 7.0*. 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. For example,
"The user is advised that the estimates and confidence intervals flagged with the letter E are considered to be of marginal quality due to high sampling variability, and should be used with caution."

Category F – Poor Quality
Estimates and confidence intervals are deemed of poor quality. The estimates contain a very high

level of instability, making them unreliable and potentially misleading. If the estimates are released, they 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.”

The table below provides the rules for assigning an estimate $\hat{Y}$and its confidence interval to a quality category (A, E or F). The rules are mainly based on sample counts.

Table 2: Release Guidelines

Type of Estimate	**Category A**	**Category E** **Marginal Quality**	**Category F** **Poor Quality**
Proportion	n ≥ 163	Not A and Not F	n < 82
Weighted count	m ≥ 163	Not A and Not F	m < 82
Mean, $\hat{Y}$	n ≥ 163 and L≤|$\hat{Y}$|	Not A and Not F	n < 82 or L>2|$\hat{Y}$|
Total, $\hat{Y}$	m ≥ 163 and L≤|$\hat{Y}$|	Not A and Not F	m < 82 or L>2|$\hat{Y}$|
Difference, $\hat{Y} = \hat{Y}_1 - \hat{Y}_2$	$\hat{Y}_1$ and $\hat{Y}_2$ are Category A	Not A and Not F	$\hat{Y}_1$ or $\hat{Y}_2$ is Category F

Notation:
n: Domain sample size. For proportions, n represents the unweighted count of the number of respondents included in the denominator of the proportion; there are no sample size requirements for the numerator of a proportion. For means, n represents the unweighted count of the number of respondents that contribute to the calculation of the mean (including respondents with values of zero).
m: Unweighted count of the number of respondents with nonzero values that contribute to the estimate
L: Length of the 95% confidence interval of $\hat{Y}$. The length of the confidence interval is used for quantitative variables such as income (as opposed to dichotomous or categorical variables).
| |: absolute value

The rules in Table 2 depend on the type of estimate. Proportions and weighted counts are estimates based on dichotomous or categorical variables. An example of a weighted count is the estimated number of drinks consumed. On the other hand, the rules for means and totals apply to quantitative variables, such as income. Estimates of the difference between two variables include estimates of change between two survey cycles, and estimates of the difference between two domains.

In addition to the rules specified by Table 2, there are two conditions that indicate that a confidence interval is of poor quality. The quality of an estimate and its confidence interval should be categorized as poor if either of the following two conditions is true:

- Length of the 95% confidence interval is zero; i.e., L=0. (An exception is if the estimate is based on a census rather than a sample, or if the estimate corresponds to a calibration control total; see *Chapter* 9 for more information on the calibration.)

- The lower bound or upper bound of the 95% confidence interval is not a plausible value for the estimate. This is an indication that the assumptions about the distribution of the estimate are violated. For example, the lower bound for the estimated number of drinks consumed should not be negative.

8.5.2 Release Guidelines based on Confidentiality

Section 8.5.1 covered the release guidelines based on quality in terms of sampling error. Another aspect to consider to determine which estimates can be released is the confidentiality. In order to make sure that the identity of respondents is protected, it is required that at least 5 respondents contributed to each released estimate. This would mean, for instance, that the unweighted count of the number of respondents included in the numerator of a proportion is at least 5.

9.0 Weighting

For the microdata files, a final set of person weights were assigned to each record to represent the number of sampled persons that the record represents. Interim household weights had to first be calculated in order to calculate these person weights.

The weighting for the CADS files consisted of several steps, beginning with household weight:

1. calculation of initial weight: each household selected represents multiple other households within the strata;
2. dropping out-of-scope records,
3. adjustments for non-responding households (key questions missing),
4. adjustments to make the household estimates consistent with known provincial totals obtained from demographic projections regularly produced by Statistics Canada.

Person weight calculation starts with the household weights in step 4:

5. calculation of weight for the selection of the person in the household, based on age selection
6. adjustments to make the population estimates consistent with known province, age groups, sex and Census Metropolitain Area (CMA) totals from the population projections.

Here is the description for each step of the weighting procedure:

9.1 Weighting Procedures

1. Calculate initial weight

Each unit in the sample was assigned a basic weight, W_1**,** equal to the inverse of its probability of selection within each province.

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

There were 22,116 sampled units with assigned weights.

2. Remove out-of-scope cases

Out-of-scope units, such as those with an address corresponding to a business, institution, seasonal or collective dwelling, were removed. There were 1,261 units identified as out-of-scope.

If out-of-scope units,

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

Else,

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

3. Adjust for non-responding households

If the person selected through the age selection of the respondent refused to participate or did not answer the questions used for weighting (age, household size, household composition and at least two questions on alcohol consumption), then the household (and person) was considered a non-respondent. There are 10,562 households that were considered non-respondents. Various variables available for all units in the sample were analyzed for inclusion in the non-response model, and the ones with the highest prediction power were kept. The in-scope weights for the 10,293 responding households were hence adjusted by household income*household type* presence/absence of telephone number. Household type represents the household composition of the dwelling – person living alone, couple without children, couple with children, one adult with children, etc.

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

4. Adjust to known external household totals

An adjustment was made to the household weights on records within each province and household size in order to make household estimates consistent with known external household counts. This corresponds to the final household weight. The adjustment factor by province*household size was defined as:

$$W_{4,i} = W_{3,i} * \left(\frac{Known\; external\; household\; count}{\sum W_3 \; for\; responding\; households\; in\; the\; sample}\right)$$

5. Calculate selected person weight

A weight was assigned to all survey respondents. The initial weight of each person is equal to the final weight of his or her household, multiplied by the inverse of the probability of having been selected in his or her household, according to age selection:

$$W_{5,i} = W_{4,i} * \left(\frac{1}{selection\; probability}\right)$$

6. Adjust to external totals

An adjustment was made to the person weights in order to make population estimates consistent with external population counts for persons 15 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. The following age groups were used: 15 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64 and 65 years old and more.
2) Population totals by CMA

The person weights obtained after this step represent the final person level weight that is available on the microdata files.

10.0 Other documentation

Questionnaires :

- English : CADS2019_Questionnaire_E.pdf
- French : ECAD2019_Questionnaire_F.pdf

Dictionnaires de données :

- Public Use Microdata File (PUMF)
  - English : CADS2019_PUMF_Cdbk.pdf
  - French : ECAD2019_FMGD_LvCd.pdf

- Master files
  - English : CADS2019_MASTER_Cdbk.pdf
  - French : ECAD2019_MAITRE_LvCd.pdf

Appendix A: *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 CADS 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, named WRMP1-WRMP1000. Essentially, the variance is estimated by calculating the value of the desired estimate using each set of bootstrap weights and then measuring the variability between the bootstrap estimates.

Statistical packages for variance estimation

For CADS, it is necessary to use bootstrap weights to compute correct estimates of the variance. 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 like CADS. These packages should not be used to calculate variance estimates, to construct confidence intervals nor to conduct statistical tests (significance tests, regression analysis, et cetera).

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 $\hat{y} \pm 1.96\sqrt{\hat{\text{var}}(\hat{y}\,)}$ 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.

1. SAS, modified Wilson confidence intervals:
```
PROC SURVEYFREQ
DATA=…. VARMETHOD=BRR;
WEIGHT WEIGHT;
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 WEIGHT;
REPWGT WRMP1-WRMP1000;
TABLES ...;
```

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.