Microdata User Guide

CANADIAN TOBACCO USE MONITORING SURVEY

ANNUAL

FEBRUARY – DECEMBER 2012

Statistics Canada Statistique Canada

Canada

TABLE OF CONTENTS

1.0 Introduction

The Canadian Tobacco Use Monitoring Survey (CTUMS) was conducted by Statistics Canada from February to December 2012 with the cooperation and support of Health Canada. This manual has been produced to facilitate the manipulation of the microdata file of the survey results.

Any questions about the data set or its use should be directed to:

Statistics Canada

Client Services
Special Surveys Division
Telephone: 613-951-3321 or toll-free 1-800-461-9050
Fax: 613-951-4527
E-mail: ssd@statcan.gc.ca

Kathleen Fowler
Special Surveys Division
2nd floor, Main Building,
150 Tunney's Pasture Driveway,
Ottawa, Ontario K1A 0T6
Telephone: 613-951-2978
Fax: 613-951-4527
E-mail: kathleen.fowler@statcan.gc.ca

Health Canada

Mark Latendresse
Controlled Substances and Tobacco Directorate
Healthy Environments and Consumer Safety Branch
Room 1605-659, AL 0301A
150 Tunney's Pasture Driveway, Tunney's Pasture
Ottawa, Ontario K1A OK9
Telephone: 613-946-9127
Fax: 613-946-8708
E-mail: mark.latendresse@hc-sc.gc.ca

2.0 Background

Statistics Canada has conducted smoking surveys on an ad hoc basis on behalf of Health Canada since the 1960s. These surveys have been done as supplements to the Canadian Labour Force Survey and as random digit dialling telephone surveys.

In February 1994, a change in legislation was passed which allowed a reduction in cigarette taxes. Since there was no survey data from immediately before this legislative change, it was difficult for Health Canada or other interested analysts to measure exactly the impact of the change.

As Health Canada wants to be able to monitor the consequences of legislative changes and anti-smoking policies on smoking behaviour, the CTUMS was designed to provide Health Canada and its partners/stakeholders with continual and reliable data on tobacco use and related issues.

Since 1999, two CTUMS files have been released every year: a file with data collected from February to June and a file with the July to December data. Additionally, there has also been a yearly summary. This process was changed for the 2011 release. There is now only an annual file released each year. The present file covers the period from February to December 2012.

3.0 Objectives

The primary objective of the survey is to provide a continuous supply of smoking prevalence data against which changes in prevalence can be monitored. The CTUMS is the only Statistics Canada survey that meets Health Canada's need for continuous coverage in time, rapid delivery of data, and sufficient detail of the most at-risk populations, namely 15 to 24 year olds. In contrast the Canadian Community Health Survey provides occasional results on a limited set of prevalence measures related to smoking.

The CTUMS allows Health Canada to look at smoking prevalence by province-sex-age group, for age groups 15 to 19, 20 to 24, 25 to 34, 35 to 44 and 45 and over. Data will continue to be collected on an on-going basis depending on availability of funds.

4.0 Concepts and Definitions

Since the CTUMS is conducted over the telephone, easy to understand terminology is used throughout the questionnaire to avoid long explanations. Some standard concepts and definitions should be used in the analysis and interpretation of this data. The survey questions were designed with these definitions in mind.

Current Smoking Status

1) Daily smoker: A person who currently smokes cigarettes every day.

2) Non-daily smoker: A person who currently smokes cigarettes, but not every day.

3) Non-smoker: A person who currently does not smoke cigarettes.

4) Current smoker: A person who currently smokes cigarettes daily or occasionally.

Smoking History

1) Former smoker: A person who has smoked at least 100 cigarettes in his life, but currently does not smoke.

2) Experimental smoker: A person who has smoked at least one cigarette, but less than 100 cigarettes, and currently does not smoke cigarettes.

3) Lifetime abstainer: A person who has never smoked cigarettes at all.

4) Ever smoker: A person who is a current smoker or a former smoker.

5) Never smoker: A person who was an experimental smoker or who is a lifetime abstainer.

Smoking Prevalence

Proportion of population which smokes cigarettes at the current time.

Age

Information about the respondent's age is obtained from two sources: from a household respondent who provided the ages of all the household members (roster age), and later, at the beginning of the interview with the selected person, directly from the individual respondent who is asked to state his/her age. The DVAGE variable is the age provided by the selected respondent or, when it is not available (e.g. refused), the roster age is used.

5.0 Survey Methodology

The CTUMS was administered between February 1st and December 31st, 2012 as a random digit dialling (RDD) survey, a technique whereby telephone numbers are generated randomly by computer. Interviewing was conducted over the telephone.

5.1 Population Coverage

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

1) residents of the Yukon, Northwest Territories and Nunavut, and
2) full-time residents of institutions.

Because the survey was conducted using a sample of telephone numbers, households (and thus persons living in households) that do not have telephone land lines were excluded from the sample population. This means that people without telephones and people with cell phones only, were excluded. People without land lines account for about 14% of the target population. However, the survey estimates have been weighted to include persons without land lines.

5.2 Stratification

In order to ensure that people from all parts of Canada were represented in the sample, each of the 10 provinces were divided into strata or geographic areas. Generally, within each province, a census metropolitan area (CMA) stratum and a non-CMA stratum was defined. In Prince Edward Island, there was only one stratum for the province. In Ontario, there was a third stratum for Toronto, and in Quebec, there was a third stratum for Montreal. CMAs are areas defined by the census and correspond roughly to the cities with populations of 100,000 or more.

5.3 Sample Design and Allocation

The sample design is a special two-phase stratified random sample of telephone numbers. The two-phase design is used in order to increase the representation in the sample of individuals belonging to the 15 to 19 and 20 to 24 age groups. In the first phase, households are selected using RDD. In the second phase, one or two individuals (or none) are selected based upon household composition.

Because the main purpose of the survey is to produce reliable estimates in all 10 provinces, an ***equal number of respondents in each province*** is targeted. The target is to get responses from 5,000 individuals aged 15 to 24 and 5,000 individuals aged 25 and over across Canada, or 500 individuals in each age group per province per cycle. The initial sample size of telephone numbers depended upon the expected response rate and the expected RDD hit rate (proportion of sampled telephone numbers which are screened in as households). To achieve the required sample sizes, two adjustments to the standard RDD methodology were introduced. First, the probabilities of selection within the household were unequal and second, households with only persons aged 25 and over present were sub-sampled. It is estimated that a total of almost 190,000 telephone numbers per year will be needed to get the 20,000 respondents per year. This assumed a 72% response rate and about 22% of households having individuals aged 15 to 24; the hit rate varies substantially by province, with an expected overall average of about 32%.

5.4 Sample Selection

The sample for the CTUMS was generated using a refinement of RDD sampling called the Elimination of Non-Working Banks (ENWB). Within each province-stratum combination, a list of working banks (area code + next five digits) was compiled from telephone company

administrative files. A working bank, for the purposes of social surveys, is defined as a bank which contains at least one working residential telephone number. Thus, all banks with only unassigned, non-working, or business telephone numbers are excluded from the survey frame.

Next, a systematic sample of banks (with replacement) was selected within each stratum. For each selected bank, a two-digit number (00 to 99) was generated at random. This random number was added to the bank to form a complete telephone number. This method allowed listed and unlisted residential numbers as well as business and non-working numbers (i.e. not currently or never in service), to have a chance of being in the sample. A screening activity aimed at removing not in service and known business numbers was performed prior to sending the sample to the computer-assisted telephone interviewing (CATI) unit.

Each telephone number in the CATI sample was dialled to determine whether or not it reached a household. If the telephone number is found to reach a household, the person answering the telephone was asked to provide information on the individual household members. The ages of the household members were used to determine who, in the household, would be selected for the tobacco use interview. Proxy interviews were not accepted.

To ensure that enough people were reached in the younger age groups, the random selection was set up such that at least one person aged 15 to 19 or 20 to 24 would be selected within a household, if they exist. The reason for this is that about 78% of all households in Canada are made up of only people aged 25 and over; another 19% consist of people 25 and over living with people in either the 15 to 19 or 20 to 24 age group; and only 3% of households contain no one aged 25 and over. If all ages were selected with equal probability and retained, the 25 and over age group would be over-represented with respect to the survey objectives. Thus, to save on the costs of additional interviews, some of the selected people in the 25 and over age group were screened out and did not receive the tobacco use interview. Two people were selected if more than one of the age groups 15 to 19, 20 to 24, and 25 and over were represented in the household. When two people in the same household were selected, they were always from different age groups. This ensured that there was no negative impact on the precision of the estimates by age group due to correlation within households. There was a small impact on the precision for the total estimates for all ages, but the sample size was sufficiently large so the impacts were minimal.

The detailed logic for the selection of individuals was as follows:

1) If everyone in the household is 15 to 19 then one person is selected at random.

2) If everyone in the household is 20 to 24 then one person is selected at random.

3) If everyone in the household is 25 and over then one person is selected at random; however, this selected person is retained for only a proportion of the cases.

4) If some household members are 15 to 19 and the rest are 20 to 24 then two people are selected at random, one from each age group.

5) If some household members are 15 to 19 and the rest are 25 and over then two people are selected at random, one from each age group; however, the person selected from the 25 and over age group is retained for only a proportion of the cases.

6) If some household members are 20 to 24 and the rest are 25 and over then two people are selected at random, one from each age group; however, the person selected from the 25 and over age group is retained for only a proportion of the cases.

7) If all three age groups are represented in the household, then a check is made to see if the 25 and over age group will be retained. If it is then two of the age groups are selected at random, if not the 15 to 19 and the 20 to 24 age groups are selected. This is a new process that started July 2009. Previously the two age groups would be selected at random and then rule 4), 5), or 6) would apply.

6.0 Data Collection

6.1 Questionnaire Design

The questionnaire design for this survey borrows heavily from the 1994 Survey on Smoking in Canada. Some questions have been added for consistency with international surveys which use the concept of smoking behaviour “in the last 30 days”.

Specifications for valid ranges and inter-question consistency were incorporated into the CATI application to the extent feasible. Additional consistency edits were done during the data processing phase.

The 2012 survey used the same questionnaire as Cycle 2 2011.

6.2 Data Collection and Editing

The interviews were conducted every month, from February through December 2012.

Data were collected using computer-assisted telephone interviewing. The CATI system has a number of generic modules which can be quickly adapted to most types of surveys. A front-end module contains a set of standard response codes for dealing with all possible call outcomes, as well as the associated scripts to be read by the interviewers. A standard approach set up for introducing the agency, the name and purpose of the survey, the survey sponsors, how the survey results will be used, and the duration of the interview was used. We explained to respondents how they were selected for the survey, that their participation in the survey is voluntary, and that their information will remain strictly confidential. Help screens were provided to the interviewers to assist them in answering questions that are commonly asked by respondents.

The CATI application ensured that only valid question responses were entered and that all the correct flows were followed. Edits were built into the application to check the consistency of responses, identify and correct outliers, and to control who gets asked specific questions. This meant that the data was already quite “clean” at the end of the collection process.

Interviewers were trained on the survey content and the CATI application. In addition to classroom training, the interviewers completed a series of mock interviews to become familiar with the survey and its concepts and definitions. Every attempt is made to ensure that the same set of interviewers is used each month. This minimizes training and yields better and more consistent data quality.

The cases were distributed to two of the Statistics Canada regional offices. The workload and interviewing staff within each office was managed by a project manager. The automated scheduler used by the CATI system ensured that cases were assigned randomly to interviewers and that cases were called at different times of the day and different days of the week to maximize the probability of contact. There were a maximum of 20 call attempts per case identified as a residential phone number; once the maximum was reached, the case was reviewed by a senior interviewer who determined if additional calls would be made. There were a maximum of 5 call attempts per case identified as an unknown phone number; if during these 5 call attempts a phone number was identified as belonging to a household the maximum was raised to 20.

7.0 Data Processing

The main output of the CTUMS two "clean" microdata files, one for the household level information and one for the person level information. This chapter presents a brief summary of the processing steps involved in producing these files.

7.1 Data Capture

As the data was collected using computer-assisted telephone interviewing, there was no need for a separate data capture system since the information was entered in the Regional Offices systems directly by the interviewers during the interview.

7.2 Editing

The first stage of survey processing was to merge the monthly files into a single file.
The first type of error treated was errors in questionnaire flow, where questions which did not apply to the respondent (and should therefore not have been answered) were found to contain answers. In this case a computer edit automatically eliminated superfluous data by following the flow of the questionnaire implied by answers to previous, and in some cases, subsequent questions.

The second type of error treated involved a lack of information in questions which should have been answered. For this type of error, a non-response or "not-stated" code was assigned to the item.

7.3 Creation of Derived Variables

A number of data items on the microdata file have been derived by combining items on the questionnaire in order to facilitate data analysis. Examples of derived variables include the average number of cigarettes smoked daily and the number of years the respondent smoked. The urban or rural character of the community where the respondent lives (DVURBAN) has been derived from the postal code. The occupational category DVNOCS10 is based on responses to questions LF_Q30 and LF_Q40 which were coded according to the 2006 National Occupational Classification for Statistics (NOC-S). The 10 occupational categories correspond to the first digit of the classification.

7.4 Weighting

The principle behind estimation in 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 simple random 2% sample of the population, each person in the sample represents 50 persons in the population.

The weighting phase is a step which calculates, for each record, what this number is. This weight appears on the microdata file, and **must** be used to derive meaningful estimates from the survey. For example, if the number of people in Canada who smoke daily is to be estimated, it is done by selecting the records referring to those individuals in the sample with that characteristic (SS_Q10 = 1) and summing the weights entered on those records.

Details of the method used to calculate these weights are presented in Chapter 11.0.

7.5 Suppression of Confidential Information

It should be noted that the public use microdata files (PUMF) may differ from the survey “master” files held by Statistics Canada. These differences usually are the result of actions taken to protect

the anonymity of individual survey respondents. The most common actions are the suppression of file variables, grouping values into wider categories, and coding specific values into the "not stated" category. Users requiring access to information excluded from the microdata files may purchase custom tabulations. Estimates generated will be released to the user, subject to meeting the guidelines for analysis and release outlined in Chapter 9.0 of this document.

Household File and Person File

Geographic Identifiers:

The survey's master data files include explicit geographic identifiers for province and stratum (census metropolitan area (CMA), non-CMA, Toronto or Montreal).The survey's public use microdata files only contain an identifier for province.

Household Age Composition:

Household age composition is available as the number of household members (capped at two) in the following age ranges: 0 to 14, 15 to 24, 25 to 44, and 45 and over.

Other Modifications to the Household File and Person File:

In order to avoid potential identification of respondents resulting from an unusual combination of characteristics, 229 records on the household and person files had a demographic variable recoded.

Additionally, when the sum of household members derived from the information about their age ranges exceeded five - the maximum value of the household size variable (HHSIZE), the age range variables (15 to 24, 25 to 44 and 45 and over) were modified. On those records, all the age ranges present in the household were maintained, but some of them had the value "two or more" replaced with "one".

There were 425 such modifications on the Household file and 411 on the Person file.

Person File Only

Geographic Identifiers:

Starting with Cycle 1 of 2002, the master data file contains the first three digits of the respondent's postal code. Since Cycle 2, 2003, the master and the PUMFs contain an urban/rural variable (DVURBAN). This variable is based on the urban/rural status of the enumeration area (defined by Statistics Canada) in which the majority of the postal codes fall. Urban areas have minimum population concentrations of 1,000 people and a population density of at least 400 people per square kilometre based on the 2001 Census population counts. All the territory outside the urban areas is considered rural.

Marital Status:

The detailed marital status variable (six categories) is available on the master file only, while on the PUMF this variable has been grouped into three categories.

Level of Education:

The detailed level of education variable has been replaced with a version of the variable where "no schooling" and "some elementary" categories have been grouped.

Age:

Cases were identified where the derived variable for the respondent's age (DVAGE) in conjunction with the number of years they have been a smoker (DVYRSSMK) and the age they had their first cigarette (PS_Q30) was greater then 85 (the maximum derived age). The number of years smoked was decreased so that the number of years smoked plus the age they had their first cigarette cannot be greater than 85.

The variable MU_Q40 was also capped, so that the age the respondent first used marijuana, cannabis or hashish could not be greater than DVAGE.

8.0 Data Quality

For the CTUMS, the response rates computed include the following.

Household File and Person File

Telephone Resolved Rate is the proportion of sampled telephone numbers that were confirmed as residential or out-of-scope (e.g. business or non-working numbers) thus were considered resolved.

$$\frac{residential\ or\ out-of-scope\ numbers}{sampled\ telephone\ numbers}$$

Hit Rate is the proportion of resolved telephone numbers that were confirmed as residential or had valid household data.

$$\frac{residential\ numbers\ or\ valid\ household\ data}{resolved\ telephone\ numbers}$$

Roster Completion Rate is the proportion of households with a complete roster containing ages for each household member; this is a necessary condition for considering a household and a person record a response.

$$\frac{households\ with\ complete\ roster}{total\ households\ (i.e.\ numbers\ resolved\ as\ residential)}$$

Household Response Rate is the proportion of households with a complete roster (ages provided for everyone in the roster) and with valid household data. *Estimated total households* include all telephone numbers resolved as residential as well as a portion of unresolved telephone numbers that are estimated to be households.

$$\frac{households\ with\ complete\ roster\ and\ valid\ household\ data}{estimated\ total\ households}$$

Person File Only

Person Response Rate is the proportion of records of selected persons with corresponding complete roster and valid household data whose records had **valid person data**.

$$\frac{persons\ with\ complete\ roster,\ with\ valid\ household\ data\ and\ valid\ person\ data}{all\ selected\ persons\ with\ complete\ household\ roster\ and\ valid\ household\ data}$$

Overall Response Rate for the survey fully reflects the response rate at the person level by combining response rates at the household and the person level.

$$Household\ Response\ Rate \times Person\ Response\ Rate$$

Telephone Resolved Rate and Hit Rate by Province

Province	Number of Telephone Numbers Generated	Total Resolved Numbers	Telephone Resolved Rate (%)	Total Number of Households	Households with Valid Roster Data	Roster Completion Rate (%)	Hit Rate (%)
Newfoundland and Labrador	24,893	23,504	94.4	6,735	6,105	90.6	27.1
Prince Edward Island	19,536	18,317	93.8	5,606	5,050	90.1	28.7
Nova Scotia	22,935	21,551	94.0	6,895	6,382	92.6	30.1
New Brunswick	26,598	25,284	95.1	6,480	5,863	90.5	24.4
Quebec	18,711	17,720	94.7	6,690	5,805	86.8	35.8
Ontario	18,324	16,943	92.5	5,514	4,259	77.2	30.1
Manitoba	17,489	16,420	93.9	5,705	4,779	83.8	32.6
Saskatchewan	17,204	15,964	92.8	5,950	4,835	81.3	34.6
Alberta	17,248	15,767	91.4	5,616	4,488	79.9	32.6
British Columbia	20,218	18,650	92.2	6,222	4,699	75.5	30.8
Canada	**203,156**	**190,120**	**93.6**	**61,413**	**52,265**	**85.1**	**30.2**

8.1 Household Response Rates – February to December 2012

A **household respondent** must complete the roster with no age refusals, and valid household data must exist. Using the new household response rate calculation there were an estimated 14,951 (22.3%) households that were non-responding, 7,757 of these households (11.6% of total households) refused participation.

Household Response Rate by Province

Province	Estimated Total Number of Households	Number of Responding Households	Household Response Rate (%)
Newfoundland and Labrador	7 335	6 087	83,0
Prince Edward Island	6 107	5 036	82,5
Nova Scotia	7 446	6 364	85,5
New Brunswick	6 930	5 850	84,4
Quebec	7 137	5 782	81,0
Ontario	6 127	4 238	69,2
Manitoba	6 166	4 761	77,2
Saskatchewan	6 553	4 807	73,4
Alberta	6 278	4 466	71,1
British Columbia	6 935	4 673	67,4
Canada	**67 015**	**52 064**	**77,7**

Household Response Rate by Survey Month

Survey Month	Estimated Total Number of Households	Number of Responding Households	Household Response Rate (%)
February	6,735	6,087	90.4
March-April	5,606	5,036	89.8
May-June	6,895	6,364	92.3
July-August	6,480	5,850	90.3
September-October	6,690	5,782	86.4
November-December	5,514	4,238	76.9
Total	**5,705**	**4,761**	**83.5**

-

8.2 Person Response Rates - February to December 2012

A **person respondent** has the following characteristics:

- The telephone number of the selected person belonged to a responding household.
- The household roster was completed with no individual age refusals.
- The selected person was 15 years of age or older at the time of the interview (confirmed with the selected person).
- The selected person answered the key questions on smoking habits, at minimum.

There were 32,229 households in which household data were collected but nobody was selected to continue with the CTUMS. (See Section 5.4 (Sample Selection), for more information). Of the remaining households, 16,422 had one person selected while 3,413 had two people selected. The refusal rate at the person level was 3.4%.

Person Response Rate by Province

Province	Total Persons Selected	Total Persons Responding	Person Response Rate (%)
Newfoundland and Labrador	2,505	1,971	78.7
Prince Edward Island	2,425	1,975	81.4
Nova Scotia	2,646	2,154	81.4
New Brunswick	2,343	1,908	81.4
Quebec	2,414	2,001	82.9
Ontario	2,135	1,792	83.9
Manitoba	2,434	2,118	87.0
Saskatchewan	2,201	1,923	87.4
Alberta	2,147	1,792	83.5
British Columbia	1,998	1,652	82.7
Canada	**23,248**	**19,286**	**83.0**

Person Response Rate by Survey Month

Survey Month	Total Persons Selected	Total Persons Responding	Person Response Rate (%)
February	2,257	1,847	81.8
March-April	4,373	3,690	84.4
May-June	4,327	3,607	83.4
July-August	4,236	3,517	83.0
September-October	4,046	3,329	82.3
November-December	4,009	3,296	82.2
Total	**23,248**	**19,286**	**83.0**

Target Number of Respondents and Person Response Rate by Age Group

Age Group	Total Persons Selected	Total Persons Responding	Person Response Rate (%)
15 to 19	6,278	4,887	77.8
20 to 24	4,823	3,646	75.6
25 and over	12,147	10,753	88.5
Total	**23,248**	**19,286**	**83.0**

Overall Response Rate by Province

Province	Household Response Rate (%)	Person Response Rate (%)	Overall Response Rate (%)
Newfoundland and Labrador	83.0	78.7	65.3
Prince Edward Island	82.5	81.4	67.2
Nova Scotia	85.5	81.4	69.6
New Brunswick	84.4	81.4	68.7
Quebec	81.0	82.9	67.1
Ontario	69.2	83.9	58.1
Manitoba	77.2	87.0	67.2
Saskatchewan	73.4	87.4	64.1
Alberta	71.1	83.5	59.4
British Columbia	67.4	82.7	55.7
Canada	**77.7**	**83.0**	**64.4**

8.3 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 on the questionnaire 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 was 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 include extensive training of interviewers with respect to the survey procedures and CATI application, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions and testing of the CATI application to ensure that range checks, edits and question flow were all programmed correctly.

8.4 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 or individuals who responded to the survey to compensate for those who did not respond.

8.5 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. refused, don't know.

8.6 Coverage

As mentioned in Section 5.1 (Population Coverage), about 14% of households in Canada do not have telephone land lines. 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 non-telephone or cell phone only ownership.

8.7 Measurement of Sampling Error

Since it is an unavoidable fact 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 documentation outlines the measures of sampling error which Statistics Canada commonly uses and which it urges users producing estimates from this microdata file to use also.

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 large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation of an estimate, is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate.

For example, suppose that, based upon the 2002 Annual survey results, one estimates that 21.4% of Canadians are currently cigarette smokers, and this estimate is found to have standard error of 0.0039. Then the coefficient of variation of the estimate is calculated as:

$$\left(\frac{0.0039}{0.214}\right) X\, 100\% = 1.8\%$$

There is more information on the calculation of coefficients of variation in Chapter 10.0.

9.0 Guidelines for Tabulation, Analysis and Release

This chapter of the documentation outlines the guidelines to be adhered to by users tabulating, analysing, publishing or otherwise releasing any data derived from the survey microdata files. With the aid of these guidelines, users of microdata 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.

9.1 Rounding Guidelines

In order that estimates for publication or other release derived from these microdata files 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 100 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 100 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.

9.2 Sample Weighting Guidelines for Tabulation

The sample design used for the CTUMS was not self-weighting. When producing simple estimates including the production of ordinary statistical tables, users must apply the proper survey weight.

If proper weights are not used, the estimates derived from the microdata files cannot be considered to be representative of the survey population, and will not correspond to those produced by Statistics Canada.

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.

9.3 Definitions of Types of Estimates: Categorical and Quantitative

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

9.3.1 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 currently smoke cigarettes, or the proportion of daily smokers that have attempted to quit smoking 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: In the past 30 days, did you smoke any cigarettes?
R: Yes / No

Q: What was your main reason to quit smoking?
R: Health / Pregnancy or a baby in the household / Less stress in life / Cost of cigarettes / Smoking is less socially acceptable / Some other reason

9.3.2 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 cigarettes smoked 6 days ago, per person. The numerator $\left(\hat{X}\right)$ is an estimate of the total number of cigarettes smoked 6 days ago, and its denominator $\left(\hat{Y}\right)$ is the number of persons who reported smoking 6 days ago.

Examples of Quantitative Questions:

Q: Some people smoke more or less depending upon the day of the week. So, thinking back over the past seven days, starting with yesterday, how many cigarettes did you smoke: …6 days ago?
R: |_|_| cigarettes

Q: At what age did you smoke your first cigarette?
R: |_|_| years old

9.3.3 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata file 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 records having the characteristic of interest for the denominator $\left(\hat{Y}\right)$, then
c) dividing estimate a) by estimate b) $\left(\hat{X} / \hat{Y}\right)$.

9.3.4 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata file 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 cigarettes smoked 6 days ago, multiply the value reported in question WP_Q10F (number of cigarettes smoked 6 days ago) by the final weight for the record, then sum this value over all records with WP_Q10F < 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 cigarettes smoked 6 days ago,

a) estimate the total number of cigarettes smoked 6 days ago $\left(\hat{X}\right)$ 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 WP_Q10F < 96, then
c) divide estimate a) by estimate b) $\left(\hat{X} / \hat{Y}\right)$.

9.4 Guidelines for Statistical Analysis

The CTUMS 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. In order for survey estimates and analyses to be free from bias, the survey weights must be used.

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, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are poor. Approximate variances for simple estimates such as totals, proportions and ratios (for qualitative variables) can be derived using the accompanying Approximate Sampling Variability Tables.

For other 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, 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 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 = men;
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 = men;
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 and clustering of the sample's design are still not taken into account, the variance estimates calculated in this way are likely to be under-estimates.

The calculation of more precise variance estimates requires detailed knowledge of the design of the survey. Such detail cannot be given in this microdata file because of confidentiality. Variances that take the complete sample design into account can be calculated for many statistics by Statistics Canada on a cost recovery basis.

9.5 Coefficient of Variation Release Guidelines

Before releasing and/or publishing any estimate from the CTUMS users should first determine the quality level of the estimate. The quality levels are *acceptable*, *marginal* and *unacceptable*. Data quality is affected by both sampling and non-sampling errors as discussed in Chapter 8.0. However for this purpose, the quality level of an estimate will be determined only on the basis of sampling error as reflected by the coefficient of variation as shown in the table below. Nonetheless users should be sure to read Chapter 8.0 to be more fully aware of the quality characteristics of these data.

First, the number of respondents who contribute to the calculation of the estimate should be determined. If this number is less than 30, the weighted estimate should be considered to be of unacceptable quality.

For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the estimate and follow the guidelines below. These quality level guidelines should be applied to rounded weighted estimates.

All estimates can be considered releasable. However, those of marginal or unacceptable quality level must be accompanied by a warning to caution subsequent users.

Quality Level Guidelines

Quality Level of Estimate	Guidelines
1) Acceptable	Estimates have a sample size of 30 or more, and low coefficients of variation in the range of 0.0% to 16.5%. No warning is required.
2) Marginal	Estimates have a sample size of 30 or more, and high coefficients of variation in the range of 16.6% to 33.3%. Estimates should be flagged with the letter E (or some similar identifier). They should be accompanied by a warning to caution subsequent users about the high levels of error, associated with the estimates.
3) Unacceptable	Estimates have a sample size of less than 30, or very high coefficients of variation in excess of 33.3%. Statistics Canada recommends not to release estimates of unacceptable quality. However, if the user chooses to do so then estimates should be flagged with the letter F (or some similar identifier) and the following warning should accompany the estimates: "Please be warned that these estimates [flagged with the letter F] do not meet Statistics Canada's quality standards. Conclusions based on these data will be unreliable, and most likely invalid."

9.6 Release Cut-off's for the Household File

The minimum size of the estimates are specified in the table below by province for households. Estimates smaller than the minimum size given in the "Unacceptable" column must be flagged in the appropriate manner.

Table of Release Cut-offs – Household File

Province	Acceptable CV 0.0% to 16.5%	Marginal CV 16.6% to 33.3%	Unacceptable CV > 33.3%
Newfoundland and Labrador	1,500 & over	500 to < 1,500	under 500
Prince Edward Island	500 & over	250 to < 500	under 250
Nova Scotia	2,500 & over	500 to < 2,500	under 500
New Brunswick	2,000 & over	500 to < 2,000	under 500
Quebec	25,500 & over	6,500 to < 25,500	under 6,500
Ontario	49,500 & over	12,000 to < 49,500	under 12,000
Manitoba	4,000 & over	1,000 to < 4,000	under 1,000
Saskatchewan	3,500 & over	1,000 to < 3,500	under 1,000
Alberta	13,000 & over	3,000 to < 13,000	under 3,000
British Columbia	17,000 & over	4,000 to < 17,000	under 4,000
Canada	**27,500 & over**	**7,000 to < 27,500**	**under 7,000**

9.7 Release Cut-off's for the Person File

The minimum size of the estimates are specified in the table below by province and age group. Estimates smaller than the minimum size given in the "Unacceptable" column must be flagged in the appropriate manner.

Table of Release Cut-offs – Person File

Province	Age Group	Acceptable CV 0.0% to 16.5%	Marginal CV 16.6% to 33.3%	Unacceptable CV > 33.3%
Newfoundland and Labrador	All	13,500 & over	3,500 to < 13,500	under 3,500
	15 to 19	2,500 & over	500 to < 2,500	under 500
	20 to 24	3,500 & over	1,000 to < 3,500	under 1,000
	25+	15,000 & over	4,000 to < 15,000	under 4,000
Prince Edward Island	All	3,500 & over	1,000 to < 3,500	under 1,000
	15 to 19	1,000 & over	0 to < 1,000	under 0
	20 to 24	1,500 & over	500 to < 1,500	under 500
	25+	4,000 & over	1,000 to < 4,000	under 1,000
Nova Scotia	All	23,500 & over	6,000 to < 23,500	under 6,000
	15 to 19	5,000 & over	1,500 to < 5,000	under 1,500
	20 to 24	7,000 & over	2,000 to < 7,000	under 2,000
	25+	27,000 & over	7,000 to < 27,000	under 7,000
New Brunswick	All	22,500 & over	5,500 to < 22,500	under 5,500
	15 to 19	8,500 & over	2,500 to < 8,500	under 2,500
	20 to 24	7,000 & over	2,000 to < 7,000	under 2,000
	25+	25,500 & over	6,500 to < 25,500	under 6,500
Quebec	All	223,000 & over	56,000 to < 223,000	under 56,000
	15 to 19	38,500 & over	10,000 to < 38,500	under 10,000
	20 to 24	52,500 & over	14,000 to < 52,500	under 14,000
	25+	259,500 & over	66,000 to < 259,500	under 66,000
Ontario	All	441,000 & over	111,500 to < 441,000	under 111,500
	15 to 19	75,500 & over	20,000 to < 75,500	under 20,000
	20 to 24	104,500 & over	28,000 to < 104,500	under 28,000
	25+	519,000 & over	133,000 to < 519,000	under 133,000
Manitoba	All	31,500 & over	8,000 to < 31,500	under 8,000
	15 to 19	7,500 & over	2,000 to < 7,500	under 2,000
	20 to 24	12,500 & over	3,500 to < 12,500	under 3,500
	25+	35,500 & over	9,000 to < 35,500	under 9,000
Saskatchewan	All	28,000 & over	7,000 to < 28,000	under 7,000
	15 to 19	6,500 & over	1,500 to < 6,500	under 1,500
	20 to 24	9,500 & over	2,500 to < 9,500	under 2,500
	25+	32,500 & over	8,000 to < 32,500	under 8,000
Alberta	All	109,500 & over	27,500 to < 109,500	under 27,500
	15 to 19	22,500 & over	6,000 to < 22,500	under 6,000
	20 to 24	34,500 & over	9,500 to < 34,500	under 9,500
	25+	132,000 & over	33,500 to < 132,000	under 33,500
British Columbia	All	162,500 & over	41,000 to < 162,500	under 41,000
	15 to 19	28,000 & over	7,500 to < 28,000	under 7,500
	20 to 24	44,000 & over	12,000 to < 44,000	under 12,000
	25+	187,000 & over	48,000 to < 187,000	under 48,000
Canada	**All**	**265,500 & over**	**65,500 to < 265,500**	**under 65,500**
	15 to 19	**47,500 & over**	**12,000 to < 47,500**	**under 12,000**
	20 to 24	**69,000 & over**	**17,500 to < 69,000**	**under 17,500**
	25+	**314,500 & over**	**78,000 to < 314,500**	**under 78,000**

10.0 Approximate Sampling Variability Tables

In order to supply coefficients of variation (CV) which would be applicable to a wide variety of categorical estimates produced from this microdata file and which could be readily accessed by the user, a set of Approximate Sampling Variability Tables has been produced. These CV tables allow the user to obtain an approximate coefficient of variation based on the size of the estimate calculated from the survey data.

The coefficients of variation are derived using the variance formula for simple random sampling and incorporating a factor which reflects the multi-stage, clustered nature of the sample design. This factor, known as the design effect, was determined by first calculating design effects for a wide range of characteristics and then choosing from among these a conservative value (usually the 75th percentile) to be used in the CV tables which would then apply to the entire set of characteristics.

The table below shows the conservative value of the design effects as well as sample sizes and population counts by province, which were used to produce the Approximate Sampling Variability Tables for the CTUMS Household file.

Household File

Province	Design Effect	Sample Size	Population
Newfoundland and Labrador	1.22	6,087	213,429
Prince Edward Island	0.64	5,036	58,034
Nova Scotia	1.15	6,364	393,380
New Brunswick	1.08	5,850	311,077
Quebec	1.19	5,782	3,404,112
Ontario	1.13	4,238	5,095,128
Manitoba	1.16	4,761	474,366
Saskatchewan	1.15	4,807	418,338
Alberta	1.09	4,466	1,458,521
British Columbia	1.18	4,673	1,840,113
Canada	**2.87**	**52,064**	**13,666,498**

The table below shows the conservative value of the design effects as well as sample sizes and population counts by province and age group, which were used to produce the Approximate Sampling Variability Tables for the CTUMS Person file.

Person File

Province	Age Group	Design Effect	Sample Size	Population
Newfoundland and Labrador	All	1.72	1,971	429,516
	15 to 19	1.21	499	28,607
	20 to 24	1.25	349	31,362
	25+	1.30	1,123	369,548
Prince Edward Island	All	1.70	1,975	120,890
	15 to 19	1.23	476	9,884
	20 to 24	1.30	321	10,263
	25+	1.41	1,178	100,742
Nova Scotia	All	1.78	2,154	795,519
	15 to 19	1.36	534	56,273
	20 to 24	1.27	382	64,612
	25+	1.41	1,238	674,635
New Brunswick	All	1.93	1,908	631,717
	15 to 19	2.89	458	44,561
	20 to 24	1.75	374	48,254
	25+	1.46	1,076	538,901
Quebec	All	1.88	2,001	6,686,906
	15 to 19	1.33	551	474,022
	20 to 24	1.28	420	521,550
	25+	1.34	1,030	5,691,335
Ontario	All	2.01	1,792	11,143,676
	15 to 19	1.19	451	856,552
	20 to 24	1.26	373	948,215
	25+	1.55	968	9,338,909
Manitoba	All	1.85	2,118	1,012,663
	15 to 19	1.43	566	87,463
	20 to 24	1.71	406	92,176
	25+	1.39	1,146	833,024
Saskatchewan	All	1.78	1,923	855,033
	15 to 19	1.34	484	71,494
	20 to 24	1.25	334	79,133
	25+	1.45	1,105	704,406
Alberta	All	1.78	1,792	3,114,487
	15 to 19	1.31	456	238,178
	20 to 24	1.24	328	281,572
	25+	1.47	1,008	2,594,737
British Columbia	All	1.96	1,652	3,889,985
	15 to 19	1.29	422	275,378
	20 to 24	1.28	305	327,760
	25+	1.52	925	3,286,846
Canada	**All**	4.91	19,286	28,680,393
	15 to 19	3.03	4,897	2,142,412
	20 to 24	2.88	3,592	2,404,897
	25+	3.88	10,797	24,133,083

All coefficients of variation in the Approximate Sampling Variability Tables are approximate and, therefore, unofficial. Estimates of actual variance for specific variables may be obtained from Statistics Canada on a cost-recovery basis. Users interested in calculating actual variance estimates may obtain upon request, free of charge, bootstrap weights with programs that compute variance estimates for various statistics.

Since the approximate CV is conservative, the use of actual variance estimates may cause the estimate to be switched from one quality level to another. For instance a *marginal* estimate could become *acceptable* based on the exact CV calculation.

Remember: If the number of observations on which an estimate is based is less than 30, the weighted estimate should be considered unacceptable and should be flagged in the appropriate manner, regardless of the value of the coefficient of variation for this estimate. This is because the formulas used for estimating the variance do not hold true for small sample sizes.

10.1 How to Use the Coefficient of Variation (CV) Tables for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the Approximate Sampling Variability Tables for estimates of the number, proportion or percentage of the surveyed population possessing a certain characteristic and for ratios and differences between such estimates.

Rule 1: Estimates of Numbers of Persons Possessing a Characteristic (Aggregates)

The coefficient of variation depends only on the size of the estimate itself. On the Approximate Sampling Variability Table for the appropriate geographic area, locate the estimated number in the left-most column of the table (headed "Numerator of Percentage") and follow the asterisks (if any) across to the first figure encountered. This figure is the approximate coefficient of variation.

Rule 2: Estimates of Proportions or Percentages of Persons Possessing a Characteristic

The coefficient of variation of an estimated proportion or percentage depends on both the size of the proportion or percentage and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a sub-group of the population. For example, the proportion of former smokers that quit for current health problems is more reliable than the estimated number of former smokers that quit for current health problems. (Note that in the tables the coefficients of variation decline in value when reading from left to right).

When the proportion or percentage is based upon the total population of the geographic area covered by the table, the CV of the proportion or percentage is the same as the CV of the numerator of the proportion or percentage. In this case, Rule 1 can be used.

When the proportion or percentage is based upon a subset of the total population (e.g. those in a particular sex or age group), reference should be made to the proportion or percentage (across the top of the table) and to the numerator of the proportion or percentage (down the left side of the table). The intersection of the appropriate row and column gives the coefficient of variation.

Rule 3: Estimates of Differences Between Aggregates or Percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately. That is, the standard error of a difference $\left(\hat{d} = \hat{X}_1 - \hat{X}_2\right)$ is:

$$\sigma_{\hat{d}} = \sqrt{\left(\hat{X}_1\alpha_1\right)^2 + \left(\hat{X}_2\alpha_2\right)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d}$. This formula is accurate for the difference between separate and uncorrelated characteristics, but is only approximate otherwise.

Rule 4: Estimates of Ratios

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied. This would apply, for example, to the case where the denominator is the number of smokers and the numerator is the number of daily smokers.

In the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of daily smokers as compared to the number of non-smokers, the standard error of the ratio of the estimates is approximately equal to the square root of the sum of squares of each coefficient of variation considered separately multiplied by $\hat{R}$. That is, the standard error of a ratio $\left(\hat{R} = \hat{X}_1 / \hat{X}_2\right)$ is:

$$\sigma_{\hat{R}} = \hat{R}\sqrt{\alpha_1{}^2 + \alpha_2{}^2}$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of $\hat{R}$ is given by $\sigma_{\hat{R}} / \hat{R}$. The formula will tend to overstate the error if $\hat{X}_1$ and $\hat{X}_2$ are positively correlated and understate the error if $\hat{X}_1$ and $\hat{X}_2$ are negatively correlated.

Rule 5: Estimates of Differences of Ratios

In this case, Rules 3 and 4 are combined. The CVs for the two ratios are first determined using Rule 4, and then the CV of their difference is found using Rule 3.

10.1.1 Examples of Using the CV Tables for Categorical Estimates

The following examples based on the 2002 Annual data are included to assist users in applying the foregoing rules. Please note that the data for these examples are different than the results obtained from the current survey and are only to be used as a guide.

Example 1: Estimates of Numbers of Persons Possessing a Characteristic (Aggregates)

Suppose that a user estimates that during the reference period 5,414,335 persons were current smokers (DVSST1 = 1) in Canada. How does the user determine the coefficient of variation of this estimate?

1) Refer to the Person coefficient of variation table for CANADA – All Ages.

Canadian Tobacco Use Monitoring Survey 2002 - February to December - Person File

Approximate Sampling Variability Tables for Canada - All Ages

NUMERATOR OF PERCENTAGE	ESTIMATED PERCENTAGE													
('000)	0.1%	1.0%	2.0%	5.0%	10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%	50.0%	70.0%	90.0%
1	197.2	196.3	195.3	192.3	187.1	181.9	176.4	170.8	165.0	159.0	152.8	139.5	108.0	62.4
2	139.4	138.8	138.1	135.9	132.3	128.6	124.8	120.8	116.7	112.5	108.0	98.6	76.4	44.1
3	113.8	113.3	112.7	111.0	108.0	105.0	101.9	98.6	95.3	91.8	88.2	80.5	62.4	36.0
4	98.6	98.1	97.6	96.1	93.6	90.9	88.2	85.4	82.5	79.5	76.4	69.7	54.0	31.2
5	88.2	87.8	87.3	86.0	83.7	81.3	78.9	76.4	73.8	71.1	68.3	62.4	48.3	27.9
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
75	********	22.7	22.5	22.2	21.6	21.0	20.4	19.7	19.1	18.4	17.6	16.1	12.5	7.2
80	********	21.9	21.8	21.5	20.9	20.3	19.7	19.1	18.5	17.8	17.1	15.6	12.1	7.0
85	********	21.3	21.2	20.9	20.3	19.7	19.1	18.5	17.9	17.2	16.6	15.1	11.7	6.8
90	********	20.7	20.6	20.3	19.7	19.2	18.6	18.0	17.4	16.8	16.1	14.7	11.4	6.6
95	********	20.1	20.0	19.7	19.2	18.7	18.1	17.5	16.9	16.3	15.7	14.3	11.1	6.4
100	********	19.6	19.5	19.2	18.7	18.2	17.6	17.1	16.5	15.9	15.3	13.9	10.8	6.2
125	********	17.6	17.5	17.2	16.7	16.3	15.8	15.3	14.8	14.2	13.7	12.5	9.7	5.6
150	********	16.0	15.9	15.7	15.3	14.8	14.4	13.9	13.5	13.0	12.5	11.4	8.8	5.1
200	********	13.9	13.8	13.6	13.2	12.9	12.5	12.1	11.7	11.2	10.8	9.9	7.6	4.4
250	********	12.4	12.4	12.2	11.8	11.5	11.2	10.8	10.4	10.1	9.7	8.8	6.8	3.9
300	********	********	11.3	11.1	10.8	10.5	10.2	9.9	9.5	9.2	8.8	8.1	6.2	3.6
350	********	********	10.4	10.3	10.0	9.7	9.4	9.1	8.8	8.5	8.2	7.5	5.8	3.3
400	********	********	9.8	9.6	9.4	9.1	8.8	8.5	8.3	8.0	7.6	7.0	5.4	3.1
450	********	********	9.2	9.1	8.8	8.6	8.3	8.1	7.8	7.5	7.2	6.6	5.1	2.9
500	********	********	8.7	8.6	8.4	8.1	7.9	7.6	7.4	7.1	6.8	6.2	4.8	2.8
750	********	********	********	7.0	6.8	6.6	6.4	6.2	6.0	5.8	5.6	5.1	3.9	2.3
1000	********	********	********	6.1	5.9	5.8	5.6	5.4	5.2	5.0	4.8	4.4	3.4	2.0
1500	********	********	********	********	4.8	4.7	4.6	4.4	4.3	4.1	3.9	3.6	2.8	1.6
2000	********	********	********	********	4.2	4.1	3.9	3.8	3.7	3.6	3.4	3.1	2.4	1.4
3000	********	********	********	********	********	3.3	**3.2**	**3.1**	3.0	2.9	2.8	2.5	2.0	1.1
4000	********	********	********	********	********	********	2.8	2.7	2.6	2.5	2.4	2.2	1.7	1.0
5000	********	********	********	********	********	********	**2.5**	2.4	2.3	2.2	2.2	2.0	1.5	0.9
6000	********	********	********	********	********	********	********	2.2	2.1	2.1	2.0	1.8	1.4	0.8
7000	********	********	********	********	********	********	********	********	2.0	1.9	1.8	1.7	1.3	0.7
8000	********	********	********	********	********	********	********	********	********	1.8	1.7	1.6	1.2	0.7
9000	********	********	********	********	********	********	********	********	********	********	1.6	1.5	1.1	0.7
10000	********	********	********	********	********	********	********	********	********	********	1.5	1.4	1.1	0.6
12500	********	********	********	********	********	********	********	********	********	********	********	1.2	1.0	0.6
15000	********	********	********	********	********	********	********	********	********	********	********	********	0.9	0.5

NOTE: FOR CORRECT USAGE OF THESE TABLES PLEASE REFER TO MICRODATA DOCUMENTATION

2) The estimated aggregate (5,414,335) does not appear in the left-hand column (the "Numerator of Percentage" column), so it is necessary to use the figure closest to it, namely 5,000,000.

3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, namely, 2.5%.

4) So the approximate coefficient of variation of the estimate is 2.5%. The finding that there were 5,414,335 (to be rounded according to the rounding guidelines in Section 9.1) current smokers in the reference period is publishable with no qualifications.

Example 2: Estimates of Proportions or Percentages of Persons Possessing a Characteristic

Suppose that the user estimates that 2,865,929 / 12,436,728 = 23.0% of men currently smoke in Canada in the reference period. How does the user determine the coefficient of variation of this estimate?

1) Refer to the Person coefficient of variation table for CANADA (see above). The CANADA level table should be used because it is the smallest table that contains the domain of the estimate, all men in Canada.

2) Because the estimate is a percentage which is based on a subset of the total population (i.e. men), it is necessary to use both the percentage (23.0%) and the numerator portion of the percentage (2,865,929) in determining the coefficient of variation.

3) The numerator, 2,865,929, does not appear in the left-hand column (the "Numerator of Percentage" column) so it is necessary to use the figure closest to it, namely 3,000,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the percentage closest to it, 25.0%.

4) The figure at the intersection of the row and column used, namely 3.1% is the coefficient of variation to be used.

5) So the approximate coefficient of variation of the estimate is 3.1%. The finding that 23.0% of men currently smoke can be published with no qualifications.

Example 3: Estimates of Differences Between Aggregates or Percentages

Suppose that a user estimates that 2,548,406 / 12,814,359 = 19.9% of women currently smoke in Canada, while 2,865,929 / 12,436,728 = 23.0% of men currently smoke in Canada. How does the user determine the coefficient of variation of the difference between these two estimates?

1) Using the Person CANADA coefficient of variation table (see above) in the same manner as described in Example 2 gives the CV of the estimate for women as 3.2%, and the CV of the estimate for men as 3.1%.

2) Using Rule 3, the standard error of a difference $\left(\hat{d} = \hat{X}_1 - \hat{X}_2\right)$ is:

$$\sigma_{\hat{d}} = \sqrt{\left(\hat{X}_1\alpha_1\right)^2 + \left(\hat{X}_2\alpha_2\right)^2}$$

where $\hat{X}_1$ is estimate 1 (men), $\hat{X}_2$ is estimate 2 (women), and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

That is, the standard error of the difference $\hat{d}$ = 0.230 – 0.199 = 0.031 is:

$$\sigma_{\hat{d}} = \sqrt{[(0.230)(0.031)]^2 + [(0.199)(0.032)]^2}$$
$$= \sqrt{(0.00005) + (0.00004)}$$
$$= 0.009$$

3) The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d}$ = 0.009 / 0.031 = 0.290.

4) So the approximate coefficient of variation of the difference between the estimates is 29.0%. The difference between the estimates is considered marginal and Statistics Canada recommends this estimate not be released. However, should the user choose to do so, the estimate should be flagged with the letter E (or some similar identifier) and be accompanied by a warning to caution subsequent users about the high levels of error associated with the estimate.

Example 4: Estimates of Ratios

Suppose that the user estimates that 237,261 women currently smoke in the age group 15 to 19, while 220,511 men currently smoke in the age group 15 to 19. The user is interested in comparing the estimate of women versus that of men in the form of a ratio. How does the user determine the coefficient of variation of this estimate?

1) First of all, this estimate is a ratio estimate, where the numerator of the estimate ($\hat{X}_1$) is the number of women currently smoking in the age group 15 to 19. The denominator of the estimate ($\hat{X}_2$) is the number of men currently smoking in the age group 15 to 19.

2) Refer to the Person coefficient of variation table for CANADA – 15-19.

Canadian Tobacco Use Monitoring Survey 2002 - February to December - Person File

Approximate Sampling Variability Tables for Canada - 15-19

NUMERATOR OF PERCENTAGE ('000)	ESTIMATED PERCENTAGE													
	0.1%	**1.0%**	**2.0%**	**5.0%**	**10.0%**	**15.0%**	**20.0%**	**25.0%**	**30.0%**	**35.0%**	**40.0%**	**50.0%**	**70.0%**	**90.0%**
1	95.8	95.3	94.9	93.4	90.9	88.3	85.7	83.0	80.2	77.3	74.2	67.8	52.5	30.3
2	67.7	67.4	67.1	66.0	64.3	62.5	60.6	58.7	56.7	54.6	52.5	47.9	37.1	21.4
3	********	55.0	54.8	53.9	52.5	51.0	49.5	47.9	46.3	44.6	42.9	39.1	30.3	17.5
4	********	47.7	47.4	46.7	45.5	44.2	42.9	41.5	40.1	38.6	37.1	33.9	26.2	15.2
5	********	42.6	42.4	41.8	40.7	39.5	38.3	37.1	35.9	34.6	33.2	30.3	23.5	13.6
6	********	38.9	38.7	38.1	37.1	36.1	35.0	33.9	32.7	31.5	30.3	27.7	21.4	12.4
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
95	********	********	********	9.6	9.3	9.1	8.8	8.5	8.2	7.9	7.6	7.0	5.4	3.1
100	********	********	********	9.3	9.1	8.8	8.6	8.3	8.0	7.7	7.4	6.8	5.2	3.0
125	********	********	********	********	8.1	7.9	7.7	7.4	7.2	6.9	6.6	6.1	4.7	2.7
150	********	********	********	********	7.4	7.2	7.0	6.8	6.5	6.3	6.1	5.5	4.3	2.5
200	********	********	********	********	**6.4**	6.2	6.1	5.9	5.7	5.5	5.2	4.8	3.7	2.1
250	********	********	********	********	********	**5.6**	5.4	5.2	5.1	4.9	4.7	4.3	3.3	1.9
300	********	********	********	********	********	5.1	4.9	4.8	4.6	4.5	4.3	3.9	3.0	1.7
350	********	********	********	********	********	********	4.6	4.4	4.3	4.1	4.0	3.6	2.8	1.6
400	********	********	********	********	********	********	4.3	4.1	4.0	3.9	3.7	3.4	2.6	1.5
450	********	********	********	********	********	********	********	3.9	3.8	3.6	3.5	3.2	2.5	1.4
500	********	********	********	********	********	********	********	3.7	3.6	3.5	3.3	3.0	2.3	1.4
750	********	********	********	********	********	********	********	********	********	********	2.7	2.5	1.9	1.1
1000	********	********	********	********	********	********	********	********	********	********	********	2.1	1.7	1.0
1500	********	********	********	********	********	********	********	********	********	********	********	********	********	0.8

NOTE: FOR CORRECT USAGE OF THESE TABLES PLEASE REFER TO MICRODATA DOCUMENTATION

3) The numerator of this ratio estimate is 237,261. The figure closest to it is 250,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 5.6%

4) The denominator of this ratio estimate is 220,511. The figure closest to it is 200,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 6.4%.

5) So the approximate coefficient of variation of the ratio estimate is given by Rule 4, which is:

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^{\,2} + \alpha_2^{\,2}}$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. That is:

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(0.056)^2 + (0.064)^2} \\ &= \sqrt{0.003136 + 0.004096} \\ &= 0.085\end{aligned}$$

6) The obtained ratio of women currently smoking in the age group 15 to 19 versus men currently smoking in the age group 15 to 19 is 237,261 / 220,511 which is 1.08 (to be rounded according to the rounding guidelines in Section 9.1). The coefficient of variation of this estimate is 8.5%, which makes the estimate releasable with no qualifications.

10.2 How to Use the Coefficient of Variation Tables to Obtain Confidence Limits

Although coefficients of variation are widely used, a more intuitively meaningful measure of sampling error is the confidence interval of an estimate. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values. For example a 95% confidence interval can be described as follows:

If sampling of the population is repeated indefinitely, each sample leading to a new confidence interval for an estimate, then in 95% of the samples the interval will cover the true population value.

Using the standard error of an estimate, confidence intervals for estimates may be obtained under the assumption that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, the chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and about 99 out of 100 that the differences would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

Confidence intervals for an estimate, $\hat{X}$, are generally expressed as two numbers, one below the estimate and one above the estimate, as $\left(\hat{X} - k, \hat{X} + k\right)$ where k is determined depending upon the level of confidence desired and the sampling error of the estimate.

Confidence intervals for an estimate can be calculated directly from the Approximate Sampling Variability Tables by first determining from the appropriate table the coefficient of variation of the estimate $\hat{X}$, and then using the following formula to convert to a confidence interval $\left(CI_{\hat{x}}\right)$:

$$CI_{\hat{x}} = \left(\hat{X} - t\hat{X}\alpha_{\hat{x}}, \hat{X} + t\hat{X}\alpha_{\hat{x}}\right)$$

where $\alpha_{\hat{x}}$ is the determined coefficient of variation of $\hat{X}$, and

t = 1 if a 68% confidence interval is desired;
t = 1.6 if a 90% confidence interval is desired;
t = 2 if a 95% confidence interval is desired;
t = 2.6 if a 99% confidence interval is desired.

Note: Release guidelines which apply to the estimate also apply to the confidence interval. For example, if the estimate is not releasable, then the confidence interval is not releasable either.

10.2.1 Example of Using the CV Tables to Obtain Confidence Limits

A 95% confidence interval for the estimated proportion of men who currently smoke (from Example 2, Section 10.1.1) would be calculated as follows:

$\hat{X}$ = 23.0% (or expressed as a proportion 0.230)

t = 2

$\alpha_{\hat{x}}$ = 3.1% (0.031 expressed as a proportion) is the coefficient of variation of this estimate as determined from the tables.

$CI_{\hat{x}}$ = {0.230 - (2) (0.230) (0.031), 0.230 + (2) (0.230) (0.031)}

$CI_{\hat{x}}$ = {0.230 - 0.014, 0.230 + 0.014}

$CI_{\hat{x}}$ = {0.216, 0.244}

With 95% confidence it can be said that between 21.6% and 24.4% of men currently smoke.

10.3 How to Use the Coefficient of Variation Tables to Do a T-test

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The sample estimates can be numbers, averages, percentages, ratios, etc. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

Let $\hat{X}_1$ and $\hat{X}_2$ be sample estimates for two characteristics of interest. Let the standard error on the difference $\hat{X}_1 - \hat{X}_2$ be $\sigma_{\hat{d}}$.

If $t = \dfrac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$ is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance. If however, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 0.05 level. That is to say that the difference between the estimates is significant.

10.3.1 Example of Using the Coefficient of Variation Tables to Do a T-test

Let us suppose that the user wishes to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of men who currently smoke and the proportion of women who currently smoke. From Example 3, Section 10.1.1, the standard error of the difference between these two estimates was found to be 0.009. Hence,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{0.230 - 0.199}{0.009} = \frac{0.031}{0.009} = 3.44$$

Since t = 3.44 is greater than 2, it must be concluded that there is a significant difference between the two estimates at the 0.05 level of significance.

10.4 Coefficient of Variation for Quantitative Estimates

For quantitative estimates, special tables would have to be produced to determine their sampling error. Since most of the variables for the CTUMS are primarily categorical in nature, this has not been done.

As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). If the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the total number of cigarettes smoked 6 days ago would be greater than the coefficient of variation of the corresponding proportion of current smokers. Hence, if the coefficient of variation of the proportion is unacceptable (making the proportion not releasable), then the coefficient of variation of the corresponding quantitative estimate will also be unacceptable (making the quantitative estimate not releasable).

Coefficients of variation of such estimates can be derived as required for a specific estimate using a technique known as pseudo replication. This involves dividing the records on the microdata files into subgroups (or replicates) and determining the variation in the estimate from replicate to replicate. Users wishing to derive coefficients of variation for quantitative estimates may contact Statistics Canada for advice on the allocation of records to appropriate replicates and the formulae to be used in these calculations.

10.5 Coefficient of Variation Tables – Household File

Refer to CTUMS2012_HH_CVTabsE.pdf for the coefficient of variation tables for the Household file for 2012.

10.6 Coefficient of Variation Tables – Person File

Refer to CTUMS2012_PR_CVTabsE.pdf for the coefficient of variation tables for the Person file for 2012.

10.7 Mean Bootstrap Method for Variance Estimation

In order to determine the quality of the estimate and to calculate the CV, the standard deviation must be calculated. Confidence intervals also require the standard deviation of the estimate. The CTUMS uses a multi-stage survey design and calibration, which means that there is no simple formula that can be used to calculate variance estimates. Therefore, an approximate method was needed. The mean bootstrap method is used because the sample design and calibration needs to be taken into account when calculating variance estimates. The mean bootstrap method does this, and with the use of the Bootvar program, discussed in the next section, is a method that is fairly easy for users.

The CTUMS uses the mean bootstrap method described by W. Yung (Yung, W. (1997b). Variance estimation for public use microdata files. *Proceedings of Symposium 1997: New Directions in Surveys and Censuses*, Statistics Canada).

Independently, in each stratum, a simple random sample of $(n-1)$ of the n units in the sample is selected with replacement. Note that since the selection is with replacement, a unit may be chosen more than once. This step is repeated R times to form R bootstrap samples. An average initial bootstrap weight based on the R samples is calculated for each sample unit in the stratum. The entire process (selecting simple random samples, recalculating weights for each stratum) is repeated B times, where B is large, yielding B different initial bootstrap weights. The CTUMS uses $R = 20$ and $B = 250$, to produce 250 bootstrap weights.

These weights are then adjusted according to the same weighting process as the regular weights: non-response adjustment, calibration and so on. The end result is 250 final mean bootstrap weights for each unit in the sample. The variation among the 250 possible estimates based on the 250 mean bootstrap weights are related to the variance of the estimator based on the regular weights and can be used to estimate it. There are a number of reasons why a user may need to calculate the coefficient of variation of estimates with the mean bootstrap method. A few are given below.

- First, if a user wishes to have estimates at a geographic level smaller than the province (for example, at the urban or rural level), then the Approximate Sampling Variability Tables provided are not adequate. Coefficients of variation of these estimates may be obtained using "domain" estimation techniques through the Bootstrap variance program.

- Second, should a user require more sophisticated analyses such as estimates of coefficients from linear regressions or logistic regressions, the Approximate Sampling Variability Tables will not provide correct associated coefficients of variation. Although some standard statistical packages allow sampling weights to be incorporated in the analyses, the variances that are produced often do not properly take into account the design and/or calibration of the weights, whereas the Bootstrap variance program does.

- Third, for estimates of quantitative variables, separate tables are required to determine their sampling error.

10.8 Statistical Packages for Variance Estimation

Statistics Canada has developed a program that can perform bootstrap variance estimation: the Bootvar program.

The Bootvar program is available in SAS or SPSS format. It is made up of macros that compute variances for totals, ratios, differences between ratios and for linear and logistic regression.

Bootvar may be downloaded from Statistics Canada's Research Data Centre (RDC) website. Users must accept the Bootvar Click-Wrap Licence before they can read the files. There is a document on the site explaining how to adapt the system to meet users' needs.

SAS: http://www.statcan.gc.ca/rdc-cdr/bootvar_sas-eng.htm
SPSS: http://www.statcan.gc.ca/rdc-cdr/bootvar_spss-eng.htm

10.8.1 Other Packages

A survey weight variable with a corresponding set of 250 mean bootstrap weight variables are provided with the CTUMS data files in order that a full design-based approach may be taken for doing analysis with the data.

A design-based approach to analysis first involves using the survey weight variable for obtaining weighted estimates of the quantities of interest. Then, additional information about the survey design is used in order to make estimates of the variances and covariances (the variance that is estimated in a design-based approach is the variability

in an estimate due to resampling by exactly the same design from the same finite population) of these estimated quantities. In the case of the CTUMS PUMFs, this additional information is in the form of 250 mean survey bootstrap weight variables, where each mean bootstrap weight is derived from 20 independent survey bootstrap samples. The design-based estimates and variance estimates can then be used for making the inferences required in the analysis.

The form of a mean bootstrap variance estimate can be described briefly as follows:

Let $\hat{\beta}$ be the weighted estimate of the quantity of interest, β, computed using the survey weight variable w, and let $\hat{\beta}^{(b)}$ be an estimate obtained in exactly the same manner, except for substituting the b th mean bootstrap weight variable $w^{(b)}$ for the survey weight variable w, b =1,2,…250. This yields mean bootstrap estimates $\hat{\beta}^{(1)},\ldots,\hat{\beta}^{(250)}$ of β. Then the usual mean bootstrap estimate of the variance of $\hat{\beta}$ is

$$\hat{V}_B(\hat{\beta}) = \frac{20}{250}\sum_{b=1}^{250}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)^2 \qquad (1)$$

If $\hat{\beta}$ is a vector instead of a single value, such as if $\hat{\beta}$ is the set of coefficients of a model, then the matrix of estimates of the variances and covariances of the elements of $\hat{\beta}$ is $\hat{V}_B(\hat{\beta}) = \frac{20}{250}\sum_{b=1}^{250}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)\left(\hat{\beta}^{(b)} - \hat{\beta}\right)'$. (The value "20" in the formulae is due to the fact that each CTUMS mean bootstrap weight is created from 20 bootstrap samples. The value "250" in the formula is due to the fact that we have 250 different mean bootstrap weights. If either the number of bootstrap samples used to create each mean bootstrap weight variable or the number of mean bootstrap weight variables should change from 20 and 250 respectively, then the values in formula (1) would need to change.

Mean bootstrapping is just one replication approach that may be used in order to obtain design-based variance estimates with survey data. While several commercial software packages for design-based analysis offer replication approaches for variance estimation, they usually do not specify mean bootstrapping as one of these approaches. However, due to the similarity in the form of the variance estimate for the mean bootstrap and for the particular replication method called BRR with a Fay adjustment, programs that can carry out variance estimation by this latter approach with user-supplied replication weights can be used to obtain mean bootstrap variance estimates. In particular, in these software, the 250 mean bootstrap weights provided in the CTUMS PUMFs need to be designated as 250 BRR weights and the Fay adjustment factor must be given the value of $1-\sqrt{1/20} \approx 0.7764$.

In the sections below, instructions will be given for implementing mean bootstrap variance estimation with the CTUMS PUMFs data, using three different commercial software packages that can carry out some design-based analysis for BRR with a Fay adjustment:

- Stata 9 or 10,
- SUDAAN and
- WesVar.

These methods are adapted for the CTUMS from a paper by Owen Phillips "*Using bootstrap weights with Wes Var 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. In all the CTUMS cycles where mean bootstrap weights are provided, the names given to these bootstrap variables in the user documentation are **wrpp0001** to **wrpp0250** for the person level files and **wrhp0001** to **wrhp0250** for the household level files. The name of the survey weight variable is **wtpp** or **wthp** respectively.

Stata 9 or 10

Beginning with Version 9, the commercial software package Stata added some replication approaches for carrying out design-based variance estimation in its survey analysis commands. One replication approach offered is the BRR approach with a Fay adjustment, and it is this approach that would be specified when analyzing the CTUMS data.

In order to specify this approach, the following is recommended:

1. Before using any of the survey analysis commands, use a "svyset" statement to declare the data to be survey data, to designate the variables that contain information about the survey design and to specify the method for variance estimation. Settings made by "svyset" are saved with a dataset when (or if) a dataset is saved. The form of the svyset statement to be used with a CTUMS analysis dataset would have the following form:

   svyset [pweight=wtpp], vce(brr) fay(.7764) brrweight(wrpp0001-wrpp0250) mse

   Declaring **pweight=wtpp** tells Stata that the survey weight (which is often called the probability weight) is the variable **wtpp**. The option **vce(brr)** states that the variance estimation approach to use is BRR. The option **fay(.7764)** states that the BRR variance estimation approach is to use a Fay's adjustment of .7764. The option **brrweight(wrpp0001-wrpp0250)** states that the names of the BRR weight variables are **wrpp0001**, **wrpp0002**, …, **wrpp0250**. This option can also be designated as **brrweight(wrpp0*)** provided there are no variables other than the bootstrap weight variables whose names begin with "wrpp0".

   Finally, the **mse** option tells Stata to calculate the variance using squared differences between bootstrap estimates and the full-sample estimate of the quantities of interest, as shown in equation (1). If this option is not included, Stata uses squared differences between each bootstrap estimate and the mean of all the bootstrap estimates. Both approaches should yield approximately the same result.

2. There is an extensive list of survey analysis commands in Stata, which take a design-based approach in their computations. These commands, described in the Stata documentation, are implemented through the use of the "svy" prefix along with the names of other estimators. For example, **svy: mean** is the command for estimating population and subpopulation means and estimates of variability taking a design-based approach. When the **svyset** statement precedes all survey commands, the survey commands do not have to contain any information about the design-based approach to be taken. It should be noted that, even though most of the commands that allow the "svy" prefix are also the names of commands for non-survey data, what is estimated, what options are available and what can be done through post-estimation change when the "svy" prefix is added.

SUDAAN

SUDAAN is a commercial software package developed by the Research Triangle Institute specifically for analysis of data from complex sample surveys and other observational and experimental studies involving cluster-correlated data. The SAS-callable version of the software is particularly useful to people familiar with SAS. In

Release 9.0 and later, all procedures in SUDAAN can take the BRR approach with a Fay adjustment to estimate variances and covariances.

Specification of the variance estimation approach to be used by SUDAAN is done in the procedure statement for a particular procedure. Additional sample design statements provide further information required by the program. In particular, to carry out mean bootstrapping with CTUMS data, the following is required:

- specify **DESIGN=BRR** in the procedure statement
- include the following WEIGHT statement to identify the survey weight variable: **WEIGHT wtpp**;
- include the REPWGT statement to indicate the names of the mean bootstrap variables on your data file and to give the number of bootstrap samples used to produce each mean bootstrap variable (which is used to calculate the Fay adjustment). In particular, for the CTUMS PUMFs, this REPWGT statement would have the form:

  REPWGT wrpp0001-wrpp0250 / ADJFAY=20;

WesVar

WesVar is a software package produced by Westat which carries out various analyses of survey data using exclusively replication methods for variance estimation. One of the methods offered is BRR with a Fay adjustment. Quoting heavily from Phillips (2004), in WesVar, the variance estimation method is specified when creating a new WesVar data file.

The resulting file is then used to define workbooks where table and regression requests are carried out. To define a WesVar data file with mean bootstrap weights:

- move the replicate weight variables (i.e., wrpp0001 to wrpp0250) to the *Replicates* box.
- move the survey weight variable (i.e., wtpp) to the *Full sample* box.
- for the mean bootstrap, specify the *Method* as Fay and specify Fay_K=.7764.
- move analysis variables to the *Variables* box, a unique identifier to the ID box (optional), and save the file.

Phillips (2004) illustrates these instructions with an example using data from the General Social Survey, Cycle 14.

11.0 Weighting

For the microdata file, statistical weights were placed on each record to represent the number of sampled households or persons that the record represents. One weight was calculated for each household and a separate weight was calculated and provided on a different file, for each person.

The weighting for the CTUMS consisted of several steps:

- calculation of a basic weight,
- adjustments for non-response,
- an adjustment for selecting one or two persons in the household,
- dropping out-of-scope records and finally
- an adjustment to make the populations estimates consistent with known province-age-sex totals from the Census projected population counts for persons 15 years and over.

11.1 Weighting Procedures for the Household and Person Files

1. Calculate telephone weight

Each telephone number in the sample was assigned a basic weight, W_1, equal to the inverse of its probability of selection.

$$W_1 = \left(\frac{Total\ number\ of\ \ possible\ sampled\ telephone\ numbers\ \ in\ province-month}{Number\ of\ \ sampled\ telephone\ numbers\ in\ province-month} \right)$$

There were 203,156 telephone numbers in the sample with assigned weights.

2. Adjust for non-resolved telephone numbers

There were 13,036 telephone numbers that were not resolved, leaving 190,120 resolved telephone numbers. The unresolved telephone numbers were not determined to belong to a household, business or out-of-scope. Each telephone number had a flag indicating whether it was expected to be a residential, business, or unknown type of telephone number, and a flag indicating whether or not it was screened out before collection as a non-working or business number. The adjustment for the unresolved telephone numbers was done within province-month, the expected line type, and whether or not the number was sent to the field.

For each province-month-expected line type-sent,

$$W_2 = W_1 * \left(\frac{\sum W_1\ \ for\ resolved\ telephone\ numbers + \sum W_1\ \ for\ unresolved\ telephone\ numbers}{\sum W_1\ \ for\ resolved\ telephone\ numbers} \right)$$

3. Remove out-of-scope telephone numbers

Telephone numbers corresponding to businesses, out-of-service numbers, or out-of-scope numbers, such as cottage telephone numbers, were dropped after the non-resolved adjustment had been applied. Note that if household or person data existed then the telephone number was assumed to be a household. There were 128,909 out-of-scope telephone numbers and 61,413 telephone numbers belonging to a household.

4. Adjust for non-response of number of telephone lines in the household

The number of telephone lines in the household was calculated. If the number of different telephone lines within the household could not be calculated but household or person data existed, then it was imputed as one in order to retain good data. After imputation, there were 8,483 telephone numbers that were still missing the number of lines. Thus, there were 52,930 households with the number of lines calculated or imputed. The adjustment was done within province-month.

$$W_3 = W_2 * \left(\frac{\sum W_2 \ for\ households\ with\ number\ of\ lines + \sum W_2 \ for\ households\ missing\ number\ of\ lines}{\sum W_2 \ for\ households\ with\ number\ of\ lines} \right)$$

5. Calculate household weight with multiple telephone lines adjustment

Weights for households with more than one telephone line (with different telephone numbers) were adjusted downwards to account for the fact that such households have a higher probability of being selected. The weight for each household was divided by the number of distinct residential telephone lines (up to a maximum of 4) that serviced the household. The adjustment was done within province-month.

$$W_4 = \left(\frac{W_3}{Number\ of\ in-scope\ telephone\ lines\ in\ the\ household} \right)$$

6. Adjust for non-responding households

Household respondents responded to the questions on their smoking habits. If these questions were not sufficiently answered, perhaps refused or only partially answered, then the household was considered a non-respondent. There were 866 non-respondents. Thus, 52,930 in-scope household weights were used and adjusted within province-month.

$$W_5 = W_4 * \left(\frac{\sum W_4 \ for\ household\ respondents + \sum W_4 \ for\ household\ non-respondents}{\sum W_4 \ for\ household\ respondents} \right)$$

11.2 Weighting Procedures for the Household File

7. Adjust to known external household province-month totals

An adjustment was made to the household weights on records within each province and month, in order to make household estimates consistent with known external household counts. The adjustment factor for province-month (P-M) was defined as:

$$W_6 = W_5 * \left(\frac{Known\ external\ household\ count\ in\ P-M}{\sum W_5 \ for\ responding\ households\ in\ the\ sample\ in\ P-M} \right)$$

The household weights, W_6, obtained after this step, were considered final and appear on the household microdata file.

11.3 Weighting Procedures for the Person File

8. Remove households with no selected persons

There were 32,229 households where no one was selected to continue with the tobacco use survey or a selected person was not retained because of sub-selection of individuals. These households were dropped because they had no person level data. About 70% of selected respondents aged 25 and over were screened out. There were 19,835 households with selected persons. There were 16,422 households with one person selected and 3,413 with two people selected.

9. Calculate group weight

All of the in-scope responding households with completed rosters (i.e. no missing ages) were assigned group weights. From the roster, three flags were assigned to indicate the presence of a person in the following age groups: 15 to 19, 20 to 24, and 25 and over. If one or two age group categories were represented then an individual was selected from each age group present (i.e. the probability of selection of the age group was 1).Thus, the weight was not inflated. However, if three age groups were represented, then two people were selected, so the probability of selecting the age group is 2 out of the 3 groups. Thus, the weight is inflated by its inverse.

If 1 or 2 age groups were represented then $W_6 = W_5$.

If all 3 age groups were represented then $W_6 = W_5 * 3/2$.

10. Assign household weights to selected persons

The 16,422 + 2(3,413) = 23,248 selected persons are associated with in-scope responding households and keep the corresponding weight, W_6.

11. Calculate selected person sub-weight

All in-scope individuals were assigned weights. The weight is inflated by the number of people within the selected age group and the inverse of the sub-sampling factor.

$$W_7 = W_6 * \left(\frac{Number\ of\ individuals\ in\ selected\ age\ group}{Sub-sampling\ \ factor} \right)$$

The sub-sampling factor was 1 for age groups 15 to 19 and 20 to 24 where there were two age groups in the household and (3-Subsampling factor)/2 if there were three age groups in the household. The sub-sampling factor was pre-assigned for the 25 and over age group and varied from 19.1% to 28.6%, depending on the province.

12. Adjust for non-responding individuals

The Person file includes records of individual respondents who completed the questions on smoking habits and gave a date of birth corresponding to the age given in the roster. There were 3,962 non-respondents.

Thus, 19,286 in-scope individual weights were used and adjusted within province, age groups derived from the roster (15 to 19, 20 to 24, 25 to 44, 45 to 64, 65 and over) and sex.

$$W_8 = W_7 * \left(\frac{\sum W_7\ \ for\ person\ respondents + \sum W_7\ \ for\ person\ non-respondents}{\sum W_7\ \ for\ person\ respondents} \right)$$

13. 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) Monthly population totals for each province, and

2) Population totals by province, sex and the following age groups: 15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69 and 70 and over. These totals were averaged over the survey period.

The method called generalized regression (GREG) estimation was used to modify the weights to ensure that the survey estimates agreed with the external totals simultaneously along the two dimensions.

The person weights obtained after this step were considered final and appear on the person microdata file.

12.0 Questionnaire

Refer to CTUMS2012_ QuestE.pdf for the English questionnaires used in 2012.

13.0 Record Layouts with Univariate Frequencies

13.1 Record Layout with Univariate Frequencies – Household File

Refer to CTUMS2012_HH_CdBk.pdf for the record layout with univariate counts for the Household file for 2012.

13.2 Record Layout with Univariate Frequencies – Person File

Refer to CTUMS2012_ PR_CdBk.pdf for the record layout with univariate counts for the Person file for 2012.