CANADIAN TOBACCO USE MONITORING SURVEY - 2000

CYCLE 2 FILES

JULY - DECEMBER 2000

Table of Contents

1.0 Introduction 1

2.0 Background 3

3.0 Objectives 5

4.0 Concepts and Definitions 7

5.0 Survey Methodology 9
5.1 Population Coverage 9
5.2 Stratification 9
5.3 Sample Design and Allocation 10
5.4 Sample Selection 10

6.0 Data Collection 13
6.1 Questionnaire Design 13
6.2 Data Collection and Editing 13

7.0 Data Processing 15
7.1 Data Capture 15
7.2 Editing 15
7.3 Creation of Derived Variables 15
7.4 Weighting 16
7.5 Suppression of Confidential Information 16

8.0 Data Quality 17
8.1 Household Response Rates - July to December 2000 18
8.2 Person Response Rate - July to December 2000 19
8.3 Survey Errors 21
8.4 Total Non-response 21
8.5 Partial Non-response 22
8.6 Coverage 22
8.7 Measure of Sampling Error 22

9.0 Guidelines for Tabulation, Analysis and Release 25
9.1 Rounding Guidelines 25
9.2 Sample Weighting Guidelines for Tabulation 26
9.2.1 Definitions of types of estimates: Categorical vs. Quantitative 26
9.2.2 Tabulation of Categorical Estimates 27
9.2.3 Tabulation of Quantitative Estimates 28
9.3 Guidelines for Statistical Analysis 28
9.4 C.V. Release Guidelines 29

10.0 Approximate Sampling Variability Tables 31
10.1 How to use the C.V. tables for Categorical Estimates 34
10.1.1 Examples of using the C.V. tables for Categorical Estimates 36

10.2 How to use the C.V. tables to obtain Confidence Limits 39
10.2.1 Example of using the C.V. tables to obtain confidence limits 40
10.3 How to use the C.V. tables to do a t-test 40
10.3.1 Example of using the C.V. tables to do a t-test 41
10.4 Coefficients of Variation for Quantitative Estimates 41
10.5 Release cut-off's for the CTUMS - Household File 42
10.6 Release cut-off's for the CTUMS - Person File 43
10.7 C.V. Tables - Household file 44
10.8 C.V. Tables - Person File 44

11.0 Weighting 45
11.1 Weighting Procedures for Both the Household and Person File 45
11.2 Weighting Procedures for the Household File 47
11.3 Weighting Procedures for the Person File 47

12.0 Questionnaire 51

13.0 Record Layout and Univariates 53
13.1 Record Layout and Univariates - Household 53
13.2 Record Layout and Univariates - Person 53

1.0 Introduction

The Canadian Tobacco Use Monitoring Survey was conducted by Statistics Canada from July to December 2000 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

Eddy Ross
Special Surveys Group, Statistics Canada
Section D7
5th floor, Jean Talon Building
Tunney's Pasture
Ottawa, Ontario K1A 0T6
Telephone (613) 951-3240
Fax: (613) 951-0562
Email: rossedd@statcan.ca

Health Canada

Murray Kaiserman
Tobacco Control Programme
123 Slater, 7th Floor,
Ottawa, Ontario K1A OK9
Telephone: (613) 954-5851
Fax: (613) 941-1551
Email: Murray_Kaiserman@hc-sc.gc.ca

Anne Zaborski
Tobacco Control Programme
123 Slater, 7th Floor,
Ottawa, Ontario K1A OK9
Telephone: (613) 954-0152
Fax: (613) 941-1551
Email: Anne_Zaborski@hc-sc.gc.ca

IT IS IMPORTANT FOR USERS TO BECOME FAMILIAR WITH THE CONTENTS OF THIS DOCUMENT BEFORE PUBLISHING OR OTHERWISE RELEASING ANY ESTIMATES DERIVED FROM THE MICRODATA FILE OF THE CANADIAN TOBACCO USE MONITORING SURVEY.

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. The earlier surveys were about smoking only, but in recent years smoking has been one topic on broader-based health surveys. These surveys have been conducted fairly infrequently over the last 10 years.

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 took place, it was difficult for Health Canada or other interested analysts to measure exactly the impact of the change. Health Canada wants to be positioned to better respond to future changes by having a data collection vehicle in place to continuously monitor smoking behaviour.

As Health Canada wants to be positioned to better respond to future changes, the Canadian Tobacco Use Monitoring Survey was design to provide Health Canada and its partners/stakeholders with continual and predictable data on tobacco use and related issues.

This release is the fourth release of the Canadian Tobacco Monitoring Survey. In January 2000, the data collected from February to June 1999 were released and in October 2000 data for the July to December 1999 as well as a summary for the whole year 1999 were released. As well, data for the February - June 2000 were released in February 2001. This release includes the data for the period of July - December 2000.

NOTE: This fourth release also includes the data for the full year 2000. The guide for the annual data is provided separately.

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. This objective differs from that of the National Population Health Survey which collects smoking data from a longitudinal sample to measure which individuals are changing their smoking behaviour, the possible factors which contribute to change, and the possible risk factors related to starting smoking and smoking duration. Because the NPHS collects data every two years and releases the data about a year after completing the collection cycle, it does not meet Health Canada's need for continuous coverage in time, rapid delivery of data, or sufficient detail of the most at-risk populations, namely 15-24 year olds.

The Tobacco Use Monitoring Survey will allow Health Canada to look at smoking prevalence by province-sex-age group, for age groups 15-19, 20-24, 25-34, 35-44 and 45+ on a semi-annual basis. Changes in smoking prevalence of about 3% or higher will be detectable on an annual basis, within age groups at the national level. The data included in these files represent data collected for the last six months of 2000. Data will continue to be collected on an on-going basis depending on availability of funds.

4.0 Concepts and Definitions

Since the Canadian Tobacco Use Monitoring Survey 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.

5.0 Survey Methodology

The first cycle of the Canadian Tobacco Use Monitoring Survey was administered between July 4 and December 30, 2000 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 Canadian Tobacco Use Monitoring Survey 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 telephones were excluded from the sample population. People without telephones account for less than 3% of the target population. However, the survey estimates have been weighted to include persons without telephones.

5.2 Stratification

In order to ensure that people from all parts of Canada were represented in the sample, each of the ten 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 15-19 and 20-24 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 ten provinces, an ***equal number of respondents in each province*** is targeted. For the first six months of collection, the target was to get responses from 5,000 individuals age 15-24 and 5,000 individuals age 25+ across Canada, or 500 individuals in each age group per province . The initial sample size of phone 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, household with only 25+ years olds present were sub-sampled. It was estimated that a total of about 148,000 telephone numbers per year will be needed to get the 20,000 respondents per year. This assumed a 75% response rate and about 20% of households having individuals aged 15-24; the hit rate varies substantially by province, with an expected overall average of about 43%.

5.4 Sample Selection

The sample for the Canadian Tobacco Use Monitoring Survey was generated using a refinement of RDD sampling called the Elimination of Working Banks (ENWB). Within each province-stratum combination, a list of working banks (area code + next 5 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 2-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 (ie. not currently or never in service), to have a chance of being in the sample.

Each phone number in the 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 phone 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 15-19 or 20-24 would be selected within a household, if they exist. The reason for this is that about 76% of all households in Canada are made up of only people over 25 years of age; another 20% consist of people over 25 living with people in either the 15-19 or 20-24 age group; and only 4% of households contain no one aged over 25. If all ages were selected with equal probability and retained, the 25+ 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+ 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 15-19, 20-24, and 25+ 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 15-19 then one person is selected at random.
2. If everyone in the household is 20-24 then one person is selected at random.
3. If everyone in the household is 25+ 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 15-19 and the rest are 20-24 then two people are selected at random, one from each age group.
5. If some household members are 15-19 and the rest are 25+ then two people are selected at random, one from each age group; however, this selected person is retained for only a proportion of the cases.
6. If some household members are 20-24 and the rest are 25+ then two people are selected at random, one from each age group; however, this selected person is retained for only a proportion of the cases.
7. If all three age groups are represented in the household, then two age groups are selected at random and then rule 4, 5, or 6 applies.

6.0 Data Collection

6.1 Questionnaire Design

The question 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". The questionnaire contains all the same questions as for Cycle 1 of the 2000 CTUMS.

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.

6.2 Data Collection and Editing

Final testing of the CATI application took place in January 2000. Beginning in February 2000, data collection was conducted on a monthly basis.

Data were collected using Computer-Assisted Interviewing techniques (CATI). Our CATI system has a number of generic modules which can quickly be adapted to most types of surveys. A front-end module contains a set of standard response codes for dealing will all possible call outcomes, as well as the associated scripts to be read by the interviewers. A standard approach set up for introducing our 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 right into the application to check consistency of responses, identify and correct outliers, and 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 were made to ensure that the same set of interviewers is used

each month. This minimized training and yield better and more consistent data quality.

The cases were distributed to 4 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 day and different days of the week to maximize the probability of contact. There were a maximum of 17 call attempts per case; once the maximum was reached, the case was reviewed by a senior interviewer who determined if additional calls would be made.

7.0 Data Processing

The main output of the Canadian Tobacco Monitoring Survey are two "clean" microdata files one for the household level information and one for the person level information. This section presents a brief summary of the processing steps involved in producing these files.

7.1 Data Capture

As the data was collected using CATI, there was no need for a separate data capture system as 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 all monthly files into a single file. Any 'out-of-range' values on the data file were replaced with blanks. This process was designed to make further editing easier.

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 edits performed involved a lack of information in questions which should have been answered. For example, if a respondent refused to answer a question, any question, which would have had an answer if there was no refusal, is assigned a "not-stated" code.

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 are average number smoked daily, income adequacy status, etc.

7.4 Weighting

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

Weighting involves calculating how many people each respondent in the survey represents. This weight must be used to derive estimates from the microdata file. For example, if the number of people in Canada who smoke daily is estimated, it is done by selecting the records referring to people with that characteristic (Q010 = '1') and summing the weights of those records. A separate weight for households and persons is calculated every 6 months.

7.5 Suppression of Confidential Information

It should be noted that the 'Public Use' microdata files described above differ in a number of important respects from the survey 'master' files held by Statistics Canada. These differences are the result of actions taken to protect the anonymity of individual survey respondents. 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 Section 9 of this document.

Geographic Identifiers: The survey master data file includes explicit geographic identifiers for province, and stratum (CMA, non-CMA, Toronto, Montreal). It also contains the respondent's postal code. The survey public use microdata file does contain the geographic identifiers for province and stratum, but does not contain geographic identifiers (i.e. postal code) below the stratum level. Where possible, the postal code can be linked to other geographic identifiers and the data can be tabulated to this new geographic level; the link and tabulations would be performed at a cost-recovery rate.

8.0 Data Quality

For the Canadian Tobacco Use Monitoring Survey, the response rates that were computed included the following:

- Telephone Resolved Rate, where telephone numbers that were confirmed as residential, business or out of scope were considered resolved,
- Hit Rate, where resolved phone numbers that were confirmed residential, had **valid household data**, or **had valid person data** were considered to belong to a household.

The Telephone Resolved Rate and the Hit Rate apply to both the household file and the person file.

- Household Response Rate for the household file, where households with all ages provided for everyone in the roster and **valid household data** were considered a response.

This Household Response Rate applies only to the household file.

- Household Response Rate for the person file, where households that were confirmed residential or **valid person data** existed with a completed roster were considered a response,
- Roster Completion Rate, where households with a completed roster containing ages for each person in the roster were considered a response,
- Combined Household Rate for person file, where the households with a valid completed roster were considered a response,
- Person Response Rate, where records with all ages provided for everyone in the roster and valid person data exists were considered a response.

The Household Response Rate for the person file, Roster Completion Rate, Combined Household Rate and Person Response rate apply only to the person file.

Telephone Resolved Rate and Hit Rate by Province

Provinces	Total No. of Phone No. Generated	Total Resolved	Resolved Rate	Total Households	Hit Rate
Newfoundland	6,119	6,046	98.8%	2,198	36.4%
Prince Edward Island	5,912	5,738	97.1%	2,534	44.2%
Nova Scotia	6,028	5,916	98.1%	2,633	44.5%
New Brunswick	7,109	7,030	98.9%	2,170	30.9%
Quebec	5,280	5,224	98.9%	2,620	50.2%
Ontario	5,598	5,517	98.6%	2,551	46.2%

Manitoba	6,000	5,989	99.8%	2,620	43.7%
Saskatchewan	5,640	5,637	99.9%	2,320	41.2%
Alberta	5,208	5,186	99.6%	2,563	49.4%
British Columbia	5,489	5,369	97.8%	2,727	50.8%
Total	58,383	57,652	98.7%	24,936	43.3%

8.1 Household Response Rates - July to December 2000

A **household respondent** must complete the roster with no age refusals, and valid household data must exist. There were 2,943 (11.8%) households that were non-responding.

Household Response Rate by Province.

Province	Total Households	Responding Households	Household Response Rate
Newfoundland	2,198	2,040	92.8%
Prince Edward Island	2,534	2,284	90.1%
Nova Scotia	2,633	2,388	90.7%
New Brunswick	2,170	1,930	88.9%
Quebec	2,620	2,238	85.4%
Ontario	2,551	1,970	77.2%
Manitoba	2,620	2,319	88.5%
Saskatchewan	2,320	2,075	89.4%
Alberta	2,563	2,262	88.3%
British Columbia	2,727	2,487	91.2%
Canada	24,936	21,993	88.2%

Household Response Rate by Survey Month.

Survey Month	Total Households	Responding Households	Household Response Rate
July	4,157	3,674	88.4%
August	4,166	3,646	87.5%
September	4,231	3,762	88.9%
October	4,154	3,704	89.2%
November	4,151	3,615	87.1%
December	4,077	3,592	88.1%
Canada	24,936	21,993	88.2%

8.2 Person Response Rate - July to December 2000

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

- The phone number corresponding to the selected person belonged to a household.
- The roster was completed with no individual age refusals.
- The selected person must have been 15 years old or older, when the survey was conducted. The age given in the roster was verified with the date of birth given by the selected person.
- The selected person must have answered the key questions on smoking habits, at minimum.

There were 12,477 households, in which, household data was collected but nobody was selected to continue with the tobacco use survey. (See “Sample Selection” for more information (section 5.4)) Of the remaining households, 8,143 had one person selected while 1,464 had two selected people. There were 309 people that refused to complete the survey (2.8%) and 870 other non-respondents (7.8%).

Household Response Rate, Roster Completion Rate and Combined Household Response Rate by Province

Province	Total HHLDs	Total HHLD with Rosters	HHLD Response Rate	HHLD with Valid Roster Data	Roster Completion Rate	Combined HHLD Response Rate
Newfoundland	2,198	2,109	96.0%	2,049	97.2%	93.2%
Prince Edward Island	2,534	2,377	93.8%	2,286	96.2%	90.2%
Nova Scotia	2,633	2,492	94.6%	2,393	96.0%	90.9%

New Brunswick	2,170	2,025	93.3%	1,942	95.9%	89.5%
Quebec	2,620	2,345	89.5%	2,248	95.9%	85.8%
Ontario	2,551	2,107	82.6%	1,979	93.9%	77.6%
Manitoba	2,620	2,403	91.7%	2,335	97.2%	89.1%
Saskatchewan	2,320	2,145	92.5%	2,092	97.5%	90.2%
Alberta	2,563	2,345	91.5%	2,294	97.8%	89.5%
British Columbia	2,727	2,546	93.4%	2,496	98.0%	91.5%
Canada	24,936	22,894	91.8%	22,114	96.6%	88.7%

Person Response Rate by Province

Province	Total Persons Selected	Total Persons Responses	Person Response Rate
Newfoundland	1,158	1,041	89.9%
Prince Edward Island	1,274	1,141	89.6%
Nova Scotia	1,177	1,060	90.1%
New Brunswick	1,019	880	86.4%
Quebec	1,121	1,021	91.1%
Ontario	1,028	880	85.6%
Manitoba	1,121	1,017	90.7%
Saskatchewan	1,021	923	90.4%
Alberta	1,191	1,065	89.4%
British Columbia	991	894	90.2%
Canada	11,101	9,922	89.4%

Person Response Rate by Survey Month

Month	Total Persons Selected	Total Persons Responding	Person Response Rate
July	1,929	1,678	87.0%
August	1,846	1,641	88.9%
September	1,856	1,643	88.5%
October	1,787	1,614	90.3%
November	1,820	1,666	91.5%
December	1,863	1,680	90.2%

Canada	11,101	9,922	89.4%

Target Number of Respondents and Person Response Rate by Age Group

Age Group	Target Number of Respondents	Total Persons Selected	Total Persons Responding	Person Response Rate
15-19	2500	2,949	2,655	89.3%
20-24	2,500	2,525	2,208	88.7%
25 +	5,000	5,627	5,059	89.8%
Canada	10,000	11,101	9,922	89.4%

8.3 Survey Errors

The survey produces estimates based on information collected from and about a sample of individuals. 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 computer 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 included extensive training of interviewers with respect to the survey procedures and CATI application; monitoring 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 when the selected household or person could not be contacted or 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

Partial non-response to the survey occurred when the respondent refused to answer a question, or could not recall the requested information. Partial non-response is indicated by codes on the microdata file.

8.6 Coverage

As mentioned in Section 5.1 (Population Coverage), less than 3% of households in Canada do not have telephones. Individuals living in non-telephone 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 ownership.

8.7 Measure 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. 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 (C.V.) 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 survey results, one estimates that 31% of Canadians are currently cigarette smokers, and this estimate is found to have standard error of .0056. Then the coefficient of variation of the estimate is calculated as:

$$\left(\frac{.0056}{.31}\right) \; x \; 100\% \; ' \; 1.8\%$$

9.0 Guidelines for Tabulation, Analysis and Release

This section 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 tapes. 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 tapes 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 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 sampling weight.

If proper weights are not used, the estimates derived from the microdata tapes 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.

9.2.1 Definitions of types of estimates: Categorical vs. 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.

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, and 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 THIRTY DAYS, did you smoke any cigarettes ?
R: Yes / No

Q: What prompted you to quit smoking?
R: Current health problems / Smoking-related illness of friend/ Pregnancy / Doctor Advice / Concern for Future Health / Illness / Accident

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 on Saturday per person. The numerator is an estimate of the total number of cigarettes smoked on Saturdays, and its denominator is the number of persons who reported smoking on Saturday.

Example of Quantitative Question :

Q: Thinking back over the past 7 days, starting with yesterday, how many cigarettes did you smoke on Monday ?
R: |_|_| Cigarettes

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

9.2.2 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 X/Y are obtained by:

(a) summing the final weights of records having the characteristic of interest for the numerator (X),
(b) summing the final weights of records having the characteristic of interest for the denominator (Y), then
(c) dividing the numerator estimate by the denominator estimate.

9.2.3 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 on Saturdays, multiply the value reported in Q090SAT (number of cigarettes smoked on Saturday) by the final weight for the record, then sum this value over all records with Q090SAT<96 (all respondents who reported a value in this field).

To obtain a weighted average of the form X/Y, the numerator (X) is calculated as for a quantitative estimate and the denominator (Y) is calculated as for a categorical estimate. For example, to estimate the average number of cigarettes smoked on Saturday,

(a) estimate the total number of cigarettes smoked on Saturday as described above,
(b) estimate the number of people in this category by summing the final weights of all records with Q090SAT < 96, then
(c) divide estimate (a) by estimate (b).

9.3 Guidelines for Statistical Analysis

The Canadian Tobacco Use Monitoring Survey 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 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. Variances for simple estimates such as totals, proportions and ratios (for qualitative variables) are provided in the accompanying 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:

- select all respondents from the file who reported SEX=male

- Calculate the AVERAGE weight for these records by summing the original person weights from the microdata file for these records and then dividing by the number of respondents who reported SEX=male
- for each of these respondents, calculate a RESCALED weight equal to the original person weight divided by the AVERAGE weight
- 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 truly meaningful 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.

9.4 C.V. Release Guidelines

Before releasing and/or publishing any estimate from the Canadian Tobacco Use Monitoring Survey 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 section 8. 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 section 8 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 weighted rounded 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 0.0% - 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 16.6% - 33.3%. Estimates should be flagged with the letter M (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 U (or some similar identifier) and the following warning should accompany the estimates: "The user is advised that . . . (specify the data) . . . do not meet Statistics Canada's quality standards for this statistical program. Conclusions based on these data will be unreliable, and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data."

10.0 Approximate Sampling Variability Tables

In order to supply coefficients of variation 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 "look-up" 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 (C.V.) 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 to be used in the look-up tables which would then apply to the entire set of characteristics.

The table below shows the design effects, sample sizes and population counts by province for the households which were used to produce the Approximate Sampling Variability Tables - “Household” File.

Province	Design Effect	Sample Size	Population
Newfoundland	1.23	2,040	194,010
Prince Edward Island	1.05	2,284	52,338
Nova Scotia	1.08	2,388	364,066
New Brunswick	1.19	1,930	288,174
Quebec	1.18	2,238	3,038,546
Ontario	1.10	1,970	4,366,888
Manitoba	1.18	2,319	428,210
Saskatchewan	1.06	2,075	385,833
Alberta	1.16	2,262	1,108,989
British Columbia	1.20	2,487	1,570,174
Canada	2.69	21,993	11,797,228

The table below shows the design effects, sample sizes and population counts by province which were used to produce the Approximate Sampling Variability Tables - “Person” File.

Region	Age Group	Design Effect	Sample Size	Population
Newfoundland	All	2.00	1041	440881
	15-19	1.41	313	40944
	20-24	1.43	230	39223
	25+	1.38	498	360714
Prince Edward Island	All	1.74	1141	110258
	15-19	1.41	311	10280
	20-24	1.54	245	9605
	25+	1.25	585	90373
Nova Scotia	All	1.92	1060	763131
	15-19	1.54	296	63765
	20-24	1.44	217	62881
	25+	1.36	547	636485
New Brunswick	All	2.08	880	612473
	15-19	1.34	232	51203
	20-24	1.62	199	51373
	25+	1.53	449	509897
Quebec	All	2.10	1021	5985595
	15-19	1.40	253	475612
	25-64	1.44	269	503670
	25+	1.53	499	5006313
Ontario	All	2.24	880	9378456
	15-19	1.72	261	768810
	20-24	1.67	212	760472
	25+	1.43	407	7849174
Manitoba	All	1.98	1017	897059
	15-19	1.53	276	80587
	20-24	2.66	232	78127
	25+	1.37	509	738344

Saskatchewan	All	1.75	923	790414
	15-19	1.33	251	78759
	20-24	1.72	193	73865
	25+	1.33	479	637790
Alberta	All	1.97	1065	2364751
	15-19	1.51	269	222542
	20-24	2.24	232	223281
	25+	1.43	564	1918927
British Columbia	All	1.72	894	3306645
	15-19	1.67	193	270703
	20-24	2.12	179	268143
	25+	1.35	522	2767799
Canada	All	5.84	9922	24496664
	15-19	3.91	2655	2063206
	20-24	3.75	2208	2070642
	25+	4.11	5059	20515816

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. The use of actual variance estimates would allow users to release otherwise “unacceptable” estimates, i.e. estimates with coefficients of variation in the “unacceptable’” range.

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 C.V. tables for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the 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 Possessing a Characteristic (Aggregates)

The coefficient of variation depends only on the size of the estimate itself. On the 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 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 cv's decline in value 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 ($\hat{d} = \hat{X}_1 - \hat{X}_2$) is:

$$s_{\hat{d}} = \sqrt{(\hat{X}_1 a_1)^2 + (\hat{X}_2 a_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and a_1 and a_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 $s_{\hat{d}}/\hat{d}$. 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 deviation 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 R. That is, the standard error of a ratio ($R = \hat{X}_1 / \hat{X}_2$) is:

$$s_{\hat{R}} = \hat{R}\sqrt{a_1^2 + a_2^2}$$

where a_1 and a_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of R is given by s_R/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 cv's 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 C.V. tables for Categorical Estimates

The following 'real life' examples are included to assist users in applying the foregoing rules.

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

Suppose that a user estimates that during the reference period 6,095,719 persons were current smokers (DVSST1 = ‘1') in Canada. How does the user determine the coefficient of variation of this estimate?

(1) Refer to the c.v. table for CANADA.

(2) The estimated aggregate (6,095,719) 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 6,000,000.

(3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, namely, 4.3%.

(4) So the approximate coefficient of variation of the estimate is 4.3%.

The finding that there were 6,095,719 current smokers in the reference period is acceptable and no warning is required.

Example 2 : Estimates of Proportions or Percentages Possessing a Characteristic

Suppose that the user estimates that 3,072,893/12,093,257 = 25.4% of males currently smoke in Canada in the reference period. How does the user determine the coefficient of variation of this estimate?

(1) Refer to the c.v. table for CANADA. The CANADA level tables should be used because it is the smallest table that contains the domain of the estimate, all males in Canada.

(2) Because the estimate is a percentage which is based on a subset of the total population (males), it is necessary to use both the percentage (25.4%) and the numerator portion of the percentage (3,072,893) in determining the coefficient of variation.

(3) The numerator, 3,072,893 does not appear in the left-hand column (the 'Numerator of Percentage' column) so it is necessary to use the figure closet 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 figure closest to it, 25.0%.

(4) The figure at the intersection of the row and column used, namely 6.0% is the coefficient of variation to be used.

(5) So the approximate coefficient of variation of the estimate is 6.0%.

The finding that 25.4% of males currently smoke is acceptable and no warning is required.

Example 3 : Estimates of Differences Between Aggregates or Percentages

Suppose that a user estimates that 3,022,826/12,556,407 = 24.1% of females currently smoke in Canada, while 3,072,893/12,093,257 = 25.4% of males currently smoke in Canada. How does the user determine the coefficient of variation of the difference between these two estimates?

(1) Using the c.v. table for CANADA in the same manner as described in Example 2 gives the c.v. of the estimate for females as 6.0%, and the c.v. of the estimate for males as 6.0%.

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

$$s_{\hat{d}} = \sqrt{(\hat{X}_1 a_1)^2 + (\hat{X}_2 a_2)^2}$$

where $\hat{X}_1$ is estimate 1 (males) $\hat{X}_2$ is estimate 2 (females), and a_1 and a_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

That is, the standard error of the difference $\hat{d}$ = (.254 -.241) = .013 is:

$$s_{\hat{d}} = \sqrt{[(0.254)(0.060)]^2 + [(0.241)(0.060)]^2}$$

$$= \sqrt{(0.00023) + (0.00021)}$$

$$= 0.021$$

(3) The coefficient of variation of $\hat{d}$ is given by $s_{\hat{d}}/\hat{d}$ = .021/.013 = 1.614

(4) So the approximate coefficient of variation of the difference between the estimates is 161.4%. This estimate is considered unacceptable and Statistics Canada recommends not to release these estimates. However, if the user chooses to do so, this estimate must be flagged in the appropriate manner.

Example 4 : Estimates of Ratios

Suppose that the user estimates that there are 272,161 female current smokers in the age group 15-19, while 239,833 male current smokers in the age group 15-19. The user is interested in comparing the estimate of females versus that of males 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 female current smokers in the age group 15-19. The denominator of the estimate ($=\hat{X}_2$) is the number of male current smokers in the age group 15-19.

(2) Refer to the c.v. table for CANADA15-19.

(3) The numerator of this ratio estimate is 272,161. 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, 10.2%

(4) The denominator of this ratio estimate is 239,833. 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, 10.2%

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

$$a_{\hat{R}} \doteq \sqrt{a_1^2 + a_2^2}$$

where a_1 and a_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

That is,

$$a_{\hat{R}} \doteq \sqrt{(0.102)^2 + (0.102)^2}$$

$$\doteq 0.144$$

The obtained ratio of female current smokers in the age group 15-19 versus male current smokers in the age group 15-19 is 272,161/239,833 which is 1.13. The coefficient of variation of this estimate is 14.4%, which means the estimate is acceptable and no warning is required.

10.2 How to use the C.V. 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 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 ($\hat{X}$-k, $\hat{X}$+k) 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 CI:

$$CI_{\hat{X}} = [\hat{X} - t\hat{X}a_{\hat{X}}, \hat{X} + t\hat{X}a_{\hat{X}}]$$

where $a_{\hat{X}}$ 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 = 3 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 "unacceptable", then the confidence interval is "unacceptable" also.

10.2.1 Example of using the C.V. tables to obtain confidence limits

A 95% confidence interval for the estimated proportion of male current smokers (from Example 2) would be calculated as follows.

$\hat{X}$ = 25.4% (or expressed as a proportion = .254)

t = 2

$a_{\hat{X}}$ = 6.0% (.060 expressed as a proportion) is the coefficient of variation of this estimate as determined from the tables.

CI_X = {.254 - (2) (.254) (.060), .254 + (2) (.254) (.060)}

CI_X = {.254 - .030, .254 + .030}

CI_X = {.224, .284}

With 95% confidence it can be said that between 22.4% and 28.4% of males currently smoke.

10.3 How to use the C.V. 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 $\mathbf{X_1}$ and $\mathbf{X_2}$ be sample estimates for 2 characteristics of interest. Let the standard error on the difference $\hat{X}_1 - \hat{X}_2$ be $s_{\hat{d}}$.

If $\frac{\hat{X}_1 - \hat{X}_2}{s_{\hat{d}}}$ 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 characteristics are significant.

10.3.1 Example of using the C.V. tables to do a t-test

Let us suppose we wish to test, at 5% level of significance, the hypothesis that there is a difference between the proportion of male current smokers and the proportion of female current smokers. From Example 3, the standard error of the difference between these two estimates was found to be 0.043. Hence,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{s_{\hat{d}}} = \frac{.254 - .241}{.021} = \frac{.013}{.021} = 0.62$$

Since t = 0.62 is less than 2, it must be concluded that there is no significant difference between the two estimates at the 0.05 level of significance.

10.4 Coefficients 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 "unacceptable", the quantitative estimate will not be either. For example, the coefficient of variation of the total number of cigarettes smoked on Monday 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", then the coefficient of variation of the corresponding quantitative estimate will also be "unacceptable".

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 Release cut-off's for the CTUMS - Household File

The minimum size of the estimate 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.

Region	Acceptable CV < 16.15	Marginal CV in 16.5-33.0	Unacceptable CV > 33.0
Newfoundland	4,000 & +	1,000-4,000	under 1,000
Prince Edward Island	1,000 & +	0-1,000	under 0
Nova Scotia	6,000 & +	1,500-6,000	under 1,500
New Brunswick	6,500 & +	1,500-6,500	under 1,500
Quebec	57,500 & +	14,500-57,500	under 14,500
Ontario	88,000 & +	22,500-88,000	under 22,000
Manitoba	8,000 & +	2,000-8,000	under 2,000
Saskatchewan	7,000 & +	2,000-7,000	under 2,000
Alberta	20,500 & +	5,000-20,500	under 5,000
British Columbia	27,500 & +	7,000-27,500	under 7,000
Canada	53,000 & +	13,000-53,000	under 13,000

10.6 Release cut-off's for the CTUMS - Person File

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

Table of Release Cut-offs

Region	Age Group	Acceptable CV < 16.15	Marginal CV in 16.5-33.0	Unacceptable CV > 33.0
Newfoundland	All	29,000 & +	7,500-29,000	under 7,500
	15-19	6,000 & +	1,500-6,000	under 1,500
	20-24	7,500 & +	2,000-7,500	under 2,000
	25+	33,500 & +	9,000-30,500	under 9,000
Prince Edward Island	All	6,000 & +	1,500-6,000	under 1,500
	15-19	1,500 & +	500-1,500	under 500
	20-24	2,000 & +	500-2,000	under 500
	25+	6,500 & +	1,500-6,500	under 1,500
Nova Scotia	All	47,500 & +	12,500-47,500	under 12,500
	15-19	10,000 & +	3,000-10,000	under 3,000
	20-24	12,500 & +	3,500-12,500	under 3,500
	25+	53,500 & +	15,000-53,500	under 14,000
New Brunswick	All	49,000 & +	13,000-49,000	under 13,000
	15-19	9,000 & +	2,500-9,000	under 2,500
	20-24	12,000 & +	3,500-12,000	under 3,500
	25+	56,500 & +	15,000-56,500	under 15,000
Quebec	All	420,500 & +	109,000-420,500	under 109,000
	15-19	80,500 & +	22,500-80,500	under 22,500
	20-24	83,000 & +	23,000-83,500	under 23,000
	25+	506,500 & +	134,500-506,500	under 134,500
Ontario	All	802,000 & +	210,500-802,000	under 210,500
	15-19	150,000 & +	43,000-150,000	under 43,000
	20-24	170,500 & +	50,500-170,500	under 50,500
	25+	897,000 & +	241,000-897,000	under 241,000
Manitoba	All	60,000 & +	15,500-60,000	under 15,500
	15-19	13,500 & +	4,000-13,500	under 4,000
	20-24	23,000 & +	7,500-23,000	under 7,500
	25+	66,500 & +	17,500-66,500	under 17,500
Saskatchewan	All	51,500 & +	13,500-51,500	under 13,500
	15-19	13,000 & +	3,500-13,000	under 3,500
	20-24	18,000 & +	5,000-18,000	under 5,550
	25+	59,000 & +	15,500-59,000	under 15,500
Alberta	All	150,500 & +	39,000-150,500	under 39,000

	15-19	38,000 & +	10,500-38,000	under 10,500
	20-24	58,500 & +	18,000-58,500	under 18,000
	25+	163,500 & +	43,000-163,500	under 43,000
British Columbia	All	218,500 & +	56,500-218,500	under 56,500
	15-19	65,500 & +	19,500-65,500	under 19,500
	20-24	50,000 & +	14,500-50,000	under 14,500
	25+	240,000 & +	63,000-240,000	under 63,000
Canada	All	521,560 & +	103,500-415,500	under 130,000
	15-19	106,000 & +	23,500-92,000	under 27,000
	20-24	121,500 & +	23,500-93,000	under 31,000
	25+	594,500 & +	122,500-489,500	under 149,000

10.7 C.V. Tables - Household file

Refer to CTUMS_00_C2_HH_CV_ENG.PDF for the c.v. tables for the "Household" file for Cycle 2 in 2000.

10.8 C.V. Tables - Person File

Refer to CTUMS_00_C2_PR_CV_ENG.PDF for the c.v. tables for the "Person" file for Cycle 2 in 2000.

11.0 Weighting

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

The weighting for the first cycle of the Canadian Tobacco Use Monitoring Survey 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 Group-Sex totals from the Census projected population counts for persons 15 years and over.

11.1 Weighting Procedures for Both the Household and Person File

1. Calculate telephone weight

Each telephone number in the sample was assigned a basic weight, W1, equal to the inverse of its probability of selection.

$$W1 = \frac{\left(\begin{array}{l}\textit{total number of possible sampled telephone numbers}\\ \textit{in Province - Stratum}\end{array}\right)}{\left(\begin{array}{l}\textit{number of sampled telephone numbers}\\ \textit{in Province - Stratum}\end{array}\right)}$$

There were 58,383 phone numbers in the sample with assigned weights.

2. Adjust for non-resolved telephone numbers

There were 731 telephone numbers that were not resolved, leaving 57,652 resolved phoned numbers. The unresolved phone numbers were not determined to belong to a household, business or out-of scope. Each phone number had a flag indicating whether it is expected to be a residential, business, or unknown type of phone number. The adjustment for the unresolved phone numbers was done within Province-Stratum and this expected line type.

For each province-stratum-expected line type,

$$W2 = W1 * \left(\frac{\sum W1 \text{ for resolved phone numbers} + \sum W1 \text{ for unresolved phone numbers}}{\sum W1 \text{ for resolved phone numbers}} \right)$$

3. Remove out-of-scope telephone numbers

Phone numbers corresponding to businesses, out-of-service numbers, or were out-of scope, such as cottage phone numbers, were dropped after the non-response adjustment for telephone non-response has been applied. Note that if **household or person data** existed then the phone number was assumed to be a household. There were 32,716 out of scope phone numbers and 24,936 phone numbers belonging to a household.

4. Adjust for non-response of number of telephone lines

The number of phone lines in the household was calculated. If the number of different phone 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 2,042 telephone numbers that were still missing the number of lines. Thus, there were 22,894 households with the number of lines calculated or imputed. The adjustment was done within Province-Stratum.

$$W3 = W2 * \left(\frac{\sum W2 \text{ for households with number of lines} + \sum W2 \text{ for households missing number of lines}}{\sum W2 \text{ for households with number of lines}} \right)$$

5. Calculate household weight with multiple telephone adjustment

Weights for households with more than one telephone number 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 numbers that serviced the household.

$$W4 = \frac{W3}{\text{number of in-scope phone numbers in the household}}$$

11.2 Weighting Procedures for the Household File

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 901 non-respondents. Thus, the 21,993 in-scope household weights, were used and adjusted within Province-Stratum.

$$W5 = W4 * \left(\frac{\sum W4 \text{ for household respondents} + \sum W4 \text{ for household non-respondents}}{\sum W4 \text{ for household respondents}} \right)$$

7. Adjust to external known stratum totals

An adjustment was made to the household weights on records within each Province, Stratum and Month. In order to make household estimates consistent with external household counts. The adjustment factor for Province-Stratum-Month (P-S-M) was defined

$$W6 = W5 * \frac{\text{External household count in P-S-M}}{\text{sum of the weights W5 for responding households in the sample in P-S-M}}$$

The household weights, W6, obtained after this step, were considered final and appear on the household microdata file.

11.3 Weighting Procedures for the Person File

6. Adjust for non-responding households

On the person file, household respondents completed the roster with no age refusals. There were 780 non-respondents. Thus, the 22,114 in-scope household weights, were used and adjusted within Province-Stratum.

$$W5 = W4 * \left(\frac{\sum W4 \text{ for household respondents} + \sum W4 \text{ for household non-respondents}}{\sum W4 \text{ for household respondents}} \right)$$

7. 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: 15-19, 20-24, and 25+. 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 one). 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 two age groups were represented then W6 = W5.
If all 3 age groups were represented then W6 = W5 * 3/2.

8. Remove households with no selected persons

There were 12,477 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 9,956 households with selected persons. There were 8,484 households with one person selected and 1,472 with two people selected.

9. Assign household weights to selected persons

These 8,173 + 2(1,464) = 11,101 selected persons are associated with in-scope “responding households” and keep the corresponding weight, W6.

10. 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.

$$W7 = W6 * \left(\frac{Number\ of\ Individuals\ in\ selected\ age\ group}{sub\text{-}sampling\ factor} \right)$$

Where the sub-sampling factor was 1 for age groups 15-19 and 20-24. The sub-sampling factor was pre-assigned for the 25 plus and varied from 39.2% to 48.4% (depending on province).

11. Adjust for non-responding individuals

On the person file, individual respondents completed the questions on smoking habits on the tobacco use survey and gave a date of birth corresponding to the age given in the roster. There were 1,179 non-respondents. Thus, the 9,922 in-scope individual weights were used and

adjusted within province, age groups derived from the roster (15-19,20-24,25-44,45-64,65+) and sex.

$$W8 = W7 * \left(\frac{\sum W7 \text{ for person respondents} + \sum W7 \text{ for person non-respondents}}{\sum W7 \text{ for person respondents}} \right)$$

12. 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-stratum-month
2) population totals by province, sex and the following age groups: 15-19, 20-24, 25-34, 35-44, 45-54, 55-64, and 65+. 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 microdata file.

12.0 Questionnaire

Refer to CTUMS_00_C2_QUES_E.pdf for the English questionnaire for the second cycle collected in 2000.

13.0 Record Layout and Univariates

13.1 Record Layout and Univariates - Household

Refer to CTUMS_00_C2_HH_CdBk.PDF for the record layout and univariate counts for the “Household” file for Cycle 2 in 2000.

13.2 Record Layout and Univariates - Person

Refer to CTUMS_00_C2_PR_CdBk.PDF for the record layout and univariate counts for the “Person” file for Cycle 2 in 2000.