Microdata User Guide

YOUTH SMOKING SURVEY

2002

Statistics Canada Statistique Canada

Canada

Table of Contents

1.0 Introduction

The 2002 Youth Smoking Survey (YSS) was conducted by Statistics Canada from October to December 2002 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 call toll-free 1 800 461-9050
Fax: (613) 951-4527
E-mail: ssd@statcan.ca

Elizabeth Majewski
Special Surveys Division
2nd floor, Main Building, Tunney's Pasture
Ottawa, Ontario K1A 0T6
Telephone: (613) 951-4584
Fax: (613) 951-0562
E-mail: elizabeth.majewski@statcan.ca

Health Canada

Alan Diener
Office of Research, Surveillance and Evaluation
Tobacco Control Programme
123 Slater Street,
Ottawa, Ontario K1A 0K9
Telephone: (613) 957-7852
Fax: (613) 954-2292
E-mail: Alan_Diener@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 2002 YOUTH SMOKING SURVEY.

2.0 Background

Statistics Canada conducted the first national Youth Smoking Survey (YSS) in the fall of 1994. The survey had two components: children aged 10 to 14 who were surveyed at school, while youth aged 15 to 19 were interviewed at home, by telephone. The survey findings were published in 1996 by Health Canada as a research report and as a set of fact sheets.

The smoking behaviour of the 15 to 19 year olds has been monitored by the Canadian Tobacco Use Monitoring Survey, conducted for Health Canada by Statistics Canada, since 1999. Health Canada asked Statistics Canada to repeat the school portion of the 1994 Youth Smoking Survey in the fall of 2002. Besides smoking as the core content, the 2002 YSS includes questions referring to experiences with alcohol and drugs for students in grades 7 to 9 (in Quebec secondary school grades 1 to 3). Both the 1994 and the 2002 survey provided national (excluding the Yukon, Northwest Territories, and Nunavut) and provincial estimates. Health Canada has plans to repeat the survey in 2004 and is working with the territories to support them in carrying out surveys comparable to the Youth Smoking Survey.

Information on smoking, as well as the use of alcohol and drugs by children and youth, is also available from the National Longitudinal Survey of Children and Youth (NLSCY), a Statistics Canada survey that started in 1994. However, given the nature of the NLSCY, the coverage of smoking behaviour is not extensive and the cross-sectional samples are of modest size. Additionally, the Canadian Community Health Survey (CCHS) conducted in 2001, collected data on smoking for children aged 12 and over.

3.0 Objectives

The main objective of the 2002 Youth Smoking Survey (YSS) is to provide current information on the smoking behaviour of students in grades 5 to 9 (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3), and to measure changes that occurred since 1994. Additionally, the 2002 survey collected basic data on alcohol and drug use by students in grades 7 to 9 (in Quebec secondary 1 to 3). Results of the Youth Smoking Survey will help with the evaluation of anti-smoking and anti-drug use programs, as well as with the development of new programs.

The YSS collected information on the following topics:

- the prevalence of smoking among students in grades 5 to 9 (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3);
- the types of smoking behaviour among children (e.g. experimental smoking, occasional smoking, daily smoking);
- the social and demographic factors associated with smoking behaviour (e.g. what motivates children to smoke, the influence of family and friends);
- where and how children obtain cigarettes;
- attitudes and beliefs about smoking, including awareness of health risks;
- recollection and opinions on health warning messages on cigarette packages;
- experience with alcohol, drugs and medications used for non-medical purposes.

4.0 Concepts and Definitions

Definitions of smoking categories for the 1994 and for the 2002 Youth Smoking Survey (YSS) have been based on categories proposed in the paper *Summary Report of the Workshop on Data for Monitoring Tobacco Use* by Mills, Stephens, and Wilkins, which appeared in Chronic Diseases in Canada in the summer of 1994. Some minor modifications were made to the categories proposed in the paper to adapt them to the population composed mainly of 10 to 14 year olds. See Section 7.4 (Creation of Derived Variables) for more detailed information about the derived variables reflecting these definitions.

Currently smokes

Has smoked at least 100 cigarettes in his/her lifetime, and has smoked in the 30 days preceding the survey. This category includes the following sub-categories:

Currently smokes daily

Has smoked at least 100 cigarettes in his/her lifetime, and has smoked at least one cigarette per day for each of the 30 days preceding the survey.

Currently smokes occasionally

Has smoked at least 100 cigarettes in his/her lifetime, and has smoked at least one cigarette during the 30 days preceding the survey, but has not smoked every day.

Formerly smoked

Has smoked 100 or more cigarettes in his/her lifetime but has not smoked at all during the 30 days preceding the survey. This category includes the following sub-categories:

Formerly smoked daily

Has smoked 100 or more cigarettes in his/her lifetime but has not smoked at all during the 30 days preceding the survey, and has at some time smoked every day for seven days in a row.

Formerly smoked occasionally

Has smoked 100 or more cigarettes in his/her lifetime but has not smoked at all during the 30 days preceding the survey, and has never smoked every day for seven days in a row.

Never smoked

Has smoked fewer than 100 cigarettes in his/her lifetime. This category includes the following sub-categories:

Experimental smoker (beginner)

Has smoked between 1 and 99 cigarettes in his/her lifetime, and has smoked in the 30 days preceding the survey.

Past experimenter

Has smoked between 1 and 99 cigarettes in his/her lifetime, but has not smoked in the 30 days preceding the survey.

Lifetime abstainer

Has smoked less than one whole cigarette in his/her lifetime.

5.0 Survey Methodology

The 2002 Youth Smoking Survey (YSS) was administered to a sample of children in grades 5 to 9 (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3) by sampling classes from a frame of all public and private schools in Canada. The sample design consists of a two-stage stratified clustered design with schools as primary sampling units and with classes as secondary sampling units. All of the students in the selected classes were surveyed.

5.1 Population Coverage

The target population consists of all young Canadian residents aged 10 to 14 attending private or public schools in grades 5 to 9 inclusively (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3). Specifically excluded from the survey's coverage are residents of the Yukon, Northwest Territories and Nunavut, persons living on Indian Reserves and inmates of institutions. Young persons who are attending special schools (schools for the blind or for deaf-mutes) or who are attending schools located on military bases are also excluded from the target population. Furthermore, the population actually surveyed differs somewhat from the target population. The differences may be categorized as:

1) Young people enrolled in small classes (less than 10 students) and;

2) Young people living in remote areas i.e.:

   Newfoundland & Labrador above latitude of 55 degrees,
   Quebec above latitude of 51 degrees, as well as Îles de la Madeleine,
   Ontario above latitude of 51 degrees,
   Manitoba and Saskatchewan above latitude of 55 degrees,
   Alberta and British Columbia above latitude of 57 degrees and the Queen Charlotte Islands.

Both categories were not eligible to be surveyed but were still part of the target population. It is estimated that these exclusions represent approximately 2.3% of the target population.

5.2 Sample Design

5.2.1 Stratification

The sample design features three levels of stratification. First, each province constitutes a stratum. An implicit stratification by grade level (5 to 9 - in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3) is used and finally, the schools are stratified by census metropolitan area (CMA) versus non-CMA, with additional strata in Quebec (Montreal) and Ontario (Toronto). (In Quebec the English and bilingual schools were grouped together in a stratum called English regardless of their CMA status. The French schools were stratified by CMA, non-CMA and Montreal). The sample was then selected in each strata independently, meaning that some schools may be selected more than once, for different grades.

5.2.2 Sample Size and Allocation

The requirements relating to the accuracy of the results consisted of a minimum estimable proportion (0.10) combined with a maximum coefficient of variation (16.5%) with respect to province and thus for the entire grade 5 to 9 (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3) target population. It was determined that an overall sample of 20,000 respondents would be needed to attain these goals.

To determine the sample size needed in each province, it was necessary to make certain assumptions regarding the overall response rates (75%) and average class sizes (25). The overall sample size was allocated to the provinces using the power allocation method. The provincial sample was then allocated proportionally to each of the strata grades based on the enrolment figures.

5.2.3 Sample Selection

The sample of schools was selected systematically with probability proportional to school size, i.e., the total number of students for each stratum. In order to ensure better representation by school board size and school size, the school file was sorted, first by school board size and then by school size within each school board. The selection of the secondary sampling units (classes) was accomplished in the field by the interviewer who randomly selected one class in the desired grade. This translated into a final sample of 1,070 classes in 982 schools situated in 327 school boards.

5.3 Overlap with the 2002 Quebec Survey of Tobacco Use by High School Students

The Institut de la statistique du Québec (ISQ) has been conducting a survey on youth smoking every two years since 1998. This survey targets students in grades 7 to 11 (secondary school grades 1 to 5) and results in a sample of approximately 160 schools. In 2002, there was a strong possibility that the same schools could be selected to participate in the two surveys (for the same grade). This was of concern because of the possible perception that the two surveys create an unnecessary burden on the schools. To alleviate this problem, a method for controlling the overlap by slightly adjusting the YSS probabilities of selection was implemented for 2002.

5.4 Replacement for Non-consenting School Boards and Schools

In order to achieve the desired final sample size, schools from non-consenting boards were replaced with schools from consenting boards that had a similar profile in terms of enrolment size and grades taught. For some very large boards, replacement was not possible since their size made them unique and replacing them with smaller boards would impact on the integrity and representativity of the sample. For schools refusals, which usually occured during field collection, a similar approach was used. Replacements schools that refused to participate to the survey were not replaced.

5.5 Sample Size by Province and Grade

The following table shows the number of classes sampled for each province and grade.

Province	School Grade					
	5	6	7	8	9	Total
Newfoundland and Labrador	15	15	15	16	17	78
Prince Edward Island	11	10	11	11	11	54
Nova Scotia	17	17	18	18	19	89
New Brunswick	16	16	16	16	19	83
Quebec	32	30	33	31	29	155
Ontario	34	33	33	34	35	169
Manitoba	19	18	19	20	20	96
Saskatchewan	18	18	18	18	20	92
Alberta	25	25	24	25	25	124
British Columbia	25	26	25	26	28	130
Canada	**212**	**208**	**212**	**215**	**223**	**1,070**

6.0 Data Collection

Data collection for the Youth Smoking Survey (YSS) took place at schools and was supplemented by telephone interviews with parents. Interviews were conducted under the voluntary provisions of the Statistics Act.

6.1 Questionnaire Design

As comparability with the 1994 survey results was an important objective of the 2002 YSS, only minimal modifications were made to the wording of the questions that were asked of children in 1994. The layout and the style of the questionnaire were also very similar to the original version. So as not to affect responses to the questions related to smoking, the alcohol and drug use questions were added at the end of the questionnaire for the older students. The new questions about children's activities, as well as those measuring self-esteem, came from the National Longitudinal Survey of Children and Youth.

The draft questionnaire was tested in the spring of 2002 with children from various grades, with or without experience with cigarettes, with good or low marks, boys and girls, English and French speaking. The respondents completed the questionnaire and later, in one-on-one interviews, provided comments and clarified their answers.

The basic function of the parent's questionnaire was still collection of socio-demographic information about the child's family. However, the design of the questionnaire was significantly modified compared with the 1994 version. Of the 15 questions addressed to parents, 13 are standard questions used in other surveys. The parent's questionnaire was tested informally.

6.2 Data Collection

Survey collection activities in schools were conducted from October to December 2002. They included mailing an introductory letter to the selected schools, selecting the classes to participate in the survey and conducting classroom sessions during which students completed paper questionnaires. These collection activities were preceded by a lengthy school board approval process which began in June, 2002.

Only experienced Statistics Canada interviewers worked on this survey. An average assignment size was three classrooms per interviewer. If more than one classroom, up to a maximum of four, was selected in one school, all of the classrooms were assigned to the same interviewer. This procedure ensured that each school was approached by only one interviewer and a minimal number of visits were made to the school. Allocation of assignments was based on the geographic distribution of the schools relative to the interviewers' residences.

Interviewers were allowed up to four hours for training. This included reading the Interviewer's Manual, completing the review exercises, answering test questions posed by senior interviewers over the phone and discussing any data collection issues. The YSS data collection managers in the Regional Offices participated in a two day classroom training session in Ottawa and later trained the senior interviewers.

Following is a summary of the data collection process.

First Contact with School

Soon after the regional offices mailed the introductory letters to the selected schools, interviewers telephoned each school to:

- introduce the YSS to the school principal;
- obtain collaboration from the principal to participate in the survey;
- schedule an appointment for a first visit to the school; and
- verify the school address and obtain directions, if necessary.

As the list of schools from which the sample was drawn was somewhat outdated by the time the survey was conducted, procedures were in place to deal with school changes. If the school principal refused to participate, the school was replaced according to a replacement strategy (see Section 5.4).

First Visit to School

Upon arrival at the school, the interviewer introduced himself/herself to the principal and briefly outlined the collection activities. A labeled Classroom Selection Form was used to control the collection activities. The form identified the grade that was selected for the survey. If the school had more than one class for the grade selected, the interviewer used the selection grid on the label to randomly select one of the classrooms.

The interviewer listed the name, telephone number(s) and preferred language of each student in the selected class. For each student the interviewer prepared a package to take home containing an introductory letter and Parental Consent Form. The principal or class teacher was asked to distribute and control the receipt of the completed Parental Consent Forms. The interviewer explained that he/she would return to the school in one week to pick-up the completed forms. In several instances, the principal would not provide the telephone number of the students despite the interviewers best efforts to explain and justify this request. Although there were procedures in place to work around this situation, if consent forms were not returned the parents could not be contacted and students could not take part in the survey. When the telephone number was missing, even if the consent form was signed, the telephone interview with a parent was not possible.

Second Visit to School

During the second visit, the interviewer:

- picked up the completed Parental Consent Forms; and
- scheduled an appointment to return to the school in one week to conduct the classroom session.

Next, the interviewer determined which children had parental consent to participate in the survey. If a Parental Consent Form was not returned, the interviewer asked for consent over the telephone. If consent was granted a brief interview was conducted by telephone with the parent or guardian of each child. Most of these interviews were completed before the classroom session.

Classroom Session (Third Visit to School)

In preparation for the classroom session, the interviewer prepared a questionnaire for each eligible child according to the preferred language of interview noted on the Parental Consent Form and transcribed on to the Classroom Selection Form. The total number of students in the classroom and the number of students with written or verbal parental consent, as well as those students without consent had to be entered on the Classroom Selection Form for the calculation of response rates at the conclusion of the survey.

The student's name was not written on the questionnaire to maintain anonymity. It was only written on an envelope for the purpose of ensuring that the correct questionnaire was given to the student.

Once in the classroom, the interviewer:

- Introduced himself/herself to the students.
- Explained the purpose of the survey.
- Asked the teacher to distribute the envelopes to the students.
- Read aloud the introduction on the questionnaire.
- Completed the first nine questions with the students to show them how to make different types of entries.
- Explained how to complete the wheel in Question 21.
- Told students to feel free to raise their hand to ask questions.
- Told the students not to put the completed questionnaires into the envelope but to leave them face down and separate from the envelope on their desk.
- First gathered the envelopes and later collected the completed questionnaires placing them in the versapak.
- Thanked the students and the teacher for their co-operation and support.

The classroom sessions, on average, lasted 30 to 40 minutes. Teachers were asked to remain in the classroom, but were asked not to circulate among the students to protect confidentiality.

7.0 Data Processing

The main output of the Youth Smoking Survey (YSS) is a "clean" microdata file. This chapter presents a brief summary of the processing steps involved in producing this file.

7.1 Data Capture

The questionnaires were data captured at head office. All the questionnaire identification numbers were keyed in twice to avoid any errors. The quality of the captured data was checked by a random verification of 20% of the records. The error rate was below 2%.

7.2 Editing and Imputation

As in 1994, the youth questionnaire was designed with very few skip patterns. It was felt that skip patterns might not be correctly followed by the young respondents aged 10 to 14 and might result in identifying smokers as non-smokers during the classroom session, as non-smokers would require much less time to get through the questionnaire.

The questionnaire was edited using the “top-down“ logic. To accomplish this task, flows had to be determined before the edit program could be written. The critical element was establishing the smoking status of respondents, as many survey questions applied to smokers only. Similar to the 1994 survey, answers to several questions were examined to resolve inconsistencies between key questions used as indicators of smoking status.

The first stage of survey processing undertaken at head office was the replacement of any ”out-of-range” values on the data file 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 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 variable.

Some item non-response was also subject to imputation. Imputation is the process used to resolve problems of missing, invalid or inconsistent responses identified during editing. This is done by changing some of the responses or item non-responses on the record being edited to ensure that a plausible, internally coherent record is created. Further information on the imputation process is given in Chapter 8.0.

The following standard codes are used on the microdata file:

Valid skip - 6, 96 and 996

Don’t know - 7, 97 and 997

Refused - 8, 98 and 998

Not stated - 9, 99 and 999

7.3 Coding of Open-ended Questions

A few data items on the questionnaire were recorded in an open-ended format. These were items relating to health problems caused by smoking, health warnings on cigarette packages and Other *(specify)* answers.

Coding of Question 48 and Question 50b

In preparation for data capture, head office developed a coding manual of possible answers for Question 48 about health problems caused by smoking and Question 50b about health warnings on cigarette packages.

The coding manual consisting of the coding procedures, sample responses, review exercises and the code list, was developed, and formed the basis for training. The training of coders was conducted in a classroom setting.

The coding supervisor did a sample verification of the coded items to ensure that codes were assigned correctly.

Question 48 - Health Problems

Youth were asked to describe what health problems people can get if they smoke for many years. The students wrote down their responses in the space provided on the questionnaire. A maximum of eight responses were manually coded.

The code list for health problems consisted of 69, two-digit codes that were grouped into major categories. Almost all the codes were the same as in 1994. Spelling errors were to be ignored; however, some responses from the school questionnaires were difficult to code due to poor spelling and penmanship.

Question 50b - Health Warning Messages

Since the summer of 2000 there have been 16 graphic health warnings about hazards associated with tobacco use displayed on cigarette packages sold in Canada. Youth who indicated that they had seen health warning messages on cigarette packages, were asked to recall as many messages as they could. The students were asked to write, in their own words, any health warning messages that they remembered seeing on cigarette packages. These responses were each then manually coded, to a maximum of eight responses.

The codes were divided into ten thematic categories referring to the content of the messages. Once a response was attributed to a specific theme, it could be coded as an exact quotation of a warning message, as an explicit reference to the picture(s) illustrating that theme, a specific reference or a general reference to the theme. Given that a picture of a full ashtray accompanies two different warnings (addiction and second-hand smoke), references to this picture without any additional clarification were coded to a separate code. The mention of health issues not covered by the warnings were coded as “Other”. Spelling errors were to be ignored, and as with the health problems, some responses were difficult to code due to poor spelling and penmanship.

Coding of Other *(specify)* Answers

There were eight partially open-ended questions in the YSS questionnaire that contained a list of answer categories that had “Other *(specify)*” as the final category. These write-in answers were examined and either recoded or remained as “Other”. The recoding was done into existing or specially created answer categories.

7.4 Creation of Derived Variables

In order to facilitate data analysis, a number of data items (variables) on the microdata file have been derived by combining items on the questionnaire. For each derived variable, there is a note in the codebook stating which survey questions were used to create the variable. The derived variables are found on the record layout following the YSS questions that were part of the derivation. The derived variables are identified by variable names beginning with "DV". Some examples of derived variables are presented below.

DVTY1ST - Smoking status using the definition from the 1994 Youth Smoking Survey

Combines the answers to question Y_Q16 (smoked 100 or more cigarettes) and question Y_Q19 (smoked in the past 30 days) to derive the smoking status defined according to one of the typologies used by researchers in the field.

Current smoker - a person who smoked 100 or more cigarettes and smoked in the past 30 days (Y_Q16 = 1 and Y_Q19 = 02 to 06);

Former smoker - a person who smoked 100 or more cigarettes but did not smoke in the past 30 days (Y_Q16 = 1 and Y_Q19 = 01);

Never smoked - a person who did not smoke 100 or more cigarettes (Y_Q14 = 2 or Y_Q16 = 2).

DVTY2ST – Detailed smoking status using the definition from the 1994 Youth Smoking Survey

This is a more detailed typology of smokers with categories defined as follows.

Current daily smoker - smoked 100 or more cigarettes and smoked every day in the past 30 days (Y_Q16 = 1 and Y_Q19 = 06);

Current occasional smoker - smoked 100 or more cigarettes and smoked on 1 to 29 days in the past 30 days (Y_Q16 = 1 and Y_Q19 = 02 to 05);

Former daily smoker - smoked 100 or more cigarettes, did not smoke in the past 30 days, but smoked every day for at least seven days in a row in the past (Y_Q16 = 1 and Y_Q17 = 1 and Y_Q19 = 01);

Former occasional smoker - smoked 100 or more cigarettes, did not smoke in the past 30 days, and never smoked every day for at least seven days in a row (Y_Q16 = 1 and Y_Q17 = 2 and Y_Q19 = 01);

Experimental smoker (beginner) - smoked between 1 and 99 cigarettes and has smoked in the past 30 days (Y_Q16 = 2 and Y_Q19 = 02 to 06);

Past experimenter - smoked between 1 and 99 cigarettes, but has not smoked in the past 30 days (Y_Q16 = 2 and Y_Q19 = 01);

Lifetime abstainer - has never smoked a whole cigarette, but has possibly taken a few puffs (Y_Q14 = 2).

DV48_20 - Reported health problems that people can get if they smoke for many years: Cancer

Identifies all the respondents who answered cancer (unspecified type) in the open-ended Question 48. A maximum of eight responses were manually coded. Derived variable DV48_20 represents only one of the responses to this question. There are 42 manually coded responses for these derived variables (DV48_20 to DV48_88 not inclusive).

DVSELF - Score on the overall self-esteem scale

Based on the variables Y_9A to Y_9D inclusive, the objective of the General Self Score is to measure the child's self-esteem. The self-esteem scale was used in the National Longitudinal Survey of Children and Youth (NLSCY) and was based on the General Self Scale of the Marsh Self Description Questionnaire developed by H.W. Marsh. See the *Microdata User Guide, National Longitudinal Survey of Children and Youth, Cycle 4, September 2000 to May 2001, Section 9.6.1, General Self Score*. The NLSCY questions used the following five point answer categories:

1) False
2) Mostly false
3) Sometimes false / sometimes true
4) Mostly true
5) True

To produce the final scores, one was subtracted from each item so that the lowest score would be zero (0). The final score was derived by totalling the values of all items with non-missing values ranging from 0 to 16. No imputation was done for not stated values. If any values were not stated, the final score was set to "Not stated".

The 2002 YSS used the following six point answer categories:

1) False
2) Mostly false
3) Sometimes false
4) Sometimes true
5) Mostly true
6) True

In order to make the scale compatible with the NLSCY, the YSS collapsed the two middle categories, three and four, to be "Sometimes false / sometimes true". The resulting scores were not revalidated thus the full implications of collecting the questions using a six-point scale are not known.

7.5 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 children in grade 9 (in Quebec secondary 3) who smoked 100 cigarettes in their life is to be estimated, it is done by selecting the records referring to those individuals in the sample with that characteristic and summing the weights entered on those records.

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

7.6 Suppression of Confidential Information

It should be noted that the "Public Use" microdata file described above differs in a number of important respects from the survey "master" file 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 Chapter 9.0 of this document.

Since the YSS file amalgamates data from both the child and the parent/guardian interview into a single record, there is a possibility that a parent/guardian may identify his/her child's answers and consequently obtain access to the child's responses. This assumption led to the suppression of all but one of the parental variables (GP2_07 – Child's family situation).

To prevent disclosure on the child's section of the record, age and aboriginal status have been suppressed, while the weekly amount of spending money has been capped at $75. Additionally, the responses to Questions 37a (father) and 39a (mother) "I don't live with a father/mother or anyone who is like a father/mother" have been coded to "Not stated". There were also a total of 44 local suppressions on 30 records to minimize the risk of disclosure in case of unique combinations of variables.

To avoid disclosure of product brands, Question 22b (What brand do you usually smoke?) has been replaced by the derived variables DVSMOKE that describes the strength of the brand and DVLOWTAR which indicates its lower tar value. Similarly, the questions referring to the use of Ritalin (75a and 75b) and Gravol (78a and 78b) have been replaced by derived variables grouping the drugs in question with other prescription (DVPDG and DVPDGAG) and non-prescription drugs (DVNDG and DVNDGAG).

8.0 Data Quality

8.1 Response Rates

There were various levels of non-response throughout the Youth Smoking Survey (YSS). A description of these levels is presented below along with the appropriate tables.

First, some degree of non-response was noted among school boards and schools. Using the replacement strategy outlined in Section 5.4, replacements were found for the majority of school boards and schools who refused to participate in the survey. The final response rate at the school board and school level, expressed as number of classes, is presented in the table below.

Number of Classes by Province

Province	School Board Level			School Level		
	Total	Consent Given	Response Rate (%)	Total	Consent Given	Response Rate (%)
Newfoundland and Labrador	78	78	100	78	77	99
Prince Edward Island	54	54	100	54	54	100
Nova Scotia	89	89	100	89	85	96
New Brunswick	83	83	100	83	79	95
Quebec	155	150	97	150	148	99
Ontario	169	148	88	148	134	91
Manitoba	96	96	100	96	91	95
Saskatchewan	92	92	100	92	92	100
Alberta	124	91	73	91	79	87
British Columbia	130	120	92	120	116	97
Canada	**1,070**	**1,001**	**94**	**1,001**	**955**	**95**

The second component of non-response relates to the students. The response rate at the student level is derived based on the number of eligible students recorded on the Classroom Selection Form in each of the participating classes (955). Non-response at the student level can be attributed to several factors. Some parents/guardians refused to allow their child to take part in the survey, and even with parental consent some students refused to participate or the student was absent from class on the day of collection. Finally, some records were regarded as non-response because they did not contain sufficient information and could not be considered as usable. The final response rates at the student level are summarized in the table below.

Student Response Rates by Province – Master File and Public Use Microdata File

Province	Eligible Students	Usable Questionnaires	Response Rate (%)
Newfoundland and Labrador	1,862	1,574	85
Prince Edward Island	1,305	1,091	84
Nova Scotia	2,108	1,784	85
New Brunswick	2,020	1,656	82
Quebec	3,869	3,229	83
Ontario	3,343	2,583	77
Manitoba	2,000	1,534	77
Saskatchewan	2,024	1,707	84
Alberta	1,803	1,442	80
British Columbia	2,883	2,418	84
Canada	**23,217**	**19,018**	**82**

Student Response Rates by Province – Share File

Province	Usable Questionnaires	Students With Sharing Agreement	Response Rate (%)
Newfoundland and Labrador	1,574	1,501	95
Prince Edward Island	1,091	1,039	95
Nova Scotia	1,784	1,724	97
New Brunswick	1,656	1,569	95
Quebec	3,229	2,936	91
Ontario	2,583	2,265	88
Manitoba	1,534	1,475	96
Saskatchewan	1,707	1,600	94
Alberta	1,442	1,348	93
British Columbia	2,418	2,252	93
Canada	**19,018**	**17,709**	**93**

8.2 Survey Errors

The estimates derived from this survey are based on a sample of students. 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 were taken 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 the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions, procedures to ensure that data capture errors were minimized, and coding and edit quality checks to verify the processing logic.

8.2.1 The Frame

A listing of all public and private schools that provided enrolment by grade in Canada for the 1999-2000 school year was used as the sampling frame. Information on the school boards was obtained from Statistics Canada's data on educational institutions for the 2001-2002 school year. In order to preserve the integrity of the sample design, procedures were developed to handle cases where sampled schools had closed, moved or no longer taught the grade for which they were originally selected. These actions were guided by the nature of the changes which was documented by the field interviewer and communicated to head office.

When the interviewer encountered cases of schools that had closed, moved or no longer taught the grade for which they were originally selected, he/she was asked to provide the following information to head office.

1) To what school(s) the students in the selected grade have been relocated?

2) Were there other students already in that grade at that school?

3) Were students from other schools also relocated to that grade in that school *(names of schools)*?

Based on the observations obtained from the field interviewers, a decision was made as to whether to replace the school. For the 2002 YSS, 29 such instances (i.e. 3% of all surveyed classes) were encountered during collection.

8.2.2 Non-response

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Total non-response occurred because school boards or schools refused to participate in the survey, parental consent was not obtained, students refused to participate in the survey or were absent on the day of the interview. Total non-response was handled by adjusting the weight of the students who responded to the survey to compensate for those who did not respond. Non-response has two effects on data: one contributing to an increase in the sampling variance (coefficients of variation) as the effective sample size is reduced, the second being bias, since the non-respondents may differ from respondents with regard to the characteristics measured.

In most cases, partial or item 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.

The YSS data file includes information obtained from children and their parents/guardians. There are 1,055 records with parental information missing due to non-response. A substantial part of item non-response was caused by non-response to the parental component of the survey. The following questions had high non-response rates (i.e. "Don't know", "Refused" or "Not stated").

Y_Q08 - How much would you like to weigh right now? 15%

Y_Q46A to F, I, and J - What do you think about some of the things that have been said about cigarette smoking? 15% to 32%

Y_Q55 - In your school, what are the rules about smoking? 16%

Y_Q56 - Do most students who smoke obey that rule? 42%

Y_Q59 - About how much money do you usually get each week to spend on yourself or to save? 23%

Y_Q80A to F - Number of deaths due to smoking versus other causes in Canada. 32% to 46%

P_16B - What is the occupation of the other parent/guardian in the household? 22%

P_17 - What is your best estimate of the total household income for the last 12 months before taxes and deductions? 16%

Even though for the majority of the variables item non-response was treated by assigning a "Not stated" code, imputation for question Y_Q16 (Have you smoked 100 or more cigarettes in your life?) was necessary since this variable was critical for deriving smoking status. Records with "Not stated" and "Don't know" answers were imputed using a donor imputation approach. The method used to perform imputation is a hot-deck procedure called nearest-neighbour or donor imputation. For item j, a missing value Y_{jk} is replaced by the value of a respondent classified to be the "nearest" as measured by a distance function defined in terms of known auxiliary variable values such as Y_Q17 (Have you ever smoked every day for at least 7 days in a row?) and Y_Q19 (On how many of the last 30 days did you smoke one or more cigarettes?). Additional matching fields were province, grade and gender. Imputation for Y_Q16 was necessary for 362 records and donors were found for every case. A total of 157 records were imputed to "Yes" and 205 to "No". Please note that the 1994 YSS had a similar incidence of "Not stated" and "Don't know" answers, but did not use imputation.

Questions Y_Q11A (Have you ever tried cigarette smoking, even just a few puffs?) and Y_Q14 (Have you ever smoked a whole cigarette?) were also imputed for missing entries since they were pivotal in determining the valid skips. In order to reconcile, and deterministically impute these two variables, several variables were considered. They were Q11A, Q14 and Q15 as primary fields and Q16, Q17 Q18, Q19 and Q20 as secondary fields. The secondary fields were used to derive a value aimed at providing a "smoking signal" that was used to resolve inconsistencies between the primary fields. Basically, the strategy was similar to the one used for the 1994 survey, that is, looking jointly at Q11A, Q14 and Q15 which were deemed of comparable importance and reliability. Whenever Q11A and Q14 did not agree (or had missing values), Q15 was used as the tie-breaker. The secondary fields were used to break the tie for more complex situations.

8.2.3 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 (CV) 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 2.7% of children in grades 5 to 9 (in Quebec primary school grades 5 and 6 and secondary school grades 1 to 3) are currently cigarette smokers and this estimate is found to have a standard error of 0.00216. Then the coefficient of variation of the estimate is calculated as:

$$\left(\frac{0.00216}{0.027}\right) X\, 100\% = 8.0\%$$

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 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 Youth Smoking Survey (YSS) 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 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 Youth Smoking Survey 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 YSS.

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 students who ever smoked a whole cigarette or the proportion of smokers who usually buy single cigarettes from a friend or someone else are examples of such estimates. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Examples of Categorical Questions:

Q: Have you ever smoked a whole cigarette?
R: Yes / No

Q: Where do you buy them (single cigarettes)?
R: At a small grocery/corner store / In another kind of store / I buy them from a friend or someone else

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 dollars children have to spend each week. The numerator is an estimate of the total number of dollars children spend per week and its denominator is the number of persons who have spending money.

Examples of Quantitative Questions:

Q: About how much money do you usually get each week to spend on yourself or to save?
R: $|_|_|_|.0|0| (Write in amount or 0 if you get no money)

Q: If yes, what age were you when you first did this (had a drink of alcohol)?
R: I was |_|_| 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 in the past seven days prior to the survey by students in grade 9 (in Quebec secondary 3) multiply the value reported in the derived variable DVCIGWK (number of cigarettes smoked in the past seven days prior to the survey) by the final weight for the record, then sum this value over all records with DVCIGWK < 996 and GRADE = 09.

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 in the past seven days prior to the survey by students in grade 9,

a) estimate the total number of cigarettes smoked in the past seven days prior to the survey by students in grade 9 $\left(\hat{X}\right)$ as described above,

b) estimate the number of students in grade 9 (in Quebec secondary 3) $\left(\hat{Y}\right)$ in this category by summing the final weights of all records with DVCIGWK < 996 and GRADE = 09, then

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

9.4 Guidelines for Statistical Analysis

The 2002 Youth Smoking 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. 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 = male;
2) calculate the AVERAGE weight for these records by summing the original child weights from the microdata file for these records and then dividing by the number of children who reported SEX = male;
3) for each of these respondents, calculate a RESCALED weight equal to the original child weight divided by the AVERAGE weight;
4) perform the analysis for these children 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 2002 Youth Smoking 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 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 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 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 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: "Please be warned that these estimates [flagged with the letter U] 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 2002 Youth Smoking Survey

The following table provides an indication of the precision of population estimates as it shows the release cut-offs associated with each of the three quality levels presented in the previous section. These cut-offs are derived from the coefficient of variation (CV) tables discussed in Chapter 10.

For example, the table shows that the quality of a weighted estimate of 1,000 people possessing a given characteristic in Newfoundland and Labrador is marginal.

Note that these cut-offs apply to estimates of population totals only. To estimate ratios, users should not use the numerator value (nor the denominator) in order to find the corresponding quality level. Rule 4 in Section 10.1 and Example 4 in Section 10.1.1 explains the correct procedure to be used for ratios.

Province	**Acceptable CV 0.0% – 16.5%**	**Marginal CV 16.6% – 33.3%**	**Unacceptable CV > 33.3%**
Newfoundland and Labrador	1,000 & over	500 to < 1,000	under 500
Prince Edward Island	500 & over	250 to < 500	under 250
Nova Scotia	2,000 & over	500 to < 2,000	under 500
New Brunswick	1,500 & over	500 to < 1,500	under 500
Quebec	9,000 & over	2,000 to < 9,000	under 2,000
Ontario	16,500 & over	4,000 to < 16,500	under 4,000
Manitoba	3,000 & over	1,000 to < 3,000	under 1,000
Saskatchewan	2,500 & over	500 to < 2,500	under 500
Alberta	9,500 & over	2,500 to < 9,500	under 2,500
British Columbia	6,000 & over	1,500 to < 6,000	under 1,500
Canada	**10,000 & over**	**2,500 to < 10,000**	**under 2,500**

Grade	**Acceptable CV 0.0% – 16.5%**	**Marginal CV 16.6% – 33.3%**	**Unacceptable CV > 33.3%**
Grade 5	11,000 & over	3,000 to < 11,000	under 3,000
Grade 6	10,500 & over	2,500 to < 10,500	under 2,500
Grade 7	12,000 & over	3,000 to < 12,000	under 3,000
Grade 8	8,500 & over	2,000 to < 8,500	under 2,000
Gade 9	9,000 & over	2,000 to < 9,000	under 2,000
Total	**10,000 & over**	**2,500 to < 10,000**	**under 2,500**

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 2002 Youth Smoking Survey (YSS).

Province	Design Effect	Sample Size	Population
Newfoundland and Labrador	1.49	1,574	33,944
Prince Edward Island	1.66	1,091	10,087
Nova Scotia	1.43	1,784	61,566
New Brunswick	1.61	1,656	49,049
Quebec	1.65	3,229	487,440
Ontario	1.56	2,583	770,598
Manitoba	1.75	1,534	76,157
Saskatchewan	1.63	1,707	67,600
Alberta	1.82	1,442	219,143
British Columbia	1.54	2,418	251,921
Canada	**2.53**	**19,018**	**2,027,506**

Grade	Design Effect	Sample Size	Population
Grade 5	2.78	3,544	396,859
Grade 6	2.73	3,717	406,037
Grade 7	2.95	3,725	424,837
Grade 8	2.28	3,960	404,250
Grade 9	2.57	4,072	395,522
Total	**2.53**	**19,018**	**2,027,506**

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. 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 is most likely unacceptable and Statistics Canada recommends not to release such an estimate, regardless of the value of the coefficient of variation.

10.1 How to Use the Coefficient of Variation 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 Children 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 Children 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 kids who have tried smoking, even just a few puffs is more reliable than the estimated number of kids who have smoked 100 or more cigarettes in their lifetime. (Note that in the tables the coefficients of variation 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 $\left(\hat{d} = \hat{X}_1 - \hat{X}_2\right)$ 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 kids who smoked in the last 30 days and the numerator is the number of kids who smoked every day in the last 30 days.

In the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of girls who smoked at least one whole cigarette as compared to the number of boys who smoked at least one whole cigarette, 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 Coefficient of Variation Tables for Categorical Estimates

The following examples based on the 2002 Youth Smoking Survey are included to assist users in applying the foregoing rules.

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

Suppose that a user estimates that 53,937 kids were current smokers at the time of the survey. How does the user determine the coefficient of variation of this estimate?

1) Refer to the coefficient of variation table for CANADA.

2002 Youth Smoking Survey - Public Use Microdata File

Approximate Sampling Variability Tables for Canada

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	51.7	51.4	51.2	50.4	49.0	47.7	46.2	44.8	43.2	41.7	40.0	36.6	28.3	16.3
2	36.5	36.4	36.2	35.6	34.7	33.7	32.7	31.7	30.6	29.5	28.3	25.8	20.0	11.6
3	********	29.7	29.5	29.1	28.3	27.5	26.7	25.8	25.0	24.1	23.1	21.1	16.3	9.4
4	********	25.7	25.6	25.2	24.5	23.8	23.1	22.4	21.6	20.8	20.0	18.3	14.2	8.2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	********	11.5	11.4	11.3	11.0	10.7	10.3	10.0	9.7	9.3	9.0	8.2	6.3	3.7
21	********	********	11.2	11.0	10.7	10.4	10.1	9.8	9.4	9.1	8.7	8.0	6.2	3.6
22	********	********	10.9	10.7	10.5	10.2	9.9	9.5	9.2	8.9	8.5	7.8	6.0	3.5
23	********	********	10.7	10.5	10.2	9.9	9.6	9.3	9.0	8.7	8.3	7.6	5.9	3.4
24	********	********	10.4	10.3	**10.0**	9.7	9.4	9.1	8.8	8.5	8.2	7.5	5.8	3.3
25	********	********	10.2	10.1	9.8	9.5	9.2	9.0	8.6	8.3	8.0	7.3	5.7	3.3
30	********	********	9.3	9.2	9.0	**8.7**	8.4	8.2	7.9	7.6	7.3	6.7	5.2	3.0
35	********	********	8.6	8.5	8.3	8.1	7.8	7.6	7.3	7.0	6.8	6.2	4.8	2.8
40	********	********	8.1	8.0	7.8	7.5	7.3	7.1	6.8	6.6	6.3	5.8	4.5	2.6
45	********	********	********	7.5	7.3	7.1	6.9	6.7	6.4	6.2	6.0	5.4	4.2	2.4
50	********	********	********	7.1	6.9	6.7	6.5	6.3	6.1	5.9	5.7	5.2	4.0	2.3
55	********	********	********	**6.8**	**6.6**	6.4	6.2	6.0	5.8	5.6	5.4	4.9	3.8	2.2
60	********	********	********	6.5	6.3	6.2	6.0	5.8	5.6	5.4	5.2	4.7	3.7	2.1
65	********	********	********	6.2	6.1	5.9	5.7	5.6	5.4	5.2	5.0	4.5	3.5	2.0
70	********	********	********	6.0	5.9	5.7	5.5	5.4	5.2	5.0	4.8	4.4	3.4	2.0
75	********	********	********	5.8	5.7	5.5	5.3	5.2	5.0	4.8	4.6	4.2	3.3	1.9
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
125	********	********	********	********	4.4	4.3	4.1	4.0	3.9	3.7	3.6	3.3	2.5	1.5
150	********	********	********	********	4.0	3.9	3.8	3.7	3.5	3.4	3.3	3.0	2.3	1.3
200	********	********	********	********	**3.5**	3.4	3.3	3.2	3.1	2.9	2.8	2.6	2.0	1.2
250	********	********	********	********	********	**3.0**	2.9	2.8	2.7	2.6	2.5	2.3	1.8	1.0
300	********	********	********	********	********	2.8	2.7	2.6	2.5	2.4	2.3	2.1	1.6	0.9
350	********	********	********	********	********	********	2.5	2.4	2.3	2.2	2.1	2.0	1.5	0.9
400	********	********	********	********	********	********	2.3	2.2	2.2	2.1	2.0	1.8	1.4	0.8
450	********	********	********	********	********	********	********	2.1	2.0	2.0	1.9	1.7	1.3	0.8
500	********	********	********	********	********	********	********	2.0	1.9	1.9	1.8	1.6	1.3	0.7
750	********	********	********	********	********	********	********	********	********	********	1.5	1.3	1.0	0.6
1000	********	********	********	********	********	********	********	********	********	********	********	1.2	0.9	0.5
1500	********	********	********	********	********	********	********	********	********	********	********	********	********	0.4

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

2) The estimated aggregate (53,937) 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 55,000.

3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, namely, 6.8%.

4) So the approximate coefficient of variation of the estimate is 6.8%. The finding that there were 53,937 (to be rounded according to the rounding guidelines in Section 9.1) kids were current smokers at the time of the survey is publishable with no qualifications.

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

Suppose that the user estimates that 53,937 / 457,087 = 11.8% of kids who have ever tried cigarette smoking, even just a few puffs, were current smokers at the time of the survey. How does the user determine the coefficient of variation of this estimate?

1) Refer to the coefficient of variation table for CANADA.

2) Because the estimate is a percentage which is based on a subset of the total population (i.e., kids who have ever tried cigarette smoking, even just a few puffs), it is necessary to use both the percentage (11.8%) and the numerator portion of the percentage (53,937) in determining the coefficient of variation.

3) The numerator, 53,937, 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 55,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the percentage closest to it, 10.0%.

4) The figure at the intersection of the row and column used, namely 6.6% is the coefficient of variation to be used.

5) So the approximate coefficient of variation of the estimate is 6.6%. The finding that 11.8% of kids who have ever tried cigarette smoking, even just a few puffs, were current smokers at the time of the survey can be published with no qualifications.

Example 3: Estimates of Differences Between Aggregates or Percentages

Suppose that a user estimates that 30,053 / 222,601 = 13.5% of girls who have ever tried cigarette smoking, even just a few puffs, were current smokers at the time of the survey, while 23,884 / 234,487 = 10.2% of boys who have ever tried cigarette smoking, even just a few puffs, were current smokers at the time of the survey. How does the user determine the coefficient of variation of the difference between these two estimates?

1) Using the CANADA coefficient of variation table in the same manner as described in Example 2 gives the CV of the estimate for girls as 8.7%, and the CV of the estimate for boys as 10.0%.

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 (girls) , $\hat{X}_2$ is estimate 2 (boys), 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 $\hat{d}$ = 0.135 - 0.102 = 0.033 is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(0.135)(0.087)]^2 + [(0.102)(0.100)]^2} \\ &= \sqrt{(0.000138) + (0.000104)} \\ &= 0.016\end{aligned}$$

3) The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d}$ = 0.016 / 0.033 = 0.485.

4) So the approximate coefficient of variation of the difference between the estimates is 48.5%. This estimate is considered unacceptable 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 U (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 222,601 girls have tried cigarette smoking, even just a few puffs, while 234,487 boys have tried cigarette smoking, even just a few puffs. The user is interested in comparing the estimate of girls versus that of boys 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 girls who have ever tried cigarette smoking, even just a few puffs. The denominator of the estimate ($\hat{X}_2$) is the number of boys who have ever tried cigarette smoking, even just a few puffs.

2) Refer to the coefficient of variation table for CANADA.

3) The numerator of this ratio estimate is 222,601. 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, 3.5%.

4) The denominator of this ratio estimate is 234,487. 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, 3.0%.

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:

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(0.035)^2 + (0.030)^2} \\ &= \sqrt{0.001225 + 0.0009} \\ &= 0.046\end{aligned}$$

6) The obtained ratio of girls versus boys who have ever tried cigarette smoking, even just a few puffs is 222,601 / 234,487, which is 0.95:1 (to be rounded according to the rounding guidelines in Section 9.1). The coefficient of variation of this estimate is 4.6%, which is releasable with no qualifications.

Example 5: Estimates of Differences of Ratios

Suppose that the user estimates that the ratio of girls who have ever tried cigarette smoking, even just a few puffs, to boys who have ever tried cigarette smoking, even just a few puffs, is 0.79:1 for Alberta while it is 1.04:1 for Quebec. The user is interested in comparing the two ratios to see if there is a statistical difference between them. How does the user determine the coefficient of variation of the difference?

1) First calculate the approximate coefficient of variation for the Alberta ratio $\left(\hat{R}_1\right)$ and the Quebec ratio $\left(\hat{R}_2\right)$ as in Example 4. The approximate CV for the Alberta ratio is 11.4% and 4.7% for Quebec.

2002 Youth Smoking Survey - Public Use Microdata File

Approximate Sampling Variability Tables for Alberta

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	********	52.2	51.9	51.1	49.7	48.3	46.9	45.4	43.9	42.3	40.6	37.1	28.7	16.6
2	********	36.9	36.7	36.1	35.2	34.2	33.2	32.1	31.0	29.9	28.7	26.2	20.3	11.7
3	********	********	30.0	29.5	28.7	27.9	27.1	26.2	25.3	24.4	23.4	21.4	16.6	9.6
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	********	********	********	********	12.1	11.7	11.4	11.0	10.6	10.2	9.8	9.0	7.0	4.0
18	********	********	********	********	11.7	11.4	11.1	10.7	10.3	10.0	9.6	8.7	6.8	3.9
19	********	********	********	********	**11.4**	11.1	10.8	10.4	10.1	9.7	9.3	8.5	6.6	3.8
20	********	********	********	********	11.1	10.8	10.5	10.2	9.8	9.4	9.1	8.3	6.4	3.7
21	********	********	********	********	10.9	10.5	10.2	9.9	9.6	9.2	8.9	8.1	6.3	3.6
22	********	********	********	********	********	10.3	10.0	9.7	9.4	9.0	8.7	7.9	6.1	3.5
23	********	********	********	********	********	10.1	9.8	9.5	9.1	8.8	8.5	7.7	6.0	3.5
24	********	********	********	********	********	**9.9**	9.6	9.3	9.0	8.6	8.3	7.6	5.9	3.4
25	********	********	********	********	********	9.7	9.4	9.1	8.8	8.5	8.1	7.4	5.7	3.3
30	********	********	********	********	********	8.8	8.6	8.3	8.0	7.7	7.4	6.8	5.2	3.0
35	********	********	********	********	********	********	7.9	7.7	7.4	7.1	6.9	6.3	4.9	2.8
40	********	********	********	********	********	********	7.4	7.2	6.9	6.7	6.4	5.9	4.5	2.6
45	********	********	********	********	********	********	********	6.8	6.5	6.3	6.1	5.5	4.3	2.5
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
95	********	********	********	********	********	********	********	********	********	********	********	3.8	2.9	1.7
100	********	********	********	********	********	********	********	********	********	********	********	3.7	2.9	1.7
125	********	********	********	********	********	********	********	********	********	********	********	********	2.6	1.5
150	********	********	********	********	********	********	********	********	********	********	********	********	2.3	1.4

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

2002 Youth Smoking Survey - Public Use Microdata File

Approximate Sampling Variability Tables for Quebec

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	******	49.5	49.2	48.5	47.2	45.9	44.5	43.1	41.6	40.1	38.5	35.2	27.2	15.7
2	******	35.0	34.8	34.3	33.4	32.4	31.5	30.5	29.4	28.4	27.2	24.9	19.3	11.1
3	******	28.6	28.4	28.0	27.2	26.5	25.7	24.9	24.0	23.2	22.2	20.3	15.7	9.1
4	******	24.7	24.6	24.2	23.6	22.9	22.2	21.5	20.8	20.1	19.3	17.6	13.6	7.9
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
75	******	******	******	******	******	******	5.1	5.0	4.8	4.6	4.4	4.1	3.1	1.8
80	******	******	******	******	******	******	5.0	4.8	4.7	4.5	4.3	3.9	3.0	1.8
85	******	******	******	******	******	******	4.8	4.7	4.5	4.3	4.2	3.8	3.0	1.7
90	******	******	******	******	******	******	**4.7**	4.5	4.4	4.2	4.1	3.7	2.9	1.7
95	******	******	******	******	******	******	4.6	4.4	4.3	4.1	4.0	3.6	2.8	1.6
100	******	******	******	******	******	******	******	4.3	4.2	4.0	3.9	3.5	2.7	1.6
125	******	******	******	******	******	******	******	******	3.7	3.6	3.4	3.1	2.4	1.4
150	******	******	******	******	******	******	******	******	******	3.3	3.1	2.9	2.2	1.3
200	******	******	******	******	******	******	******	******	******	******	******	2.5	1.9	1.1
250	******	******	******	******	******	******	******	******	******	******	******	******	1.7	1.0
300	******	******	******	******	******	******	******	******	******	******	******	******	1.6	0.9
350	******	******	******	******	******	******	******	******	******	******	******	******	******	0.8
400	******	******	******	******	******	******	******	******	******	******	******	******	******	0.8

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

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

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

where α_1 and α_2 are the coefficients of variation of $\hat{R}_1$ and $\hat{R}_2$ respectively. That is, the standard error of the difference $\hat{d}$ = 0.79 - 1.04 = -0.25 is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{\left[(0.79)(0.114)\right]^2 + \left[(1.04)(0.047)\right]^2} \\ &= \sqrt{(0.0081) + (0.0024)} \\ &= 0.102\end{aligned}$$

3) The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d}$ = 0.102 / (-0.25) = -0.408.

4) So the approximate coefficient of variation of the difference between the estimates is 40.8%. This estimate is considered unacceptable 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 U (or some similar identifier) and be accompanied by a warning to caution subsequent users about the high levels of error, associated with the estimate.

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 Coefficient of Variation Tables to Obtain Confidence Limits

A 95% confidence interval for the estimated proportion of kids who have ever tried cigarette smoking, even just a few puffs, and were current smokers at the time of the survey (from Example 2, Section 10.1.1) would be calculated as follows:

$\hat{X}$ = 11.8% (or expressed as a proportion 0.118)

t = 2

$\alpha_{\hat{x}}$ = 6.6% (0.066 expressed as a proportion) is the coefficient of variation of this estimate as determined from the tables.

$CI_{\hat{x}}$ = {0.118 - (2) (0.118) (0.066), 0.118 + (2) (0.118) (0.066)}

$CI_{\hat{x}}$ = {0.118 - 0.016, 0.118 + 0.016}

$CI_{\hat{x}}$ = {0.102, 0.134}

With 95% confidence it can be said that between 10.2% and 13.4% of kids who have ever tried cigarette smoking, even just a few puffs, were current smokers at the time of the survey.

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 girls who have ever tried cigarette smoking, even just a few puffs, and were current smokers and the proportion of boys who have ever tried cigarette smoking, even just a few puffs, and were current smokers. From Example 3, Section 10.1.1, the standard error of the difference between these two estimates was found to be 0.016. Hence,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{0.135 - 0.102}{0.016} = \frac{0.033}{0.016} = 2.06$$

Since t = 2.06 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 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 2002 Youth Smoking Survey 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 in a week by current smokers would be greater than the coefficient of variation of the corresponding proportion of kids who are current smokers. Hence, if the coefficient of variation of the proportion is not releasable, then the coefficient of variation of the corresponding quantitative estimate will also not be 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

Master File and Public Use Microdata File:

Refer to YSS2002_PUMF_CVTabsE.pdf for the coefficient of variation tables for the public use microdata file (PUMF) for the 2002 YSS.

Share File:

Refer to YSS2002_Share_CVTabsE.pdf for the coefficient of variation tables for the share microdata file for the 2002 YSS.

11.0 Weighting

Statistical weights were placed on each record to represent the number of sampled persons that the record represents. The weighting for the Youth Smoking Survey (YSS) consisted of several steps which are described in the following paragraphs.

1) Initial Sampling Weight (School Weight)

   The first step is to calculate the initial weight (Weight1) for each selected unit (school-grade). For a given unit, this is equal to the inverse of the probability of selection within the stratum. This probability is proportional to the number of students at the school for the given grade. Note for Québec, the initial sampling weight was calculated differently since the probability of selection was modified in order to reduce the overlap with a provincial survey conducted at the same time as the YSS (see Section 5.3)

2) Adjustment for Non-response at the School Level

   Among the originally selected school-grade units, some non-response was observed. Non-response at the school level can be due to several factors such as school board refusals, school refusals or interviewers were unable to complete the interview within the allotted collection period. The school level non-response adjustment for units belonging to strata h in grade g is defined as:

$$adj_{school\_nr} = \frac{Number\ of\ selected\ schools\ in\ strata\ h\ for\ grade\ g}{Number\ of\ responding\ schools\ in\ strata\ h\ for\ grade\ g}$$

   The resulting weight from Step 2) is:

$$Weight\,2 = Weight\,1 * adj_{school\_nr}$$

3) Adjustment for the Selection of a Class (Class Weight)

   This adjustment relates to the second stage of sampling, when a class is selected at random from all the classes of the same grade in the selected school. Since only one class is selected per school-grade, the adjustment consists of multiplying the weight obtained from the preceding stage by the total number of classes in the school for this grade. This number is obtained from the Classroom Selection Form.

$$Weight3\ (Class\ Weight) = Weight2 * Number\ of\ classes$$

4) Adjustment for Class Non-response

   This adjustment takes care of the non-response at the class level. Non-response at the class level is defined as any cases where the number of classes is known (and is positive) but for which there are no responding students. The adjustment factor is defined as:

$$adj_{class\_nr} = \frac{Sum\,of\,Weight3\,for\,responding\,classes + Sum\,of\,Weight3\,for\,non-responding\,classes}{Sum\,of\,Weight3\,for\,responding\,classes}$$

   The resulting weight from Step 4) is:

$$Weight\,4 = Weight\,3 * adj_{class\_nr}$$

Note, since all the students in the selected classes are surveyed, this step also provides the student weight.

5) Adjustment for Student Non-response

This adjustment is intended to compensate for non-response at the student level. The main reasons for this type of non-response are: parental consent not obtained, student refused to participate or student was not in class on the day of the interview. The adjustment consists of multiplying the weight (Weight4) resulting from Step 4) by the following ratio.

$$adj_{student\_nr} = \frac{Number\ of\ eligible\ students\ in\ the\ selected\ class}{Number\ of\ responding\ students\ in\ the\ selected\ class}$$

The resulting weight from Step 5) is:

$$Weight\,5 = Weight\,4 * adj_{student\_nr}$$

6) Post-stratification Adjustment

For the Master File and Public Use Microdata File:

The sampling weights are adjusted to agree with the enrolment counts for certain groupings (post-strata). The enrolment counts are obtained using the most recent file available at Statistics Canada and are adjusted based on demographic projections. Post-stratification is performed by province-grade-sex. The ratio of the actual number of students to the number of students estimated by the sampling design in a given post-stratum represents the adjustment. For units belonging to post-stratum p, the post-stratification adjustment is defined as:

$$adj_{post\_strata} = \frac{Enrolment\ totals\ for\ post-stratum_p}{Sum\ of\ Weight\,5\ for\ records\ in\ post-stratum_p}$$

The final sampling weight attached to each record is the product of the adjusted student weight multiplied by adj_{post_strata}.

$$WTPP = Weight\,5 * adj_{post\_strata}$$

For the Share File:

To produce the share file weights, in addition to a change in the post-stratification step, a non-response adjustment was applied to account for non-sharers by using non-response groups. There were 62 non-response groups in different combinations of province, sex, language, grade, stratum and smoking variables. The sharing agreement adjustment factor is defined as:

$$adj_{share} = \frac{Sum\ of\ Weight\,5\ for\ sharer\ records + Sum\ of\ Weight\,5\ for\ non\_sharer\ records}{Sum\ of\ Weight\,5\ for\ sharer\ records}$$

for every record belonging to a non-response group.

The resulting weight from Step 6) is:

$$Weight\ \ 6 = Weight\ \ 5 * adj_{share}$$

Finally, the last step in weighting the share file consists of post-stratifying to province-grade-sex counts as well as calibrating to the estimated number of kids who smoked in the last 30 days (Y_Q19) using the weighted master file records. The final share weights are called WTPS.

12.0 Questionnaires

There were three questionnaires used to collect the information for the Youth Smoking Survey (YSS).

YSS2002_Grades7to9_QuestE.pdf - contains the English questionnaire administered to students in grades 7 to 9 (in Quebec secondary school grades 1 to 3).

YSS2002_Grades5&6_QuestE.pdf - contains the English questionnaire administered to students in grades 5 and 6.

YSS2002_Parent_QuestE.pdf - contains the English questionnaire administered to parents.

Note: The questionnaire for students in grades 5 and 6 is identical to the questionnaire used for the older grades, except that it does not include questions referring to experiences with alcohol and drugs (questions 65a to 79a).

13.0 Record Layout with Univariate Frequencies

Master File:

See YSS2002_Master_CdBk.pdf for the record layout with univariate counts for the 2002 Youth Smoking Survey public use microdata file.

Public Use Microdata File:

See YSS2002_PUMF_CdBk.pdf for the record layout with univariate counts for the 2002 Youth Smoking Survey public use microdata file.

Share File:

See YSS2002_Share_CdBk.pdf for the record layout with univariate counts for the 2002 Youth Smoking Survey share microdata file.