

**Microdata User Guide**

**National Graduates Survey**

**Class of 2005**

Public Use Microdata File



Statistics  
Canada

Statistique  
Canada

Canada



## Table of Contents

<b>1.0</b>	<b>Introduction.....</b>	<b>5</b>
<b>2.0</b>	<b>Background.....</b>	<b>7</b>
<b>3.0</b>	<b>Objectives.....</b>	<b>9</b>
<b>4.0</b>	<b>Content.....</b>	<b>11</b>
4.1	Concepts and Definitions .....	11
4.2	Uses.....	16
<b>5.0</b>	<b>Survey Methodology.....</b>	<b>17</b>
5.1	Target Population.....	17
5.2	Survey Frame.....	17
5.2.1	CIP coding.....	18
5.2.2	Removal of duplicates .....	18
5.3	Survey Design.....	18
5.3.1	Longitudinal Sample.....	18
5.3.2	Stratification.....	18
5.4	Sample Allocation, Selection and Size .....	19
<b>6.0</b>	<b>Data Collection.....</b>	<b>21</b>
<b>7.0</b>	<b>Data Processing.....</b>	<b>23</b>
7.1	Data Capture.....	23
7.2	Editing .....	23
7.3	Coding of Open-ended Questions .....	24
7.3.1	Coding of Education Programs .....	24
7.3.2	Coding of Industry and Occupation.....	24
7.3.3	Coding of “Other – Specify” Answers.....	24
7.4	Imputation .....	24
7.5	Creation of Derived Variables.....	25
<b>8.0</b>	<b>Response Rates .....</b>	<b>27</b>
<b>9.0</b>	<b>Treatment of Non-response and Weighting .....</b>	<b>31</b>
9.1	Sampling Weight (phase 1).....	31
9.2	Non-response adjustment (phase 2).....	32
9.3	Subsampling adjustment for the PUMF (phase 3).....	32
9.4	Post-stratification.....	33
<b>10.0</b>	<b>Data Quality.....</b>	<b>35</b>
10.1	Sampling Errors .....	35
10.2	Non-sampling Errors .....	36
10.3	Non-response.....	36
10.4	Coverage.....	36
<b>11.0</b>	<b>Guidelines for Tabulation Analysis and Release .....</b>	<b>37</b>
11.1	Rounding Guidelines.....	37
11.2	Sample Weighting Guidelines for Tabulation.....	37
11.3	Definitions of Types of Estimates: Categorical and Quantitative.....	38
11.3.1	Tabulation of Categorical Estimates .....	39
11.3.2	Tabulation of Quantitative Estimates .....	39

11.4	Guidelines for Statistical Analysis .....	40
11.5	Release Guidelines .....	40
11.6	Release Cut-offs for the PUMF .....	41
<b>12.0</b>	<b>Approximate Sampling Variability Tables .....</b>	<b>43</b>
12.1	How to Use the Coefficient of Variation Tables for Categorical Estimates .....	43
12.1.1	Examples of Using the Coefficient of Variation Tables for Categorical Estimates .....	45
12.2	How to Use the Coefficient of Variation Tables to Obtain Confidence Limits .....	48
12.2.1	Example of Using the Coefficient of Variation Tables to Obtain Confidence Limits ....	49
12.3	How to Use the Coefficient of Variation Tables to Do a T-test .....	49
12.3.1	Example of Using the Coefficient of Variation Tables to Do a T-test .....	50
12.4	Coefficients of Variation for Quantitative Estimates .....	50
12.5	Coefficient of Variation Tables .....	50
<b>13.0</b>	<b>Questionnaire, Code Sheets and Documentation of Derived Variables .....</b>	<b>51</b>
<b>14.0</b>	<b>Record Layout with Univariate Frequencies .....</b>	<b>53</b>

## 1.0 Introduction

The National Graduates Survey – Class of 2005 (NGS2005) was conducted by Statistics Canada from May to September 2007. This manual has been produced to facilitate the manipulation of the public use microdata file.

The public use microdata file, or PUMF, contains a reduced list of variables compared to the master file. The need to preserve the confidentiality of respondents dictated that many variables that could have been used to identify individuals (including all geographic information) be removed from the file. In addition, all continuous variables such as those relating to income, student loans or age at graduation, were converted to categorical variables, and many existing categorical variables were grouped into a smaller number of categories. Finally, local suppression was used where necessary to further protect confidentiality. Every effort was made to preserve the analytical utility of the data during this process.

It is also important to note that this PUMF contains fewer records than the master file. As an initial measure of diminishing the risk of disclosure, a subsample of the records from the master file was drawn. The PUMF therefore is made up of 16,081 records, or roughly half the number in the master. Users should be aware that estimates produced using the subsample may not correspond exactly to those produced by Statistics Canada using the master file.

Users requiring access to information excluded from the microdata files may purchase custom tabulations.

This document retains most of the content from the original user guide for the NGS master microdata file for informational purposes. Notes have been added to indicate where full content is not applicable to the PUMF.

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

### Statistics Canada

Client Services  
Centre for Education Statistics  
Room SC-2000 B, Main Building  
150 Tunney's Pasture Driveway  
Ottawa, Ontario  
K1A 0T6

Telephone: (613) 951-7608 or call toll-free 1 800 307-3382

Fax: (613) 951-4441

E-mail: [educationstats@statcan.gc.ca](mailto:educationstats@statcan.gc.ca)



## 2.0 Background

In 1978, Statistics Canada conducted a survey on the labour market experiences of 1976 graduates from universities and community colleges in Canada. In 1984, a similar survey, the National Graduates Survey (NGS) of 1982 graduates was sponsored jointly by the Department of the Secretary of State and Employment and Immigration Canada. The 1984 NGS expanded on the content of the previous survey and extended the population base to include completers of trade/vocational programs in addition to graduates from community colleges and universities.

Since these two surveys in 1978 and 1984, a series of graduate surveys has been completed on the labour market experiences of graduates from universities and community colleges in Canada.

The following is a summary of the graduate surveys conducted by Statistics Canada.

Graduation Year	Survey Year	Survey Name
1976	1978	Survey of 1976 Graduates of Post-Secondary Programs
1982	1984	Survey of 1982 Graduates (S82G) (also known as the National Graduates Survey or NGS)
1982	1987	Follow-up of 1982 Graduates (F82G) (also known as the Follow-up of Graduates or FOG)
1986	1988	Survey of 1986 Graduates (S86G)
1986	1991	Follow-up of 1986 Graduates (F86G)
1990	1992	Survey of 1990 Graduates (S90G)
1990	1995	Follow-up of 1990 Graduates (F90G)
1995	1997	Survey of 1995 Graduates (S95G)
1995	2000	Follow-up of 1995 Graduates (F95G)
2000	2002	National Graduates Survey - Class of 2000 (NGS2000)
2000	2005	Follow-up of Graduates Survey - Class of 2000 (FOG2000)
2005	2007	National Graduates Survey - Class of 2005 (NGS2005)

The survey contains data on: the link between education experience and labour market outcomes; information regarding the job held in the week prior to the interview and the first job after graduation; financial and loan information; additional education pursued after graduation; and socio-economic background.

In comparison to the NGS2000 questionnaire, the main changes that were made to the NGS2005 questionnaire are the following:

- a question on the language of instruction of the program completed in 2005 was added;
- a question on components of the program taken outside of Canada was added, as well as the duration if applicable;
- in addition to information on the job held in the week prior to the interview, information was collected on the first job after graduation rather than on all jobs held between graduation and the week prior to the interview;
- details on unpaid leave periods were dropped;

- the number of questions on educational programs taken after graduation was reduced; and
- with the exception of the two first questions, the module regarding job-related training was dropped.

Note, however, that for confidentiality reasons, information specific to graduates who took trade/vocational programs, graduates who lived in the United States, and information on components of programs taken outside of Canada is not available on the PUMF.



### 3.0 Objectives

The survey's primary objective is to obtain information on the labour market experiences of graduates entering the labour market, focusing on employment, occupations and the relationship between jobs and education.

The survey's key data objectives are:

- To obtain information for labour market analysis of a key youth group at an important time, focusing on education, training, employment, occupations and geographic mobility. The data and analysis will be useful for policy development.
- To obtain information on the exposure of graduates to additional learning opportunities.
- To extend available information required to improve occupational supply and demand projection models for various occupational categories.
- To obtain data regarding longer-term labour market experiences of graduates, with special emphasis on employment and occupations, for use in counselling on career and post-secondary education course selection.
- To obtain information on labour market experiences of members of target groups (such as women, native people and the disabled), which permits longitudinal and comparative analysis useful in the formulation of job equity policies.
- To gain a better understanding of school-work transitions and returns to human capital.
- To gain a better understanding of post-secondary education financing.
- To obtain more detailed information on knowledge and skills.



## 4.0 Content

The following table describes the content of each section of the National Graduates Survey – Class of 2005 (NGS2005) questionnaire.

Section	Content
Program confirmation (PR)	Graduates are asked to give information about the program they graduated from in 2005.
Activities before graduation (AB)	Contains information about the graduate's activities (i.e. employment, education, etc.) prior to graduating in 2005.
Graduates who live/lived in the United States (US and MU)*	Identifies graduates who moved to the United States (US) after graduation and obtains information on their activities in the US and also about their return to Canada, if applicable.
Activities last week (LF)	Asks about the graduate's labour force activity the week before the interview.
First job after graduation (EM)	Contains information about the first job held by the graduate after graduation.
First job in the United States (FU)*	Contains information about the first job in the United States, if applicable and not already collected in sections LF or EM.
Education programs (ED) Education program description (EP)	Asks about completed and uncompleted educational programs taken after graduation.
Student loans (SL)	Asks questions about student loans and finances.
Higher education (HE)	Contains information about the intentions of graduates to pursue a Master's degree or a Ph.D.
Demographic characteristics (DE and DEM)	Asks general questions such as marital status, number of dependant children, income, and disabilities.

\*Note: this information is not available on the PUMF.

## 4.1 Concepts and Definitions

### Graduation date

For the purpose of this survey, the graduation date is the year and month in which the graduate completed the requirements of his/her program. To complete the requirements of the program, graduates must have written and passed the last exam, submitted the last paper, report or project for a program or defended a thesis. The variables PR\_D11Y and PR\_D11M in the master codebook contain the graduation date. These variables are not available on the PUMF.

**Graduates who moved to live in the United States**

Graduates who live in the United States, or lived in the United States since their graduation but have returned to Canada, are included in the survey. They may have moved to attend school, to work or to accompany a partner or spouse. Anyone who visited or vacationed in the United States temporarily is not considered to have moved. These variables are not available on the PUMF.

**Transition after completing post-secondary studies**

A number of modules in the survey are devoted to obtaining information on the graduate's activities after completing his/her post-secondary studies. The information found in these modules allows for a detailed analysis on the graduate's transition after completing his/her post-secondary studies.

- The LF module asks about the graduate's labour force activities during the week prior to the interview (i.e., employed, unemployed, or not in the labour force). Detailed information on the job held in the week prior to the interview is also collected.
- The EM module obtains information about the first employer the graduate worked for after graduation, and detailed information about the job held with this employer (or equivalent information if the respondent was self-employed).
- The ED and EP modules collect information on completed and uncompleted educational programs taken after graduation when these programs lead towards a diploma, certificate or degree that would take someone three months or more to complete if taken full-time.

**Main job**

The job involving the greatest number of hours per week.

**Paid worker**

A person who works for others (i.e. works for an employer). Payment may be in cash (salary, wages, tips, commissions) or "payment in kind" (payment in goods or services rather than money). Such employer-employee relationships almost always involve some legal obligations on the part of the employer, must deduct and remit income tax and Canada/Québec Pension Plan premiums, etc.

**Self-employed**

A person who works directly for himself/herself. A self-employed person may or may not have a business, farm or professional practice. Examples of self-employed persons with a business would be: a man with his own barber shop, or a woman with her own medical practice.

Examples of self-employed persons without a business include:

- a cleaning person working for a number of people in their homes;
- a freelance writer, a paper carrier;
- a general handyman;
- a caregiver who works for a number of people.

**Unpaid family worker**

An unpaid family worker is someone who worked without pay on a farm or in a business owned and operated by another family member living in the same household. The work done must

contribute directly to the operation of a family farm or family business. This variable is not available on the PUMF.

### **Permanent job**

A permanent job is one that is expected to last as long as the employee wants it and as long as business conditions permit. That is, the employer did not hire the employee on the understanding the job would end at a specified time in the near future. Sometimes permanent jobs are referred to as indeterminate, since they have no pre-specified date of termination.

### **Non-permanent job**

A job that is not permanent is one that has a predetermined date on which it will end or will end as soon as a specified project is completed. The employer has hired the employee on the understanding that the job will end at this specified time in the near future.

**Seasonal job** - This occurs in industries where employment levels rise and fall with the seasons (seasonal employment).

Examples: farming, fishing, logging and the tourist industry.

**Temporary, term or contract job (non-seasonal)** - A job in which there was a definite indication from the employer before the job was accepted that the job would terminate at a specified point in time, or at the end of a particular task or project.

**Casual job** - Is one of the following:

- respondent has work hours that vary substantially from one week to the next;
- respondent is called to work by the employer when the need arises, not on a pre-arranged schedule; or
- respondent does not usually get paid for time not worked and there is no indication from the employer that he/she will be called to work on a regular, long-standing basis.

### **Number of (paid) hours worked per week**

Serves to separate the employed into full-time (30 hours of work or more per week) and part-time (less than 30 hours of work per week) workers.

Number of paid hours usually worked is asked of employees.

Number of hours usually worked is asked of self-employed persons.

### **Wages or salary**

For employees, this refers to wages before deductions by the employer for taxes, employment insurance (EI), government pension plans (CPP/QPP), union dues, etc. (referred to as “other deductions”). Most pay cheques are received weekly or every two weeks but some respondents only know their salaries/wages before taxes and deductions on a monthly or annual basis. The respondent may choose any reporting period, which makes it easier for him/her to give accurate data.

**Bonuses** - In some situations, wages are paid in the form of both regular pay cheques and periodic bonuses based on work performance. In these cases, the bonus is averaged over the period for which it applies and included with the wages or salary reported.

**Tips and commissions** - Tips, bonuses or commissions are averaged over the period for which they apply and included with the wages or salary reported. This applies to weekly, bi-weekly, semi-monthly, monthly and yearly wages.

### **Government sponsored student loan**

A loan sponsored by the federal government or any provincial/territorial government, which enables the respondent to finance his/her studies.

As of March 2001, Canada Student Loans come directly from the Government of Canada through the National Student Loans Service Centre. The loan is either deposited or mailed to the individual.

From August 1, 1995 up to March 2001, Canada Student Loans were issued by banks, Credit Unions and Caisses Populaires but were guaranteed by the government.

“Student loan” applies to any education, not just the program from which the respondent graduated. It could include undergraduate and graduate programs.

### **Scholarships, awards, fellowships, prizes**

Merit-based (i.e. based on individual achievements) financial assistance to help students continue their studies. These may be awarded by governments or by private donors. Scholarships, awards, fellowships and prizes apply to any education, not just the program from which the respondent graduated. It could include undergraduate and graduate education.

### **Grants, bursaries**

Financial assistance to students which is need-based and/or targeted for specific purposes.

A grant is a gift (usually a sum of money) made by a government or corporation (as an educational or charitable foundation) to a beneficiary on the condition that certain terms be accepted or certain engagements fulfilled which are required by the sponsor.

A bursary refers to a monetary award to assist a student in the pursuit of his/her studies based on financial need and satisfactory achievement.

Grants and bursaries apply to any education, not just the program from which the respondent graduated. It could include undergraduate and graduate education.

### **Income**

The income information is for the income received from all sources by the graduate in the calendar year 2006. It is not limited to monies that are taxable.

It includes:

- income from wages and salaries;
- net income from self-employment;
- regular Employment Insurance benefits as well as those for sickness, maternity or paternity leave, adoption, job creation, work sharing, retraining and benefits to self-employed fisherman;
- retraining and retirement benefits received under the Human Resources and Social Development Canada employment insurance program;

- payments from provincial or municipal programs for persons in need such as Social Assistance or welfare;
- spousal support or child support;
- scholarships, grants, bursaries or fellowships;
- money from the Canada or Quebec Pension Plan;
- Canada Child Tax Benefits or provincial child tax benefits or credits;
- interest from Canadian and foreign sources;
- foreign dividends;
- taxable dividends received from Canadian corporations;
- net rental income;
- rents for leased farm land;
- regular income from an estate or trust fund;
- cash dividends from life insurance policies;
- pensions from deferred profit sharing plans and other private pension plans; and
- money from parents, guardians or others that does not have to be repaid.

It excludes:

- monies received from student loans or any other loan;
- income tax refunds;
- tax-free Registered Retirement Savings Plan withdrawals used for purchasing a home;
- proceeds from the sale of property, businesses, financial assets or personal belongings;
- loans repaid to the graduate as a lender; and
- refund of contributions to work-related pension plans.

## 4.2 Uses

Following from previous surveys, this survey extends the existing base of information on the labour-market experiences of recent graduates. Information derived from the survey has the potential to shed light on many areas of current interest. The following are examples of uses to which the survey's data is applied.

- The survey data can be used to update the occupational supply and demand models and the student flow model. These models project supplies of labour by occupation and industry, especially in highly-skilled and highly-qualified categories.
- Job equity programs will receive important labour market related information on designated groups such as women, aboriginal peoples, persons with disabilities and visible minorities.
- The survey provides concrete information regarding graduates' labour market experiences during the two years after graduation. This information can be used to aid post-secondary education course selection and career counselling.



## 5.0 Survey Methodology

The National Graduates Survey – Class of 2005 (NGS2005) is a longitudinal survey designed to collect data from Canadian graduates. However, for the 2005 cohort of graduates, the follow-up of graduates (FOG) planned for the 2010 was cancelled. Therefore, the 2005 survey is now considered a cross-sectional survey.

### 5.1 Target Population

The target population of the NGS2005 consists of all graduates from a recognized public post-secondary Canadian institution who completed the requirements of an admissible program or obtained a diploma some time in 2005, and who were living in Canada or the United States at the time of the survey (with the exception of American citizens living in the United States at the time of the survey).

These graduates include:

- graduates of university programs that lead to bachelor's, master's or doctoral degrees, or that lead to specialized certificates or diplomas;
- graduates of post-secondary programs (that is, programs that normally require a secondary school completion or its equivalent for admission) in Colleges of Applied Arts and Technology (CAAT), Collèges d'enseignement général et professionnel (CEGEP in Quebec), community colleges, technical schools or similar institutions; and
- graduates of skilled trades (that is, pre-employment programs that are normally three months or more in duration). A trade/vocational school is a public educational institution that offers courses to prepare people for employment in a specific occupation such as heavy equipment operator, automotive mechanic or upholsterer. Many community colleges and technical institutes offer certificates or diplomas at the trade level.

The survey excludes:

- graduates from private post-secondary institutions (for example, computer training and commercial secretarial schools);
- graduates who completed "continuing education" courses at universities and colleges (unless they led to a degree or diploma); and
- graduates in apprenticeship programs.

### 5.2 Survey Frame

The survey frame for the 2005 graduates was created by Statistics Canada's Centre for Education Statistics from a list of all graduates from universities, colleges and trade/vocational schools in Canada.

It should be noted that graduates of an "Attestation of College Studies" in Quebec were included for the first time in the NGS frame. This should be kept in mind when comparing college data with previous cycles, especially at the provincial level for Quebec, since graduates of those programs have characteristics different from graduates of three-year technical programs.

Data on graduates were provided through two sources: the main source of information was from the individual institutions and provincial co-ordinating bodies, while the second source of graduate data came from the Postsecondary Student Information System (PSIS), which is maintained by the Centre for Education Statistics.

Where the PSIS data could not be extracted, files of graduates, preferably in machine-readable form, were requested from the institutions or provincial co-ordinating bodies. The same information that is submitted to the PSIS was requested for each graduate: his/her name, permanent address and telephone number, local address and telephone number, qualification obtained in 2005, major field of study, date of birth, student number, immigration status, gender, mother tongue, graduation date and whether the program taken was a co-op program.

### **5.2.1 CIP coding**

A standard Classification of Instructional Programs (CIP) code was assigned to all graduates on the frame. This coding process is mostly automated as it is already a regular process for PSIS, but some of the cases were coded manually. The CIP code was required to derive the field of study variable used for stratification. It was also used to eliminate from the frame graduates from programs that are not part of the target population.

### **5.2.2 Removal of duplicates**

A verification of duplicates was done on the survey frame. Duplicates consist of two or more records on the frame that refer to the same person and that are classified in the same stratum (see Section 5.3.2 for the stratum definition). When duplicates were found, only one record was kept on the survey frame for that person. Note that when a person graduated in two different programs (programs falling into two different strata), both records of this person were kept on the survey frame. However, if both records were selected in the sample, that person was contacted only once.

## **5.3 Survey Design**

The NGS2005 uses a stratified simple random sample design. The sample selection of graduates within strata is done without replacement and using a systematic method.

### **5.3.1 Longitudinal Sample**

The survey involves a longitudinal design with graduates being interviewed at two different times: at two years and five years after graduating from post-secondary institutions in Canada. The sample design has been developed using a "funnel-shaped" approach, therefore only graduates that respond to the initial interview will be traced for the follow-up interview.

### **5.3.2 Stratification**

Three variables are used for stratification; geographical location of the institution, level of certification and field of study. There are 13 geographical locations: the ten provinces and the three northern territories. There are 5 levels of certification: trade/vocational certificate or diploma, college diploma, bachelor's degree, master's degree, and doctorate. Finally, there are 12 fields of study: categories 01 to 12 of the primary groupings of the Classification of Instructional Programs (CIP). Details about the field of study can be found in Appendix A. The combination of these three variables makes for a possibility of 780 strata in total. However, there are not graduates in every possible strata and therefore, the final number of strata created was 506.

## 5.4 Sample Allocation, Selection and Size

The sample is designed to yield estimates of a minimal proportion of 5.5% with a maximum coefficient of variation (CV) of 16.5% for any of the NGS2005's marginal. A marginal is defined as: i) a given field of study regardless of the province of institution or ii) a given province of institution regardless of the field of study; and that for each of the five levels of certification. The marginal's CVs are then allocated to each stratum (or cell in a table) to obtain the cells or strata's CV using a raking-ratio algorithm. The last step consists of converting the CV's into sample sizes.

Note that the expected non-response and out-of-scope rates were taken into account when establishing the sample sizes.

The table below presents the distribution of the population and the sample size by province/territory and level of certification. The population sizes represent the number of graduates on the final frame.

**Population and Sample Size by Province / Territory and Level of Certification**

<b>Province / Territory by Level of Certification</b>	<b>Population Size</b>	<b>Sample Size</b>
<b>Newfoundland and Labrador</b>	<b>4,206</b>	<b>2,319</b>
Trade/vocational	62	38
College diploma	903	659
Bachelor's degree	2,757	1,186
Master's degree	459	411
Doctorate	25	25
<b>Prince Edward Island</b>	<b>1,554</b>	<b>1,470</b>
Trade/vocational	0	0
College diploma	830	830
Bachelor's degree	691	607
Master's degree	27	27
Doctorate	6	6
<b>Nova Scotia</b>	<b>12,497</b>	<b>4,680</b>
Trade/vocational	0	0
College diploma	3,304	1,618
Bachelor's degree	7,399	1,820
Master's degree	1,711	1,159
Doctorate	83	83
<b>New Brunswick</b>	<b>7,779</b>	<b>3,559</b>
Trade/vocational	3	3
College diploma	2,834	1,441
Bachelor's degree	4,356	1,554
Master's degree	545	520
Doctorate	41	41
<b>Quebec</b>	<b>116,340</b>	<b>12,378</b>
Trade/vocational	35,107	3,566
College diploma	23,017	2,296
Bachelor's degree	44,521	2,794
Master's degree	12,439	2,466
Doctorate	1,256	1,256

<b>Ontario</b>	<b>141,155</b>	<b>10,792</b>
Trade/vocational	3,994	1,766
College diploma	49,705	2,743
Bachelor's degree	68,366	2,453
Master's degree	12,932	2,119
Doctorate	1,711	1,711
<b>Manitoba</b>	<b>10,630</b>	<b>4,185</b>
Trade/vocational	693	597
College diploma	3,292	1,255
Bachelor's degree	5,885	1,656
Master's degree	660	577
Doctorate	100	100
<b>Saskatchewan</b>	<b>9,375</b>	<b>4,645</b>
Trade/vocational	1,214	951
College diploma	3,039	1,330
Bachelor's degree	4,377	1,705
Master's degree	638	552
Doctorate	107	107
<b>Alberta</b>	<b>31,290</b>	<b>7,091</b>
Trade/vocational	991	762
College diploma	9,247	2,284
Bachelor's degree	17,734	2,172
Master's degree	2,846	1,401
Doctorate	472	472
<b>British Columbia</b>	<b>44,664</b>	<b>9,152</b>
Trade/vocational	4,876	2,318
College diploma	14,944	2,441
Bachelor's degree*	19,848	2,194
Master's degree*	4,328	1,785
Doctorate*	464	414
<b>Yukon</b>	<b>138</b>	<b>138</b>
Trade/vocational	49	49
College diploma	57	57
Bachelor's degree	32	32
<b>Northwest Territories</b>	<b>204</b>	<b>198</b>
Trade/vocational	77	71
College diploma	127	127
<b>Nunavut</b>	<b>94</b>	<b>94</b>
Trade/vocational	23	23
College diploma	50	50
Bachelor's degree	21	21
<b>Canada</b>	<b>375,275</b>	<b>60,701</b>
Trade/vocational	47,088	10,144
College diploma	111,350	17,131
Bachelor's degree	175,987	18,194
Master's degree	36,585	11,017
Doctorate	4,265	4,215

\*Note: One university in British Columbia provided the list of graduates for the academic year 2005-2006 instead of the calendar year 2005. As a result, the 2006 graduates were removed from the frame and most 2005 graduates for that institution could not be sampled as they were not included in the frame. However, the number of graduates for 2005 could be estimated for each level of certification and these estimates are included in the British Columbia population sizes.

## 6.0 Data Collection

Project supervisors and Senior interviewers from the Statistics Canada Regional Offices came to head office for a two-day classroom training seminar. Presentations on subject matter and methodology were made, along with mock interviews. Project supervisors and Senior interviewers then conducted a 2-day training of interviewers in the Regional Offices, assisted with an interactive tutorial and mock interviews.

Interviewers collected the data using a computer-assisted telephone interviewing method (CATI). They were instructed to make all reasonable attempts to obtain interviews with the selected graduates. Proxy response was not allowed. For graduates who refused to participate, a letter was sent from the Regional Office to the dwelling address stressing the importance of the survey and the graduate's cooperation. This was followed by a second call from the interviewer. For cases in which the timing of the interviewer's call was inconvenient, an appointment was arranged to call back at a more convenient time. For cases in which there was no one home, numerous call backs were made. If graduates had moved, various tracing methods were used to locate them.

The collection period was scheduled to run from May 8<sup>th</sup> to August 31<sup>st</sup>, 2007. Collection was extended to allow interviewers to contact respondents and collect data up to September 15<sup>th</sup>. After collection, it was discovered that many of the 2005 graduates from the Holland College (the only college in Prince Edward Island) were not included in the survey frame. It was decided to prepare a supplementary sample and return to collection in order to obtain a sufficient number of college graduates for Prince Edward Island. Collection ran from October 24<sup>th</sup> to December 31<sup>st</sup>, 2007 in the Halifax Regional Office.



## 7.0 Data Processing

This chapter presents a brief summary of the processing steps involved in producing the microdata file.

### 7.1 Data Capture

Responses to survey questions are captured directly by the interviewer at the time of the interview using a computerized questionnaire. The computerized questionnaire reduces processing time and costs associated with data entry, transcription errors, and data transmission. The response data are transmitted over a secure line to Ottawa.

Some editing is done directly at the time of the interview. Where the information entered is out of range (too large or small) of expected values, or inconsistent with previous entries, the interviewer is prompted, through message screens on the computer, to modify the information. However, for some questions interviewers have the option of bypassing the edits and of skipping questions if the graduate does not know the answer or refuses to answer. Therefore, the response data are subjected to further edit processes once they arrive in head office.

### 7.2 Editing

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 graduate (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 questions.

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

The third type of editing performed was related to inconsistencies in some of the responses received. In a situation where an inconsistency was found, depending on the nature of the inconsistency, various actions could be taken. The inconsistent variable (or one of the variables involved) could either be changed to “not stated”, corrected or left unchanged. For example, if a respondent had been in a job for two years and said that he had taken nine periods of unpaid leave (an unpaid leave period being defined as at least four consecutive weeks), the number of unpaid leave periods was changed to “not stated”. If a respondent reported an hourly salary of 35,000 dollars, the “hourly” was changed to “annually”. However, in situations where it was not possible to determine which variable was most likely to be wrong, no action was taken.

One of the changes that was made to the NGS2005 questionnaire was to collect only information about the first employer after graduation (EM module), instead of information about all employers (up to six) since graduation other than last week’s employer. Due to the fact that there was no question asking if the first employer after graduation was the same as last week’s employer, respondents who had had at least one other employer since graduation, but whose first employer was the same as last week’s employer, had to repeat the same information that was already provided in the LF module. This situation led to a misunderstanding and many of these respondents provided information about the other employer they had had even though it was not their first employer after graduation. A global correction was made to solve this problem: if the start date with the employer in the EM module was after the start date of last week’s job, the information provided in the LF module was copied in the EM module.

For quantitative variables such as financial variables, editing which includes outlier detection was performed. These variables include reported information on earnings, income, and student loans. Potential outliers were identified and manual investigations were made on these cases to confirm their outlier status. Outliers were changed to “not stated” or replaced by a more plausible value when a realistic value could be deduced from the other variables.

## **7.3 Coding of Open-ended Questions**

A few data items on the questionnaire were recorded by interviewers in an open-ended format. These were items relating to the type of education programs taken before and after graduation in 2005, as well as questions relating to the graduates’ industry and occupation. These open-ended questions were coded using various standard classifications (see Sections 7.3.1 and 7.3.2). An additional type of coding performed is called “Other – Specify” coding (see Section 7.3.3).

### **7.3.1 Coding of Education Programs**

Field of study program descriptions were coded using the Classification of Instructional Programs (CIP 2000). Programs were coded at the six-digit level. See Appendix A for details on the code set.

### **7.3.2 Coding of Industry and Occupation**

For each job held by the graduate in the reference periods, the questionnaire collected information on the name of the employer, the kind of business, industry or service the employer was in, the kind of work done and the usual duties or responsibilities of the graduate in the job. This information was used to assign industry and occupation codes to each job using the North American Industry Classification System (NAICS) 2002 and the National Occupational Classification for Statistics (NOC-S) 2001. See Appendix B and C for details on the code sets. For the user’s convenience, the NAICS and the NOC-S variables have been grouped in their own section in the codebook.

### **7.3.3 Coding of “Other – Specify” Answers**

“Other – Specify” coding was done on questions that contained a list of answer categories that had “Other - Specify” as the final category. If the write-in was reflected in one of the existing categories, the response was recoded into the appropriate one. New categories were added if there were a large number of write-ins which could be categorized together. The latter occurred for questions relating to the reason for looking for another job (LF\_Q60), the reason for not wanting to move to improve job or career prospects (LF\_Q96), the two main sources of funding for postsecondary education and for programs taken after graduation (SL\_Q01 and EP\_Q34), and the reason for not wanting to pursue a master’s degree, for not wanting to pursue a Ph.D. and for not wanting to become a university professor (HE\_Q01B, HE\_Q03B and HE\_Q05 respectively). Responses that could not be coded into an existing category or into new categories were coded as “Other”.

## **7.4 Imputation**

No imputation was done for the National Graduates Survey – Class of 2005.



## 7.5 Creation of Derived Variables

### Combining Items

A number of variables have been derived by combining questions on the questionnaire in order to facilitate data analysis. For example, six questions from the Activities Last Week (LF) section are used to derive labour force status in the week prior to the interview (LFSTAT). These included:

LF\_Q02 - [Last week], were you enrolled full-time or part-time [in any credit courses at an educational or training institution]?

LF\_Q03 - Last week, did you work at a job or a business?

LF\_Q05 - Were you absent from work [last week] because of a temporary layoff?

LF\_Q07 - Last week, did you have a job to start at a definite date in the future?

LF\_Q10 - Last week, were you looking for a job?

LF\_Q11 - [Last week], were you looking for a job at which you would usually work 30 or more hours per week?

### Where to find the Derived Variables on the File

For a list of the derived variables available on the PUMF and a description of how they were derived, see Appendix D.



## 8.0 Response Rates

This chapter describes the response rates for the National Graduates Survey – Class of 2005 (NGS2005). Survey response rates are measures of the effectiveness of the population being sampled and the collection process. They are also a good indicator of the quality of the estimates produced.

In-scope records are records that met all criteria in the target population as defined in Section 5.1. A respondent is a person for whom there is usable minimal information on the questionnaire. Cases where the graduates did not go far enough in the questionnaire or where crucial questions (e.g. diploma or degree obtained, employment status) were not answered, were deemed non-responding units.

Table 8.1 presents the collection results for the NGS2005. The following two types of response rates are presented in that table:

Response Rate – Master File =

$$\frac{\text{Number of responding graduates on Master File}}{\text{Number of in-scope graduates}}$$

Response Rate – Share File =

$$\frac{\text{Number of responding graduates who agreed to share their data}}{\text{Number of in-scope graduates}}$$

**Table 8.1 Response Rate by Province / Territory and Level of Certification – Unweighted**

Province / Territory by Level of Certification	Total Sample Size	In-scope Sample Size	Responding Graduates		Response Rate (%)	
			Master	Share	Master	Share
<b>Newfoundland and Labrador</b>	<b>2,319</b>	<b>2,247</b>	<b>1,577</b>	<b>1,530</b>	<b>70.2</b>	<b>68.1</b>
Trade/vocational	38	38	28	25	73.7	65.8
College diploma	659	646	467	447	72.3	69.2
Bachelor's degree	1,186	1,145	776	761	67.8	66.5
Master's degree	411	398	290	282	72.9	70.9
Doctorate	25	20	16	15	80.0	75.0
<b>Prince Edward Island</b>	<b>1,470</b>	<b>1,425</b>	<b>982</b>	<b>957</b>	<b>68.9</b>	<b>67.2</b>
College diploma	830	812	530	511	65.3	62.9
Bachelor's degree	607	582	429	423	73.7	72.7
Master's degree	27	27	20	20	74.1	74.1
Doctorate	6	4	3	3	75.0	75.0
<b>Nova Scotia</b>	<b>4,680</b>	<b>4,478</b>	<b>2,987</b>	<b>2,911</b>	<b>66.7</b>	<b>65.0</b>
College diploma	1,618	1,589	1,095	1,061	68.9	66.8
Bachelor's degree	1,820	1,719	1,140	1,111	66.3	64.6
Master's degree	1,159	1,102	710	697	64.4	63.2
Doctorate	83	68	42	42	61.8	61.8
<b>New Brunswick</b>	<b>3,559</b>	<b>3,426</b>	<b>2,320</b>	<b>2,251</b>	<b>67.7</b>	<b>65.7</b>
Trade/vocational	3	3	2	2	66.7	66.7
College diploma	1,441	1,424	917	874	64.4	61.4
Bachelor's degree	1,554	1,469	1,041	1,023	70.9	69.6
Master's degree	520	491	335	327	68.2	66.6
Doctorate	41	39	25	25	64.1	64.1

Province / Territory by Level of Certification	Total Sample Size	In-scope Sample Size	Responding Graduates		Response Rate (%)	
			Master	Share	Master	Share
<b>Quebec</b>	<b>12,378</b>	<b>12,001</b>	<b>8,375</b>	<b>8,066</b>	<b>69.8</b>	<b>67.2</b>
Trade/vocational	3,566	3,519	2,391	2,293	67.9	65.2
College diploma	2,296	2,264	1,702	1,647	75.2	72.7
Bachelor's degree	2,794	2,719	1,934	1,867	71.1	68.7
Master's degree	2,466	2,373	1,589	1,526	67.0	64.3
Doctorate	1,256	1,126	759	733	67.4	65.1
<b>Ontario</b>	<b>10,792</b>	<b>10,354</b>	<b>7,155</b>	<b>6,827</b>	<b>69.1</b>	<b>65.9</b>
Trade/vocational	1,766	1,726	1,264	1,185	73.2	68.7
College diploma	2,743	2,692	1,854	1,739	68.9	64.6
Bachelor's degree	2,453	2,357	1,636	1,581	69.4	67.1
Master's degree	2,119	2,015	1,360	1,305	67.5	64.8
Doctorate	1,711	1,564	1,041	1,017	66.6	65.0
<b>Manitoba</b>	<b>4,185</b>	<b>4,088</b>	<b>2,907</b>	<b>2,766</b>	<b>71.1</b>	<b>67.7</b>
Trade/vocational	597	581	382	351	65.7	60.4
College diploma	1,255	1,238	871	828	70.4	66.9
Bachelor's degree	1,656	1,614	1,188	1,143	73.6	70.8
Master's degree	577	558	401	382	71.9	68.5
Doctorate	100	97	65	62	67.0	63.9
<b>Saskatchewan</b>	<b>4,645</b>	<b>4,502</b>	<b>3,034</b>	<b>2,922</b>	<b>67.4</b>	<b>64.9</b>
Trade/vocational	951	887	543	515	61.2	58.1
College diploma	1,330	1,312	923	876	70.4	66.8
Bachelor's degree	1,705	1,667	1,139	1,117	68.3	67.0
Master's degree	552	534	358	348	67.0	65.2
Doctorate	107	102	71	66	69.6	64.7
<b>Alberta</b>	<b>7,091</b>	<b>6,857</b>	<b>4,646</b>	<b>4,503</b>	<b>67.8</b>	<b>65.7</b>
Trade/vocational	762	732	386	359	52.7	49.0
College diploma	2,284	2,207	1,482	1,433	67.1	64.9
Bachelor's degree	2,172	2,133	1,495	1,466	70.1	68.7
Master's degree	1,401	1,351	973	946	72.0	70.0
Doctorate	472	434	310	299	71.4	68.9
<b>British Columbia</b>	<b>9,152</b>	<b>8,432</b>	<b>5,366</b>	<b>5,103</b>	<b>63.6</b>	<b>60.5</b>
Trade/vocational	2,318	2,202	1,333	1,242	60.5	56.4
College diploma	2,441	2,293	1,434	1,346	62.5	58.7
Bachelor's degree	2,194	2,028	1,318	1,273	65.0	62.8
Master's degree	1,785	1,562	1,053	1,025	67.4	65.6
Doctorate	414	347	228	217	65.7	62.5
<b>Yukon</b>	<b>138</b>	<b>135</b>	<b>84</b>	<b>77</b>	<b>62.2</b>	<b>57.0</b>
Trade/vocational	49	48	30	30	62.5	62.5
College diploma	57	56	35	29	62.5	51.8
Bachelor's degree	32	31	19	18	61.3	58.1
<b>Northwest Territories</b>	<b>198</b>	<b>192</b>	<b>110</b>	<b>100</b>	<b>57.3</b>	<b>52.1</b>
Trade/vocational	71	68	36	32	52.9	47.1
College diploma	127	124	74	68	59.7	54.8

Province / Territory by Level of Certification	Total Sample Size	In-scope Sample Size	Responding Graduates		Response Rate (%)	
			Master	Share	Master	Share
<b>Nunavut</b>	<b>94</b>	<b>89</b>	<b>45</b>	<b>37</b>	<b>50.6</b>	<b>41.6</b>
Trade/vocational	23	23	13	11	56.5	47.8
College diploma	50	47	22	17	46.8	36.2
Bachelor's degree	21	19	10	9	52.6	47.4
<b>Canada</b>	<b>60,701</b>	<b>58,226</b>	<b>39,588</b>	<b>38,050</b>	<b>68.0</b>	<b>65.3</b>
Trade/vocational	10,144	9,827	6,408	6,045	65.2	61.5
College diploma	17,131	16,704	11,406	10,876	68.3	65.1
Bachelor's degree	18,194	17,483	12,125	11,792	69.4	67.4
Master's degree	11,017	10,411	7,089	6,858	68.1	65.9
Doctorate	4,215	3,801	2,560	2,479	67.4	65.2

A subsample of the NGS master file, consisting of 16,081 records, was selected for the PUMF. For confidentiality reasons, a provincial breakdown of records on the PUMF cannot be provided.



## 9.0 Treatment of Non-response and Weighting

The National Graduates Survey – Class of 2005 (NGS2005) is a probability survey. As is the case with any probability survey the sample is selected to represent a reference population - the graduate population - at a specific date within the context of the survey as accurately as possible. Each unit in the sample must therefore represent a certain number of units in the population. If the frame used was perfect (covering exactly the population of interest) and all selected units were traced, contacted and completed the survey, then the design weight assigned to each unit would represent accurately and exactly the number of graduates in the target population. In this situation, using this weight would yield unbiased estimates. However, this is not the case when surveys are faced with non-response and imperfect frames. Weight adjustments are traditionally used to compensate for these different issues. Response patterns have to be studied carefully to appropriately correct for non-response.

It was observed that non-response did not occur randomly or uniformly within the population since different response rates were obtained for different sub-populations. For example, the table in Chapter 8.0 shows that graduates with doctorates tend to have lower response rates than graduates with other degrees. The use of appropriate techniques will correct non-response bias that may be introduced.

The chosen technique for the NGS2005 was based on response homogeneous groups (RHG). RHGs were developed with the premise of identifying sample units with similar response probabilities. In other words, it is assumed that graduates pertaining to a given RHG are equally likely to respond to the survey in a similar fashion. Many factors, among them gender and age, are traditionally known to be factors associated with different non-response patterns. Analyses were completed and the RHGs were identified.

As indicated in Section 1, the NGS2005 PUMF represents a sub-sample of the NGS2005 Master File. The subsampling strategy retained was conducted by selecting a subsample directly from the NGS2005 respondents using homogeneity groups based on the master final weight. This led, obviously, to adding a third phase in the weighting process.

The NGS2005 PUMF can then be considered as a three-phase survey. The first phase being the selection of the original sample for NGS2005 and the responding units to NGS2005 being the second phase sample. This approach is based on the underlying assumption that the second phase sample represents a subsample of the first phase sample. One should note that in practice, the second phase is considered a Bernoulli sample with selection probabilities being the response probabilities observed in RHG. Finally, the PUMF units (the subsample) is considered the third phase sample.

### 9.1 Sampling Weight (phase 1)

At the time of selection, an initial design weight was assigned to each graduate as the inverse of its probability of selection. Since the NGS2005 design is stratified with simple random sampling within strata, the probability of selection of the graduate  $i$  in stratum  $h$  is:

$$\pi_{ih}^{phase1} = \frac{n_h}{N_h}$$

where,  $n_h$  and  $N_h$  denote respectively the sample and population size of stratum  $h$ .

Therefore, the first phase weight is:

$$w_{ih}^{phase1} = \frac{1}{\pi_{ih}^{phase1}}$$

## 9.2 Non-response adjustment (phase 2)

After the calculation of the first-phase weight, a non-response adjustment (second phase adjustment) was applied on the sample units. The sample was divided into two groups: resolved units and unresolved units. The group of resolved units contains the survey respondents and the out-of-scope units identified at collection (e.g. graduates living overseas at the time of collection). The group of unresolved units contains the rest of the sample, i.e., the non-respondents. For simplicity, we use the term non-response adjustment but in fact, it is an unresolved adjustment. For the purpose of this adjustment, response homogeneity groups (RHGs) were formed. RHGs are determined through a combination of logistic regressions to predict the probability of being a resolved unit and then using a clustering procedure based on the modelled probability of being a resolved unit. For building the logistic regression model, explanatory variables such as gender, age, country of residence, field of study, level of certification and province of study were used.

For graduate  $i$  in RHG  $g$  the non-response adjustment is:

$$\pi_{ig}^{phase2} = \frac{\sum_i w_{ih}^{phase1} I_{ig}}{\sum_i w_{ih}^{phase1} I_{ig} I_{ir}}$$

where  $I_{ig}$  equals 1 if graduate  $i$  is in RHG  $g$ ; equals 0 otherwise.

$I_{ir}$  equals 1 if graduate  $i$  is resolved and in RHG  $g$ ; equals 0 otherwise.

The master weight consists of multiplying the first-phase weight and the non-response adjustment.

For graduate  $i$  the master weight is:

$$w_i^{phase2} = w_{ih}^{phase1} \times \pi_{ig}^{phase2}$$

Note that, after this step, all resolved units (i.e. respondents and out-of-scope units) received a master weight. However, the NGS2005 Master File contains the respondents only.

## 9.3 Subsampling adjustment for the PUMF (phase 3)

The PUMF subsample was selected from NGS2005 respondents by using homogeneity groups based on the weight on the Master File (weight after phase 2).

Since the selection was done randomly within the homogeneity groups, the selection probability of graduates  $i$  within group  $c$  is:

$$\pi_{ic}^{phase3} = \frac{n_c}{N_c}$$

where  $n_c$  and  $N_c$  represent respectively the subsample size and the population size of group  $c$ .



Therefore, the weight of graduate  $i$  after phase 3 is :

$$w_i^{phase3} = w_i^{phase2} \times \frac{1}{\pi_{ic}^{phase3}}$$

## 9.4 Post-stratification

Following the PUMF subsampling, the sum of the weights (after phase 3) is slightly different in comparison to the sum of the weights of the NGS Master File (weight after phase 2). An adjustment called “post-stratification” is made to the weights to make sure that the sum of the final weights on the PUMF is the same as on the Master File. The post-strata were created by cross-tabulating the certification level (3) and the field of study (10). In total, there were 30 post-strata.

For graduate  $i$  in post-strata  $p$ , the post-stratification adjustment was calculated using the following formula :

$$\pi_{i,p} = \frac{\text{Sum of weights on NGS2005 Master File in post - strata p}}{\text{Sum of weights after phase 3 in post - strata p}}$$

Therefore, the final weight on the PUMF for graduate  $i$  is:

$$w_i^{PUMF} = w_i^{phase3} \times \pi_{i,p}$$



## 10.0 Data Quality

This chapter provides the user with information about the various factors affecting the quality of the survey data. There are two main types of errors: sampling errors and non-sampling errors. A sampling error is the difference between an estimate derived from a sample and the one that would have been obtained from a census that used the same procedures to collect data from every person in the population. All other types of errors such as frame coverage, response, processing and non-response are non-sampling errors. Many of these errors are difficult to identify and quantify. These are discussed in Section 10.2.

### 10.1 Sampling Errors

The estimates derived from the National Graduates Survey – Class of 2005 (NGS2005) are based on a sample of graduates and not from a complete enumeration (census). This difference is the sampling error of the estimates.

The basis for measuring sampling error 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 measure, known as the coefficient of variation (CV) of an estimate, is obtained by expressing the standard error of the estimate as a percentage of the estimate. This measure allows for better quality comparisons between different types of estimates. The smaller the CV, the smaller the sampling variability, meaning smaller CVs are more desirable. The CV depends on the size of the sample on which the estimate is based, the population size and on the distribution of the sample, i.e. the sampling fraction of the units of the domains being estimated. The following diagram presents the characteristics of some CVs and the Statistics Canada guidelines for release.

Note that for the NGS2005, the error due to non-response has been incorporated into the sampling error. As described in Section 10.2 the use of the Generalized Estimation System (GES) takes into account the non-response variability into the estimates variability.

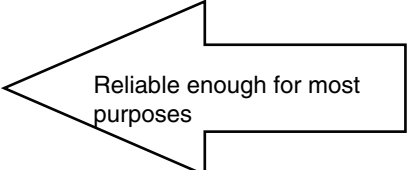
#### Characteristics

0.0% - 1.0%	Excellent
1.0% - 5.0%	Very Good
5.0% - 10.0%	Good
10.0% - 16.5%	Moderate

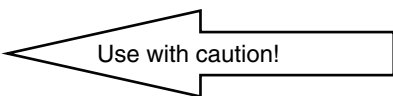
16.6% - 33.3%
---------------

33.4% +
---------

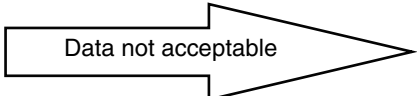
#### Guidelines for Release



Reliable enough for most purposes



Use with caution!



Data not acceptable

## 10.2 Non-sampling Errors

There are many sources of non-sampling errors that are not related to sampling, but may occur at almost any phase of a survey operation. Interviewers may misunderstand survey instructions, graduates may make a mistake in answering the questions, responses may be recorded in the questionnaire incorrectly or errors may be made in the processing or tabulating of the data. For the NGS2005, quality assurance measures were implemented at each phase of the data collection to monitor the quality of the data. These measures included precise interviewer training with respect to the survey procedures and questionnaire, observation of interviews to detect questionnaire design problems or misinterpretation of instructions and coding and edit quality checks to verify the processing logic. Chapter 7.0 outlines data processing procedures. Other kinds of non-sampling error are more easily quantifiable, especially non-response and population frame under-/over-coverage, the topics of the next two sections.

## 10.3 Non-response

Non-response, if not appropriately corrected, is a type of error that can lead to bias in the survey estimates. For the NGS2005, non-response significantly reduced the number of usable records. Biased estimates can occur when unusable units have significantly different characteristics from the usable ones. In Chapter 8.0, non-response rates were computed for basic domains to describe its extent. Extensive studies were completed on non-response to construct the proper adjustment weights for the NGS2005. Since the use of the final weights will yield the appropriate estimates of the population counts and ensure that non-respondents are incorporated and accounted for, it stresses the importance of using the final weights in any tabulations or analysis using the NGS2005 data. Any estimation done without the use of weights may produce biased or incorrect results.

Note that the census of graduates in some strata does not mean that no errors occurred and that the resulting variance will be zero in these strata. As mentioned in the previous section, the variance due to non-response is accounted for in the calculation of the final weight. Consequently, the resulting CVs reflect the global quality of the estimates even for units collected from a census.

## 10.4 Coverage

Coverage is an indication of how a survey frame covers the target population. There could be over-coverage if the survey frame contains units that should not have been included, such as deaths, duplicates, or incorrect date of graduation captured on the file. There could also be under-coverage, if the survey frame missed some units that should have been included.

For the NGS2005, there was some under-coverage for graduates of colleges in some provinces. Data required to build the frame could not be obtained from a few institutions and therefore, graduates from those institutions were not included on the frame. Consequently, they could not be selected nor represented in any tabulation. It is estimated that approximately 10,000 college graduates in Ontario and 5,000 college graduates in Alberta are missing from the NGS2005 population. No adjustment was made at the weighting stage to compensate for this under-coverage.

## 11.0 Guidelines for Tabulation Analysis and Release

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

### 11.1 Rounding Guidelines

In order that estimates for publication or other release derived from the National Graduates Survey – Class of 2005 (NGS2005) microdata file correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates:

- a) Estimates in the main body of a statistical table are to be rounded to the nearest hundred units using the normal rounding technique. In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, in normal rounding to the nearest 100, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last 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 ratio) 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.

### 11.2 Sample Weighting Guidelines for Tabulation

The NGS2005 uses a stratified simple random sample design without replacement of graduates within strata. When producing simple estimates, including the production of ordinary statistical tables, users must use the final weight associated with the graduates concerned by the analysis. If final weights are not used, the estimates derived from the microdata file cannot be considered to be representative of the survey population and will not correspond to those produced by Statistics Canada. The final weight assigned to a given responding graduate reflects the number of graduates in the NGS2005's population he/she represents.

For any analysis dealing with correlation analysis or any other statistics where a significance measure is required, it is recommended that an adjusted weight be used. This weight is obtained by multiplying the final weight by the sample size and dividing this total by the total estimated population. This produces a mean weight of 1 and a sum of weights equal to the sample size.

The benefit of this adjusted weight is that an overestimation of the significance (which is very sensitive to sample size) is avoided while maintaining the same distributions as those obtained when using the demographic weight. The disadvantage is that the numerator is not weighted up to the target population and the coefficient of variance is no longer useful as a measure of data quality.

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.

## 11.3 Definitions of Types of Estimates: Categorical and Quantitative

The NGS2005 PUMF has been set up so that the graduate is the unit of analysis. The final weight that can be found on each record is called WEIGHTPF in the codebook.

### 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 or the proportion of self-employed graduates working at a job last week is an example 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: Last week, did you work at a job or a business?  
R: Yes / No

Q: At your (main) job last week, were you a paid worker or self-employed?  
R: Paid worker / Self-employed / Unpaid family worker

### 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 hours worked per week at a job. The numerator is an estimate of the total number of hours worked per week and its denominator is the number of graduates working.

#### Examples of Quantitative Questions:

Q: How many (paid) hours a week do you usually work at this job?  
R: |\_|\_|\_| hours

Q: How much do you now owe for all your government-sponsored student loans?  
R: |\_|\_|\_|\_|\_| dollars

### 11.3.1 Tabulation of Categorical Estimates

Estimates of the number of graduates 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 ( $\hat{X}$ ),
- b) summing the final weights of records having the characteristic of interest for the denominator ( $\hat{Y}$ ), then
- c) dividing estimate a) by estimate b) ( $\hat{X} / \hat{Y}$ ).

### 11.3.2 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 hours worked by graduates in their main job in the week before they were surveyed multiply the value reported in question LF\_Q79 (hours worked per week) by the final weight for the record, then sum this value over all records with LFSTAT = 1 (employed) and LF\_Q79 < 996.

To obtain a weighted average of the form  $\hat{X} / \hat{Y}$ , the numerator ( $\hat{X}$ ) is calculated as for a quantitative estimate and the denominator ( $\hat{Y}$ ) is calculated as for a categorical estimate. For example, to estimate the average number of hours worked by graduates in their main job in the week before they were surveyed,

- a) estimate the total number of hours ( $\hat{X}$ ) as described above,
- b) estimate the number of graduates ( $\hat{Y}$ ) in this category by summing the final weights of all records with LFSTAT = 1 and LF\_Q79 < 996, then
- c) divide estimate a) by estimate b) ( $\hat{X} / \hat{Y}$ ).

## 11.4 Guidelines for Statistical Analysis

The NGS2005 is based upon a sample design with stratification and different probabilities of selection, depending on the stratum and non-uniform non-response patterns. Using data from such surveys presents problems to analysts because the survey design items mentioned above affect the estimation and variance calculation procedures that should be used. For all types of analysis, final weights are strongly suggested.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures may differ from that which is appropriate in a sample survey framework, with the result that, while in many cases the estimates produced by the packages are correct, the variance estimates that are calculated are poor. Approximate variances for simple estimates such as totals, proportions and ratios (for qualitative variables and for common domains) can be derived using the accompanying Approximate Sampling Variability Tables (see Chapter 12.0). Also, for the NGS2005 PUMF, approximate release cut-offs have been calculated and are presented in Section 11.6.

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.

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.

## 11.5 Release Guidelines

Before releasing and/or publishing any estimate from the NGS2005, users should first determine 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 10.0.

First, the number of graduates (unweighted) who contribute to the calculation of the estimate should be determined. If this number is less than 30, the weighted estimate should be considered of unacceptable quality and more importantly too small for disclosure. Users are invited to read the document Statistics Canada Quality Guidelines available on Statistics Canada web site.

Once this criterion is met, users must determine the coefficient of variation of the estimate and follow the guidelines below. All estimates can be considered releasable. However, those of marginal or unacceptable quality level must be accompanied by a warning to caution subsequent users. These quality level guidelines should be applied to weighted rounded estimates.



### Quality Level Guidelines

Quality Level of Estimate	Guidelines
1) Acceptable	<p>Estimates have:</p> <ul style="list-style-type: none"> <li>a sample size of thirty graduates or more, and</li> <li>low coefficients of variation in the range of 0.0% to 16.5%.</li> </ul> <p>No warning is required.</p>
2) Marginal	<p>Estimates have:</p> <ul style="list-style-type: none"> <li>a sample size of thirty graduates or more, and</li> <li>high coefficients of variation in the range of 16.6% to 33.3%.</li> </ul> <p>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.</p>
3) Unacceptable	<p>Estimates have:</p> <ul style="list-style-type: none"> <li>a sample size of less than thirty graduates, or</li> <li>very high coefficients of variation in excess of 33.3%.</li> </ul> <p>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:</p> <p>“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.”</p>

## 11.6 Release Cut-offs for the PUMF

The following table provides an indication of the precision of population estimates as it shows the release cut-offs associated with a CV of 16.5% and a CV of 33.3% (correspond to quality levels presented in the previous section). These cut-offs are derived from the coefficient of variation (CV) tables discussed in Chapter 12.0.

For example, the table shows that the quality of a weighted estimate of 500 college level graduates possessing a given characteristic 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 12.1 and Example 4 in Section 12.1.1 explain the correct procedure to be used for ratios.

<b>Domain</b>	<b>CV of 16.5% Min X</b>	<b>CV of 33.3% Min X</b>
Canada (all respondents)	1,255	309
College Level (CERTLEVP=1)	1,178	292
Bachelor Level (CERTLEVP=2)	1,441	356
Master/Doctorate Level (CERTLEVP=3)	626	156

## 12.0 Approximate Sampling Variability Tables

In order to supply coefficients of variation (CV) that 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 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 sample design and the adjustment for nonresponse. 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 75<sup>th</sup> percentile) to be used in the CV tables, which would then apply to the entire set of characteristics.

All coefficients of variation in the Approximate Sampling Variability Tables are approximate and therefore unofficial.

**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 releasing such an estimate, regardless of the value of the coefficient of variation.

### 12.1 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 Persons Possessing a Characteristic (Aggregates)**

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

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

The coefficient of variation of an estimated proportion or percentage depends on both the size of the proportion or percentage, and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a sub-group of the population. For example, the proportion of working persons who are self-employed is more reliable than the estimated number of self-employed persons. (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 covered by the table, the CV of the proportion or percentage is the same as the CV of the numerator of the proportion or percentage. In this case, Rule 1 can be used.

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

**Rule 3: Estimates of Differences Between Aggregates or Percentages**

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

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

where  $\hat{X}_1$  is estimate 1,  $\hat{X}_2$  is estimate 2, and  $\alpha_1$  and  $\alpha_2$  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 working persons and the numerator is the number of self-employed persons.

In cases where the numerator is not a subset of the denominator, for example, the ratio of the number of self-employed males as compared to the number of self-employed females, 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 ( $\hat{R} = \hat{X}_1 / \hat{X}_2$ ) is:

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

where  $\alpha_1$  and  $\alpha_2$  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.

### 12.1.1 Examples of Using the Coefficient of Variation Tables for Categorical Estimates

The following examples based on the NGS2005 PUMF are included to assist users in applying the above rules.

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

Suppose that a user estimates that 27,836 graduates had difficulties repaying their student loans. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the coefficient of variation table for Canada.
- 2) The estimated aggregate (27,836) 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 30,000.
- 3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, in this case 3.2%.
- 4) So the approximate coefficient of variation of the estimate is 3.2%. The finding that there were 27,836 graduates (to be rounded according to the rounding guidelines in Section 11.1) who had difficulties repaying their student loans is publishable with no qualifications.

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

Suppose that the user estimates that  $10,615 / 27,836 = 38.1\%$  of graduates who had difficulties repaying their student loans are married or in common-law relationships. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the coefficient of variation table for Canada.

Because the estimate is a percentage based on a subset of the total population (i.e., graduates who had difficulties repaying their student loans), it is necessary to use both the percentage (38.1%) and the numerator portion of the percentage (10,615) in determining the coefficient of variation.

- 2) The numerator, 10,615, 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 10,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the percentage closest to it, 40.0%.
- 3) The figure at the intersection of the row and column, 4.5%, is the coefficient of variation to be used.
- 4) So the approximate coefficient of variation of the estimate is 4.5%. The finding that 38.1% of graduates who had difficulties repaying their student loans are married or in common-law relationships can be published with no qualifications.

#### **Example 3: Estimates of Differences Between Aggregates or Percentages**

Suppose that a user estimates that  $3,681.4 / 10,141.9 = 36.3\%$  of male graduates who had difficulties repaying their student loans are married or in common-law relationships,

while  $6,933.4 / 17,693.7 = 39.2\%$  of female graduates who had difficulties repaying their student loans are married or common-law. 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 men as 7.5%, and the CV of the estimate for women as 5.4%.
- 2) Using Rule 3, the standard error of a difference ( $\hat{d} = \hat{X}_1 - \hat{X}_2$ ) is:

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

where  $\hat{X}_1$  is estimate 1 (women),  $\hat{X}_2$  is estimate 2 (men), and  $\alpha_1$  and  $\alpha_2$  are the coefficients of variation of  $\hat{X}_1$  and  $\hat{X}_2$  respectively.

That is, the standard error of the difference  $\hat{d} = 0.392 - 0.363 = 0.029$  is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(0.392)(0.054)]^2 + [(0.363)(0.075)]^2} \\ &= \sqrt{(0.000448) + (0.000741)} \\ &= 0.034\end{aligned}$$

- 3) The coefficient of variation of  $\hat{d}$  is given by  $\sigma_{\hat{d}} / \hat{d} = 0.034 / 0.029 = 1.194$ .
- 4) So the approximate coefficient of variation of the difference between the estimates is 119%. The difference between the estimates 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 30,803 males supervised other employees at their main job last week, while 38,524 females supervised other employees at their main job last week. The user is interested in comparing the estimate of men versus women 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 male graduates who supervised other employees at their main job last week. The denominator of the estimate ( $\hat{X}_2$ ) is the number of female graduates who supervised other employees at their main job last week.
- 2) Refer to the coefficient of variation table for Canada.
- 3) The numerator of this ratio estimate is 30,803. The figure closest to it is 30,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely 3.2%.

- 4) The denominator of this ratio estimate is 38,524. The figure closest to it is 40,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, 2.7%.
- 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  $\alpha_1$  and  $\alpha_2$  are the coefficients of variation of  $\hat{X}_1$  and  $\hat{X}_2$  respectively.  
That is:

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(0.032)^2 + (0.027)^2} \\ &= \sqrt{0.0010 + 0.0007} \\ &= 0.042\end{aligned}$$

- 6) The obtained ratio of male graduates versus female graduates who supervised other employees at their main job last week is 30,803 / 38,524, which is 0.80 (to be rounded according to the rounding guidelines in Section 11.1). The coefficient of variation of this estimate is 4.2%, which makes the estimate releasable with no qualifications.

#### Example 5: Estimates of Differences of Ratios

Suppose that the user estimates that the ratio of male to female graduates who supervised other employees at their main job last week, is 0.73 at the Bachelor certification level (CERTLEVP = 2) and 1.01 at the Master/Doctorate level (CERTLEVP = 3). 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 Bachelor ratio ( $\hat{R}_1$ ) and the Master/Doctorate ratio ( $\hat{R}_2$ ) as in Example 4. The approximate CV for the Bachelor ratio is 6.1%, and 10.6% for the Master/Doctorate ratio.
- 2) Using Rule 3, the standard error of a difference ( $\hat{d} = \hat{R}_1 - \hat{R}_2$ ) is:

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

where  $\alpha_1$  and  $\alpha_2$  are the coefficients of variation of  $\hat{R}_1$  and  $\hat{R}_2$  respectively. That is, the standard error of the difference  $\hat{d} = 1.01 - 0.73 = 0.27$  is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(1.01)(0.106)]^2 + [(0.73)(0.061)]^2} \\ &= \sqrt{(0.011382) + (0.002008)} \\ &= 0.116\end{aligned}$$

- 3) The coefficient of variation of  $\hat{d}$  is given by  $\sigma_{\hat{d}} / \hat{d} = 0.116 / 0.27 = 0.425$ .

- 4) So the approximate coefficient of variation of the difference between the estimates is 42.5%. The difference between the estimates 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.

## 12.2 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 difference would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

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

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

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

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.

### 12.2.1 Example of Using the Coefficient of Variation Tables to Obtain Confidence Limits

A 95% confidence interval for the estimated proportion of graduates who are married or in common-law relationships among those who had difficulties repaying their student loans (from Example 2, Section 12.1.1) would be calculated as follows:

$$\hat{X} = 38.1\% \text{ (or expressed as a proportion 0.381)}$$

$$t = 2$$

$\alpha_{\hat{x}} = 4.5\%$  (0.045 expressed as a proportion) is the coefficient of variation of this estimate as determined from the tables.

$$CI_{\hat{x}} = \{0.381 - (2) (0.381) (0.045), 0.381 + (2) (0.381) (0.045)\}$$

$$CI_{\hat{x}} = \{0.381 - 0.034, 0.381 + 0.034\}$$

$$CI_{\hat{x}} = \{0.347, 0.416\}$$

With 95% confidence, it can be said that between 34.7% and 41.6% of graduates who have difficulties repaying their student loans are married or in common-law relationships.

### 12.3 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  $t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$  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. In other words, the difference between the estimates is significant.

### 12.3.1 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 male and female graduates who are married or in common-law relationships among those who had difficulties repaying their student loans. From Example 3, Section 12.1.1, the standard error of the difference between these two estimates was found to be 0.034. Hence,  $0.392 - 0.363 = 0.029$ .

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_d} = \frac{0.392 - 0.363}{0.034} = \frac{0.029}{0.034} = 0.837$$

Since  $t = 0.837$  is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance.

## 12.4 Coefficients of Variation for Quantitative Estimates

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

## 12.5 Coefficient of Variation Tables

Approximate Sampling Variability Tables are available in Appendix E.

## 13.0 Questionnaire, Code Sheets and Documentation of Derived Variables

Please refer to the files listed below for the National Graduates Survey – Class of 2005 (NGS2005).

### **Questionnaire:**

NGS2005\_QuestE.doc

NGS2005\_QuestE.pdf

### **Code Sheets:**

#### **Classification of Instructional Programs (CIP)**

Appendix A - CIP Aggregate\_PUMF.doc

Appendix A - CIP Aggregate\_PUMF.pdf

#### **North American Industry Classification System (NAICS) 2002**

Appendix B - NAICS\_PUMF.doc

Appendix B - NAICS\_PUMF.pdf

#### **National Occupational Classification for Statistics (NOC-S) 2001**

Appendix C - NOC-S\_PUMF.doc

Appendix C - NOC-S\_PUMF.pdf

#### **Documentation of derived variables**

Appendix D - Documentation of Derived Variables\_PUMF.doc

Appendix D - Documentation of Derived Variables\_PUMF.pdf

#### **Approximate Sampling Variability Tables by various domains**

Appendix E - CV\_Tables\_PUMF.doc

Appendix E - CV\_Tables\_PUMF.pdf



## **14.0 Record Layout with Univariate Frequencies**

See NGS2005\_PUMF\_CdBkE.doc or NGS2005\_PUMF\_CdBk\_E.pdf for the record layout with univariate counts for the public use microdata file.