



## **Microdata User Guide**

# **Survey of Earned Doctorates 2003-2004**



Statistics  
Canada

Statistique  
Canada

Canada



## Table of Contents

<b>1. Introduction.....</b>	<b>5</b>
<b>2. Background.....</b>	<b>6</b>
<b>3. Objectives .....</b>	<b>7</b>
<b>4. Concepts and Definitions .....</b>	<b>8</b>
4.1 Content .....	8
4.2 Concepts and Definitions.....	9
<b>5. Survey Methodology .....</b>	<b>10</b>
5.1 Population Coverage .....	10
5.2 Sample Design .....	10
<b>6. Data Collection .....</b>	<b>11</b>
6.1 Distribution and Collection of Survey Materials and Lists .....	11
6.2 Supervision and Control .....	12
6.3 Non-response .....	13
<b>7. Data Processing .....</b>	<b>14</b>
7.1 Data Capture .....	14
7.2 Editing.....	14
7.3 Coding of Open-ended Questions .....	15
7.3.1 Coding of SED files of study.....	15
7.3.2 Coding of Industry and Occupation.....	15
7.3.3 Coding of “Other – Specify” answers .....	15
7.4 Creation of Derived Variables.....	16
7.5 Weighting.....	16
<b>8 Data Quality.....</b>	<b>17</b>
8.1 Response Rates .....	17
8.2 Survey Errors.....	17
8.2.1 Survey Frame.....	17
8.2.2 Data Collection .....	17
8.2.3 Data Processing .....	18
8.2.4 Non-response.....	18
<b>9 Guidelines for Tabulation, Analysis, and Release.....</b>	<b>20</b>
9.1 Rounding Guidelines .....	20
9.2 Weighting Guidelines for Tabulation.....	21
9.3 Definitions of Types of Estimates: Categorical and Quantitative.....	21
9.3.1 Categorical Estimates .....	21
9.3.2 Quantitative Estimates .....	21
9.3.3 Tabulation of Categorical Estimates.....	22

9.3.4	Tabulation of Quantitative Estimates .....	22
9.4	Guidelines for Statistical Analysis.....	23
9.5	Coefficient of Variation Release Guidelines .....	23
9.6	Release Cut-off's for the 2003-04 Survey of Earned Doctorates .....	25
<b>10.</b>	<b>Approximate Variability Table .....</b>	<b>26</b>
10.1	How to Use the Coefficient of Variation Table for Categorical Estimates .....	26
10.1.1	Examples of Using the Coefficient of Variation Table for Categorical Estimates .....	28
10.2	How to Use the Coefficient of Variation Table to Obtain Confidence Limits .....	32
10.2.1	Example of Using the Coefficient of Variation Table to Obtain Confidence Limits .....	33
10.3	How to Use the Coefficient of Variation Table to Do a T-test.....	33
10.3.1	Example of Using the Coefficient of Variation Table to Do a T-test.....	34
10.4	Coefficients of Variation for Quantitative Estimates .....	34
10.5	Table of Approximate Variability.....	35
<b>11.</b>	<b>Questionnaire and Code Sheets.....</b>	<b>36</b>
<b>12.</b>	<b>Record Layout with Univariate Frequencies.....</b>	<b>37</b>

## **1. Introduction**

The 2003-2004 Survey of Earned Doctorates (SED) was conducted by Statistics Canada from September 2003 to September 2004 with the cooperation of all Canadian doctoral-degree granting institutions. This manual has been produced to facilitate the manipulation of the microdata file of the survey results.

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

Statistics Canada  
Client Services  
Centre for Education Statistics  
Room SC-2200 Main Building  
Tunney's Pasture  
Ottawa, Ontario            K1A 0T6

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

Fax: (613) 951-9040

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

## **2. Background**

The Survey of Earned Doctorates (SED) was developed to obtain integrated information on the characteristics and plans of doctoral graduates. SED provides a vehicle for future research and analysis of the relationship between the educational paths that doctoral recipients have taken and their plans for future work or study.

The survey provides data about the graduates' postsecondary academic path, the characteristics of their programs of study, funding sources for their graduate studies, financial debt load and labour market and mobility plans following graduation.

### **3. Objectives**

The Survey of Earned Doctorates (SED) is an annual census of doctorate recipients in Canada that was conducted for the first time on a national basis during the 2003-2004 academic year. The basic purpose of this survey is to gather data about all doctoral graduates in Canada to inform government, associations, universities and other stakeholders on the characteristics and plans of these very highly qualified graduates as they leave their doctoral programs.

These data are important in improving graduate education by providing governmental and private agencies with the information necessary to make program and policy decisions. Data about an institution's own doctorate recipients are also provided to, and used by, research offices of institutions who participate in the survey.

The survey's key data objectives are:

- To evaluate the impact of the various sources of institutional funding;
- To gather information on the retention of doctoral students in Canada;
- To gain a better understanding of postgraduate education financing and debt level;
- To allow labour market planners to assess the additions to the domestic stock of highly qualified human resources in various fields;
- To allow an examination of the path to receipt of doctoral degrees and the impact of foreign students.

## 4. Concepts and Definitions

### 4.1 Content

This section outlines the content areas of the questions for the Survey of Earned Doctorates.

<b>Section</b>	<b>Content</b>
PART A – EDUCATION	<p>Respondents are asked to provide the title their dissertation, the field of study and field code.</p> <p>Asks about the sources of financial support during graduate school and financial debt load.</p> <p>Obtains detailed information about the degrees received and any other postsecondary diplomas or certificates attained.</p>
PART B – POSTGRADUATION PLANS	<p>Provides information on respondent's postgraduate plans and intended country of residence and how definite these are.</p> <p>Asks about the respondent's main activity after graduation.</p> <p>Obtains information about the respondent's postgraduate labour force activity such as occupation and sector of industry.</p> <p>Asks about the respondents expected or actual salary after graduation or source of financial support for postdoctoral study or research.</p>
PART C – BACKGROUND INFORMATION	<p>Collects general information such as marital status, number of dependent children, parents' education, citizenship and ethno-cultural origin.</p>
CONTACT INFORMATION	<p>Obtains the phone number and e-mail address of the respondent in addition to the name and address of a contact person for follow-up purposes.</p>
DATA SHARING AGREEMENT	<p>Asks the respondent to share their answers with their doctoral institution.</p>



## **4.2 Concepts and Definitions**

### **Research-oriented versus non research-oriented doctoral degree**

SED focuses on all doctoral recipients, whether the degree is research-oriented or not. Typically, a degree is research-oriented if it requires the completion of a dissertation or thesis or equivalent project of original work (e.g., musical composition) and is not exclusively intended as a degree for the practice of a profession. In Canada, all doctoral degrees are included; this means that professional degrees such as MD, DDS, and other doctorates such as Doctor of Theology (ThD), Doctor of Music (Mus. Doc.), etc. are also included. Question A1b has been included to allow comparisons with the Survey of Earned Doctorates conducted in the United States, which only collects information on *research* doctorate.

### **Total time to complete the doctoral degree**

The amount of time to complete the doctoral degree represents the total time elapsed between the date when the respondent started his/her doctoral degree and the date when the degree was awarded as provided in Question A8, whether or not the respondent interrupted his/her studies during that period. The variable DV\_TIMPH in the codebook contains the total time to complete the doctoral degree.

### **Sources of financial support**

Two questions on the questionnaire provide information on the respondent's sources of financial support during graduate school. The first question (A5) asks the respondent to indicate ALL his/her sources of financial support amongst a checklist of 28 items. The second question (A6) asks the respondent to indicate which of the checked items in A5 was the primary source (most important) and which was the secondary source (second most important)

### **Definite versus indefinite postgraduation plans**

The questionnaire included a number of questions on the graduates' postgraduation plans (Section B). Whether the graduate's plans were definite or not was determined by the answer provided in Question B3. Plans are definite if respondent has answered (1) "Am returning to, or continuing in, same employment/position as prior to PhD completion"; or (2) "Have signed contract or made definite commitment for other work or study. Variable DV\_DEFPL (postgraduation plans are definite) has been derived from Question B3.

## **5. Survey Methodology**

The 2004 Survey of Earned Doctorates (SED) was designed to collect data from all individuals graduating with a doctorate degree from July 1, 2003 to June 30, 2004.

### **5.1 Population Coverage**

The population includes all doctoral graduates from all public Canadian postsecondary education institutions who obtained their degree during the reference period of July 1, 2003 to June 30, 2004. Specifically excluded are graduates from institutions that did not participate in the survey during the reference period.

### **5.2 Sample Design**

SED was designed as a census; thus, questionnaires were distributed to all eligible respondents at participating institutions.

## **6. Data Collection**

Data collection for SED was conducted from September 2003 to September 2004. Some portions of the collection process were performed by participating institutions; the remainder were performed by Statistics Canada.

### **6.1 Distribution and Collection of Survey Materials and Lists**

There were three phases to the distribution and collection of survey materials (a covering letter, a paper questionnaire, and a pre-paid return envelope) and associated lists of graduates.

#### **Summer Phase**

Students who graduated between July 2003 and September 2003 were mailed the package of survey materials in September 2003 by their respective institutions. Some institutions directed students to send the completed questionnaire to Statistics Canada via the return envelope; some other institutions preferred that students seal the completed questionnaire in the return envelope and return it to the institution; the remaining institutions offered both options.

Institutions also provided a list of all graduates receiving questionnaires to Statistics Canada. The list was used at Statistics Canada for telephone follow-up of tardy respondents. The following information was supplied on the lists, where available:

- Name
- Current address
- Home and work telephone
- Gender
- Language of choice for communication
- Date the questionnaire was sent.

Lists were transmitted to Statistics Canada via a number of secure methods.

## October 2003 to June 2004 Phase

Data collection procedures for this phase were identical to the summer phase, except that institutions were asked to:

- distribute the survey materials to students at the time of application for graduation or when the doctorate was awarded; and
- send a list of graduates who received survey materials to Statistics Canada monthly.

## End of Survey Cycle Phase

At the end of the survey cycle, each institution was asked to provide a complete and final list of all graduates for the reference period to Statistics Canada. The final lists from all participating institutions were used to establish the frame for the survey. Depending on an institution's procedures, this final list may not reflect all the students listed on an institution's prior (monthly) lists. For example, in some institutions the doctorate students received the SED questionnaire at the time of their thesis defence. At that time, the institution provided a monthly list of those who had received the questionnaire. However, some of those students did not deposit their final thesis and graduate until the next survey cycle year. These respondents were deemed to be out-of-scope for 2003-2004. In addition, questionnaires that were received during the collection period, but which did not correspond to a graduate on the final list were also deemed to be out-of-scope.

The final list of graduates did not include the full contact information supplied in the monthly lists, but was specified to include other information (fields marked with an asterisk were mandatory):

- \*full graduate name
- \*program name
- major field of study
- \*graduation date
- \*gender
- \*birth date
- immigration status
- citizenship
- country of birth
- visa/foreign
- aboriginal/visible minority
- mother tongue

For the 2003/2004 survey, the majority of institutions did not provide the non-mandatory fields.

## 6.2 Supervision and Control

Questionnaires were distributed to the graduates by the participating institutions. Institutions received a manual outlining the SED concepts and collection procedures. Once a questionnaire was completed by a student and sealed in the STC envelope provided as part of the survey materials, it could only be opened by Statistics Canada, regardless of whether the completed questionnaires were collected by the institution or were sent directly by the respondent to Statistics Canada. Follow-up interviews were done by Statistics Canada clerical staff familiar with the SED concepts and procedures. Clerical staff were supervised by senior operational staff, who periodically monitored the follow-up interviews and reviewed completed documents. The senior operational staff in turn worked with subject-matter staff at the Centre for Education Statistics to resolve any remaining issues and to monitor the collection process.

### **6.3 Non-response**

Operational personnel were instructed to make all reasonable attempts to obtain completed questionnaires from graduates. Follow-ups were discontinued after five attempts were made without success, unless prior arrangements were already made which were then respected. For cases where the timing of calls was inconvenient, an appointment was scheduled with the respondent to call back at a more convenient time.

#### Failed edits

There were a few critical items on the questionnaire that were especially important for research and analysis purposes. If the respondent left any of these fields blank, operational personnel contacted the respondent to try to obtain the missing information.

Where no response was received due to incorrect or out-of-date contact information, attempts were made to trace the respondent both in and outside Canada. Efforts were made to contact traced respondents located in time-zones that permitted calls within the working hours of the operational staff.

## **7. Data Processing**

The main output of the Survey of Earned Doctorates is a microdata file of the survey responses with derived variables and survey weights. An additional file of population totals is also produced to allow user-supplied software to calculate the estimates of error due to non-response.

### **7.1 Data Capture**

Capture of survey data was accomplished using computers located in Statistics Canada's head office. During this process answers from all completed questionnaire were keyed into a database. Value/range edits were built into the program to prevent errors and promote data integrity.

Manual coding and grooming — The first step was to prepare the completed questionnaires for data capture. Each questionnaire was carefully reviewed for accuracy and completeness of responses. Where needed, codes were assigned and noted on the questionnaire. Responses had to be legible for keying, and were made clearer when necessary.

Direct Data Entry — Data capture was performed by traditional keying on microcomputers. Verification of the accuracy of the keyed data was done by having a different operator capture non-text fields of all the questionnaires a second time. Quality control was achieved by comparing the two sets of captured information against the paper questionnaire.

### **7.2 Editing**

The data editing phase of processing involves the identification of erroneous or inconsistent values in the survey data, and the modification of such conditions.

The first type of error treated were errors in questionnaire flow. Conflicting questionnaire information would sometimes indicate that a respondent had answered questions in a section that in fact did not apply to him. In these cases, the superfluous data was eliminated. An example of this type of error would be when a respondent indicated in Question B3 that he did not plan to work or study within the next year but then answered Question B4 which should have been skipped.

The second type of errors treated involved editing the survey records according to pre-specified edit rules to check for logical inconsistencies in the survey data. In these cases a strategy was established to eliminate the minimum amount of data, establishing priorities so as to eliminate the values that were most likely to be in error. An example of this would be when a respondent indicated that he was Canadian by birth in Question C8, while also indicating that he was a landed immigrant in Question C9 and providing the year that he became a landed immigrant in question C10. In this case, the responses to Questions C9 and C10 were considered more likely to be correct and therefore Question C8 was recoded.

The third type of error dealt with assigning a code of 'not stated' whenever the respondents did not provide any information to questions that should have been answered.

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

### **7.3.1 Coding of SED fields of study**

All fields of study provided on the questionnaire (Questions A2 and A8) were coded using the Classification of Instructional Programs (CIP – 2000) in addition to the SED fields of study codes as provided on page 8 of the questionnaire. Most fields of study 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 respondents who indicated that they would be employed upon graduation, the questionnaire collected information on the kind of business, industry or service the employer was in, the kind of work done and the main activities of the respondent in the job (Questions B5 to B7). This information was used to assign industry and occupation codes using the North American Industry Classification System (NAICS) 2002 and the National Occupational Classification for Statistics (NOC-S) 2001. See Appendices B and C for details on the code sets.

### **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. This type of write-in occurs when a question has a list of possible response categories, as well as the option of stating another response not already listed. In this situation the text was captured and then manually reviewed. If the write-in was reflected in one of the existing categories, the response was recoded into the appropriate category. Responses that could not be coded into an existing category were left as a true “Other -Specify”.

## 7.4 Creation of Derived Variables

A number of data items on the microdata file have been derived from information collected on the questionnaire. In some cases, the derived variables are straightforward and involve collapsing of categories. In other cases, two or more variables have been combined to create a new variable. The following is a list of the derived variables for the SED

Variable	Description
DV_AGEPH	Age at graduation (doctorate)
DV_MARST	Recoded marital status
DV_CITIZ	Recoded citizenship
DV_TIMPH	Total time to complete doctoral degree (doctorate only)
DV_ACTLM	Any activity limitation
DV_DEFPL	Postgraduate plans are definite
DV_EDUMO	Recoded educational attainment of mother
DV_EDUFA	Recoded educational attainment of father
DV_DIP1	First additional postsecondary degree, diploma or certificate
DV_DIP2	Second additional postsecondary degree, diploma or certificate
DV_DIP3	Third additional postsecondary degree, diploma or certificate
DV_DIP4	Fourth additional postsecondary degree, diploma or certificate

## 7.5 Weighting

Although the Survey of Earned Doctorates was a census, weights were calculated to adjust for non-responding graduates from participating institutions. Thus, weighted estimates from the SED represent the entire population – both respondents and non-respondents. The weights were created by the following process:

1. During the collection period, response rates for population subgroups were continually monitored to identify any emerging patterns.
2. Universities were strongly encouraged to provide additional information on characteristics of graduates at the end of the survey cycle to create the final frame (see 6.1, above). This additional information was available for every graduate, regardless of response status.
3. The additional information from the universities was matched against respondents (and thus, non-respondents) to identify factors affecting response. For example, females may be more likely to respond than males, or those graduates remaining in Canada may be more likely to respond than graduates who leave Canada after graduation.
4. Using the additional factors, graduates were placed into groups of similar graduates (both responding and non-responding). These groups were also called weighting classes. Each graduate was placed into only one weighting class.
5. A weight (DWEIGHT) was derived for each respondent. This weight is equal to the total number of graduates in a particular weighting class, divided by the total number of responding graduates in that weighting class. All respondents in the same weighting class received the same weight.

The effectiveness of the weighting procedure is highly dependent on the availability of additional characteristics of graduates from the universities. See Section 8.2.4 for further information.



## **8 Data Quality**

### **8.1 Response Rates**

In total, 3327 graduates were asked to complete a questionnaire. Fifty-five students refused to complete the questionnaire, questionnaires from 13 respondents had too few questions answered to be usable, and an additional 1278 could not be contacted, yielding a response rate of  $1981 / 3327 = 60\%$ .

### **8.2 Survey Errors**

Errors may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered on the questionnaire and errors may be introduced in the processing and tabulation of the data. These are all examples of non-sampling errors.

Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates. Considerable time and effort were devoted to try to reduce the non-sampling errors in the survey. Quality assurance measures were implemented at each step of the data collection and processing cycle to monitor the quality of the data. These measures include the use of highly-skilled personnel, extensive training with respect to the survey procedures and questionnaire, observation of personnel to detect problems of questionnaire design or misunderstanding of instructions, procedures to ensure that data capture errors were minimized and coding and edit quality checks to verify the processing logic.

#### **8.2.1 Survey Frame**

The frame for the SED target population was created starting from a list of Canadian postsecondary institutions granting doctoral degrees. This list was compiled and is kept up-to-date by the Centre for Education Statistics of Statistics Canada. Every listed institution was invited to participate in this survey. Institutions with no doctoral graduates for the survey reference year were excluded from the target population. Institutions were also asked to supply monthly lists of new graduates and a final annual list of graduates to Statistics Canada. The final lists were used as the survey frame.

While SED is intended to cover all institutions offering doctoral degrees, nine institutions either: could not be contacted; were contacted but did not participate; or were contacted and agreed to participate but encountered handling problems while distributing questionnaires. It is estimated that around 9% of all graduates, or approximately 300 doctoral students, graduated from these institutions. No adjustment was made for these graduates, the majority of which came from one institution. Although most cross-sectional characteristics are not expected to be affected, directly comparing published levels across years will not be possible.

#### **8.2.2 Data Collection**

The SED Institutional Contact Manual and associated materials and software applications were provided to all participating institutions. Support was provided to institutions via e-mail and a toll-free telephone line. Contact information was also provided on the questionnaire for respondents to communicate directly with Statistics Canada if questions arose. Statistics Canada operational staff assigned to the survey received training, including a description of the background and objectives of the survey, as well as a glossary of terms and definitions.

The collection period ran from September 2003 to September 2004. Data collection practices varied by institution, as detailed in Section 6.1. The effect of the difference in collection procedures is unknown; however, it is expected that the timing of distribution of survey materials by institutions affected response rates. For example, two students could complete the work required for the graduate degree at the same time, but one student's institution could distribute the survey materials when the student's thesis was defended, while the other student's institution distributed the survey materials when the institution's governing body confirmed the degree. In this scenario, the latter student would be mailed the questionnaire many months after the other student received the questionnaire. Given the mobility of graduating students, the proportion of students who could not be followed-up would be greater for the latter institution than the first institution.

### **8.2.3 Data Processing**

The Survey of Earned Doctorates is unique in that the files that make up the survey's frame – the final lists provided by institutions – are received at the end of the collection period, not prior to collection as is customary. Thus, questionnaires can be completed by persons not on the final lists, and the monthly lists may not correspond exactly to the final lists.

The ongoing nature of the graduation of doctorates means that some respondents who graduated early in the reference period may receive the previous year's questionnaire, and respondents who graduated late in the reference period may receive the questionnaire for the following survey cycle. These cases were resolved manually.

### **8.2.4 Non-response**

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial, or item, non-response (failure to answer just one or some questions) to total non-response. Total, or unit, non-response occurs because the respondent could not be contacted, the respondent refused to participate in the survey, or the questionnaire was insufficiently completed. For the 2003-04 SED, 13 records with partial non-response were coded as non-response because they had insufficient data.

#### **Total Non-response**

Total non-response was handled by assigning a weight to respondents. (See Section 7.5 for additional details.) The total non-response rate of 40% for the SED decreases the reliability of estimates based on the survey data. The existence of non-response in a census survey creates variance and potential bias in the estimated characteristics. The degree to which an estimate of a characteristic is affected depends on how similar SED respondents and non-respondents are with respect to this characteristic, and the extent to which dissimilarities are accounted for by the weights. The estimation methodology used in SED assumes that all persons within a weighting class – both respondents and non-respondents – have the same propensity to respond and that this propensity is independent of the characteristics measured by the survey. The validity of these assumptions determines the quality of the survey estimates and may vary from one characteristic to another.

For the 2003-04 SED, the auxiliary information on the frame that could be used to create weighting classes was very limited. Thus, it was not possible to construct weighting classes to adjust for all of the expected sources of non-response bias. In particular, estimates of error do not account for the potential bias introduced by the lower proportion of responding graduates among those who had moved outside of Canada. Data users

are advised to apply caution in extrapolating results from the 2003/2004 SED to the population of graduates who moved out of Canada immediately after graduation.

**Partial Non-response**

In most cases, partial non-response to the survey occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information. During data processing, more unknown answers were generated due to data inconsistencies or, more often, due to a path of the questionnaire that was skipped during collection. After data processing, unknown answers were coded as “Not stated”. No imputation was performed.

A number of questions were identified as critical for analysis and were used in assigning a “partial or complete” status to the questionnaire. The following table presents the rate of non-response (not stated) for these items.

**Non-response rate for critical questions**

<b>Question</b>	<b>Description</b>	<b>Rate (%)</b>
<b>A6</b>	Primary source of financial support during graduate school	2.1
<b>A8 Doctorate degree</b>	Month graduate started doctorate degree	0.5
	Year graduate started doctorate degree	0.6
	Month graduate received doctorate degree	0.4
	Year graduate received doctorate degree	0.4
	Field of study for doctorate degree	0.1
<b>A15</b>	Owed money directly related to undergraduate education	1.3
	Owed money directly related to graduate education	1.0
<b>B1</b>	Country intend to live after graduation	1.6
<b>B3</b>	Status of postgraduate plans (in the next year)	1.5
<b>B4</b>	Description of postgraduate plans	2.4
<b>C1</b>	Sex of respondent	0.3
<b>C7</b>	Year of birth	0.7
	Month of birth	0.9

The other variables on the microdata file exhibit similar rates, with exception of question B14. The rate for this variable is almost 10%. Careful analysis of this variable is recommended.

## **9 Guidelines for Tabulation, Analysis, and Release**

This chapter of the documentation outlines the guidelines to be adhered to by users tabulating, analysing, publishing or otherwise releasing any data derived from the survey microdata files. With the aid of these guidelines, users of microdata should be able to produce the same figures as those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

### **9.1 Rounding Guidelines**

In order that estimates for publication or other release derived from these microdata files correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates:

- a) Estimates in the main body of a statistical table are to be rounded to the nearest hundred units using the normal rounding technique. In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, in normal rounding to the nearest 100, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last digits are between 50 and 99 they are changed to 00 and the preceding digit is incremented by 1.
- b) Marginal sub-totals and totals in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units using normal rounding.
- c) Averages, proportions, rates and percentages are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves to one decimal using normal rounding. In normal rounding to a single digit, if the final or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1.
- d) Sums and differences of aggregates (or ratios) are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units (or the nearest one decimal) using normal rounding.
- e) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used resulting in estimates to be published or otherwise released which differ from corresponding estimates published by Statistics Canada, users are urged to note the reason for such differences in the publication or release document(s).
- f) Under no circumstances are unrounded estimates to be published or otherwise released by users. Unrounded estimates imply greater precision than actually exists.

## **9.2 Weighting Guidelines for Tabulation**

Although a census, weights are used in the Survey of Earned Doctorates (SED). Users should apply the proper weight when producing estimates.

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

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

## **9.3 Definitions of Types of Estimates: Categorical and Quantitative**

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

### **9.3.1 Categorical Estimates**

Categorical estimates are estimates of the number, or percentage of the surveyed population possessing certain characteristics or falling into some defined category. The number of graduating visa students in Biological Sciences or the proportion of graduates of Biological Sciences who are visa students are examples of such estimates. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Examples of Categorical Questions:

Q: In what country do you intend to live after graduation (within the next year)?

A: In Canada / U.S. / Other country / Not stated

Q: Do you intend to take a “postdoc” position?

A: Yes / No/Not stated

### **9.3.2 Quantitative Estimates**

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

An example of a quantitative estimate is the average time to complete a doctoral degree among individuals who have received a loan from any source. The numerator is an estimate of the total time taken by graduates receiving a loan from any source, and its denominator is an estimate of the number of graduates who have received a loan from any source.

Example of a Quantitative Question:

Q: After coursework and exams, how many months or years did you work on your dissertation or thesis (non-course related preparation or research, writing, and defense)?

A: Months |\_|\_| OR Years |\_|\_|

### **9.3.3 Tabulation of Categorical Estimates**

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

- a) summing the final weights of records having the characteristic of interest for the numerator  $(\hat{X})$ ,
- 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)  $(\hat{X} / \hat{Y})$ .

### **9.3.4 Tabulation of Quantitative Estimates**

Estimates of quantities can be obtained from the microdata file by multiplying the value of the variable of interest by the final weight for each record, then summing this quantity over all records of interest. 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 obtain an estimate of the average number of months taken to complete a doctoral degree in Engineering:

- a) Consider only the records where the values were reported, i.e., where the number of months, DV\_TIMPH, doesn't equal 999 and where PHDNUMB is between 300 and 399 (Engineering);
- b) For each record, multiply the value reported in derived variable DV\_TIMPH by DWEIGHT, the final weight for the record;
- c) Add up the products calculated in step b);
- d) Add up the values of DWEIGHT over all records where PHDNUMB is between 300 and 399 (Engineering);
- e) Divide the sum from step c) over the sum in step d).

## 9.4 Guidelines for Statistical Analysis

While SED was a census, non-response is a significant source of error in the estimates produced from this survey. Weight adjustments are used to compensate for the non-response, and a variance estimator is used to estimate the errors introduced by non-response. In order for survey estimates and analyses to be as free from bias as possible, the weights must be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures may differ from that which is appropriate in a survey framework, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are poor. Approximate variances for simple estimates such as totals, proportions and ratios (for qualitative variables) can be derived using the accompanying Approximate Variability Table.

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 and the sum of the weights equals the number of respondents.

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

- 1) select all respondents from the file who reported SEX = men;
- 2) calculate the AVERAGE weight for these records by summing DWEIGHT, the weight from the microdata file for these records and then dividing by the number of respondents who reported SEX = men;
- 3) for each of these respondents, calculate a RESCALED weight equal to the original person weight divided by the AVERAGE weight;
- 4) perform the analysis for these respondents using the RESCALED weight.

The calculation of more precise variance estimates requires the use of additional information. These figures are given in a separate file, suitable for use by users with access to software that can support a stratified simple random sample survey designs. Although the SED is not a sample survey, the design information and variance estimate the error due to non-response. To obtain these estimates of error, the software must offer the option of a stratified simple random sample design. (The SED weighting classes can be considered as strata.) If such software is used, DWEIGHT, rather than the rescaled weight described above, is the appropriate weight variable to use. Users may contact Statistics Canada for advice on the appropriate formulae to be used in these calculations.

## 9.5 Coefficient of Variation Release Guidelines

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

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

unacceptable quality.

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

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

**Quality Level Guidelines**

<b>Quality Level of Estimate</b>	<b>Guidelines</b>
1) Acceptable	<p>Estimates have a sample size of five or more, and coefficients of variation in the range of 0.0% to 16.5%.</p> <p>No warning is required.</p>
2) Marginal	<p>Estimates have a sample size of five or more, and coefficients of variation in the range of 16.6% to 33.3%.</p> <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 a sample size of less than five, or coefficients of variation in excess of 33.3%.</p> <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>



## 9.6 Release Cut-off's for the 2003-04 Survey of Earned Doctorates

The following table provides an indication of the precision of population estimates as it shows the release cut-offs associated with each of the three quality levels presented in the previous section. These cut-offs are derived from the coefficient of variation (CV) table discussed in Chapter 10.0.

For example, the table shows that the coefficient of variation of a weighted estimate of 25 graduates in Engineering is between 16.6% and 33.3%. This CV is of marginal quality.

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

Field of Study	Acceptable CV 0.0% – 16.5%	Marginal CV 16.6% – 33.3%	Unacceptable CV > 33.3%
Agricultural Sciences	30 & over	8 to < 30	under 8
Biological Sciences	30 & over	8 to < 30	under 8
Health Sciences	25 & over	7 to < 25	under 7
Engineering	30 & over	9 to < 30	under 9
Computer & Info. Science	25 & over	11 to < 25	under 11
Mathematics	35 & over	11 to < 35	under 11
Astronomy	none	6 to < 15	under 6
Atmospheric Science/Meteorology	none	9 to < 11	under 9
Chemistry	30 & over	9 to < 30	under 9
Geological & related Sciences	30 & over	9 to < 30	under 9
Physics	25 & over	9 to < 25	under 9
Miscellaneous Physical Sciences	30 & over	9 to < 30	under 9
Psychology	25 & over	9 to < 25	under 9
Social Sciences	30 & over	9 to < 30	under 9
Humanities	25 & over	7 to < 25	under 7
Education	30 & over	6 to < 30	under 6
Professional Fields	30 & over	9 to < 30	under 9
Not Stated	none	none	all
<b>All Graduates</b>	<b>15 &amp; over</b>	<b>8 to &lt; 15</b>	<b>under 8</b>

## 10. Approximate Variability Table

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

As SED is not a sample survey, the estimates of CV do not represent an estimate of sampling variability. In SED, the CV estimates the error due to non-response. However, the coefficients of variation are derived using the variance formula for stratified simple random sampling. Coefficients of variation were calculated for a wide range of characteristics; then, from amongst these, a conservative value (the 75<sup>th</sup> percentile) was used in the CV tables which would then apply to the entire set of characteristics.

The Approximate Variability Table can be used for all graduates as well as graduates within almost all major fields of study. There were too few graduates in Astronomy or Atmospheric Science/Meteorology to produce any estimates with an acceptable CV.

All coefficients of variation in the Approximate Variability Table are approximate and, therefore, unofficial. More accurate variance estimates for specific point estimates may be obtained from Statistics Canada on a cost-recovery basis, or calculated with the appropriate software. Since the approximate CV is conservative, the use of a variance estimate calculated for a specific point estimate may cause the estimate to be switched from one quality level to another. For instance a *marginal* estimate could become *acceptable* based on the exact CV calculation, or the reverse could occur. Users may contact Statistics Canada for advice on the appropriate formulae to be used in these calculations

**Remember:** If the number of observations on which an estimate is based is less than five, 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.

### 10.1 How to Use the Coefficient of Variation Table for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the Approximate Variability Table 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 Variability Table, 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 graduates in the Life Sciences who are married is more reliable than the estimated number of graduates in the Life Sciences who are married (Note that in the table the coefficients of variation decline in value reading from left to right).

When the proportion or percentage is based upon the total population of graduates, 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 field of study), 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 graduates in the Life Sciences and the numerator is the number of graduates in the Life Sciences who are married.

In the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of female graduates in Education as compared to the number of male graduates in Education, 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.

### **10.1.1 Examples of Using the Coefficient of Variation Table for Categorical Estimates**

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

A user estimates that 1520 graduates had no debt at the time of graduation. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the coefficient of variation table.
- 2) The estimated aggregate (1520) 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 1500.
- 3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, namely, 1.6%.
- 4) So the approximate coefficient of variation of the estimate is 1.6%. The finding that there were 1520 (to be rounded according to the rounding guidelines in Section 9.1) graduates who owed no debt at the time of their

graduation is publishable with no qualifications.

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

A user estimates that  $402 / 945 = 42.5\%$  graduates in the Life Sciences had no debt at the time of graduation. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the coefficient of variation table.
- 2) Because the estimate is a percentage which is based on a subset of the total population (i.e., graduates in the Life Sciences who have no debt at graduation), it is necessary to use both the percentage (42.5%) and the numerator portion of the percentage (402) in determining the coefficient of variation.
- 3) The numerator, 402, 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 400.
- 4) The percentage estimate, 42.5%, is half-way between the 35% and 50% column headings, so it is necessary to use the lower figure, 35%.
- 5) The figure at the intersection of the row and column used, namely 4.1% is the coefficient of variation to be used.

So the approximate coefficient of variation of the estimate is 4.1%. The finding that 42.5% of graduates in the Life Sciences had no debt at the time of their graduation can be published with no qualifications.

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

A user estimates that  $402 / 945 = 42.5\%$  graduates in the Life Sciences had no debt at the time of graduation, while  $230 / 599 = 38.4\%$  in the Social Sciences had no debt at the time of graduation. How does the user determine the coefficient of variation of the difference between these two estimates?

- 1) Using the coefficient of variation table in the same manner as described in Example 2, gives the CV of the estimate for graduates in the Life Sciences who had no debt at the time of their graduation as 4.1%.

Using the coefficient of variation table in the same manner as described in Example 2, gives the CV of the estimate for graduates in the Social Sciences who had no debt at the time of graduation as 5.5%.

- 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 (Life Sciences),  $\hat{X}_2$  is estimate 2 (Social Sciences),

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  $\hat{d} = 0.425 - 0.384 = 0.041$  is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(0.425)(0.041)]^2 + [(0.384)(0.055)]^2} \\ &= \sqrt{(0.0003036) + (0.00046054)} \\ &= 0.027\end{aligned}$$

- 3) The coefficient of variation of  $\hat{d}$  is given by  $\sigma_{\hat{d}} / \hat{d} = 0.027 / 0.041 = 0.658$ .
- 4) So the approximate coefficient of variation of the difference between the estimates is 65.8%. The finding that the rate of graduates who had no debt at the time of graduation is greater by 4.1% in the Life Sciences as compared to the Social Sciences is considered of unacceptable quality and Statistics Canada recommends this finding 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

A user estimates that 891 graduates who had no debt at the time of graduation are males while 629 are females. The user is interested in comparing the estimate of males with no debt versus the estimate of females with no debt 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 males who have no debt at the time of their graduation. The denominator of the estimate ( $\hat{X}_2$ ) is the number of females who have no debt at the time of their graduation.
- 2) Refer to the coefficient of variation table.
- 3) The numerator of this ratio estimate is 891. The figure closest to it is 1000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 2.3%.
- 4) The denominator of this ratio estimate is 629. The figure closest to it is 750. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 2.9%.
- 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:

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(0.023)^2 + (0.029)^2} \\ &= \sqrt{0.000529 + 0.000841} \\ &= 0.037\end{aligned}$$

The obtained ratio of males with no debt versus females with no debt is 891 / 629 which is 1.42:1 (to be rounded according to the rounding guidelines in Section 9.1). The coefficient of variation of this estimate is 3.7%, which makes the estimate releasable with no qualifications.

### Example 5: Estimates of Differences of Ratios

A user estimates the ratio of male to female graduates in the Life Sciences with no debt at the time of graduation is 203 versus 199 or 1.02:1, while the ratio for graduates in the Social Sciences with no debt at the time of graduation is 98 versus 132 or 0.74:1. The user is interested in comparing both 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 Life Sciences ratio ( $\hat{R}_1$ ) and the Social Sciences ratio ( $\hat{R}_2$ ) as in Example 4. The approximate CV for the Life Sciences ratio is 8.5% and 11.9% for Social Sciences.
- 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  $\hat{d} = 1.02 - 0.74 = 0.28$  is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(1.02)(0.085)]^2 + [(0.74)(0.119)]^2} \\ &= \sqrt{(0.00752) + (0.00775)} \\ &= 0.124\end{aligned}$$

- 3) The coefficient of variation of  $\hat{d}$  is given by  $\sigma_{\hat{d}} / \hat{d} = 0.124 / 0.28 = 0.443$ .
- 4) So the approximate coefficient of variation of the difference between the estimates is 44.3%. The finding that the ratio of male to female graduates with no debt is greater by 0.28:1 in the Life Sciences as compared to the

Social Sciences is considered of unacceptable quality and Statistics Canada recommends this finding not be released. However, should the user choose to do so, the estimate should be flagged with the letter U (or some similar identifier) and be accompanied by a warning to caution subsequent users about the high levels of error associated with the estimate.

## 10.2 How to Use the Coefficient of Variation Table to Obtain Confidence Limits

Although coefficients of variation are widely used, a more intuitively meaningful measure of 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 surveying of the population is repeated indefinitely, and the population showed the same response mechanisms and patterns, with each sample leading to a new confidence interval for an estimate, then in 95% of the surveys 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 Variability Table 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: These t-values are not exact, but are suitable approximations given the release guidelines.



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

### **10.2.1 Example of Using the Coefficient of Variation Table to Obtain Confidence Limits**

A 95% confidence interval for the estimated proportion of graduates in the Life Sciences who have no debt at the time of graduation (Example 2, Section 10.1.1) would be calculated as follows:

$$\hat{X} = 42.5\% \text{ (or expressed as a proportion 0.425)}$$

$$t = 2$$

$\alpha_{\hat{x}}$  = 4.1% (0.041 expressed as a proportion) is the coefficient of variation of this estimate as determined from the table.

$$CI_{\hat{x}} = \{0.425 - (2) (0.425) (0.041), 0.425 + (2) (0.425) (0.041)\}$$

$$CI_{\hat{x}} = \{0.425 - 0.035, 0.425 + 0.035\}$$

$$CI_{\hat{x}} = \{0.39, 0.46\}$$

With 95% confidence it can be said that between 39% and 46% of graduates in the Life Sciences had no debt at the time of graduation.

### **10.3 How to Use the Coefficient of Variation Table 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  $\sigma_{\hat{d}}$ .

$$\text{If } t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$$

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. That is to say that the difference between the estimates is significant.

### **10.3.1 Example of Using the Coefficient of Variation Table 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 graduates in the Life Sciences who have no debt at the time of graduation and the proportion of graduates in the Social Sciences who have no debt at the time of graduation. From Example 3, Section 10.1.1, the standard error of the difference between these two estimates was found to be 0.027. Hence,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{0.425 - 0.384}{0.027} = \frac{-0.041}{0.027} = 1.5$$

Since  $t = 1.5$  is between -2 and 2, it must be concluded that there is no significant difference between the two estimates at the 0.05 level of significance.

### **10.4 Coefficients of Variation for Quantitative Estimates**

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

As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). If the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the total amount of money owed for government student loans would be greater than the coefficient of variation of the corresponding number of individuals. Hence, if the estimated number of individuals is not releasable, then the corresponding quantitative estimate will likely not be releasable.

Coefficients of variation can be derived as required for specific point estimates. Users wishing to derive coefficients of variation for quantitative estimates may use statistical software that can handle a stratified simple random sample survey designs, together with the supplemental design file. Users may contact Statistics Canada for advice on the appropriate formulae to be used in these calculations.

### 10.5 Table of Approximate Variability

Numerator of Percentage	Percentage											
	5	10	15	25	35	50	65	75	85	90	95	
5	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0
10	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0
12	26.0	26.0	26.0	26.0	26.0	26.0	26.0	26.0	26.0	26.0	26.0	26.0
15	24.0	23.8	23.6	23.6	23.6	23.6	23.6	23.5	23.5	23.4	23.4	23.4
20	21.0	21.0	21.0	21.0	21.4	21.9	21.6	21.3	19.8	19.8	19.8	19.8
25	18.0	18.0	18.0	17.9	17.9	17.9	17.9	17.9	17.9	17.9	17.9	17.9
30	16.5	16.5	16.5	16.4	16.4	16.3	16.3	16.3	16.3	16.3	16.3	16.3
40	14.6	14.6	14.5	14.5	14.5	14.5	14.4	14.4	14.4	14.4	14.4	14.4
50	13.2	13.1	13.1	13.0	13.0	13.0	12.8	12.6	12.6	12.6	12.6	12.6
60	11.7	11.6	11.6	11.6	11.5	11.5	11.5	11.3	11.3	11.3	11.3	11.2
70	11.1	11.1	11.1	11.1	11.1	11.0	11.0	11.0	11.0	11.0	11.0	11.0
80	10.0	10.0	9.9	9.9	9.8	9.8	9.8	9.8	9.8	9.8	9.8	9.8
90	9.3	9.3	9.3	9.2	9.2	9.2	9.0	9.0	9.0	9.0	9.0	9.0
100	9.0	9.0	9.0	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9
125	7.8	7.7	7.7	7.7	7.7	7.7	7.5	7.5	7.5	7.5	7.5	7.5
150	6.9	6.9	6.9	6.9	6.9	6.8	6.8	6.8	6.8	6.8	6.8	6.8
175	6.5	6.5	6.5	6.5	6.4	6.4	6.4	6.4	6.3	6.3	6.3	6.3
200	6.0	6.0	6.0	6.0	6.0	6.0	5.9	5.9	5.8	5.8	5.8	5.8
250	*****	5.5	5.5	5.5	5.5	5.5	5.4	5.4	5.4	5.4	5.4	5.4
300	*****	4.9	4.8	4.8	4.8	4.8	4.8	4.7	4.7	4.6	4.6	4.6
400	*****	4.1	4.1	4.1	4.1	4.0	4.0	4.0	4.0	4.0	4.0	4.0
500	*****	3.8	3.8	3.8	3.8	3.8	3.7	3.6	3.5	3.4	3.3	3.3
750	*****	*****	*****	2.9	2.8	2.7	2.7	2.7	2.7	2.7	2.6	2.6
1000	*****	*****	*****	2.3	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
1500	*****	*****	*****	*****	*****	1.6	1.6	1.6	1.6	1.6	1.6	1.6
2000	*****	*****	*****	*****	*****	*****	1.2	1.2	1.2	1.2	1.2	1.2
2500	*****	*****	*****	*****	*****	*****	*****	0.9	0.9	0.8	0.8	0.8
3000	*****	*****	*****	*****	*****	*****	*****	*****	0.6	0.5	0.4	0.4

## **11. Questionnaire and Code Sheets**

Please refer to the files listed below for the Survey of Earned Doctorates 2003-2004.

### **Questionnaires:**

Survey of Earned Doctorates-E.pdf  
Enquête titulaire dun doctorat-F.pdf

### **Code Sheets:**

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

Appendix A - CIP.pdf

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

Appendix B - NAICS.pdf    OR    Appendix B - NAICS.doc

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

Appendix C - NOC-S.pdf    OR    Appendix C – NOC-S.doc

## **12. Record Layout with Univariate Frequencies**

See SED\_Codebook\_master\_e.pdf OR SED\_Codebook\_master\_e.doc for the record layout with univariate counts.

