Back to referring page
Français

Catalogue no. 89M0019GPE

ETHNIC DIVERSITY SURVEY

(2002)

User's Guide to the Public Use Microdata File

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Social and Aboriginal Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-5979).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our Web site.

National inquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Depository Services Program inquiries	**1 800 700-1033**
Fax line for Depository Services Program	**1 800 889-9734**
E-mail inquiries	**infostats@statcan.ca**
Web site	**www.statcan.ca**

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1 800 263-1136.

Statistics Canada
Social and Aboriginal Statistics Division

Ethnic Diversity Survey (2002)

User's Guide to the Public Use Microdata File

Published by authority of the Minister responsible for Statistics Canada

May 2005

Catalogue no. 89M0019GPE

Frequency: Occasional

Ottawa

Cette publication est disponible en français (n^{o} 89M0019GPE au catalogue)

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Symbols

The following standard symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published

Table of contents

1.0 Introduction

The Ethnic Diversity Survey (EDS) was conducted jointly by Statistics Canada and the Department of Canadian Heritage. This survey has two objectives. First, the data will help us to better understand how people's backgrounds affect their participation in Canada's social, economic and cultural life. Secondly, the information that is gathered will help us to better understand how Canadians of different ethnic origins interpret and report their ethnicity. The information collected in the survey will also be used to inform policy and program development in the Department of Canadian Heritage.

This document was developed to facilitate the use of the Ethnic Diversity Survey's Public Use Microdata File (PUMF). It describes the survey's methodology, data quality and other issues related to data analysis and dissemination.

Any questions about the EDS PUMF or its use should be directed to:

Client Services
Social and Aboriginal Statistics Division
Statistics Canada
Jean Talon Building, 7th floor
Tunney's Pasture
Ottawa, Ontario
K1A 0T6

Telephone: (613) 951-5979
Fax: (613) 951-0387
Email: sasd-dssea@statcan.ca

2.0 Survey Methodology

This chapter presents a brief description of the methods used in the Ethnic Diversity Survey (EDS). It also deals with the main aspects of data quality as well as data analysis and dissemination guidelines. This information is aimed at helping users understand the strengths and limitations of the data and to assist them in using it properly.

2.1 Sampling Frame

A sampling frame provides access to the population that must be included in a survey. The Ethnic Diversity Survey was a post-censal survey, which means that its sample was obtained from the census. EDS respondents were selected based on the responses that they had provided to certain questions in the last census, which took place on May 15, 2001. Thus, the EDS sampling frame was created from the list of persons who had provided responses on the long questionnaire (2B) in the 2001 Census.

2.2 Target Population

The EDS target population consisted of persons aged 15 and older living in private dwellings in Canada's ten provinces. Just as in the census, Canadian citizens, landed immigrants and non-permanent residents (holders of student, work or ministerial permits, refugee status claimants and family members living in Canada with them) were part of the target population. However, the following groups were excluded:

- persons under 15 years of age;
- persons living in collective dwellings (hotels, nursing homes, hospitals, military or work camps, prisons, residences for senior citizens, etc.);
- Indian reserves;
- persons who declared an Aboriginal ethnic origin or Aboriginal identity on the 2001 Census;
- the territories and remote areas.

The target population of the EDS represents 23,092,643 persons in the Canadian population. Among them, 57,242 persons were selected for the survey. In total, 42,476 respondents participated in the EDS.

To ensure confidentiality, a sub-sample was draw from the EDS respondents for the EDS PUMF. The EDS PUMF includes 41,695 respondents representing the same population of 23,092,643 persons.

2.3 Reference Period and Data Collection

The Ethnic Diversity Survey's reference period corresponds to that of the data collection, which took place between April and August 2002.

Statistics Canada regional office employees collected EDS data using *Blaise* software and the *computer-assisted telephone interview* (CATI) method. The use of this technology allows for the collection of high quality data at a reasonable cost for surveys with long and complex questionnaire designs such as the EDS.

The average length of an EDS interview was between 35 and 45 minutes, depending on the respondent's characteristics. In addition to being conducted in English and French, interviews were administered in the following seven non-official languages: Mandarin, Cantonese, Italian, Punjabi, Portuguese, Vietnamese and Spanish. Interviews conducted in languages other than English and French were generally longer than interviews conducted in English or French.

Proxy (or third person) responses were not permitted; the person who had been pre-selected had to be reached and interviewed.

2.4 Sample Design

The Ethnic Diversity Survey was a probabilistic survey, which means that a random sample was selected to represent the target population. The EDS used a two-phase stratified sampling design. Phase I, the 2001 Census, consisted of distributing the long questionnaire to one out of five households in Canada, on average. Phase II consisted of selecting a sample of Phase I respondents on the basis of the responses given to the 2001 Census questions on ethnic origin, place of birth and place of birth of parents.

In order to meet the objectives of the survey, it was fundamental to target two main groups to ensure sufficient counts of these sub-populations in the sample. These two main groups included persons reporting "Canadian" ethnic origin in the 2001 Census and persons belonging to non-Canadian, non-British or non-French ethnic groups.

In Phase II, responses to the 2001 Census ethnic origin question were divided up into two main categories:

- **CBFA+**
  where C = Canadian, B = British Isles, F = French and A+ = American, Australia and/or New-Zealander

- **Non-CBFA+**
  All other responses containing at least one origin other than CBFA+

The CBFA+ category was subdivided according to whether or not the response included a "Canadian" origin. The Non-CBFA+ category was subdivided into European origins (for example, German, Italian, Dutch, Portuguese) and non-European origins (for example, Chinese, Jamaican, Lebanese, Iranian). The Non-CBFA category was then further subdivided depending on whether or not the response included a Canadian origin.

Next, the questions on birthplace and birthplace of parents were used to establish the respondent's generational status. **Generation 1** included respondents who were born outside Canada. **Generation 2** included respondents who were born in Canada but who had at least one parent who was born outside Canada. **Generation 3** included respondents born in Canada to two Canadian-born parents. Where necessary, the groups or strata created using generational status were collapsed in order to ensure a sufficiently high number of persons in each stratum.

As a result, the EDS target population was divided among the following strata:

- **CBFA+**
  - Canadian only
    - Generation 1 et 2
    - Generation 3 et plus
  - Canadian with BFA+
    - Generation 1 et 2
    - Generation 3 et plus
  - BFA+ only
    - Generation 1 et 2
    - Generation 3 et plus

- **Non-CBFA+**
  - Other European with Canadian
    - Generation 1 et 2
    - Generation 3 et plus
  - Other non-European with Canadian
    - All generations
  - Other European
    - Generation 1
    - Generation 2
    - Generation 3 et plus
  - Other non-European
    - Generation 1
    - Generation 2 et plus

(Multiple ethnic origin responses which contained both a European and a non-European group were classified with other non-European origins for the purposes of stratification).

This stratification greatly simplified the sample strategy since it eliminated the need to classify hundreds of single and multiple ethnic origins as well as the many birthplaces reported in the census.

2.5 Sample Size and Selection

In light of the Ethnic Diversity Survey's objectives and of the need for data on certain sub-populations, particularly persons in the first and second generations, the EDS sample distribution was established at 1/3 for CBFA+ and at 2/3 for non-CBFA+. This distribution both ensured that persons with non-Canadian, non-British, non-French origins would be well represented in the sample and that the EDS would still include, and be relevant to, all people in Canada.

In order to support data analysis, the goal of the survey was to obtain at least 40,000 respondents. A sample size of approximately 57,200 persons was required since a response rate of 70% was expected. The coefficient of variation (c.v.) was used as a measure of reliability. For each stratum, the initial sample size was determined by a minimum proportion of 4%, a maximum c.v. of 12.5%, a design effect of 1.2 and expected response rates for each stratum which were estimated based on results from the September 2001 EDS Pilot Test.

Once the strata were determined, the sampling frame was ordered by province, electoral district, enumeration area and household in order to ensure a good geographic distribution of the sample and to reduce the likelihood that people from the same household would be selected for an interview. A systematic sample was then selected independently in each stratum. (A systematic sampling entails the selection of units from a list, according to a selection interval that has been established beforehand.)

The final EDS sample included 57,242 persons. Of these, 42,476 responded to the survey. This represents a total response rate of 75.6%, if the 1,057 persons classified as being outside the scope of the survey are taken into account. From this rate, 73.1% were complete responses and 3.2% were partial responses. In general, when persons started an interview, they answered all the survey questions. Complete responses, therefore, represent 96.8% of all survey respondents. Total non-response represents 24.4% of the sample. The response rate by stratum, defined earlier, varies between 72% and 80%. As might have been expected, first generations provided a lower rate of response, at 73% compared to 77% for the second and third generations and more.

Of the 42,476 EDS respondents, 781 were not included in the EDS PUMF because there was a risk of disclosure associated with those respondents. However, some adjustments were done to the survey weights to ensure that the EDS PUMF is still representative of the EDS target population. As such the EDS PUMF includes 41,695 respondents, representing a population of 23,092,643 persons.

2.6 Questionnaire Processing

After the coding of "Other – Specify" responses, edits and verifications of all sampling units were performed. These edits and verifications included reviewing remarks and notes written by the interviewers in the questionnaire; verifying each answer in order to identify missing, invalid or inconsistent entries; and verification of outcome codes assigned by the interviewers to each questionnaire. A master file containing "clean" data and weights from the survey was then created.

2.6.1 Definition of Response Status

One of the preliminary steps of the weighting process is to verify outcome codes in order to assign a response status to each sampling unit. In order to do this, there must be a record for each person selected in the sample. Then, each record is assigned one of the following statuses:

i) complete response
ii) partial response
iii) total non-response
iv) out of scope

There is a **complete response** when all, or almost all, of the data are collected for a sampling unit. In the EDS, a record received the complete response status when a valid response was provided in at least **80%** of the mandatory questions (at least 31 of the 39 questions asked of all respondents) **and** when at least one valid response was provided in one of the following three questions: ethnic ancestry (ID_Q010), ethnic identity (ID_Q100) and place of birth (BK_Q010).

There is a **partial response** when only some data are collected for a sampling unit. In the EDS, a record received the partial response status when a valid response was provided in at least 31% of mandatory questions (at least 12 of the 39 questions asked of all respondents) but to not more than 79.9% of mandatory questions **or** when a valid response was provided in less than 31% of mandatory questions **and** a valid response was provided in all of the three following questions: ethnic ancestry (ID_Q010), ethnic identity (ID_Q100) and place of birth (BK_Q010).

There is **total non-response** when no data (or almost no data) are gathered for a sampling unit. In the EDS, a record received the total non-response status when a valid response was provided for less than **31%** of mandatory questions **and** no valid response was provided in any of the three following questions: ethnic ancestry (ID_Q010), ethnic identity (ID_Q100) and place of birth (BK_Q010).

A unit is **out of scope** when it is in the survey frame but, according to information collected during the survey, the unit is not part of the target population. In the EDS, persons who were out of scope were persons who were deceased; younger than 15 years of age; had moved to one of the three territories; were living on an Indian reserve; were living in a collective dwelling; etc.

Units considered to be “total non-response” or “out of scope” were removed from the final EDS data file. Only respondents who were assigned the “complete response” or “partial response” status are included on the EDS master file.

2.6.2 Verification of validity of responses

Another part of the EDS questionnaire processing stage consisted of verifying the validity of each response provided by survey respondents. Consistency and question flow verifications were performed. The first objective of these verifications was to detect errors, weakness and inconsistencies in survey data in order to correct them. For example, if a respondent reported speaking a language at work that had not been reported previously in the knowledge of languages question, the language at work was considered to be invalid or inconsistent and, when possible, was corrected based on supplementary data and interviewer notes.

As part of the verification process, each response to each question was classified according to the categories described below:

- **Valid response**

The respondent provided an answer to a question that he/she was supposed to answer. A valid response differs from responses "Don't know" and "Refused".

It should be mentioned that some valid responses were assigned to the code "uncodeable" during processing. "Uncodeable" responses are responses which could not be assigned to an existing code by the interviewer during collection or by a survey expert during processing for one of the following three reasons:

(1) The response was unintelligible. For example, an ethnic ancestry of "xyzlocan" would have been considered unclassifiable and coded to "Uncodeable".

(2) The response could potentially be assigned to more than one code and it could not be determined which code was the correct one. For example, a place of birth response of "Albertville" would have been coded to "Uncodeable" if it was not possible to determine whether this response indicated a city in Canada, France or the United States.

(3) The response that was provided to the ethnic ancestry or ethnic identity question indicated that the respondent may have understood the question to some extent but did not provide a response considered to be the appropriate type of response to that question. For example, "Uncodeable" responses to the ethnic ancestry question included responses which indicated the respondent's family name, immigrant status or general comments about their family history.

- **Don't know**

The respondent did not know the answer to the question. In the EDS data file, a "Don't know" response is coded as "9". In the case of a two-digit variable, the code is "99"; for a three-digit variable, the code is "999", etc.

- **Refused**

The respondent refused to answer the question. In the EDS data file, a refusal is coded as "8". In the case of a two-digit variable, the code is "98"; for a three-digit variable, the code is "998", etc.

- **Not applicable**

The respondent did not have to answer the question because a particular response was given to the corresponding filter question. A filter question is the first question in a group of questions and is used to screen out respondents for whom the subsequent questions would be irrelevant. In the EDS file, a respondent for whom a question was "Not Applicable" is assigned to code "7". In the case of a two-digit variable, the code is "97"; for a three-digit variable, the code is "997", etc.

- **Not asked**

On the EDS data file, there are two types of responses which are assigned to the final code for "Not Asked", which is represented by a code of "6", or in the case of a a two-digit or three-digit variable, code "96" or code "996", etc. These two types of responses are "Not asked" and "Path unknown".

Not asked: The respondent was supposed to answer the question, but the question was not asked. This is usually the result of a special circumstance or system error. For example, if a respondent refused to identify other members of his household in the EDS Entry Module, but then later mentioned that he had a wife and this was written in the survey notes by the interviewer, this respondent's marital status was imputed as "Married". However, because this imputation was done during processing (after data collection) the respondent would not have been asked any of the questions regarding his wife during the survey. Thus, the code "Not asked" was assigned to all questions about his wife.

Path unknown: It is unknown whether or not the respondent was supposed to answer the question because both the question and the earlier filter question were without an answer. This situation is usually the result of a filter question that was not asked. In the EDS file, a question with an unknown flow was originally coded "-1" but on the final data file was aggregated with codes "6", "96", "996" etc.

During analysis, users will need to define their estimation domain (total population) for each variable. It will be important to consider whether or not “Don’t know”, “Refused”, “Not applicable” and “Not asked” codes should be included or excluded. The inclusion or exclusion of each of these codes depends on the objective of the analysis. However, users who would like to account for partial non-response during data analysis should include the codes “Don’t know” and “Refused” in the domain of each variable and should exclude the codes “Not applicable” and “Not asked”.

Analysts who wish to produce the same figures as those already published by Statistics Canada in the analytical report “Ethnic Diversity Survey: portrait of a multicultural society” (released on September 29, 2003) should generally exclude counts for “Don’t Know”, “Refused”, “Not Asked” and “Not Applicable” from their totals. Although calculations varied according to the issue under investigation, the percentages included in that report were usually calculated with a denominator which was equal to the sum of all valid responses only.

2.7 The Aboriginal Population

People who reported Aboriginal origins or identities in the 2001 Census were not part of the population targeted for the Ethnic Diversity Survey (EDS). These people were excluded from the EDS target population. The main reasons for the exclusion of the Aboriginal population were respondent burden (these people were already covered by another post-censal survey: the Aboriginal Peoples Survey) and the data collection method (telephone interviews are not suitable for data collection on most reserves). (For more information on the Aboriginal Peoples Survey, please refer to http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3250&lang=en&db=IMDB&dbg=f&adm=8&dis=2 or contact client services using the contact information provide on page 3 of this User Guide).

There were, however, EDS respondents (810 in the full dataset and 793 in the PUMF) who reported an Aboriginal ancestry in questions ID_Q010, ID_Q020 and/or an Aboriginal identity in question ID_Q100. There may be a number of explanations for this, including proxy reporting in the Census, the reporting of only a Canadian origin in the 2001 Census ethnic origin question and the fluidity of the concept of ethnicity.

Prior to collection, it was decided that any EDS respondent who reported Aboriginal ethnic ancestry or identity in the survey would be considered out of scope. The Blaise application was thus built to screen for Aboriginal answers to the ethnic ancestry and ethnic identity questions and to consider these responses out of scope. If one of a pre-determined list of Aboriginal ancestries or identities were reported, the application went directly to the last question of the survey.

Nevertheless, during data processing it was decided to keep any respondents who had reported Aboriginal ancestry or Aboriginal identity in the EDS data file. More than half of these persons had answered all EDS questions because they had not been screened out by Blaise. It was considered appropriate to keep these respondents in the survey sample because they were not covered by the Aboriginal Peoples Survey. As well, specific analysis of the characteristics of this population is planned.

In the release of Ethnic Diversity Survey data that occurred on September 29, 2003, respondents with any Aboriginal ancestry or identity were excluded from the analysis for the release because this population was originally to be excluded from the EDS target population. However, no Aboriginal ancestry or identity variables could be included on the EDS PUMF because such variables would have presented a risk for disclosure. As a result, analysis of the PUMF data may result in slightly different results than those presented in the release day data.

3.0 Estimation

In a probabilistic sample such as the EDS sample, estimation is based on the principle that each person included in the sample represents not only him/herself but also a number of other persons who were not included in the sample. For example, in a simple random sample of 2% of the population, each person represents 50 members of the population (him/herself and 49 others). The number of persons represented by a given respondent is known as the respondent's weight or weighting factor.

A weighting factor is included in the EDS microdata file:

WGT_PUMF : This is the weight for analysis with respect to persons, that is, for calculating estimates of the number of persons (included in the target population) with one or more of specified characteristics. WGT_PUMF should be used to calculate all estimates. For example, to estimate the number of persons who are black, it is necessary to sum the WGT_PUMF values for all records that include this characteristic (VISMIND=4).

3.1 Weighting

As noted above, EDS 2002 is a survey of individuals, and the microdata file contains responses to the questionnaire and related information provided by 41,695 respondents. Calculating the weight for the PUMF was a four-stage process:

1) Calculating the initial weight

The first stage was the assignment of an initial weight based on the sampling design. The initial weight is the inverse of the probability of inclusion in the sample. For the 2002 EDS, the initial weight was the product of two components: the 2001 Census weight and the inverse of the person's probability of selection for EDS (sampling weight). Following this calculation, individuals selected by mistake and those missed during sample selection were taken into consideration and the appropriate weight adjustments were applied to the initial weight.

2) Correction for non-response

The second stage of the weighting process was adjustment for non-response. This stage consisted of applying a correction factor to the initial weights to compensate for the effects of non-response. The "response propensity model" was used. This method predicts the probability of responding to the survey using a logistic regression model with a set of independent variables.

To apply such a method, the EDS could employ a considerable number of variables, namely the variables of the long census questionnaire. With this method, the probabilities of predicted responses that result from the model are used to classify individuals into groups of approximately the same size so that individuals who have similar predicted probabilities are in the same class. The inverse value of the weighted response rate of each group is used as an adjustment factor for this group. The initial weights of respondents are multiplied by this adjustment factor.

Different models were thus developed successively for persons who were not contacted and contacted persons who did not respond. Approximately ten classes of roughly the same size were obtained for each logistic regression model.

3) Post-stratification

The third step in the weighting process was an *a posteriori* stratification, also called post-stratification. This method ensures that the sum of respondents' weights from the EDS corresponds to 2001 Census tabulations for each variable used. More specifically, in the case of the EDS, post-stratification was done by cross tabulating region, stratum, age group and sex.

Moreover, in order to ensure that the final survey estimates for other selected variables agreed with their known census distributions, the weights were adjusted for different geographic levels by using the raking ratio estimation method. This adjustment was made separately for religion, generation, mother tongue and visible minorities.

The weights corrected for non-response were then adjusted using the ratio of the census count to the sample count. The weights obtained after this stage were used to produce all estimates for the different releases of EDS data.

4) Additional adjustments of weights for the PUMF

Since the PUMF is a sub-sample of EDS respondents, a fourth step was necessary in the weighting process. This step consists of additional adjustments made to the weights of units in the PUMF to take into account the units removed. To do this, the weight obtained in step 3 was first multiplied by the sub-sampling weight. Then, a new post-stratification adjustment was made in order to re-adjust to the Census counts.

3.2 Weighting Guidelines

Thus, the final weight assigned to each respondent underwent numerous adjustments so that respondents would better represent the target population. Weighting of the data ensured that the EDS PUMF sample is representative of the target population even if the sampling ratio differs widely from one individual to another. **The use of the weights is then essential for all analyses that use the survey data.**

Users should not disseminate any unweighted total or perform analyses based on unweighted survey results. Sampling rates and non-response rates vary considerably from one stratum to another, and non-response rates also vary according to demographic characteristics. Clearly, therefore, unweighted sample counts cannot be considered as representative of the population targeted by the survey.

3.3 Types of Estimation

Using EDS data, two types of “simple” estimates can be calculated: qualitative estimates (estimates of numbers or proportions of persons with certain attributes or characteristics) and quantitative estimates (estimates of quantities or averages). Section 7.4 deals with more complex estimates and analyses.

3.3.1 Qualitative Estimates

Qualitative estimates are estimates of the number or percentage of persons in the population targeted by the survey who have a certain characteristic or fall into a defined category. The values of these variables represent a quality rather than a quantity. An example of a qualitative estimate is the number or proportion of persons who reported “High school diploma” as the highest level of schooling completed.

Qualitative estimates can be obtained by summing the final weights of all records that contain the characteristic(s) of interest. Proportions and ratios of the form $\hat{Y}/\hat{W}$ are obtained by following the steps below:

(i) sum the final weights of records containing the characteristic of interest $\hat{Y}$;
(ii) sum the final weights of records containing the characteristic of interest $\hat{W}$;
(iii) divide the result obtained in (i) by the result obtained in (ii), namely $\hat{Y}/\hat{W}$.

3.3.2 Quantitative Estimates

Quantitative estimates are estimates of totals or means, medians or other measures of central tendency representing quantities. The number of weeks or hours worked is an example of a quantitative estimate.

This type of estimate can be obtained by multiplying the value of the variable of interest by the final weight of the corresponding record and summing this amount for all records selected. To obtain a weighted average of the form $\hat{Y}/\hat{W}$, the numerator ($\hat{Y}$) is calculated in the same way as a quantitative estimate and the denominator ($\hat{W}$) in the same way as a qualitative estimate. For example, to estimate the average number of hours worked by respondents, proceed as follows:

(i) estimate the total number of hours worked by respondents ($\hat{Y}$) by multiplying the number of hours worked by each respondent by its corresponding final weight, then sum this value for all respondents;

(ii) estimate the number of respondents ($\hat{W}$) by summing the final weights for all records corresponding to a respondent;

(iii) divide (i) by (ii), namely $\hat{Y}/\hat{W}$.

3.4 Guidelines for Analysis

As explained in detail in Section 2.4, EDS respondents do not constitute a simple random sample of the target population. The survey is based on a complex sampling design. Consequently, the selection of respondents was done according to unequal probabilities.

Survey weights must therefore be used in making estimates and analyses so that insofar as possible, the over- or under-representation of some groups in the unweighted file can be taken into consideration. The use of data from such a complex survey can pose problems for analysts, since the choice of methods of estimation and variance calculation depends on the sampling design and selection probabilities. A number of analysis methods integrated into statistical packages allow the use of weights, but the meaning and definition of these weights often differ from those that apply in the context of a sample survey. Therefore, while the estimates made using these packages are often accurate, the variances calculated are practically meaningless.

In many methods of analysis (such as linear regression, logistic regression, estimation of rates or proportions and analysis of variance), the application of current software packages can be made more meaningful by standardizing the weights that appear in the records so that the average weight is equal to 1. The results produced by traditional packages are thus more reasonable, because even though they do not always reflect the stratification and clustering in the sampling design, they take account of selection with unequal probabilities. This standardizing can be done by dividing each weight by the overall average weight before proceeding to the analysis.

For example, for an analysis of all respondents who are “black” according to the visible minority variable, the procedure to follow is as follows:

- from the file, select all respondents who were black (VISMIND=4);
- calculate the average value of WGT_PUMF for all these records;
- for each of these respondents, calculate a “working” weight equal to WGT_PUMF/ average weight;
- carry out the analysis for these respondents using the “working” weight.

Section 4 gives a more detailed description of sampling variability and data reliability, and “Appendix - Variance estimation” contains the rules for obtaining the approximate variance for estimating the sampling variability of a large number of qualitative estimates of proportions.

4.0 Guidelines on data dissemination and reliability

It is important for the user to become familiar with the content of this section before publishing or otherwise disseminating any estimate calculated using the EDS microdata file.

This section of the document gives guidelines that users of the microdata file must follow. Users will thus be able to obtain figures which are consistent with those produced by Statistics Canada and which conform to established guidelines on rounding and dissemination. The guidelines fall into three major categories: minimum sample size for producing estimates; sampling variability and rounding.

It must be noted that results obtained using the EDS microdata file might be slightly different from the ones published using the analytical file because of the sub-sampling and the other methods applied to ensure confidentiality of the data.

4.1 Minimum Sample Size for Producing Estimates

The user must determine the number of records in the microdata file that provided the data entering into the calculation of a particular estimate. If the number is less than or equal to 10, the weighted estimate must generally not be disseminated, regardless of its approximate coefficient of variation. If the estimate is nevertheless disseminated, this must be done with considerable caution, and the user should clearly indicate that the estimate is based on an insufficient number of records. Please note that suppressed data must be included in the totals (if the totals are greater than 10).

4.2 Sampling Variability

Estimates drawn from the survey are based on a sample of individuals. Different figures might have been obtained if a complete census had been conducted using the same questionnaire, interviewers, supervisors, data processing methods, etc. The difference between an estimate produced from a sample and one produced from a complete enumeration conducted in similar conditions is called the sampling error of the estimate.

Errors unrelated to sampling can occur at almost any stage of a survey. Interviewers may misunderstand the instructions, respondents may make mistakes when answering questions, responses may be miskeyed into the computer or errors may occur when the data are processed or totalled. These are all examples of non-sampling errors.

If there are a large number of observations, random errors have little effect on estimates based on data from the survey. However, errors that occur systematically bias estimates. Much time and effort has been devoted to reducing non-sampling errors. At each stage of the data collection and processing cycle, quality assurance measures were applied in order to control data quality. These measures included using highly skilled interviewers, giving them intensive training in survey methods and the questionnaire, observing interviewers in order to detect problems caused by the questionnaire design or a failure to understand the instructions, adopting procedures to minimize data capture errors, and implementing coding controls and edits to confirm the processing logic.

4.2.1 Non-sampling Errors

The effect of non-response on survey results is a major source of non-sampling error. Non-response may be either partial (not answering one or more questions) or total. There is total non-response when the interviewer is unable to locate the respondent or the respondent cannot provide the desired information (perhaps because of a language problem) or refuses to participate in the survey. Cases of total non-response are treated by correcting the weight applied to those persons who responded to the survey to compensate for those who did not.

In most cases, there was partial non-response to the survey where the respondent misunderstood or misinterpreted a question or was unable to remember the information requested. In EDS, no responses were imputed to compensate for partial non-response, and the question was assigned the response code “Not stated.”

4.2.2 Sampling Errors

Since estimates based on a sample survey inevitably contain sampling errors, good statistical methods require researchers to inform users of the magnitude of this type of error. Although it is not possible to obtain an exact measure of the sampling error of an estimate as defined above using the sample data alone, it is possible to estimate a statistical measure of this error, namely the standard error, using these data. Based on the standard error, confidence intervals can be obtained for estimates (not taking the effects of non-sampling errors into account) on the assumption that the distribution of the estimates around the true value of the population is normal. In these conditions, the chances that the deviation between an estimate based on the sample and the true value for the population is less than one standard deviation are 68 in 100, while the chances that it is less than two standard deviations are approximately 95 in 100, and it is virtually certain that it is less than three standard deviations.

Since the absolute size of the sampling error of an estimate is often less important than its relative size (compared to the estimate itself), the standard error is not always the best measure of sampling error. For example, a standard error of 10 for an estimate of 20 would generally indicate that the quality of the estimate is poor, while the same standard error for an estimate of 1,000 would generally indicate that the estimate is good. Consequently, the size of the sampling error is often expressed in relation to the size of the estimate, in the form of a coefficient of variation (CV). The coefficient of variation of an estimate is obtained by dividing the standard error of the estimate by the estimate itself and expressing the resulting fraction as a percentage. In the above example, the CV of the first estimate is 50% (10/20), while that of the second is 1% (10/1,000).

Guidelines for Dissemination of Estimates

Before disseminating and/or publishing estimates based on the microdata file, the user must calculate the coefficient of variation associated with the estimates, should consult the table below and follow the guidelines corresponding to the value of the coefficient of variation of the estimate.

For more information about the bootstrap method and how to calculate coefficient of variation with the EDS data, please refer to “Appendix - Variance estimation”.

Coefficient of variation - Guidelines for dissemination

Estimator's Level of Quality	Symbol to Use	Condition	Guidelines
Acceptable	Requires no symbol	0.0% ≤ **c.v.** ≥ 16.5%	The estimate can be used with no restriction
Marginal	E	16.6% ≤ **c.v.** ≥ 33.3%	The estimate must be used carefully as it is associated with a high level of error
Unacceptable	F	**c.v. ≥** 33.4 %	This information should not be disseminated. However, if the user chooses to do so, he must disseminate the information with the following warning: "We inform the user that … <specify the data> … does not meet Statistics Canada's quality standards. The conclusions drawn from this data would not be reliable."

Note: The sampling variability guidelines should be applied to rounded estimates. For more information, consult the publication *Statistics Canada Quality Guidelines* (No. 12-539-XIE in the Statistics Canada catalogue).

4.3 Rounding

To ensure that results published using the EDS microdata file will conform to established dissemination guidelines, the user is strongly advised to follow the rounding guidelines. Disseminating unrounded estimates could be misleading, since such estimates might appear to be more precise than they actually are.

4.3.1 Rounding Guidelines

1) Estimates of totals that appear in the body of a statistical table should be rounded to the nearest ten by the traditional rounding method (see definition in Section 4.3.2).

2) Partial and grand totals in statistical tables should be calculated from their unrounded components, then rounded to the nearest ten by the traditional rounding method.

3) Averages, proportions, rates and percentages should be calculated from unrounded components, then rounded to one decimal by the traditional rounding method.

4) Sums and differences of aggregates or ratios should be calculated from their corresponding unrounded components, then rounded to the nearest ten or the nearest decimal using the traditional rounding method.

5) Because of technical or other constraints, a rounding method other than traditional rounding may be used. In this case, the estimates obtained may differ from the corresponding estimates produced by

Statistics Canada. If so, the user is strongly advised to state the reason for these differences in the document disseminated.

4.3.2 Traditional Rounding Method

According to the traditional rounding method, if the first or only figure to be suppressed falls between 0 and 4, the last figure retained does not change. If the first or only figure to be suppressed falls between 5 and 9, the value of the last figure retained is increased by one unit (1).

For example, the figure 8,499 rounded to the nearest thousand would be 8,000, while the figure 8,500 rounded to the nearest thousand would be 9,000.

5.0 Other EDS data products

Additional information on the Ethnic Diversity Survey may be obtained from Statistics Canada website at www.statcan.ca.

Specifically, general survey information (such as that included in this User Guide) is available at:

http://www.statcan.ca/cgibin/imdb/p2SV.pl?Function=getSurvey&SDDS=4508&lang=en&db=IMDB&dbg=f&adm=8&dis=2

Selected analytical results from the survey are located at:

http://dissemination.statcan.ca/Daily/English/030929/d030929a.htm
http://www.statcan.ca:8096/bsolc/english/bsolc?catno=89-593-X&CHROPG=1

Users may also wish to apply for access to the EDS analytical file, which is a microdata file considerably more detailed than the EDS PUMF. The EDS analytical file includes all content from the survey (raw data and derived variables), including detailed geographic identifiers and some 2001 Census information for EDS respondents. Access to this file is only available from within Statistics Canada's Research Data Centres (RDCs) which are located at selected universities across Canada (for more information, please refer to the webpage http://www.statcan.ca/english/rdc/index.htm). Access to the EDS analytical file is granted through application to Social Sciences and Humanities Research Council using the application located at:

http://www.sshrc.ca/web/apply/program_descriptions/ciss_reseach_data_e.asp

Custom tabulations of EDS data are available from Statistics Canada at a price that reflects the resources required to produce them. To purchase custom tabulations or for additional information on the EDS PUMF or any other EDS products, please contact:

Client Services
Social and Aboriginal Statistics Division
Statistics Canada
Jean Talon Building, 7th floor
Tunney's Pasture
Ottawa, Ontario
K1A 0T6

Telephone: (613) 951-5979
Fax: (613) 951-0387
Email: sasd-dssea@statcan.ca