Ethnic Diversity Survey - Methodology and Data Quality

The information that follows presents a brief description of the methodology of the survey and of the main aspects of data quality. This information will help users to better understand the strengths and the limits of the data, as well as the best way to use them. This information will be particularly important in the comparison of the data to those of other surveys or other data sources.

Methodology

The Ethnic Diversity Survey (EDS) is a postcensal survey, which means that it uses the microdata of the 2001 Census to identify its target population and select its sample. This approach offers a number of advantages, the most important of which is that it permits the study of certain restricted and dispersed subpopulations, such as first and second generation Canadians, which otherwise would have been impossible or too costly to obtain.

Target Population

The target population for the survey consists of persons aged 15 years or older living in private households in the ten Canadian provinces. Canadian citizens, landed immigrants, holders of student, work or ministerial permits, and refugee status claimants all form part of the target population.

The following groups are excluded from the Ethnic Diversity Survey:

- persons under 15 years of age;
- persons living in collective dwellings;
- Indian reserves;
- persons declaring an Aboriginal origin or identity on the 2001 Census;
- persons residing in territories and remote areas.

These groups were excluded for various reasons. The survey did not permit proxy responses since it would be impossible to respond to questions of identity or attitude through a third person, and consequently persons under 15 years of age were excluded. Concerning collective dwellings, the principal reason for exclusion was that three-quarters of the persons living in collective dwellings are found in institutions (hospitals, prisons, treatment centres for handicapped persons, retirement homes, hostels for young offenders, etc.) and did not receive a long Census form (2B). In non-institutional collective dwellings, (campgrounds, rooming houses and lodging houses, school residences, ships, shelters, etc.) the population is highly mobile and good coverage would be very costly in a survey. Indian reserves and Aboriginals were excluded because of the response burden, since both these groups were covered by the Aboriginal Peoples Survey in 2001- 2002. Finally, remote areas, including the Yukon, Nunavut and the Northwest Territories, were excluded from the survey for operational and budgetary reasons.

Reference period

The reference period for the survey was the period of data collection, which took place between April and August 2002.

General methodology

The data for the Ethnic Diversity Survey were collected by the regional offices of Statistics Canada, using *Blaise* software and the computer-assisted telephone interview (CATI) method. The average length of an interview was 35 to 40 minutes, but this could vary with the respondent's situation. As mentioned earlier, proxy (or third person) responses were not permitted, so that only the person selected was traced and interviewed. In addition to the two official languages, interviews were conducted in seven non-official languages: Mandarin, Cantonese, Italian, Punjabi, Portuguese, Vietnamese and Spanish.

The data received from the interviews were checked to ensure the validity, consistency and completeness of the questionnaires. Wherever possible, automated controls were integrated into the collection mode to minimize errors and correct them with the respondent's assistance. Subsequently, potential errors were corrected with the help of notes made by the interviewers. Control and edit rules were developed to identify and correct inconsistencies for each question and each potential path within the questionnaire. In addition to these checks, missing responses to geographic (i.e., province of residence, census metropolitan area) and demographic (age, sex, marital status and relation to members of household) variables were imputed in a deterministic way.

Sample plan and selection

The sampling plan used for the survey was a two-phase stratified plan, based on the long questionnaires of the 2001 Census. Phase I, the 2001 Census, was to distribute the long questionnaire to one household in five in Canada. The population sampled in Phase II was selected on the basis of the responses given in the long form to questions on ethnic origin, place of birth, and place of birth of parents. Responses to the ethnic origin question were divided up to form the two main categories of interest: CBFA+ (Canadian or British or French or Americans or Australians and/or New Zealanders) and Non-CBFA+ (all other responses containing at least one origin other than CBFA+). The non-CBFA+ category was divided into European origins (for example, German, Italian, Dutch, Portuguese) and non-European (for example, Chinese, Jamaican, Lebanese, Iranian).

Finally, questions on the birthplace of respondents and their parents were used to establish the respondent's generational status. The first generation includes respondents born outside Canada. The second generation includes respondents born in Canada with at least one parent born outside Canada. The third-plus generation includes respondents born in Canada to two Canadian-born parents. Where necessary, the strata formed by generations were consolidated to obtain a sufficiently high number of persons in the stratum. Therefore, respondents to the EDS were distributed among the following strata:

CBFA+
Canadian only - Generation 1 and 2
Canadian only – Generation 3 and more
Canadian with BFA+ – Generations 1 and 2
Canadian with BFA+ – Generations 3 and more
BFA+ – Generation 1 and 2
BFA+ – Generation 3 and more

Non-CBFA+
Other Europeans with Canadian – Generation 1 and 2
Other Europeans with Canadian – Generation 3 and more
Other Europeans – Generation 1
Other Europeans – Generation 2
Other Europeans – Generation 3 and more
Other non-Europeans with Canadian – All generations
Other non-Europeans – Generation 1
Other non-Europeans – Generation 2 and more

At the time the sample was selected, imputed Census data were unavailable. Thus, responses to the variables used to stratify the survey frame were not always at hand. Therefore an additional stratum was created for persons who did not respond to one or another of these questions to give everyone in the target population the opportunity to be selected in the sample.

In light of the goals of the survey and the data requirements for certain subpopulations, the sample distribution was established at one-third for CBFA+ and two-thirds for non-CBFA+. This distribution ensured that a sufficient number of persons would be obtained in the categories of interest, especially among the foreign-born.

In each stratum, sample size was determined by the anticipated response rate, of a minimal proportion to be estimated with a certain coefficient of variation (c.v.). A pilot test conducted in September 2001 made it possible to fine-tune some of the assumptions used in the distribution of the size of the sample, particularly those concerning the response rate in each stratum and the changes in responses relative to the Census.

Once constructed, the survey frame was ordered by province, electoral district, enumeration area and household to ensure a good geographic distribution of the sample and reduce the number of persons of the same household selected for an interview. (It should be noted that one of the sampling constraints was to avoid, as far as possible, conducting more than one interview per household.) A systematic sample was then selected independently in each stratum. The final sample included 57,242 persons. Of that number, 42,476 responded to the survey. This corresponds to an overall response rate of 75.6% if the 1,057 persons classified as being outside the scope of the survey were taken into account.

Weighting and estimation

Since the Ethnic Diversity Survey is a survey based on a probability sampling plan, each respondent represents a certain number of other persons of the population who are not part of the sample. This number of persons is known as the weight.

The weight adjustment of the EDS was performed in three steps. The first step was the calculation of the initial weight that was obtained by multiplying the weight of the respondent in the 2001 Census by the sampling weight. The latter corresponded to the reciprocal of the person's probability of selection.

The second step was the adjustment due to non-response. This step consisted of applying a correction factor to the initial weights to compensate for the effects of non-response. The "response propensity model" method consisted of using a predicted probability of response obtained with the aid of a logistical regression model.

The third step was *a posteriori* stratification, known also as post-stratification. The aim of this method was to ensure that the sum of the weights of respondents corresponded to 2001 Census tabulations for each variable used. This adjustment was defined by the crossing of certain variables (region, stratum, age group and sex).

The raking ratio estimation method was also used. This method guarantees that the final estimates based on the sample agree simultaneously with the known distributions of a certain number of variables. Adjustment of weights was performed separately for religion, generation and mother tongue. Data on visible minorities were crossed-classified by regions.

The final weight assigned to each respondent underwent numerous adjustments so as to represent the target population as correctly as possible. Weighting of the data ensured that the sample of the EDS is representative of the target population, even if the sampling ratio differs widely from one stratum to the other. As a result, use of the weights is essential for any analysis of the survey data.

Estimation of variance

The Ethnic Diversity Survey used the bootstrap method to estimate variance. This re-sampling method drew 500 independent samples (with replacement) based on the initial sample, corrected as per the survey sampling plan.

Initial weights were calculated for each of the samples and adjustments to the weights performed using the same steps as for the initial sample (non-response, post-stratification and raking ratio estimation). In this way, the component of variance generated at each step of the weighting could be taken into account. Estimators of interest were then calculated for each sample. Empirical variance was calculated on all the samples to produce an estimate of the variance of the estimators in question.

The Ethnic Diversity Survey uses the coefficient of variation (c.v.) as a quality indicator measurement. WesVar software, developed by Westat, was used as a tool to calculate coefficients of variation.

Data quality

Despite every effort to ensure that the data obtained are of superior quality, errors can be produced at practically every stage of the survey process. These errors divide into two major categories: sampling errors, and errors not due to sampling.

Sampling error comes from the fact that the estimates were obtained from a sample, rather than from a census of the entire population performed under the same conditions. In the Ethnic Diversity Survey, this error was measured using the c.v. This number, expressed as a percentage, corresponds to the standard error (or square root of the variance of the estimator) divided by the estimator itself. The smaller the c.v., the smaller the variability of the sample and the more accurate the estimators. The EDS uses the following measurements:

(i) when the c.v. is greater or equal to 33.4 %, the estimator is considered "unacceptable" and the symbol "F" appears beside the corresponding estimator;

(ii) when the c.v. falls between 16.6 % and 33.3 %, the estimator is considered "poor" and the symbol "E" for caution appears beside the estimator;

(iii) when the c.v. is 16.5 % or less, the estimator is considered "acceptable", it can be used without restriction and no indication appears beside it.

Other types of error are not due to sampling and may arise at any stage of a survey. This type of error includes primarily errors in coverage, non-response, response and processing. Response errors occur when the respondents, and in some cases the interviewers, misinterpret a question and an inexact answer is entered. Measures have been adopted to reduce this type of error. Interviewers received specialised training on all subject matter contained in the survey.

As far as possible, standard questions taken from other Statistics Canada surveys were used in the questionnaire. In developing the new content for the survey, an advisory committee provided guidance in the form of discussions and recommendations. As well, content development took into consideration the results of a series of qualitative studies. A pilot test was carried out in September 2001 and other quality studies were performed before the final questionnaire was prepared.

As for processing errors, these can occur during coding, when the responses in letters are converted into numerical codes, or during any other manipulation of the data. Other types of non-sampling errors, such as non-response and coverage errors, are easier to identify and to quantify.

Any survey is subject to a certain percentage of non-response. Total non-response is when a person cannot be interviewed at all. Partial non-response is when only a part of the questionnaire is completed. Non-response errors depend on the type and importance of differences likely to exist between the characteristics of respondents and non-respondents. As a rule, the more marked these differences, the more they affect the precision of estimates.

Measures were taken at each stage of the survey to minimize non-response error in order to reduce the bias attributable to non-response. To increase the rate of participation, a letter providing information on the survey and requesting the participation of the person selected was sent before collection was started. A tracing operation, performed by experienced interviewers, located and contacted respondents who could not be reached at the beginning of the survey at the telephone numbers originally provided. In the regional offices, senior interviewers were tasked with converting refusals or difficult cases into complete responses. Efforts were continually made throughout the survey to reach response objectives, not only in overall terms but also, as far as possible, within each stratum.

Objectives set were achieved and in some cases exceeded by each regional office. As indicated earlier, the total response to the survey was 75.6%. Rates of response per stratum, as defined earlier, ranged from 72% to 80%. As might be expected, those of the first generation posted the lowest response rate, 73% compared to 77% for the second generation and higher. Partial non-response accounted for only 3.2% of responses, which means that, generally, when a person began the interview, all the questions were answered. Complete responses were thus 96.8% of the total.

A coverage error occurs when there is a variance between the target population and the sampled population, such as when a person in the survey frame was missed in the survey frame, mistakenly included, or counted twice. Using the Census as a survey frame helped reduce this kind of error. (The net under-coverage error of the Census is around 3%.)

It should also be noted that in selecting the sample for the EDS, a coding error slipped into the survey frame for some census subdivisions. This problem mainly affects two Atlantic provinces, Nova Scotia and New Brunswick. A detailed study of the characteristics of persons not covered, using census data, shows that the sample remained representative of the target population for the majority of the survey variables of interest. Adjustments similar to those used for correction of non-response were made to the weights to reduce potential bias due to this error in coverage. Although the bias created could not be completely eliminated, the quality of the data at the national level was not affected. However, all identification of provinces in the Atlantic Region was removed from the final database. The lowest geographical level for the Atlantic Region is the indicator of Census Metropolitan Area (CMA) and the non-CMAs.

Confidentiality and rounding

The purpose of rounding the results of survey data is first to establish a degree of consistency with the level of accuracy of the sampling plan. It also serves to protect the confidentiality of the information supplied.

For the EDS, the figures presented in tables will be rounded off to the nearest thousand. This method rounds all figures, including totals, up or down to the nearest multiple of 10. This provides a certain protection against the possibility that the data will be associated with a readily recognizable individual. Since the totals are rounded separately, they do not necessarily report the sum of the values in each category. Similarly, the sum of percentages, which are calculated from the rounded data, is not necessarily exactly 100 percent.

Finally, all the data based on 10 or fewer individuals have been suppressed to ensure even greater confidentiality of the responses of the persons interviewed. In all cases, however, the suppressed data have been included in the totals.