9 GUIDELINES FOR RELEASE (DATA QUALITY)

Microdata users should apply the rules for assessing data quality, below, to all estimates they produce, and retain only those that satisfy the release criteria. Estimates that do not satisfy the release criteria are not reliable.

Introduction

The guidelines for release and publication make use of the concept of *sampling variability* to determine whether estimates obtained from the microdata files are reliable. Sampling variability is the error in the estimates caused by the fact that we survey a sample rather than the entire population. The concept of *standard error* and the related concept of *coefficient of variation* and *confidence interval* provide an indication of the magnitude of the sampling variability.

The standard error and coefficient of variation do not measure any systematic biases in the survey data which might affect the estimate. Rather, they are based on the assumption that the sampling errors follow a normal probability distribution.

Subject to this assumption, it is possible to estimate the extent to which different samples that have the same design and the same number of observations would give different results. This indicates the margin of error that is likely to be included in the estimates derived from our single sample.

For a detailed description of the measures of sampling variability, see A. Satin and W. Shastry, *Survey Sampling: A Non-Mathematical Guide*, Statistics Canada, Catalogue Number 12-602E.

Survey errors

The survey produces estimates based on information collected from and about a sample of individuals. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used in the survey. The difference between the estimates obtained from the sample and those resulting from a complete count taken under similar conditions is called the sampling error of the estimate.

Errors which are not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered on the computer 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 are made to reduce non-sampling errors in the survey. Quality assurance measures are implemented at each step of the data collection and processing cycle to monitor the quality of the data. These measures include the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions, procedures to ensure that data capture errors are minimized and coding and edit quality checks to verify the processing logic.

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Total non-response occurs because the interviewer is either unable to contact the respondent, no member of the household is able to provide the information, or the respondent refuses to participate in the survey. Total non-response is handled by adjusting the weight of households who respond to the survey to compensate for those who do not respond (refer to Chapter 6).

In most cases, partial non-response to the survey occurs when the respondent does not understand or misinterprets a question, refuses to answer a question, cannot recall the requested information, or cannot provide proxy information.

Since it is an unavoidable fact that estimates from a sample survey are subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error. This chapter of the documentation outlines the measures of sampling error which Statistics Canada commonly uses and which it urges users producing estimates from these microdata files to use also.

The basis for measuring the potential size of sampling errors is the standard error of the estimates derived from survey results.

Hypothesis tests provided by statistical software packages

The Canadian Travel Survey is based upon a complex design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures differ from that which is appropriate in a sample survey framework, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, analysis of variance), a method exists which can make the application of standard packages more meaningful. If the weights on the records are rescaled so that the average weight is one (1), then the results produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by dividing each weight by the overall average weight before the analysis is conducted.

In order to provide a means of assessing the quality of tabulated estimates, Statistics Canada has produced a set of Approximate Sampling Variability Tables (commonly referred to as "C.V. Tables") for the Canadian Travel Survey. These tables can be used to obtain approximate coefficients of variation for categorical-type estimates and proportions. Refer to Chapter 10 for more details.

Minimum sizes of estimates for release

Before releasing and/or publishing any estimate from these microdata files, users should first determine the number of respondents who contribute to the calculation of the estimate. If this number is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the rounded estimate and follow the guidelines given in Table 5.

When the unweighted estimate is satisfactory, the user should look at the weighted estimate to see if it satisfies the release criteria. The release cutoffs for the weighted estimates of the CTS can be found in the Data Quality Report (Chapter 10).

Using the coefficient of variation (CV)

The standard error of an estimate is frequently expressed as a percentage of the estimate itself, in which case it is called the coefficient of variation. Whereas the standard error is measured in the same units as the estimate, the coefficient of variation is simply a ratio. This makes it easier to use as a criterion for the reliability of any estimate.

For example, suppose that, based upon the survey results, one estimates that 25.9% of Canadians aged 15 or more took at least one trip in March 1996, and this estimate is found to have standard error of 0.009. Then the coefficient of variation of the estimate is calculated as:

$$\left(\frac{.009}{.259}\right) x \ 100\% \ = \ 3.47\%$$

The coefficients of variation (C.V.) are derived using the variance formula for simple random sampling and incorporating a factor which reflects the multi-stage, clustered nature of the sample design. This factor, known as the design effect, was determined by first calculating design effects for a wide range of characteristics and then choosing from among these a conservative value to be used in the look-up tables which would then apply to the entire set of characteristics.

To have more information on the design effects, sample sizes and population counts by province that were used to produce the Approximate Sampling Variability Tables for person-weights, refer to the Data Quality Report (Chapter 10).

Note that Approximate Sampling Variability Tables are also available for person-trip weights, household-trip weights, person-night weights and expenditures weights.

In the Data Quality Report provided with this Guide, a set of Approximate Sampling Variability Tables has been provided to give microdata users some approximate coefficients of variation for groups of estimates at a time, such as all estimates pertaining to a particular province. In most cases, these will be adequate to determine whether an

TABLE 5. Acceptable levels of the coefficient of variation

Approximate coefficient of variation (%)	**Restriction on use**
0.0 - 16.5	ACCEPTABLE. Estimates can be considered for general unrestricted release.
16.6 - 25.0	MARGINAL. Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates.
25.1 - 33.3	CONFIDENTIAL. Estimates can be considered for general unrestricted release only when exact coefficients of variation have been calculated on a cost-recovery basis and are found to be acceptable. Otherwise, such estimates should not be used or released.
33.4 or greater	UNACCEPTABLE. Estimates should not be used or released under any circumstances.

estimate can be released. The Data Quality Report (Chapter 10) explains how to obtain the approximate CV from the tables, depending on whether the estimate is a simple population count or a percentage, difference or ratio of population subgroups. In the case of numeric totals or means, the CV is generally larger than the CV of the population count on which it is based.

All coefficients of variation in the Approximate Sampling Variability Tables are approximate and, therefore, unofficial. Estimates of actual variance for specific variables may be obtained from Statistics Canada on a cost-recovery basis. The use of actual variance estimates would allow users to release otherwise unreleaseable estimates, i.e., estimates with coefficients of variation in the 'confidential' range.

Remember: if the number of observations on which an estimate is based is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. This is because the formulas used for estimating the variance do not hold true for small sample sizes.