Section 7

Data quality

7.1 Non-sampling errors

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 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. 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, edits to ensure that data entry errors are minimized and coding and edit quality checks to verify the processing logic.

7.2 Sampling errors

The Labour Force Survey collects information from a sample of households. Somewhat different figures might have been obtained if a complete census had been taken using the same questionnaires, interviewers, supervisors, processing methods, etc. as those actually used in the Labour Force Survey. The difference between the estimates obtained from the sample and those that would give a complete count taken under similar conditions is called the sampling error of the estimate, or sampling variability. Approximate measures of sampling error accompany Labour Force Survey products and users are urged to make use of them while analysing the data.

Three interpretation methods can be used to evaluate the precision of the estimates: the standard error, and two other methods also based on standard error: confidence intervals and coefficients of variation.

7.2.1 Interpretation using standard error

The sampling error, or standard error, is a measure that quantifies how different the sample estimate might be from the census value. It is based on the idea of selecting several samples, although in a survey only one sample is drawn and information is collected on units in that sample. Using the same sampling plan, if a large number of samples were to be drawn from the same population, then about 68% of the samples would produce a sample estimate that is within one standard error of the census value and in about 95% of the samples it will be within two standard errors of the census value.

When looking at changes, for example month to month changes, two thirds of the time (68%) a change greater than the sampling error indicates a real change. The larger the change compared to the standard error, the better the chance that we are observing a real change, as opposed to a change due to sampling variability. At the 95% confidence level, the change in the estimate must be greater than twice the sampling error in order to ensure that change is real.

Movements in estimates that are smaller than the sampling error are less likely to reflect a real change and more likely to be due to sampling variability. While the above is true for monthly movements, one can have more confidence in a series of consecutive movements in the same direction, even though some of the monthly movements may be smaller than the sampling error.

To illustrate, let us say that between two months the published estimate for total employment increases by 40,000, and that the associated standard error for the movement estimate is 27,200. Since the increase is larger than the standard error, the probability is at least two out of three (68%) that the increase of 40,000 in employment is a real change. To reach a 95%

confidence level, the standard error has to be doubled. Because the increase of 40,000 in employment is smaller than twice the standard error (54,400), it is impossible to state with a 95% confidence level that there was an increase in employment.

7.2.2 Interpretation using confidence intervals

Confidence intervals provide another way of looking at the variability inherent in estimates of sample surveys. To illustrate how to calculate the confidence interval, let us say that one month the published estimate for total employment rose by 16,000 to reach 16,500,000. The associated standard error for the movement estimate is 27,200. Using the standard error to build the confidence intervals, we can say that:

- There are approximately two chances in three (68%) that the real value of the movement between the two months falls within the range -11,200 to +43,200 (16,000 + or – one standard error).
- There are approximately nine chances in ten (90%) that the real value of the movement between the two months falls within the range -27,520 to +59,520 (16,000 + or – 1.6 times the standard error).
- There are approximately nineteen chances in twenty (95%) that the real value of the movement between the two months falls within the range -38,400 to +70,400 (16,000 + or – two standard errors).

7.2.3 Interpretation using coefficient of variation

Sampling variability can also be expressed relative to the estimate itself. The standard error as a percentage of the estimate is called the coefficient of variation (CV) or the relative standard error. The CV is used to give an indication of the uncertainty associated with the estimates. For example, if the CV is 7%, then in 68% of the samples the census value will lie within plus or minus 7% or one CV and in 95% of the samples the census value will lay within plus or minus 14% or two times the CV of the estimate.

Small CV's are desirable because they indicate that the sampling variability is small relative to the estimate. The CV depends on the size of the estimates, the sample size the estimate is based on, the distribution of the sample, and the use of auxiliary information in the estimation procedure. The size of the estimates is important because the CV is the sampling error expressed as a percentage of the estimate. The smaller the estimate the larger the CV (all other things being equal). For example, when the unemployment rate is high the CV may be small. If the unemployment rate falls due to improved economic conditions then the corresponding CV will become larger. Typically, of similar estimates, the one with largest sample size will yield the smaller CV. This is because the sampling error is smaller.

Also, estimates referring to characteristics that are more clustered will have a higher CV. For example, persons employed in forestry, fishing, mining, oil and gas in Canada are more clustered geographically than employed women aged 55 to 64 years in Ontario. The latter will have a smaller sampling variability although the estimates are of approximately the same size.

Finally estimates referring to age and sex are usually more reliable than other similar estimates because the LFS sample is calibrated to post-censal population projections of various age and sex groupings. Continuing the previous example, persons employed part-time in Alberta will have a larger sampling variability than employed men aged 35 to 44 years in British Columbia although the estimates are of similar size.

7.2.4 Variability of monthly estimates

To look up an approximate measure of the CV of an estimate of a monthly total, please consult table 7.1, which gives the size of the estimate as a function of the geography and the CV. The rows give the geographic area of the estimate while the columns indicate the resulting level of accuracy in terms of the CV, given the size of the estimate. To determine the CV for an estimate of size X in an area A, look across the row for area A, find the first estimate that is less than or equal to X. Then the title of the column will give the approximate CV. For example, to determine the sampling error for an estimate of 36.0 thousand unemployed in Newfoundland and Labrador in August 2010, we find the closest but smaller estimate of 25.7 thousand giving a CV of 5%. Therefore, the estimate of 36,000 unemployed in Newfoundland and Labrador has a CV of roughly 5%.

Table 7.1 is supplied as a rough guide to the sampling variability. The sampling variability is modeled so that, given an estimate, approximately 75% of the actual CVs will be less than or equal to the CVs derived from the

table. There will, however, be 25% of the actual CVs that will be somewhat higher than the ones given by the table.

Table 7.1 can also be used with either seasonally adjusted estimates, or with estimates that have not been seasonally adjusted. Studies have shown that LFS standard errors for seasonally adjusted data are close to those for unadjusted data.

The CV values given in table 7.1 are derived from models based on LFS sample data for the 47-month period from January 2007 through November 2010 inclusive. It is important to bear in mind that these values are approximations.

7.2.5 Variability of annual estimates

To look up an approximate measure of the CV of an estimate of an annual average, please consult table 7.2, which gives the size of the estimate as a function of the geography and the CV. The rows give the geographic level of the estimate while the columns indicate the resulting level of accuracy in terms of the CV, given the size of the estimate. To determine the CV for an estimate of size X in an area A, look across the row for area A, find the first estimate that is less than or equal to X. Then the title of the column will give the approximate CV. For example, to determine the sampling error for an annual average estimate of 51.2 thousand unemployed in Newfoundland and Labrador in 2010, we find the closest but smaller estimate of 16.3 thousand giving a CV of 2.5%. Therefore, the estimate of 51,200 unemployed in Newfoundland and Labrador has a CV of roughly 2.5%.

Table 7.2 is supplied as a rough guide to the sampling variability. The sampling variability is modeled so that, given an estimate, approximately 75% of the actual CVs will be less than or equal to the CVs derived from the table. There will, however, be 25% of the actual CVs that will be somewhat higher than the ones given by the table.

The CV values given in table 7.2 are derived from a model based on LFS sample data for the 5-year period from December 2005 through November 2010 inclusive. It is important to bear in mind that these values are approximations.

7.2.6 Sampling variability tables for the territories

The CV values given in table 7.3 for the Yukon and Northwest Territories are derived from models based on LFS sample data for the 48-month period from December 2006 through November 2010 inclusive. For Nunavut, they are based on LFS sample data for the 35-month period from January 2008 through November 2010 inclusive.

Table 7.1
CVs for estimates of monthly totals for Canada and the provinces

	Coefficient of variation								
	1.0%	2.5%	5.0%	7.5%	10.0%	15.0%	20.0%	25.0%	30.0%
Canada	1,137.7	339.8	150.4	88.1	50.4	29.6	19.8	14.4	11.1
Newfoundland and Labrador	243.2	65.4	25.7	14.3	8.4	4.7	3.0	2.2	1.6
Prince Edward Island	70.6	20.7	8.8	5.1	3.0	1.7	1.2	0.8	0.6
Nova Scotia	244.0	71.9	31.0	18.0	10.5	6.1	4.1	3.0	2.3
New Brunswick	205.4	60.2	25.8	15.0	8.7	5.0	3.4	2.4	1.9
Quebec	1,038.5	302.8	130.2	75.3	43.5	25.1	16.7	12.1	9.3
Ontario	1,152.3	333.5	143.2	82.6	47.3	27.3	18.1	13.1	10.0
Manitoba	185.6	57.1	26.5	15.8	8.9	5.3	3.6	2.6	2.0
Saskatchewan	164.9	50.1	22.9	13.6	7.6	4.5	3.0	2.2	1.7
Alberta	519.9	160.7	75.5	45.3	25.0	15.0	10.2	7.5	5.8
British Columbia	737.8	213.7	92.6	53.6	30.2	17.5	11.6	8.4	6.4

Note(s): Estimates are in thousands.

Table 7.2
CVs for estimates of annual averages for Canada and the provinces

	Coefficient of variation								
	1.0%	2.5%	5.0%	7.5%	10.0%	15.0%	20.0%	25.0%	30.0%
Canada	438.1	123.7	54.4	31.3	16.7	9.6	6.3	4.5	3.4
Newfoundland and Labrador	55.7	16.3	7.8	4.6	2.3	1.4	0.9	0.7	0.5
Prince Edward Island	19.3	6.0	3.0	1.8	1.0	0.6	0.4	0.3	0.2
Nova Scotia	69.8	21.3	10.4	6.3	3.2	2.0	1.3	1.0	0.7
New Brunswick	58.4	18.0	8.8	5.4	2.8	1.7	1.1	0.8	0.6
Quebec	314.9	94.7	45.3	27.2	14.1	8.5	5.7	4.2	3.2
Ontario	407.6	115.7	51.8	30.0	15.7	9.1	6.0	4.3	3.3
Manitoba	70.3	21.0	10.2	6.1	3.1	1.9	1.3	0.9	0.7
Saskatchewan	60.8	17.6	8.3	4.9	2.5	1.5	1.0	0.7	0.5
Alberta	201.6	59.3	28.5	17.0	8.5	5.1	3.4	2.5	1.9
British Columbia	212.2	63.6	30.8	18.6	9.4	5.7	3.8	2.8	2.1

Note(s): Estimates are in thousands.

Table 7.3
CVs for estimates for the territories, 3-month moving average and annual averages

	Coefficient of variation								
	2.0%	3.5%	5.0%	7.5%	10.0%	15.0%	20.0%	25.0%	30.0%
3 month moving averages									
Yukon	21.8	9.6	4.2	2.2	1.0	0.5	0.3	0.2	0.2
Northwest Territories	27.8	12.4	5.1	2.7	1.1	0.6	0.4	0.2	0.2
Nunavut	11.1	5.6	2.6	1.5	0.7	0.4	0.3	0.2	0.1
Annual averages									
Yukon	24.6	11.9	5.6	3.1	1.5	0.8	0.5	0.4	0.3
Northwest Territories	31.3	15.0	6.8	3.8	1.7	1.0	0.6	0.4	0.3
Nunavut	15.8	8.1	3.9	2.3	1.1	0.7	0.4	0.3	0.2

Note(s): Estimates are in thousands.

For more accurate measures of variability, please contact Client Services at 1 866 873-8788 or e-mail us at *labour@statcan.gc.ca*.

7.2.7 Variability of rates

Estimates that are rates and percentages are subject to sampling variability that is related to the variability of the numerator and the denominator of the ratio. The various rates given are treated differently because some of the denominators are calibrated figures that have no sampling variability associated with them.

7.2.8 Unemployment rate

The unemployment rate is the ratio of X, the total number of unemployed in a group, to Y, which is the total number of participants in the labour force in the same group. Here the group may be a province or CMA and/or it may be an age-sex group. For example, in September 2009, there were approximately 39,100 unemployed persons in Newfoundland and Labrador and 252,300 participants in the labour force, giving an unemployment rate of 15.5%.

The CV for the unemployment rate can be estimated with the following formula:

$$[CV(X/Y)]^2 = [CV(X)]^2 + [CV(Y)]^2 - 2p[CV(X)]\,[CV(Y)]$$

where CV(X) would be the CV for the total number of unemployed in a specific geographic or demographic subgroup and CV(Y) would be the CV for the total number of participants in the labour force in the same subgroup. The correlation coefficient, denoted *p*, mesures the amount of linear association between X and Y (respectively, the number of unemployed and the number of participants in the labour force in the same subgroup). The value of *p* ranges between -1 and 1. For example, a strong positive linear association would indicate that unemployment counts generally increases as the total number of participants in the labour force increases. Note that we can expect a larger CV for the unemployment rate when *p* is negative since in this

case, the third term on the right side of the equation above becomes positive.

When *p* is not available, the most conservative approach is to take *p* = -1, which leads to the simplified formula:

CV(X/Y) = CV(X) + CV(Y)

Note that this will likely lead to an overestimation of the CV(X/Y).

In the previous example, the CVs of the monthly estimates for the unemployment count and the total number of participants in the labour force in Newfoundland and Labrador are respectively 5% and 2.5% from Table 7.1. An approximation of the CV for the unemployment rate of 15.5% using the above formula would be:

5.0% + 2.5% = 7.5%

7.2.9 Participation rate and employment rate

The participation rate represents the number of persons in the labour force expressed as a percentage of the total population size. The employment rate is the total number of employed divided by the total population size. For both the above rates, the numerator and the denominator represent the same geographic and demographic group.

For Canada, the provinces, CMAs and some age-sex groups the LFS population estimates are not subject to sampling variability because they are calibrated to independent sources. Therefore, in the case of the participation rate and the employment rate of these geographic and demographic groups, the CV is equal to that of the contributing numerator.

Subgroups of Canada, the provinces and age-sex groups are called domains; for example, persons employed in agriculture in Manitoba are a domain. To determine the CV of rates in the case of domains, the variability of both the numerator and the denominator have to be taken into account because the denominator is no longer a controlled total and is subject to sampling variability. Therefore, for participation rates and employment rates of domains, the CV can be determined similar to the unemployment rate. The totals in the numerator and denominator for the relevant rate should reflect the same domain or subgroup.

7.2.10 Variability of estimate of change

The difference of estimates from two time periods gives an estimate of change that is also subject to sampling variability. An estimate of year-to-year or month-to-month change is based on two samples which may have some households in common. Hence, the CV of change depends on the CV of the estimates for both periods and the sample overlap, *p*, between the periods. The following formula can be used to approximate the estimate of change:

$$CV(Y_2 - Y_1) = \sqrt{1-\rho}\,\frac{\sqrt{Y_1^2 CV(Y_1)^2 + Y_2^2 CV(Y_2)^2}}{(Y_2 - Y_1)}$$

where Y_1 and Y_2 are the estimates for the two periods. The value of *p* is the correlation coefficient between Y_1 and Y_2. The value of *p* ranges between -1 and 1, with 1 being the perfect positive linear association. One can generally use the sample overlap to approximate the correlation coefficient as follows:

- For the provinces: use *p* = 5/6 for month-to-month changes, and *p* = 0 for year-to-year changes.

Empirical studies at Statistics Canada have shown that for the provinces, a *p* value equal to 5/6 is a good approximation for estimates of employment, but for estimates of unemployment, a *p* value of 0.45 would yield a better approximation for month-to-month changes.

When comparing the annual averages of two years, the CV of the annual estimates (table 7.2) should be used. For month-to-month change, seasonally adjusted estimates should be used in conjunction with the CVs of the monthly estimates from table 7.1. Note that the above formula gives approximate estimates of the sampling variability associated to an estimate of change.

7.2.11 Guidelines on data reliability

Household surveys within Statistics Canada generally use the following guidelines and reliability categories in interpreting CV values for data accuracy and in the dissemination of statistical information.

Category 1 - If the CV is ≤ 16.5% - no release restrictions: data are of sufficient accuracy that no special warnings to users or other restrictions are required.

Category 2 - If the CV is > 16.5% and ≤ 33.3% - release with caveats: data are potentially useful for some purposes but should be accompanied by a warning to users regarding their accuracy.

Category 3 - If the CV > 33.3% - not recommended for release: data contain a level of error that makes them so potentially misleading that they should not be released in most circumstances. If users insist on inclusion of Category 3 data in a non-standard product, even after being advised of their accuracy, the data should be accompanied by a disclaimer. The user should acknowledge the warnings given and undertake not to disseminate, present or report the data, directly or indirectly, without this disclaimer.

7.3 Release criteria

Statistics Canada is prohibited by law from releasing any data which would divulge information obtained under the Statistics Act that relates to any identifiable person, business or organization without the prior knowledge or the consent in writing of that person, business or organization. Various confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.

The LFS produces a wide range of outputs that contain estimates for various labour force characteristics. Most of these outputs are estimates in the form of tabular cross-classifications. Estimates are rounded to the nearest hundred and a series of suppression rules are used so that any estimate below a minimum level is not released.

The LFS suppresses estimates below the levels presented in the table 7.4.

Table 7.4
Minimum size for release, Canada, provinces and territories

	Minimum size for release
	thousands
Canada	**1.5**
Newfoundland and Labrador	0.5
Prince Edward Island	0.2
Nova Scotia	0.5
New Brunswick	0.5
Quebec	1.5
Ontario	1.5
Manitoba	0.5
Saskatchewan	0.5
Alberta	1.5
British Columbia	1.5
Yukon	0.2
Northwest Territories	0.2
Nunavut	0.2