10 DATA QUALITY REPORT

Introduction

The coefficient of variation of a statistic is largely a product of the total survey sample size and the importance of the sub-population in the total Canadian population. It also depends on the level of non-response and the particular sample design. For general purposes, this report includes a table on the sample sizes per province and a table of monthly response rates by province.

This chapter indicates how to obtain the approximate coefficient of variation for a statistic from the Approximate Sampling Variability Tables for the Canadian Travel Survey.

Sample size

The following table shows the number of household members, in the LFS sampled rotations who were eligible for the Canadian Travel Survey supplement.

TABLE 6. Monthly sample sizes by province, 1996

	Jan	Feb	Mar	Apr	May	Jun
Nfld	577	590	585	879	898	888
P.E.I.	462	416	440	451	439	460
N.S.	1110	1046	1072	1095	1100	1117
N.B.	997	1004	990	1018	1010	1032
Que	3217	3278	3328	3326	3313	3282
Ont	4861	5019	4960	5071	5174	5131
Man	1121	1185	1149	1219	1158	1171
Sask	1055	1059	1054	1022	1079	1016
Alta	1222	1218	1190	1241	1292	1233
B.C.	1506	1441	1494	1528	1530	1499
Canada	16128	16256	16262	16850	16993	16829

TABLE 6. (continued) ***Monthly sample sizes by province, 1996***

	Jul	Aug	Sep	Oct	Nov	Dec
Nfld	887	924	930	949	937	918
P.E.I.	455	448	463	425	442	468
N.S.	1124	1148	1132	1106	1153	1134
N.B.	1042	1046	1076	1079	1051	1061
Que	3391	3475	3588	3466	3433	3502
Ont	5209	5156	5213	5145	5109	5109
Man	1242	1200	1198	1256	1194	1212
Sask	1026	1093	1087	1025	1100	1088
Alta	1282	1328	1240	1309	1279	1220
B.C.	1595	1607	1552	1612	1587	1536
Canada	17253	17425	17479	17372	17285	17248

Response rates

The following tables summarize the response rates to the 1996 Canadian Travel Survey. The response rates shown in these table reflect the proportion of people eligible for the Canadian Travel Survey who have reported information. These response rates are not cumulative, that is, they don't take into account those people who would have been eligible for CTS but have been non-respondents to LFS. This is because those individuals who don't respond to the LFS are not even asked if they would like to answer the Canadian Travel Survey. Thus they cannot be considered as non-respondents to the CTS.

TABLE 7. Monthly response rates by province, per cent, 1996

	Jan	Feb	Mar	Apr	May	Jun
Nfld	94.1	95.4	92.0	92.0	91.1	88.0
P.E.I.	94.4	93.3	94.0	92.2	94.3	92.0
N.S.	95.0	95.0	92.0	93.0	92.0	92.0
N.B.	94.0	92.4	92.0	94.0	91.0	89.1
Que	95.0	94.0	94.1	93.0	93.0	92.0
Ont	93.0	91.0	91.2	90.3	90.3	88.4
Man	91.0	91.0	90.3	93.0	91.1	88.0
Sask	91.0	89.2	89.3	89.0	87.0	85.3
Alta	94.0	93.0	93.0	92.3	90.0	89.0
B.C.	93.1	91.0	89.0	91.0	90.3	87.0
Canada	93.4	92.1	92.0	92.0	91.0	89.0

TABLE 7. (continued) ***Monthly response rates by province, per cent, 1996***

	Jul	Aug	Sep	Oct	Nov	Dec
Nfld	90.0	93.0	91.2	94.3	93.0	94.2
P.E.I.	92.3	94.0	96.0	94.4	96.2	96.4
N.S.	92.0	92.2	92.3	93.3	93.1	94.0
N.B.	91.0	90.0	92.0	92.0	89.2	91.0
Que	93.0	94.0	94.3	93.4	94.0	94.0
Ont	89.4	90.0	90.4	91.1	91.0	90.0
Man	88.2	88.2	92.0	92.1	91.2	90.0
Sask	88.0	87.4	89.1	90.2	88.4	87.2
Alta	90.4	88.0	92.2	93.0	89.0	90.2
B.C.	89.2	90.0	91.0	91.4	90.3	90.4
Canada	90.3	91.0	92.0	92.2	91.4	91.1

Design effect

The next table shows the design effects, sample sizes and population counts by province which were used to produce the Approximate Sampling Variability Tables for person-weights. Note that although the CTS contains different sample and population sizes for each month, the design effects remain constant throughout the months. For this reason, the design effects, sample size, and population size for only one month are presented.

TABLE 8. Design effects, March 1996

Province	Design effect	Sample size	Population
Nfld	2.0	536	453272
P.E.I.	2.0	412	105966
N.S.	2.0	986	734177
N.B.	2.0	906	599604
Que	2.0	3133	5858168
Ont	2.0	4523	8820143
Man	2.5	1037	853597
Sask	2.0	941	751563
Alta	2.5	1103	2097529
B.C.	2.5	1330	3008863
Atlantic provinces	2.0	2840	1893019
Man & Sask	2.5	1978	1605160
Alta & B.C.	2.5	2433	5106392
Canada	3.0	14907	23282882

Release cut-off's for the CTS

The minimum size of the estimate (using person weights) at the provincial, regional and Canada levels are specified in the table below. Estimates smaller than the minimum size given in the Unacceptable column may not be released under any circumstances. Note that only one table of release cut-offs is presented below. This table

represents the release cut-offs for March 1996. More tables could be made available for other months when certain provinces used more than two LFS rotation groups for the sample.

TABLE 9. Sample Table of Release Cut-offs, March 1996

Province	Acceptable	Marginal	Confi-dential	Unaccep-table
Nfld	54,500 +	25,500-54,499	15,000-25,499	under 15,000
P.E.I.	16,000 +	7,500-15,999	4,500-7,499	under 4,500
N.S.	51,000 +	23,000-50,999	13,000-22,999	under 13,000
N.B.	45,000 +	20,500-44,999	11,500-20,499	under 11,500
Que	134,000 +	59,000-133,999	33,500-58,999	under 33,500
Ont	141,000 +	62,000-140,999	35,000-61,999	under 35,000
Man	69,500 +	31,500-69,499	18,000-31,499	under 18,000
Sask	54,500 +	24,500-54,499	14,000-24,499	under 14,000
Alta	161,000 +	73,500-160,999	42,000-73,499	under 42,000
B.C.	194,500 +	88,000-194,499	50,000-87,999	under 50,000
Atlantic prov.	47,500 +	21,000-47,499	12,000-20,999	under 12,000
Man & Sask	71,000 +	32,000-70,999	18,000-31,999	under 18,000
Alta & B.C.	185,500 +	82,500-185,499	47,000-82,499	under 47,000
Canada	171,000 +	74,500-170,999	42,000-74,999	under 42,000

Obtaining approximate CV's from the tables

Approximate coefficients of variation (CV's) are shown in the Approximate Sampling Variability Tables at the end of this section and on the CD-ROM. Before applying the criterion of the coefficient of variation, first follow the guidelines based on sample sizes described in Chapter 9 (Guidelines for release).

The following rules and examples should enable the user to determine the approximate coefficients of variation from the Sampling Variability Tables for estimates of the number of the surveyed population possessing a certain characteristic. The 'real life' examples are included to assist users in applying the rules. These examples use variables which require person weights in order to create estimates.

Rule 1: Estimates of Numbers Possessing a Characteristic (Aggregates)

The coefficient of variation depends only on the size of the estimate itself. On the Sampling Variability Table for the appropriate geographic area, locate the estimated number in the left-most column of the table (headed "Estimate") and follow the asterisks (if any) across to the first figure encountered. This figure is the approximate coefficient of variation.

Example using rule 1:

Suppose that a user estimates that 6,032,234 persons took at least one trip in March 1996. How does the user determine the coefficient of variation of this estimate?

- Refer to the CV table for CANADA.

- The estimated aggregate (6,032,234) does not appear in the left-hand column (the 'Numerator of Percentage' column), so it is necessary to use the figure closest to it, namely 6,000,000.

- The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, namely, 2.3%.

- So the approximate coefficient of variation of the estimate is 2.3%. The finding that there were 6,032,234 persons who took at least one trip in March 1996 is publishable with no qualifications.

Rule 2: Estimates of Proportions or Percentages Possessing a Characteristic

The coefficient of variation of an estimated proportion or percentage depends on both the size of the proportion or percentage and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a sub-group of the population. For example, the proportion of "persons aged 15 or more who took at least one trip in the reference month" is more reliable than the estimated number of "persons aged 15 or more who took at least one trip in the reference month". (Note that in the tables the CV's decline in value reading from left to right).

When the proportion or percentage is based upon the total population of the geographic area covered by the table, the CV of the proportion or percentage is the same as the CV of the numerator of the proportion or percentage. In this case, Rule 1 can be used.

When the proportion or percentage is based upon a subset of the total population (e.g. those in a particular sex or age group), reference should be made to the proportion or percentage (across the top of the table) and to the numerator of the proportion or percentage (down the left side of the table). The intersection of the appropriate row and column gives the coefficient of variation.

Example using rule 2:

Suppose that the user estimates that 2,951,511/ 6,032,234=49% of those persons who travelled in March took at least one same-day trip.

How does the user determine the coefficient of variation of this estimate?

- Refer to the table for CANADA.

- Because the estimate is a percentage which is based on a subset of the total population (i.e., travellers who took at least one same-day trip in March), it is necessary to use both the percentage (49%) and the numerator portion of the percentage (2,951,511) in determining the coefficient of variation.

- The numerator, 2,951,511, does not appear in the left-hand column (the 'Numerator of Percentage' column) so it is necessary to use the figure closest to it, namely 3,000,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the figure closest to it, 50.0%.

- The figure at the intersection of the row and column used, namely 2.8%, is the coefficient of variation to be used.

- So the approximate coefficient of variation of the estimate is 2.8%. The finding that 49% of persons who travelled in March and took at one same-day trip can be published with no qualifications.

Rule 3: Estimates of Differences Between Aggregates or Percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately. That is, the standard error of a difference ($\hat{d} = \hat{X}_1 - \hat{X}_2$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}}/\hat{d}$. This formula is accurate for the

difference between separate and uncorrelated characteristics, but is only approximate otherwise.

Example using rule 3:

Suppose that a user estimates that 2,951,511/ 6,032,234=49% of persons who travelled in March took at least one same-day trip, while 3,998,785/ 6,032,234=66.3% of persons who travelled in March took at least one overnight trip. (Note that a person could take both a same-day trip and an overnight trip in the same month, hence the estimates overlap). How does the user determine the coefficient of variation of the difference between these two estimates?

- Using the CANADA CV table in the same manner as described in example 2 gives the CV of the estimate for travellers who took at least one same-day trip as 2.8%, and the CV of the estimate for travellers who took at least one overnight trip as 2.1%.

- Using rule 3, the standard error of a difference ($\hat{d} = \hat{X}_1 - \hat{X}_2$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

That is, the standard error of the difference $\hat{d}$ = (.663 - .490) = .173 is:

$$\sigma_{\hat{d}} = \sqrt{[(.49)(.028)]^2 + [(.660)(.021)]^2}$$

$$= \sqrt{(.00001882384) + (.0001920996)}$$

$$= .0195022$$

- The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}}/\hat{d}$ = .019/.173 = 0.11.

- So the approximate coefficient of variation of the difference between the estimates is 11%. This estimate can be released without restrictions.

Rule 4: Estimates of Ratios

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied. This would apply, for example, to the case where the denominator is the number of persons who took at least one trip in the reference month and the numerator is the number of "persons who took at least one business trip in the reference month".

In the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of "persons who took at least one business trip in the reference month" as compared to the number of "persons who took at least one trip for pleasure during the reference month", the standard deviation of the ratio of the estimates is approximately equal to the square root of the sum of squares of each coefficient of variation

considered separately multiplied by R. That is, the standard error of a ratio ($\hat{R} = \hat{X}_1 / \hat{X}_2$) is:

$$\sigma_{\hat{R}} = \hat{R}\sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

The coefficient of variation of $\hat{R}$ is given by $\sigma_{\hat{R}}/\hat{R}$. The formula will tend to overstate the error, if $\hat{X}_1$ and $\hat{X}_2$ are positively correlated and understate the error if $\hat{X}_1$ and $\hat{X}_2$ are negatively correlated.

Example using rule 4:

Suppose that the user estimates that 3,998,785 March travellers took at least one overnight trip, while 2,951,511 March travellers took at least one same-day trip. The user is interested in comparing the estimate of overnight travellers versus that of same-day travellers in the form of a ratio. How does the user determine the coefficient of variation of this estimate?

- First of all, this estimate is a ratio estimate, where the numerator of the estimate (= $\hat{X}_1$) is the number of March travellers who took at least one overnight trip. The denominator of the estimate (= $\hat{X}_2$) is the number of March travellers who took at least one same-day trip.

- Refer to the table for CANADA.

- The numerator of this ratio estimate is 3,998,785. The figure closest to it is 4,000,000.

The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 3.1%.

- The denominator of this ratio estimate is 2,951,511. The figure closest to it is 3,000,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 3.6%.

- So the approximate coefficient of variation of the ratio estimate is given by rule 4, which is,

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the coefficients of variation of $\bar{X}_1$ and $\bar{X}_2$ respectively.

That is ,

$$\alpha_{\hat{R}} = \sqrt{(.031)^2 + (.036)^2}$$

$$= 0.047$$

- The obtained ratio of March 1996 travellers who took at least one overnight trip versus March 1996 travellers who took at least one sameday trip is 3,998,785/2,951,511 which is 1.35:1. The coefficient of variation of this estimate is 4.7%, which is releasable with no qualifications.

Rule 5: Estimates of Differences of Ratios

In this case, Rules 3 and 4 are combined. The CV's for the two ratios are first determined using Rule 4,

and then the CV of their difference is found using Rule 3.

Using C.V. tables to obtain confidence limits

Although coefficients of variation are widely used, a more intuitively meaningful measure of sampling error is the confidence interval of an estimate. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values. For example a 95% confidence interval can be described as follows:

If sampling of the population is repeated indefinitely, each sample leading to a new confidence interval for an estimate, then in 95% of the samples the interval will cover the true population value.

Using the standard error of an estimate, confidence intervals for estimates may be obtained under the assumption that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, the chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and about 99 out 100 that the differences would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

Confidence intervals for an estimate, $\hat{X}$, are generally expressed as two numbers, one below the estimate and one above the estimate, as ($\hat{X}$-k, $\hat{X}$+k) where k is determined depending upon the level of confidence desired and the sampling error of the estimate.

Confidence intervals for an estimate can be calculated directly from the Approximate Sampling Variability Tables by first determining from the appropriate table the coefficient of variation of the estimate $\hat{X}$, and then using the following formula to convert to a confidence interval CI:

$$CI_X = [\hat{X} - t\hat{X}\alpha_{\hat{X}} , \hat{X} + t\hat{X}\alpha_{\hat{X}}]$$

where $\alpha_{\hat{X}}$ is the determined coefficient of variation of $\hat{X}$, and

t = 1 if a 68% confidence interval is desired
t = 1.6 if a 90% confidence interval is desired
t = 2 if a 95% confidence interval is desired
t = 3 if a 99% confidence interval is desired.

Note: Release guidelines which apply to the estimate also apply to the confidence interval. For

example, if the estimate is not releasable, then the confidence interval is not releasable either.

Example of using C.V. tables to obtain confidence limits

A 95% confidence interval for the estimated proportion of persons who travelled in March and took at least one same-day trip (from example using rule 2, Chapter 10) would be calculated as follows.

$\hat{X}$ = 49% (or expressed as a proportion = .49)

t = 2

$\alpha_{\hat{X}}$ = 2.8% (.028 expressed as a percentage) is the coefficient of variation of this estimate as determined from the tables.

CI_X = {.49 - (2) (.49) (.028), .49 + (2) (.49) (.028)}

CI_X = {.49 - .027, .49 + .027}

CI_X = {.463, .517}

With 95% confidence it can be said that between 46.3% and 51.7% of persons who travelled in March took at least one same-day trip.

Using C.V. tables to do t-tests

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The sample estimates can be numbers, averages, percentages, ratios, etc. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

Let $\mathbf{X_1}$ and $\mathbf{X_2}$ be sample estimates for two characteristics of interest. Let the standard error on the difference $\hat{X}_1 - \hat{X}_2$ be $\sigma_{\hat{d}}$.

If $t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$ is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance. If however, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 0.05 level. That is to say that the characteristics are significant.

Example of using C.V. tables to do a t-test

Let us suppose we wish to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of travellers in March who took at least one same-day trip and the proportion of travellers in March who took at least one overnight trip. From the example using rule 3, the standard error of the difference between these two estimates was found to be = .019. Hence ,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{.49 - .663}{.019} = \frac{-.173}{.019} = -9.10.$$

Since t = -9.10 is less than -2, it must be concluded that there is a significant difference between the two estimates at the 0.05 level of significance.

Cvs for quantitative estimates

For quantitative estimates, special tables would have to be produced to determine their sampling error. Since most of the variables for the Canadian Travel Survey are primarily categorical in nature, this has not been done. These tables are included in the documentation and may be used with the analysis of variables that use person-trip weights, household-trip weights, person-night weights and expenditures weights.

As a general rule, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). Note that if the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the total number of trips taken in March would be greater than the coefficient of variation of the corresponding proportion of persons who took at least one trip in March. Hence if the coefficient of variation of the proportion is not releasable, then the coefficient of variation of the corresponding quantitative estimate will also not be releasable.

Coefficients of variation of such estimates can be derived as required for a specific estimate using a technique known as pseudo replication. This involves dividing the records on the microdata files into subgroups (or replicates) and determining the variation in the estimate from replicate to replicate. Users wishing to derive coefficients of variation for quantitative estimates may contact Statistics Canada for advice on the allocation of records to

appropriate replicates and the formulae to be used in these calculations.