

## Minimal sample size requirements for estimates of proportions.

April 2003

Didier Garriguet, DMES-SSMD

The purpose of this document is to help users to determine the required sample sizes to estimate proportion based on the NLSCY data for different domains of interest.

With the help of computer simulations, we've calculated variances, coefficients of variation and also confidence intervals at the 95% level for different proportions ranging from 1 to 50%. These were based on the cross-sectional population for cycle 4 using bootstrap weights. Note that the use of bootstrap weights from the longitudinal population can produce slightly different estimates for predicting cycle 5 and 6 sample sizes. In practical terms, we simulated a dichotomous variable in proportions of 1, 5, 10, 15, 20, 30, 40 and 50%. In doing so, we obtained a good approximation for the complete spectrum of proportions since knowing one proportion, we also know the corresponding 100% minus the calculated proportion. The variance and standard error will remain the same, but not the c.v. An approximate c.v. is obtained by dividing the standard error by the proportion. However the user should note that for disclosure issues, for a dichotomous variable, both variables should be publishable simultaneously. You should always ensure the quality of the smaller proportion. For a given repetition, the observed proportion in the random sample can be different from that of the targeted proportion. We therefore use the mean of 100 repetitions to be able to account for that variability.

We projected the cycle 5 and 6 populations by assuming a uniform response rate of 90% for each year and repeated the exercise on those estimated populations. With the mean from a 100 repetitions we can adjust for the randomness in the selection of respondents and for non-response.

We've studied numerous domains of estimation, in particular for various geographical levels. We calculated proportions for each province, for regions –Atlantic provinces (Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick), Québec, Ontario, Prairies (Manitoba, Saskatchewan, Alberta) and British Columbia; and for Canada as a whole. For demographic characteristics, we used individual ages, and the following age groups at cycle 4.

### *Age groups available*

0-1	2-7	5-6	9-11	14-15
0-3	2-8	6-13	10-11	16-17
0-5	4-11	6-7	10-15	
0-6	4-5	6-8	10-17	
2-3	4-6	7-8	12-13	
2-5	4-7	7-9	12-17	

The file [\\Lhs5\Method\nlscy\\_c4\variance\bootstrap\Tableau cv excel\table cv-size eng.xls](\\Lhs5\Method\nlscy_c4\variance\bootstrap\Tableau cv excel\table cv-size eng.xls) contains the results of these simulations. In selecting one or more field, we obtain the coefficient of variation and the confidence interval for a particular domain. The fields are:

<b>Province:</b>	The province or ATLANTIC or PRAIRIES for these specific regions or CANADA for the country as a whole.
<b>C4 Age:</b>	Age at cycle 4. Can take values from 0 to 18 and different age groups.
<b>C5 Age:</b>	Age at cycle 5. Can take values from 2 to 20 and different age groups.
<b>C6 Age:</b>	Age at cycle 6. Can take values from 4 to 22 and different age groups.
<b>Target prop. :</b>	The theoretical proportion used to simulate a variable. Can take the values 1%, 5%, 10%, 15%, 20%, 30%, 40% and 50%.
<b>Cycle:</b>	C4 (observed), C5 (simulated response rate), C6 (simulated response rate) for every cycle.
<b>Yhat:</b>	The mean from 100 calculated proportions. Should be close to Target prop.
<b>n:</b>	The average sample size of the specified domain from 100 repetitions.
<b>Bs_var:</b>	The mean of 100 variances for the specified domain.
<b>Bs_sd:</b>	The mean of 100 standard errors for the specified domain.
<b>Bs_cv:</b>	The mean of 100 coefficients of variation for the specified domain.
<b>Cil95:</b>	The mean of 100 95% confidence interval lower boundaries.
<b>Ciu95:</b>	The mean of 100 95% confidence interval upper boundaries.

Note that according to the sampling design for cycle 5, there shouldn't be any 6-7 year-old kids selected (4-5 year-olds at cycle 4). The results are only projections.

For example, to estimate the proportion of 3 year-old boys or girls in Newfoundland and Labrador, we select province « NEWFOUNDLAND & LABRADOR », C4 Age « 4 » and Target prop. « 50% » (since a reasonable estimate of that proportion should be close to 1 out of 2). For every cycle, the sample size drops from 472 to 425 and then to 382 kids. The coefficients of variation are stable ranging from 5,64% to 6,26% in cycles 4 through 6. The confidence intervals reiterate the stability of the variance estimate of that variable even with the smaller sample size.

Another way to use the table is to select only one province and one age group and to look at the variability of the coefficient of variation to determine the proportion and the sample size to get a reliable estimate. If we fix a coefficient of variation threshold of 16,5%, we can see, for example, in selecting province of Quebec for the age group 0-5 year-olds, we get estimates for proportions above 5% in every cycle.

Finally, one might be interested in knowing all the domains where the estimates are above a certain level of reliability. A selection on the field bs\_cv with or without province and/or with or without age groups will identify the domains where the condition is met.

Two examples are available in PowerPoint format in the file (from cycle 3):

[\\Lhs5\Method\nlscy\\_c4\variance\bootstrap\Tableau cv excel\Examples.ppt](\\Lhs5\Method\nlscy_c4\variance\bootstrap\Tableau cv excel\Examples.ppt).

In closing, as a reference, many surveys in Statistics Canada use the following quality standards:

- 1) An estimate is said **acceptable** if the sample size is at least 30 and the coefficient of variation is lower than 16,5%
  - 2) An estimate is said **marginal** if the sample size is at least 30 and the coefficient of variation is between 16,5% and 33,33%. This estimate should be accompanied by a warning to emphasise the high level of error.
  - 3) An estimate is said **unacceptable** if the sample size is lower than 30 or if the coefficient of variation is greater than 33,33%. This estimate should not be released.
- Reminder: Statistics Canada Quality Level Guidelines