

APPENDIX:

VARIANCE ESTIMATION FOR THE ETHNIC DIVERSITY SURVEY

The variability or variance of an estimate is a good indicator of the quality of that estimate. An estimate with an overly large variance is deemed unreliable. In order to quantify what is considered an overly large variance, the EDS uses the coefficient of variation (CV), which is a relative measure of variability. The CV is more useful than the variance when comparing the accuracy of estimates from samples of different sizes or scales.

The following section contains examples that may answer some of the questions that tend to come up when analyzing data.

- 1) How does one determine the CV for a given estimate?
- 2) Is the difference observed between the two estimates (percentage or proportion) statistically significant?
- 3) How does one determine the CV when the observed percentage is greater than 50%?
- 4) How does one determine the CV when only one sub-sample (domain) of the population answered a question?

Question 1. How does one determine the CV for a given estimate?

For the EDS PUMF, there are two ways to estimate the CV associated with an estimate. Users may calculate the CV by using the bootstrap weights included in the PUMF (see section 1a) or the Excel tool containing pre-calculated CV approximations for certain domains (see section 1b).

Although the first method is more precise for estimating variance, the Excel tool can be used to obtain virtually equivalent CVs for proportions, and to do so more quickly.

A) Bootstrap method

An efficient way of estimating variance using survey data from a complex sample plan, such as the EDS is to use one of the resampling techniques, like the bootstrap. In order to apply this technique, the estimate of interest $\hat{\theta}$ should be calculated from survey data, then this estimate should be recalculated for each of the 500 sets of bootstrap weights (located in the file BSW.txt). The next step is the calculation of the variability between the

estimates obtained through the following formula, which corresponds to the bootstrap variance for this estimate:

$$\hat{V}_B(\hat{\theta}) = \frac{1}{500} \sum_{i=1}^{500} (\hat{\theta}_{Bi} - \hat{\theta})^2$$

where $\hat{\theta}_{Bi}$ is the estimate based on the bootstrap weights for bootstrap sample i .

The CV for the estimate can be determined from the following formula:

$$CV(\hat{\theta}) = \frac{\sqrt{\hat{V}_B(\hat{\theta})}}{\hat{\theta}}$$

Software like WESVAR, developed by WESTAT and SUDAAN, can be used to directly estimate variance by the bootstrap method. Software such as SAS, SPSS and Stata do not have this method directly available. Statistics Canada developed SAS macros (called BOOTVAR) to apply the bootstrap method in order to obtain an accurate estimate of the variance.

Users are free to use whichever software they wish to estimate variance as long as they ensure that the bootstrap method is applied using the weights provided with the EDS PUMF.

The BOOTVAR macros program is included in the EDS PUMF product. User could refer to the following document for instructions on using the BOOTVAR:

Appendix – User’s guide for the BOOTVAR

For more information on the bootstrap method, users should consult the following documents:

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury Press, USA.

Rao, J.N.K., C.F.J. Wu and K. Yue. 1992. “Some recent work on resampling methods for complex surveys.” *Survey methodology* (Statistics Canada, Catalogue 12-001). Ottawa: Statistics Canada, 18, 2: 209-217

K.F. Rust , J.N.K. Rao, “Variance estimation for complex surveys using replication techniques”, *Statistical Methods in Medical Research*, 5, 1996, p. 281-310

Statistics Canada. 2003, *Survey methods and practices*, 12-587-XPE

Users may refer to the following documents for additional information on using the bootstrap method in WesVar, SUDAAN or Stata:

Piérard, E., Buckley, N., Chowman, J. “Bootstrapping made easy: A STAT ADOO File”. *The Research Data Centres information and technical bulletin*. Volume 1, number 1, spring 2004, 20-36 (Statistics Canada, no. 12-002-XIE in the catalogue).

Phillips, O. “Using bootstrap weights with WesVar and SUDAAN”. *The Research Data Centres information and technical bulletin*. Volume 1, number 2, fall 2004, 6-15 (Statistics Canada, no. 12-002-XIE in the catalogue).

Research Triangle Institute. 2001. *SUDAAN User’s Manual, Release 8.0*. Research Triangle Institute, Research Triangle Park, NC.

Westat. 2002. *WesVar 4.2 User’s Guide*. Westat, USA.

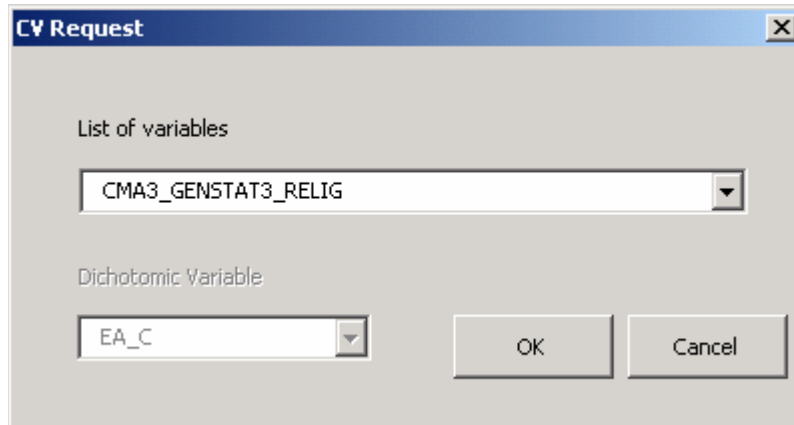
B) Excel tool

Approximate coefficients of variation (CV) can be obtained for EDS estimates by using a simple interactive tool. This tool is part of the EDS PUMF product. It is in the form of an Excel spreadsheet.

Important notice to EDS PUMF users

The Excel tool presented in this section was designed only to estimate CVs and the variance in simple cross-tabulation tables. With statistical methods that require the measurement of significance (e.g. regression analysis), users will have to resort to the previously presented bootstrap method.

To use the CV estimation tool for the EDS, open **FindCV.xls**. A window may appear. If so, click on “Enable Macros”. An Excel spreadsheet will appear, and you should be able to see the survey title and the “**CV requested**” button. Click on “**CV requested**” to open the application. The following screen will appear:



The screenshot shows a dialog box titled "CV Request". It has a standard Windows-style title bar with a close button (X) in the top right corner. The dialog contains two dropdown menus. The first is labeled "List of variables" and has "CMA3_GENSTAT3_RELIG" selected. The second is labeled "Dichotomic Variable" and has "EA_C" selected. At the bottom right of the dialog are two buttons: "OK" and "Cancel".

Step 1: Selecting the type of domain

The estimation domain is simply the sub-category of the total sample chosen to produce a given estimate (e.g., Montreal Catholics). Based on the classification variables used in the analysis, select the appropriate domain in Table 1. Selecting one of the last two domains will activate the drop box for dichotomous variables. The dichotomous variables are listed in Table 2.

Table 1: Available domains

CMA3, GENSTAT3 and RELIG
CMA3, AGES, SEX and VISMIND
CMA3, AGES, SEX and PBSLCT
CMA3, GENSTAT3 and VISMIND
CMA3, AGES, SEX and GENSTAT3
CMA3, GENSTAT3 and Dichotomous Variable
CMA3, AGES, SEX and Dichotomous Variable

Table 2: Available dichotomous variables

EA_C	EA_NOR	EA_FIL	EA_OCR
EA_F_C	EA_SWD	EA_JPN	EA_OLT
EA_QUE	EA_HNG	EA_VTN	EA_OOT
EA_F	EA_POL	EA_JAM	L1_ENG
EA_ENG	EA_ROM	EA_AME	L1_FRE
EA_IRS	EA_RUS	EA_REG	L1_GER
EA_SCT	EA_UKR	EA_OWE	L1_ITA
EA_WEL	EA_GRK	EA_ONE	L1_POL
EA_BRT	EA_ITL	EA_OEE	L1_POR
EA_AUS	EA_SPN	EA_OSE	L1_PUN
EA_BEL	EA_POR	EA_OOE	L1_SPA
EA_DUT	EA_JEW	EA_OAF	L1_TAG
EA_GER	EA_LEB	EA_OAR	L1_UKR
EA_SWS	EA_EIN	EA_OWA	L1_ARA
EA_DAN	EA_PNJ	EA_OSA	L1_DUT
EA_FIN	EA_CHN	EA_OEA	L1_CHIN

Note: The dichotomous variables starting with the prefix EA_ have not been put in the microdata file. Each of these variables represents a specific ancestry or specific group of ancestries as defined in EAC1 through EAC8. For example, if the user wants to have approximations of CVs for respondents who reported an English ancestry (i.e. at least one of the variables EAC1 through EAC8 is equal to 06), he should select the dichotomous variable EA_ENG in the Excel tool. If analyse by ethnic ancestry are planned, it is suggested to first derive the associated dichotomous variable based on EAC1 through EAC8 then use this variable in the analyse.

Example 1

We want to calculate the CV for the proportion of French speaking people whose generational status is first generation and whose first language is Italian.

In the CV Requested window, select CMA3_GENSTAT3_Dichotomous Variable and dichotomous variable L1_ITA.


Step 2: Selecting the desired items in the estimation domain

After clicking on OK, the Results sheet appears in the following form.

	A1	CMA3											
	A	B	C	D	E	F	G	H	I	J	K	L	M
	CMA3	GENSTAT3	L1_ITA	Target P	Simulated P	N	n	Variance	Standard deviation	CV	INF	SUP	
1													
2	TOTAL	TOTAL	TOTAL	1%	0.997386648	23092640	41695	0.005129349	0.071480315	7.1645	0.85728523	1.137488066	
3	TOTAL	TOTAL	TOTAL	5%	4.992491684	23092640	41695	0.025641243	0.159931031	3.203	4.679028663	5.305965604	
4	TOTAL	TOTAL	TOTAL	10%	9.985484036	23092640	41695	0.048263488	0.219461432	2.1975	9.555339628	10.41562844	
5	TOTAL	TOTAL	TOTAL	15%	15.01544388	23092640	41695	0.067460731	0.259544539	1.7295	14.50673659	15.52415118	
6	TOTAL	TOTAL	TOTAL	20%	20.00817715	23092640	41695	0.084945282	0.291242117	1.4565	19.4373426	20.5790117	
7	TOTAL	TOTAL	TOTAL	25%	25.01779336	23092640	41695	0.100212396	0.316404853	1.265	24.39763985	25.63794688	
8	TOTAL	TOTAL	TOTAL	30%	30.01932272	23092640	41695	0.111734826	0.334095208	1.1125	29.36449611	30.67414933	
9	TOTAL	TOTAL	TOTAL	35%	34.99156344	23092640	41695	0.12098186	0.347527453	0.9925	34.31040963	35.67271725	
10	TOTAL	TOTAL	TOTAL	40%	39.98273106	23092640	41695	0.127946591	0.357484798	0.894	39.28206085	40.68340126	
11	TOTAL	TOTAL	TOTAL	50%	49.98612441	23092640	41695	0.134985245	0.367268635	0.735	49.26627788	50.70597093	
12	TOTAL	TOTAL	Italian	1%	1.089180785	543310	1209	0.122616024	0.345016284	31.7255	0.412948868	1.7654127	
13	TOTAL	TOTAL	Italian	5%	4.99832151	543310	1209	0.543818201	0.732938733	14.648	3.561761593	6.434881428	
14	TOTAL	TOTAL	Italian	10%	9.93003553	543310	1209	1.05853876	1.025195237	10.3175	7.920652865	11.9394182	
15	TOTAL	TOTAL	Italian	15%	14.98355073	543310	1209	1.47763427	1.212731698	8.0905	12.6065966	17.36050485	
16	TOTAL	TOTAL	Italian	20%	19.87757456	543310	1209	1.849377899	1.358042545	6.831	17.21581118	22.53933795	
17	TOTAL	TOTAL	Italian	25%	24.87797428	543310	1209	2.170236394	1.472205733	5.92	21.99245105	27.76349752	
18	TOTAL	TOTAL	Italian	30%	29.72354395	543310	1209	2.413552308	1.552877957	5.228	26.67990315	32.76718475	
19	TOTAL	TOTAL	Italian	35%	34.81981704	543310	1209	2.59976281	1.611465205	4.6305	31.66134523	37.97828884	
20	TOTAL	TOTAL	Italian	40%	39.85401224	543310	1209	2.747085406	1.656620555	4.1595	36.60703596	43.10098853	
21	TOTAL	TOTAL	Italian	50%	49.72689172	543310	1209	2.835662886	1.683231189	3.3865	46.42775859	53.02602485	
22	TOTAL	1st generation	TOTAL	1%	0.993236829	5273330	10686	0.013814806	0.117168614	11.8053	0.763586346	1.222867312	
23	TOTAL	1st generation	TOTAL	5%	4.994070534	5273330	10686	0.067163875	0.258880986	5.1847	4.486663802	5.501477265	
24	TOTAL	1st generation	TOTAL	10%	9.995230197	5273330	10686	0.127135429	0.356325421	3.5665	9.296832372	10.69362802	
25	TOTAL	1st generation	TOTAL	15%	14.99065655	5273330	10686	0.179670329	0.423633037	2.8262	14.16033574	15.82097725	
26	TOTAL	1st generation	TOTAL	20%	20.01168607	5273330	10686	0.226241579	0.475387185	2.3755	19.07992718	20.94344495	
27	TOTAL	1st generation	TOTAL	25%	25.00881223	5273330	10686	0.263298445	0.512844355	2.0512	24.0016373	26.01198717	
28	TOTAL	1st generation	TOTAL	30%	30.01888545	5273330	10686	0.296432653	0.544180819	1.8133	28.95229105	31.08547986	
29	TOTAL	1st generation	TOTAL	35%	35.0103029	5273330	10686	0.320313493	0.565695795	1.6155	33.90153914	36.11906665	
30	TOTAL	1st generation	TOTAL	40%	40.00510997	5273330	10686	0.342058775	0.584582082	1.4612	38.85932909	41.15089085	
31	TOTAL	1st generation	TOTAL	50%	49.99906677	5273330	10686	0.356076392	0.596392961	1.1925	48.83013657	51.16799697	
32	TOTAL	1st generation	Italian	1%	0.993780227	300070	683	0.177695766	0.413416077	41.5487	0.183484715	1.804075739	

Every column in the output table has a specific meaning. In our example:

- CMA3 – Selected domain variable
- GENSTAT3 – Selected domain variable
- L1_ITA – Selected domain variable
- P Target – Target proportion in the simulation
- P Simulated – Real proportion obtained in the simulation
- N – Population size (rounded to the nearest tenth)
- n – Sample size
- Variance – Proportion estimate variance
- Standard deviation – Proportion estimate standard deviation
- CV – Coefficient of variation
- INF – Lower limit of the 95% confidence interval
- SUP – Upper limit of the 95% confidence interval

Once the Results sheet appears, the desired items have to be chosen in the estimation domain. This is done by clicking on the scroll-down list  for column “CMA3” and selecting the area for which the estimates are wanted. This will filter the data and retain only those lines in the table that contain estimates for the specified geographic area. If user

wants to list all geographic areas, select “(all)” so they will all be listed, or choose “TOTAL” to save only overall estimates (i.e., estimate at Canada level). The same thing should be done with the columns representing the other variables in the domain.

Then, select the desired proportion by clicking on “P Target”. For instance, if the objective is to obtain a CV for a proportion of 23% (which is not on the list), select “(all)” in the list in order to save all proportions. Thus, by using the CVs that correspond to the proportions of 20% and 25% for a given domain, the desired CV would be between these two limits.

Example 1 (Continued)

In the Results sheet, select “TOTAL” for variable CMA3, “1st generation” for GENSTAT3 and Italian for “L1_ITA”.

Then determine the target proportion. It is easier to determine the desired proportion with a frequency table. Weight WGT_PUMF will give the following results for languages spoken:

Table 3: Languages spoken by people whose generational status is 1st generation and whose first language is Italian

English only	0.76%
French only	0.30%
Non-official language	17.75%
English and French	0.00%
English and non-official language(s)	53.34%
French and non-official language(s)	5.98%
English and French and non-official language(s)	21.75%
Non-official languages	0.12%

According to the table, 28.03% of people whose first language is Italian and whose generational status is first generation speak French. Since 28.03% is not in the P Target list, we will select “(all)”, which will leave us with the following Excel sheet:

	A	B	C	D	E	F	G	H	I	J	K	L
1	CMA3	GENSTAT3	L1_ITA	Target P	Simulated P	N	n	Variance	Standard deviation	CV	INF	SUP
32	TOTAL	1st generation	Italian	1%	0.993780227	330070	683	0.177695766	0.413416077	41.6487	0.183484715	1.804075739
33	TOTAL	1st generation	Italian	5%	5.11211527	330070	683	0.90368122	0.946081827	18.5161	3.257794889	6.966435652
34	TOTAL	1st generation	Italian	10%	10.16291098	330070	683	1.686177372	1.296133846	12.7593	7.622488647	12.70333332
35	TOTAL	1st generation	Italian	15%	15.12928023	330070	683	2.387484773	1.543402864	10.2078	12.10421062	18.15434984
36	TOTAL	1st generation	Italian	20%	20.12691574	330070	683	2.982589379	1.725453826	8.5777	16.74502624	23.50880524
37	TOTAL	1st generation	Italian	25%	25.10009202	330070	683	3.463937076	1.860047801	7.416	21.45439833	28.74578571
38	TOTAL	1st generation	Italian	30%	30.13074369	330070	683	3.881452662	1.968783525	6.5376	26.27192798	33.9895594
39	TOTAL	1st generation	Italian	35%	35.2163414	330070	683	4.206450092	2.049903937	5.8241	31.19852968	39.23415311
40	TOTAL	1st generation	Italian	40%	40.14810871	330070	683	4.434033126	2.104588252	5.2451	36.02311573	44.27310168
41	TOTAL	1st generation	Italian	50%	50.12918307	330070	683	4.64027567	2.153052829	4.2966	45.90919952	54.34916661

This gives a CV for the estimate between 6.5376% and 7.4160%.

Note to users: The actual proportion for the CV (P Simulated), the coefficient of variation (CV) and the confidence interval (INF and SUP) are only approximate values based on the “P Target” that is closest to the estimate obtained. Interpolation can be used to calculate a more accurate CV and confidence interval.

Example 1 (Continued)

The proportion of people whose first spoken language is Italian, whose generational status is first generation and who speak French comes to 28.03%. Thus, we looked at “P Target” of 25% and 30%. This gave:

P Target	P Simulated	CV	INF	SUP
25%	25.1001	7.416	21.4544	28.7458
30%	30.1307	6.5376	26.2719	33.9896

Linear interpolation based on where 28.03% was located between 25% and 30%, gave:

P Target	P Simulated	CV	INF	SUP
28.03%	28.03	6.9044	24.2602	31.7999

The new CV of 6.9044%, for instance, was calculated as follows:

$$7.416 + (6.5376 - 7.416) * (28.03 - 25.1001) / (30.1307 - 25.1001)$$

Step 3: Quality rules

Certain quality rules were applied to the CV calculations. When the number of individuals (non-weighted) in a cell is less than or equal to ten, that cell is deleted, along with its associated results. Moreover, there are guidelines for disseminating the estimates.

Table 4: Guidelines for disseminating estimates

Category	Coefficient of variation (%)	Recommendations
Acceptable	0.0 – 16.5	This estimate can be used with no restriction.
Marginal	16.6 – 33.3	The estimate must be used carefully as it is associated with a high level of error. Every time this level occurs, the symbol “E” should be attached to the estimate in question. In Excel, cells containing a CV between 16.6 and 33.3 are marked in yellow.
Unacceptable	Over 33.3	If the value obtained for the CV is over 33.3, this information should not be disseminated. However, if the user chooses to do so, the estimate should be disseminated with the following warning: “We inform the user that ... <specify the data > ... does not meet Statistics Canada’s quality standards. The conclusions drawn from this data would not be reliable”. Also, the symbol “F” should be tagged onto the estimate in question. In Excel, cells containing a CV higher than 33.3 are marked in red.

It should be mentioned that some simulated proportions are quite far from the target proportion. In most cases, this is because of the small number of observations in the cell in question. Thus, it is very likely that all of the simulated proportions in this domain will be far from the target value and that the corresponding CVs will be marked in red.

Step 4: Saving the results

The Results sheet contents are replaced with every new search. To save the results of the current search, copy the results that are to be saved and paste them into another Excel file, then save this new file.

Question 2. Is the difference observed between two estimates statistically significant?

This question is best answered with an example.

Example 2

We want to know whether there is a significant difference between the proportion of French-speaking people among those whose generational status is second generation and whose first language is Italian compared to the proportion of those who speak French and whose generational status is first generation and whose first language is Italian.

Table 5: Languages spoken by people whose generational status is second generation and whose first language is Italian

English only	2.53%
French only	0.0%
Non-official language	0.0%
English and French	0.26%
English and non-official language(s)	50.33%
French and non-official language(s)	0.16%
English and French and non-official language (s)	46.53%
Non-official languages	0.0%

Table 5 indicates that 46.95% of people whose generational status is second generation and whose first language is Italian speak French. The corresponding proportion for the first generation in example 1 is 28.03%. Is the difference between the two proportions statistically significant?

The CV (6.9044%) and the confidence interval (24.2602% to 31.7999%) are already known for the first generation. Users need only establish the CV and the confidence interval for the second generation by repeating the same steps as before, but this time choosing “2nd generation” in the “GENSTAT3” column and setting “P Target” as close to 46.95% as possible.

The CVs of people whose generational status is second generation, whose first language is Italian and who speak French is 5.7765%. The confidence interval is between 41.6914% and 52.2086%, with a 95% confidence threshold.

The two confidence intervals have to be compared in order to determine whether the difference between the two estimates is statistically significant.

1st generation: Between 24.2602% and 31.7999%

2nd generation: Between 41.6914% and 52.2086%

The method for determining whether the difference between the two estimates is statistically significant is explicit. If the two intervals overlap, we cannot confirm whether the two estimates are different (or, in more technical terms, with a confidence level of 95%, we *cannot* dismiss the null hypothesis whereby there is no statistical difference between the two estimates). However, if the two intervals do not overlap, it is possible to confirm that the two percentages are different, with a confidence level of 95% (in more technical terms, we *can* dismiss the null hypothesis whereby there is no statistical difference between the two estimates).

In summary, given the CVs and the confidence intervals, it is possible to confirm that the proportion of French-speaking people among those whose first language is Italian and whose generational status is first generation is much lower than the proportion of French-speaking people among those whose first language is Italian and whose generational status is second generation.

Question 3. How does one determine a CV when the estimate is higher than 50%?

First, apply the formula to calculate a coefficient of variation:

$$CV = \text{Standard Error} / \text{Estimation} * 100$$

Assume that we are interested in a specific domain, with a proportion higher than 50%. As the table shows, no CV has been calculated for proportions above 50%. However, the desired CVs can easily be calculated through the complementary proportion, as follows:

- We want the CV for the B proportion that is higher than 50%.
- We use the CV for the complementary A proportion for which $A=100-B$
- We must work **in a same domain** for proportions A and B.
- Thus, we have:

$$CV_A = \text{Standard Error}_A / \text{Estimation}_A * 100$$

- We must isolate the standard error in the formula and calculate the standard error from the CV and the estimate in the table.

$$\text{Standard Error}_A = (CV_A * \text{Estimation}_A) / 100$$

- Since **the standard error for A is the same as for its complement B**, we have to use the starting formula to find the CV for B.

$$CV_B = Standard\ Error_A / Estimation_B * 100$$

Example 3

We want to calculate the CV for the percentage of English-speaking people among those whose generational status is first generation and whose first language is Italian.

In table 3, we note that the proportion of English-speaking people among those whose generational status is first generation and whose first language is Italian is 75.85%. Thus, the complementary proportion is 24.15%.

The CV associated with this proportion is 7.6379%. Thus, the standard error is 1.8446. Hence, the CV associated with the estimate is 2.4318%.

Question 4. How does one determine a CV when only one sub-group answered a question? (e.g. questions only applicable to immigrants or people who are member of a club or an organization)

This scenario is different from the previous ones inasmuch as the respondents were separated from the rest of the population ahead of time by identifying with a particular characteristic.

If the sub-group in question corresponds to a domain that falls among those listed in the Excel application, the approach is the same as in response to question 1. For instance, we are looking for the CV of the proportion of immigrants who arrived in Canada before 1991. Here, the sub-group is immigrants (GENSTAT3 = 1).

However, if the sub-group does not correspond to a domain listed in the Excel application, the proportion of those respondents out of all respondents is the one to use, not the proportion out of the sub-group. For instance, the application can be used to find the CV associated with the proportion of people in the total population who are part of a team or sports club, but we cannot get the CV associated with the proportion of people who are part of a team or sport club out of those who have indicated that they are part of a group or organization because this domain is not available in the Excel application. To calculate the CV associated with the second proportion, it would be necessary to use the bootstrap method described in question 1a).

The previous case showed that there are several estimation domains, and it is important to distinguish between them to obtain a CV for a sub-group in the population. Essentially, it is a matter to ensuring that the denominator for the proportion corresponds to the value of N in the Results sheet.