**New Release Rules for Quality for CSDID Surveys**

Elisabeth Neusy, SSMD

(June 23, 2021)

1. **Introduction**

There has long been concern with the use of CV-based rules for proportions because coefficients of variation (CVs) are not an appropriate measure of quality for proportions. A full review and revision of the release rules was therefore conducted. The new rules involve releasing 95% confidence intervals (CI) to report quality, and include rules for issuing quality warnings and applying suppressions.

This document presents the new release rules. Section 2 explains why confidence intervals are recommended, and Section 3 presents rules for issuing warnings and applying suppressions for quality. Section 4 describes SAS macros that were developed to implement the new rules and provides examples. Finally, Section 5 distinguishes between rules for quality and rules for confidentiality.

1. **Reporting Quality through Confidence Intervals**

As mentioned, the new rules involve releasing 95% confidence intervals to report the quality of estimates in terms of their sampling error. This recommendation was approved as a best practice by the Methods and Standards Committee in 2017 ([MSC 2017](http://icn-rci.statcan.ca/888/888c/888c04/888c04_2017/888c04_20171120-eng.html)). The 95% confidence intervals should be released with the estimates, in the same tables as the estimates.

There are many advantages in using confidence intervals to report quality. They are appropriate for all types of estimates, such as estimated proportions, means, totals and differences. As well, they provide an objective measure of sampling error. This is contrast to the CV-based rules where quality is reported using subjective terms such as ‘acceptable’ or ‘marginal’, which may be misinterpreted by data users. For example, data users may wrongly assume that ‘acceptable’ quality means the estimate is of sufficiently reliable quality for all possible uses. Lastly, confidence intervals express the sampling error in a form that is easy to interpret. The Canadian public should be familiar with the notion of interval estimates since poll results are described by the media as being “accurate to plus or minus the margin of error, 19 times out of 20”.

1. **Release Rules for Issuing Quality Warnings and Applying Suppressions**

Even with the release of 95% confidence intervals, release rules are needed for issuing warnings or applying suppressions for quality reasons. The *Scientific Review Committee on Social Statistics* was consulted for social surveys ([TCHS 2018](https://gcdocs.gc.ca/statcan/llisapi.dll/link/2002680), [SRCSS 2019](https://gcdocs.gc.ca/statcan/llisapi.dll/link/5695939)). The Committee approved a general framework with the following three release categories:

Category A: Release with no warning. Data users should use the 95% confidence interval to assess whether the quality of the estimate is sufficient. Note that the ‘A’ is not a quality indicator; it should not be released.

Category E: Release with a quality warning. Data users should use the 95% confidence interval to assess whether the quality of the estimate is sufficient.

Category F: Suppress for quality reasons. The estimate and/or its confidence interval are deemed of such poor quality that they are not fit for any use.

The rules for assigning an estimate to a release category depends on the type of estimate. Note that these rules identify suppressions for quality; a different set of rules should be applied to identify suppressions for confidentiality.

* 1. **Release Rules for Estimated Proportions and Estimated Counts**

Estimated proportions and estimated counts are computed from binary variables. Estimated counts are estimates of the total number of persons/households with a characteristic of interest; in other words, they are the weighted sum of a binary variable (e.g., estimated number of immigrants). Estimated proportions are estimates of the proportion of persons/households with a characteristic of interest (e.g., estimated proportion of immigrants in the general population). Estimated counts and proportions can also be computed from categorical variables: that is, estimates of the number or proportion of persons/household who belong to a category.

The release rules for estimated proportions and estimated counts are based on sample size. The thresholds for the rules are survey specific because the level of quality achieved by a sample size depends on the design of the survey. For example, Tables 1 and 2 provide the release rules for the Canadian Tobacco and Nicotine Survey (CTNS) 2020; the rules in Table 2 are used whenever all respondents that contribute to the estimate are aged 15 to 24, otherwise the rules in Table 1 are used.

Table 1: General rules for proportions and counts for CTNS 2020

|  |  |  |
| --- | --- | --- |
| Sample Size (n) | Release Category | Action |
| n ≥ 100 | A | Release with no warning; users should use CI as quality indicator |
| 50 ≤ n < 100 | E | Release with quality warning; users should use CI |
| n < 50 | F | Suppress the estimate and its CI for quality reasons |

Table 2: Rules for proportions and counts for CTNS 2020 when all respondents are aged 15 to 24

|  |  |  |
| --- | --- | --- |
| Sample Size (n) | Release Category | Action |
| n ≥ 70 | A | Release with no warning; users should use CI as quality indicator |
| 35 ≤ n < 70 | E | Release with quality warning; users should use CI |
| n < 35 | F | Suppress the estimate and its CI for quality reasons |

For estimated proportions, n is defined as the unweighted count of the number of respondents in the denominator (not the numerator) of the proportion. For estimated counts, n is defined as the unweighted count of the number of respondents with nonzero values that contribute to the estimate.

The new release rules suppress estimated proportions based on small sample sizes. Estimated proportions based on small sample sizes are unreliable and very unstable. The confidence intervals are unstable as well, and there is a greater risk of poor performance by the confidence interval methods. Sample size is a better measure of quality for estimated proportions than CVs. In general, CVs tend to be too conservative for small proportions, and too lax for large proportions.

Methodology will provide one set of thresholds for each survey. There may be situations where Methodology provides a second set of thresholds for a specific subpopulation of a survey; clear instructions will be included whenever two sets of thresholds are provided. CTNS 2020 is an example of a survey with two sets of thresholds: a separate set of thresholds was provided for estimates based on respondents aged 15 to 24 because a different frame and sample design was used for this age group, which allows for substantially lower thresholds and more estimates to be released. The 15 to 24 age group is of particular interest to the client.

* 1. **Release Rules for Means and Totals of Quantitative Variables**

The release rules for the estimated mean and total of quantitative variables are based on sample size and on the CV of the estimate. The sample size thresholds are the same as those for estimated proportions and estimated counts. The CV thresholds are 50% for suppressing, and 25% for issuing a quality warning; looser thresholds are used than those used for the past CV-based rules because confidence intervals are released under the new rules and data users have information about the sampling error of the estimate. Tables 3 and 4 provide the release rules for CTNS 2020; the rules in Table 4 are used whenever all respondents that contribute to the estimate are aged 15 to 24, otherwise the rules in Table 3 are used.

Table 3: General rules for means and totals for CTNS 2020

|  |  |  |
| --- | --- | --- |
| Sample Size & CV | Release Category | Action |
| n≥100 and CV≤25% | A | Release with no warning; users should use CI as quality indicator |
| Otherwise | E | Release with quality warning; users should use CI |
| n<50 or CV>50% | F | Suppress the estimate and its CI for quality reasons |

Table 4: Rules for means and totals for CTNS 2020 when all respondents are aged 15 to 24

|  |  |  |
| --- | --- | --- |
| Sample Size & CV | Release Category | Action |
| n≥70 and CV≤25% | A | Release with no warning; users should use CI as quality indicator |
| Otherwise | E | Release with quality warning; users should use CI |
| n<35 or CV>50% | F | Suppress the estimate and its CI for quality reasons |

For estimated means, n is defined as the unweighted count of the number of respondents that contribute to the estimate including values of zero. For estimated totals, n is defined as the unweighted count of the number respondents with nonzero values that contribute to the estimate.

* 1. **Summary of the Release Rules for Quality**

Tables 5 and 6 provide a summary of the rules specified by Tables 1 to 4 for CTNS 2020. The rules in Table 6 are used whenever all respondents that contribute to the estimate are aged 15 to 24, otherwise the rules in Table 5 are used. Note that the definition for n is given in Sections 3.1 and 3.2 for each type of estimate.

Table 5: General rules for CTNS 2020

|  |  |  |  |
| --- | --- | --- | --- |
| Type of Estimate | Category A\* (no warning) | Category E (warning) | Category F (suppress) |
| Proportion | n ≥ 100 | 50 ≤ n < 100 | n < 50 |
| Count | n ≥ 100 | 50 ≤ n < 100 | n < 50 |
| Mean | n≥100 and CV≤25% | Otherwise | n<50 or CV>50% |
| Total | n≥100 and CV≤25% | Otherwise | n<50 or CV>50% |

Table 6: Rules for CTNS 2020 when all respondents are aged 15 to 24

|  |  |  |  |
| --- | --- | --- | --- |
| Type of Estimate | Category A\* (no warning) | Category E (warning) | Category F (suppress) |
| Proportion | n ≥ 70 | 35 ≤ n < 70 | n < 35 |
| Count | n ≥ 70 | 35 ≤ n < 70 | n < 35 |
| Mean | n≥70 and CV≤25% | Otherwise | n<35 or CV>50% |
| Total | n≥70 and CV≤25% | Otherwise | n<35 or CV>50% |

\* Note that ‘A’ is not a quality indicator; it should not be released with the estimate. The 95% confidence interval is the quality indicator.

The above release rules should suppress most estimates and confidence intervals of poor quality. There are also two conditions that indicate that a confidence interval is of poor quality. An estimate and its confidence interval should be assigned to release category F if either of the following two conditions are true:

* The lower bound of the 95% confidence interval is equal to the upper bound of the interval; in other words, the confidence interval is of length zero. (Exceptions are if the estimate is based on a census rather than a sample, or if the estimate corresponds to a calibration control total.)
* The lower bound or upper bound of the 95% confidence interval is not a plausible value for the estimate. For example, the lower bound for an estimated proportion is negative.
  1. **Release Rules for Differences**

In order to assign a release category for an estimated difference between two estimates, the analyst must first determine the release category of each of the two estimates using the rules described in Sections 3.1 to 3.3. Next, the release category of the estimated difference or the estimate of change is assigned the lower release category of the two estimates; this can be specified as follows:

* If one or both estimates are category F estimates, then assign the estimated difference to category F and suppress it
* Otherwise, if one or both estimates are category E estimates, then assign the estimated difference to category E
* If both estimates are category A estimates, then assign the estimated difference to category A

1. **Implementing the New Release Rules**
   1. **SAS Macros for Implementing the New Release Rules**

Methodology has developed general macros that use SAS procedures to produce estimates and 95% confidence intervals of the following:

* Proportions and/or counts of a categorical variable using the SAS procedure surveyfreq
* Means and/or totals of a quantitative variable using the SAS procedure surveymeans
* Differences of means or proportions, and perform t-tests

Methodology has also developed an example of a main program that calls the different general macros to produce estimates. The program uses CTNS data and can be used as a template for other surveys.

The thresholds for the release categories are specified once in the main program; the release category of an estimate is then automatically assigned by the general macros. Methodology proposes to provide the program code that is required to specify the thresholds of each survey. The CTNS 2020 thresholds described in Section 3.1-3.3 are specified as follows in the main program:

**%macro** ***Thresholds***;

if agegroup in (**1**,**2**) then do; \* thresholds for respondents aged 15-24;

limit\_sup=**35**; \* threshold for suppressions (F);

limit\_warn=**70**; \* threshold for warnings (E);

end;

else do; \* general thresholds;

limit\_sup=**50**; \* threshold for suppressions (F);

limit\_warn=**100**; \* threshold for warnings (E);

end;

**%mend**;

* 1. **Examples Illustrating the Application of the New Release Rules**

Tables 7 to 11 illustrate the application of the release rules specified for CTNS 2020. The estimates and counts are fictional. Table 7 contains a column with the number of respondents in the numerator of the proportion: the counts in this column are struck through to emphasize that the number of respondents in the numerator are not used to determine quality; rather, it is the number of respondents in the denominator of the proportion that are used. The last column of Tables 7 to 10 indicates which rules from Sections 3.1 and 3.2 were used to assign the Release Category.

Table 7: Estimated Smoking Rate

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Province | Age Group | Estimate | # Respondents | | Release Category | Rules from Section 3.1 |
| ~~Numerator~~ | Denominator |
| 10 | 15-24 | 30% | ~~35~~ | 80 | A | Table 2 |
| 10 | 25+ | 55% | ~~45~~ | 90 | E | Table 1 |
| 10 | Total | 45% | ~~80~~ | 170 | A | Table 1 |

Table 8: Estimated Number of Smokers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Province | Age Group | Estimate | # Respondents who smoke | Release Category | Rules from Section 3.1 |
| 10 | 15-24 | 10,000 | 35 | E | Table 2 |
| 10 | 25+ | 20,000 | 45 | F | Table 1 |
| 10 | Total | 30,000 | 80 | E | Table 1 |

Table 9: Estimated mean number of cigarettes smoked per day

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Province | Age Group | Estimate | CV | # Respondents who contribute to the mean | Release Category | Rules from Section 3.2 |
| 11 | 15-24 | 5 | 55% | 80 | F | Table 4 |
| 11 | 25+ | 10 | 20% | 90 | E | Table 3 |
| 11 | Total | 8 | 30% | 170 | E | Table 3 |

Table 10: Estimated total number of cigarettes smoked in one day

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Province | Age Group | Estimate | CV | # Respondents who smoked one or more cigarettes | Release Category | Rules from Section 3.2 |
| 11 | 15-24 | 50,000 | 25% | 70 | A | Table 4 |
| 11 | 25+ | 100,000 | 15% | 80 | E | Table 3 |
| 11 | Total | 150,000 | 20% | 150 | A | Table 3 |

Table 11: Estimated difference between the estimates for persons aged 15 to 24 and the estimates for persons aged 25 and over (Rules from Section 3.4)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Province | Estimated Difference | Release Category | | |
| Estimate for 15-24 | Estimate for 25+ | Estimated Difference |
| 10 | Smoking Rate (Table 7) | A | E | E |
| 10 | Number of Smokers (Table 8) | E | F | F |
| 11 | Mean number of cigarettes smoked per day (Table 9) | F | E | F |
| 11 | Total number of cigarettes smoked in one day (Table 10) | A | E | E |

1. **Suppressions for Confidentiality Reasons**

The release rules presented in Sections 3 and 4 are release rules for quality reasons only. A separate set of rules are required for suppressions for confidentiality reasons. For example, the CTNS 2020 confidentiality rule for proportions (vetting rule) requires at least 5 ones and at least 5 zeros in the numerator of the proportion. Table 12 illustrates the application of this confidentiality rule using fictional data.

Table 12: Example of the application of the CTNS 2020 confidentiality rule for proportions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Province | Question | Response | # Respondents in numerator | Estimate | Suppress for confidentiality? |
| 12 | Have you ever smoked? | Yes | 100 | 97% | Yes |
| No | 3 | 3% | Yes |
| 13 | Have you ever smoked? | Yes | 5 | 60% | No |
| No | 5 | 40% | No |