SSC Annual Meeting, May 2021
Proceedings of the Survey Methods Section

INFERENCE FOR CENSUS LONG-FORM WEIGHTED COUNTS

Sarah-Anne Savard[1]

ABSTRACT

Confidence intervals for statistics obtained from survey data are normally constructed with the Wald procedure. However, it is known to be inadequate for proportions close to 0 or 1, and in the presence of small samples. This is also true for weighted counts, which are the most widely used statistics in Census disseminated products. In this paper, we describe the main findings from a research project undertaken at Statistics Canada to improve the inference for counts, and we apply these findings to the Census long-form context. We present simulation results that show that the proposed method has better properties than the usual method for constructing confidence intervals for weighted counts.

KEY WORDS: Confidence interval; count; official statistics.

RÉSUMÉ

Les intervalles de confiance pour des statistiques obtenues à partir de données d'enquête sont habituellement construits avec la méthode de Wald. Cependant, cette méthode n'est pas appropriée pour des proportions se rapprochant de 0 ou de 1, et pour de petites tailles d'échantillon. C'est également le cas pour les comptes pondérés, qui sont très répandus dans les produits diffusés au Recensement. Dans cet article, nous décrivons les principaux résultats d'un projet de recherche entrepris à Statistique Canada pour améliorer l'inférence pour les comptes, et nous appliquons ces résultats au contexte du questionnaire détaillé du Recensement. Nous présentons des résultats de simulations qui montrent que la méthode proposée a de meilleures propriétés que la méthode habituelle pour construire des intervalles de confiance pour les comptes pondérés.

MOTS CLÉS : Intervalle de confiance; compte; statistiques officielles.

1. INTRODUCTION

In Canada, the Census of Population is conducted every five years and the 2021 cycle is currently in production. The Census program has two main components: a census that enumerates the entire Canadian population and collects basic demographic information, and a sample survey where further social and economic information is collected for approximately 25% of dwellings. This second part is called the long form, after the name of the questionnaire used to collect information from the respondents.

One of the main changes over the past cycles is that the dissemination strategy of quality indicators has been greatly improved for 2021. Among other additions, confidence intervals will now be available for the majority of estimates produced with long-form data. The confidence interval was chosen as the main quality indicator for accuracy because it allows users to easily make a correct statistical inference given that it is constructed with an appropriate method.

It is worth noting that the dissemination process of the Census is characterized by a large volume of estimates, many of which are for very small domains and rare characteristics. Moreover, the majority of statistics produced with Census long-form data are weighted counts, or estimations of a population total, which is simply the number of units having a certain characteristic. As counts are closely related to proportions, and since it is known that the usual method for producing

[1] Sarah-Anne Savard, 100 promenade Tunney's Pasture, Ottawa, ON, Canada, K1A 0T6, sarah-anne.savard@statcan.gc.ca

confidence intervals is not always appropriate for proportions, there was a need to investigate and develop a new method to produce confidence intervals with good coverage properties for weighted counts that could be used for the 2021 Census.

The objective of this paper is to describe the application of a new method for constructing confidence intervals for weighted counts for the 2021 Census. The paper does not provide details of the development of the proposed method. Those are to be published in a separate paper. The paper is organized as follows. Section 2 gives a brief overview of the Census long-form estimation strategy. Section 3 explains the problem of using Wald confidence intervals for weighted counts. This section also includes the main result of the research project that was undertaken to address this problem. Finally, section 4 describes the simulation study that was carried out to evaluate the method in the census long-form context and provides some results of its performance.

2. OVERVIEW OF THE CENSUS LONG-FORM ESTIMATION STRATEGY

This section describes the estimation procedure associated with the Census long-form survey. In the greater part of the country, the long-form sample is selected with a systematic stratified design and the sampling unit is the dwelling. The sampling fraction is equal to ¼ and the design weight is equal to 4 for every household. The weights are first adjusted for sample coverage and total non-response using available auxiliary information. The resulting adjusted weights are calibrated to known totals from the census to ensure coherence and to reduce the variability of estimates. These adjustments to the weights are performed simultaneously at two geographical levels: the aggregated dissemination area (ADA), which comprises between 5,000 and 15,000 persons, and the super-ADA, which is an aggregation of ADAs. Since every person living in a selected dwelling is included in the sample, estimation can be carried out at both the person and household levels using a single set of weights.

The estimation strategy is completely different in remote areas, where all dwellings are taken to be part of the long-form sample. In these areas, the total non-response is compensated by an imputation procedure, no weight adjustments are performed and all households have a final weight of 1.

Variance estimation is carried out with a replication method that uses a small number of replicates, the Partially Balanced Repeated Replication-epsilon or PBRR-epsilon method (Devin and Verret, 2016), which is based on Fay's method (Dippo, Fay and Morganstein, 1984). The replicates are adjusted with the same methodology as the main set of weights. In 2021, a set of 32 replicates will be used to produce the variance estimates used for the construction of the confidence intervals included in the disseminated products.

3. PROPOSED METHOD: WILSON TYPE CONFIDENCE INTERVALS

3.1 Problems identified with the Wald-type interval for weighted counts

Confidence intervals are traditionally obtained using the Wald method. This method assumes that the point estimator follows a normal distribution. For a given parameter of interest θ, the Wald confidence interval is given by $\hat{\theta} \mp z\sqrt{\widehat{Var}\left(\hat{\theta}\right)}$, where $\hat{\theta}$ is the point estimate produced with survey weights, z is the normal quantile at a given level and $\widehat{Var}\left(\hat{\theta}\right)$ is the estimated variance of $\hat{\theta}$. For small sample sizes, the quantile t is used instead: it is obtained from the Student-T distribution with the appropriate number of degrees of freedom.

For some estimators, it is known that the Wald procedure results in confidence intervals that have poor coverage. For example, for a proportion close to 0 or 1, and for small sample sizes, the coverage of the usual confidence interval is lower than expected. The case of confidence intervals for proportions estimated from complex survey data was studied (Neusy and Mantel, 2016) and a more appropriate method to produce these confidence intervals was identified and is used by household surveys at Statistics Canada.

A weighted count is conceptually close to a proportion as a proportion is defined as the ratio of a weighted count to a population size. The weighted count has a lower bound of 0 and an upper bound equal to the number of units in the population. Basic simulations show that the usual confidence interval for weighted counts has the same undercoverage problem as for proportions: when the population total to be estimated is very small, or when it is close to the population size, the coverage of the Wald and Student confidence intervals drops below the nominal level. In this paper, we mainly address the problem of very small counts.

3.2 Proposed confidence intervals

A research project was undertaken at Statistics Canada to improve the method to produce confidence intervals for weighted counts from complex survey data (Neusy, 2021; Hidiroglou, 2020; Neusy, Savard, Hidiroglou and Martin, 2021), so as to obtain confidence intervals with proper coverage in the 2021 Census long-form disseminated products. By mimicking Wilson's approach for proportions (Wilson, 1927), the following general formula for estimating confidence intervals for weighted counts was obtained:

$$\frac{\hat{Y} + Nt^2/2n_e}{1 + t^2/n_e} \pm \frac{t\sqrt{\hat{Y}\left(N - \hat{Y}\right) + N^2t^2/4n_e}}{\sqrt{n_e}(1 + t^2/n_e)} \tag{1}$$

where $\hat{Y}$ is the weighted count estimated with design weights, N is the population size, n_e is the effective sample size and t is the Student quantile with an appropriate number of degrees of freedom. The effective sample size is given by:

$$n_e = \frac{n}{\widehat{Deff}_{SRSWR}\left(\hat{Y}\right)} \tag{2}$$

where n is the sample size and $\widehat{Deff}_{SRSWR}\left(\hat{Y}\right)$ is the estimated design effect with respect to simple random sampling with replacement (SRSWR), which is the sampling design assumed by the Wilson method. The estimated design effect is given by:

$$\widehat{Deff}_{SRSWR}\left(\hat{Y}\right) = \frac{\widehat{Var}\left(\hat{Y}\right)}{\widehat{Var}_{SRSWR}\left(\hat{Y}\right)} = \widehat{Var}\left(\hat{Y}\right)\frac{n}{\hat{Y}\left(N - \hat{Y}\right)} \tag{3}$$

Different definitions could be used for the population term, N, and the sample size term, n. The research project showed that for the Census long form, where the estimation uses calibrated weights for arbitrary domains, denoted as d, the population and sample size terms should be based on the calibration groups of interest. The calibration groups of interest are all the calibration groups that could potentially contain units having the characteristic of interest. Suppose that we denote the population size in the calibration groups of interest as $\widetilde{N}_I$ and the corresponding sample size as $\tilde{n}_I$. Then the confidence interval is given by:

$$\frac{\tilde{Y}_d + \widetilde{N}_I t^2/2n_e}{1 + t^2/n_e} \pm \frac{t\sqrt{\tilde{Y}_d\left(\widetilde{N}_I - \tilde{Y}_d\right) + \widetilde{N}_I^{\,2} t^2/4n_e}}{\sqrt{n_e}(1 + t^2/n_e)} \tag{4}$$

where $\tilde{Y}_d$ is the weighted count in domain d estimated with calibrated weights, and where n_e is calculated using formulas (2) and (3), but with $\widetilde{N}_I$ and $\tilde{n}_I$ in place of N and n.

Confidence intervals based on (4) cannot be produced by the existing Census tabulation system. This system was built under the assumption that all the design information is included in the replicate weights and it cannot identify the calibration group that contains any given domain. Although it would have been possible to include this requirement in the system, this would have taken too much time to implement as it would have meant rethinking the structure of the system.

One possible solution that was identified was to use an approximation of the confidence interval instead of formula (4). At the early stage of the project, an approximation that used an instrumental variable and the size of the domain was first proposed and tested. However, this approximation did not improve the coverage over the Wald interval in a number of studied scenarios.

After observing that using the population size and the sample size at the national level in formula (4) yielded very good results, the following approximation was proposed:

$$\tilde{Y}_d + t^2 \frac{1}{2} \frac{\widehat{Var}(\tilde{Y}_d)}{\tilde{Y}_d} \mp \sqrt{t^2 \widehat{Var}(\tilde{Y}_d) + \left(t^2 \frac{1}{2} \frac{\widehat{Var}(\tilde{Y}_d)}{\tilde{Y}_d} \right)^2} \tag{5}$$

This approximate interval is simple to implement in the existing system as it only depends on the estimate, its estimated variance, and the quantile from the Student-T distribution. As will be shown in section 4, the confidence interval given by (5) has good properties for the estimates based on the Census long form, where the sampling fraction is uniform across the strata. Unfortunately, this was not the case in more general simulation studies where the sampling fraction was different across strata (Neusy, 2021).

4. SIMULATION STUDY IN THE CENSUS LONG-FORM CONTEXT

4.1 Description of the simulations

A simulation study was undertaken to evaluate the properties of the confidence intervals in a context that recreates the Census long-form weighting and estimation strategy. In order to do so, we used a pseudo-population that was created from 2016 Census long-form respondents' data. It has a household-person hierarchical structure and allows for the estimation of characteristics from real data. The pseudo-regions in this population are about the same size as the calibration groups used in the Census long-form weighting process.

The results presented in section 4.2 were obtained with data from a pseudo-region composed of 7,804 persons in 3,270 households. This region corresponds to a single calibration group. To replicate long-form sampling and weighting, the following steps were repeated 500 times. A sample of households was selected with a simple random sampling without replacement (SRSWOR) design and a sampling fraction of ¼. The resulting sample includes all the persons living in the selected dwellings. Knowing that the long-form response rate is typically very high because it is a mandatory survey, no non-response was simulated. Calibration was performed with the long-form methodology: the calibration constraints are the total number of persons and the total number of households in the calibration group, as well as other totals from the Census that are selected with an automated procedure. For variance estimation, a set of 100 PBRR-epsilon replicates was created and the replicate weights were calibrated with the same calibration constraints used for the main weight.

Two sets of variables of interest were simulated with each set covering the entire range of the proportion underlying the count (from 1% to 99% of units having the characteristic). For the first set, denoted as x, the variables of interest are assigned independently to the persons regardless of their household, to simulate person-level characteristics with no intra-cluster correlation between members of the same household. For the second set, denoted as y, the variables of interest are equal for all the persons that belong to the same household. These correspond to household-level characteristics, which is the extreme case of perfect intra-cluster correlation between the members of the same household. In this simulation, the estimation is always performed at the person-level; the purpose of creating the second set is to study the effect of intra-household correlation on the properties of the confidence intervals.

Domains of various sizes were also simulated to assess the coverage of the confidence intervals in different scenarios. Some of the simulated domains are assigned at the person level whereas others are defined at the household level. The domains cover a certain percentage of persons (or of households) in the population, for example 1%, 10% or 30%.

For each combination of a domain and a variable of interest, three confidence intervals were produced: the Wald interval with a Student quantile, the proposed interval from formula (4) and the approximate interval from formula (5). To obtain the empirical coverage of the confidence intervals, the proportion of intervals that include the known population count was calculated. At the 95% confidence level, an appropriate confidence interval should contain the population parameter for 95% of the simulated samples.

4.2 Results

The graphs presented in Figure 1 show the coverage of confidence intervals for weighted counts for two domains assigned at the person level (the variable *d01pp* is a domain of about 1% of persons and the variable *d10pp* is a domain of about 10%

of persons) and for both sets of variables of interest (person-level on the left and household-level on the right). The x-axis indicates the proportion underlying the count, that is the count divided by the population size.

The red curve represents the coverage of the Student interval, where the degrees of freedom are defined as the minimum between the number of households that contribute to the estimate and the number of replicates used for variance estimation. It can be seen from the graphs that the coverage of the Student interval is too low for counts that have an underlying proportion close to 0, and especially when the size of the domain is small such as in the first row. The blue curve corresponds to the theoretical Wilson interval from formula (4) and the orange line to the approximate Wilson interval from formula (5). These two Wilson-type confidence intervals perform better than the Student interval and their coverage is always close to or above 95%. The two Wilson-type intervals perform similarly to each other; it seems to be appropriate to use the approximate interval instead of the theoretical interval.

Figure 1 – Coverage of different confidence intervals of counts estimated in random domains covering 1% (top) and 10% (bottom) of persons in the population, for variables of interest that are assigned independently to persons (left) and for variables of interest that are the same for every member of the same household (right)

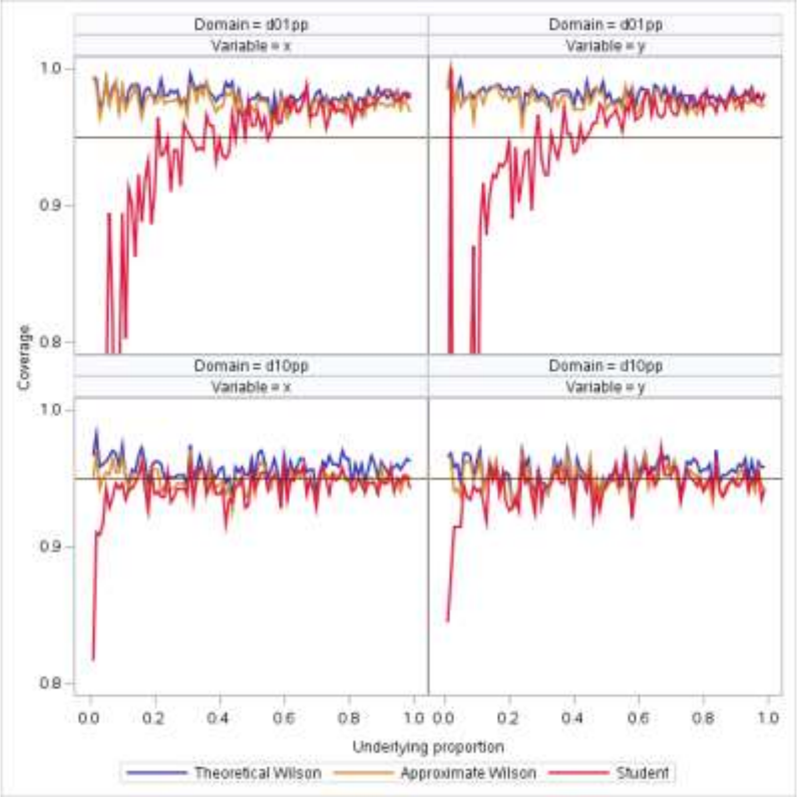
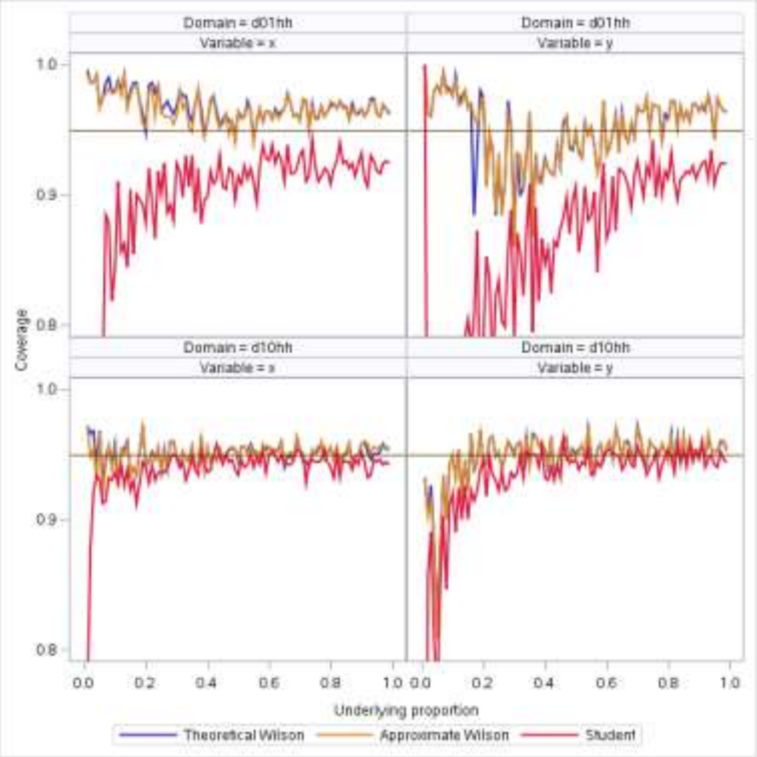


Figure 2 shows the coverage of the three confidence intervals for domains assigned at the household level (the variable *d01hh* is a domain of about 1% of households and the variable *d10hh* is a domain of about 10% of households) and for both sets of variables of interest (person-level on the left and household-level on the right).

The coverage of both Wilson-type intervals, represented by the blue and the orange curves, is close to or above 95% in the graphs on the left. For the smaller domain (top row), their coverage is much better than that of the Student interval. In the graphs on the right, which correspond to the extreme case of variables of interest with perfect intra-household correlation, the Wilson-type intervals perform better than the Student interval but there is undercoverage. In the top graph, the coverage is good for smaller counts; this is because we have to use a slightly different definition when the estimate is 0. The undercoverage appears to be related to a bias in the estimated design effect in formula (3). Some solutions were explored (e.g., Franco et al., 2019) but they were too complicated to implement in the Census tabulation system.

Although there is still room for improvement, the approximate Wilson interval will be used to produce confidence intervals for counts in the 2021 Census disseminated products as it performs better than the Student interval and has good coverage in most scenarios.

Figure 2 – Coverage of different confidence intervals of counts estimated in random domains covering 1% (top) and 10% (bottom) of households in the population, for variables of interest that are assigned independently to persons (left) and for variables of interest that are the same for every member of the same household (right)



ACKNOWLGEGMENT

The author would like to thank Elisabeth Neusy, Vincent Martin, Mike Hidiroglou, Emily Campling, Guylaine Dubreuil and Michelle Simard for their review and continuous support.

Disclaimer

The content of this article represents the views of the author and does not necessarily represent those of Statistics Canada.

REFERENCES

Devin, N. and Verret F. (2016). The Development of a Variance Estimation Methodology for Large-Scale Dissemination of Quality Indicators for the 2016 Canadian Census Long Form Sample. In JSM Proceedings.

Dippo, C. S., Fay, R. E. & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 489-494.

Franco, C., Little, R.J.A., Louis, T.A. and Slud, E.V. (2019). Comparative Study of Confidence Intervals for Proportions in Complex Surveys. Journal of Survey Statistics and Methodology, 7, 334-364.

Hidiroglou, M. A. (2020). Wilson Intervals for Proportions and Counts. Statistics Canada internal document.

Neusy, E. (2021). Wilson Confidence Intervals for Proportions and Totals of Binary Variables Using Complex Survey Data. Statistics Canada internal document.

Neusy E., Savard S.-A., Hidiroglou M., Martin V. (2021). Modified Wilson Confidence Intervals for Estimated Counts with Application to Census 2021 Long Form Estimation. Statistics Canada internal document.

Neusy E., and Mantel H. (2016). Confidence Intervals for Proportions Estimated from Complex Survey Data. Proceedings of the Survey Methods Section. SSC Annual Meeting, June 2016.

Wilson, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association, 22, 209-212