Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Improving the coverage of confidence intervals with respect to degrees of freedom: Application to the Canadian Census

by Marie-Hélène Toupin and Vincent Martin

Release date: June 29, 2026

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Survey Methodology, June 2026
Vol. 52, No. 1, pp. 187-211
Statistics Canada, Catalogue No. 12-001-X

Improving the coverage of confidence intervals with respect to degrees of freedom: Application to the Canadian Census

Marie-Hélène Toupin and Vincent Martin[1]

Abstract

Confidence intervals are very often constructed based on a probability distribution that uses a certain number of degrees of freedom as a parameter. This is the case with the Student and the modified Wilson confidence intervals, discussed in this article, which use quantiles from the Student distribution where the number of degrees of freedom is generally unknown. For the length of a confidence interval to be representative of the reliability of an estimate, the actual coverage rate must match the nominal rate. To that end, the number of degrees of freedom in the probability distribution used in practice to calculate the confidence interval must be estimated as precisely as possible. An approximate rule is often used, although it tends to overestimate the actual number of degrees of freedom. In this article, a more precise version of degrees of freedom, derived from the Satterthwaite approximation, is obtained in the context of the Canadian Census of Population. The sampling design is equivalent to a simple random design without replacement, cluster-stratified, and the variance estimation method is an adaptation of the balanced repeated replication method. An explicit expression of the degrees of freedom is obtained under these conditions, enabling the factors influencing them to be identified. For comparison, the degree of freedom formula is also established for the conventional variance estimator. A simulation study shows that using this version of degrees of freedom corrects the undercoverage problem observed with the approximate rule, showing the importance of accurately assessing this number.

Key Words: Balanced repeated replication method; Canadian Census of Population; Confidence intervals; Coverage; Degrees of freedom; Small area estimation.

1. Introduction

In Canada, the Census of Population takes place every five years. It consists of two questionnaires: a short form, which the entire population completes, and a long form, which is sent to a sample of households and covers more specific topics, such as daily activities, sociocultural information, education and labour market activities. In the 2021 Census, one-quarter of households were randomly selected to complete the long-form questionnaire.

For the release of 2021 Census data, Statistics Canada used 95% confidence intervals as reliability indicators for the estimates derived from the long-form questionnaire. For these intervals to accurately reflect reliability, the nominal confidence level must be observed. This means that the actual coverage rate of the confidence intervals should be at or above the selected confidence level.

Compliance with this nominal confidence level depends in particular on the type of confidence interval used, which must be suitable for the estimated parameter. In this article, two types of confidence intervals are examined. First, the Student confidence interval, developed for continuous variables of interest, may not be suitable for certain types of statistics, in particular when proportions are very low or very high. Second, the lesser-known modified Wilson confidence interval was specifically designed for estimating proportions or counts. This type of interval was used for the release of count estimates from the 2021 Canadian Census

1. Marie-Hélène Toupin, Senior Methodologist, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: marie-helene.toupin@statcan.gc.ca; Vincent Martin, Senior Methodologist, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: vincent.martin@statcan.gc.ca.

of Population long-form questionnaire. In both cases, calculating the confidence interval involves a generally unknown parameter: the number of degrees of freedom.

Both the variance estimation method and the degrees of freedom influence the coverage of confidence intervals. This article specifically focuses on the impact of degrees of freedom. Heeringa, West and Berglund (2010) explain the role of degrees of freedom: "Probability distributions such as the Student t, χ^2, and F play a critical role in the construction of confidence intervals for population values or as the reference distributions for formal tests of hypotheses concerning population parameters. Included in the quantities that define the shape of these distributions are degrees of freedom parameters. The degrees of freedom are indices of how precisely the true variance parameters of the reference distribution have been estimated from the sample design. Sample designs with large numbers of degrees of freedom for variance estimation enable more precise estimation of the true variance parameters of the reference distribution. [...] Precise determination of the degrees of freedom for variance estimation available under complex sample designs used in practice is difficult."

A common practice to bypass the fact that the number of degrees of freedom is unknown is to assume that the number is very large. In this case, the quantile of the Student distribution is replaced by the quantile of the standardized normal distribution. Another commonly used technique is to determine the number of degrees of freedom according to an approximate rule (or rule of thumb). For example, in a stratified design, a commonly used rule is the number of primary sampling units minus the number of strata. Both practices tend to overestimate the number of degrees of freedom, and this can lead to significant undercoverage.

The approach proposed in this article is to use the Satterthwaite (1946) approximation to estimate the degrees of freedom more accurately than what is obtained with the approximate rule. The general formula, valid for a given sample design and variance estimator, is used. Based on this formula, specific terms are derived for the Canadian long-form census questionnaire, based on a stratified cluster sample without replacement. The formulas, obtained to estimate a total over a domain, can be generalized for any linear estimator.

The more imprecise the variance estimator, the smaller the number of degrees of freedom. In this article, two variance estimation methods will be examined: the conventional variance estimator and a variant of the balanced repeated replication resampling method described by Devin and Verret (2016).

The specific formula resulting from the Satterthwaite approximation presents several advantages. First, the resulting algebraic expression enables a deeper understanding of the factors that influence degrees of freedom. Second, a simulation study shows that it corrects the undercoverage issues observed with the approximate rule for small domains. This is of particular interest because, in trying to best meet users' needs, Statistics Canada aims to disseminate information for the smallest domains possible.

This article is structured as follows. Section 2 details the Satterthwaite approximation, which provides a general formula for calculating degrees of freedom. The notation used is introduced in Section 3. Section 4 defines the Student and the modified Wilson confidence intervals. The variance estimation method used for the census is described in Section 5. Section 6 establishes the theoretical formula for the number of degrees of freedom for two specific variance estimation methods and proposes an estimator for this number. A comprehensive simulation study is then presented in Section 7. Lastly, Section 8 concludes the article.

2. Satterthwaite approximation

This section presents a formula for degrees of freedom in a general context. This formula was originally derived by Satterthwaite (1946) and proposed by Rust (1986) as a way of obtaining an expression for degrees of freedom. More details on the Satterthwaite approximation are provided in Valliant and Rust (2010) and Kott (2020).

Let $\hat{\theta}$ be an unbiased estimator of a population parameter θ and $v(\hat{\theta})$ be a variance estimator $\text{Var}_P(\hat{\theta})$ under the sample design P. Let $v(\hat{\theta})$ also depend on some random process *. For example, for a replication variance estimation method, * would represent the replicate creation process. Let $\text{Var}_{P,*}(\hat{\theta})$ be the variance of $\hat{\theta}$ in the sample design P and random process *.

Let $v(\hat{\theta})$ be an unbiased variance estimator $\text{Var}_P(\hat{\theta})$. If $dl \times v(\hat{\theta}) / \text{Var}_P(\hat{\theta})$ is a chi-square distribution, then the number of degrees of freedom associated with $v(\hat{\theta})$ is given by

$$dl = \frac{2\left\{\text{Var}_P(\hat{\theta})\right\}^2}{\text{Var}_{P,*}\left\{v(\hat{\theta})\right\}}. \tag{2.1}$$

The Satterthwaite approximation, described in equation (2.1), is a generic expression applicable to various sample designs, estimators and variance estimation methods. The conventional assumption under which $dl \times v(\hat{\theta}) / \text{Var}_P(\hat{\theta})$ is a chi-square distribution, which is at the origin of this equation, may not be verified when complex survey data are involved. The value of dl given in equation (2.1) will be called the "theoretical number of degrees of freedom" in the remainder of this article. The algebraic expression of dl is derived in Section 6 for two variance estimation methods, in the specific context described in the next section.

3. Framework and notation

The rest of the article assumes a sample design similar to that used for the 2021 Census long-form questionnaire. First, households, the primary sampling units, are grouped into a certain number of strata. One in four households is then randomly selected without replacement in each stratum. Lastly, all the individuals from the selected households are part of the sample.

Let U be a certain stratified population with H strata. Assume that this population is made up of N households containing M individuals. Let N_h be the number of households in stratum h; M_{hi} be the number of individuals in household i of stratum h; and y_{hij} be the variable of interest for individual j in household i of stratum h, $h = 1,\ldots,H, i = 1,\ldots,N_h$, $j = 1,\ldots,M_{hi}$. The following is defined

$$N = \sum_{h=1}^{H} N_h \quad \text{and} \quad M = \sum_{h=1}^{H} M_h,$$

where $M_h = \sum_{i=1}^{N_h} M_{hi}$. Let $D \subseteq U$ be a certain domain and z_{Dhij} the indicator of individual j's membership in the domain. Note $M_D \leq M$, the number of individuals in the population belonging to this domain, i.e.,

$$M_D = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} z_{Dhij}.$$

The total of y_{hij} over domain D of all the individuals in household i of stratum h is noted as

$$t_{Dhi} = \sum_{j=1}^{M_{hi}} y_{hij} z_{Dhij}. \tag{3.1}$$

The parameter of interest is total Y_D over domain D, i.e.,

$$Y_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} t_{Dhi}.$$

A simple random sample without replacement s_h, composed of n_h households containing m_h individuals, is selected in stratum h. Let I_{hi} be the sample inclusion indicator of household i in stratum h. The number of individuals in sample s over domain D, whose value is random, is noted as

$$m_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} I_{hi} m_{Dhi},$$

where $m_{Dhi} = \sum_{j=1}^{M_{hi}} z_{Dhij}$. Subsequently, only the case where $m_D \geq 1$ is considered. Let also n_D be the number of households with at least one individual belonging to domain D in sample s, i.e.,

$$n_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} I_{hi} J_{Dhi}, \tag{3.2}$$

where J_{Dhi} equals 1 if $m_{Dhi} > 0$, and otherwise 0. The Narain-Horvitz-Thompson estimator of Y_D is

$$\hat{Y}_D = \sum_{h=1}^{H} w_h \sum_{i=1}^{N_h} I_{hi} t_{Dhi}, \tag{3.3}$$

where $w_h = N_h / n_h$. For logistical reasons, this type of estimator is used for Census of Population estimates. Let σ_{Dh}^2 and m_{Dh4} be, respectively, the variance and the fourth central moment of totals t_{Dhi} in stratum h, which are fixed in relation to the sample design, i.e.,

$$\sigma_{Dh}^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (t_{Dhi} - \mu_{Dh})^2 \quad \text{and} \quad m_{Dh4} = \frac{1}{N_h} \sum_{i=1}^{N_h} (t_{Dhi} - \mu_{Dh})^4, \tag{3.4}$$

where

$$\mu_{Dh} = \frac{1}{N_h} \sum_{i=1}^{N_h} t_{Dhi}.$$

The exact form of the variance of $\hat{Y}_D$ in relation to sample design P is

$$\mathrm{Var}_P(\hat{Y}_D) = \sum_{h=1}^{H} \frac{N_h^2 (N_h - n_h)}{n_h (N_h - 1)} \sigma_{Dh}^2. \tag{3.5}$$

Since variances σ_{Dh}^2, $h = 1, \ldots, H$, are generally unknown, $\mathrm{Var}_P(\hat{Y}_D)$ must be estimated. The conventional unbiased estimator for this variance is

$$v_{cl}(\hat{Y}_D) = \sum_{h=1}^{H} N_h \left(\frac{N_h - n_h}{n_h} \right) S_{Dh}^2, \quad \text{where} \quad S_{Dh}^2 = \frac{1}{n_h - 1} \left\{ \sum_{i=1}^{n_h} t_{Dhi}^2 - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} t_{Dhi} \right)^2 \right\}. \tag{3.6}$$

4. Construction of a confidence interval

There are several ways to build a confidence interval. The Student confidence interval is based on the distribution of the pivot

$$t_{\text{pivot}} = \frac{\hat{Y}_D - Y_D}{\sqrt{v(\hat{Y}_D)}}, \tag{4.1}$$

where $v(\hat{Y}_D)$ is an unbiased estimator of $\text{Var}_P(\hat{Y}_D)$. Assuming that t_{pivot} follows a Student distribution with a certain number of degrees of freedom $dl > 0$, a confidence interval $1-\alpha$ is then obtained by using two values from the distribution of t_{pivot}, such as $P(t_1 < t_{\text{pivot}} < t_2) = 1-\alpha$. The two bounds of the confidence interval (BI the lower and BS the upper) are then deduced algebraically such that $P\big(BI < Y_D < BS\big) = 1-\alpha$. Choosing t_1 and t_2 symmetrically leads to the Student confidence interval of level $1-\alpha$ that takes the form

$$\hat{Y}_D \pm t_{dl,1-\alpha/2}\sqrt{v(\hat{Y}_D)}, \tag{4.2}$$

where $t_{dl,1-\alpha/2}$ is the Student quantile with dl degrees of freedom. As a special case, the Wald confidence interval, which is widely used, is obtained when the number of degrees of freedom is very large. When a variable of interest is dichotomous, the Student confidence interval may not be appropriate. It can, for instance, generate bounds that fall outside the logical limits for a count. For example, when $\hat{Y}_D$ is close to zero, the lower bound given in equation (4.2) may become negative, resulting in undercoverage. This problem is especially pronounced when the sample size is small.

The modified Wilson confidence interval was originally proposed by Wilson (1927) for estimating a proportion in a simple random sample with replacement. Kott and Carr (1997) then adapted this method to make it valid for estimating proportions derived from complex survey data, such as for sample designs without replacement. More recently, Neusy, Savard, Hidiroglou and Martin (2021) extended this approach to apply it to estimating counts. Unlike the Student confidence interval, the modified Wilson confidence interval bounds are generally asymmetric around estimate $\hat{Y}_D$, and the method guarantees an interval always within logical limits. This asymmetry results from the use of a different pivot than that shown in equation (4.1); for more details, see Neusy et al. (2021). Furthermore, their simulations have shown that this type of interval often improves coverage compared with other methods, notably the Student confidence interval.

Let $\mathcal{H}_D \subseteq \{1,\ldots,H\}$ be all the indices of the strata that are not empty, i.e., those where at least one individual is part of the domain. $M_I \leq M$ and $m_I \leq m$ are defined so that

$$M_I = \sum_{h\in\mathcal{H}_D} M_h \quad \text{and} \quad m_I = \sum_{h\in\mathcal{H}_D} m_h.$$

Let $\widehat{EP}(\hat{Y}_D)$ be the estimated design effect compared with the simple random design with replacement, i.e., $\widehat{EP}(\hat{Y}_D) = v(\hat{Y}_D)\, m_I \big/ \hat{Y}_D(M_I - \hat{Y}_D)$. The modified Wilson confidence interval of level $1-\alpha$ is given by

$$\frac{\hat{Y}_D + M_I a_D/2}{1+a_D} \pm \frac{\sqrt{\hat{Y}_d(M_I - \hat{Y}_D)\, a_D + M_I^2 a_D^2/4}}{1+a_D}, \tag{4.3}$$

where $a_D = t^2_{dl,1-\alpha/2} / m_e$ and $m_e = m_1 / \widehat{EP}(\hat{Y}_D)$ is the effective sample size. Note that the estimated design effect $\widehat{EP}(\hat{Y}_D)$ is not defined for extreme cases $\hat{Y}_D = 0$ and $\hat{Y}_D = M_1$. These extreme cases are simply excluded from the simulations in Section 7.

In the case of the Student confidence interval, assuming that t_{pivot} follows a Student distribution comes from assuming that $dl \times v(\hat{\theta}) / \text{Var}_P(\hat{\theta})$ follows a chi-square distribution. Therefore, the Satterthwaite approximation provides the appropriate number of degrees of freedom to use in this context. However, this correspondence is not necessarily assured for the modified Wilson confidence interval, where the pivot assumptions differ, although this same value of dl is generally used in this case.

Since the variance estimation influences the degrees of freedom, it is crucial to pay special attention to the method used to obtain this estimate. The variance estimation method used to estimate the long-form census questionnaire is described in detail in the following section.

5. Partially balanced repeated replication epsilon (PBRR-ε) method

The balanced repeated replication (BRR) method, formalized by McCarthy (1966), is a replication variance estimation technique, originally intended designed for a stratified random sample design with replacement involving exactly two primary units per stratum. For more details, see Wolter (2007). The method used for the estimation of the long-form census questionnaire, called PBRR-ε and described in Devin and Verret (2016), is an adaptation of the BRR method that is valid for complex sample designs such as the one used for the census. This variance estimation technique is presented here.

5.1 Restructuring of households in the sample

Since the number of households by stratum n_h is generally greater than 2 in the case of the census, the households in the sample must be restructured to apply the PBRR-ε method. Wolter (2007) suggests two solutions for the case where $n_h > 2$: divide each stratum into two clusters and choose one per replicate or create several artificial strata of two households and select one per replicate. The first solution is computationally advantageous but is costly in terms of loss of information, while the second results in less information loss but is more costly from a computational standpoint. Devin and Verret (2016) have adopted a compromise between these two approaches, as illustrated in steps 1 and 2 of Figure 5.1. The composition of the balancing groups (Step 3 in Figure 5.1, in red) will be detailed in Section 5.3.

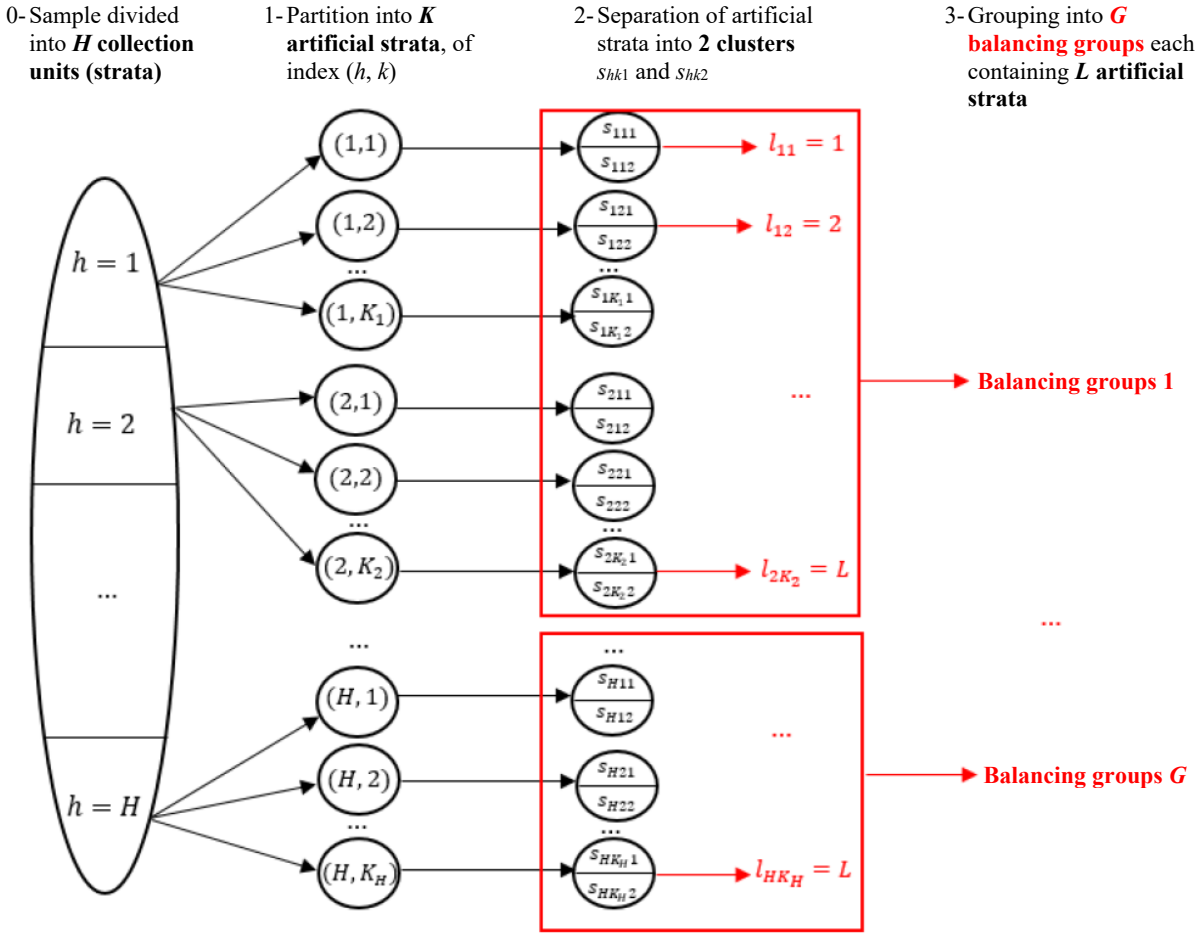
First, the n_h households in each stratum h are randomly divided into K_h groups known as "artificial strata". Each artificial stratum (h, k), with $k = 1, \ldots, K_h$, contains n_{hk} households. Thus, all of the n households in the sample are divided into $K = \sum_{h=1}^{H} K_h$ artificial strata. This distribution is performed so that all sizes n_{hk} of the artificial strata are roughly equal. Second, the households in each artificial stratum (h, k) are randomly divided into two groups, called "clusters", of approximately equal size. Let s_{hk1} and s_{hk2} be all the households respectively belonging to clusters 1 and 2 of artificial stratum (h, k). Thus, $s_h = \bigcup_{k=1}^{K_h} (s_{hk1} \cup s_{hk2})$.

Let $t_{Dhk1} = \sum_{i \in s_{hk1}} t_{Dhi}$ and $t_{Dhk2} = \sum_{i \in s_{hk2}} t_{Dhi}$ be the totals over domain D for clusters 1 and 2 of artificial stratum (h, k), where t_{Dhi} is defined in equation (3.1). Each replicate r, $r = 1, \ldots, R$, consists in selecting only one of the two clusters (s_{hk1} or s_{hk2}) for each artificial stratum (h, k), and then calculating $\hat{Y}_D^{(r)}$, an estimation of the total Y_D, via

$$\hat{Y}_D^{(r)} = \sum_{h=1}^{H} \sum_{k=1}^{K_h} \left(w_{hk1}^{(r)} t_{Dhk1} + w_{hk2}^{(r)} t_{Dhk2} \right), \tag{5.1}$$

where $w_{hk1}^{(r)}$ and $w_{hk2}^{(r)}$ are called replicate weights and are detailed below.

Figure 5.1 Steps for creating artificial strata, clusters and balancing groups for the PBRR-ε method



Note: Partially balanced repeated replication (PBRR).

5.2 Definition of the replicate weights

In the conventional case, the replicate weights are equivalent to 0, i.e., $2w_h$ depending on the selected cluster. Thus, only half of the sample contributes to the estimation of the total for each replicate, with the other half being given a zero weight. A working definition of the replicate weights is then $(1 + \delta_{hk}^{(r)})\, w_h$ or $(1 - \delta_{hk}^{(r)})\, w_h$, with

$$\delta_{hk}^{(r)} = \begin{cases} 1 & \text{if } s_{hk1} \text{ is selected in the rth replicate} \\ -1 & \text{if } s_{hk2} \text{ is selected in the rth replicate.} \end{cases}$$

As Rao and Shao (1999) point out, the significant perturbation of the replicate weights in relation to the original weights w_h can pose some problems, such as when the replicate weights are calibrated. To mitigate this effect, a perturbation parameter $\varepsilon \in (0,1)$ is introduced and the replicate weights become $(1-\varepsilon\delta_{hk}^{(r)})\, w_h$ or $(1+\varepsilon\delta_{hk}^{(r)})\, w_h$. This variant, called Fay's BRR method or BRR-ε, is detailed by Judkins (1990) and enables all the households in the sample to participate in each replicate. Devin and Verret (2016) suggest using $\varepsilon = \sqrt{1/2}$; this value, employed for the estimation of the long-form census questionnaire, will be used hereafter.

Two additional modifications are applied to the replicate weights. First, as proposed by Rao and Shao (1999), to avoid bias related to potential differences in size between clusters s_{hk1} and s_{hk2}, adjustment factors $F_{hk1} = n_{hk2} / n_{hk1}$ and $F_{hk2} = n_{hk1} / n_{hk2}$ are introduced, where n_{hk1} and n_{hk2} are the respective sizes of clusters s_{hk1} and s_{hk2}.

Second, in our context of a large sampling fraction, the finite population correction factor $(1-n_h / N_h) = (1-1/w_h)$, as suggested by Wolter (2007), is also applied. Thus, the replicate weights used in the PBRR-ε method are written

$$w_{hk1}^{(r)} = \left(1+\delta_{hk}^{(r)}\,\varepsilon\sqrt{F_{hk1}\left(1-\frac{1}{w_h}\right)}\right) w_h \quad \text{and} \quad w_{hk2}^{(r)} = \left(1-\delta_{hk}^{(r)}\,\varepsilon\sqrt{F_{hk2}\left(1-\frac{1}{w_h}\right)}\right) w_h.$$

It is important to note that replicate weights $w_{hk1}^{(r)}$ and $w_{hk2}^{(r)}$ thus defined observe the following consistency over the total sum of the weights, i.e.,

$$\sum_{k=1}^{K_h} (n_{hk1} w_{hk1}^{(r)} + n_{hk2} w_{hk2}^{(r)}) = N_h.$$

5.3 Reduction in the number of replicates

The conventional balanced repeated replication method considers all possible $R = 2^K$ half-samples, and this is very costly from a computational standpoint. The "balanced" method reduces this number by selecting replicates in a structured way, such that $\sum_{r=1}^{R} \delta_{hk}^{(r)} = 0$ and $\sum_{r=1}^{R} \delta_{hk}^{(r)} \delta_{h'k'}^{(r)} = 0$ for all distinct pairs of artificial strata $(h,k) \neq (h',k')$. In practice, values of $\delta_{hk}^{(r)} \in \{-1,+1\}$ are determined using an $R \times R$ Hadamard matrix. See the article by Hedayat and Wallis (1978) for more details on Hadamard matrices. According to Wolter (2007), the balanced method sets R at the smallest multiple of 4 greater than K, thus reducing the number of replicates while ensuring a variance estimator equivalent to the conventional estimator for linear statistics. However, when K is large, implementation remains complex from a computational standpoint.

The partially balanced repeated replication (PBRR) method, defined by Wolter (2007), further reduces the number of replicates. Used for the estimation of the long-form census questionnaire (Devin and Verret, 2016), this method consists of randomly forming $G \geq 2$ "balancing groups" by grouping the artificial strata

of several strata to obtain an $L = K / G$ equal number of artificial strata per group. Each artificial stratum (h, k) is given a random index $l_{hk} \in \{1, \ldots, L\}$ within its group. The creation of the balancing groups and the numbering of the artificial strata from 1 to L are depicted in Figure 5.1 (Step 3, in red). The number of R replicates is the smallest multiple of 4 greater than L. The replicates are structured as in the BRR method, but within each balancing group, while meeting the following three conditions:

1) $\sum_{r=1}^{R} \delta_{hk}^{(r)} = 0$ for all artificial strata (h, k)
2) $\sum_{r=1}^{R} \delta_{hk}^{(r)} \delta_{h'k'}^{(r)} = 0$ for all pairs of artificial strata $(h, k) \neq (h', k')$ such that $l_{hk} \neq l_{h'k'}$
3) $\delta_{hk}^{(r)} = \delta_{h'k'}^{(r)}$ for all pairs of artificial strata $(h, k) \neq (h', k')$ such that $l_{hk} = l_{h'k'}$ and all $r = 1, \ldots, R$.

The first two conditions ensure that the replicates are balanced for each group, while the third condition structures the replicates in the same way within the a balancing group. Lastly, the variance estimator of the PBRR-ε method is given by

$$v_{PBRR}(\hat{Y}_D) = \frac{1}{R\varepsilon^2} \sum_{r=1}^{R} (\hat{Y}_D^{(r)} - \hat{Y}_D)^2, \tag{5.2}$$

where $\hat{Y}_D^{(r)}$ is defined in equation (5.1). The estimator $v_{PBRR}(\hat{Y}_D)$ is not equivalent to the conventional estimator v_{cl} given in equation (3.6). However, as demonstrated in Appendix A, $E_*\{v_{PBRR}(\hat{Y}_D)\} = v_{cl}$, where E_* refers to the expectation under the random replicate creation process. This result implies that the estimator $v_{PBRR}(\hat{Y}_D)$ is unbiased for $\mathrm{Var}_P(\hat{Y}_D)$, since v_{cl} is also unbiased.

6. Degrees of freedom

In this section, the formula for the theoretical number of degrees of freedom is obtained for the conventional variance estimator $v_{cl}(\hat{Y}_D)$ and for $v_{PBRR}(\hat{Y}_D)$, as part of a stratified cluster (household) sample without replacement. The algebraic expressions obtained in propositions 6.1 and 6.2 are derived directly from equation (2.1). They are obtained for the estimator of total $\hat{Y}_D$ but can be generalized to any linear estimator. Estimated and approximate versions of the number of degrees of freedom are also presented.

6.1 Degrees of freedom with the conventional variance estimation method

The following proposition provides the theoretical number of degrees of freedom associated with $v_{cl}(\hat{Y}_D)$, the conventional variance estimator defined in equation (3.6).

Proposition 6.1: Let $V_1 = \mathrm{Var}_P(\hat{Y}_D)$ be as defined in equation (3.5) and

$$V_{2,cl} = \sum_{h=1}^{H} \frac{N_h^3 (N_h - n_h)^3}{n_h^3 (n_h - 1)(N_h - 1)^2 (N_h - 2)(N_h - 3)} \times \left[m_{Dh4} (N_h - 1)\{N_h(n_h - 1) - (n_h + 1)\} - \sigma_{Dh}^4 \{N_h^2 (n_h - 3) + 6N_h - 3(n_h + 1)\} \right].$$

The theoretical number of degrees of freedom associated with $v_{cl}(\hat{Y}_D)$ is $dl_{cl} = 2V_1^2 / V_{2,cl}$.

The result of proposition 6.1, the proof of which can be found in Appendix B, can also be applied to other types of variance estimators that are equivalent to $v_{cl}(\hat{Y}_D)$. This is the case with the jackknife and bootstrap method when the number of replicates is very large.

6.2 Degrees of freedom with the PBRR-ε method

The algebraic expression of the theoretical number of degrees of freedom is now derived when the variance estimator is $v_{PBRR}(\hat{Y}_D)$, as provided in equation (5.2).

Proposition 6.2: Let $V_1 = \mathrm{Var}_P(\hat{Y}_D)$ be as defined in equation (3.5) and

$$\begin{aligned} V_{2,PBRR} &= \sum_{h=1}^{H} \frac{N_h^3 (N_h - n_h)^2}{n_h^4 (N_h - 1)(N_h - 2)(N_h - 3)} \left\{ N_h (N_h + 1) B_h - (N_h - 1)(n_h^2 + 2A_h) \right\} m_{Dh4} \\ &+ \sum_{h=1}^{H} \frac{N_h^3 (N_h - n_h)^2}{n_h^4 (N_h - 1)(N_h - 2)(N_h - 3)} \left\{ 2A_h (N_h^2 - 3N_h + 3) + \frac{n_h^2 (N_h^2 - 3)}{N_h - 1} - 3N_h (N_h - 1) B_h \right\} \sigma_{Dh}^4 \\ &+ \frac{G-1}{K-1} \sum_{h \neq h'} \frac{N_h^2 N_{h'}^2 (N_h - n_h)(N_{h'} - n_{h'})}{n_h n_{h'} (N_h - 1)(N_{h'} - 1)} \sigma_{Dh}^2 \sigma_{Dh'}^2, \end{aligned}$$

where

$$A_h = \sum_{k=1}^{K_h} n_{hk}^2 \quad \text{and} \quad B_h = \sum_{k=1}^{K_h} \left(\frac{n_{hk1}^2}{n_{hk2}} + \frac{n_{hk2}^2}{n_{hk1}} \right). \tag{6.1}$$

The theoretical number of degrees of freedom associated with $v_{PBRR}(\hat{Y}_D)$ is

$$dl_{PBRR} = \frac{2V_1^2}{V_{2,PBRR}}. \tag{6.2}$$

The proof of proposition 6.2 is presented in Appendix C. Recall that G and K respectively represent the number of balancing groups and the number of artificial strata associated with the use of the PBRR-ε method. Since $v_{PBRR}(\hat{Y}_D)$, as defined in equation (5.2), is expressed as a sum of squares, it is reasonable to assume that this estimator is distributed according to a chi-square rule when the sample size is large enough.

As shown in Appendix A, $E_*\{v_{PBRR}(\hat{Y}_D)\} = v_{cl}(\hat{Y}_D)$. Consequently, variance $V_{2,PBRR}$ in the denominator of dl_{PBRR} is

$$V_{2,PBRR} = \mathrm{Var}_P\{v_{cl}(\hat{Y}_D)\} + E_P\left[\mathrm{Var}_*\{v_{PBRR}(\hat{Y}_D)\}\right] \geq \mathrm{Var}_P\{v_{cl}(\hat{Y}_D)\} = V_{2,cl},$$

where $V_{2,cl}$ is as defined in proposition 6.1. As a result, $dl_{PBRR} \leq dl_{cl}$. Thus, dl_{cl} is an upper bound for dl_{PBRR}, and the difference between the two represents the loss in degrees of freedom because of the use of the PBRR-ε method compared with the conventional method.

Equation (6.2) does not allow the factors that influence the number of degrees of freedom to be easily identified. For an easier interpretation, assume that all quantities within each stratum are equal, i.e., $N_h = N_{h'}$, $n_h = n_{h'}$, $\sigma_{Dh}^2 = \sigma_{Dh'}^2$, $m_{Dh4} = m_{Dh'4}$, $B_h = B_{h'}$ and $A_h = A_{h'}$, for every $h \neq h' \in \{1, \ldots H\}$. Assume also that the stratum size is very large, i.e., $N_h \to \infty$. Under these conditions, the formula for dl_{PBRR} is reduced to

$$dl_{PBRR} \approx \frac{2H}{B_h\kappa_{2D} + 2A_h/n_h^2 + (G-1)(H-1)/K - 1},$$

where $\kappa_{2D} = m_{Dh4}/\sigma_{Dh}^4 - 3 \geq -2$ is the excess kurtosis coefficient of totals t_{Dhi}. This simplified formula shows, on the one hand, that a large sample size n_h contributes to increasing dl_{PBRR}. On the other hand, dl_{PBRR} decreases as κ_{2D} increases. A large excess kurtosis indicates that the distribution of totals t_{Dhi} has a pronounced peak around the mean and heavy distribution tails. When the domain size is small, the probability of observing zero totals increases, and this contributes to increasing κ_{2D} and therefore to decreasing dl_{PBRR}.

This simplified formula also helps identify ways to optimize the PBRR-ε method. First, imposing artificial stratum sizes n_{hk} that are as equal as possible, as well as selecting cluster sizes n_{hk1} and n_{hk2} that are as similar as possible within these strata, reduces the loss of degrees of freedom by minimizing the values of A_h and B_h. Second, maintaining a small number of balancing groups G and a large number of artificial strata K also helps reduce this loss. However, choosing a low G and large K involves a greater number of replicates R. Thus, as expected, a large number of replicates leads to a larger number of degrees of freedom.

6.3 Estimation of dl_{PBRR}

Calculating dl_{PBRR} requires that parameters σ_{Dh}^2 and m_{Dh4} be known, as defined in equation (3.4), for each stratum $h = 1, \ldots, H$. These parameters are generally unknown. Let $\bar{t}_{Dhi} = \sum_{i \in s_h} t_{Dhi}/n_h$. Unbiased estimators of σ_{Dh}^2 and m_{Dh4}, as reported by Gerlovina and Hubbard (2019), are respectively given by

$$\hat{\sigma}_{Dh}^2 = \frac{1}{n_h - 1}\sum_{i \in s_h}(t_{Dhi} - \bar{t}_{Dhi})^2$$

and

$$\hat{m}_{Dh4} = \frac{1}{(n_h-1)(n_h-2)(n_h-3)}\left\{(n_h^2 - 2n_h + 3)\sum_{i \in s_h}(t_{Dhi} - \bar{t}_{Dhi})^4 - \frac{3(2n_h-3)(n_h-1)}{n_h}\hat{\sigma}_{Dh}^2\right\}.$$

Thus, an estimator of dl_{PBRR}, denoted as $\widehat{dl}_{PBRR}$, consists of the formula for the theoretical number of degrees of freedom dl_{PBRR}, given in equation (6.2), for which σ_{Dh}^2 and m_{Dh4} are simply replaced with $\hat{\sigma}_{Dh}^2$ and $\hat{m}_{Dh4}$, respectively. An empirical version of dl_{cl}, the expression of which also requires knowing σ_{Dh}^2 and m_{Dh4}, could be obtained in the same way.

6.4 Approximate rules

In practice, degrees of freedom are often determined using an approximate rule. The standard rule, which is generally used, consists of subtracting the number of strata from the number of primary sampling units. As reported by Valliant and Rust (2010), this rule is a special case for the theoretical number of degrees of freedom obtained in proposition 6.1, the precision of which is based on several assumptions. Among these,

it is assumed that the sampling fractions are negligible in each stratum, that the variable of interest follows a normal distribution and that the variances for each stratum are equivalent. When these assumptions are not met, the standard rule may result in a substantial overestimation of the actual number of degrees of freedom. This problem is all the more likely when the sample size is small and can be particularly significant for sample designs with a single degree of freedom.

Because of operational constraints, a slightly different approximate rule is used for the estimation of the 2021 Census long-form questionnaire. It consists of

$$dl_{App} = \min\left(n_D, R\right), \tag{6.3}$$

where R is the number of replicates for the PBRR-ε variance estimation method, and n_d is as defined in equation (3.2).

In the next section, four versions of the degrees of freedom are compared in a simulation study, i.e., dl_{PBRR}, $\widehat{dl}_{PBRR}$, dl_{cl} and dl_{App}.

7. Simulations

In this section, the results of a simulation study are presented to assess the potential gain from using the theoretical number of degrees of freedom dl_{PBRR} as defined in equation (6.2) compared with the approximate version dl_{App} given in equation (6.3). The efficiency of the estimator $\widehat{dl}_{PBRR}$ and the reduction in the number of degrees of freedom caused by using v_{PBRR} compared with v_{cl} are also assessed. The simulation scenarios are built to be as realistic as possible in relation to the long-form census questionnaire estimates.

7.1 Description of simulation scenarios

First, a fictitious population was generated. Made up of $H = 22$ strata of varying sizes, this population consists of $N = 4{,}030$ households with a total of $M = 11{,}227$ individuals. For each scenario described below, 3,000 independent samples were selected from the fictitious population. These samples were drawn using a simple random design without replacement, with a 1/4 sampling ratio in each of the 22 strata.

To apply the PBRR-ε method, a total of 198 artificial strata were created by dividing, for each of the 22 strata, the selected households into 9 artificial strata of approximately equal size; $G = 2$ balancing groups were formed, each containing $L = 99$ artificial strata. The number of replicates was set at $R = 100$. This restructuring into 99 artificial strata and the use of 100 replicates are consistent with the procedure used for variance estimation in the 2021 Census. The value of ε was set at $\sqrt{1/2}$.

Domain D was generated at the household level, i.e., the indicator of membership in D is identical for all members of a household. In other words, $z_{Dhij} = z_{Dhi} \in \{0, 1\}$. Three domains with different sizes were

generated. Thus, z_{Dhi} was randomly generated independently for each household in the population with inclusion probabilities of 3%, 5% or 10%. Table 7.1 shows the sizes of the three domains.

Table 7.1
Number of individuals and households in the fictitious population in domain *D*

Domain	Number of individuals	Number of households
3%	338	121
5%	570	204
10%	1,136	428

Note that additional simulations were also performed for a domain where z_{Dhij} was independently generated for each individual in a household. The results of these simulations are similar to those obtained with the domains in Table 7.1; this is why they are not presented here.

For the Census of Population, intra-household dependence is often observed, meaning that the variables of interest for members of a household are dependent. This is the case, for example, for language-related variables. Thus, all the simulation scenarios were implemented first without, and then with, intra-household dependence. First, for scenarios with intra-household independence, the variable y_{hij} is generated independently of $y_{hij'}$, for all $j \neq j'$. Second, for scenarios with intra-household dependence, the variable of interest (dichotomous or continuous) was randomly generated using a normal copula and, by considering a dependence such as Spearman's rho between y_{hij} and $y_{hij'}$, for all $j \neq j'$, is 0.95, which corresponds to a Pearson correlation coefficient of 0.954. See Nelsen (2006) for more details on copulas. No dependence between members of separate households was considered.

For each of the 3,000 samples, calibration was first performed at the level of each stratum. Calibration weights $\tilde{w}_{hi}$, $h = 1, \ldots, H, i = 1, \ldots, N_h$, were obtained to minimize the sum of the relative differences under constraints

$$\sum_{i=1}^{N_h} \tilde{w}_{hi} I_{hi} = N_h \quad \text{and} \quad \sum_{i=1}^{N_h} \sum_{j=1}^{M_h} \tilde{w}_{hi} I_{hi} = M_h.$$

These two constraints ensure that the sum of the weights at the household level is equal to the actual number of households N_h in stratum h and that the sum of the weights at the individual level is equal to the actual number of individuals M_h in stratum h. The estimator $\tilde{Y}_D$ of the total Y_D used for the simulations is obtained by replacing the original weights with the calibration weights in equation (3.3), i.e.,

$$\tilde{Y}_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \tilde{w}_{hi} I_{hi} t_{Dhi}.$$

The variance of $\tilde{Y}_D$ was estimated using the PBRR-ε method, as described in Section 5. All replicate weights were also calibrated to represent as accurately as possible the procedures applied during the census. It is difficult to assess the impact of calibration on the number of degrees of freedom, since the calibration estimators are not equivalent to the Horvitz-Thompson linear estimator, even asymptotically. Although there is no guarantee that this impact is negligible, no adjustment has been made in the theoretical formulas of degrees of freedom to account for calibration in the simulations. This exercise is left for future consideration. Calibration was nonetheless included in the simulations to accurately replicate the census operational procedures.

For each scenario, the 95% confidence intervals were calculated using the four versions of the degrees of freedom, i.e., dl_{PBRR}, $\widehat{dl}_{PBRR}$, dl_{cl} and dl_{App}. The coverage of the confidence intervals is estimated by the proportion of the 3,000 samples for which the confidence interval contains the true value of the population total Y_D.

Note that the value of dl_{PBRR}, as well as the value of dl_{cl}, is fixed for all 3,000 samples in the same scenario. The parameters σ^2_{Dh} and m_{Dh4} that are involved in the calculation of dl_{PBRR} are obtained at the population level, while the other elements, such as n_h and N_h, depend on the sampling design, which is fixed for the 3,000 replicates. Conversely, the values of $\widehat{dl}_{PBRR}$ and dl_{App} vary from sample to sample because they depend on quantities calculated at the sample level – the number of households in domain n_D for dl_{App} and $\hat{\sigma}^2_{Dh}$ and $\hat{m}_{Dh4}$ for $\widehat{dl}_{PBRR}$.

In some cases where dl_{PBRR} was very small, coverage was significantly greater than the nominal rate (close to 100%). One possible explanation is that the assumption under which $dl \times v(\hat{\theta}) / \mathrm{Var}_P(\hat{\theta})$ has a chi-square distribution may not be valid for some limiting cases. Thus, dl_{PBRR}, $\widehat{dl}_{PBRR}$ and dl_{cl} were bounded below 2.5 degrees of freedom for the simulation study.

Different types of variables of interest were generated: continuous and dichotomous variables. Section 7.2 presents the results for estimating a total of continuous variables, while Section 7.3 presents the results for a count estimation.

7.2 Simulation results for continuous variables

The following three continuous probability distributions were considered: the normal distribution with mean 5 and variance 4, the Weibull distribution with form parameter 0.5 and scale parameter 1, and the Gamma distribution with form parameter 0.5 and scale parameter 1. The distribution of these three variables at the population level is shown in Figure 7.1 with intra-household independence. Additional simulations were also performed for exponential, Fisher and Student variables. The results are similar to those reported in this section and are not presented here.

Figure 7.1 Distribution of the three continuous variables for the 11,227 individuals in the population, scenarios without intra-household dependence

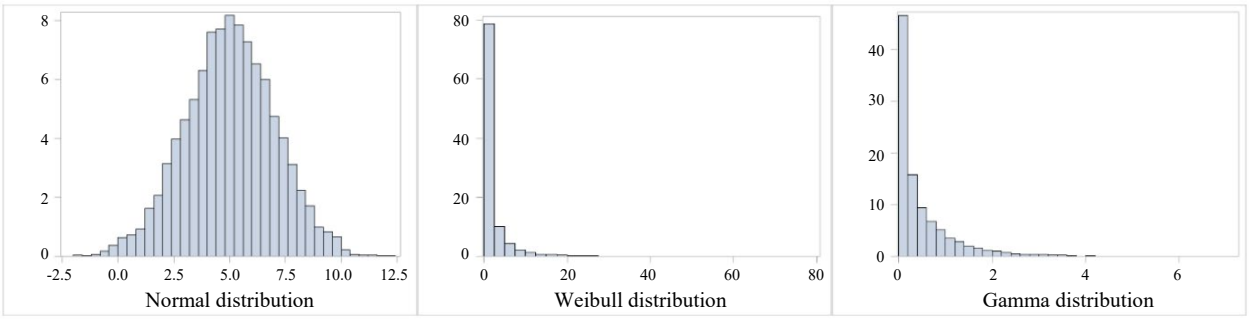


Table 7.2 (intra-household independence) and Table 7.3 (intra-household dependence) present dl_{PBRR} and dl_{cl} (calculated at the population level) and the means $\overline{\widehat{dl}}_{PBRR}$ and $\overline{dl}_{App}$ obtained from the 3,000 samples of $\widehat{dl}_{PBRR}$ and dl_{App}, respectively, for each scenario. These tables also compare the estimated coverage of the Student confidence intervals, described in equation (4.2), as well as the mean relative lengths expressed as a percentage of those intervals, i.e., the length divided by the absolute value of $\hat{Y}_D$, obtained for each scenario.

As expected, dl_{PBRR} increases with domain size. In addition, the presence of intra-household dependence appears to reduce the number of degrees of freedom. Indeed, dl_{PBRR} is smaller under scenarios with intra-household dependence (Table 7.3) compared with scenarios under intra-household independence (Table 7.2). Moreover, dl_{PBRR} is greater when a normal variable is present. This is consistent with the fact that the value of the excess kurtosis κ_{2D} is smaller for the normal variable than for the Weibull and Gamma variables.

Table 7.2
Estimated coverage of 3,000 replicated Student confidence intervals over total continuous variables with intra-household independence

Type of variable	Domain size	Degrees of freedom				Mean relative length of confidence intervals (%)				Estimated coverage (%)			
		dl_{PBRR}	$\overline{\widehat{dl}}_{PBRR}$	dl_{cl}	$\overline{dl}_{App}$	dl_{PBRR}	$\widehat{dl}_{PBRR}$	dl_{cl}	dl_{App}	dl_{PBRR}	$\widehat{dl}_{PBRR}$	dl_{cl}	dl_{App}
Normal	3%	28.2	26.3	39.2	30.3	72.7	73.1	71.8	72.6	93.9	94.1	93.7	94.0
	5%	44.1	40.4	68.4	50.7	54.4	54.6	53.9	54.2	93.7	93.7	93.4	93.7
	10%	67.8	63.2	129.1	99.0	37.2	37.3	36.9	37	95.2	95.2	95.0	95.0
Weibull	3%	5.8	11.3	6.4	30.3	121.9	113.9	119.1	100.9	94.4	93.1	94.1	91.6
	5%	9.5	12.2	11.1	50.7	90	89.7	88.2	80.6	93.4	92.8	93.0	91.3
	10%	11.0	16.6	12.8	99.0	64.5	63.8	63.4	58.2	93.7	92.6	93.5	91.3
Gamma	3%	19.3	18.4	23.9	30.3	87.1	87.8	85.9	85.1	93.9	94.0	93.6	93.5
	5%	21.3	21.2	27.6	50.7	68.6	69.0	67.7	66.3	93.8	93.8	93.4	93.2
	10%	32.3	32.6	43.9	99.0	47.3	47.4	46.8	46.1	94.5	94.5	94.3	94.1

Note: Partially balanced repeated replication (PBRR).

Table 7.3
Estimated coverage of 3,000 replicated Student confidence intervals over total continuous variables with intra-household dependence

Type of variable	Domain size	Degrees of freedom				Mean relative length of confidence intervals (%)				Estimated coverage (%)			
		dl_{PBRR}	$\overline{\widehat{dl}}_{PBRR}$	dl_{cl}	$\overline{dl}_{App}$	dl_{PBRR}	$\overline{\widehat{dl}}_{PBRR}$	dl_{cl}	dl_{App}	dl_{PBRR}	$\overline{\widehat{dl}}_{PBRR}$	dl_{cl}	dl_{App}
Normal	3%	20.1	20.2	26.2	30.3	76.4	76.8	75.3	74.9	94.5	94.5	94.1	94.1
	5%	35.5	33.6	51.9	50.7	57.6	57.8	57.0	57.0	93.4	93.5	93.1	93.2
	10%	57.5	53.8	100.5	99.0	38.8	38.8	38.4	38.4	94.7	94.7	94.4	94.4
Weibull	3%	2.5	6.1	2.5	30.3	238.1	181.8	238.1	136.2	93.6	88.1	93.6	83.8
	5%	2.5	7.0	2.6	50.7	203.8	151.2	199.9	114.5	95.2	88.1	95.0	84.3
	10%	3.0	10.6	3.4	99.0	137.7	112.2	128.2	86	95.6	87.7	94.6	85.2
Gamma	3%	3.5	8.0	4.0	30.3	163.2	138.4	154.1	113.6	94.4	90.2	93.4	87.6
	5%	5.1	10.2	6.1	50.7	113.8	105.3	108.8	89.6	94.1	91.6	93.0	89.3
	10%	9.9	16.5	11.7	99.0	69.9	68.7	68.5	62.1	94.0	93.2	93.6	91.8

Note: Partially balanced repeated replication (PBRR).

The empirical version of the theoretical number of degrees of freedom $\widehat{dl}_{PBRR}$ tends to be fairly close on average to dl_{PBRR} under the normal distribution. Conversely, the results are not as good for the Weibull and Gamma variables, especially with intra-household dependence. The empirical version $\widehat{dl}_{PBRR}$ appears to overestimate the true value of dl_{PBRR} in these cases.

Note that the difference between dl_{cl} and dl_{PBRR} shows the loss in degrees of freedom associated with the use of the PBRR-ε variance estimation method compared with the conventional method. This loss is particularly significant when the number of degrees of freedom is large. For example, for the normal distribution with intra-household dependence and for the 10% domain, the degrees of freedom decrease from 100.5 to 57.5. However, when the number of degrees of freedom is large (e.g., greater than 30), this loss has a limited impact on coverage (from 94.4% to 94.7% in the previous example). By contrast, when the number of degrees of freedom is low, the absolute difference between dl_{cl} and dl_{PBRR} is less significant, but it affects coverage to a greater extent. For example, for the Gamma distribution with intra-household dependence and the 3% domain, a decrease in degrees of freedom from 4 to 3.5 results in an increase in estimated coverage of 1 percentage point, from 93.4% to 94.4%.

In terms of approximate degrees of freedom, the type of variable and the presence of intra-household dependence have no effect on dl_{App}. Only the size of the domain influences the value of dl_{App}. In all cases, dl_{App} is larger on average than dl_{PBRR}. As a result, the relative mean lengths of the confidence intervals are almost always smaller with dl_{App} compared with dl_{PBRR}. This also implies that the estimated coverage for dl_{PBRR} is generally equal to or greater than the estimated coverage for dl_{App}. The estimated coverage with

dl_{PBRR} is generally close to 95% (from 93.4% to 95.5%), while the estimated coverage with dl_{App} is more variable and is less than or equal to 95% (from 83.8% to 95.0%).

For the largest domain (10%), the observed coverage is generally close to the nominal threshold for both dl_{PBRR} and dl_{App}. The undercoverage problems with dl_{App} are noted especially for small domains. The performance of dl_{App} is fairly good most of the time with intra-household independence (Table 7.2). However, with intra-household dependence (Table 7.3), dl_{App} shows significant undercoverage for the Weibull and Gamma variables. In these cases, using dl_{PBRR} significantly improves the estimated coverage compared with dl_{App}. For example, for the Gamma variable, for the 3% domain with intra-household dependence, the estimated coverage increases from 87.6% with dl_{App} to 94.4% with dl_{PBRR}, and the mean relative length of the confidence intervals increases from 113.6% to 163.2%, an increase of 49.6 percentage points. This particular example clearly illustrates the danger of overestimating the number of degrees of freedom. It implies that the estimates are more accurate than they really are, because the confidence intervals are too short. The objective of using a lower bound of 2.5 degrees of freedom for the theoretical number of degrees of freedom was met since no significant overcoverage was observed. Lastly, there is no indication that the calibration performed had a significant impact on the degrees of freedom, since the estimated coverage remains close to 95% in all scenarios with the theoretical formula of degrees of freedom. This suggests that the theoretical expressions developed remain appropriate even with calibration, at least in the context being studied.

7.3 Simulation results for a count estimation

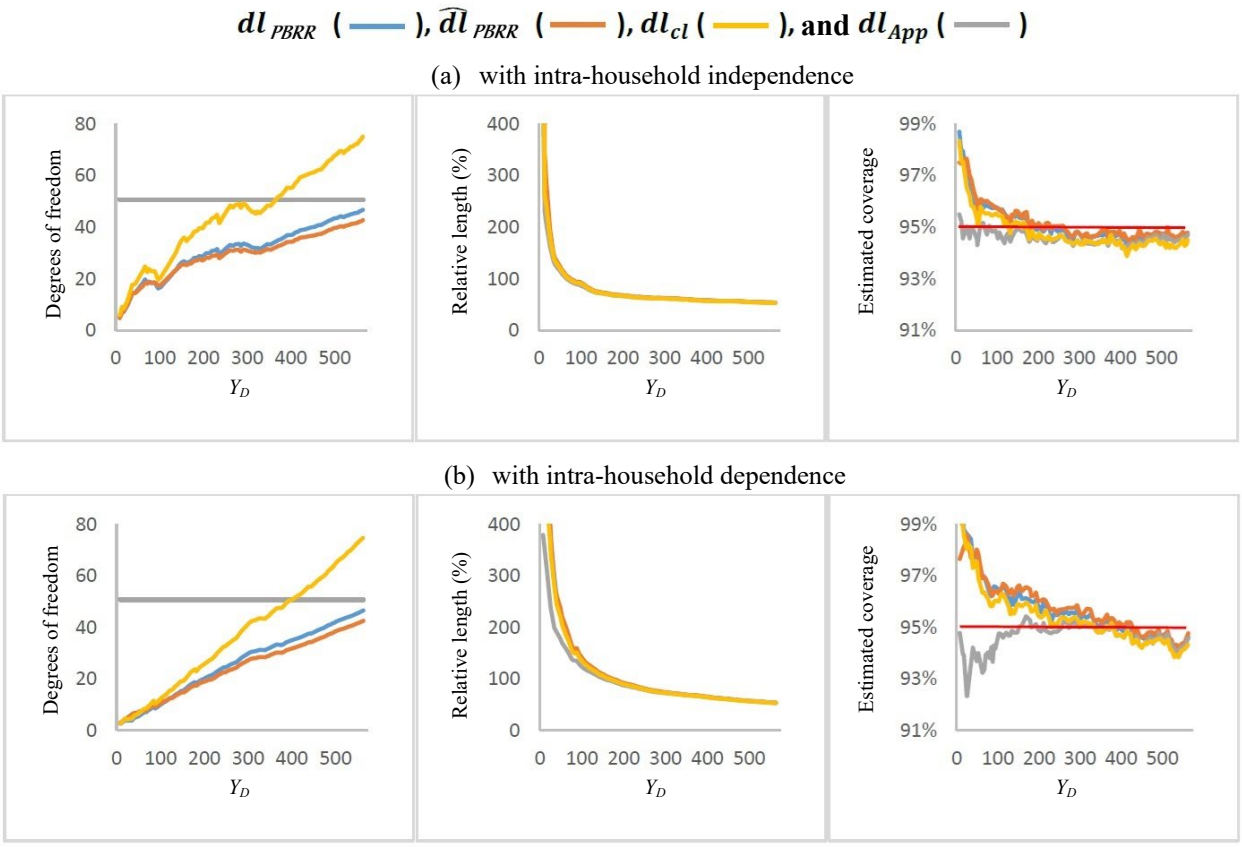
Most estimates produced from the responses to the long-form census questionnaire are counts. This section presents a simulation study for count estimation. The possible values of the variable of interest y_{hij} are 0 and 1, and they were generated based on a Bernoulli distribution with $P(y_{hij}=1 \mid z_{hij}=1)=p$ for all the individuals in the population. For each of the three possible domains, 99 dichotomous variables corresponding to $p=1,\ldots,$ 99% were generated.

Modified Wilson confidence intervals, as given in equation (4.3), were calculated using dl_{PBRR}, $\widehat{dl}_{PBRR}$, dl_{cl} and dl_{App}. Here, all the strata contain at least one individual from the domain, for the three generated domains. The modified Wilson confidence interval is thus calculated with $M_I = M$ and $m_I = m$. Note that for some simulated samples (when p is very small and the domain size is small), the estimated count was zero. In these cases, the variance estimator $v_{PBRR}(\hat{Y}_D)$ is also zero. For example, for the 5% domain, this was observed for about 0.3% of all simulated samples. For simplicity, these cases were removed from the simulation study and coverage was estimated from the other samples.

Figure 7.2 shows the results for the 5% domain; the results for the 3% and 10% domains are similar and are not presented here. Figure 7.2 shows the theoretical numbers of degrees of freedom dl_{PBRR} and dl_{cl}, as

well as $\overline{\widehat{dl}}_{PBRR}$ and $\overline{dl}_{App}$, i.e., the mean values of $\widehat{dl}_{PBRR}$ and dl_{App} for the 3,000 samples. The figure also compares the mean relative length and the estimated coverage of confidence intervals based on count Y_D.

Figure 7.2 Degrees of freedom dl_{PBRR}, $\overline{\widehat{dl}}_{PBRR}$, dl_{cl} and $\overline{dl}_{App}$ (left), mean relative length (centre) and estimated coverage (right) of the modified Wilson confidence intervals for estimating count Y_D for the 5% domain



Note: Partially balanced repeated replication (PBRR).

In terms of the number of degrees of freedom (Figure 7.2, left), dl_{PBRR} can be seen to increase as the count Y_D increases, and the value of dl_{PBRR} is smaller with intra-household dependence than with independence for a given value of Y_D. The approximate number of degrees of freedom dl_{App}, for its part, is influenced neither by the size of the count at the population level nor by the presence of intra-household dependence. These initial observations are similar to what was observed in Section 7.2 for continuous variables. These behaviours therefore do not depend on the nature of the variable of interest.

In all cases, the estimator $\widehat{dl}_{PBRR}$ appears to be fairly reliable in estimating the theoretical value dl_{PBRR}. The estimated mean number of degrees of freedom $\widehat{dl}_{PBRR}$ is close to the theoretical value dl_{PBRR} for all the scenarios. The difference between dl_{cl} and dl_{PBRR} increases when the count Y_D increases. Thus, the loss of

degrees of freedom associated with the use of the PBRR-ε method, compared with the conventional method, is greater when counts are large.

With respect to the estimated coverage (Figure 7.2, right), the coverage obtained with the estimated degrees of freedom $\widehat{dl}_{PBRR}$ is quite similar to that observed with the theoretical degrees of freedom dl_{PBRR}. With intra-household independence and where the count Y_D is large, the estimated coverage with the four versions of degrees of freedom is similar, generally close to the nominal rate of 95%. By contrast, with intra-household dependence, the estimated coverage with dl_{App} falls below 95% when the count is small. Conversely, dl_{PBRR} and $\widehat{dl}_{PBRR}$ show no significant undercoverage in all the scenarios considered. When the count is low, with both intra-household dependence and independence, overcoverage is observed for dl_{PBRR} and $\widehat{dl}_{PBRR}$, reinforcing their usefulness for small domain estimation. This overcoverage could be explained by the fact that calibration was not taken into account in the theoretical formula for the number of degrees of freedom.

With intra-household independence, the mean relative lengths (Figure 7.2, middle) of the confidence intervals are very similar for all four versions of the degrees of freedom. However, when Y_D is small and there is intra-household dependence, the relative lengths of the confidence intervals are slightly shorter for dl_{App} than for the other three alternatives. This is consistent with cases where dl_{App} results in undercoverage.

8. Conclusion

For the 2021 Census long-form questionnaire, Statistics Canada published confidence intervals to reflect the reliability of estimates. The length of these intervals is a good indicator of reliability, provided their coverage satisfies the nominal rate. This article proposes using a more precise formula for the degrees of freedom by using the Satterthwaite approximation to correct the undercoverage problems observed with the approximate rule. The general formula described in equation (2.1), which can be adapted to different estimators, sample designs and variance estimation methods, shows that the more unstable the variance estimator, the smaller the number of degrees of freedom. However, obtaining an explicit formula can be difficult, particularly for complex estimators such as quantiles.

Explicit expressions of the theoretical number of degrees of freedom are obtained, under a simple random stratified design without cluster replacement, for two variance estimation methods: the conventional method and the PBRR-ε method. The algebraic expressions that are obtained, derived to estimate a total over a domain, can be generalized to any linear estimator. The formula for degrees of freedom was obtained for the conventional method. It also applies to any variance estimation method using replicates that is identical to the conventional variance estimator in the linear case. This is particularly the case for the bootstrap method, which is algebraically equivalent to the conventional estimator when there is an infinite number of replicates.

A simplified version of the formula associated with the PBRR-ε method shows the factors influencing the degrees of freedom. To limit the loss of degrees of freedom, it is recommended to favour a large number of artificial strata and a small number of balancing groups, which implies a greater number of replicates.

The simulations show that using the theoretical number of degrees of freedom corrects the undercoverage caused by the approximate rule in many situations, in particular for small domains and with intra-household dependence. In certain limiting cases, overcoverage may occur with theoretical degrees of freedom; this could be explained by a violation of the assumptions underlying the Satterthwaite approximation. Because overcoverage is less problematic than undercoverage, this supports the use of the theoretical number of degrees of freedom rather than the approximate rule. An empirical version of the theoretical degrees of freedom that consists in replacing the second and fourth central moments with estimates yields good results for estimating a count, but in some cases tends to overestimate the actual number of degrees of freedom for continuous variable totals.

Lastly, while the use of the theoretical number of degrees of freedom has advantages, it is much more complex to calculate than the approximate rule. Currently, the census release system does not allow for its implementation. This would require a significant modification to the system.

Acknowledgements

We would like to thank Claude Girard, François Verret and Jean-François Beaumont from Statistics Canada, as well as the reviewers of *Survey Methodology*, for their careful review and valuable feedback, which improved this version of the article. Thank you also to Statistics Canada's Methodology Research Block Fund for the financial support that made this project possible.

Appendix

A. Proof that $E_*\left\{v_{PBRR}(\hat{Y}_D)\right\} = v_{cl}$

First, an alternative form of $v_{PBRR}(\hat{Y}_D)$ that will facilitate later calculations is obtained. Let $\tilde{w}_h = w_h(w_h - 1)$ and $\Delta_{hk} = \sqrt{n_{hk1}n_{hk2}}(\bar{t}_{Dhk1} - \bar{t}_{Dhk2})$, where $\bar{t}_{Dhk1} = t_{Dhk1}/n_{hk1}$ and $\bar{t}_{Dhk2} = t_{Dhk2}/n_{hk2}$. The following is obtained:

$$v_{PBRR}(\hat{Y}_D) = \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}\sqrt{\tilde{w}_h\tilde{w}_{h'}}\Delta_{hk}\Delta_{h'k'}\frac{1}{R}\sum_{r=1}^{R}\delta_{hk}^{(r)}\delta_{h'k'}^{(r)}.$$

Let $l_{hk} \in \{1, \ldots, L\}$ be the index for artificial stratum (h, k) in its balancing group, as shown in Figure 5.1. Because of conditions 2 and 3 described in Section 5.3, the following is obtained:

$$\frac{1}{R}\sum_{r=1}^{R}\delta_{hk}^{(r)}\delta_{h'k'}^{(r)} \;=\; \begin{cases}1 & \text{when } l_{hk}=l_{h'k'}\\ 0 & \text{otherwise}\end{cases}.$$

It follows that $\sum_{r=1}^{R}\delta_{hk}^{(r)}\delta_{h'k'}^{(r)}/R=\mathrm{I}(l_{hk}=l_{h'k'})$, where $\mathrm{I}(l_{hk}=l_{h'k'})=1$ when $l_{hk}=l_{h'k'}$ and $\mathrm{I}(l_{hk}=l_{h'k'})=0$ when $l_{hk}\neq l_{h'k'}$. Hence,

$$v_{PBRR}(\hat{Y}_D) \;=\; \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}\sqrt{\tilde{w}_h\tilde{w}_{h'}}\;\Delta_{hk}\Delta_{h'k'}\mathrm{I}(l_{hk}=l_{h'k'}).$$

Note that this sum contains exactly $K+2L\times\binom{G}{2}=GK$ non-zero elements, where G is the number of balancing groups, K is the number of total artificial strata and $L=K/G$. When $h=h'$, $\mathrm{I}(l_{hk}=l_{h'k'})=1$ when $k=k'$, whereas $\mathrm{I}(l_{hk}=l_{h'k'})=0$ when $k\neq k'$. Hence,

$$v_{PBRR}(\hat{Y}_D) \;=\; \sum_{h=1}^{H}\sum_{k=1}^{K_h}\tilde{w}_h\Delta_{hk}^2 \;+\; \sum_{h\neq h'=1}^{H}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}\sqrt{\tilde{w}_h\tilde{w}_{h'}}\;\Delta_{hk}\Delta_{h'k'}\mathrm{I}(l_{hk}=l_{h'k'}). \tag{A.1}$$

Random process $*$ can be separated into two independent random processes:

1- the separation of n_h sample units from stratum h into K_h artificial strata and the distribution of n_{hk} units from the artificial stratum (h,k) into two clusters s_{hk1} and s_{hk2}
2- the creation of balancing groups with the numbering of $l_{hk}\in\{1,\ldots,L\}$ and the selection of clusters via $\delta_{hk}^{(r)}$.

Note $E_*=E_{12}$ to represent these two processes. The result of equation (A.1) gives

$$E_*\left\{v_{PBRR}(\hat{Y}_D)\right\} \;=\; \sum_{h=1}^{H}\sum_{k=1}^{K_h}\tilde{w}_h E_1(\Delta_{hk}^2) \;+\; \sum_{h\neq h'=1}^{H}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}\sqrt{\tilde{w}_h\tilde{w}_{h'}}\;E_1(\Delta_{hk}\Delta_{h'k'})\;E_2\left\{\mathrm{I}(l_{hk}=l_{h'k'})\right\}.$$

First, $E_1\left(\Delta_{hk}\Delta_{h'k'}\right)=E_1(\Delta_{hk})\,E_1(\Delta_{h'k'})$ for $h\neq h'$ and $E_1(\Delta_{hk})=E_1(\Delta_{h'k'})=0$. Thus, $E_1(\Delta_{hk}\Delta_{h'k'})=0$. Second, a little algebra results in

$$E_1(\Delta_{hk}^2) \;=\; \frac{n_{hk}}{n_h(n_h-1)}\left(n_h\sum_{i=1}^{n_h}t_{Dhi}^2 \;-\; \left(\sum_{i=1}^{n_h}t_{Dhi}\right)^2\right) \;=\; n_{hk}S_{Dh}^2.$$

Therefore,

$$E_*\left\{v_{PBRR}(\hat{Y}_D)\right\} \;=\; \sum_{h=1}^{H}\sum_{k=1}^{K_h}\tilde{w}_h n_{hk}S_{Dh}^2 \;=\; \sum_{h=1}^{H}N_h\left(\frac{N_h-n_h}{n_h}\right)S_{Dh}^2 \;=\; v_{cl}.$$

B. Proof of proposition 6.1

The result is a direct consequence of equation (2.1). We have

$$V_{2,cl} \;=\; \sum_{h=1}^{H}N_h^2\left(\frac{N_h-n_h}{n_h}\right)^2\mathrm{Var}_P(S_{Dh}^2).$$

The form of $\mathrm{Var}_P(S_{Dh}^2)$ can be found in Cho, Cho and Eltinge (2005).

C. Proof of proposition 6.2

Using the result of equation (2.1), $V_{2,PBRR} = \mathrm{Var}_{P,*}\left\{v_{PBRR}(\hat{Y}_D)\right\}$ can be developed algebraically. Since $v_{PBRR}(\hat{Y}_D)$ is unbiased for $\mathrm{Var}_P(\hat{Y}_D)$ given in equation (3.5), we have

$$V_{2,PBRR} = E_{P,*}\left\{(v_{PBRR}(\hat{Y}_D))^2\right\} - \left(\sum_{h=1}^{H}\frac{N_h n_h \tilde{w}_h}{(N_h-1)}\sigma_{Dh}^2\right)^2.$$

The result of equation (A.1) gives $E_{P,*}\left\{(v_{PBRR}(\hat{Y}_D))^2\right\} = S_1 + 2S_2 + S_3$, where

$$S_1 = \sum_{h,h'=1}^{H}\tilde{w}_h\tilde{w}_{h'}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}E_{P,1}(\Delta_{hk}^2\Delta_{h'k'}^2)$$

$$S_2 = \sum_{h_1=1}^{H}\sum_{h_2\neq h_2'=1}^{H}\tilde{w}_{h_1}\sqrt{\tilde{w}_{h_2}\tilde{w}_{h_2'}}\sum_{k_1=1}^{K_{h_1}}\sum_{k_2=1}^{K_{h_2}}\sum_{h_2'=1}^{K_{h_2'}}E_{P,1}(\Delta_{h_1k_1}^2\Delta_{h_2k_2}\Delta_{h_2'k_2'})\,E_2\left\{\mathrm{I}(l_{h_2k_2}=l_{h_2'k_2'})\right\}$$

and

$$S_3 = \sum_{h_1\neq h_1'}\sum_{h_2\neq h_2'}\sqrt{\tilde{w}_{h_1}\tilde{w}_{h_1'}\tilde{w}_{h_2}\tilde{w}_{h_2'}}$$

$$\times\sum_{k_1=1}^{K_{h_1}}\sum_{k_1'=1}^{K_{h_1'}}\sum_{k_2=1}^{K_{h_2}}\sum_{k_2'=1}^{K_{h_2'}}E_{P,1}(\Delta_{h_1k_1}\Delta_{h_1'k_1'}\Delta_{h_2k_2}\Delta_{h_2'k_2'})\,E_2\left\{\mathrm{I}(l_{h_1k_1}=l_{h_1'k_1'},l_{h_2k_2}=l_{h_2'k_2'})\right\}.$$

As in Appendix A, the random process * was separated into two independent processes; indices 1 and 2 respectively refer to the random processes of creating artificial strata and creating balancing groups.

First, note that $E_{P,1}(\Delta_{hk})=0$ and that Δ_{hk} and $\Delta_{h'k'}$ are independent when $h\neq h'$. Second, $E_{P,1}(\Delta_{hk_1}\Delta_{hk_2})=0$ as soon as $k_1\neq k_2$, since

$$E_{P,1}(\bar{t}_{Dhk_11}\bar{t}_{Dhk_21}) = E_{P,1}(\bar{t}_{Dhk_11}\bar{t}_{Dhk_22}) = E_{P,1}(\bar{t}_{Dhk_12}\bar{t}_{Dhk_21}) = E_{P,1}(\bar{t}_{Dhk_12}\bar{t}_{Dhk_22}).$$

These arguments will be used in the following developments.

Development of S_1 : Since Δ_{hk} and $\Delta_{h'k'}$ are independent when $h\neq h'$, we have

$$S_1 = \sum_{h=1}^{H}\tilde{w}_h^2\left\{\sum_{k=1}^{K_h}E_{P,1}(\Delta_{hk}^4)+\sum_{k\neq k'}E_{P,1}(\Delta_{hk}^2\Delta_{hk'}^2)\right\} + \sum_{h\neq h'}\tilde{w}_h\tilde{w}_{h'}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}E_{P,1}(\Delta_{hk}^2)\,E_{P,1}(\Delta_{h'k'}^2).$$

Development of S_2 : Since $\Delta_{h_2k_2}$ and $\Delta_{h_2'k_2'}$ are independent for $h_2\neq h_2'$ and since $E_{P,1}(\Delta_{h_2k_2}) = E_{P,1}(\Delta_{h_2'k_2'})=0$, we have $E_{P,1}(\Delta_{h_1k_1}^2\Delta_{h_2k_2}\Delta_{h_2'k_2'})=0$ in all cases and thus, $S_2=0$.

Development of S_3 : For $h_1 \neq h_1'$ and $h_2 \neq h_2'$, the only case where $E_{P,1}(\Delta_{h_1k_1}\Delta_{h_1'k_1'}\Delta_{h_2k_2}\Delta_{h_2'k_2'})$ is not zero is when $h_1 = h_2$, $h_1' = h_2'$, $k_1 = k_2$ and $k_1' = k_2'$. Since $E_2\{\mathrm{I}(l_{hk} = l_{h'k'})\} = (G-1)/(K-1)$, we get

$$S_3 = \frac{G-1}{K-1}\sum_{h\neq h'}\tilde{w}_h\tilde{w}_{h'}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}E_{P,1}(\Delta_{hk}^2)\,E_{P,1}(\Delta_{h'k'}^2).$$

A long algebraic development results in $E_{P,1}(\Delta_{hk}^2\Delta_{hk'}^2) = n_{hk}n_{hk'}g_2(h)$,

$$E_{P,1}(\Delta_{hk}^2) = \frac{n_{hk}N_h}{N_h-1}\sigma_{Dh}^2 \quad \text{and} \quad E_{P,1}(\Delta_{hk}^4) = \left(\frac{n_{hk2}^2}{n_{hk1}}+\frac{n_{hk1}^2}{n_{hk2}}\right)g_1(h)+3n_{hk}^2g_2(h),$$

where

$$g_1(h) = N_h^2\frac{(N_h+1)\,m_{Dh4}-3(N_h-1)\,\sigma_{Dh}^4}{(N_h-1)\,(N_h-2)\,(N_h-3)} \quad \text{and} \quad g_2(h) = N_h\frac{(N_h^2-3N_h+3)\,\sigma_{Dh}^4-(N_h-1)\,m_{Dh4}}{(N_h-1)\,(N_h-2)\,(N_h-3)}.$$

Therefore, we get

$$S_1 = \sum_{h=1}^{H}\tilde{w}_h^2\left\{\left\{g_1(h)\,B_h+g_2(h)\,(2A_h+n_h^2)\right\}\right\} + \sum_{h\neq h'}\frac{\tilde{w}_h\tilde{w}_{h'}n_hn_{h'}N_hN_{h'}}{(N_h-1)\,(N_{h'}-1)}\sigma_{Dh}^2\sigma_{Dh'}^2$$

and

$$S_3 = \frac{G-1}{K-1}\sum_{h\neq h'}\frac{\tilde{w}_h\tilde{w}_{h'}n_hn_{h'}N_hN_{h'}}{(N_h-1)\,(N_{h'}-1)}\sigma_{Dh}^2\sigma_{Dh'}^2$$

where A_h and B_h are as defined in equation (6.1). Consequently,

$$V_{2,PBRR} = \sum_{h=1}^{H}\tilde{w}_h^2\left\{\left\{g_1(h)\,B_h+g_2(h)\,(2A_h+n_h^2)\right\}-\frac{N_h^2n_h^2}{(N_h-1)^2}\sigma_{Dh}^4\right\}+\frac{G-1}{K-1}\sum_{h\neq h'}\frac{\tilde{w}_h\tilde{w}_{h'}n_hn_{h'}N_hN_{h'}}{(N_h-1)\,(N_{h'}-1)}\sigma_{Dh}^2\sigma_{Dh'}^2.$$

Finally, an algebraic development gives the desired result.

References

Cho, E., Cho, M.J. and Eltinge, J. (2005). The variance of the variance of samples from a finite population. *International Journal of Pure and Applied Mathematics*, 21(1).

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. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Chicago, United States.

Gerlovina, I. and Hubbard, A.E. (2019). Computer algebra and algorithms for unbiased moment estimation of arbitrary order. *Cogent Mathematics & Statistics*, 6(1).

Hedayat, A. and Wallis, W. (1978). Hadamard matrices and their applications. *The Annals of Statistics*, 6(6), 1184-1238.

Heeringa, S.G., West, B.T. and Berglund, P.A. (2010). *Applied Survey Data Analysis*. Chapman and hall/CRC.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, Statistics Sweden, 6(3), 223-239.

Kott, P.S. (2020). *The Degrees of Freedom of a Variance Estimator in a Probability Sample*. RTI Press.

Kott, P.S. and Carr, D.A. (1997). Developing an estimation strategy for a pesticide data program. *Journal of Official Statistics*, 13(4), 367-383.

McCarthy, P.J. (1966). Replication: An approach to the analysis of data from complex surveys. *Vital and Health Statistics, Data Evaluation and Methods Research*, Series 2, 14, 1-38.

Nelsen, R.B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. New York: Springer, 2nd ed.

Neusy, E., Savard, S.-A., Hidiroglou, M.A. and Martin, V. (2021). Modified Wilson intervals for estimated counts with application to Census 2021 long form estimation. Presentation to the Advisory Committee on Statistical Methods, May 2021. Internal document, Statistics Canada.

Rao, J.N.K. and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, Oxford University Press, 86(2), 403-415.

Rust, K.F. (1986). Efficient replicated variance estimation. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 81-87.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110-114.

Valliant, R. and Rust, K.F. (2010). Degrees of freedom approximations and rules-of-thumb. *Journal of Official Statistics*, 26(4), 585-602.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer Science & Business Media.