Using Bootstrap Variance Calculations for a Survey with a Simple Design: The Case of the 2005 National Survey of the Work and Health of Nurses

Gildas Kleim[1], Yves Bélanger[2]
[1]Statistics Canada, 18-A R.H. Coats building, 100, Tunney's Pasture driveway, Ottawa, ON, K1A 0T6, Canada
[2]Statistics Canada

Abstract

Rao and Wu (1988) proposed a bootstrap method aimed for use in complex sample designs involving stratification, multiple stages and unequal probability sampling. They also presented a bootstrap procedure for stratified simple random sampling without replacement (WOR). Even though formulas for variance estimates for such a simple design are well known, it may happen nevertheless that clients request bootstrap weights to be produced. The usual bootstrap procedure used by Statistics Canada, and referred to as the Rao-Wu-Yue bootstrap, assumes sampling with replacement (WR) and thus leads to an overestimation of the variance for WOR sampling, especially when the stratum sampling fractions are high. In this paper, we describe and evaluate an implementation of the bootstrap procedure for a simple design using the 2005 National Survey of the Work and Health of Nurses.

KEY WORDS: Variance estimation, Bootstrap, Simple Design

1. Introduction

The Canadian National Survey of the Work and Health of Nurses (NSWHN) was conducted by telephone in October 2005. Its aim was to produce reliable information concerning the work-related health issues for nurses, who comprise the largest occupation within the Canadian health sector. The survey was based on a stratified simple random sampling of nurses selected from a frame built using administrative files provided by provincial regulatory organizations.

Following data collection, Statistics Canada clients requested bootstrap weights to be produced for the purpose of variance estimation, although simple variance estimation formulas for such a design were available. However, the usual bootstrap procedure used by Statistics Canada, referred to as the Rao-Wu-Yue bootstrap (1992), assumes sampling with replacement (WR) within strata and thus leads to an overestimation of the variance for without replacement (WOR) sampling designs, especially when the stratum sampling fractions are high. This paper describes the investigation leading up to the solution that was finally adopted.

The methodology of the NSWHN is described in Section 2. Section 3 deals with the way variances were estimated using the usual textbook formulas and Section 4 describes how the bootstrap was initially implemented, the problems that occurred, and the solutions that were found. A conclusion is given in Section 5.

2. The 2005 National Survey of the Work and Health of Nurses

2.1 Target population

The NSWHN is a survey of nurses for which the target population is all nurses employed in the nursing field in Canada in 2005. Specifically, there are three classes of nurses in Canada: Registered Nurses (RN), Licensed Practical Nurses (LPN) and Registered Psychiatric Nurses (RPN). While the RNs and LPNs are found throughout the entire country, the RPNs are unique to the four western-most provinces. Excluded from the survey's coverage are retired nurses, unemployed nurses and nurses currently working in a field other than nursing.

2.2 Frame

For each type of nurses in each province, there exists a regulating body that collects and maintains membership information (there are roughly 26 provincial nursing organizations across Canada). A nurse must be registered in a given province in order to work there. Each organization sent Statistics Canada its membership list and Statistics Canada compiled the lists to create the survey frame.

To ensure that we could create the most up-to-date version of the frame possible, Statistics Canada received membership files from the provincial organizing bodies in the summer and fall of 2005. This ensured that we were using the most up-to-date registration lists and avoided significant difficulties with out-of-date information as well as under- or over- coverage.

Further, as a result of the way the frame was created, duplicate records had to be removed, and some imputation was done when part of the stratification

Table 1: Population and Sample Size by Province and Type of Nurse for the 2005 NSWHN

	Registered Practical Nurses		Licensed Practical Nurses		Registered Psychiatric Nurses		Total	
Province	**Population**	**Sample**	**Population**	**Sample**	**Population**	**Sample**	**Population**	**Sample**
Newfoundland & Labrador	5,371	1,172	2,620	897	-	-	7,991	2,069
Prince Edward Island	1,478	693	609	333	-	-	2,087	1,026
Nova Scotia	9,110	1,142	3,222	911	-	-	12,332	2,053
New Brunswick	7,934	1,136	2,760	875	-	-	10,694	2,011
Québec	65,281	1,513	16,842	1,374	-	-	82,123	2,887
Ontario	95,815	1,653	24,874	1,348	-	-	120,689	3,001
Manitoba	11,105	1,097	2,462	849	931	468	14,498	2,414
Saskatchewan	8,482	1,062	2,183	978	1,130	527	11,795	2,567
Alberta	26,239	1,228	5,280	1,041	1,169	511	32,688	2,780
British Columbia	28,765	1,224	4,781	1,008	2,023	803	35,569	3,035
Territories	1,291	414	235	186	-	-	1,526	600
Canada	**260,871**	**12,334**	**65,868**	**9,800**	**5,253**	**2,309**	**331,992**	**24,443**

information was missing. Finally, a pilot survey was conducted in the spring of 2005, based on an earlier version of the membership files, and the nurses who were selected for this pilot study were excluded from selection in the main survey to reduce the effects of respondent burden and conditioning.

2.3 Sample Design

The NSWHN sample was based upon a stratified design employing probability sampling. The design principles were the same for each province. Each combination of province of registration and type of nurse can be thought of as a separate survey; that is, the expected quality of the estimates for each combination had to be better than a fixed minimum. To achieve this goal the primary stratification for NSWHN involved creating a separate stratum for each combination of province of registration and type of nurse.

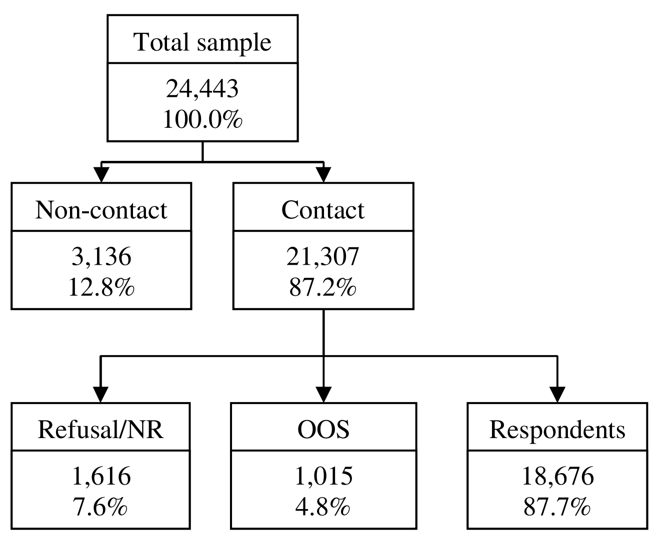
Within any given stratum described previously, the frame was further stratified to meet domain requirements identified by the client. The client expressed interest in obtaining reliable estimates for the following 11 domains: age group (4 groups), place of work (4 groups), and employment status (3 groups). To this end, the frame was further stratified; however, the secondary stratification was more complex than the primary stratification: some strata had to be collapsed in order to keep the sample size reasonable.

Once the secondary stratification was completed, a simple random sample (SRS) of nurses was drawn from each ultimate stratum. The sample size of eligible nurses for NSWHN was determined so as to meet the statistical precision requirements of the clients. The sample allocation also tried to improve the quality of national level estimates by allocating additional sample to larger provinces. Table 1 shows the population and sample sizes of the NSWHN.

The initial sample selected consisted of 24,443 nurses from the 331,992 nurses on the frame. While conducting the survey, 4,752 cases were lost because of non-contact or refusal. Further, 1,015 individuals were finally considered as out of the scope (OOS) of the survey, since it was established during the interview that they were part of one of the exclusions described in 2.1 above, leaving 18,676 final respondents.

Figure 1 shows a breakdown of the total sample following collection. We were unable to contact 12.8% of the initial sample (or 3,136 cases). Of the cases that were contacted, 87.7% responded to the survey.

Figure 1: Breakdown of the Total Sample after Collection



2.4 Final Sample and Weighting Process

The starting point for weighting in stratum h was the theoretical design weight (N_h/n_h) where N_h is the number of nurses according to the frame and n_h is the size of the sample initially selected from that stratum. The strata were used as nonresponse groups because:

1. They contain information about the province of registration and type of nurse as well as the domains of interest identified by the clients.

2. Using the strata as nonresponse groups does not complicate the textbook formula for variance estimation.

The weighting was done in two steps within each stratum:

Step 1: The weights of the non-contacted cases and the nonrespondents cases for which status (in or out of scope) could not be determined were re-distributed across the remaining sample.

Step 2: The weights of the refusals/nonrespondents that were known to be in-scope were re-distributed across the in-scope respondents

Table 2 shows however that most of the nonresponse occurred before it could be determined whether individuals were in-scope or out of the scope of the survey.

Further, it is straightforward to show that, within a stratum, a SRSWOR of n_h individuals followed by a Bernoulli sampling among the selected individuals, resulting in a sample of r_{1h} respondents, is equivalent to a SRSWOR of r_{1h} individuals among N_h individuals, under the assumption that all individuals in the stratum have the same probability of responding.

For the purpose of variance estimation, selecting a sample of individuals and getting the information as to whether they belong to the target population or not can be seen as the first phase of a two-phase sample design. This information can then be used to stratify the sample further, each first-phase respondent being classified into one of two new strata (in-scope or out-of-scope) created within each initial sampling stratum.

The subsequent nonresponse can then be viewed as a second phase of sampling. It is the nonresponse that occurred after it was determined whether a selected person was in-scope or out of the scope of the survey. Of course, such nonresponse happened only for individuals who had previously been considered to be within the scope of the survey: no further question was asked of the respondents once it was determined that they were not part of the survey's target population.

Table 2: Distribution of weights according to final status

Final status	**Number of records**	**Sum of design weights**	**Sum of weights after Step 1**	**Sum of weights after Step 2**
Non-contact & refusal (unknown)	4,590	62,822	0	0
Refusal (in-scope)	162	2,422	3,012	0
Out-of-scope	1,015	13,919	17,063	17,063
Respondent	18,676	252,829	311,917	314,929
Total	**24,443**	**331,992**	**331,992**	**331,992**

3. Variance Estimation using Usual Formulas

The initial sample was drawn using a stratified simple random sampling without replacement (SRSWOR) of nurses as described in section 2. Then some nonresponse occurred during collection, which led to a smaller sample than previously expected.

Nonresponse usually occurred before it could be determined whether the selected person was in-scope or out of the scope of the survey. This first nonresponse mechanism is usually assumed to behave like a stratified Bernoulli sampling, in which each individual has an unknown but constant probability of being a respondent within each stratum (Särndal, Swensson and Wretman, 1992, Chapter 15).

Similarly to the first nonresponse mechanism, this second nonresponse mechanism can be assumed to behave like a Bernoulli sampling design. This nonresponse would therefore constitute a second phase of sample selection.

As a result, for the purpose of variance estimation, the final sample is considered to be the result of a two-phase selection process: the first phase being the combination of the random selection of the initial sample and the nonresponse that occurred before it could be determined whether individuals were in-scope or out-of-scope of the survey, and the second phase representing the nonresponse that occurred after it was determined that individuals were in-scope or out of the scope of the survey. The assumption was made for the second phase, similar to the first phase, that all individuals in a second-phase stratum had the same probability of responding.

The Generalized Estimation System (GES) designed by Statistics Canada (2005) was used to generate variance estimates based on this two-phase approach. These estimates were then used as a reference to evaluate the results obtained using the bootstrap, which means that variance estimates for a series of totals were calculated using both methods and compared to determine whether or not the bootstrap provided results similar to the GES.

4. Variance Estimation using the Bootstrap

Since acceptable variance estimates for the survey design assumed here can be easily calculated using textbook formulas, it is legitimate to question the relevance of using the bootstrap, which necessitates more programming work, and is much more computer intensive. It must be mentioned that Statistics Canada's clients for the 2005 NSHWN are accustomed to working with other surveys conducted by the Agency, for which the bootstrap is the usual method used to produce variance estimates. Therefore, automated computer programs were already available to process the bootstrap weights, and using another method would have required a learning period, which was not available because of a short timeline.

4.1 Bootstrap Variance Estimator

Initially proposed by Efron (1979) for the i.i.d. case, the bootstrap was then adapted to the case of survey data by several authors. The version known as the Rao-Wu-Yue bootstrap, and described in a *Survey Methodology* paper in 1992 is the version usually used at Statistics Canada. In the case of the 2005 NSHWN it has been applied in the following way:

For the b'th bootstrap:

1) Within each stratum, select a sample of n_h-*1* units with replacement from the initial sample (n_h being the number of individuals selected in stratum h).

2) Calculate the initial weight of the i^{th} individual of the h^{th} stratum, as
$$w_{hi}^{(b)} = w_{hi} \frac{n_h}{n_h - 1} t_{hi}^{(b)}$$
where $t_{hi}^{(b)}$ is the number of times the individual was selected in the bootstrap sample and w_{hi} is the sampling weight.

3) Recalculate the nonresponse adjustments and apply them to the $w_{hi}^{(b)}$ the same way as for survey weights,

Repeat steps 1 to 3, B times. In our case, B=500.

Calculate $\hat{\theta}^{(b)}$, b=1, ...B, the estimates of θ using the weights created in 3).

Calculate the bootstrap variance estimate of $\hat{\theta}$ as

$$v_{BS}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta})^2$$

Table 3: Number of strata and percentage of the population by sampling fraction, for Canada level and Prince Edward Island

	Canada		Prince Edward Island	
Sampling fraction	**# of strata**	**% of population**	**# of strata**	**% of population**
Less than 10%	99	83%	0	0%
10% to less than 30%	126	13%	0	0%
30% to less then 60%	117	4%	26	97%
60% to less then 100%	34	<1%	1	2%
100%	13	<1%	1	1%
Total	**389**	**100%**	**28**	**100%**

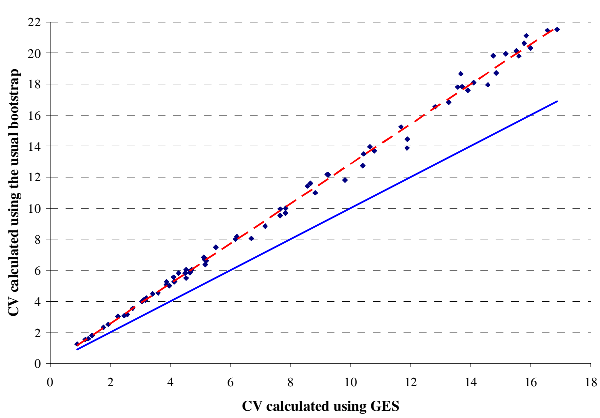
It must be stressed that the form of the Rao-Wu-Yue bootstrap described above assumes that sampling was done with replacement, which is not a major issue for most surveys, since the sampling fractions are usually low, and therefore ignorable. In the case of the 2005 NSWHN however, the sampling fractions could be as high as 30% in some strata, and this caused concerns about the quality of the variance estimates produced using this bootstrap approach for the NSHWN.

As an illustration, Table 3 shows the distribution of the strata by sampling fraction, and the percentage of nurses contained in these strata, for Canada and Prince Edward Island, which is the smallest province in Canada. At the national level, most strata had a sampling fraction higher than 10%, and a census had even been taken in some of them, but the vast majority of nurses were located in strata with a low sampling fraction, that is, less than 10%. On the contrary in Prince Edward Island, no stratum had a sampling fraction lower than 30%. For this reason, this province will serve as an example of the impact of high sampling fractions on variance estimates calculated using the particular bootstrap approach described above.

In the three graphics below, each dot represents the coefficient of variation (CV) of an estimate of a total in Prince Edward Island. A total of 74 estimates were used, such as the number of nurses being daily smokers, or the number of nurses having used at least one medication during the month before the interview. The x-coordinate is the variance estimate calculated using GES, while the y-coordinate is the variance estimate calculated using the bootstrap.

The plain line indicates the reference situation, that is, when the CVs calculated using GES and the CVs calculated using the bootstrap are equal. The dotted line is a regression curve that has been fitted to the variance estimates calculated as indicated above. In the ideal situation, the plain and dotted lines would be confounded since the GES estimates are being considered the “gold standard”.

Graph 1: Comparison of the CVs calculated using the usual Rao-Wu-Yue bootstrap to the CVs calculated using GES for estimates of totals in Prince Edward Island



4.2 Taking High Sampling Fractions into Account

Although the results obtained were considered satisfactory for estimates calculated at the national level (not shown here), Graph 1 shows that the variance is obviously over-estimated for estimates calculated in domains where the sampling fraction is high, and thus a WR assumption untenable, such as in a small province.

This sampling fraction is therefore simply not ignorable in this case, and for this reason, the equation used in step 2) of the process described above needs to be replaced by a formula that accounts for the fact that the sampling was actually WOR. The nonresponse adjustments are then to be applied as indicated in subsection 4.1.

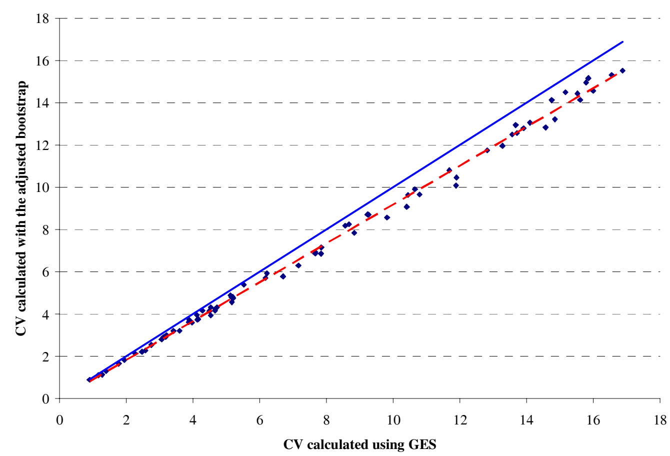
The formula is given in section 4.1 of Rao and Wu (1988) in terms of the survey data, but it can also be expressed in terms of the weights as:

$$w_{hi}^{(b)} = w_{hi}\left\{1-\sqrt{1-f_h} + \frac{n_h}{n_h-1}\sqrt{1-f_h}\cdot t_{hi}^{(b)}\right\}$$

where f_h is the sampling fraction n_h/N_h in stratum h.

Graph 2 shows the results obtained using this method for the same set of estimates of totals in Prince Edward Island. It appears that the method used now systematically results in estimates of the variance that are lower than those produced by GES. An explanation is provided in the next section.

Graph 2: Comparison of the CVs calculated using the Rao-Wu bootstrap to the CVs calculated using GES for estimates of totals in Prince Edward Island



4.3 Impact of Nonresponse on Variance Estimates

The variance in the 2005 NSWHN is generated by two sources: the selection of the sample and the nonresponse. It is known however (see Shao and Steel (1999)), that the bootstrap is simply not able to take the whole variance that occurs because of nonresponse into account. This variance can be ignored when the sampling fraction is low, because it is generally negligible when compared to the variance that occurs because of the sampling process.

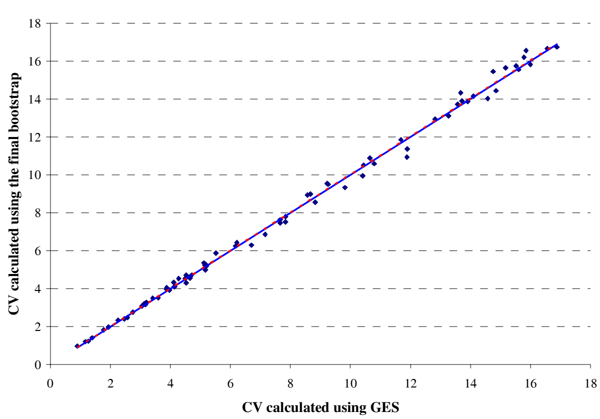
In the case of the NSWHN, however, the fact that some strata had a high sampling fraction, as previously mentioned, makes the variance due to nonresponse in the survey difficult to ignore.

Considering the fact that the initial NSWHN sample is a SRSWOR, and that both nonresponse steps have been assumed to be equivalent to Bernoulli mechanisms, a good approximation of the variance can be obtained by assuming that the sample was in fact a SRSWOR of the r_h final respondents within each stratum. In that case, the bootstrap replicates can be drawn from the final pool of respondents and out-of-scope cases, using the final adjusted weight. Each bootstrap replicate is then calculated using the following equation:

$$w_{hi}^{(b)} = w_{hi}^{adj}\left\{1-\sqrt{1-f_{rh}} + \frac{r_h}{r_h-1}\sqrt{1-f_{rh}}\cdot t_{hi}^{(b)}\right\}$$

where w_{hi}^{adj} is the final weight, i.e., after nonresponse adjustments, and $f_{rh} = r_h / N_h$
where r_h = number of respondents + number of out-of-scope cases.

Graph 3 Comparison of the CVs calculated using the Rao-Wu bootstrap on the final weights to the CVs calculated using GES for estimates of totals in Prince Edward Island



Graph 3 shows that this method provides accurate results for estimates at the provincial level, as the regression line is now confounded with the line corresponding to equal CVs.

Although this may lead to further problems in surveys where the response rates differ from one second phase stratum to another, it is a fact that, in the NSWHN, few individuals (less than 1%) ended the interview once it was established that they were part of the target population of the survey. Therefore, the supplementary variance occurring because of this nonresponse can be considered negligible compared to the variance caused by the sampling itself and the initial nonresponse.

5. Conclusion

We have described in this paper an example of a successful use of a bootstrap variance estimation method, in a situation where its implementation was not straightforward. The practical solution proposed was to the complete satisfaction of the clients. However, since the evaluation relied partially on empirical evidence, further research is continuing in order to find a solution that has a stronger theoretical justification. We are confident that the approach proposed by Kim, Navarro and Fuller (2006) for bootstrapping a two-phase survey can be adapted to the case of the NSWHN, given the relative simplicity of the NSWHN design (e.g., the fact that second-phase strata were built within first-phase strata).

Acknowledgements

The authors would like to acknowledge Jean-François Beaumont, Claude Girard, Lenka Mach, Georgia Roberts for their comments.

References

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics*, 7, 1-26.

Kim J.K., Navarro, A. and Fuller, W.A. (2006). "Replication Variance Estimation for Two-phase Stratified Sampling", *Journal of the American Statistical Association*, 101, pp. 312-320.

Rao, J.N.K. and Wu, C.F.J. (1988). "Resampling Inference with Complex Survey Data", *Journal of the American Statistical Association*, 83, pp.231-241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys", *Survey Methodology*, 18, pp.209-217.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) "Model Assisted Survey Sampling", New York: Springer-Verlag.

Shao, J. and Steel, P. (1999) "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions", *Journal of the American Statistical Association*, 94, pp. 254-265.

Statistics Canada (2005) "GES v4.3 User's Guide", Internal Document.