The Rao-Wu Rescaling Bootstrap: From theory to practice

Claude Girard
Senior Methodologist – Statistics Canada
claude.girard@statcan.gc.ca

Context

At Statistics Canada, variance estimation for complex surveys is mainly carried out using replication methods. The two replication methods which have been used in the last decade are the delete-one Primary Sampling Unit (PSU) jackknife and, more recently, the bootstrap. As Valliant (2007) rightly points out there are several variants of the bootstrap introduced by Efron (1979) in the i.i.d. case which are being used in survey sampling and we have to be clear which one is being referred to; at Statistics Canada we use solely the Rao-Wu rescaling bootstrap for production. Even though it was introduced in 1988 (see Rao and Wu (1988)), first implementations occurred only in the late 1990s. The Rao-Wu bootstrap is performed on both with and without replacement designs for which it yields sensible variance estimates for a variety of estimators including percentiles, a claim not matched by the delete-one PSU jackknife. Because of this, surveys which initially relied on the jackknife often make the switch to the bootstrap when the occasion arises, like during a survey re-design. The reader can get a good overview of what has been tried in survey sampling with regard to the bootstrap from Rust and Rao (1996), Shao (2003) and Lahiri (2003). The Rao-Wu variant of the original bootstrap procedure is appealing to survey users because it is simple to implement, it yields adequate variance estimates when the sample sizes are small (which is common with stratified designs) and it comes in the form of bootstrap weights.

For any given methodology, making the transition from theory to practice is a challenge, and the Rao-Wu bootstrap is no exception. After several years of use at Statistics Canada, strange variance estimates were unearthed in some surveys, mainly longitudinal surveys. These surveys are particularly challenging from a variance estimation standpoint because of the myriad ways one can conduct inferences: cross-sectionally (or by wave/cycle) and/or longitudinally (or across waves/cycles), *etc*. An in-depth investigation of the implementation strategy used up to that point unveiled several flaws which have since been addressed. The new implementation strategy reflects a better overall understanding of the bootstrap and the issues that may arise in variance estimation as we go from theory to practice.

Many of the issues we had to ponder and address would require too lengthy of a preamble to be mentioned in this paper; in fact, they already have been documented in the Statistics Canada working paper Girard (2007) with an addendum in the later working paper Girard *et al*. (2009). (Either or both can be obtained by contacting the author.) What we propose in this paper is to give a flavour of what is involved in applying the Rao-Wu bootstrap to real surveys by presenting some of the issues that may arise through case studies. In the process we have tried to collect the known theoretical relevant results (which are scattered in the literature) and also help the bootstrap user with their interpretation, hence to make the transition from theory to practice smoother.

Case study I: The set up featuring Simple Random Sampling Without Replacement (SRSWOR) with sampling fraction $f \cong 0$ and estimating the mean.

This first case study describes the most natural setting for the bootstrap introduced by Efron (1979); it is as close to the classic i.i.d. case we can get in a survey setting. (The early writings of Bradley Efron *e.g.*, Efron (1979) and Efron (1981), are enlightening and ought to be read by all bootstrap enthusiasts, as they provide great insights into the technique; the author's favourite exposé is Efron and Lepage (1992).) To understand where the Rao-Wu variant comes from, it is important to revisit Efron's bootstrap.

Suppose a sample *s* of size *n* has been obtained from a Simple Random Sample With Replacement (SRSWR) design or equivalently, for all practical purposes, from a SRSWOR with sampling fraction $f \cong 0$**,** and that the estimate $\hat{\bar{y}}$

of the population mean of *y* was obtained. Efron's bootstrap is usually implemented by selecting a large number *B* of samples of size *n with replacement from the observed sample*. (These samples are customarily called bootstrap replicates.) In an ideal world we would be able to use all bootstrap replicates and not just *B* of them, but in practice we cannot afford it. Indeed, a standard combinatorics argument shows that there are $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$ *distinct* bootstrap replicates to be obtained from a sample of size *n* which usually far exceeds what computers can handle efficiently. (To give an idea of its magnitude, observe that with *n*=50 this number has 28 digits.)

Using standard SRSWR results (*e.g.*, Result 3.3.4 in Särndal *et al*. (1992)) one can show that the following holds for Efron's bootstrap estimator $\hat{\bar{y}}_{BOOT}$:

$$
\begin{aligned}
&E_{bs}\left(\hat{\bar{y}}_{BOOT}\middle|s\right)=\hat{\bar{y}}\\
&V_{bs}\left(\hat{\bar{y}}_{BOOT}\middle|s\right)=\frac{n-1}{n}\hat{V}\left(\hat{\bar{y}}\right)
\end{aligned}
\qquad (1)
$$

The underscript "BOOT" indicates that expectation and variance are taken with respect to the bootstrap sampling mechanism.

Thus, the bootstrap variance is not quite unbiased for the usual variance estimate of the mean, a bias important enough for some of the small sample sizes encountered in survey sampling (mainly with stratified designs). One can describe the Rao-Wu rescaling bootstrap, introduced in Rao and Wu (1988), as a way to achieve unbiasedness on both fronts through a suitable linear transform of Efron's estimator *i.e.*, find:

$$
\hat{\bar{y}}_{R-W} = \alpha\hat{\bar{y}}_{BOOT} + \beta \qquad (2)
$$

such that:

$$
\begin{aligned}
&E_{BOOT}\left(\hat{\bar{y}}_{R-W}\middle|s\right)=\hat{\bar{y}}\\
&V_{BOOT}\left(\hat{\bar{y}}_{R-W}\middle|s\right)=\hat{V}\left(\hat{\bar{y}}\right)
\end{aligned}
\qquad (3)
$$

They also opened the door to re-sampling any number *m* of units and thus to the possibility of a choice which outperforms the traditional *n*.

Note: Since the remainder of this paper is exclusively about the Rao-Wu rescaling bootstrap, we are going to refer to it hereinafter simply as the bootstrap.

Because in practice *B* is chosen to be much less than $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$, the variance calculated is actually a Monte Carlo approximation:

$$
V_{BOOT}\left(\hat{\bar{y}}_{BOOT}\middle|s\right)\cong v_{BOOT}\left(\hat{\bar{y}}_{BOOT}\middle|s\right)\equiv\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{y}}_{BOOT(b)}-\hat{\bar{y}}\right)^{2} \qquad (4)
$$

It is important to realize that the choice of *B* will affect the magnitude of the error introduced in the ensuing estimate v_{BOOT} by the bootstrap implementation. The only thing that stops us in practice from taking a very large *B* is computer efficiency of the ensuing calculations (and the fact usually observed that the increased precision in v is not worth the additional replicates used). It is interesting to put some of these things about the bootstrap in perspective by looking at the delete-one PSU jackknife. The delete-one PSU jackknife is implemented by taking out in turn one unit at a time, and re-weighting appropriately the remaining units; this yields in all *n* jackknife replicates.

Unless n is itself very large, one can use *all* jackknife replicates in practice, something we cannot do with the bootstrap as we just saw.

In practice, many surveys use as many as 500 or even 1,000 replicates. (Given the size of the output files, this is typically the most we can afford while allowing the computation of variance estimates in a timely fashion when needed.) This may appear large compared to what is typically advocated in the i.i.d. case but we have to consider multi-purpose surveys in which domain estimation which involves a subset only of the sampled units; a large B ensures stability over all attempted analyses. Yet, some survey statisticians prefer (and have the luxury) to custom-fit B to the task at hand and refrain from adopting a "universal" value for B. This is typically the case of surveys which only have just a few and well-defined analytical objectives *i.e.*, their output tables are standard and established in advance. In such a situation, for a given sample and estimator, the survey statistician first sets out a small value for B, say 50, and computes the ensuing (main) variance estimate. They he/she augments B to say 100, 150 and so on, and plots the variance estimates obtained as in Figure 1. Typically, a plateau emerges and where it begins is the number of replicates the survey statistician ends up using in that specific context *i.e.*, for that one sample and estimator.

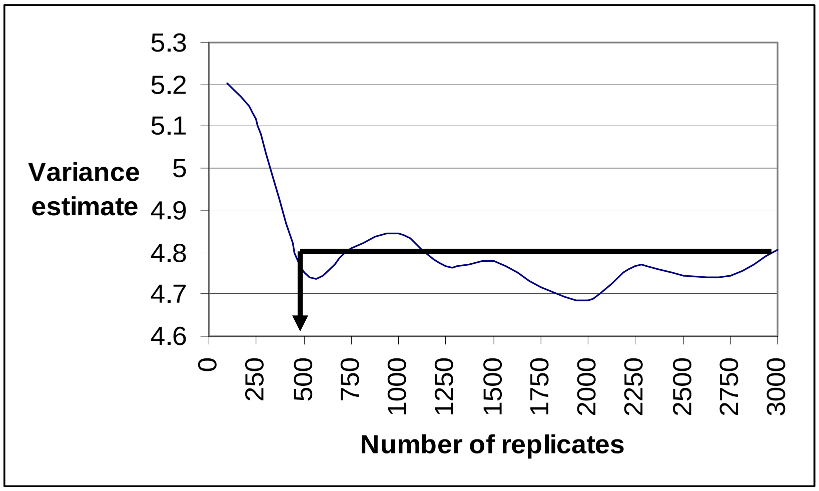


Figure 1: Plotting the bootstrap variance estimates for a given sample and estimator with ever increasing number B of replicates until a plateau emerges; where the plateau begins on the left is the number of replicates (*i.e.*, B=500) to be used in *that* specific context.

With regard to the value of m to choose, some hindsight is provided by the follow-up paper Rao, Wu and Yue (1992) which introduces the so-called bootstrap weights. From a practical perspective, the initial Rao-Wu methodology was not totally satisfactory because it yielded a "one-piece" bootstrap estimator of the mean, and not one which is built from the "weights up" as it is usually the case in survey sampling. This is especially useful to carry out domain estimation, nonresponse and post-stratification methodologies, and obtain estimates of the median. Rao, Wu and Yue (1992) show that one can express the bootstrap estimator for the mean (see equation 2.4 of Rao and Wu (1988)) in the form:

$$\hat{\bar{y}}_{BOOT(b)} = \sum_k w_b(k) y_k / N \tag{5}$$

provided the b^{th} bootstrap weight of unit k, $w_b(k)$, are defined as:

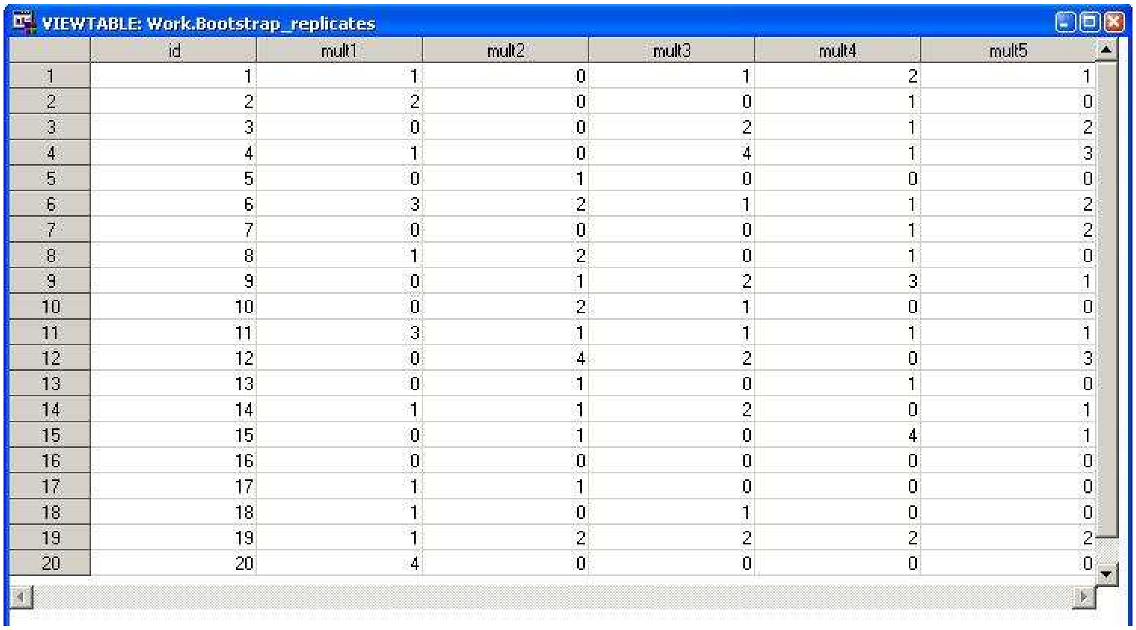
$$w_b(k) = w(k)\left[1 - \sqrt{\frac{m}{n-1}} + \sqrt{\frac{m}{n-1}}\frac{n}{m} mult_b(k)\right] \tag{6}$$

The multiplicity factor $mult_b(k)$ indicates how many times unit *k* was chosen in the *b*th replicate under the with-replacement scheme of the bootstrap and $w(k)$ represents the weight assigned to unit *k*. Observe that the choice of *m*=*n*-1 simplifies significantly the algebraic expression of (6) to

$$w_b(k) = w(k)\frac{n}{n-1}mult_b(k) \qquad (7)$$

To make that choice final in the mind of practitioners, it only remained to observe that a choice of *m*>*n*-1 may lead to some negative values for the weights whereas *m*<*n*-1 (including Rao and Wu (1988) own suggestion of $m \cong n-3$ for large enough *n*) does not yield any significant gain through simulations (and requires a large *n* to begin with).

An example helps to get a feel for the replication procedure and the bootstrap weights. Consider SRSWOR with *N*=20,000 and *n*=20; as a result $f = 0.001 \cong 0$ and $w_k = 1{,}000$ for all *k*. The following snapshot, taken from SAS®, illustrates the bootstrap replicates obtained with *B*=5.



VIEWTABLE: Work.Bootstrap_replicates

	id	mult1	mult2	mult3	mult4	mult5
1	1	1	0	1	2	1
2	2	2	0	0	1	0
3	3	0	0	2	1	2
4	4	1	0	4	1	3
5	5	0	1	0	0	0
6	6	3	2	1	1	2
7	7	0	0	0	1	2
8	8	1	2	0	1	0
9	9	0	1	2	3	1
10	10	0	2	1	0	0
11	11	3	1	1	1	1
12	12	0	4	2	0	3
13	13	0	1	0	1	0
14	14	1	1	2	0	1
15	15	0	1	0	4	1
16	16	0	0	0	0	0
17	17	1	1	0	0	0
18	18	1	0	1	0	0
19	19	1	2	2	2	2
20	20	4	0	0	0	0

Observe first that summing the 20 values of any of the "mult" columns of the snapshot yields 19, as it should, since *n*-1=19 units were selected with replacement. One surprising feature is that while technically no multiplicity value in this case can exceed 19 (which can only occur if multiplicities of 0 are observed *simultaneously* for all other units in the same column, as they must sum to 19), no multiplicity greater than 4 was actually observed. This is because, for a *given* unit *k*, the observed multiplicity is governed by the binomial distribution Bin (*n*-1,1/*n*), though the joint distribution over all units is multinomial. (Observe that *k* is subject to *n*-1 independent trials, each with a probability 1/*n* of being selected; the selection is the success here.) The following table gives the frequencies for large *n*:

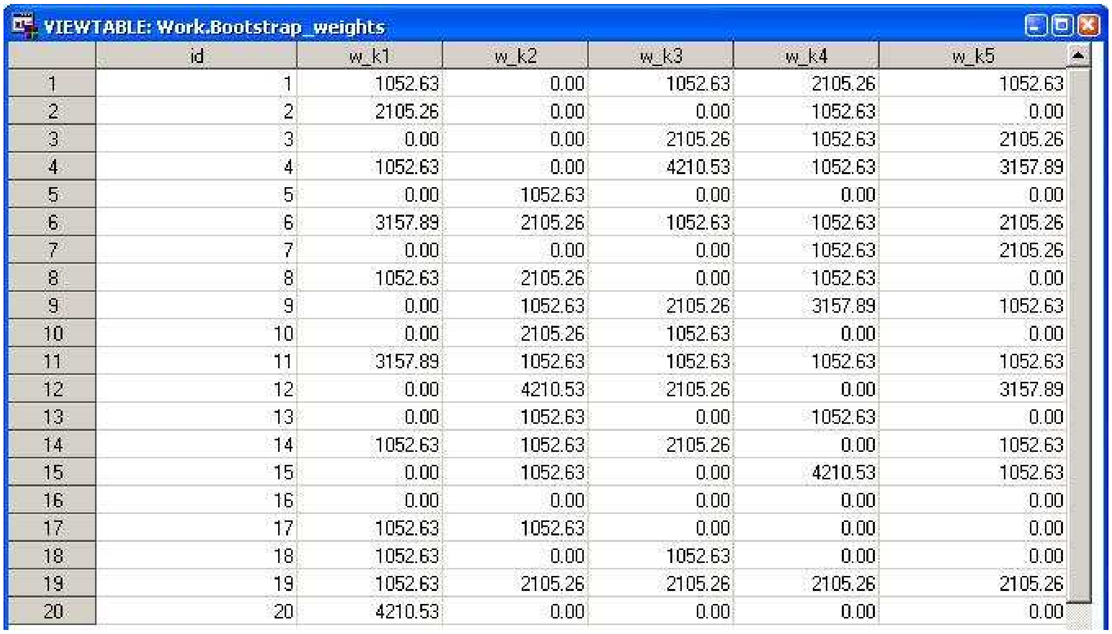
mult(*k*)	Relative frequency (%)
0	36.79
1	36.79
2	18.39
3	6.13
4	1.53
5	0.3
6 to *n*	1.6

Table 1: Expected relative frequencies of the multiplicities for a given unit *k*

In essence, Table 1 says that for the values of B used in practice (*i.e.*, $B \leq 1{,}000$), no multiplicity over 5 has a reasonable chance of ever being observed.

Note: For variance purposes, the stochastic behaviour of the multiplicities should be looked at as a whole *i.e.*, over all k in s at once, and not for a given unit as we just did. Actually, it is because the multiplicities over all k in s are multinomially distributed that their sum always is 19, for if they were *independent* realizations of binomials (19,1/20), the sum would be *random* (though with expected value of 19).

It is instructive to see what the file depicted in the Snapshot above looks like once the multiplicities are turned into bootstrap weights using (7):



	id	w_k1	w_k2	w_k3	w_k4	w_k5
1	1	1052.63	0.00	1052.63	2105.26	1052.63
2	2	2105.26	0.00	0.00	1052.63	0.00
3	3	0.00	0.00	2105.26	1052.63	2105.26
4	4	1052.63	0.00	4210.53	1052.63	3157.89
5	5	0.00	1052.63	0.00	0.00	0.00
6	6	3157.89	2105.26	1052.63	1052.63	2105.26
7	7	0.00	0.00	0.00	1052.63	2105.26
8	8	1052.63	2105.26	0.00	1052.63	0.00
9	9	0.00	1052.63	2105.26	3157.89	1052.63
10	10	0.00	2105.26	1052.63	0.00	0.00
11	11	3157.89	1052.63	1052.63	1052.63	1052.63
12	12	0.00	4210.53	2105.26	0.00	3157.89
13	13	0.00	1052.63	0.00	1052.63	0.00
14	14	1052.63	1052.63	2105.26	0.00	1052.63
15	15	0.00	1052.63	0.00	4210.53	1052.63
16	16	0.00	0.00	0.00	0.00	0.00
17	17	1052.63	1052.63	0.00	0.00	0.00
18	18	1052.63	0.00	1052.63	0.00	0.00
19	19	1052.63	2105.26	2105.26	2105.26	2105.26
20	20	4210.53	0.00	0.00	0.00	0.00

We have observed in the previous snapshot on multiplicities that the column-sums (*i.e.*, over *all* units of the sample) were all 19, a fact guaranteed by the multinomial distribution underlying the bootstrap re-sampling. There is an analogue for bootstrap weights: the sum the values of *any* column weight "w_k" yields 20,000, which is the population size. Because all 5 such sums here yield the *same* value of 20,000, we obtain a bootstrap variance estimate for the estimated size of the population, $v_{BOOT}\left(\hat{N}\right)$, of zero as we are entitled to here. Therefore, the bootstrap "figures out" what is known and what is not through the multinomial distribution. In contrast, suppose a domain D of interest corresponds in the sample to the first eighteen records in the snapshot above depicting the bootstrap weights. The bootstrap variance estimate $v_{BOOT}\left(\hat{N}_D\right)$ of the estimator of the domain's size will be nonzero (as it should since the domain's size is unknown) because the 5 sums will not be identical; the guarantee of getting the same total for all replicates offered by the multinomial does no apply here because not all of the sample was used to build the estimate $\hat{N}_D$.

This case study is important because it clearly shows what assumptions were made to obtain the so-called bootstrap weights, namely they are obtained by expressing the bootstrap estimator of the mean as a weight-based estimator. While it is inviting to turn the bootstrap weight to fit any case, in the end the farther we stray away from these assumptions, the more tenuous the guarantees the theory can offer.

<u>Case study II:</u> Stepping into the jackknife's no man's land or SRSWOR with sampling fraction $f \cong 0$ and estimating the median.

As we mentioned at the beginning, one reason to prefer the bootstrap over the jackknife is the possibility of obtaining correct variance estimates for the median (and other percentiles). While it is now common knowledge that the bootstrap "works" for the median, users are generally not aware of the guarantees offered by the theoretical developments in this field. For instance, very few users know that the guarantees are asymptotic in nature, that some kind of convergence is involved. While the theory says that the bootstrap variance estimator "converges" in both the mean and median cases, it does not say that it "converges" at the *same speed* in *both* cases. And this is something users need to be aware of, since the rate of convergence will dictate which sample sizes are reasonably covered by the theory from those which are not.

What is usually understood by the sentence "the bootstrap works for an estimator" is that the (sequence of) bootstrap variance estimator (indexed by *n*) is consistent for the true variance of that estimator *i.e.*, it converges to that value in probability. For our needs, it suffices to say that convergence in probability describes the fact that, as *n* increases, the distribution of probabilities on the bootstrap variance estimator gets tighter and tighter around the true variance. In that sense, the bootstrap works for SRSWOR with negligible *f* for both the mean and the median (of a population), while the jackknife does *not* work for the median (though it works for the mean).

Note: The bootstrap is being evaluated here solely for its capacities to yield correct variance estimators in the cases which matter most to users through the notion of consistency; this is how the jackknife, its predecessor, is evaluated for the same tasks. But the bootstrap is more than a method to produce correct variance estimators. Indeed, in many situations the bootstrap distribution successfully mimics the sampling distribution of the estimator (in a sense we will make precise later, where we will also show a potential use for that in practice), which is not something the jackknife was ever designed to do.

Let us look now at several graphs describing the notion of the consistency. We performed the following simulation which exploits the equivalent statement that the estimator is consistent if the ratio of bootstrap variance to the true variance converges to 1 as sample size *n* increases; the key steps are:

1) Create a population of very large size (as to have $f \cong 0$ for the sample sizes considered in Step 2 below) with a normally distributed variable of interest, say.
2) Draw 500 samples under SRSWOR for each of the sizes *n*=50, *n*=100, *n*=500 and *n*=1,000.
3) Obtain from Monte Carlo simulations the "true" variance for the mean and for the median.
4) Bootstrap each sample using one hundred replicates (*i.e.*, *B*=100) and obtain for each estimator an estimate of the variance $v_{BOOT,n}$; we thus get, for each *n*, 500 variance estimates.
5) For each *n* and estimator, compute the ratio of the variance $v_{BOOT,n}$ to the corresponding Monte Carlo variance V_{MC}.
6) For each *n* and estimator, obtain the histogram of the 500 ratios.

First, the case of the mean. Since the bootstrap is consistent in this case, we expect to find the 500 ratios spread out to some extent around 1 in the histograms below, but getting tighter and tighter around 1 as *n* increases. The continuous line is the same in all histograms of the mean: it is that of the best fit obtained for *n*=50, displayed to facilitate comparisons across sample sizes.

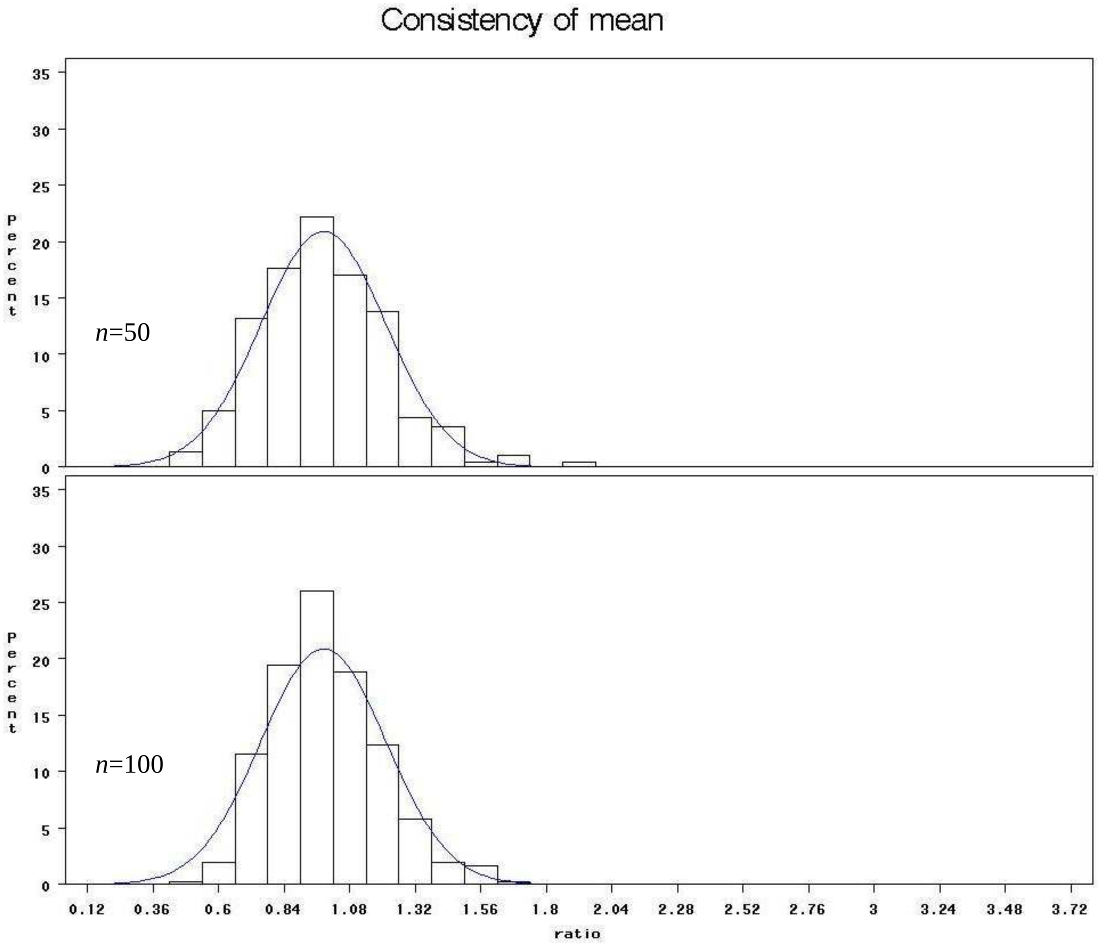


Figure 2: Histogram of 500 ratios of bootstrap variance estimates to the true variance for *n*=50 (top) and *n*=100 (bottom) in the case of the mean. To facilitate comparisons across values of *n*, the best fit line is that of *n*=50.

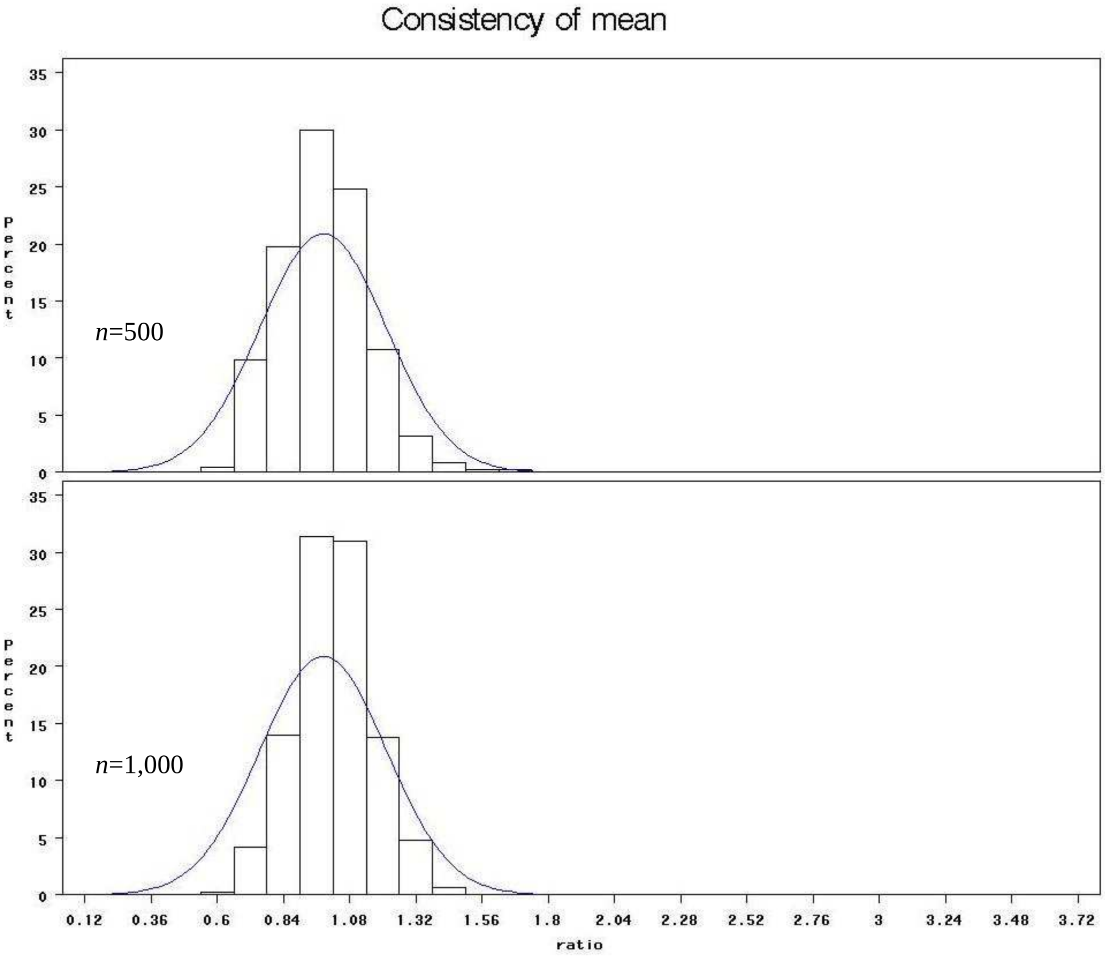


Figure 3: Histogram of 500 ratios of bootstrap variance estimates to the true variance for *n*=500 (top) and *n*=1,000 (bottom) in the case of the mean. To facilitate comparisons across values of *n*, the best fit line is that of *n*=50.

As we can see, the histograms do get tighter and tighter around 1 as *n* increases as the theory asserts. To be fully covered, the rate of convergence depends on both *n* and *B*, the number of replicates used. As mentioned above, we fixed *B* to 100 which was deemed a large enough value to illustrate the idea while allowing the simulations to run in a timely fashion.

Let us now examine the corresponding plots in the case of the median.

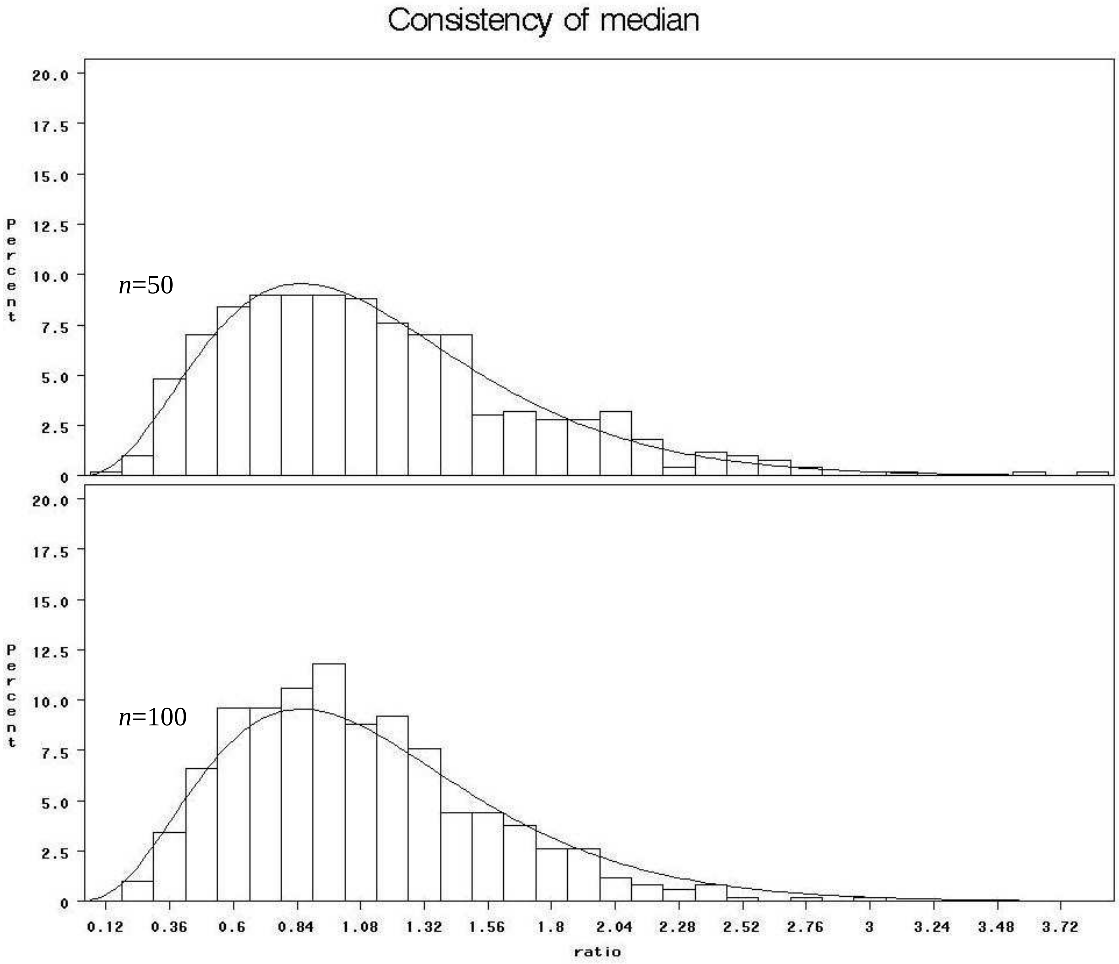


Figure 4: Histogram of 500 ratios of bootstrap variance estimates to the true variance for *n*=50 (top) and *n*=100 (bottom) in the case of the median. To facilitate comparisons across values of *n*, the best fit line is that of n=50.

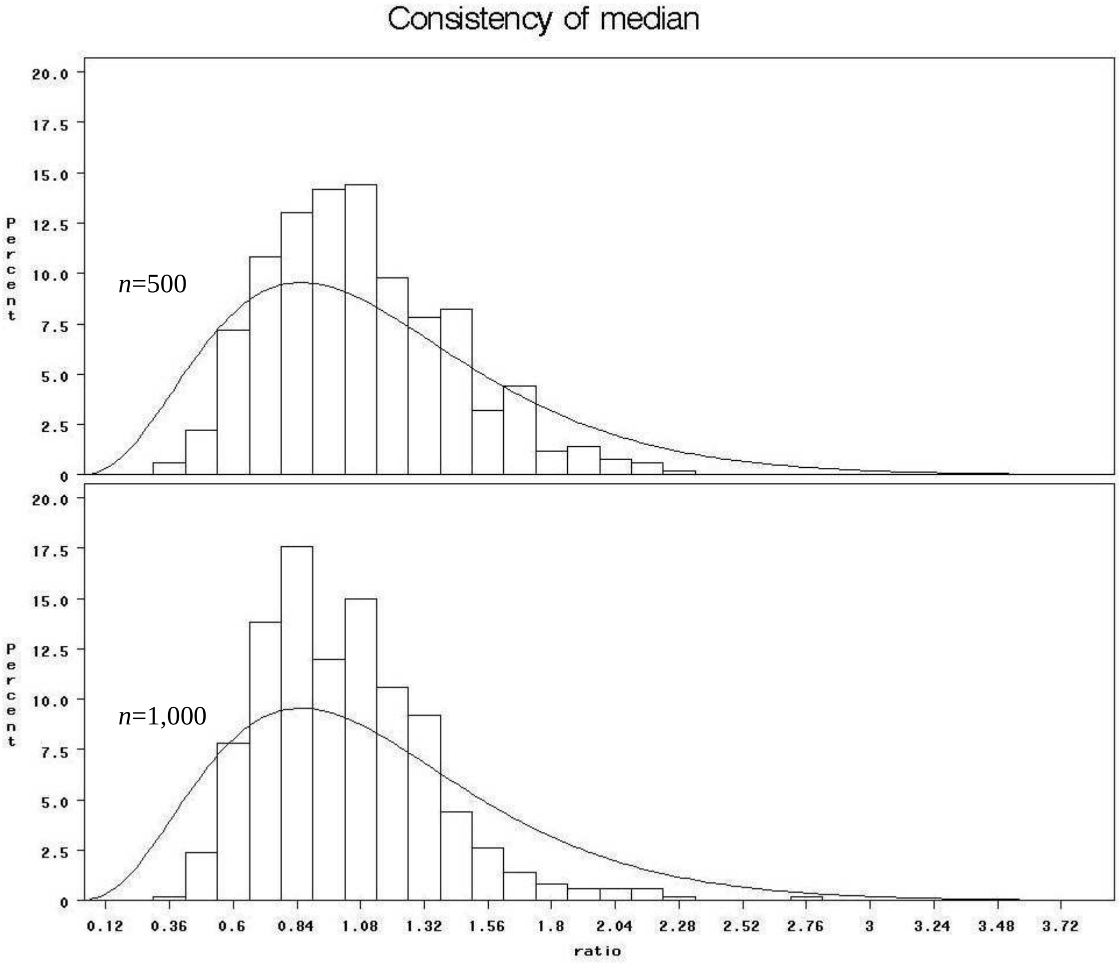


Figure 5: Histogram of 500 ratios of bootstrap variance estimates to the true variance for *n*=500 (top) and *n*=1,000 (bottom) in the case of the median. To facilitate comparisons across values of *n*, the best fit line is that of *n*=50.

While the best fit curve makes it clear that the histograms get tighter and tighter around 1 as *n* increases (as theory says they should), it is also clear that they do at a slower rate than for the mean. In practice this means that the experience a survey statistician has gathered bootstrapping the mean, for instance, is to be extended *with care* to more complex estimators like the median (and smallish domains), even if the bootstrap is known to work for them too. To emphasize this, we could just say that the bootstrap *works very nicely* for the mean and *works* for the median.

For the reader concerned with the spread in the histograms about the median (with some ratios well over 2 even when *n*=1,000), let us take a look at the median with the jackknife. As we have mentioned already, the jackknife is known to be inconsistent for the median. Actually, the theory tells us to expect that the ratio will behave like a random variable with distribution $\left(\chi_2^2/2\right)^2$ (see, for instance, Efron (1979) on page 6). Consequently, it does not converge in probability to 1. Instead of illustrating this through simulations, we preferred to transform the ratio of the jackknife variance to the true variance, $v_{JACK,n}/V_{MC}$, into $2\sqrt{v_{JACK,n}/V_{MC}}$ so that we get a histogram that ought to compare directly (for large *n*) with the chi-square distribution with 2 degrees of freedom (see Figure 6).

Indeed, observe that if

$$v_{JACK,n}/V_{MC} \xrightarrow{Law} \left(\chi_2^2/2\right)^2 \qquad (8)$$

then

$$2\sqrt{v_{JACK,n}/V_{MC}} \xrightarrow{Law} \chi_2^2 \qquad (9)$$

since the square root function is continuous.

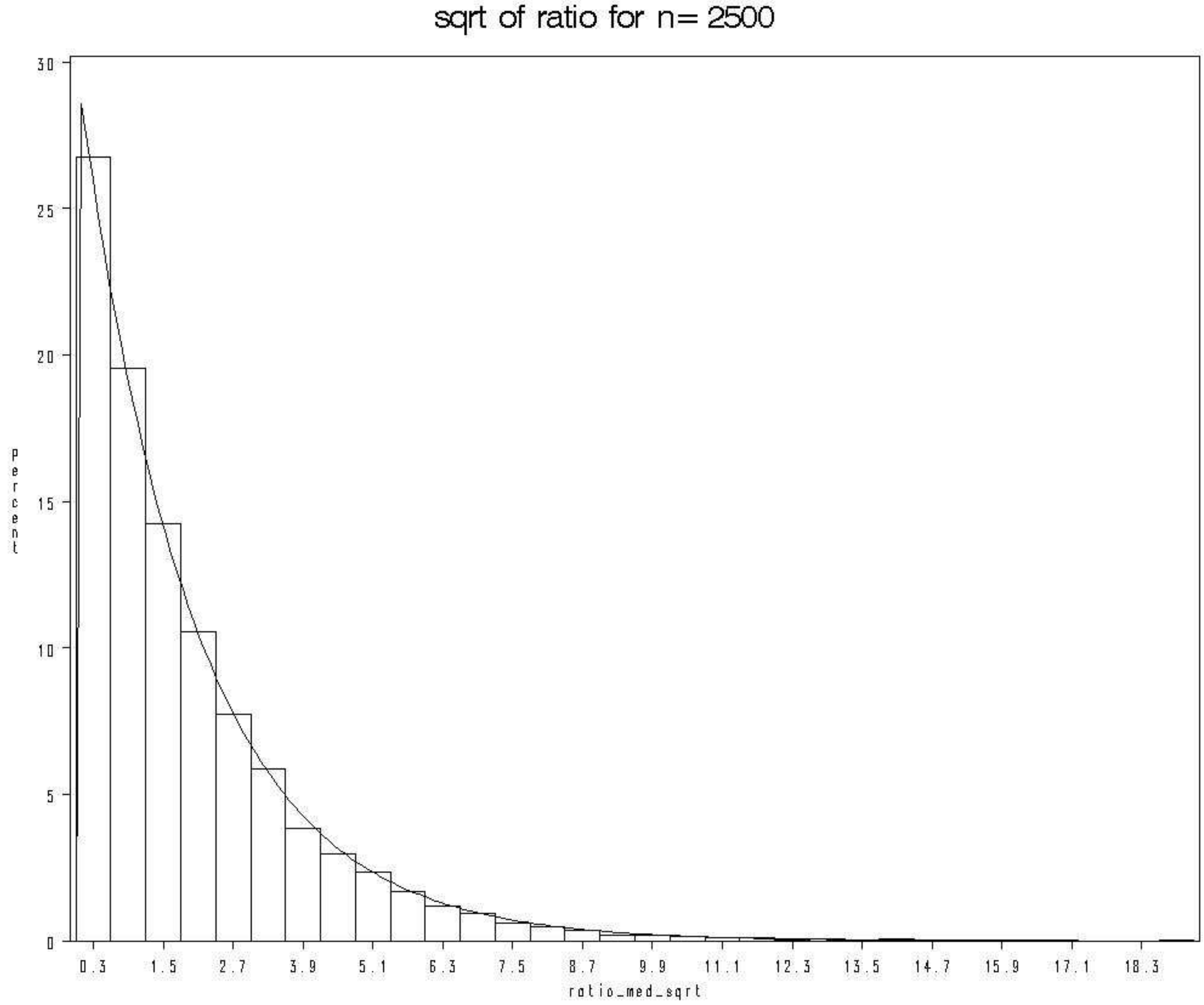


Figure 6: Histogram of $2\sqrt{v_{JACK,n}/V_{MC}}$ for the jackknife with n=2,500 and the chi-square distribution with 2 degrees of freedom as the continuous line.

The histogram clearly shows what happens when a variance estimator is known to be inconsistent for the true variance: even when large sample sizes are involved, the tails refuse to go away and thus will not allow the mass to collect in just one spot.

This case study is a reminder that estimating the variance of a complex estimator from a small sample still is a challenge, even if we have a technique like the bootstrap to carry out the calculations involved (whereas with classical techniques we never usually reach that point). In other words, understanding how an estimator, which uses sample information in a non-straightforward way, "responds" to the complex structure introduced in the sample by the sample design is a difficult task (even for the bootstrap), especially if the information content is minimal *i.e.*, small sample size. From that point of view, the more complex the interaction between estimator and sample design, the more information we need to assess it appropriately.

Case study III: Venturing into WOR territory with SRSWOR and non-negligible sampling fraction *f*.

With this case we are getting closer to the settings many users of survey data encounter in practice.

Even though the bootstrap weights are not provided for this case in Rao, Wu and Yue (1992), many users readily make the extension from the negligible sampling fraction case described above by showing that the bootstrap estimator for the mean (see equation 4.1 of Rao and Wu (1988)) can be expressed in the form

$$\hat{\bar{y}}_b = \sum_k w_b(k) y_k / N \qquad (10)$$

provided the b^{th} bootstrap weight of unit *k*, $w_b(k)$, is defined as

$$w_b(k) = w(k)\left[1 - \sqrt{1-f} + \sqrt{1-f}\,\frac{n}{n-1} mult_b(k)\right] \qquad (11)$$

Even though the bootstrap is carried out using a with-replacement scheme, we can still hope to track down the variance of the original without-replacement sampling scheme by introducing into the equations the sampling fraction *f*. The bootstrap weights in this case come with a surprise: if a unit has multiplicity zero *i.e.*, it was not chosen in a given replicate by the with-replacement design, then it gets a non-zero bootstrap weight nonetheless. This in itself is not problematic, *provided* all units of the sample are listed in the replicates file. To drive home this point, consider SAS® which is the software used at Statistics Canada. (It is not an issue if you do not know SAS®, as no programming is involved.) Let us look at how the replicates file depicted in the snapshot above was obtained; to simplify things, let us focus on the first two units of that file *i.e.*, id=1 or 2:

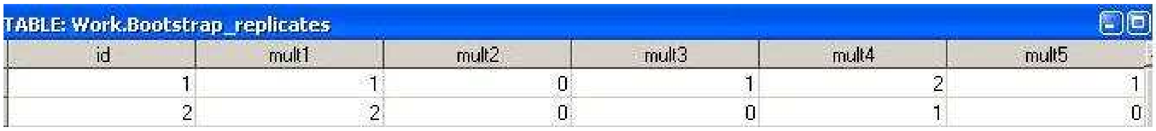
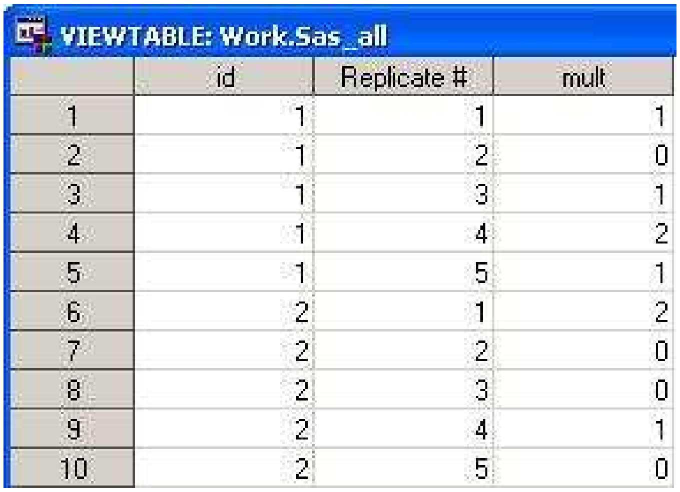


TABLE: Work.Bootstrap_replicates

id	mult1	mult2	mult3	mult4	mult5
1	1	0	1	2	1
2	2	0	0	1	0

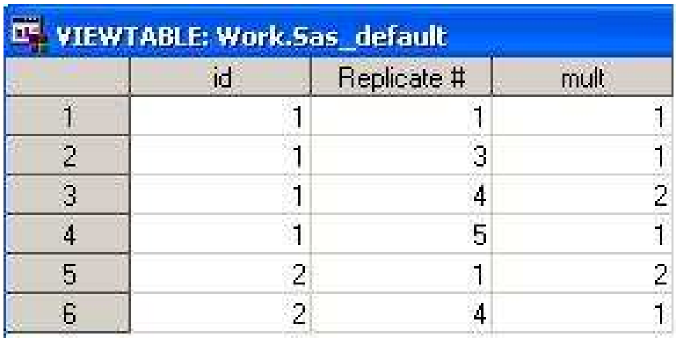
This is not directly an output of the SAS® PROC SURVEYSELECT procedure used to draw the replicates from the observed sample; it is actually the transpose of the following "rows-file" which *can* be obtained from SAS® PROC SURVEYSELECT:



VIEWTABLE: Work.Sas_all

	id	Replicate #	mult
1	1	1	1
2	1	2	0
3	1	3	1
4	1	4	2
5	1	5	1
6	2	1	2
7	2	2	0
8	2	3	0
9	2	4	1
10	2	5	0

But this is not the procedure's *default* output, which rather is:



	id	Replicate #	mult
1	1	1	1
2	1	3	1
3	1	4	2
4	1	5	1
5	2	1	2
6	2	4	1

In other words, in the default output file the units with zero multiplicity are *not* provided (they are only outputted under the option OUTALL of the PROC SURVEYSELECT statement). Consequently, if the default file is used to compute the bootstrap weights and then transposed, the zero-multiplicities will have missing bootstrap weight values in the "squared file" which will be set presumably to zero afterwards to avoid releasing missing values.

This subtle point of the implementation of the bootstrap in the case of non-negligible f has led to erroneous early (and still persistent to this day) assessments to the effect that the bootstrap did not quite work for the mean, contrary to what was expected. This is likely because the key contributors which are the units with zero multiplicity were missing in the simulations performed due to the use of the default options for the output of the selection procedure in SAS®. While individually the missing bootstrap weights tend to be small compared to the non-zero multiplicities records, there are many of them around (recall from Table 1 that about 40% of all units in a given replicate have a multiplicity of zero). If the zero multiplicities are handled the right way then the ensuing bootstrap variance estimates *do* behave for the mean like we expect them to.

To explore on our own the consistency of the bootstrap in the case of the median, we redid the simulations described in Case Study III but this time using f=1/2 instead of $f \cong 0$. To accomplish this we actually had to change populations with each increased sample size n, as n and N are linked by f=1/2. In other words, we considered a sequence of increasing sample and population sizes such that f=1/2 in all cases. We have obtained basically the same histograms (not shown here) as in the negligible f scenario.

Case study IV: Nonresponse and the bootstrap featuring SRSWOR and uniform nonresponse.

Nonresponse is unfortunately a key feature of day-to-day surveying activities, but one usually missing from practical discussions involving the bootstrap.

Nonresponse is a threat to a survey for at least the following two reasons: a) the loss of analytical power that comes with a smaller effective sample size; b) the possibility of bias due to differences in the profile of respondents and nonrespondents.

The first issue is usually addressed at sampling where the loss is anticipated and compensated by sending in the field a larger sample than what is truly needed. The second issue is addressed at the weighting stage using a suitable corrective methodology to thwart the effects nonresponse would have on the estimates if ignored. One of these methodologies involves the construction of Response Homogeneous Groups (RHGs). RHGs are typically constructed using the response propensity score from a logistic model (see Little (1986) and Eltinge and Yansaneh (1997)) or the cross-classification method. Basically, the score method gathers respondents with similar predicted response propensities which are obtained from a logistic model fitted to the collected data. One feature of this approach is the controlled number of RHGs which the method gives rise to. With the cross-classification approach (often known under the name of the software commonly used to implement it, KnowledgeSeeker®), groups are formed by crossing the (possibly many) variables found to "explain" the nonresponse. Since crossing variables is multiplicative, this approach often gives rise to a large number of RHGs, especially if used routinely in production.

The proliferation of RHGs has a soothing effect because it gives the impression that the most has been done to thwart nonresponse: aggressive measures were taken. The issue is that with the growing number of RHGs comes smaller and smaller RHGs which, in turn, often results in unstable variance estimates. Consequently, a trade-off has to be reached between efforts to reduce bias and the instability in the variance estimates these efforts will bring. But how? This can be accomplished using the so-called bootstrap histograms which most users are not aware exist.

The widely accepted way to implement the bootstrap in presence of nonresponse is to compute, within each replicate and RHG determined based on the sample, a nonresponse adjustment using bootstrap weights of respondents and nonrespondents. (Consequently, the construction of the RHGs themselves is not revisited by the bootstrap procedure: we take the RHGs as they stand from examination of the nonresponse patterns in the sample.)

To illustrate the idea of how bootstrap diagrams can be used to see if the construction of RHGs undertaken will lead or not to stable estimates, consider the following example. Suppose a sample of size 100 was drawn under SRSWOR and that uniform nonresponse yielded 75 respondents. Under the assumption of uniform nonresponse, a single RHG would do but we will nonetheless create two RHGs, one big which represents 90% of the sample, and one small (the remaining 10%), just for the sake of comparison. Indeed, aside from their sizes, the two RHGs should behave the same with respect to the nonresponse mechanism. As mentioned above, the bootstrap is conducted by re-sampling *s* and re-computing all 1,000 nonresponse adjustments for each RHG by using the 1,000 sets of bootstrap weights. We then plot the 1,000 nonresponse adjustments obtained for each RHG.

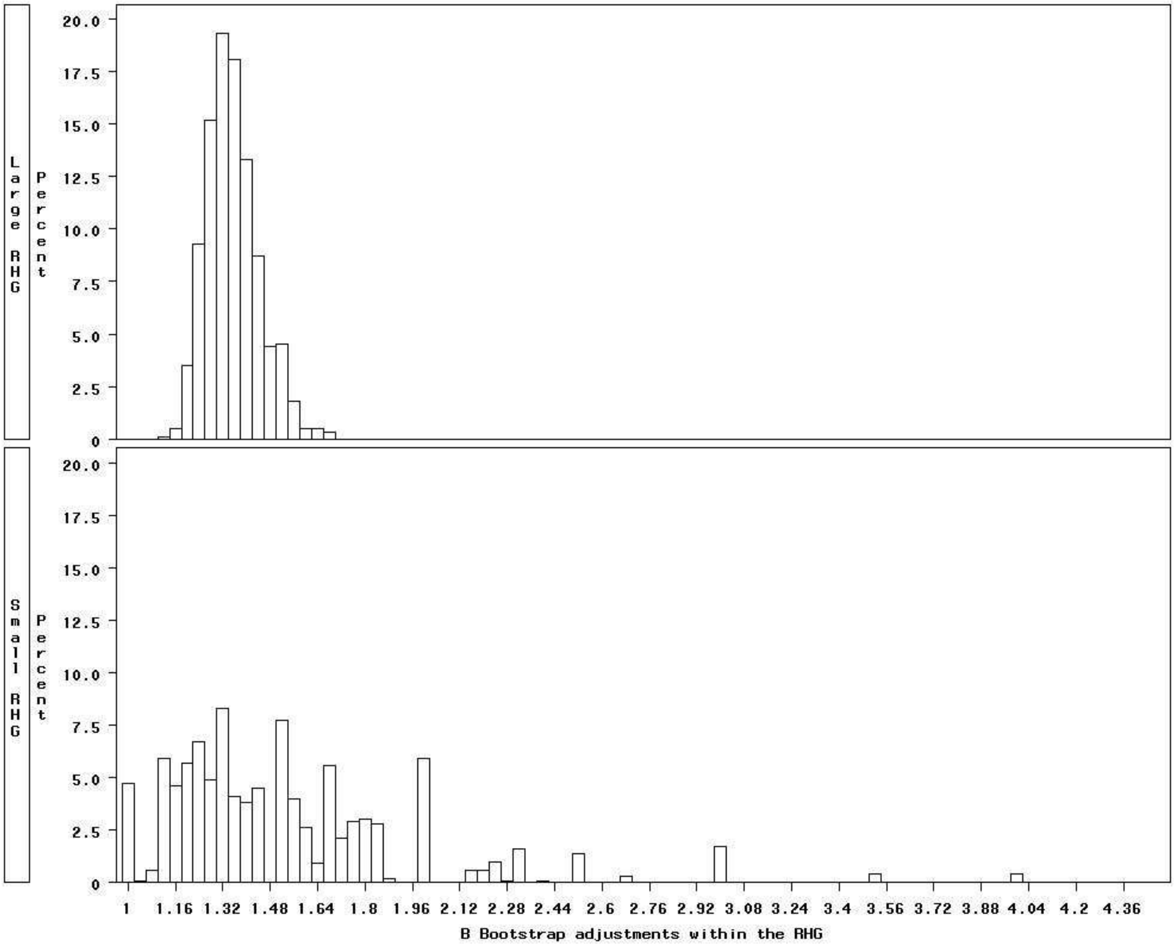


Figure 7: Histogram of bootstrap nonresponse adjustments by RHG, one big and one small, in the case of uniform nonresponse in a SRSWOR sample.

Again, the two RHGs ought to behave the same, except possibly for the effect their size may have on the computed nonresponse adjustments as a whole. In the large RHG the adjustments are gathered tight around the nonresponse adjustment obtained from *s* whereas in the small RHG they are all over the place. It is clear that the nonresponse methodology has given rise to unstable adjustments in the small RHG case, and all this is seen from the bootstrap histogram.

A similar case can be made for post-stratification, for instance, where the creation of too many post-strata may jeopardize stability. In this case, as in the case of nonresponse, instead of relying on a rule of thumb about the appropriate size of the groups usually employed to avoid abusive use of these methodologies, one can use instead the diagnostic feature of the bootstrap for that very same purpose.

But *even* if this bootstrap is implemented correctly in the presence of nonresponse, and the methodology used has been scrutinized through the bootstrap histogram technique just depicted, it is a fact that it nonetheless fails to capture *all* of the variance due to the design and response mechanism.

Let us use the simplest of all cases to illustrate this: SRSWOR, uniform nonresponse and the mean. In this case it is well-known that the true variance under total response is:

$$V\left(\hat{\bar{y}}\right)=\frac{1-n/N}{n}S_U^2 \tag{12}$$

Under the assumed nonresponse the usual estimator of that variance is (*e.g.* Särndal *et al*. 1992, Section 15.6.1):

$$\hat{V}\left(\hat{\bar{y}}\right)=\frac{1-r/N}{r}s_r^2 \tag{13}$$

where s_r^2 is the variability observed in the *y*-values among the respondents.

It is interesting to observe that in this simple case, the net effect on variance of the mean of the sampling and occurring nonresponse is known: it amounts to a sampling of *r* units straight from the population, as if the original sampling of *n* units from *N* never occurred. Of course, in practice we almost never know what form the variance takes for any given estimator when nonresponse comes into the picture, and this is the very reason why replication methods are used in practice.

Of the myriad ways to express the total estimated variance (13) as the sum of two algebraic entities, only a few make conceptually any sense. And of the latter, one is most useful to understand what part of (13) this bootstrap does capture and what part is does not: it is Fay's inverted approach (see Fay (1991) and Shao and Steel (1999) for a description of the idea, and Haziza (2005), Section 5.3, for the worked out calculations underlying our development here). In our scenario Fay's inverted approach yields the following decomposition of the variance estimator (13):

$$\hat{V}\left(\hat{\bar{y}}\right)=\underbrace{\frac{1-n/N}{r}s_r^2}_{\hat{V}_I}+\underbrace{\frac{n-r}{Nr}s_r^2}_{\hat{V}_{II}}=\frac{1-r/N}{r}s_r^2 \tag{14}$$

The component $\hat{V}_I$ is an estimate of the following expectation:

$$V_I=E_rV_D\left(\hat{\bar{y}}\middle|r\right) \tag{15}$$

Notes: The subscript *r* refers to the nonresponse mechanism and *D* refers to the design, and the variable *r* which appears in the condition refers to the observed set of respondents. (Taking the design expectation *given* the set of respondents, and not vice-versa, constitute the heart of Fay's inverted approach.) Also, the subscripts *I* and *II* simply mean “one” and “two”, and not “first-phase” and “two-phase” as one could be led to think since nonresponse is an important case of the two-phase variance estimation framework (*e.g.*, Section 9.8 of Särndal *et al.* (1992)).

Since we only have one set of respondents at our disposal, V_I is estimated simply by estimating the design variance $V_D\left(\hat{\bar{y}}\middle|r\right)$ given the respondents. But this is precisely what the bootstrap is providing: an estimate of the design variance given the respondents. Indeed, heuristically, to the (Rao-Wu) bootstrap the response status of a unit is nothing else than a 0-1 value that has been assigned (deterministically) to each unit of the sample: the bootstrap does not “perceive” any other selection mechanism than the sample design it has been subject to. Therefore, since no replication of any kind of the nonresponse mechanism has been attempted by the (Rao-Wu) bootstrap, the response status of a unit is just like a post-strata indicator: a status completely determined once the sample has been selected. This is not to say that the (Rao-Wu) bootstrap does not capture *any* of the variance due to nonresponse. It simply captures the variance due to nonresponse as if nonresponse were a deterministic process.

Therefore, all the (Rao-Wu) bootstrap can hope to provide us with is a numerical value which would match exactly $\hat{V}_I$ in equation (14) if no Monte Carlo approximation was needed, which means that $\hat{V}_{II}$ is completely missed out. This begs the question of when exactly is $\hat{V}_{II}$ small enough as to be ignored. An easy calculation shows that for a given *r*:

$$\frac{Missed}{Total}=\frac{\frac{n-r}{Nr}s_r^2}{\frac{N-r}{Nr}s_r^2}\cong 0 \qquad (16)$$

whenever:

$$\frac{n}{N}\cong 0 \qquad (17)$$

The idea is that when the sampling fraction is very small, the total variance is dominated *in magnitude* by the sampling and *not* the occurring nonresponse. As a result, when the sampling fraction *f* is near 0, whatever contribution to the variance the bootstrap picks up due to nonresponse (by treating it as a deterministic mechanism given *s*), in addition to the contribution it captures from the sample design, is more than enough to get in the end a very reasonable variance estimate of the total variance.

At the other end of the spectrum this suggests that when we have a census *i.e.*, *f*= $n/N=1$ and nonresponse, then the bootstrap captures nothing of the total variance. One often hears practitioners say that the bootstrap fails when we have nonresponse as we get near to a census as they observe the bootstrap variance estimate deviates more and more from the total variance to be captured. It is not that the bootstrap *fails*: it still works, even then. It works at capturing the variance introduced by the sample design; if there are other factors at work, other mechanisms such as nonresponse, *and* these are not fully accounted by the design *then* the bootstrap will understandably miss them. One could say that to apply the bootstrap in the presence of nonresponse in census-like conditions is an ill-application of the method since there is hardly any design variance to be picked up: the total variance is dictated by the occurring nonresponse which the bootstrap we use is not designed to capture.

Another way of seeing this comes directly from the bootstrap weights (11):

$$w_b(k)=w(k)\left\{1-\sqrt{1-f}+\sqrt{1-f}\,\frac{n}{n-1}mult_b(k)\right\}$$

Indeed, if *f*=1, then the bootstrap weights become:

$$w_b(k) = w(k)\left\{1 - 0 + 0\frac{n}{n-1}mult_b(k)\right\} = w(k) \tag{18}$$

In other words, the bootstrap weight of a unit is the same in all replicates: the randomness introduced by the bootstrap re-sampling has been completely neutralized.

In general it is not easy to say how much of the variance is missed by the bootstrap in presence of nonresponse since an explicit decomposition like (14) cannot usually be obtained (*e.g.*, the median and/or non-uniform nonresponse). But one thing is nonetheless clear: for any given estimator, the larger the sampling fraction the larger the component to the variance missed by this bootstrap.

In addition to the case of SRSWOR followed by uniform nonresponse, another instance of the two-phase framework for variance estimation is of particular interest to users: the two-stage sampling design. To illustrate, we will exploit a SRSWOR design at both stages *i.e.*, example 4.3.1 of Särndal *et al.* (1992):

The true variance of the total (equation 4.3.22 of Särndal *et al.* (1992)) is:

$$V_{2st}(\hat{t}) = N_I^2 \frac{1 - n_I / N_I}{n_I} S_{tU_I}^2 + \frac{N_I}{n_I} \sum_{U_I} N_i^2 \frac{1 - n_i / N_i}{n_i} S_{yU_i}^2 \tag{19}$$

where U_I is the list of clusters, N_I is the number of clusters in U_I, n_I is the number of clusters that were selected at first-stage and $S_{tU_I}^2$ is the variance of the N_I clusters, and the subscript *i* refers to their inner-cluster counterparts.

The first component measures the contribution to variance due to unequal totals across clusters and the second component measures just how spread the *y*-values are in each of the cluster.

An unbiased variance estimator is (equation 4.3.23 of Särndal *et al.* (1992)):

$$\hat{V}_{2st}(\hat{t}) = \underbrace{N_I^2 \frac{1 - n_I / N_I}{n_I} S_{ts_I}^2}_{\hat{V}_A} + \underbrace{\frac{N_I}{n_I} \sum_{s_I} N_i^2 \frac{1 - n_i / N_i}{n_i} S_{ys_i}^2}_{\hat{V}_B} \tag{20}$$

Simulations show clearly that by bootstrapping this design, the bootstrap variance estimate seeks to match $\hat{V}_A$ *i.e.*, $\hat{V}_A = E_{boot}(v_{bs} | s_I \ \& \ s_i)$. However, tempting as it is to assume that the first component $\hat{V}_A$ of (20) is unbiased for the "matching" first component of (19) it is incorrect. It is in fact interesting to see that $\hat{V}_A$ is unbiased for a variance term V_A which corresponds to a mixture of the across-clusters and inner-clusters components of the true variance. In other words, from a bootstrap perspective, the decomposition of the total variance provided by (19) is not "natural", but the following decomposition is:

$$V_{2st}(\hat{t}) = V_A + V_B \tag{21}$$

where $\hat{V}_A$ is unbiased for V_A (and so is $\hat{V}_B$ for V_B incidentally); calculations show that these are explicitly:

$$V_A = N_I^2 \frac{1 - n_I / N_I}{n_I} S_{tU_I}^2 + N_I \frac{1 - n_I / N_I}{n_I} \sum_{U_I} N_i^2 \frac{1 - n_i / N_i}{n_i} S_{yU_i}^2 \tag{22}$$

$$V_B = \sum_{U_I} N_i^2 \, \frac{1-n_i / N_i}{n_i} \, S_{yU_i}^2 \qquad (23)$$

An alternate form for V_A is of interest:

$$V_A = N_I^2 \, \frac{1-n_I / N_I}{n_I} \left[S_{tU_I}^2 + \frac{1}{N_I} \sum_{U_I} N_i^2 \, \frac{1-n_i / N_i}{n_i} \, S_{yU_i}^2 \right] \qquad (24)$$

This last expression suggests that when we compute $\hat{V}_A$ (or when we bootstrap the first-stage sample), we are trying to capture the variance in totals across clusters *and* the average over the clusters of the inner-cluster variances in *y*-values. This form also makes the comparison of V_A to the total variance V_{2st} more direct:

$$\frac{V_A}{V_{2st}} = \frac{S_{tU_I}^2 + 1/ N_I \sum_{U_I} N_i^2 \, \frac{1-n_i / N_i}{n_i} \, S_{yU_i}^2}{S_{tU_I}^2 + \left[N_I^2 \, \frac{1-n_I / N_I}{n_I} \right]^{-1} \frac{N_I}{n_I} \sum_{U_I} N_i^2 \, \frac{1-n_i / N_i}{n_i} \, S_{yU_i}^2} \qquad (25)$$

Indeed, further algebraic simplifications of (25) show that this ratio is about 1 whenever $n_I << N_I$ *i.e.,* $f_I \cong 0$ which is basically what we had concluded above in the case of nonresponse. To summarize, when it comes to the variance captured by the bootstrap under a two-phase or two-stage design, the first-phase/first-stage sampling fractions dictates it all.

Case study V: Routinely performed survey data analysis.

In social surveys users are typically handed out a file of survey data along with bootstrap weights for them to investigate “on their own”, as opposed to business surveys where estimates are often obtained in close collaboration with the survey methodologist and constitute a standard output product of the survey. Consequently, variance estimates are often obtained with the barest minimum knowledge of the bootstrap method, if not of variance estimation in itself. Indeed, it is the author’s experience that users see the bootstrap weights at best as an annoyance. It thus surprises them when they realize that the bootstrap can actually be an ally through its diagnostic capacities.

In the previous case study we gave an example of the bootstrap diagnostic capacities with the creation of RHGs, but this was addressed primarily to survey samplers since it occurs before the bootstrap weights are released to users. The diagnostic we describe here shows that there is also something in the bootstrap for users themselves. We use for illustrative purposes the problematic of routinely built confidence intervals and proportions.

Here we go beyond the bootstrap as a technique to estimate the sampling variance and look at its ability to reproduce, in many contexts, the sampling distribution of the estimator at hand. More precisely, it is known that for SRSWOR with $f \cong 0$ and the mean that:

$$\hat{\bar{y}}_{BOOT} - \hat{\bar{y}} \xrightarrow{Law} \hat{\bar{y}} - \bar{y} \qquad (26)$$

Some exploit this fact to propose new practices (*e.g.*, construct confidence intervals using bootstrap percentiles) but here we will use it to *question* an existing practice. The extent to which (26) holds for just about any estimator and sample design other than the one considered here is debatable and must therefore be used with caution.

Either through software or from their own programming, users routinely construct confidence intervals assuming normality, whether they realize it explicitly or not. This is particularly true when mass production of estimates in a timely fashion is under way after the collection has ended.

Consider the situation where a SRSWOR sample of size 100 is observed and the bootstrap methodology is applied to yield bootstrap weights such that the following 95% confidence intervals (based on the normality assumption) were obtained for the proportion of two separate characteristics, one estimated at 3% (true *p*=5%) and the other at 22% (true *p*=20%):

(-0.36%, 6.36%)

and

(13.89%, 30.11%)

Something is clearly wrong with the confidence interval involving the smaller proportion: the negative lower bound does not make sense. In practice, though, this is likely to be treated lightly by simply rounding the lower bound to zero and have the originally computed confidence interval be released as: (0.0%, 6.36%). To explain why things have gone awry with the smaller proportion it is natural to suspect the normality assumption is at fault, but how to confirm that? Through the bootstrap. What follows is the histogram of the bootstrap estimates for each estimated proportion which underlay the bootstrap variances from which the confidence intervals were obtained.

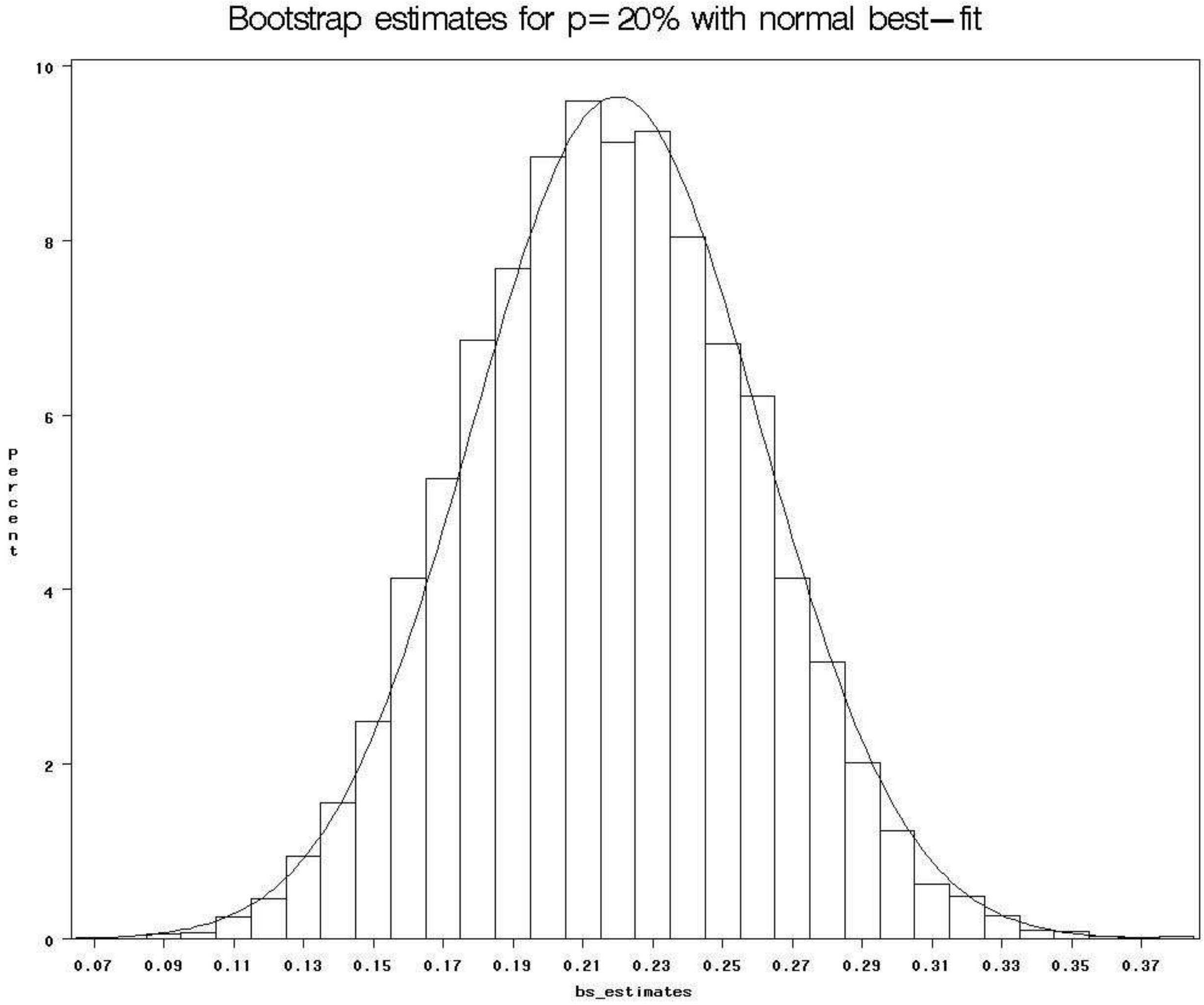


Figure 8: Histogram of the bootstrap estimates for the estimated proportion $\hat{p}$ =22%.

The histogram is well-fitted overall by the normal curve and thus the assumption of normality made about the sampling distribution of the estimator is reasonably met here. In this case, the left tail goes quickly to zero, well before the origin is encountered: the normal curve, despite it left tail which goes to minus infinity is a reasonable model for the sampling distribution which only admits after all positive values.

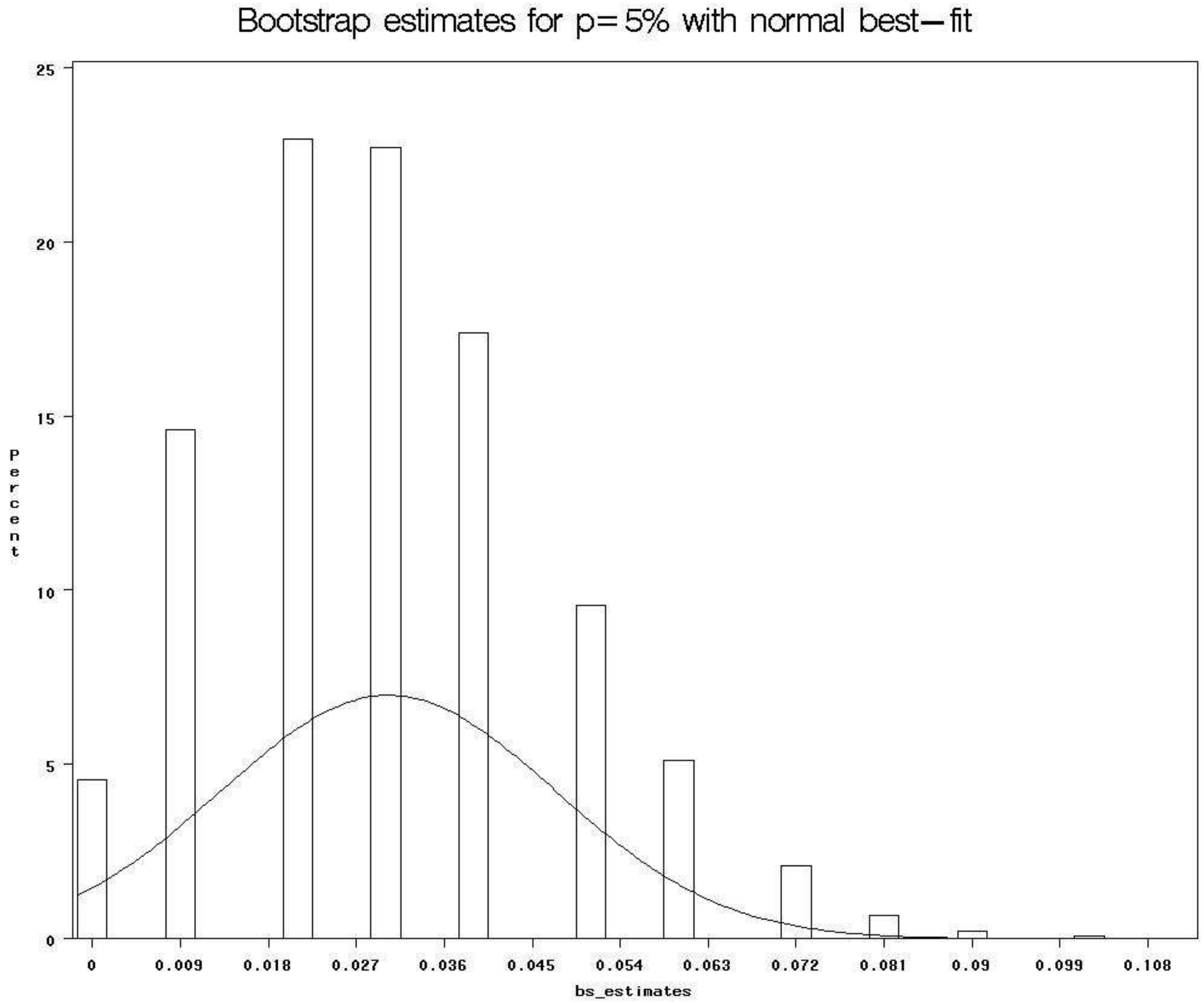


Figure 9: Histogram of the bootstrap estimates for the estimated proportion $\hat{p}$ =3%.

In the problematic case of the smaller proportion of the two, it is clear that the histogram and the best-fit normal curve are far apart: it is not reasonable in this case to assume normality. We can actually see why a negative lower bound was obtained: the best-fit normal curve’s left tail extends well into the negative axis before it can be reasonably assumed to be zero as it was the case for the larger proportion we have just looked at.

In summary, bootstrap histograms should be used at all steps, whether by the survey sampler to check on the adjustments introduced or by the user in the analysis performed. In either case, if the histogram looks strange with say some replicated values standing clearly apart from the rest, then a deeper investigation of the methodology under consideration should be undertaken. In other words, it is not required to know what theory says the histogram ought to approximate to get a valid diagnostic procedure: any peculiar feature of the histogram in itself should be treated as suspicious and investigated on these grounds.

Concluding remarks

There are several aspects of the implementation of the Rao-Wu rescaling bootstrap that require finesse: as with any methodology, going from theory to practice is sometimes tricky. In this paper we have presented only a few issues since a more complete picture would have required a lengthy description of related survey situations (especially about longitudinal surveys); see Girard (2007) and Girard *et al.* (2009) for more details. It is our experience that when common (and sometimes not so common!) pitfalls are avoided, and its limitations kept well in mind, the implementation of the Rao-Wu bootstrap method goes smoothly and is overall versatile and easy to use.

Acknowledgements

The author is indebted to Sarah Franklin, Yves Lafortune (who, incidentally, spared the author the trouble of deriving equations (22) and (23) by providing his own), Professor J.N.K. Rao and Wesley Yung for discussions that contributed in making this a better paper.

References

Note: No effort was made to retrace the original contributions behind the references made in this paper. The author has used sources which were particularly readable and provided a central view, rather than quote original results from many separate sources.

Efron, B. (1979) Bootstrap methods: another look at the jackknife (The 1977 Rietz lecture). The Annals of Statistics Vol. 7 No.1 pp.1-26

Efron, B. 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 68.3. pp. 589-599

Efron, B. and R. Lepage (1992). Introduction to the bootstrap. Part of Exploring the limits of the bootstrap, edited by Raoul Lepage and Lynne Billard. Wiley, New-York.

Eltinge, J.L. and I.S. Yansaneh. 1997. Diagnostics for formation of nonresponse adjustment cells with an application to income nonresponse in the U.S. Consumer Expenditure Survey. Survey Methodology, 23, pp.33-40.

Fay, B.E. 1991. A design-based perspective on missing data variance. Proceedings of the 1991 Annual Research Committee, US Bureau of Census, pp. 429-440.

Girard, C. 2007. How to avoid getting all tied up bootstrapping a survey: A walk-through featuring the National Longitudinal Survey of Children and Youth. Statistics Canada Working Paper HSMD-2007-001E.

Girard, C., M. Ferland, S. Franklin and M. Tremblay (co-editors). 2009. Rethinking the NLSCY weighting methodology: the lessons learned and the innovations introduced. Statistics Canada Working Paper HSMD-2009-002E.

Haziza, D. 2005. Inférence en presence d'imputation simple dans les enquêtes: un survol. Journal de la Société Française de Statistique tome 146, no.4.

Little, R.J.A. 1986. Survey nonresponse adjustments for estimates of means. International Statistical Review, 54, pp.139-157.

Rao, J.N.K. and C.F.J. Wu. 1988. Re-Sampling Inference With Complex Survey Data. JASA, March Vol. 83 No. 401 pp. 231-241.

Rao, J.N.K., C.F.J. Wu and K. Yue. 1992. Some recent work on re-sampling methods for complex surveys. Survey Methodology 18 pp. 209-217.

Rust, and J.N.K. Rao. 1996. Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research. 5. pp. 283-310.

Särndal, C.-E., B. Swensson and J. Wretman. 1992. Model Assisted Survey Sampling. Springer. New-York.

Shao, J. and P. Steel. 1999. Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions. JASA, March, Vol. 94, No. 445 pp. 254-265.

Valliant, R. 2007. An overview of the pros and cons of linearization versus replication in establishment surveys. Proceedings of the Third International Conference on Establishment Surveys (ICES III).