

The Finite Population Bootstrap - from the Maximum Likelihood to the Horvitz-Thompson Approach

Andreas Quatember
Johannes Kepler University

Abstract

The finite population bootstrap method is a computer-intensive alternative to estimate the sampling distribution of a sample statistic. In one possible approach, generation of an artificial population, the so-called “bootstrap population”, becomes a necessary step between the original sample drawn and the resamples needed to mimic this distribution. For this, the main problem is how to create a bootstrap population that is adequately close-to-real for resampling. For this process, the bootstrap population need not be generated in reality. After an overview of different methods of the finite population bootstrap, this paper presents an approach, based on the idea behind the Horvitz-Thompson estimator, which allows not only whole units in the bootstrap population but also parts of whole units. In a simulation study, this method is compared with a more heuristic technique, taken from the bootstrap literature.

Keywords: survey methodology, sampling theory, bootstrap, variance estimation.

1. Introduction

The bootstrap technique was originally published by [Efron \(1979\)](#) for the problem of estimating the sampling distribution of a statistic $\hat{\theta}$, depending on a random sample and an unknown probability distribution of a variable y under study, on the basis of the observed sample. This procedure can be described in the following way (cf. [Efron 1979](#), p.3):

1. An i.i.d. sample of size n is drawn to observe an empirical distribution of the study variable.
2. From this empirical distribution, a bootstrap i.i.d. resample of size n is considered.
3. The sampling distribution of the statistic of interest is approximated by the theoretical bootstrap distribution of it.

This bootstrap distribution equals the sampling distribution of the statistic if the empirical distribution of the variable equals its probability distribution. [Efron \(1979\)](#) considers as “the difficult part of the bootstrap procedure ... the actual calculation of the bootstrap distribution” (p.4). Three methods are possible: The direct theoretical calculation, an approximation

by Taylor series expansion, and a Monte Carlo approximation. The latter has turned out to be most common. In this case, B resamples of the same size as that of the original sample are drawn with replacement from the empirical distribution of y , which can be seen as the Maximum-Likelihood (ML) estimator of the underlying probability distribution (cf. [Chao and Lo 1994](#), pp.391). Within each of the B bootstrap samples s_1, \dots, s_B , the estimator $\hat{\theta}_b$ of parameter θ is calculated in the same way as the statistic $\hat{\theta}$ for the sample s ($b = 1, \dots, B$). For large B , the distribution of $\hat{\theta}_b$ is interpreted as an estimate of the sample distribution of $\hat{\theta}$. Hence, the theoretical variance $V(\hat{\theta})$ of $\hat{\theta}$ is estimated by the Monte Carlo (MC) variance given by

$$V_{MC}(\hat{\theta}) = \frac{1}{B-1} \cdot \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2 \quad (1)$$

with

$$\bar{\theta} = \frac{1}{B} \cdot \sum_{b=1}^B \hat{\theta}_b. \quad (2)$$

The statistic $\bar{\theta}$ is the mean value of the estimators $\hat{\theta}_b$ in the B bootstrap samples. For approximately normal sampling distributions, this variance estimator can be used for the calculation of an approximate confidence interval. For large B and non-normal sampling distributions, a confidence interval can be calculated by applying the percentile method (cf. [Efron 1981](#), pp.317ff).

With increasing memory space of computers, an application of this method to finite population sampling became desirable. Such an effort has to consider complex sampling designs consisting of complex estimators and sampling schemes drawing the sample units without replacement, and arbitrary sample inclusion probabilities at various stages of the sampling process (cf., as an example, the discussion of the sampling design of the Austrian PISA survey in [Quatember and Bauer 2012](#)). For this purpose, different approaches are available in the relevant literature (cf., for instance, [Wolter 2007](#), p.200ff). One of them rescales the observations in the resamples drawn with replacement from the original sample in a way that the bootstrap variance (1) approximates the actual variance under a given sampling design (cf. [Rao and Wu 1988](#)). Another approach is to use the with-replacement bootstrap technique and adjust its bootstrap variance estimator to the parameter by an adequate choice of the size of the resamples (cf. [McCarthy and Snowden 1985](#)). [Sitter \(1992a\)](#) presented the “Mirror-Match Method”, in which subsamples of the original sample are drawn repeatedly according to the original sampling plan with a subsample size chosen to appropriately match the original variance of the estimator. [Antal and Tillé \(2011\)](#) discuss another approach, in which different with- and without-replacement resampling procedures are mixed in a way that the bootstrap variance estimator calculated from resamples of the same size as that of the original sample under this mixture of resampling schemes, equals the interesting variance.

Furthermore, the finite population bootstrap approach, which is considered a natural extension of the technique by [Efron \(1979\)](#) to finite population sampling, generates an artificial population, the “bootstrap population”, from the observed sample data. For this problem, the finite population U of N elements takes over the role of the unknown probability distribution in the i.i.d. bootstrap. The population elements are characterized by their values y_k of y and x_k of a possible auxiliary variable x ($k = 1, \dots, N$). [Gross \(1980\)](#) was the first to adapt the original bootstrap method to the specific case of a simple random sample without replacement (SI), but only with the restriction of integer design weights $\frac{N}{n} \in \mathbb{N}$. For this purpose, from an SI sample s , a set-valued finite population estimator U_G^* of size $N_G^* = N$ of the true population U of size N is generated by replicating each sample value y_k exactly $\frac{N}{n}$ times (cf. p.184) providing a variable” y^* denoting these “clones” of the sample values. Hence, the bootstrap population U_G^* can be interpreted as the finite population with the ML regarding the sample drawn (cf. [Chao and Lo 1994](#), p.396). For $\frac{N}{n} = \frac{2,400}{400} = 6$, for example, the bootstrap population U_G^* comprises six units of each sample value y_k resulting in a population of total size $N_G^* = 2,400$ ($k = 1, \dots, n$).

This entire process can be seen as the application of the idea behind the unbiased Horvitz-Thompson (H-T) estimator of the total t of y , that is

$$t_{H-T} = \sum_{k=1}^n y_k \cdot \frac{1}{\pi_k} \quad (3)$$

with first-order sample inclusion probabilities π_k , to the problem of generating an adequate bootstrap population. For SI sampling, Eq.(3) results in

$$t_{H-T} = \sum_{k=1}^n y_k \cdot \frac{N}{n}. \quad (4)$$

Obviously, this estimation strategy can be described by the generation of a so-called “pseudo-population” U_{H-T}^* , for which each sample value y_k is replicated exactly $\frac{N}{n}$ times (cf. [Quatember 2014](#), pp.20ff). Compared to the bootstrap population U_G^* from above with $\frac{N}{n} \in \mathbb{N}$, the pseudo-population U_{H-T}^* of the H-T process allows not only to contain $\lfloor \frac{N}{n} \rfloor$ whole units ($\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$) with the same value y_k of variable y but also an $(\frac{N}{n} - \lfloor \frac{N}{n} \rfloor)$ -piece of a unit with that value when $\frac{N}{n}$ is not an integer ($\forall k \in s$). For $\frac{N}{n} = \frac{2,600}{400} = 6.5$, for example, the H-T pseudo-population U_{H-T} comprises six whole and one half unit of each sample value y_k ($k = 1, \dots, n$).

After U_G^* is generated, B resamples of size n are drawn from U_G^* following the original sampling method. In other words, the resamples are no i.i.d. samples of size n from the original sample s . Instead, the resampling process from U_G^* follows a multivariate hypergeometric distribution with parameters N , n , and N times 1 (cf., for instance [Ranalli and Mecatti 2012](#)). Hence, each of the n sample values y_1, \dots, y_n has the same probability $\frac{1}{n}$ of being chosen as the first value in the resample of same size n . After the first draw, the same value already drawn at the first step has a probability of $\frac{\frac{N}{n}-1}{N-1} = \frac{N-n}{n(N-1)}$ for being chosen also as the second element of the resample. The other $n-1$ values of y in s , not selected as the first resample element, have a probability of $\frac{\frac{N}{n}}{N-1} = \frac{N}{n(N-1)}$ and so on. Generally, a value y_k observed in s has a probability

$$\frac{N - n \cdot h_{k,j-1}}{n \cdot (N - j + 1)} \quad (5)$$

of being selected at the j -th step of a resample selection from U_G^* ($j = 1, \dots, n$). In (5), $h_{k,j-1}$ denotes the number of times the value y_k was already selected in the first $j-1$ steps of the process to generate a resample ($h_{k,0} = 0 \forall k \in s$).

This shows that the bootstrap population U_G^* does not have to be generated in reality. The resample process from U_G^* might as well be carried out by applying the probability mechanism described above directly to the sample s . This was also discussed by [Ranalli and Mecatti \(2012\)](#) as a resource and time saving alternative to the physical generation of the bootstrap population. These resamples form the basis for the estimation of the sampling distribution of the estimator $\hat{\theta}$ (for example, the H-T estimator t_{H-T}) for parameter θ (for instance, the total t) in SI sampling. For this purpose, in each of the B resamples s_b , the estimator $\hat{\theta}_b$ has to be calculated in the same way as $\hat{\theta}$ was calculated in the original sample s ($b = 1, \dots, B$).

Considering, for instance, the estimation of parameter t of variable y by Formula (4), this means that within each SI resample s_b of size n , an estimate t_{H-T_b} is calculated using the replication variable y^* in U_G^* :

$$t_{H-T_b} = \sum_{k=1}^n y_k^* \cdot \frac{N}{n}.$$

The MC variance (1) of these B estimates serves as an estimator of the variance $V(t_{H-T})$ of t_{H-T} under the SI sampling scheme. This variance estimator is approximately unbiased in large samples (cf., for instance [Sitter 1992b](#), p.139).

Obviously, for general applicability in the practice of survey sampling, this idea has to be extended to

- i) non-integer design weights, and
- ii) general probability sampling, including stratification and clustering, ensuring a bootstrap population with the same structure as the original population (cf. [Chao and Lo 1994](#), pp.398ff).

The techniques to generate a bootstrap population U^* proposed in the relevant literature for different sampling schemes deviate more or less from the ML principle applied by [Efron \(1979\)](#) and [Gross \(1980\)](#) in the generation of the bootstrap population (cf., for instance [Bickel and Freedman 1984](#); [Kuk 1989](#); [Sitter 1992b](#); [Chao and Lo 1994](#); [Booth, Butler, and Hall 1994](#)).

General probability proportional to size random sampling without replacement (π PS), where the selection process is carried out, for instance, by systematic selection of elements ordered randomly in a list (cf. [Särndal, Swensson, and Wretman 1992](#), p.96f), is commonly used in large-scale surveys such as the PISA survey (cf. [OECD 2012](#), ch.4). This sampling method has a positive impact on the efficiency of the HT estimator of t , when the study variable y and the size variable x are approximately proportionally related. But, the estimation of the variance of the HT estimator may be hard. In particular, the calculation of the second-order inclusion probabilities, needed for the HT variance estimator, can be cumbersome or even impossible. [Holmberg \(1998\)](#) proposed a bootstrap approach to estimate the variance for general π PS sampling. The total of size variable x in U is denoted as t_x . Under the restriction $x_k \cdot n \leq t_x \forall k \in U$, the design weight $\frac{1}{\pi_k} = \frac{t_x}{x_k n}$ of survey unit k is decomposed into an integer part $\lfloor \frac{t_x}{x_k n} \rfloor$ and the “rest” $\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor$. To generate the pseudo-population U_H^* , the values y_k and x_k of each unit k are jointly replicated $\lfloor \frac{t_x}{x_k n} \rfloor$ times and, independently from each other, randomly once more with probability $\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor$. This process creates a bootstrap population U_H^* of size N_H^* with an expected value of $E(N_H^*) = N$. After U_H^* is generated, the sample inclusion probabilities π_k have to be recalculated according to the variable x^* consisting of the replicated sample values of x , before the resampling process can start. Then, a number of B resamples of size n are drawn from U_H^* according to π PS sampling. The estimation of the parameter under study is done in each of these bootstrap samples in the same way as it was done in the original one. For large N and n , the bootstrap variance estimator (1), for example, achieves approximate unbiasedness with respect to the variance of the H-T estimator of t (cf. [Holmberg 1998](#), p.381).

[Barbiero and Mecatti \(2010\)](#) aimed to simplify the procedure presented for π PS sampling by [Holmberg \(1998\)](#) and, at the same time, improve its efficiency with respect to the estimation of the variance of the H-T estimator of t . They propose to make “a more complete use of the auxiliary information” ([Barbiero and Mecatti 2010](#), p.62) available for size variable x , in particular of its total t_x . According to these authors, the following understandable properties should apply to a bootstrap algorithm with respect to the estimation of a total t of variable y (cf. [Barbiero and Mecatti 2010](#), pp.60ff):

1. Given the sample s , in a bootstrap population U^* , the total t_{x^*} of variable x^* should be equal to the total t_x of x in U .
2. The total t_{y^*} of variable y^* in U^* should be equal to the H-T estimator t_{H-T} of t calculated in the original sample s .
3. For given s , over all B resamples s_b , the H-T estimator of the total t_{y^*} of y^* in U^* should have an expectation equal to the H-T estimator of t in the original sample s .

Obviously, these properties are desirable for an efficient estimation of $V(t_{HT})$ by $V_{MC}(t_{HT})$. For different bootstrap methods in the literature dealing with the generation of bootstrap populations, these three properties hold only for $\frac{1}{\pi_k} \in \mathbb{N} \forall k \in s$. Hence, [Barbiero and Mecatti](#)

(2010) at least proposed an “ x -balanced π PS-bootstrap”, where after replicating each sample unit k a number of $\lfloor \frac{t_x}{x_k n} \rfloor$ times, further units are iteratively added to the bootstrap population U_{BM}^* from a list where these units are sorted in decreasing order of their $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$ -value or their ratio $\frac{1}{\pi_k} \cdot \frac{1}{\lfloor \frac{t_x}{x_k n} \rfloor + 1}$ until the minimum difference of t_{x^*} and t_x is achieved. Considering the ratios, for the same $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$ -values, elements with a higher integer part $\lfloor \frac{t_x}{x_k n} \rfloor$ of their design weight $\frac{t_x}{x_k n}$ have a higher probability of again being added to U_{BM}^* , as compared with elements with a lower integer part. After U_{BM}^* is generated, the probabilities π_k have also to be recalculated before the resampling process can start.

But, these proposals for non-integer design weights also result in bootstrap populations not guaranteeing a size $N_{BM}^* = N$ for SI sampling, when $\frac{1}{\pi_k} \notin \mathbb{N}$ (cf. [Ranalli and Mecatti 2012](#), p.4095). For $\lfloor \frac{t_x}{x_k n} \rfloor \in \mathbb{N}$ and SI sampling, the methods of [Holmberg \(1998\)](#) and [Barbiero and Mecatti \(2010\)](#) reduce to the original concept proposed by [Gross \(1980\)](#).

2. The proposed bootstrap method

All the methods described in the introductory section are more or less heuristic when it comes to the generation of a bootstrap population in the presence of non-integer design weights (cf. [Rao and Wu 1988](#), pp.237). They all try to establish a bootstrap population to start the resampling process from it, which includes solely integer numbers of replications of the original sample values and, as a consequence, also of the total number of units in the bootstrap population.

In the following, a procedure is proposed, which is a direct application of the idea behind the H-T estimator of a total as it was described below Eq.(4) to the bootstrap population problem. It complements the proposals of [Holmberg \(1998\)](#) and [Barbiero and Mecatti \(2010\)](#) for the problem of non-integer design weights. This Horvitz-Thompson based bootstrap approach (HTB) also allows non-integer numbers of replications of the sample values of y and x to generate the bootstrap population U_{HTB}^* . Let each unit k be replicated exactly $\frac{1}{\pi_k} = \frac{t_x}{x_k n}$ times. In this way, a bootstrap population U_{HTB}^* is generated which contains not only $\lfloor \frac{t_x}{x_k n} \rfloor$ whole units with values y_k and x_k but also an additional $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$ -piece of a unit with these values when $\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor > 0$ applies ($k \in s$). In this way, U_{HTB}^* has an expected size N_{HTB}^* of $E(N_{HTB}^*) = \sum_s \frac{1}{\pi_k} = N$. For SI sampling with $\pi_k = \frac{n}{N}$, this means that a bootstrap population with size $N_{HTB}^* = N$ is guaranteed. In the resampling process based on the bootstrap population U_{HTB}^* , a whole unit k belonging to this population has a resample inclusion probability proportional to its original x -value. But, for a $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$ -piece of a unit, this probability is proportional to $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$ times x . Hence, after the generation of U_{HTB}^* as a set-valued estimator of U , the design weights of the elements will not have to be recalculated.

From the point of view of the underlying probability mechanism, the value y_k of the original sample s ($k = 1, \dots, n$) has a probability of

$$\frac{t_x - n \cdot h_{k,j-1} \cdot x_k}{n \cdot (t_x - \sum_{s_{b_{j-1}}} x_i)} \quad (6)$$

for being selected into the b -th resample at the j -th step of the process of drawing n resampling units ($j = 1, \dots, n$) when $\frac{t_x}{x_k n} - h_{k,j-1} \cdot x_k > 0$ applies. Otherwise, for the j -th draw, its inclusion probability is set to zero. In Eq. (6), $h_{k,j-1}$ denotes the number of times y_k was already chosen within the first $j - 1$ steps of the selection of n units for resample s_b . Furthermore, $s_{b_{j-1}}$ denotes the subset of the resample s_b after the $(j - 1)$ -th draw. Applying this probability mechanism in the resampling process can replace the resource consuming physical generation of the bootstrap population U_{HTB}^* . For $x_k = 1 \forall k \in U$ and $\frac{N}{n} \in \mathbb{N}$, the method reduces to the strategy of the SI technique of [Gross \(1980\)](#) as discussed under Section 1.

In each of the resamples drawn, the original estimator $\hat{\theta}$ of parameter θ under study is calculated. In the case of the estimation of total t , for instance, the estimator can be the ordinary H - T estimator, ratio or regression estimator, or other calibrated estimators based on poststratification or iterative proportional fitting (cf., for instance [Alfons and Templ 2013](#), p.20f).

For the proposed HTB technique, regarding the three desired properties for efficient variance estimation, as mentioned in Section 1 (cf. [Barbiero and Mecatti 2010](#), pp.60ff), the following applies:

1. The total t_{x^*} of size variable x^* in U_{HTB}^* is given by: $t_{x^*} = \sum_{k=1}^n x_k \cdot \frac{1}{\pi_k} = t_x$.
2. For the total t_{y^*} of variable y^* in U_{HTB}^* , $t_{y^*} = \sum_{k=1}^n y_k \cdot \frac{1}{\pi_k} = t_{H-T}$ applies.
3. The expected value of the H-T estimator of the total t_{y^*} of y^* in U_{HTB}^* yields $E^*(\sum_{k=1}^n y_k^* \cdot \frac{1}{\pi_k}) = t_{y^*} = t_{H-T}$ with E^* denoting the expectation over all resamples, given s and the sampling design.

Clearly, for usual design weights, the proposed HTB method it is not expected to perform much better than, for instance, the technique of [Holmberg \(1998\)](#). Nevertheless, it is shown that the three desirable properties regarding estimation quality mentioned by [Barbiero and Mecatti \(2010\)](#) always hold for this method. The proposed method to generate the bootstrap population might as well seem more understandable in terms of educational reasons than the heuristic methods from the literature, because it follows the same idea as the one behind the widely used H-T estimator when it comes to the composition of the bootstrap population. Moreover, it is not necessary that the bootstrap population needed for the resampling process is physically generated, which often may be cumbersome. The resampling can be done directly from the original sample applying the probability mechanism behind the sampling scheme. Additionally, the HTB bootstrap can still be used in situations, where other methods fail because of first-order sample inclusion probabilities π_k of the population units which are close to one. For a π PS sample, this might provide resample inclusion probabilities that are outside the acceptable range (see Section 3).

3. Simulation study

A simulation study was undertaken to compare exemplarily the performance of the proposed H-T based technique (HTB) with the method (H) presented by [Holmberg \(1998\)](#). For this purpose, the Swedish MU 281 population as described in Appendix B of [Särndal *et al.* \(1992\)](#), formed the basis. This specific population was also used by [Holmberg \(1998\)](#) as basis for a simulation study. It consists of all but the largest three municipalities Stockholm, Gothenburg and Malmö.

Different sets of study variables y and auxiliary size variables x were used. For all the variables, in the simulation results, almost the same pattern appeared. Hence, as a typical example, the simulation results for study variable SS82 ($= y$), that is the number of social-democratic seats in municipal council, are presented. The total t of y in the population is 6,193. 10,000 simulations were conducted according to a π PS sampling scheme with size variable

- i) P75 ($\equiv x_1$, the number of inhabitants in the municipality in 1975),
- ii) $x_2 = 1 + \frac{x_1}{100}$, and
- iii) $x_3 = 1 \forall k \in U$.

As the values of x_1 widely differ, so do the first-order sample inclusion probabilities $\pi_k = \frac{x_k n}{t_x}$. For size variable x_2 , these probabilities are much closer. With size variable x_3 , the π PS method reduces to an SI scheme.

The parameter to be estimated is the variance of the H-T estimator t_{H-T} of t . For each simulation, the chosen sample sizes were $n_1 = 40$ and, if possible, $n_2 = 100$. The chosen number B of bootstrap resamples was $B = 300$.

The simulation results are reported in Tables 1 and 2 for the two simulated sample sizes, as also for all three π PS sampling schemes where possible. Furthermore, the results are shown exemplarily for one of these setups in the form of boxplots in Figures 1 and 2. For $n_2 = 100$, with y and x_1 , no π PS design could be carried out because $\frac{x_{1k}n}{t_{x_1}}$ was greater than one for at least one $k \in U$. Moreover, for each setup, the relative simulation bias (in percent)

$$rb_{sim}(V) = \frac{E_{sim}(V) - V_{sim}(t_{H-T})}{V_{sim}(t_{H-T})} \cdot 100 \quad (7)$$

was computed as an indicator of its performance. The terms $E_{sim}(V)$ in Eq.(7) and $sd_{sim}(V)$ in Tables 1 and 2 denote the simulation mean values and standard deviations, respectively, of the MC bootstrap variance estimates V according to Eq.(1) with respect to the bootstrap method HTB (V_{HTB}) or H (V_H) within the 10,000 simulations. Furthermore, the term $V_{sim}(t_{H-T})$ in Eq.(7) denotes the variance of the H-T estimates within the 10,000 simulations as the reference value because for π PS sampling, the true variance of the H-T estimator cannot be calculated exactly. This reference term is substituted by the known variance $V(t_{SI})$ in the simulations of the SI method. In the tables that follow, the simulated standard deviations $sd_{sim}(N^*)$ of the unbiased 10,000 respective bootstrap population sizes N^* are also presented for both bootstrap methods. Eventually, the percentage coverage rates of approximate confidence intervals in the simulations using the variance estimates can be found in the tables for both approaches.

Table 1: Simulation results for three different sampling designs (sample size $n_1 = 40$)

Sampling scheme	π PS (y, x_1)		π PS (y, x_2)		SI (y)	
	HTB	H	HTB	H	HTB	H
$rb_{sim}(V)$	+1.09	-1.18	+0.71	-2.40	-1.90	-2.18
$sd_{sim}(V)$	182,044	180,328	9,662	9,688	20,867	21,055
$sd_{sim}(N^*)$	36.34	36.28	6.53	7.06	0	1.00
coverage (in %)	92.61	92.34	94.23	94.23	93.72	93.64

Table 2: Simulation results for two different sampling designs (sample size $n_2 = 100$)

Sampling scheme	π PS (y, x_2)		SI (y)	
	HTB	H	HTB	H
$rb_{sim}(V)$	-2.80	-0.81	2.14	-0.68
$sd_{sim}(V)$	1,936	1,975	4,054	4,084
$sd_{sim}(N^*)$	3.28	4.99	0	3.90
coverage (in %)	94.98	94.70	94.54	94.30

As expected, no major improvement is found with respect to the performance of the variance estimator in comparison to the performance of the one presented by (Holmberg 1998). Nevertheless, in all the simulations, the HTB method strongly tends to perform slightly better with respect to relative bias and standard deviation. Figure 1 shows, as an example, the boxplots regarding the π PS sampling design with auxiliary variable x_2 and sample size $n = 100$.

A π PS sampling with auxiliary variable x_1 is defined only for $n \leq 49$ because for all elements k of the population U , $\frac{x_{1k}n}{t_{x_1}} \leq 1$ has to apply. Whereas the HTB method is applicable for all sample sizes $n \leq 49$, method H does not work for sample sizes close to the upper limit of 49, because the inclusion probabilities have to be recalculated for a given bootstrap population U_H^* . Depending on the drawn sample s , this process may yield resample inclusion probabilities

outside the admissible range. In the HTB method, for the resamples drawn from the bootstrap population, no recalculation of the original first-order inclusion probabilities π_k is required. Hence, it is sufficient that the original probabilities are within the admissible range.

Allowing not only integer numbers of clones of the original sample values in the bootstrap population has also an impact on the size N_{HTB}^* of the bootstrap population. While for both the methods, size N^* is unbiased for N , the standard deviation of N^* is smaller for the HTB method in almost all simulation results. The difference between these standard deviations increases with less differing original first-order inclusion probabilities. This is shown in both the tables as well as in Figure 2.

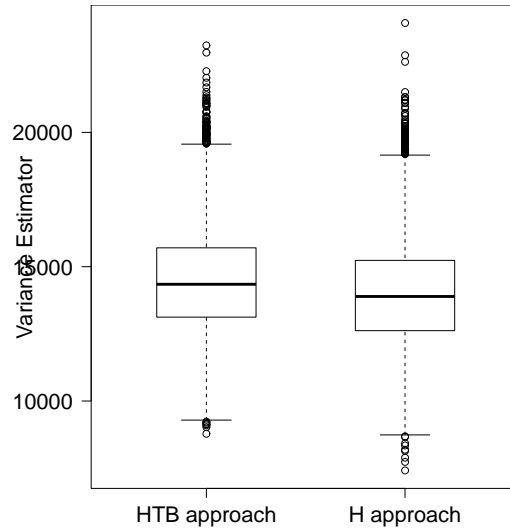


Figure 1: Boxplots of simulated variance estimates calculated according to the HTB and the H approaches for π PS sampling with size variable x_2 ($n_2 = 100$)

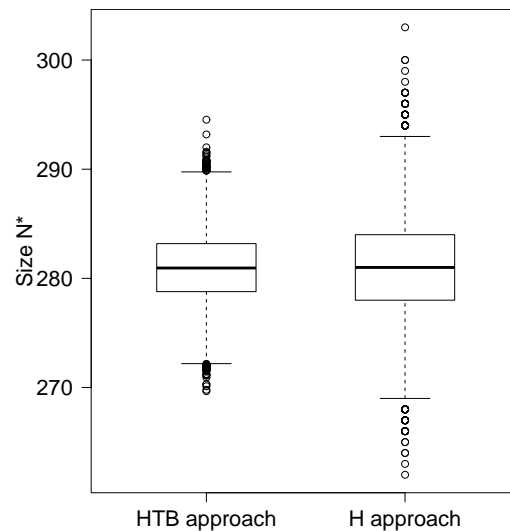


Figure 2: Boxplots of simulated bootstrap population sizes N^* calculated according to the HTB and H approaches for π PS sampling with size variable x_2 ($n_2 = 100$)

Eventually, in all the simulated cases, the coverage rates of the usual approximate confidence intervals, calculated by using the variance estimates of the HTB method are closer to the desired 95% level than the intervals calculated by using the variance estimates of the H method.

4. Conclusion

The H-T approach to the generation of the bootstrap population presented herein applies the idea behind the H-T estimator to the generation of a bootstrap population for finite populations. Overall, as expected, the simulation results indicate that the bootstrap estimator of the variance of a total based on this bootstrap population U_{H-T}^* is slightly more efficient than the one proposed by Holmberg (1998). For the proposed method, the three properties, considered desirable by Barbiero and Mecatti (2010) for efficient variance estimation, hold. This approach has an effect on the precision of the variance estimates. Applying this approach, the size of the bootstrap population, a variable unbiased for the true size of the original population, has a smaller standard deviation as compared to that of the approach by Holmberg. Furthermore, this method, unlike other methods in the literature, does not require the recalculation of the inclusion probabilities in general π PS sampling. This also means that the method proposed here can be applied even in situations where other methods fail.

In practice, the generation of the bootstrap population will not have to be processed physically. The whole resampling procedure can be carried out using the probability mechanism behind the process.

However, further studies including other populations than the one used here, and topics such as the optimum number B of resamples, or the estimation of other parameters than totals, are necessary to examine the suitability of this method in greater detail.

Acknowledgements

The author is grateful to the Associate Editor and two learned referees for their valuable comments and suggestions that led to improvement of the paper.

References

- Alfons A, Templ M (2013). “Estimation of Social Exclusion Indicators from Complex Surveys: The R Package *laeken*.” *Journal of Statistical Software*, **54**.
- Antal E, Tillé Y (2011). “A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population.” *Journal of the American Statistical Association*, **106**.
- Barbiero A, Mecatti F (2010). “Bootstrap algorithms for variance estimation in π PS sampling.” In P Mantovan, P Secchi (eds.), *Complex Data Modeling and Computationally Intensive Statistical Methods*, pp. 57–69. Springer, Milan.
- Bickel PJ, Freedman DA (1984). “Asymptotic Normality and the Bootstrap in Stratified Sampling.” *The Annals of Statistics*, **12**.
- Booth JG, Butler RW, Hall P (1994). “Bootstrap Methods for Finite Populations.” *Journal of the American Statistical Association*, **89**.
- Chao MT, Lo SH (1994). “Maximum Likelihood Summary and the Bootstrap Method in Structured Finite Populations.” *Statistica Sinica*, **4**, 389–406.
- Efron B (1979). “Bootstrap Methods: Another Look at the Jackknife.” *Annals of Statistics*, **7**, 1–26.
- Efron B (1981). “Censored Data and the Bootstrap.” *Journal of the American Statistical Association*, **76**.
- Gross S (1980). “Median Estimation in Sample Surveys.” *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 181–184.

- Holmberg A (1998). "A Bootstrap Approach to Probability Proportional-to-Size Sampling." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 378–383.
- Kuk AYC (1989). "Double Bootstrap Estimation of Variance under Systematic Sampling with Probability Proportional to Size." *Journal of Statistical Computation and Simulation*, **31**, 73–82.
- McCarthy PJ, Snowden CB (1985). "The Bootstrap and Finite Population Sampling." *National Center of Health Statistics: Data Evaluation and Methods Research*, **2**.
- OECD (ed.) (2012). *PISA 2009 Technical Report*. PISA, OECD Publishing. Available on [June 12, 2014]: <http://dx.doi.org/10.1787/9789264167872-en>.
- Quatember A (2014). *Datenqualit?t in Stichprobenerhebungen*. Springer Spektrum, Berlin.
- Quatember A, Bauer A (2012). "Genauigkeitsanalysen zu den ?sterreich-Ergebnissen der PISA-Studie 2009." In Eder, F (ed.), *PISA 2009 - Nationale Zusatzanalysen*, pp. 534–550. M?nster, Waxmann Verlag.
- Ranalli MG, Mecatti F (2012). "Comparing Recent Approaches for Bootstrapping Sample Survey Data: A First Step Towards A Unified Approach." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 4088–4099.
- Rao JNK, Wu CFJ (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association*, **83**.
- S?rndal CE, Swensson B, Wretman J (1992). *Model Assisted Survey Sampling*. Springer, New York.
- Sitter RR (1992a). "A Resampling Procedure for Complex Survey Data." *Journal of the American Statistical Association*, **87**.
- Sitter RR (1992b). "Comparing Three Bootstrap Methods for Survey Data." *The Canadian Journal of Statistics*, **20**.
- Wolter KM (2007). *Introduction to Variance Estimation*. Springer, New York.

Affiliation:

Andreas Quatember
Institute of Applied Statistics
Johannes Kepler University
A-4040 Linz, Austria
E-mail: andreas.quatember@jku.at
URL: <http://www.jku.at/ifas>