Sufficient bootstrapping

Article *in* Computational Statistics & Data Analysis · April 2011

DOI: 10.1016/j.csda.2010.10.010 · Source: DBLP

2 authors, including:

Sarjinder Singh
Texas A&M University - Kingsville
411 PUBLICATIONS **2,821** CITATIONS

SEE PROFILE

Computational Statistics and Data Analysis 55 (2011) 1629–1637

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sufficient bootstrapping

Sarjinder Singh *, Stephen A. Sedory

Department of Mathematics, Texas A&M University -Kingsville, Kingsville, TX 78363, USA

ARTICLE INFO

Article history:
Received 4 March 2010
Received in revised form 26 June 2010
Accepted 7 October 2010
Available online 20 October 2010

Keywords:
Bootstrapping
Sufficient bootstrapping
Estimation of mean
Resampling
Distinct units

ABSTRACT

In this paper, we introduce an idea we refer to as sufficient bootstrapping, which is based on retaining only distinct individual responses, and also develop a theoretical framework for the techniques. We demonstrate through numerical illustrations that the proposed sufficient bootstrapping may be better than the conventional bootstrapping in certain situations. The expected gain by the sufficient bootstrapping has been computed for small and large sample sizes. The relative efficiency shows that there could be significant gain by the sufficient bootstrapping and it could reduce computational burden. Variance expressions for both the conventional and sufficient bootstrapping sample means are derived. Here the word "sufficient" is being used in the sense that it is "sufficient to take just one of any duplicated items in the bootstrap sample" and is not tightly connected to sufficiency in terms of any likelihood perspective. R code for comparing bootstrapping and sufficient bootstrapping are provided. A huge scope of further studies is suggested.

1. Introduction

Bradley Efron is a statistician best known for proposing the bootstrap re-sampling technique, which has had a major impact in the field of statistics and virtually every area of statistical application. The bootstrap was one of the first computer-intensive statistical techniques, replacing traditional algebraic derivations with data-based computer simulation. He received numerous awards on this and related contributions in the field of statistics as cited in the free online encyclopedia entitled Wikipedia. It was named bootstrapping because it involves resampling from the original data set. The bootstrap is a form of a larger class of methods that resample from the original data set and is thus are also called resampling procedure. For details, one could also refer to the books on bootstrapping by Efron and Tibshirani (1993) and Chernick (1999).

Casella (2003) provides an introduction to the Silver Anniversary of the Bootstrap. Efron (2003) discusses a second thought on bootstrapping. Davison et al. (2003) have a critical review on recent developments in bootstrap methodology during the year 2003. Beran (2003), Lele (2003), Shao (2003), Lahiri (2003) and Politis (2003) explain the impact of bootstrap on statistical algorithms and theory, estimating functions, sample surveys, small area estimation and time series, respectively. Ernst and Hutson (2003) and Rueda et al. (1998, 2005, 2006) discussed the application of bootstrapping for quantile estimation. Holmes (2003) and Soltis and Soltis (2003) discuss applications of bootstrapping in phylogenetic trees and phylogeny reconstruction respectively. Holmes et al. (2003) provide an overview of a conversation on bootstrap between Bradley Efron and other good friends. Horowitz (2003) discussed the use of bootstrap in econometrics and Hall (2003) discussed a short prehistory of bootstrap. Johnson (2001) has given a nice introduction to bootstrapping. Likewise, Hesterberg (2008) taught a very valuable course entitled, "*Bootstrap methods and permutation tests*"during the conference of Statisticians at San Antonio, TX. Also a course taught by Kolenikov (2009) on "*Bootstrap for complex survey data*"at the Joint Statistical Meeting; Washington, DC has been found to be a good source for learning and updating.

* Corresponding author. Tel.: +1 320 308 5423.
E-mail address: sarjinder@yahoo.com (S. Singh).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.10.010

Table 2.1
Comparison between conventional and sufficient bootstrapping.

Conventional bootstrapping			Sufficient bootstrapping		
12	12	12	12		
12	12	15	12		15
12	12	21	12		21
12	15	12	12	15	
12	15	15	12	15	
12	15	21	12	15	21
12	21	12	12	21	
12	21	15	12	21	15
12	21	21	12	21	
15	12	12	15	12	
15	12	15	15	12	
15	12	21	15	12	21
15	15	12	15		12
15	15	15	15		
15	15	21	15		21
15	21	12	15	21	12
15	21	15	15	21	
15	21	21	15	21	
21	12	12	21	12	
21	12	15	21	12	15
21	12	21	21	12	
21	15	12	21	15	12
21	15	15	21	15	
21	15	21	21	15	
21	21	12	21		12
21	21	15	21		15
21	21	21	21		

Table 2.2
Conventional bootstrapping for mean.

Sample means, $\bar{y}_{\text{boot}}$	12	13	14	15	16	17	18	19	21	Sum
Frequency	1	3	3	4	6	3	3	3	1	27
$p^{(1)}_{\text{boot}}$	1/27	3/27	3/27	4/27	6/27	3/27	3/27	3/27	1/27	1

2. Idea of sufficient bootstrapping

We introduce the idea of "sufficient" bootstrapping which may help to reduce the computational burden, and may result in better inference for certain cases than the use of conventional bootstrapping due to Efron (1978). The conventional bootstrapping can be seen as a special case of simple random sampling with replacement (SRSWR) where the sample size n becomes equal to the population size N. For simplicity, consider a population consisting of $N = 3$ units, say $Y_1 = 12$, $Y_2 = 15$ and $Y_3 = 21$. One easily computes, the population mean $\overline{Y} = 16$, population standard deviation $\sigma_y = 3.74$ and population coefficient of variation $c_y = \left(\frac{\sigma_y}{\overline{Y}}\right) \times 100\% = 23.39\%$.

Consider the situation of selecting a sample of $n = 3$ units using SRSWR sampling from the given population of $N = 3$ units. Obviously, the total number of SRSWR samples will be $N^n = 3^3 = 27$. In Table 2.1, we list all 27 of the possible SRSWR samples. In the first sample, the first unit Y_1 is selected three times as $Y_1 = 12$, $Y_1 = 12$ and $Y_1 = 12$. In the conventional bootstrapping, we use the same information about unit Y_1 three times, whereas in the proposed sufficient bootstrapping, we keep only the distinct unit $Y_1 = 12$. In the second sample, the first unit $Y_1 = 12$ is selected twice and unit $Y_2 = 15$ is selected only once. Conventional bootstrapping considers $Y_1 = 12$, $Y_1 = 12$ and $Y_2 = 15$ as a bootstrapping sample, whereas in the proposed sufficient bootstrapping we keep only distinct units $Y_1 = 12$ and $Y_2 = 15$. For details about distinct units in SRSWR sampling refer to Raj and Khamis (1958). Now, we do the same thing for all 27 possible samples listed in Table 2.1 to construct sufficient bootstrapping samples from the conventional bootstrapping samples.

From the conventional bootstrapping, the frequency distribution tables for sample means, sample standard deviations and coefficient of variations (CV) are given, respectively, in Tables 2.2–2.4. From Table 2.2, the expected value of the conventional bootstrapping mean is given by

$$E(\bar{y}_{\text{boot}}) = \sum_{\text{boot}} p^{(1)}_{\text{boot}} \bar{y}_{\text{boot}} = 16 = \overline{Y}. \tag{2.1}$$

Thus, in this case the conventional bootstrapping sample mean is an unbiased estimator of the population mean.
From Table 2.3, the expected value of the conventional bootstrapping sample standard deviation is given by

$$E(s_{\text{boot}}) = \sum_{\text{boot}} p^{(2)}_{\text{boot}} s_{\text{boot}} = 3.33. \tag{2.2}$$

Table 2.3
Conventional bootstrapping for standard deviation.

Sample standard deviation, s_{boot}	0	1.73	3.46	4.58	5.20	Sum
Frequency	3	6	6	6	6	27
$p^{(2)}_{\text{boot}}$	3/27	6/27	6/27	6/27	6/27	1

Table 2.4
Conventional bootstrapping for coefficient of variation.

Sample CV, $\widehat{c}_{\text{boot}}$	0	12.37	13.32	18.23	20.38	28.64	28.87	34.64	Sum
Frequency	3	3	3	3	3	6	3	3	27
$p^{(3)}_{\text{boot}}$	3/27	3/27	3/27	3/27	3/27	6/27	3/27	3/27	1

Table 2.5
Sufficient bootstrapping for mean.

Sample means, $\bar{y}_{\text{sb}}$	12	13.5	15	16	16.5	18	21	Sum
Frequency	1	6	1	6	6	6	1	27
$p^{(4)}_{\text{sb}}$	1/27	6/27	1/27	6/27	6/27	6/27	1/27	1

Table 2.6
Sufficient bootstrapping for standard deviation.

Sample standard deviation, s_{sb}	0	2.12	4.24	4.58	6.36	Sum
Frequency	3	6	6	6	6	27
$p^{(5)}_{\text{sb}}$	3/27	6/27	6/27	6/27	6/27	1

The relative bias in the conventional bootstrapping sample standard deviation is given by

$$\text{RB}(s_{\text{boot}}) = \frac{E(s_{\text{boot}}) - \sigma_y}{\sigma_y} \times 100\% = -10.962\%. \tag{2.3}$$

Thus, the conventional bootstrapping underestimates the standard deviation.

From Table 2.4, the expected value of the conventional bootstrapping sample coefficient of variation is given by

$$E(\widehat{c}_{\text{boot}}) = \sum_{\text{boot}} p^{(3)}_{\text{boot}} \widehat{c}_{\text{boot}} = 20.57. \tag{2.4}$$

Thus, in this case the relative bias in the conventional bootstrapping sample coefficient of variation is given by

$$\text{RB}(\widehat{c}_{\text{boot}}) = \frac{E(\widehat{c}_{\text{boot}}) - c_y}{c_y} \times 100\% = -12.56\%. \tag{2.5}$$

In this example, conventional bootstrapping also underestimates the coefficient of variation.

Now, from the sufficient bootstrapping, the frequency distribution tables for sample means, sample standard deviations and coefficient of variations (CV) are, respectively, given in Tables 2.5–2.7.

The suffix "*sb*" stands for sufficient bootstrap.

From Table 2.5, the expected value of the sufficient bootstrapping mean is given by

$$E(\bar{y}_{\text{sb}}) = \sum_{\text{sb}} p^{(4)}_{\text{sb}} \bar{y}_{\text{sb}} = 16 = \overline{Y}. \tag{2.6}$$

Thus, in this case the sufficient bootstrapping sample mean is also an unbiased estimator of the population mean.

From Table 2.6, the expected value of the sufficient bootstrapping sample standard deviation is given by

$$E(s_{\text{sb}}) = \sum_{\text{sb}} p^{(5)}_{\text{sb}} s_{\text{sb}} = 3.84. \tag{2.7}$$

Thus, the relative bias in the sufficient bootstrapping sample standard deviation is given by

$$\text{RB}(\widehat{s}_{\text{sb}}) = \frac{E(s_{\text{sb}}) - \sigma_y}{\sigma_y} \times 100\% = 2.67\%. \tag{2.8}$$

From Eqs. (2.3) and (2.8), it is clear that the relative bias in the proposed sufficient bootstrapping estimator of standard deviation is much less than that based on conventional bootstrapping. Also note that sufficient bootstrapping overestimates the standard deviation, while conventional bootstrapping underestimates the standard deviation.

Table 2.7
Sufficient bootstrapping for coefficient of variation.

Sample CV, $\widehat{c}_{sb}$	0	15.71	23.57	28.64	38.57	Sum
Frequency	3	6	6	6	6	27
$p_{sb}^{(6)}$	3/27	6/27	6/27	6/27	6/27	1

From Table 2.7, the expected value of the sufficient bootstrapping sample coefficient of variation is given by

$$E(\widehat{c}_{sb}) = \sum_{sb} p_{sb}^{(6)} \widehat{c}_{sb} = 23.66. \tag{2.9}$$

Thus, the relative bias in the sufficient bootstrapping sample coefficient of variation is given by

$$\mathrm{RB}(\widehat{c}_{sb}) = \frac{E(\widehat{c}_{sb}) - c_y}{c_y} \times 100\% = 1.15\%. \tag{2.10}$$

Interestingly, for this particular example, the estimator of the coefficient of variation based on the proposed sufficient bootstrapping has less relative bias than that based on the conventional bootstrapping method.

From Tables 2.2 and 2.5, the distributions of sample means based on conventional and sufficient bootstrapping differ. From Tables 2.3 and 2.6, the distributions of sample standard deviations remain similar. Sufficient bootstrapping is unbiased while estimating the population mean, is less biased and overestimates the standard deviation and coefficient of variation in this particular example. In the same way, all the conventional bootstrapping methods available in Efron and Tibshirani (1993) and Chernick (1999) can be compared with the proposed sufficient bootstrapping method. It is expected that some of the methods will show improvements, as in this example the estimates of standard deviation and coefficient of variation show improvements from the relative bias point of views.

In the following sections, we also develop theoretical evidences which show that the proposed sufficient bootstrapping can perform better than the conventional bootstrapping.

3. Conventional bootstrapping

Let $s = (y_1, y_2, \ldots, y_n)$ be an original sample of length n. Let $s_b = (y_{b_i} : i = 1, 2, \ldots, n)$ be the bth bootstrapping sample such that $b = 1, 2, \ldots, n^n$. Obviously, the conventional bootstrapping mean based on the bth sample is given by:

$$\bar{y}_b = \frac{1}{n} \sum_{i \in s_b} y_{b_i}. \tag{3.1}$$

It is well known that the conventional bootstrapping sample mean in (3.1) is unbiased for the mean of the original sample and the variance of the conventional bootstrapping sample mean is given by

$$V(\bar{y}_b) = \frac{1}{n} s_y^2 \tag{3.2}$$

where

$$s_y^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \bar{y}_n)^2. \tag{3.3}$$

4. Proposed sufficient bootstrapping

Let $s_b^{(\nu)} = (y_{i(b)}^{(\nu)} : i = 1, 2, \ldots, \nu)$ be the bth bootstrapping sample, where $b = 1, 2, \ldots, n^n$ and which we assume has ν distinct units. Then, the sufficient bootstrap sample mean is given by:

$$\bar{y}_b^{(\nu)} = \frac{1}{\nu} \sum_{i \in s_\nu} y_{i(b)}^{(\nu)}. \tag{4.1}$$

Then, we have the following theorems:

Theorem 4.1. *The sufficient bootstrapping sample mean is unbiased for the mean of the original sample.*

Proof. Let E_2 denote the expected value for the given number of distinct units ν and E_1 denote the expected value over all bootstrapping samples, we have:

$$E(\bar{y}_b^{(\nu)}) = E_1 E_2 \left(\frac{1}{\nu} \sum_{i \in s_\nu} y_{i(b)}^{(\nu)} \right) = E_1 \left(\frac{1}{n} \sum_{i \in s} y_i \right) = \bar{y}_n$$

which proves the theorem. □

Theorem 4.2. *The variance of the sufficient bootstrapping sample mean is given by:*

$$V(\bar{y}_b^{(\nu)}) = \left[E_d\left(\frac{1}{\nu}\right) - \frac{1}{n}\right] s_y^2 \tag{4.2}$$

where E_d denotes the expected value over all possible distinct units.

Proof. Let V_2 denote the variance for the given number of distinct units ν and V_1 denote the variance over all possible bootstrapping samples, we have:

$$\begin{aligned} V(\bar{y}_b^{(\nu)}) &= E_1 V_2[\bar{y}_b^{(\nu)}|\nu] + V_1 E_2[\bar{y}_b^{(\nu)}|\nu] \\ &= E_1\left[\left(\frac{1}{\nu} - \frac{1}{n}\right) s_y^2\right] + V_1(\bar{y}_n) \\ &= \left(\frac{1}{\nu} - \frac{1}{n}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n}\right) s_y^2 \\ &= \left(\frac{1}{\nu} - \frac{1}{n}\right) s_y^2. \end{aligned} \tag{4.3}$$

The variance expression in (4.3) is the conditional variance for the given number of distinct units ν, thus taking expected value E_d, on both sides of (4.3), over all possible number of distinct units, we have the theorem. □

In the next section, we compare the proposed sufficient bootstrapping estimator $\bar{y}_b^{(\nu)}$ with the conventional bootstrapping estimator $\bar{y}_n$.

5. Comparison of conventional and sufficient bootstrapping

Following Feller (1957), we define an analog of the distribution of the number of distinct units, ν, in the case of sufficient bootstrapping as:

$$P(\nu = t) = \frac{1}{n^n}\binom{n}{t}\sum_{r=1}^{t}(-1)^t\binom{t}{r}(t-r)^n \tag{5.1}$$

where $t = 1, 2, \ldots, n$. Following Pathak (1961), we have

$$E_d\left(\frac{1}{\nu}\right) = \frac{1}{n^n}\sum_{l=1}^{n} l^{(n-1)}. \tag{5.2}$$

Thus, under the distribution (5.1), it is easy to verify that the variance of the sufficient bootstrapping estimator can be written as:

$$V(\bar{y}_b^{(\nu)}) = \left[\frac{1}{n^n}\sum_{l=1}^{n-1} l^{n-1}\right] s_y^2. \tag{5.3}$$

Thus, the percent relative efficiency of the sufficient bootstrapping estimator over the conventional bootstrapping estimator is given by:

$$\mathrm{RE} = \frac{V(\bar{y}_b)}{V(\bar{y}_b^{(\nu)})} \times 100\% = \left(\frac{n^{n-1}}{\sum_{l=1}^{n-1} l^{n-1}}\right) \times 100\%. \tag{5.4}$$

Table 5.1 provides the percent relative efficiency for different sample sizes.

Under the Feller (1957) distribution, as the sample size increases from 3 to 100, the percent relative efficiency increases from 540.00% to 18 923.44%, thus the use of sufficient bootstrapping may lead to more efficient results than the conventional bootstrapping.

Remark. We should be clear that in a sufficient bootstrap sample units never appear more than once, but some distinct units may have the same value, so it is possible that values appear more than once in a sample. That is, any values that are repeated in a sufficient bootstrap sample must have come from different units. For example, suppose there is a repeated value in the original sample, say $y_1 = 34$, $y_2 = 56$, $y_3 = 34$. If the conventional bootstrap sample consists of $y_1 = 34$, $y_2 = 56$, $y_3 = 34$, then the sufficient bootstrap sample also consists of $y_1 = 34$, $y_2 = 56$, $y_3 = 34$. If the conventional bootstrapping sample

Table 5.1
Percent Relative Efficiency (RE).

n	RE	n	RE	n	RE	n	RE
3	540.00	10	1741.24	45	7 754.70	80	13 768.60
4	711.11	15	2600.15	50	8 613.83	85	14 627.76
5	882.77	20	3459.18	55	9 472.95	90	15 486.89
6	1054.37	25	4318.25	60	10 332.09	95	16 346.03
7	1226.04	30	5177.35	65	11 191.22	100	17 205.17
8	1397.75	35	6036.46	70	12 050.35	105	18 064.31
9	1569.48	40	6895.58	75	12 909.49	110	18 923.44

consists of $y_1 = 34, y_1 = 34, y_1 = 34$, then the sufficient bootstrap consists of one distinct unit based on distinct index number $y_1 = 34$. If the conventional bootstrapping sample consists of $y_1 = 34, y_1 = 34, y_3 = 34$, the the sufficient bootstrap consists of two distinct units based on index numbers $y_1 = 34, y_3 = 34$. This is especially important when estimating a proportion with the proposed estimator where the outcome variable is a Bernoulli variate which only takes on the values 0 and 1.

In the next section, we perform a simulation study when the bootstrap samples are significantly reduced for different sample sizes.

6. Simulation study

In the simulation study, we consider two different situations: (a) Estimation of mean, variance, standard deviation and coefficient of variation of a quantitative variable, and (b) Estimation of a proportion of a qualitative variable.

6.0.1. Quantitative variable

We generate a random sample s of size n from the Beta distribution $B(1.2, 1.6)$ using R-function *rbeta*(n, *alpha*, *beta*). The true value of the parameters of interest of the population are given by:

$$\text{Population mean: } \overline{Y} = \frac{\alpha}{\alpha+\beta} = \theta_1 \text{ (say)} \tag{6.1}$$

$$\text{Population variance: } \sigma_y^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \theta_2 \text{ (say)} \tag{6.2}$$

$$\text{Population standard deviation: } \sigma_y = \sqrt{\sigma_y^2} = \theta_3 \text{ (say)} \tag{6.3}$$

$$\text{Coefficient of variation: } C_y = \frac{\sigma_y}{\overline{Y}} \times 100\% = \theta_4 \text{ (say).} \tag{6.4}$$

In a given sample s of n units, we assigned labels $i = 1, 2, \ldots, n$ to the observed y_i values in the same order as they are selected from the population. (Note this labeling is not necessary, but it helps while dealing with a Bernoulli variate as discussed in the next section). We used the R-function *runif* (n, 1, n) to select boots = 500 conventional bootstrapping samples each of size n units from the given sample s. Then based on these 500 boots, we computed the relative bias and mean squared errors of the estimators of θ_j, $j = 1, 2, 3, 4$, (defined in Eqs. (6.1)–(6.4)) as:

$$\text{RB}(\widehat{\theta}_j)_b = \frac{\frac{1}{\text{boots}}\sum_{b=1}^{\text{boots}} \widehat{\theta}_{j(\text{b})} - \theta_j}{\theta_j} \times 100\% \tag{6.5}$$

and

$$\text{MSE}(\widehat{\theta}_j)_b = \frac{1}{\text{boots}}\sum_{b=1}^{\text{boots}} (\widehat{\theta}_{j(\text{b})} - \theta_j)^2 \tag{6.6}$$

where $\widehat{\theta}_{j(\text{b})}$ is the conventional bootstrap estimator of θ_j. (Refer to R-code for details). The R-software contains the R-function, *unique*, which is very useful for generating sufficient bootstrapping samples. From each conventional bootstrapping sample we generated the corresponding sufficient bootstrapping sample by retaining the distinct units in the sample using the R-function *unique*. As before, based on 500 sufficient bootstrapping samples, the relative bias and the mean squared error of the corresponding estimators are computed as:

$$\text{RB}(\widehat{\theta}_j)_{\text{sb}} = \frac{\frac{1}{\text{boots}}\sum_{sb=1}^{\text{boots}} \widehat{\theta}_{j(\text{sb})} - \theta_j}{\theta_j} \times 100\% \tag{6.7}$$

Table 6.1
Percent relative bias and relative efficiency.

	$n = 10$	$n = 20$	$n = 30$	$n = 100$	$n = 200$
$\mathrm{RB}(\widehat{\theta}_{1(b)})$	−0.758074	−0.480335	0.350712	0.814050	0.813332
$\mathrm{RB}(\widehat{\theta}_{1(\mathrm{sb})})$	−0.777020	−0.503935	0.323479	0.780930	0.778240
$\mathrm{RE}(\widehat{\theta}_{1(b)}, \widehat{\theta}_{1(\mathrm{sb})})$	135.818	137.964	138.599	140.036	141.213
$\mathrm{RB}(\widehat{\theta}_{2(b)})$	−10.642579	−16.74531	−19.94702	−31.38589	−36.14765
$\mathrm{RB}(\widehat{\theta}_{2(\mathrm{sb})})$	0.518253	−0.368783	−0.121571	−0.406664	1.573797
$\mathrm{RE}(\widehat{\theta}_{2(b)}, \widehat{\theta}_{2(\mathrm{sb})})$	110.916	128.942	131.822	138.652	139.147
$\mathrm{RB}(\widehat{\theta}_{3(b)})$	−8.951192	−13.48752	−16.05619	−24.75360	−28.90311
$\mathrm{RB}(\widehat{\theta}_{3(\mathrm{sb})})$	−2.974640	−4.627741	−5.293538	−7.812855	−8.23082
$\mathrm{RE}(\widehat{\theta}_{3(b)}, \widehat{\theta}_{3(\mathrm{sb})})$	119.719	133.938	135.154	139.766	139.671
$\mathrm{RB}(\widehat{\theta}_{4(b)})$	−3.095421	−5.154346	−6.593556	−9.608855	−10.00058
$\mathrm{RB}(\widehat{\theta}_{4(\mathrm{sb})})$	1.926808	2.221759	2.304471	4.217639	6.779138
$\mathrm{RE}(\widehat{\theta}_{4(b)}, \widehat{\theta}_{4(\mathrm{sb})})$	130.958	138.404	138.930	139.028	140.703

and

$$\mathrm{MSE}(\widehat{\theta}_j)_{\mathrm{sb}} = \frac{1}{\mathrm{boots}} \sum_{sb=1}^{\mathrm{boots}} (\widehat{\theta}_{j(\mathrm{sb})} - \theta_j)^2 \tag{6.8}$$

where $\widehat{\theta}_{j(\mathrm{sb})}$ is the sufficient bootstrap estimator of θ_j. Then the percent relative efficiency (RE) of the sufficient bootstrapping estimator with respect to the conventional boot strapping estimator is computed as:

$$\mathrm{RE}(\widehat{\theta}_{j(\mathrm{b})}, \widehat{\theta}_{j(\mathrm{sb})}) = \frac{\mathrm{MSE}(\widehat{\theta}_j)_b}{\mathrm{MSE}(\widehat{\theta}_j)_{sb}} \times 100\%. \tag{6.9}$$

We generate 1000 populations each of size n from the same Beta distribution, $B(1.2, 1.6)$. The averages of the percent relative biases $\mathrm{RB}(\widehat{\theta}_j)_b$ and $\mathrm{RB}(\widehat{\theta}_j)_{\mathrm{sb}}$ and the percent relative efficiency $\mathrm{RE}(\widehat{\theta}_{j(\mathrm{b})}, \widehat{\theta}_{j(\mathrm{sb})})$ obtained over the 1000 populations are reported in Table 6.1. The R-code used in the simulation is given in the Appendix-A (see Appendix).

Discussion of results: For $n = 10$ and 20, the magnitude of percent relative bias in the estimator of mean, θ_1, remains slightly higher for sufficient bootstrapping than conventional bootstrapping, but the relative efficiency remains 135.818% to 137.964%. As soon as the sample size becomes 30 or above, the percent relative bias in sufficient bootstrapping becomes smaller than that in the case of conventional bootstrapping. The percent relative efficiency value goes on increasing from 138.599% to 141.213% as the sample size increases from 30 to 200. In Table 5.1, it is shown that the RE of the estimator of mean increases very fast if the number of distinct units follows the Feller (1957) distribution. Thus, more gain in relative efficiency is expected by the use of sufficient bootstrapping in situations where the distinct units will follow the Feller (1957) distribution. It is interesting to note that the percent relative bias in the estimator of variance θ_2 remains significantly lower in magnitude when using sufficient bootstrapping rather than conventional bootstrapping for all the sample sizes we considered. The percent relative efficiency of the estimator of variance θ_2 increases from 110.916% to 139.147% as the sample size increases from 10 to 200. The magnitude of the percent relative bias also remains smaller with sufficient bootstrapping when considering the problem of estimating the standard deviation θ_3. The percent relative efficiency increases from 119.719% to 139.671% as the sample size increases from 10 to 200. It is more interesting to note that conventional bootstrapping underestimates the coefficient of variation for all the sample sizes considered between 10 to 200, while sufficient bootstrapping gives overestimates. Further, note that the magnitude of the percent relative bias in the estimator of coefficient of variation θ_4 remains smaller than that in case of conventional bootstrapping. Also, the relative efficiency of sufficient bootstrapping, while estimating the coefficient of variation θ_4, increases from 130.958% to 140.703% as the sample size increases from 10 to 200. Thus, we conclude that the use of the proposed sufficient bootstrapping may, in many situations, lead to greater efficiencies than the conventional bootstrapping.

6.0.2. Qualitative variable

In this case we consider the problem of estimating a population proportion $p = \theta_5$ (say). For given values of p and n, we generated a sample s of n Bernoulli variate x using the R-function *rbinom*$(n, 1, p)$. Then, from the given sample s consisting of only 0 and 1 values, we selected boots = 500 bootstrapping samples. For each bootstrapping sample, we estimated the proportion p, or θ_5, as:

$$\widehat{p}_b = \frac{\text{Numbers of 1s in the given bootstrap sample}}{n} = \widehat{\theta}_{5(\mathrm{b})}. \tag{6.10}$$

Let ν_b be the number of distinct units in the bth bootstrapping sampling. Again we used the R-function *unique* to create a sufficient bootstrapping sample from the given bootstrapping sample. Then, we estimated the proportion p from the

Table 6.2
Percent relative bias and relative efficiency.

n	p	$\mathrm{RB}(\widehat{\theta}_5)_b$	$\mathrm{RB}(\widehat{\theta}_5)_{\mathrm{sb}}$	$\mathrm{RE}(\widehat{\theta}_{5(b)}, \widehat{\theta}_{5(\mathrm{sb})})$
30	0.1	0.507800	0.520609	137.099
	0.3	0.863711	0.898878	139.963
	0.5	−0.633599	−0.622563	138.290
	0.7	0.332019	0.352395	138.051
	0.9	0.023311	0.018557	137.123
60	0.1	−1.514633	−1.459951	139.392
	0.3	−1.010478	−0.986605	140.992
	0.5	−0.242599	−0.231660	139.998
	0.7	0.068433	0.075120	140.695
	0.9	−0.111774	−0.111521	138.271
150	0.1	0.308986	0.290529	140.171
	0.3	−0.247084	−0.224526	140.083
	0.5	0.025989	0.030562	140.517
	0.7	0.120009	0.126941	141.044
	0.9	0.276125	0.271552	140.478

sufficient bootstrapping sample as:

$$\widehat{p}_{sb} = \frac{\text{Numbers of 1s in the given sufficient bootstrap sample}}{v_b} = \widehat{\theta}_{5(\mathrm{sb})}. \tag{6.11}$$

Then based of these 500 bootstrapping samples, we used (6.5)–(6.8) to compute the percent relative biases and mean squared errors of the estimators of $\widehat{\theta}_{5(\mathrm{b})}$ and $\widehat{\theta}_{5(\mathrm{sb})}$ respectively. The percent relative efficiency of the estimator $\widehat{\theta}_{5(\mathrm{sb})}$ with respect to the estimator $\widehat{\theta}_{5(\mathrm{b})}$ was computed using (6.9).

We generated such samples of sizes n from 1000 populations from the same Bernoulli distribution $B(n, 1, p)$. The averages of the percent relative biases $\mathrm{RB}(\widehat{\theta}_5)_b$ and $\mathrm{RB}(\widehat{\theta}_5)_{\mathrm{sb}}$ and the percent relative efficiency $\mathrm{RE}(\widehat{\theta}_{5(\mathrm{b})}, \widehat{\theta}_{5(\mathrm{sb})})$ obtained over the 1000 such populations are reported in the Table 6.2. The R-code used in the simulation is given in the Appendix-B (see Appendix).

Discussion of results: While estimating population proportion, we also consider a smaller number of bootstrapping samples 500. As the value of p increases from 0.1 to 0.9 for different sample sizes between 30 to 150, the values of the percent relative biases are mixed for both the conventional and the proposed sufficient bootstrapping, but remain close to each other. The percent relative efficiency of sufficient bootstrapping remains between 137% to 139% for sample size $n = 30$. Also for sample size $n = 150$, the percent relative efficiency of sufficient bootstrapping remains 140% to 141% depending on the value of p. Thus, we conclude that sufficient bootstrapping may also perform better than conventional bootstrapping when considering the problem of estimating population proportion.

7. Further study

The idea of sufficient bootstrapping is easily extendable in all directions where bootstrapping has been applied. Obvious examples are: sufficient bootstrapping for the distribution of correlation coefficient, ratio estimator, regression estimator, product estimator, median, mode, quantiles and percentiles, etc. under different sampling schemes. It is not possible to tabulate or list all possible studies of sufficient bootstrapping. Further, researchers may investigate where the proposed sufficient bootstrapping method proves to be better than conventional bootstrapping.

Acknowledgements

The authors are thankful to the Editor Professor Dr. Stanley P. Azen, an Associate Editor and two learned referees for valuable comments on the original version of the paper. The authors are also thankful to, the "R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org ", for using R coding in the simulation study.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2010.10.010.

References

Beran, R., 2003. The impact of the bootstrap on statistical algorithms and theory. Statistical Science 18 (2), 175.
Casella, G., 2003. Introduction to the silver anniversary of the bootstrap. Statistical Science 18 (2), 133.
Chernick, M.R., 1999. Bootstrap Methods. A Practitioner's Guide. Wiley, NY.
Davison, A.C., Hinkley, D.V., Young, G.A., 2003. Recent developments in bootstrap methodology. Statistical Science 18 (2), 141.

Efron, B., 1978. Controversies in the foundations of statistics. American Mathematical Monthly 85, 231–246.
Efron, B., 2003. Second thoughts on the bootstrap. Statistical Science 18 (2), 135.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall.
Ernst, M.D., Hutson, A.D., 2003. Utilizing a quantile function approach to obtain exact bootstrap solutions. Statistical Science 18 (2), 231.
Feller, W., 1957. An Introduction to Probability Theory and its Applications, vol. 1. John Wiley and Sons, New York.
Hall, P., 2003. A short prehistory of the bootstrap. Statistical Science 18 (2), 158.
Hesterberg, T., 2008. Bootstrap methods and permutation tests. In: A Course at the Conference of Statisticians at San Antonio. Texas. During March 2008.
Holmes, S., 2003. Bootstrapping phylogenetic trees: Theory and methods. Statistical Science 18 (2), 241–255.
Holmes, S., Morris, C., Tibshirani, R., 2003. Bradley Efron: a conversation with good friends. Statistical Science 18 (2), 288.
Horowitz, J.L., 2003. The bootstrap in econometrics. Statistical Science 18 (2).
Johnson, R.W., 2001. An introduction to the bootstrap. Teaching statistics. Journal of the Royal Statistical Society, Series D 23 (2), 49–54.
Kolenikov, S., 2009. Bootstrap for complex survey data. In: A Course Taught at the Joint Statistical Meeting. Washington, DC.
Lahiri, P., 2003. On the impact of bootstrap in survey sampling and small-area estimation. Statistical Science 18 (2), 199.
Lele, S.R., 2003. Impact of bootstrap on the estimating functions. Statistical Science 18 (2), 185.
Pathak, P.K., 1961. On the evaluation of moments of distinct units in a sample. Sankhya, A 23, 415–420.
Politis, D.N., 2003. The impact of bootstrap methods on time series analysis. Statistical Science 18 (2), 219.
Raj, D., Khamis, S.H., 1958. Some remarks on sampling with replacement. Annals of Mathematical Statistics 29, 550–557.
Rueda, M., Arcos, A., Artes, E., 1998. Quantile interval estimation in a finite population using a multivariate ratio estimator. Metrika 47, 203–213.
Rueda, M., Martinez-Miranda, , Arcos, A., 2005, 2006. Bootstrap confidence intervals for finite population quantiles in the presence of auxiliary information. Model Assisted Statistics and Applications 1, 279–290.
Shao, J., 2003. Impact of the bootstrap on sample surveys. Statistical Science 18 (2), 191.
Soltis, P.S., Soltis, D.E., 2003. Applying the bootstrap in phylogeny reconstruction. Statistical Science 18 (2), 256–267.