Technical Report RM 705, Department of Statistics and Probability

Michigan State University

A REVIEW OF THE BOOTSTRAP METHOD FOR SAMPLING WITHOUT REPLACEMENT

P.K. Pathak and A. Alin

September 10, 2014

Abstract

The bootstrap method is a uniquely ubiquitous tool with great potential for statistical analyses when closed form solutions are unavailable. There is a great deal of literature on it, empirical as well as theoretical, when the underlying variables are independent and identically distributed (IID case). Although there have been adhoc efforts to extend the bootstrap in the nonIID case, there remains a pressing need to develop a parallel set of analogous results in the nonIID case. In this paper, we present a brief account of the bootstrap method in the specific context of sampling without replacement when the independence assumption is violated. We furnish a few illustrative numerical examples based on real data to demonstrate potentials and challenges of the bootstrap method when sampling from finite populations.

KEY WORDS: bootstrap, finite population, sampling without replacement

*A. Alin is Associate Professor, Department of Statistics, Dokuz Eylul University, Izmir, TURKEY and Visiting Associate Professor, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824; P.K. Pathak is Professor, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824. P.K. Pathak's work was supported in part by the NIH Grant RC102753.

1. Introduction

A time-honored problem in statistical inference is the estimation of population parameters. For example, sample proportion, mean, median, ratio of means, variance, correlation etc. are generally used as estimators of the corresponding population parameters. In order to evaluate performance characteristics of these estimators and to ascertain their margin of error, one needs to calculate their standard errors. For linear-type estimators, a closed form variance formula is easily derived. For nonlinear-type estimators, only large-sample asymptotic expressions are generally available and made use of. The bootstrap method of estimating the performance characteristics of nonlinear-type estimators in the IID case has received considerable attention during the recent past. Here we focus on the nonIID case in the context of sampling from a finite population. To do so, we first introduce an artifice of a super-population model, a technique commonly used in sampling from finite populations.

For simplicity in exposition, consider an infinite population (random variable) Ω defined on the real line. Let A_1 and A_2 respectively denote its mean and variance. Let $\mathbf{X_N}$ denote a simple random sample of size N from Ω. Without loss of generality, let $\mathbf{X_N}$=$(X_1, \ldots, X_N)$ in which the X_i denotes the ith X-variate value from Ω. Let $\hat{\theta}_N = \hat{\theta}(\mathbf{X_N})$ be a linear-type estimator of an unknown parameter θ. Then for a large class of estimators, such as the sample mean, variance etc.

$$V(\hat{\theta}_N)) \approx \frac{A_2}{N}$$

Now let $\mathbf{X}_n$ be a simple random sample without replacement (SRSWOR) of size n from $\mathbf{X_N} = (X_1, \ldots, X_N)$. Then $\mathbf{X}_n$ is also a simple random sample with replacement (SRSWR) of size n from Ω. Let $\hat{\theta}_n$ be the corresponding estimator of θ, based on $\mathbf{X_n}$. Then

$$V(\hat{\theta}_n)) \approx \frac{A_2}{n}$$

A remarkable property of a large class of linear-type estimators is the following Rao-Blackwellization identity:

$$E(\hat{\theta}_n|\mathbf{X_N}) = \hat{\theta}_\mathbf{N}$$

It can be shown that if θ is an estimable parameter, such a $\hat{\theta}_n$ always exists. In the sequel, we will refer to such estimators as the RB-estimator. The Rao-Blackwell theorem (1945) entails the following variance decomposition formula:

$$V(\hat{\theta}_n) = V(\hat{\theta}_N) + E(\hat{\theta}_n - \hat{\theta}_N)^2$$

Therefore

$$\begin{aligned} E[E\{(\hat{\theta}_n - \hat{\theta}_N)^2|\mathbf{X_N}\}] &= E(\hat{\theta}_n - \hat{\theta}_N)^2 = V(\hat{\theta}_n) - V(\hat{\theta}_N) \\ &\approx \frac{A2}{n} - \frac{A2}{N} = \left(1 - \frac{n}{N}\right)\frac{A_2}{n} \end{aligned} \tag{1.1}$$

It is worth noting that $E\{(\hat{\theta}_n - \hat{\theta}_N)^2|\mathbf{X_N}\}$ is the variance of the RB-estimator $\hat{\theta}_n$ under SRSWOR of size n from $\mathbf{X_N}$ while A_2/n represents its approximate variance under SRSWR. The above equation shows that in general the variance of an estimator under SRSWOR can be expected to be less than the corresponding estimator under SRSWR by a factor of $(1 - n/N)$.

When an estimator is not an RB-estimator, theoretical calculations of this kind become more complex. Bootstrap method pioneered by Efron (1979) in the IID case is a popular resampling technique which is used to estimate performance characteristics of any statistic for which theoretical derivation is complex. This method replaces rigorous mathematical calculations by computer calculations. In the IID case, the bootstrap has made great empirical and theoretical strides, e.g. see Singh (1981), Bickel and Freedman (1981), Shao and Tu (1995). Chernick (1999) and Hall (2003) provide an excellent overview of the history of the bootstrap. The bootstrap resampling methods are now an indispensable statistical tools

in current state of data analysis. Nonetheless much remains to be done in the nonIID case as compared to the IID case.

Efron's bootstrap assumes that the observed sample consists of independent and identically distributed variates. When sampling from a finite population, this assumption is satisfied when the underlying target population is either virtually infinite or when we have an SRSWR sample from a finite population. Efron's bootstrap method, if used in the nonIID case, is likely to furnish inaccurate estimates. There is a pressing need for further refinement of resampling methodology in the nonIID case. The primary focus of this article is to give a brief account of resampling methods in the context of SRSWOR from finite populations. Section 2 includes the bootstrap methodology for this case as well as the IID case. A few illustrative examples are presented in Section 3.

2. The Bootstrap Method

2.1 The Naive Bootstrap For IID Samples

The naive bootstrap can be thought of as a two-stage sampling design. At the first stage, a simple random sample $\mathbf{X_n}$ is drawn from the target population $\mathbf{X_N} = (X_1, X_2, ..., X_N)$, and at the second stage an SRSWR sample of size n, $\mathbf{X_n^*} = (X_1^*, \ldots, X_n^*)$ is drawn from $\mathbf{X_n}$, in which the X_i^* represents the X-variate values associated with the units in $\mathbf{X_n^*}$. The second stage is repeated a large number of times, say about $B = 1000$. The conditional variability of bootstrap estimates $\hat{\theta}_n^* = \hat{\theta}(\mathbf{X_n^*})$ given $\mathbf{X_n}$ is used as an estimate of the variability of $\hat{\theta}$. Listed below is a bootstrap algorithm for estimating the standard error of $\hat{\theta}$ for an SRSWR sample $\mathbf{X_n}$ from $\mathbf{X_N}$.

Let $\mathbf{X_n^{*b}} = (X_1^{*b}, \ldots, X_n^{*b})$ represent X-variate values of the units in the bth random sample drawn by SRSWR from the original sample $\mathbf{X_n}$, and $\hat{\theta}_n^{*b}$ represent the bootstrap parameter estimate calculated from this resample.

1. Select a random sample $\mathbf{X_n^*}$ from $\mathbf{X_n}$. Let $\mathbf{X_n^{*1}} = (X_1^{*1}, \ldots, X_n^{*1})$ be n X-variate values from this first bootstrap sample.

2. Calculate the bootstrap parameter estimate $\hat{\theta}^{*1}$ from $\mathbf{X_n^{*1}}$.

3. Repeat Steps 1 and 2, B times.

4. Estimate the standard error of $\hat{\theta}$ by the standard deviation of $\hat{\theta}^*$, using the following formula:

$$\hat{\sigma}^* = \sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\hat{\theta}}^*)^2}{B-1}} \text{ where } \bar{\hat{\theta}}^* = \frac{\sum_{b=1}^{B}\hat{\theta}^{*b}}{B} \tag{2.1}$$

Let us turn now to the case of sampling from finite populations. Suppose that $\mathbf{X_n}$ is an SRSWOR sample from $\mathbf{X_N}$. Because of sampling without replacement, successive sample units in $\mathbf{X_n}$ are negatively correlated. The correlation between any pairs of units equals $-V(X_1)/(N-1)$. Now consider applying the naive bootstrap method in this context. This entails simple random sampling with replacement, violating the dependency among the observations in the original sample. If we resample without replacement of size n, then the set of bootstrap sample units is identical to the corresponding set of the original first stage sample of units, evidentally a degenerate outcome. There is a need here to modify the naive bootstrap resampling scheme so that resampling from the observed sample reflects the key characteristics of SRSWOR from a finite population. At a minimum, the inclusion probability of each unit in the sample should be closest to (n/N) and the inclusion probability of each pair of units should be closest to $(n/N)(1-n/N)$.

One of the most common problems in survey sampling is the estimation of population mean μ. Sample mean $\bar{X}$ is a commonly used estimator with a closed form formula for its standard error for infinite as well as finite populations. Consider SRSWR, in this case for large n and sampling from a population with mean μ and standard deviation σ, the sample mean $\bar{X}_n$ is asymptotically normally distributed with mean μ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. For SRSWOR, under certain mild conditions [Erdös and Renyi (1959), Hajek

(1960), and Mitra and Pathak (1984)], the sample mean $\bar{X}_n$ of an SRSWOR sample of size n is asymptotically normally distributed with mean μ and standard error:

$$\sigma_{\bar{X}}^{WOR} = \sqrt{V(\bar{X}_n|(X_1,\ldots,X_N)} = \sqrt{\frac{1}{n}\left(1-\frac{n}{N}\right)S_N^2} \tag{2.2}$$

where $S_N^2 = N\sigma^2/(N-1)$ is the population variance with divisor $(N-1)$. The ratio n/N is called the sampling ratio/fraction. $\sigma_{\bar{X}_n}$ and $\sigma_{\bar{X}_n}^{WOR}$ are asymptotically equal when the sampling ratio is small and approaches zero. Consider the bootstrap estimate of sample mean $\bar{X}_n^*$ calculated from bootstrap samples drawn with replacement. Then the conditional distribution of $\bar{X}_n^*$ given $\mathbf{X_n}$ is asymptotically normal with mean $\bar{X}_n$ and standard deviation $\hat{\sigma}_{\bar{X}_n} = s_n/\sqrt{n}$ where $s_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is the sample variance, an unbiased estimator of finite population variance S_N^2 . As noted by Shao (2003), unless $n/N \to 0$, the empirical distribution of $\bar{X}_n^*$ does not provide valid asymptotic approximation to the distribution of $\bar{X}_n$ when $\mathbf{X_n^*}$ is SRSWR sample and the original sample $\mathbf{X_n}$ is SRSWOR sample. What one needs is a bootstrap sample for which the probability of inclusion of each unit and each pair of units from the original sample to be approximately n/N and $[n(n-1)]/[N(N-1)]$ respectively. This can be done in a number of ways. There are a number of resampling methods in the current literature specifically proposed for finite populations, see for example, Gross (1980), Bickel and Freedman (1984), Chao and Lo (1985), McCarthy and Snowden (1985), Sitter (1992) and Boot et al.(1994). As an illustrative example, listed below is an algorithm for the method proposed by Gross (1980) and Chao and Lo (1985). For simplicity, we first assume that the population size N is a multiple of sample size n, i.e. $N = kn$ where k is an integer.

1. Create a virtual population of size N by replicating each $X_1,\ldots,X_n$ k times.

2. Draw a simple random sample of size n without replacement from this virtual population.

3. Calculate the bootstrap parameter estimate $\hat{\theta}^{*1}$ for this sample.

4. Repeat steps 2 and 3, B times.

5. Estimate the standard error of $\hat{\theta}$ by the standard deviation of $\hat{\theta}^*$, using Eq (2.1).

Chao and Lo (1985) state that if $\bar{X}_n^*$ is the bootstrap sample mean based on the above approach, the conditional distribution of $\bar{X}_n^*$ given $\mathbf{X_n}$, is asymptotically normal with mean $\bar{X}_n$ and standard deviation

$$\sigma_{\bar{X}_n}^{WOR*} = \sqrt{\frac{1}{n}(1-\frac{n}{N})\frac{N(n-1)}{(N-1)n}s_n^2} \tag{2.3}$$

Note that $\sigma_{\bar{X}_n}^{WOR*}$ is asymptotically the same as $\sigma_{\bar{X}_n}^{WOR}$ (Shao, 2003). Thus, the empirical distribution of $\bar{X}_n^*$ provides an asymptotically valid approximation to the distribution of $\bar{X}_n$. The approach proposed by McCarthy and Snowden (1985) is, on the other hand, based on selecting a sample of size $n^* = (n-1)/(1-n/N)$ from $\mathbf{X_n}$ with replacement so that the variance of $\bar{X}_n^*$ will be approximately the same as that of $\sigma_{\bar{X}_n}^{WOR}$. They studied properties of their method for other statistics as well, including the ratio estimator.

For simplicity we earlier assumed N to be an integer multiple of n. Consider now the the case when the population size is not an integer multiple of n. There are a number of ways this can be handled. Here we propose a minor modification to the Chao-Lo approach (1985). If N is not a multiple of n, use the integer part $k = [N/n]$, without rounding up and replicate as in the original approach. For the remaining observations in the virtual population of size N, draw an SRSWOR sample of size $N - nk$ from $\mathbf{X_n}$ so that we have of N virtual observations in all. We study two approaches for this method: 1) adding the same set of observations of size $N - nk$ to each bootstrap sample $\mathbf{X_n^*}$, or 2) for each bootstrap sample, adding a new set of observations of size $N - nk$ drawn by SRSWOR.

The following section includes numerical comparison of methods proposed by Chao and Lo (1985), McCarthy and Snowden (1985), naive bootstrap and our modified approach for the case where N is not an integer multiple of n. Moreover, we have also estimated the standard error of $\bar{X}_n$ by multiplying the naive bootstrap estimator of standard error by

$\sqrt{(1-n/N)}$. This is based on Eq.(1.1) in which A_2/n replaced by the naive bootstrap variance estimator. As illustrative examples, we consider the estimation of sample mean for which a closed form variance formula is known, as well as the estimattion of median and correlation coefficient for which closed form variance formulae are unavailable.

3. Numerical Results

We use the LSAT-GPA and Score data sets from Efron (1993). The population size for the LSAT-GPA data is 82. Efron used the Score data as a sample. But for this study, we assume that it is a finite population of size of 88. Only Algebra (ALG) and Statistics (STA) variables from the Score data set are used for numerical calculations. For each of these populations we have drawn SRSWOR sample of size 15. Table 1 is for parameters and Table 2 - Table 6 include summary statistics for bootstrap samples. The bootstrap summary results are obtained from $10,000$ simulations and $2,000$ bootstrap resamples, i.e. for each population, $10,000$ different samples were drawn, and for each of these samples $2,000$ bootstrap resamples were drawn. For the standard errors of median and correlation coefficient estimators, there are no closed form formulae. Hence, we ran separate $1,000,000$ simulations to estimate the true standard error in each case.

Concerning bias, all methods show similar performances. Relative error column is obtained as (Estimated standard error-True standard error)/True standard error. Naive bootstrap overestimates the true standard error all the time. The original Chao-Lo method underestimates the true standard error for the sample mean, while the McCarthy-Snowden approach comes closest to the true standard error followed by our second modified approach to the Chao-Lo method. On the other hand, for median and correlation coefficient, the Chao-Lo method and our first modified approach outperform the McCarthy-Snowden method. The standard error estimate that we get by multiplying the naive bootstrap variance estimator by finite sample correction factor gives us a smaller relative error compared to the Chao-Lo

approach. The simulations also seem to suggest that the general approach based on the Rao-Blacwell theorem and the naive bootstrap furnishes satisfactory results for simple random sampling without replacement.

4. Concluding Remarks

The bootstrap is an immensely popular computer-intensive resampling technique to evaluate the performace characteristics of statistical methods. Its theoretical and empirical justifications are largely based on the assumption that the observed data consists of IID observations. On the other hand, most sample surveys in practice are based on samples selected without replacement. Thus a naive application of the bootstrap resampling techniques in sampling from finite populations raises interesting issues about its validity. In this article we briefly touched upon the need for refinements of the bootstrap in the context of sample surveys. In the case of simple random sampling without replacement, we discussed how the naive bootstrap can be modified to make it applicable in the context of simple random sampling without replacement. There is a great potential for the bootstrap technology in sample surveys and much remains to be done.

Table 1: Summary statistics for LSAT, GPA, Algebra, and Stat data (Efron, 1993)

	GPA	LSAT	ALG	STA
μ	3.133714	597.548800	50.602270	42.306820
M	3.140239	597.500000	50.000000	40.000000
σ	0.186727	38.48814	10.62478	17.255590
$\sigma_{\bar{X}}^{WOR}$	0.043580	8.982801	2.498587	4.057927
σ_{m}^{WOR}	0.054622	11.102005	2.539547	4.219431
ρ	0.759998		0.664736	
$\sigma_{\hat{\rho}}^{WOR}$	0.118439		0.152822	

Table 2: Bootstrap statistics for GPA data (Efron, 1993)

	$\bar{X}$			m		
	$E(\bar{X}_n)$	$\sigma^*_{\bar{X}_n}$	Rel. Err	$E(m)$	σ^*_m	Rel. Err.
Chao and Lo(1985)	3.135418	0.042019	-0.035	3.140692	0.057873	0.059
Proposed method 1	3.135415	0.042417	-0.027	3.140655	0.058512	0.071
Proposed method 2	3.135407	0.042650	-0.021	3.140681	0.058817	0.077
Naive Bootstrap	3.135423	0.046660	0.071	3.140667	0.064598	0.182
McCarthy and Snowden (1985)	3.135432	0.043828	0.005	3.140701	0.060982	0.116
$\sqrt{(1-n/N)}\sigma^*_{\bar{X}_n(naive)}$		0.042177	-0.032			

Table 3: Bootstrap statistics for LSAT data (Efron, 1993)

	$\bar{X}$			m		
	$E(\bar{X}_n)$	$\sigma^*_{\bar{X}_n}$	Rel. Err	$E(m)$	σ^*_m	Rel. Err.
Chao and Lo(1985)	597.545263	8.501012	-0.054	597.350201	11.300989	0.018
Proposed method 1	597.529890	8.584461	-0.044	597.327145	11.419015	0.029
Proposed method 2	597.544045	8.628486	-0.039	597.343727	11.470277	0.033
Naive Bootstrap	597.546463	9.440499	0.051	597.318733	12.560376	0.131
McCarthy and Snowden (1985)	597.543479	8.868013	-0.012	597.338412	11.883407	0.070
$\sqrt{(1-n/N)}\sigma^*_{\bar{X}_n(naive)}$		8.533465	-0.050			

Table 4: Bootstrap statistics for ALG data (Efron, 1993)

	$\bar{X}$			m		
	$E(\bar{X}_n)$	$\sigma^*_{\bar{X}_n}$	Rel. Err	$E(m)$	σ^*_m	Rel. Err.
Chao and Lo(1985)	50.579120	2.338481	-0.064	50.639679	2.781551	0.095
Proposed method 1	50.580308	2.378387	-0.048	50.642924	2.836133	0.117
Proposed method 2	50.580839	2.383586	-0.046	50.645280	2.841580	0.119
Naive Bootstrap	50.581198	2.597346	0.040	50.656265	3.127341	0.231
McCarthy and Snowden(1985)	50.580501	2.514615	0.006	50.656005	2.894557	0.140
$\sqrt{(1-n/N)}\sigma^*_{\bar{X}_n(naive)}$		2.365647	-0.053			

Table 5: Bootstrap statistics for STA data (Efron, 1993)

	$\bar{X}$			m		
	$E(\bar{X}_n)$	$\sigma^*_{\bar{X}_n}$	Rel. Err	$E(m)$	σ^*_m	Rel. Err.
Chao and Lo(1985)	42.250435	3.817832	-0.059	40.425025	4.764754	0.129
Proposed method 1	42.256931	3.883556	-0.043	40.440078	4.865164	0.153
Proposed method 2	42.249008	3.891133	-0.041	40.435715	4.874323	0.155
Naive Bootstrap	42.249924	4.240786	0.045	40.494368	5.411338	0.282
McCarthy and Snowden(1985)	42.251100	4.106820	0.012	40.496215	4.996672	0.184
$\sqrt{(1-n/N)}\sigma^*_{\bar{X}_{n(naive)}}$		3.862482	-0.053			

Table 6: Bootstrap statistics for correlations (Efron, 1993)

	LSAT-GPA			ALG-STA		
	$E(\bar{X}_n)$	$\sigma^*_{\bar{X}_n}$	Rel. Err	$E(m)$	σ^*_m	Rel. Err.
Chao and Lo(1985)	0.737771	0.112980	-0.046	0.637011	0.147596	-0.034
Proposed method 1	0.737542	0.114214	-0.036	0.636542	0.150250	-0.017
Proposed method 2	0.737576	0.114820	-0.031	0.636524	0.150519	-0.015
Naive Bootstrap	0.735830	0.126337	0.067	0.634132	0.164498	0.076
McCarthy and Snowden (1985)	0.737427	0.115922	-0.021	0.635330	0.157298	0.029

References

Bickel, P.J., Freedman, D.A., 1984. Asymptotic normality and the bootstrap in stratified random sampling. *The Annals of Statistics,* 12, 470-482.

Chao, M.T., Lo, S.H., 1985. A bootstrap method for finite populations. *Sankhyā, Ser. A,* 47, 399-405.

Chernick, M.R., 2008. *Bootstrap Methods: A guide for practitioners and researchers, 2nd Edition.* Wiley, Hoboken, New Jersey.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics,* 7, 1-26.

Erdös, P., Rényi, A., 1959. On the central limit theorem for samples from a finite population. *Publications of Institute of Mathematics, Hungarian Academy of Sciences A*4, 49–61.

Gross, S., 1980. Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 181-184.

Hájek, J., 1960. Limiting distributions in simple random sampling from a finite population. *Publications of Institute of Mathematics, Hungarian Academy of Sciences A*5, 361–374.

Hall, P., 2003. A short prehistory of the bootstrap. *Statistical Science,* 18, 158–167.

McCarthy, P.J., Snowden, C.B., 1985. The bootstrap and finite population sampling. *in Vital Health Statistics* (Ser., 2, No. 95), Public Health Service Publication, Washington, DC: U.S. Goevernment Printing Office, 85-1369.

Mitra, S.K., Pathak, P.K., 1984. The nature of simple random sampling *Ann. Statist.*, 12, 1536–1542.

Rao, C.R., 1945. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Cal. Math. Soc.,* 37, 81–91 (Republished in S. Kotz & N. Johnson [eds.], *Breakthroughs in Statistics:* 1889–1990, vol. 1).

Singh, K., 1981. On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics,* 9, 1187–1195.

Sitter, R.R., 1992. A resampling procedure for complex survey data. *Journal of the American Statistical Association,* 87, 755-764.

Shao, J., Tu, D., 1995. *The Jacknife and Bootstrap.* Springer, New York.