THE BAYESIAN BOOTSTRAP AND MULTIPLE IMPUTATION FOR UNEQUAL PROBABILITY SAMPLE DESIGNS

Michael P. Cohen, National Center for Education Statistics
555 New Jersey Avenue NW, Washington DC 20208-5654

Key Words: Missing data, Nonresponse, Probability proportional to size, Proper imputation, Variance estimation

Abstract: Efron's bootstrap, Rubin's Bayesian bootstrap, the finite population bootstrap of Gross, and the finite population Bayesian bootstrap of Lo are described. The finite population Bayesian bootstrap is generalized to account for sampling with unequal probabilities. The connection between the Bayesian versions of the bootstrap and multiple imputation is discussed.

1. Introduction

Multiple imputation is a method for handling missing data designed to capture the extra uncertainty due to the missing data elements. One of the main practical messages about multiple-imputation procedures is (Rubin, 1987, p. 126):

> Draw imputations following the Bayesian paradigm as repetitions from a Bayesian posterior distribution of the missing values under the chosen models for nonresponse and data, or an approximation to this posterior distribution that incorporates appropriate between-imputation variability.

The drawing of sample elements (not necessarily imputations) from a Bayesian posterior distribution is exactly what the *Bayesian bootstrap* (Rubin, 1981) accomplishes. Like the "regular" bootstrap of Efron (1979, 1982), the Bayesian bootstrap is relatively simple yet the bootstrap sampling is computationally intensive.

This paper explores the Bayesian and other forms of the bootstrap with multiple imputation in mind. Much of the discussion will also be pertinent to those who are interested in the bootstrap for other purposes such as estimating the sampling variance. Our ultimate aim is to develop a form of the Bayesian bootstrap that is appropriate for unequal probability sample designs in finite populations. This paper contains the initial results of this investigation.

For book-length treatments of the bootstrap and related methods, we recommend Efron and Tibshirani (1993) and Shao and Tu (1995). Chapter 6 of the latter reference is on applications to sample surveys.

The organization of this paper is as follows: This introduction is Section 1. We discuss the original form of the bootstrap in Section 2. In Section 3 the simplest version of the Bayesian bootstrap is introduced. Section 4 introduces the bootstrap for finite populations (both Bayesian and non-Bayesian). Section 5 treats the Bayesian bootstrap for unequal probability sample designs. Section 6 relates the Bayesian versions of the bootstrap to multiple imputation. Concluding remarks are provided in Section 7.

2. Efron's Bootstrap

Let us begin by considering a simple random sample with replacement of sample size n. Let $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ be the observed values. The most standard form of the bootstrap, due to Efron (1979), involves selecting *bootstrap samples* $\mathrm{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ by randomly sampling n times with replacement from the original sample $x_1, x_2, \ldots, x_n$. The idea is to select a large number B of these bootstrap samples $\mathrm{x}^{*1}, \mathrm{x}^{*2}, \ldots, \mathrm{x}^{*B}$. If $\hat{\theta}(\mathrm{x})$ is an estimate based on the original sample x, then $\hat{\theta}(\mathrm{x}^{*b})$ is the corresponding estimate based on the b^{th} bootstrap sample. The estimate $\hat{\theta}(\mathrm{x})$ could be, for instance, the mean or the median. By studying how the $\hat{\theta}(\mathrm{x}^{*1}), \hat{\theta}(\mathrm{x}^{*2}), \ldots, \hat{\theta}(\mathrm{x}^{*B})$ are distributed, we learn about the distribution of $\hat{\theta}(\mathrm{x})$. In particular, Efron (1982, p. 28) suggests that the standard error of $\hat{\theta}(\mathrm{x})$ be estimated by

$$\widehat{\mathrm{SE}}[\hat{\theta}(\mathrm{x})] = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} [\hat{\theta}^{*b}(\mathrm{x}) - \hat{\theta}^{*\cdot}(\mathrm{x})]^2 \right\}^{\frac{1}{2}}$$

where $\hat{\theta}^{*\cdot}(\mathrm{x}) = \sum_{b=1}^{B} \hat{\theta}^{*b}(\mathrm{x})/B$.

The early justification for the bootstrap was based on heuristic reasoning. Work by Bickel and Freedman (1981), Singh (1981), and others, though, demonstrates that the bootstrap has remarkable large sample (asymptotic) properties. Confidence intervals may be constructed based on the bootstrap that outperform those based on the normal approximation.

3. The Bayesian Bootstrap

Rubin (1981) introduced a variation on the bootstrap with a Bayesian justification. Like

Efron's, this *Bayesian bootstrap* was first developed for simple random sampling with replacement (with sample size n). Each Bayesian bootstrap sample $\mathbf{x}^{*b}$ is selected by a two-step procedure, as follows (adapted from Rubin, 1987, p. 124):

Step 1. Draw n uniform random numbers between 0 and 1, and let their ordered values be $a_1, a_2, \ldots, a_n$; also let $a_0 = 0$ and $a_n = 1$.

Step 2. Draw each of the n values in $\mathbf{x}^{*b} = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$ by drawing from $x_1, x_2, \ldots, x_n$ with probabilities $(a_1 - a_0), (a_2 - a_1), \ldots, (1 - a_{n-1})$; that is, independently n times, draw a uniform number u, and select x_i if $a_{i-1} < u \leq a_i$.

Although a bit more involved than the Efron bootstrap, this procedure is easily computerized.

To demonstrate the Bayesian nature of the procedure, we suppose that the data vector $\mathbf{x}$ can assume at most K distinct values. This restriction can be eliminated by nonparametric Bayesian arguments, but given that K can be arbitrarily large, we see no need to do so here. Let $\mathbf{d} = (d_1, d_2, \ldots, d_K)$ be the vector of these distinct values. Define the vector of probabilities $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ by

$$\Pr(x_i = d_k|\boldsymbol{\lambda}) = \lambda_k, \quad \sum \lambda_k = 1.$$

The $x_1, x_2, \ldots, x_n$ given $\boldsymbol{\lambda}$ are assumed to be independent and identically distributed. Rubin (1981) showed that the Bayesian bootstrap procedure is equivalent to assuming that the prior distribution of $\boldsymbol{\lambda}$ is the (improper) distribution

$$\Pr(\boldsymbol{\lambda}) = \prod_{k=1}^{K} \lambda_k^{-1} \quad \text{if} \sum \lambda_k = 1 \text{ and 0 otherwise.}$$

The posterior distribution of $\boldsymbol{\lambda}$, that is, the conditional distribution of $\boldsymbol{\lambda}$ given the data $\mathbf{x}$, is described by

$$\Pr(\boldsymbol{\lambda}|\mathbf{x}) \propto \prod_{k=1}^{K} \lambda_k^{n_k - 1}$$

where the n_k are the number of x_i, $i = 1, 2, \ldots, n$, equal to d_k; $\sum n_k = n$; and $\sum \lambda_k = 1$. This posterior distribution can be recognized as a $(K-1)$-dimensional Dirichlet distribution.

Lo (1987) showed that the Bayesian bootstrap has the same desirable large sample properties as Efron's bootstrap.

4. Finite Population Bootstraps

The first finite population bootstrap (FPB) was suggested by Gross (1980). To describe it, let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be a sample from a finite population $(Y_1, Y_2, \ldots, Y_N)$, $n \leq N - 1$. The sample is assumed to be a simple random sample, either with or without replacement. We have switched from using x to y to describe the sample in accord with survey sampling notational conventions. The key to the Gross FPB method is to first create an FPB population of size N from which the FPB samples are drawn. We shall discuss here only the simple case where the population size is an integer multiple of the sample size; that is, $N = kn$ for some integer k. In this case, the FPB population is created by replicating the sample $(y_1, y_2, \ldots, y_n)$ k times. Each FPB sample is produced by simple random sampling without replacement from the FPB population to obtain $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$. There has been much recent research on extending the FPB to more complex sample designs. Consult Chapter 6 of Shao and Tu (1995) for further information.

The corresponding finite population Bayesian bootstrap (FPBB) was developed by Lo (1988). It is based on a form of sampling that is a generalization of something called a "Pólya urn scheme." Consider an urn containing a finite number of balls. Select a ball from the urn at random; it is then replaced and another ball just like it is added to the urn. Continue this process until a fixed number, say m, of balls have been selected. Such a sample is called a Pólya sample of size m. An urn containing $z_1, z_2, \ldots, z_n$ will be denoted by $\mathsf{urn}\{z_1, z_2, \ldots, z_n\}$.

Each replication of the FPBB is formed as follows (adapted from Lo, 1988, p. 1686):

Step 1. Draw a Pólya sample of size $N - n$, denoted by $y_1^*, y_2^*, \ldots, y_{N-n}^*$, from the $\mathsf{urn}\{y_1, y_2, \ldots, y_n\}$.

Step 2. Form the FPBB population $y_1, y_2, \ldots, y_n, y_1^*, y_2^*, \ldots, y_{N-n}^*$.

Unlike other methods studied thus far, Lo's FPBB in effect resamples the population outside of the sample, rather than resampling the sample itself.

5. Unequal Probability Bayesian Bootstrap

In survey sampling, it is commonly the case that units are selected with unequal probabilities. Let π_i denote the probability that unit i is selected into the sample and set $w_i = 1/\pi_i$. Then w_i is called the *weight* of unit i and can be thought of as the number of units in the population that unit i represents. The procedure below is proposed to extend Lo's FPBB to this unequal probability of selection situation.

Step 1. Draw a sample of size $N-n$, denoted by $y_1^*, y_2^*, \ldots, y_{N-n}^*$, as follows:
Determine y_k^* by drawing from $y_1, y_2, \ldots, y_n$ in such a way that y_i is selected with probability

$$\frac{w_i - 1 + \ell_{i,k-1}\frac{N-n}{n}}{N-n+(k-1)\frac{N-n}{n}}$$

where $\ell_{i,k-1}$ = number of bootstrap selections of y_i among $y_1^*, y_2^*, \ldots, y_{k-1}^*$. Set $\ell_{i,0}=0$ and note that $\sum_{i=1}^{n} \ell_{i,k-1} = k-1$.

Step 2. Form the FPBB population
$y_1, y_2, \ldots, y_n, y_1^*, y_2^*, \ldots, y_{N-n}^*$.

The properties of this procedure are under investigation. Use will be made of the ideas of Lo (1988, 1993a, 1993b) and Walker and Muliere (forthcoming).

6. Multiple Imputation

Let us now discuss multiple imputation and how it is related to the Bayesian forms of the bootstrap. In doing so, we need to transform our notation somewhat. The goal of imputation is to produce, to the extent possible, a *complete sample* by producing imputed values for the cases that did not respond (or, for some other reason, the data are missing). The goal of multiple imputation is to do this in a way that does not cause the error to be underestimated. Lo's FPBB is described in terms of bootstrapping to the entire population, but for the application to imputation the "population" we want to bootstrap to is just the complete sample. The "sample" is the set of respondents; let's say there are r of them. There are $m = n - r$ cases for which data are missing — corresponding to the $N-n$ cases outside the sample in Lo's FPBB. This situation can be represented schematically, where the notation on the left is that used in Lo's FPBB and that on the right is used in the application to imputation:

$$\begin{aligned} N &\Longrightarrow n \\ n &\Longrightarrow r \\ N-n &\Longrightarrow m = n-r. \end{aligned}$$

We shall treat r and m as fixed.

One source of randomness in a bootstrap sample comes from being sampled and responding or not responding (viewed here as a single process). Let $\mathcal{I}_j$ denote an indicator variable that is 1 if unit j was sampled *and* responded and is 0 otherwise. We assume that the probability of not responding is the same for all cases in the sample (in reality, this condition could at most be expected to hold within an imputation cell). Let $\mathcal{I}$ denote the vector of $\mathcal{I}_j$ values, $j = 1, \ldots, n$.

Another source of randomness comes from the bootstrap process itself. Let $c_j^{*b} = \#\left\{y_i^{*b} = y_j\right\}$ be the number of times the response of respondent j is used in bootstrap replicate b (including the one time it is used to represent itself). Let $\mathbf{c}^{*b}$ denote the vector of c_j^{*b} values, $j = 1, \ldots, n$. The randomness in a bootstrap replicate sample can now be represented by the two vectors $\mathcal{I}$ and $\mathbf{c}^{*b}$. The vector $\mathbf{c}^*$ will denote the particular $\mathbf{c}^{*b}$ vector used for the actual imputations.

Using the properties of conditional expectation and variance, we can represent the variance of an estimator $\hat\theta = \hat\theta(\mathcal{I}, \mathbf{c}^*)$ by

$$\begin{aligned} &\operatorname{var}\hat\theta(\mathcal{I}, \mathbf{c}^*) \\ &\quad = \operatorname{var}_{\mathcal{I}} \mathrm{E}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right] + \mathrm{E}_{\mathcal{I}} \operatorname{var}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right]. \end{aligned}$$

The first term in this decomposition is essentially a sampling variance term, whereas the second term is essentially an imputation variance term. But $\operatorname{var}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right]$ in the second term can be estimated by

$$\frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat\theta(\mathcal{I}, \mathbf{c}^{*b}) - \hat\theta(\mathcal{I}, \cdot)\right\}^2$$

where $\hat\theta(\mathcal{I}, \cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat\theta(\mathcal{I}, \mathbf{c}^{*b})$, so this estimates

$\mathrm{E}_{\mathcal{I}} \operatorname{var}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right]$ unbiasedly as well.

On the other hand, $\mathrm{E}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right]$ in the first term can be estimated by $\hat\theta(\mathcal{I}, \cdot)$. If $\hat\theta$ is linear, though, $\mathrm{E}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right]$ reduces to $\hat\theta(\mathcal{I}, \bar{\mathbf{c}})$ where $\bar{\mathbf{c}} = \mathrm{E}_*\left(\mathbf{c}^*|\mathcal{I}\right)$. By Taylor series arguments, $\hat\theta(\mathcal{I}, \bar{\mathbf{c}})$ should also work as an approximation for nonlinear but sufficiently "smooth" $\hat\theta$. Whichever estimator is used, one has to estimate its sampling variance by one of the usual techniques to get an estimate of

$$\operatorname{var}_{\mathcal{I}} \mathrm{E}_* \left[\hat\theta(\mathcal{I}, \mathbf{c}^*)|\mathcal{I}\right].$$

By combining the various pieces, one can now estimate $\operatorname{var}\hat\theta(\mathcal{I}, \mathbf{c}^*)$.

The idea of decomposing the variance into two terms and using multiple imputations to estimate the "imputation" term originated with Rubin (see Rubin, 1987). The variance decomposition given here is different in that the first term treats the sample size as r rather than n.

7. Concluding Remarks

The Bayesian bootstrap provides a simple mechanism for sampling from a posterior distribution. In cases studied thus far, it has the same desirable large sample properties as the "regular" bootstrap and is not difficult to computerize. Particularly with multiple imputation in mind, the challenge is to develop the Bayesian bootstrap for the kinds of complex sample designs that arise in practice and especially for unequal probability sampling. Multiple "hot-deck" imputations can then be drawn by Bayesian bootstrap selections of respondents.

In this paper we have very briefly summarized some of the different versions of the bootstrap with initial results on developing a Bayesian bootstrap for unequal probability sample designs for application to multiple imputation. A substantial amount of work remains before this idea is fully developed.

Acknowledgment: The author thanks Lynn Kuo of the University of Connecticut for helpful comments and references.

REFERENCES

Bickel, P. J., and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Annals of Statistics* **9** 1196–1217.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7** 1–26.

——(1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Gross, S. (1980). Median estimation in sample surveys, presented at the 1980 Joint Statistical Meetings.

Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap, *Annals of Statistics* **15** 360–375.

——(1988). A Bayesian bootstrap for a finite population, *Annals of Statistics* **16** 1684–1695.

——(1993a). A Bayesian bootstrap for censored data, *Annals of Statistics* **21** 100–123.

——(1993b). A Bayesian method for weighted sampling, *Annals of Statistics* **21** 2138–2148.

Rubin, D. (1981). The Bayesian bootstrap, *Annals of Statistics* **9** 130–134.

——(1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Shao, J., and Tu, D. (1995). *The Jackknife and the Bootstrap.* New York: Springer.

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

Walker, S., and Muliere, P. (forthcoming). Beta-Stacy processes and a generalization of the Polya-urn scheme, *Annals of Statistics.*