Finite Population Unequal Probability Bayesian Bootstraps and Multiple Imputation

Mike Cohen

American Institutes for Research

University of Connecticut, Storrs, 12 May 2018
Conference on Bayesian Modeling, Computation, and Applications
in Honor of Lynn Kuo

Outline

- Efron's Bootstrap
- Rubin's Bayesian Bootstrap
- Gross's Finite Population Bootstrap
- Lo's Finite Population Bayesian Bootstrap
- Unequal Probability Bayesian Bootstrap
- Connections to Multiple Imputation
- Recent Work

Efron's Bootstrap, Efron (1979, 1982), Singh (1981), Bickel and Freedman (1981)

- Observed values $\mathbf{x} = (x_1, x_2, \ldots, x_n)$
- Bootstrap values $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ sampled n times with replacement from $\mathbf{x}$.
- Do B bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$.
- $\hat{\theta}(\mathbf{x}), \quad \hat{\theta}(\mathbf{x}^{*b})$
- $\hat{\theta}(\mathbf{x}^{**}) = \sum_{b=1}^{B} \hat{\theta}(\mathbf{x}^{*b})/B$
- $\hat{\mathrm{SE}}[\hat{\theta}(\mathbf{x})] = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} [\hat{\theta}(\mathbf{x}^{*b}) - \hat{\theta}(\mathbf{x}^{**})]^2 \right\}^{\frac{1}{2}}$

Rubin's Bayesian Bootstrap (BB), Rubin (1979, 1982), Lo (1987)

- *Step 1.* Draw $n-1$ uniform[0, 1] r.v.'s. Let their ordered values be $a_1, a_2, \ldots, a_{n-1}$. Let $a_0 = 0;\ a_n = 1$.
- *Step 2.* Draw each of the n values in $\mathbf{x}^{*b} = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$ independently from $x_1, x_2, \ldots, x_n$ with probabilities $(a_1 - a_0), (a_2 - a_1), \ldots, (a_n - a_{n-1})$.

Rubin's BB: Why Bayesian?

- Vector of probabilities $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$.
  x_i i.i.d., $\Pr(x_i = d_k) = \lambda_k$, $\sum \lambda_k = 1$.
- Rubin(1981) showed BB like assuming (improper) prior

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

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

  where $n_k = \#\{x_i = d_k\}$.
- Lo (1987) showed BB has same desirable large sample properties as Efron's bootstrap.

Finite Population Bootstrap (FPB) of Gross (1980)

- Finite population bootstrap (FPB), Gross (1980)
- Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be sample from population

  $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N),\ n \leq N - 1$
- Simple random sampling, either with or without replacement
- Assume for simplicity $N = kn$, integer k.
- Create FBP population $\mathbf{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_N^*)$ with k copies of sample.
- Each FPB sample $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$ is a simple random sample without repacement from $\mathbf{Y}^*$.
- See Chapter 6 of Shao and Tu (1995) for extensions.

Finite Population Bayesian Bootstrap (FPBB) of Lo (1988)

"Pólya Urn Scheme"

- An urn contains a finite number of balls.
- Select ball from urn at random
- Ball is replaced and another ball just like it is also added to urn.
- Continue until a fixed number, say m, of balls is selected.
- An urn containing $z_1, z_2, \ldots, z_n$ will be denoted by **urn**$\{z_1, z_2, \ldots, z_n\}$.

Finite Population Bayesian Bootstrap (FPBB) of Lo (1988), continued

Calculation of a FPBB Replicate

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

Lo's FPBB resamples the population outside the sample rather than resampling the sample itself.

Unequal Probability Bayesian Bootstrap (UPBB)

Unequal Probability Sampling

- In survey sampling, it is common to select units with unequal probabilities. For example, if x_i is a measure of size (say, number of employees or total revenue) of business establishment i, we might select i into the sample with a probability proportional to x_i.
- Let π_i be the probability that unit i is selected into the sample.
- The (base) weight is $w_i = 1/\pi_i$.
- If x_i is the number of employees at establishment i and establishment i is selected into the sample, then w_i can be thought of as the number of employees in the *population* that estabishment i represents. ($\sum_{\mathcal{S}} w_i x_i$ estimates number of employees in population.)

Unequal Probability Bayesian Bootstrap (UPBB), cont.

Calculation of a UPBB Replicate

Each replication of UPBB is formed in two steps (Cohen, 1997):

- *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 $y_1, y_2, \ldots, y_n$ 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 UPBB population $y_1, y_2, \ldots, y_n, y_1^*, y_2^*, \ldots, y_{N-n}^*$.

UPBB: Why is it Bayesian?

- Vector of probabilities $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$.
  $\Pr(y_i = d_k) = \lambda_k, \quad \sum \lambda_k = 1.$
- From Dong, Elliott, and Raghunathan (2014), assume (improper) prior

  $\Pr(\boldsymbol{\lambda}) = \prod_{k=1}^{K} \lambda_k^{-1} \quad$ if $\sum \lambda_k = 1$ and 0 otherwise.
- Let $z_k = \sum_{j=1}^{n} (w_j - 1) \times I(y_j = d_k)$ and $n_k = \#\{y_i = d_k\}$.
- Likelihood:

  $\Pr(y_1, \ldots, y_n | \boldsymbol{\lambda}) \propto \prod_{k=1}^{K} \lambda_k^{z_k}.$
- Posterior:
  $\Pr(y_1^*, y_2^*, \ldots, y_{N-n}^* | y_1, y_2, \ldots, y_n) \propto \prod_{k=1}^{K} \frac{\Gamma(z_k + n_k)}{\Gamma(z_k)}$

Lo's FPBB and Multiple Imputation (MI)

- Goal of imputation is to produce a *complete sample*.
- Multiple imputation repeats the imputation process to assess variability.
- Correspondence between Lo's FPBB and multiple imputation:

Lo's FPBB	Multiple Imputation
population size N	sample size n
sample size n	number of respondents r
size of nonsample $N-n$	number of missing $m=n-r$

Multiple Imputation (MI)

- Let $\mathcal{I}_j$ denote the indicator that is 1 if unit j was sampled ***and*** responded, 0 otherwise. Let $\boldsymbol{\mathcal{I}}$ be the vector of $\mathcal{I}_j$ values, $j = 1, \ldots, n$.
- Let $c_j^{*b} = \#\{y_i^{*b} = y_j\}$ be the number of times respondent j is used in bootstrap replicate b ($c_j^{*b} \geq 1$). Let $\mathbf{c}^*$ be the vector of c_j^{*b} values for a specific b.
- Then

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

Dong, Elliott, and Raghunathan (2014)

- Gives a nonparametric method to generate synthetic populations accounting for complex sampling (not MI).
- Uses UPBB.
- Notes that draws from "weighted" Pólya urn can be produced using function wtpolyap in R package *polypost*.

Zhou, Elliott, and Raghunathan (2016)

- Treats MI.
- Uses UPBB.
- Considers two-stage designs (e.g., sampling schools, then students within the school).
- Meeden (1999) considered equal-probability two-stage designs, showed step-wise Bayes.
- Zhou et al. treats MI in unequal probability two-stage setting.
- See also Zhou's 2014 University of Michigan Ph. D. dissertation.

References I

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, **9**, 1196–1217.

Cohen, M. P. (1997), "The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs," *1997 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 635–638.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014), "A Nonparametric Method to Generate Synthetic Populations to Adjust for Complex Sampling Design Features," *Survey Methodology*, **40**, 1, 29–46.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, **7**, 1–26.

References II

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

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.

Lo, A. Y. (1988), "A Bayesian Bootstrap for a Finite Population," *Annals of Statistics*, **16**, 1684–1695.

Meeden, G. (1999), "A Noninformative Bayesian Approach for Two-Stage Cluster Sampling," *Sankhyā: The Indian Journal of Statistics, Series B*, **61**, 1, 133–144.

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

References III

Rubin, D. (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.

Zhou, H. (2014), "Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap," University of Michigan Ph. D. Dissertation, available at: https://deepblue.lib.umich.edu/handle/2027.42/108807

References IV

Zhou, H., Elliott, M. R., and Raghunathan T. E. (2016), "Multiple Imputation in Two-Stage Cluster Samples Using the Weighted Finite Population Bayesian Bootstrap," *Journal of Survey Statistics and Methodology*, **4**, 139–170.

Thank You!