Better Bootstrap Confidence Intervals

by Bradley Efron

Gregory Imholte

University of Washington, Department of Statistics

May 3, 2012

We wish to make a confidence interval for some parameter $\theta \equiv T(F)$ (e.g. $\theta = E_F X$), based on data

$$X_i \stackrel{\text{i.i.d.}}{\sim} F \in \mathcal{F}.$$

- Exact intervals
- Normal approximation ($\hat{\theta} \pm 1.96 \times se(\hat{\theta})$)
- Approximations based on series expansions
- Bootstrap confidence intervals

Exact interals and series expansions are hard, solutions differ from problem to problem. Normal approximation can be poor.

For simplicity, we start by assuming that $\hat{\theta} \sim f_\theta$ estimates $\theta \in \Theta$. We make the following assumptions regarding our estimator. For our family $\{f_\theta : \theta \in \Theta\}$, there exists a monotone increasing transformation h and constants z_0 and a such that

$$\hat{\phi} = h(\hat{\theta}) \qquad \phi = h(\theta)$$

satisfy

$$\hat{\phi} = \phi + \sigma_\phi(Z - z_0) \qquad Z \sim N(0,1)$$

with

$$\sigma_\phi = 1 + a\phi$$

- The constant z_0 is the bias correction constant
- The constant a is the acceleration constant

Let $\hat{G}(s) = P_{\hat{\theta}}\{\hat{\theta}^* < s\}$ denote the bootstrap distribution function. We can either estimate this via Monte Carlo (i.e., resample data x^* with replacement and compute $\hat{\theta}^*$ over and over), or use $\hat{\theta}^* \sim f_{\hat{\theta}}$.

Lemma

Under the conditions in the previous slide the correct central confidence interval of level 1 - 2α for θ is

$$\left[\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1-\alpha]))\right]$$

where

$$z[\alpha] = z_0 + \frac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}$$

Sketch of proof:

$$P\left(z^{(\alpha)} - z_0 \leq \frac{\hat{\phi} - \phi}{1 + a\phi} \leq z^{(1-\alpha)} - z_0\right) = 1 - 2\alpha \qquad (1)$$

Apply the monotone decreasing *involution* $r(x) = \frac{\hat{\phi} - x}{1 + ax}$ to the inside to get

$$P\left(\frac{\hat{\phi} + z^{(\alpha)} + z_0}{1 - a(z^{(\alpha)} + z_0)} \leq \phi \leq \frac{\hat{\phi} + z^{(1-\alpha)} + z_0}{1 - a(z^{(1-\alpha)} + z_0)}\right) = 1 - 2\alpha \qquad (2)$$

So we know the endpoints of an exact central interval on the ϕ scale.

The transformation h is monotone increasing, so the bootstrap cdf of $\hat{\phi}$ is $P_{\hat{\phi}}(\hat{\phi}^* < h(s)) = \hat{H}(h(s)) = \hat{G}(s) = P_{\hat{\theta}}(\hat{\theta}^* < s)$, and

$$\hat{H}(x) = \Phi\left(\frac{x - \hat{\phi}}{\sigma_{\hat{\phi}}} + z_0\right)$$
$$\hat{H}^{-1}(\alpha) = \left[\Phi^{-1}(\alpha) - z_0\right]\sigma_{\hat{\phi}} + \hat{\phi}$$

and it turns out that

$$\hat{H}^{-1}(\Phi(z[\alpha])) = \frac{\hat{\phi} + z^{(\alpha)} + z_0}{1 - a(z^{(\alpha)} + z_0)}.$$

This gives us that

$$\left[\hat{H}^{-1}(\Phi(z[\alpha])), \hat{H}^{-1}(\Phi(z[1-\alpha]))\right]$$

matches the interval on the previous slide.

Finally, we note the relationship between bootstrap cdfs:

$$\alpha = \hat{H}(h(s))$$
$$h^{-1}\left(\hat{H}^{-1}(\alpha)\right) = s = \hat{G}^{-1}(\alpha)$$

This implies that the following events are equivalent:

$$h(\theta) = \phi \in \left[\hat{H}^{-1}(\Phi(z[\alpha])), \hat{H}^{-1}(\Phi(z[1-\alpha]))\right]$$
$$\theta \in \left[\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1-\alpha]))\right]$$

Note that the form of the transformation h never comes into play. In a sense, the method automatically selects a transformation that brings $\hat{\theta}$ to normality, computes an exact 95% interval, and then transforms backwards to reach the θ scale again.

We require estimates of a and z_0 to make this work:

We estimate the quantity z_0 as follows. Recalling that $\hat{\phi} = \phi + \sigma_\phi(Z - z_0)$,

$$P_\theta(\hat{\theta} < \theta) = P_\phi(\hat{\phi} < \phi) = P(Z < z_0) = \Phi(z_0),$$

so that $z_0 \approx \Phi^{-1}(\hat{G}(\hat{\theta}))$.

To estimate a, we leverage some interesting properties of the score transformation. For any smooth one-to-one function m, if $\phi = m(\theta)$ then

$$\begin{aligned} \frac{\partial}{\partial \phi} \log f(X, m^{-1}(\phi)) &= \frac{\partial}{\partial m^{-1}(\phi)} \log f(X, m^{-1}(\phi)) \frac{\partial m^{-1}(\phi)}{\partial \phi} \\ &= \frac{\partial}{\partial \theta} \log f(X, \theta) \frac{1}{m'(\theta)}. \end{aligned}$$

For another smooth one-to-one function g, transforming $Y = g(X)$

$$\frac{\partial}{\partial \phi} \log f(g^{-1}(Y), h^{-1}(\phi)) \left| \frac{dX}{dY} \right| = \frac{\partial}{\partial \theta} \log f(X, \theta) \frac{1}{h'(\theta)}$$

The Jacobian doesn't depend on parameters, and $g^{-1}(Y) = X$.

For $\dot{l}_\phi(\hat{\phi})$ and $\dot{l}_\theta(\hat{\theta})$, the previous results say that

$$\dot{l}_\phi(\hat{\phi}) = \dot{l}_\theta(\hat{\theta})/h'(\theta)$$

The skew of a random variable X is defined as $\mu_3(X)/\mu_2(X)^{3/2}$, so

$$SKEW(\dot{l}_\phi) = SKEW(\dot{l}_\theta).$$

Efron shows that $SKEW(\dot{l}_\theta)/6 \approx a$ by appealing to properties of $\dot{l}_\phi$ being a transformation of a standard normal. Thus we have all the components we need to form the BC_A interval in Lemma 1: $a, z_0, \hat{G}$.

For now, a diversion: consider a single interval endpoint $\hat{\theta}[\alpha]$ that is intended to have one-sided coverage α:

$$Prob(\theta \leq \hat{\theta}[\alpha]) \approx \alpha.$$

A procedure is *first-order* or *second-order* accurate if, respectively

$$\begin{aligned} Prob(\theta \leq \hat{\theta}[\alpha]) &= \alpha + O(n^{-1/2}), \\ Prob(\theta \leq \hat{\theta}[\alpha]) &= \alpha + O(n^{-1}). \end{aligned}$$

A procedure is *first-order* or *second-order* correct if, respectively

$$\begin{aligned} \hat{\theta}[\alpha] &= \hat{\theta}_{EX}[\alpha] + O_p(n^{-1}), \\ \hat{\theta}[\alpha] &= \hat{\theta}_{EX}[\alpha] + O_p(n^{-3/2}). \end{aligned}$$

Generally, n^{th} order correctness implies n^{th} order accuracy.

Main result: the BC_A interval is *second-order correct*. The proof involves a lot of "straightforward" expansions of quantile functions, estimators, and distribution functions. Still working on that.

We also extend the BC_A method to multiparameter families $\mathcal{G} = \{g_\eta : \eta \in \Lambda \subset \mathbb{R}^k\}$, to estimate $\theta = t(\eta)$ for some one to one function $t : \Lambda \to \mathbb{R}$. The notation ∇t denotes the gradient of t with respect to η.

The idea is to use a one-dimensional subfamily of $\mathcal{G}$ that will "stand-in" for the whole family, but which one?. Assume that our estimator is the MLE $t(\hat{\eta})$.

When estimating θ with an unbiased estimator $\hat{\theta}$, the CR bound is $\nabla t^T I^{-1}(\eta)\nabla t$. For an arbitary non-zero vector $d \in \mathbb{R}^k$, consider the one-parameter subfamily $\mathcal{G}_d = \{g_{\hat{\eta}+\tau d} : \tau \in \mathbb{R}\}$.

- $I_\tau(\tau) = d^T I(\hat{\eta} + d\tau)d$
- $\frac{d\theta}{d\tau} = d^T \nabla t(\hat{\eta} + d\tau)$
- Consider the CR bound at $\tau = 0$, $\frac{d^T \nabla t(\hat{\eta}) \nabla^T t(\hat{\eta}) d}{d^T I(\hat{\eta}) d}$, as a function of d.
- From linear algebra, this is maximized by choosing $\hat{\delta} \equiv I^{-1}(\eta)\nabla t(\hat{\eta})$, with maximum value $\nabla t^T I^{-1}(\eta)\nabla t$! In other words, estimation in this family is no easier than in the full family.

The subfamily $\mathcal{G}_{\hat{\delta}}$ is called least favorable, and is due to Stein (1956). On this one-parameter family, we can operate with machinery previously derived. We also extend the method to a non-parametric case!

We typically imagine bootstrap samples as resampling the data with replacement, but another way to think about it is to consider the sample space $\hat{\chi} = (x_1, x_2, \ldots, x_n)$ fixed, and consider only distributions F supported on $\hat{\chi}$.

These distributions consist of all possible ways of shuffling around the mass on the points of $\hat{\chi}$, hence F is an n-category multinomial family, and is also an exponential family. This implies all the multivariate results can be extended to the non-parametric case.

Moving forward:

- Work out proof of main theorem (2nd order correctness).
- Fill in some details of non-parametric and multi-parameter derivations of formulas
- Evaluate method on some “hard” problems (e.g. ratios of means)
- Make some pictures!