JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

Amy Braverman

(Jet Propulsion Laboratory)

The Bootstrap

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Outline

Whether Frequentist or Bayesian, we need to know how the sampling distribution of $\hat{\theta}$ depends on the (realized) value (of $\Theta =$) θ.

Can we get this information without making assumptions about the distribution*, $f_{\hat{\theta}|\Theta}(\hat{\theta}|\theta)$?

* Please forgive the abuse of notation: using $\hat{\theta}$ to represent both a random variable and its realized value.

Outline

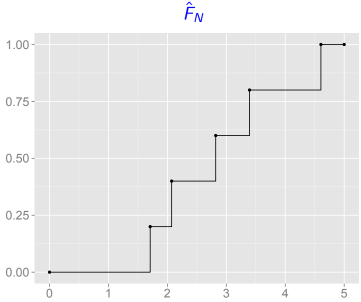
Introduction to Bootstrap:

- The ordinary Bootstrap.
- The M-out-of-N Bootstrap.
- Other Boostraps.

Material in this section based substantially on notes provided by Professors Anirban DasGupta of Purdue University and Charlie Geyer of the University of Minnesota, and on the Politis, Romano, and Wolf 1999 book, *Subsampling*.

The ordinary Bootstrap

- Say we have $Y_1, \ldots, Y_N$, an iid sample from distribution F (previously called $F_X(x)$ in Part 3).
- We have a statistic, $\hat{\theta} = g(Y_1, \ldots, Y_N)$, and need to know its sampling distribution.
- Note that the sample itself defines a distribution (the empirical distribution function) that puts mass $1/N$ on each realized value, $y_1, \ldots, y_N$. Call this $\hat{F}_N$.



Example:

$$\mathbf{y} = (2.82, 1.71, 2.07, 4.60, 3.39)^T.$$

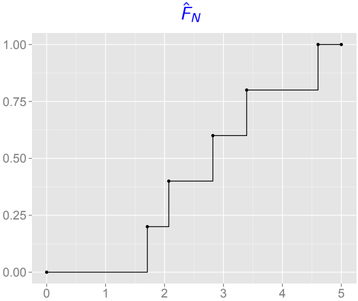
The ordinary Bootstrap

- A resample is a sample of size N from $\hat{F}_N$ drawn *with* replacement. Denote the resample by $Y_1^*, \ldots, Y_N^*$.
- Compute $\hat{\theta}^* = g(Y_1^*, \ldots, Y_N^*)$.
- Repeat the resampling B times and obtain $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
- The CDF of $\hat{\theta}$ is estimated by $H_{Boot,N}(t)$:

$$H_{Boot,N}(t) = P_{\hat{F}_N}(\hat{\theta}^* \leq t) \approx \frac{1}{B}\sum_{b=1}^{B} 1(\hat{\theta}_b^* \leq t),$$

where $P_{\hat{F}_N}(\cdot)$ is the probability defined by the empirical distribution function of the sample, and $1(\cdot) = 1$ if its argument is true and zero otherwise.

The ordinary Bootstrap



Example:

$$\mathbf{y} = (2.82, 1.71, 2.07, 4.60, 3.39)^T,\ \hat{\theta}(\mathbf{Y}) = 2.82,$$

$$\mathbf{y}_1^* = (2.82, 4.60, 1.71, 2.07, 4.60)^T,\ \hat{\theta}_1^*(\mathbf{Y}) = 2.82,$$
$$\mathbf{y}_2^* = (2.07, 2.07, 2.82, 1.71, 4.60)^T,\ \hat{\theta}_2^*(\mathbf{Y}) = 2.07,$$
$$\mathbf{y}_3^* = (3.39, 3.39, 1.71, 2.82, 1.71)^T,\ \hat{\theta}_3^*(\mathbf{Y}) = 2.82,$$
$$\vdots$$
$$\mathbf{y}_B^* = (3.39, 1.71, 3.39, 1.71, 2.82)^T,\ \hat{\theta}_B^*(\mathbf{Y}) = 2.82.$$

The PDF (or PMF) of $\hat{\theta} = g(Y_1, \ldots, g(Y_N))$ is approximated by the histogram of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

The ordinary Bootstrap

Compare $H_{Boot,N}(t)$ to the true CDF of $\hat{\theta}$, $H_N(t)$:

$$H_{Boot,N}(t) = P_{\hat{F}_N}(g(Y_1^*, \dots, Y_N^*) \leq t) \approx \frac{1}{B}\sum_{b=1}^{B} 1(\hat{\theta}_b^* \leq t),$$

$$H_N(t) = P_F(g(Y_1, \dots, Y_N) \leq t).$$

Two sources of error using $H_{Boot,N}(t)$ in place of $H_N(t)$:

- Treating $Y_1^*, \dots, Y_N^*$ as if they were draws from F rather than $\hat{F}_N$,
- The Monte Carlo approximation of $P_{\hat{F}_N}(g(Y_1^*, \dots, Y_N^*) \leq t)$.

The Monte Carlo error can be driven down by making B large, but what about $\hat{F}_N$ versus F?

The ordinary Bootstrap

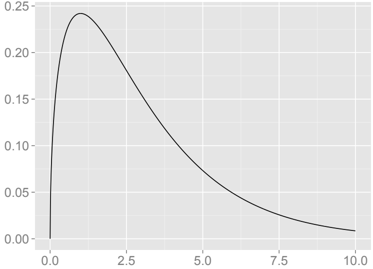
- $\hat{F}_N$ converges (in various senses) to F as $N \to \infty$ so Y_n^* are more and more like Y_n as N gets large.

- Does this guarantee that $H_{Boot,N}(t) \to H_N(t)$ for all t? (This is called statistical consistency.) This is what we mean by “works".

- No. It depends on various things including the transformation $g(\cdot)$ and properties of F.

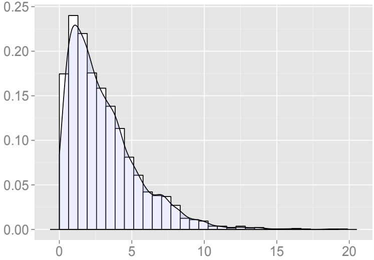
The ordinary Bootstrap

$\chi^2(3)$



$q_{.50} = 2.366$

$N = 3000$ iid draws from $\chi^2(3)$

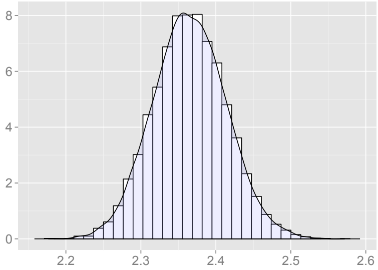


$q_{.50} = 2.415$

The ordinary Bootstrap

True distribution of the sample median, $\hat{\theta}$

(Based on 10,000 samples of size 3000)

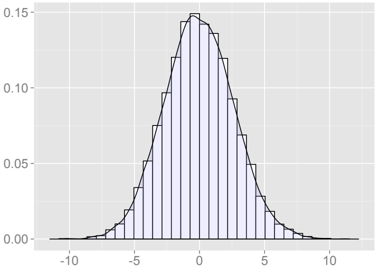


$$E(\hat{\theta}) = 2.365,$$

$$var(\hat{\theta}) = .002$$

True distribution of $\sqrt{N}(\hat{\theta} - q_{.50})$

$N = 3000, B = 10,000$



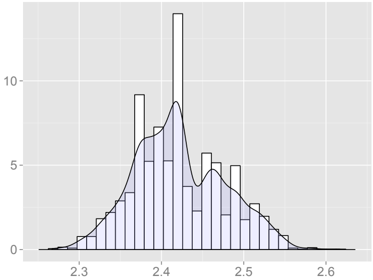
$$E\left(\sqrt{N}(\hat{\theta} - q_{.50})\right) = -.049,$$

$$var\left(\sqrt{N}(\hat{\theta} - q_{.50})\right) = 7.118$$

The ordinary Bootstrap

Bootstrap distribution of the sample median, $\hat{\theta}^*$

$N = 3000, B = 3000$

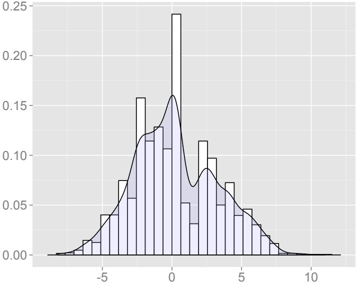


$$E(\hat{\theta}^*) = 2.421,$$

$$var(\hat{\theta}^*) = .033$$

Bootstrap distribution of the sample median, $\sqrt{N}\left(\hat{\theta}^* - \hat{\theta}\right)$

$N = 3000, B = 3000$



$$E\left(\sqrt{N}(\hat{\theta}^* - \hat{\theta})\right) = .257,$$

$$var\left(\sqrt{N}(\hat{\theta}^* - \hat{\theta})\right) = 9.063$$

The ordinary Bootstrap

Some examples where the Bootstrap “works" (is statistically consistent):

- Sample mean, if F has finite variance.
- Sample skewness, if F has six finite moments and positive variance.
- Sample kurtosis, if F has eight finite moments and positive variance.
- Sample correlation coefficient between two random variables, if each variable has at least four finite moments.
- t-statistic, if F has four finite moments.
- Sample quantiles (but larger N required).

The ordinary Bootstrap

When does the Bootstrap *not* work?

- F has infinite variance.
- Badly behaved $g(\cdot)$ (not continuous, not differentiable etc).
- The support of F depends on θ (e.g., uniform distribution on $[0, \theta]$).
- Same sorts of situation where the CLT fails.

In many of these sorts of cases, there *is* a variant of the Bootstrap that does “work": the *M*-out-of-*N* bootstrap.

M-out-of-N Bootstrap

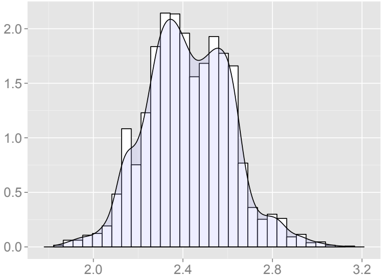
The M-out-of-N bootstrap:

- Resample size = $M < N$ can correct statistical consistency.
- The condition: $M/N \to 0$ as $M, N \to \infty$.
- How to choose M? $M = N^{1/2}, N^{2/3}$ as a rule-of-thumb.

M-out-of-N Bootstrap

M/N bootstrap distribution of $\hat{\theta}^*$

$N = 3000, B = 3000, M = \lfloor N^{2/3} \rfloor$

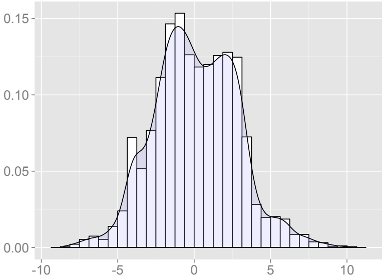


$$E(\hat{\theta}^*) = 2.424,$$

$$var(\hat{\theta}^*) = .035$$

M/N bootstrap distribution of $\sqrt{M}(\hat{\theta}^* - \hat{\theta})$

$N = 3000, B = 3000, M = \lfloor N^{2/3} \rfloor$



$$E\left(\sqrt{M}(\hat{\theta}^* - \hat{\theta})\right) = .111,$$

$$var\left(\sqrt{M}(\hat{\theta}^* - \hat{\theta})\right) = 7.342$$

Other Bootstraps

There are many other bootstrap schemes out there for both iid and dependent data:

- parametric bootstrap
- residual bootstrap
- smoothed bootstrap
- wild bootstrap

- non-overlapping block bootstrap
- moving-block bootstrap
- circular bootstrap
- sieve bootstrap
- frequency-domain bootstrap
- wavestrap

Bibliography

- *An Introduction to the Bootstrap* by Bradley Efron and R.J. Tibshirani, Chapman and Hall, 1979.
- *Resampling Methods for Dependent Data* by S.N. Lahiri, Springer, 2003.
- *Subsampling* by Dimitris N. Politis, Joseph P. Romano, and Michael Wolf, Springer, 1999.

Next

Subsampling: a related technique that works under less restrictive conditions.