MONTE CARLO AND THE BOOTSTRAP

Jim Booth
Department of Statistics
University of Florida
jbooth@stat.ufl.edu

BACKGROUND

- The most important applications of the bootstrap are
  1. Variance estimation (standard errors)
  2. Confidence intervals
- Folklore – many more resamples needed for confidence intervals than standard errors
- But simplest c.i. formula is

  $$\hat{\theta} \pm \mathrm{MOE}$$

  where

  $$\mathrm{MOE} = 2\mathrm{SE}$$
- Approximating c.i. endpoint is the same problem as approximating the standard error!!

OUTLINE

- Plug-in rule
- Resampling
- Exact bootstrap computation
- Monte Carlo approximation
- Rule-of-thumb for number of resamples
- Application to bootstrap variance
- Efron’s argument
- Application to confidence intervals

BOOTSTRAP $\equiv$ PLUG-IN RULE

- Sample: $Y_1, \ldots, Y_n$ i.i.d. F
- Plug-in rule: estimate $T(F)$ by $T(\hat{F})$
- e.g. empirical CDF:

$$\hat{F}(y) = \text{proportion of sample values} \leq y$$

$\hat{F}$ assigns probability $1/n$ to each sample value

References:

Hall (1992), Efron and Tibshirani (1993), Davison and Hinkley (1997)

- Consider functionals of the form

$$T(F) = E_F\left\{h(Y;F)\right\}$$

- The bootstrap estimate of $T(F)$ is

$$T(\hat{F}) = E_{\hat{F}}\left\{h(Y^*;\hat{F})\right\}$$

where Y^* denotes a *resample* - a sample drawn from $\hat{F}$.

e.g. Let $\hat{\theta}$ be an estimate of a population characteristic (parameter) $\theta = \theta(F)$ based on the sample. The variance of $\hat{\theta}$ in repeated sampling is

$$\sigma^2(F) = E_F\left[\left\{\hat{\theta} - E_F(\hat{\theta})\right\}^2\right]$$

EXAMPLE

Scores on 5 different math tests by 80 first year students
(Mardia, Kent and Bibby, 1979)

- Suppose the scores for the 80 students are a random sample from a population with unknown variance-covariance matrix, Σ.
- The proportion of the total variation explained by the first principal component is $\theta = \lambda_1 / tr(\Sigma)$. where λ_1 is the largest eigenvalue of Σ.
- An estimate of θ is $\hat{\theta} = \hat{\lambda}_1 / tr(S)$ where $\hat{\lambda}_1$ is the largest eigenvalue of the sample variance-covariance matrix, S.

Q : How do you estimate the standard error $\hat{\theta}$?

A : Square root of bootstrap variance (Efron and Tibshirani, 1993)

RESAMPLING

- In the case of the empirical CDF, $\hat{F}$, assigns probability $1/n$ to each value in the original sample.
- Operationally, we draw a resample

  $$Y^* = (Y_1^*, \ldots, Y_n^*)$$

  by sampling *with replacement* from $Y = (Y_1, \ldots, Y_n)$.
- Let x_j denote the number of times Y_j occurs in the resample. Then,

  $$(x_1, \ldots, x_n) \sim Mult\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$$

Drawing a resample is essentially equivalent to generating a symmetric multinomial vector.

EXACT BOOTSTRAP COMPUTATION

$$T(\hat{F}) = E_{\hat{F}}\left\{h(Y^*;\hat{F})\right\} = \sum_{y^* \in S} h(y^*;\hat{F}) \frac{n!}{\prod_{i=1}^{n} x_i!} \left(\frac{1}{n}\right)^n$$

where

$$S = \text{set of all possible bootstrap samples}$$

$$\#S = \binom{2n-1}{n-1}$$

n	$\#S$
8	6,435
10	92,378
15	77×10^6

Proof: How many ways can *n* indistinguishable balls be arranged in *n* cells? (Feller,V1,II.5)

MONTE CARLO APPROXIMATION

Let $Y_1^*, \ldots, Y_m^*$ denote independent resamples from $\hat{F}$. By the LLN,

$$\frac{1}{m}\sum_{r=1}^{m} h(Y_r^*;\hat{F}) \longrightarrow E_{\hat{F}}\left\{h(Y^*;\hat{F}\right\} = T(\hat{F})$$

as $m \to \infty$, w.p. 1.

“Monte Carlo techniques are often required for the extraction process, but that is not essential to the basic idea of the bootstrap.”

Efron (1994, Stat. Sci.)

BOOTSTRAP VARIANCES

Booth and Sarkar, 1998, *American Statistician*

$$\begin{aligned}
\sigma^2 = E_F\left[\left\{\hat\theta - E_F(\hat\theta)\right\}^2\right] &= \text{true variance} \\
\hat\sigma^2 = E_{\hat F}\left[\left\{\hat\theta^* - E_{\hat F}(\hat\theta^*)\right\}^2\right] &= \text{bootstrap variance} \\
\hat\sigma^2_m = \frac{1}{m}\sum_{r=1}^{m}(\hat\theta^*_r - \hat\theta^*_{\cdot})^2 &= \text{MC approximation}
\end{aligned}$$

where $\hat\theta^*_{\cdot}$ is the average of the $\hat\theta^*_r$'s.

Theorem: approx/exact bootstrap ratio is chisquared:

$$m\frac{\hat{\sigma}_m^2}{\hat{\sigma}^2} \sim \chi^2(m)$$

Proof: Sampling distribution of $\hat{\theta}$ (and therefore the resampling distribution of $\hat{\theta}^*$) is approximately normal.

Corollary: Relative error is normally distributed:

$$\frac{\hat{\sigma}_m^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \sim \sqrt{\frac{2}{m}}Z, \quad \text{where} \quad Z \sim N(0,1)$$

Proof: m is big!

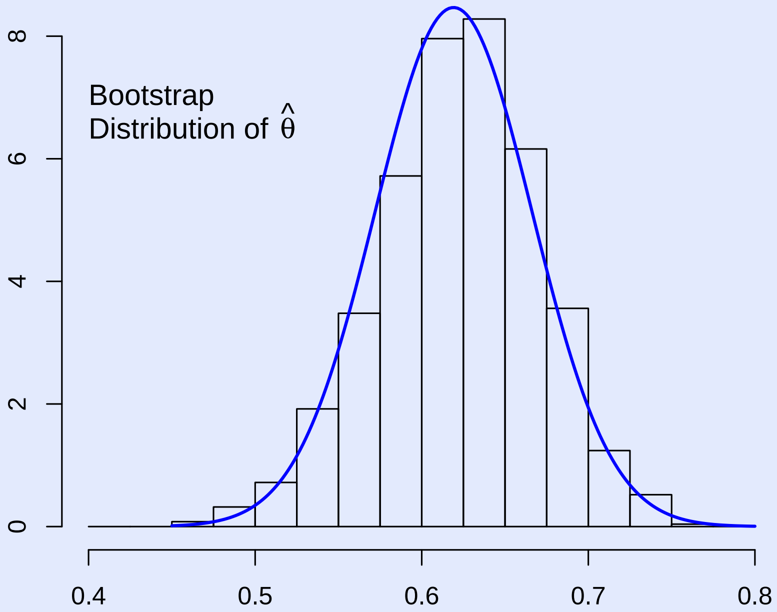

Figure 1: Bootstrap distribution of proportion of variance explained by the first principle component for the open and closed book data

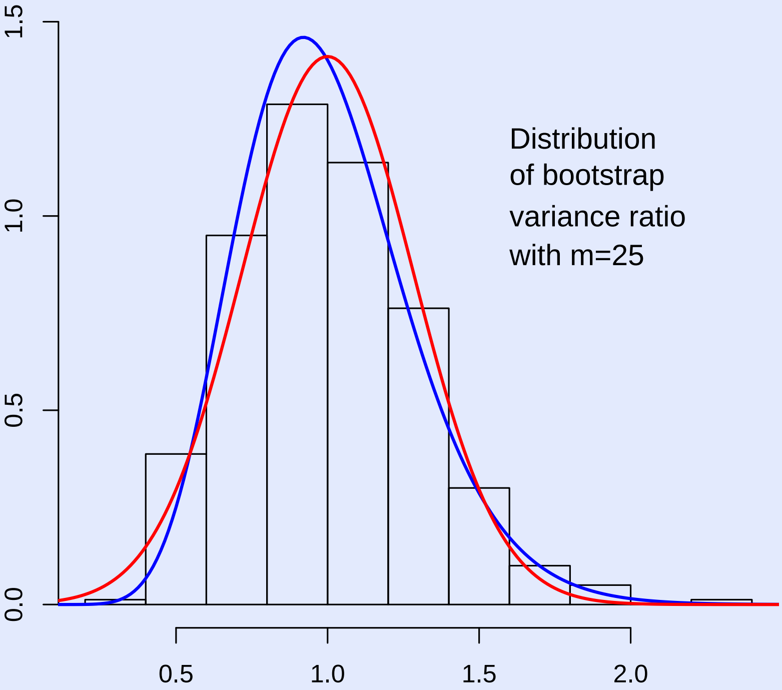

Figure 2: Distribution of 400 bootstrap variance estimates, each based on 25 resamples, relative to the “exact” bootstrap variance.

RULE-OF-THUMB

Informal: Same answer before and after lunch!

Formal: 10% margin-or-error for the bootstrap variance (with 95% confidence).

$$\begin{aligned} 1-2\alpha &\leq P\left\{-\delta < \frac{\hat{\sigma}_m^2-\hat{\sigma}^2}{\hat{\sigma}^2} < \delta\right\} \\ &= 1-2\Phi(-\delta\sqrt{m/2}) \end{aligned}$$

$$\boxed{m \geq 2(z_\alpha/\delta)^2}$$

e.g. $2\alpha = .05,\ \delta = .1 \Rightarrow m \geq 800$

"...there is little improvement past $m = 100$. In fact, m as small as 25 gives reasonable results." Efron (1987, JASA)

P-VALUE JUSTIFICATION

Consider a test of $H_0 : \theta = \theta_0$ based on

$$Z = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}} \sim \text{ approx } N(0,1)$$

- Significance at 5% level attained if $Z \approx 2$.
- $100\delta\%$ relative error requirement is

$$\frac{Z}{\sqrt{1+\delta}} < Z_m = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_m} < \frac{Z}{\sqrt{1-\delta}}$$

δ	m	Probable range of p-values
.1	800	(.039, .062)
.2	200	(.028, .074)
.5	30	(.006, .110)

COEFFICIENT OF VARIATION

- 10% margin-of-error for the bootstrap variance (with 95% confidence) is equivalent to

$$2\,COV_{\hat{F}}(\hat{\sigma}^2_m) = 2\frac{\mathrm{SD}_{\hat{F}}(\hat{\sigma}^2_m)}{\hat{\sigma}^2} \leq .1$$

- By the Corollary, $COV_{\hat{F}}(\hat{\sigma}^2_m) = \sqrt{2/m}$.
- Notice that this is the COV *conditional* on the sample. We are concerned about the impact of Monte Carlo variability.

BOOTSTRAP STANDARD ERROR

Using the delta method:

$$COV_{\hat{F}}(\hat{\sigma}_m) \approx \frac{1}{2} COV_{\hat{F}}(\hat{\sigma}_m^2)$$

10% margin-of-error for the bootstrap variance

$$\equiv$$

5% margin-of-error for the bootstrap standard error

$$\boxed{2\,COV_{\hat{F}}(\hat{\sigma}_m) \leq .05}$$

UNCONDITIONAL C.O.V.

Efron (1987, JASA) suggested choosing m based on the **unconditional** COV:

$$COV_F(\hat{\sigma}_m) \approx \sqrt{COV_F(\hat{\sigma})^2 + COV_{\hat{F}}(\hat{\sigma}_m)^2}$$

$$\approx \sqrt{COV_F(\hat{\sigma})^2 + \frac{1}{2m}}$$

"For values of $COV(\hat{\sigma}) \geq .1$, typical in practice, there is little improvement past $m = 100$. In fact, m as small as 25 gives reasonable results."

UNCONDITIONAL P-VALUE

$$Z_m = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_m} = \left(\frac{\hat{\theta} - \theta_0}{\hat{\sigma}} \right) \frac{\hat{\sigma}}{\hat{\sigma}_m} \sim \frac{Z}{\sqrt{\chi^2_m / m}}$$

- Unconditionally, Z_m has a t-distribution with df$= m$, whereas Z has a standard normal distribution. Doesn't this suggest low values m (such as $m = 25$) are OK?
- No! ± LUNCH TEST

CONFIDENCE INTERVALS

- Naive "bootstrap" c.i.: $\hat{\theta} \pm 2\hat{\sigma}$

- MC approximation: $\hat{\theta} \pm 2\hat{\sigma}_m$

- Want MC error small relative to interval halfwidth; i.e.

$$2COV_{\hat{F}}(\text{halfwidth}) = 2COV_{\hat{F}}(\hat{\sigma}_m) < \varepsilon$$

- Rule-of-thumb: $\varepsilon = .05$

5% margin-of-error in endpoints of bootstrap confidence intervals relative to halfwidth

PERCENTILE METHOD

- Denote the α-quantile of the sampling distribution of $\hat{\theta}$ by $\theta_\alpha = \theta_\alpha(F)$; i.e.

$$\alpha = P_F(\hat{\theta} \leq \theta_\alpha)$$

- The bootstrap estimate of θ_α is $\hat{\theta}_\alpha = \theta_\alpha(\hat{F})$ satisfying

$$\alpha = P_{\hat{F}}(\hat{\theta}^* \leq \hat{\theta}_\alpha)$$

- A Monte Carlo approximation of $\hat{\theta}_\alpha$ is

$$\hat{\theta}_{\alpha,m} = \hat{\theta}^*_{(k)}$$

  where $k = (m+1)\alpha$ and $\hat{\theta}^*_{(k)}$ is the kth order statistic from the versions of $\hat{\theta}_\alpha$ computed using m resamples.

  e.g. $m = 999$, $\alpha = .025 \longrightarrow k = 25$

- Efron's original percentile method confidence interval with nominal coverage level $1-2\alpha$ is $(\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha})$.

Efron (1987) suggested choosing m for bootstrap confidence intervals based on the **conditional** COV:

$$COV_{\hat{F}}(\text{halfwidth}) = COV_{\hat{F}}(\hat{\theta}_{\alpha,m} - \hat{\theta}) \approx \sqrt{\frac{\alpha(1-\alpha)}{m z_\alpha^2 \phi(z_\alpha)^2}}$$

- Numerical comparison for 95% confidence intervals

	$2COV_{\hat{F}}(\text{halfwidth})$	
m	Naive	Percentile
200	.10	.19
800	.05	.10
3,000	.026	.05

WHY USE T-TABLES?

MC error should be less than t-table adjustment!!

$$z_{.025} = 1.960 \qquad \frac{\text{error}}{\text{halfwidth}} = .08$$

$$t_{.025,15} = 2.131$$

DOUBLE BOOTSTRAP

- Approximation of double bootstrap confidence intervals involves nested levels of resampling; i.e. resamples from each resample.
- Equal allocation of such as 1000(1000) is not optimal – 5000(200) is better!
- One million nested resamples fails to attain 5% margin-of-error for the endpoints of a 95% interval.

Booth and Hall, 1993; Booth and Presnell, 1998

SUMMARY

- If the *bootstrap variance* is to be used for inference (e.g. hypothesis testing, confidence intervals), then more resamples are required than generally thought. $m = 800$ not $m = 50$!
- Construction of *percentile method* bootstrap confidence intervals also requires more resamples than generally thought. $m = 3,000$ not $m = 1,000$!
- Theoretical benefits of the *double bootstrap* such as second-order correctness and coverage accuracy are swamped by Monte Carlo error in practice. Even $m = 1,000,000$ is not enough!
- Simulation studies that only emphasize coverage accuracy are misleading. (± LUNCH TEST)
- Importance sampling can reduce the number of resamples for confidence intervals.