This bootstrap, time: MCMC, next MCMC.

Real: IS ch.12 First (ARS132)
 Draft [14 Mar 17]

If take-home final is 0 available at course website (absolute due date Fri 24 Mar 2017 5pm)

do time

Please complete online course evaluation form for cns 132

The bootstrap

Brad Efron (1979)

(1938-)

So far we have always assumed a parametric sampling distribution for the data. Ex. Pos.

\( (i=1, \ldots, n) \sim \text{Bernoulli}(\theta) \)

Case Study

Ex. NB19

case study

\( (i=1, \ldots, n) \sim \text{Exponential} (\theta) \)

Ex. ER10

\( (i=1, \ldots, n) \sim \text{Tr}(\mu, \sigma^2) \)

Is it possible to draw valid inferences without assuming a specific parametric form for the sampling distribution? A: Yes, in some cases.
This part of statistics is called nonparametric inference, & it has both frequentist & Bayesian versions. Efron developed the bootstrap as one frequentist nonparametric approach.

\[ \text{Math Fact: Every probability distribution is uniquely characterized by its cumulative distribution function } F. \]

\[ \text{Def. If a random variable taking values on the real number line } \mathbb{R}. \]

The CDF of \( \mathbb{Z} \) is defined to be \( F_{\mathbb{Z}}(y) = P(\mathbb{Z} \leq y). \)

CDFs of continuous RVs look like this:

\[
\begin{align*}
0 & \leq F_{\mathbb{Z}}(y) \leq 1 \\
& \text{is nondecreasing in } y
\end{align*}
\]
Setting $F_\xi(y) = p$, if $F_\xi$ is continuous it will have an inverse: $y = F_\xi^{-1}(p)$

Ex. $m = F_\xi^{-1}(\frac{1}{2})$ is the median of the distribution of $\xi$:

Case I: A random sample of $n=100$ final exam scores from the population of all UCLA students who took introductory chemistry in 2013.

Can we estimate the population median & build a 95% confidence interval for $m$, without assuming a specific parametric form for $F$? A: yes
There are actually two frequentist non-parametric methods to solve this problem, one specific to the median (or some other percentile/quantile) and one that's much more general. Specify method: 

Suppose the population CDF is continuous.

Take your random sample $X_1, \ldots, X_n$ from $F$ and sort it from smallest to largest. Let's call the sorted dataset $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. 

\[ \text{minimum order statistic: } X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \leq \text{maximum of } X \]
What's the probability that any given data value \( z_i \) falls below the population median?

A: \( P(z_i < m) = P(z_i > m) = \frac{1}{2} \)

What's the probability that \( z_{(i)} \) falls above the median?

Let \( S = \# \text{ of } z_i > m \)

Then \( S \) follows the Binomial distribution with parameters \( n \) and \( \frac{1}{2} \), so

\[ S \sim \text{Binomial}(n, \frac{1}{2}) \]

Similarly,

\[ P(z_{(i)} < m) = P(S = n-1) = (n-1) \left( \frac{1}{2} \right)^{n-1} \left( \frac{1}{2} \right) = n \left( \frac{1}{2} \right)^n 
\]

\[ = P(S = 1) \]
So you can find a $100(1 - \alpha)%$ CI for $m$ just by finding the two integers $I_{low}$ and $I_{high}$ such that

\[ P\left( I_{low} < m < I_{high} \right) = \frac{\alpha}{2} \]

and

\[ P\left( I_{low} > m \right) = \frac{\alpha}{2} \]

UCLA Case Study: Let's compute that

\[ \text{binom}( .025, 101, 0.5 ) = 41 \]

\[ \text{binom}( .975, 101, 0.5 ) = 60 \]

So $P\left( \frac{41}{100} < m < \frac{60}{100} \right) = 95\%$. As our point estimate of $m$, we will use the sample median: $\hat{m} = I_{(50)} = 81$, with $95\%$ CI for $m (77, 83)$ (asymmetric).
The bootstrap method sampling model is \((Y_i; F) \sim F\) (i = 1, ..., n) cont.

The quantity of interest is \(m = F^{-1}(\frac{1}{2}) = \text{population median}\).

The best frequentist estimate of \(F\) on the basis of \(\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)\) is the empirical CDF \(\tilde{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(\tilde{y}_i \leq y)\).

Example: \(\tilde{y} = (1, 2, 2, 9)\)

\(F_n\) is a step function with jumps at the data values.
Sample |
---
the observed students

Final exam score

Actual

$F_0$ | $F_1$
---|---
$F_0$ | $F_1$

$n = 101$

Median $m = 81$

CDF $F = ?$

Hypothetical

IDD

This was our basic frequentist diagram earlier in the course.

Efron had an interesting idea: why not pretend the sample = the population and take repeated samples from this sample?
"P" = sample of observed students

bootstrap sample

final exam score

n = 101

median \( \hat{m}_1 = 81 \)

CDF = ECDF

\( F_n \)

median \( \hat{m}_2 \)

\[ \begin{pmatrix} Y_1^* \\ \vdots \\ Y_n^* \end{pmatrix} \]

h = 101

n = 101

\[ \begin{pmatrix} 80 \\ 79 \\ 82 \end{pmatrix} \]

4

\( m \)

\( m^* \)

approx. 95% bootstrap CI for m

This approach to building bootstrap CIs is called the percentile method; fancier small-n refinements exist but we don't have time to look at them.
It has recently been shown (Draper, 2010) (also see Hjort (1986)) that the bootstrap is actually a special case of a Bayesian nonparametric method based on Dirichlet process priors on CDFs.

A single Bayesian model comparison method

\[ \text{DIC} = \text{Deviance Information Criterion} \]

(Spiegelhalter et al. (2002))

DIC is one of a number of "information criteria" (AIC, BIC, ...) that trade off (model goodness of fit) \( \text{I} \) against (model complexity) \( \text{II} \). There is a tug of war between \( \text{I} \) and \( \text{II} \).
You can make a model fit the data arbitrarily well by increasing model complexity, but the resulting model won't do a good job of predicting new data because you have overfit it to the current dataset - DIC tries to resolve this by of way:

\[ \text{DIC} = \hat{D} + p_D \]

\( \hat{D} \) is an estimate of the deviance of your current model, defined to be:

\[ -2 \log \text{ (likelihood)} \]

we want the log likelihood (LL) to be big (the model fits the data well), so we want the deviance \((-2\hat{D})\) to be small.
$p_0$ is an estimate of the effective number of parameters in your current model: $\mathbb{E} \left( \frac{1}{p_0} \right) \sim \mathcal{N}(\mu, \sigma^2)$.

$p_0 = 2$ for this model $(\mu, \sigma^2)$.

We want $p_0$ to be small to avoid needlessly complicated models. But you can't simultaneously have $\tilde{I}$ and $p_0$ small because increasing model complexity (if you do it right) will increase the log(likelihood) and decrease the deviance $\hat{D}_L$ is a compromise between $\tilde{I}$ and $p_0$.  

$y$

$m_2$ is better than $m_1$ if $\text{DIC}(m_2) < \text{DIC}(m_1)$

How much smaller? See DIC FAQ page or course website.