Mathematical Biostatistics Bootcamp: Lecture 12, Bootstrapping

Brian Caffo

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

October 16, 2012

Table of contents

1. Table of contents
2. The jackknife
3. The bootstrap principle
4. The bootstrap

The jackknife

- The jackknife is a tool for estimating standard errors and the bias of estimators
- As its name suggests, the jackknife is a small, handy tool; in contrast to the bootstrap, which is then the moral equivalent of a giant workshop full of tools
- Both the jackknife and the bootstrap involve *resampling* data; that is, repeatedly creating new data sets from the original data

The jackknife

- The jackknife deletes each observation and calculates an estimate based on the remaining $n - 1$ of them
- It uses this collection of estimates to do things like estimate the bias and the standard error
- Note that estimating the bias and having a standard error are not needed for things like sample means, which we know are unbiased estimates of population means and what their standard errors are

The jackknife

- We'll consider the jackknife for univariate data
- Let $X_1, \ldots, X_n$ be a collection of data used to estimate a parameter θ
- Let $\hat{\theta}$ be the estimate based on the full data set
- Let $\hat{\theta}_i$ be the estimate of θ obtained by *deleting observation i*
- Let $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i$

Continued

- Then, the jackknife estimate of the bias is

$$(n-1)\left(\bar{\theta} - \hat{\theta}\right)$$

(how far the average delete-one estimate is from the actual estimate)

- The jackknife estimate of the standard error is

$$\left[\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2\right]^{1/2}$$

(the deviance of the delete-one estimates from the average delete-one estimate)

Example

- Consider the data set of 630 measurements of gray matter volume for workers from a lead manufacturing plant
- The median gray matter volume is around 589 cubic centimeters
- We want to estimate the bias and standard error of the median

Example

The gist of the code

```
n <- length(gmVol)
theta <- median(gmVol)
jk <- sapply(1 : n,
             function(i) median(gmVol[-i])
             )
thetaBar <- mean(jk)
biasEst <- (n - 1) * (thetaBar - theta)
seEst <- sqrt((n - 1) * mean((jk - thetaBar)^2))
```

Example

Or, using the `bootstrap` package

```
library(bootstrap)
out <- jackknife(gmVol, median)
out$jack.se
out$jack.bias
```

Example

- Both methods (of course) yield an estimated bias of 0 and a se of 9.94
- Odd little fact: the jackknife estimate of the bias for the median is always 0 when the number of observations is even
- It has been shown that the jackknife is a linear approximation to the bootstrap
- Generally do not use the jackknife for sample quantiles like the median; as it has been shown to have some poor properties

Pseudo observations

- Another interesting way to think about the jackknife uses pseudo observations
- Let
$$\text{Pseudo Obs} = n\hat{\theta} - (n-1)\hat{\theta}_i$$
- Think of these as "whatever observation i contributes to the estimate of θ"
- Note when $\hat{\theta}$ is the sample mean, the pseudo observations are the data themselves
- Then the sample standard error of these observations is the previous jackknife estimated standard error.
- The mean of these observations is a bias-corrected estimate of θ

Example: Tom's notes

The bootstrap

- The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistics
- For example, how would one derive a confidence interval for the median?
- The bootstrap procedure follows from the so called bootstrap principle

The bootstrap principle

- Suppose that I have a statistic that estimates some population parameter, but I don't know its sampling distribution
- The bootstrap principle suggests using the distribution defined by the data to approximate its sampling distribution

The bootstrap in practice

- In practice, the bootstrap principle is always carried out using simulation
- We will cover only a few aspects of bootstrap resampling
- The general procedure follows by first simulating complete data sets from the observed data with replacement
  - This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution
- Calculate the statistic for each simulated data set
- Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error

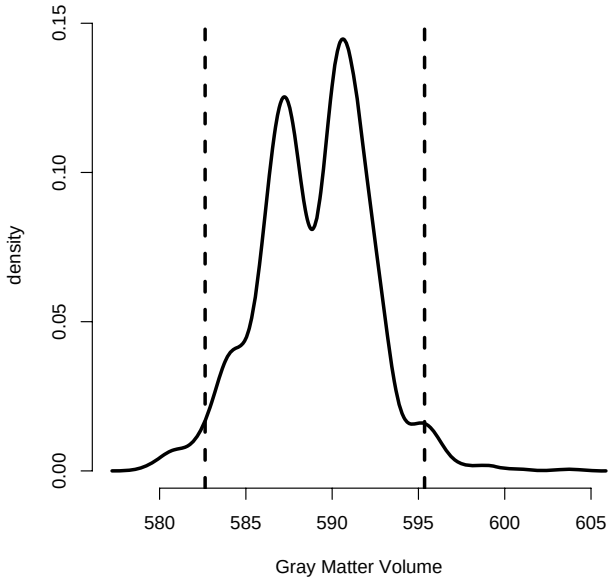
Example

- Consider again, the data set of 630 measurements of gray matter volume for workers from a lead manufacturing plant
- The median gray matter volume is around 589 cubic centimeters
- We want a confidence interval for the median of these measurements

- Bootstrap procedure for calculating for the median from a data set of n observations
  - *i*. Sample n observations **with replacement** from the observed data resulting in one simulated complete data set
  - *ii*. Take the median of the simulated data set
  - *iii*. Repeat these two steps B times, resulting in B simulated medians
  - *iv*. These medians are approximately draws from the sampling distribution of the median of n observations; therefore we can
    - Draw a histogram of them
    - Calculate their standard deviation to estimate the standard error of the median
    - Take the 2.5^{th} and 97.5^{th} percentiles as a confidence interval for the median

Example code

```
B <- 1000
n <- length(gmVol)
resamples <- matrix(sample(gmVol,
                           n * B,
                           replace = TRUE),
                    B, n)
medians <- apply(resamples, 1, median)
sd(medians)
[1] 3.148706
quantile(medians, c(.025, .975))
    2.5%    97.5%
582.6384 595.3553
```


density
0.00
0.05
0.10
0.15
580
585
590
595
600
605
Gray Matter Volume

Notes on the bootstrap

- The bootstrap is non-parametric
- However, the theoretical arguments proving the validity of the bootstrap rely on large samples
- Better percentile bootstrap confidence intervals correct for bias
- There are lots of variations on bootstrap procedures; the book "An Introduction to the Bootstrap" by Efron and Tibshirani is a great place to start for both bootstrap and jackknife information

```
library(boot)
stat <- function(x, i) {median(x[i])}
boot.out <- boot(data = gmVol,
                 statistic = stat,
                 R = 1000)
boot.ci(boot.out)
Level     Percentile            BCa
95%   (583.1, 595.2 )   (583.2, 595.3 )
```