Stat 704 Data Analysis I
Bootstrap

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Table of contents

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

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

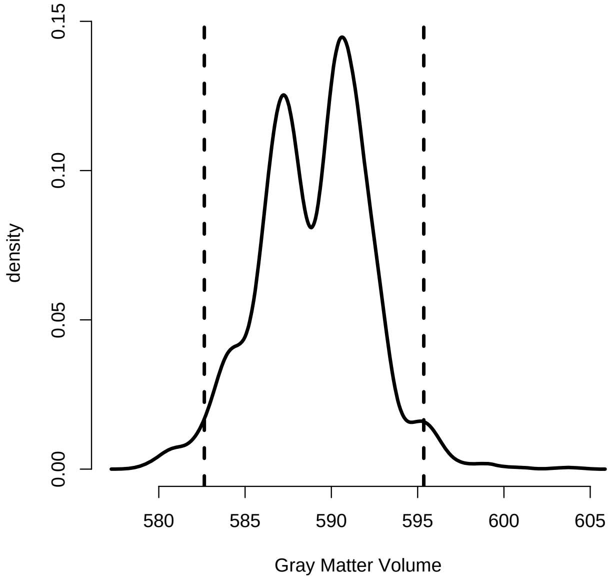
Example

- 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 confidence interval 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 drawn 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 )
```