Lecture 12

Ciprian M. Crainiceanu

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

September 8, 2020

Table of contents

Outline

1. The jackknife
2. Introduce the bootstrap principle
3. Outline the bootstrap algorithm
4. Example bootstrap calculations
5. Discussion

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 $\hat{\theta}_n$ be the estimate based on the full data set
- Let $\hat{\theta}_{i,n}$ be the estimate of θ obtained by *deleting observation i*
- Let $\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{i,n}$

Continued

- Then, the jackknife estimate of the bias is

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

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

- The jackknife estimate of the standard error is

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

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

Jackknife: intuition

Consider the case when the estimator $\hat{\theta}_n$ has bias of order n:

$$E(\hat{\theta}_n) = \theta + \frac{b}{n}$$

Then $E(\hat{\theta}_{i,n}) = \theta + \frac{b}{n-1}$ and

$$E(\bar{\theta}_n) = \theta + \frac{b}{n-1}$$

with the Jackknife estimator of the bias

$$(n-1)\{E(\bar{\theta}_n) - E(\hat{\theta}_n)\} = (n-1)\left\{\frac{b}{n-1} - \frac{b}{n}\right\} = \frac{b}{n}$$

Jackknife: bias correction

For all estimators with a bias of the type $E(\hat{\theta}_n) = \theta + \frac{b}{n}$

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

is unbiased

$$\mathrm{ps}_{i,n} = \hat{\theta}_n - (n-1)(\bar{\theta}_{i,n} - \hat{\theta}_n) = n\hat{\theta}_n - (n-1)\hat{\theta}_{i,n}$$

are called pseudo-values; Jackknife: treat pseudo-values as independent
If $\hat{\theta}_n = \bar{X}_n$ then $\mathrm{ps}_{i,n} = X_i$.

Jackknife: variance estimation

Assume that the variance of $\hat{\theta}_n$ is of the type σ^2/n
Then an estimator of σ^2/n is

$$\frac{1}{n}\{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{ps}_{i,n}-\bar{\mathrm{ps}}_n)^2\} \approx \frac{1}{n}\{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{ps}_{i,n}-\hat{\theta}_n)^2\}$$

As $(\mathrm{ps}_{i,n}-\hat{\theta}_n)=(n-1)(\hat{\theta}_n-\hat{\theta}_{i,n})$ the Jackknife formula for the variance estimator follows

Pseudo observations

- Another interesting way to think about the jackknife uses pseudo observations
- Let
$$\mathrm{ps}_{i,n} = n\hat{\theta}_n - (n-1)\hat{\theta}_{i,n}$$
- Think of these as "whatever observation i contributes to the estimate of θ"
- When $\hat{\theta}_n$ 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

- 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
- 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
- Do not use the jackknife for sample quantiles like the median; it has been shown to have some poor properties

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

Example: nonparametric bootstrap

- Given a vector of n subjects (rows, ...)
- "Nonparametric bootstrap": B resamples "with replacement"; each resample is done exactly n times

```
a=c("John","Sarah","Gina","Victor","Jimmy")
for (i in 1:6)
{print(sample(a,replace=TRUE))}
```

```
[1] "John"   "Jimmy"  "Sarah"  "Sarah"  "Victor"
[1] "Jimmy"  "Sarah"  "Victor" "Victor" "Victor"
[1] "Sarah" "Jimmy" "Sarah" "John"   "Gina"
[1] "Victor" "Jimmy"  "Gina"   "Sarah"  "Jimmy"
[1] "John"   "Sarah"  "John"   "Victor" "Gina"
[1] "Sarah"  "John"   "Victor" "Jimmy"  "Victor"
```

Set theory: an example

PID	BMI	SEX	AGE
1	22	1	45
2	27	0	57
3	31	1	66
4	24	1	49
5	23	0	33
6	18	0	40
7	21	0	65
8	26	1	59
9	34	1	65
10	20	0	42

Q: Construct a bootstrap CI for the difference in the mean BMI of women and men

Example: nonparametric bootstrap

```
data_bmi=read.table(file=file.name,header=TRUE)
attach(data_bmi)

women_bmi<-BMI[SEX==1]
men_bmi<-BMI[SEX==0]
n_women<-length(women_bmi)
n_men<-length(men_bmi)

B_boot<-10000
mean_diff=rep(NA,B_boot)
for (i in 1:B_boot)
{#Begin bootstrap
mw<-mean(women_bmi[sample(1:n_women,replace=TRUE)])
mm<-mean(men_bmi[sample(1:n_men,replace=TRUE)])
mean_diff[i]<-mw-mm
}#End bootstrap
```

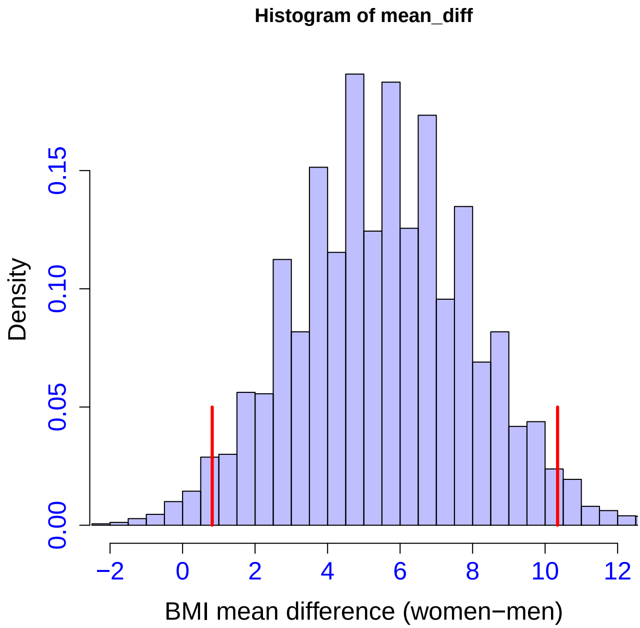
Example: nonparametric bootstrap

```
mBoot<-mean(mean_diff)
sdBoot<-sd(mean_diff)

CI1<-c(mBoot-1.96*sdBoot,mBoot+1.96*sdBoot)
CI2<-quantile(mean_diff,probs=c(0.025,0.975))

>CI1
[1]  0.8176955 10.3444245

>CI2
 2.5% 97.5%
  0.8  10.4
```


Histogram of mean_diff
Density
0.00
0.05
0.10
0.15
−2
0
2
4
6
8
10
12
BMI mean difference (women−men)

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

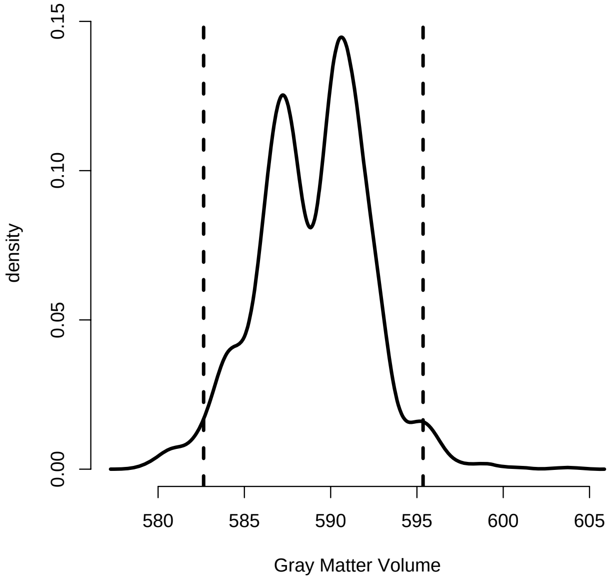
Example

- 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
  1. Sample n observations **with replacement** from the observed data resulting in one simulated complete data set
  2. Take the median of the simulated data set
  3. Repeat these two steps B times, resulting in B simulated medians
  4. 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
```


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

Notes on the bootstrap

- This bootstrap procedure is non-parametric
- Essentially: 1) sample with replacement the population to create a similar population *of the same size*; 2) apply whatever procedure you want to this resampled population; 3) repeat; 4) aggregate results; 5) report results.
- Theoretical arguments proving the validity of the bootstrap rely on large samples
- 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 )
```

Bradley Efron

- Bootstrap: Efron B (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics. 7, 1–26
- One of the most influential methods in Statistics
- A fundamental method based on understanding randomization
- B. Efron is Professor of Statistics at Stanford University