Lecture 6: Bootstrap II

Pratheepa Jeganathan

04/15/2019

Recall

- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in aspirin-intake group to placebo group.

Complete Enumeration

- Number of bootstrap samples $\boldsymbol{X}^* = \{x_1^*, \cdots, x_n^*\}$ is n^n.
- Not all these sets are different.
  - For example with $n = 3$, $\{x_1, x_1, x_3\}$ is same as $\{x_3, x_1, x_1\}$.
- Group together bootstrap samples that generate the same subset.
- Characterize each bootstrap sample by its weight vector $(k_1, \cdots, k_n)$, where k_i is the number of times x_i appears in the bootstrap sample. Thus, $k_1 + \cdots k_n = n$.
- Let the space of compositions of n into at most n parts be $\mathcal{C}_n = \{\boldsymbol{k} = (k_1, \cdots, k_n),\ k_1 + \cdots + k_n = n,\ k_i \geq 0,\ k_i \text{ integer}\}$.

Complete Enumeration

- Size of the space $\mathcal{C}_n$ is $|\mathcal{C}_n| = \binom{2n-1}{n-1}$.
  - Each component in the vector $(k_1, \cdots, k_n)$ is considered to be a box. There are n boxes to contain n balls in all.
  - We want to count the number of ways of separating the n balls into the n boxes.
  - Put $(n-1)$ separators of $|$ to make boxes and n balls.
    - For example, if $n = 3$ and $\boldsymbol{X} = \{x_1, x_2, x_3\}$.
    - o|o|o corresponds to $\boldsymbol{X}^{*1} = \{x_1, x_2, x_3\}$.
    - oo||o corresponds to $\boldsymbol{X}^{*2} = \{x_1, x_1, x_3\}$.
  - $2n-1$ positions from which to choose $n-1$ bars positions.
- Each bootstrap sample corresponds to sampling weight $\boldsymbol{k} = (k_1, \cdots, k_n) \sim \text{Multinomial}\,(n, \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \cdots, p_n)$ and $p_i = \dfrac{1}{n} \;\; \forall i$.

The exhaustive bootstrap

- The exhaustive bootstrap distribution of a statistic $T(\boldsymbol{X})$
  - compute each of the $\binom{2n-1}{n-1}$ statistics and
  - associate a weight $\boldsymbol{k} \sim \text{Multinomial}(n, \boldsymbol{p})$ with it.
- The shift from space of possible resamples with replacement to $\mathcal{C}_n$ gives substantial savings.
  - For example with $n = 10$, the number of enumerations reduce from $10^{10} \approx 1 \times 10^{10}$ to 92378.

Compare bootstrap method using Monte Carlo simulations and exhaustive bootstrap

- **W** Chapter 3.8: The data are LSAT scores (for entrance to law school) and GPA. This data were used to illustrate the bootstrap by Bradly Efron, the inventor of the bootstrap.

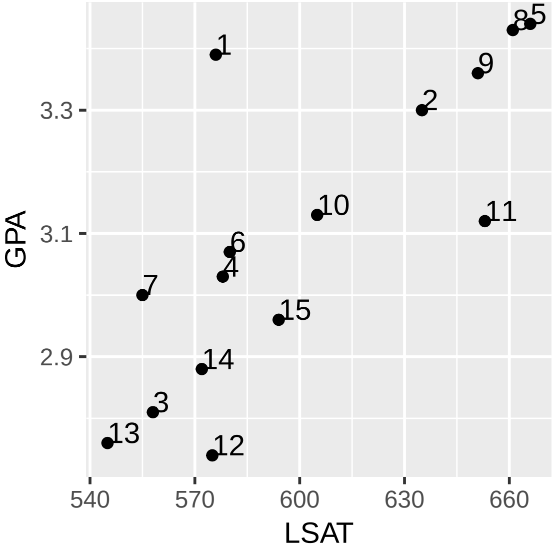
```
library(bootstrap);
data(law)
t(law)
```

```
##           1     2      3      4      5      6   7      8
## LSAT 576.00 635.0 558.00 578.00 666.00 580.00 555 661.00
## GPA    3.39   3.3   2.81   3.03   3.44   3.07   3   3.43
##          11     12     13     14     15
## LSAT 653.00 575.00 545.00 572.00 594.00
## GPA    3.12   2.74   2.76   2.88   2.96
```

Example (scatterplot)

```r
library(ggplot2)
ggplot(data = law, aes(x= LSAT, y= GPA))
```

Example (scatterplot)



Example (Plug-in estimate of the correlation coefficient)

```
theta.hat = cor(law$LSAT, law$GPA)
theta.hat
```

```
## [1] 0.7763745
```

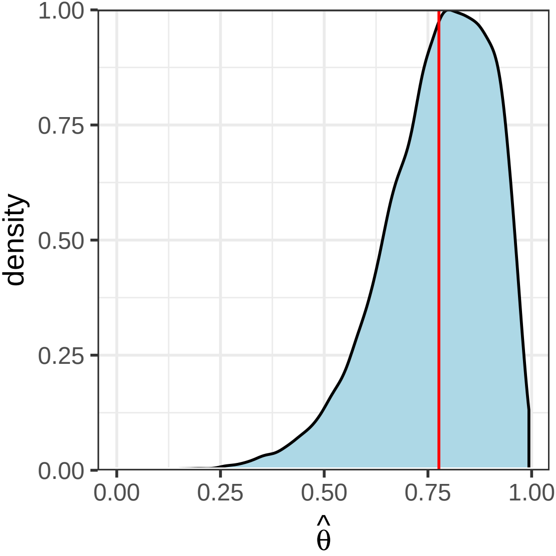
Example (bootstrap replicates)

```
draw.bootstrap.samples = function(df){
  n = dim(df)[1]
  ind = sample(n, replace = TRUE)
  cor.bootstrap.replicate = cor(df[ind, "LSAT"], df[ind, "G
  return(cor.bootstrap.replicate)
}
R = 10000
theta.hat.star = replicate(R, draw.bootstrap.samples(law))
```

Example (bootstrap approximation for the sampling distribution of plug-in estimator)

```
theta.hat.star.df = data.frame(theta.hat.star = theta.hat.s
ggplot(theta.hat.star.df) +
  geom_density(aes(x = theta.hat.star, y = ..scaled..),
    fill = "lightblue") +
  geom_hline(yintercept=0, colour="white", size=1) +
  theme_bw() +
  ylab("density") +
  xlab(bquote(hat(theta))) +
  geom_vline(xintercept = theta.hat, col = "red")+
  scale_y_continuous(expand = c(0,0))
```

Example (bootstrap approximation for the sampling distribution of plug-in estimator)



Example (standard error using bootstrap)

```r
sd(theta.hat.star)
```

```
## [1] 0.1318886
```

Example (the exhaustive bootstrap distribution of the plug-in estimate)

- Create matrix of all $\binom{2n-1}{n-1}$ enumerations.

```
library(partitions)
n = 15
allCompositions = compositions(n, n)
```

Example (the exhaustive bootstrap distribution of the plug-in estimate)

```r
allCompositions[,1:10]
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]   15   14   13   12   11   10    9    8    7     6
##  [2,]    0    1    2    3    4    5    6    7    8     9
##  [3,]    0    0    0    0    0    0    0    0    0     0
##  [4,]    0    0    0    0    0    0    0    0    0     0
##  [5,]    0    0    0    0    0    0    0    0    0     0
##  [6,]    0    0    0    0    0    0    0    0    0     0
##  [7,]    0    0    0    0    0    0    0    0    0     0
##  [8,]    0    0    0    0    0    0    0    0    0     0
##  [9,]    0    0    0    0    0    0    0    0    0     0
## [10,]    0    0    0    0    0    0    0    0    0     0
## [11,]    0    0    0    0    0    0    0    0    0     0
## [12,]    0    0    0    0    0    0    0    0    0     0
## [13,]    0    0    0    0    0    0    0    0    0     0
```

Example (the exhaustive bootstrap distribution of the plug-in estimate)

- Check number of compositions

```r
dim(allCompositions)[2] == choose((2*n-1), (n-1))
```

```
## [1] TRUE
```

Example (the exhaustive bootstrap distribution of the plug-in estimate)

- Compute $\binom{2n-1}{n-1}$ bootstrap replicates.

```
library(parallel)
nCompositions = dim(allCompositions)[2]
t.start = proc.time()
enumData = mclapply(1:nCompositions, function(i) {
    ind = allCompositions[,i]
    law.list = lapply(1:n,function(j) matrix(rep(law[j,], i
    newLaw = do.call(rbind, law.list)
    c(cor(unlist(newLaw[,1]),unlist(newLaw[,2])),dmultinom(
    }, mc.cores = 4)
proc.time() - t.start

enumData = t(simplify2array(enumData))
colnames(enumData) = c("ex.theta.hat.star","weight")
save(enumData,file = "enumData.Rdata")
```

Example (the exhaustive bootstrap distribution of the plug-in estimate)

```
load("enumData.Rdata")
ex.theta.hat.star.df =
  data.frame(ex.theta.hat.star =
      enumData$ex.theta.hat.star)
ggplot(ex.theta.hat.star.df) +
  geom_density(aes(x = ex.theta.hat.star, y = ..scaled..),
    fill = "lightblue") +
  geom_hline(yintercept=0, colour="white", size=1) +
  theme_bw() +
  ylab("density") +
  xlab(bquote(hat(theta))) +
  geom_vline(xintercept = theta.hat, col = "red") +
  ggtitle("The exhaustive bootstrap distribution of the plu
```

Gray Codes to speed up the enumeration

- We can speedup enumeration by changing only one coordinates at the time using Gray codes.
- Suggested reading:
  - **Re:DH1994**: Diaconis and Holmes (1994). Gray Codes for Randomization Procedures.

References for this lecture

W Chapter 3.8 (The bootstrap and the jackknife).

Li:C2016: Seiler (2016). Lecture Notes on Nonparametric Statistics - bootstrap example.

Li:H2004: Holmes (2004). Lecture Notes on Complete Enumeration.