Bootstrap

CSU, Chico Math 314

2018-11-06

outline

- Recap
- Sampling Distributions Revisited
  - Sample Median
- Bootstrap
  - Lite Example
  - Mechanics
- Example
- Pointers
- Reference Material
  - How to Use `boot`
  - Theory
- Homework
- References

outline

- Recap
- Sampling Distributions Revisited
  - Sample Median
- Bootstrap
  - Lite Example
  - Mechanics
- Example
- Pointers
- Reference Material
  - How to Use `boot`
  - Theory
- Homework
- References

recap, standard error

We can calculate the standard error of the sample mean analytically,

$$Var(\bar{X}) = \frac{\sigma^2}{n}.$$

The Central Limit Theorem allows us to easily estimate a population mean using this standard error.

outline

- Recap
- Sampling Distributions Revisited
    - Sample Median
- Bootstrap
    - Lite Example
    - Mechanics
- Example
- Pointers
- Reference Material
    - How to Use boot
    - Theory
- Homework
- References

Sampling Distribution Revisited

All statistics have a sampling distributions. The sample mean is relatively easy to calculate, and thus statistics (the discipline) most often estimates population means.

Sample Median

Assume we have $X_1, \ldots, X_n \sim_{iid} F$, with probability density function f. Define m such that $P(X \leq m) = 0.5$. It can be shown that the sampling distribution of the sample median $\tilde{X}$ is appropximately normal when the sample size is sufficiently large,

$$\frac{\tilde{X} - m}{4nf^2(m)} \dot{\sim} N(0, 1). \tag{1}$$

Sample Median, notes

Note that we want a statement like Equation (1) where we don't have to assume we know F/f – the Central Limit Theorem gives us this for the sample mean.

outline

- Recap
- Sampling Distributions Revisited
  - Sample Median
- Bootstrap
  - Lite Example
  - Mechanics
- Example
- Pointers
- Reference Material
  - How to Use `boot`
  - Theory
- Homework
- References

Bootstrap

"[The] bootstrap allows us to estimate the sampling distribution of a statistic empirically without making assumptions about the form of the population, and without deriving the sampling distribution explicitly" Fox and Weisberg [2010].

Some Fake Data, sample mean

Assume we know F. We'll generate some fake data from F, calculate the true standard error for the sample mean and compare it to the bootstraped standard error.

```
x <- rnorm(101, mean=0, sd=3)
3/sqrt(101) # true standard error

## [1] 0.2985112

# load bootstrap library
suppressMessages(library(boot))
b <- boot(x, sample_mean, R=1000) # run bootstrap
sd(b$t) # estimated standard error

## [1] 0.2827387
```

Bootstrap Mechanics, in words

Assume we have a random sample of size n. Bootstrapping re-samples, R times with replacement, n observations from our original sample and calculates the statistic of interest from each re-sample.

Bootstrap Mechanics, in math

Assume $X_1, \ldots, X_n \sim_{iid} F$, and that interest lies in the statistic $T = t(\mathbf{X})$ which estimates θ. Bootstrapping proceeds as follows:

1. Randomly select with replacement $X^*_{r1}, \ldots, X^*_{rn}$ from the original sample
2. Calculate $T^*_r = t(\mathbf{X}^*_r)$
3. Repeat steps 1-2 R times.

Bootstrap Estimates

The distribution of T_r^* about the original estimate T is analogous to the sampling distribution of the estimator T about the population parameter θ.

The population is to the sample
as the sample is to the bootstrap samples.

outline

- Recap
- Sampling Distributions Revisited
    - Sample Median
- Bootstrap
    - Lite Example
    - Mechanics
- Example
- Pointers
- Reference Material
    - How to Use `boot`
    - Theory
- Homework
- References

Carnivora Brain Weight, sample mean

Consider the data set `ape::carnivora`. Suppose we are interested in the mean brain weight of the order Carnivora.

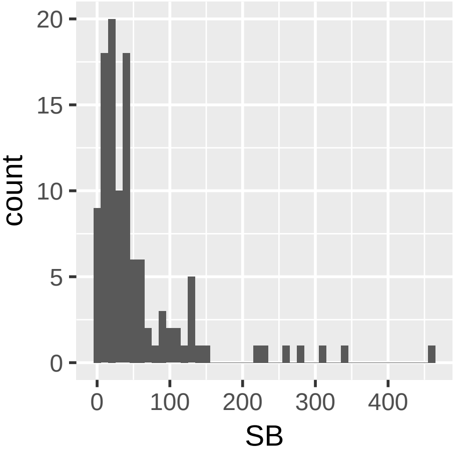
```r
suppressMessages({library(ape)
    data(carnivora)
    library(ggplot2)})
anyNA(carnivora$SB)

## [1] FALSE

p <- ggplot(data=carnivora, aes(x=SB)) +
    geom_histogram(binwidth=10)
```

Carnivora Brain Weight, sample mean

Plot the data!



Carnivora Brain Weight, sample mean

A 95% confidence interval using the proper standard error.

```
with(carnivora, {
    n <- length(SB)
    mean(SB) + qt(c(0.025, 0.975), n-1)*sd(SB)/sqrt(n)
})

## [1] 42.39558 70.46334
```

Carnivora Brain Weight, sample mean

A 95% confidence interval using a bootstrap estimated standard error

```
sample_mean <- function(d, i) {
    mean(d[i])
}
with(carnivora, {
    b <- boot(data=SB, statistic=sample_mean, R=2000)
    ci <- boot.ci(b, conf=0.95, type="norm")
    ci$normal
})

##      conf
## [1,] 0.95 42.98298 70.38129
```

Carnivora Brain Weight

But doesn’t the median seem more appropriate for these data?

Carnivora Brain Weight, sample median

A 95% confidence interval using the proper standard error.

```
# doh
```

Carnivora Brain Weight, sample median

A 95% confidence interval using a bootstrap estimated standard error

```
sample_median <- function(d, i) {
    median(d[i])
}
with(carnivora, {
    b <- boot(data=SB, statistic=sample_median, R=2000)
    ci <- boot.ci(b, conf=0.95, type="bca") # use this typ
    ci$bca[4:5] # bca is recommended
})

## [1] 24.3 39.3
```

outline

Recap

Sampling Distributions Revisited

- Sample Median

Bootstrap

- Lite Example
- Mechanics

Example

Pointers

Reference Material

- How to Use `boot`
- Theory

Homework

References

Bootstrap Pointers

Some notes on using the bootstrap.

1. Just like CLT, bootstrap relies on large sample sizes
2. Lots of variations on the bootstrap exist. In order, try
   2.1 `boot.ci(..., type=‘‘bca’’)`
   2.2 `boot.ci(..., type=‘‘basic’’, h=log, hdot=function(x) 1/x)`
   2.3 `boot.ci(..., type=‘‘perc’’)`
   2.4 `boot.ci(..., type=‘‘basic’’)`
3. Use $R = max(2000, 2 * n)$ replications, if not more.

outline

Recap

Sampling Distributions Revisited
- Sample Median

Bootstrap
- Lite Example
- Mechanics

Example

Pointers

Reference Material
- How to Use `boot`
- Theory

Homework

References

R's package boot

R's package boot is awesome. http://www.mayin.org/ajayshah/KB/R/documents/boot.html http://www.statmethods.net/advstats/bootstrapping.html
And of course

```
?boot # find arguments parallel, ncpus
```

Mathematics of Bootstrap

These are some of the more simple discussions of the bootstrap.

http://statweb.stanford.edu/~tibs/sta305files/FoxOnBootingRegInR.pdf

http://stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf

outline

- Recap
- Sampling Distributions Revisited
  - Sample Median
- Bootstrap
  - Lite Example
  - Mechanics
- Example
- Pointers
- Reference Material
  - How to Use `boot`
  - Theory
- Homework
- References

Bootstrap by Groups

Here are two ways to perform bootstrap by group, the first is more similar to what we've done in class. Study this gist alongside the lab tomorrow. There's a quiz on Friday about this gist. https://gist.github.com/roualdes/1de1c9a4a26581ba18a7ae9b96019970

outline

- Recap
- Sampling Distributions Revisited
    - Sample Median
- Bootstrap
    - Lite Example
    - Mechanics
- Example
- Pointers
- Reference Material
    - How to Use `boot`
    - Theory
- Homework
- References

references I

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

J. Fox and S. Weisberg. *An R Companion to Applied Regression*. SAGE Publications, 2010. ISBN 9781452235752. URL https://books.google.com/books?id=l9eiNeME8ukC.