Bootstrap Sampling: An Introduction

BIOS 6611

CU Anschutz

Week 6

1. Motivation

2. The Bootstrap

3. Evaluating Bootstrap Performance

Motivation

Sampling Distributions

In classical statistics, we often assume a distribution for the population, and use this distributional assumption to derive sampling distributions for a statistic.

Recall, the **sampling distribution** of a *statistic* is the probability distribution we would observe based on random sampling. It is an assumption of the theoretical, underlying distribution.

For example, we have discussed the following:

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right)$
- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
- $\frac{X-\mu}{\sigma} \sim N(0,1)$
- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

Unknown Distribution of the Sampling Statistic

What if the *underlying distribution is unknown* for a given statistic?

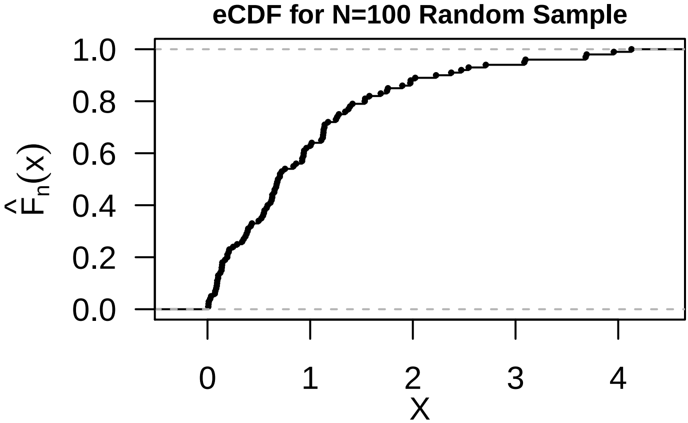
How could we estimate the sampling distribution for a statistic in this case?

It is even possible to estimate the distribution of a random variable?

The Empirical Cumulative Distribution Function

We can still describe the distribution of a sampling statistic. Assuming *random sampling*, our sample is a representation of the population.

The **empirical cumulative distribution function** is our best "estimate" of the population distribution:



The Plug-In Principle

If our goal is to estimate the sampling distribution for some statistic, we need to know:

- The underlying population (which may be unknown!)
- The sampling procedure (e.g., sampling with or without replacement)
- The statistic, e.g., $\bar{X}$

Using the plug-in principle is a natural and frequent approach in statistics. Think $\bar{X}$ for μ and s^2 for σ^2.

As another example, we know $\bar{X}$ calculated from i.i.d. observations has a standard error of $\frac{\sigma}{\sqrt{n}}$. When σ is unknown we plug in the estimator s to obtain our standard error estimate: $\frac{s}{\sqrt{n}}$.

The Empirical Set-Up

Let F and f denote the cdf and pdf for some unknown distribution with $x_1, x_2, \ldots, x_n$ a random sample from this distribution. Without making further assumptions about the distribution, we can use the empirical distribution:

$$\hat{F}(s) = \frac{1}{n} \left\{ \text{number of points} \leq s \right\}, \text{note this is a discrete function}$$

$$\hat{f}(s) = \frac{1}{n} \{ \text{number of points} = s \}$$

Recall, for a discrete R.V. the mean of $X \sim F$ is $E_F\left(X\right) = \mu_F = \sum_x x f(x)$. When we do not know F, we can plug in $\hat{F}$ (i.e., empirical distribution):

$$E_{\hat{F}}(X) = \mu_{\hat{F}} = \sum_x x \hat{f}(x) = \sum_{i=1}^{n} x_i \left(\frac{1}{n} \right) = \bar{x}$$

$$Var_{\hat{F}}(X) = \sigma^2_{\hat{F}} = E_{\hat{F}}[(X - \mu_{\hat{F}})^2] = \sum_{i=1}^{n} (x_i - \bar{x})^2 \left(\frac{1}{n} \right)$$

The Bootstrap

The Bootstrap Idea

The original sample approximates the population from which it was drawn. So resamples from this sample approximate what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, approximates the sampling distribution of the statistic, based on many samples.[1]

This resampling will create an empirical distribution which we use as the bootstrap distribution for our statistic of interest. Specifically, this repeated sampling is known as *Monte Carlo sampling*.

[1] from pg. 100 of our *Chihara and Hesterberg* textbook

Sketching the General Bootstrap Concept

Original
Sample

Bootstrap
Distribution

Sample
Statistic

Simple Example

To illustrate the general bootstrap idea in this slide deck, consider the sample $(1, 3, 4, 6)$. How many bootstrap samples are there?

What is the probability the mean is 1?

What is the probability the maximum is 6?

Verify Results in R

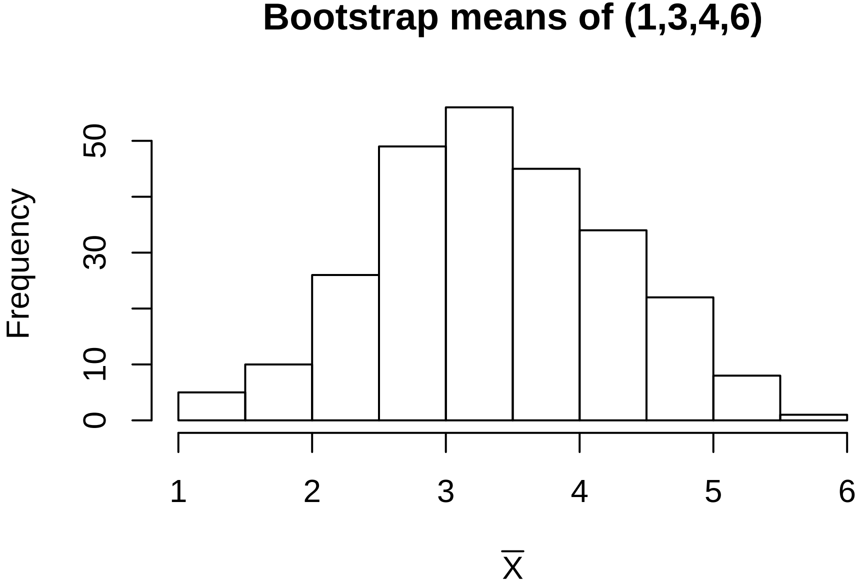
```r
dat <- c(1,3,4,6)
boot <- expand.grid(dat,dat,dat,dat) #all possible combos

#Pr(mean is 1 in bootstrap sample)
boot.mean <- apply(boot,MARGIN=1,mean)
mean(boot.mean==1); 1/256
```

```
## [1] 0.00390625

## [1] 0.00390625
```

```r
#Pr(max is 6 in bootstrap sample)
boot.max <- apply(boot, MARGIN=1,max)
mean(boot.max == 6); 1-(3^4)/(4^4)
```

```
## [1] 0.6835938

## [1] 0.6835938
```

Histogram of Sample Means



Comparing the Bootstrap and Original Samples

Our original sample of $(1, 3, 4, 6)$ has a mean and standard deviation of

```
mean(dat)
## [1] 3.5
sd(dat)
## [1] 2.081666
```

The bootstrap distribution of our sample mean has a mean and **bootstrap standard error** of

```
mean(boot.mean)
## [1] 3.5
sd(boot.mean) #bootstrap standard error
## [1] 0.9031535
```

The **bootstrap SE** of a statistic is the standard deviation of the bootstrap distribution of that statistic.

Bootstrap Sampling for a Single Population

When $n = 4$ we only had 256 possible combinations and could define the entire sampling space. As n increases we can't exhaustively explore this (e.g., $n = 30$ has $30^{30} = 2.0589 \times 10^{44}$ combinations).

Fortunately, we can leverage *sampling with replacement* to estimate our bootstrap distribution:

Given a sample of size n from a population,

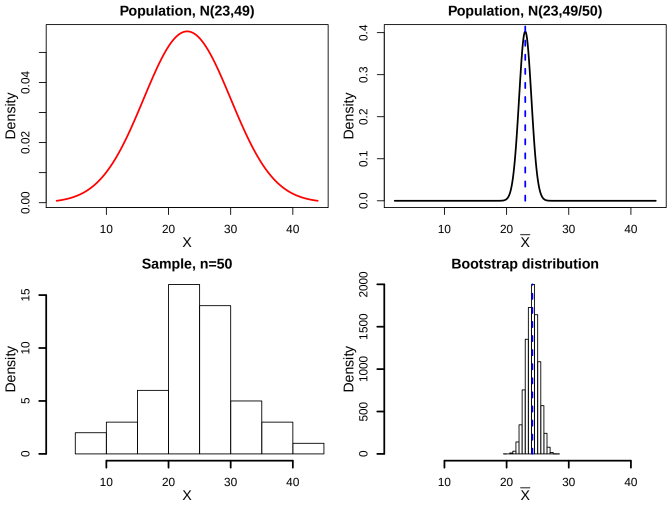
1. Draw a resample of size n with replacement from the sample. Compute a statistic that describes the sample, such as the sample mean.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Example with Known Sampling Distribution

Consider a sample of size 50 drawn from $N(23, 7^2)$. Let's estimate the bootstrap distribution for the sample mean:

```r
set.seed(515)
dat <- rnorm(n=50, mean=23, sd=7)

B <- 10000 # number of bootstraps
my.boot <- numeric(B) # initialize vector for results

for(i in 1:B){
  # resample with replacement:
  x <- sample(dat, size=50, replace=TRUE)

  # compute mean, store in my.boot:
  my.boot[i] <- mean(x)
}
```

Known Sampling Distribution Figures



- Shape and spread of the bootstrap distribution are comparable to that of the sampling distribution
- Centers of the bootstrap and sampling distributions are different (23 for population, 24.15 for bootstrap mean)
- Comparing the center of the bootstrap distribution to the observed statistic gives a measure of **bias**

Evaluating Bootstrap Performance

Bias

The bias of an estimator $\hat{\theta}$ is $Bias\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta$.

The bootstrap estimate of the bias is $Bias_{boot}\left[\hat{\theta}^*\right] = E\left[\hat{\theta}^*\right] - \hat{\theta}$, where $E[\hat{\theta}^*]$ is the mean of the bootstrap distribution and $\hat{\theta}$ is the sample estimate.

If an estimator, $\hat{\theta}$, tends to over or under estimate the true parameter value, θ, then it is biased. An estimator is unbiased if the bias is zero.

Bootstrap Performance

For most common estimators and under fairly general distribution assumptions:

- **Center:** the bootstrap distribution is NOT an accurate estimator for the center of the sampling distribution
- **Spread:** the spread of the bootstrap distribution does reflect the spread of the sampling distribution
- **Skewness:** the skewness of the bootstrap distribution does reflect the skewness of the sampling distribution
- **Bias:** the bootstrap can be used to estimate the bias of the sampling distribution

Thus, bootstrap sampling is useful for studying the *sampling behavior of estimators* (e.g. SE, skewness, bias) and *obtaining confidence intervals* for a parameter. It is not used to improve estimators.

How Many Bootstrap Samples?

For good accuracy, generally 10^4 or more.

"In large samples, clearly the bootstrap [is preferred]. In small samples, the classical procedure may be preferred. If the sample size is small, then skewness cannot be estimated accurately from the sample, and it may be better to assume skewness = 0 in spite of the bias, rather than to use an estimate that has high variability."[2]

Some parameters are not estimated well with bootstrap sampling, such as quantiles (e.g., median) or ones that depend heavily on a small number of observations from the larger sample.

[2] "Bootstrap" review paper by Tim Hesterberg (2011)