Lecture 3: The Bootstrap

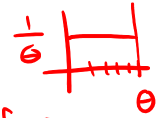
STAT GR5206 *Statistical Computing & Introduction to Data Science*

Gabriel Young
Columbia University

May 31, 2017

$X_1, \dots, X_n$ iid Unif$(0, \theta)$

$\bar{X}$



Sampling Dist'n – the probability dist'n of a statistic

$\max(X) = \theta$

The Bootstrap Principle

- If we could repeat an experiment over and over again, we could actually find a very goood approximation to the sampling distribution.
- Grocery example: If I had 1000 years of data, run the regression model on each year to see how estimates change.

The Bootstrap Principle

- If we could repeat an experiment over and over again, we could actually find a very goood approximation to the sampling distribution.
- Grocery example: If I had 1000 years of data, run the regression model on each year to see how estimates change.
- Often too expensive or time-consuming.
- Bradley Efron's Idea: Use computers to **simulate** replication.
- Instead of repeatedly obtaining new, independent datasets from the *population*, we repeatedly obtain datasets from the *sample* itself, the original dataset. 1 dataset

Bootstrap Methods

"Pull yourself up by your bootstraps!"

To get a bootstrap estimate,

1. Resample from the original data n times *with replacement* (note an original data observation could be in the new sample more than once),
2. Use the new dataset to compute a bootstrap estimate,
3. Repeat this to create B new datasets, and B new estimates.

$n=4$

Data	Data 1	Data 2	Data 3	. . . Data B
1	7	3	1	
7	7	7	1	
3	2	2	1	
2	1	1	7	
$\bar{X} = \frac{(1+7+3+2)}{4}$	$\bar{X}^{(1)}$	$\bar{X}^{(2)}$	$\bar{X}^{(4)}$	$\bar{X}^{(B)}$

Bootstrap Methods

Formally, you have original data $(x_i)_{i=1}^n$ and you are interested in estimating a population parameter Θ from the data. Label the estimate $\hat{\Theta}$. $\bar{x}$

Procedure

new bootstrap sample

1. For $b = 1, \ldots, B$,
   - Create a new dataset $\mathcal{B}_b = (x_i^{(b)})_{i=1}^n$ by sampling from original dataset *with replacement*. (sample() does this)
   - Use the new dataset to find an estimate $\hat{\Theta}^{(b)}$.

Bootstrap Methods

Formally, you have original data $(x_i)_{i=1}^n$ and you are interested in estimating a papulation parameter Θ from the data. Label the estimate $\hat{\Theta}$.

Procedure

1. For $b = 1, \ldots, B$,
   - Create a new dataset $\mathcal{B}_b = (x_i^{(b)})_{i=1}^n$ by sampling from original dataset *with replacement*.
   - Use the new dataset to find an estimate $\hat{\Theta}^{(b)}$.
2. The collection $(\hat{\Theta}^{(b)} - \hat{\Theta})_{b=1}^B$ estimates the sampling distribution of $\hat{\Theta} - \Theta$.

mean of origin

"truth" real Θ

mean of original

not random

distributional property

histograms will look similar!

Standard error

standard deviation of an estimator

(q_L, q_U) $\quad (\hat{\Theta}^{(b)} - \hat{\Theta}) \overset{D}{\approx} \hat{\Theta} - \Theta$ $\quad SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

Example: Gaussian Random Variables

You sample $n = 100$ data points, $x_1, \ldots, x_{100} \sim N(\mu, 1)$. (Recall, Lab 1.)

```
> n <- 100
> vec <- rnorm(n, mean = mu)
> head(vec)
```

```
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
[6] -0.8204684
```

- What's a good estimator for μ?

Example: Gaussian Random Variables

You sample $n = 100$ data points, $x_1, \ldots, x_{100} \sim N(\mu, 1)$. (Recall, Lab 1.)

```
> n <- 100
> vec <- rnorm(n, mean = mu)
> head(vec)
```

```
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
[6] -0.8204684
```

- What's a good estimator for μ?

```
> mean(vec)
```

```
[1] 0.1088874
```

Set $\hat{\mu} = 0.11$. Recall, $\hat{\mu} \sim N(\mu, 1/100)$.

Example: Gaussian Random Variables

You sample $n = 100$ data points, $x_1, \ldots, x_{100} \sim N(\mu, 1)$. (Recall, Lab 1.)

```
> n <- 100
> vec <- rnorm(n, mean = mu)
> head(vec)
```

```
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
[6] -0.8204684
```

- What's a good estimator for μ?

```
> mean(vec)
```

```
[1] 0.1088874
```

Set $\hat{\mu} = 0.11$. Recall, $\hat{\mu} \sim N(\mu, 1/100)$.

- How can we estimate $Var(\hat{\mu})$?

Example: Gaussian Random Variables

We'll use the bootstrap to estimate the variance! For $b = 1 : B$,

- Resample $x_1, \ldots, x_{100}$ *with replacement* to get $x_1^{(b)}, \ldots, x_{100}^{(b)}$.
- Compute $\hat{\mu}^{(b)} = \frac{1}{100} \sum_{i=1}^{100} x_i^{(b)}$.

if $X_1, \ldots, X_n \overset{iid}{\sim}$ normal dist

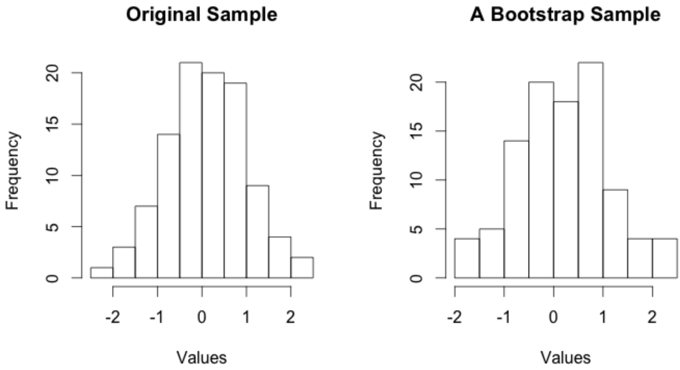
mean $= \mu$, var $= \sigma^2$

$\bar{X} \overset{D}{=} N(\mu, \frac{\sigma^2}{n})$

```
> B <- 1000
> estimates <- vector(length = B)
> for (b in 1:B) {
+   new_sample <- sample(vec, size = n, replace = TRUE)
+   estimates[b] <- mean(new_sample)
+ }
> head(estimates)
```

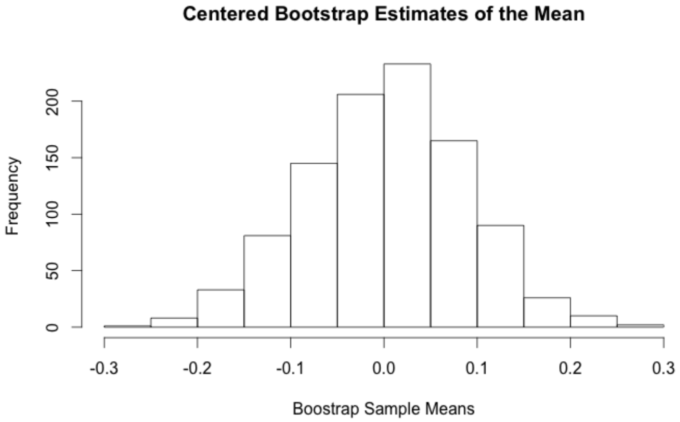
```
[1] 0.12250487 0.10894538 0.21117547 0.05405239 0.16694190
[6] 0.13804749
```

A histogram of the original sample and a histogram of a single resampled bootstrap sample.



Example: Gaussian Random Variables

The **Bootstrap Distribution of the Statistic**. Recall $(\hat{\mu}^{(b)} - \hat{\mu})_{b=1}^{B}$ approximates the sampling distribution of $\hat{\mu} - \mu$.



Example: Gaussian Random Variables

We'll use the bootstrap to estimate the variance!

Estimating the Variance

```
> var(estimates)
```

```
[1] 0.007380355
```

True variance: $Var(\hat{\mu}) = \frac{\sigma^2}{n} = \frac{1}{100} = 0.01$.

Bootstrapping Summary

Bootstrapping is very flexible!

- Bootstrapping gives you a distribution over estimators.
- This can be used to:
  - Approximate more complicated metrics (medians, quantiles, etc.).
  - Approximate distributional properties.
  - Create confidence intervals.
- By resampling $(x_i, y_i)_{i=1}^n$ pairs, we could create bootstrap estimators for linear model regression parameters.

Optional Reading

- Chapter 6 (The Bootstrap) in Advanced Data Analysis from an Elementary Point of View.