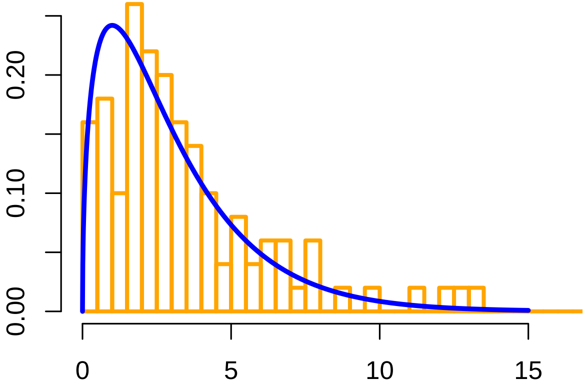
Bootstrapping
18.05 Spring 2022



Announcements

Announcements

- R Quiz tomorrow
- Pset 11 (not to be turned in) on confidence intervals is on MITx week 14.
- Will post R Studio 10 (not to be turned in) on the bootstrap.

Agenda

- Bootstrap terminology
- Bootstrap principle
- Empirical bootstrap (also called nonparametric bootstrap)
- Parametric bootstrap

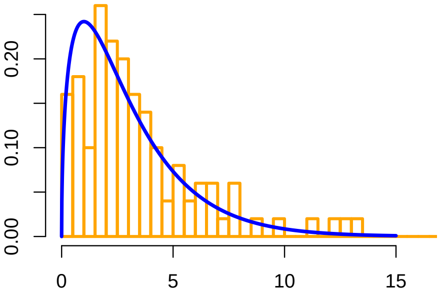
Empirical distribution of data

This is the actual distribution of the data –not the underlying distribution from which it was drawn.

Example 1. Data: 1, 2, 2, 3, 8, 8, 8. (Assume independently drawn)

x^*	1	2	3	8
$p^*(x^*)$	$1/7$	$2/7$	$1/7$	$3/7$

Example 2.



The true and empirical distribution are approximately equal.

Resampling

- Sample (size 6): 1 2 1 5 1 12
- Resample (size m): Randomly choose m numbers with replacement from the original sample.
- Resample probabilities = empirical distribution: $P(1) = 1/2$, $P(2) = 1/6$ etc.
- E.g. resample (size 10): 5 1 1 1 12 1 2 1 1 5
- A bootstrap (re)sample is always the same size as the original sample:
- Bootstrap sample (size 6): 5 1 1 1 12 1

Bootstrap principle for the mean

Setup

- Distribution F with mean μ.
- Data $x_1, x_2, \ldots, x_n \sim F$, with mean $\overline{x}$
- $F^* =$ empirical distribution (resampling distribution) of the data.
- $x_1^*, x_2^*, \ldots, x_n^*$ resample same size with mean $\overline{x}^*$.

Bootstrap Principle: (**really holds for any statistic**)

1. $F^* \approx F$
2. $\overline{x}^* \approx \overline{x} \approx \mu$
3. The variation of $\overline{x}$ is approximated by the variation of $\overline{x}^*$

Key: We can resample as many times as we want to get an accurate estimate of the variation of $\overline{x}^*$

Empirical percentile bootstrap confidence intervals

Use the data to estimate the variation of estimates based on the data!

- Data: $x_1, \ldots, x_n$ drawn from a distribution F.
- Estimate a feature θ of F by a statistic $\hat{\theta}$.
- Generate many bootstrap samples $x_1^*, \ldots, x_n^*$.
- Compute the statistic θ^* for each bootstrap sample.
- The $1-\alpha$ percentile bootstrap confidence interval is

$$[\theta^*_{\alpha/2},\, \theta^*_{1-\alpha/2}],$$

  where $\theta^*_{\alpha/2}$ is the $\alpha/2$ **quantile** for θ^*.

Principle Distibution of $\theta^* \approx$ distribution of $\hat{\theta}$.

Empirical basic bootstrap confidence intervals

Won't do this in class.

Use the data to estimate the variation of estimates based on the data!

- Data: $x_1, \ldots, x_n$ drawn from a distribution F.
- Estimate a feature θ of F by a statistic $\hat{\theta}$.
- Generate many bootstrap samples $x_1^*, \ldots, x_n^*$.
- Compute the statistic θ^* for each bootstrap sample.
- Compute the bootstrap difference

$$\delta^* = \theta^* - \hat{\theta}.$$

- The $1 - \alpha$ basic bootstrap confidence interval is

$$[\hat{\theta} - \delta^*_{\alpha/2}, \hat{\theta} - \delta^*_{1-\alpha/2}],$$

  where $\delta^*_{\alpha/2}$ is the $\alpha/2$ **critical value** for δ^*.

  Principle $\delta^* = \theta^* - \hat{\theta} \approx \hat{\theta} - \theta = \delta$

Which bootstrap is best?

The percentile bootstrap is a little simpler and, empirically, in many settings it it appears to be slightly better than the basic bootstrap.

But, there are more sophisticated bootstrap methods that are generally better than both.

The principles are the same, but some tweaks improve performance.

Concept question: which stat is easiest

Consider finding bootstrap confidence intervals for

I. the mean **II.** the median **III.** 47th percentile.

Which is easiest to find?

(a) I (b) II (c) III (d) I and II

(e) II and III (f) I and III (g) I and II and III

Board question: empirical bootstrap

Data: 3 8 1 8 3 3

Bootstrap samples (each column is one bootstrap trial):

8	8	1	8	3	8	3	1
1	3	3	1	3	8	3	3
3	1	1	8	1	3	3	8
8	1	3	1	3	3	8	8
3	3	1	8	8	3	8	3
3	8	8	3	8	3	1	1

(a) Compute a bootstrap 80% percentile confidence interval for the mean.

(b) Compute a bootstrap 80% percentile confidence interval for the median.

Percentile empirical bootstrapping in R

```
x = c(30,37,36,43,42,43,43,46,41,42)   # original sample
n = length(x)       # sample size
xbar = mean(x)      # sample mean
n_boot = 5000        # number of bootstrap samples to use

# Generate nboot empirical samples of size n and organize
in a matrix
tmp_data = sample(x, n*n_boot, replace=TRUE)
bootstrap_sample = matrix(tmp_data, nrow=n, ncol=n_boot)

# Compute bootstrap means xbar*
xbar_star = colMeans(bootstrap_sample)

# Find the 0.1 and 0.9 quantiles and make the bootstrap 80%
confidence interval
ci = quantile(xbar_star, c(0.1, 0.9))
```

Parametric bootstrapping

Use the estimated parameter to estimate the variation of estimates of the parameter!

- Data: $x_1, \ldots, x_n$ drawn from a parametric distribution $F(\theta)$.
- Estimate θ by a statistic $\hat{\theta}$.
- **Generate many bootstrap samples from $F(\hat{\theta})$.**
- Compute the statistic θ^* for each bootstrap sample.
- Compute the bootstrap difference

$$\delta^* = \theta^* - \hat{\theta}.$$

- Use the critical values of δ^* to approximate those of

$$\delta = \hat{\theta} - \theta.$$

- Set a confidence interval $[\hat{\theta} - \delta^*_{\alpha/2}, \hat{\theta} - \delta^*_{1-\alpha/2}]$

Parametric sampling in R

```
# Data from binomial(15, θ) for an unknown θ
x = c(3, 5, 7, 9, 11, 13)
binom_size = 15     # known size of binomial
n = length(x)      # sample size
theta_hat = mean(x)/binom_size     # MLE for θ
n_boot = 5000       # number of bootstrap samples to use

# nboot parametric samples of size n; organize in a matrix
tmp_data = rbinom(n*n_boot, binom_size, theta_hat)
bootstrap_sample = matrix(tmp_data, nrow=n, ncol=n_boot)

# Compute bootstrap means theta_hat* and differences delta*
theta_hat_star = colMeans(bootstrap_sample)/binom_size
delta_star = theta_hat_star - theta_hat

# Find quantiles and make the bootstrap confidence interval
d = quantile(delta_star, c(0.1, 0.9))
ci = theta_hat - c(d[2], d[1])
```

Board question

Data is taken from a Binomial(8, θ) distribution. After 6 trials, the results are

$$6 \quad 5 \quad 5 \quad 5 \quad 7 \quad 4$$

(a) Estimate θ.

(b) Write out the R code to generate data of 100 parametric bootstrap samples and compute an 80% confidence interval for θ.

(Try this without looking at your notes.)