Agenda

- Bootstrap terminology
- Bootstrap principle
- Empirical bootstrap
- Parametric bootstrap
Empirical distribution of data

Data: $x_1, x_2, \ldots, x_n$ (independent)

**Example 1.** Data: 1, 2, 2, 3, 8, 8, 8.

<table>
<thead>
<tr>
<th>$x^*$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^<em>(x^</em>)$</td>
<td>1/7</td>
<td>2/7</td>
<td>1/7</td>
<td>3/7</td>
</tr>
</tbody>
</table>

**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 \( m \) samples with replacement from the original sample.

- Resample probabilities = empirical distribution: 
  \[ P(1) = \frac{1}{2}, \ P(2) = \frac{1}{6} \text{ etc.} \]

- 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

- Data \( x_1, x_2, \ldots, x_n \sim F \) with true mean \( \mu \).
- \( F^* = \) empirical distribution (resampling distribution).
- \( x_1^*, x_2^*, \ldots, x_n^* \) resample same size data

Bootstrap Principle: (really holds for any statistic)

1. \( F^* \approx F \) computed from the resample.
2. \( \delta^* = \bar{x}^* - \bar{x} \approx \bar{x} - \mu = \) variation of \( \bar{x} \)

Critical values: \( \delta_{1-\alpha/2}^* \leq \bar{x}^* - \bar{x} \leq \delta_{\alpha/2}^* \)

then \( \delta_{1-\alpha/2}^* \leq \bar{x} - \mu \leq \delta_{\alpha/2}^* \) so

\[
\bar{x} - \delta_{\alpha/2}^* \leq \mu \leq \bar{x} - \delta_{1-\alpha/2}^*
\]
Empirical 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 $\theta$ of $F$ by a statistic $\hat{\theta}$.
- Generate many bootstrap samples $x_1^*, \ldots, x_n^*$.
- Compute the statistic $\theta^*$ for each bootstrap sample.
- Compute the **bootstrap difference**

\[ \delta^* = \theta^* - \hat{\theta}. \]

- Use the quantiles of $\delta^*$ to approximate quantiles of

\[ \delta = \hat{\theta} - \theta \]

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

(By $\delta_{\alpha/2}$ we mean the $\alpha/2$ quantile.)
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

answer: G. The program is essentially the same for all three statistics. All that needs to change is the code for computing the specific statistic.
Board question

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

Compute a bootstrap 80% confidence interval for the mean.

Compute a bootstrap 80% confidence interval for the median.
Solution: mean

\[ \bar{x} = 4.33 \]

\[ \bar{x}^*: \quad 4.33, 4.00, 2.83, 4.83, 4.33, 4.67, 4.33, 4.00 \]

\[ \delta^*: \quad 0.00, -0.33, -1.50, 0.50, 0.00, 0.33, 0.00, -0.33 \]

Sorted \[ \delta^*: \quad -1.50, -0.33, -0.33, 0.00, 0.00, 0.00, 0.33, 0.50 \]

So, \[ \delta^*_{0.9} = -1.50, \quad \delta^*_{0.1} = 0.37. \]

(For \[ \delta^*_{0.1} \] we interpolated between the top two values —there are other reasonable choices. In R see the \texttt{quantile()} function.)

80% bootstrap CI for mean: \[ [\bar{x} - 0.37, \bar{x} + 1.50] = [3.97, 5.83] \]
Solution: median

\[ x_{0.5} = \text{median}(x) = 3 \]

\[ x^{*}_{0.5}: \quad 3.0, 3.0, 2.0, 5.5, 3.0, 3.0, 3.0, 3.0 \]

\[ \delta^*: \quad 0.0, 0.0, -1.0, 2.5, 0.0, 0.0, 0.0, 0.0 \]

Sorted

\[ \delta^*: \quad -1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.5 \]

So, \( \delta^*_{0.9} = -1.0, \quad \delta^*_{0.1} = 0.5. \)

(For \( \delta^*_{0.1} \) we interpolated between the top two values – there are other reasonable choices. In R see the \text{quantile()} \ function.)

80% bootstrap CI for median: \[ [\bar{x} - 0.5, \bar{x} + 1.0] = [2.5, 4.0] \]
Empirical bootstrapping in R

```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
nboot = 5000  # number of bootstrap samples to use

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

# Compute bootstrap means xbar* and differences delta*
xbarstar = colMeans(bootstrapsample)
deltastar = xbarstar - xbar

# Find the .1 and .9 quantiles and make
# the bootstrap 80% confidence interval
d = quantile(deltastar, c(.1, .9))
ci = xbar - c(d[2], d[1])
```

```text
January 1, 2017     11 / 16
```
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 $\theta$ by a statistic $\hat{\theta}$.
- **Generate many bootstrap samples from** $F(\hat{\theta})$.
- Compute the statistic $\theta^*$ for each bootstrap sample.
- Compute the bootstrap difference

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

- Use the quantiles of $\delta^*$ to approximate quantiles of

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

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

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

# nboot parametric samples of size n; organize in a matrix
tmpdata = rbinom(n*nboot, binomSize, thetahat)
bootstrapsample = matrix(tmpdata, nrow=n, ncol=nboot)

# Compute bootstrap means thetahat* and differences delta*
thetahatstar = colMeans(bootstrapsample)/binomSize
deltastar = thetahatstar - thetahat

# Find quantiles and make the bootstrap confidence interval
d = quantile(deltastar, c(.1,.9))
CI = thetahat - c(d[2], d[1])
Data: 6 5 5 5 7 4 \sim \text{binomial}(8, \theta)

1. Estimate \theta.

2. Write out the R code to generate data of 100 parametric bootstrap samples and compute an 80% confidence interval for \theta.

(Try this without looking at your notes. We’ll show the previous slide at the end)
Preview of linear regression

- Fit lines or polynomials to bivariate data
- Model: \( y = f(x) + E \)
  \( f(x) \) function, \( E \) random error.
- Example: \( y = ax + b + E \)
- Example: \( y = ax^2 + bx + c + E \)
- Example: \( y = e^{ax+b+E} \) (Compute with \( \ln(y) = ax + b + E \).)
18.05 Introduction to Probability and Statistics
Spring 2014

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.