Bootstrapping

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

Agenda

- Empirical bootstrap
- Parametric bootstrap

Resampling

Sample (size 6): 1 2 1 5 1 12

Resample by choosing k uniformly between 1 and 6 and taking the k^{th} element.

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

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 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*

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

- Use the quantiles of δ^* to approximate quantiles of

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

- Set a confidence interval $[\hat{\theta} - \delta^*_{1-\alpha/2}, \hat{\theta} - \delta^*_{\alpha/2}]$
  ($\delta_{\alpha/2}$ is the $\alpha/2$ *quantile*.)

Concept question

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

Data: 3 8 1 8 3 3

Bootstrap samples (each column is one bootstrap trial):

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

Compute a 75% confidence interval for the mean.

Compute a 75% confidence interval for the median.

Solution

$\bar{x} = 4.33$

$\bar{x}^*$:

```
3.17 3.17 4.67 5.50 3.17 2.67 3.50 2.67
```

δ^*:

```
-1.17 -1.17 0.33 1.17 -1.17 -1.67 -0.83 -1.67
```

So, $\delta^*_{.125} = -1.67$, $\delta^*_{.875} = 0.75$. (For $\delta^*_{.875}$ we took the average of the top two values –there are other reasonable choices.)

Sort:

```
-1.67 -1.67 -1.17 -1.17 -1.17 -0.83 0.33 1.17
```

75% CI: $[\bar{x} - 0.75,\ \bar{x} + 1.67] = [3.58\ 6.00]$

Parametric bootstrapping

Use the data to estimate a parameter. Use the parameter to estimate the variation of the parameter estimate.

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

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

- Use quantiles of δ^* to approximate quantiles of

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

- Use the quantiles to define a confidence interval.

Parametric sampling in R

```
# an arbitrary array from binomial(15, theta) for an
unknown theta
x = c(3, 5, 7, 9, 11, 13)
binomSize = 15
n = length(x)
thetaHat = mean(x)/binomSize
parametricSample = rbinom(n, binomSize, thetaHat)
print(parametricSample)
```

Board question

Data: `6 5 5 5 7 4` $\sim$ binomial(8,θ)

1. Estimate θ.

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

(You will want to make use of the R function `quantile()`.)

Preview of linear regression

- Fit lines or polynomials to bivariate data
- Model: $y = f(x) + E$
  $f(x)$ function, E random error.
  item Example: $y = ax + b + E$
- Example $y = ax^2 + bx + c + E$
- Example $y = \mathrm{e}^{ax+b+E}$