Lecture 5: The Jackknife and Bootstrap

Pratheepa Jeganathan

04/12/2019

Recall

- Testing location parameter.
- Assumptions on F, either continuous cdf or symmetric continuous cdf.
- Estimators of location parameters.
- Distribution-free confidence intervals for location parameters.
- Measures of robustness of estimators
  - robustness to the observed data (sensitivity, breakdown point).
  - robustness to the theoretical distribution underlying the data (influence functions).
- Location parameters as statistical functionals.
- Approximates the standard error of a plug-in estimator using influence function.

The jackknife

- Asymptotic connection between jackknife estimate of variance of an estimator and influence function.
  - Influence function with $\epsilon = \frac{-1}{n-1}$, and $\hat{F}$ provides the jackknife estimate of variance as $n \to \infty$

$$\begin{aligned}\hat{L}(z) &= \lim_{n\to\infty} \frac{T\left(\left(1 - \frac{-1}{n-1}\right)\hat{F} + \frac{-1}{n-1}\delta_z\right) - T\left(\hat{F}\right)}{\frac{-1}{n-1}} \\ &= \lim_{n\to\infty} \frac{T\left(\hat{F}_{(i)}\right) - T\left(\hat{F}\right)}{\frac{-1}{n-1}},\end{aligned} \tag{1}$$

    where $\hat{F}_{(i)}$ is the empirical cdf with i-th observation removed.
  - $\tau^2 = \int L(z)^2 \, dF(z)$.
  - $\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{L}(z)\right)^2$ provides $\mathbb{V}\left(T\left(\hat{F}\right)\right) = \frac{\hat{\tau}^2}{n}$.

The jackknife

- Suppose $\boldsymbol{X} = (X_1, \cdots, X_n)^T \sim F$ a random sample.
  - $\theta = T(F)$, a parameter of interest.
  - $\hat{\theta} = s(\boldsymbol{x})$, an estimate from $\boldsymbol{x} = (x_1, \cdots, x_n)^T$, the observed data.
  - $s(\boldsymbol{x})$ may not be a plug-estimate $T\left(\hat{F}\right)$.
- The jackknife method can be used for estimating the bias and standard error of $\hat{\theta} = s(\boldsymbol{x})$.
  - Let $\boldsymbol{X}_{(i)}$ be a random sample with i-th observation removed.
  - Let $\hat{\theta}_{(i)} = s\left(\boldsymbol{x}_{(i)}\right)$ be an estimate of θ with i-th observation removed.
  - Define $\hat{\theta}_{(\cdot)} = \dfrac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n}$
  - The jackknife estimate of $\widehat{\text{bias}}_{\text{jack}} = (n-1)\left(\hat{\theta}_{(\cdot)} - \hat{\theta}\right)$
  - The jackknife estimate of standard error $\widehat{\text{se}}_{\text{jack}} = \left[\dfrac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right]^{1/2}$

The jackknife

- Jackknife is easier to compute (for $n \approx 200$) than the bootstrap to estimating standard error and bias of an estimator but is less efficient than the bootstrap.
- Jackknife estimate may be inconsistent (example, jackknife estimate of variance of median).

The bootstrap

- Computer-based resampling procedure to access the statistical accuracy.
- Computes standard error or bias of a statistic or sampling distribution of a statistic or confidence intervals of parameters.
- No need a mathematical expression for the statistical accuracy such as bias or standard error.

The bootstrap

- $\boldsymbol{X} = (X_1, \cdots, X_n)^T \sim F$.
- $\theta = T(F)$, a parameter of interest.
- $\hat{\theta} = s(\boldsymbol{x})$, an estimate from $\boldsymbol{x} = (x_1, \cdots, x_n)^T$.
- Let $\hat{F}$ be the empirical distribution of the observed values x_i, $\hat{F}(t) = \frac{\sum_{i=1}^{n} I(x_i \leq t)}{n}$.
- A bootstrap sample $\boldsymbol{X}^* = (X_1^*, \cdots, X_n^*)$, a random sample of size n drawn with replacement from $\hat{F}$ (total of distinct bootstrap samples $\binom{2n-1}{n-1}$) see illustration.
- A bootstrap replication of $\hat{\theta}$ is $\hat{\theta}^* = s(\boldsymbol{x}^*)$.
- Bootstrap replicates $\hat{\theta}^{*1}, \cdots, \hat{\theta}^{*R}$, where R is the number of bootstrap samples.

The bootstrap (illustration)

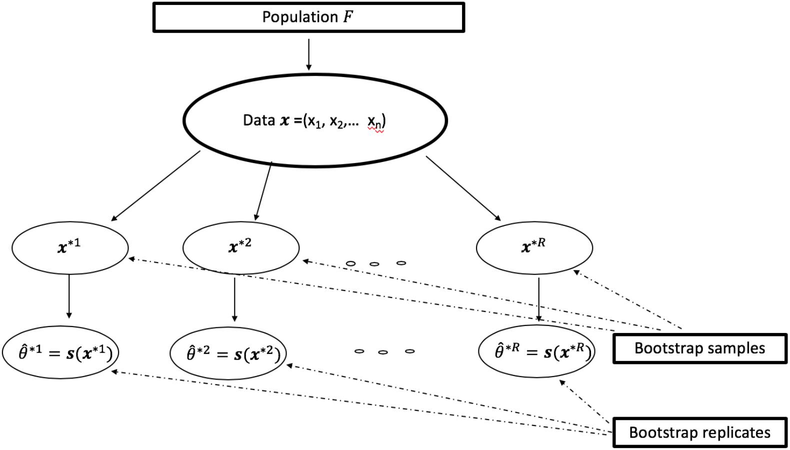


Figure 1: Bootstrap method

The bootstrap method for estimating standard error

- Draw R bootstrap samples $\boldsymbol{X}^{*1}, \cdots, \boldsymbol{X}^{*R}$ each with n values with replacement.
- Evaluate bootstrap replicate $\hat{\theta}^{*r} = s\left(\boldsymbol{x}^{*r}\right)$.
- Estimate the standard error $\mathrm{se}\left(\hat{\theta}\right)$

$$\widehat{\mathrm{se}}_{\mathrm{boot}}\left(\hat{\theta}\right) = \left[\frac{\sum_{r=1}^{R}\left(\hat{\theta}^{*r} - \hat{\theta}^{*}(\cdot)\right)^2}{R-1}\right]^{1/2},$$

where $\hat{\theta}^{*}(\cdot) = \dfrac{\sum_{r=1}^{R}\hat{\theta}^{*r}}{R}$.

- How large should be R? The rules of thumb: $R = 50$ is often enough to give a good estiamte of $\mathrm{se}\left(\hat{\theta}\right)$ (much larger values of R are required for confidence intervals).

Bootstrap percentile confidence interval

- Order bootstrap replicates $\hat{\theta}^*_{(1)}, \cdots, \hat{\theta}^*_{(R)}$.
- Let $m = [\alpha/2 \times R]$, $[u]$ is the largest integer less than or equal to u.
- Approximate $(1-\alpha)\,100\%$ confidence interval for θ is $\left(\hat{\theta}^*_{(m)}, \hat{\theta}^*_{(R-m)}\right)$.
- Choose $R = 1000$ or larger than 1000.

One-sample location problem

- $\boldsymbol{X} = (X_1, \cdots, X_n)^T \sim F$.
- $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ - observed random sample.
- Hypothesis, $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_a : \theta > \theta_0$.
- Let $T(\boldsymbol{X})$ be a test statistic (need not be an estimate of a parameter θ). Let $T(\boldsymbol{x}) = \hat{\theta} = \bar{\boldsymbol{x}}$.
- We need to take bootstrap samples from $\{x_1 - \hat{\theta} + \theta_0, \cdots, x_n - \hat{\theta} + \theta_0\}$.
- Then

$$\text{P-value} = \frac{\#\{\hat{\theta}^{*r} \geq \hat{\theta}\}}{R}.$$

Assessing the error in bootstrap estimates

- Bootstrap estimates are not exact (nearly unbiased but can have substantial variance).
- Two sources of variability
  - Sampling variability: we have only a sample of size n rather than the entire population.
  - Bootstrap resampling variability: we only take R bootstrap samples rather than total of $\binom{2n-1}{n-1}$ distinct bootstrap samples.

Two sources of variability (illustration)

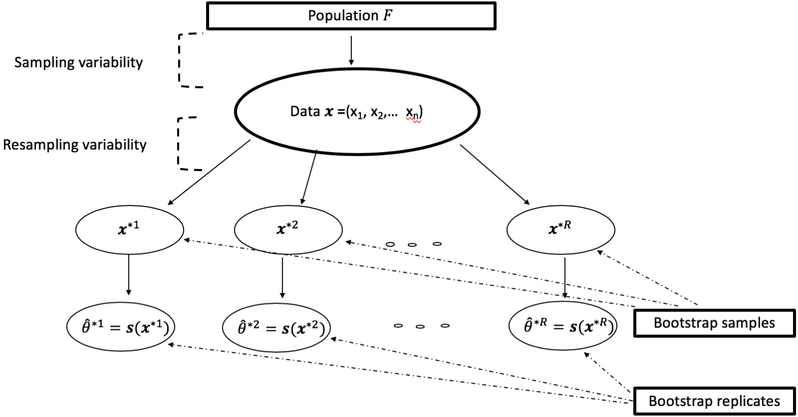


Figure 2: The sampling and resampling variability

Example (Hypothesis)

- Hypothesis: Small aspirin doses would prevent heart attacks in healthy middle-aged men.

Example (Experiment)

- Controlled, randomized, double-blinded study (both physicians and subjects were blinded the assignment).
- One half of the subjects received aspirin and other half received placebo.
- Define $X_i = 1$ if heart attack is observed and $X_i = 0$ otherwise.

```r
labels = c("nattacks","nsubjects")
aspirin = c(104, 11037)
placebo = c(189, 11034)
data = data.frame(aspirin, placebo)
rownames(data) = labels
data
```

```
##           aspirin placebo
## nattacks      104     189
## nsubjects   11037   11034
```

Example (Estimation)

- $\hat{\theta} =$ Ratio of rate of heart attacks in the aspirin group to placebo group.
- $\text{H}_0 : \theta = 1$ versus $\text{H}_a : \theta < 1$

```
ratio = function(r) {r[1]/r[2]}
theta.hat = ratio(data$aspirin)/ratio(data$placebo)
theta.hat
```

```
## [1] 0.550115
```

This indicates that in this sample the aspirin-takers only have 55% as many heart attacks as placebo-takers.

Example

- What is the uncertainty of $\hat{\theta}$?
- Use bootstrap to access the statistical accuracy.

```
sample.aspirin = c(rep(1,
  times = data["nattacks","aspirin"]),
  rep(0,
    times = (data["nsubjects","aspirin"] -
      data["nattacks","aspirin"])))

table(sample.aspirin)
```

```
## sample.aspirin
##     0     1
## 10933   104
```

Example

```r
sample.placebo = c(rep(1,
  times = data["nattacks","placebo"]),
  rep(0,
    times = (data["nsubjects","placebo"] -
        data["nattacks","placebo"])))

table(sample.placebo)
```

```
## sample.placebo
##     0     1
## 10845   189
```

Example (bootstrap samples)

- Draw bootstrap samples and compute bootstrap replicates.

```
bootstrap.sample = function() {
  boot.sam.aspirin = sample(sample.aspirin,
    replace = TRUE)
  boot.sam.placebo = sample(sample.placebo,
    replace = TRUE)
  h.rate.aspirin = sum(boot.sam.aspirin)/length(boot.sam.as
  h.rate.placebo = sum(boot.sam.placebo)/length(boot.sam.pl
  return(h.rate.aspirin/h.rate.placebo)
}
```

Example (bootstrap replicates)

- $R = 1000$ bootstrap samples and $\hat{\theta}^{*r}$.

```
R = 10000
theta.boot = replicate(R, bootstrap.sample())
```

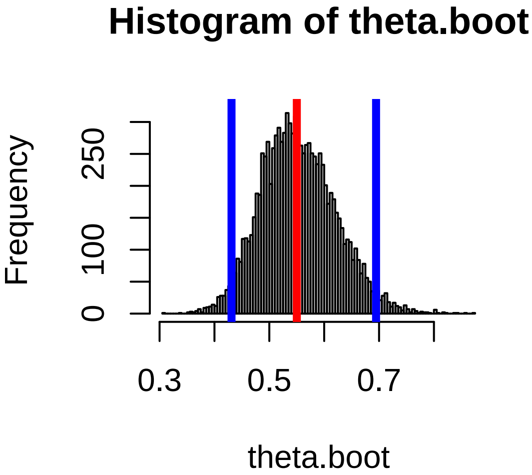
Example

```
hist(theta.boot, breaks=100)

# observed value of ratio of
# heart attack rates
abline(v=theta.hat, col = "red",lwd = 4)

# 95% bootstrap percentile confidence
# interval for true ratio of heart attack rates
theta.lower = sort(theta.boot)[R*.025]
theta.upper = sort(theta.boot)[R*.975]
abline(v=c(theta.lower, theta.upper),
```

Example


Histogram of theta.boot
Frequency
0
100
250
0.3
0.5
0.7
theta.boot

Example

- Confidence interval for true value θ.
- The true value of θ lies in the interval

```
quantile(theta.boot, probs = c(.025, .975))
```

```
##      2.5%     97.5%
## 0.4312508 0.6945525
```

with 95% confidence.

- We can conclude that aspirin is significantly beneficial for preventing heart attacks in healthy middle-aged men.

References

References for this lecture

W Chapter 3 (The bootstrap and the jackknife).

ET Chapter 1 (aspirin-intake example), Chapter 6 (The bootstrap estimate of standard error), Chapter 8.2 (one-sample problem), Chapter 11 (The jackknife), Chapter 13.3 (percentile intervals), Chapter 19.1 (assessing the error in bootstrap estimates).

KM Chapter 2.4.

HWC Chapter 8.4.

Li:C2016: Seiler (2016). Lecture Notes on Nonparametric Statistics - bootstrap example.