Bootstrap For Complex Data

Mark M. Fredrickson (`mfredric@umich.edu`)

Computational Methods in Statistics and Data Science (Stats 406)

Bootstrap Review

- A sample of n units, $X_i \overset{\text{iid}}{\sim} F$

Bootstrap Review

- A sample of n units, $X_i \overset{\text{iid}}{\sim} F$
- A **statistic** $T(X_1, \ldots, X_n)$

Bootstrap Review

- A sample of n units, $X_i \overset{\text{iid}}{\sim} F$
- A **statistic** $T(X_1, \ldots, X_n)$
- Estimate $F(x)$ with $\hat{F} = (1/n)\sum_{i=1}^{n} I(X_i \leq x)$

Bootstrap Review

- A sample of n units, $X_i \overset{\text{iid}}{\sim} F$
- A **statistic** $T(X_1, \ldots, X_n)$
- Estimate $F(x)$ with $\hat{F} = (1/n) \sum_{i=1}^{n} I(X_i \leq x)$
- **Inversion method** for $\hat{F}$ implies **draw from $X_1, \ldots, X_n$, uniformly**.

Bootstrap Review

- A sample of n units, $X_i \overset{\text{iid}}{\sim} F$
- A **statistic** $T(X_1, \ldots, X_n)$
- Estimate $F(x)$ with $\hat{F} = (1/n)\sum_{i=1}^{n} I(X_i \leq x)$
- **Inversion method** for $\hat{F}$ implies **draw from** $X_1, \ldots, X_n$**, uniformly**.
- **Bootstrap sample** $X_1^*, X_2^*, \ldots, X_n^*$ is taken **with replacement** from original sample.

Bootstrap Review

- A sample of n units, $X_i \stackrel{\text{iid}}{\sim} F$
- A **statistic** $T(X_1, \ldots, X_n)$
- Estimate $F(x)$ with $\hat{F} = (1/n) \sum_{i=1}^{n} I(X_i \leq x)$
- **Inversion method** for $\hat{F}$ implies **draw from** $X_1, \ldots, X_n$**, uniformly**.
- **Bootstrap sample** $X_1^*, X_2^*, \ldots, X_n^*$ is taken **with replacement** from original sample.
- Estimate **sampling distribution** of T using draws from $T^* = T(X_1^*, \ldots, X_n^*)$

Bootstrap Review

- A sample of n units, $X_i \overset{\text{iid}}{\sim} F$
- A **statistic** $T(X_1, \dots, X_n)$
- Estimate $F(x)$ with $\hat{F} = (1/n)\sum_{i=1}^{n} I(X_i \leq x)$
- **Inversion method** for $\hat{F}$ implies **draw from** $X_1, \dots, X_n$**, uniformly**.
- **Bootstrap sample** $X_1^*, X_2^*, \dots, X_n^*$ is taken **with replacement** from original sample.
- Estimate **sampling distribution** of T using draws from $T^* = T(X_1^*, \dots, X_n^*)$
- Use the bootstrap distribution to form **confidence intervals** (several methods)

Example: Student academic test scores

ELSEVIER

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Research in Social Stratification and Mobility 30 (2012) 97–112

Research in Social Stratification and Mobility

http://elsevier.com/locate/rssm

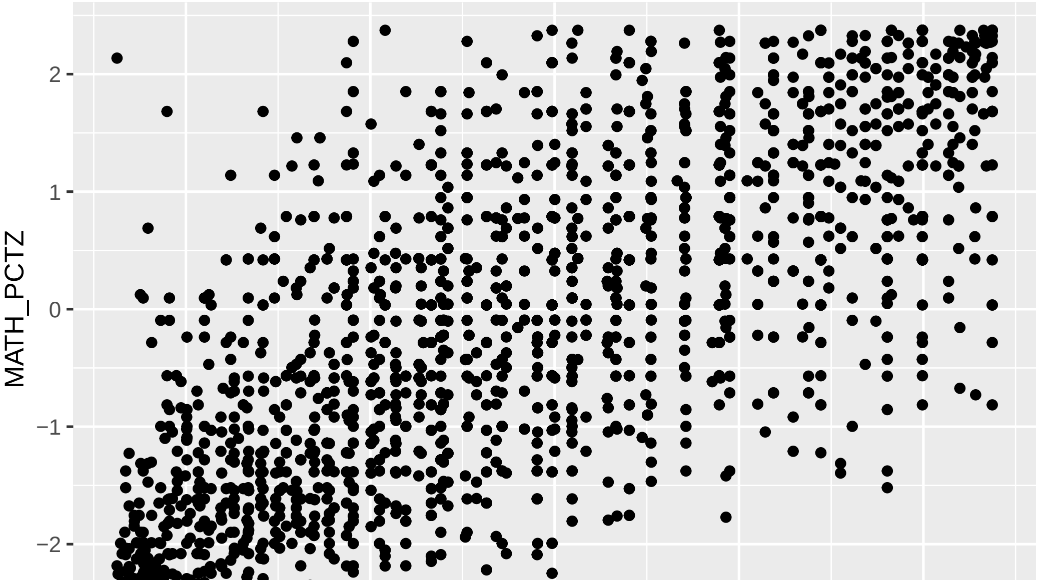
Differences between Hispanic and non-Hispanic families in social capital and child development: First-year findings from an experimental study

Adam Gamoran [a,*], Ruth N. López Turley [b], Alyn Turner [a], Rachel Fish [a]

[a] University of Wisconsin-Madison, , United States
[b] Rice University, , United States

Received 15 March 2011; received in revised form 5 August 2011; accepted 9 August 2011

Reading Test Scores



Test Score Correlation

```
> library(boot)
> median_boot <- function(x, index) {
+   xstar <- x[index, ]
+   cor(xstar$READ_PCTZ, xstar$MATH_PCTZ)
+ }
> cor_boot <- boot(gamoran, statistic = median_boot, R = 100)
```

Confidence Intervals

```
> boot.ci(cor_boot, type = "basic")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates

CALL :
boot.ci(boot.out = cor_boot, type = "basic")

Intervals :
Level      Basic
95%   ( 0.6924,  0.7443 )
Calculations and Intervals on Original Scale
Some basic intervals may be unstable
```

Bootstrap for Dependence and Structure

Bootstrap basics

Recall our usual setup for the bootstrap:

$$X_i \stackrel{\text{iid}}{\sim} F$$

so we use $\hat{F}$ instead of F in the inversion method.

Bootstrap basics

Recall our usual setup for the bootstrap:

$$X_i \overset{\text{iid}}{\sim} F$$

so we use $\hat{F}$ instead of F in the inversion method.

But what if we think there is dependence in the X_i values? What if we don't think they all come from the same distribution?

Bootstrap basics

Recall our usual setup for the bootstrap:

$$X_i \overset{\text{iid}}{\sim} F$$

so we use $\hat{F}$ instead of F in the inversion method.

But what if we think there is dependence in the X_i values? What if we don't think they all come from the same distribution?

We'll consider two cases:

- Bootstrap for **time series** (stochastic processes)
- **Stratified bootstrap** for different groups

Time series models

Suppose we are interested in the following **stochastic process**:

$$X(t) = \rho X(t-1) + e(t), \quad e(t) \sim F, E(e(t)) = 0, \forall t$$

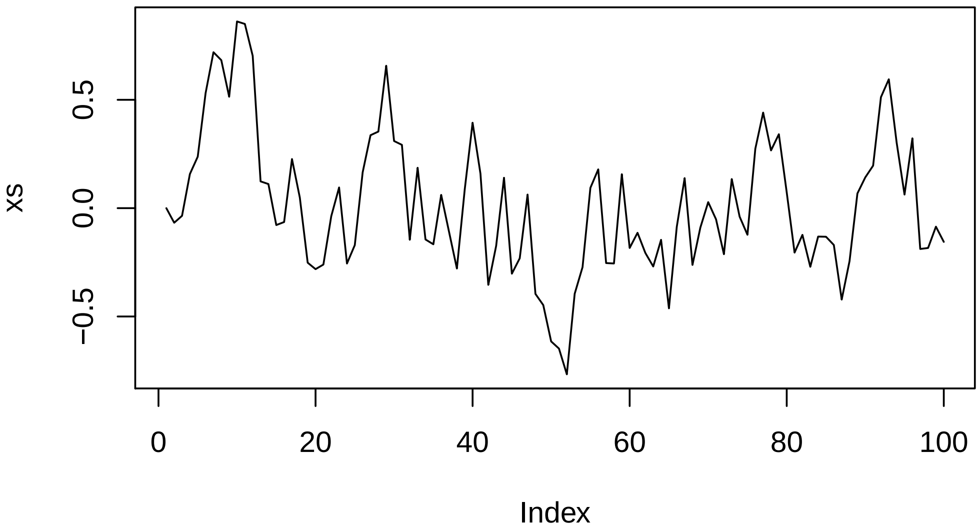
and we are interested in estimating ρ.

Suppose we are interested in the following **stochastic process**:

$$X(t) = \rho X(t-1) + e(t), \quad e(t) \sim F, E(e(t)) = 0, \forall t$$

and we are interested in estimating ρ. Here is some simulated example data:

```
> rho <- 0.75
> k <- 100
> xs <- numeric(k)
> xs[1] <- 0
> for (i in 2:k) {
+     xs[i] <- rho * xs[i - 1] + (rbeta(1, 2, 2) - 0.5)
+ }
```


xs
0.5
0.0
−0.5
0
20
40
60
80
100
Index

Estimating ρ

One way we can estimate ρ is to perform **OLS on $X(t)$ given $X(t-1)$**:

```
> (est <- lm(xs[2:k] ~ xs[1:(k - 1)] - 1)) # no intercept term

Call:
lm(formula = xs[2:k] ~ xs[1:(k - 1)] - 1)


Coefficients:
xs[1:(k - 1)]
         0.71
```

What is the variation? Can we get bootstrap confidence intervals?

Parametric bootstrap

Here we need **bootstrap samples with dependence**:

$$X(0)^*, X(1)^*, \ldots, X(t)^* \quad \text{such that: } X(s)^* = \rho X(s-1)^* + e(s)^*$$

Parametric bootstrap

Here we need **bootstrap samples with dependence**:

$$X(0)^*, X(1)^*, \ldots, X(t)^* \quad \text{such that: } X(s)^* = \rho X(s-1)^* + e(s)^*$$

We have already seen **parametric bootstrapping** in which we use **a model** to generate the bootstrap samples.

Idea: starting from the observed $X(0)$, draw from the **estimated distribution of e** to create a **bootstrap series**:

$$X(1)^* = \hat{\rho} X(0) + e(1)^*, \ldots, X(t)^* = \hat{\rho} X(t-1)^* + e(t)^*$$

```
> rhohat <- coef(est)
> et_hat <- xs[2:k] - predict(est) # estimate residuals
> est_rho_stat <- function(x, index) {
+   estar <- x[index]
+   new_series <- numeric(k)
+   new_series[1] <- xs[1] # starts at the same point
+   for (i in 2:k) {
+     new_series[i] <- rhohat * new_series[i - 1] + estar[i]
+   }
+   coef(lm(new_series[2:k] ~ new_series[1:(k - 1)] - 1))
+ }
```

CIs for ρ

```
> boot_rho <- boot(et_hat, est_rho_stat, R = 1000)
> boot.ci(boot_rho, type = "perc")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_rho, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.5263,  0.8143 )
Calculations and Intervals on Original Scale
```

Two Sample Problems

In the previous example, all the $e(t)$ were **identical**, but the $X(t)$ were not **independent**.

Two Sample Problems

In the previous example, all the $e(t)$ were **identical**, but the $X(t)$ were not **independent**.

What about problems were you have **independence but not identical** data?

Two Sample Problems

In the previous example, all the $e(t)$ were **identical**, but the $X(t)$ were not **independent**.

What about problems were you have **independence but not identical** data?

Two sample problems consider the situation of having two independent samples

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F \quad Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} G$$

and estimating quantities such as the **difference of means**:

$$\Delta = \mathrm{E}(X) - \mathrm{E}(Y)$$

(notice: if $F = G$, $\Delta = 0$, though not necessarily the converse.)

Students in Cities

One aspect of the Gamoran et. al study is that one portion of the students where located in San Antonio, TX while another portion was located in Phoenix, AZ.

Students in Cities

One aspect of the Gamoran et. al study is that one portion of the students where located in San Antonio, TX while another portion was located in Phoenix, AZ.

We might not be willing to believe that those two populations have the same distribution function.

Students in Cities

One aspect of the Gamoran et. al study is that one portion of the students where located in San Antonio, TX while another portion was located in Phoenix, AZ.

We might not be willing to believe that those two populations have the same distribution function.

In fact, we might be interested in knowing if the two populations have the same distribution (or same mean, etc). In particular, we'll focus on $\Delta = \mathrm{E}\,(\text{San Antonio}) - \mathrm{E}\,(\text{Phoenix})$.

Students in Cities

One aspect of the Gamoran et. al study is that one portion of the students where located in San Antonio, TX while another portion was located in Phoenix, AZ.

We might not be willing to believe that those two populations have the same distribution function.

In fact, we might be interested in knowing if the two populations have the same distribution (or same mean, etc). In particular, we'll focus on $\Delta = \mathrm{E}\left(\text{San Antonio}\right) - \mathrm{E}\left(\text{Phoenix}\right)$.

We'll bootstrap in the two groups separately to get estimates of means and then combine them.

Estimating the difference of means

The boot function has a strata (groups) argument we can use.

```
> mean_diff <- function(x, index) {
+     xstar <- x[index, ] # boot will handle stratification for us
+     mean(xstar$READ_PCTZ[xstar$PH.AZ], na.rm = TRUE) -
+         mean(xstar$READ_PCTZ[!xstar$PH.AZ], na.rm = TRUE)
+ }
> gam.boot <- boot(gamoran,
+                  statistic = mean_diff,
+                  strata = gamoran$PH.AZ,
+                  R = 1000)
```

CI

```
> (gbci <- boot.ci(gam.boot, type = "basic"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = gam.boot, type = "basic")

Intervals :
Level      Basic
95%   ( 0.1398,  0.3782 )
Calculations and Intervals on Original Scale
```

Interpreting Results

We saw that the **95% confidence interval for Δ did not include zero**:

```
0.1398 0.3782
```

Interpreting Results

We saw that the **95% confidence interval for Δ did not include zero**:

```
0.1398 0.3782
```

In other words, we can **reject the hypothesis** that the cities have **the same average reading score**.

Interpreting Results

We saw that the **95% confidence interval for Δ did not include zero**:

```
0.1398 0.3782
```

In other words, we can **reject the hypothesis** that the cities have **the same average reading score**.

Additionally, we can **reject** any hypothesis that says that **San Antonio has a higher average reading score**. (All at $\alpha = 0.05$ level).

Interpreting Results

We saw that the **95% confidence interval for Δ did not include zero**:

```
0.1398 0.3782
```

In other words, we can **reject the hypothesis** that the cities have **the same average reading score**.

Additionally, we can **reject** any hypothesis that says that **San Antonio has a higher average reading score**. (All at $\alpha = 0.05$ level).

Let's be clear that this is not a **causal conclusion**, merely that on average San Antonio students tend to score lower.

Summary

- Key assumption for bootstrap is that sample is **independent, identically distributed**.

Summary

- Key assumption for bootstrap is that sample is **independent, identically distributed**.
- Bootstrap statistics **resample rows**

Summary

- Key assumption for bootstrap is that sample is **independent, identically distributed**.
- Bootstrap statistics **resample rows**
- For **dependent or structured data**, we can still bootstrap by **including the dependence/structure**.

Summary

- Key assumption for bootstrap is that sample is **independent, identically distributed**.
- Bootstrap statistics **resample rows**
- For **dependent or structured data**, we can still bootstrap by **including the dependence/structure**.
- For dependent data, generate samples following dependence.

Summary

- Key assumption for bootstrap is that sample is **independent, identically distributed**.
- Bootstrap statistics **resample rows**
- For **dependent or structured data**, we can still bootstrap by **including the dependence/structure**.
- For dependent data, generate samples following dependence.
- For structured data, incorporate the structure.