Bootstrap

Stat 510

Setup

$$X_1, X_2, \ldots X_n \sim F$$

- How to estimate F? Empirical Distribution
- How to estimate properties of F, $T(F)$? Plug-in

$$T_n = g(X_1, X_2, \ldots X_n)$$

↑ statistic

Sampling Distribution

$$X_1, X_2, \ldots X_n \sim F$$

$$T_n = g(X_1, X_2, \ldots X_n)$$

↑ STATISTIC

- WHAT IS THE DISTRIBUTION OF T_n?
  - $\mathbb{E}[T_n] = ?$
  - $\mathbb{V}[T_n] = ?$ ← $T_n \pm 2 \cdot se(T_n)$
  - ETC

Easy Example

$$X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n) \quad \checkmark$$

"HARD" EXAMPLE

$$X_1, X_2, \ldots X_n \sim Exp(\lambda)$$

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

$$S_n^2 \sim ?$$

ALSO MEDIAN, QUANTILE, ETC

Simulation

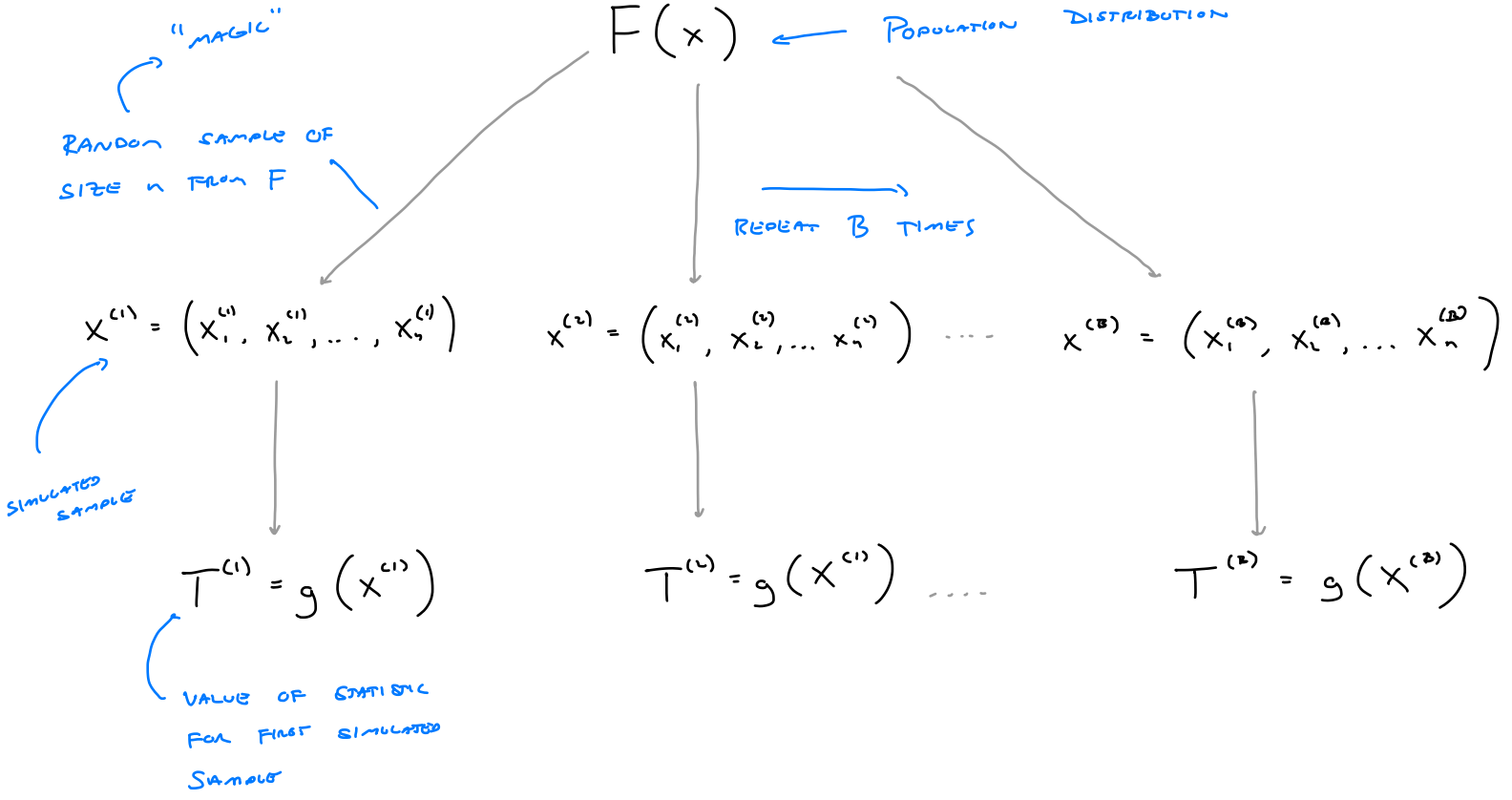
1. * Simulate data $X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)} \sim F$
2. Calculate statistic $T_n^{(i)} = g\left(X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)}\right)$

Repeat B times $\longrightarrow$ $T_n^{(1)}, T_n^{(2)}, T_n^{(3)}, T_n^{(4)}, \ldots T_n^{(B)}$

Use empirical distribution as estimate of sampling distribution

$\hat{F}_{T_n}$ estimates F_{T_n}

* magic


"MAGIC"
RANDOM SAMPLE OF SIZE n FROM F
F(x)
POPULATION DISTRIBUTION
REPEAT B TIMES
$X^{(1)} = (X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)})$
$X^{(2)} = (X_1^{(2)}, X_2^{(2)}, \ldots X_n^{(2)})$ ....
$X^{(B)} = (X_1^{(B)}, X_2^{(B)}, \ldots X_n^{(B)})$
SIMULATED SAMPLE
$T^{(1)} = g(X^{(1)})$
$T^{(2)} = g(X^{(2)})$ ....
$T^{(B)} = g(X^{(B)})$
VALUE OF STATISTIC FOR FIRST SIMULATED SAMPLE

PROBLEM

IN PRACTICE, WE DON'T KNOW F

WE HAVE DATA $x_1, x_2, \ldots, x_n$

IDEA

INSTEAD OF SIMULATING DATA FROM $F(x)$

SIMULATE DATA FROM $\hat{F}(x)$

EMPIRICAL DISTRIBUTION

BOOTSTRAP

BOOTSTRAP RESAMPLE

1. GENERATE DATA $X_*^{(i)} = \left(X_{1,*}^{(i)}, X_{2,*}^{(i)}, \ldots, X_{n,*}^{(i)}\right) \sim \hat{F}_n$
2. CALCULATE STATISTIC $T_*^{(i)} = g\left(X_*^{(i)}\right)$

BOOTSTRAP REPLICATE

REPEAT B TIMES $\longrightarrow$ $T_*^{(1)}, T_*^{(2)}, T_*^{(3)}, T_*^{(4)}, \ldots T_*^{(B)}$

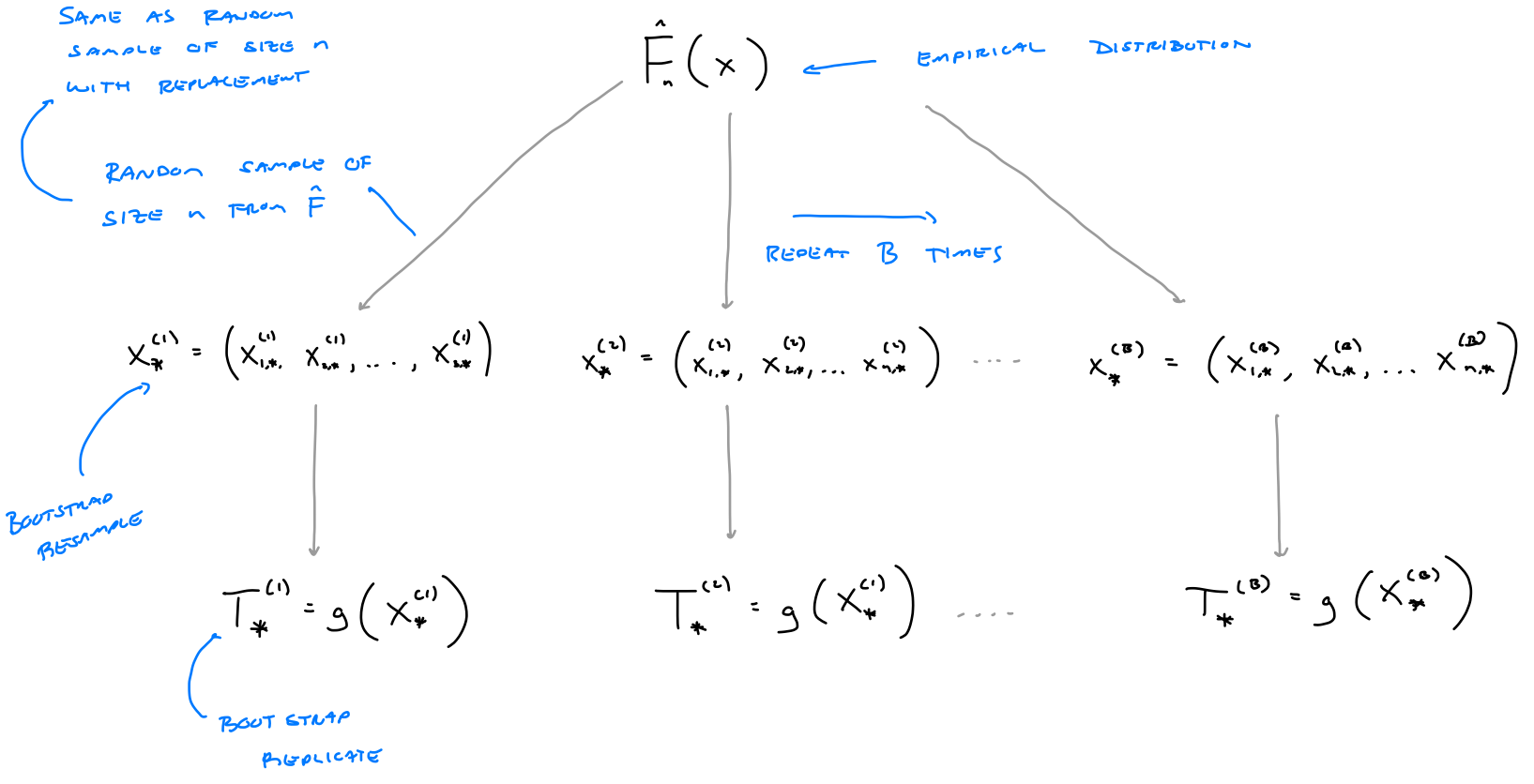
USE EMPIRICAL DISTRIBUTION AS ESTIMATE OF SAMPLING DISTRIBUTION

Bootstrap Resample

- A sample of size n with replacement from the original sample.
  - Same as a random sample of size n from $\hat{F}_n(x)$

Bootstrap Replicate

- A statistic calculated on a bootstrap resample


SAME AS RANDOM SAMPLE OF SIZE n WITH REPLACEMENT
RANDOM SAMPLE OF SIZE n FROM $\hat{F}$
$\hat{F}_n(x)$
EMPIRICAL DISTRIBUTION
REPEAT B TIMES
$X_*^{(1)} = (X_{1,*}^{(1)}, X_{2,*}^{(1)}, \ldots, X_{n,*}^{(1)})$
$X_*^{(2)} = (X_{1,*}^{(2)}, X_{2,*}^{(2)}, \ldots X_{n,*}^{(2)})$ ....
$X_*^{(B)} = (X_{1,*}^{(B)}, X_{2,*}^{(B)}, \ldots X_{n,*}^{(B)})$
BOOTSTRAP RESAMPLE
$T_*^{(1)} = g(X_*^{(1)})$
$T_*^{(2)} = g(X_*^{(2)})$ ....
$T_*^{(B)} = g(X_*^{(B)})$
BOOTSTRAP REPLICATE

WANT TO ESTIMATE $\mathbb{V}_F[T_n]$

IF POSSIBLE USE $\mathbb{V}_{\hat{F}_n}[T_n]$

PLUG-IN ESTIMATOR

"SAMPLE VARIANCE"

FOR EXAMPLE $\hat{\mathbb{V}}[\bar{X}] = \hat{\sigma}^2/n$

WITH BOOTSTRAP USE $V_{BOOT} = \frac{1}{B}\sum_{b=1}^{B}\left(T_n^{*(b)} - \frac{1}{B}\sum_{r=1}^{B} T_n^{*(r)}\right)^2$

("SAMPLE") VARIANCE OF BOOTSTRAP REPLICATES

Bootstrap Standard Error

$$\hat{SE}_{BOOT} = \sqrt{V_{BOOT}}$$

WITH EST AND SE ...

CONFIDENCE INTERVALS!

Bootstrap CI

1. Normal
2. Pivotal
3. Percentile

"Normal" Bootstrap Interval

$(1-\alpha) \times 100\%$ CI

$$T_n \pm z_{\alpha/2} \cdot \hat{se}_{boot}$$

"Good" only if T_n approx normal

"Pivotal" Bootstrap Interval

$(1-\alpha) \times 100\%$ CI

$$\theta = T(F) \qquad \hat{\theta} = T(\hat{F})$$

$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$ ← Bootstrap replicates of $\hat{\theta}$

$$C_n = \left(2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}^*_{\alpha/2} \right)$$

$1-\alpha/2$ sample quantile of bootstrap replicates

$\alpha/2$ sample quantile of bootstrap replicates

"Percentile" Boot strap interval

$(1-\alpha) \times 100\%$ CI

$$\theta = T(F) \qquad \hat{\theta} = T(\hat{F})$$

$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$ ← Boot strap replicates of $\hat{\theta}$

$$C_n = \left(\theta^*_{\alpha/2}, \; \theta^*_{1-\alpha/2} \right)$$

IN R

- WRITE FUNCTION THAT
  - RESAMPLES DATA
  - CALCULATES STATISTIC
- REPLICATE!
- CALCULATE INTERVAL