# Statistical Theory

# In the Age of the Computer:

## BOOTSTRAP METHODS

Bradley Efron
(Stanford)

# $n = 15$ Observed Lifetimes:

| .143 | .182 | .256 | .260 | .270 | .437 |
|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |

| .509 | .611 | .712 | 1.04 | 1.09 | 1.15 |
|---|---|---|---|---|---|
| $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |

| 1.46 | 1.88 | 2.08 |
|---|---|---|
| $X_{13}$ | $X_{14}$ | $X_{15}$ |

Sample Mean

$\bar{x} = .804$

Sample Median

$\hat{\theta} = .611$

Estimated Standard Error

$\hat{\sigma} = .155$

$\hat{\sigma} = ??$

$\hat{\sigma} = \left[\sum (x_i - \bar{x})^2 / n^2\right]^{\frac{1}{2}}$

# We have

DATA $\underset{\sim}{y}$

e.g. $\underset{\sim}{y} = (x_1, x_2, \cdots, x_{15})$ random sample of 15 lifetimes

PARAMETER Of Interest $\theta$

e.g. the true expected lifetime
or true median lifetime.

We Want to estimate

$\theta$ from $\underset{\sim}{y}$

# Two Basic Questions

Question 1: What statistic $\hat{\theta}(\underset{\sim}{y})$ should we use to estimate $\theta$?

Question 2: How accurate is $\hat{\theta}$ as an estimate of $\theta$?

# Maximum Likelihood Theory

Answer 1: Use $\hat{\theta}(\underset{\sim}{y})$ the MLE

Answer 2: Standard Error of $\hat{\theta}$ is approximately

$$\hat{\sigma} = \frac{1}{\sqrt{\text{Fisher Info}}}$$

# Bootstrap is a more general way to answer Q2.

- Less Parametric Modelling
  (even nonparametric)
- More Computation
  (x100 or 1000)
- Automatic
  (Algorithm)

# The Simplest Situation

$F \xrightarrow{\text{Random Sample}} (X_1, X_2, \cdots X_n) = \underset{\sim}{Y}$

($F$: true distribution)

Statistic of Interest: $\bar{X} = \sum_1^n X_i / n$

$\hat{\theta}(\underset{\sim}{Y})$

Simple formula for standard error:

$$\sigma(F) = [\mu_2(F)/n]^{1/2}$$

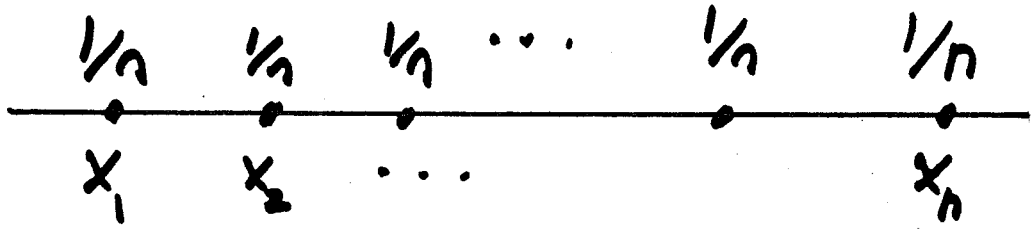
Where

$\mu_2(F) = 2^{nd}$ Central Moment of $F$

$$= E_F[X - E_F\{X\}]^2$$

# Estimating $\sigma(F)$

Empirical Distribution: $\hat{F}$:




$$\mu_2(\hat{F}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$$

Gives estimated standard error

$$\sigma(\hat{F}) = [\mu_2(\hat{F})/n]^{1/2}$$

$$= \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 / n^2\right]^{1/2}$$

a nice simple formula.

# For More Complicated Statistics

e.g. $\hat{\theta}(\underset{\sim}{y})$ = Sample Median

- No Simple Formula for
  $\sigma(F) \equiv$ Standard Error of $\hat{\theta}(\underset{\sim}{y})$

- Can't "Substitute $\hat{F}$ for $F$"

- **Bootstrap**: computer algorithm for finding numerical value of

$$\sigma(\hat{F}) \equiv \hat{\sigma}$$

$\hat{F}$: Empirical

$\hat{\sigma}$: Bootstrap Estimate of Standard error

# Bootstrap Sampling

"$\hat{F} \rightarrow \underset{\sim}{y}^* = (x_1^*, x_2^*, \cdots x_n^*)$"

Means that you

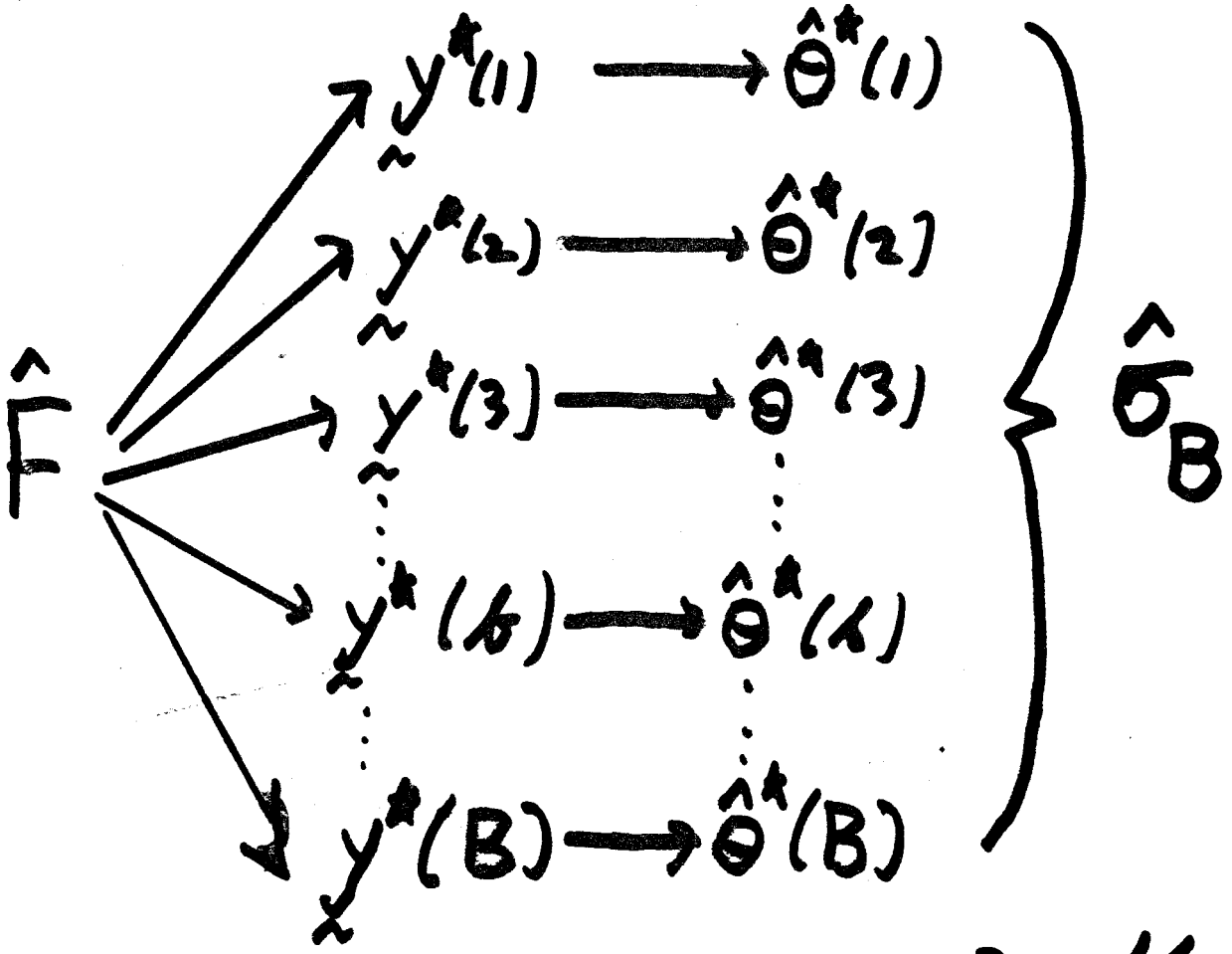
"Draw a random sample of size n from $\hat{F}$"

Equivalent: $\underset{\sim}{y}^* = (x_1^*, x_2^*, \cdots x_n^*)$ is a random sample of size n drawn with replacement from

$$\{x_1, x_2, \cdots, x_n\}$$

Definition: $\underset{\sim}{y}^*$ is a Bootstrap Sample

# BOOTSTRAP ALGORITHM




$$\hat{\sigma}_B = \left\{ \frac{\sum [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1} \right\}^{1/2}$$

As $B \to \infty$, $\hat{\sigma}_B \to \hat{\sigma} = \sigma(\hat{F})$,

the bootstrap estimate of standard error.

Usually $B = 100$ is plenty:

Lifetime Data, $\hat{\theta}$ = Sample Median

| $B$: | 25 | 50 | 100 | 200 | 1000 |
|---|---|---|---|---|---|
| $\hat{\sigma}_B$: | .23 | .22 | .23 | .25 | .25 |

# Results for the 15 lifetimes

B = 100 Bootstraps

Sample Mean $\hat{\sigma}_{100} = .156$ → $\hat{\sigma}_{\infty} = .155$

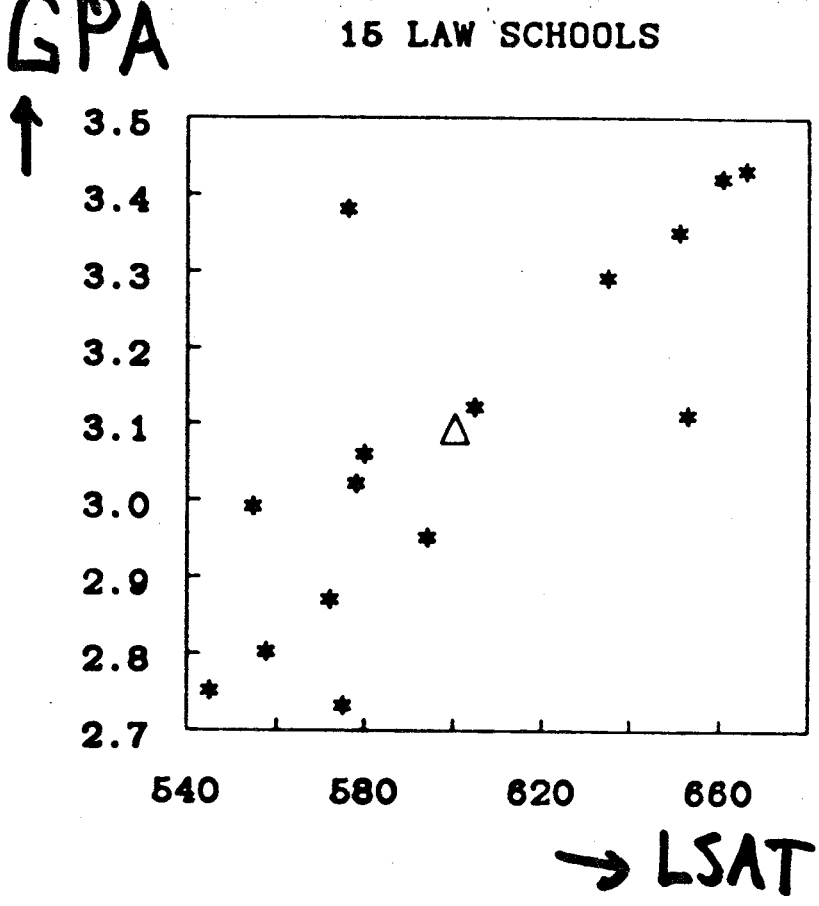
$\left\{ \Sigma (x_i - \bar{x})^2 / n^2 \right\}^{1/2}$

Sample Median $\hat{\sigma}_{100} = .229$

"boot (x, 100, "median")"

$n = 15$ $\underset{\sim}{y} = (x_1, x_2, \cdots, x_{15})$

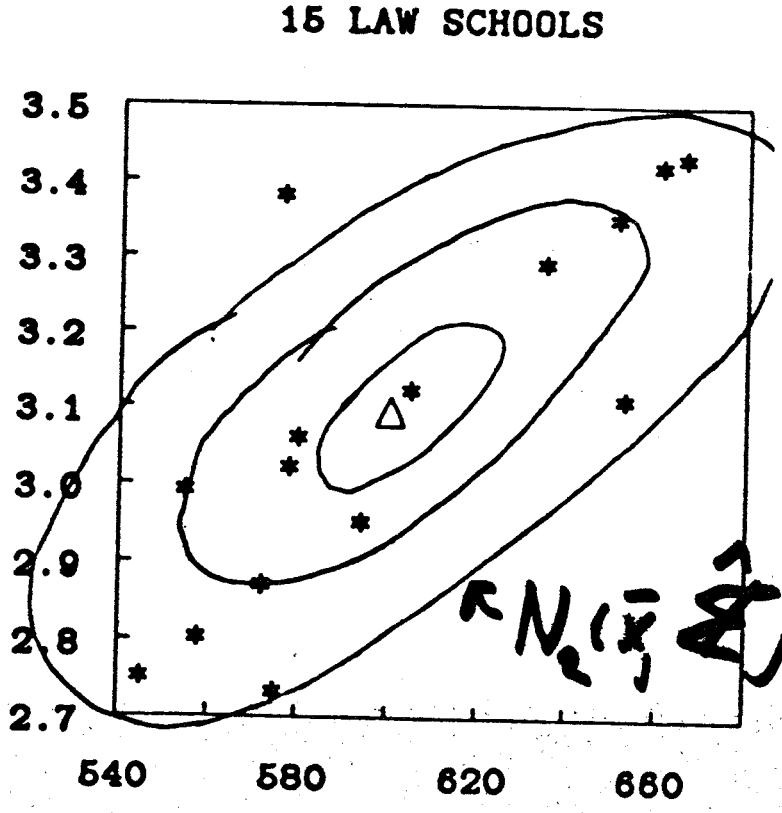
$x_i = (LSAT_i, GPA_i)$




$\hat{\theta}(\underset{\sim}{y})$ = Correlation Coefficient

$= .776$

$\hat{\sigma}_{1000} = .127$

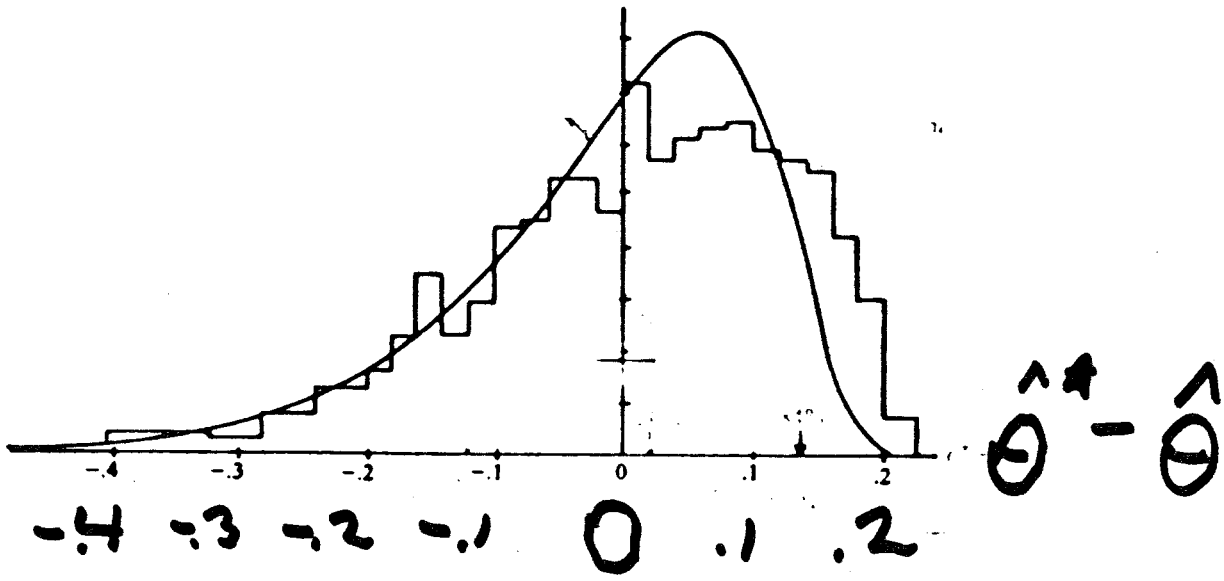



$\hat{\sigma}_{NORM} = \frac{1 - \hat{\theta}^2}{\sqrt{12}}$

$= .115$

$\nwarrow N_2(\bar{x}, \hat{\Sigma}) = \hat{F}_{NORM}$

# Histogram of $B=1000$ $\hat{\theta}^*$'s




- $\hat{\sigma}_{1000} = .127$ Compared to $\hat{\sigma}_{NORM} = .115$
- Smooth curve is $B = \infty$ bootstraps starting from $\hat{F}_{NORM} \sim N_2(\bar{x}, \hat{\Sigma})$

  $\hat{F}_{NORM} \longrightarrow \underset{\sim}{y}^* \longrightarrow \hat{\theta}(\underset{\sim}{y}^*)$

- Both histograms are very skewed to the left

# A More Complicated Statistic

n=88 students each took five tests:

| | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | |
|---|---|---|---|---|---|---|
| Student 1 | 77 | 82 | 67 | 67 | 81 | $x_1$ |
| Student 2 | 44 | 56 | 55 | 61 | 36 | $x_2$ |
| Student 3 | 17 | 51 | 52 | 35 | 31 | $x_3$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Student 88 | 0 | 40 | 21 | 9 | 14 | $x_n$ |
| | Mech. (closed) | Vectors (C) | Alg. (open) | Analysis (O) | Stats (O) | |

$$\underset{88\times 5}{X} \rightarrow X_c \rightarrow \underset{5\times 5}{G} = X_c'X_c \rightarrow \text{eigenvectors } \hat{\theta}$$

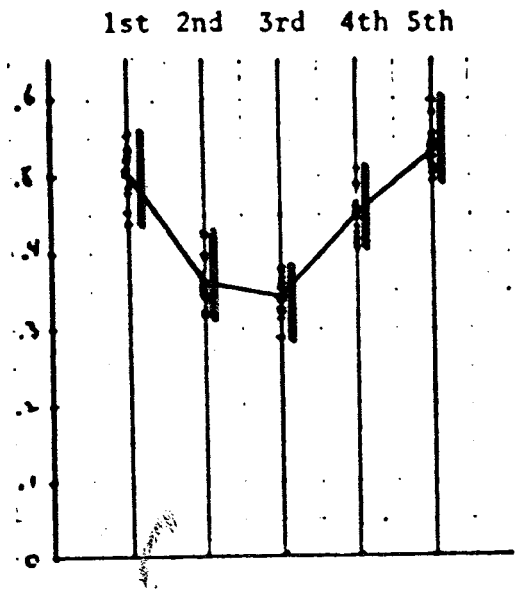
1st eigenvector: (.51, .37, .35, .45, .53)

2nd eigenvector: (.75, .21, -.08, -.30, -.55)

} How Verrallic?

# B=10 Bootstraps of the Test Data

First Eigenvector Components

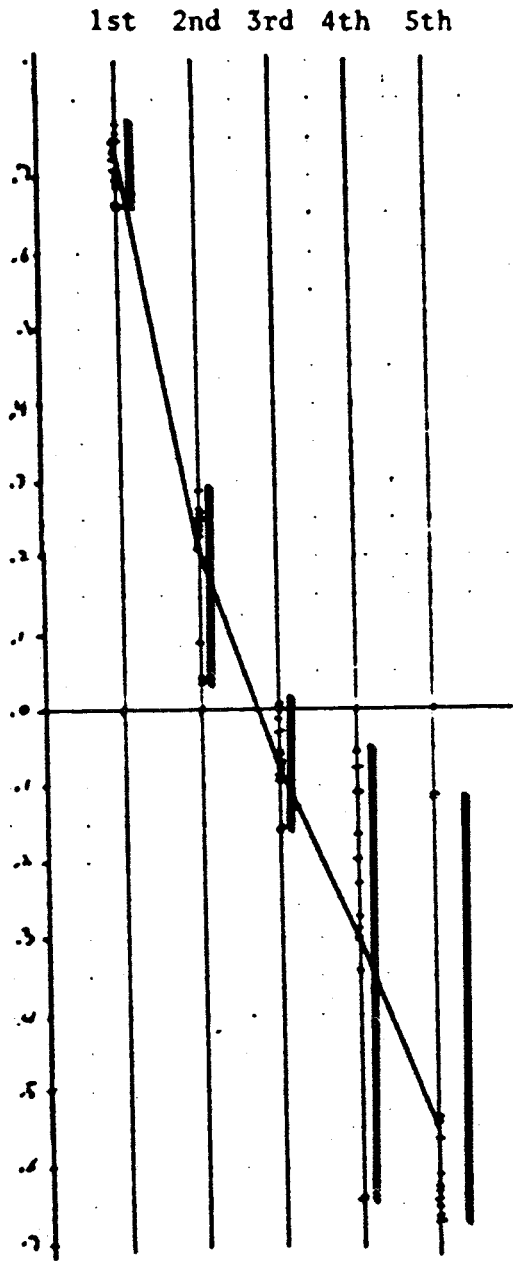



1st Eigenvector

Values of bootstrap eigenvalue components indicated by dashes.

2nd Eigenvector →

Second Eigenvector Components




- Blue lines indicate range of boot values
- 2nd More variable than 1st

"The history of science exhibits a steady tendency to eliminate intellectual effort in the solution of individual problems, by developing comprehensive formulas which can resolve by rote a whole class of them."

... Ernest Nagel, 1955

## References:


Efron and Tibshirani "Bootstrap Methods..."
Statistical Science (86) Vol 1, No 1, p 54-77

Efron "Computers and the theory of statistics..." SIAM review (79) Vol 21, No 4 p460-480