Bootstrap Methods

B. Efron
and
R. Tibshirani

Efron and Tibshirani: "Bootstrap Methods..."
Statistical Science (86) Vol 1, #1, p54-77

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

$\hat{\sigma} = .155$

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

$\hat{\sigma} = ??$

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 θ

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

We Want to estimate

θ from $\underset{\sim}{y}$

Two Basic Questions

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

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

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: $\hat{\theta}(\underset{\sim}{y}) = \bar{x} = \sum_1^n x_i / n$

Simple formula for standard error:

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

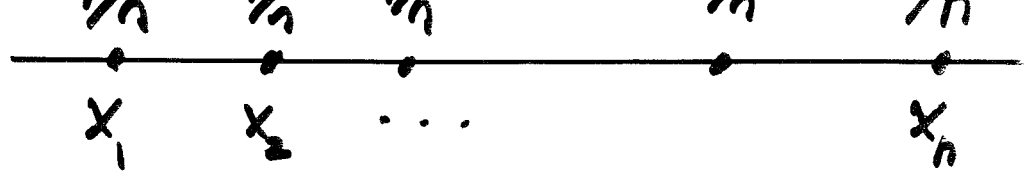
Where

$\mu_2(F) = 2^{nd}$ 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 <u>estimated</u> standard error

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

$\hat{F}$ for F

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

a nice simple formula.

For More Complicated Statistics

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

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

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

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

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

Empirical ↗ ↖ Bootstrap Estimate of Standard error

Bootstrap Sampling

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

Means that you

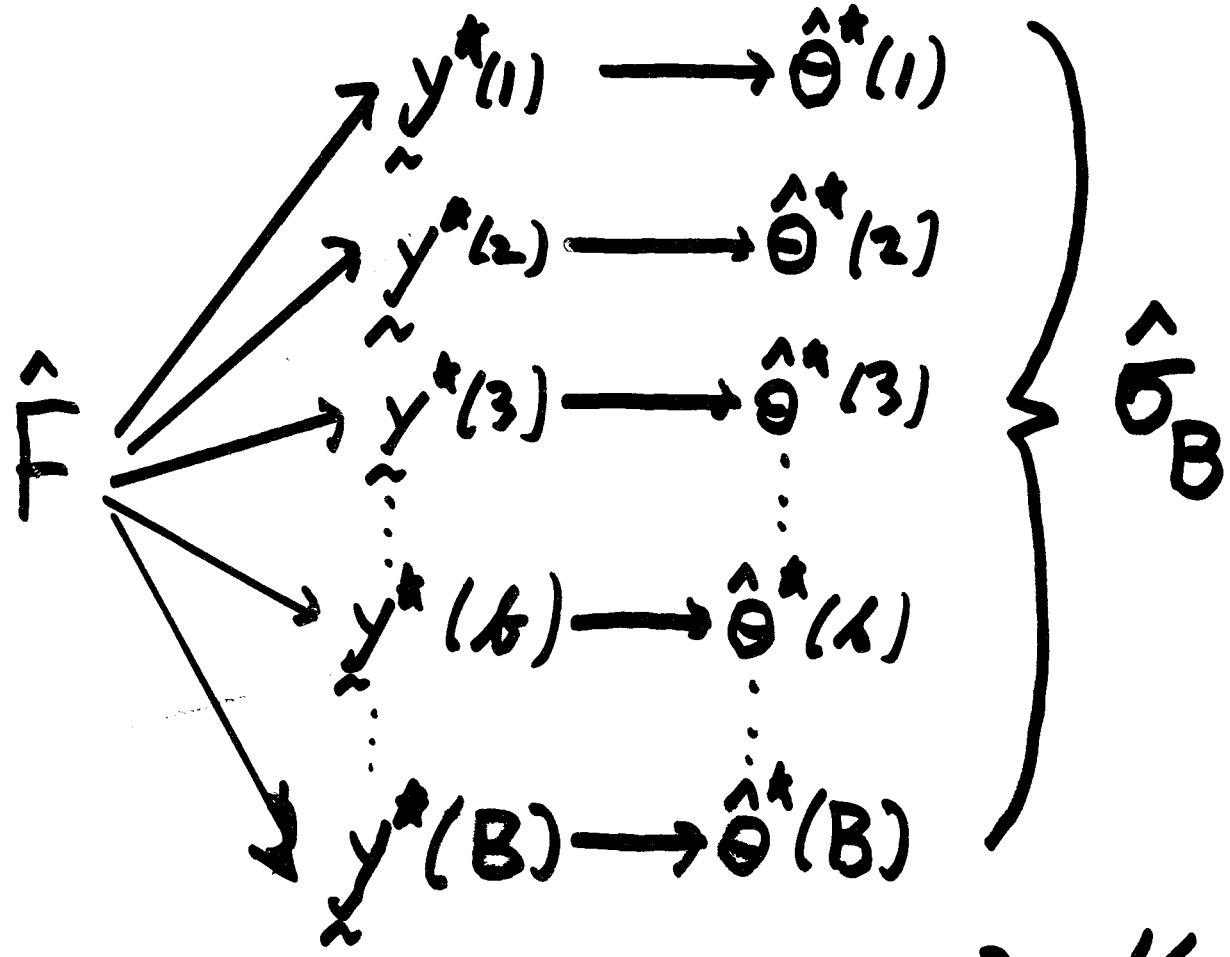
"Draw a random sample of size n from $\hat{F}$"

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

the bootstrap estimate of standard error.

Bootstrap Sample Sizes

How Big "B"? In general

$$\frac{SD\{\hat{\sigma}_B\}}{E\{\hat{\sigma}_B\}} = CV(\hat{\sigma}_B) \downarrow \text{ B.}$$

As $B \to \infty$, $CV(\hat{\sigma}_B) \downarrow CV(\hat{\sigma}_\infty)$,

Coeff of var. for ideal bootstrap.

Example: $F = N(0,1)$, $n = 20$, "$\hat{\theta}$" $= \bar{x}$

Then $\hat{\sigma}_\infty = \left[\frac{\sum (x_i - \bar{x})^2}{20^2}\right]^{1/2}$ has $CV(\hat{\sigma}_\infty) = .15$

"Table 9" shows $\hat{\sigma}_{100}$ almost as good as $\hat{\sigma}$ in this case.

Table 9

$\hat{\sigma}_{25}$	$\hat{\sigma}_{50}$	$\hat{\sigma}_{100}$	$\hat{\sigma}_{200}$	$\hat{\sigma}_{\infty}$ ← $\hat{\sigma}$
.29	.27	.26	.25	.25
.24	.22	.21	.21	.20
.21	.18	.17	.16	.15
.15	.11	.09	.07	.05
.14	.10	.07	.05	0

Coefficient of Variation of $\hat{\sigma}_B$

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

* For Standard Errors

Results for the 15 lifetimes

B = 100 Bootstraps

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

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

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

"boot (y, 100, "median")"

nk

Simulation Study

- $F = N_2\left(0, \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}\right)$ • $n = 15$ • $\hat{\theta}$ = corr coeff
- True standard error is $\sigma(F, n, \hat{\theta}) = .218$

$\hat{\sigma}$	$E\{\hat{\sigma}\}$	$(Var\{\hat{\sigma}\})^{1/2}$	$MSE^{1/2}$
Boot, B=128	.206	.066	.067
Boot, B=512	.206	.063	.064
Jackknife	.223	<u>.085</u>	.085
Delta Method	<u>.175</u>	.058	.072
Normal theory $(1-\hat{\theta}^2)/\sqrt{12}$	.217	.056	.058

Table 2.

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)	Stat (O)	

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

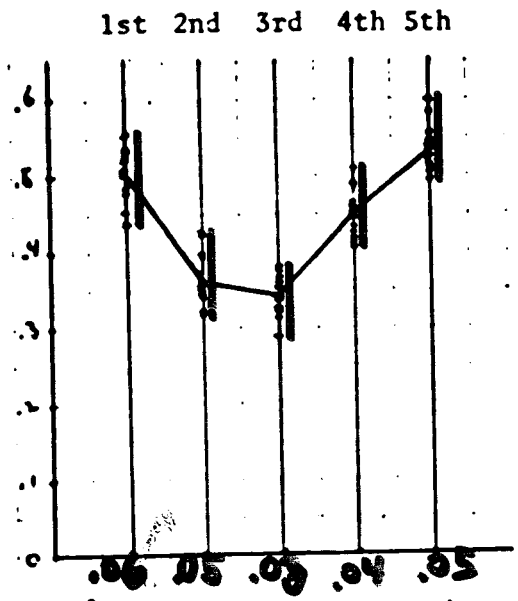
1st eigenvector: (.51, .37, .35, .45, .53)

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

} How Varimax?

B=10 Bootstraps of the Test Data

First Eigenvector Components

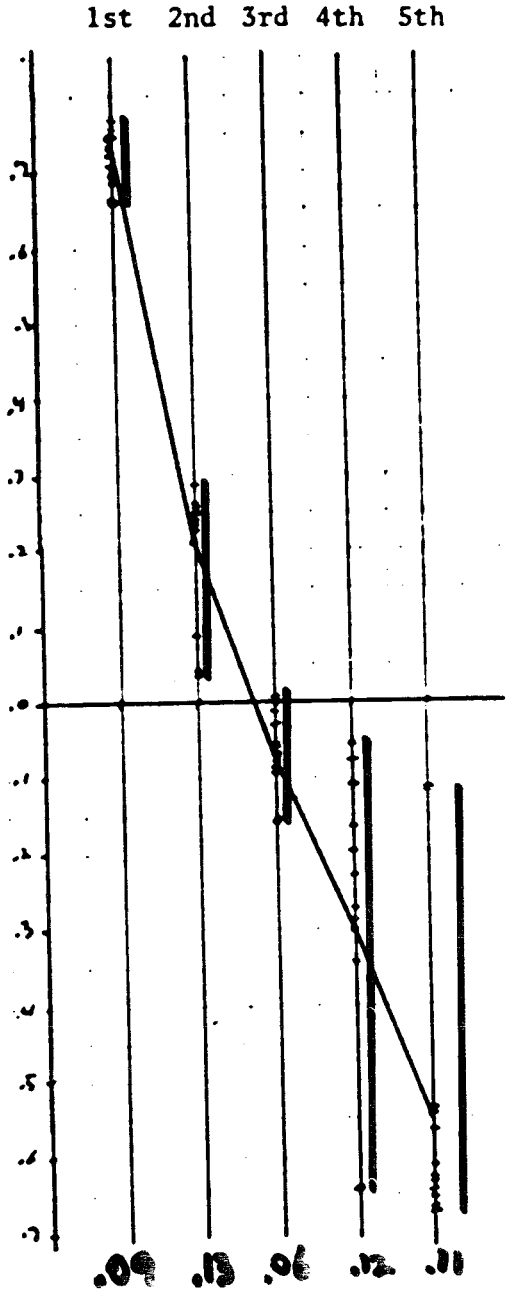


1st Eigenvector

Values of bootstrap eigenvalue components indicated by dashes.

2nd Eigenvector →

Second Eigenvector Components



B=100

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

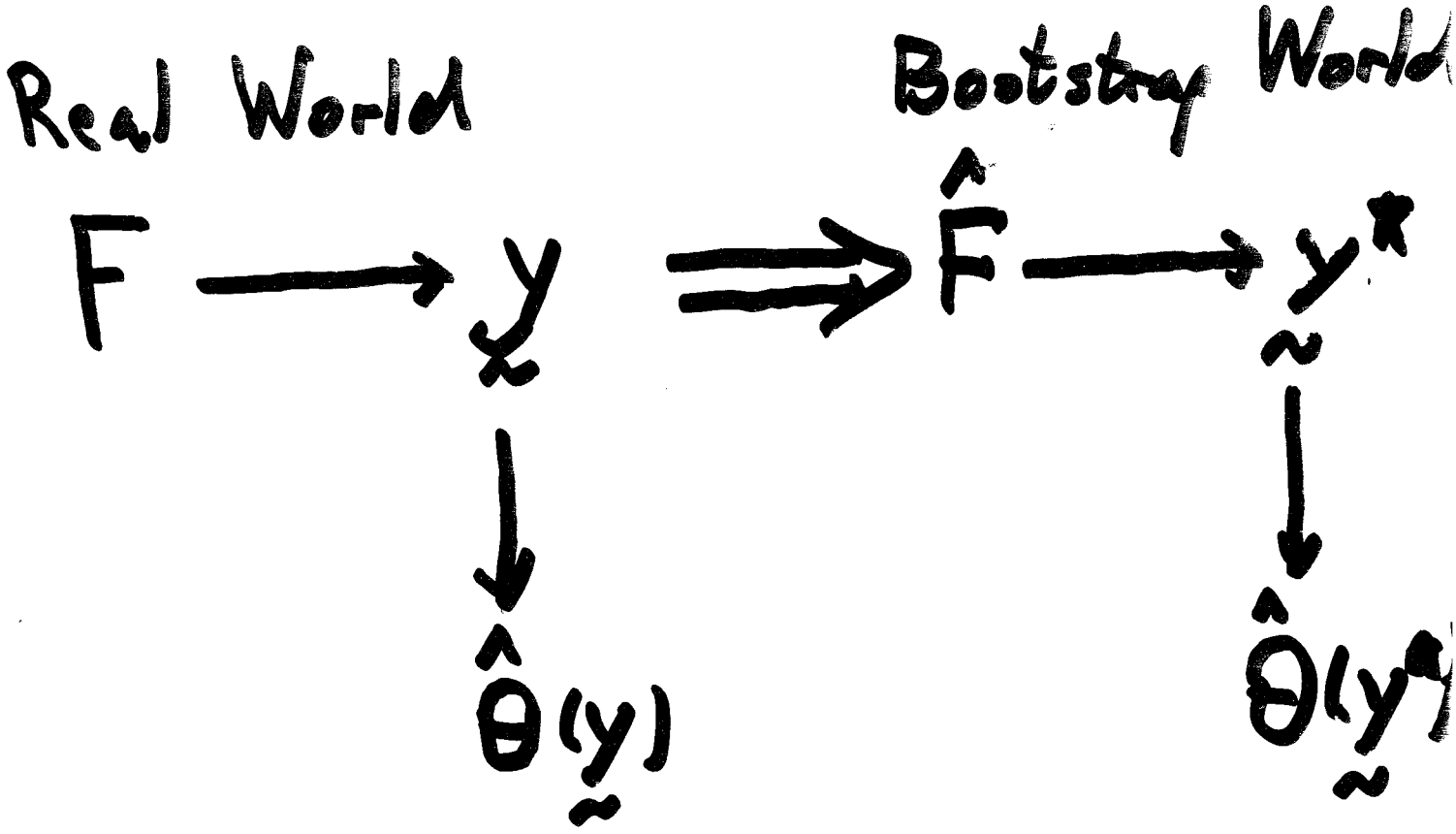
MORAL

- Most of the familiar methods of getting standard errors are approx. versions of $\hat{\sigma} = \sigma(\hat{F})$.

- Bootstrap evaluates $\sigma(\hat{F})$ directly by brute force Monte Carlo

- Mathematical complications of step $\underset{\sim}{y} \rightarrow \hat{\theta}$ don't bother the bootstrap.

Our Story so far....

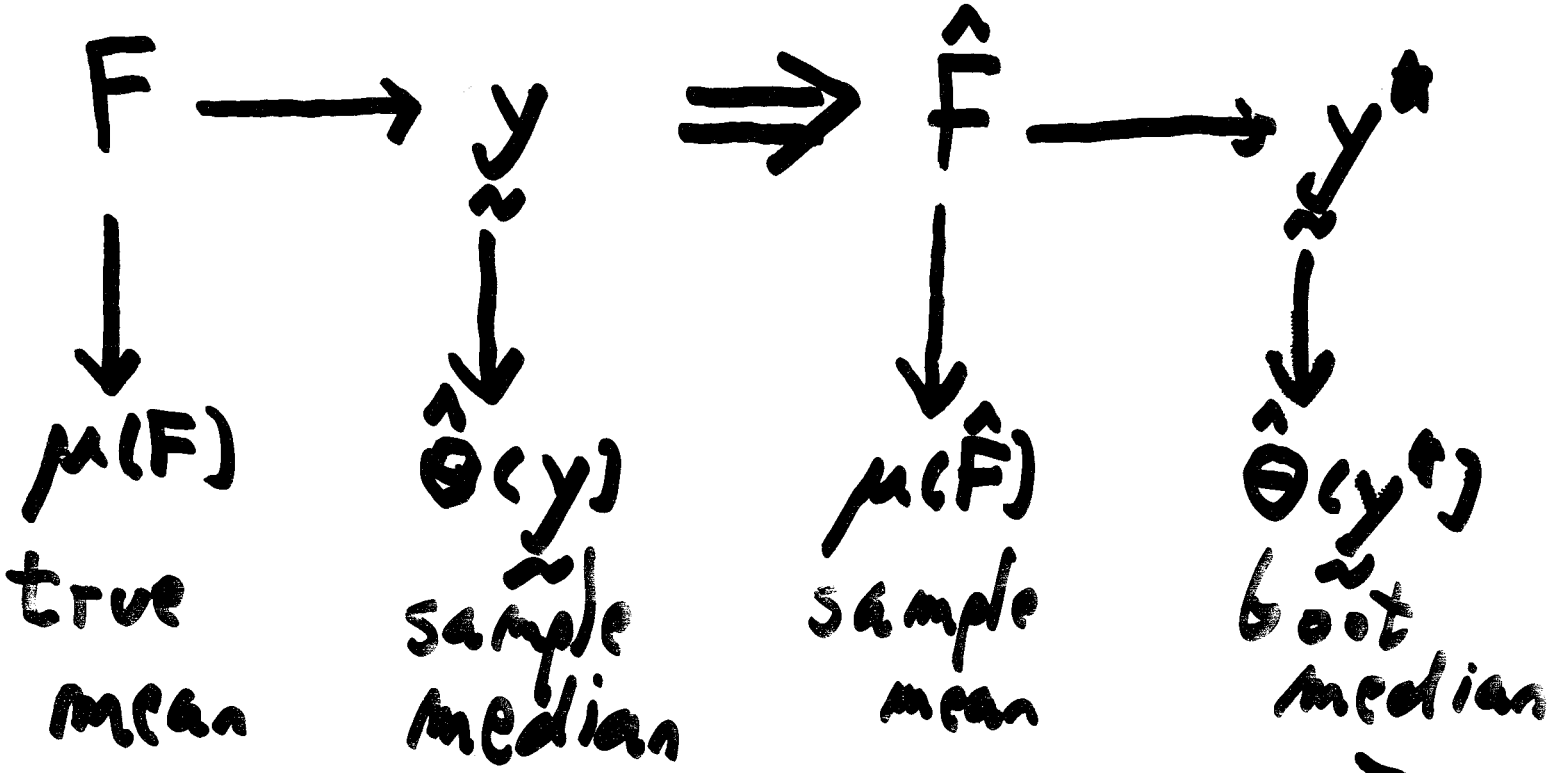


Estimate $\sigma = SD_F\{\hat{\theta}(y)\}$ by

$\hat{\sigma} = SD_{\hat{F}}\{\hat{\theta}(y^*)\}$

Usually need Monte Carlo to evaluate $\hat{\sigma}$

More Complicated Measures of Error



$$\text{BIAS} = E_F\{\hat{\theta}(\underset{\sim}{y}) - \mu(F)\} = ?$$

$$\widehat{\text{BIAS}} = E_{\hat{F}}\{\hat{\theta}(\underset{\sim}{y}^*) - \mu(\hat{F})\}$$

$$= \hat{\theta}^*(\cdot) - \mu(\hat{F})$$

$\hat{\theta}^*(\cdot)$: Average Boot Median

$\mu(\hat{F})$: sample mean

OR... $T = \frac{\bar{X} - \mu}{\bar{\sigma}}$ ← true mean of F

90% Conf. Int. for μ is

$$[\bar{x} - \bar{\sigma} T^{.95}, \bar{x} - \bar{\sigma} T^{.05}]$$

↑ %ile of T ↗

Instead of assuming $T^{.05} = t_{.14}$

can estimate T %ile by

bootstrapping $T^* = \frac{\bar{x}^* - \bar{x}}{\bar{\sigma}^*}$ ← true mean of $\hat{F}$

Estimate $P_F\{T < c\}$ by $P_{\hat{F}}\{T^* < c\}$

More Complicated Data Structures

So far always assume

$$F \longrightarrow (x_1, x_2, \cdots x_n) = \underset{\sim}{y}$$

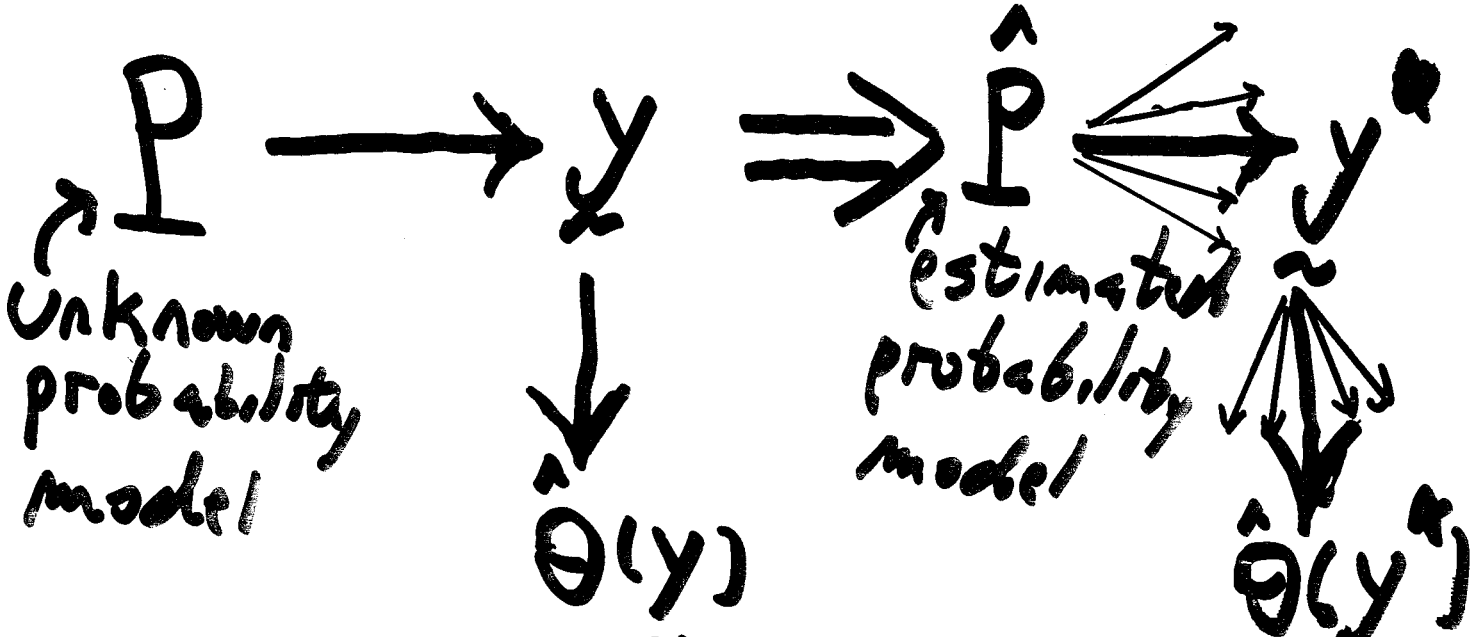
by simple random sampling from F.

Often have more general ways of generating data from models,

$$P \longrightarrow \underset{\sim}{y}$$

- two sample
- k sample
- ANOVA
- Regression
- Time Series
- Finite Pop.
- Censoring
- Missing Data
- Spatial Sampling

Schematic Picture:



Can still use bootstrap idea:

Estimate $\sigma(P) = SD_P\{\hat{\theta}(y)\}$

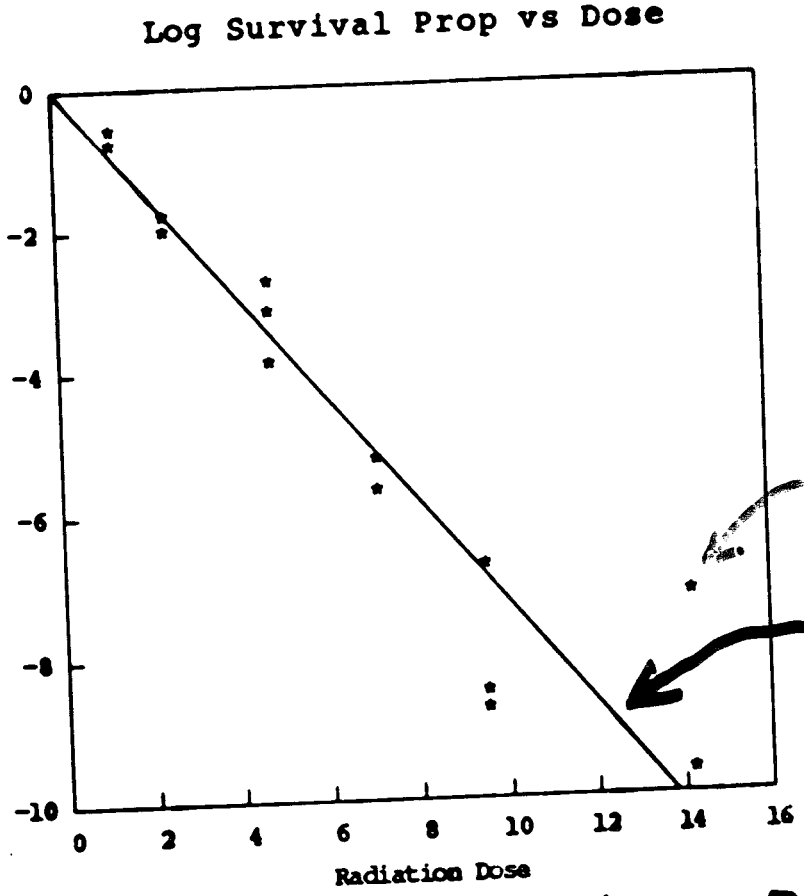
by $\hat{\sigma} = \sigma(\hat{P}) = SD_{\hat{P}}\{\hat{\theta}(y^*)\}$

Using Monte Carlo to evaluate $\hat{\sigma}$.

Crucial Step is "$\Longrightarrow$"

Estimating entire prob. model P

L.A.D. Regression



- 14 points (x,y) $\begin{cases} x = \text{radiation dose} \\ y = \text{log survival prop.} \end{cases}$
- Fit regression line $y_i = \beta x_i + e_i$ by
  LAD: minimize $\Sigma |y_i - \beta x_i|$
- $\hat{\beta} = -.725 \pm ?$

sampling distribution of errors

random sample from F

empirical distribution of $\hat{e}_i = y_i - \hat{\beta} x_i$

random sample from $\hat{F}$

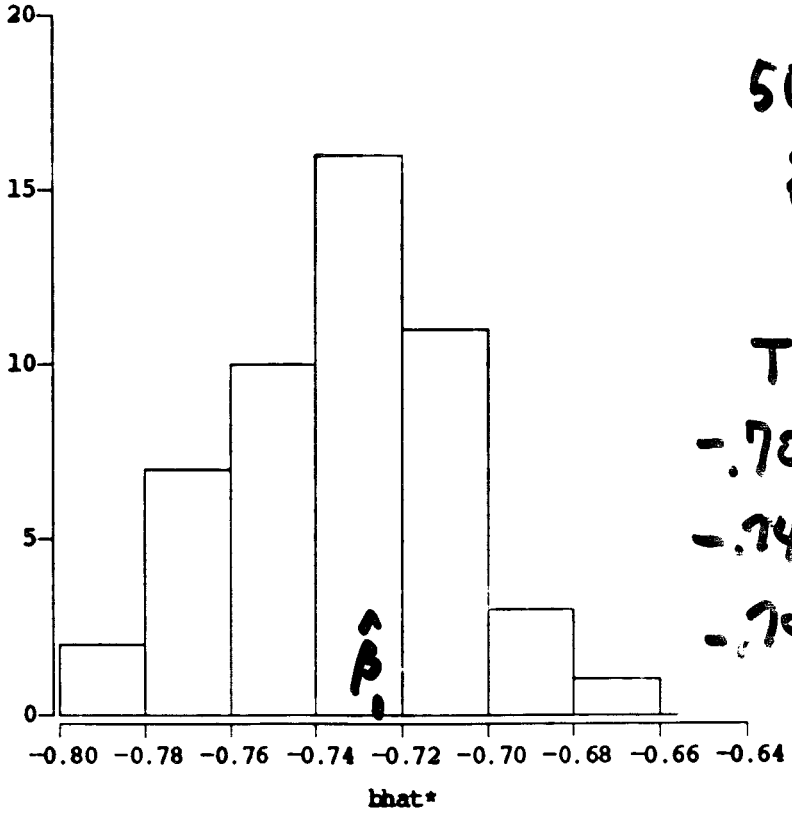
$$P = (\beta, F) \rightarrow y_i = \beta x_i + e_i \Rightarrow \hat{P} = (\hat{\beta}, \hat{F}) \rightarrow y_i^* = \hat{\beta} x_i + e_i^*$$

$i = 1, 2, \cdots, 14$ $\qquad$ $i = 1, 2, \cdots, 14$

$\hat{\beta}$ $\qquad$ $\hat{\beta}^*$

LAD from original data $\qquad$ LAD from bootstrap data

histogram of bootstraps



50 bootstraps gave

$\hat{\sigma}_{50} = .027$

The 1st 10 $\hat{\beta}^*$'s:

−.707, −.762, −.725, −.723,
−.749, −.720, −.744, −.725,
−.745, −.770 ~ ~ ~

Sunspot Data

$\underset{\sim}{y} = (x_{1770}, x_{1771}, \cdots, x_{1889})$

x_{1770}: # Sunspots in 1770 ... x_{1889}: # Sunspots in 1889

Probability Model $X_i = \phi X_{i-1} + \epsilon_i$

$\epsilon_i \overset{iid}{\sim} F$ for $i = 1771, \cdots 1889$

(x_{1770} a fixed covariate)

$$P = (\phi, F)$$

Usual Estimate of ϕ :

Minimize $\sum [x_i - \hat{\phi} x_{i-1}]^2$

$\hat{\phi} = .815 \pm .053$ ← $1/\sqrt{\text{Fisher Info.}}$

Assuming $F = N(0, \sigma^2)$

Obvious Estimate of F:

$\hat{F}$: Empirical Distribution of the n values $\hat{\epsilon}_i = x_i - \hat{\phi} x_{i-1}$

(Hope)

So .... $\underline{\hat{P}} = (\hat{\phi}, \hat{F})$

Generating Bootstrap Sunspot Data

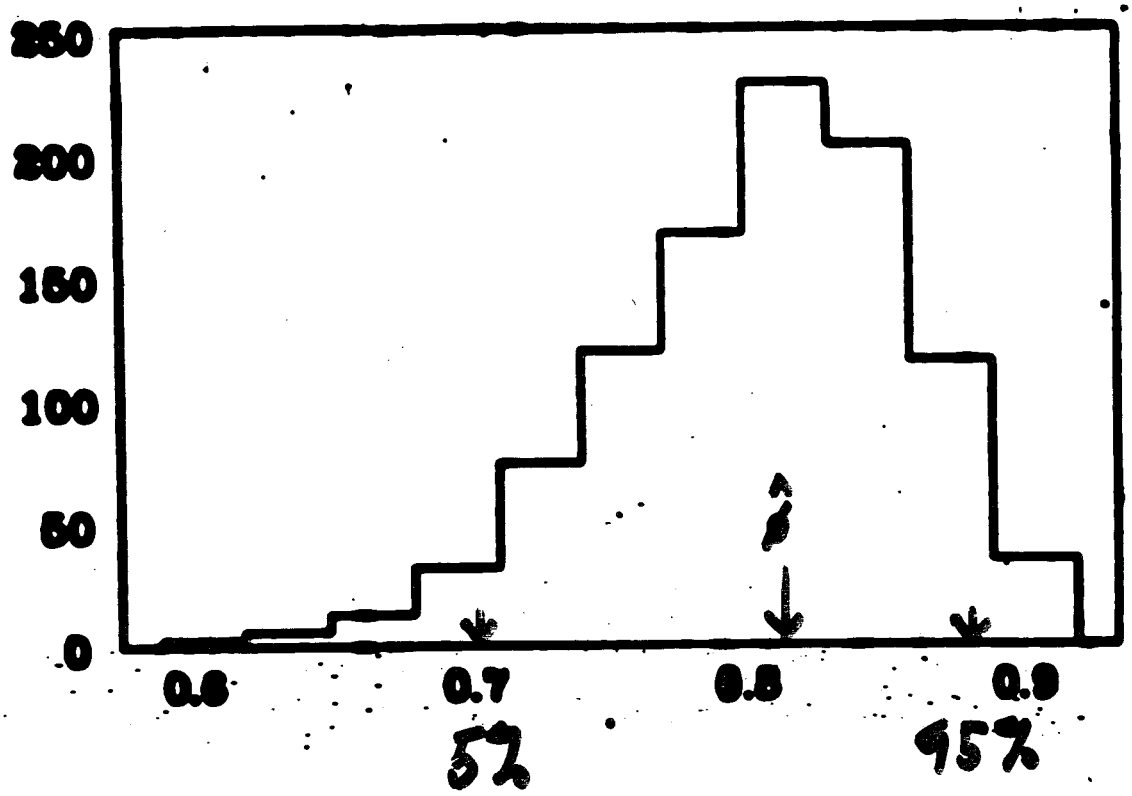
- Start with first value x_{1770}
- $X_i^* = \hat{\phi} X_{i-1}^* + \epsilon_i^*$ $\quad i = 1771, \ldots 1869$

  where $\epsilon_i^* \sim \hat{F}$
- $\hat{\phi}^*$ $\quad$ Minimize $\sum [x_i^* - \hat{\phi}^* x_{i-1}^*]^2$

$B = 1000$ Bootstraps of $\hat{\phi}^*$ gave

$$\hat{\sigma}_{1000} = .055$$

Almost the same as Normal Theory



However Histogram of bootstrap $\hat{\phi}$'s very nonnormal.

Suggests that $\hat{\phi} \pm z^{(\alpha)} \hat{\sigma}$

NOT a good confidence interval

"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