

The Bootstrap, and Bayes Too

Bradley Efron

Stanford

(1)

2

Standard Errors

- $F \xrightarrow{\text{i.i.d.}} \mathbf{x} = (x_1, x_2, \dots, x_n) \rightarrow \hat{\theta} = t(\mathbf{x})$
- How variable is $\hat{\theta}$?
- *Bootstrap Sample* $\hat{F} \rightarrow \mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$

$$\begin{array}{ccc} \mathbf{x}^{*1}, & \mathbf{x}^{*2}, & \dots, & \mathbf{x}^{*B} \\ \downarrow & \downarrow & & \downarrow \\ \hat{\theta}^{*1} = t(\mathbf{x}^{*1}) & \hat{\theta}^{*2} = t(\mathbf{x}^{*2}) & & \hat{\theta}^{*B} = t(\mathbf{x}^{*B}), \end{array}$$

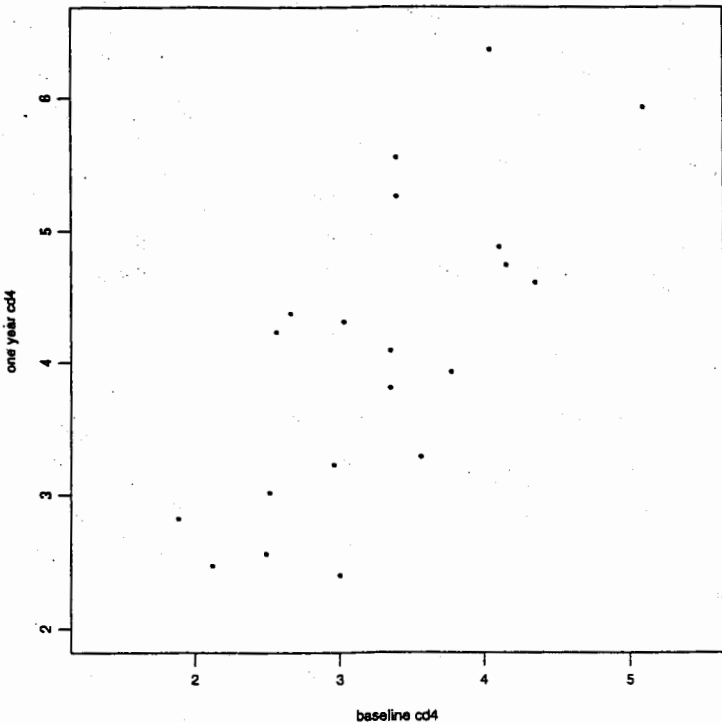
the bootstrap replications

- Empirical Standard Deviation of $\hat{\theta}^*$ values is Bootstrap estimate of standard error

3

4

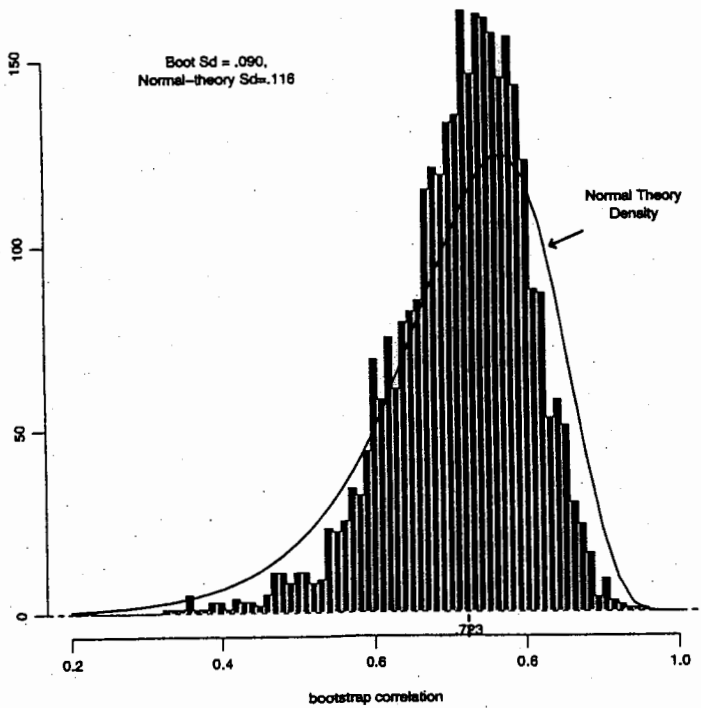
CD4 counts for 20 patients, at baseline and one year, Sample Correlation = .723



CD4 Example

- $n = 20$ bivariate points, CD4 measurements $x = (\text{baseline}, \text{one year})$.
- F is unknown bivariate distribution.
- $\hat{\theta} = \text{Sample Correlation Coefficient}$
- $\hat{\theta} = .723 \pm ?$
- Normal Theory Standard Error = .116
- $B = 3200$ bootreps gave $\hat{s}e_B = .090$.

3200 nonparametric bootstrap replications for the cd4 correlation coefficient



Bootstrap Histogram

- Histogram of 3200 $\hat{\theta}^*$'s is narrower than Normal-theory density

- Don't need $B = 3200$ for \hat{se} :

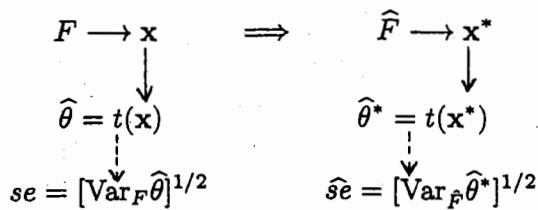
B: 100 200 400 800 1600 3200

\hat{se} : .082 .085 .090 .089 .091 .090

- *Later Use histogram for confidence intervals*

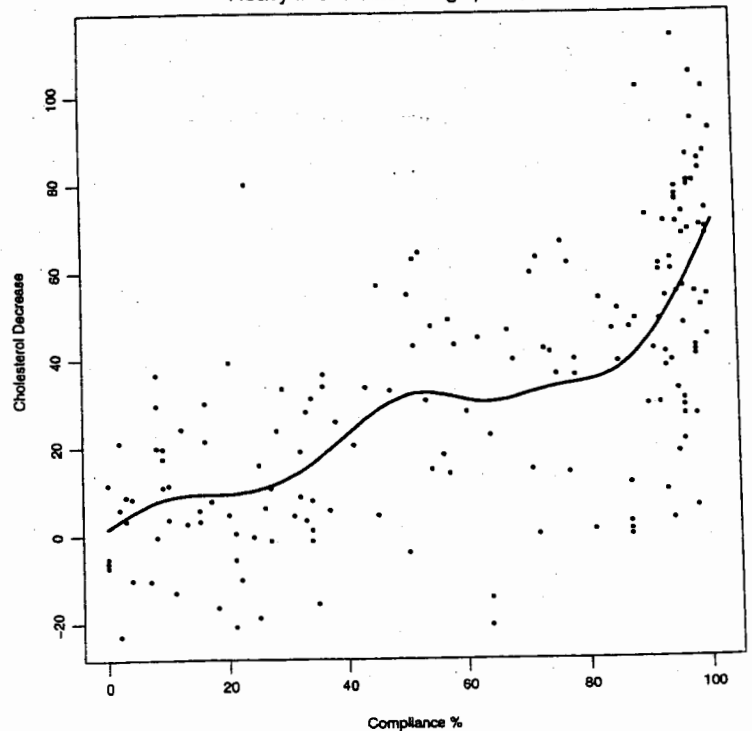
Real World

Bootstrap World

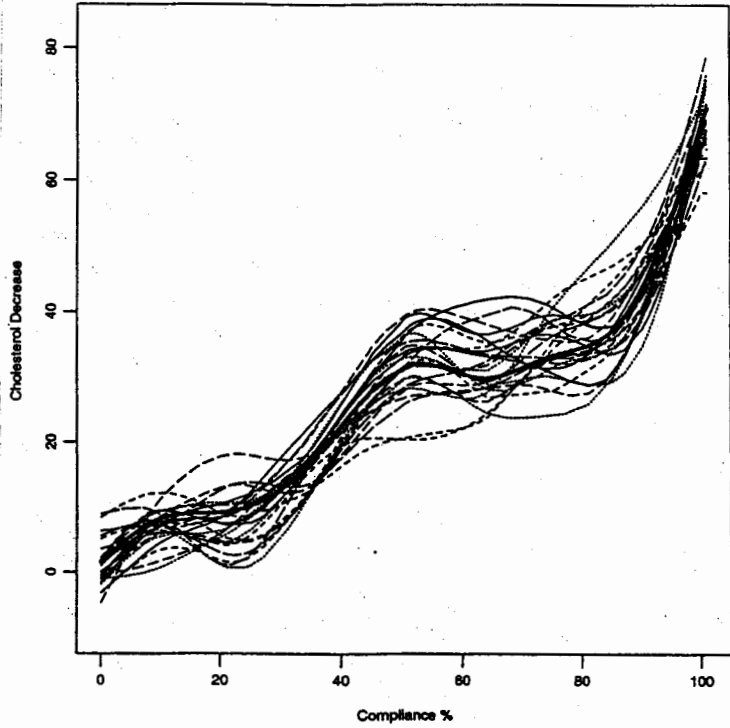


- Inference "Plug in" F for \hat{F}
- Parametric Bootstrap Take " \hat{F} " to be MLE among bivariate normal F .
- Works for "complicated" statistics $t(\cdot)$

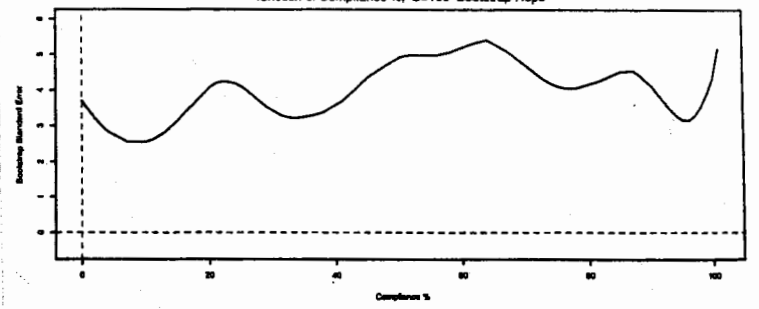
Cholesterol Decrease vs Compliance, 165 Subject; Heavy line is smoothing spline, df=7



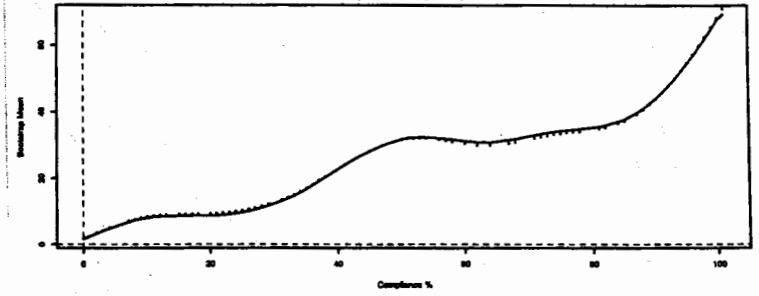
First 25 Nonparametric Bootstrap Replications of the Smoothing Spline



Bootstrap Standard Errors for Smoothing Spline fit, as a function of Compliance %; B=100 Bootstrap Reps



Boot mean vs Compliance (solid curve) Compared with original curve (points)



Bias Estimation

$$\begin{array}{ccc}
 F \rightarrow x & \Rightarrow & \hat{F} \rightarrow x^* \\
 \downarrow & & \downarrow \\
 \theta = t(F) & & \hat{\theta} = t(\hat{F}) \\
 \downarrow & & \downarrow \\
 \hat{\theta} = t(x) & & \hat{\theta}^* = t(x^*) \\
 \downarrow & & \downarrow \\
 \beta_F = E_F\{\hat{\theta}\} - \theta & & \hat{\beta} = E_{\hat{F}}\{\hat{\theta}^*\} - \hat{\theta}
 \end{array}$$

• Boot estimate

$$\hat{\beta}_B = \frac{1}{B} \sum \hat{\theta}^{*b} - \hat{\theta}$$

The Standard Intervals

- Observe $x \rightarrow \hat{\theta}, \hat{\sigma}$
MLE Estimated Sterr

(For corr $\hat{\sigma} = (1 - \hat{\theta}^2) / \sqrt{n - 3}$)

- Standard 90% Interval

$$\boxed{\hat{\theta} \pm 1.645 \cdot \hat{\sigma}}$$

- symmetric about $\hat{\theta}$

- Assumes $\hat{\theta} \sim N(\theta, \hat{\sigma}^2)$
Normal Unbiased constant sterr

- Accuracy $.05 + c/\sqrt{n}$

Bootstrap Intervals

(Efron '87 *JASA*)

- *Percentile Method* $\theta \in [\hat{\theta}^{*(.05)}, \hat{\theta}^{*(.95)}]$,

5th and 95th percentiles of bootstrap histogram.

- *BCA* Instead use modified percentiles

$$\Phi\left(z_0 + \frac{z_0 \pm 1.645}{1 - a(z_0 \pm 1.645)}\right)$$

bias correction acceleration

(Need $B > 1000$ bootreps)

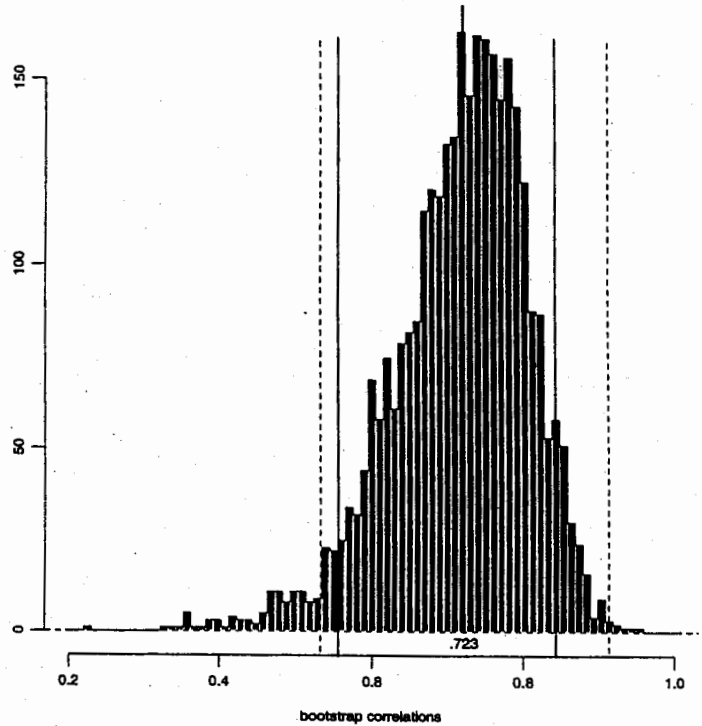
Transformation Invariant

Second Order Accurate $.05 + c/n$

(4)

14

Bootstrap BCA (Solid) and Standard (Dashed)
90% Confidence intervals for the CD4 correlation



15

Bootstrap-T Intervals

- Define "generalized student's t "

$$T = (\hat{\theta} - \theta) / \hat{\sigma}$$

- 90% interval

$$\theta \in [\hat{\theta} - \hat{\sigma}T^{(.95)}, \hat{\theta} - \hat{\sigma}T^{(.05)}]$$

- Bootstrap estimate of T percentiles:

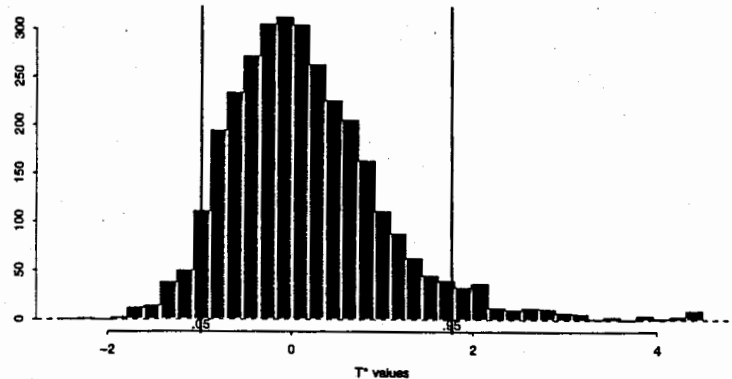
$$x^* \rightarrow (\hat{\theta}^*, \hat{\sigma}^*) \rightarrow T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$$

- Bootstrap- T Interval

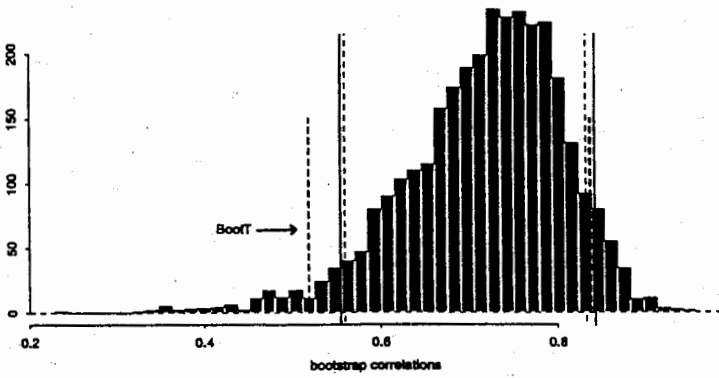
$$[\hat{\theta} - \hat{\sigma}T^{*(.95)}, \hat{\theta} - \hat{\sigma}T^{*(.05)}]$$

16

Bootstrap- T distribution for cd4 correlation data



Compare BCA (solid), ABC (dashed), and Boot-T 90% Limits



ABC Method

- Approximates BCA endpoints analytically using numerical second derivatives of $t(\mathbf{x})$
- Requires only $2n + 4$ recomputations of " $\hat{\theta}$ "
- Works for smooth statistics $\hat{\theta} = t(\mathbf{x})$
- Standard: $(\hat{\theta}, \hat{\sigma})$ • ABC: $(\hat{\theta}, \hat{\sigma}, \hat{z}_0, \hat{a}, \hat{c})$
- Need to write $\hat{\theta}^*$ in "resampling form", i.e. as function of bootstrap weights

Discrete Situations

- Sample Space $\mathcal{X} = \{1, 2, 3, \dots, K\}$
- True Distribution $\mathbf{p} = (p_1, p_2, \dots, p_K)$
- Random Sample $(x_1, x_2, \dots, x_n) \rightarrow \hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_K),$

$$\hat{p}_k = \#\{x_i = k\} / n$$

- Multinomial $\hat{\mathbf{p}} | \mathbf{p} \sim (\mathbf{p}, \mathbb{X} / n)$
- \uparrow \uparrow
 mean cov

where

$$\mathbb{X}_{ii} = p_i(1 - p_i) \text{ and } \mathbb{X}_{ij} = -p_i p_j$$

Objective Bayes Inference

- Prior $g(\mathbf{p}) \propto \prod_{k=1}^K p_k^{a-1}$ with $a \rightarrow 0.$
- Posterior $\boxed{\mathbf{p} | \hat{\mathbf{p}} \sim (\hat{\mathbf{p}}, \hat{\mathbb{X}} / (n + 1))}$
- Bootstrap $\mathbf{p} \xrightarrow{\text{real}} \hat{\mathbf{p}} \xrightarrow{\text{boot}} \hat{\mathbf{p}}^*$
- and $\boxed{\mathbf{p}^* | \hat{\mathbf{p}} \sim (\hat{\mathbf{p}}, \hat{\mathbb{X}} / n)}$
- For $\theta = t(\mathbf{p}) : \theta | \hat{\mathbf{p}} \sim \hat{\theta}^* | \hat{\mathbf{p}}.$

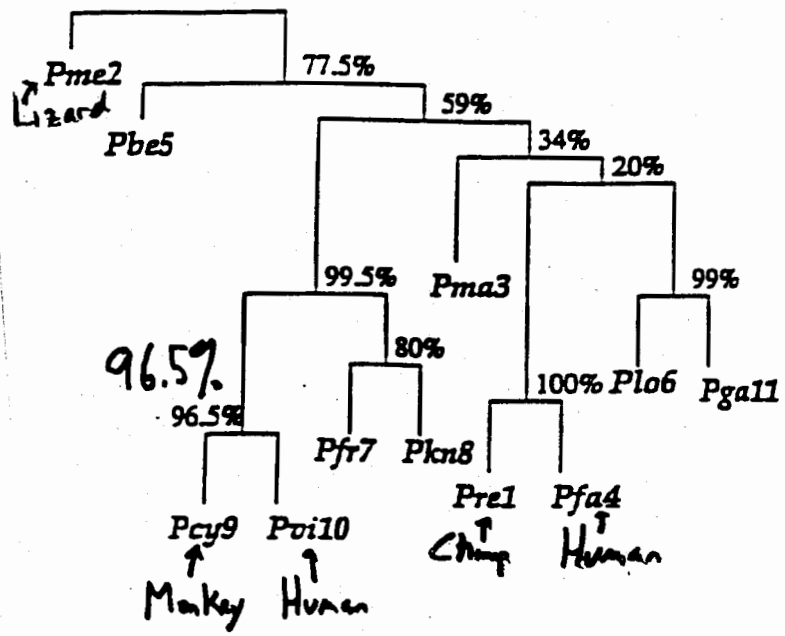
The Malaria Data

(Efron, Halloran, & Holmes, PNAS '96)

Species	Site: 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1 Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	A	T	T	C	T	T	A	G	A	T	A
2 Pse (Lizard)	T	C	T	A	A	A	A	G	A	T	T	A	T	A	T	A	G	A	T	T	A
3 Psa (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T	T	A	G	A	T	T	A
4 Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	T	A
5 Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	C	A	C	A	A	A	T	A
6 Plo (Bird)	T	T	T	A	A	G	A	A	A	A	A	C	T	T	C	A	C	A	A	T	C
7 Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T	T	A	G	A	T	T	A
8 Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T	T	A	G	A	T	T	A
9 Pcy (Monkey)	C	T	T	A	T	G	A	A	A	A	T	T	C	T	T	A	G	A	T	T	A
10 Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T	T	C	G	A	T	T	A
11 Pgs (Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T	C	A	A	A	T	T	C

- $x = 11 \times 221$ Data Matrix
- Rows are 11 species of malaria parasite.
- Columns are aligned sites on DNA.
- Evolution?

Phylogenetic Tree

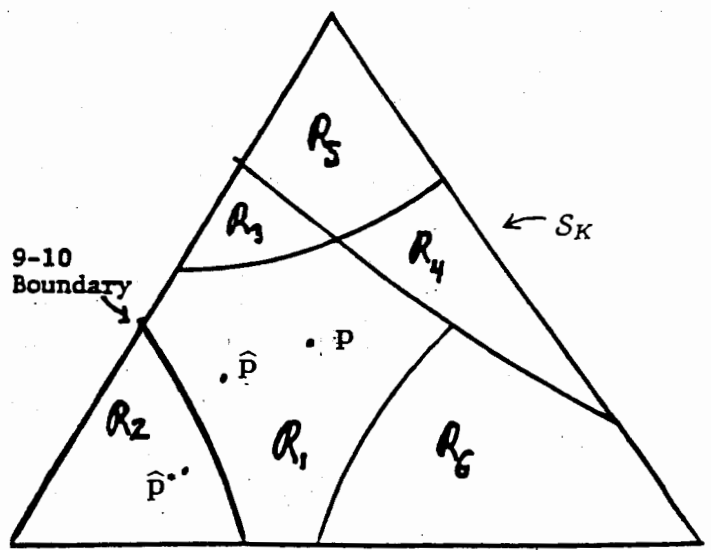


How Accurate Is The Tree? (Felsenstein, 1985)

- Original data matrix x is 11×221
- $x \rightarrow$ tree where " \rightarrow " is the tree-building algorithm "hclust"
- Bootstrap data matrix x^* : choose 221 columns of x , randomly and with replacement
- $x^* \rightarrow$ tree*
- Of 200 boot trees, 193 had "9-10"
- "Confidence value". $193/200 = 96.5\%$
- BC_a : 93.8%

Multinomial Representation

- Columns: Independent Multinomials, $K = 4^{11}$
- K -simplex partitioned into various tree regions



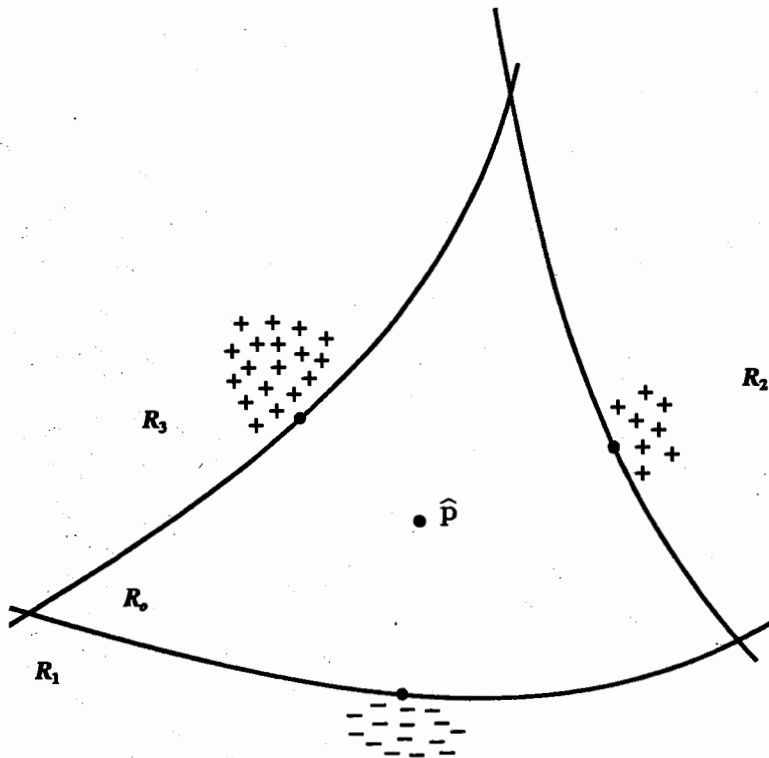
• BC_a : 93.8%

The Problems of Regions

(Efron and Tibshirani, '98 *Annals*)

- Observe \hat{p} , with expectation p
- Sample space partitioned into regions
- $\hat{p} \in R_0$
- How confident that $p \in R_0$?
- "BCA" modification of Felsenstein bootstrap

THE PROBLEM OF REGIONS



Uninformative Priors and Nuisance Parameters

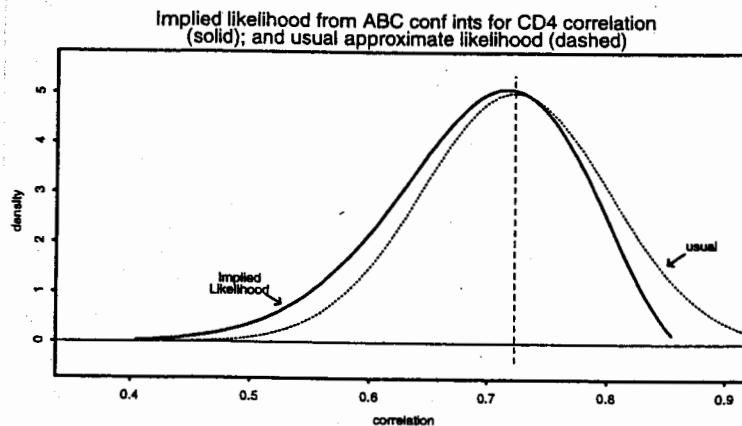
(Efron, '93 *Biometrika*)

- *Simple Case*: $L(\theta|x) = cf_{\theta}(x)$
- *Nuisance Parameters*: $L(\theta|x) = c \int f_{\theta,v}(x) \pi_0(\theta, v|x) dv$

where $\pi_0(\theta, v)$ is "uninformative prior".

- *Welch-Peers* $\pi_0 \Rightarrow$ A posteriori Conf. Intervals.

Boot Conf. Ints. \Rightarrow "Implied Likelihood" (Fiducial)



Some References

Efron and Tibshirani: "Introduction to the bootstrap", Chapman & Hall, 1993.

(statlib@lib.stat.cmu.edu with one-line message send bootstrap.funs from S)

Efron and Tibshirani: "The Problem of Regions" *Annals*, '98, 1687-1718.

Efron and DiCiccio: "Bootstrap Confidence Intervals" *Stat. Science* '96, 189-228.

Efron, Halloran, Holmes: "Bootstrap confidence levels for phylogenetic trees", *PNAS*, '96, 13429-13434.

Felsenstein: "Confidence Limits on Phylogenies: an approach using the bootstrap", *Evolution*, 1985, 783-791.

Efron: "Bayes and Likelihood Calculations from Confidence Intervals", *Biometrika*, 1993, 3-26.

Hypothetical Bayes Analysis for CD4 correlation: Implied Likelihood combined with Expert Prior $N(.5, .1^2)$ to give posterior distribution

