Computers, Bootstraps, and Statistics

Bradley Efron

Royal Statistical Society

September 16, 1994

Some Important Post-War Developments

- Nonparametric/Robust
- Kaplan-Meier/Proportional Hazards
- Empirical Bayes/James-Stein
- Logistic Regression/GLIM
- Jackknife/Bootstrap
- EM/Gibbs

Confidence Intervals

EXACT: Binomial, Poisson, Student-t, Fieller

STANDARD APPROXIMATE:

$$\widehat{\theta} \pm z^{(\alpha)}\widehat{\sigma}.$$

- Makes coverage errors of order $O(1/\sqrt{n})$ in each tail.

BETTER APPROXIMATE: Bootstrap, Saddlepoint

- Coverage errors $O(1/n)$ in each tail ("Second Order Accurate")
- Corrections $O(\widehat{\sigma}/\sqrt{n})$ to standard endpoints
- Student-t corrections are $O(\widehat{\sigma}/n)$

The Student Score Data

$n = 22$ students each took 5 tests:

A	B	C	D	E	A	B	C	D	E
63	63	65	70	63	42	60	54	49	33
53	61	72	64	73	40	63	53	54	25
51	67	65	65	68	23	55	59	53	44
44	69	53	53	53	49	49	45	48	39
42	69	61	55	45	17	53	57	43	51
31	49	62	63	62	39	46	46	32	43
44	61	52	62	46	48	38	41	44	33
49	41	61	49	64	46	40	47	29	17
30	69	50	52	45	30	34	43	46	18
36	59	51	45	51	12	30	32	35	21
56	40	56	54	35	?5	26	15	20	20

- $F \xrightarrow{\text{i.i.d.}} \mathbf{x}_{22\times 5} = (x_1, x_2, \cdots, x_{22})$
- $\Sigma\!\!\!\!-\!\!\!\!| = \text{Cov}_F\{x\}$
- Parameter of interest

$$\begin{aligned} \Theta &= \text{Maximum eigenvalue of } \Sigma\!\!\!\!-\!\!\!\!| \\ &\equiv s(F) \end{aligned}$$

Point Estimate for θ

- $\widehat{F}$ = empirical distribution (nonparametric MLE)

OR

- $\widehat{F}_{\text{norm}} = N_5(\widehat{\mu}, \widehat{\Sigma})$ (normal-theory MLE)
- $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})(x_i - \widehat{\mu})'$ $[\widehat{\mu} = \bar{x}]$

Both give the same MLE for θ,

$$\begin{aligned}\widehat{\theta} &= \text{MaxEigenvalue}(\widehat{\Sigma}) \\ &= s(\widehat{F}) = s(\widehat{F}_{\text{norm}})\end{aligned}$$

Bootstrap Confidence Intervals

- Bootstrap extends point estimate to an accuracy estimate. Relies on "bootstrap samples":

- $\widehat{F} \xrightarrow{\text{i.i.d.}} \mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$

- Bootstrap replication $\widehat{\theta}^* = s(\widehat{F}^*)$

$$\mathbf{x}^* \longrightarrow \widehat{F}^* \longrightarrow \widehat{\Sigma}^* \longrightarrow \widehat{\theta}^*$$

$$\begin{array}{cccc} \mathbf{x}^*(1), & \mathbf{x}^*(2), & \mathbf{x}^*(3), \cdots, & \mathbf{x}^*(B) \\ \searrow & \searrow & \searrow & \searrow \\ \widehat{\theta}^*(1) & \widehat{\theta}^*(2) & \widehat{\theta}^*(3), \cdots, & \widehat{\theta}^*(B) \end{array}$$

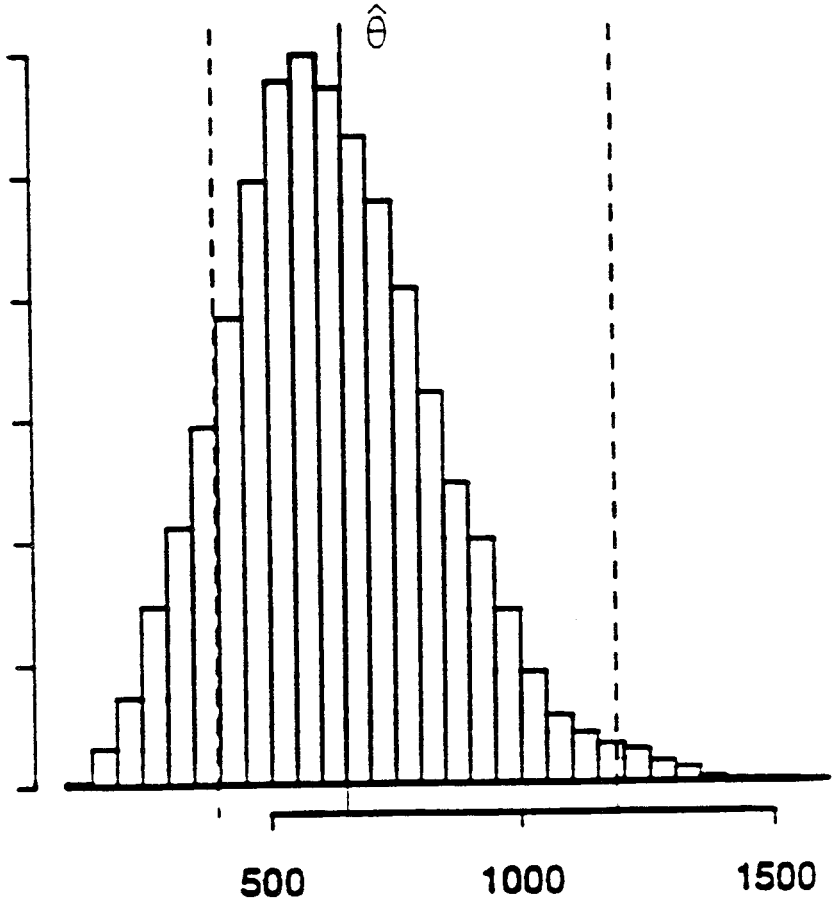
- Bootstrap standard error is empirical sterr of $\widehat{\theta}^*$ values ($B \approx 200$)

- We can use histogram of $\widehat{\theta}^*$ values to get second-order accurate confidence intervals for θ ($B \approx 2000$)

2000 Bootstrap Replications

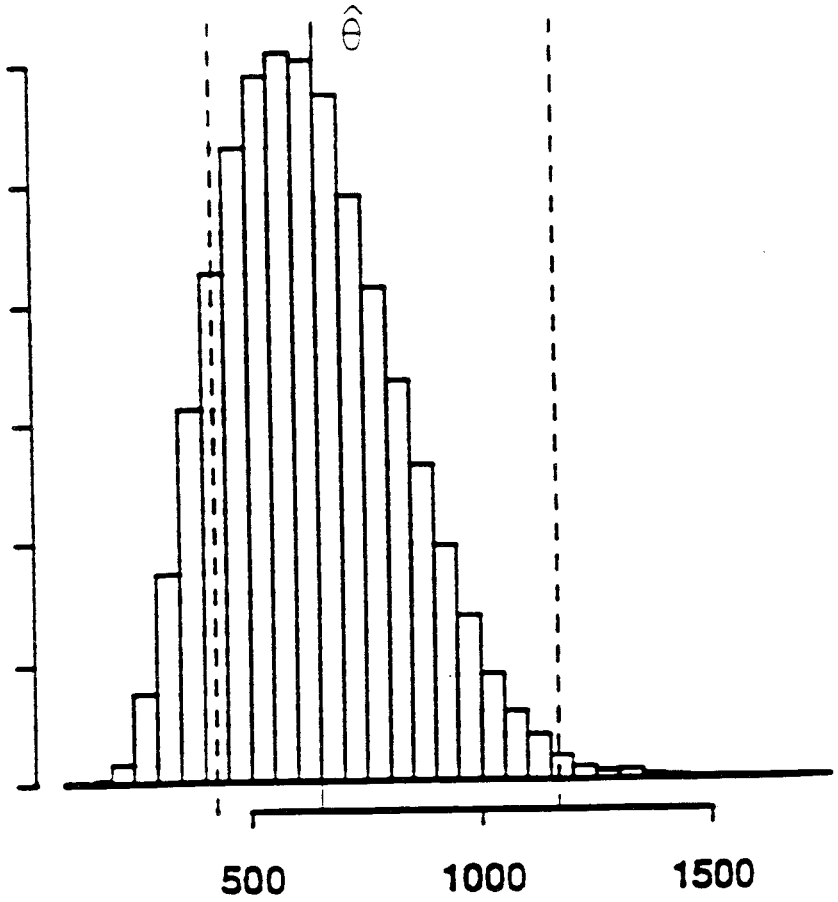
Nonparametric $\widehat{F}$ = empirical distribution

- Histogram skewed



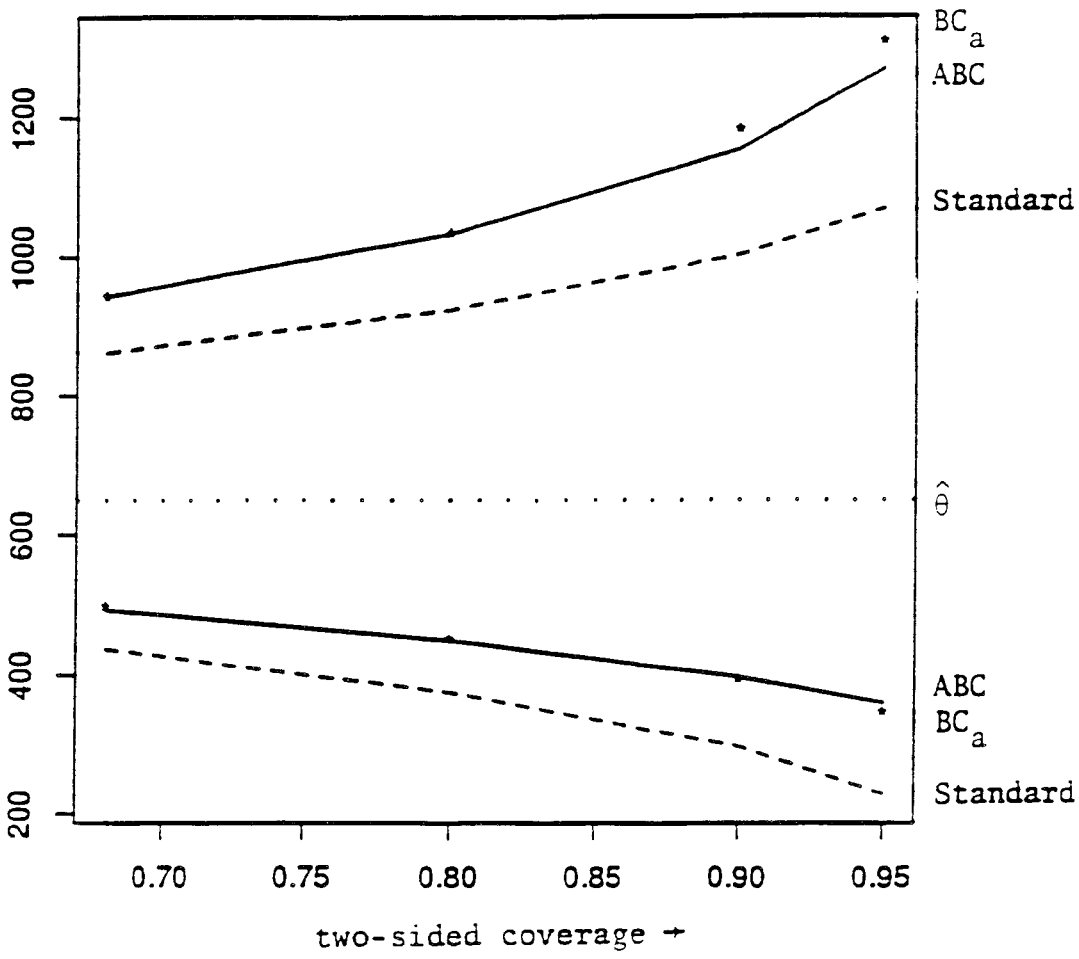
boot sterr=209.9

Parametric $\widehat{F} = N_5(\widehat{\mu}, \widehat{\Sigma})$



boot sterr=195.2

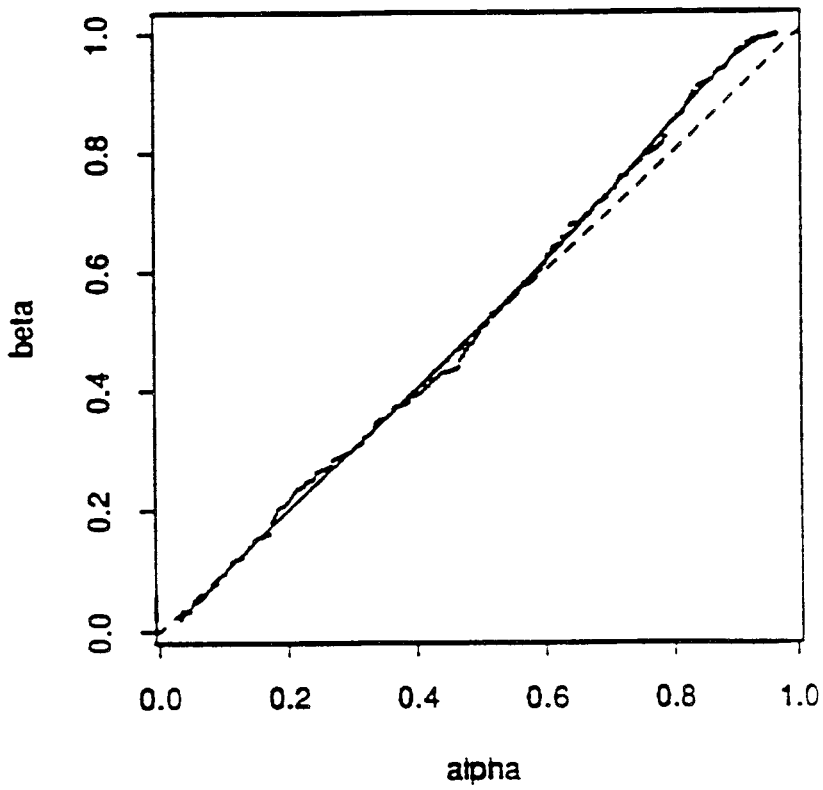
Nonparametric Intervals



- ABC is analytic version
- Can see the second-order corrections
- Efron and Tibshirani, "Introduction to the Bootstrap", Chapman & Hall

Calibration

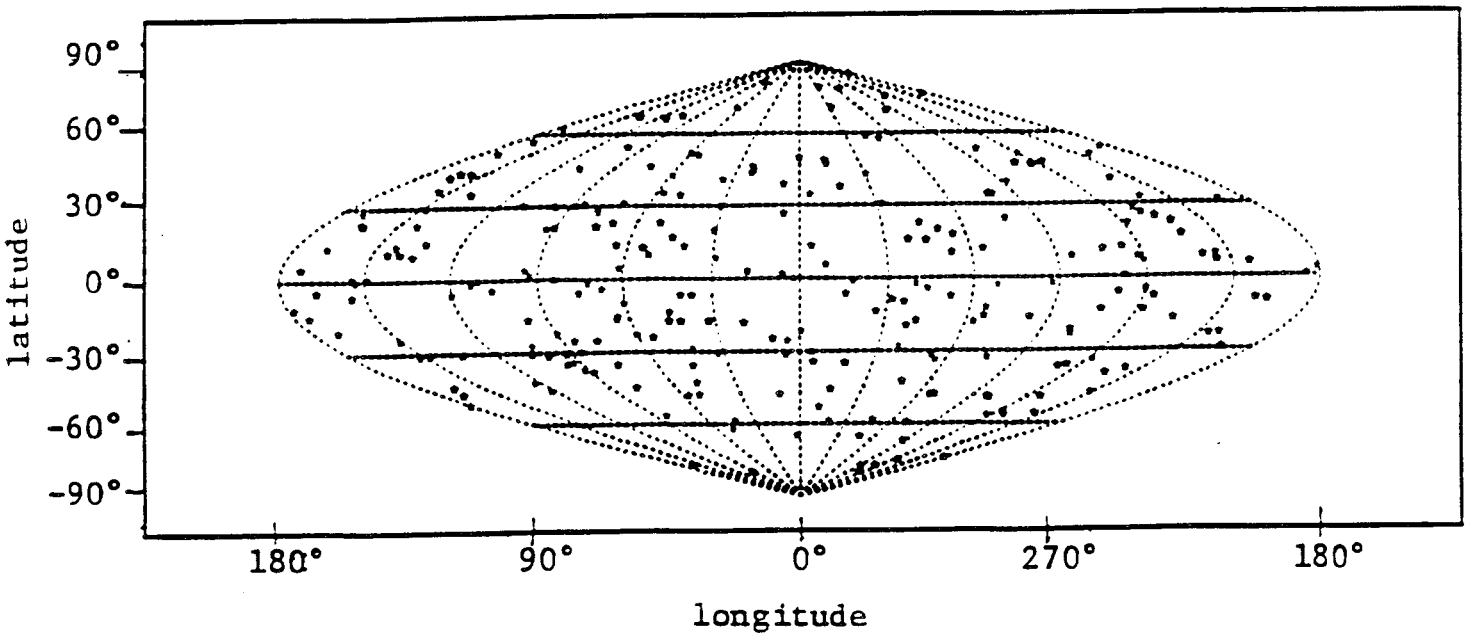
- $\alpha(\beta) = \text{Prob}_F\{\theta < \hat{\theta}[\beta]\}$ ← Calibration curve
- ↑ actual level α ↑ nominal level β
- $\hat{\alpha}(\beta) = \text{Prob}_{\hat{F}}\{\hat{\theta} < \hat{\theta}^*[\beta]\}$ ← Bootstrap calibration curve
- Nonparametric



- $\beta = .90$ gives $\alpha = .80$
- Can get 3rd Order Accuracy

The Gamma-Ray Burst Data

- **BATSE Recorded $n = 260$ bursts in its 1st year:**



- **Question:** Are bursts isotropic?
- Median angular error 5.66°
- Additional 325 bursts since then (but flawed)

The All-Angles Test

- Let $\mathbf{a} = (a_1, a_2, a_3, \cdots, a_N)$ be ordered pariwise angles

- $N = \binom{260}{2} = 33670$

- **Monte Carlo:** $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_{260}^*) \sim$ Uniform

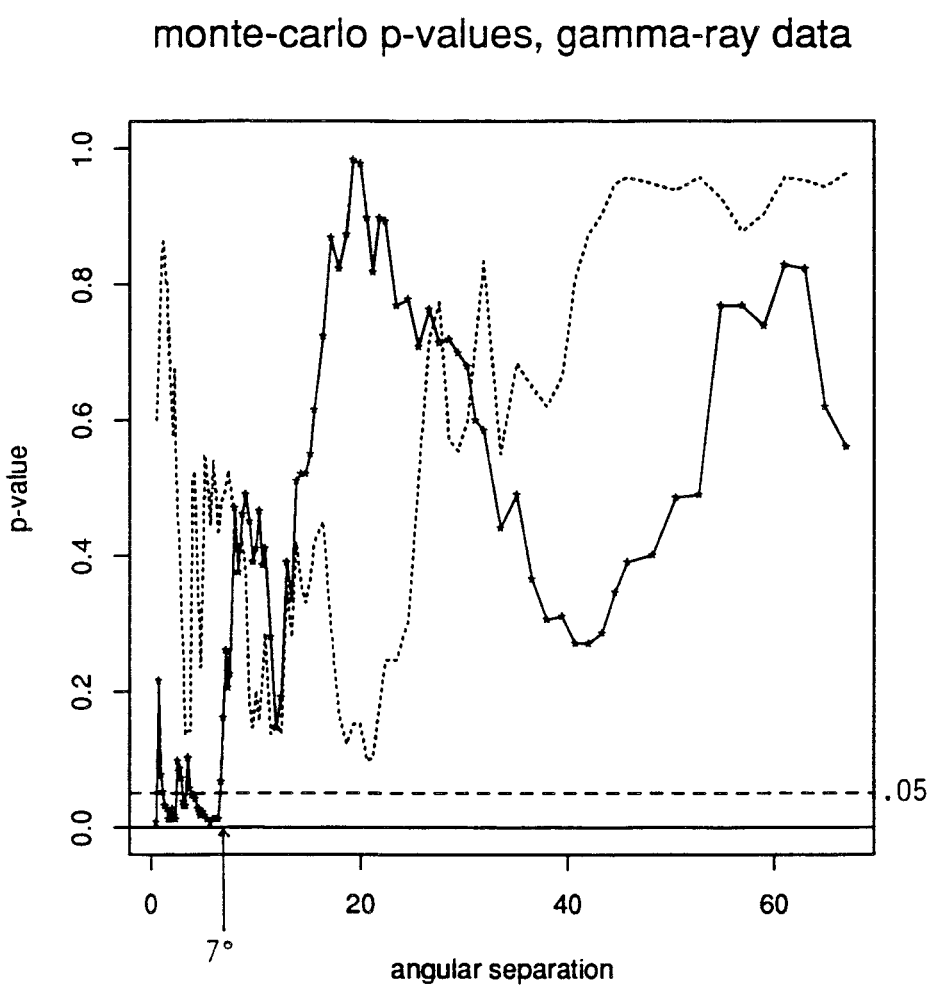
  gives $\mathbf{a}^* = (a_1^*, a_2^*, \cdots, a_N^*)$

- Computed $\mathbf{a}^*(1), \mathbf{a}^*(2), \cdots, \mathbf{a}^*(400)$

- P-value for kth coordinate is

  $p(k) =$ Proportion of a_k^* values less than a_k

- Plot versus $m(k) =$ Median $\{a_k^*(b), b = 1, 2, \cdots, 400\}$



- $p(k)$ small for $M(k) < 7°$
- But not for bigger data set!
- Barnard Monte Carlo tests

Simultaneous Test For Original 260

- Consider set of ordered angles $k \in K$
- Let $S(K) = \sum_{k \in K} \log(p(k))$
- Compare $S(K)$ with the 400 $S^*(K)$ (letting each $\mathbf{x}^*$ play the role of original data set)
- Gives simultaneous p-value $P(K)$;

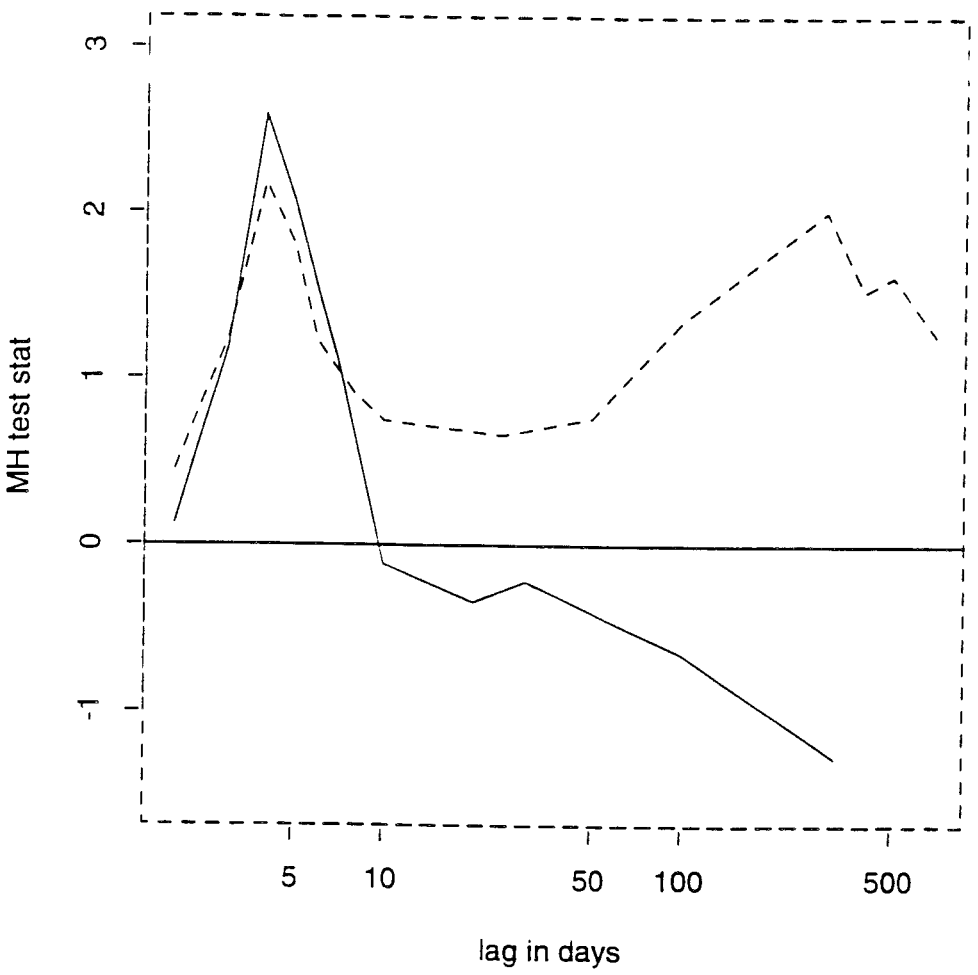
$\text{Max}_K[m(k)]$	$P(K)$
3.4°	.004
5.2°	.004
7.6°	.004
11.0°	.006
15.7°	.011
22.4°	.026

Time-Space Clustering

- Bursts ordered in time $t_1 < t_2 < t_3 \cdots < t_{260} \cdots$
- Want to test for clustering in (time, angle) space
- "Mantel-Haenszel" Test:
- Consider event at time t_j
- Assign score s_{jk} to each burst $k \geq j$
- Score measures how close in angle burst is to bursts occurring within previous d days
  - Compare actual score $S_j = s_{jj}$ with

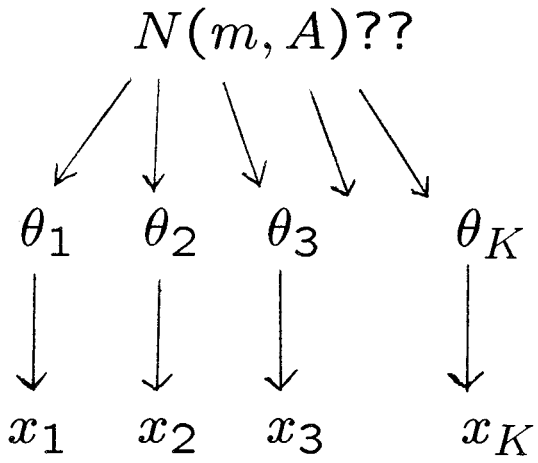
  $m_j = \text{mean}\{s_{jk}, k \geq j\}$

  - $\Sigma_j (S_j - m_j)/v_j^{1/2} \doteq N(0,1)$ under null hypothesis



- Statistic quite significant for $d = 4$ days
- Also for $d > 300$ days, all data
- Efron and Petrosian, June 1994, *JASA*

Empirical Bayes



$x_i \sim N(\theta_i, 1)$

- $E\{\theta_1|X_1\} = m + \frac{A}{A+1}(x_1 - m)$

James-Stein estimator

- $\widehat{\theta}_1 = \widehat{m} + \frac{\widehat{A}}{\widehat{A}+1}(x_1 - \widehat{m})$

  $\widehat{m} = \bar{x}$ and $\widehat{A} = (K-3)/\Sigma(x_k - \bar{x})^2$

The Ulcer Data

- 41 2×2 Tables

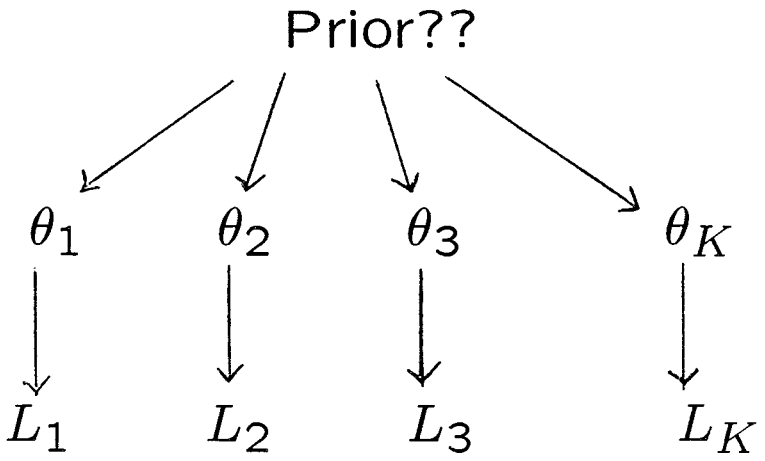
	Success	Failure
Treatment	a	b
Central	c	d

$$\widehat{\theta} = \log(ad/bc)$$

City	a	b	c	d	$\widehat{\theta}$
1	7	8	11	2	−1.84
3	5	29	4	35	0.41
8	1	15	13	3	−4.17
13	9	12	7	17	0.60
23	14	54	13	61	0.20

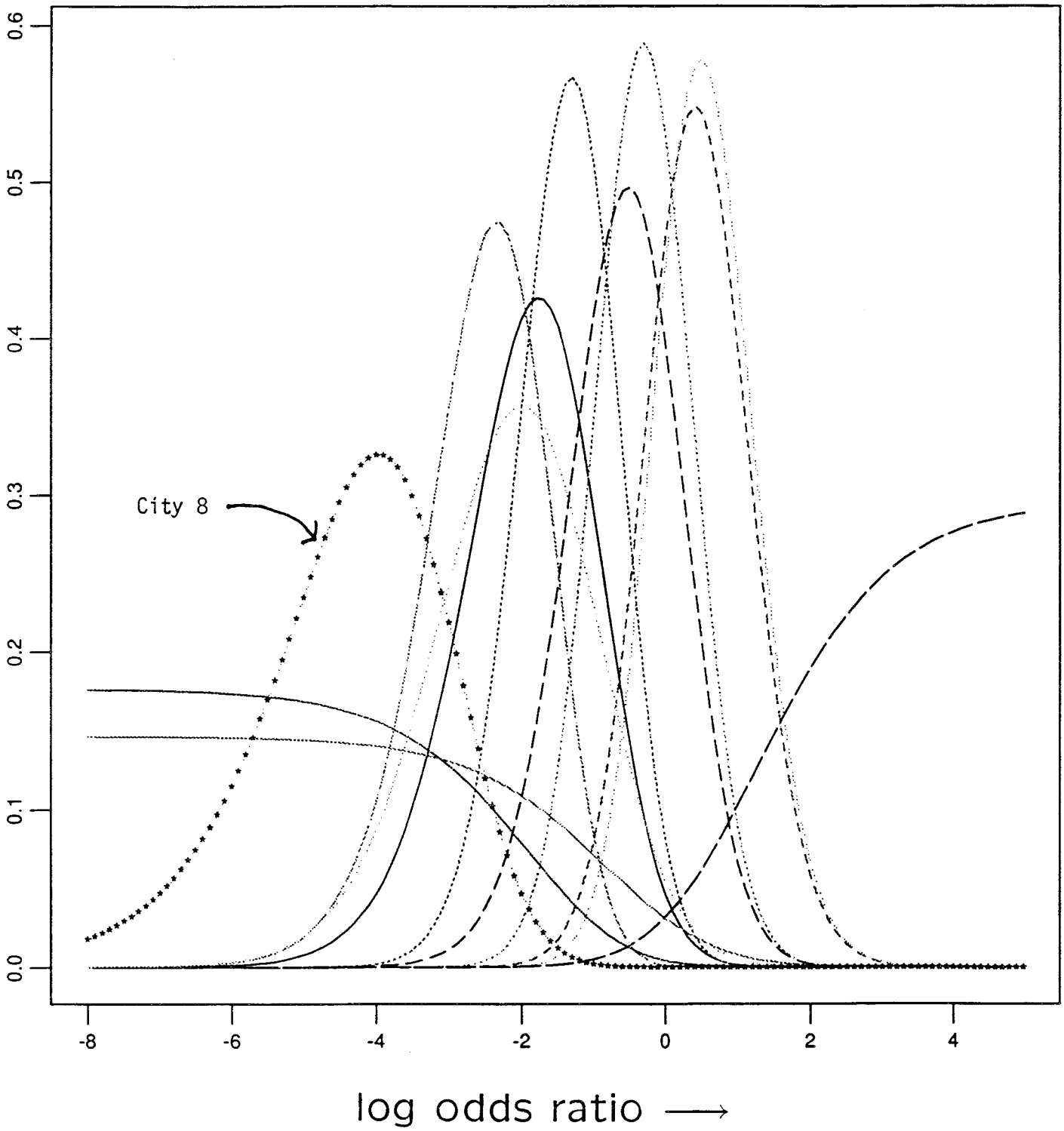
- Θ_k = True log odds ratio in City k

- $L_k = L(\theta_k | a_k, b_k, c_k, d_k)$ = hypergeometric likelihood



- **Empirical Bayes** Aposteriori interval for θ_1 given L_1 and "other" data $L_2, L_3, \cdots, L_K$

The First 12 of the 41 Likelihoods



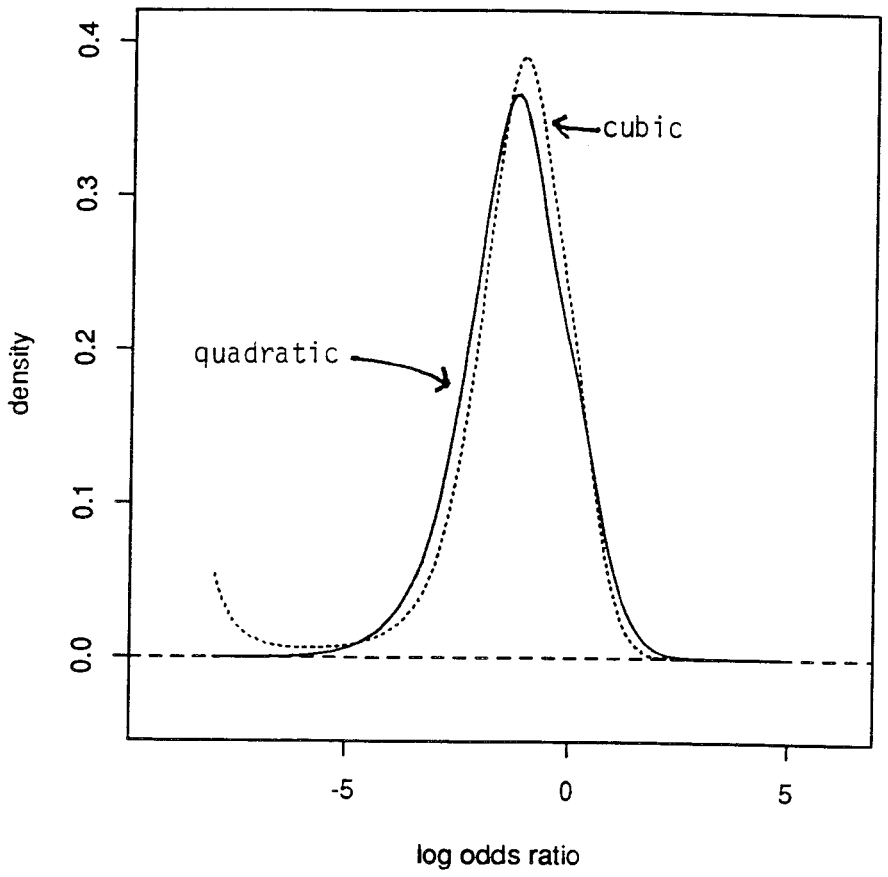
Special Exponential Families of Priors

- **Prior Density Family** $g_\eta(\theta) = g_0(\theta)e^{\eta' t(\theta) - c(\eta)}$

- η = unknown hyperparameter vector

- $t(\theta)$ vector of sufficient statistics,
  e.g. $t(\theta) = (\theta, \theta^2, \theta^3)$

- $g_0(\theta)$ = carrier • $c(\eta)$ = normalizing constant

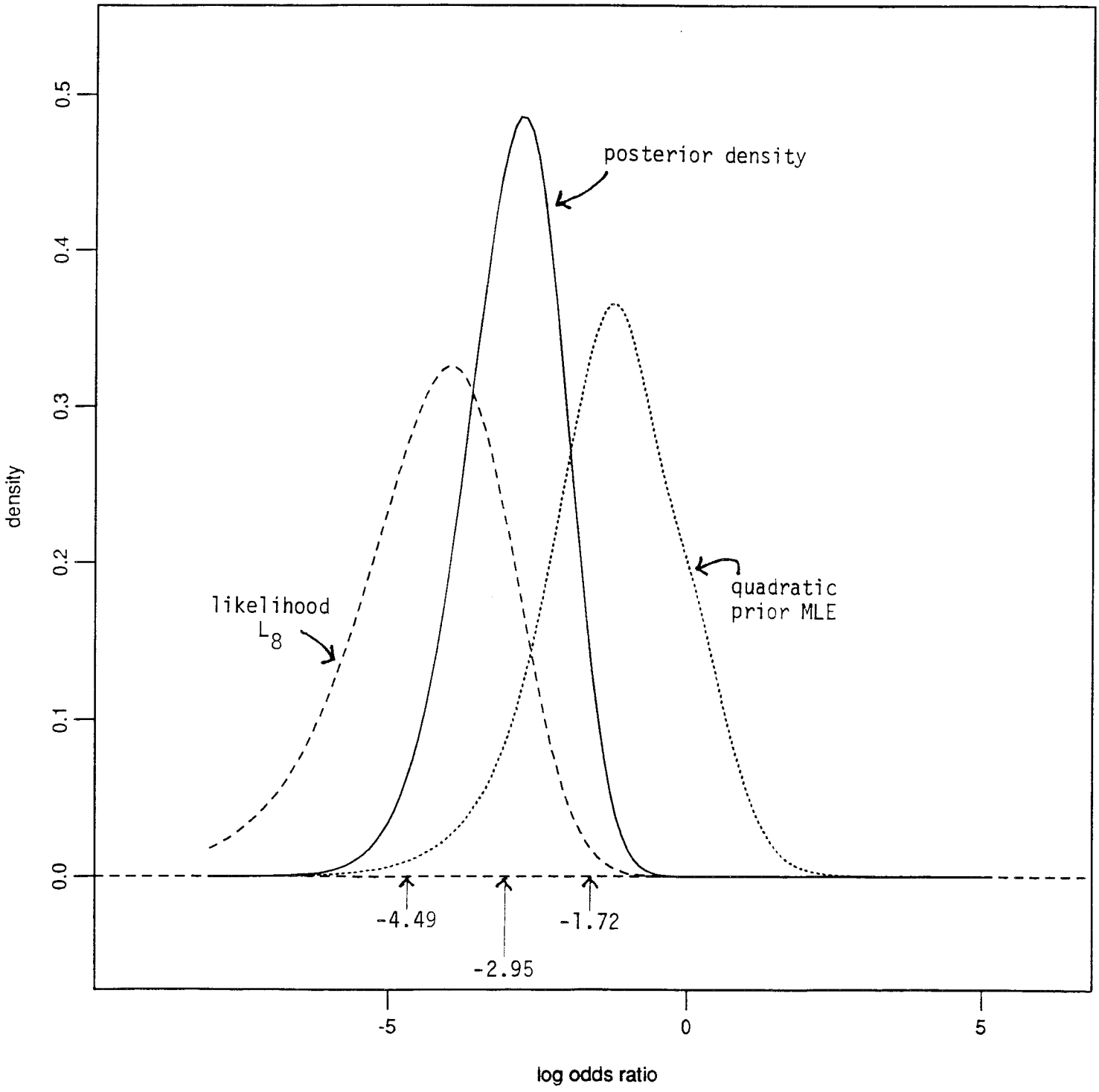


- quadratic
  $t(\theta) = (\theta, \theta^2)$

- cubic
  $t(\theta) = (\theta, \theta^2, \theta^3)$

- MLE priors $g_{\hat{\eta}}(\theta)$, with $g_0(\theta) = \bar{L}(\theta)$

Empirical Bayes Inference For City 8

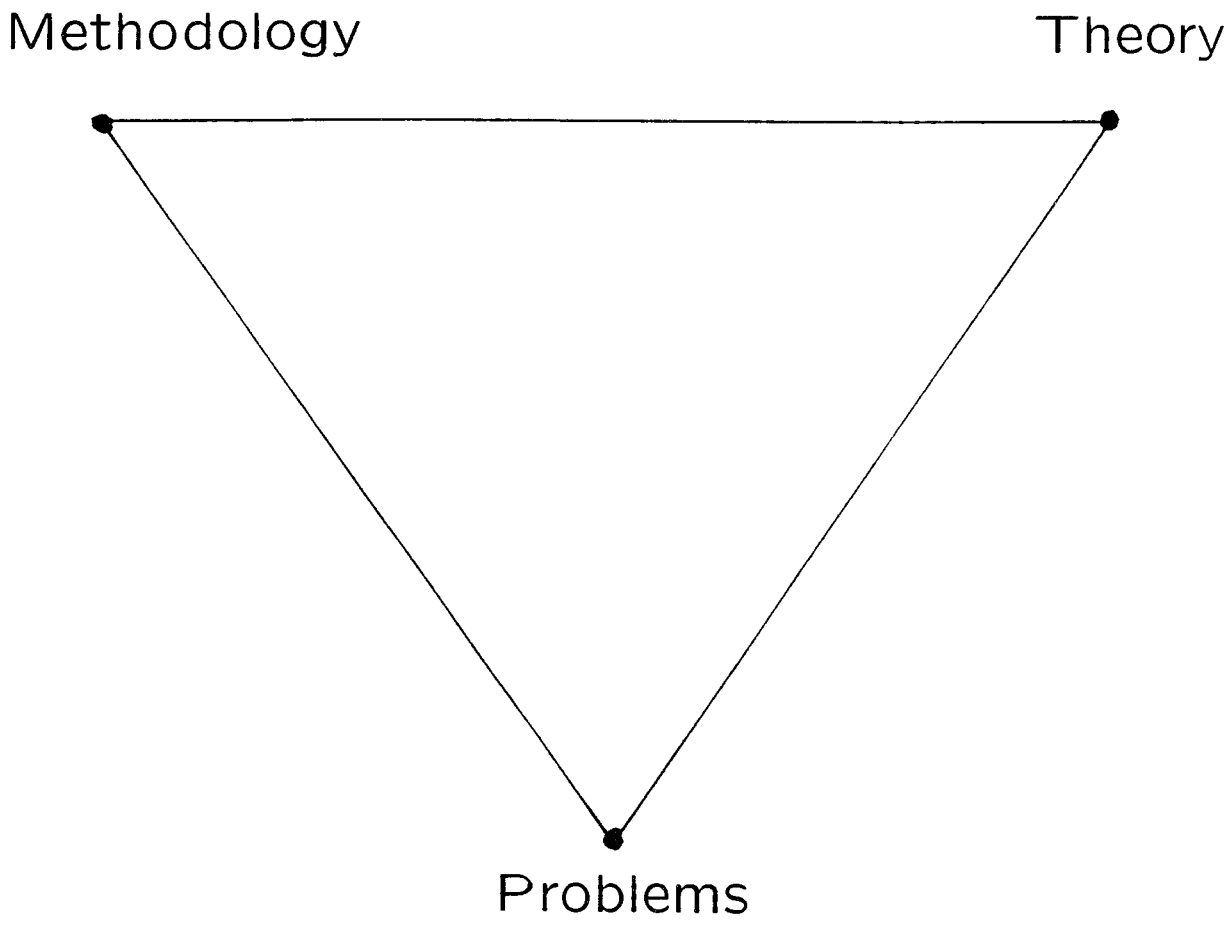


- **Bias-Correction** Correct for bias of MLE prior vis-a-vis vague hyperprior analysis.

- **Nuisance Parameters** Can get good approximate likelihood $L_k(\theta_k|x_k)$ free of nuisance parameters.

- **Numerical** No special matheamtical forms required. Uses ABC algorithm.

  Efron "Empirical Bayes Methods for Combining Likelihooods"


Methodology
Theory
Problems