

Discussion of

“... Indirect Evidence” by Brad Efron

and

“*f*-Divergences...” by Michael Jordan

L. Brown

For O-Bayes Conference, June 6, 2009

- These are two quite different papers
- Both are deep, challenging and worthy of extensive study  
(in combination with the papers on which they are based)

# Apology!

- Only a brief time is allotted to my discussion
- I'll mostly describe some things I've discovered about Bayes and empirical Bayes approaches to the standard hierarchical normal model as in Efron's p2, 8 – 10 & 26
- Thus leaving to others to respond to the heart of Efron's presentation *eg, Cai, Jin (2007,2009), Zhao (today)*
- But I will conclude with a question for Jordan

I'll begin by  
Considering Efron and Morris' Famous Baseball Players

---

## Eighteen Baseball Players (Efron and Morris, 1977)

Name	hits/AB	Observed Avg	"TRUTH"	James-Stein
1. Clemente	18/45	.400	.346	0.290
2. F. Robinson	17/45	.378	.298	0.286
3. F. Howard	16/45	.356	.276	0.281
4. Johnstone	15/45	.333	.222	0.277
⋮	⋮	⋮	⋮	⋮
14. Petrocelli	10/45	.222	.264	0.254
15. E. Rodriguez	10/45	.222	.226	0.254
16. Campaneris	9/45	.200	.286	0.249
17. Munson	8/45	.178	.316	0.244
18. Alvis	7/45	.156	.200	0.239
Grand Average		.265	.265	0.265

## How E & M got their J-S estimates

Let  $p_i$  denote true ability of  $i$ -th player.

Transform:

$$\text{Avg}_i \rightarrow X_i : X_i \approx N(\theta_i, \sigma_i^2)$$

$$\text{where } \sigma_i^2 = \frac{1}{4n_i} = \frac{1}{4 \times 45} \text{ and } \theta_i = \arcsin \sqrt{p_i}.$$

Apply J-S to  $\{X_i\}$ :

$$\tilde{\theta}_i = \bar{X} + \left( 1 - \frac{p-3}{\sum (X_i - \bar{X})^2 / \sigma_i^2} \right) (X_i - \bar{X}).$$

Invert transform:  $\tilde{p}_i = \sin^2 \tilde{\theta}_i.$

go back to table

J-S is Empirical Bayes  
Stein (1962), Efron & Morris (1970s)

- Write:

$$X_i \sim N(\theta_i, \sigma_i^2) \text{ with } \theta_i \sim N(\mu, \tau^2).$$

- Marginally:

$$X_i \sim N(\mu, \sigma_i^2 + \tau^2)$$

- So can estimate  $\mu$  by  $\bar{X}$  and  $\tau^2$  by (say)  $\tilde{\tau}^2$ . Then,
- plug into formula for posterior mean for known  $\mu, \tau^2$  to get  $\tilde{\theta}$ .
- *The estimate used here is  $\tilde{\tau}^2 = \sum (X_i - \bar{X})^2 / (p - 3) - \sigma_i^2$ . This is reasonable, although it's not the MLE.*

## (Hierarchical) Bayes Analysis

- Prior (on hyperparameters)

$$d\mu d\tau^2$$

- This is not Jeffreys prior. This is not the Reference prior?

- But, this **is** nearly the prior on the edge of admissibility:

*If one instead takes  $\mu = 0$  rather than  $\mu$  uniform, then this corresponds to “harmonic” prior density of Stein (1973, 1981):*

$$\text{prior}(\boldsymbol{\theta}) \propto \|\boldsymbol{\theta}\|^{2-p}.$$

- The above prior gives an estimator and risk function similar to that of J-S estimator.

## Analysis of 2005 Data

I applied these ideas to all batters in 2005 having at least 11 at-bats in each half season.

The estimators were computed from their record for the first half season

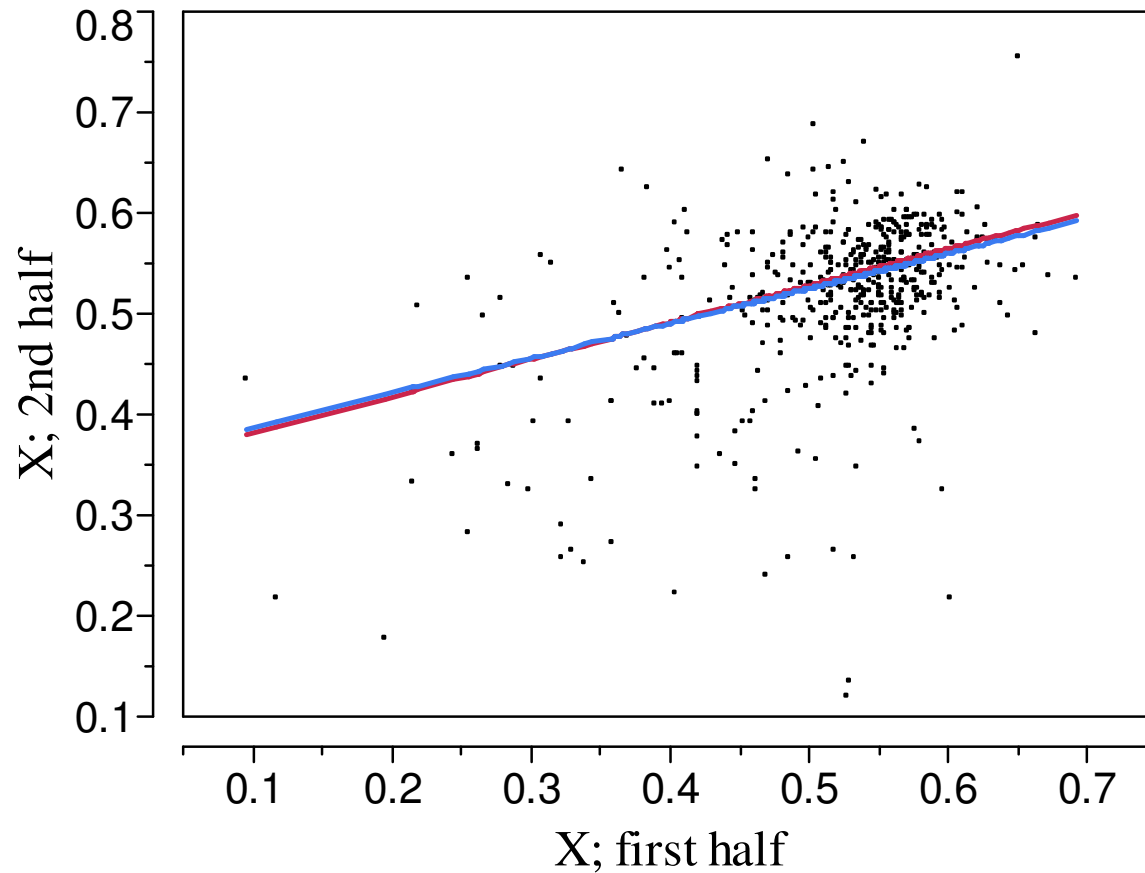
The estimators were then compared to their performance in the second half. The total squared prediction error was computed.

## Remarks About the Various Analyses

- Estimators doing a bad job:
  - The Bayes estimator
  - The E-B estimator when  $\mu, \tau^2$  are estimated by ML
- Estimators doing a pretty good job:
  - The E-B estimator when  $\mu, \tau^2$  are estimated by Method of Moments (but only if this is done in the “right” way)
  - The J-S estimator
- Estimator doing the best job:
  - A nonparametric E-B estimator (Brown, Greenshtein (2009))
- Next page illustrates that J-S Estimator and the (weighted) Linear Regression Estimator are clones of each other (Stigler (1990), Brown (2007)).

*This “explains” Clemente and Munson effects*

Scatterplot of Transformed  $BA^s$  for all batters (2005) by half season  
minimum 11 at bats each half of season



\_\_\_\_\_ = **Wt'd Linear Regression;**      \_\_\_\_\_ = **J-S estimator**

## Question for Jordan:

Jim Berger's informal description (yesterday) of Bayesian statistics was that it is a way of “representing uncertainty in the problem through [prior) distributions”

Rob Kass (yesterday) tried to describe the commonality of all statisticians through their desire to probabilistically represent the uncertainty involved in data applications.

There is uncertainty *implicit* in the problems elegantly analyzed in your paper. But it is not *explicit* in the formulation or in the criteria applied within this formulation.

I think it is fair to caricature your formulation as follows (but feel free to disagree):

The data is fixed. The goal is to find a rule,  $\gamma$ , (within some class of rules) that “best” describes this data in a computationally feasible way. Presumably the results obtained here will be applied to some future data.

What is the mechanism generating this future data? And how well will the good rules generated on the observed data work on future data?