BAYES, WHY BOTHER?

Thomas A. Louis, PhD

Department of Biostatistics
Johns Hopkins Bloomberg SPH

Research & Methodology
U. S. Census Bureau

Outline

- Quite basic examples of when to bother with Bayes
- Coda

Historical Controls

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	50	50	100

- Fisher's exact one-sided P = 0.121
- But, pathologists get excited:
  - "The 3 tumors are Biologically Significant"
- Statisticians protest:
  - "But, they aren't Statistically Significant"

Include Historical Data

- Possibly, the pathologist has historical information for the same species/strain, same Lab, recent time period with 0 tumors in 450 control rodents
- S/he has the following table in mind:

Pooled Analysis			
	C	E	Total
Tumor	**0**	3	3
No Tumor	500	47	547
	500	50	550

- Fisher's exact one-sided P $\doteq$ .0075
- Convergence between biological and statistical significance!
- The Bayesian formalism can be used to bring in the history, in general giving it partial credit

Bringing in history

- Structure the approach before seeing the data, by identifying relevant experiments
- Use the Bayesian formalism
  - Control rates are drawn from a Beta(μ, M)
  - Use all the data to estimate μ and M
    (or to produce the joint posterior distribution)
  - Give the historical data weight equivalent to a sample size of $\widehat{M}$ with rate $\hat{\mu}$
- Female, Fisher F344 Male Rats, 70 historical experiments (Tarone 1982)

Tumor	N	$\widehat{M}$	$\hat{\mu}$	$\frac{\widehat{M}}{N}$
Lung	1805	513	.022	28.4%
Stromal Polyp	1725	16	.147	0.9%

- Adaptive down-weighting of history

Reverend Thomas Bayes

To find a method for:

"… the probability that an event has to happen, in given circumstances…"

Bayes Rule:

$$\Pr(\theta|Y) \propto \Pr(Y|\theta)\Pr(\theta)$$

© http://www-history.mcs.st-andrews.ac.uk/PictDisplay/Bayes.html

Bayesian Analysis

1. Design a study & collect data
2. Specify a statistical model
   - The "data model" (ok, the model)
   - A prior distribution and possibly a hyper-prior
     Bayesians need to make these explicit
3. Use Bayes' theorem to produce the Posterior Distribution
4. Do something with it, possibly structured by a loss function
   - $\mathbf{(\ldots)^2}$**:** Posterior Mean
   - $|\ldots|$**:** Posterior median
   - $\mathbf{0/1 + c \times}$ **volume:** Tolerance Interval (CI)
   - **0/1:** Hypothesis Test/Model Choice

Bayesian Analysis

1. Design a study & collect data
2. Specify a statistical model
   - The "data model" (ok, the model)
   - A prior distribution and possibly a hyper-prior
     Bayesians need to make these explicit
3. Use Bayes' theorem to produce the Posterior Distribution
4. Do something with it, possibly structured by a loss function
   - $(\ldots)^2$: Posterior Mean
   - $|\ldots|$: Posterior median
   - **0/1 + c $\times$ volume:** Tolerance Interval (CI)
   - **0/1:** Hypothesis Test/Model Choice

- Steps 1 & 2 depend on scientific/policy knowledge and goals
- Steps 3 & 4 are governed by the rules of probability

Bayesian Analysis

1. Design a study & collect data
2. Specify a statistical model
   - The "data model" (ok, the model)
   - A prior distribution and possibly a hyper-prior
     Bayesians need to make these explicit
3. Use Bayes' theorem to produce the Posterior Distribution
4. Do something with it, possibly structured by a loss function
   - $(\ldots)^2$: Posterior Mean
   - $|\ldots|$: Posterior median
   - **0/1 + c $\times$ volume:** Tolerance Interval (CI)
   - **0/1:** Hypothesis Test/Model Choice

- Steps 1 & 2 depend on scientific/policy knowledge and goals
- Steps 3 & 4 are governed by the rules of probability
- Step 3 does not depend on what you are going to do in Step 4

Evidence, then decisions

Bother with Bayes when you want

- Excellent Bayesian performance
  - Phase I/II studies
- Excellent Frequentist performance
  - Use priors and loss functions as tuning parameters
- To strike an effective Variance/Bias trade-off
- Full uncertainty propagation
- To design, conduct and analyze complex studies
- **Sometimes it isn't worth the bother**
- **Sometimes you are (almost) forced into it**

Bother when you are (almost) forced into it, at least to generate a procedure

- (Adaptive) Design including monitoring
- Non-linear and complex models
- Diagnostic Tests
- Missing Data/Measurement error
- Small number of clusters
- Complex systems & Complex Goals
- Large "P" relative to "N"
- Smoothing & dimension reduction via penalties
- Spatial models, small area estimates
  - Data at different spatio-temporal scales
- Multiplicity
- .....

Procedure Generation: Binomial CIs

Intervals produced by the Bayesian formalism
can have excellent frequentist performance

The Beta-binomial Model

- Y is the number of events in n trials; θ the event probability

$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Conjugate Beta prior distribution:

$$g(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}, \; a, b > 0$$

- Mean: $\mu = \frac{a}{a+b}$
- Variance: $\tau^2 = \frac{\mu(1-\mu)}{M+1} = \frac{\mu(1-\mu)}{a+b+1}$
- M = a + b is the precision and is like a prior sample size

Posterior Distribution
Shrinkage and Variance Reduction

$$E(\theta \mid Y) = \mu_n = B_n \mu + (1 - B_n)\left(\frac{Y}{n}\right)$$

$$\begin{aligned} V(\theta \mid Y) &= \frac{\mu_n(1-\mu_n)}{M+n+1} \\ B_n &= \frac{M}{M+n} \end{aligned}$$

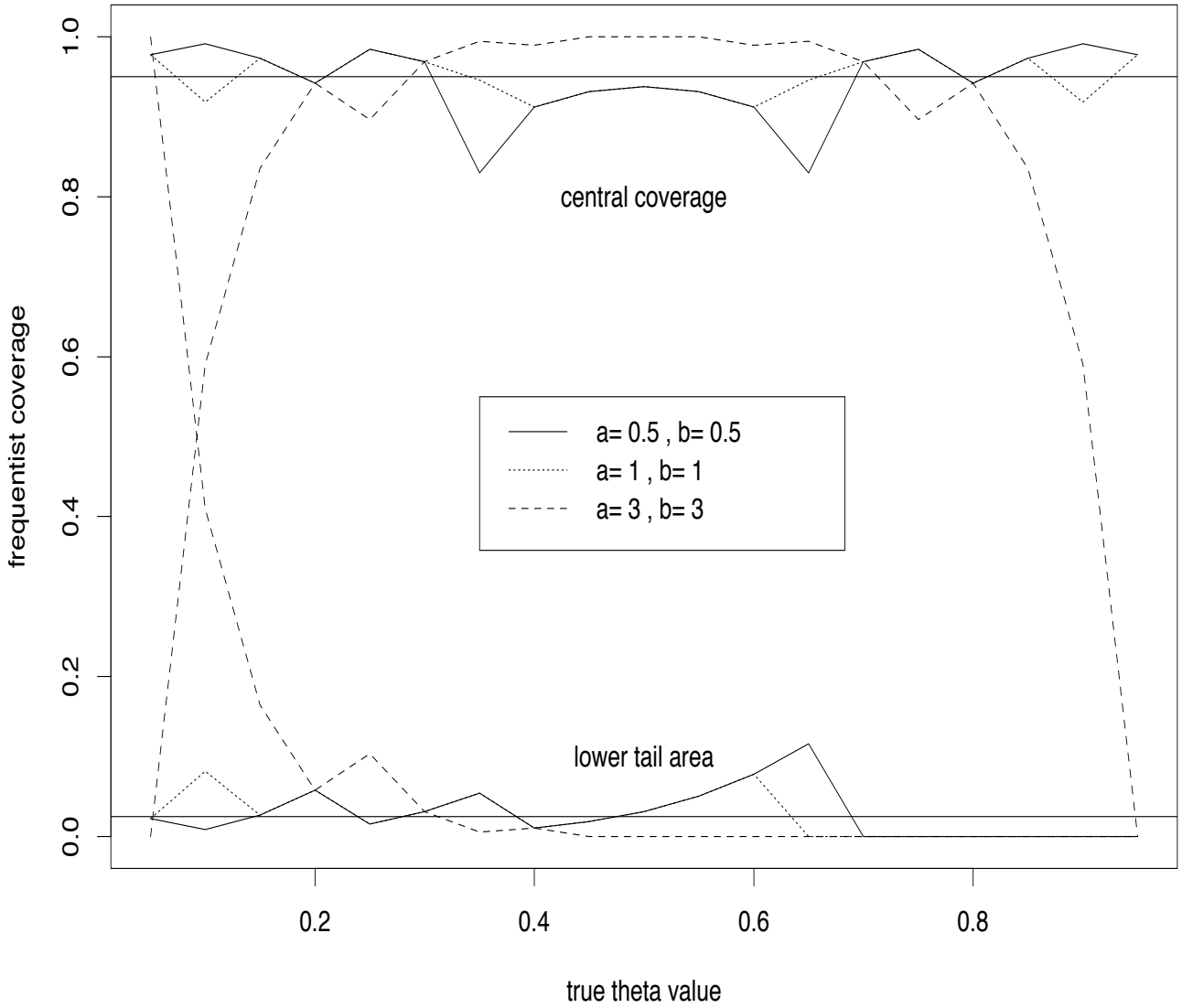
- As $n \to \infty, B_n \to 0$ (weight on the MLE $\to 1$)

Beta Priors for Binomial CIs

a	b	μ	M	B_5	B_{20}	comments
0.5	0.5	.50	1.0	17	5	Jeffreys (U-shaped)
1.0	1.0	.50	2.0	29	9	uniform
3.0	3.0	.50	6.0	55	23	symmetric, informative

- CI via the Highest Posterior Density (HPD) region (horizontal line drawing)
- **The computer doesn't know it's doing a Bayesian computation**

Binomial CI, frequentist coverage: $n = 5$



Design

- Everyone is a Bayesian in the design phase
- All evaluations are "preposterior," integrating over both the data (a frequentist act) and the parameters (a Bayesian act)
- A frequentist designs to control frequentist risk over a range of parameter values
- A Bayesian designs to control preposterior (Bayes) risk
- Bayesian design is effective for both Bayesian and frequentist goals and analyses

Bayesian Design to Control Frequentist CI Length

- Variance of a single observation: σ^2
- L is the desired maximal total length (distance from the low endpoint to the high endpoint) of the CI
- For two-sided coverage probability $(1-\alpha)$:

$$\mathbf{n(\sigma, L, \alpha) = 4Z^2_{1-\alpha/2}\left(\frac{\sigma}{L}\right)^2}$$

- If we don't know σ^2, then CI length is, itself, a random variable and uncertainty related to it must be accommodated

Bayesian Design to Control Frequentist CI Length

- Variance of a single observation: σ^2
- L is the desired maximal total length (distance from the low endpoint to the high endpoint) of the CI
- For two-sided coverage probability $(1-\alpha)$:

$$\mathbf{n(\sigma, L, \alpha) = 4Z^2_{1-\alpha/2}\left(\frac{\sigma}{L}\right)^2}$$

- If we don't know σ^2, then CI length is, itself, a random variable and uncertainty related to it must be accommodated
- To find a suitable sample size, we can,
  - do a series of "what ifs" or a "worst case"
  - put a distribution on σ^2 (ideally developed from other, similar studies) and use it to incorporate uncertainty in its value

Frequentist CI Length: The Bayesian approach

- Background data or prior elicitation provide a prior distribution (G) for σ^2
- Using G, select the sample size (n) to satisfy either,

$$E_G(\text{CI length}|n) \leq L$$

- Or, more relevant for a single study,

$$pr_G(\text{CI length} > L|n) \leq \gamma$$

Frequentist CI Length: The Bayesian approach

- Background data or prior elicitation provide a prior distribution (G) for σ^2
- Using G, select the sample size (n) to satisfy either,

$$E_G(\text{CI length}|n) \leq L$$

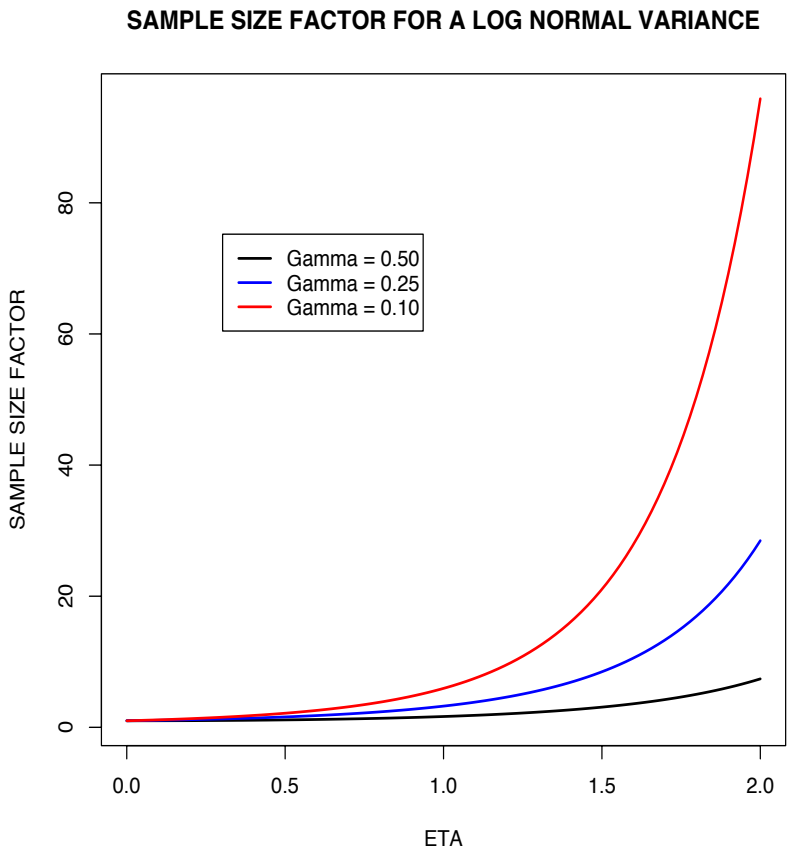
- Or, more relevant for a single study,

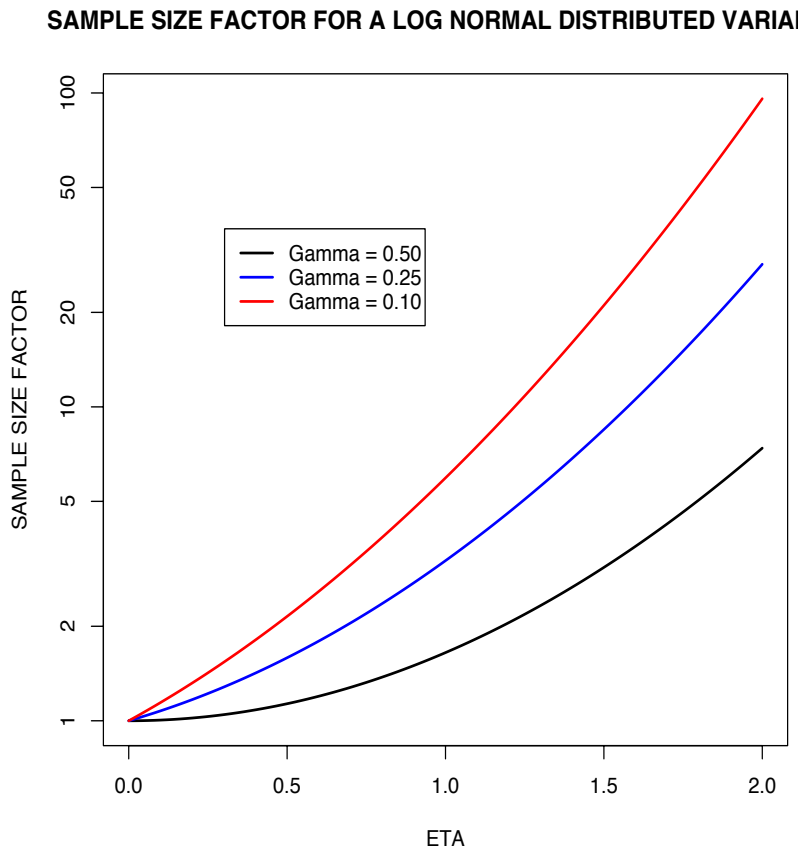
$$pr_G(\text{CI length} > L|n) \leq \gamma$$

- Similarly, for testing find n so that,

$$pr_G(\text{Power} < 0.80|n) \leq \gamma$$

CI Length: sample size factor for a prior coefficient of variation (η) relative to knowing σ^2 ($\eta = 0$)





Cluster Randomized Trials

- Develop an informative prior distribution for the between-cluster variance using studies thought to have a similar variance component, and us it

Design: to find the required number of clusters for a stand-alone analysis

Analysis: to conduct a Bayesian analysis for the between cluster variance for a study with a small number of clusters that can't/shouldn't stand alone

Adaptive Design & Allocation

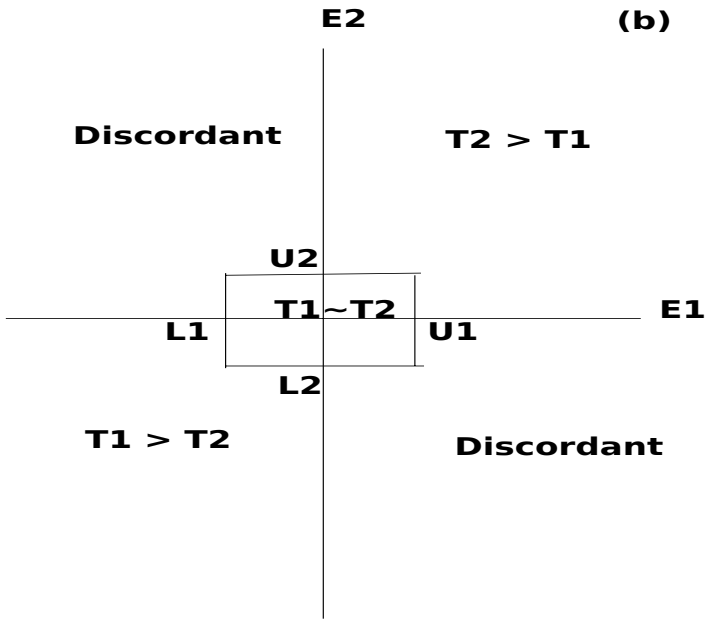
- Stopping rules
- Adaptive dosing
- Adaptive allocation
  - On baseline covariates, balancing
  - Allocation on treatment comparisons

Addressing non-standard and otherwise challenging goals

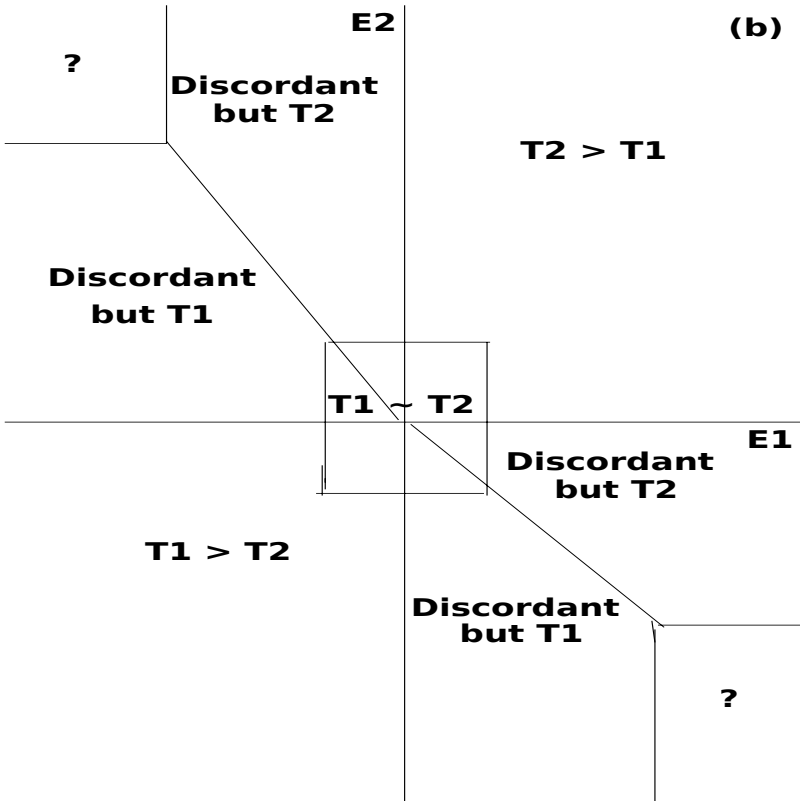
Bayesians have a corner on the market,
at least wrt to procedure-generation

- Regions for parameters
  - Bio-equivalence & non-Inferiority
  - Inherently bivariate treatment comparisons
- Ranks and Histograms
- Non-linear models
- Adaptive design
- Threshold utilities, for example in allocating federal funds

Combining endpoint-specific, univariate regions



Inherently bivariate regions



Compound sampling

- Multiple draws from the prior confers a degree of objectivity by supporting use of the data to estimate the prior
  - "Empirical Bayes" or "Bayes empirical Bayes"

$$
\begin{aligned}
\theta_1, \ldots, \theta_K \quad & iid \quad N(\mu, \tau^2) \\
[Y_k \mid \theta_k] \quad & ind \quad N(\theta_k, \sigma_k^2) \\
[\theta_k \mid Y_k] \quad & \sim \quad N\left(\mu + (1 - B_k)(Y_k - \mu), (1 - B_k)\sigma_k^2\right) \\
B_k \quad & = \quad \frac{\sigma_k^2}{\sigma_k^2 + \tau^2}
\end{aligned}
$$

When $\sigma_k^2 \equiv \sigma^2$

$$
\begin{aligned}
\hat{\mu} \quad & = \quad \bar{Y} \\
S^2 \quad & = \quad \frac{1}{K-1} \sum_k (Y_k - \bar{Y})^2 \\
\hat{\tau}^2 \quad & = \quad (S^2 - \hat{\sigma}^2)^+
\end{aligned}
$$

- Yes, it is a random effects ANOVA

Ranking Standardized Mortality Ratios, SMRs

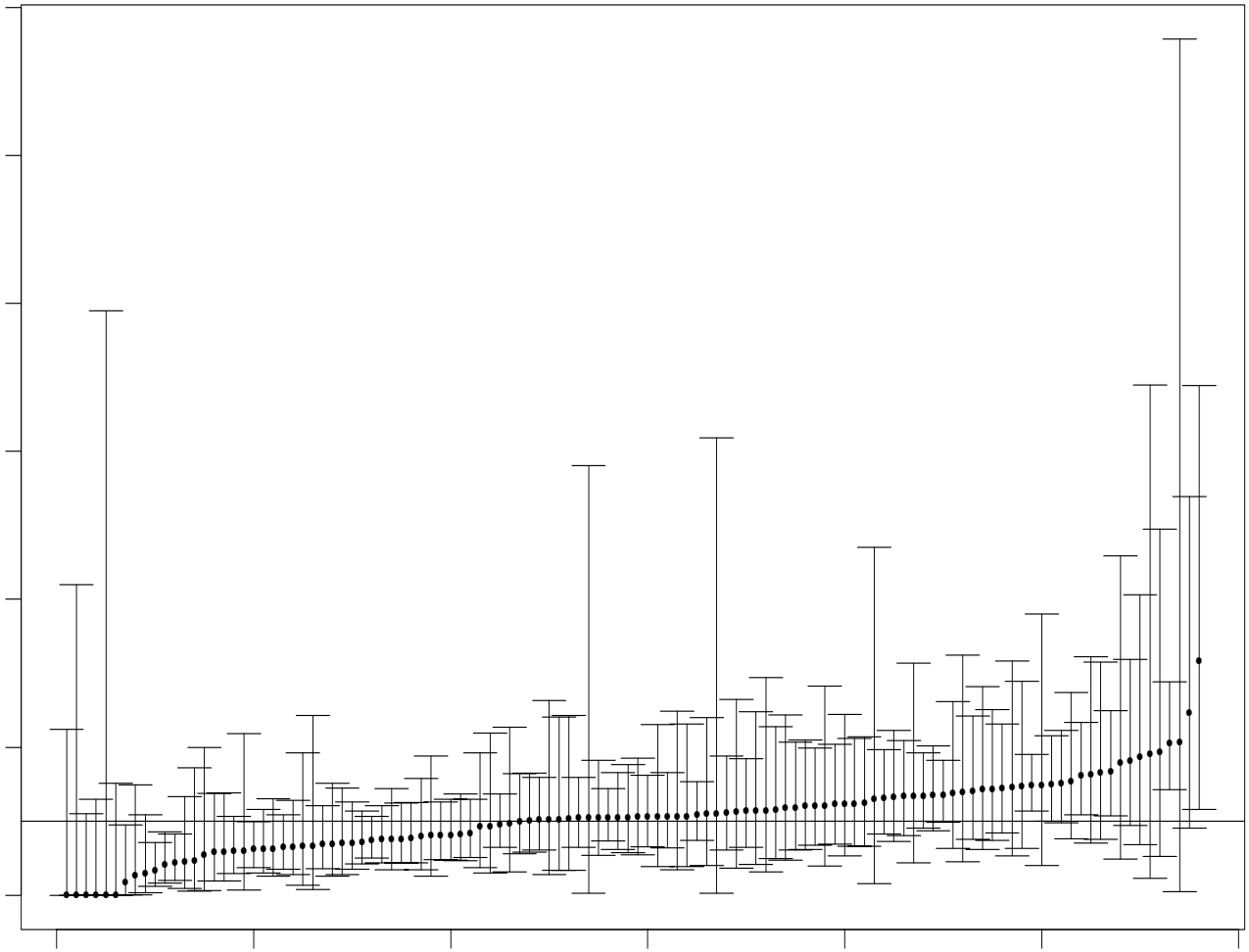
$$\text{SMR} = \frac{\text{observed deaths}}{\text{expected deaths}}$$

- Expecteds from a case mix adjustment model
- Rank 3459 dialysis providers using 1998 USRDS data
- Large and small providers, so standard errors of the estimated SMRs vary considerably
- Ranging from 1 patient per year to 355 patients per year

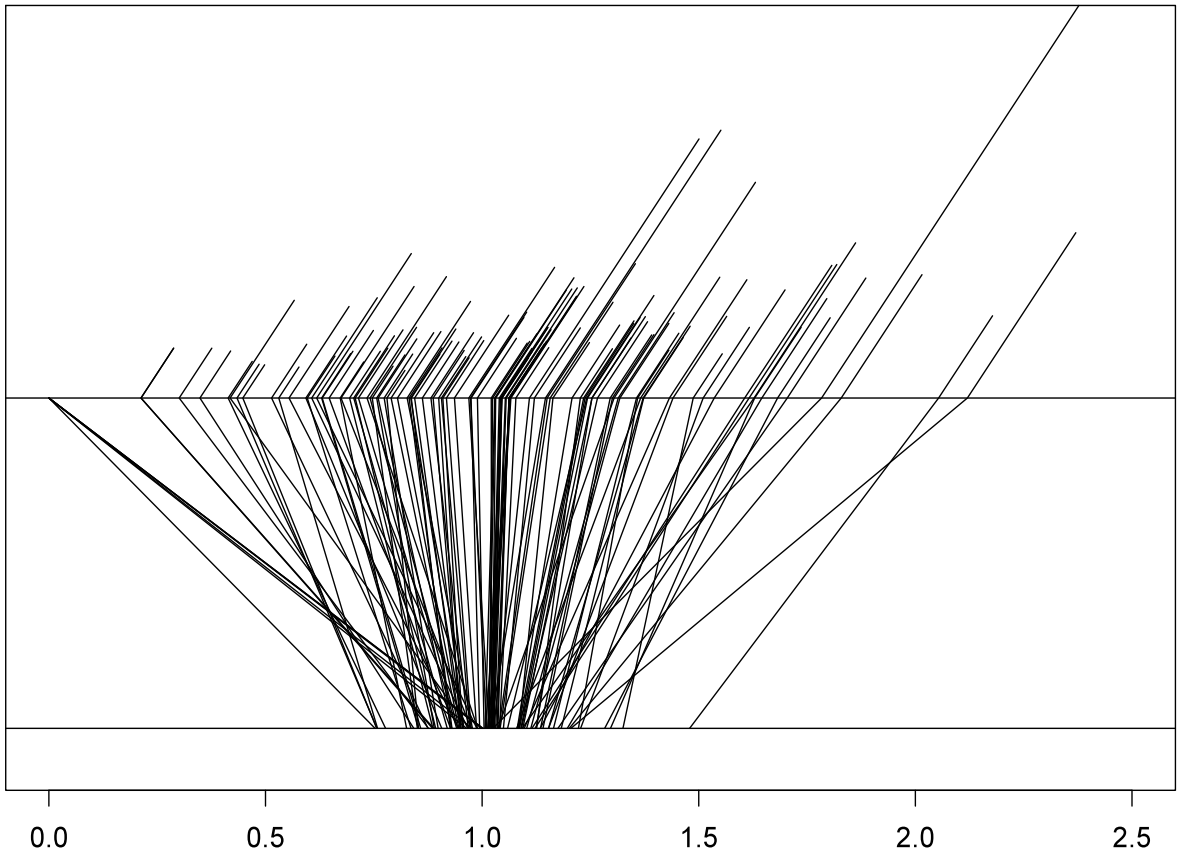
The Ranking Challenge

- Ranking estimated SMRs is inappropriate, if the SEs vary over providers
  - Unfairly penalizes or rewards providers with relatively high variance
- Hypothesis test based ranking: $H_0 : \text{SMR}_{unit} = 1$
  - Unfairly penalizes or rewards providers with relatively low variance
- Therefore, need to trade-off signal and noise
- However, even the optimal estimates can perform poorly

MLEs and exact CIs



Shrinkage Plot: MLEs, SEs and Posterior Means (PMs)



- Ranked MLEs are different from ranked PMs

Optimal Ranks/Percentiles

- The ranks are,

$$\begin{aligned} R_k(\boldsymbol{\theta}) &= \text{rank}(\theta_k) = \sum_{j=1}^{K} I_{\{\theta_k \geq \theta_j\}} \\ P &= R/(K+1) \end{aligned}$$

- The smallest θ has rank 1 and the largest has rank K
  The optimal SEL estimator is,

$$\bar{R}_k(\mathbf{Y}) = E_{\boldsymbol{\theta}|\mathbf{Y}}[R_k(\boldsymbol{\theta}) \mid \mathbf{Y}] = \sum_{j=1}^{K} \text{pr}(\theta_k \geq \theta_j \mid \mathbf{Y})$$

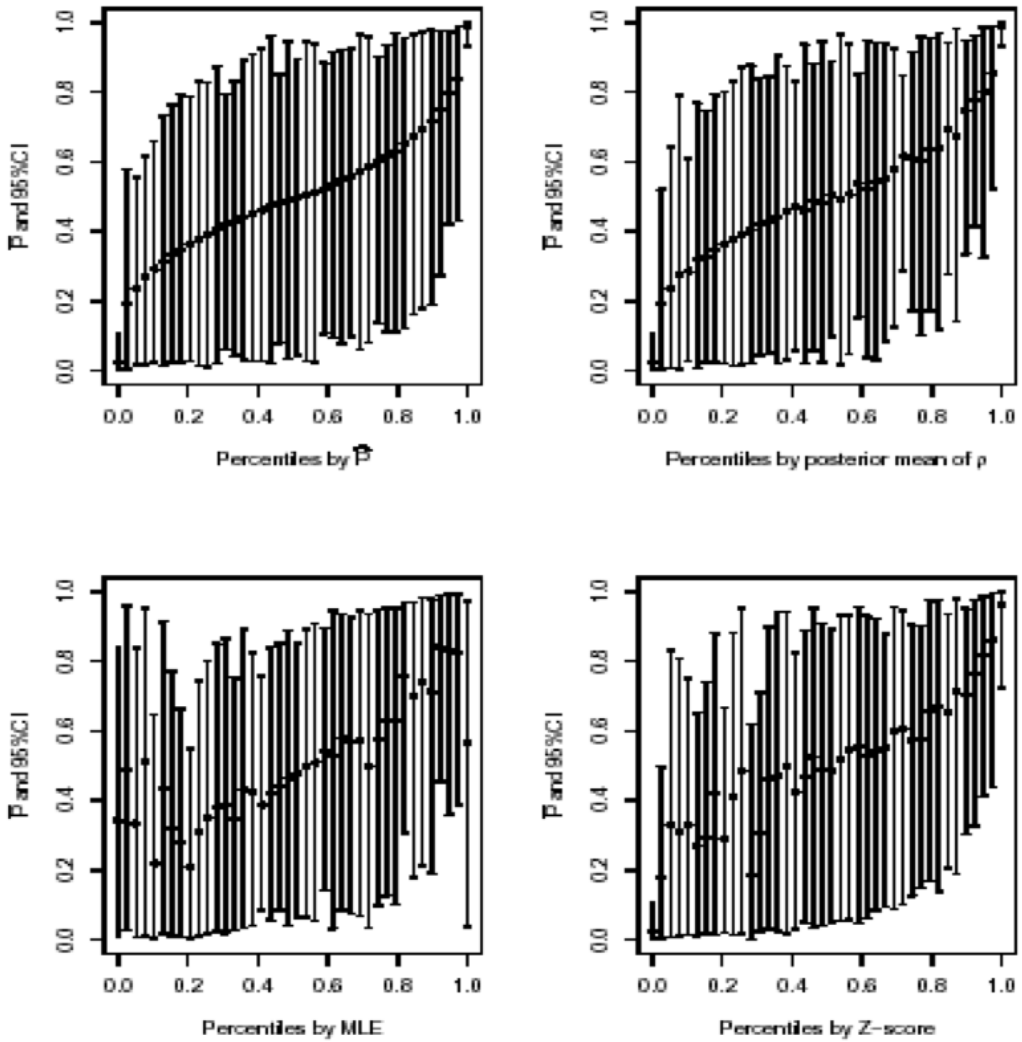
  Optimal integer ranks are, $\hat{R} = \text{rank}(\bar{R})$

$$\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y})); \ \hat{P}_k = \hat{R}_k/(K+1)$$

- Other loss functions, for example P (above γ)/(below γ) are more relevant in genomics and other applications wherein the goal is to identify the extremes

Relations among percentiling methods

1998 USRDS data



Histogram Estimates

- The setup,

$$\begin{aligned} \theta_1, \ldots, \theta_K & \quad iid \quad G \\ Y_k|\theta_k & \sim f_k(y|\theta_k) \\ \mathbf{G_K(t|\theta)} & = \frac{\mathbf{1}}{\mathbf{K}} \sum \mathbf{I}_{\{\theta_k \le t\}} \end{aligned}$$

- G_K is the "EDF" of the θ_k **operating in this dataset**
  - There is a connection with finite-population inference
- The optimal SEL estimate is:

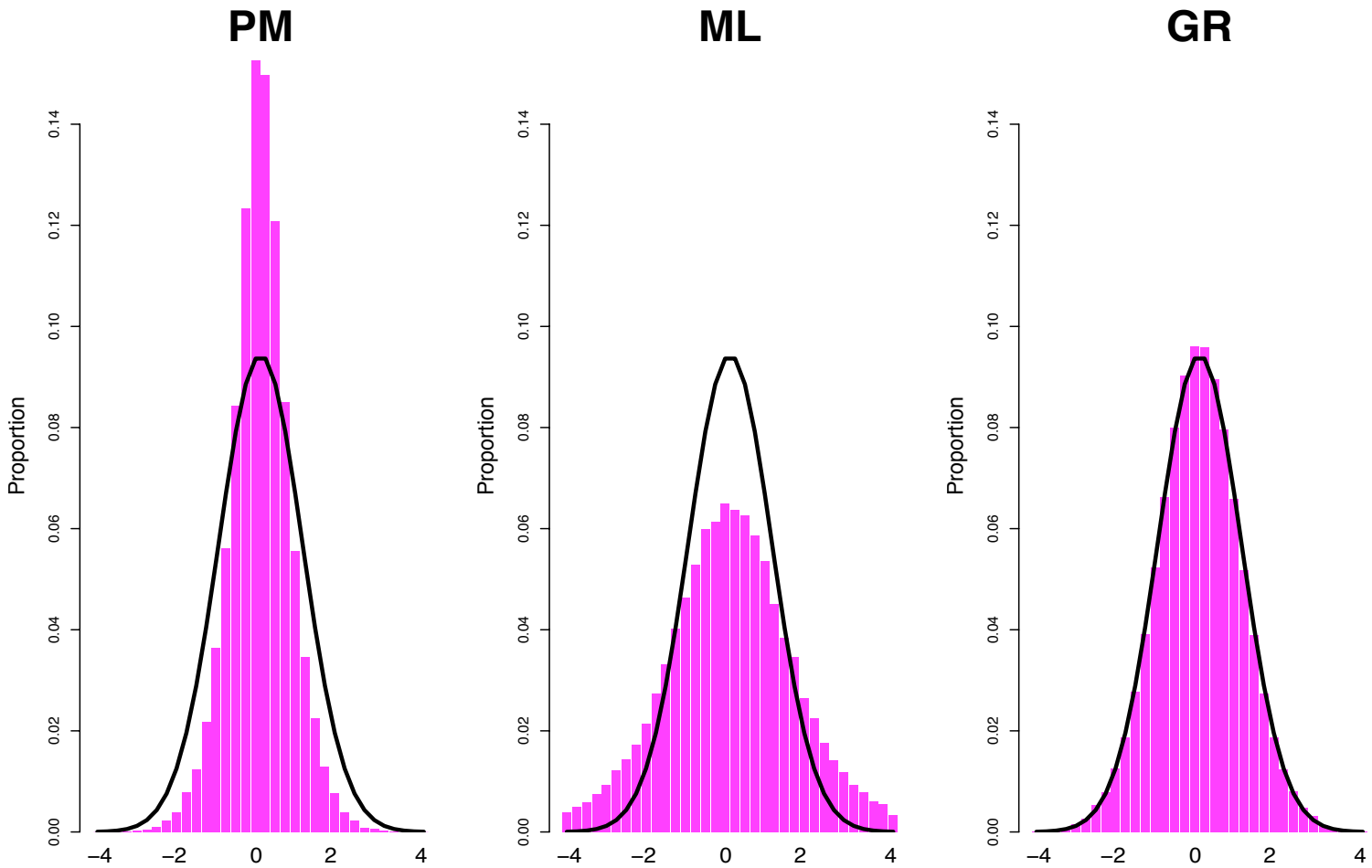
$$\mathbf{\bar{G}_K(t|Y)} = E[G_K(t;\theta)|\mathbf{Y}] = \frac{1}{K}\sum P(\theta_k \le t|\mathbf{Y})$$

The optimal discrete SEL estimate is:

$$\hat{G}_K(t \mid \mathbf{Y}) : \text{mass } 1/\text{K at } \hat{U}_j = \bar{G}_K^{-1}\left(\frac{2j-1}{2K} \mid \mathbf{Y}\right)$$

Gaussian Simulations: GR $= \hat{G}_K$

Need to get the spread right



Getting the spread right

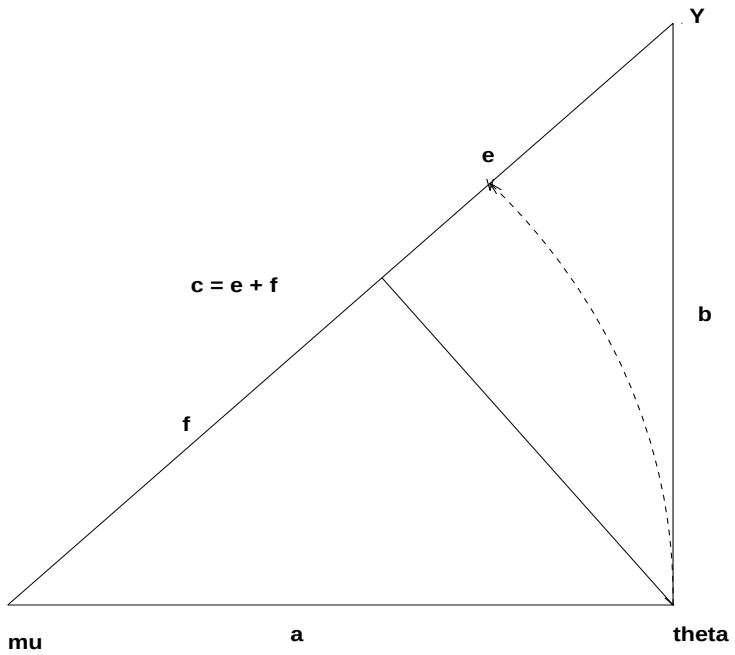
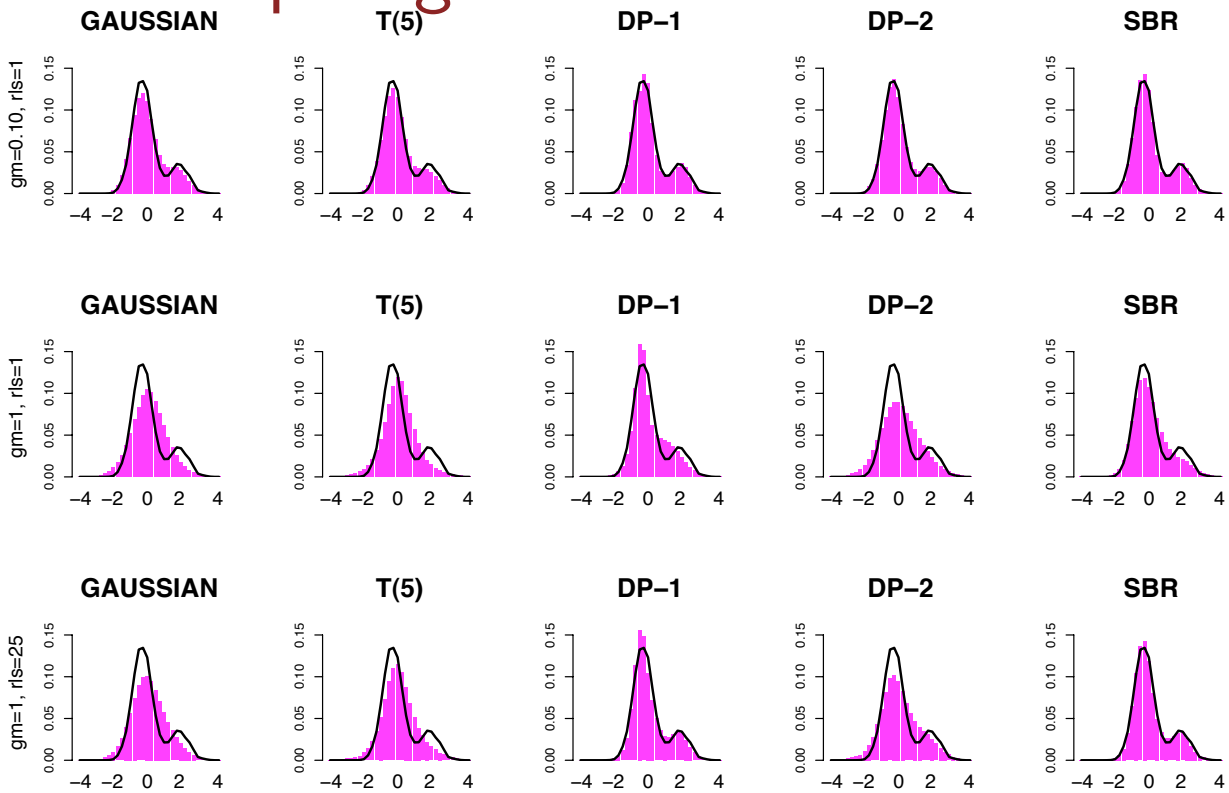


Figure 3: A triangle demonstration of the value of shrinkage

Gaussian mixtures, $\tau^2 = 1.25$:

Need to get the shape right



- Rows are σ_k^2 scenarios
  - constant, $\sigma_k^2 << 1$; constant, $\sigma^2 \equiv 1$; σ_k^2 variable, GM $= 1$
- Columns: Gaussian, T_5, DP-1, DP-2, SBR

Estimation relative to a threshold loss function

- Consider a mathematically tractable example in the spirit of Title I of the Elementary and Secondary Education Act
- Let θ be the true poverty rate for a single area, "A" be the amount allocated to the area, N the population size, and $\mathbf{Y}$ denote all data
- For a threshold $T \geq 0$, consider the societal loss function

Condition	Eligible for Concentration Funds?	Loss
$\theta \geq T$	yes	$N \times (\theta - A)^2$
$\theta < T$	no	$N \times A^2$

  - It would be more appropriate to replace squared-error by absolute error in the last column

- Using the Bayesian formalism, the optimal allocation value is,

$$A_T(\mathbf{Y}) = N \times \{E(\theta \mid \theta \geq T, \mathbf{Y}) \times \mathrm{pr}(\theta \geq T \mid \mathbf{Y})\}$$

Threshold loss, continued

- The allocation formula as a function of the true poverty rate (θ) has a threshold, but the optimal allocation value is a continuous function of the data
  - This may be difficult politically, but it is what it is!
  - Similarly, agencies (e.g., the CMS) can issue penalties or rewards as a continuous function of the posterior probability of exceeding a threshold
- For $T > 0$, $N^{-1}A_T(\mathbf{Y})$ is not the "center" of the the posterior distribution for θ, and

$$\frac{A_T(\mathbf{Y})}{N} \leq E(\theta \mid \mathbf{Y})$$

- The Bayesian formalism is almost essential in coming up with an effective allocation

Bayesians have had many successes, but there are challenges and a long way to go

- Continued development of semi- and non-parametric methods, especially in multivariate settings
- Evaluation of robustness and sensitivity
- Choice of hyper-prior and when to use empirical Bayes
- Computing innovations
- Reporting standards
- Model criticism, including evaluation of complex systems quantifying the contributions of the prior and the likelihood in producing the posterior
- Brad Efron has commented,

Bayesians have had many successes, but there are challenges and a long way to go

- Continued development of semi- and non-parametric methods, especially in multivariate settings
- Evaluation of robustness and sensitivity
- Choice of hyper-prior and when to use empirical Bayes
- Computing innovations
- Reporting standards
- Model criticism, including evaluation of complex systems quantifying the contributions of the prior and the likelihood in producing the posterior
- Brad Efron has commented,

"Bayesians get all the glory,
but frequentists do all the hard work"

The (Bayesian) Future is Bright

- The benefits of Bayesian structuring are substantial, but validity and effectiveness require expertise and care
  - The approach is by no means a panacea
- Computing has enabled accommodating complex data and implementing models
  - Enabling collaboration on challenging and important applications
- Success has and will depend on "anchored flexibility"
  - Eclecticism is (almost) always necessary, however it is essential to have a point of view, a framework, aids to navigation
- Keep in mind that traditional values still apply,

The (Bayesian) Future is Bright

- The benefits of Bayesian structuring are substantial, but validity and effectiveness require expertise and care
  - The approach is by no means a panacea
- Computing has enabled accommodating complex data and implementing models
  - Enabling collaboration on challenging and important applications
- Success has and will depend on "anchored flexibility"
  - Eclecticism is (almost) always necessary, however it is essential to have a point of view, a framework, aids to navigation
- Keep in mind that traditional values still apply,

  Space-age techniques will not rescue stone-age data!

#questions?

Literature

- Carlin BP, Louis TA (2009). Bayesian Methods for Data Analysis, 3^{rd} edition. Chapman & Hall/CRC, Boca Raton, FL.
- Carvalho B, Louis TA, Irizarry RA (2009). Quantifying Uncertainty in Genotype Calls. Bioinformatics, 26: 242-249.
- Li Q, Fallin D, Louis TA, Lasseter VK, McGrath JA, Avarmopoulos D, Wolyniec PS, Valle D, Liang K-Y, Pulver AE, Ruczinski I (2010). Detection of SNP-SNP Interactions in Trios of Parents with Schizophrenic Children. Genetic Epidemiology, 34: to appear.
- Lin R, Louis TA, Paddock S, Ridgeway G (2006). Loss Function Based Ranking in Two-Stage, Hierarchical Models. Bayesian Analysis, 1: 915-946.
- Lin R, Louis TA, Paddock S, Ridgeway G (2009). Ranking USRDS, provider-specific SMRs from 1998–2001. Health Services Outcomes and Research Methodology, 9: 22-38.
- Louis TA, Li Q, Carvalho, B, Fallin MD, Irizarry RA, Ruczikski I (2010). Association Tests that Accommodate Genotyping Errors. Bayesian Statistics, 9, Oxford University Press, to appear.
- Mwambi HG, Louis TA, Edmore R (2010). Joint Inference Regions for Longitudinally Observed Bivariate Clinical Measurements. In preparation.
- Paddock SM, Ridgeway G, Lin R, Louis TA (2006). Flexible Prior Distributions for Triple-Goal Estimates in Two-Stage Hierarchical Models. Computational Statistics and Data Analysis, 50: 3243-3262.
- Tarone RE (1982). The use of historical control information in testing for a trend in proportions. *Biometrics 38:* 215-220.

The general hierarchical model

$$
\begin{aligned}
[\boldsymbol{\theta} \mid \boldsymbol{\eta}] &\sim g(\cdot|\boldsymbol{\eta}) \quad \text{Prior} \\
[\mathbf{Y}|\boldsymbol{\theta}] &\sim f(\mathbf{y}|\boldsymbol{\theta}) \quad \text{Likelihood} \\
g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})}{f_G(\mathbf{y}|\boldsymbol{\eta})} \quad \text{Posterior} \\
f_G(\mathbf{y}|\boldsymbol{\eta}) &= \int f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta} \quad \text{Marginal}
\end{aligned}
$$

Or, Bayes empirical Bayes via a hyper-prior (H),

$$
g(\boldsymbol{\theta}|\mathbf{y}) = \int g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta})h(\boldsymbol{\eta}|\mathbf{y})d\boldsymbol{\eta}
$$