WHY BOTHER WITH BAYES?

Thomas A. Louis, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
tlouis@jhu.edu; www.biostat.jhsph.edu/~tlouis/

Expert Statistical Consultant
Center for Drug Evaluation & Research
U.S. Food & Drug Administration
Thomas.Louis@fda.hhs.gov

Brief Bio

- PhD Columbia, Mathematical Statistics, 1972
- Then, Postdoc@Imperial, BU Math Dept., Harvard Biostatistics
  U of Minn. Biostatistics (Head), RAND/DC, JHU, JHU/Census
  JHU, JHU/FDA (Emeritus, but not retired) Whew!
- IBS president, editor of *JASA/ACS* and *Biometrics*
- Considerable involvement in clinical trials
- Lots of Bayes involvement, both methods development and applications
  - Carlin & Louis (2009). *Bayesian Methods for Data Analysis, 3rd ed.* Chapman & Hall/CRC Press
- Visit my vita for full information

Outline of a, somewhat grand, tour

- Examples of when to bother with Bayes
- Methods
- Applications
  - Clinical
  - Epi
  - Policy
- Summary

I'll present a subset of the following slides

Preamble

- Check out these references[1,2,3]
- Efron (1986)[4] is a must-read. Quoting Efron, then Dennis Lindley;

  Efron: 'A prime requirement for any statistical theory intended for scientific use is that it reassures oneself *and others* that the data have been interpreted fairly."
  Lindley: "The objective element is the data: interpretation of the data is subjective, ..."
- Two more quotes,
  - Herman Rubin (1970): "A good Bayesian does better than a non-Bayesian, but a bad Bayesian gets clobbered."[5]
  - Tom Louis (2019): 'Pure' Bayes pairs nicely with Port, but when you leave port for the high seas of applications, some degree of impurity is usually necessary. Bayesians who engage in important studies use the paradigm as the aid to navigation, not as a straightjacket. The goal is to do a good job, and one can't be (too) doctrinaire.

---

[1] Carlin BP, Louis TA (2009). *Bayesian Methods for Data Analysis, 3rd ed.* Chapman & Hall/CRC.

[2] Gelman, et al. (2013). *Bayesian Data Analysis, 3rd ed.* Chapman & Hall/CRC.

[3] O'Hagan A (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, doi.org/10.1080/00031305.2018.1518265.

[4] Efron B (1986). Why isn't everyone a Bayesian? (with discussion) The American Statistician, 40: 1–11.

[5] Reported by IJ Good.

Confidence interval for a binomial probability

- Confidence intervals produced by the Bayesian formalism can have excellent frequentist performance, indeed as good or better than a 'frequentist' approach
- Here's the frequentist model,
  - Parameter: p = probability of an event
  - Data model: $Y \sim$ binomial(n, p)
  - Estimate: $\hat{p}^{freq} = \frac{Y}{n}$, (the MLE, the direct estimate)
- Add the Bayesian structure,
  - Prior: p is generated by a Beta(a, b) distribution with mean $\mu = a/(a+b)$, and 'effective sample size:' M = a + b
  - Posterior: Beta(a + Y, b + n - Y) with mean a weighted average of μ and $\hat{p}^{freq}$, thereby shrinking the latter towards the former

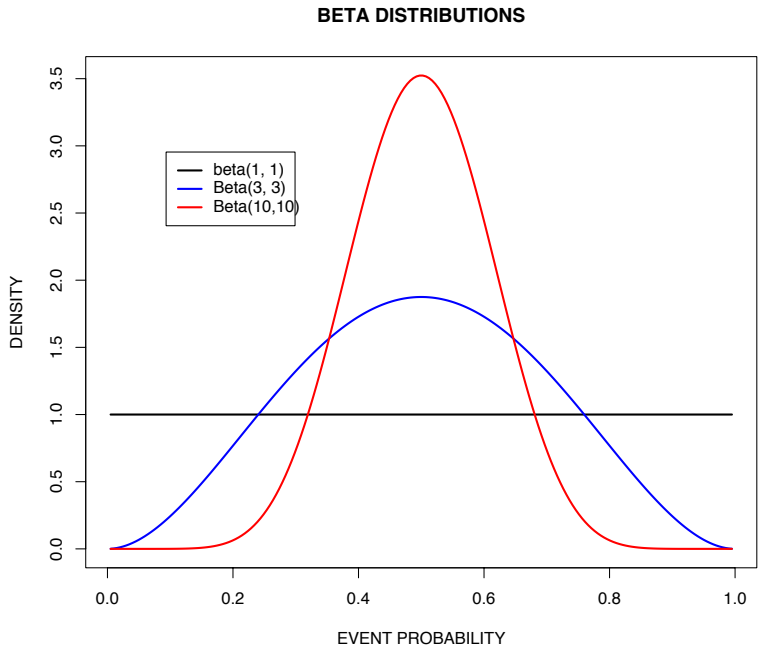
$$\hat{p}^{Bayes} \quad = \quad (1 - D_n)\mu + D_n \hat{p}^{freq}$$

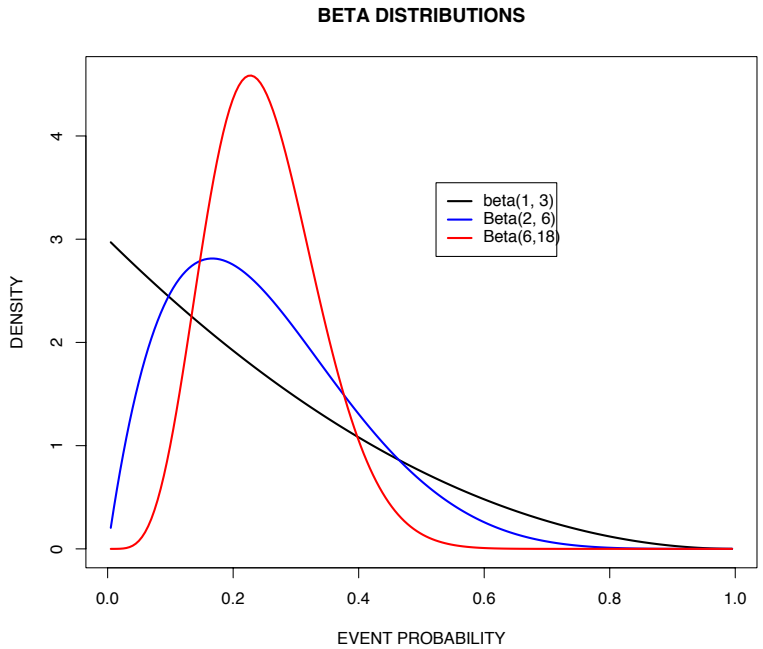
- With weight on $\hat{p}^{freq}$,

$$D_n \quad = \quad \frac{n}{M+n},$$

  which increases towards 1.0 as n increases
- Stabilize by shrinkage, but as n gets large, increase the weight on $\hat{p}^{freq}$

Beta Distributions and producing a CI



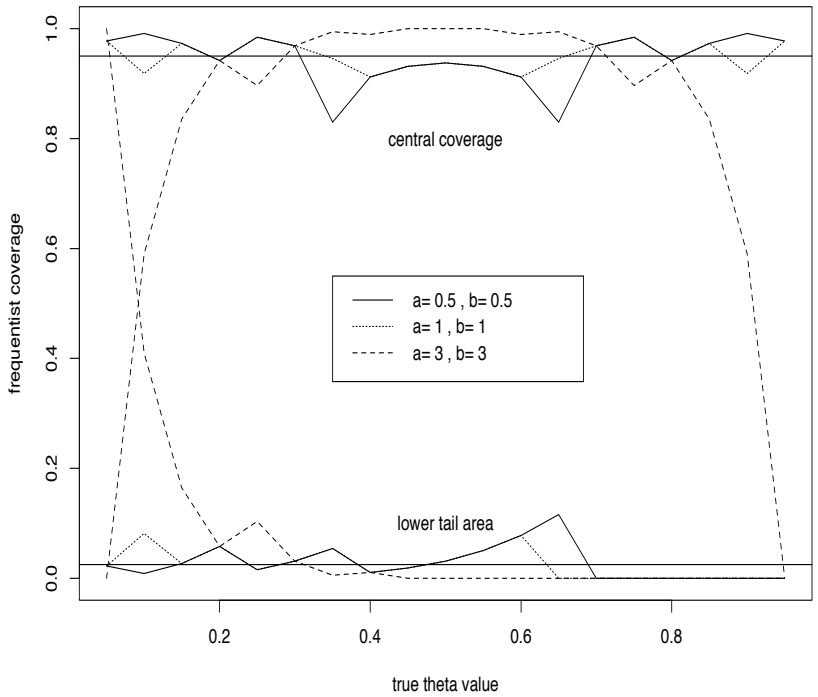


Producing a CI

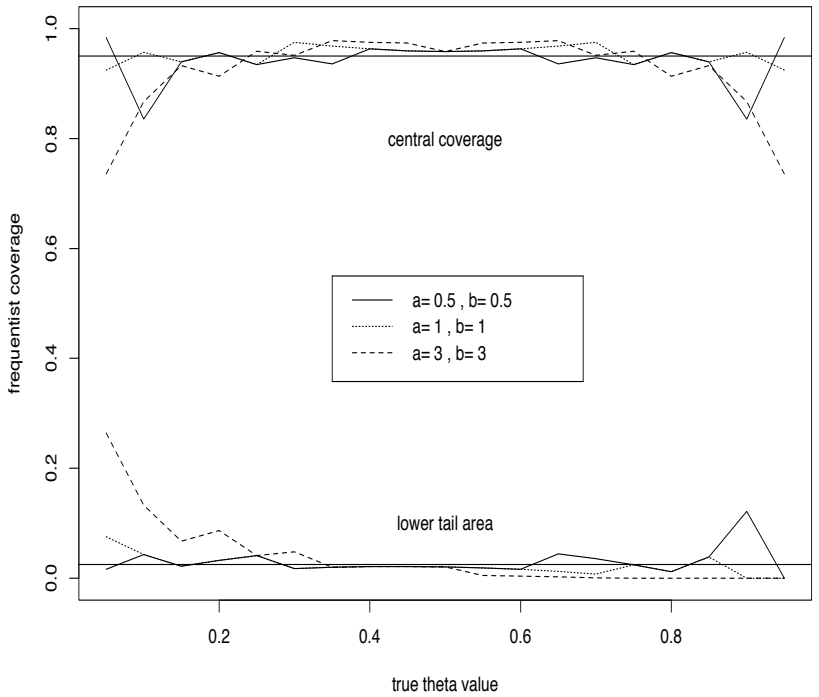
- Either cut off 2.5% tails on each side or find the highest posterior density (HPD) interval, produced by drawing a horizontal line on the posterior density so that the resultant interval has the desired probability (e.g., 0.95)
- HPD automatically deals with shape; the R function **binom.bayes** and others do the computations

Frequentist Coverage: Nominal level is 95%

$n = 5$

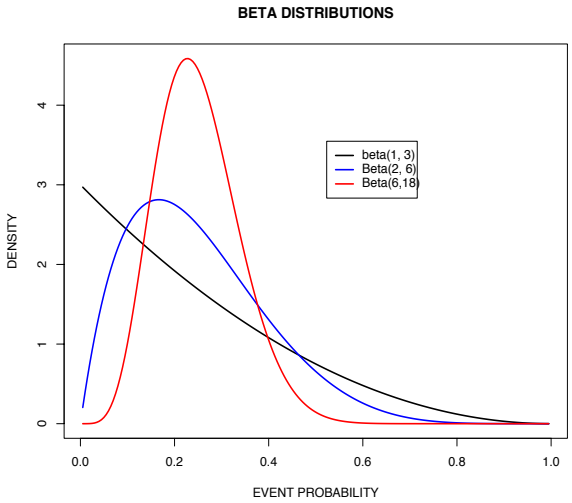


$n = 20$



- $a = b = 0.5$ and $a = b = 1.0$ work well;
  for $n = 5$, $a = b = 3$ performs poorly near 0 and 1
- It's worth discussing whether all CIs for the binomial distribution be Bayes-generated with either $a = b = 0.5$ or $a = b = 1.0$
  - Unless there is credible evidence for using an informative prior

CI after observing 0 events in n trials[6,7]



n	lower	upper
10	0	2.38/n
20	0	2.66/n
50	0	2.85/n
100	0	2.92/n
200	0	2.96/n
1000	0	2.99/n

- Using the black curve, a horizontal line produces 0 as the left-hand limit

  `binom.bayes(0,n, conf.level=.95, prior.shape1=1, prior.shape2=1)`
  - A frequentist CI for 0 events is similar, but 1 of n, etc. is challenging
- When $n \geq 100$ the 95% CI with $a = b = 1$ is very close to $[0, 3/n)$
- The '3' is a threshold for the number of successes in a study replication that would seem unusual after having observed 0 successes in the initial study
- The sample size to ensure 95% power at p^* to reject $H_0 : p = 0$ is $n = 3/p^*$
  - To detect $p^* = 10^{-5}$ (occupational risk) requires $n = 300{,}000$

[6] Louis TA (1981). Confidence intervals for a binomial parameter after observing no successes. *The American Statistician*, 35: 154.

[7] Manu P, Louis TA, Lane TJ, Gottlieb L, Engel P, Rippey RM (1988). Unfavuorable outcomes of drug therapy: Subjective probability versus confidence intervals. *J. Clin. Pharm. and Therapeutics*, 13: 213–217.

Historical Controls

- Data from the current experiment:

	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
- Important: Complete pooling gives too much credit to history, and the Bayesian formalism should be used to structure partial pooling

Bringing in history

- Before seeing the current data, identify relevant experiments
- Use the Bayesian formalism
  - Control rates (θ_k) are drawn from a Beta(μ, M)

$$\begin{aligned} E(\theta) &= \mu \\ V(\theta) &= \frac{\mu(1-\mu)}{M+1} \end{aligned}$$

  - Use all the data to estimate $(\mu, M) \rightarrow (\hat{\mu}, \hat{M})$ (or to produce their joint posterior distribution)
  - Use Beta($\hat{\mu}, \hat{M}$), better still mix over the full posterior
- Female, Fisher F344 Male Rats, 70 historical experiments[8]

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
- Judgment is required as to what historical data are sufficiently relevant

[8] Tarone RE (1982). The use of historical control information in testing for a trend in proportions. *Biometrics 38:* 215–220.

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

Bayes's grave

Non-conformist section of Bunhill Fields, London

Bayesian Analysis

1. Design a study & collect data
2. Specify a statistical model
   - The 'data model' (the likelihood)
   - 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 doesn't depend on what you are going to do in Step 4

Bayesian Analysis

1. Design a study & collect data
2. Specify a statistical model
   - The 'data model' (the likelihood)
   - 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 doesn't depend on what you are going to do in Step 4

Evidence, then decisions

Bayes Infographic (FYI)

('paras' are parameters)

- The go-forward model:

Prior Distribution	$\Longrightarrow$	paras	$\Longrightarrow$	Observed Data
G	$\Longrightarrow$	paras $\sim G$	$\Longrightarrow$	$Y \sim F(y \mid paras)$

- Fundamental Bayesian computation: Condition on the observed data and update the probability distribution for the paras
- The go-back model:

Observed Data	$\Longrightarrow$	Posterior: paras $\mid$ Data	$\Longrightarrow$	Inferences
Y	$\Longrightarrow$	paras $\sim G(\cdot \mid Y)$	$\Longrightarrow$	Inferences, e.g., E(paras $\mid Y$)

- With f and g density or mass functions, it is always the case that;

$$\text{(Posterior Odds)} = \text{(Prior Odds)} \times \text{(Likelihood Ratio)}$$

$$\frac{g(\text{paras} \mid Y)}{g(\text{paras}^* \mid Y)} = \frac{g(\text{paras})}{g(\text{paras}^*)} \times \frac{f(Y \mid \text{paras})}{f(Y \mid \text{paras}^*)}$$

Bother with Bayes when you want (FYI)

- Excellent Bayesian performance
- Excellent Frequentist performance
  - Use priors and loss functions as tuning parameters
- To strike an effective Variance/Bias trade-off
- To propagate full uncertainty
- To design, conduct and analyze complex studies
- To address non-standard goals such as ranking
- **Sometimes it isn't worth the bother**
- **Sometimes you are (almost) forced into it**
  - To incorporate prior information (duh)
  - To formally combine evidence
  - To analyzing complex systems & address complex goals
  - To develop spatial and network models
  - To deal with a small number of clusters
  - To accommodate complex measurement error
  - To handle complex . . . . . .
  - To avoid Rod Little's 'inferential schizophrenia' in design-based analyses[9]

[9] Little RJ (2012). Calibrated Bayes: an alternative inferential paradigm for official statistics (with discussion). *Journal of Official Statistics*, 28: 309-372 .

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 $(1-\alpha)$:

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

- 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,

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

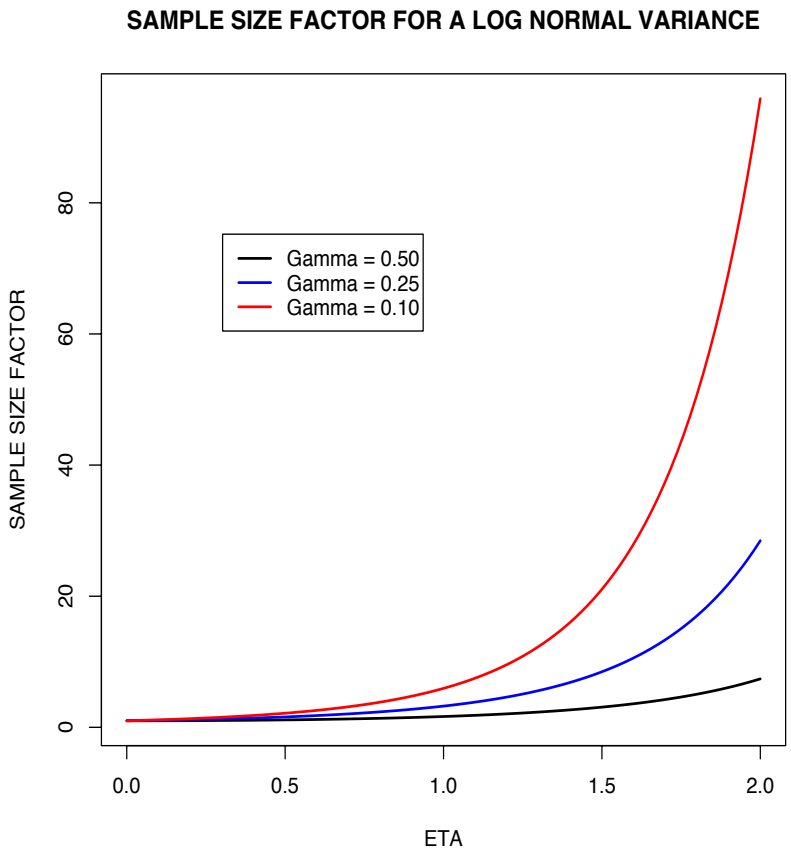
- Or, more relevant for a single study,

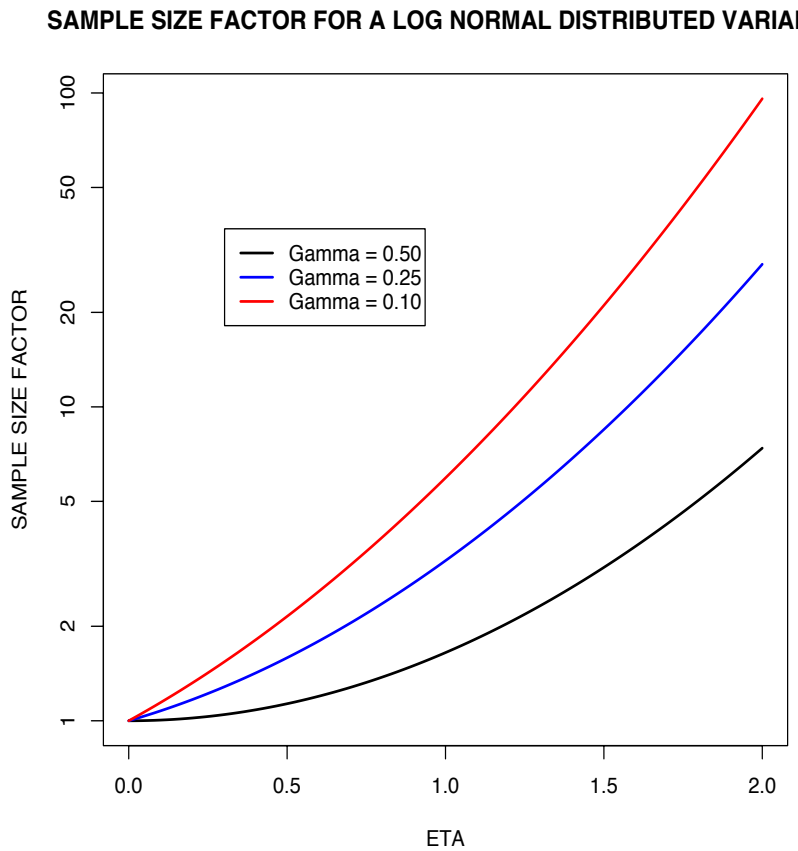
$$pr_G(\text{CI length} > L|n) \leq \gamma$$

- 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 ($\eta = 0$)





The basic Gaussian/Gaussian model

$$\begin{array}{rrcl} \text{Prior:} & \theta \sim G & = & N(\mu, \tau^2) \\ \text{Sampling distn.:} & [Y \mid \theta] \sim f & = & N(\theta, \sigma^2) \\ \text{Marginal distn.:} & f_G & = & N(\mu, \sigma^2 + \tau^2) \end{array}$$

Overdispersion

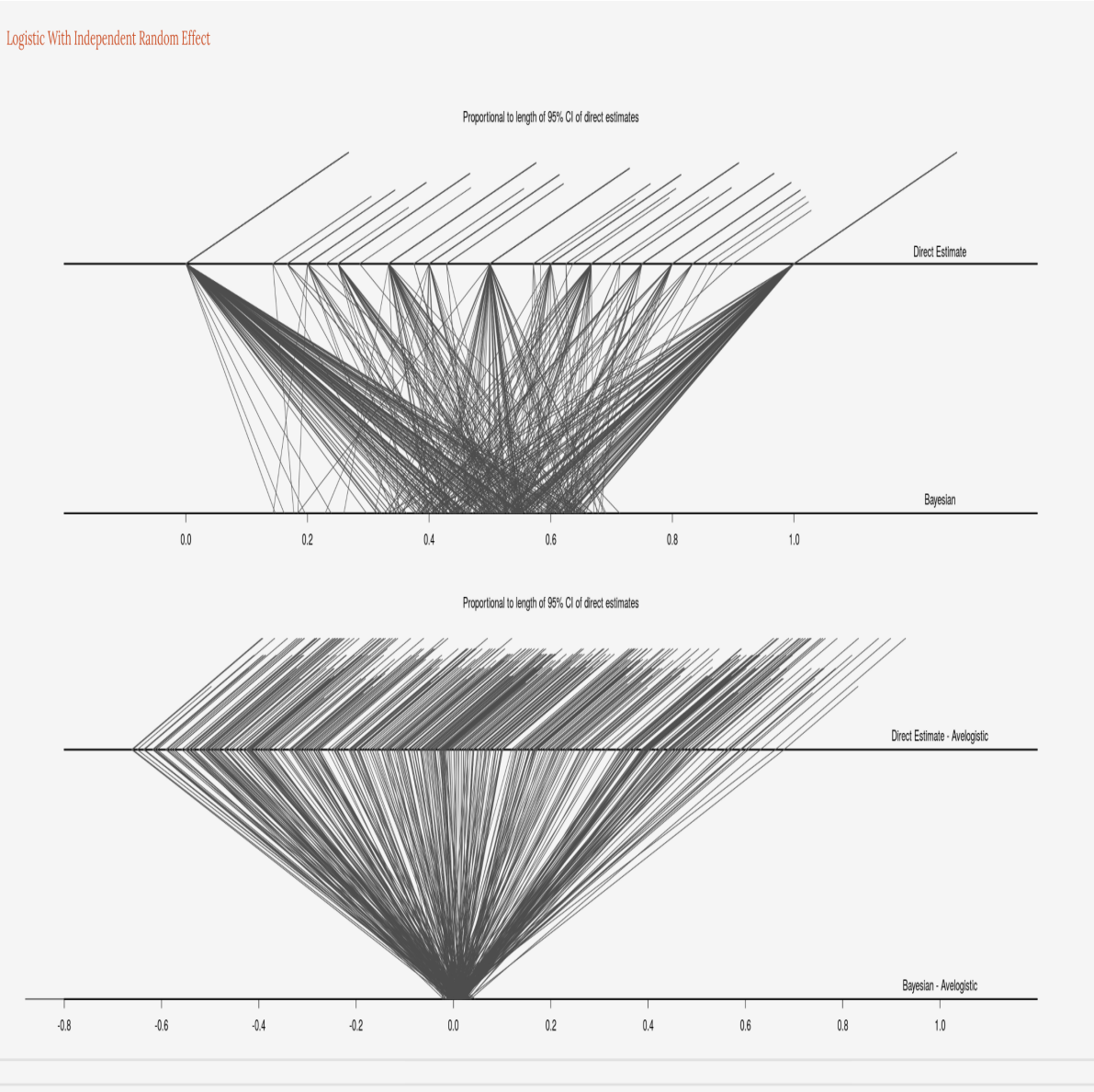
- For known (μ, τ^2, σ^2), the posterior is also Gaussian:

$$\begin{aligned} E(\theta|Y) &= B\mu + (1-B)Y = \mu + (1-B)(Y-\mu) \\ V(\theta|Y) &= (1-B)\sigma^2 = B\tau^2 \\ B &= \frac{\sigma^2}{\sigma^2 + \tau^2} \end{aligned}$$

- Shrinkage & variance reduction
  - Larger σ^2 produces greater shrinkage
  - Larger τ^2 produces less shrinkage

Nchelenge Zambia Malaria Prevalence: Independent RE model with covariates

Residuals shrink towards 0



Basic Estimates & Confidence intervals

You might not need to bother

- Estimate a population mean based on an iid sample

$$\begin{aligned} \hat{\mu} &= \bar{X}_n \\ \text{CI:} &\quad \bar{X}_n \pm Z\hat{\sigma}/\sqrt{n} \end{aligned}$$

- Yes, it's Bayes with a flat prior, but so what?
- A frequentist can use a BC_a CI to (almost) avoid parametric assumptions

So, why bother with Bayes?

Not so basic: what if we know $\mu \geq 0$

- The maximum likelihood estimate is $\hat{\mu}^{mle} = \max(\bar{X}_n, 0)$, but there are likelihood-based including Bayesian alternatives that can perform better
- So, consider a Bayesian CI with either a flat (likelihood) or an informative prior on $[0, \infty)$
- The posterior mean ($\hat{\mu}^{pm}$) is a worthy competitor to the MLE,

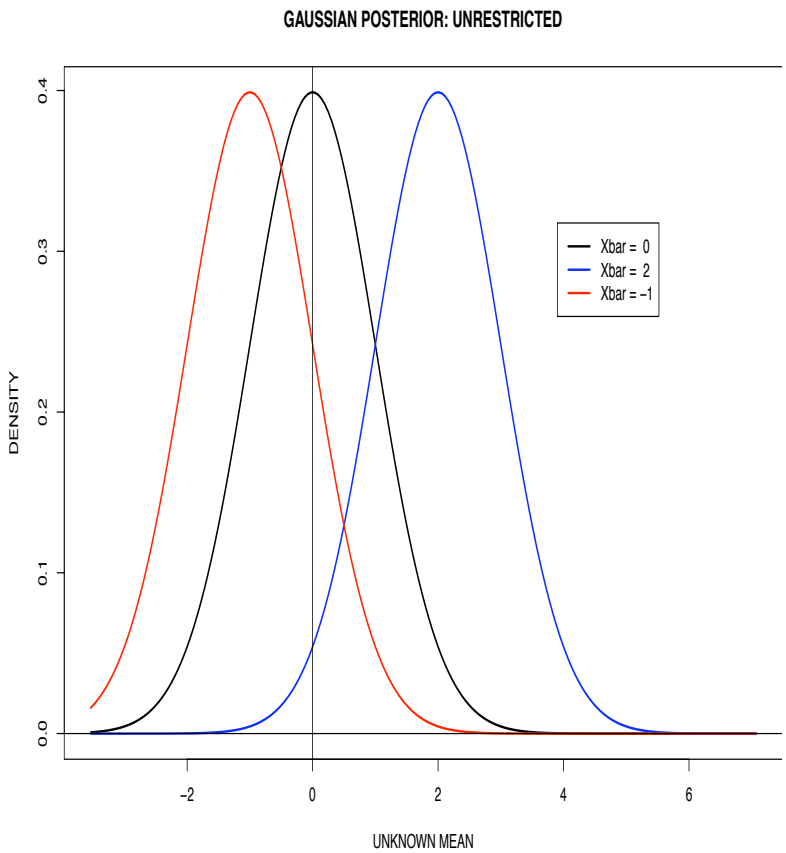
$$\mu^{pm} = E(\mu \mid \text{data}) = \int_0^\infty u \cdot g(u \mid \text{data}) du$$

- Good estimates strike an effective variance/bias trade-off and so have small mean squared error (MSE)

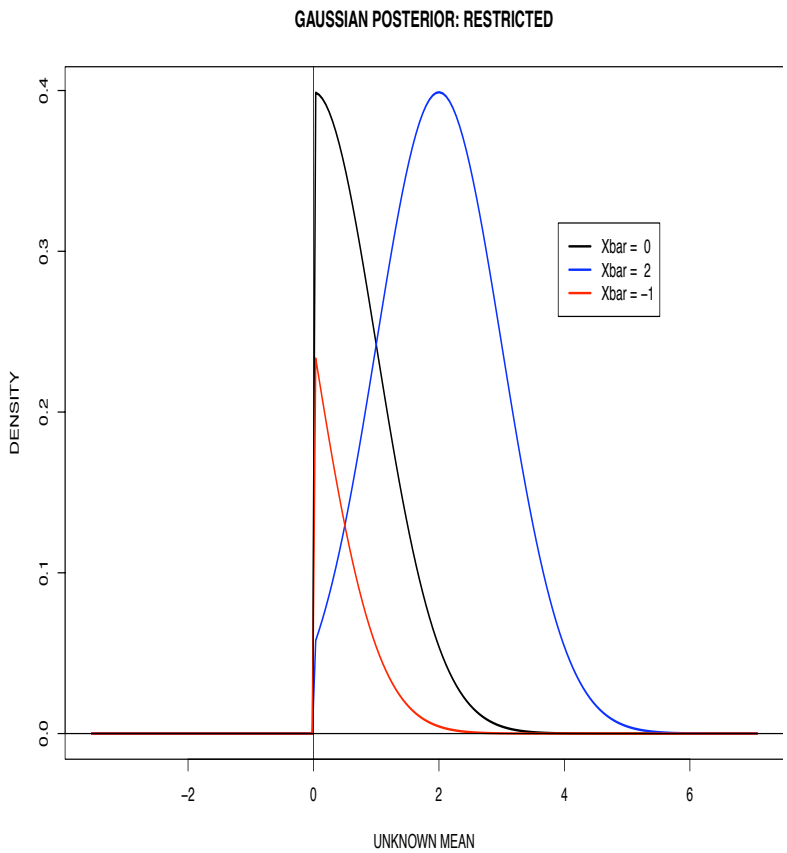
$$\text{MSE} = \mathbf{E}(\hat{\mu} - \mu)^2 = \mathbf{V}(\hat{\mu}) + (\text{BIAS})^2$$

Unrestricted and Restricted Posterior Distributions

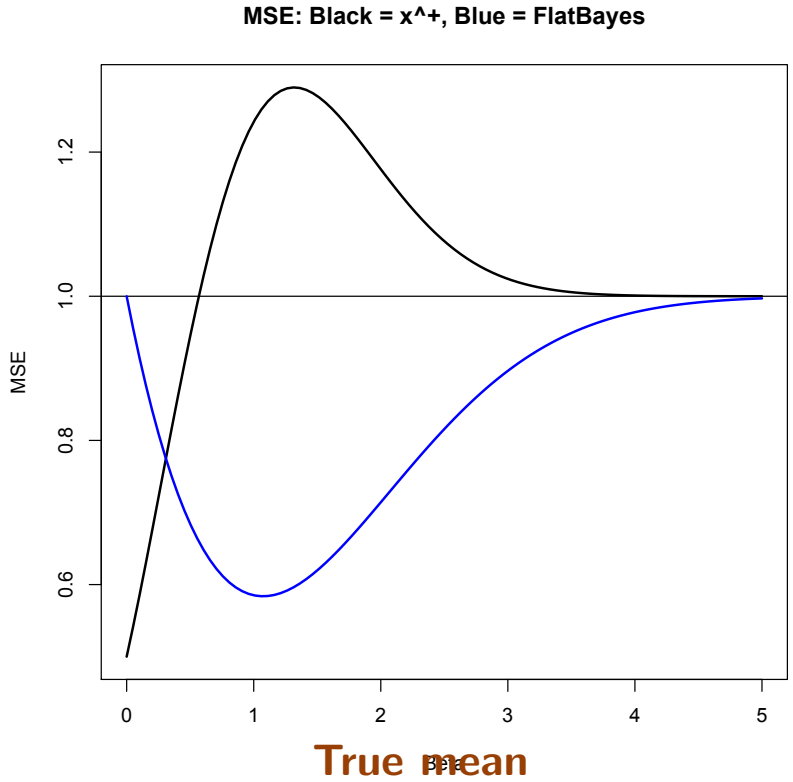
Unrestricted



Restricted



MSE for $\hat{\mu}^{mle}$ and $\hat{\mu}^{pm}$ when the parameter is ≥ 0



- If you want to improve MSE for the Bayes estimate near 0, use a prior that gives more weight near 0
- No Free Lunch: You pay for this with degraded performance for large μ
  - If the prior is $HN(0, \tau^2)$, MSE will increase quadratically
  - For a fix, use something like that in[10]

[10] Eberly L, Louis TA (2004). Bayes/frequentist compromise decision rules for Gaussian Sampling. *J. Statistical Planning & Inference*, 121: 191-207.

Multiple draws, compound sampling

Empirical Bayes (EB) and Bayes empirical Bayes (BEB)[11]

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

- 'Shrinkage' and **Variance Reduction**
  - For unequal σ_k^2, posterior variance flattening
- Generalizes to complicated models: regression structure in the prior, spatial or network models, non-conjugate priors
  - Need to use MCMC to do the computations
- Generalizes $H_0 : \theta_1 = \ldots = \theta_K$, by posing that the θ_k come from the same distribution rather than all being equal
  - $\tau^2 = 0$ produces the usual H_0

[11] Efron B (2019). Bayes, Oracle Bayes, and Empirical Bayes (with discussion). *Statistical Science*, 34: 177–235.

Objectivity conferred by compound sampling

- Multiple draws from the prior provide information on it
  $\Longrightarrow$ Empirical Bayes or Bayes empirical Bayes (BEB)
- When $\sigma_k^2 \equiv \sigma^2$

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

- $\hat{\tau}^2$ measures 'unexplained variation' not necessarily 'inexplicable variation'
  - A more general $\mathbf{X}_k\boldsymbol{\beta}$ mean model can reduce $\hat{\tau^2}$ and move direct estimates closer to the unit-specific prior mean
  - However, saturating the mean model gets back to the direct estimates and a 'sweet spot model' is most effective
- Unequal σ_k^2 requires recursion to produce the marginal MLE
- BEB (hyper-prior Bayes) brings in the uncertainty in the prior parameters by integrating over the posterior hyper-prior, generally, requiring MCMC or other computer-intensive approaches
- Subjectivity/Judgment required in choice of the data model, form of the prior, relevant data, ...

Age-specific rate of bone loss[12]

Woman/age–specific, slope estimates

- Positive values are 'loss' and a positive trend indicates a loss rate that increases with age
- Short follow-up, so estimated slope and residual variance are imprecise
- Empirical Bayes (EB) calms variation and improves woman-specific predictions

Woman/age-specific estimated slopes

Individual estimates

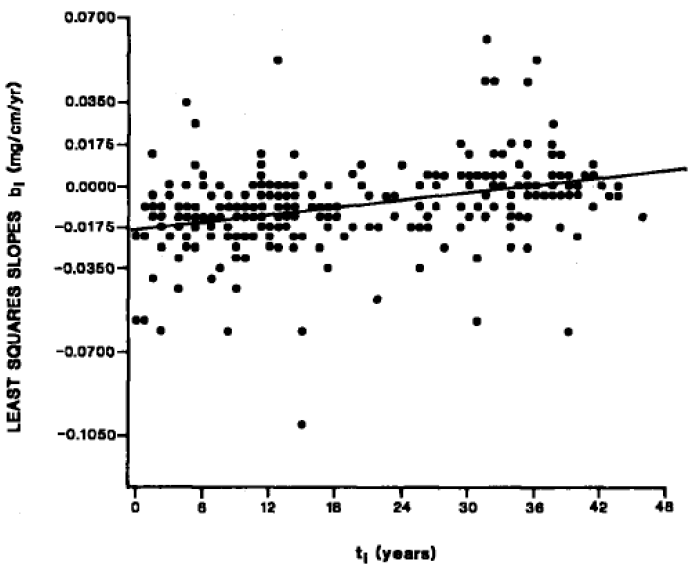


Figure 2. Individual least squares estimates of rate of bone loss b_i vs. t_i, where the t_i are suitably chosen points in the follow-up intervals.

EB estimates

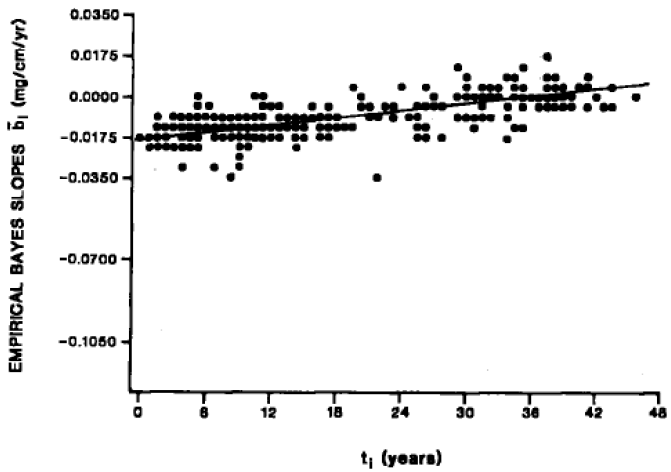


Figure 4. Individual empirical Bayes estimates of rate of bone loss $\bar{b}_i$ versus t_i.

[12] Hui, Berger (1983). Empirical Bayes estimation of rates in longitudinal studies. *JASA*, 78: 753–759.

Stabilizing woman-specific, estimated residual variance

- Directly estimated residual variance estimates are shrunken towards a common value; the degree of shrinkage depends on the precision of the direct estimate
  - Lower precision produces greater shrinkage
- The distribution of the shrunken estimates isn't chi-square, but a fully Bayesian analysis uses the joint posterior distribution of all parameters to produce woman-specific future predictions, prediction intervals, and other inferences that are more precise than using only each woman's data
- This full probability processing is an advantage of the Bayesian approach

General Point: In this and other contexts, there are three analytic strategies;

Lump: *They are all women*, so combine the data and use the population-level slopes and residual variances for each woman

Split: *Each woman is absolutely unique*, so infer/predict for each woman using only her data

Compromise: *Each woman is unique, but they are all women*, so compromise between 'Lump' and 'Split' with the amount of shrinkage towards population values depending on precision of the woman-specific estimates

The general hierarchical model

$$
\begin{aligned}
\boldsymbol{\theta} &= && \text{A vector of parameters} \\
[\boldsymbol{\theta} \mid \boldsymbol{\eta}] &\sim g(\boldsymbol{\theta}|\boldsymbol{\eta}) && [\text{Prior}|\boldsymbol{\eta}], \text{ e.g., } \textit{iid } \theta_k \\
[\mathbf{Y}|\boldsymbol{\theta}] &\sim f(\mathbf{y}|\boldsymbol{\theta}) && \text{Likelihood} \\
f_G(\mathbf{y}|\boldsymbol{\eta}) &= \int f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} && [\text{Marginal}|\boldsymbol{\eta}] \\
g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\boldsymbol{\eta})}{f_G(\mathbf{y}|\boldsymbol{\eta})} && [\text{Posterior}|\mathbf{y}, \boldsymbol{\eta}]
\end{aligned}
$$

The general hierarchical model

$$
\begin{aligned}
\boldsymbol{\theta} &= && \text{A vector of parameters} \\
[\boldsymbol{\theta} \mid \boldsymbol{\eta}] &\sim g(\boldsymbol{\theta}|\boldsymbol{\eta}) && [\text{Prior}|\boldsymbol{\eta}],\ \text{e.g., } \textit{iid}\ \theta_k \\
[\mathbf{Y}|\boldsymbol{\theta}] &\sim f(\mathbf{y}|\boldsymbol{\theta}) && \text{Likelihood} \\
f_G(\mathbf{y}|\boldsymbol{\eta}) &= \int f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} && [\text{Marginal}|\boldsymbol{\eta}] \\
g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\boldsymbol{\eta})}{f_G(\mathbf{y}|\boldsymbol{\eta})} && [\text{Posterior}|\mathbf{y}, \boldsymbol{\eta}] \\
[\boldsymbol{\eta} \mid h] &\sim h(\boldsymbol{\eta}) && \text{Hyper-prior} \\
g(\boldsymbol{\theta} \mid h) &= \int g(\boldsymbol{\theta}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} && [\text{Prior}|h],\ \text{e.g., exchangeable } \theta_k \\
g(\boldsymbol{\theta}|\mathbf{y}) &= \int g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) h(\boldsymbol{\eta}|\mathbf{y}) d\boldsymbol{\eta} && \text{Full Posterior}
\end{aligned}
$$

The general hierarchical model

$$
\begin{aligned}
\boldsymbol{\theta} &= && \text{A vector of parameters} \\
[\boldsymbol{\theta} \mid \boldsymbol{\eta}] &\sim g(\boldsymbol{\theta}|\boldsymbol{\eta}) && [\text{Prior}|\boldsymbol{\eta}], \text{ e.g., } \textit{iid}\ \theta_k \\
[\mathbf{Y}|\boldsymbol{\theta}] &\sim f(\mathbf{y}|\boldsymbol{\theta}) && \text{Likelihood} \\
f_G(\mathbf{y}|\boldsymbol{\eta}) &= \int f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} && [\text{Marginal}|\boldsymbol{\eta}] \\
g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}|\boldsymbol{\eta})}{f_G(\mathbf{y}|\boldsymbol{\eta})} && [\text{Posterior}|\mathbf{y}, \boldsymbol{\eta}] \\
[\boldsymbol{\eta} \mid h] &\sim h(\boldsymbol{\eta}) && \text{Hyper-prior} \\
g(\boldsymbol{\theta} \mid h) &= \int g(\boldsymbol{\theta}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} && [\text{Prior}|h], \text{ e.g., exchangeable } \theta_k \\
g(\boldsymbol{\theta}|\mathbf{y}) &= \int g(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) h(\boldsymbol{\eta}|\mathbf{y}) d\boldsymbol{\eta} && \text{Full Posterior}
\end{aligned}
$$

- Bayes empirical Bayes (BEB) combines evidence by integrating wrt $h(\boldsymbol{\eta}|\mathbf{y})$, importing uncertainty in $\boldsymbol{\eta}$
- Hyper-prior Bayes is just 'Bayes' with a different prior
- The model can be enhanced via covariates in the prior
- Can add a hyper-hyper-prior, . . . , but I leave that to epistemology

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
  - Alternative language ballots
- Ranks and Histograms
- Non-linear models
- Adaptive design
- Threshold utilities, for example in allocating federal funds

Being in a complex region ($\mathcal{R}$)

- Section 203 of the U. S. voting rights act mandates that a state or political subdivision must provide language assistance to voters,
  - if more than 5% of voting age citizens are members of a single language minority group
  - **and** do not 'speak or understand English adequately enough to participate in the electoral process"
  - **and** if the rate of those citizens who have not completed the fifth grade is higher than the national rate of voting age citizens who have not completed the fifth grade

  A political subdivision is **also covered**,
  - if more than 10,000 of the voting age citizens are members of a single language minority group, do not 'speak or understand English adequately enough to participate in the electoral process,"
  - **and** the rate of those citizens who have not completed the fifth grade is higher than the national rate of voting age citizens who have not completed the fifth grade.
- Every 5 years the Census Bureau must transmit determinations to the Department of Justice
- Bayesian structuring is essential for combining evidence, stabilizing estimates and computing summaries such as,

$$pr(\boldsymbol{\theta} \in \mathcal{R} \mid \text{data})$$

Comparing two treatments: New (N) vs Current Standard (S)

Frequentist

- Do a hypothesis test or CI and make a decision

Bayes

- Regions for the latent truth: N better than or equal to S; N worse than S
  - Can also include an indifference region
- Find the posterior probability of regions and make decisions
- Using 0 as the threshold, decide in favor of N, if $pr(N - S > 0 \mid data) > 0.98$
- Can find a sample size that controls the probability of mis-classification
- Can adjust the prior distribution to satisfy a frequentist criterion such as, for a specific $(N - S) \leq d < 0$,

$$\text{pr(decide in favor of } N \mid d) \leq 0.05$$

- However, if you trust the original prior, use it!
  - Compute frequentist properties, but don't rigidly adhere to them

Non-inferiority assessment

- (Inferior, nonInferior, Superior) regions for $(N - S)$ based on the true, underlying treatment relations
  - For a cure rate, 'Superior' will be positive values
  - For a death rate, 'Superior' will be negative values
- Use the posterior distribution to compute the probability of each region and use these to inform decisions
- Conduct sensitivity analyses by varying the,
  - nonInf threshold
  - prior distribution, e.g., (pessimistic, equipoise, optimistic)
  - data model

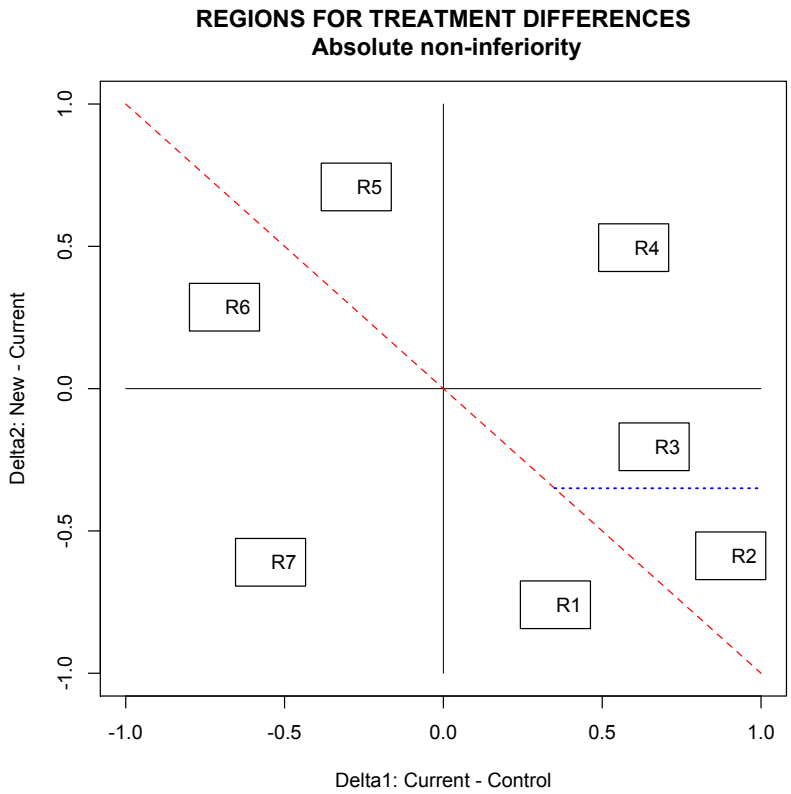
Framework for three treatment, non-inferiority assessment

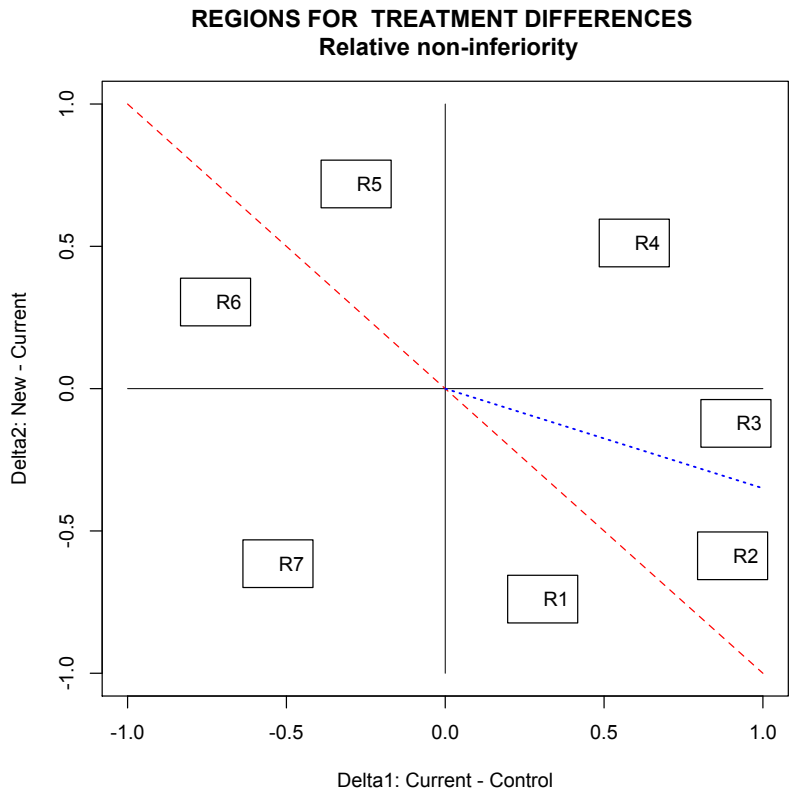
Decision Regions

R1: Current better than Control, New worse than Control
R2: Current better than Control, New better than Control, but 'inferior'
R3: Current better than Control, New better than Control and 'non-inferior'
R4: Current better than Control, New better than Current
R5: Current worse than Control, New better than Current and better than Control
R6: Current worse than Control, New better than Current but worst than Control
R7: All bets are off: Current worse than Control, New worse than Control

- Boundaries are defined by the true, underlying attributes (treatment effects, side effects) with no account for statistical uncertainty
- The Bayesian posterior distribution provides a window to this latent world
  - R2 accommodates differential side-effects,
  - If the new and current have similar side-effects, R2 can be empty
- Regions are best determined via a utility function

The seven regions (R1–R7)





- pr(Region | data) can be obtained no matter how complicated the region, either by computation or simulation {e.g., Markov Chain Monte Carlo (MCMC)}


I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT YEARS AGO.

From xkcd

Formal approach to region occupancy

- Inferring into which of several, possibly multivariate, regions true, underlying treatment effects fall is a fundamental goal
- Ingredients include formulating the decision problem, optimizing the decision, evaluating properties, and determining a sample size that achieves desired Bayesian (or frequentist) goals
- Here is a Bayesian approach, focusing on a scalar parameter (e.g, the difference in treatment effects) and two regions
  - A more comprehensive modeling entails a joint prior distribution for all parameters, producing a joint posterior distribution for them, then extracting the posterior for the treatment effect or effects
- We introduce notation for a three region categorization; with D a generic region,

$$(D_0, D_1, D_2) = (\textit{'Inferior', 'nonInferior', 'Superior'}).$$

- And, produce two regions by combining D_1 and D_2,

$$(D_0, D_{1:2}) = (\textit{'Inferior', 'Superior or nonInferior'})$$

$$D_{1:2} = D_1 \cup D_2.$$

Notation for the two region case

$\theta =$ Parameter of interest (e.g., treatment effect)

$G =$ Prior distribution for θ

$\mathbf{D} =$ Regions defined by values of θ

$\mathbf{X}_n =$ Data as a r.v., $\mathbf{x}_n$ observed value

$f(\mathbf{x}_n \mid \theta) =$ The data likelihood.

$\delta(\mathbf{x}_n) =$ 0 or 1 according as the decision is that $\theta \in D_0$, or $\theta \in D_{1:2}$

$L(\delta, \theta, c) =$ Loss function, $c \geq 0$

$\pi(\mathbf{x}_n) =$ the posterior distribution of region membership,

$$\begin{aligned} \pi(\mathbf{x}_n) &= \text{pr}(\theta \in D_{1:2} \mid \mathbf{x}_n) = \int_{D_{1:2}} g(\theta \mid \mathbf{x}_n) d\theta = \frac{\int_{D_{1:2}} f(\mathbf{x}_n \mid \theta) g(\theta) d\theta}{\int_{\Theta} f(\mathbf{x}_n \mid u) g(u) du} \\ &= \frac{\int_d^{\infty} f(\mathbf{x}_n \mid \theta) g(\theta) d\theta}{\int_{-\infty}^{\infty} f(\mathbf{x}_n \mid u) g(u) du}, \quad \text{when } D_{1:2} = [d, \infty). \end{aligned}$$

Decision Structure

- Decisions (inferences) are structured by a loss and resulting risk function
- If the goal is to minimize the probability of incorrect decision, then the obvious (and correct) decision rule is to decide $\theta \in D_{1:2}$, if $\pi(\mathbf{x}_n) \geq 0.5$
- However, the consequences of declaring that $\theta \in D_{1:2}$ when in fact $\theta \in D_0$ can be different from those for declaring that $\theta \in D_0$ when in fact $\theta \in D_{1:2}$, and the following loss function addresses this more general case
- With $\delta = 0$ or 1 according as the decision is $\theta \in D_0$ or $\theta \in D_{1:2}$,
  - Loss $= 0$, when θ is correctly classified
  - Loss $= 1$, when $\delta = 0$ is an incorrect classification
  - Loss $= c \geq 0$, when $\delta = 1$ is an incorrect classification.
- Specifically,

$$L(\delta, \theta, c) \quad = \quad c \cdot \delta I_{\{\theta \in D_0\}} + \{1 - \delta\} I_{\{\theta \in D_{1:2}\}}$$

- The decision depends on data, producing,

$$L(\delta(\mathbf{x}_n), \theta, c) \quad = \quad \mathbf{c} \cdot \delta(\mathbf{x}_n) I_{\{\theta \in D_0\}} + \{1 - \delta(\mathbf{x}_n)\} I_{\{\theta \in D_{1:2}\}}$$

Posterior Bayes Risk (conditional expected loss)

$$\begin{aligned} R_G(\mathbf{x}_n, \delta(\mathbf{x}_n), c) &= E_G\{L(\delta(\mathbf{x}_n), \theta, c) \mid \mathbf{x}_n\} \\ &= c \cdot \delta(\mathbf{x}_n)\{1 - \pi(\mathbf{x}_n)\} + \{1 - \delta(\mathbf{x}_n)\}\pi(\mathbf{x}_n). \end{aligned}$$

- The minimizing $\delta(\mathbf{x}_n)$ is,

$$\ddot{\delta}(\mathbf{x}_n) = 1, \iff \pi(\mathbf{x}_n) \geq \frac{c}{1+c} = \pi^*(c) \quad \left(c = \frac{\pi^*}{1-\pi^*}\right).$$

- So, decide,
  - $\theta \in D_{1:2}$, if $\pi(\mathbf{x}_n) \geq \pi^* = c/(1+c)$
  - $\theta \in D_0$ otherwise
- This formulation can justify a π^* value, where c is the relative cost of mistakenly declaring $\theta \in D_{1:2}$ versus mistakenly declaring $\theta \in D_0$
- **For example, justifying $\pi^* = 0.98$ requires $c = 49$**, a very (very) large relative cost of a Type I versus Type II error
  - $c = 49$ would be an extreme relative cost in many contexts

Optimal posterior and pre-posterior risk (FYI)

- Using $\ddot{\delta}(\mathbf{x}_n)$, the optimal posterior risk is,

$$\begin{aligned} R_G(\mathbf{x}_n, \ddot{\delta}(\mathbf{x}_n), c) &= \min[\pi(\mathbf{x}_n), c \cdot \{1 - \pi(\mathbf{x}_n)\}] \\ &= \begin{cases} \pi(\mathbf{x}), & \pi(\mathbf{x}_n) < \pi^* = \frac{c}{1+c} \\ c \cdot \{1 - \pi(\mathbf{x}_n)\}, & \pi(\mathbf{x}_n) \geq \pi^* = \frac{c}{1+c} \end{cases} \end{aligned}$$

- The risk with no information (no data, only the prior) is,

$$R_G^{(0)}(c) = \min(\pi, , c \cdot \{1 - \pi\})$$

- The *pre-posterior, optimal risk (Bayes Risk)* is the expectation using the marginal distribution of $\mathbf{X}_n$.

$$R_G(c) = E_G\left\{\min[\pi(\mathbf{X}_n), c \cdot \{1 - \pi(\mathbf{X}_n)\}]\right\} \quad (1)$$

Finding the required sample size (FYI)

- Pre-posterior risk (equation 1) structures finding a sample size that produces acceptable performance
- Evaluation can be by computation or simulation, using the marginal distribution of $\mathbf{X}_n$ (the distribution produced by integrating over the prior distribution)
- For the Gaussian model with known variance, $\bar{x}_n$ is sufficient, and with $\theta \sim N(\mu, \tau^2), [\bar{x}_n \mid \theta] \sim N(\theta, \sigma^2/n)$, the posterior distribution of θ is,

$$\begin{aligned} g(\theta \mid \bar{x}_n) &= N\left\{\mu + (1 - B_n)(\bar{x}_n - \mu), (1 - B_n)\frac{\sigma^2}{n}\right\} \\ B_n &= \frac{\sigma^2}{\sigma^2 + n\tau^2} \end{aligned}$$

- For $D_{1:2} = [d, \infty)$ and $\Phi(\cdot)$ the normal cdf,

$$\pi(\bar{x}_n) = 1 - \Phi\left\{\frac{d - \mu - (1 - B_n)(\bar{x}_n - \mu)}{\frac{\sigma}{\sqrt{n}}(1 - B_n)^{.5}}\right\} = \Phi\left\{\frac{\mu + (1 - B_n)(\bar{x}_n - \mu) - d}{\frac{\sigma}{\sqrt{n}}(1 - B_n)^{.5}}\right\}$$

- $\mu = 0$ is 'equipoise' producing *a priori* a 50/50 chance of being in D_0 or $D_{1:2}$.
- $\mu = 0.675$ is 'optimistic' producing *a priori* $\text{pr}(D_{1:2}) \approx 0.63$.
- $\mu = -0.675$ is 'pessimistic' producing *a priori* $\text{pr}(D_{1:2}) \approx 0.37$.

Results (via computation, not simulation)

Scenario	Sample Size (n) 10	20	50	100	300
$\mu = 0$ (equipoise)					
$c = 1, \pi^* = 0.50$	18	13	**9**	6	4
$c = 3, \pi^* = 0.75$	29	22	14	10	6
classification error	26	19	12	9	5
$c = 49, \pi^* = 0.98$	48	43	34	24	15
classification error	48	43	**32**	24	15
$\mu = 0.675$ (optimism)					
$c = 1, \pi^* = 0.50$	11	8	5	4	2
$c = 3, \pi^* = 0.75$	22	16	10		
classification error	17	13	8		
$\mu = -.675$ (pessimism)					
$c = 1, \pi^* = 0.5$	10		5		
$c = 3, \pi^* = 0.75$	14		8		
classification error	13		7		

100×pre-posterior risk: Rows led by values of c and π^* report risk computed with the $c-$value used to produce the optimal rule; rows led by 'classification error' report performance of the same rule, but evaluated with the $(1, 1)$ loss function. All entries are for $d = 0, \sigma^2 = 4, \tau^2 = 1$.

Comments (FYI)

- $\mu = 0$ is 'equipoise' producing *a priori* a 50/50 chance of being in D_0 or $D_{1:2}$.
- $\mu = 0.675$ is 'optimistic' producing *a priori* $\text{pr}(D_{1:2}) \approx 0.63$.
- $\mu = -0.675$ is 'pessimistic' producing *a priori* $\text{pr}(D_{1:2}) \approx 0.37$.
- The classification rule based on $c = 1$, minimizes classification error and has risk smaller than the classification error associated with a rule generated with a different $c-$value, for example, $25.7 > 17.9; 32.4 > 8.8$.
- Not surprisingly, classification error increases with c because the optimal rule gives increasingly discrepant costs to the two types of error.
- For a given scenario, the optimal risk for the risk function used to compute the rule and the classification error are quite close, with the discrepancy decreasing as c increases.
- The method can be used to find the necessary sample size, either by computing for a fine grain of $n-$values, and identifying the sample size that works, or implementing an interval-halving search
- Results show that for the risk to be below 10% for $(\mu = 0, c = 3)$ requires $n \approx 100$, producing a risk of 10.3 and a classification error of 8.9.

Robustness evaluation (FYI)

- In the foregoing $\pi(\mathbf{x}_n)$ is generated by the *working model* (e.g., the assumed prior G and data model f)
- If it is different from the *true model*, it generates *Regret* relative to the truly optimal rule
- Regret can be evaluated using equation (1) with the decision rule determined by the working model, but the distribution for $\mathbf{x}_n$ and the function $\pi(\mathbf{x}_n)$ produced by the true model,

$$\textit{Regret} \quad = \quad \text{(Working model risk) - (Bayes Risk)}$$

- Or, alternatively compute

$$\text{Relative Regret} \quad = \quad \frac{\text{Regret}}{\text{Bayes risk}} = \frac{\text{Working model risk}}{\text{Bayes Risk}} - 1.0 = \text{RelRisk} - 1.0.$$

- With $\tilde{\pi}(\mathbf{x}_n)$ the posterior under the true model and $\tilde{\delta}(\mathbf{x}_n)$ the optimal rule for it, $\tilde{E}$ computes under the true model for $\mathbf{X}$,

$$\begin{aligned} \text{Regret} \quad &= \quad \tilde{E}(\min[\pi(\mathbf{X}_n), c \cdot \{1 - \pi(\mathbf{X}_n)\}]) - \tilde{E}(\min[\tilde{\pi}(\mathbf{X}_n), c \cdot \{1 - \tilde{\pi}(\mathbf{X}_n)\}]) \geq 0 \\ &= \quad \tilde{E}\{\tilde{\pi}(\mathbf{X}_n) I_{\{\pi(\mathbf{X}_n) < \pi^* < \tilde{\pi}(\mathbf{X}_n)\}} + c(1 - \tilde{\pi}(\mathbf{X}_n)) I_{\{\tilde{\pi}(\mathbf{X}_n) < \pi^* < \pi(\mathbf{X}_n)\}}\} \end{aligned}$$

Frequentist risk and regret (FYI)

- Frequentist risk and regret is a special case with $\theta \equiv \theta_0$ fixed at a single value (equivalently, a prior with a point mass at θ_0)
- For this working model, the optimal rule is $\tilde{\delta}(\mathbf{x}_n) \equiv I_{\{\theta_0 \in D_{1:2}\}}$
- For example, let $c = 1$, $D_{1:2} = [0, \infty)$ and $\theta > 0$, and a working model wherein $\pi(\mathbf{x}_n) \geq 0.5 \iff \bar{x}_n \geq 0$. Then, the regret is the probability of mis-classification, and with $\theta_0 > 0$ we have, $\text{pr}(\pi(\mathbf{x}_n) < 0.5 \mid \theta_0) = \text{pr}(\bar{x}_n < 0 \mid \theta_0) = \Phi\left(-\frac{\sqrt{n}\theta_0}{\sigma}\right)$, which is (1 - Power).
- With $\theta_0 = 0$, $c = 1$, the Type I error is always $\alpha = 0.5$. For a general c, $\alpha = 1/(1 + c)$ and so to produce Type I error $= \alpha_0$, use $c = (1 - \alpha_0)/\alpha_0$.
- This relation shows that selecting the nominal α_0 can be justified by the loss function in equation (1). For $\alpha_0 = 0.05$, $c = 19$ $(= .95/.05)$, a 19:1 penalty for false rejection relative to false non-rejection.
- Going in the other direction, $\pi^*(c) = 0.98$ produces $c = 49$, a large penalty.
- In general, the Type I error associated with an informative prior and a loss function determined value of c will not be close to a traditional α_0, and forcing equality by changing c will degrade Bayesian performance

Bayesian Monitoring: The BLOCK HF trial[13]

- Intention-to-treat was the primary analysis for all outcomes
- The trial used an adaptive Bayesian design allowing a maximum of 1200 patients, featuring sample size re-estimation and two interim analyses with pre-specified, adaptive rules for stopping enrollment or terminating follow-up
- These rules addressed patient safety, futility, and eventual trial success
- The safety stopping rule, assessed at each interim analysis, was based on the posterior probability of an increased risk of primary endpoints in patients with BiV pacing relative to RV pacing
- Enrollment and follow-up termination was based on the predictive probability of passing the primary objective (PP_0) or on futility (PPR), projected to when all subjects had been followed for at least 12 months
- Low information priors were used

[13] Curtis et al. (2013). Biventricular Pacing for Atrioventricular Block and Systolic Dysfunction. *NEJM*, 368: 1585–1593.

BLOCK-HF decision table

Decision Boundaries					
	Conclude objective is met and stop study early	Conclude that sample size is sufficient to continue	Determine that sample size is insufficient but elect not to increase sample size	Conclude that sample size must be increased in increments of 175	Stop study for safety
First Interim Analysis	$PP_0 > 0.99$	$0.90 \leq PP_0 \leq 0.99$	$PRR > 0.9$	$PP_0 < .90$ and $PRR \leq 0.9$	P ($\theta >$ 0|data,prior) $\geq$ 0.90
Sample Size Re-estimation Phase	N/A	$0.90 \leq PP_0$	$PRR > 0.9$	$PP_0 < .90$ and $PRR \leq 0.9$	N/A
Second Interim Analysis	$PP_0 > 0.99$	If neither the outcome in column 2 nor the outcome in column 6 occurs, then the study will continue with the current sample size.			P ($\theta >$ 0|data,prior) $\geq$ 0.90

Post-hoc, Bayesian Monitoring of the CPCRA-TOXO trial[14,15,16]

(CPCRA = Community Programs for Clinical Research on AIDS)

- Eligibility
  - Either an AIDS defining illness or CD4 < 200
  - Or, a positive titre for *toxoplasma gondii*
- Originally designed with four treatment groups
  - Active & placebo clindamycin, 2:1
  - Active & placebo PYRImethamine, 2:1
- The clindamycin arm was stopped after a few months, so consider PYRI vs Placebo

---

[14] Chaloner, Church, Louis, Matts (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician,* 42: 341-353.

[15] Carlin, Chaloner, Church, Louis, Matts (1993). Bayesian approaches for monitoring clinical trials, with an application to toxoplasmic encephalitis prophylaxis. *The Statistician*, 42: 355-367.

[16] Brownstein, Louis, O'Hagan, Pendergast (2019). The role judgement in statistical inference and evidence-based decision-making. *The American Statistician*, doi.org/10.1080/00031305.2018.1529623.

After-the-fact analysis of the Toxo Trial[17]

The DSMB monitored it in real time

- Elicited priors from three HIV/AIDS clinicians, one PWA conducting AIDS research, and one AIDS epidemiologist
- Used the Cox model and adjusted for baseline CD_4
- 'Stopped" when the posterior probability of benefit or the posterior probability of harm got sufficiently high
- Used a variety of prior distributions, including an equally-weighted mixture of the five elicited priors

[17] Jacobson, et al. (1994). Primary prophylaxis with pyrimethamine for toxoplasmic encephalitis in patients with advanced human immunodeficiency virus disease: Results of a randomized trial. *The Journal of Infectious Diseases*, 169: 384–394.

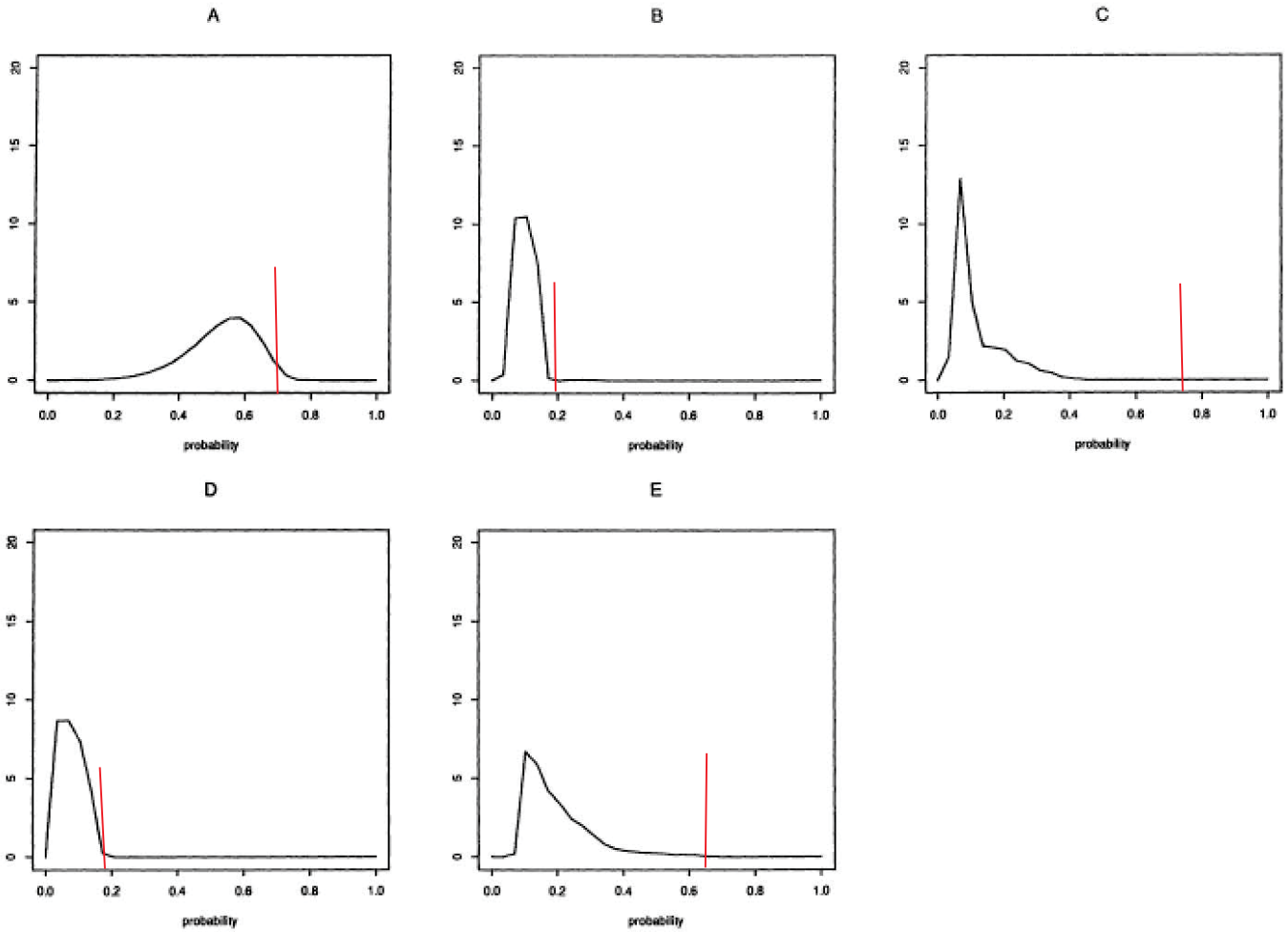
Elicitation

Asked about potential observables

- P = pr[event in two years]
- P_0 = best guess for the placebo
  - mode, median, mean
- Then, distribution of $[P_{pyri} \mid P_0]$
  - percentiles
  - draw a picture
- Then, convert to a Cox model-relevant parameter:

$$\theta = \beta_1 = \log(1 - P_0) - \log(1 - P_{pyri})$$

Elicited Priors



- Red line is at the best guess for the two-year rate under placebo

Actual TOXO Monitoring

- At its meeting on 12/31/91, the DMC recommended stopping due to:

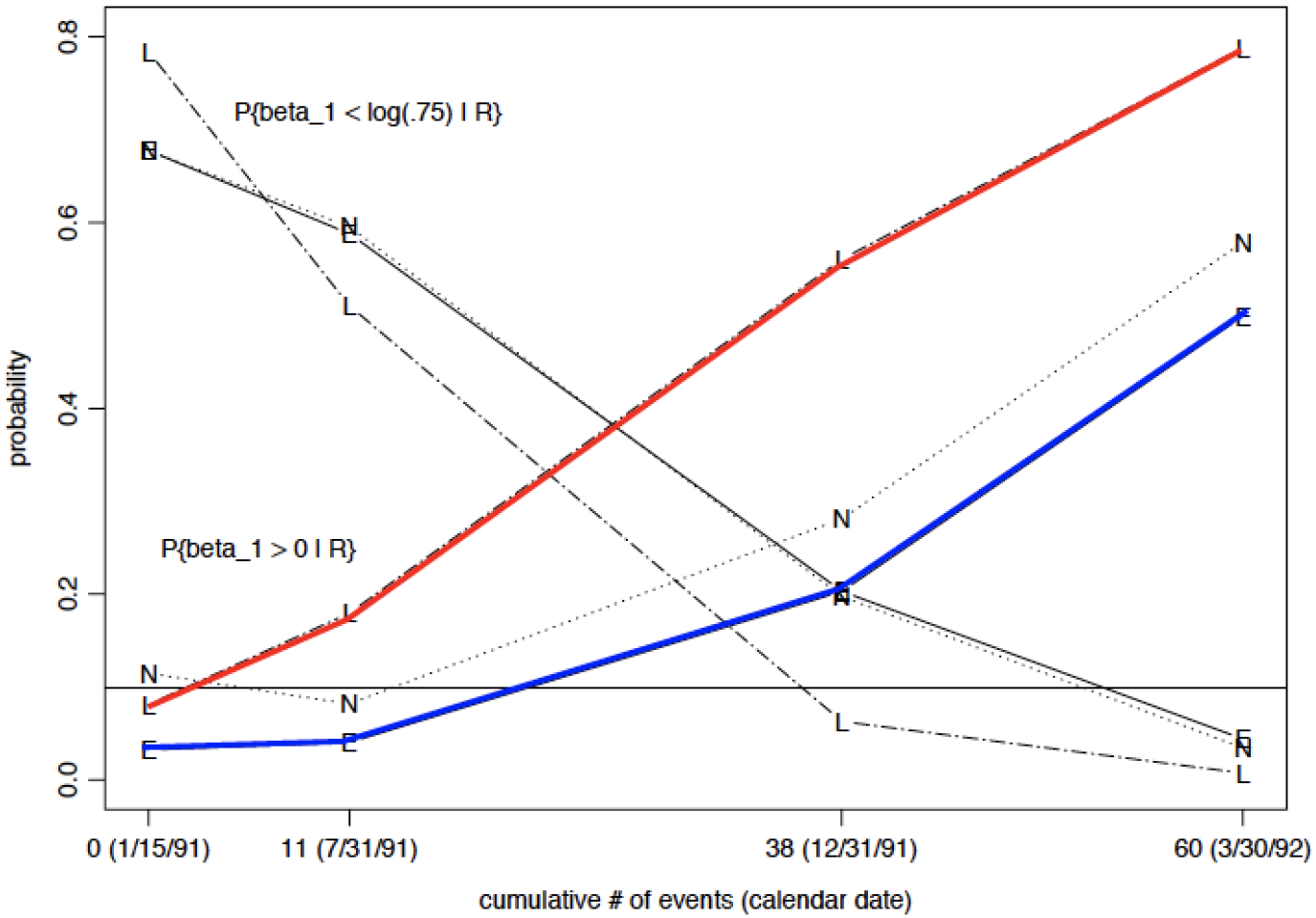
  Futility: The pyrimethamine group had not shown significantly fewer TE events, and the low overall TE rate made a statistically significant difference unlikely to emerge

  Harm: There was an increase in the number of deaths in the pyrimethamine group relative to the placebo

Mixture prior ⇒ Posterior Probabilities of regions

(Bayes with the mixture prior takes longer to stop)

E = exact; N = normal approximation; L = likelihood



Observations

- The elicited priors are very far from the eventual data because elicitees believed that TE was common in the patient population and Pyrimethamine would have a substantial prophylactic effect
- Consequently, the likelihood-based ('flat prior" Bayes) analysis gave an earlier warning than did the Bayesian assessments due to,

$$\text{High: } \text{pr}(\theta > 0 \mid \text{data}) \quad \& \quad \text{Low: } \text{pr}(\theta < \log(0.75) \mid \text{data})$$

Likely Harm Unlikely Benefit

- Eventually, the data overwhelmed the elicited priors

Observations

- The elicited priors are very far from the eventual data because elicitees believed that TE was common in the patient population and Pyrimethamine would have a substantial prophylactic effect
- Consequently, the likelihood-based ('flat prior" Bayes) analysis gave an earlier warning than did the Bayesian assessments due to,

$$\text{High: } \text{pr}(\theta > 0 \mid \text{data}) \quad \& \quad \text{Low: } \text{pr}(\theta < \log(0.75) \mid \text{data})$$

Likely Harm Unlikely Benefit

- Eventually, the data overwhelmed the elicited priors

If the elicited priors had been used in the actual monitoring, would it have been ethical to wait so that these representatives of PWAs, clinicians and HIV/AIDS researchers were convinced?

Prior partitioning: Backwards Bayes

- Motivated by Mosteller&Wallace[18], Carlin&Louis[19] consider identifying prior distributions that in the light of observed data would lead to various decisions, using the CPCRA/TOXO trial[20] as an example
- Partitioning uses the Bayesian framework to put bounds on priors leading to specific decisions; a stakeholder can decide if the boundaries are so extreme in one direction that the decision is the same for most priors
  - This is 'backwards Bayes'
- The approach is similar to threshold utility analysis and in the same spirit as sensitivity analysis for non- or weakly- identified parameters
- Partitioning can be completely unconstrained, or restricted by moment or percentile restrictions, or based on regions for parameters in a parametric prior
- 'Pure' or nearly pure Bayesians find this use of the Bayesian formalism close to apostasy, but it can be effective in quantifying the strength of evidence provided by a data set
- The following figure display regions, conditional on the observed data, where there is or is not a prior distribution that permits or does not permit rejecting H_0

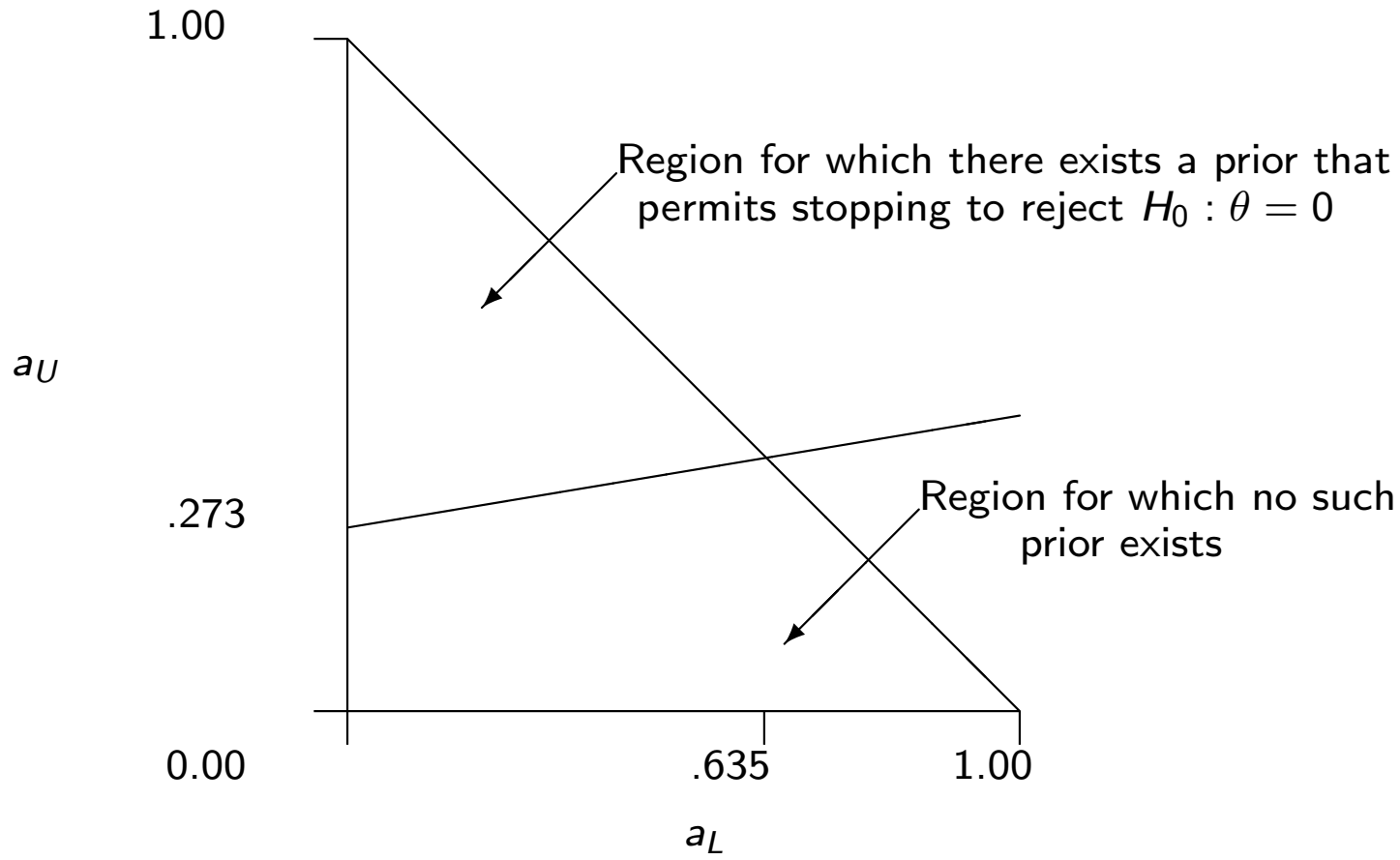
[18] Mosteller, Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Addison–Wesley.

[19] Carlin, Louis . (1995) Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials. In, *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (eds. D Berry, K Chaloner, J Geweke), 493–503. Wiley, New York.

[20] Jacobson, et al. (1994). Primary prophylaxis with pyrimethamine for toxoplasmic encephalitis in patients with advanced human immunodeficiency virus disease: Results of a randomized trial. *The Journal of Infectious Diseases*, 169: 384–394.

Prior tail area regions where there is/(is not) a prior distribution that permits rejecting $H_0: \theta = 0$

(Conditional on the TOXO trial data)



- $a_L = \text{pr}_{prior}\{\theta < \log(0.75)\}$; $a_U = \text{pr}_{prior}(\theta > 0)$

Priors: On what? (and consequences)

- $S(t) = \text{pr}(\text{event time} > t)$ ~Beta(1, 1) (flat)
- If $S(t) = S(t \mid \lambda) = e^{-\lambda t}$,
  - Prior on $\lambda = -\frac{1}{t}\log\{S(t \mid \lambda)\}$ (exponential with hazard t)
  - Prior on $S(2t) =$ prior on $S^2(t)$ $\{$is Beta$\left(\frac{1}{2}, 1\right)\}$
- Going the other way, a flat prior on λ induces an improper prior with density proportional to $\frac{1}{st}$ on $S = e^{-\lambda t}$
- Etc.
- Morals:
  - Explore consequences of priors
  - Elicit priors for features that an expert might know something about

Prior for two studies

- Have two studies, each with a treatment effect wrt independent comparators that are statistically identical
- (θ_1, θ_2) = treatment effects for studies 1 and 2, with '1' preceding '2'
  - The θs could be log(odds)

$$\begin{aligned} \eta &= \frac{\theta_1 + \theta_2}{2}, \quad \delta = \frac{\theta_2 - \theta_1}{2} \\ \eta &= \text{the average treatment effect} \\ \delta &= \text{the between study difference in treatment effects} \\ \theta_1 &= \eta - \delta, \quad \theta_2 = \eta + \delta \end{aligned}$$

- Priors

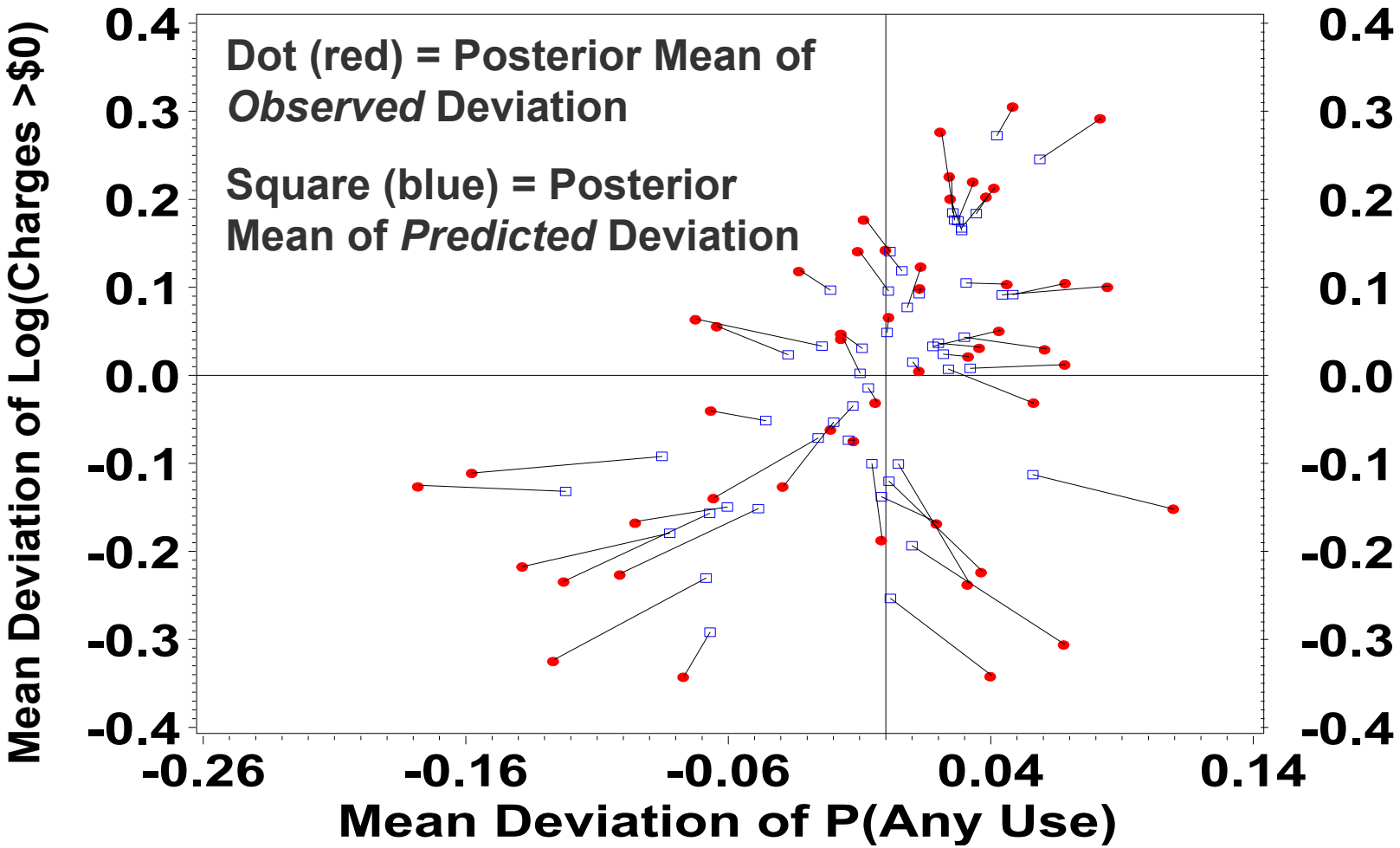
$$\begin{aligned} \eta &\sim (\mu, \tau^2), \quad \delta \sim (0, \xi^2) \\ \text{cov}(\theta_1, \theta_2) &= \text{cov}\{(\eta - \delta), (\eta + \delta)\} = \tau^2 - \xi^2 \\ \rho = \text{cor}(\theta_1, \theta_2) &= \frac{\tau^2 - \xi^2}{\tau^2 + \xi^2} \end{aligned}$$

- If $\tau^2 \neq \xi^2, (\theta_1, \theta_2)$ are correlated, and information on θ_1 produces an updated prior for θ_2 even though there is no direct information
- A small ξ allows δ to be stochastically small (similar treatment effects) while retaining appropriate uncertainty on η
  - For binomial responses, need to use MCMC, but we can do that!

Don't trust your intuition

- Shrinkage 'towards the mean' can be
  - 'Away from' if the distribution is multi-modal, univariate
  - 'Away from' or 'beyond' when evaluating the univariate consequences of bivariate shrinkage
  - 'Almost anything' for models with correlated random effects
- Here are health services, multivariate measurement error, and malaria prevalence examples

Observed and Predicted Deviations for *Primary Care Service*: Log(Charges>$0) and Probability of Any Use of Service



Don't trust your intuition

Multivariate Measurement Error: Simulation example

- $\mathbf{X}_t$ and $\mathbf{X}_o$ are vector regressors
  - For example, one coordinate is the exposure of interest and the other is a potential confounder, or data on 6 dietary components
- Measurement error, especially correlated error, can confound confounding adjustments and standard measurement error adjustments
- Formal modeling appropriately accounts for the measurement error process, commonly producing non-intuitive adjustments
- Information on the joint measurement error distribution is necessary

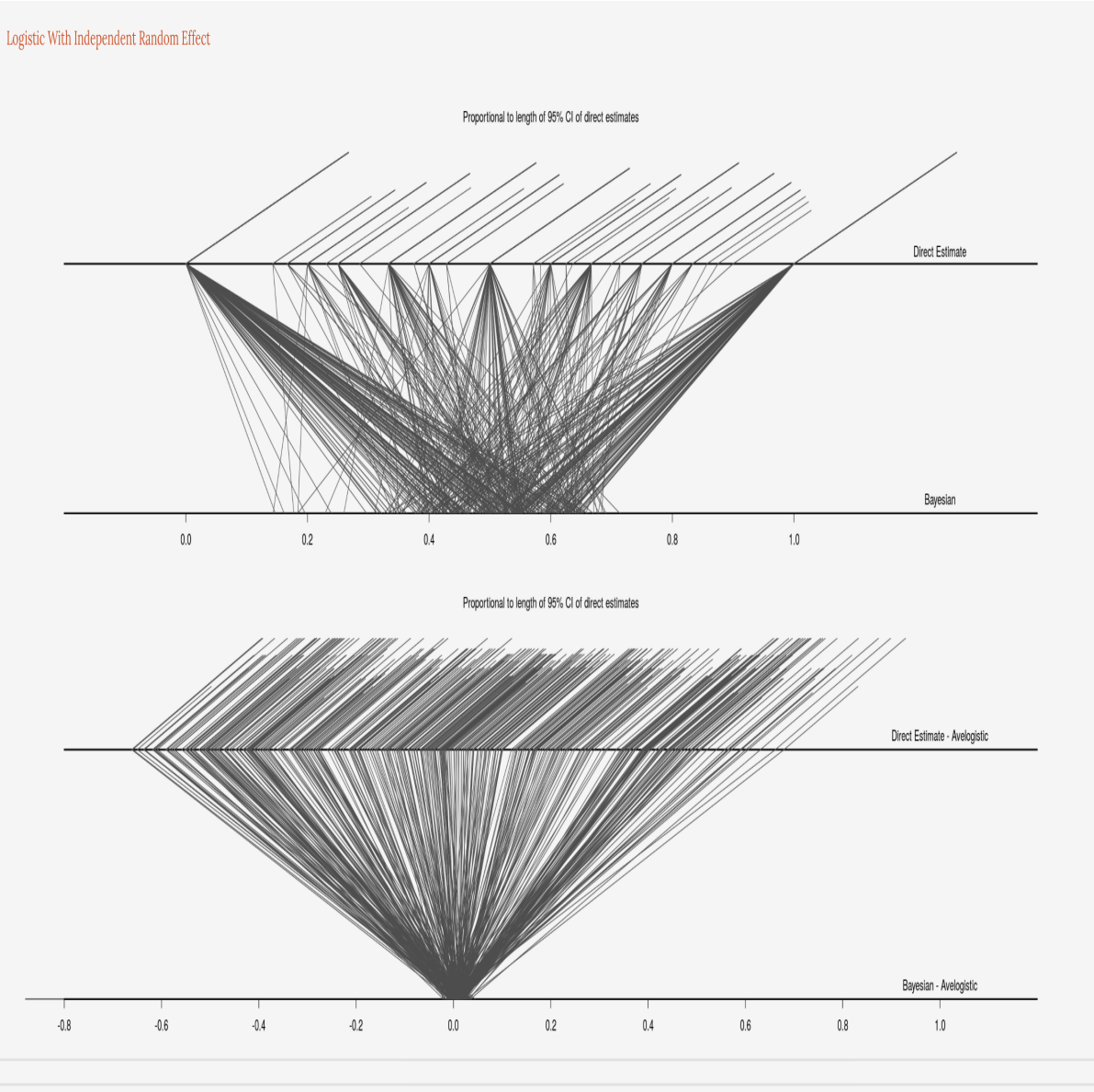
	Coefficients ($\times 10^4$)			
Regressor	Unadj.	Univ adj.	Mult. adj.	True
sodium	7	19	23	21
potassium	7	14	-20	-15
calcium	3	7	11	11
caffeine	-19	-30	-31	-30
alcohol	903	1474	1528	1528
bmi	1348	1443	1645	1657

Measurement Error: High, Moderate, Low

- De-attenuation AND crossing 0

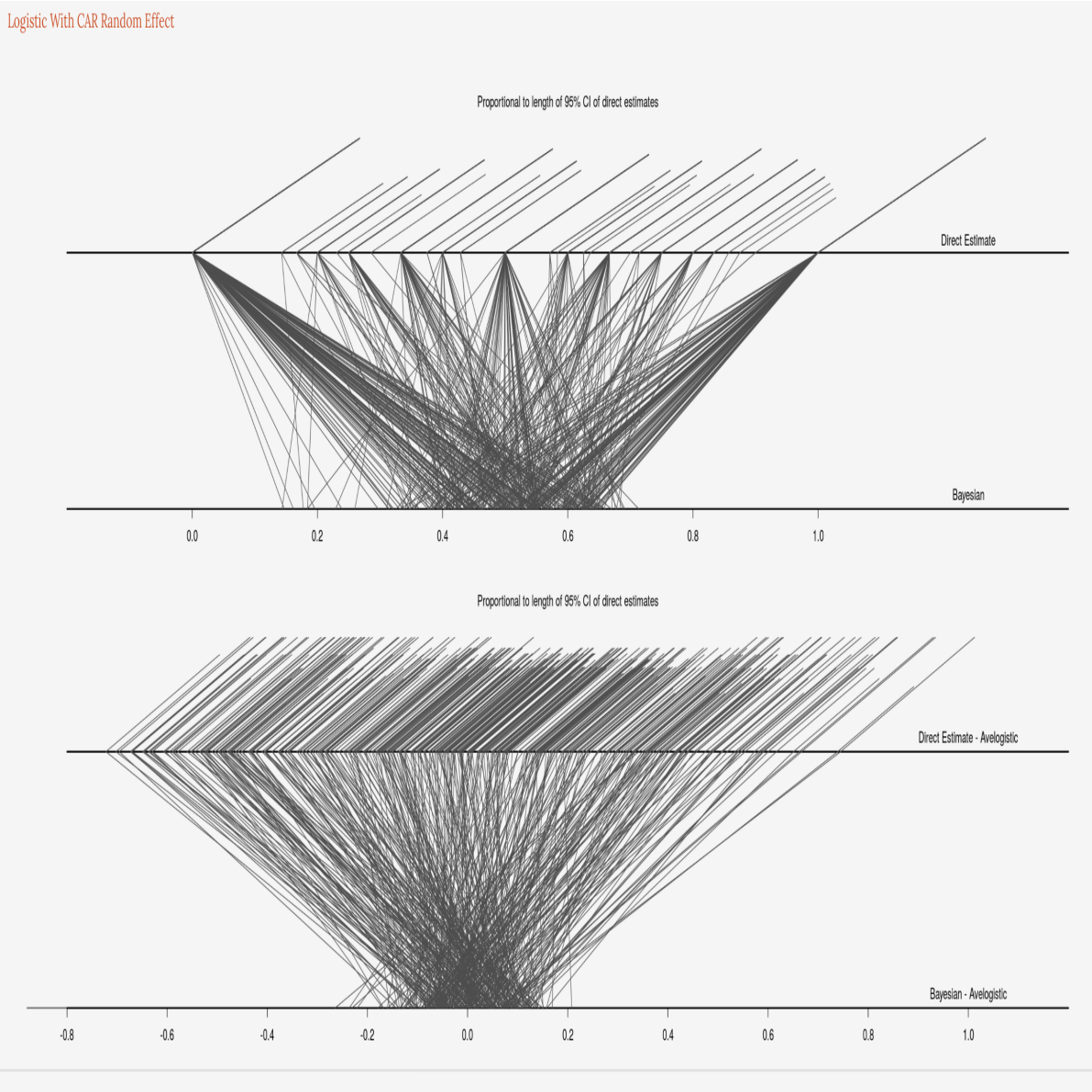
Nchelenge Zambia Malaria Prevalence: Independent RE model with covariates

Residuals shrink towards 0



Nchelenge Zambia Malaria Prevalence: Conditional AutoRegressive model with covariates

Residuals can cross over 0



Informative sample size in Bayesian analysis

The deck may be stacked against high-variance units

- The posterior mean (PM_k) for a Gaussian/Gaussian, Bayesian model is:

$$\begin{aligned} \text{PM}_k &= B_k \mathbf{X}_k \beta + (1 - B_k) Y_k \\ B_k &= \sigma_k^2/(\sigma_k^2 + \tau^2) \end{aligned}$$

  a weighted average of the direct estimate (Y_k) and a regression prediction ($\mathbf{X}_k\beta$) with larger B_k for the relatively unstable direct estimates

- $\hat{\beta}^{mle}$ gives more weight to the units with relatively stable direct estimates; the high B_k units that 'care about' the regression model have less influence, and if the model is mis-specified, PM_k will be unfair to them
- Giving them relatively more weight will pay variance, but can improve MSE[21,22]

[21] Jiang, Nguyen, Rao (2011). Best Predictive Small Area Estimation. *JASA*, 106: 732-745

[22] Chen, Jiang, Nguyen (2015). Observed Best prediction for small area counts. *Journal of Survey Statistics and Methodology*, 3: 136–161.

Informative sample size in Bayesian analysis

The deck may be stacked against high-variance units

- The posterior mean (PM_k) for a Gaussian/Gaussian, Bayesian model is:

$$\begin{aligned} \text{PM}_k &= B_k \mathbf{X}_k \beta + (1 - B_k) Y_k \\ B_k &= \sigma_k^2/(\sigma_k^2 + \tau^2) \end{aligned}$$

  a weighted average of the direct estimate (Y_k) and a regression prediction ($\mathbf{X}_k\beta$) with larger B_k for the relatively unstable direct estimates

- $\hat{\beta}^{mle}$ gives more weight to the units with relatively stable direct estimates; the high B_k units that 'care about' the regression model have less influence, and if the model is mis-specified, PM_k will be unfair to them
- Giving them relatively more weight will pay variance, but can improve MSE[21,22]

Hospital Profiling

- **Practice makes perfect:** Small hospitals may have poorer performance than larger, for example their performance for riskier patients is worse, and giving more weight to the higher volume hospitals when estimating the risk-adjustment creates some unfairness

Small Area Estimates (SAEs) & Subgroups

- The true regression slopes may depend on population size, and predictions/inferences for the smaller domains will be degraded if $\hat{\beta}$ is the MLE

[21] Jiang, Nguyen, Rao (2011). Best Predictive Small Area Estimation. *JASA*, 106: 732-745

[22] Chen, Jiang, Nguyen (2015). Observed Best prediction for small area counts. *Journal of Survey Statistics and Methodology*, 3: 136–161.

MSE Comparisons

mmle vs mMssBias vs α_{opt} compromise

- Compromise: $\alpha_{opt} \times$ MMLE $+ (1 - \alpha_{opt}) \times$ ObservedBestPredictor
- α_{opt} minimizes MSE(α) as a function of the estimated SSqbias increment associated with mmle weights relative to mMSSbias weights:

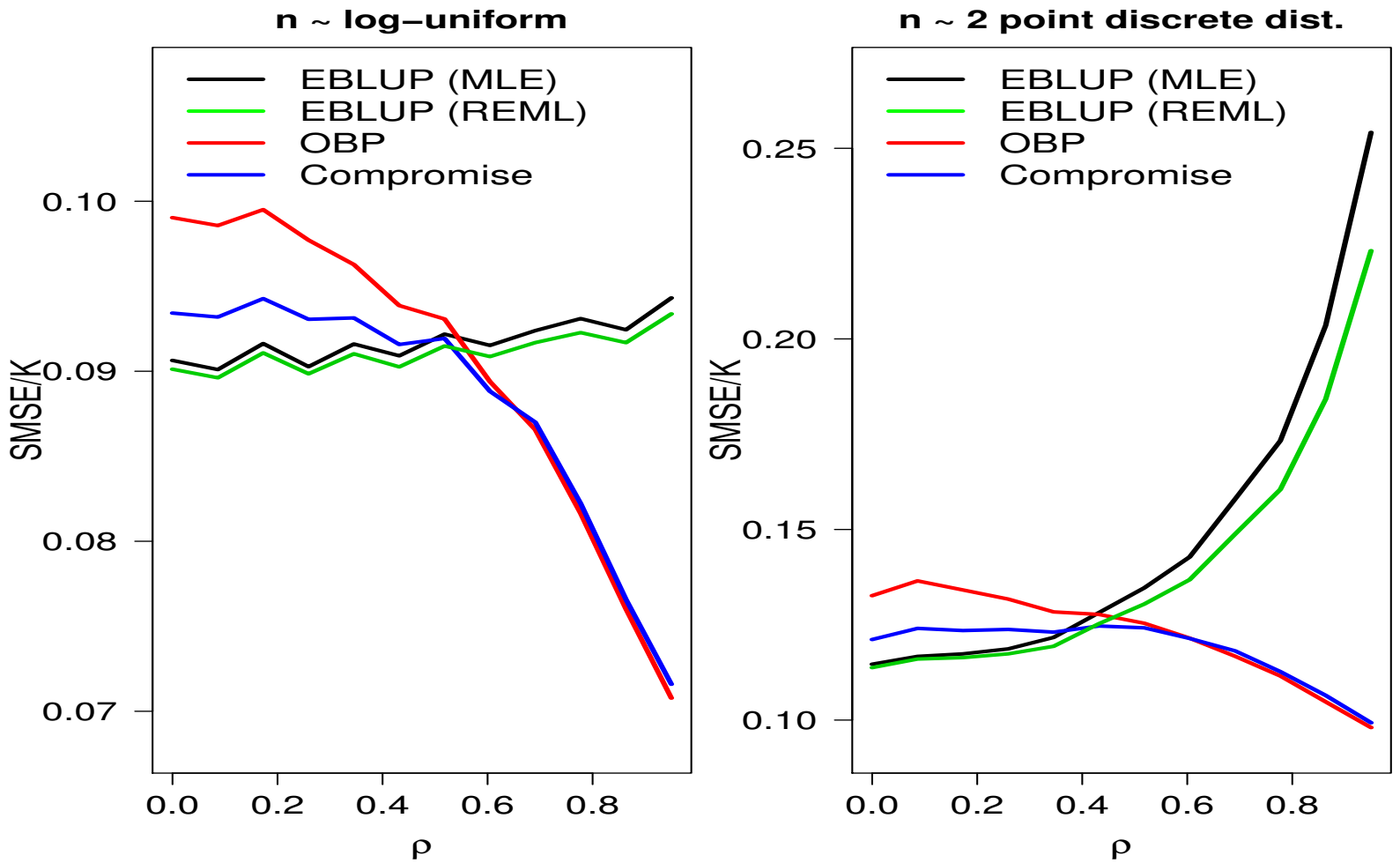
$$\hat{\Delta}^2 = \sum_k \left(\hat{\theta}_k^{mmle} - \hat{\theta}_k^{obp} \right)^2$$

- For the Gaussian model, assuming (incorrectly) that the B_k don't change

$$\hat{\Delta}^2 = \sum_k B_k^2 \left\{ \mathbf{x}_k \left(\hat{\boldsymbol{\beta}}^{mmle} - \hat{\boldsymbol{\beta}}^{obp} \right) \right\}^2$$

Gaussian/Gaussian (best case for MMLE)

$$\rho \approx \text{cor}(\theta_k, n_k)$$



Ranking: A non-standard goal

Ranking Standardized Mortality Ratios, SMRs

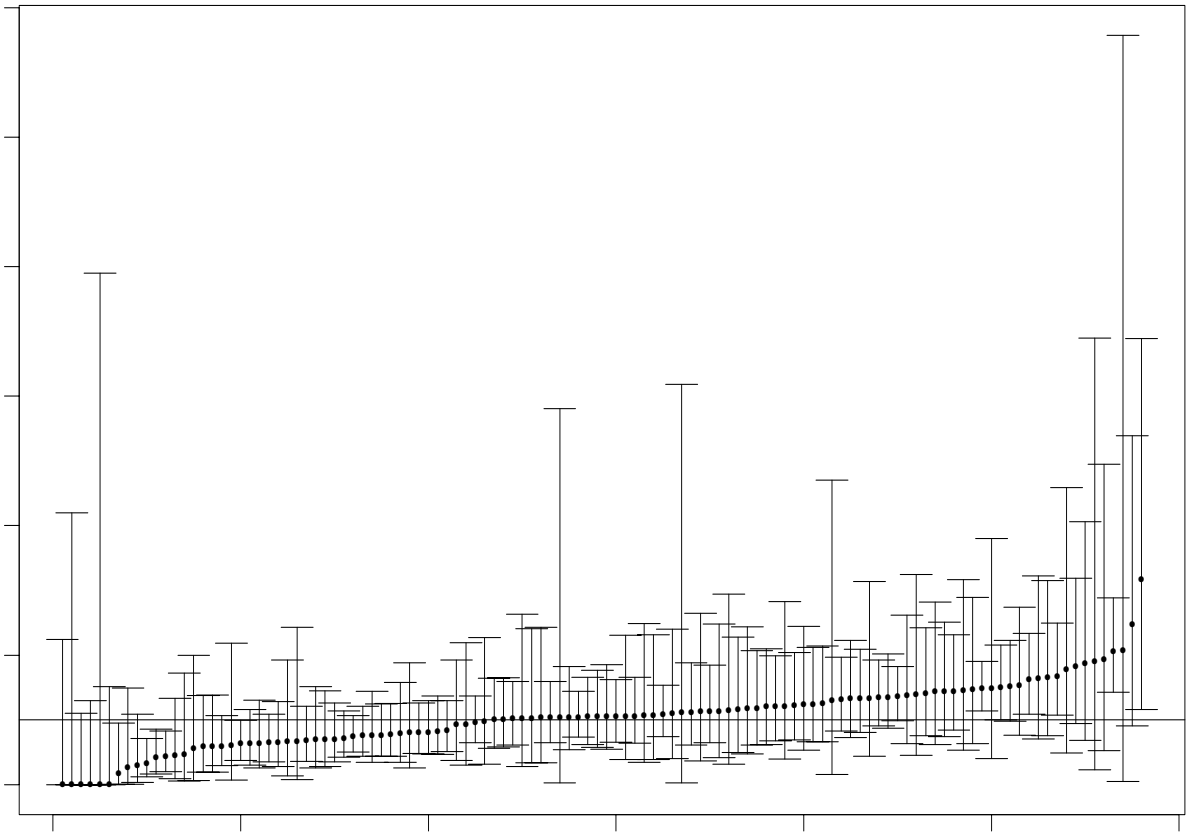
$$\text{SMR} = \frac{\text{observed deaths}}{\text{expected deaths}}$$

- Expecteds from a case mix adjustment model
- Rank 3459 dialysis providers using 1998 USRDS data
- Large and small providers, treating from 1 to 355 patients per year
- So, the expected deaths and standard errors of the estimated SMRs have a very broad relative range

The Ranking Challenge

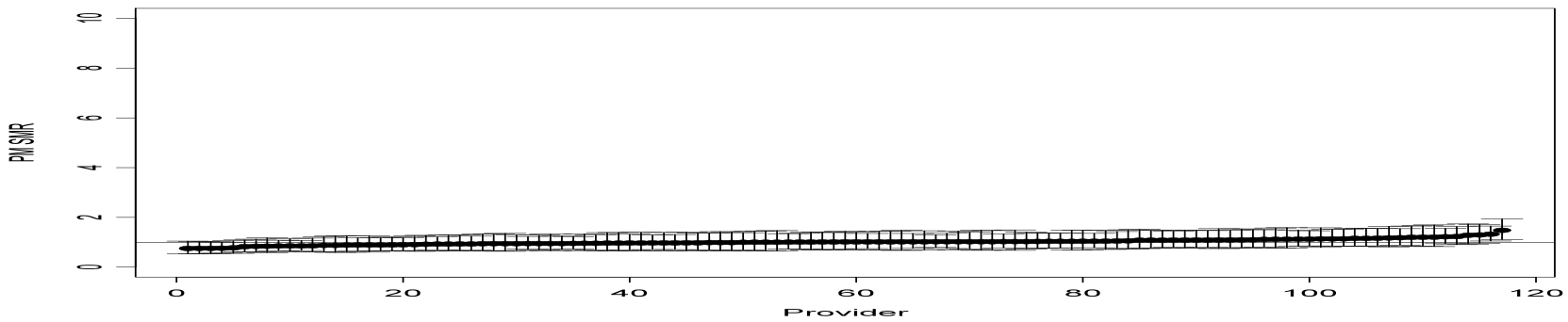
- Ranking estimated SMRs is inappropriate, if the SEs vary over providers
  - Unfairly penalizes or rewards providers with relatively high variance
- Hypothesis test based ranking: $H_0 : \text{SMR}_{unit} = 1$
  - Unfairly penalizes or rewards providers with relatively low variance
- Therefore, need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

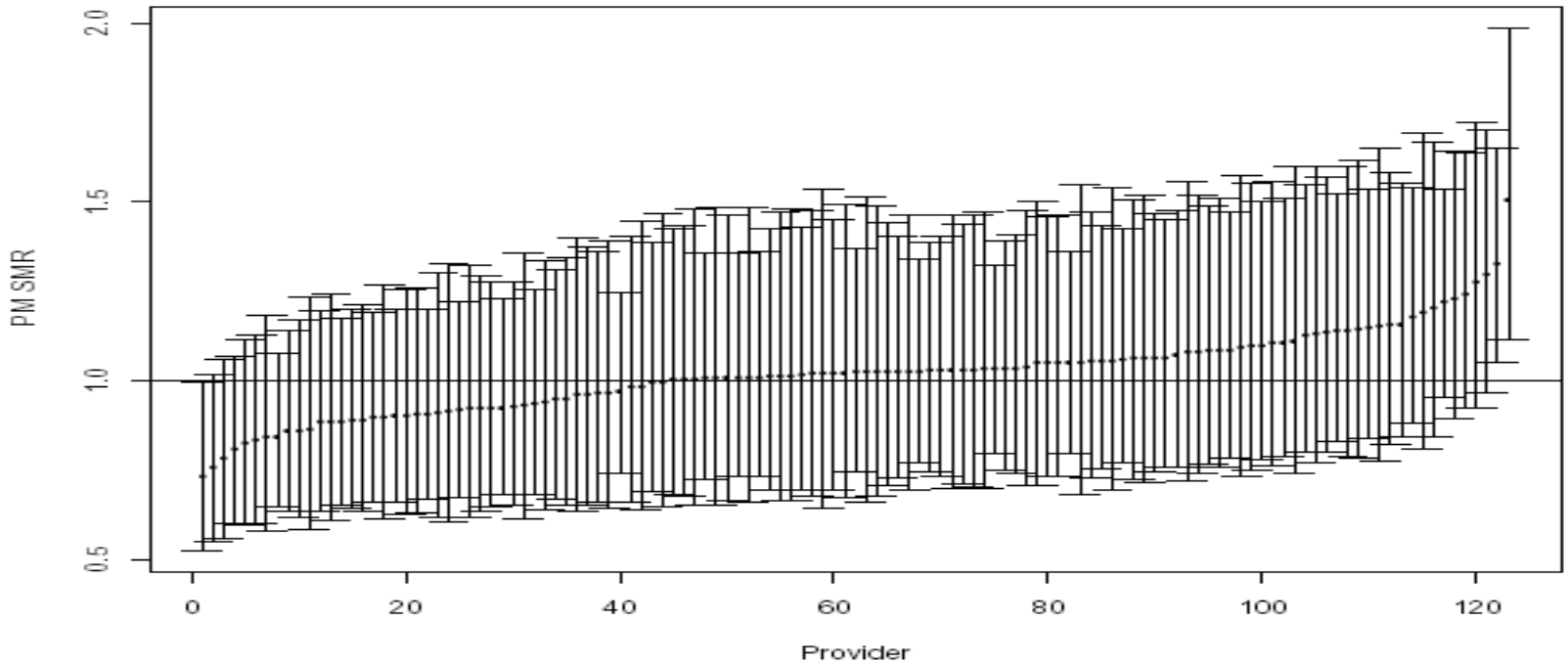
USRDS, SMRs: MLEs and exact CIs (1/40, ordered MLEs)



- Sampling variability has a wide range over units
- Difficult to trade-off signal and noise 'by hand'

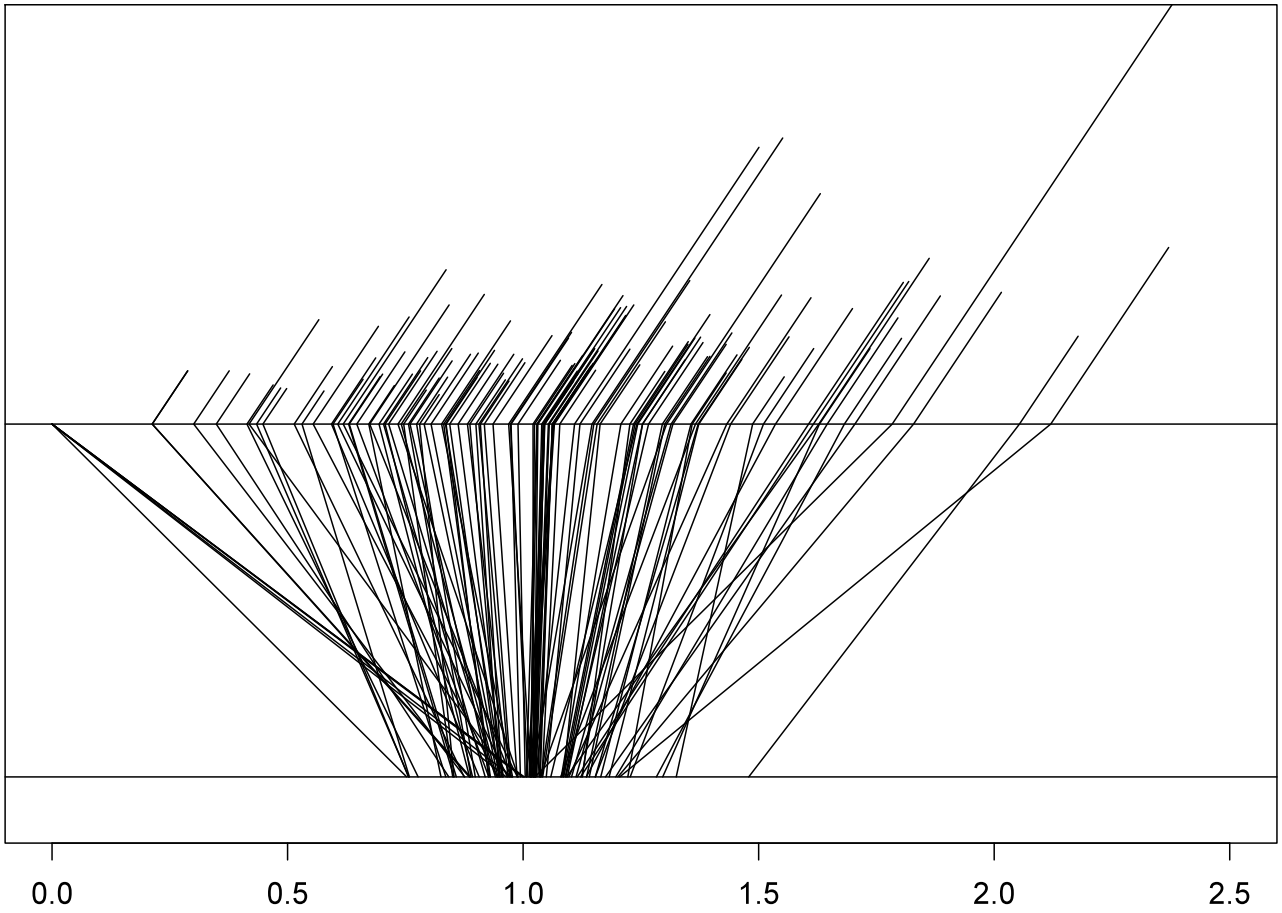
Posterior distribution: original and stretched scale





$\hat{\rho}^{mle}$, $\hat{\rho}^{pm}$, SE($\hat{\rho}^{mle}$) using USRDS dialysis data [23]

middle = MLE :: whisker = SE :: bottom = Posterior Mean



- Ranks for $\hat{\rho}^{mle}$ are different from those for $\hat{\rho}^{pm}$

[23] Lin R, Louis TA, Paddock S, Ridgeway G (2009). Ranking USRDS, provider-specific SMRs from 1998–2001. *Health Services Outcomes & Research Methodology*, 9: 22-38. DOI 10.1007/s10742-008-0040-0.

Optimal Ranks/Percentiles[24]

- 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, $\hat{R} = \text{rank}(\bar{R})$

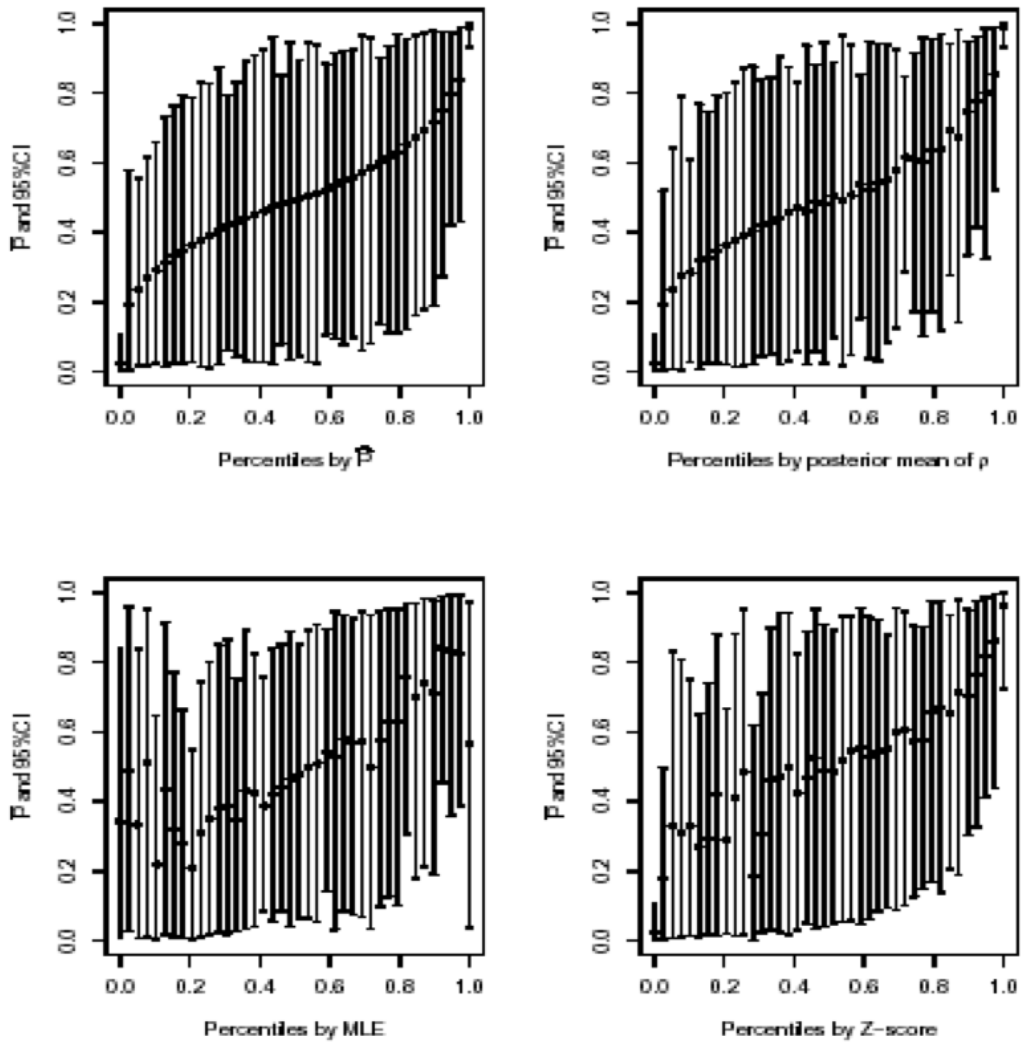
$$\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y})); \; \hat{P}_k = \hat{R}_k/(K+1)$$

- 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

[24] Shen W, Louis TA (1998). Triple-goal estimates in two-stage hierarchical models. *J. Royal Statistical Society, Ser. B*, 60: 455-471.

Relations among percentiling methods

1998 USRDS data



Performance Comparisons: Gaussian/Gaussian with σ_k^2

Lin et al. (2006) *Bayesian Analysis*

Variation in σ_k^2	Percentiles computed from: Optimal	Posterior Mean log(SMR)	Posterior Mean SMR	MLE SMR
None	516	516	516	516
medium	517	517	534	582
high	522	525	547	644

SEL performance: $10^4 \times E(P^{est} - P^{true})^2$

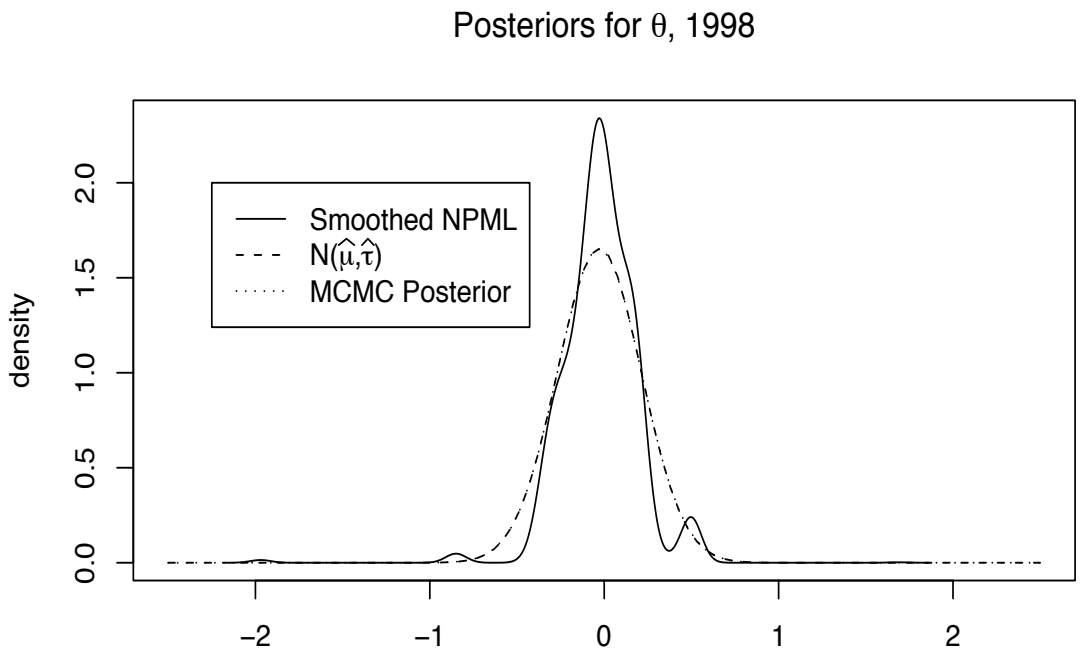
(the no-information value is **833**)

Robustness

When K is 'not small,' can use a (smooth) non-parametric or semi-parametric prior

Robustness: with large K, can us a (smooth) NP Prior

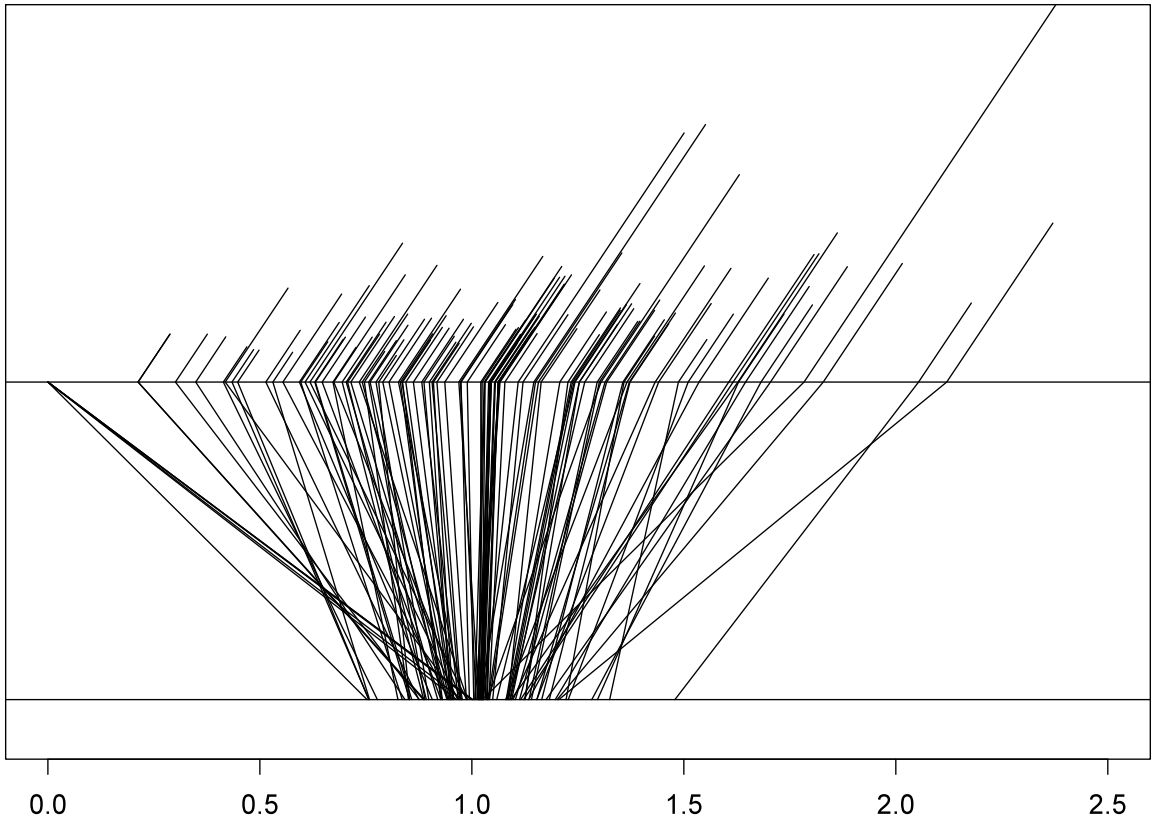
Smoothed NP - - - - - Parametric



Shrinkage can be controversial

Ash et al. (2012), COPSS White Paper

- For example the SMR for the center with greatest uncertainty is pulled all the way back to 1.0, 'hiding' the poor performance
- It is especially controversial when sample size might be informative in that low volume (high variance) units tend to perform relatively poorly (practice makes perfect) and that shrinkage masks this feature



Classification (above γ)/(below γ) loss (FYI)

Lin et al. (2006) *Bayesian Analysis*

- In some contexts interest is in identifying the several highest (top 0.1%) or lowest true ranks
- Examples include SNP identification, dialysis center performance, poverty rates in small areas, ...
- For $0 < \gamma < 1$, minimize a normalized false detection rate (FDR), denoted 'OC'

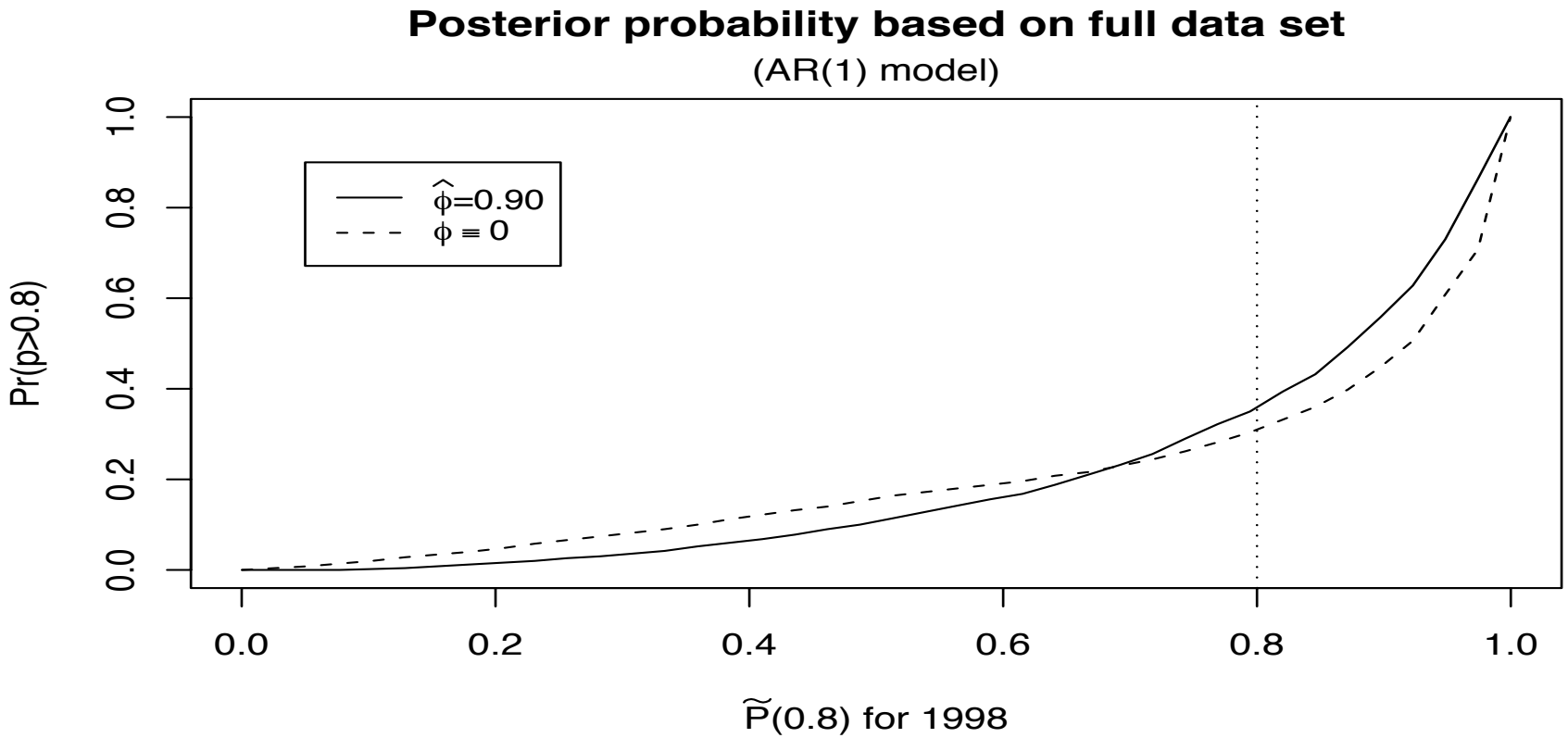
$$\mathbf{OC}(\gamma \mid \mathbf{Y}) = \frac{\mathrm{pr}(\mathbf{P} > \gamma \mid \mathbf{P}_k^{\mathbf{est}} \leq \gamma, \mathbf{Y})}{\mathbf{1} - \gamma}$$

- For optimal estimates, let

$$\begin{aligned} \pi_k(\gamma) &= \mathrm{pr}(P_k > [\gamma K]) \quad \text{(see below for an efficient computation)} \\ \tilde{P}_k(\gamma) &= \mathrm{rank}\{\pi_k(\gamma)\}/(K+1) \end{aligned}$$

- As do the $\bar{R}_k(\mathbf{Y})$, the $\pi_k(\gamma)$ quantify the strength of the ranking signal

$\pi_k(0.8 \mid \mathbf{Y})$ versus $\tilde{P}_k(0.8)$ for 1998



- Optimal percentiles and posterior probabilities computed with the single year model ($\phi \equiv 0$) and the AR1 model ($\hat{\phi} = 0.90$)

Histogram Estimates[26]

The setup:

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

- Note the finite population goal
- The optimal SEL estimate is:

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

- The optimal discrete SEL estimate is:

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

- An empirical version of Efron's Oracle, see[25]

---

[25] Efron B (2019). Bayes, Oracle Bayes, and Empirical Bayes (with discussion). *Statistical Science,* 34: 177–235.

[26] Shen W, Louis TA (1998). Triple-goal estimates in two-stage hierarchical models. *J. Royal Statistical Society, Ser. B,* 60: 455-471.

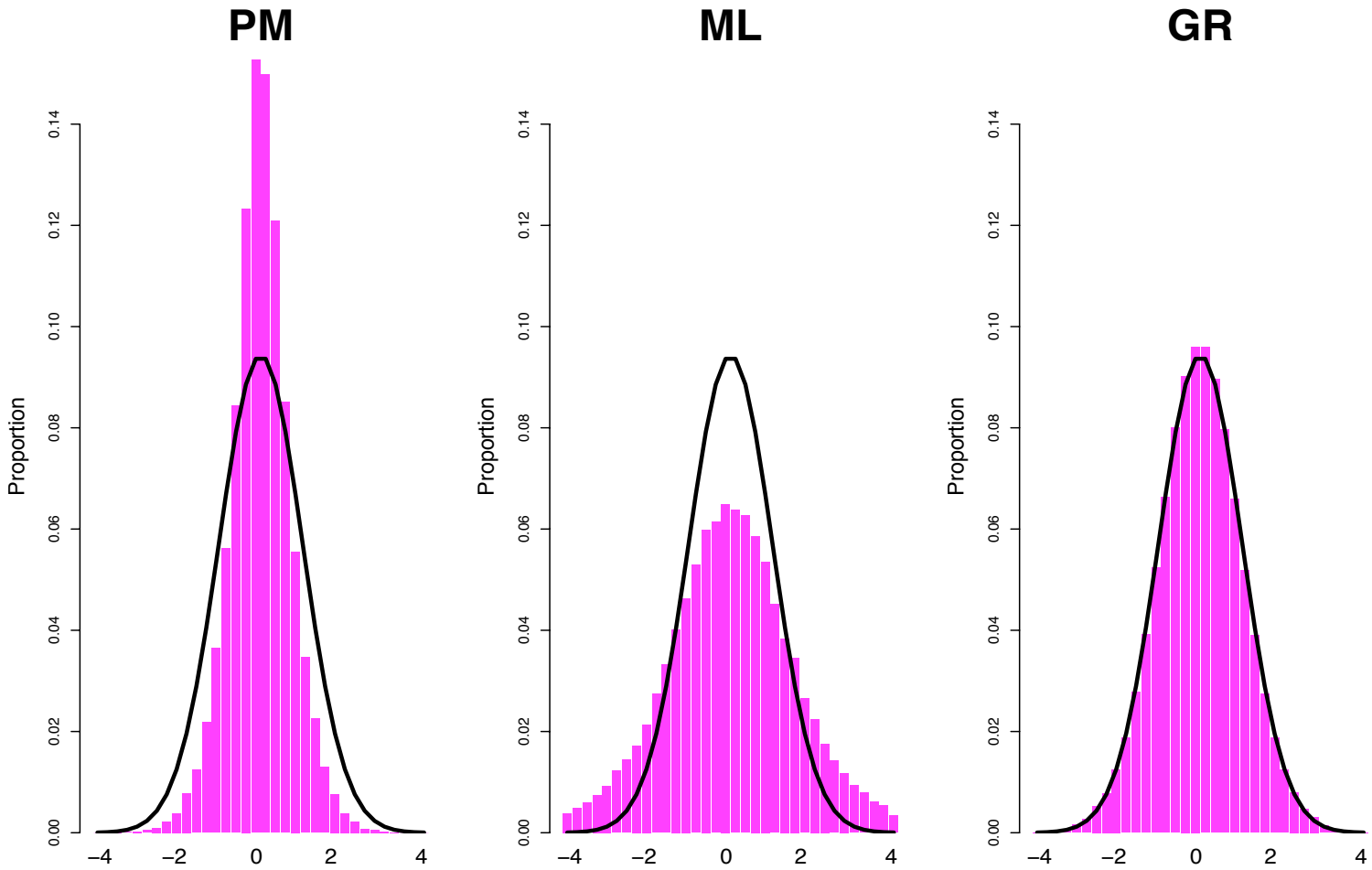
Associated mean and variance produced by $\bar{G}_K$ (and approximately those by $\widehat{G}_K$)

- Let, $\theta_k^{pm} = E(\theta \mid \mathbf{Y})$

$$\begin{aligned} \textbf{mean} &= \int t d\bar{G}_K(t) = \frac{1}{K}\sum \theta_k^{pm} = \theta_\bullet^{pm} \\ \textbf{variance} &= \int t^2 d\bar{G}_K(t) - (\theta_\bullet^{pm})^2 \\ &= \frac{1}{K}\sum V(\theta_k \mid \mathbf{Y}) + \frac{1}{K}\sum \left(\theta_k^{pm} - \theta_\bullet^{pm}\right)^2 \end{aligned}$$

- The histogram of the θ_k^{pm} is under-dispersed because it represents the second term, but not the first term
- So, use a histogram based on the mass points for $\widehat{G}_K$
- If the model is correct, $\bar{G}$ and $\hat{G}$ are consistent estimates of G with appropriate location, spread and shape

Gaussian Simulations: GR $= \hat{G}_K$, Need to get the spread right



Getting the spread right

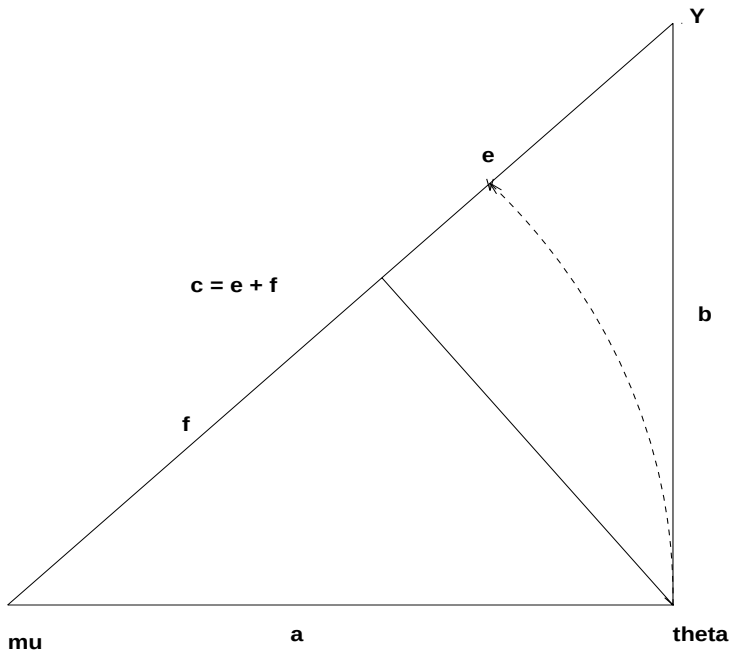


Figure 3: A triangle demonstration of the value of shrinkage

A log-normal prior

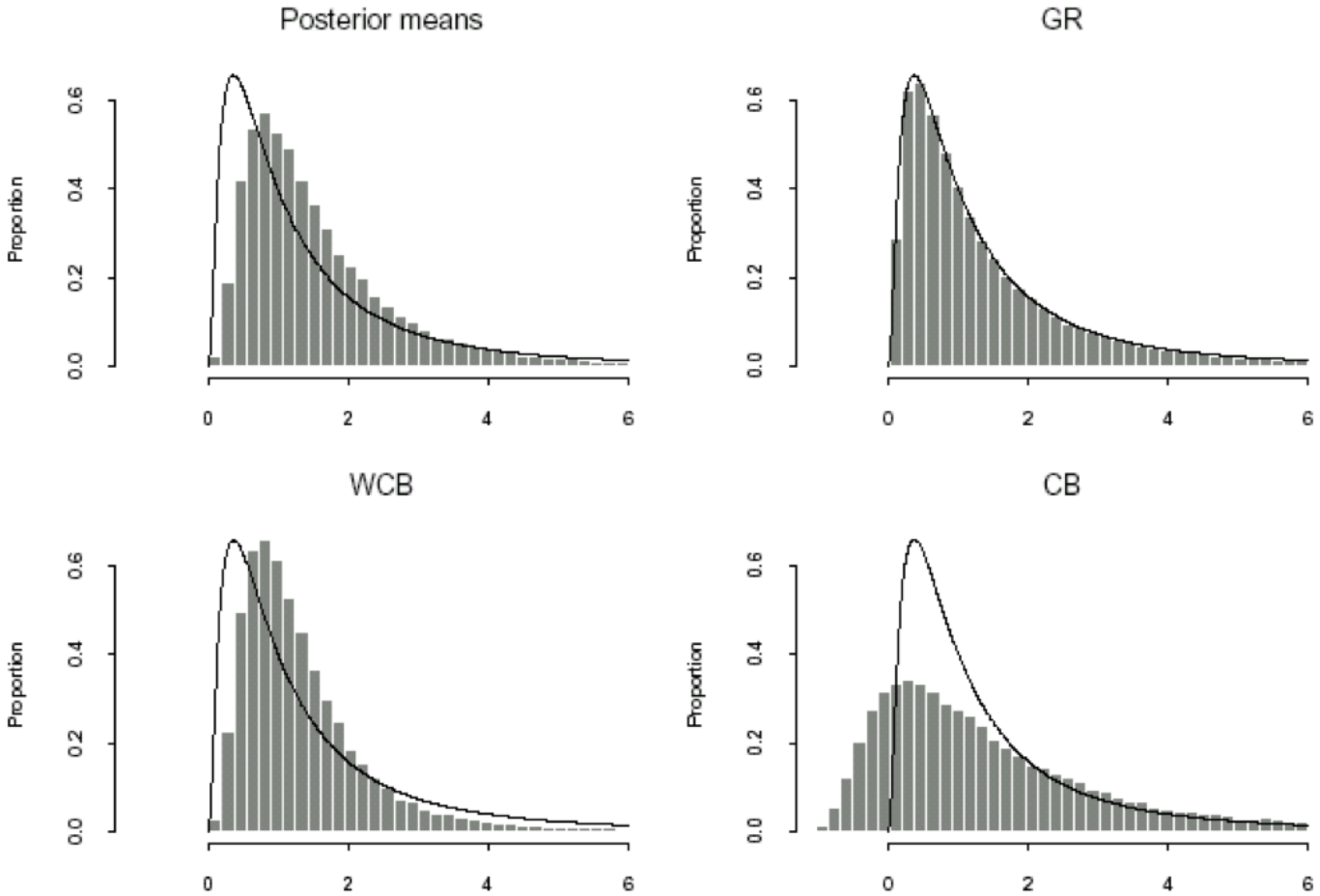
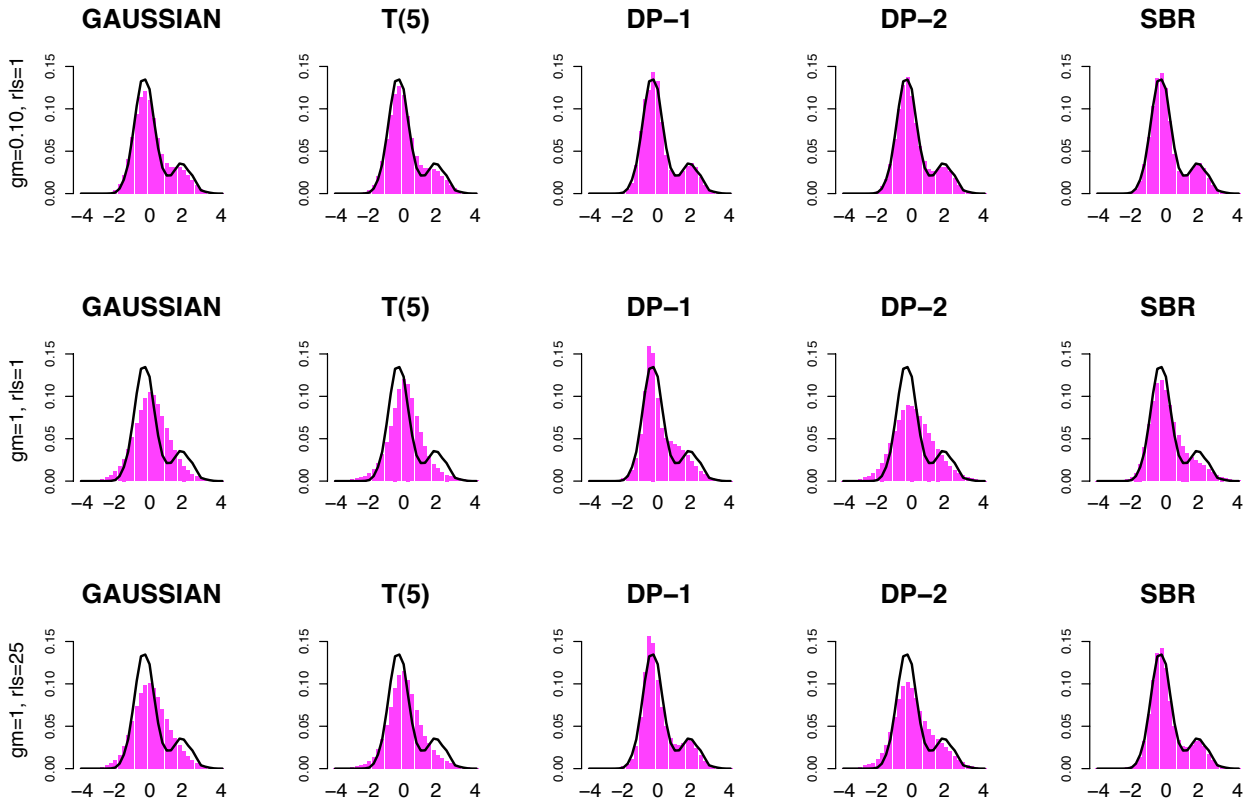


Figure 6: PM, GR and CB for a log-normal prior

Gaussian mixtures, prior variance, $\tau^2 = 1.25$

Columns are modeling priors

- $\sigma^2 << 1.0$
- $\sigma^2 \equiv 1.0$
- σ_k^2, GM $= 1.0$



DP-1 and DP-2 = Dirichlet process priors; SBR = Smoothing by Roughening

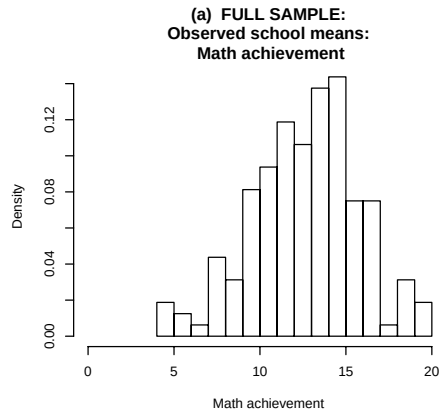
Math Achievement[27] (FYI)

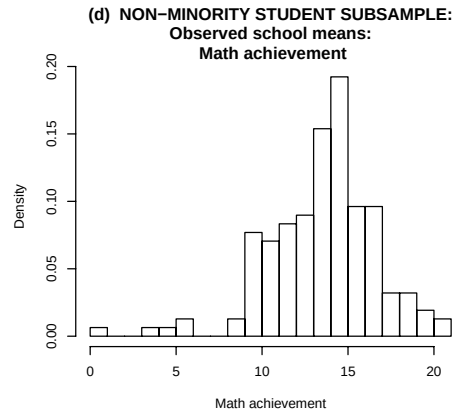
- Data are math achievement scores for 7185 students in 160 schools from the dataset 'MathAchieve' in the `R` package `nlme`
- Histograms produced by a Bayesian model to produce school-level effects using a Gaussian model for student scores conditional on the school effect and either a Gaussian or a Dirichlet Process (DP) prior for the school effects
- Histograms are for the full sample and the non-minority student sub-sample
  - 'Direct' are school-level effects w/o shrinkage (a flat prior on them)
  - 'Bayes Gaussian' are school-level effects via a Gaussian prior and 'histogrammed'
  - 'Bayes/DP' are school-level effects via a Dirichlet Process prior and 'histogrammed'

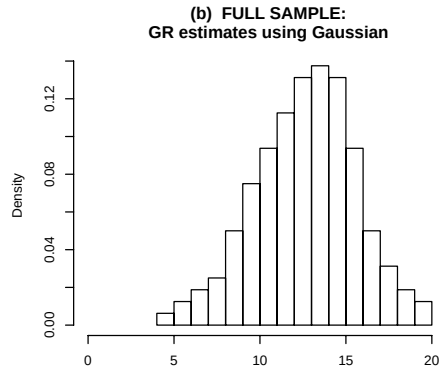
[27] Paddock SM, Ridgeway G, Lin R, Louis TA (2006). Flexible Prior Distributions for Triple-Goal Estimates in Two-Stage Hierarchical Models. Computational Statistics and Data Analysis, 50: 3243-3262.

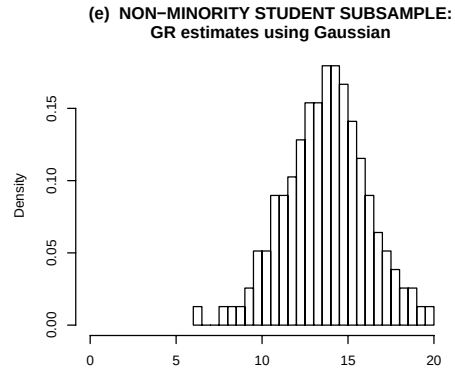
Histogram estimates for math achievement (FYI)

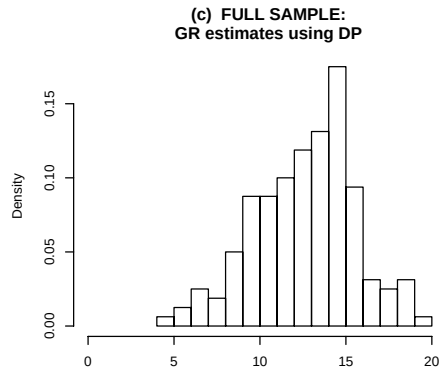
- Direct
- Bayes/Gaussian
- Bayes/DP

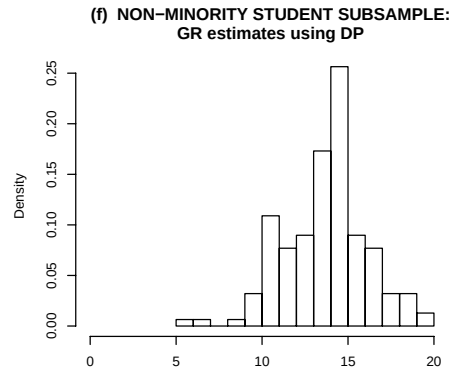












The triple-goal, GR Estimates[28] (FYI)

- To produce $\hat{\theta}$ with a histogram that is a good estimate of the empirical distribution of the underlying θ-values, use

$$\begin{aligned} \theta_k^{gr} &= U_{\hat{R}_k} \\ \hat{U}_j &= \bar{G}_K^{-1}\left(\frac{2j-1}{2K} \mid \mathbf{Y}\right) \end{aligned}$$

  with $\hat{R}_k$ the optimal, integer ranks $\{\hat{R}_k, \tilde{R}_k(\gamma), \ldots\}$

- The θ^{gr} are triple-goal:
  - Ranking them produces optimal ranks
  - Their histogram is optimal
  - SEL for estimating individual θs is higher than for the posterior means, but the penalty is small and GR estimates retain much of the Bayes advantage over MLEs
- They allow one set of estimates to be released and used for all three goals
- They support subgroup identification

[28] Shen W, Louis TA (1998). Triple-goal estimates in two-stage hierarchical models. *J. Royal Statistical Society, Ser. B*, 60: 455-471.

Bayes & Multiplicity

A basic case, see also (Bayes) FDR, etc.

- The prior to posterior mapping doesn't 'know' about multiple comparisons
- With additive $(1 + 1 = 2)$, component-specific losses, each comparison is optimized separately with no accounting for the number of comparisons
- However, empirical Bayes or Bayes empirical Bayes links the components because the posterior 'borrows information"
- The consequent shrinkage towards the overall mean controls multiplicity
- The Bayesian structure 'calms' the multiplicity

Shrinkage controls multiplicity: The K-ratio

RE ANOVA

$$
\begin{aligned}
&\bullet\ \theta_1, \ldots, \theta_K && iid && N(\mu, \tau^2) \\
&\bullet\ [Y_{ik} \mid \theta_k] && ind && N(\theta_k, \sigma^2) \\
&\bullet\ [\theta_k \mid Y_{.k}] && \sim && N\left(\mu + (1-B)(Y_{.k} - \mu), (1-B)\frac{\sigma^2}{n}\right) \\
&\qquad\qquad F && = && 1/\hat{B}
\end{aligned}
$$

Compare columns 1 and 2 (can compare all columns):

$$Z_{1vs2}^{Bayes} = Z_{1vs2}^{freq}\left\{\frac{(F-1)^+}{F}\right\}^{\frac{1}{2}} = \left(\frac{\sqrt{n}(Y_{.1}-Y_{.2})}{\hat{\sigma}\sqrt{2}}\right)\left\{\frac{(F-1)^+}{F}\right\}^{\frac{1}{2}}$$

- The Z-score is damped by the value of the F-statistic; larger F damps less
- If $H_0: \theta_1 = \theta_2 = \ldots = \theta_K$ is true, the overall type I error is controlled because F will be close to 1.0

One-sided, type I error using the posterior distribution

True $B = 1 \quad (\tau^2 = 0)$

K	single test	$K-1$ indep. contrasts	pr$(\hat{B} = 1)$ $\times 100$
10	0.00116	0.01038	56.3
20	0.00050	0.00943	54.3
30	0.00028	0.00796	53.5
50	0.00012	0.00562	52.7
100	0.00003	0.00267	51.9
500	0.00000	0.00009	50.8
1000	0.00000	0.00001	50.6

Comments

- The magnitude of F continuously adjusts the test statistic
- For large K, under the global null hypothesis ($H_0: \theta_1 = \theta_2 = \ldots = \theta_K$, equivalently $\tau^2 = 0$),
  $pr(F \leq 0) \approx 0.5$ and so pr(all $Z_{ij} = 0$) ≈ 0.5
- The family-wise rejection rate is much smaller than 0.5, thus controlling the type I error
- 'Scoping' is important because the type and number of components in the analysis determines the value of $\hat{\mu}$ and $\hat{B}$
- If collective penalties are needed, use a multiplicity-explicit, non-additive loss function (e.g., $1 + 1 = 2.5$)

Non-Additive Loss (FYI)

- Unit penalties for single errors + an extra penalty for making two errors

$$
\begin{aligned}
\text{Parameters:} \quad & \theta_1, \theta_2 \in \{0, 1\} \\
\text{Probabilities:} \quad & \pi_{ij} = pr[\theta_1 = i, \theta_2 = j] \\
\text{Decisions:} \quad & a_1, a_2 \in \{0, 1\} \\
\text{Loss}(a, \boldsymbol{\theta}) : \quad & a_1(1-\theta_1) + (1-a_1)\theta_1 \\
+ \quad & a_2(1-\theta_2) + (1-a_2)\theta_2 \\
+ \quad & \boxed{\gamma(1-\theta_1)(1-\theta_2)a_1 a_2}
\end{aligned}
$$

Optimal Decision Rule (FYI)

Decision Rule

$$\pi_{1+} \leq .5, \pi_{+1} \leq .5 \quad a_1 = 0, a_2 = 0$$

$$\pi_{1+} \leq .5, \pi_{+1} > .5 \quad a_1 = 0, a_2 = 1$$

$$\pi_{1+} > .5, \pi_{+1} \leq .5 \quad a_1 = 1, a_2 = 0$$

$$\pi_{1+} > \pi_{+1} > .5 \quad a_1 = 1$$

$$a_2 = \begin{cases} 0, & \text{IF } (2\pi_{+1} - 1) < \gamma\pi_{00} \\ 1, & \text{IF } (2\pi_{+1} - 1) \geq \gamma\pi_{00} \end{cases}$$

Bayes in the regulatory context

Visit, FDA Impact Story: Using Bayesian Hierarchical Models

- Frequentist properties can be assessed, but timing is key
- At the outset of an investigation, if there is little prior information, the frequentist properties of the full investigation are relevant
- However, a second, well-controlled study with the (possibly discounted) results of the first study used as an informative prior that gives relatively high probability to a non-null region, will produce an inflated type I error
  - If you trust the prior, compute the Type I error, but don't pay much attention to it
- Timing of is also important in a frequentist analysis; part-way through a study the *conditional* type I error will not be 0.05.

Need a trusted process

- A trusted and reproducible protocol/process is needed for developing prior distributions, making decisions, etc.
- The particulars will differ from the frequentist criteria currently used by the FDA, but the goals are the same:
  - Valid design, conduct and analysis
  - A trusted, transparent process for evaluating sponsor-produced designs and results

The Bayesian Approach to Design and Analysis

- Potential benefits are substantial, but effectiveness requires expertise and care
- It is very effective in generating procedures that can be evaluated for both Bayes and frequentist properties
- Analyses are guided by the laws of probability, which is especially valuable when addressing complex, non-linear models and utilities
- All (identified) uncertainties are transported to the posterior distribution
- Induces probabilistic relations amongst data sources, and combining evidence occupies the middle ground between 'complete pooling' and 'no relation,'
  - Bayes & Frequentist

    H_{00}: Unit-specific values are equal

    H_A: The unit-specific values are unrelated
  - Uniquely Bayes (the key to combining evidence)

    H_0: Unit-specific values come from the same probability distribution; they are different, but are 'siblings'

The Bayesian Approach to Design and Analysis

- Potential benefits are substantial, but effectiveness requires expertise and care
- It is very effective in generating procedures that can be evaluated for both Bayes and frequentist properties
- Analyses are guided by the laws of probability, which is especially valuable when addressing complex, non-linear models and utilities
- All (identified) uncertainties are transported to the posterior distribution
- Induces probabilistic relations amongst data sources, and combining evidence occupies the middle ground between 'complete pooling' and 'no relation,'
  - Bayes & Frequentist

    H_{00}: Unit-specific values are equal
    H_A: The unit-specific values are unrelated
  - Uniquely Bayes (the key to combining evidence)

    H_0: Unit-specific values come from the same probability distribution; they are different, but are 'siblings'
- Warning: The approach will not rescue poor data or a poor data model
  - e.g., a model that fails to address selection effects, confounding, ...

The Bayesian Approach to Design and Analysis

- Potential benefits are substantial, but effectiveness requires expertise and care
- It is very effective in generating procedures that can be evaluated for both Bayes and frequentist properties
- Analyses are guided by the laws of probability, which is especially valuable when addressing complex, non-linear models and utilities
- All (identified) uncertainties are transported to the posterior distribution
- Induces probabilistic relations amongst data sources, and combining evidence occupies the middle ground between 'complete pooling' and 'no relation,'
  - Bayes & Frequentist

    H_{00}: Unit-specific values are equal
    H_A: The unit-specific values are unrelated
  - Uniquely Bayes (the key to combining evidence)

    H_0: Unit-specific values come from the same probability distribution; they are different, but are 'siblings'
- Warning: The approach will not rescue poor data or a poor data model
  - e.g., a model that fails to address selection effects, confounding, ...
- Closing mantra: There are no free lunches in statistics, but there are a large number of reduced-price meals, many based on Bayesian recipes

Supplementary Slides

Bayes and Subgroups[29,30,31]

- The SOLVD studies of left ventricular dysfunction examined the impact of the drug Enalapril in a group of patients with congestive heart failure and low ejection fraction
- In total, 2569 patients were enrolled in the treatment trial with 1285 patients being assigned to the treatment arm and 1284 patients being assigned to the placebo arm
- At the scheduled end of the study, 510 patients had died in the placebo group while 452 had died in the Enalapril group
- We created 12 subgroups
  $\{\text{gender} \times (\text{age} \leq 65 \text{ vs} > 65) \times (\text{ejfr } 6\text{-}22,\ 23\text{-}29,\ 30\text{-}35)\}$

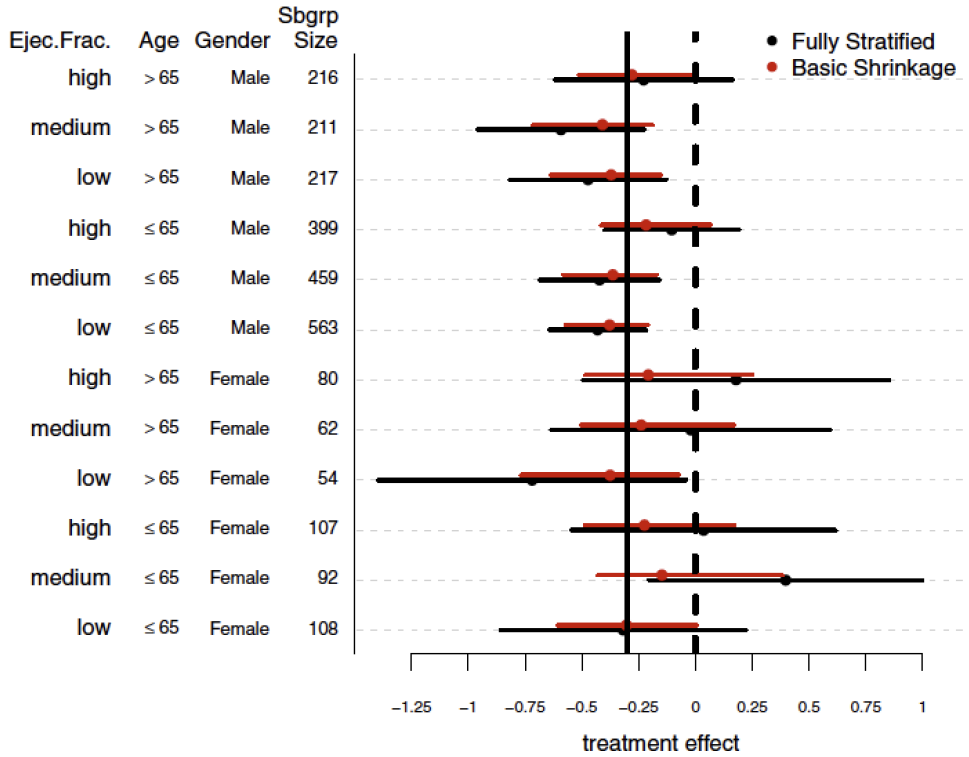
[29] Henderson NC, Louis TA, Wang C, Varadhan R (2016). Bayesian Analysis of Heterogeneous Treatment Effects for Patient-Centered Outcomes Research. *Health Services and Outcomes Research Methodology*, 16: 213-233. doi.10.1007/s10742-016-0159-3.

[30] Wang C, Louis TA, Henderson N, Weiss CO, Varadhan R (2018). BEANZ: An R Package for Bayesian Analysis of Heterogeneous Treatment Effect with a Graphical User Interface. *Journal of Statistical Software*, 85: doi: 10.18637/jss.v085.i07.

[31] BEANZ at: https://www.research-it.onc.jhmi.edu/dbb/custom/A6/ and at http://cran.r-project.org.

Basic subgroup results for the log(hazard ratio)

- Substantial shrinkage for the basic model with the SD for the between-subgroup RE, $\omega \sim$ Half-N(100)
- Little 'enthusiasm' for subgroup effects



Black: Frequentist estimates and CIs

Red: Standard Bayes estimates and credible intervals

Solid vertical: Overall treatment effect

Sensitivity analysis wrt the between-subgroup SD (ω)

- $b|Z| \sim$ Half-N(b^2)

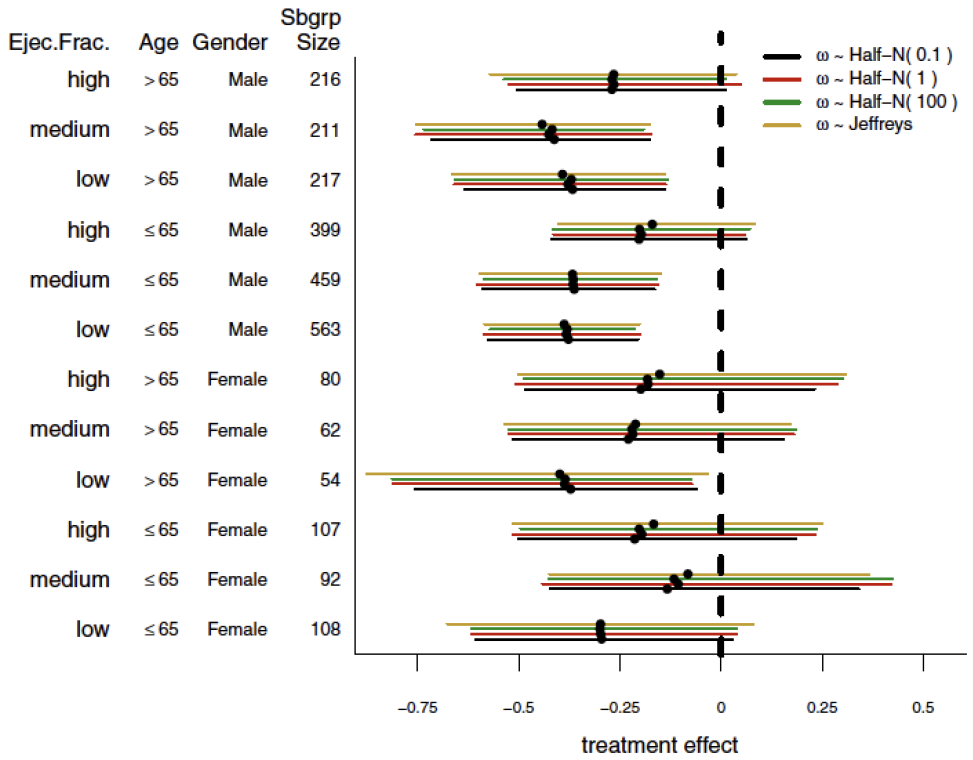


Fig. 3 Basic shrinkage model—sensitivity to choice of prior. SOLVD data. Posterior means and associated credible intervals for the following choices of the prior for ω: $\omega \sim$ Half $-$ Normal(0.1), $\omega \sim$ Half $-$ Normal(1), $\omega \sim$ Half $-$ Normal(100), and $\omega \sim$ Jeffreys. The approximate Jeffreys prior for ω^2 employed here is $p(\omega^2) \propto \omega^{-2}$ for $\omega^2 \geq 0.005$ and $p(\omega^2) = 200$ otherwise

Stratified, basic Bayes, extended Dixon-Simon[32]

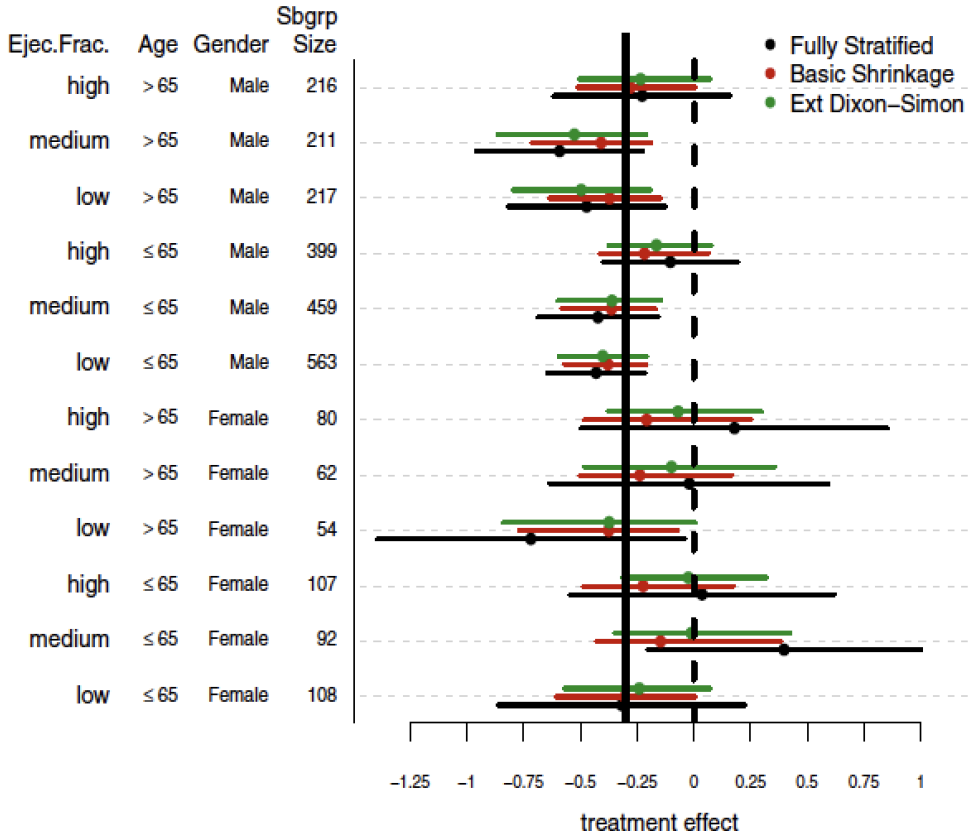


Fig. 4 Extended Dixon–Simon model. SOLVD data. Posterior means and credible intervals for each of the 12 subgroups defined by the variables: gender, age, and baseline ejection fraction. Point estimates and uncertainty intervals from the basic shrinkage model and from the fully stratified frequentist analysis are also shown

[32] Jones, H, Ohlssen, D, Neuenschwander, B, Racine, A, Branson, M (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 8: 129–143.

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

Design and Analysis for Cluster Randomized Studies

Setting

- Compare two weight loss interventions
- Randomize clinics in pairs, one to A and one to B
- Compute clinic-pair-specific comparisons combine over pairs
- How to design and how to analyze, especially with a small number of clinics?

The equal sample size, unpaired case

- There are K clusters
- Within-cluster sample sizes are $n_k \equiv n$
- $V_{ind} = \mathrm{V}(\text{treatment comparison})$, when assuming independence
- Adjust this by the between-clinic variance component, equivalently by ρ, the Intra-class Correlation Coefficient (ICC):

$$
\begin{aligned}
V_{icc} &= V_{ind} \times [1 + \rho\,(n-1)] = V_{ind} \times [\text{design effect}] \\
\rho &= \frac{\tau^2}{\sigma^2 + \tau^2} \quad \text{(the ICC)} \\
\tau^2 &= \text{the between-clinic variance} \\
\sigma^2 &= \text{single-observation variance}
\end{aligned}
$$

Design and Analysis Considerations

- In the paired-clinic case, to compute $V_{icc} = V(\text{treatment comparison})$, need to account for the following variances:
- Individual measurement (σ^2)
  - The trial will provide sufficient information
- Between-clusters: within (τ_w^2) and between (τ_b^2) cluster pairs with $(\tau^2 = \tau_w^2 + \tau_b^2)$

The need for an informative prior

- With a small number of clusters, the trial will provide little information on τ^2 and even less information on $\gamma = \tau_b^2/(\tau_w^2 + \tau_b^2)$
- Without informative priors, an 'honest' computation of posterior uncertainty (one that integrates over uncertainty in τ^2 and γ) will be so large as to make results essentially useless
- Therefore, either don't do the study or use informative priors to bring in outside information
- Fortunately, other weight loss studies provide credible and informative prior information on τ^2, but not so for γ
    - For γ, we need to rely primarily on expert opinion and sensitivity analysis

A Bayesian Model

- Use an informative, data-based prior for τ^2 and a small-mean, small-variance prior for γ

$$\begin{aligned} \tau^2 &\sim \text{IG: } = \tau^2_{50} \text{ with } \tau^2_{95} = 2 \times \tau^2_{50} \\ [\gamma \mid \epsilon, M] &\sim \text{Beta}(\epsilon, M) \\ E(\gamma) &= \epsilon, V(\gamma) = \epsilon(1-\epsilon)/M \end{aligned}$$

- Take the 'best estimates' of (σ^2, ρ) from other cluster-randomized studies of weight change and obtain $\sigma^2 \approx (0.34)^2$, likely $\hat{\rho}$: (0.006, 0.010, 0.050)
- $\Rightarrow 10^4 \times \tau^2 = (7.0, 11.7, 60.8)$, $\boldsymbol{\tau^2_{50} = 11.7 \times 10^{-4}, \tau^2_{95} = 23.4 \times 10^{-4}}$
- Use $\epsilon \approx 0.10$ and a relatively large $M = 15$
  - The 90^{th} percentile is approximately 0.20
  - Conservative in that there is little gain from pairing

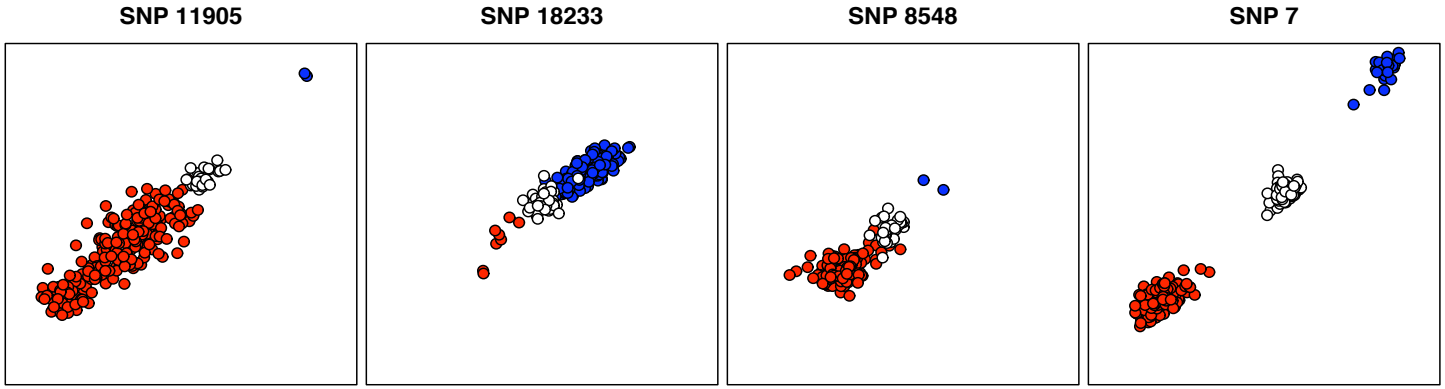
Measurement Uncertainty & Full Probability Modeling

Trend Tests that Accommodate Genotyping Errors[33]

- A standard GWAS evaluates SNP-specific association with a phenotype (disease status), for example by ranking SNP-specific Z-scores
- However, due to technical and biological factors, genotype 'calls" (AA, AB, BB) can be uncertain
- These errors can produce invalid or inefficient inferences
- Most calling algorithms produce a 'best' call along with a call-specific uncertainty measure
- Many recommend not using the call if uncertainty is too large

[33] Louis TA, Carvalho BS, Fallin MD, Irizarry RA, Li Q, Ruczinski I (2011). Association Tests that Accommodate Genotyping Errors. pp. 393–420 in, *Bayesian Statistics 9.* (JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, M West, Eds.), Oxford University Press, Oxford UK.

Sense/Antisense Information



- Genotype calls ranging from 'difficult' to 'easy'
- Most SNPs are 'easy' (like #7), some are uncertain, some are essentially hopeless

Quantify and Retain Genotype Uncertainty

- A test statistic should take genotype uncertainty into account via a vector of genotype posterior probabilities
  - Deterministic calls have a 1 in a single position
- Efficiency is measured by the correlation between the true genotype and genotype probabilities

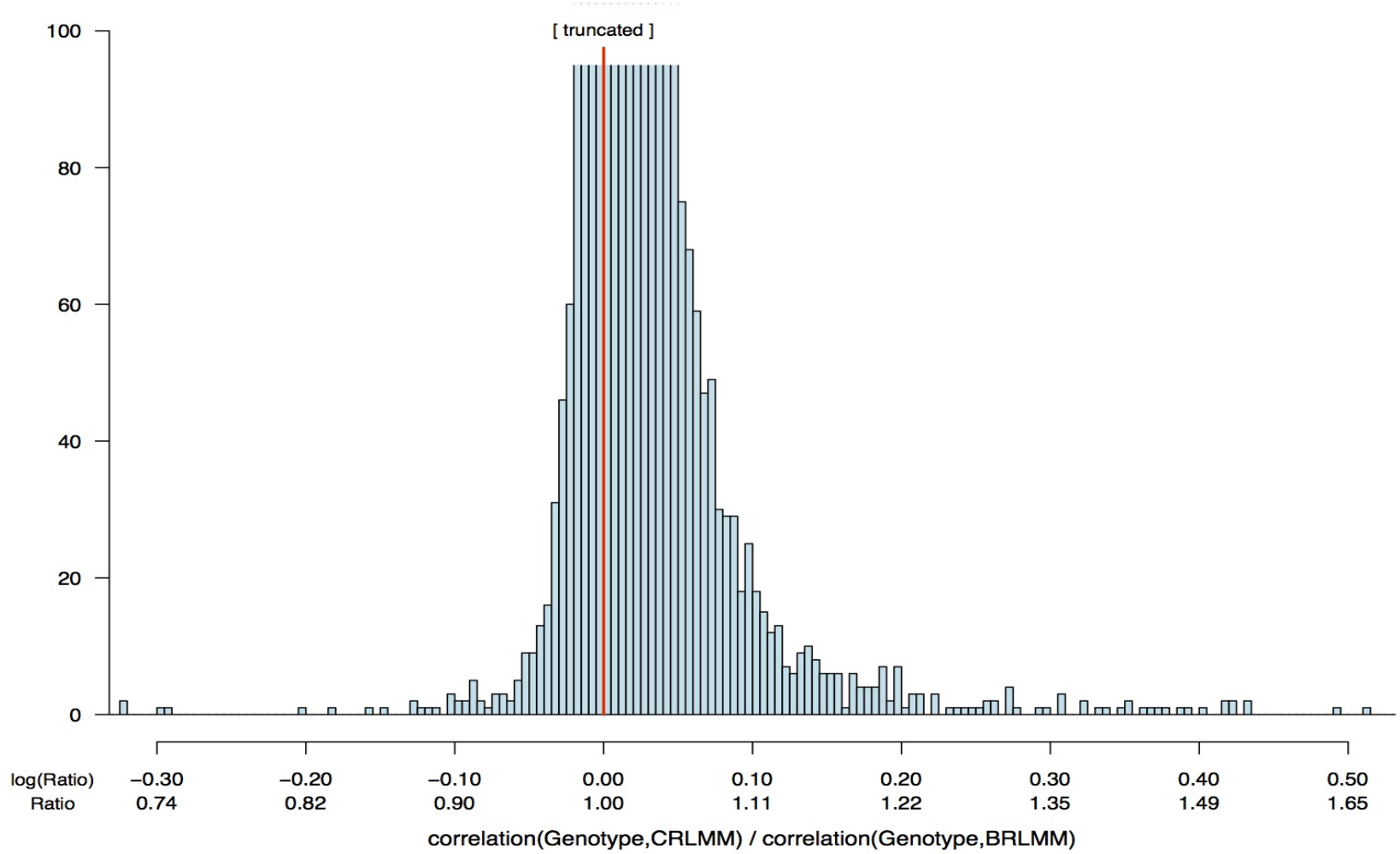
HapMap Gold Standard

- Compute correlations between the gold standard and standard 'best' calls, probability vector calls and 'best' as the mode of the probability vector
- Posterior probabilities from Carvalho et al. (2009)[34]

[34] Carvalho B, Louis TA, Irizarry RA (2009). Quantifying Uncertainty in Genotype Calls. *Bioinformatics*, 26: 242-249.

Correlation(Bayes, Gold)/Correlation(Standard, Gold)

Lesson: Build an uncertainty model and use Bayesian processing



Correlation(Bayes, Gold)/Correlation(ModalBayes, Gold)

Lesson: Percolate uncertainty all the way through