Case Studies in Calibrating Hierarchical Model Priors

Kert Viele
Corrine Elliott
Joe Marion
DIA Bayesian KOL
February 22, 2019

Hierarchical Models need no introduction

- Workhorse of Bayesian inference
  - powerful way to combine information
  - synthesizes within and across group variation
  - better than treating groups separately
- Focus here on basic situation
  - $\theta_1, \ldots, \theta_G \sim N(\eta, \tau)$ with priors on η and τ
  - θ_g might determine mean, rate, or HR of group
  - What prior should we pick for η and τ?

Noninformative priors

- With larger effective sample sizes….
  - large nested sociology models
  - meta-analyses
- noninformative priors are useful
  - Gelman (2006) and subsequent literature
  - notes importance of prior mass around 0 for across group variability.

Informative priors

- In clinical trials we often have a few groups…
  - basket trials (3-6 groups) often with small sample sizes within each
  - hierarchical borrowing of information from a small number of historical trials
- In these situations the prior matters
  - usually want to choose the prior to reflect subjective information and/or
  - choose the prior to obtain desirable operating characteristics

Two case studies

- We will show two case studies where informative priors are desired

- Discuss possible ways of calibrating prior to achieve desirable performance

Case Study 1 - ADEPT

(David Brody, Lindsay Oberman, Thaddeus Haight)

- Investigate strategies for magnetic brain stimulation to treat depression
  - MADRS endpoint
  - 16 arms in 4x2x2 structure
    - four methods for targeting, lateral/bilateral, two protocols
- We expect additive structure, but want robustness

	Truth additive	Truth not additive
Additive Model	Achieve efficiency (low posterior variance) from additive model	BAD biases from lack of fit
”Free” Model	Decent fit, but lose efficiency through extra parameters	Appropriate fit, although with higher variance

Can we bridge the gap here?

Possible options include model averaging, etc.

Operating Characteristics for additive and separate analyses

	Truth is additive	Truth is no additive
Additive Model	Posterior SD per arm 0.06 Estimates are unbiased	Posterior SD per arm can be extremely large from lack of fit Estimates can have large biases
"Free" Model No constraints	Posterior SD 0.10 Estimates are unbiased	Posterior SD 0.10 Estimates are unbiased

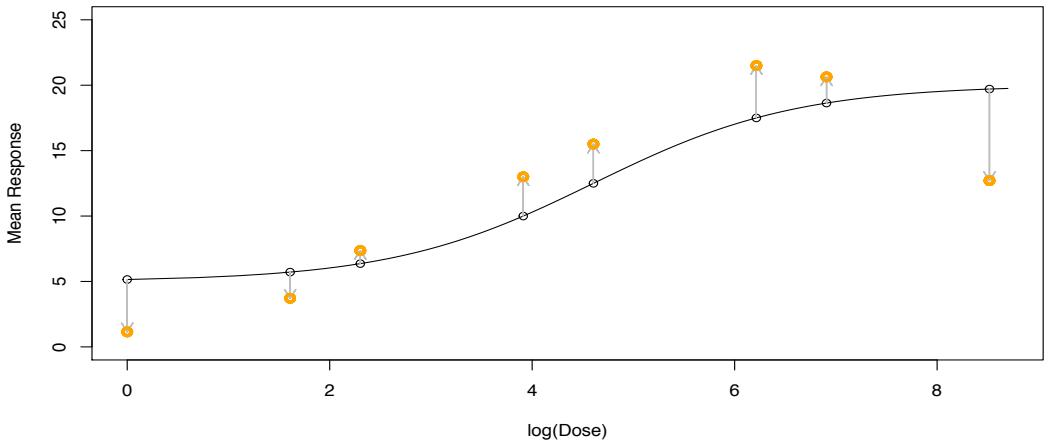
We want the posterior SD to be close to 0.06 when the truth is additive while avoiding bad biases when the truth is not additive

Deviation based models

- The additive model would have the form
  - Y_{ai} = Targeting + Laterality + Protocol + error
    = Model(a) + error
    - Y_{ai} is i^{th} observation from arm a
- Allow deviations from the additive model
  - Y_{ai} = Model(a) + ζ_a + error
  - $\zeta_1, \ldots, \zeta_A \sim N(0, \tau)$ with $\sum \zeta_a = 0$
  - $\tau = 0$ results in the additive model
  - τ large results in separate estimates per arm

Dose Ranging Deviation Models

- Doses 1,…,D each with parameter θ_d
  - $\theta_d = \text{Model(d)}$ for example Model(d)=Emax(d)
  - Generalize to $\theta_d = \text{Model(d)} + \zeta_d$
  - $\zeta_1, \ldots, \zeta_D \sim N(0, \tau)$ with $\sum \zeta_d = 0$

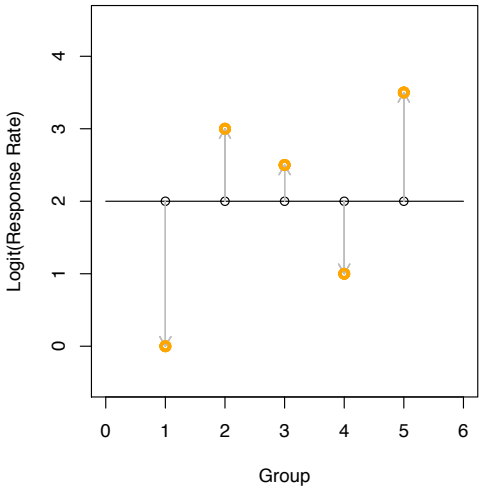


Black curve is base emax model
Grey arrows are $\zeta_1, \ldots, \zeta_D$
Orange points are $\theta_1, \ldots, \theta_D$

Model allows slight deviations from emax, U-shapes, etc.

Basket trials are Deviation Models

- Standard dichotomous basket trial, G groups
  - $Y_{gi} \sim Bern(p_g)$
  - $\theta_g = logit(p_g)$
  - $\theta_1, \ldots, \theta_G \sim N(\eta, \tau)$ with priors on η and τ
- Can be reparameterized
  - $\theta_g = \eta + \zeta_g$
  - $\zeta_1, \ldots, \zeta_G \sim N(0, \tau)$ with $\sum \zeta_g = 0$
  - priors on η and τ
  - same induced prior on $p_1, \ldots, p_G$
  - The "Model" is the pooled model

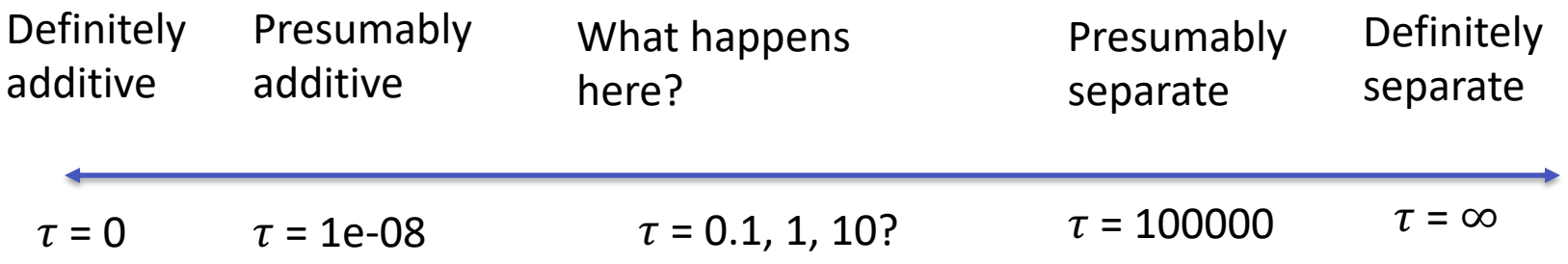


Prior on τ?

- We choose $\tau^2 \sim \text{IGamma}(\alpha, \beta)$
  - in some situations, IGamma is problematic
  - here behavior around $\tau = 0$ may be a feature
- Choice of α and β?
  - want to achieve desirable operating characteristics
    - prefer near additive data to result in small zetas
    - prefer non additive data to smoothly fit separately
  - this reflects our subjective prior belief
    - we think its additive, but could be convinced otherwise

Prior on τ?

- We know $\tau = 0$ is an additive model
- We know τ large is a separate model
- Where is the dividing line?
  - What about $\tau = 0.1$? 0.01, 10.0?
  - Need a prior which has meaningful probability of being either additive or separate



We tried multiple fixed τ

- If you write code for $\tau^2 \sim \text{IGamma}(\alpha, \beta)$
  - $\tau^2 \sim \text{IGamma}(10000, 10000 * X)$ is essentially a point mass at τ^2= X
  - Considered τ^2= 0.001, 0.01, 0.1, and 1
- We simulated multiple datasets
  - Random data from a true additive model
  - Random data from a separate model, mostly null with a single effective arm (a ”nugget” scenario)
  - more not shown here

Results for Additive Dataset

- Found posterior mean for each prior and arm
- Computed mean squared error for the 16 arms and average posterior standard deviation
  - these were consistent across arms

Prior	MSE	Posterior SD	
Additive $\tau^2 = 0$	0.004	0.061	
$\tau^2 = 0.001$	0.004	0.066	Close to additive
$\tau^2 = 0.01$	0.006	0.083	
$\tau^2 = 0.1$	0.010	0.097	
$\tau^2 = 1$	0.011	0.100	Close to separate
Separate τ^2 large	0.011	0.100	

Results for Separate Dataset

- Found posterior mean for each prior and arm
  - MSE and posterior SD for arm 16 (the nugget) computed separately from arms 1-15 (nulls)

Prior	MSE 1-15	Post SD 1-15	MSE 16	Post SD 16	
Additive $\tau^2 = 0$	34.8	0.570	859.1	0.567	
$\tau^2 = 0.001$	34.7	0.570	857.1	0.567	Close to additive
$\tau^2 = 0.01$	34.0	0.569	839.2	0.566	
$\tau^2 = 0.1$	26.2	0.549	654.3	0.566	
$\tau^2 = 1$	0.017	0.100	0.062	0.100	Close to separate
Separate τ^2 large	0.010	0.100	0.001	0.099	

A first attempt…

- We have a good neighborhood
  - Values around $\tau^2 = 0.001$ emulate additive
    - maybe could go a little lower
  - Values around $\tau^2 = 1$ emulate separate
    - maybe could go a little higher
- Let's try a prior anchored in this area
  - Started with IGamma(0.2,0.0002)
    - 20th, 80th percentiles are 0.00075, 0.95762

A first mediocre attempt…

- The additive data fit disappointed….
  - not very close to additive model

Prior	MSE	Posterior SD
Additive $\tau^2 = 0$	0.004	0.061
$\tau^2 = 0.001$	0.004	0.066
$\tau^2 = 0.01$	0.006	0.083
$\tau^2 = 0.1$	0.010	0.097
$\tau^2 = 1$	0.011	0.100
Separate τ^2 large	0.011	0.100
IG(0.2,0.0002)	0.006	0.087

A first mediocre attempt…

- The separate fit was better

Prior	MSE 1-15	Post SD 1-15	MSE 16	Post SD 16
Additive $\tau^2 = 0$	34.8	0.570	859.1	0.567
$\tau^2 = 0.001$	34.7	0.570	857.1	0.567
$\tau^2 = 0.01$	34.0	0.569	839.2	0.566
$\tau^2 = 0.1$	26.2	0.549	654.3	0.566
$\tau^2 = 1$	0.017	0.100	0.062	0.100
Separate τ^2 large	0.010	0.100	0.001	0.099
IG(0.2,0.0002)	0.006	0.103	0.014	0.102

What went wrong?

- Easier to see lack of fit than lack of parsimony
  - Non-additive data fit by additive model produces huge residuals, obviously wrong
  - Additive data fit by a separate model is fit well, just with larger posterior standard deviation
- So let's hedge further toward additive
  - consistent with our subjective belief of additivity unless the data strongly suggest otherwise

Second attempt….

- We had 20th and 80th percentile at 0.001 and 1
- Switch to 50th and 90th percentiles
  - τ^2 ~ IGamma(0.23, 0.000035)
  - 50th percentile 0.0010
  - 90th percentile 1.1706
- We also generated new datasets…

Second attempt works better
Additive Data Results

Prior	MSE	Posterior SD
Additive $\tau^2 = 0$	0.008	0.061
$\tau^2 = 0.001$	0.008	0.066
$\tau^2 = 0.01$	0.010	0.083
$\tau^2 = 0.1$	0.020	0.097
$\tau^2 = 1$	0.021	0.100
Separate τ^2 large	0.021	0.100
IG(0.23,0.000035)	0.008	0.067

We might consider further fine tuning toward additivity, but this works well.

Second attempt works better
Separate Data Results

Prior	MSE 1-15	Post SD 1-15	MSE 16	Post SD 16
Additive $\tau^2 = 0$	34.2	0.571	867.3	0.567
$\tau^2 = 0.001$	34.1	0.571	865.3	0.567
$\tau^2 = 0.01$	33.4	0.570	847.4	0.567
$\tau^2 = 0.1$	25.7	0.549	652.6	0.566
$\tau^2 = 1$	0.019	0.098	0.127	0.099
Separate τ^2 large	0.018	0.098	0.008	0.098
IG(0.23,0.000035)	0.018	0.098	0.009	0.098

First case study conclusions

- Need to understand how choice of τ is related to trial goals
  - here we want a progression between additive and separate fits
- Find range of τ which is sensible
- Place prior mass in that range
- May need to "hedge" based on trial goals
  - here preference toward additive model
- Note prior is not informative toward trial goals (assumption on additive/separate, not treatment effect)

Second Case Study

- Oncology example altered for anonymity
- Time to event study testing HR<1
- Four subgroups tested as secondaries
  - some variation in groups
  - one outlying group with observed HR>1
  - observed HRs are not (0.40, 0.65, 0.75, 1.20)
    - observed log HRs are not (-0.92, -0.43, -0.29, 0.18)
    - but those are ok for motivation and discussion
- FDA asked for a hierarchical model analysis

Modeling

- Y_{ija} is i^{th} observation in group j on arm a
  - $Y_{ija} \sim Exp(\lambda_j)$ for a=control
  - $Y_{ija} \sim Exp(\lambda_j \exp(\theta_j))$ for a=treatment
- $\theta_1, \ldots, \theta_4$ are the log hazard ratios
  - Standard hierarchical model
  - $\theta_1, \ldots, \theta_4 \sim N(\eta, \tau)$
  - $\eta \sim N(0, 0.5)$
  - $\tau^2 \sim \text{IGamma}(\alpha, \beta)$
  - $\lambda_1, \ldots, \lambda_4$ have independent weak priors

Issues

- This is post hoc
  - We know the data and the spread and the issue with the outlying group
  - No optimal solution
  - We will consider multiple priors
    - What is a good range?

Again consider fixed τ

- Suppose $\eta = (-0.288)$
  - close to ~~observed~~ with exp(-0.288) = 0.75
- We model $\theta_1, \ldots, \theta_4 \sim N(\eta, \tau)$
  - The highest and lowest of 4 such draws are expected to be at $\eta \pm 1.03\tau$
    - Note observed HRs tend to be spread FURTHER apart

τ	Expected Range of True log HRs from 4 groups	Exponentiating Range
0.10	(-0.39,-0.18)	(0.68,0.83)
0.25	(-0.55,-0.03)	(0.58,0.97)
0.50	(-0.80,0.23)	(0.45,1.25)
1.00	(-1.32,0.74)	(0.27,2.10)

We have found our range…

τ	Expected Range of True log HRs from 4 groups	Exponentiating Range
0.10	(-0.39,-0.18)	(0.68,0.83)
0.25	(-0.55,-0.03)	(0.58,0.97)
0.50	(-0.80,0.23)	(0.45,1.25)
1.00	(-1.32,0.74)	(0.27,2.10)

- These span a good range of τ.
  - 0.10 and 0.25 are tighter than we observed
  - 0.50 about equal (keeping in mind observed HR should be wider than true HRs)
  - 1.00 wider than observed by decent amount

Location Weight parameterization

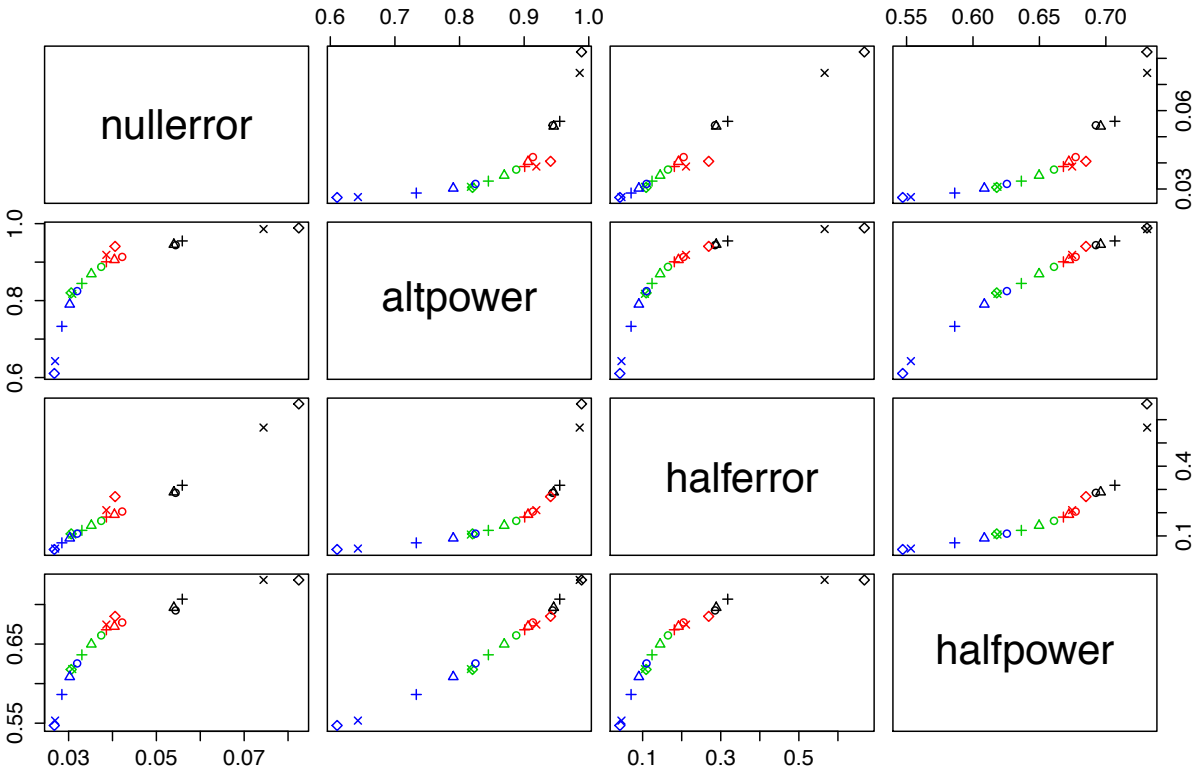
- A diversion…
- Instead of $\tau^2 \sim \text{IGamma}(\alpha, \beta)$
  - Weight = 2α effective sample size of prior
  - Location = sqrt(β/α) "location" used very loosely
    - for large effective weights, related to mean of precision
    - for small effective weights, just supplies ordering, with small locations resulting in smaller τ
  - Inverse Transform
    - α = Weight/2 β=Location^2*Weight/2

Why reparameterize?

Weight is the effective sample size of the prior

Location drives the operating characteristics

Colors indicate location, point indicates weight



"Base" prior

- We established a range of τ from (0.10,1.00)
- Use "location"=0.1, weight=0.5
  - 25th and 75th percentiles 0.10 to 0.97
- Created a grid of priors
  - included smaller weights of 0.1 and 0.3
  - larger and smaller locations of 0.05 and 0.25
  - 9 possibilities
- We ran all 9 and presented to FDA

Theme from the case studies

- Understanding what an individual τ implies is vital for finding a good range of τ values
- Depending on setting, prior may be adjusted up or down as needed
- Be careful not to make the prior inadvertently informative
  - I've seen basket trials where the prior is so weak borrowing can't happen!
  - In first case study want to allow both additive and separate models
  - In second case study don't want to guarantee strong borrowing (or weak borrowing).

THANK YOU!!!

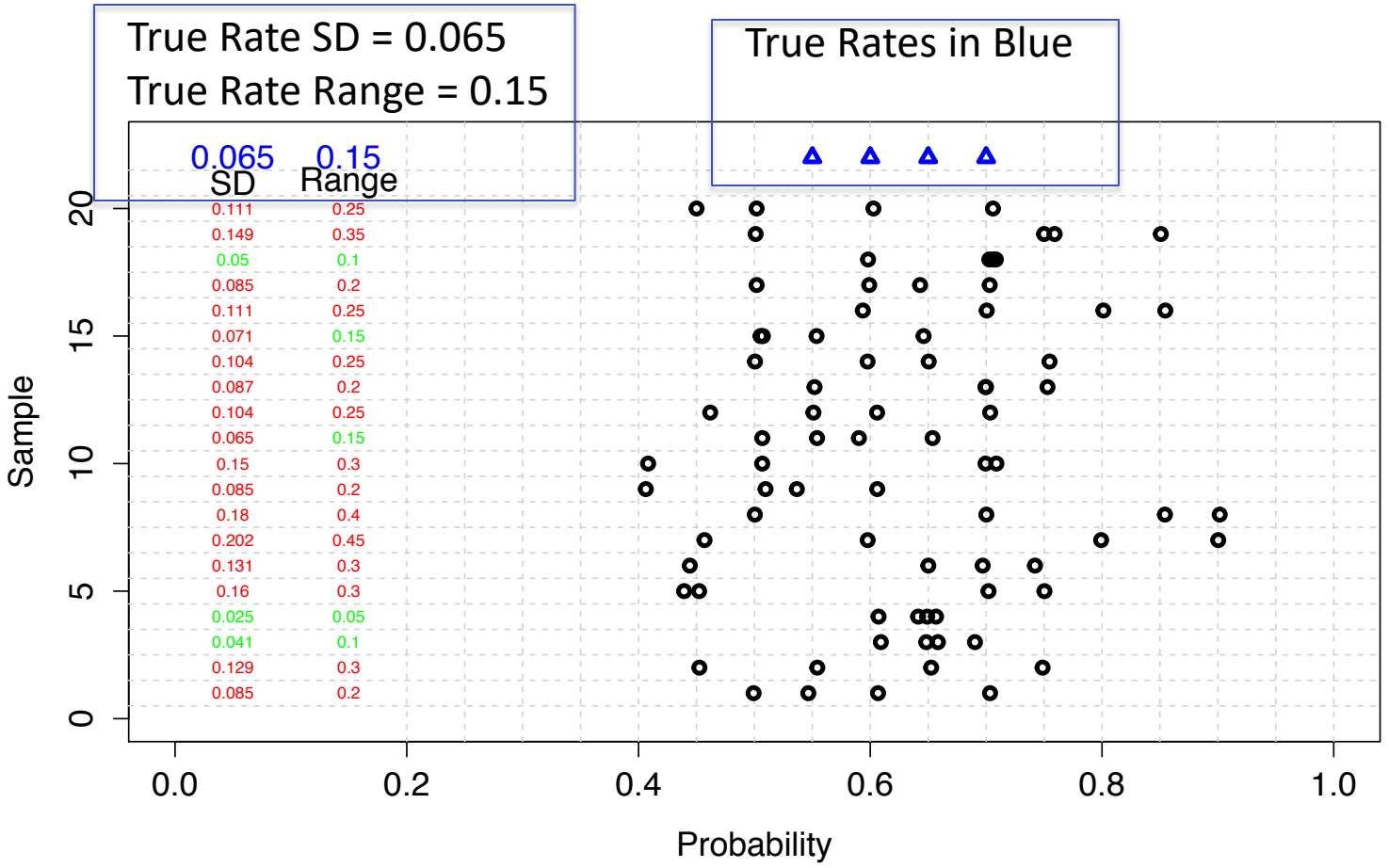
- Thanks to my collaborators on the case studies!
- Thanks to Fanni Natanegara and Mathangi Gopalakrishnan for the invitation and organization
- Thanks to colleagues at Berry Consultants for continuing discussions

Backup slides

Justification for borrowing

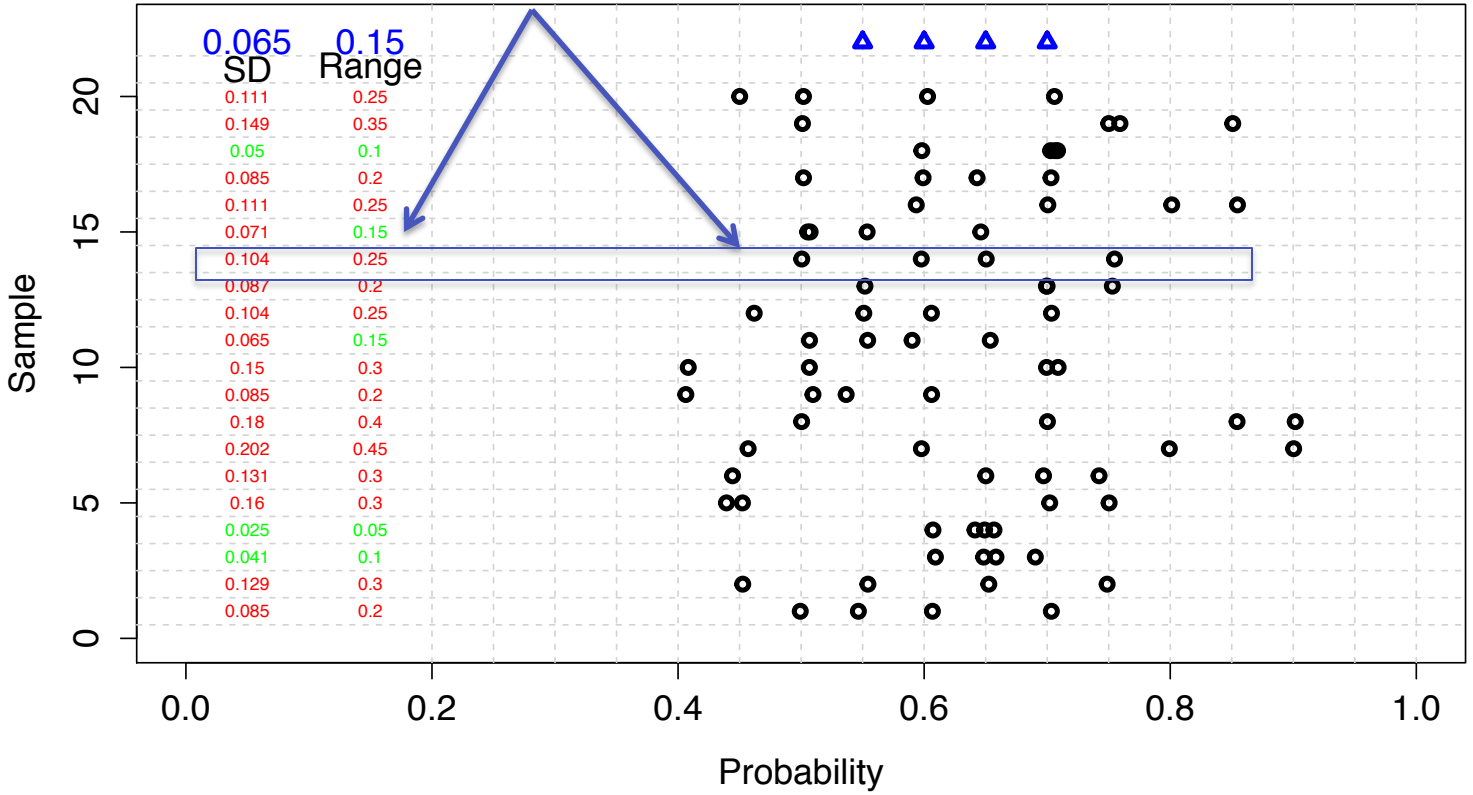
- Theoretical back to Stein (1950s)
  - this theory does not require *scientific* "commonality" among the groups.
  - Sample rates in all groups essentially consists of the true parameter + random noise
  - the random noise is common to all groups.
- Random noise increases variability
  - Consider 4 group example
  - true rates 0.55, 0.60, 0.65, 0.70 (sd=0.065, range 0.15)
  - n=20 per sample

Samples from Multiple Groups



Samples from Multiple Groups

Each row has one random set of sample proportions
Red and Green numbers show sample SD and Range
(red if bigger than truth)

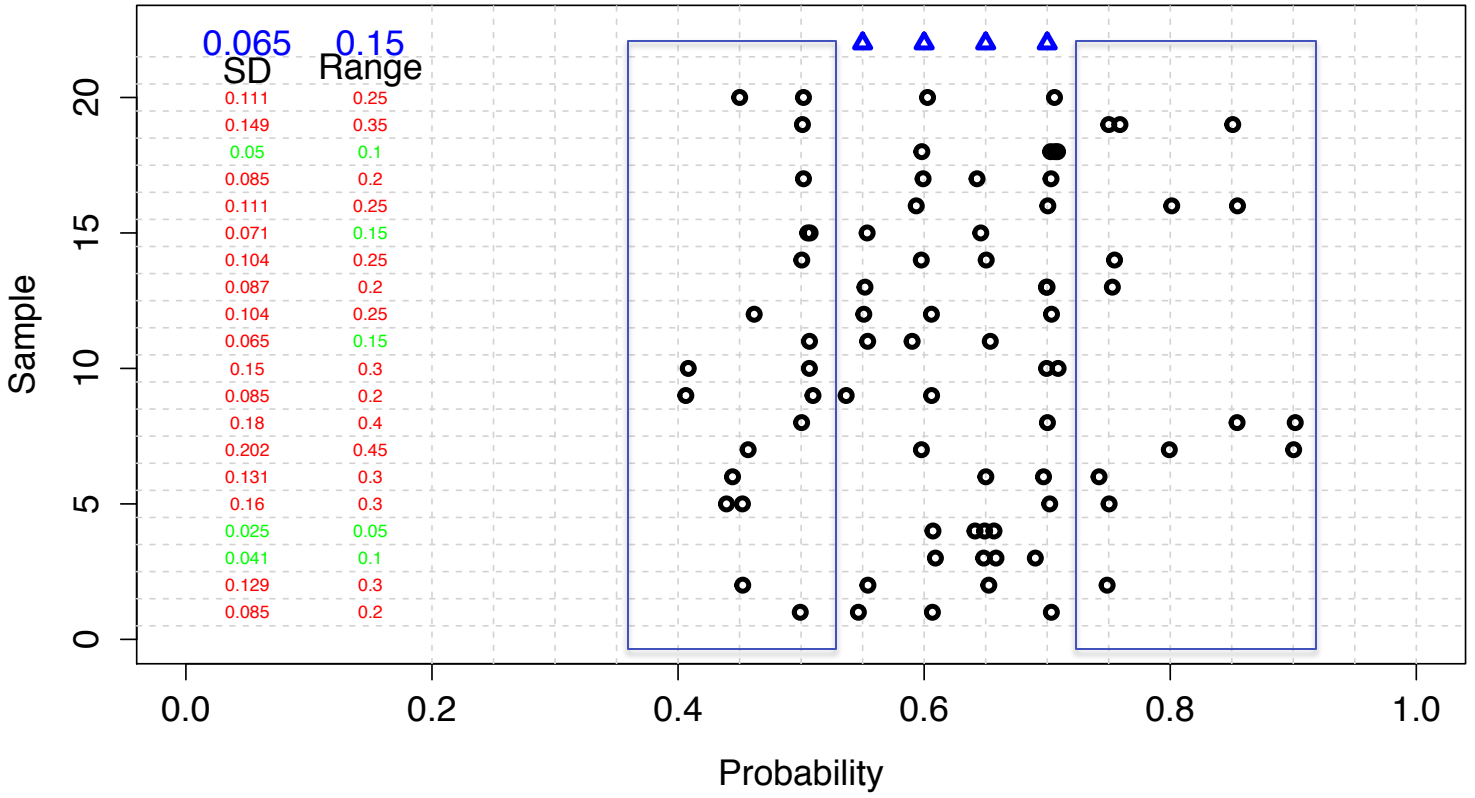


Samples from Multiple Groups

Most SDs are bigger than truth, most ranges are bigger than truth

Average SD = 0.116 > 0.065 Average Range = 0.258 > 0.150

Average minimum = 0.494 < 0.55 Average maximum = 0.752 > 0.70



Justification for borrowing

- Basic point
  - adding noise to a system always increases the variation.
  - when you see multiple groups
    - the lowest of the low is biased low (it's likely worse, but also "unlucky")
    - the highest of the high is biased high (it's likely better, but also "lucky").
  - Better estimates can be obtained by pushing estimating together ("shrinkage")
    - try to remove the "luck"