Bayesian Inference and Decision Theory

Unit 7: Hierarchical Bayesian Models

v7.2

Learning Objectives for Unit 7

- Specify hierarchical models (aka multi-level models) for comparing samples from subgroups of a population
  - Each sample consists of iid observations from a subgroup of a larger population
  - Parameters for the groups are viewed as a sample from a larger population of groups
  - Information can be shared across groups
- Explain the benefits of Bayesian hierarchical models for complex multi-parameter problems
- Apply techniques from previous units to inference in hierarchical models
- Apply techniques from previous units to evaluate structural assumptions in hierarchical models

Hierarchical (or Multi-Level) Models

- Address tension between realism and statistical power
  - Realistic models often have too many parameters for reliable frequentist inference
  - This often leads to aggregation, which can introduce bias for sub-populations
- Use structural assumptions to achieve better statistical power without sacrificing realism
  - Parameters are often related to each other by structure of the problem
  - Hierarchical models exploit the relationship
  - "Borrow strength" from data used to estimate related parameters
  - Flexibly adapt dimensionality of model to exploit the information in the data

Example: Rat Tumors
(Gelman et al, Sections 5.1-5.3)

- Studies are commonly performed on rats to evaluate effects of drugs
- The observations come from a set of 71 studies in which rats were given a substance and evaluated for the presence of tumors*
  - Y_s is the number of rats out of n_s rats in the s^{th} study which developed tumors
  - Model for Y_s:
    - Tumors occur independently with probability θ_s
    - Y_s is drawn from a Binomial(n_s, θ_s) distribution
  - The question of interest is whether the probabilities θ_s are different for different studies
- Modeling choices:
  - Pool all data and estimate a single tumor probability (not supported by the data)
  - Independent analyses for each of the 71 studies (low statistical power, especially for studies with small sample sizes)
  - Hierarchical model – gives us benefits of both approaches without drawbacks of either

*Tarone, R. E. (1982) The Use of Historical Control Information in Testing for a Trend in Proportions. *Biometrics* **38**, 215-220

Pooled Model of Tumor Probability

- Pooled model:
  - All studies have the same probability θ of each rat developing tumor
  - Uniform beta(1,1) prior distribution for θ
  - Observations: 267 out of 1739 rats have tumors
  - Posterior distribution is also a beta distribution with:

$$\alpha^* = 1 + \sum_s Y_s$$

Total number of rats with tumors +1

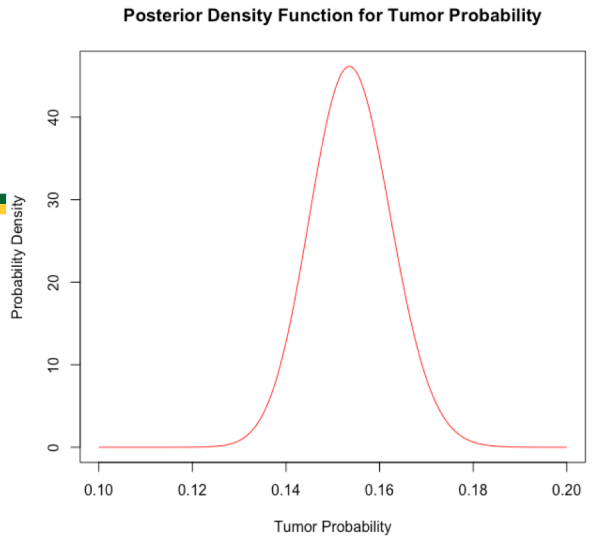
$$\beta^* = 1 + \sum_s n_s - Y_s$$

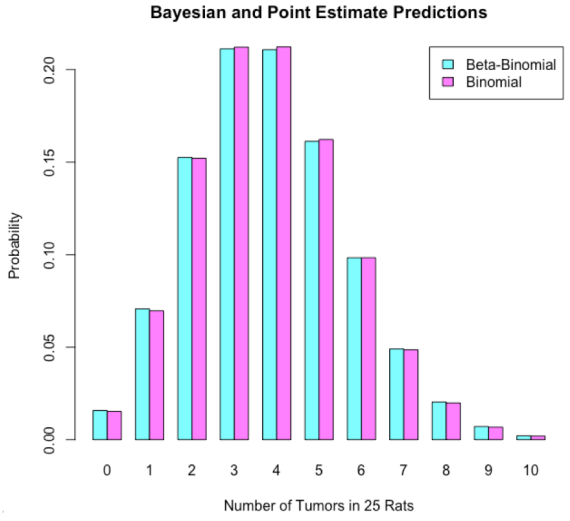
Total number of rats without tumors +1

- Posterior expected probability a rat will have a tumor

$$E\left[\theta \middle| \underline{Y}\right] = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{1 + \sum_s Y_s}{2 + \sum_s n_s} = \frac{268}{1741} = 0.154$$

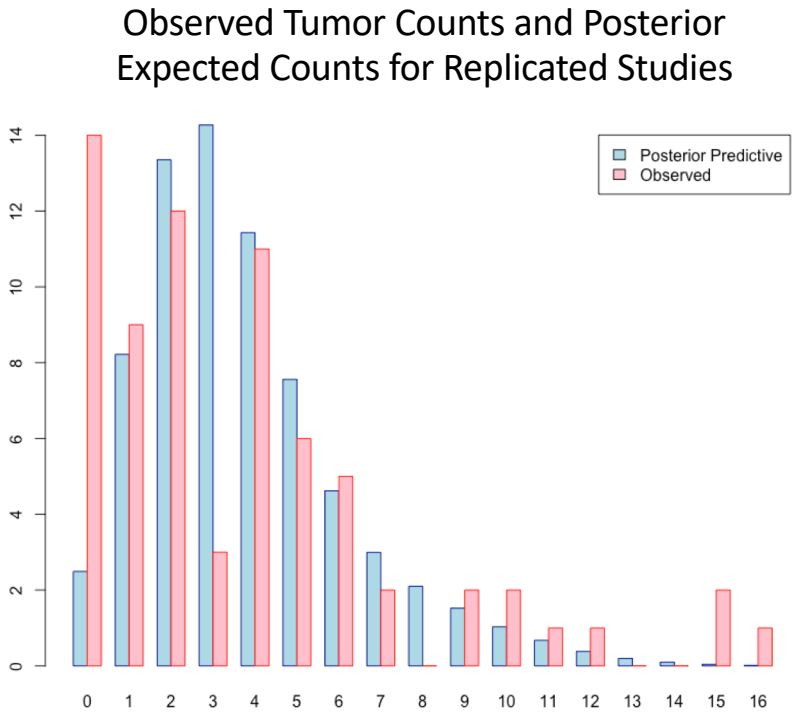
- Posterior predictive distribution for a new study of size n is beta-binomial with probability 0.154, size n, and overdispersion 1741
  - Nearly identical to binomial distribution



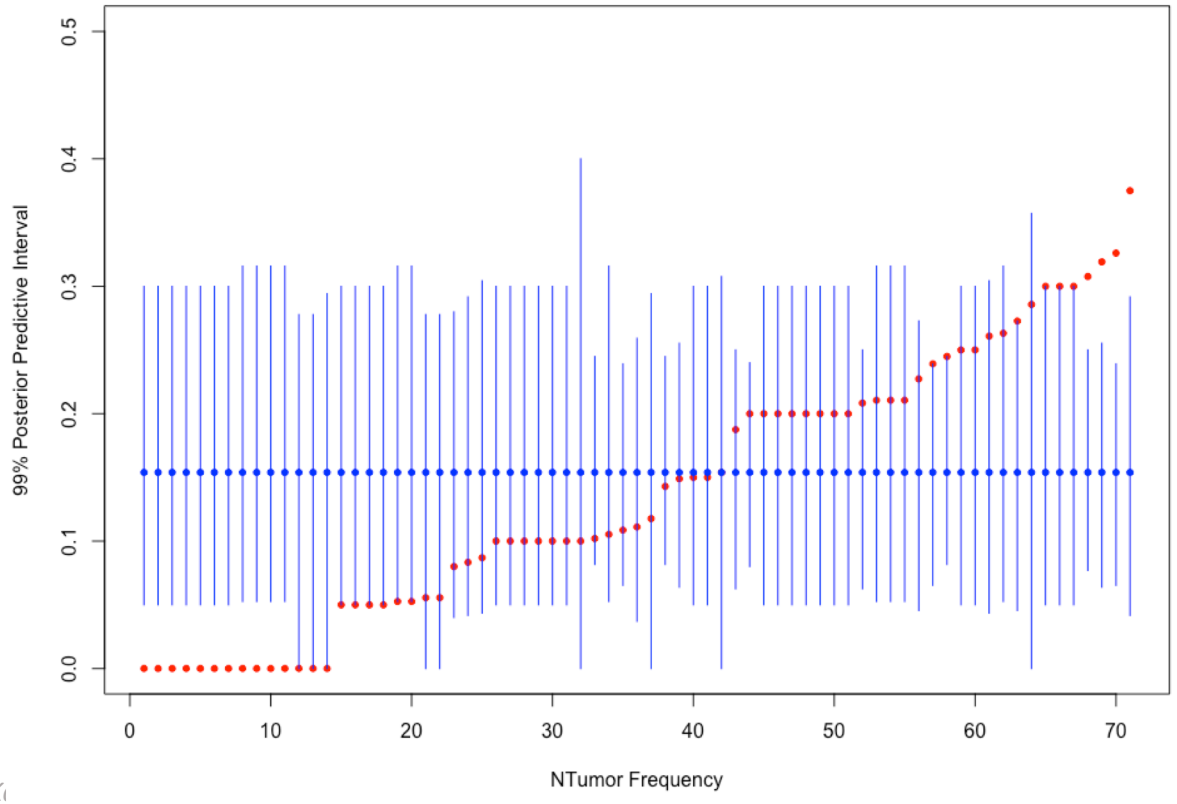


Posterior Predictive Evaluation of Pooled Model

- Posterior predictive distribution for a new study of size n is beta-binomial with probability 0.154, size n, and overdispersion 1741
- Compare observed tumor counts in the 71 studies with posterior expected counts for 71 new studies with same sample sizes
- There are more studies with zero or more than 12 tumors than expected by the posterior predictive analysis
- This suggests that different studies may have different tumor probabilities
  - Some very small and some larger probabilities



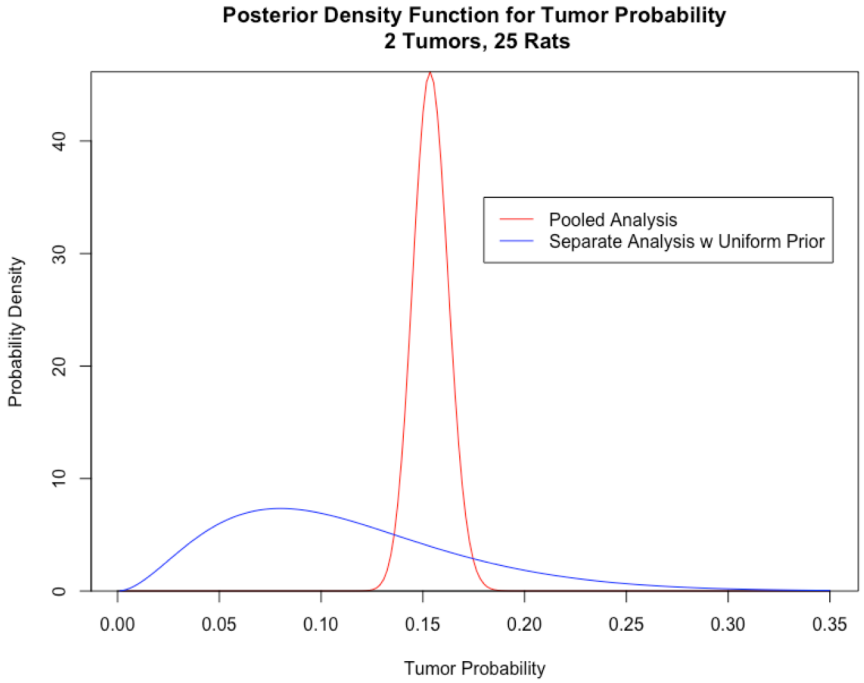
Sample Frequencies and Posterior Predictive Intervals for Sample Frequencies



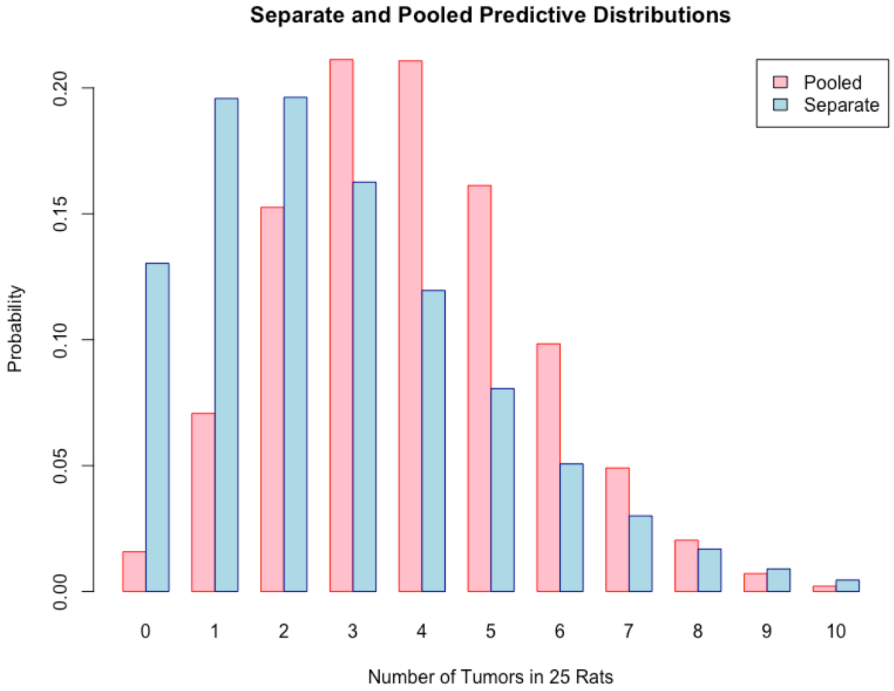
- We expect 7 out of 71 studies to lie outside the 90% predictive interval
- Sample frequencies for 15 out of 71 studies lie outside 90% posterior predictive interval
- The pooled model is a poor fit to the observations

Pooled and Separate Analyses

- Posterior density functions for pooled and separate analysis of study #24 (2 tumors in 25 rats)



- Posterior predictive distributions for pooled and separate analysis of study #24 (2 tumors in 25 rats)



Hierarchical Model for Rat Tumor Example

- The data $Y = (Y_1, Y_2, \ldots, Y_{71})$:
  - Y_s are independent draws from a Binomial(n_s, Θ_s) distributions
    - n_s is number of rats and Θ_s is tumor probability for study s
  - The tumor probabilities Θ_s may vary due to differences in rats and experimental conditions
- The parameters $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_{71})$:
  - Conditional on hyperparameters $\mathrm{A} = \alpha$ and $\mathrm{B} = \beta$, the tumor probabilities $\Theta_1, \Theta_2, \ldots, \Theta_{71}$ are independent draws from a Beta(α, β) distribution
- Hyperparameters A and B are drawn from a distribution $g(\alpha, \beta)$

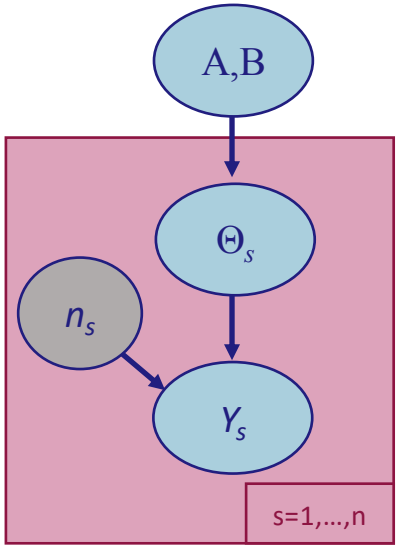


Plate representation

Graphical Representation of Model Structure

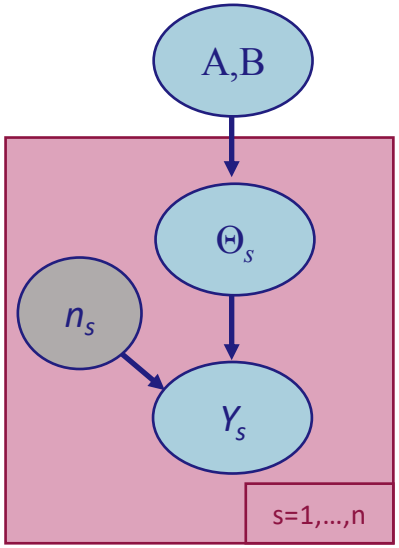
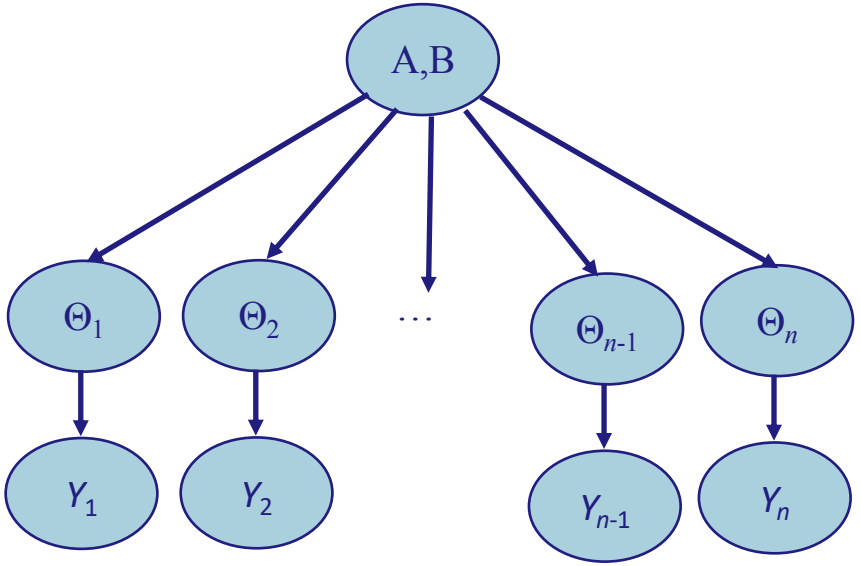


Plate representation

Hyperparameters

Parameters

Observations



"Unrolled" Graph

$(A,B) \sim g(A,B)$

$\Theta_s \sim \text{Beta}(A,B)$

$Y_s \sim \text{Binomial}(n_s, \Theta_s)$

An Equivalent Model

- X_{si} are independent Bernoulli (0/1) random variables indicating whether i^{th} rat in study s develops a tumor
- $Y_s = \Sigma_i X_{si}$ is sufficient statistic for Θ_s

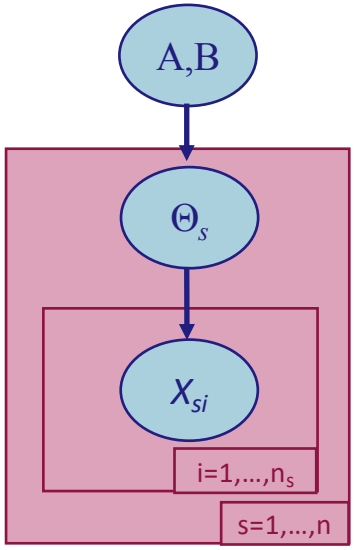
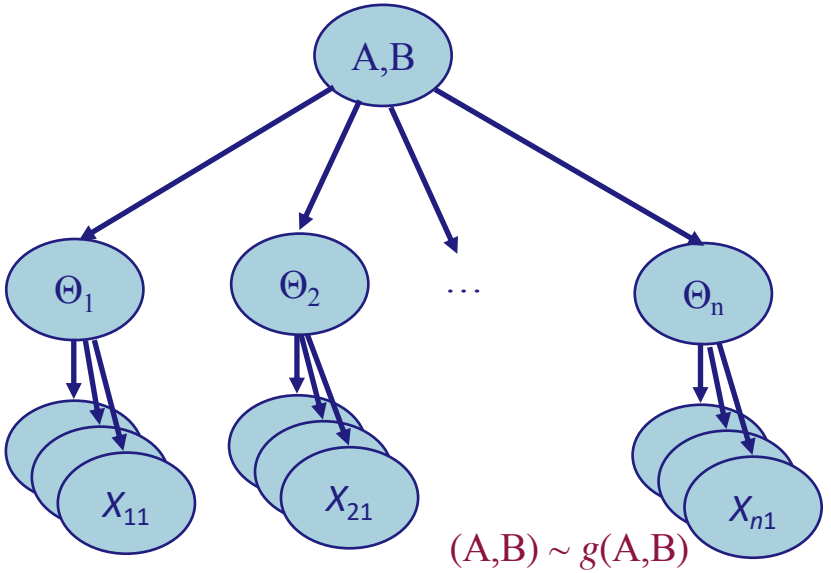


Plate representation



"Unrolled" Graph

$(A,B) \sim g(A,B)$

$\Theta_i \sim \text{Beta}(A,B)$

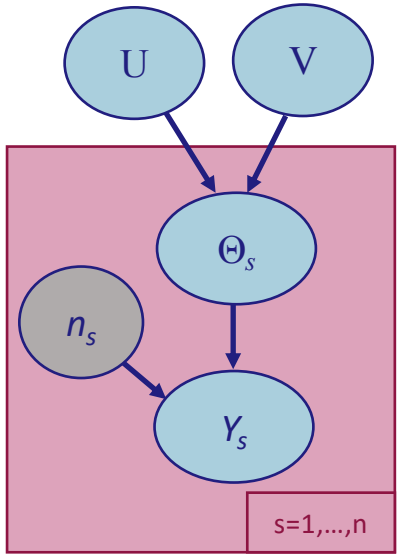
$X_{ij} \sim \text{Bernoulli}(\Theta_i)$

Hyperparameter Prior

- For ease of interpretation and efficiency of computation, reparameterize the model as
  - $U = \mathrm{A}/(\mathrm{A}+\mathrm{B})$ expected tumor probability
  - $V = \mathrm{A} + \mathrm{B}$ virtual count for population prior distribution
  - $\Theta_1, \Theta_2, \ldots, \Theta_n$ ~iid Beta$(UV, (1-U)V)$
- There is no conjugate family for U, V
- We assume U, V independent with prior distribution:
  - U ~ Uniform
  - V ~ Gamma(1,20)
    - Expected value is 20
    - P(V < 90) = 0.99
- Uncertainty in U and V allows the studies to share information
  - If U and V were given then the Θ_s would be independent
  - Each study contributes to posterior distribution of (U, V), which then affects distribution for other Θ_s

***Rationale**: A and B are highly correlated a posteriori, making MCMC sampling less efficient*

***Rationale**: Prior distribution for (U,V) is weakly informative, dominated by the likelihood and yields a proper posterior distribution*



Full Conditional Distributions for (Approximate) Gibbs Sampling

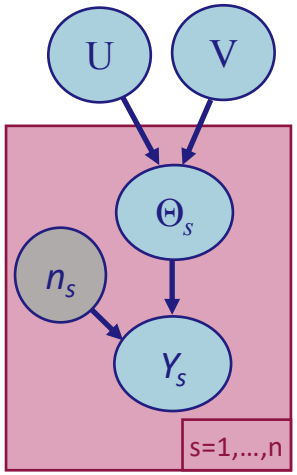
- Conditional on U and V and $Y_{1:71}$ the Θ_s are independent of each other, and have Beta$(UV + Y_s, (1-U)V + n_s - Y_s)$ distributions
- We do not have a closed form distribution for U given V and $\Theta_1, \Theta_2, \ldots, \Theta_n$
  - Define discrete grid of L points $0 < u_1 < \ldots < u_L < 1$
  - Approximate posterior distribution as

    $P_u(u_j|v, \theta_{1:71}) \propto \frac{1}{L} \prod_{i=1}^{71} f(\theta_i|u_j v, (1-u_j)v)$,

    where $f(\theta|\alpha, \beta)$ is the Beta density function
- We do not have a closed form distribution for V given U and $\Theta_1, \Theta_2, \ldots, \Theta_n$
  - Define discrete grid of points $0 < v_1 < \ldots < v_M < \infty$

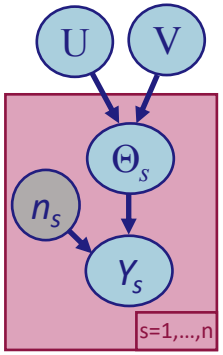
    $P_v(v_j|u, \theta_{1:71}) \propto g(v_j) \prod_{i=1}^{71} f(\theta_i|uv_j, (1-u)v_j)$,

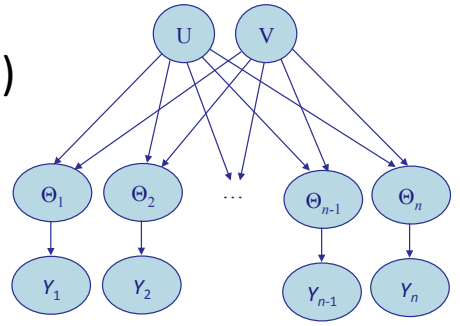
    where $f(\theta|\alpha, \beta)$ is the Beta density function and $g(v)$ is the gamma density function



Gibbs Sampling Procedure

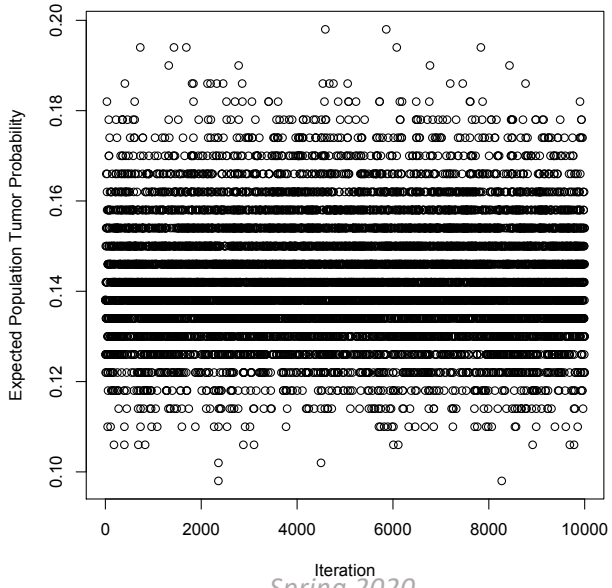
- Initialize $u^{(0)}, v^{(0)}$ (we do not need to initialize θ's)
- For $k = 1$ to desired sample count:
  - Let $\alpha = u^{(k-1)}v^{(k-1)}$ and $\beta = (1 - u^{(k-1)})v^{(k-1)}$
  - For each study $s = 1, \ldots, n$, sample $\theta_s^{(k)}$ from Beta$(\alpha + y_s, \beta + n_s - y_s)$ distribution
  - Sample $u^{(k)}$ from discrete approximation $P_u(u \mid v^{(k-1)}, \theta_1^{(k)}, \ldots, \theta_n^{(k)})$
  - Sample $v^{(k)}$ from discrete approximation $P_v(v \mid u^{(k)}, \theta_1^{(k)}, \ldots, \theta_n^{(k)})$
- Perform MCMC diagnostics on the sample
- Use sample to approximate posterior quantities of interest

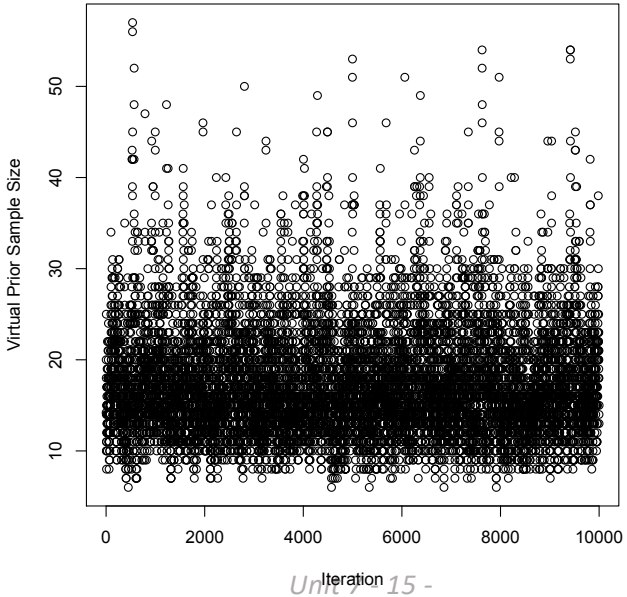




Diagnostics for 10,000 Gibbs Samples

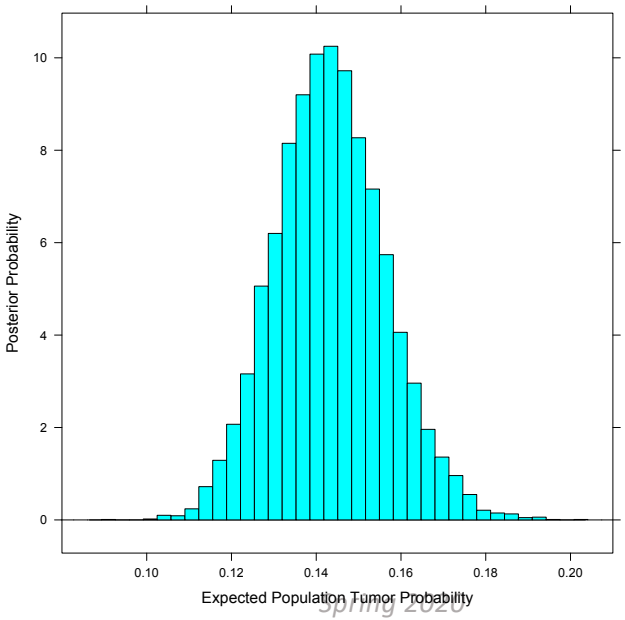
- Autocorrelations for U and V are moderate to high
  - First-order autocorrelation is 0.50 for U and 0.79 for V
- Effective sample size is 3230 for U, 1207 for V, and an average of 8483 for the Θ_s
- Traceplots appear stationary

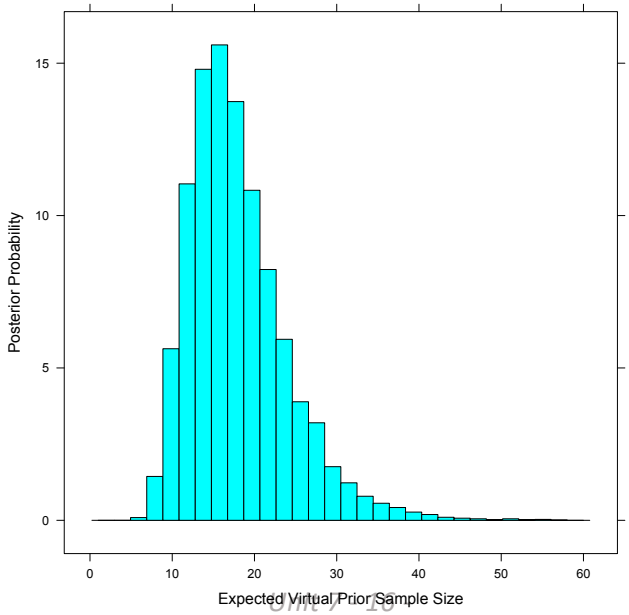




Population Hyperparameter Posterior Distributions

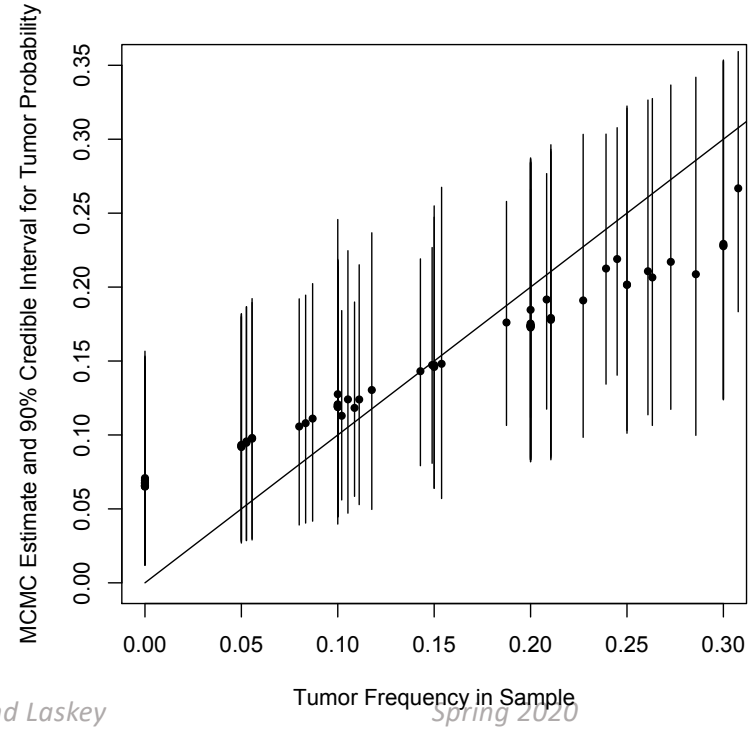
- MC estimate of posterior mean of U is 0.144
  - Average of sample frequencies y_s/n_s is 0.138
  - Population pooled tumor proportion is 0.154
- MC estimate of posterior mean of V is 17.6
  - Considerable information is shared between studies
  - There are 71 studies, with average sample size of 24.5

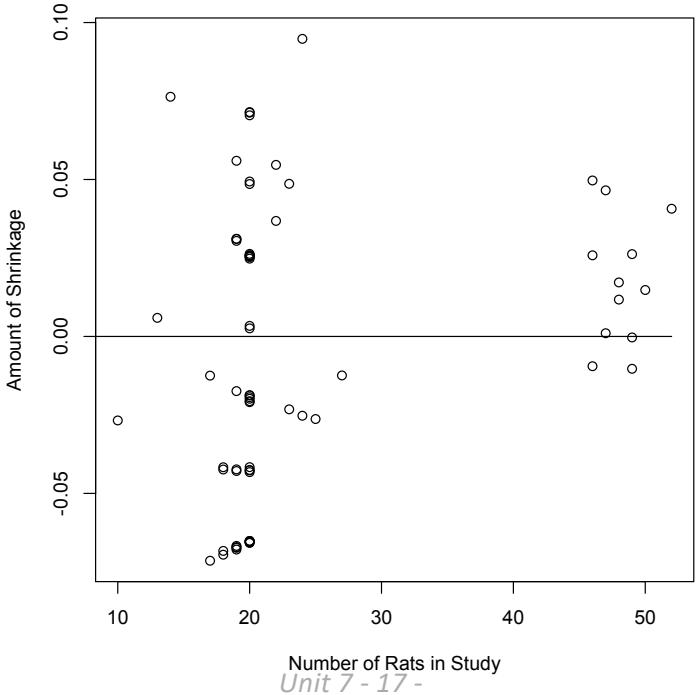




Shrinkage

- Posterior mean tumor probability "shrinks" sample proportions toward population mean
- Shrinkage tends to be greater for smaller studies
- Smaller studies "borrow strength" from other studies

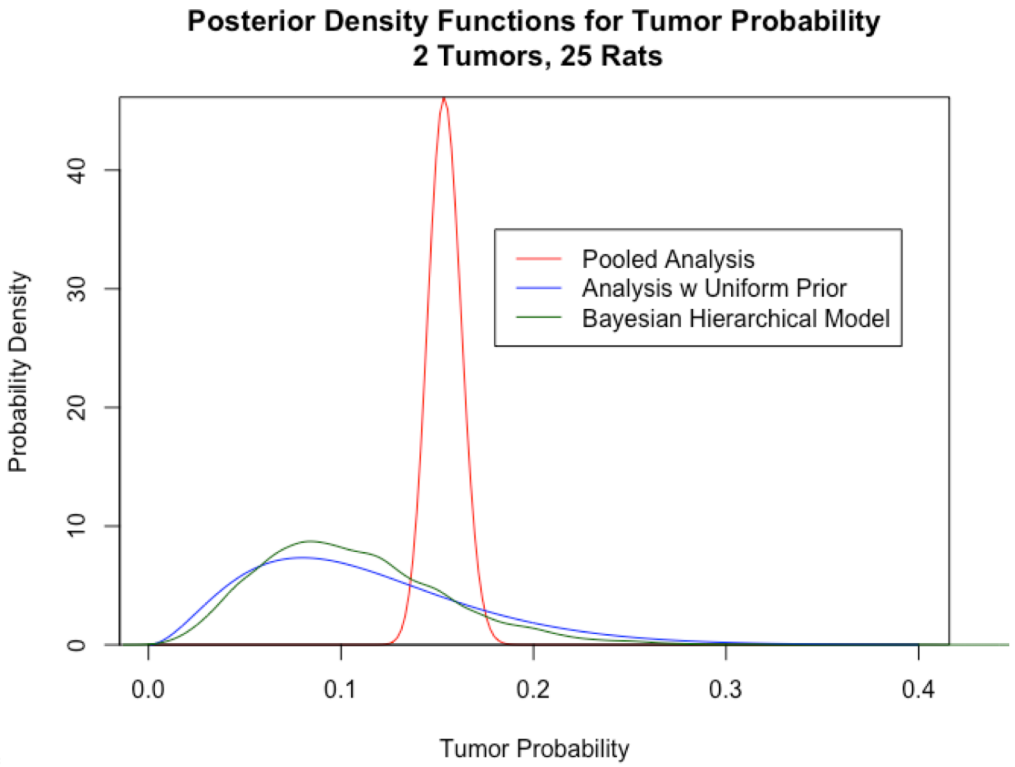


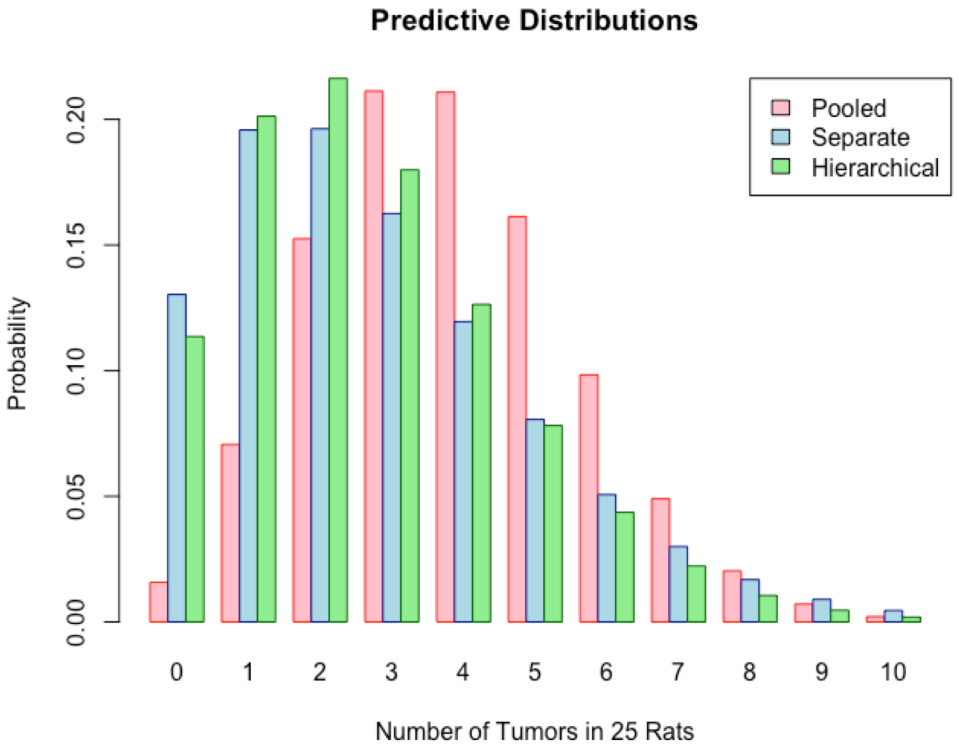


Shrinkage and Order Reversals

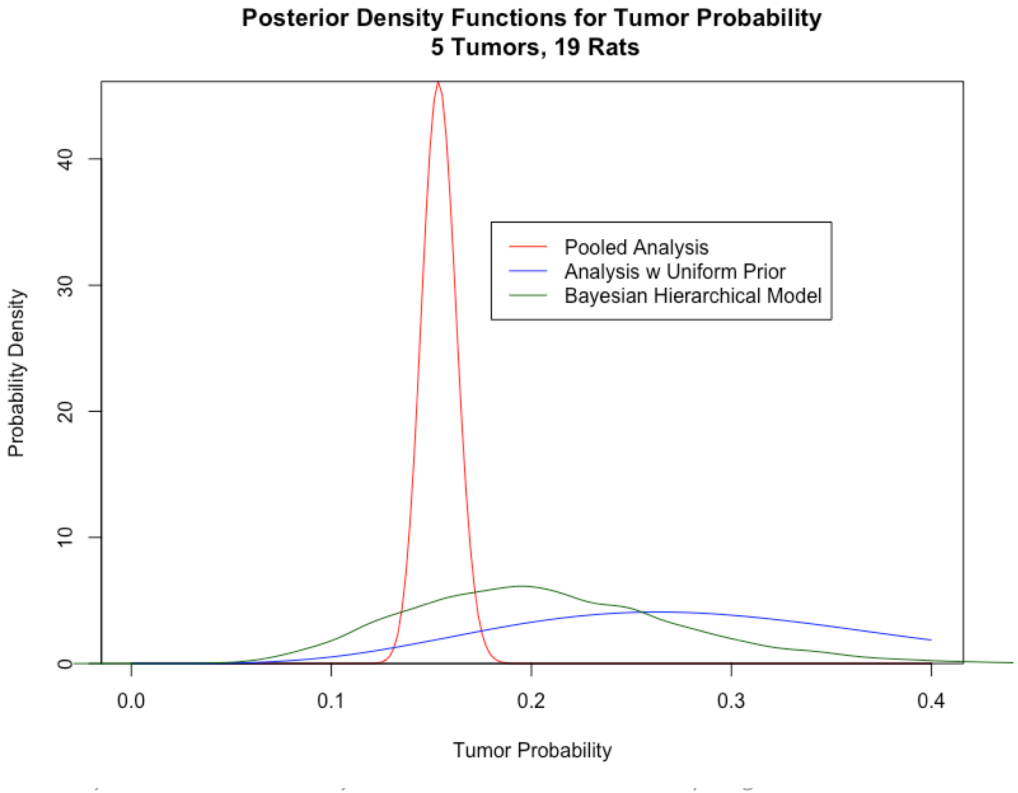
- Shrinkage can change the rank ordering of point estimators
- Example:
  - Study number 58 had 12 tumors in 49 rats
    - Sample frequency is 12/49 = 0.245
    - Estimated posterior mean is $E[\Theta_{58} \mid Y] = 0.219$
  - Study number 71 had 4 tumors in 14 rats
    - Sample frequency is 4/14 = 0.286
    - Estimated posterior mean is $E[\Theta_{71} \mid Y] = 0.209$
  - Sample frequencies for both studies shrink toward the population mean but smaller study shrinks more
  - This causes a reversal of ranks: Study 71 has a larger sample proportion but a smaller estimated posterior mean than Study 58
- This makes sense: sample proportions for smaller studies are more variable, so extreme values are more likely to be due to sample fluctuation

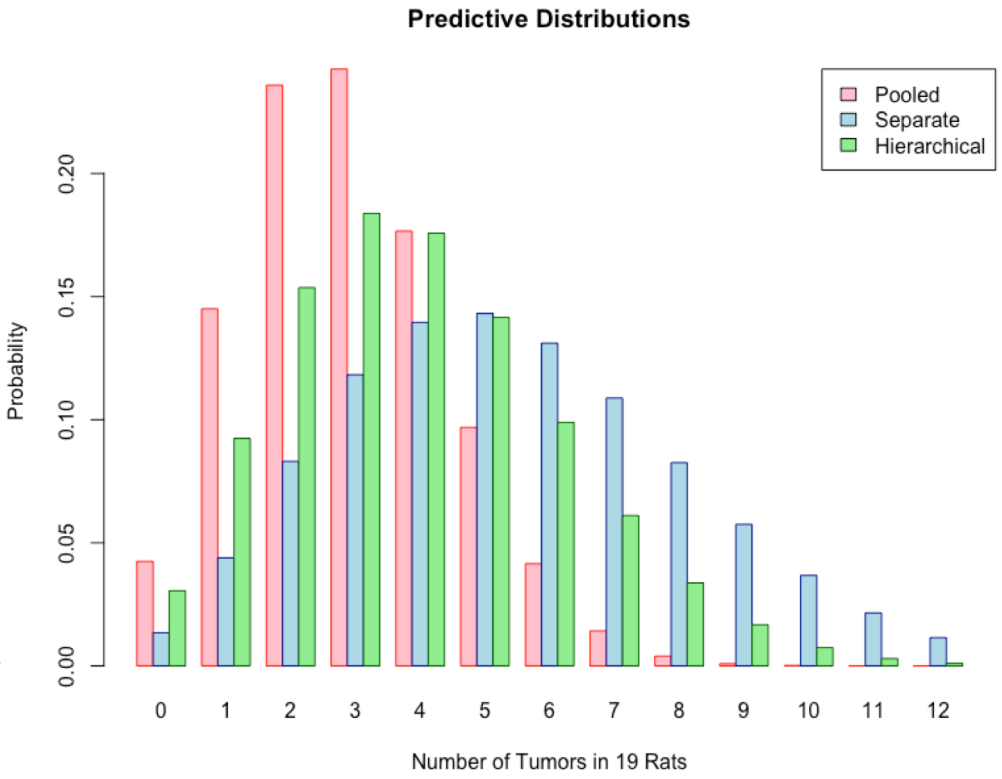
Pooled, Separate, and Hierarchical Analyses for Study #24 (Observed Frequency 0.08)



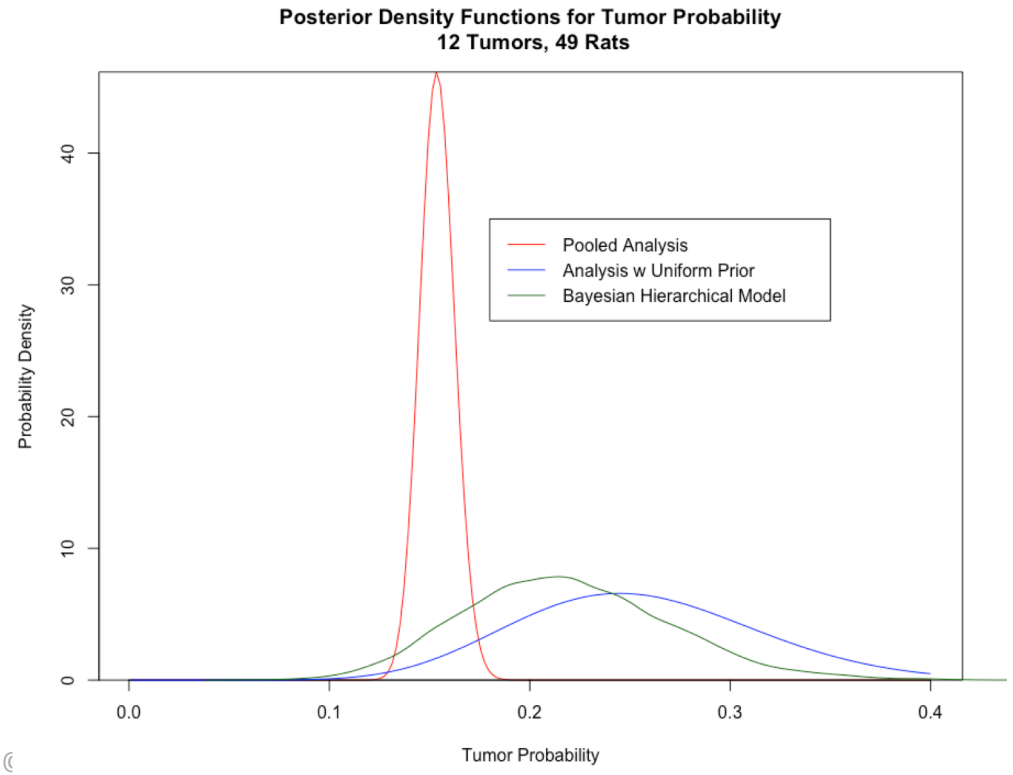


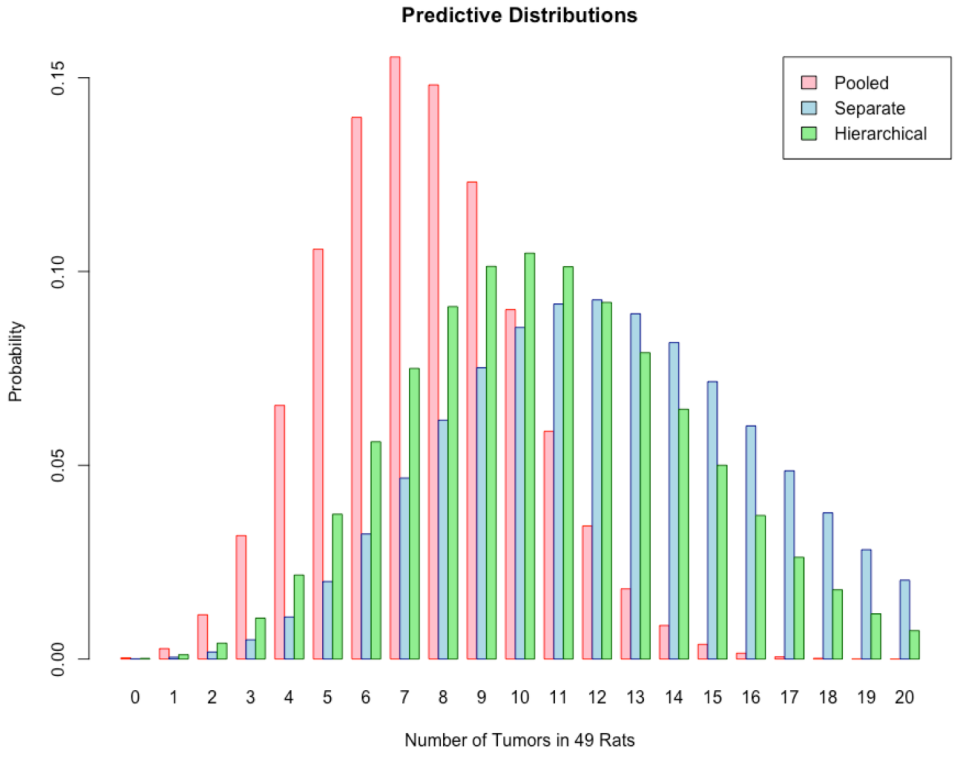
Pooled, Separate, and Hierarchical Analyses for Study #62 (Observed Frequency 0.263)



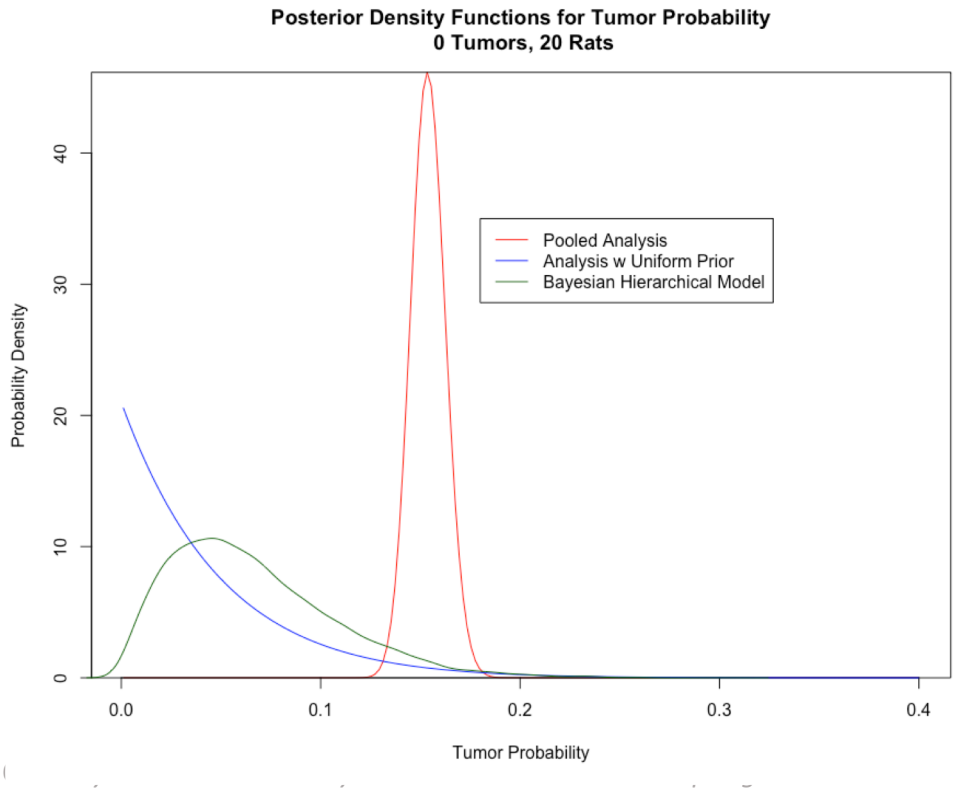


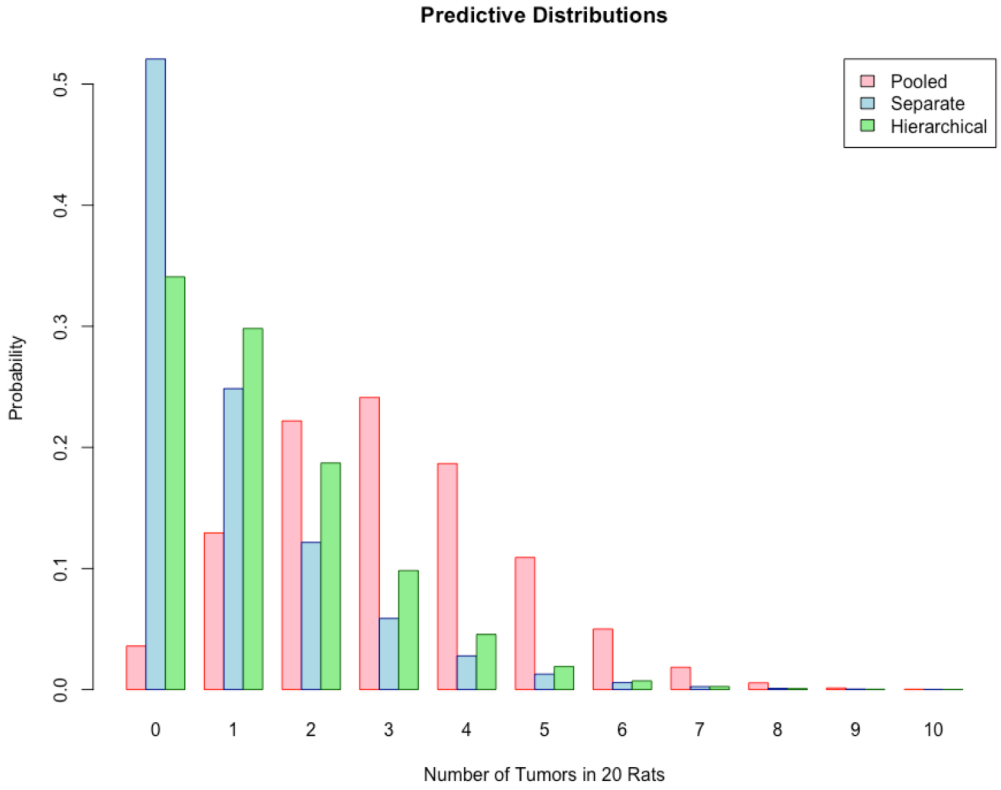
Pooled, Separate, and Hierarchical Analyses for Study #58 (Observed Frequency 0.245)

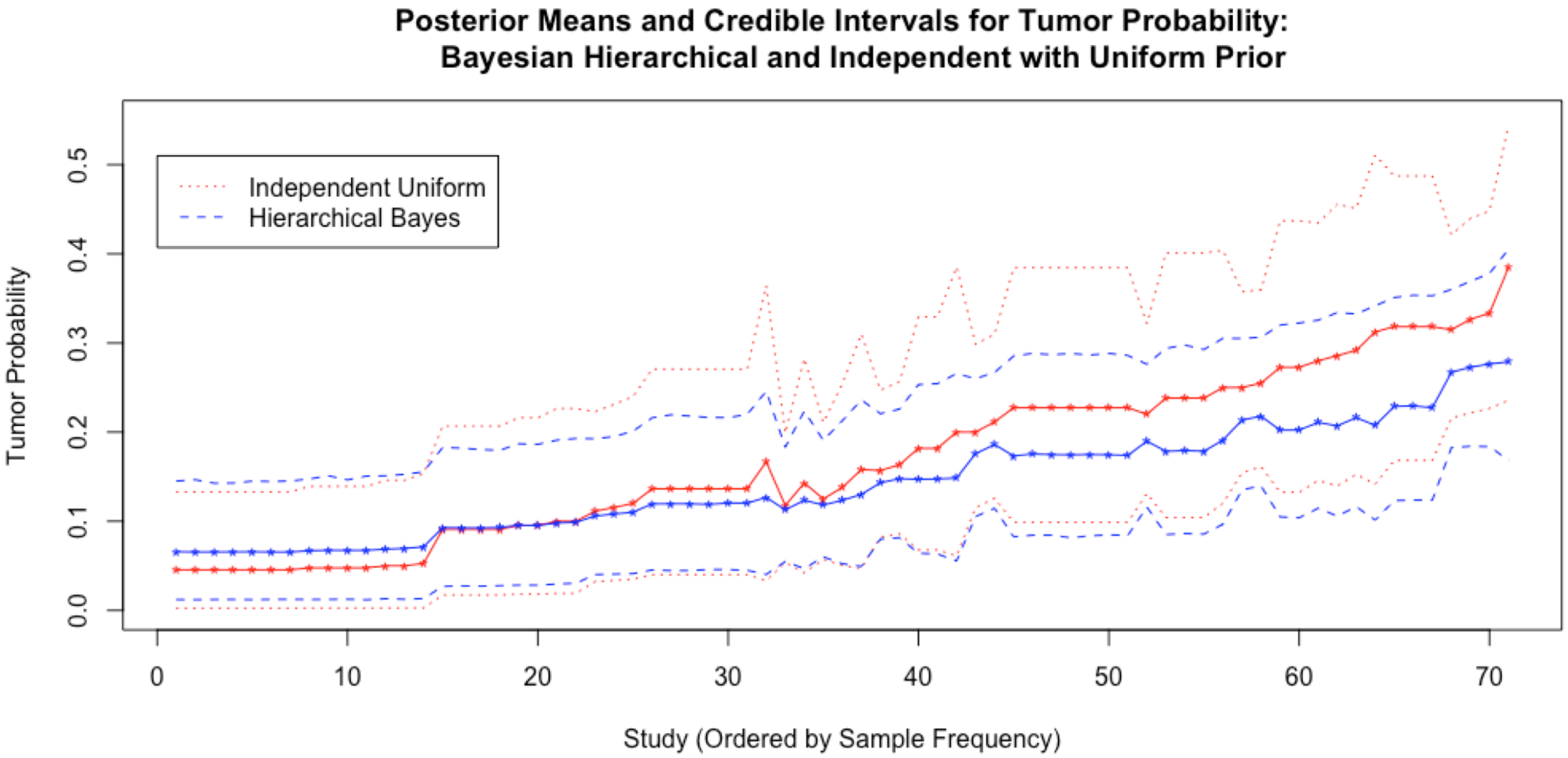




Pooled, Separate, and Hierarchical Analyses for Study #1 (Observed Frequency 0)





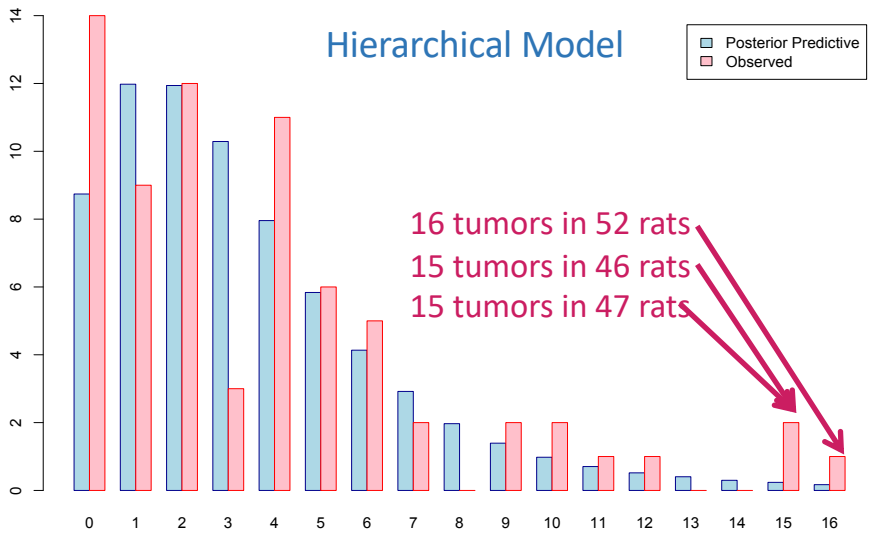


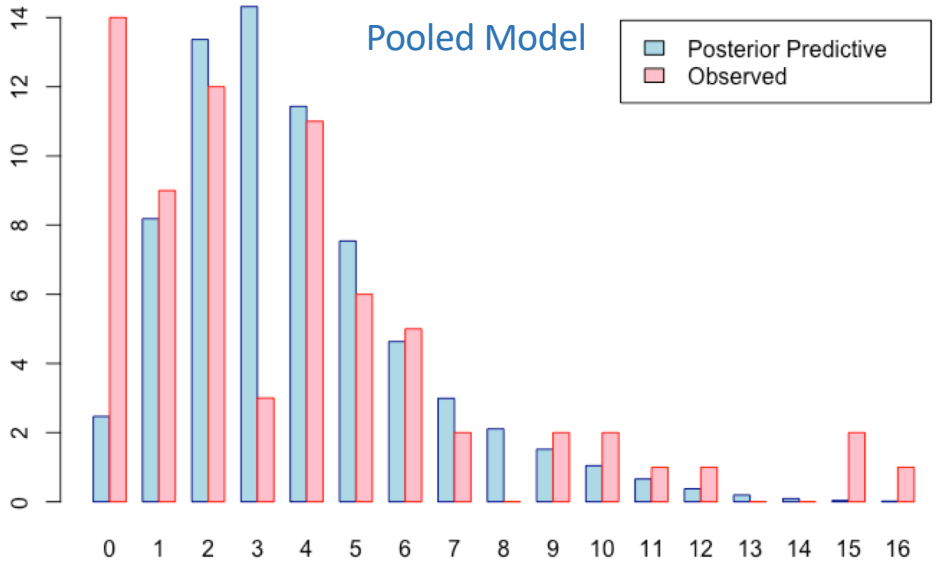
90% Credible intervals for hierarchical analysis are generally narrower

Uniform prior shrinks (slightly) toward 0.5; Bayesian hierarchical model shrinks (more) toward 0.14

Posterior Predictive Model Check

- Hierarchical model can be evaluated by examining the posterior predictive distribution
  - For each simulated $(U, V, \Theta_{1:71})$ from MC sampler, simulate a hypothetical replication of all 71 studies by drawing Y_s ~ Binomial(n_s, Θ_s)
- Posterior predictive sample still has more zeros and more high tumor counts than expected
  - But less so than for pooled model
  - We might consider a model with two (or more) "clusters" of studies





Rat Model in JAGS

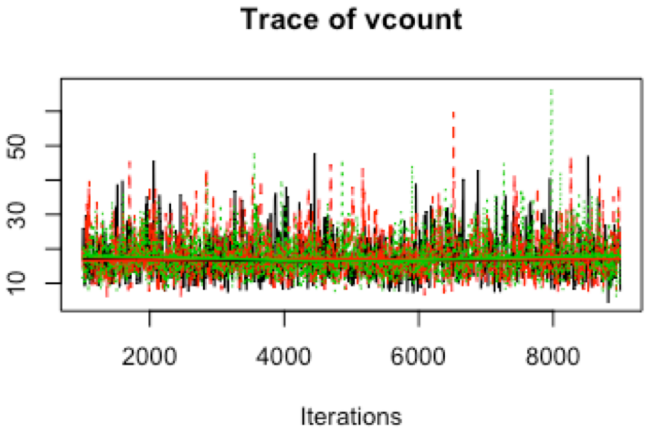
The JAGS Model:

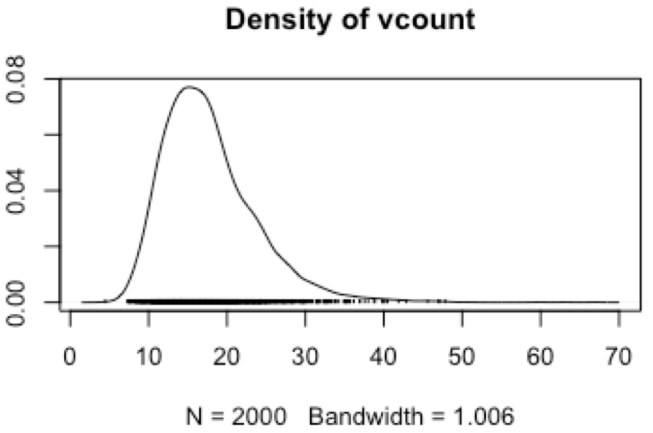
```
model {
    for(i in 1:nS) {
        nt[i]~dbinom(theta[i],nr[i]) # binomial data
        theta[i]~dbeta(alpha,beta)   # beta prior distribution
    }
    alpha<-vfreq*vcount         # virtual tumor count
    beta<-(1-vfreq)*vcount      # virtual non-tumor count
    vfreq~dbeta(1,1)            # virtual tumor frequency is uniform
    vcount~dgamma(1,1/20)       # virtual count
}
```

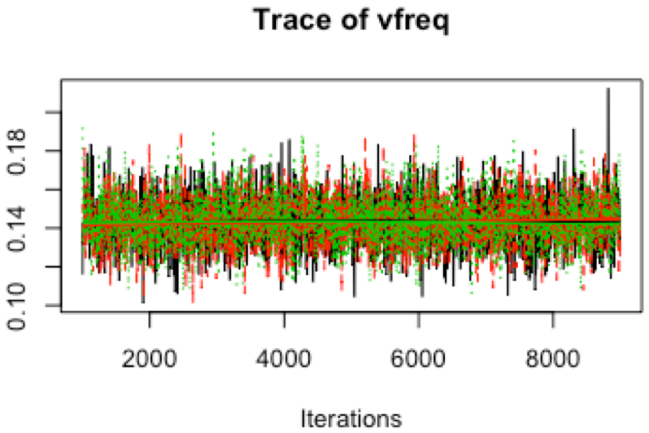
To Run The Model from R:

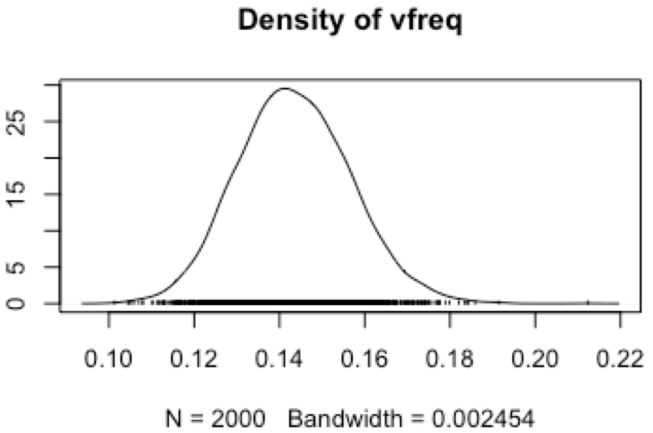
```
# Read in the data and define the variables
rats <- read.table ("rats.txt", header=T)
nS = nrow(rats)      # Number of studies
nt <- rats$tumors    # Number of tumors in each study
nr <- rats$numRats   # Number of rats in each study

require(R2jags)        # JAGS

theta.hat <- nt/nr          # Frequency of tumors in each study

#Gibbs sample

rat.data <- list("nS","nt","nr")

rat.params <- c("vfreq","vcount","theta")

rat.inits <- rt.inits <- function(){
 list("vfreq"=c(0.5), "vcount"=c(20),"theta"=array(mean(nt/nr),nS))
 }

# The jags function takes data and starting values as input. It automatically writes
# a jags script, calls the model, and saves the simulations for easy access in R.
rat.fit <- jags(data=rat.data, inits=rat.inits, rat.params, n.chains=3,
                n.iter=9000, n.burnin=1000,model.file="rats.model.jags",
                n.thin=4)

rat.fit.mcmc <- as.mcmc(rat.fit)  # change to mcmc object
summary(rat.fit.mcmc)             # show summary of mcmc object
uv.mcmc =                         # extract chains for tumor prob and vcount
  rat.fit.mcmc[,match(c("vcount","vfreq"),varnames(rat.fit.mcmc))]
plot(uv.mcmc)                     # plot chains for tumor prob and vcount
```

JAGS Output



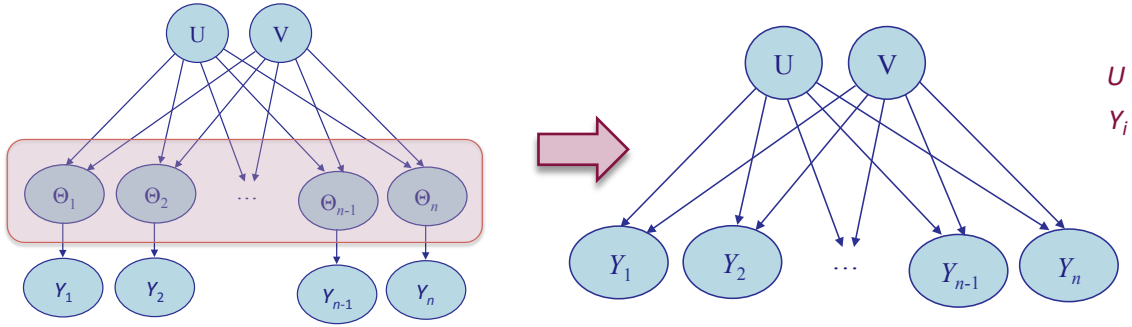






Collapsing the Model: Integrating Out the Tumor Probabilities

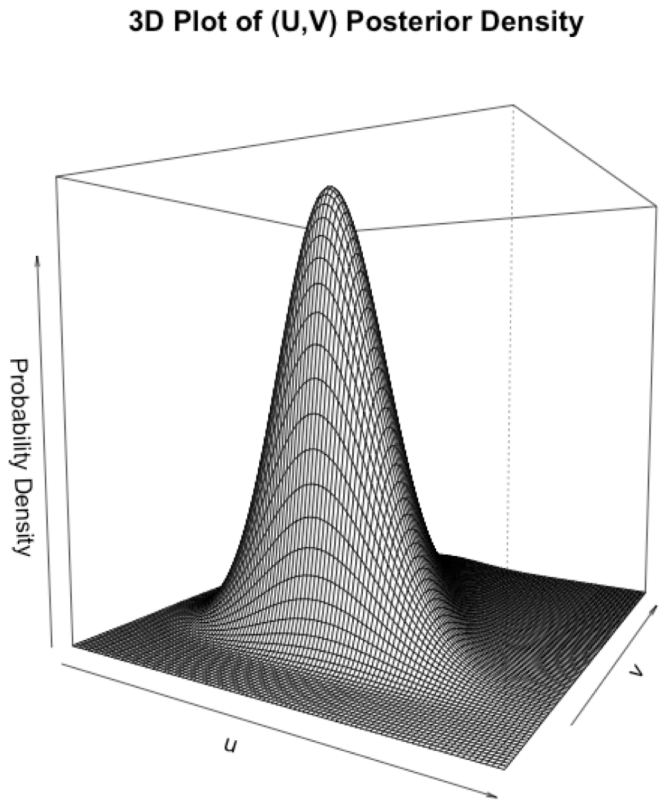
- We can estimate $P(U, V \mid Y)$ directly by marginalizing out the Θ_s
- Conditional on $U = u$ and $V = v$, observations Y_s are independent beta-binomial random variables with size n_s, probability u, and overdispersion v
- We can use numerical methods to estimate the posterior density $g(u, v|\underline{y}) \propto f\left(\underline{y}\middle|u, v\right) g(u, v)$
  - $f\left(\underline{y}\middle|u, v\right)$ is the product of 71 beta-binomial likelihood functions
  - $g(u, v)$ is the the product of a uniform density for u and a gamma(1,20) density for v

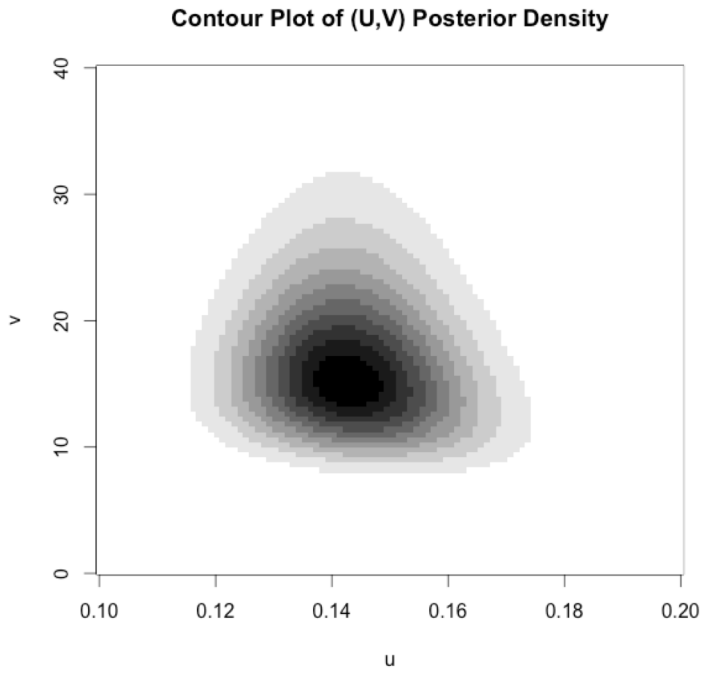


U ~ Uniform; *V* ~ Gamma(1,20)

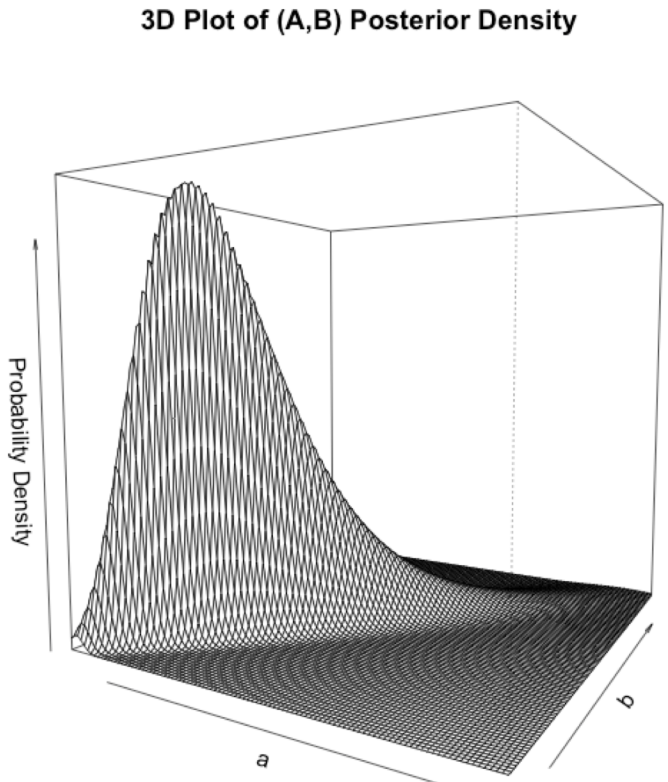
$Y_i \sim_{\text{indep}}$ Beta-Binomial(n_i, uv, $(1-u)v$)

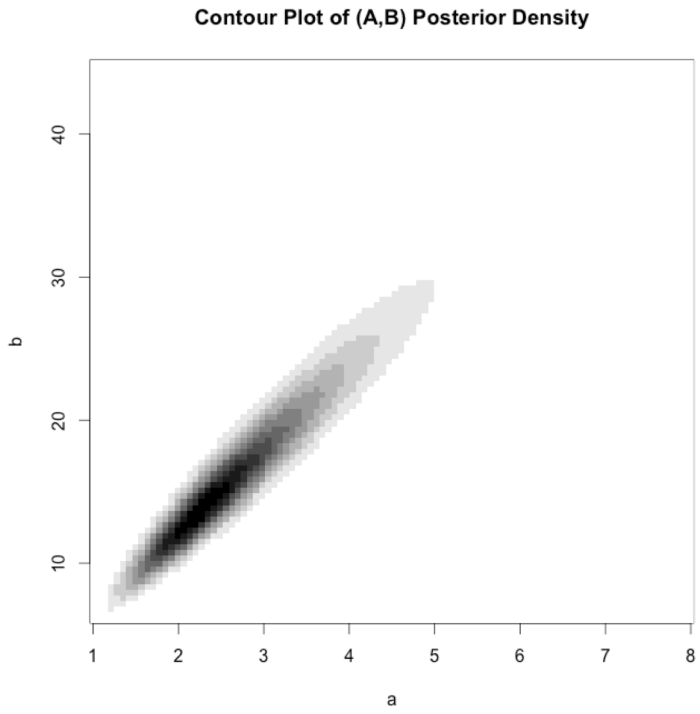
Posterior Distribution: (U,V) Parameterization



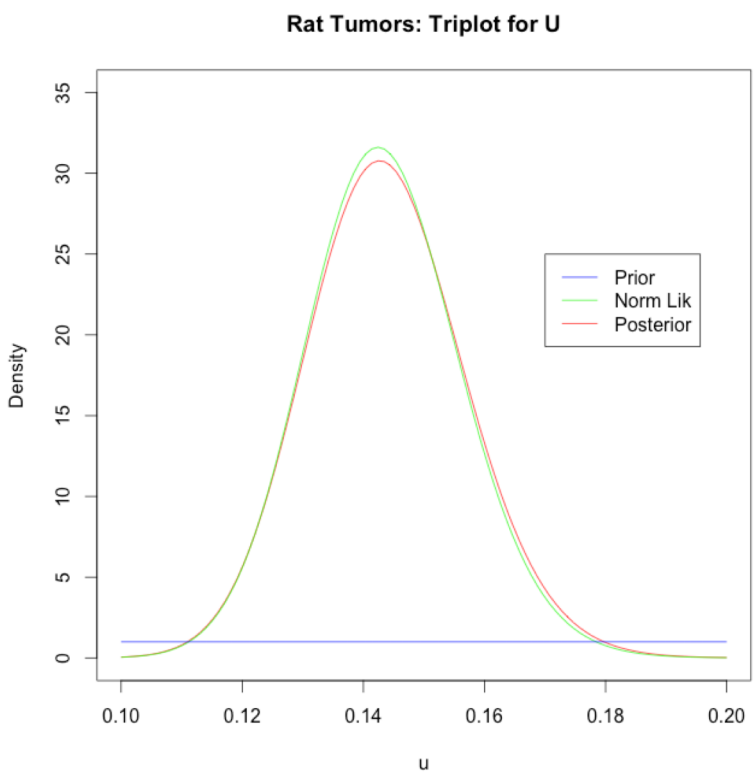


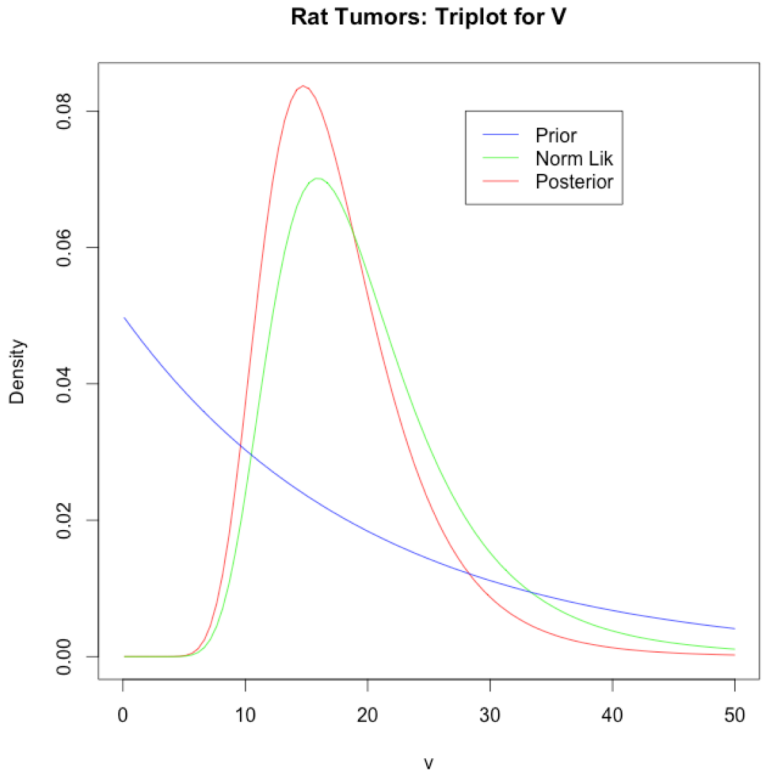
Posterior Distribution: (A,B) Parameterization



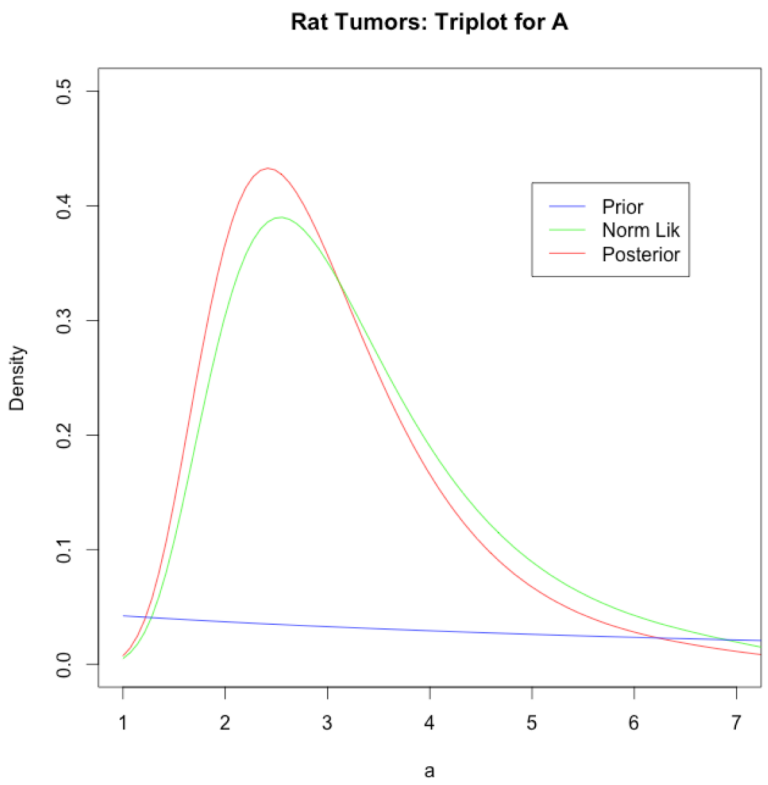


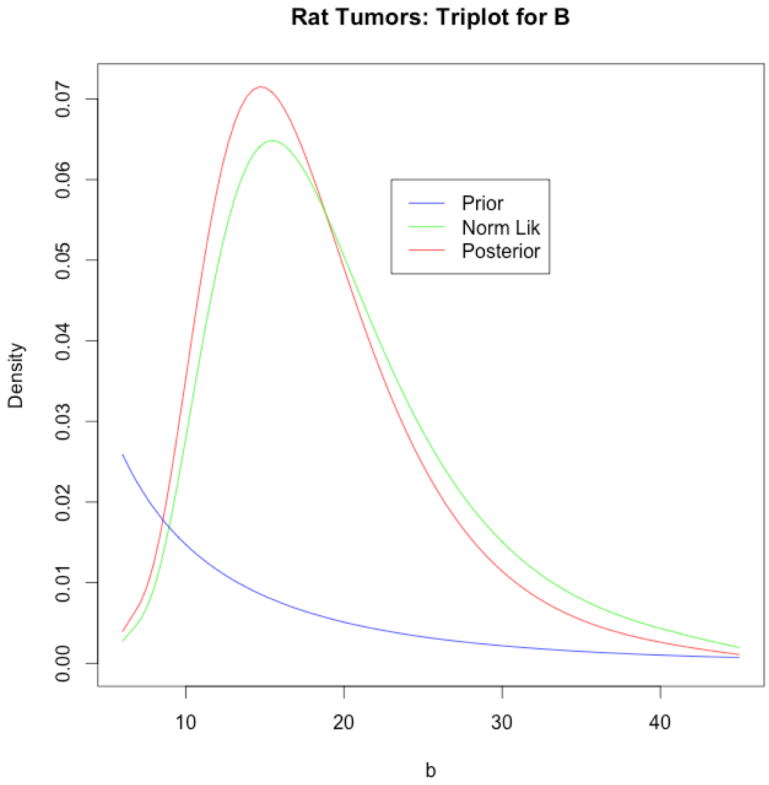
Triplots for Marginal Posterior Distributions for U and V



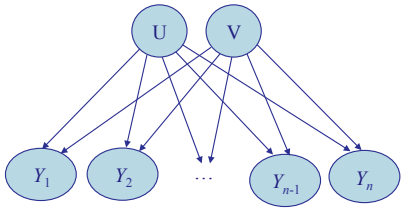


Triplots for Marginal Posterior Distributions for A and B





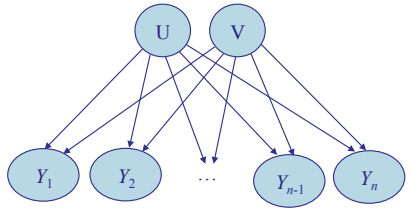
Collapsed Gibbs Sampling Procedure



- Initialize $u^{(0)}$ and $v^{(0)}$
- For k=1 to desired sample count:
  - Sample $u^{(k)}$ from numerical approximation to $P\left(u \left| v^{(k-1)}, y_{1:71} \right.\right) \propto f\left(y_{1:71} \left| u, v^{(k-1)} \right.\right)$
  - Sample $v^{(k)}$ from numerical approximation to $P\left(v \left| u^{(k)}, y_{1:71} \right.\right) \propto f\left(y_{1:71} \left| u^{(k)}, v \right.\right) g(v)$
  - Sample $\theta_{1:71}^{(k)}$ if desired (we may not need to sample the θ's because we have an exact distribution given u, v and $y_{1:71}$)
- Perform MCMC diagnostics on the sample
- Use sample to approximate posterior quantities of interest
  - Example: expected tumor probability and 90% interval for each study
    - For each $(u^{(k)}, v^{(k)})$ use beta distribution (or the MC sample if we sampled the θ's) to compute expected values and 90% intervals for tumor probabilities $\theta_{1:71}$ given $(u^{(k)}, v^{(k)})$
    - Use sample averages to estimate unconditional expected tumor probabilities and 90% intervals for $\theta_{1:71}$

$f(y_{1:71}|u, v)$ is product of beta-binomial mass functions for each y_s
$g(v)$ is a gamma density function

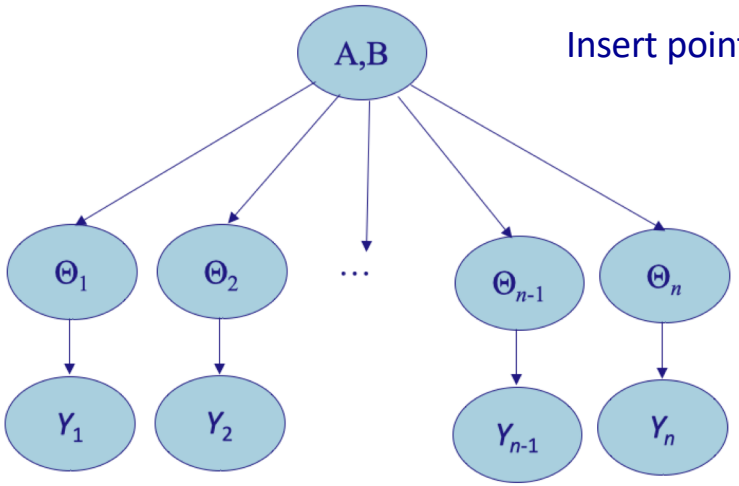
Collapsed Gibbs Sampling: Summary



- Collapsed Gibbs sampling integrates out $\Theta_{1:71}$, and does Gibbs (or MH) sampling on U and V
  - We can also sample the Θ's given (U, V) or use the exact beta distributions to estimate posterior quantities for the Θ's
- Collapsed sampling is an example of a general principle called "Rao Blackwellization"
  - Rao Blackwell Theorem: replacing a Monte Carlo step by an exact integral with the correct expectation reduces the variance of the resulting estimator
  - If the exact integral can be calculated efficiently, the resulting estimator will have lower variance for equivalent computation cost

Empirical Bayes

- Empirical Bayes is an approach that borrows ideas from both frequency and Bayesian viewpoint
  - If we knew α and β (or u and v) then the Θ_s would be independent and identically distributed Beta random variables
  - Empirical Bayes uses the data to construct point estimates of the hyperparameters, then uses the estimates to perform conjugate Bayesian updating on all the studies
- Provides some of the advantages of Bayesian updating at much lower computational cost



Insert point estimate of α and β

Given A=α and B=β:
$\Theta_s \sim_{iid}$ Beta(α, β)
Given A=α and B=β and $\underline{Y}$:
$\Theta_s \sim_{indep}$ Beta($\alpha+Y_s$, $\beta+n_s-Y_s$)

Empirical Bayes Inference - Calculations

- We can think of Y_s/n_s as an estimate of Θ_s and use moment matching to estimate the hyperparameters
- The sample mean and sample variance of observed frequencies $Y_1/n_1, \ldots, Y_{71}/n_{71}$ are 0.138 and 0.0109, respectively
- We look for a Beta distribution with this mean and variance
  - $E[\Theta_s] = \alpha/(\alpha+\beta) = 0.138$
  - $V[\Theta_s] = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1)) = 0.0109$
- We obtain $\alpha = 1.37$ and $\beta = 8.54$
  - Expected value of Θ_s is 0.138; median is 0.114
  - 90% prior credible interval for Θ_s is [0.016, 0.344]
- Posterior distribution for a study with Y_s tumors in n_s rats is Beta(1.37+ Y_s, 8.54 + n_s - Y_s)

Solve for α and β:

$\alpha/(\alpha+\beta) = 0.138$

$\beta/(\alpha+\beta) = 1-0.138 = 0.862$

$0.0109 = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$

$= (0.138)(0.862)/(\alpha+\beta+1)$

$(\alpha+\beta+1) = (0.138)(0.862)/0.0109 = 10.91$

$\alpha+\beta = 9.91$

$\alpha = 9.91*0.138 = 1.37$

$\beta = 9.91*0.862 = 8.54$

Comparison: Gibbs sampling estimate of hyperparameters for the fully Bayesian model:

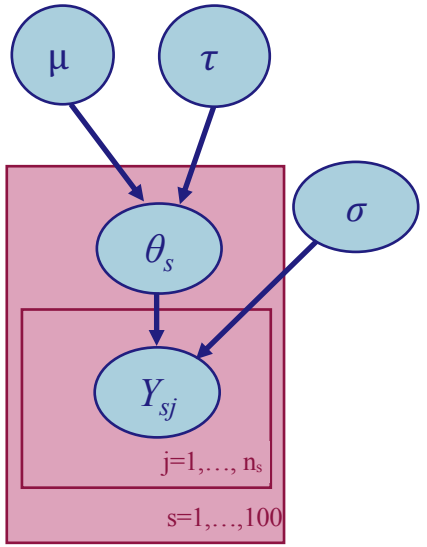
- $E[U] = 0.144$
- $E[V] = 17.6$
- $E[A] = 2.55$
- $E[B] = 15.25$

Remarks on Empirical Bayes

- Empirical Bayes method allows estimates for sub-populations to borrow strength from data for other sub-populations
- Strict Bayesians criticize empirical Bayes for using the data twice
  - Empirical Bayesian analyses can be prone to overfitting
- Estimation method for prior hyperparameters requires care and judgment
- A carefully done empirical Bayesian analysis can provide many of the advantages of a full Bayesian analysis without the computational overhead of MCMC

Example: Math Test Scores

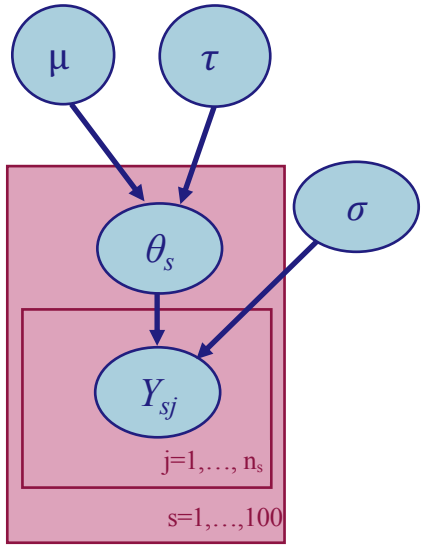
- Chapter 8 of Hoff applies a hierarchical model to analyze math test scores in 100 schools
  - There are n_s students in school s
  - Test scores Y_{sj} are normally distributed with school-specific mean θ_s and common standard deviation σ
  - θ_s is normally distributed with mean μ and standard deviation τ
- Weakly informative prior distributions for μ, σ and τ
  - μ is normally distributed with mean 50 and standard deviation 25
    - 95% credible interval for μ is about [0, 100]
  - $1/\sigma^2$ is gamma with shape 1/2 and scale 1/50
    - Expected value is 1/100. Test was designed to have nationwide variance 100. Therefore, within-school variance should be no more than 100
    - Small shape parameter means prior is weakly concentrated around this value
  - $1/\tau^2$ is gamma with shape 1/2 and scale 1/50
    - Also weakly concentrated around a precision of 1/100



Conditional Posterior Distributions for Math Test Hierarchical Model

- Given $(\theta_{1:100}, \tau)$, μ is normally distributed with
  - mean $\frac{100\theta/\tau^2+50/25^2}{100/\tau^2+1/25^2}$ and
  - standard deviation $(100/\tau^2 + 1/25^2)^{-1/2}$
- Given $(\theta_{1:100},\mu)$, $1/\tau^2$ has a gamma distribution with
  - shape ½+100/2 =50.5
  - scale $\left(50 + \frac{1}{2}(\sum(\theta_s - \mu)^2)\right)^{-1}$
- Given $(y_{1:100}, \theta_{1:100})$, $1/\sigma^2$ has a gamma distribution with
  - shape ½+$\frac{1}{2}\sum_{s=1}^{100} n_s$
  - scale $\left(50 + \frac{1}{2}\left(\sum_{s=1}^{100}\sum_{i=1}^{n_s}(y_{si} - \theta_s)^2\right)\right)^{-1}$
- Given $(\mu, \tau, \sigma, y_{s,j}, j = 1{:}\, n_s)$, each θ_s is normally distributed with
  - mean $\frac{\sum_{i=1}^{n_s} y_{si}/\sigma^2+\mu/\tau^2}{n_s/\sigma^2+1/\tau^2}$ and
  - standard deviation $(n_s/\sigma^2 + 1/\tau^2)^{-1/2}$

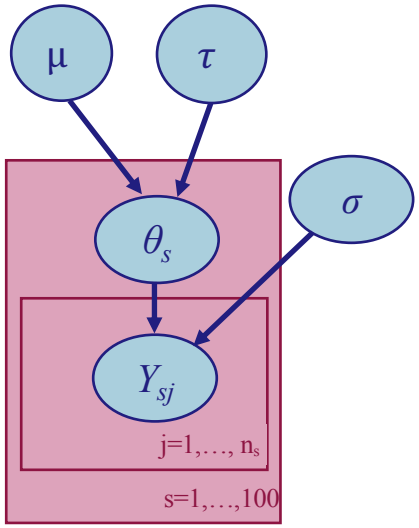
These formulas apply the normal and gamma semi-conjugate updating equations given in Unit 6 for the 2-parameter normal model.



Results: Math Test Scores

- Hoff shows results of Gibbs sampling with 5000 samples
- Diagnostics show that samples for all parameters are stationary and effective sample sizes are all above 2500
- Posterior distribution shrinks sample means for schools toward population mean, with smaller samples shrinking more

Data and code available at https://pdhoff.github.io/book/



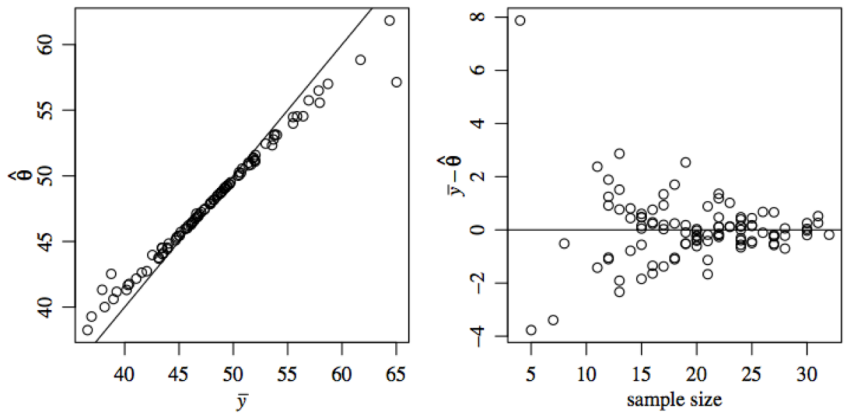


Fig. 8.8. Shrinkage as a function of sample size.

Fitting the Math Scores Example in JAGS

The JAGS Model:

```
model {
  for(i in 1:nS) {
    theta[i] ~ dnorm(mu, lambda)       # observation mean for school i
  }
  for (j in 1:n) {
    score[j] ~ dnorm(theta[school[j]],rho)  # observation for student
  }

  mu ~ dnorm(50,0.0016)    # normal prior distribution for school means
  lambda ~ dgamma(0.5,50)  # gamma prior distribution for 1/(tau^2)
  rho ~ dgamma(0.5,50)     # gamma prior distribution for 1/(sigma^2)
}
```

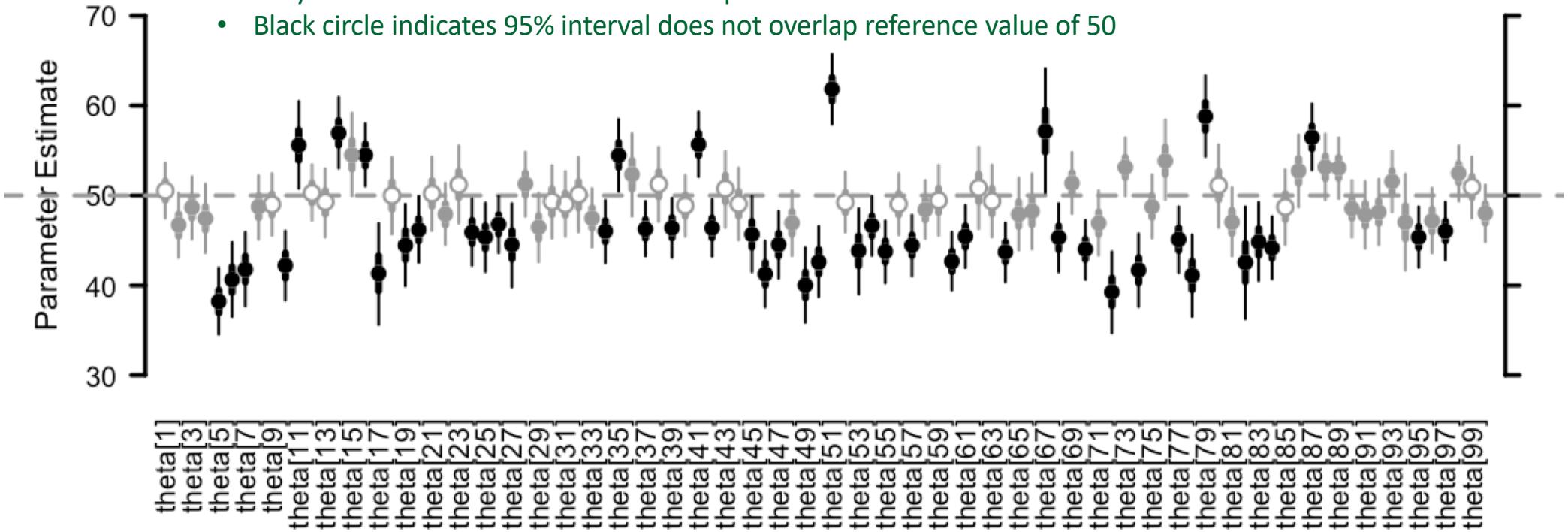
To Run The Model from R:

```
# Read in the data and define the variables
Y.school.mathscore<-dget("Y.school.mathscore") # math score data
school = Y.school.mathscore[,1]                # 1st column is school
score = Y.school.mathscore[,2]                 # 2nd column is score
nS = length(unique(school))                    # number of schools
n = length(score)                             # total number of observations

school.means = aggregate(score, by=list(school), FUN="mean")[,2]  # find means for schools
school.sd = aggregate(score, by=list(school), FUN="sd")[,2]  # find SDs for schools
school.size = aggregate(score, by=list(school), FUN="length")[,2] # number of students

library(R2jags)        # JAGS

#Gibbs sample

math.data <- list("nS","n","school","score")

math.params <- c("mu","lambda","rho","theta")

math.inits <- rt.inits <- function(){
 list("mu"=c(50), "lambda"=c(0.01),"rho"=c(0.01), "theta"=array(50,nS))
 }

# The jags function takes data and starting values as input. It automatically writes
# a jags script, calls the model, and saves the simulations for easy access in R.
math.fit <- jags(data=math.data, inits=math.inits, math.params, n.chains=2,
                 n.iter=5000, n.burnin=100,model.file="mathscore.model.jags",
                 n.thin=1)
```

Caterpillar Plot of Sampled Posterior School Means

Uses `MCMCplot` function from MCMCvis package

- White circle indicates 50% interval overlaps reference value of 50
- Gray circle indicates 95% interval overlaps reference value of 50
- Black circle indicates 95% interval does not overlap reference value of 50



Some Comments on Hierarchical Models

- We considered hierarchical models in which parameters can vary between sub-populations
- Traditionally, modelers had to choose between pooling all observations and treat as a single data set, or analyzing each sub-population separately
  - Pooling introduces bias
  - Treating them all separately increases variance of estimates, especially for sub-populations with fewer observations
- Bayesian analysis gives us advantages of pooling while mitigating the disadvantages
  - Represent variability in subpopulations with separate parameters for each study
  - "Borrow strength" from other sub-populations to inform estimates
- Practical advantages
  - Greater statistical power for sub-populations with smaller sample sizes
  - Especially important when data collection is expensive or difficult

Adding Complexity to a Hierarchical Model

- We used a simple model in which sub-population means are independent and identically distributed given hyperparameters of the prior distribution
- There are several ways to make this kind of model more complex
  - Subgroups (e.g., different breeds of rats / different categories of drugs / clusters discovered from data)
  - Covariates (e.g., weight of rat; age of rat)
  - Regression:
    - Represent E[tumor probability | covariates] as a function with parameters
    - Define prior distribution for parameters and use data to estimate posterior distribution
    - This amounts to defining a higher level of the hierarchy
- Often we look at simpler models first and assess whether a more complex model is needed for our purpose

From Data to Knowledge

- Even the largest data sets are “too small” for all the analyses we would like to do
- Good statisticians learn to:
  - Formulate models with enough statistical power to
    - Extract useful information from an existing data set
    - Plan data collection that will yield useful information
  - Make assumptions that simplify analysis without distorting conclusions
- Bayesian hierarchical models help to squeeze the most out of data
  - With careful modeling, we can specify models with enough flexibility to capture useful structure without overfitting

Top-Level Priors in Hierarchical Models

- In many problems:
  - We have fairly vague prior information about hyperparameters at top level of hierarchy
  - Structural constraints at lower levels give us statistical power
  - We do not want results to depend heavily on details of top-level hyperparameters
  - We want the likelihood function to be much more informative than the prior distribution about the top-level hyperparameters
- We used these principles in defining the prior distributions for the examples in this unit

Meta-Analysis

- Meta-analysis is an increasingly popular approach to combining the results of multiple studies
  - Frequently there are several studies that bear upon a problem of importance to decision makers
  - Analysts must summarize and integrate the results of all these studies
  - Meta-analysis provides a formal methodology for this process
- Bayesian hierarchical modeling provides a natural and theoretically principled theoretical foundation for meta-analysis
  - Gelman et al. supplementary text gives an example of meta-analysis for a medical problem

Small Area Estimation

- Surveys are designed for precise estimation of characteristics of an entire population
- We are often interested in characteristics of sub-populations
- Sample sizes in a small sub-population are often too small for accurate estimation of their characteristics
- Small-area estimation is a branch of statistics devoted to estimating characteristics of sub-populations from data from a survey of a larger population
- Bayesian hierarchical models are often used to “borrow strength” from other sub-populations to improve sub-population estimates

Summary and Synthesis

- Modern statistical problems typically involve very large data sets, many parameters of direct interest, and many nuisance parameters
- Traditional methods assume a fixed number of parameters much smaller than the number of observations
- Hierarchical models provide theory and practical tools for statistical inference in very high-dimensional parameter spaces
- We examined a simple but powerful class of hierarchical models
  - Several groups of observations
  - Parameters for groups are independent draws from conjugate distribution
  - Population hyperparameters (parameters of conjugate distribution) govern distribution of group parameters