Statistics 225
Bayesian Statistical Analysis (Part 2)

Hal Stern

Department of Statistics
University of California, Irvine
sternh@uci.edu

March 28, 2019
Hierarchical models – motivation
James-Stein inference

- Suppose $X \sim N(\theta, 1)$
  - $X$ is admissible (not dominated) for estimating $\theta$ with squared error loss
- Now $X_i \sim N(\theta_i, 1), \ i = 1, \ldots, r$
  - $X = (X_1, \ldots, X_r)$ is admissible if $r = 1, 2$ but not $r \geq 3$
  - for $r \geq 3$
    $$\delta_i = (1 - \frac{r - 2}{\sum_i X_i^2})X_i$$
    yields better estimates
  - known as James-Stein estimation
Hierarchical models – motivation
James-Stein inference (cont’d)

- The Bayes view: $X_i \sim N(\theta_i, 1)$ and $\theta_i \sim N(0, a)$
  - posterior distn: $\theta_i | X_i \sim N$
  - posterior mean is $(1 - \frac{1}{a+1})X_i$
  - need to estimate $a$; one natural approach yields James-Stein

- Summary
  - estimation results depend on loss function
  - squared-error loss do well on avg but maybe poor for one component
  - powerful lesson about combining related problems to get improved inferences
Hierarchical Models

Suppose we have data

\[ Y_{ij} \quad j = 1, \ldots, J \]
\[ i = 1, \ldots, n_j \]

such that \( Y_{ij} \quad i = 1, \ldots, n_j \) are independent given \( \theta_j \) with distribution \( p(Y|\theta_j) \). e.g. \( \text{scores for students in classrooms} \) It

might be reasonable to expect \( \theta_j \)'s to be “similar” (but not necessarily identical).

Therefore, we may perhaps try to estimate population distribution of \( \theta_j \)'s. This is achieved in a natural way if we use a prior distribution in which the \( \theta_j \)'s are viewed as a sample from a common population distribution.
Hierarchical Models

- **Key:** The observed data, $y_{ij}$, with units indexed by $i$ within groups indexed by $j$, can be used to estimate aspects of the population distribution of the $\theta_j$'s even though the values of $\theta_j$ are not themselves observed.

- **How?** It is natural to model such a problem hierarchically
  - observable outcomes modeled conditionally on parameters $\theta$
  - $\theta$ given a probabilistic specification in terms of other parameters, $\phi$, known as *hyperparameters*. 
Hierarchical Models

- Nonhierarchical models are usually inappropriate for hierarchical data. Why?
  - a single $\theta$ (i.e., $\theta_j \equiv \theta \ \forall j$) may be inadequate to fit a combined data set.
  - separate unrelated $\theta_j$ are likely to “overfit” data.
  - information about one $\theta_j$ can be obtained from others’ data.

- Hierarchical model uses many parameters but population distribution induces enough structure to avoid overfitting.
Setting up hierarchical models
Exchangeability

Recall: A set of random variables \((\theta_1, \ldots, \theta_k)\) is **exchangeable** if the joint distribution is invariant to permutations of the indexes \((1, \ldots, k)\). The indexes contain no information about the values of the random variables.

- hierarchical models often use exchangeable models for the prior distribution of model parameters
- iid random variables are one example
- seemingly non-exchangeable r.v.’s may become exchangeable if we condition on all available information (e.g., regression analysis)
Setting up hierarchical models

Exchangeable models

- Basic form of exchangeable model
  - \( \theta = (\theta_1, \ldots, \theta_k) \) are independent conditional on additional parameters \( \phi \) (known as hyperparameters)

\[
p(\theta | \phi) = \prod_{j=1}^{k} p(\theta_j | \phi)
\]

- \( \phi \) referred to as hyperparameter(s) with hyperprior distn \( p(\phi) \)
- implies \( p(\theta) = \int p(\theta | \phi) p(\phi) d\phi \)
- work with joint posterior distribution, \( p(\theta, \phi | y) \)

- One objection to exchangeable model is that we may have other information, say \( (X_j) \). In that case may take

\[
p(\theta_1, \ldots, \theta_J | X_1, \ldots, X_J) = \prod_{i=1}^{J} p(\theta_i | \phi, X_i)
\]
Setting up hierarchical models

- Model is usually specified in nested stages
  - sampling distribution of data \( p(y|\theta) \)
    (first level of hierarchy)
  - prior (or population) distribution for \( \theta \) is \( p(\theta|\phi) \)
    (second level of hierarchy)
  - prior distribution for \( \phi \) (hyperprior) is \( p(\phi) \)
  - Note: more levels are possible
  - hyperprior at highest level is often diffuse but improper priors
    must be checked carefully to avoid improper posterior distributions.
Setting up hierarchical models

▸ Inference
▸ Joint distn:

\[ p(y, \theta, \phi) = p(y|\theta, \phi)p(\theta|\phi)p(\phi) = p(y|\theta)p(\theta|\phi)p(\phi) \]

▸ Posterior distribution

\[ p(\theta, \phi|y) \propto p(\phi)p(\theta|\phi)p(y|\theta) = p(\theta|y, \phi)p(\phi|y) \]

▸ often \( p(\theta|\phi) \) is conjugate for \( p(y|\theta) \)
▸ if we know (or fix) \( \phi \): \( p(\theta|y, \phi) \) follows from conjugacy
▸ then need inference for \( \phi \): \( p(\phi|y) \)
Computational approaches for hierarchical models

- Marginal model

\[ p(y|\phi) = \int p(y|\theta)p(\theta|\phi) d\theta \]

do inference only for \( \phi \) (e.g. marginal maximum likelihood)

- this is the approach that is often used in traditional random effects models
- no inference for \( \theta \)
Computational approaches for hierarchical models

- Empirical Bayes

\[ p(\theta|y, \hat{\phi}) \propto p(y|\theta)p(\theta|\hat{\phi}) \]

- estimate \( \phi \) (often using marginal maximum likelihood)
- inference for \( \theta \) conditional on the estimated \( \phi \)
- underestimates the uncertainty about \( \theta \)
Computational approaches for hierarchical models

- Hierarchical Bayes (a.k.a. full Bayes)

\[ p(\theta, \phi | y) \propto p(y | \theta)p(\theta | \phi)p(\phi) \]

inference for \( \theta \) and \( \phi \)
  - full posterior distribution of \( \theta \) and \( \phi \) is obtained
  - this is the approach we rely on
Hierarchical models and random effects
Animal breeding example

Consider the following mixed linear model commonly used in animal breeding studies

\[ Y = X\beta + Zu + e \]

\(X = \) design matrix for fixed effects
\(Z = \) design matrix for random effects
\(\beta = \) fixed effects parameters
\(u = \) random effects parameters
\(e = \) individual variation \(\sim N(0, \sigma^2_e I)\)

\[ Y | \beta, u, \sigma^2_e \sim N(X\beta + Zu, \sigma^2_e I) \]

\[ u | \sigma^2_a \sim N(0, \sigma^2_a A) \]

(can also think of \(\beta\) as random with \(p(\beta) \propto 1\))
Hierarchical models and random effects
Animal breeding example

- Marginal model (after integrating out $u$)

$$ Y|\beta, \sigma^2_a, \sigma^2_e \sim N(X\beta, \sigma^2_a ZAZ' + \sigma^2_e I) $$

- Note: the separation of parameters into $\theta$ and $\phi$ is somewhat ambiguous here:
  - model specification suggests $\phi = \{\sigma^2_a\}$ and $\theta = \{\beta, u, \sigma^e\}$
  - marginal model suggests $\phi = \{\beta, \sigma^2_a, \sigma^2_e\}$ and $\theta = \{u\}$
Hierarchical models and random effects
Animal breeding example

- Empirical Bayes (known as REML/BLUP)
  We can estimate $\sigma^2_a$, $\sigma^2_e$ by marginal
  (restricted?) maximum likelihood ($\hat{\sigma}^2_a$, $\hat{\sigma}^2_e$).
  Then
  \[
p(u, \beta|y, \hat{\sigma}^2_a, \hat{\sigma}^2_e) \propto p(y|\beta, u, \hat{\sigma}^2_e)p(u|\hat{\sigma}^2_a)
  \]
  (a joint normal distn)

- Hierarchical Bayes

  \[
p(\beta, \sigma^2_a, \sigma^2_e, \mu|y) \propto p(y|\beta, u, \sigma^2_e)P(u|\sigma^2_a)p(\beta, \sigma^2_a, \sigma^2_e)
  \]
Computation with hierarchical models

- Two cases
  - conjugate case ($p(\theta|\phi)$ conjugate prior for $p(y|\theta)$)
    - approach described below
  - non-conjugate case
    - requires more advanced computing
    - problem-specific implementations

- Computational strategy for conjugate case
  - write $p(\theta, \phi|y) = p(\phi|y)p(\theta|\phi, y)$
  - identify conditional posterior density of $\theta$ given $\phi$, $p(\theta|\phi, y)$ (easy for conjugate models)
  - obtain marginal posterior distribution of $\phi$, $p(\phi|y)$
  - simulate from $p(\phi|y)$ and then $p(\theta|\phi, y)$
Computation with hierarchical models
The marginal posterior distribution $p(\phi|y)$

- Approaches for obtaining $p(\phi|y)$
  - integration $p(\phi|y) = \int p(\theta, \phi|y)d\theta$
  - algebra - for a convenient value of $\theta$

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}$$

- Sampling from $p(\phi|y)$
  - easy if known distribution
  - grid if $\phi$ is low-dimensional
  - more sophisticated methods (later)
Normal-normal hierarchical model

- Data model
  - \( y_j | \theta_j \sim N(\theta_j, \sigma_j^2), j = 1, \ldots, J \) (indep)
  - \( \sigma_j^2 \)'s are assumed known for now
    (can release this assumption later)
  - motivation: \( y_j \) could be a summary statistic
    with (approx) normal distn from the \( j \)-th study
    (e.g., regression coefficient, sample mean)

- Prior distn
  - need a prior distn \( p(\theta_1, \ldots, \theta_J) \)
  - if exchangeable, then model \( \theta \)'s as iid given
    parameters \( \phi \)
Normal-normal hierarchical model: motivation

- Can think of this data model as a one-way ANOVA model (especially if $y_j$ is a sample mean of $n_j$ obs in group $j$). Typical ANOVA analysis begins by testing:

  $H_0 : \theta_1 = \ldots = \theta_J$
  $H_a : \text{not } H_0$

- If we don’t reject $H_0$, we might prefer to estimate each $\theta_j$ by the pooled estimate,

  $$\bar{y}_. = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} y_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$$

- If we reject $H_0$, we might use separate estimates, $\hat{\theta}_j = y_j$ for each $j$.

- Alternative: compromise between complete pooling and none at all, e.g., a weighted combination,

  $$\theta_j = \lambda_j y_j + (1 - \lambda)\bar{y}_.$$ where $\lambda_j \in (0, 1)$
Normal-normal hierarchical model

Constructing a prior distribution

(a) The pooled estimate $\hat{\theta} = \bar{y}$ is the posterior mean if the $J$ values $\theta_j$ are restricted to be equal, with a uniform prior density on the common $\theta$; i.e. $p(\theta) \propto 1$.

(b) The unpooled estimate $\hat{\theta}_j = y_j$ is the posterior mean if the $J$ values $\theta_j$ have independent uniform prior densities on $(-\infty, \infty)$; i.e. $p(\theta_1, \ldots, \theta_J) \propto 1$.

(c) The weighted combination is the posterior mean if the $J$ values $\theta_j$ are iid $N(\mu, \tau^2)$.

Note: (a) corresponds to (c) with $\tau^2 = 0$
(b) corresponds to (c) with $\tau^2 \to \infty$
Normal-normal hierarchical model

- Data model \( p(y_j|\theta_j) \sim N(\theta_j, \sigma_j^2), j = 1, \ldots, J \)
  \( \sigma_j^2 \)'s assumed known

- Prior model for \( \theta_j \)'s is normal (conjugate)

\[
p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2)
\]

i.e. \( \theta_j \)'s conditionally independent given \((\mu, \tau)\)

- Hyperprior distribution \( p(\mu, \tau) \)
  - noninformative distribution for \( \mu \) given \( \tau \), i.e., \( p(\mu|\tau) \propto 1 \)
    (this won’t matter much because the combined data from all \( J \) experiments are highly informative about \( \mu \))
  - more on \( p(\tau) \) later

\[
p(\mu, \tau) = p(\tau)p(\mu|\tau) \propto p(\tau)
\]
Normal-normal model: computation

- Joint posterior distribution:

\[ p(\theta, \mu, \tau | y) \]
\[ \propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta) \]
\[ \propto p(\tau) \prod_{j=1}^{J} N(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} N(y_j | \theta_j, \sigma_j^2) \]
\[ \propto p(\tau) \frac{1}{\tau^J} \exp \left[ -\frac{1}{2} \sum_{j} \frac{1}{\tau^2} (\theta_j - \mu)^2 \right] \exp \left[ -\frac{1}{2} \sum_{j} \frac{1}{\sigma_j^2} (y_j - \theta_j)^2 \right] \]

- Factors that depend only on \( y \) and \{\sigma_j\} are treated as constants because they are known
- Posterior distn is a distn on \( J + 2 \) parameters
- Can compute using MCMC (later) or
- Hierarchical computation:
  1. \( p(\theta_1, \ldots, \theta_J | \mu, \tau, y) \)
  2. \( p(\mu | \tau, y) \)
  3. \( p(\tau | y) \)
Normal-normal model: computation
Conditional posterior distn of $\theta$ given $\mu, \tau, y$

- Treat $(\mu, \tau)$ as fixed in previous expression
- Given $(\mu, \tau)$, the $J$ separate parameters $\theta_j$ are independent in their posterior distribution
- $\theta_j | y, \mu, \tau \sim N(\hat{\theta}_j, V_j)$ with

  \[
  \hat{\theta}_j = \frac{1}{\sigma_j^2} y_j + \frac{1}{\tau^2} \mu
  \]

  \[
  V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}
  \]

- Result from simple normal-normal conjugate analysis
- $\hat{\theta}_j$ is weighted average of hyperprior mean and data
Normal-normal model: computation
Marginal posterior distribution of $\mu, \tau$ given $y$

- We can analytically integrate the full posterior distn $p(\theta, \mu, \tau|y)$ over $\theta$
  
  $$p(\mu, \tau|y) = \int p(\theta, \mu, \tau|y) \, d\theta$$

- An alternative is to use the marginal model
  
  $$p(\mu, \tau|y) \propto p(y|\mu, \tau)p(\mu, \tau)$$

- Marginal model
  
  $$p(y|\mu, \tau) = \prod_{j=1}^{J} \int N(\theta_j|\mu, \tau)N(\bar{y}_j|\theta_j, \sigma_j^2) \, d\theta_j$$
  
  quadratic in $y_j$

  $$\Rightarrow \quad y_j|\mu, \tau \sim \text{Normal}$$

  $$E(y_j|\mu, \tau) = E(E(y_j|\theta_j, \mu, \tau)) = E(\theta_j) = \mu$$

  $$\text{Var}(y_j|\mu, \tau) = E(\text{Var}(y_j|\mu, \tau, \theta_j)) + \text{Var}(E(y_j|\mu, \tau, \theta_j))$$

  $$= E(\sigma_j^2) + \text{Var}(\theta_j) = \sigma_j^2 + \tau^2$$
Normal-normal model: computation
Marginal posterior distribution of $\mu, \tau$ given $y$

End result is

$$p(\mu, \tau | y) \propto p(\tau) \prod_{j=1}^{J} N(y_j | \mu, \sigma_j^2 + \tau^2)$$

$$\propto p(\tau) \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp \left( -\frac{(y_j - \mu)^2}{2(\sigma_j^2 + \tau^2)} \right)$$

Note: in non-normal models, it is not generally possible to integrate over $\theta$ and rely on the marginal model, so that more elaborate computational methods are needed
Normal-normal model: computation

Posterior distribution of $\mu$ given $\tau, y$

- Instead of sampling $(\mu, \tau)$ on a grid, factor the distribution:
  
  $$p(\mu, \tau|y) = p(\tau|y)p(\mu|\tau, y)$$

- $p(\mu|\tau, y)$ is obtained by looking at $p(\mu, \tau|y)$ and thinking of $\tau$ as known:
  
  $$\Rightarrow p(\mu|\tau, y) \propto \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2)$$

- This is the posterior distn corresponding to a normal sampling distribution with a noninformative prior density on $\mu$

- Result: $\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$ with
  
  $$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} y_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}}$$
Normal-normal model: computation

Posterior distribution of \( \tau \) given \( y \)

- \( p(\tau|y) \) can be found in two equivalent ways
  - integrate \( p(\mu, \tau|y) \) over \( \mu \)
  - use algebraic form \( p(\tau|y) = p(\mu, \tau|y)/p(\mu|\tau, y), \) which must hold for any \( \mu \)

- Choose the second option, and evaluate at \( \mu = \hat{\mu} \) (for simplicity):

\[
p(\tau|y) \propto \prod_{j=1}^{J} \frac{N(y_j|\hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)}
\]

\[
\propto V_{\mu}^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp \left( -\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right)
\]

- Note that \( V_\mu \) and \( \hat{\mu} \) are both functions of \( \tau \)
- Compute \( p(\tau|y) \) on a grid of values of \( \tau \)
Normal-normal model: computation

Summary

To simulate from joint posterior distribution \( p(\theta, \mu, \tau|y) \):

1. draw \( \tau \) from \( p(\tau|y) \) (grid approximation)
2. draw \( \mu \) from \( p(\mu|\tau, y) \) (normal distribution)
3. draw \( \theta = (\theta_1, \ldots, \theta_J) \) from \( p(\theta|\tau, y) \)
   (independent normal distribution for each \( \theta_j \))

Choice of \( p(\tau) \)

- \( p(\tau) \propto 1 \) - proper posterior distribution
- \( p(\log \tau) \propto 1 \) - improper posterior distribution
  (equivalent to \( p(\tau^2) \propto 1/\tau^2 \) but this common noninformative prior for variances doesn’t work in this case)
- discuss further on the next slide

Then illustrate with SAT coaching example (add to slides or do separately)
Normal-normal model: computation

Hyperprior distribution

- Non-informative or weakly informative prior distributions for $\tau$
  - $p(\tau) \propto 1$ - yields a proper posterior distribution ($J > 2$); can be thought of as limit of $U(0, A)$; sometimes useful to use $U(0, A)$ with $A$ determined by context of problem
  - $p(\log \tau) \propto 1$ - yields an improper posterior distribution; why??
    - this is a common noninformative prior for variances
    - here $1/\tau^2$ assigns infinite mass near $\tau = 0$ and the data can never rule out $\tau = 0$ because the $\theta_j$’s are not observable
    - can contrast with $\sigma^2$ in usual normal model where data (assuming all $y$’s are not equal) rules out $\sigma^2 = 0$
  - $p(\tau) = \text{inverse-gamma}(\epsilon, \epsilon)$ - proper prior distribution; but does not yield a proper posterior in the limit as $\epsilon \to 0$ so choice of $\epsilon$ matters
  - $p(\tau) \propto (1 + \tau^2/A^2\nu)^{-(\nu+1)/2}$ - known as half-t; distn of absolute value of a mean zero $t$ distribution with scale parameter $A$ and degrees of freedom $\nu$ (see Gelman 2006)
Beta-binomial example

- Series of toxicology studies
- Study $j$: $n_j$ exchangeable individuals $y_j$ develop tumors
- Model specification:
  - $y_j|\theta_j \sim \text{Bin}(n_j, \theta_j), j = 1, \ldots, J$ (indep)
  - $\theta_j, j = 1, \ldots, J | \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ (iid)
  - $p(\alpha, \beta)$ – to be specified later, hopefully ”non” or ”weakly” informative
- Marginal model:
  - can integrate out $\theta_j, j = 1, \ldots, J$ in this case

\[
p(y|\alpha, \beta) = \int \cdot \int \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1 - \theta_j)^{\beta-1} \binom{n_j}{y_j} \theta_j^{y_j}(1 - \theta_j)^{n_j - y_j} d\theta_1 \cdot d\theta_J
\]

\[
= \prod_{j=1}^{J} \binom{n_j}{y_j} \frac{\Gamma(\alpha + \beta) \Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + n_j)}
\]

- $y_j, j = 1, \ldots, J$ are ind
- distn of $y_j$ is known as beta-binomial distn
Beta-binomial example

- Conditional distn of $\theta$'s given $\alpha, \beta, y$
  - $p(\theta|\alpha, \beta, y) = \prod_j \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$
  - independent conjugate analyses
  - find this by algebra or by inspection of $p(\theta, \alpha, \beta|y)$
  - analysis is thus reduced to finding (and simulating from) $p(\alpha, \beta|y)$

- Marginal posterior distn of $\alpha, \beta$

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta) \Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + n_j)}$$

- could derive from marginal distn on previous slide
- could also derive from joint posterior distn
- not a known distn (on $\alpha, \beta$) but easy to evaluate
Beta-binomial example

- Hyperprior distn $p(\alpha, \beta)$
  - First try: $p(\alpha, \beta) \propto 1$ (flat, noninformative?)
  - equivalent to $p(\alpha/(\alpha + \beta), \alpha + \beta) \propto (\alpha + \beta)$
    (relevant because $\alpha/(\alpha + \beta)$ is the mean and $1/(\alpha + \beta)$ is roughly proportional to variance)
  - equivalent to $p(\log(\alpha/\beta), \log(\alpha + \beta)) \propto \alpha\beta$
  - check to see if posterior is proper
    - consider diff’t cases (e.g., $\alpha \to 0$, $\beta$ fixed)
    - if $\alpha, \beta \to \infty$ with $\alpha/(\alpha + \beta) = c$,
      then $p(\alpha, \beta|y) \propto$ constant (not integrable)
    - this is an improper distn
  - contour plot would also show this
    (lots of probability extending out towards infinity)
Beta-binomial example

- Hyperprior distn $p(\alpha, \beta)$
  - Second try: $p(\alpha/(\alpha + \beta), \alpha + \beta) \propto 1$
    (flat on prior mean and precision)
    - more intuitive, these two params are plausibly independent
    - equivalent to $p(\alpha, \beta) \propto 1/(\alpha + \beta)$
    - still leads to improper posterior distn
  - Third try: $p(\log(\alpha/\beta), \log(\alpha + \beta)) \propto 1$
    (flat on natural transformation of prior mean and variance)
    - equivalent to $p(\alpha, \beta) \propto 1/(\alpha \beta)$
    - still leads to improper posterior distn
  - Fourth try: $p(\alpha/(\alpha + \beta), (\alpha + \beta)^{-1/2}) \propto 1$
    (flat on prior mean and prior s.d.)
    - equivalent to $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$
    - "final answer" - proper posterior distn
    - equivalent to $p(\log(\alpha/\beta), \log(\alpha + \beta)) \propto \alpha \beta (\alpha + \beta)^{-5/2}$ (this will come up later)
Beta-binomial example

- **Computing**
  - later consider more sophisticated approaches
  - for now, use grid approach
    - simulate $\alpha, \beta$ from grid approx to posterior distn
    - then simulate $\theta$’s using conjugate beta posterior distn
  - convenient to use $(\log(\alpha/\beta), \log(\alpha + \beta))$ scale because contours ”look better” and we can get away with smaller grid
- Illustrate with rat tumor data (add slides or do separately?)