Hierarchical models

Anita Wang

2017 - 2018

Constructing a parameterized prior distribution

- we consider the problem of estimating a parameter θ using data from a small experiment and a prior distribution constructed from similar previous (or historical) experiments.
- Mathematically, we will consider the current and historical experiments to be a random sample from a common population.

Example

- In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents
- Suppose the immediate aim is to estimate θ, the probability of tumor in a population of female laboratory rats that receive a zero dose of the drug
- The data show that 4 out of 14 rats developed tumor
- It is natural to assume a binomial model for the number of tumors, given θ.
- For convenience, we select a prior distribution for θ from the conjugate family, $\theta \sim \text{Beta}(\alpha, \beta)$.

Analysis with a fixed prior distribution

- From historical data, suppose we knew that the tumor probabilities θ among groups of female lab rats follow an approximate beta distribution, with known mean and standard deviation.
- we could find values for α, β that correspond to the given values for the mean and standard deviation.
- assuming a Beta(α, β) prior distribution for θ yields a Beta(α + 4, β + 10) posterior distribution for θ.

Approximate estimate of the population distribution

- Typically, the mean and standard deviation of underlying tumor risks are not available.
- historical data are available on previous experiments on similar groups of rats.
- the historical data were in fact a set of observations of tumor incidence in 70 groups of rats. In the jth historical experiment, let the number of rats with tumors be y_j and the total number of rats be n_j.
- We model the y_j's as independent binomial data, given sample sizes n_j and study-specific means θ_j. Assuming that the beta prior distribution with parameters (α, β) is a good description
- of the population distribution of the θ_j 's in the historical experiments

Table: tumor incidence

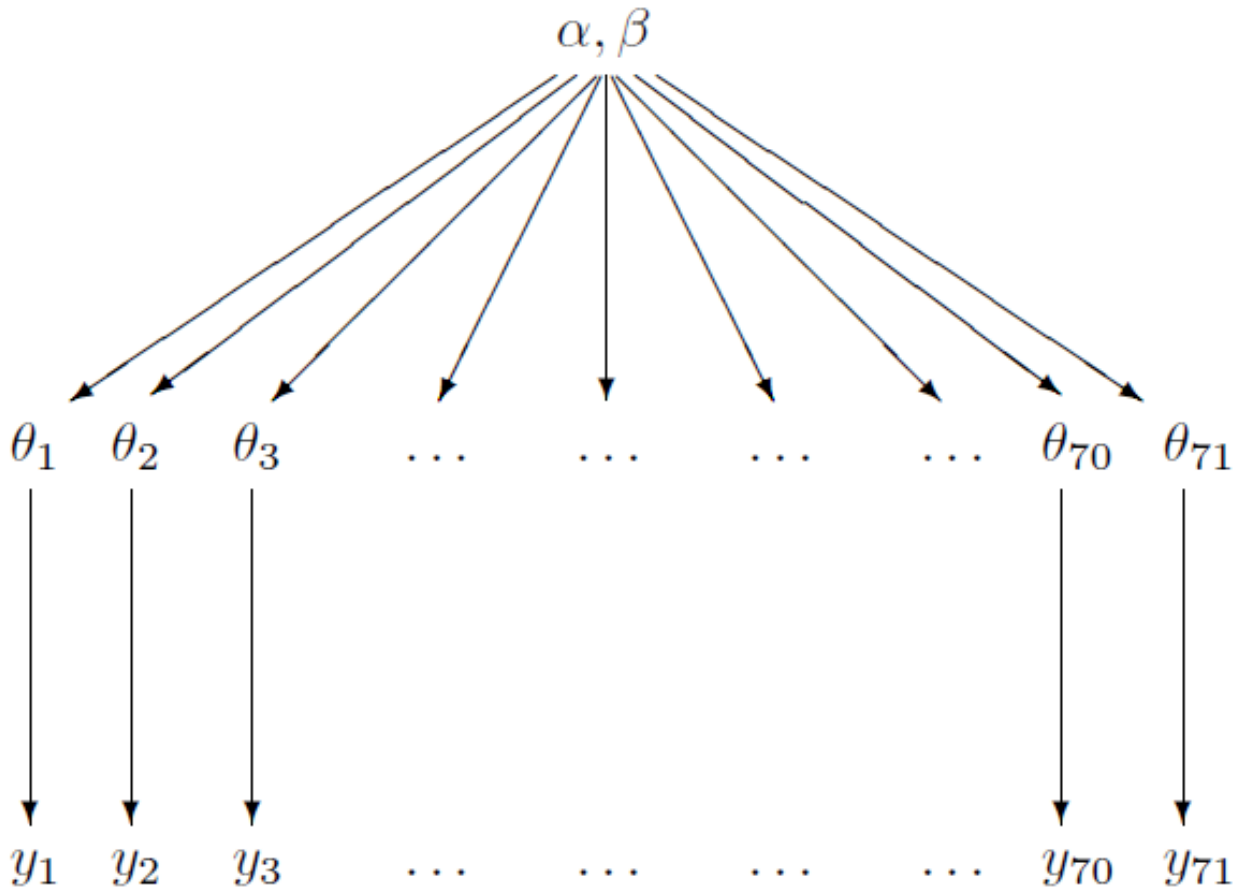
Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:

4/14

Structure of the hierarchical model



A starting point

- The observed sample mean and standard deviation of the 70 values $\frac{y_j}{n_j}$ are 0.136 and 0.103.
- If we set the mean and standard deviation of the population distribution to these values, we can solve for α and β. The resulting estimate for (α, β) is (1.4, 8.6).
- This is not a Bayesian calculation because it is not based on any specified full probability model. The estimate (1.4, 8.6) is simply a starting point from which we can explore the idea of estimating the parameters of the population distribution.

Exchangeability

- Generalizing from the example of the previous section, consider a set of experiments j = 1, . . . , J, in which experiment j has data (vector) y_j and parameter (vector) θ_j , with likelihood p(y_j |θ_j).
- A useful assumption in building models, if no information, other than the data y is available to distinguish any of the μ_j's from any of the others, and no ordering of grouping of the parameters can be made, one must assume symmetry among the parameters in the prior.

Exchangeability

- For example, in the rat tumor example, we have no prior reason to assume that $\mu_{70} < \mu_{71}$ is more likely than $\mu_{70} > \mu_{71}$. In fact, for the information given, the order that the groups are listed in is meaningless.
- So for this problem, it seems reasonable to have the distribution on the μ_j's be exchangeable, i.e. the distribution $p(\mu_1, \ldots \mu_J)$ should be invariant under permutations of the indices $(1, \ldots, J)$. If J = 3, then the distributions

$p(\theta_1, \theta_2, \theta_3), p(\theta_1, \theta_3, \theta_2), p(\theta_2, \theta_1, \theta_3), p(\theta_2, \theta_3, \theta_1),$
$p(\theta_3, \theta_1, \theta_2), p(\theta_3, \theta_2, \theta_1)$

are all of the same form.

Objections to exchangeable models

- it is natural to object to exchangeability on the grounds that the units actually differ.
- For example, the 71 rat tumor experiments were performed at different times, on different rats, and presumably in different laboratories. Such information does not, however, invalidate exchangeability
- Note that exchangeability does not imply independence.
- However all iid models are exchangeable.

Exchangeability

- One way of getting exchangeable distribution is to take a mixture of iid distributions.

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

- As ϕ is usually unknown, the distribution on μ must average over the uncertainty in ϕ.

$$p(\theta) = \int \left[\prod_{j=1}^{J} p(\theta_j|\phi) \right] p(\phi) d\phi$$

- All models of this form are exchangeable. To think of the μ_j's as draws from a superpopulation model that is determined by the hyperparameter ϕ.

Exchangeability

- One way of thinking of exchangeability is in terms non-informativeness or ignorance about the random variables.
- In the rat example, we have no preferences for different orderings of the theta's.
- In the rat problem, for example, the model like

$$\theta_i \sim Beta(\alpha, \beta) \text{ iid}$$
$$\alpha \sim U(0, 20)$$
$$\beta \sim U(0, 20)$$

Bayesian treatment of the hierarchical model

- the key 'hierarchical' part of these models is that φ is not known and thus has its own prior distribution, p(φ).
- Suppose we have the following hierarchical model

$$y|\theta,\phi \sim p(y|\theta)$$
$$\theta|\phi \sim p(\theta|\phi)$$
$$\phi \sim p(\phi)$$

- The joint prior distribution is

$$p(\varphi, \theta) = p(\varphi)p(\theta|\varphi),$$

and the joint posterior distribution is

$$p(\varphi, \theta|y) \propto p(\varphi, \theta)p(y|\varphi, \theta)$$
$$= p(\varphi, \theta)p(y|\theta),$$

The hyperprior distribution

- In order to create a joint probability distribution for (φ, θ), we must assign a prior distribution to φ.
- If little is known about φ, we can assign a diffuse prior distribution, but we must be careful when using an improper prior density to check that the resulting posterior distribution is proper
- In most real problems, one should have enough substantive knowledge about the parameters in φ
- In the rat tumor example, the hyperparameters are (α, β), which determine the beta distribution for θ.

Posterior predictive distributions

- There are two posterior predictive distributions that might be of interest to the data analyst:
  - the distribution of future observations $\tilde{y}$ corresponding to an existing θ_j
  - the distribution of observations $\tilde{y}$ corresponding to future θ_j 's ($\tilde{\theta}$) drawn from the same superpopulation.
- In the rat tumor example, future observations can be (1) additional rats from an existing experiment, or (2) results from a future experiment.

Bayesian analysis of conjugate hierarchical models

- we present an approach that combines analytical and numerical methods to obtain simulations from the joint posterior distribution, $p(\theta, \varphi|y)$
- the population distribution, $p(\theta|\varphi)$, is conjugate to the likelihood, $p(y|\theta)$.
- Three-step:
  1. $p(\varphi, \theta|y) \propto p(\varphi)p(\theta|\varphi)p(y|\theta, \varphi)$
  2. Conditional posterior
     $p(\theta|\varphi,y) \propto p(\theta|\varphi)p(y|\theta, \varphi)$
     This will be easy if a conjugate prior is used.
  3. Marginal posterior
     $p(\varphi|y) = \int p(\theta, \varphi|y)d\theta$. .

Bayesian analysis of conjugate hierarchical models

- For many standard models, the marginal posterior distribution of φ can be computed using the conditional probability,

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

- The difficulty is that the denominator $p(\theta|\varphi, y)$, a function of both θ and φ for fixed y, has a normalizing factor that depends on φ as well as y.
- One must be careful with the proportionality 'constant' in Bayes' theorem, especially when using hierarchical models, to make sure it is actually constant

Drawing simulations from the posterior distribution

- Simulating a draw from the joint posterior distribution, p(θ, φ|y):
  1. Sample $\phi_1, \ldots, \phi_m$ from p(φ|y)
  2. Sample $\theta_1, \ldots, \theta_m$ from p(θ|φ_i, y), given the drawn value of φ.
  3. If necessary, sample $\tilde{y}$. The form of this draw depends on whether the θ of interest is one corresponding to the dataset or a new one. It might be necessary first to draw a new value $\tilde{\theta}$ given φ.

  As usual, the above steps are performed L times in order to obtain a set of L draws.

Application to the model for rat tumors

- the data from experiments j = 1, . . . , J, J = 71, are assumed to follow independent binomial distributions:

$$y_j \sim \text{Bin}(n_j , \theta_j),$$

- The parameters θ_j are assumed to be independent samples from a beta distribution:

$$\theta_j \sim \text{Beta}(\alpha, \beta),$$

- we shall assign a noninformative hyperprior distribution to reflect our ignorance about the unknown hyperparameters.

Joint, conditional, and marginal posterior distributions

- The joint posterior distribution of all parameters is

$$p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta)$$

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

- Given (α, β), the components of θ have beta densities, and the joint density is

$$p(\theta|\alpha, \beta, y)$$

$$= \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

Joint, conditional, and marginal posterior distributions

- Marginal posterior

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

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

‘noninformative’ hyperprior distribution

- One reasonable choice of diffuse hyperprior density is uniform on $(\alpha/\alpha+\beta\ ,\ (\alpha+\beta)^{-1/2})$,

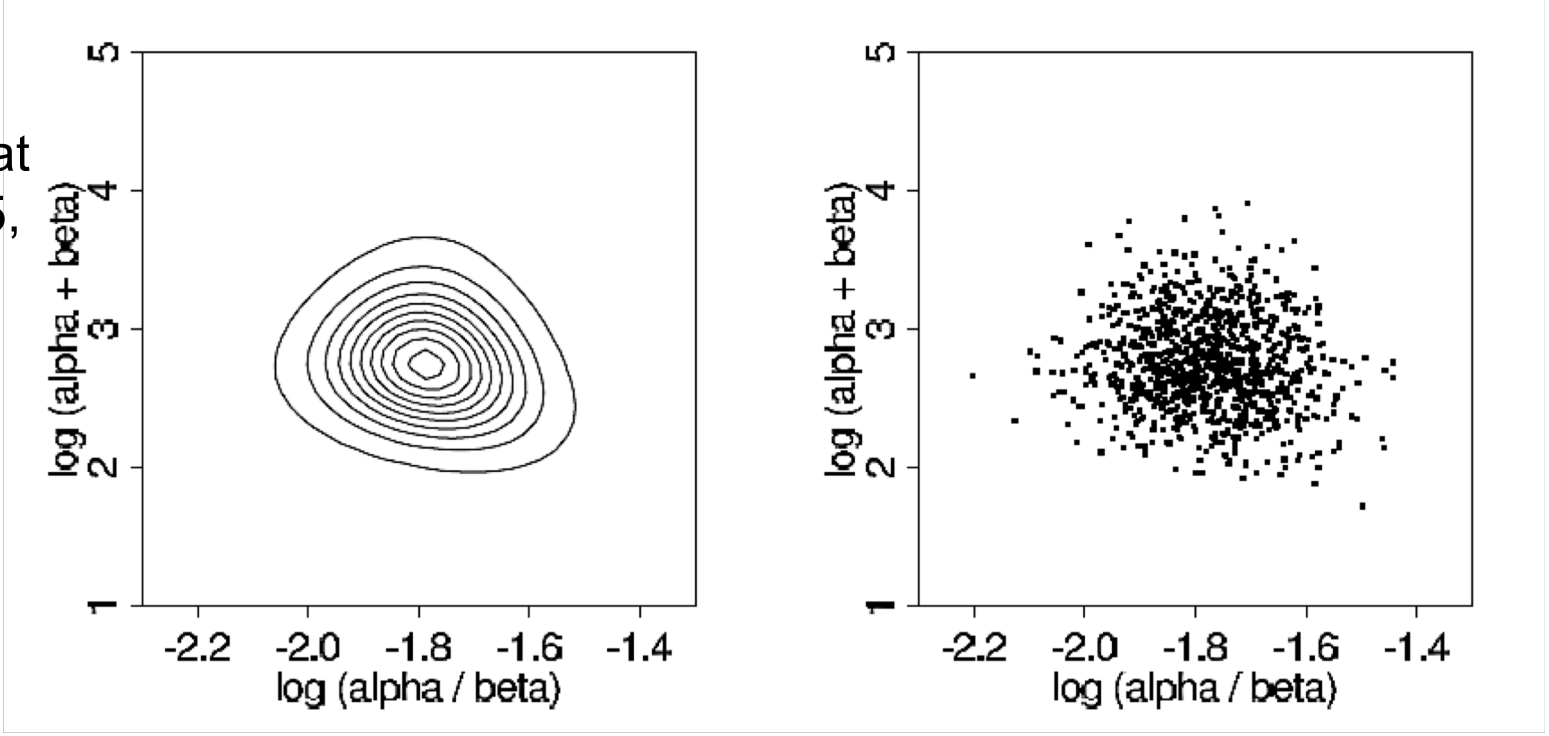
$$p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$$

- Before assigning a hyperprior distribution, we reparameterize in terms of $\text{logit}(\alpha/\alpha+\beta\ ) = \log(\alpha/\beta\ )$ and $\log(\alpha+\beta)$, which are the logit of the mean and the logarithm of the ‘sample size’ in the beta population distribution for θ.
- We use the logistic and logarithmic transformations to put each on a $(-\infty,\infty)$ scale.

the marginal posterior density of the hyperparameters

- the mode ((-1.75, 2.8) and (α, β) = (2.4, 14.0)) is not far from the point estimate (as we would expect)
- important parts of the marginal posterior distribution lie outside the range of the graph.

Contour lines are at 0.05, 0.15, . . . , 0.95 times the density at the mode



posterior moments

- E(α|y) is estimated by

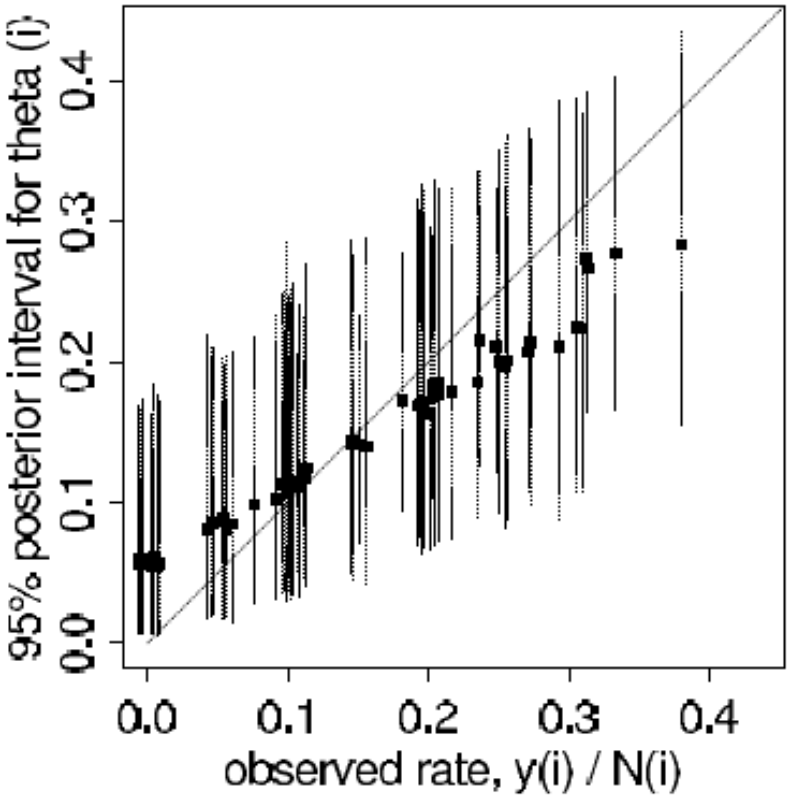
$$\sum_{\log\left(\frac{\alpha}{\beta}\right),\log(\alpha+\beta)} \alpha p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\,|y\right)$$

- we compute E(α|y) = 2.4 and E(β|y) = 14.3.

Sampling from the joint posterior distribution

- We draw 1000 random samples from the joint posterior distribution of $(\alpha, \beta, \theta_1, \ldots, \theta_J)$
  1. Simulate 1000 draws of $(\log(\alpha/\beta), \log(\alpha+\beta))$ from their posterior distribution displayed in Figure
  2. For $l = 1, \ldots, 1000$:
     a) Transform the lth draw of $(\log(\alpha/\beta), \log(\alpha+\beta))$ to the scale (α, β) to yield a draw of the hyperparameters from their marginal posterior distribution
     b) For each $j = 1, \ldots, J$, sample θ_j from its conditional posterior distribution, $\theta_j|\alpha, \beta, y \sim \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$.

Displaying the results



Posterior medians and 95% intervals of rat tumor rates, θ_j, (plotted vs. observed tumor rates y_j/n_j), based on simulations from the joint posterior distribution. The 45◦ line corresponds to the unpooled estimates, $\hat{\theta}= y_i/n_i$.

Thanks