ST 559 Project

Funfay Jen

2021-06-11

Exercise 5.9.5

Theorem 0.1 (Mixtures of independent distributions)**.** *Suppose the distribution of $\theta = (\theta_1, ..., \theta_J)$ can be written as a mixture of independent and identically distributed components:*

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

it follows that the covariances cov(θ_i, θ_j) are all nonnegative.

Before we embark on the proof, it is worthwhile to gain some intuition into the nature of the sought-after property. The purpose of this problem is to convince ourselves that even though the individual θ_j's are exchangeable, they are marginally correlated. Indeed, the exchangeability and independence in θ_i's is conditional on the hyperparameter ϕ. The reason for positive marginal correlation is that even though θ_i's are conditional independent with zero conditional covariance, there is another component of the covariance coming from the covariance of the conditional expectation of θ_i's, as opposed to θ_i's themselves. It is that piece of covariance that contributes positively, since the conditional expectation of θ_i are not independent as they are all conditioned on the same hyperparameter ϕ, and the process of averaging over θ_i "washes off" the independence. With this intuition, the proof follows rather straightforwardly. However, one lemma on the iterated covariance formula that parallels that of iterated variance will prove useful for the proof.

Lemma 0.2 (Law of Total Covariance)**.** *If X, Y, and Z are random variables on the same probability space, and the covariance of X and Y is finite, then*

$$\operatorname{cov}(X, Y) = \operatorname{E}(\operatorname{cov}(X, Y \mid Z)) + \operatorname{cov}(\operatorname{E}(X \mid Z), \operatorname{E}(Y \mid Z))$$

Proof of Lemma 0.2. The law of total covariance can be proved using the law of total expectation: First,

$$\operatorname{cov}(X, Y) = \operatorname{E}[XY] - \operatorname{E}[X] \operatorname{E}[Y]$$

from a simple standard identity on covariances. Then we apply the law of total expectation by conditioning on the random variable Z:

$$= \operatorname{E}\left[\operatorname{E}[XY \mid Z]\right] - \operatorname{E}\left[\operatorname{E}[X \mid Z]\right] \operatorname{E}\left[\operatorname{E}[Y \mid Z]\right]$$

Now we rewrite the term inside the first expectation using the definition of covariance:

$$= \operatorname{E}\big[\operatorname{cov}(X, Y \mid Z) + \operatorname{E}[X \mid Z] \operatorname{E}[Y \mid Z]\big] - \operatorname{E}\left[\operatorname{E}[X \mid Z]\right] \operatorname{E}\left[\operatorname{E}[Y \mid Z]\right]$$

Since expectation of a sum is the sum of expectations, we can regroup the terms:

$$= \mathrm{E}[\mathrm{cov}(X, Y \mid Z)] + \mathrm{E}[\mathrm{E}[X \mid Z]\,\mathrm{E}[Y \mid Z]] - \mathrm{E}[\mathrm{E}[X \mid Z]]\,\mathrm{E}[\mathrm{E}[Y \mid Z]]$$

Finally, we recognize the final two terms as the covariance of the conditional expectations $\mathrm{E}[X \mid Z]$ and $\mathrm{E}[Y \mid Z]$:

$$= \mathrm{E}\left[\,\mathrm{cov}(X, Y \mid Z)\right] + \mathrm{cov}\left(\,\mathrm{E}[X \mid Z], \mathrm{E}[Y \mid Z]\right)$$

■

Now we can use the lemma by recognizing θ_i as X, θ_j as Y, and Φ as Z. Thus:

Proof of Theorem 0.1.

$$\begin{aligned}
\mathrm{cov}(\theta_i, \theta_j) &= \mathrm{E}(\mathrm{cov}(\theta_i, \theta_j|\Phi)) + \mathrm{cov}(E(\theta_i|\Phi), \mathrm{E}(\theta_j|\Phi)) \\
&= 0 + \mathrm{cov}(\mathrm{E}(\theta_i|\Phi), \mathrm{E}(\theta_i|\Phi)) \\
&= \mathrm{var}(\mathrm{E}(\theta_i|\Phi)) \\
&\geq 0
\end{aligned}$$

■

where the second equality comes from the condtional independence and exchangeabiity.

We close this section by summarizing the moral of the story: the introduction of hierarchy in the model is one way to incorporate positive correlation between the parameters in different experiments, a reasonable assumption in many cases. A non-hierarchical model will necessarily imply that the parameters in different experiments are uncorrelated and independent. We now use such models to replicate the examples in chapter 5.

Rat Example

Data

We use the data on p. 102 which documents the rates of tumor incidence in historical control groups and current group of rates.

Inferential Questions

The immediate aim is to estimate θ, the probability of tumor in a population of rats. For this course project, we aim to compare hierarchical and non-hierarchical approaches, and see where in the spectrum hierarchical models fit.

Models

We explore three models: separate models for each experiment, complete pooling, and hierarchical models. We will see that the hierarchical models lie in between the two other extremes. In all three kinds of models, there are some shared modeling components as follows:

The data from experiments $j = 1, \ldots, J, J = 71$, are assumed to follow independent binomial distributions: $y_j \sim \mathrm{Bin}(n_j, \theta_j)$, with the number of rats, n_j , known. The parameters θ_j are assumed to be independent samples from a beta distribution: $\theta_j \sim \mathrm{Beta}(\alpha, \beta)$, Now we highlight the differences in them.

Pooled Model We can assume that all experiments have the same success probability, θ. This means that $\theta_1 = \ldots = \theta_J = \theta$. In other words, it's the added constraint that makes it a pooled model. The prior distribution on θ can be anything, but in this project we can assume it to be uniform, i.e., Beta(1,1).

Separate Models To have a separate model for θ for each experiment, simply assume a fixed prior distribution on θ. That means that we do not need to any any complexities to the shared model structures described earlier. Like in the pooled model, we can assume the prior on θ to be uniform, i.e., Beta(1,1). By having a fixed prior, we have a fixed pair of parameters (α, β), or in the notation of the theoretical exercise 5.9.5, a fixed Φ, which ensures that the correlation between θ_i and θ_j is 0 for different i and j.

Before we describe the hierarchical modeling, we note that we can also replace the Beta(1,1) distribution by Beta(1.4, 8.6) in both the pooled model and separate models which comes from estimating (α, β) (see explanations on p. 103). We are not exploring this option here since the hierarchical modeling is more powerful.

Hierarchical Models To have a hierarchical model, that is to induce some correlation between experiments, but not so strong as to degenerate into a pooled model, we invoke hierarchy. That is, we assume the prior distribution itself follows a distribution. As worked out on p. 110, we assign a noninformative hyperprior distribution to reflect our ignorance about the unknown hyperparameters. Thus, we assume a non-informative hyperprior density given by

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

In this way, as worked out by exercise 5.9.5, we are modeling a positive correlation between the success probabilities in distinct experiments.

Our models are now complete.

Simulation/Computation The posterior simulations of the pooled model and the separate models are straightforward because we can leverage our knowledge about Beta-Binomial conjugacy. The posterior simulation for the hierarchical models is more involved due to the variability in the hyperparameters. The following formula will prove useful in pinning the hyperparameters down.

Posterior Distributions of (α, β) in the Hierarchical Models

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

The posterior $p(\alpha, \beta|y)$ is useful because for a fixed pair of (α, β), we know the properties of the model very well: it is as if we have separate models, and we can calculate quantities such as posterior probability densities and posterior quantiles of θ, which is the parameter of interest. Any of these quantities can be regarded as a *function* of (α, β), and we can obtain posterior expectations of such quantities by weighting them with $p(\alpha, \beta|y)$. Computationally, we just need to sample (α, β) from the posterior $p(\alpha, \beta|y)$ and perform a simple averaging, making it easier to compute quantities about the parameters of interest.

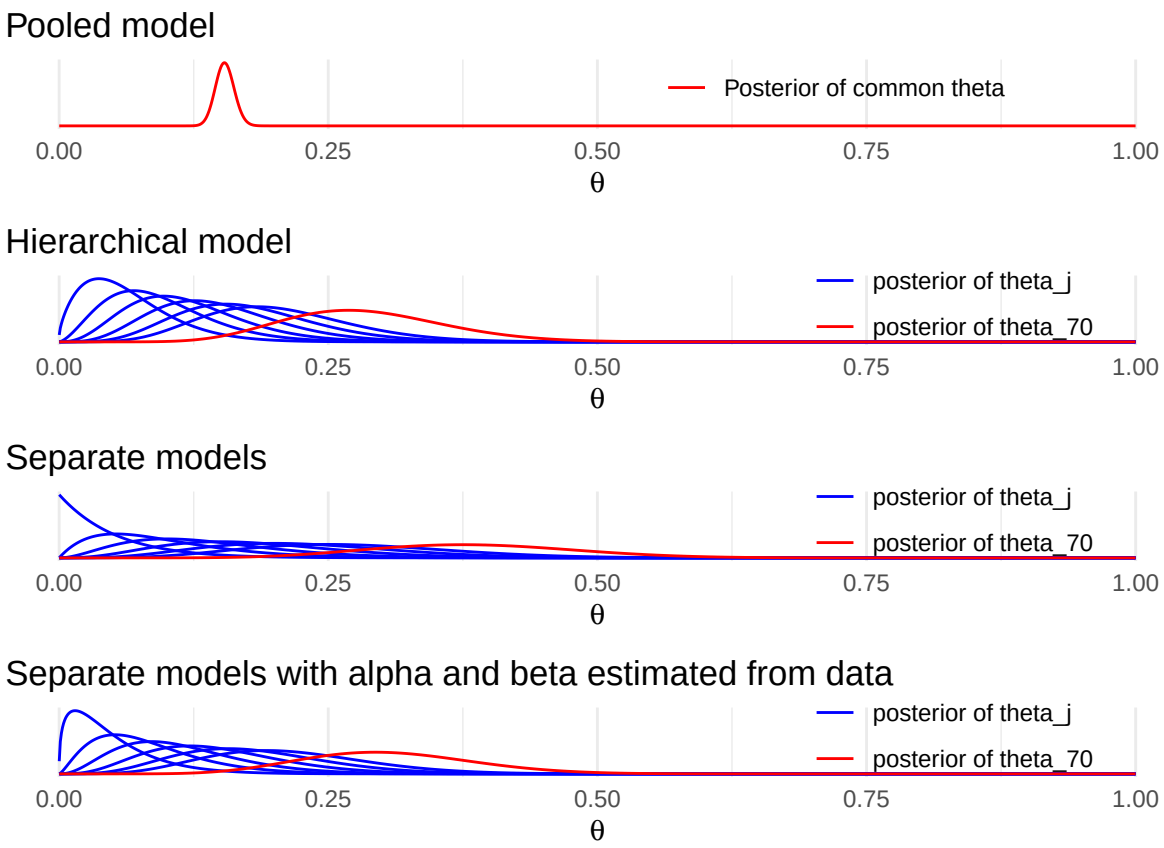
On the other hand, an important point to note is that the posterior probability densities and posterior quantiles of θ can also be found "directly" by obtaining the marginal posterior density of *theta*; however, they can be harder to derive. One can in fact show that taking the expectations of the conditional (conditioned on (α, β)) posterior density with respect to $p(\alpha, \beta|y)$ is equivalent to the "direct" way via finding the marginal posterior density of *theta*. However, the posterior quantiles of θ cannot be exactly obtained via the same method. That is, the quantiles of the posterior distribution of θ is NOT the same as the *expected* quantiles. We nonetheless use those quantities due to the computational convenience, hoping they are not far off from the exact values. Even if they are far off, we argue that either method of obtaining quantiles for θ is reasonable, just that the expected quantiles might be slightly less common than quantiles of expectations (the "direct" way).

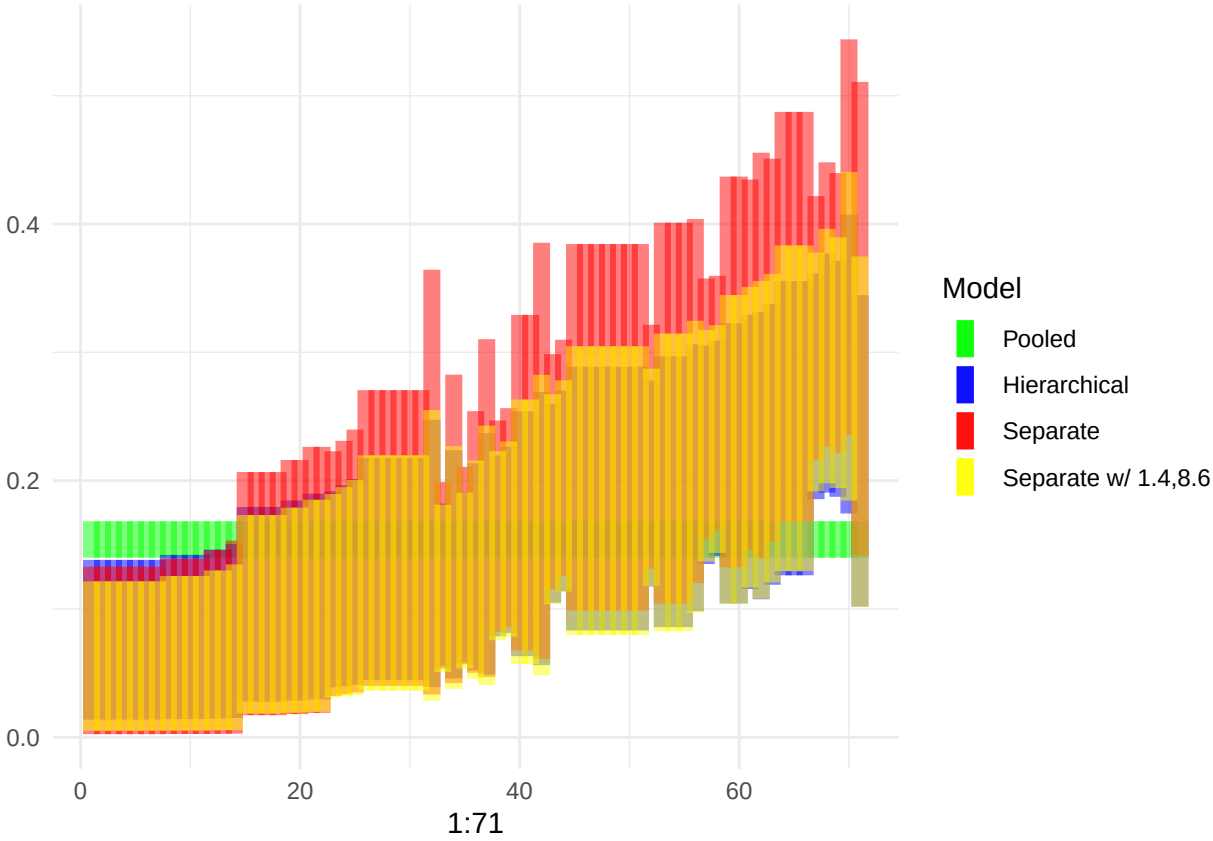
Findings

We plot the posterior distribution of θ for the pooled, hierarchical, and separate models in order (each curve represents an individual experiment, and to make the plot less cluttered, we only plot every 10th experiment for the separate and hierarchical models). We see that the pooled model yields a distribution that is the most constrained. However, this overconfidence might be unwarranted, since there might not be any guarantee that the true success rates are the same across experiments. The separate model yields distributions that have the most uncertainty for each experiment, which is due to the fact that the sample size for is smaller when each experiment is considered separately than all of them lumped together. The hierarchical models yield a middleground, because there is some moderate amount of pooling of information across experiments when computing/simulating the posterior hyperparameters (α, β), which narrows down our inference, but definitely not as much as the pooled model, which pools information even more strongly.

Note that our findings are in contradiction with the explanations on p. 113 in the text. One reason for this is because we have used a uniform (Beta(1,1)) prior for the separate model, whereas the text used the Beta(1.4, 8.6) prior, which will yield less uncertainty. This is confirmed with the amended graph. However, we note that the hierarchical model still shows less uncertainty. It's possible that this might be due to the fact that the quantiles are not computed exactly as mentioned earlier; but we do believe that even exact quantiles should be narrower than the separate models. At least in this case, the posterior draws of (α, β) are larger than (1.4, 8.6), which makes the intervals narrower. Whether this always happens needs to be studied further. (We guess yes because hierarchical models represent some positive correlation, which is less than full correlation as in the pooled model, and more than no correlation at all as in the separate models.)

Lastly, we plot the posterior 95% credible intervals for each experiment, where the indices are labeled from 1 to 71, corresponding to the experiments.




0.4
0.2
0.0
0
20
40
60
1:71
Model
Pooled
Hierarchical
Separate
Separate w/ 1.4,8.6