Group Comparisons and Hierarchical Modeling

A common problem in statistics is the comparison of populations, especially population means.

We begin by considering the comparison of two population means and then generalize to any number of means.

Suppose we have independent random samples from two different populations. Initially we assume

$$Y_{i1} \sim N(\mu + \delta, \sigma^2), \quad i = 1, \ldots, n_1,$$

$$Y_{i2} \sim N(\mu - \delta, \sigma^2), \quad i = 1, \ldots, n_2.$$

In this parameterization, μ represents the average of the two population means, and δ represents how much each population mean differs from this average.

A convenient prior for the three parameters is

$$p(\mu, \delta, \sigma^2) = p_1(\mu)p_2(\delta)p_3(\sigma^2),$$

where p_1 is $N(\mu_0, \gamma_0^2)$, p_2 is $N(\delta_0, \tau_0^2)$ and p_3 is inverse-gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$.

This prior leads to known full conditionals for each of the three parameters. The full conditional for each of μ and δ is normal, and that of σ^2 is inverse-gamma.

This knowledge allows exploration of the posterior via the Gibbs sampler.

In the case where we compare more than two population means, we'll use a *hierarchical model*.

First, let's consider a general case where we want to compare the parameters of m different populations.

The levels of the hierarchy are as follows:

- A prior distribution, call it p, for a parameter $\boldsymbol{\psi}$.

- A between-groups level wherein

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m | \boldsymbol{\psi} \sim \text{ i.i.d. } p(\boldsymbol{\theta}|\boldsymbol{\psi}).$$

- A within-groups level such that

$$Y_{1,j}, \ldots, Y_{n_j,j} | \boldsymbol{\theta}_j \sim \text{ i.i.d. } f(y|\boldsymbol{\theta}_j).$$

  for $j = 1, \ldots, m$.

One version of the hierarchical normal model assumes that we have independent random samples from m different normal populations, all of which have the same variance σ^2.

The previous hierarchical model is assumed with

$$f(y|\theta_j, \sigma^2) \equiv N(\theta_j, \sigma^2).$$

In this model only the mean parameter varies across groups, and so we have only a distribution for $\theta_j|\psi$, which is

$$p(\theta|\psi) \equiv N(\mu, \tau^2),$$

where $\psi = (\mu, \tau^2)$.

Finally, we need a prior for σ^2, μ and τ^2. We assume that these parameters are a priori independent with

$$1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2),$$

$$1/\tau^2 \sim \text{gamma}(\eta_0/2, \eta_0\tau_0^2/2)$$

and

$$\mu \sim N(\mu_0, \gamma_0^2).$$

The posterior has the form

$$p(\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2 | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \propto$$

$$\prod_{j=1}^{m} \prod_{i=1}^{n_j} f(y_{i,j} | \theta_j, \sigma^2) \prod_{j=1}^{m} p(\theta_j | \mu, \tau^2)$$

$$\times p_1(\sigma^2) p_2(\mu) p_3(\tau^2).$$

This result depends on the assumption that the conditional distribution of the data given $\theta_1, \ldots, \theta_m$, σ^2, μ and τ^2 *is the same as that of the data given* $\theta_1, \ldots, \theta_m$ *and* σ^2.

To apply Gibbs sampling we need the full conditionals. This is possible to do given the form of the normal hierarchical model and the prior chosen for (σ^2, μ, τ^2).

Let's derive the full conditionals. Let $N = \sum_{j=1}^{m} n_j$.

$$p(\theta_1, \ldots, \theta_m, \sigma^2, \mu, \tau^2 | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \propto$$

$$\sigma^{-N} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{i,j} - \theta_j)^2 \right) \right]$$

$$\times \tau^{-m} \exp\left(-\frac{1}{2\tau^2} \sum_{j=1}^{m} (\theta_j - \mu)^2 \right)$$

$$\times (\sigma^2)^{-\nu_0/2-1} e^{-\nu_0 \sigma_0^2/(2\sigma^2)}$$

$$\times (\tau^2)^{-\eta_0/2-1} e^{-\eta_0 \tau_0^2/(2\tau^2)}$$

$$\times \exp\left(-\frac{1}{2\gamma_0^2} (\mu - \mu_0)^2 \right).$$

By inspection, we can recognize that the full conditional of σ^2 is inverse-gamma with parameters $(N + \nu_0)/2$ and

$$\frac{1}{2}\left[\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2 + \nu_0\sigma_0^2\right].$$

Likewise the full conditional of τ^2 is inverse-gamma with parameters $(m + \eta_0)/2$ and

$$\frac{1}{2}\left[\sum_{j=1}^{m}(\theta_j - \mu)^2 + \eta_0\tau_0^2\right].$$

The basic trick in deriving the full conditionals of $\theta_1, \ldots, \theta_m$ and μ is "completing the square."

Notice that the only parts of the posterior that depend on μ are the red piece and the last purple one.

The product of these two pieces, as a function of μ, is proportional to

$$\exp\left[-\frac{1}{2}\left(\frac{m}{\tau^2}(\mu^2 - 2\mu\bar{\theta}) + \frac{1}{\gamma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right],$$

where $\bar{\theta} = \sum_{j=1}^{m} \theta_j/m$. Completing the square, the last quantity is proportional to

$$\exp\left[-\frac{1}{2\tilde{\sigma}^2}(\mu - \tilde{\mu})^2\right],$$

where

$$\tilde{\mu} = \left(\frac{m\bar{\theta}}{\tau^2} + \frac{\mu_0}{\gamma_0^2}\right)\left(\frac{m}{\tau^2} + \frac{1}{\gamma_0^2}\right)^{-1}$$

and

$$\tilde{\sigma}^2 = \left(\frac{m}{\tau^2} + \frac{1}{\gamma_0^2}\right)^{-1}.$$

It follows that the full conditional of μ is $N(\tilde{\mu}, \tilde{\sigma}^2)$.

Using the same technique, the full conditional of θ_j (for each j) is normal with mean

$$\left(\frac{n_j \bar{y}_j}{\sigma^2} + \frac{\mu}{\tau^2}\right)\left(\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

and variance

$$\left(\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}.$$

Example 15 Pain threshold and hair color

Studies conducted at the University of Melbourne indicate that there may be a difference between the pain thresholds of blonds and brunettes.

Men and women of various ages were divided into four categories according to hair color: light blond, dark blond, light brunette, and dark brunette.

The purpose of the experiment was to determine whether hair color is related to the amount of pain produced by common types of mishaps and assorted types of trauma.

Each person in the experiment was given a pain threshold score based on his or her performance in a pain sensitivity test (the higher the score, the higher the persons pain tolerance).

Data Summary

	Light Blond	Dark Blond	Light Brunette	Dark Brunette
Mean	59.2	51.2	42.5	37.4
SD	8.5	9.3	5.4	8.3
n	5	5	4	5

We'll analyze the data using our hierarchical normal model. I'd like the priors to be more or less noninformative.

I therefore want η_0, τ_0, ν_0 and σ_0 to be small. I'll take each of these to be 0.5.

To make the prior for μ noninformative, we should take γ_0 to be large. I'll use $\gamma_0 = 500$ and

$$\mu_0 = \frac{1}{19} \sum_{j=1}^{4} n_j \bar{y}_j = 47.84.$$

I used Gibbs sampling with $100,000$ iterations.

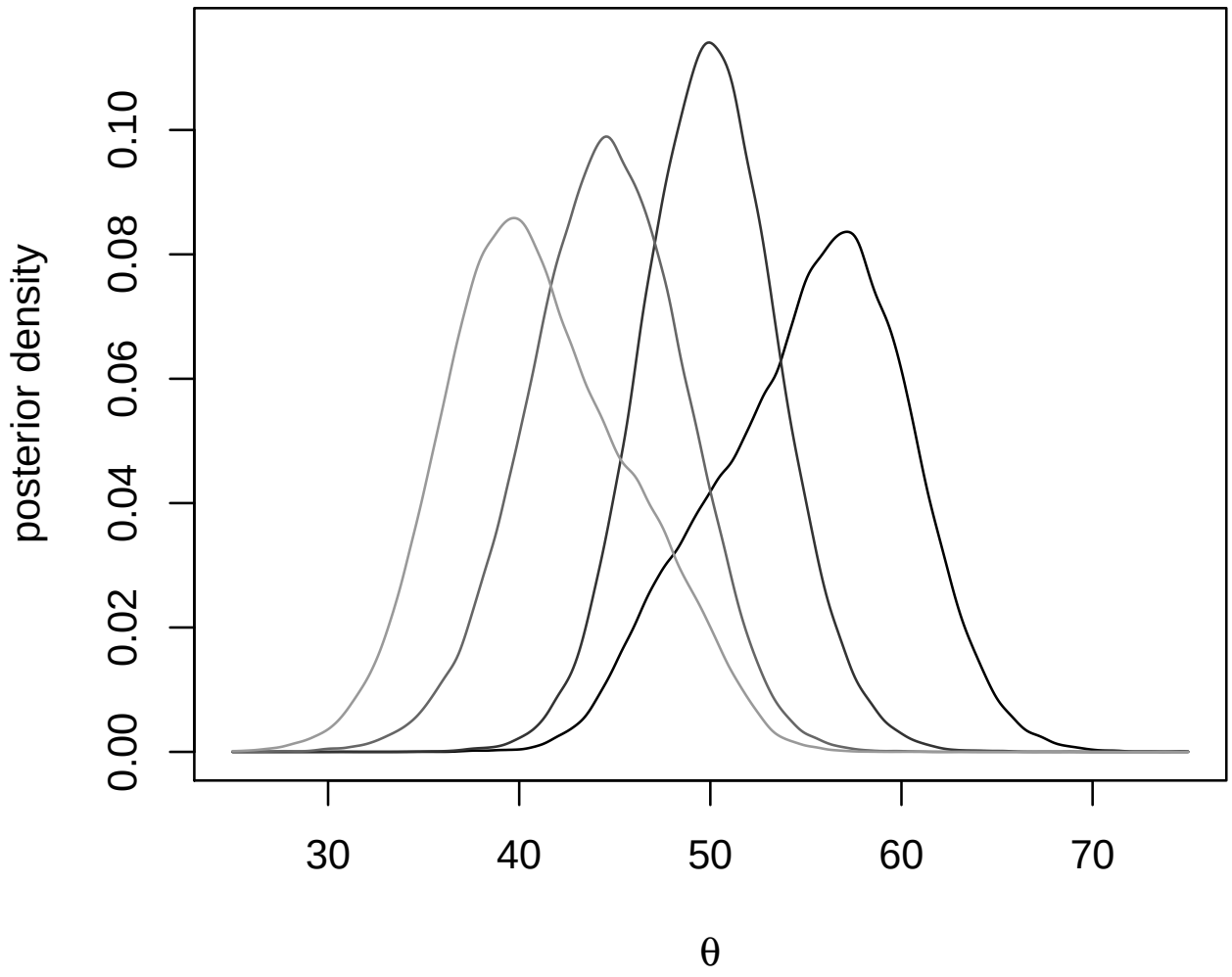
Statistics from Gibbs Output

	θ_1	θ_2	θ_3	θ_4
Mean	55.2	50.0	44.5	41.0
SD	4.97	3.58	4.09	4.80

Note that each $\widehat{\theta}_j$ is closer to the overall mean than is $\bar{y}_j$. This is called shrinkage.

	σ	μ	τ
Median	8.77	47.7	7.08
IQR	2.66	5.63	6.52

Kernel estimates of posterior densities of θ_js



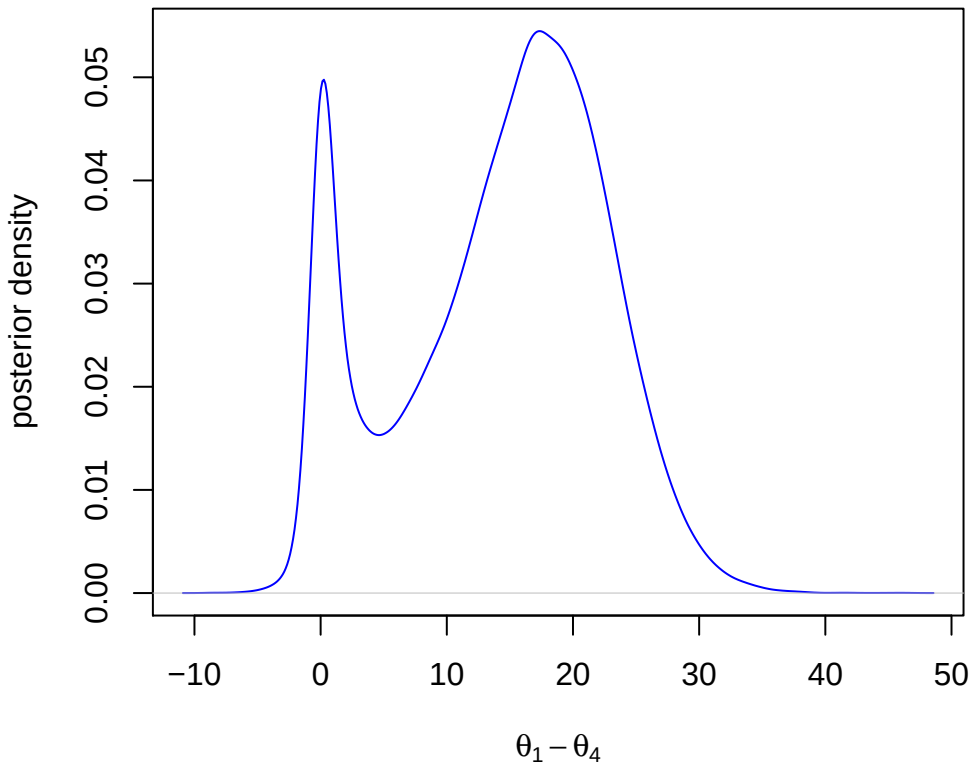
The modes, from right to left, correspond to θ_1, θ_2, θ_3 and θ_4, respectively.

Do light blondes tolerate pain better than dark brunettes, i.e., is the true θ_1 larger than the true θ_4?

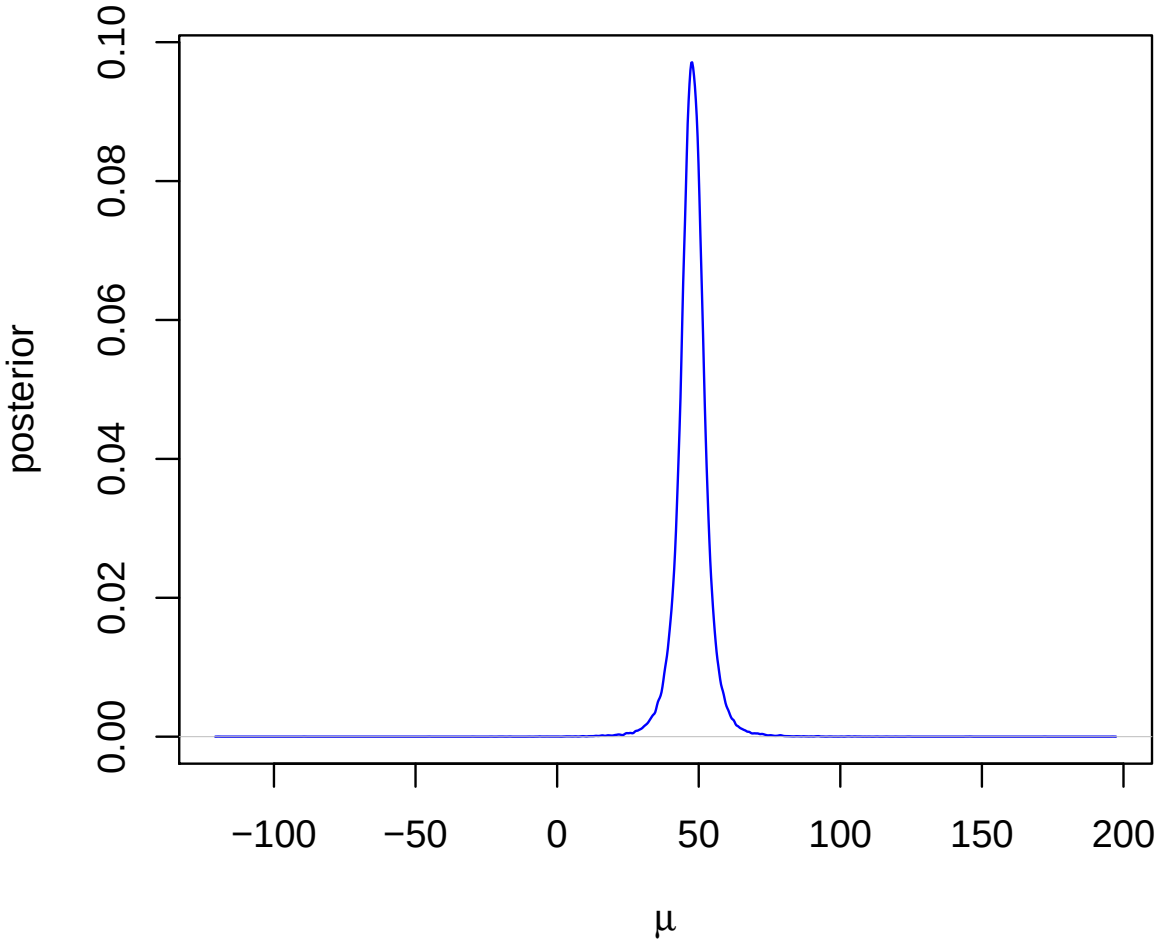
Out of 100,000 observations, $\theta_1 > \theta_4$ 96,230 times, and so

$$P(\theta_1 > \theta_4 | \boldsymbol{y}) \approx 0.9623.$$

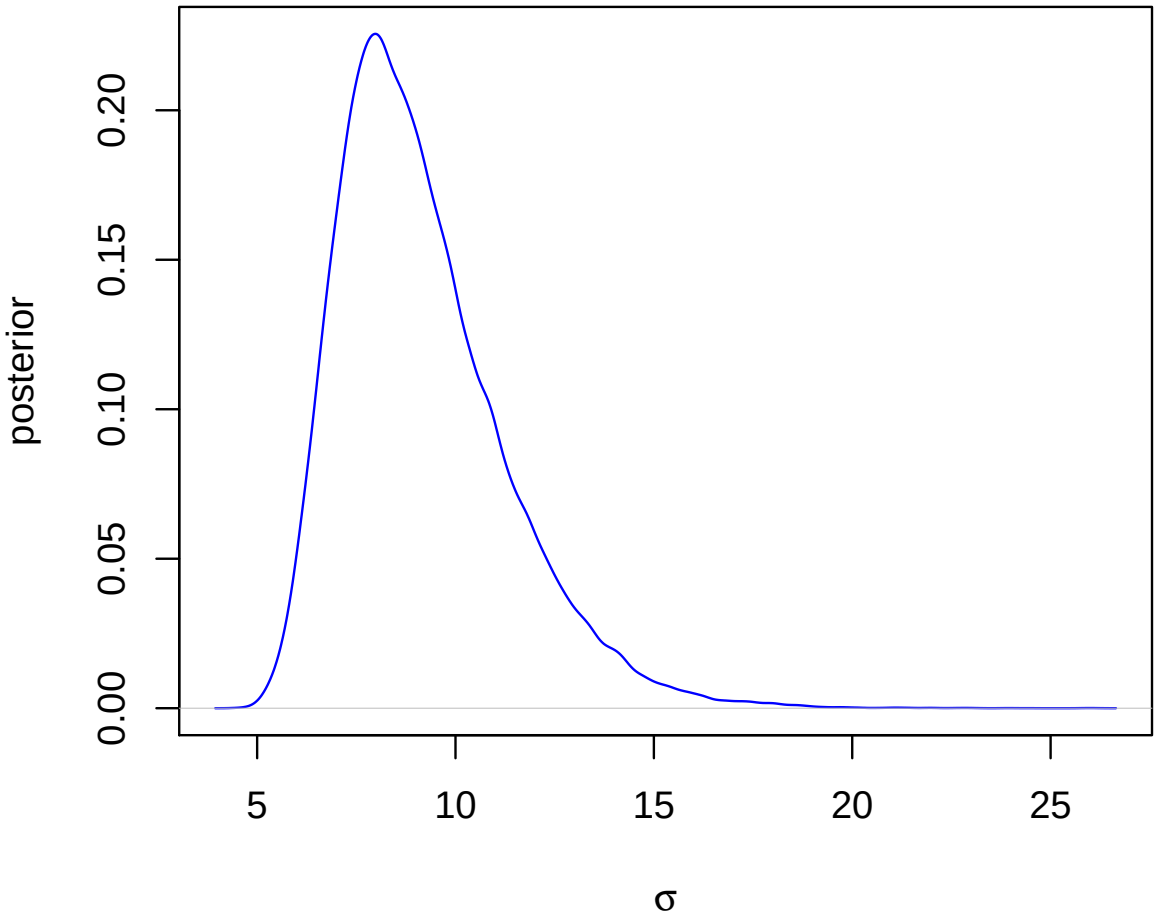
Kernel estimate of posterior density of $\theta_1 - \theta_4$



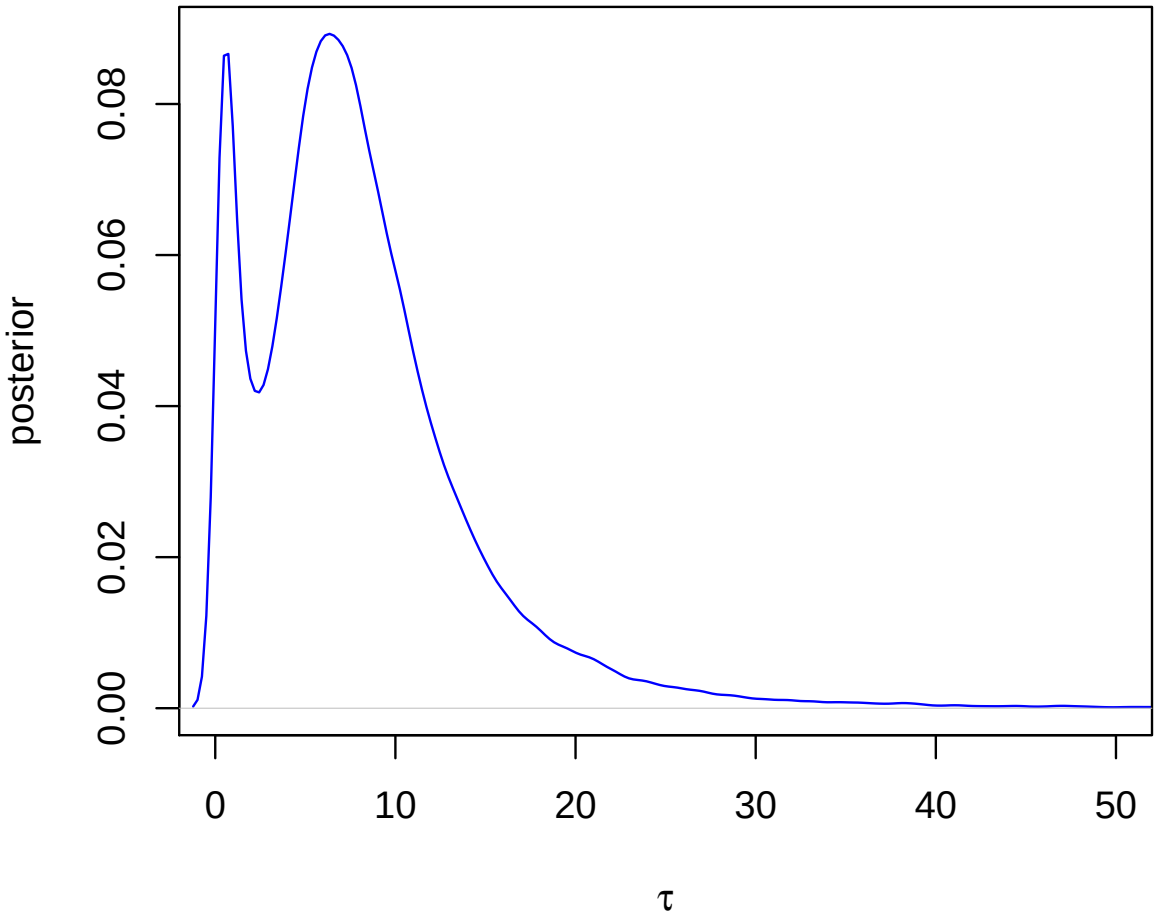
Kernel estimate of posterior of μ



Kernel estimate of posterior of σ



Kernel estimate of posterior of τ



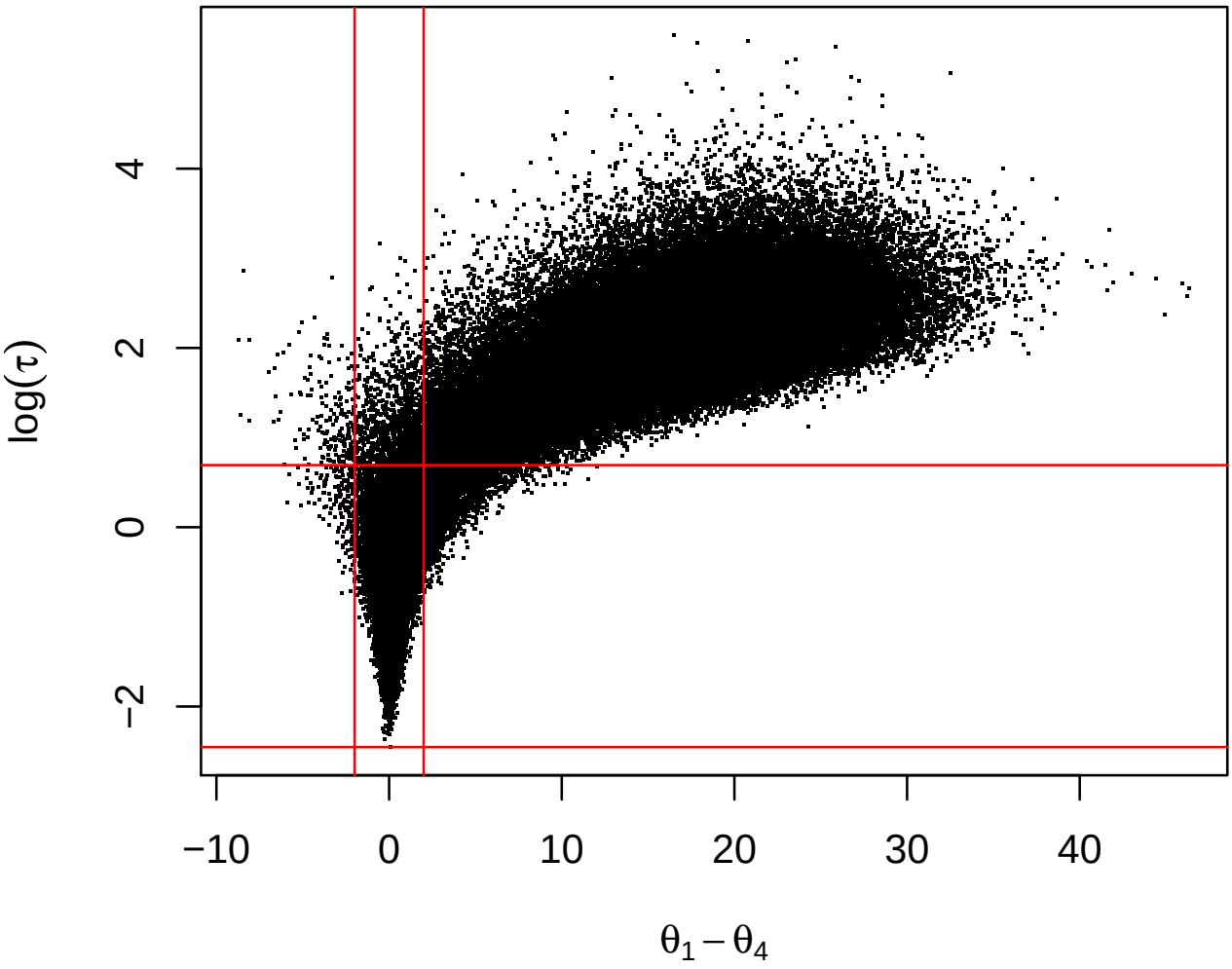
The distribution of τ is important because *small values of τ suggest that there is little difference between the θ_js.*

Note that the posterior on the previous page is somewhat conflicted. On the one hand it says there is about a 68% chance that τ is larger than 5.

But the mode at 0.6 suggests that τ *just might be fairly small.*

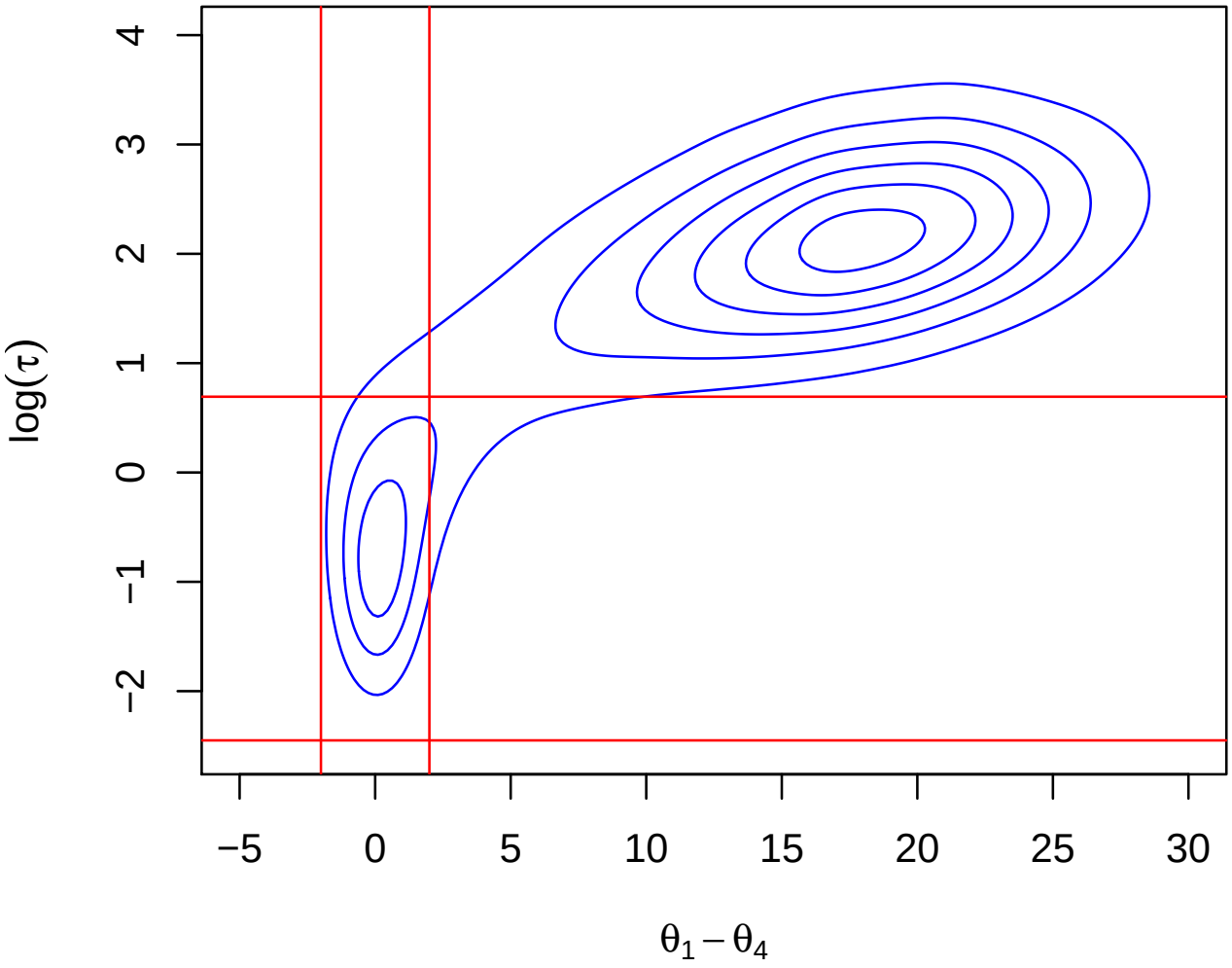
Not surprisingly, the mode at 0.6 in the distribution of τ is responsible for the mode at 0 in the distribution of $\theta_1 - \theta_4$. (See the graphs on the next two pages.)

Scatterplot of $\log(\tau)$ *vs.* $\theta_1 - \theta_4$

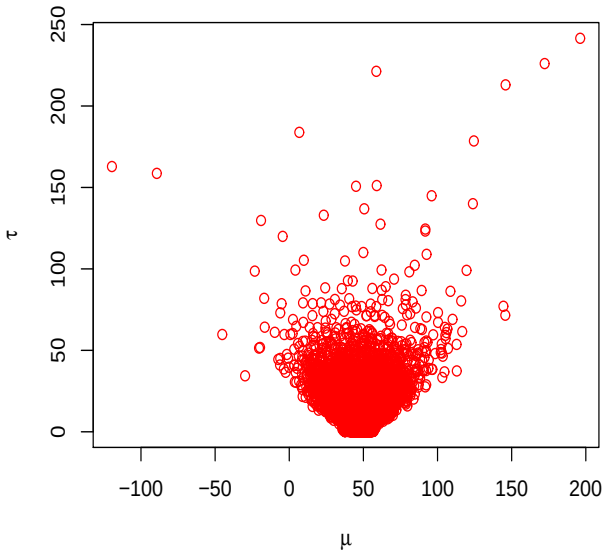


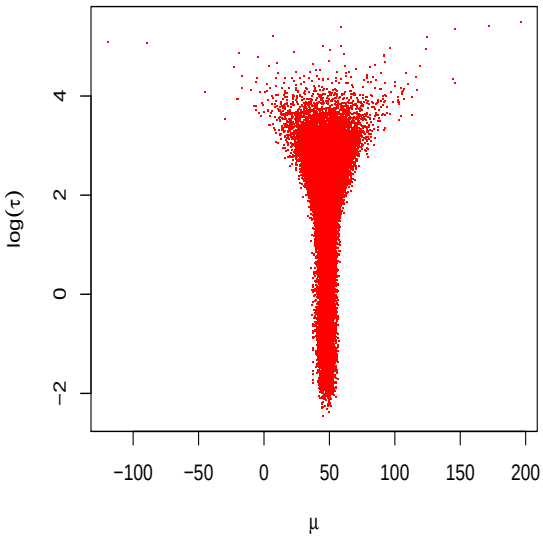
This graph and the one on the next page show that there is a close correspondence between the smaller modes in the graphs on pp. 195N and 198N.

Bivariate kernel estimate for data on previous page



Scatterplots for Gibbs sampling output: τ *vs.* μ *and* $\log(\tau)$ *vs.* μ





The distribution of μ is definitely long-tailed relative to the normal.

Verify this using the fact that the median and interquartile range of the μ distribution are 47.7 and 5.63, respectively.

An analysis of the output showed that each autocorrelation function decreased monotonically with lag, and no first lag autocorrelation was larger than 0.6.

It turns out that kernel density estimates are *highly robust to this level of autocorrelation,* and so the kernel estimates are essentially as reliable as they would be if we had 100,000 *independent* observations.

See

Hart, J.D. (1996). Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* **6** 115-142.

Empirical Bayes and the normal hierarchical model

To illustrate the crux of the empirical Bayes idea, we assume that the data are structured as on pp. 184-85N, *except that σ^2 is known*.

In this case, $\bar{Y}_1, \ldots, \bar{Y}_m$ are sufficient statistics, and their joint distribution is

$$p(\bar{y}_1, \ldots, \bar{y}_m | \theta_1, \ldots, \theta_m) \propto$$

$$\exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{m} n_j (\bar{y}_j - \theta_j)^2 \right).$$

As before, we assume that $\theta_1, \ldots, \theta_m$ are i.i.d. from $N(\mu, \tau^2)$, *but now we will not have a prior for μ or τ^2.*

In empirical Bayes,

- μ and τ^2 are estimated from the observations $\bar{y}_1, \ldots, \bar{y}_m$, and then

- the estimated "prior" for the θ_js is employed in usual Bayesian fashion to obtain estimates of $\theta_1, \ldots, \theta_m$.

Let's see how this works.

To perform the first step, we first find the marginal density of the data.

The joint density of $\bar{y}_1, \ldots, \bar{y}_m$ and $\theta_1, \ldots, \theta_m$ is

$$p(\bar{y}_1, \ldots, \bar{y}_m | \theta_1, \ldots, \theta_m)$$

$$\times \exp\left(-\frac{1}{2\tau^2} \sum_{j=1}^{m} (\theta_j - \mu)^2\right).$$

To find the marginal of $\bar{Y}_1, \ldots, \bar{Y}_m$, we need to integrate out $\theta_1, \ldots, \theta_m$, which can be done separately for each θ_i.

Using our beloved "complete the square" trick, we find that

$$\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{n_j(\bar{y}_j-\theta_j)^2}{\sigma^2}+\frac{(\theta_j-\mu)^2}{\tau^2}\right)\right] d\theta_j =$$

$$\exp\left[-\frac{1}{2}\left(\frac{n_j\bar{y}_j^2}{\sigma^2}+\frac{\mu^2}{\tau^2}\right)\right]$$

$$\times \exp\left(\frac{B^2}{2A}\right)\int_{-\infty}^{\infty} \exp\left[-\frac{A}{2}\left(\theta_j-\frac{B}{A}\right)^2\right] d\theta_j,$$

where

$$A=\frac{n_j}{\sigma^2}+\frac{1}{\tau^2} \quad \text{and} \quad B=\frac{n_j\bar{y}_j}{\sigma^2}+\frac{\mu}{\tau^2}.$$

It follows that the integral is

$$\exp\left[-\frac{1}{2}\left(\frac{n_j\bar{y}_j^2}{\sigma^2}+\frac{\mu^2}{\tau^2}\right)\right]\exp\left(\frac{B^2}{2A}\right)\frac{\sqrt{2\pi}}{\sqrt{A}}\propto$$

$$\frac{1}{\sqrt{2\pi(\tau^2+\sigma^2/n_j)}}\exp\left[-\frac{1}{2}\frac{(\bar{y}_j-\mu)^2}{(\tau^2+\sigma^2/n_j)}\right].$$

So, marginally, $\bar{Y}_1,\ldots,\bar{Y}_m$ are independent with

$$\bar{Y}_j\sim N\left(\mu,\tau^2+\frac{\sigma^2}{n_j}\right).$$

Now find the maximum likelihood estimators of μ and τ^2. Call these $\widehat{\mu}$ and $\widehat{\tau}^2$.

Now we estimate each θ_j as we would in a Bayesian approach with prior

$$\prod_{j=1}^{m}\frac{1}{\widehat{\tau}}\exp\left(-\frac{1}{2\widehat{\tau}^2}(\theta_j-\widehat{\mu})^2\right).$$

The θ_js are independent, given the data, and (using the result on p. 64N) the posterior mean of θ_j is

$$\frac{\widehat{\tau}^2\,\bar{y}_j + \widehat{\mu}\left(\frac{\sigma^2}{n_j}\right)}{\widehat{\tau}^2 + \left(\frac{\sigma^2}{n_j}\right)}. \tag{9}$$

When the sampling variance of $\bar{Y}_j$ is smaller than $\widehat{\tau}^2$, then more weight is placed on $\bar{y}_j$ than $\widehat{\mu}$. Otherwise more weight is placed on $\widehat{\mu}$.

We can see that this makes sense by looking at the extreme case where $\tau^2 = 0$.

In this case all θ_js are equal to μ, and we definitely want to estimate each θ_j by the same weighted average of all the $\bar{y}_j$s.

The estimate (9) is a *shrinkage estimator.* Instead of using $\bar{y}_j$ as the estimate of θ_j, this estimate is shrunk towards the overall weighted mean $\widehat{\mu}$.

This is an example of *borrowing information*. Information from all samples is used in estimating each θ_j.

Robbins (1955) is usually credited with the advent of empirical Bayes methods.

An introduction to the idea may be found in Casella (1985), *The American Statistician*.

Dealing with unequal variances

Unequal variances are easily dealt with by using exactly the same model as before but with

$$Y_{1,j} \ldots, Y_{n_j,j} \text{ i.i.d. } N(\theta_j, \sigma_j^2),$$

given θ_j and σ_j^2.

We use the same type of hierarchical model for $\sigma_1^2, \ldots, \sigma_m^2$ as we do for the means. Assume that

$$\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_m^2} \text{ are i.i.d. gamma}(\nu_0/2, \nu_0\sigma_0^2/2).$$

We would then have a prior π for ν_0 and σ_0^2. The resulting posterior is like that on p. 187N but with the blue term replaced by

$$\left(\prod_{j=1}^{m} \sigma_j^{-n_j}\right) \exp\left[-\frac{1}{2}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{(y_{i,j}-\theta_j)^2}{\sigma_j^2}\right)\right],$$

the first purple term replaced by $\pi(\nu_0, \sigma_0^2)$, and the whole thing multiplied by the new term

$$\prod_{j=1}^{m} (\sigma_j^2)^{-\nu_0/2-1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma_j^2}\right).$$

Hoff proposes a choice for the prior π on p. 144. Unfortunately, a simple conjugate prior for ν_0 does not exist.