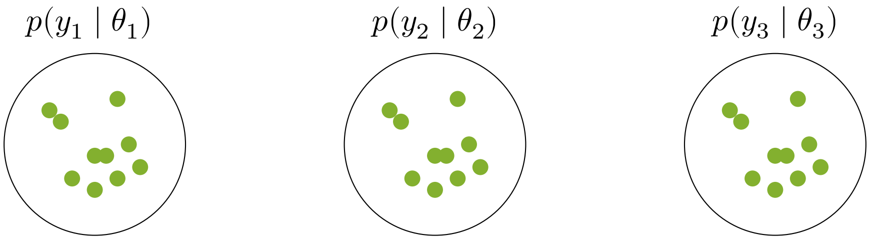
Hierarchical Models

Introduction

Suppose we want to compare the distributions of several groups.For example, suppose we are looking at a multi-center study. In particular, let's say there are 8 hospitals in the study, and at each we are able to look at the average response of patients to a new treatment. In this case, we have two populations: the hospitals and the patients. The population of patients is nested within the population of hospitals. This is an example of a hierarchical situation.

Whenever we have groups dividing a population, and we want to compare between those groups, we can use a hierarchical model approach. In particular, for the jth group, we can define our likelihood as $p(y_i|\theta_j)$. Notice here that we are saying θ_j is the expected mean in group j, and this mean will differ between groups, but we also expect those means to be related to one another. We can set a prior on θ_j such that we treat them as coming from a common population.



$y_j =$ The observed data in the j^{th} group

$\theta_j =$ The mean of the j^{th} group

(The parameter of interest)

Let's consider a similar example. In 1993, Donald Rubin investigated a hypothetical situation relating to hospital ratings. In particular, he divided the country into 8 geographic regions. In each region, a random sample of hospitals were taken. The problems at each hospital were rated on a scale from 0 to 1000, with 0 being no problems and 1000 being completely problematic. Ratings greater than 150 were considered "worse than typical." The data is as follows:

Region	n_j	y_j	$SE(y)$
1	18	176	13
2	30	152	10
3	12	141	16
4	26	151	11
5	35	143	9
6	24	145	11
7	28	162	10
8	10	156	18

n_j is the number of observations in region j
y_j is the average rating in region j
$J = 8$ different groups to compare

What is the true average rating in region j?
In other words,
we want to determine θ_j for $j = 1, 2, \dots, J$

Let θ_j be the true mean rating in region j. Hows can we go about comparing the mean ratings in each area?

- Pool all regions together and carryout and ANOVA to see if they are different
- Use sample mean from each group and conduct pairwise analysis to compare between groups

We could determine if θ_j is significantly different from 150, the average value cutting off worse than typical from better than typical.

$$\begin{aligned} \hat{\theta}_1 &= \bar{y}_1 = 176 \\ \text{95 \% CI:} \quad &= \bar{y}_1 \pm 2 \cdot SE(\bar{y}_1) \\ \text{(Region 1)} &= 176 \pm 2 \cdot 13 = (150.5, 201.5) \end{aligned}$$

Alternatively, we could perform an F-test from an ANOVA to determine if the eight means are significantly different from each other.

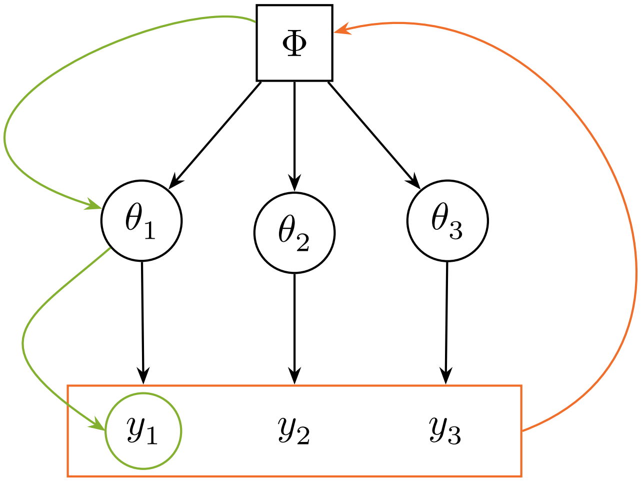
$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J$$
$$H_a : \text{at least one } \theta_j \text{ is different}$$
$$\text{Grand Mean:} \quad \hat{\theta}_j = \bar{y}_{..} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + \cdots + n_8\bar{y}_8}{n_1 + n_2 + \cdots + n_8} = 152.6$$

For every region we would use the grand mean of all hospitals for our estimate of θ_j

Using these two approaches, we would arrive at different estimates of the mean. Let's just consider θ_1 for the time being.

$$\left.\begin{array}{l} \text{Using the ANOVA approach,} \\ \text{our estimated mean would be:} \\ \qquad\qquad \hat{\theta}_1 = 152.6 \\ \text{Using the data from just region 1,} \\ \text{our estimated mean would be:} \\ \qquad\qquad \hat{\theta}_1 = 176 \end{array}\right\} \begin{array}{l} \text{The true value of } \theta_1 \\ \text{is likely somewhere} \\ \text{between these two} \\ \text{estimates} \end{array}$$

We probably believe the true estimate is somewhere between these two. A hierarchical model will allow us to incorporate information both from other regions and just from the jth region to better estimate the truth.



In general, for a hierarchical model, we will extend the inference we have done so far to include **another level of distributions**. In particular:

Up until now:

$$\begin{array}{ll} y|\theta \sim N(\theta, \sigma^2) & y|\theta \sim \text{Bin}(n, \theta) \\ \theta \sim N(\mu, \tau^2) & \theta \sim \text{Beta}(\alpha, \beta) \end{array}$$

Hierachical Model:

$p(y_j \vert \theta_j)$	Likelihood
$p(\theta_j \vert \phi)$	Prior
$p(\phi)$	Hyperprior

Notice here, that we have now put a distribution on the hyperparameters. In doing so, we can borrow strength from other groups to help estimate the mean in the jth group.

Before considering different cases of hierarchical models, we need to revisit the concept of exchangeability. Suppose we have J different groups. For the jth group, we have the likelihood $p(y_j|\theta_j)$ and prior $p(\theta_j)$. Assuming we have no further information by which to distinguish the J groups, we assume symmetry among these parameters in their priors. **We do this by modeling the priors as being exchangeable:**

$$\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_J)$$

$$p(\vec{\theta}) = \left.\prod_{j=1}^{J} p(\theta_j|\phi)\right\} \begin{array}{c}\scriptstyle\text{independent}\\ \scriptstyle\text{and}\\ \scriptstyle\text{identically}\end{array} \text{ distributed priors on } \theta_j's$$

With the hyperparameters now having a distribution of their own, this can be updated as follows:

Overall Prior:

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

General Comments

For a hierarchical model, we have the likelihood $p(y|\theta)$ and prior $p(\theta|\phi) = p(\theta|\phi)p(\phi)$. In this case, the **joint posterior distribution** is:

$$\begin{aligned} p(\theta_1, \dots, \theta_J, \phi|y) &\propto p(\vec{y}|\vec{\theta})p(\vec{\theta}|\phi)p(\phi) \\ &= \left[\prod_{j=1}^{J} p(y_j|\theta_j)\; p(\theta_j|\phi)\right] p(\phi) \end{aligned}$$

When using an uninformative prior for ϕ, it is important to be sure that the posterior is proper.

Ultimately, we will go about doing inference in a similar manner as to what we did when there were two unkown variables for the normal model:

1. Determine the joint posterior: $p(\theta, \phi|y) \propto p(y|\theta)p(\theta|\phi)p(\phi)$
2. Obtain the conditional posterior of θ given ϕ, or $p(\theta|\phi, y)$
3. Obtain the marginal posterior of ϕ, or $p(\phi|y)$

If we then are interested in a new value of y (i.e. $\tilde{y}$), we can get that by taking a draw from $p(\phi|y)$. Then for each draw of $\phi|y$, we can draw $p(\theta|\phi^\star, y)$. From here, we cam then draw $\tilde{y}$ from the likelihood given $\theta^\star$ and repeat this process many times. This is equivalent to a draw from the posterior predictive distribution.

Binomial Model

Suppose our data are discrete and arise from a binomial distribution. We will begin by discussing the hierarchical model for this case. Suppose there are J different groups for which was are comparing the probability of success.

$$\theta_j = \text{probability of success in group } j$$

In this case, our likelihood would be:

$$y_j|\theta_j \sim \text{Bin}(n_j, \theta_j), \quad \text{for } j = 1, ..., J$$

where y_j is the number of successes in group j and n_j is the number of trials in group j. The prior for θ_j is:

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

For each θ_j individually, we could perform inference as discussed in Chapter 2. Here, though, we **will place a prior on** (α, β), which will allow groups to influence one another. Thus, we also have $p(\alpha, \beta)$, which we still need to determine an appropriate form for. **By putting a distribution on (α, β) we are treating them as random!**

The overall goal of inference is to examine the joint posterior distirbution $p(\theta_1, ..., \theta_J, \alpha, \beta|y)$. Similar to the methods of Multiple Parameter Models, we will break this down into step to make inference more managabe and interpretable.

Hierarchical Model:

$$\begin{aligned} &y_j|\theta_j \sim \text{Bin}(n_j, \theta_j) \\ &\theta_j|\alpha, \beta \sim \text{Beta}(\alpha, \beta) \\ &\alpha, \beta \sim p(\alpha, \beta) \end{aligned}$$

Single Parameter Model:

$$\begin{aligned} &y|\theta_j \sim \text{Bin}(n, \theta) \\ &\theta \sim \text{Beta}(\alpha, \beta) \end{aligned}$$

1. Determine the joint posterior: $p(\theta_1, ..., \theta_J, \alpha, \beta|y)$
2. Factor the joint posterior into two pieces:
   (a) $p(\theta_1, ..., \theta_J|\alpha, \beta, y)$ **Conditional Posterior**
   (b) $p(\alpha, \beta|y)$ **Marginal Posterior**
3. Determine an appropriate prior for (α, β)

Lets begin by determining the joint posterior:

$$\begin{aligned} p(\theta_1, ..., \theta_J, \alpha, \beta|y) &\propto p(\vec{y}|\theta_1, ..., \theta_J, \cancel{\alpha, \beta})\ p(\theta_1, ..., \theta_J, \alpha, \beta) \\ &= p(\vec{y}|\theta_1, ..., \theta_J)\ p(\theta_1, ..., \theta_J|\alpha, \beta)\ p(\alpha, \beta) \\ &= \left[\prod_{j=1}^{J} p(y_j|\theta_j) p(\theta_j|\alpha, \beta)\right] p(\alpha, \beta) \quad \text{by exchangablity} \\ &= \left[\prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j - y_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\right] p(\alpha, \beta) \\ &= p(\alpha, \beta) \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^J \prod_{j=1}^{J} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1} \quad \textbf{Joint Posterior} \end{aligned}$$

Now, let's determine $p(\theta_1, ..., \theta_J|\alpha, \beta, y)$:

$$
\begin{aligned}
p(\theta_1, ..., \theta_J|\alpha, \beta, y) \propto \prod_{j=1}^{J} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1} &\propto \prod_{j=1}^{J} p(\theta_j|\alpha, \beta, y_j) \\
&= \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)}\theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1} \\
\theta_j|\alpha, \beta, y_j \sim \text{Beta}(y_j+\alpha, n_j-y_j+\beta) \quad &\textbf{Conditional Posterior}
\end{aligned}
$$

And finally $p(\alpha, \beta|y)$:

$$
\begin{aligned}
p(\alpha, \beta|y) &\propto \int p(\theta_1, ..., \theta_J, \alpha, \beta|y)d\theta_1...d\theta_J \\
&= \int \cdots \int p(\alpha, \beta)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^J \left[\prod_{j=1}^{J} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}\right] d\theta_1...d\theta_J \\
&= p(\alpha, \beta)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^J \int \cdots \int \left[\prod_{j=1}^{J} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}\right] d\theta_1...d\theta_J \\
&= p(\alpha, \beta)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^J \prod_{j=1}^{J} \frac{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)}{\Gamma(n_j+\alpha+\beta)} \quad \textbf{Marginal Posterior}
\end{aligned}
$$

Determining a useful prior for hyper-parameters can be diffcult, particularly since there is little intuition regarding these parameters. Skipping the details as to how this prior is obtained, an uninformative prior can be determined to be $p(\alpha, \beta|y) \propto (\alpha+\beta)^{-5/2}$. this leads to an updatedmarginal posterior for (α, β):

$$
\begin{aligned}
p(\alpha, \beta) &\propto (\alpha+\beta)^{-5/2} \\
p(\alpha, \beta|y) &\propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^J \prod_{j=1}^{J} \frac{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)}{\Gamma(n_j+\alpha+\beta)} \cdot (\alpha+\beta)^{-5/2}
\end{aligned}
$$

This posterior distribution is not a nice convenient form that we can readily use for inference. Because **this is not a standard distribution**, we would need to use a computational method to approximate this distribution. Generally, we will develop a method to **draw samples** from the posterior distribution. We will use those samples as our best **approximation for the true distribution**, and we can empirically determine the mean, posterior interval, and any other quantities of interest from there.

Binomial Example Consider an extension to the binomial example we discussed in section 2.1.6. In particular, people from the greater Rochester area received a Facebook sidebar advertisement for Restaurant Good Luck. Facebook users were targeted based on their address, being in either the City of Rochester, Henrietta, Pittsford, Brighton, Fairport, Penfield, or Webster. Good Luck is interested in determining the proportion of people who click the link based on where they live. This information can help better determine the areas that people who most likely will eat at Good Luck live. We have data from this study as follows:

Area	n_j	y_j
Rochester	73	20
Henrietta	32	5
Pittsford	52	18
Brighton	21	8
Fairport	24	6
Penfield	19	4
Webster	34	10

$$
\begin{aligned}
&\theta_j = \text{ probability of clicking on the ad in area } j\\
&y_j|\theta_j \sim \text{Bin}(n_j, \theta_j)\\
&\theta_j|\alpha, \beta \sim \text{Beta}(\alpha, \beta)\\
&J = 7\\
&p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}
\end{aligned}
$$

Let's consider posterior inference for this scenario using an uninformative prior.

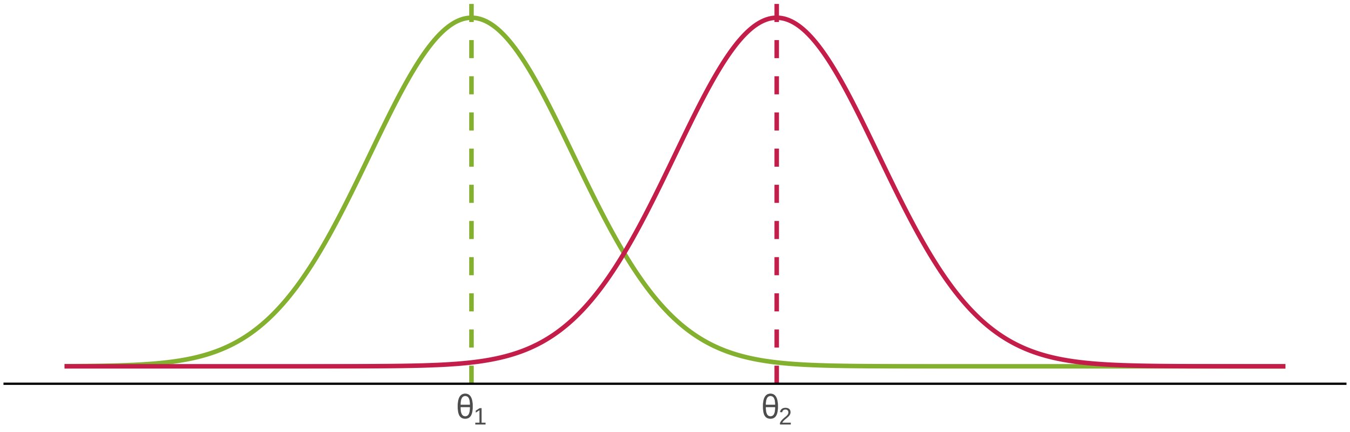
$$
\begin{aligned}
\theta_j|\alpha, \beta, y_j &\sim \text{Beta}(y_j+\alpha, n_j-y_j+\beta)\\
\alpha, \beta &\sim p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}\\
p(\alpha, \beta|y) &\propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^J \prod_{j=1}^{J} \frac{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)}{\Gamma(n_j+\alpha+\beta)} \cdot (\alpha+\beta)^{-5/2}\\
\theta_1|\alpha, \beta, y_1 &\sim \text{Beta}(20+\alpha, 73-20+\beta) \quad \theta_2|\alpha, \beta, y_2 \sim \text{Beta}(5+\alpha, 32-5+\beta)
\end{aligned}
$$

Find a way to sample (α, β) from the posterior distribution. We can then sample θ_j from the conditional posterior distribution. Finally, we can sample y_j from the likelihood given θ_j. This will give us a sample of y_j that is consistent with the data we have observed. (MCMC Sampling; Chapter 5)

Normal Model

Now, let's consider the example of our data being normally distributed, with unknown mean and known variance. In this case:

$$
y_{ij}|\theta_j \sim \mathcal{N}(\theta_j.\sigma^2), \quad \text{for } i=1,...,n_j, \quad j=1,...,J
$$



If instead of dealing with each y_ij individually, we look at $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n_j} y_{ij}$, then we have σ^2/n_j and our likelihood is:

$$
\bar{y}_j|\theta_j \sim \mathcal{N}(\theta_j, \frac{\sigma^2}{n_j}), \quad \text{for } j=1,...,J
$$

The last piece of information we need is the overall mean for al the data this quantity is given as follows:

$$\bar{y}_{..} = \frac{\sum_j \sum_i y_{ij}}{\sum_j n_j} \quad \text{(Grand Mean)}$$

As previously discussed with our motivating example, if we want to estimate θ_j, we likely will want a posterior estimate somewhere between $\bar{y}_j$ and $\bar{y}_{..}$. In particular, we want something of the form:

$$\hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j)\bar{y}_{..}$$

λ_j will be a quantity between 0 and 1. This weighting of the posterior mean will depend on choice of prior used. Lets consider a few options:

Mean Weights

1. If we let each θ_j have an independent uniform distribution on $(-\infty, \infty)$, then the posterior mean θ_j is:

   $$\theta_j | \mu, \tau^2 \sim \mathcal{N}(0, \infty)$$

   ✗ No sharing of information

2. Another option would be to restrict all θ_j to equal $\theta^\star$ and then place a uniform prior on θ. This would look like:

   $$\theta_j | \mu, \tau^2 \sim \mathcal{N}(\mu, 0)$$

   ✗ Complete sharing of information

3. What if we used $\theta_j \sim \mathcal{N}(\mu, \tau^2)$, for exchangable θ_j?

   $$\begin{aligned} &\bar{y}_j | \theta_j \sim \mathcal{N}(\theta_j, \frac{\sigma^2}{n_j}) \\ &\theta_j | \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2) \\ &p(\mu, \tau^2) \end{aligned}$$

   ✓ Posterior mean of each θ_j will be a weighted average of $\bar{y}_j$ and $\bar{y}_{..}$

Posterior Distribution(s)

If we us this prior for θ, we also would want to place priors on μ and τ in order to make this a true hierarchical model. Using an uninformative uniform prior for $\mu | \tau$ gives:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

Having an uninformative prior on μ is fine since the data has plenty enough information for this. We will discuss the prior for τ in more detail later.

In this case, the joint posterior can be written as follows:

$$\begin{aligned} p(\theta_1, ..., \theta_J, \mu, \tau | y) &\propto p(\bar{y}_1, ..., \bar{y}_J | \theta_1, ..., \theta_J) p(\theta_1, ..., \theta_J | \mu, \tau) p(\mu, \tau) \\ \text{by exchangability} \rightarrow &= \left[\prod_{j=1}^{J} p(\bar{y}_j | \theta_j) p(\theta_j | \mu, \tau) \right] p(\mu, \tau) \\ &\propto \left[\prod_{j=1}^{J} \exp\left(\frac{-(\bar{y}_j - \theta_j)^2}{2\sigma^2} \right) \cdot \frac{1}{\tau} \exp\left(\frac{-(\theta_j - \mu)^2}{2\tau^2} \right) \right] p(\tau) \end{aligned}$$

As with other scenarios, we will consider this joint posterior based on the decomposed pieces.

$$\begin{aligned}p(\theta_1,...,\theta_J,\mu,\tau|y) &\propto p(\theta_1,...,\theta_J|\mu,\tau,y)p(\mu,\tau|y)\\ &= \left[\prod_{j=1}^{J} p(\theta_j|\mu,\tau,y)\right]p(\mu,\tau|y)\end{aligned}$$

Let's begin with $p(\theta_1,...,\theta_J|\mu,\tau,y)$

$$p(\theta_1,...,\theta_J|\mu,\tau,y) \propto \prod_{j=1}^{J}\left[\exp\left\{\frac{-(\bar{y}_j-\theta_j)^2}{2\sigma^2}\right\}\exp\left\{\frac{-(\theta_j-\mu)^2}{2\tau^2}\right\}\right]$$

$$p(\theta_j|\mu,\tau,y) \propto \exp\left\{\frac{-(\bar{y}_j-\theta_j)^2}{2\sigma^2}-\frac{(\theta_j-\mu)^2}{2\tau^2}\right\}$$

→ From here, combine the two parts, drop parts without θ_j
and work into a recognizable form as we did in the notes on single parameter models

$$p(\theta_j|\mu,\tau,y) = \mathcal{N}\left(\frac{\frac{\bar{y}_j}{\sigma_j^2}+\frac{\mu}{\tau^2}}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}},\frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}\right)$$ **Conditional Posterior**

Now let's consider $p(\mu,\tau|y)$ **Marginal Posterior**.

$$p(\mu,\tau|y) = \int\cdots\int p(\theta_1,...,\theta_J,\mu,\tau|y)d\theta_1...d\theta_J$$

$$p(\mu,\tau|y) \propto p(y|\mu,\tau)p(\mu,\tau) \propto p(y|\mu,\tau)\cancel{p(\mu|\tau)}p(\tau)$$

$$p(\mu,\tau|y) \propto \left[\prod_{j=1}^{J}p(\bar{y}_j|\mu,\tau)\right]p(\tau)$$

It follows from this that this marginal posterior will have a
normal distribution. As such, it is easier to use the law of total
expectation to find the mean & variance than to solve this integral.

$$\begin{aligned}\mathbb{E}(\bar{y}_j|\mu,\tau) &= \mathbb{E}\left[\mathbb{E}(\bar{y}_j|\theta_j)|\mu,\tau\right]\\ &= \mathbb{E}(\theta_j|\mu,\tau) = \mu\end{aligned}$$

$$\begin{aligned}\mathbb{V}(\bar{y}_j|\mu,\tau) &= \mathbb{E}\left[\mathbb{V}(\bar{y}_j|\theta_j)|\mu,\tau\right]+\mathbb{V}\left[\mathbb{E}(\bar{y}_j|\theta_j)|\mu,\tau\right]\\ &= \mathbb{E}\left[\sigma_j^2|\mu,\tau\right]+\mathbb{V}\left[\theta_j|\mu,\tau\right] = \tau^2+\sigma_j^2\end{aligned}$$

$$p(\mu,\tau|y) \propto p(\tau)\prod_{j=1}^{J}\left[\left(\sigma_j^2+\tau\right)^{-1/2}\exp\left\{\frac{-(\bar{y}_j-\mu)^2}{2(\sigma_j^2+\tau^2)}\right\}\right]$$

Decomposing this into its two pieces gives us the the posterior distribution of $\mu|\tau$ as:

$$p(\mu,\tau|y) = p(\mu|\tau,y)p(\tau|y) \qquad p(\mu|\tau,y) \propto p(\mu,\tau|y)$$

$$\propto \prod_{j=1}^{J} \exp\left\{\frac{-(\bar{y}_j-\mu)^2}{2(\sigma_j^2+\tau^2)}\right\} = \exp\left\{-\sum_{j=1}^{J} \frac{(\bar{y}_j-\mu)^2}{2(\sigma_j^2+\tau^2)}\right\}$$

$$p(\mu|\tau,y) \sim \mathcal{N}\left(\frac{\sum_{j=1}^{J} \frac{\bar{y}_j}{\sigma_j^2+\tau^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}, \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}\right) = \mathcal{N}(\hat{\mu},\nu_\mu)$$

And the posterior distribution of $\tau|\mu$ is as follows:

$$p(\tau|y) \propto \int p(\mu,\tau|y)d\mu$$

$$p(\tau|y) = \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \qquad p(\mu,\tau|y) = p(\mu|\tau,y)p(\tau|y)$$

$$\propto \frac{p(\tau)\prod_{j=1}^{J}\left[\left(\sigma_j^2+\tau\right)^{-1/2}\exp\left\{\frac{-(\bar{y}_j-\mu)^2}{2(\sigma_j^2+\tau^2)}\right\}\right]}{\nu_\mu^{-1/2}\exp\left\{\frac{-(\mu-\hat{\mu}_j)^2}{2\nu_\mu}\right\}}$$

$$p(\tau|y) \propto p(\tau)\ \nu_\mu^{-1/2}\prod_{j=1}^{J}\left[\left(\sigma_j^2+\tau\right)^{-1/2}\exp\left\{\frac{-(\bar{y}_j-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right\}\right]$$

For the prior $p(\tau)$, we could use an uninformative approach and do $p(\tau) \propto c$. While this will lead to a proper posterior, it is often better to use something that is somewhat informative. (i.e. determine a "best guess" or an upper bound for τ)

To draw inference from this joint posterior distribution, we again would need to use some sort of computational approximation. We could draw a sample from $p(\tau|y)$. Then draw from $p(\mu|\tau,y)$ and finally from $p(\theta_1,...,\theta_J|\mu,\tau,y)$.

Procedure

1. Sample τ from $p(\tau|y)$ using MCMC Methods
2. Use the sample of τ to sample μ from $p(\mu|\tau,y)$
3. Use the sample of (μ,τ) to draw samples of θ_j from $p(\theta_j|\mu,\tau,y)$