Demo: Bayesian Hierarchical Models (Draft Version)

Paul Gustafson

2024-09-11

This R Markdown demo resides at github.com/paulgstf/Bayes-demos

The beginning

Consider a statistical problem where $Y_i \sim N(\theta_i, 1)$, independently for $i = 1, \ldots, 10$. So we can think of the data vector $Y = (Y_1, \ldots, Y_{10})$ estimating the parameter vector $\theta = (\theta_1, \ldots, \theta_{10})$. And we will use the phrasing of "units," in the sense that θ_i describes the unknown state of the i-th unit, with Y_i being the observable quantity for said unit.

For a particular value of θ we simulate 5000 independent realizations of the data vector Y:

```
theta.tr <- seq(from=0.5, to=5, length=10)

### each row of y.mat is a simulated realization of the data vector
y.mat <- mvrnorm(5000, mu=theta.tr, Sigma=diag(10))
```

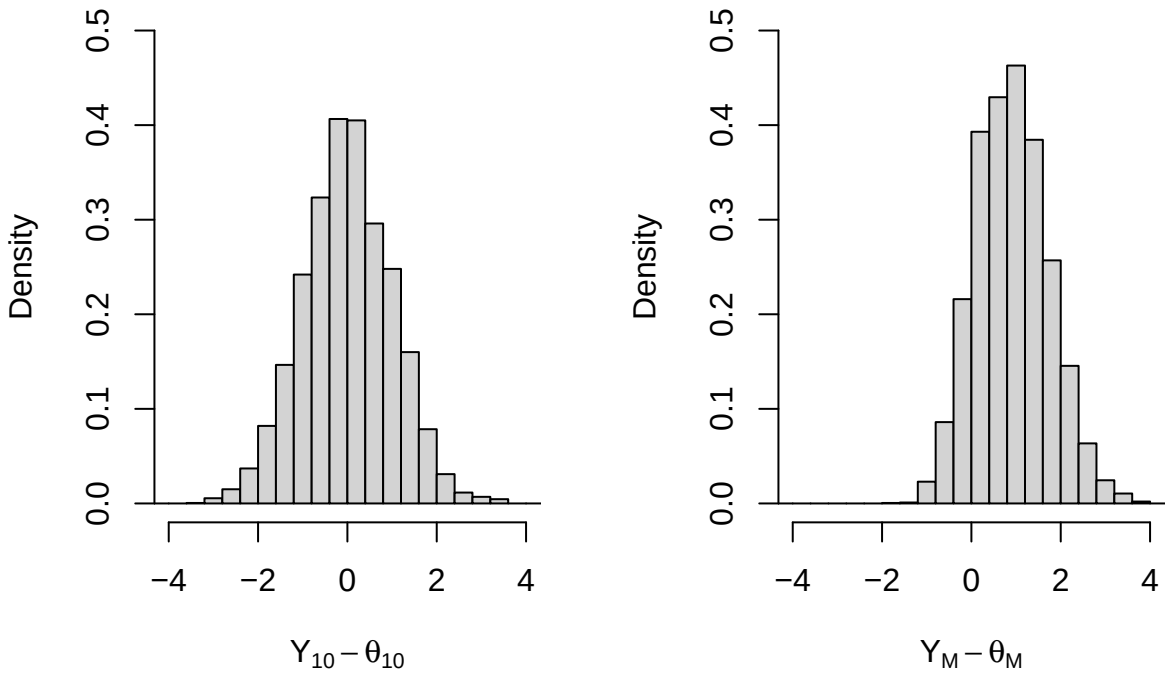
For this illustration, the largest element of θ is $\theta_{10} = 5$. But due to sampling variation, Y_{10} may, or may not, be the largest element of Y. To pay some heed to the largest element, Let $M = M(Y)$ be its index. So formally $M = \text{argmax}_i\{Y_i\}$, and $Y_M = \max_i\{Y_i\}$. In summarizing the 5000 realizations of M, we see indeed that in almost half of the realizations it *isn't* the biggest element of θ that produces the biggest element of Y.

```
### for each simulated data vector, determine M
m <- t(apply(y.mat, 1, order))[,10]

table(m)
```

```
## m
##    2    3    4    5    6    7    8    9   10
##    1    9   10   33  109  295  638 1376 2529
```

Our next move is to examine the distribution, across our repeated sampling, of (i), the error incurred by Y_{10} as an estimator of θ_{10}, and (ii), the error incurred by Y_M as an "estimator" of θ_M. Now (i) is well understood: this error is $Y_{10} - \theta_{10}$, which has mean zero and variance one with respect to repeated sampling of Y given θ. However, (ii) is more nuanced. The error is $Y_M - \theta_M$. But θ_M, in being a function of *both* the parameter vector θ and the data vector Y, is not an estimand in the traditional sense (of being a function of parameters only). Nonetheless, we can examine how well the largest element of Y estimates *the corresponding mean that spawned it,* which is exactly what $Y_M - \theta_M$ describes.

```
par(mfrow=c(1,2))

br <- seq(from=-8, to=8, by=0.4)
hist(y.mat[,10]- theta.tr[10],
  xlab=expression(Y[10]-theta[10]),
  breaks=br, prob=T, main="", xlim=c(-4,4), ylim=c(0,0.5))

hist(y.mat[cbind(1:5000,m)]- theta.tr[m],
  xlab=expression(Y[M]-theta[M]),
  breaks=br, prob=T, main="", xlim=c(-4,4), ylim=c(0,0.5))
```



The left panel above simply provides empirical confirmation that a $N(0,1)$ error is incurred when Y_i estimates θ_i, for $i = 10$ (or for that matter, for any i). In contrast, the right panel

shows a substantial positive bias when we conceive of Y_M as an estimator of θ_M.

Upon reflection, seeing $E(Y_M - \theta_M) > 0$ makes sense. For any j, Y_j is more likely to become $Y_M = \max_i\{Y_i\}$ if its realized error, $Y_j - \theta_j$, is positive. Across repeated sampling then, Y_M is more likely to fall above θ_M, rather than below.

Note that this narrative is focusing on the largest element of Y, for ease of explanation. Equally, though, the smallest element of Y will underestimate the mean that spawned it. And the phenomenon would carry over, albeit less strongly, to larger and smaller elements of Y, not just *the* largest and *the* smallest.

Notwithstanding the non-traditional sense of θ_M as an estimand of interest, we wonder if a bias like $E(Y_M - \theta_M) > 0$ could have implications for good statistical estimation of θ as a whole. The bias does cast doubt on the *prima facie* thought that since the elements of Y are independent of one another given θ, the estimation of θ_i should be driven entirely by Y_i, with $(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_{10})$ being irrelevant for this purpose. Since observing that Y_i is large relative to its peers suggests it is more likely an overestimate of θ_i, we wonder if some commensurate tweak to our estimation procedure is warranted? While (the vector) $Y = (Y_1, \ldots, Y_{10})$ is the obvious estimator of (the vector) $\theta = (\theta_1, \ldots, \theta_{10})$, could it be that something less obvious is actually better?

Bayesian analysis

We posit that the path to betterment starts with bringing a prior distribution to the table. A simple, conjugate prior would be $\theta_i \overset{\text{iid}}{\sim} N(\mu, \tau^2)$. Presuming μ and τ to be *hyperparameters* (i.e., the user specifies values for them), our posterior distribution for $(\theta|Y = y)$ has independent and normal components, with

$$\begin{aligned} E(\theta_i|y, \mu, \tau^2) &= \frac{\tau^2}{1+\tau^2} y_i + \frac{1}{1+\tau^2}\mu, \\ Var(\theta_i|y, \mu, \tau^2) &= \frac{\tau^2}{1+\tau^2}. \end{aligned} \tag{1}$$

Here the posterior mean of θ_i, which we regard as an estimator of θ_i, is a convex combination of the datapoint y_i and the prior mean μ. And a pleasing intuition is that for this combination, the weight given to the datapoint decreases as the prior variance τ^2 decreases. A common and useful phrasing would be that the estimate of θ_i is "shrunk" toward the prior mean μ, with the amount of "shrinkage" governed by the strength of prior assertion, i.e., more shrinkage when τ is smaller.

Taking stock, a user-provided assertion about the way in which the components of θ_i are similar to one another somehow binds together the ten inferential tasks of estimating the ten elements of θ. This path becomes more useful, however, if we *let the data speak* about the extent to which the components of θ_i are similar. This is achieved by treating μ and τ as *unknown parameters* rather than known hyperparameters. The ensuing construct is often

referred to as a *hierarchical model*, since the specification is made in stages: first we describe the uncertainty about Y given θ, followed by the uncertainty about θ given (μ, τ), and then finally the uncertainty about (μ, τ). More formally, with I observations (so $I = 10$ in our running example), the joint posterior distribution of all unknown parameters takes the form

$$f(\theta, \mu, \tau | y) \propto \left\{ \prod_{i=1}^{I} f(y_i | \theta_i) \right\} \left\{ \prod_{i=1}^{I} f(\theta_i | \mu, \tau) \right\} f(\mu, \tau), \tag{2}$$

where the last term is a prior density for (μ, τ). While terminology can vary, we will refer to θ as first-stage parameters, and (μ, τ) as second-stage parameters.

Using this joint posterior on (θ, μ, τ), we can do inference as per the usual recipe. For instance, the posterior mean $E(\theta_i | Y = y)$ is an obvious point estimator for θ_i. In terms of computing posterior quantities, we go the Monte Carlo route, since (2) doesn't quite have a closed-form representation. But we gain some closed-form insight by leaning on (1), and the law of iterated expectations, to arrive at:

$$E(\theta_i | Y = y) = \left(1 - \hat{b}\right) \hat{a} + \hat{b} y_i,$$

where

$$\hat{a} = \frac{E\left\{\mu (1 + \tau^2)^{-1} | Y = y\right\}}{E\left\{(1 + \tau^2)^{-1} | Y = y\right\}}, \tag{3}$$

$$\hat{b} = 1 - E\left\{(1 + \tau^2)^{-1} | Y = y\right\}. \tag{4}$$

To be clear, here $\hat{a}$ and $\hat{b}$ are sample statistics depending on all elements of y, as formed by expectations with respect to the marginal posterior distribution of (μ, τ) given $Y = y$. Via $\hat{a}$ and $\hat{b}$, the data decide how much to tweak from y_i to a presumed better estimate of θ_i. Note from the form of (3) that $\hat{a}$ can be regarded as (an admittedly convoluted) posterior estimate of μ, Note from the form of (4) that $\hat{b}$ will necessarily be between zero and one. Indeed then, the posterior means of θ arise from shrinking the components of y toward an estimate of μ. Since this estimate of μ is based on the data from all the units, it is common to refer to "borrowing-of-strength." That is, we "borrow" information from the other units, in order to give a more refined estimate of θ_i for the i-th unit. Moreover, noting that $y_i - E(\theta_i | Y = y) = (1 - \hat{b})(y_i - \hat{a})$, the magnitude of the additive shift is proportional to $|y_i - \hat{a}|$, i.e., the more outlying elements of y are shrunk more.

Under the hood, we write an R function to generate a Monte Carlo sample from (2). (Under the hood meaning we won't dwell on all the details here, but the function is available in the code generating this report.) In brief, this function, **posterior.A()**, takes the data vector y and the choice of prior density $p(\mu, \tau)$ as inputs. It then provides a sample of user-specified

size as output. To say slightly more, it turns out we can draw exact Monte Carlo samples from a very good approximation to the posterior distribution, and we can use *importance weighting* to correct for the discrepancy between this approximation and the actual posterior distribution.

In the following example analysis we use a prior distribution for (μ, τ) with independent components $\mu \sim N(0, 10^2)$ and $\tau \sim \text{Exponential}(0.1)$. These can be regarded as being very weakly informative. To wit, note that the rate parameter of 0.1 would imply only about a twofold decline in prior density when contrasting a tenfold change from $\tau = 1$ to $\tau = 10$. We encode this prior specification as an R function which can be given as an argument to **posterior.A()**.

```
### log prior density for mu,tau jointly

lg.pri.jnt <- function(mu, tau, mn.mu=0, sd.mu=10, rt.tau=0.1) {
  dnorm(mu, mean=mn.mu, sd=sd.mu, log=T) +
  dexp(tau, rate=rt.tau, log=T)
}

### for later purposes, log prior density for tau marginally
lg.pri.tau <- function(tau, rt.tau=0.1) {
  dexp(tau, rate=rt.tau, log=T)
###
}
```

We illustrate the inferential scheme using a data vector simulated under the same parameter values used earlier, i.e., $\theta = (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)$.

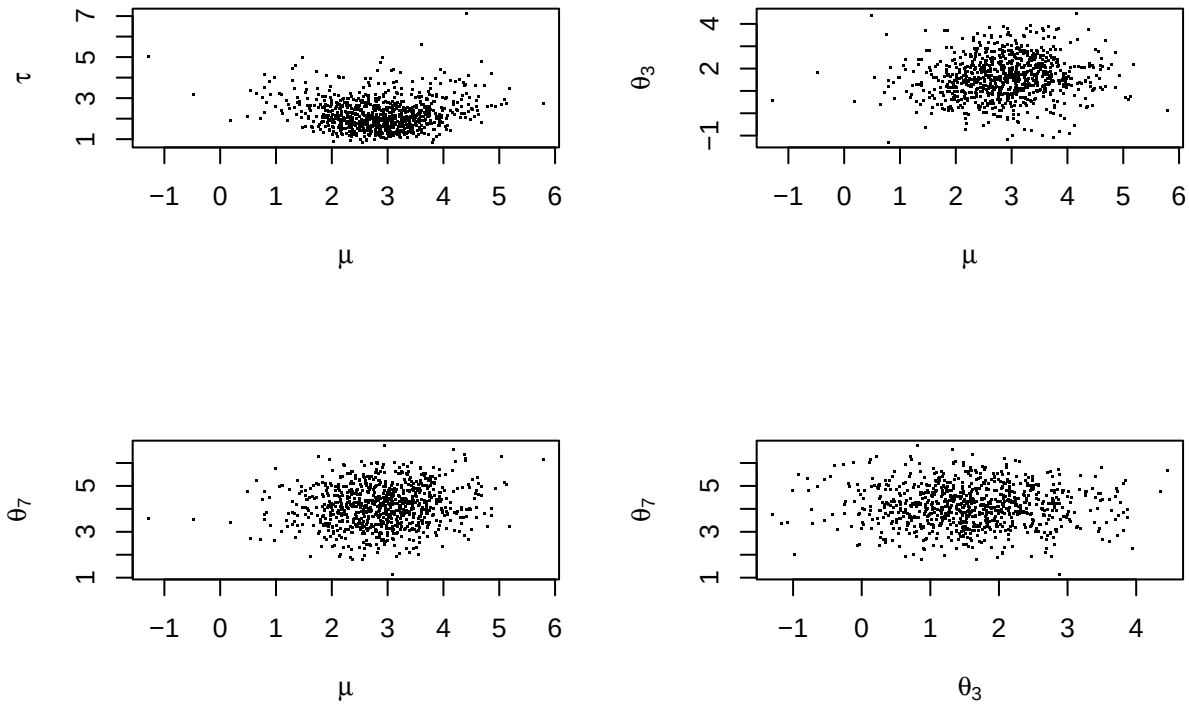
```
theta.tr <- seq(from=0.5, to=5, length=10)
set.seed(17)
y.test <- rnorm(10, mean=theta.tr)

round(y.test,2)
```

```
##  [1] -0.52  0.92  1.27  1.18  3.27  2.83  4.47  5.72  4.76  5.37
```

We compute a joint posterior over (θ, μ, τ) given the data $Y = y$:

```
ans.test <- posterior.A(y.test, lg.pri.jnt=lg.pri.jnt)
```

To underscore that we now "possess" the joint posterior distribution of (θ, μ, τ), here are a few of the resultant bivariate marginal distributions:



More pointedly, our MC output applied to (3) and (4) gives:

```
b.hat <- 1-mean(1/(1+ans.test$tau^2))
a.hat <- mean(ans.test$mu/(1+ans.test$tau^2))/(1-b.hat)

c(a.hat, b.hat)
```

```
## [1] 2.916 0.792
```

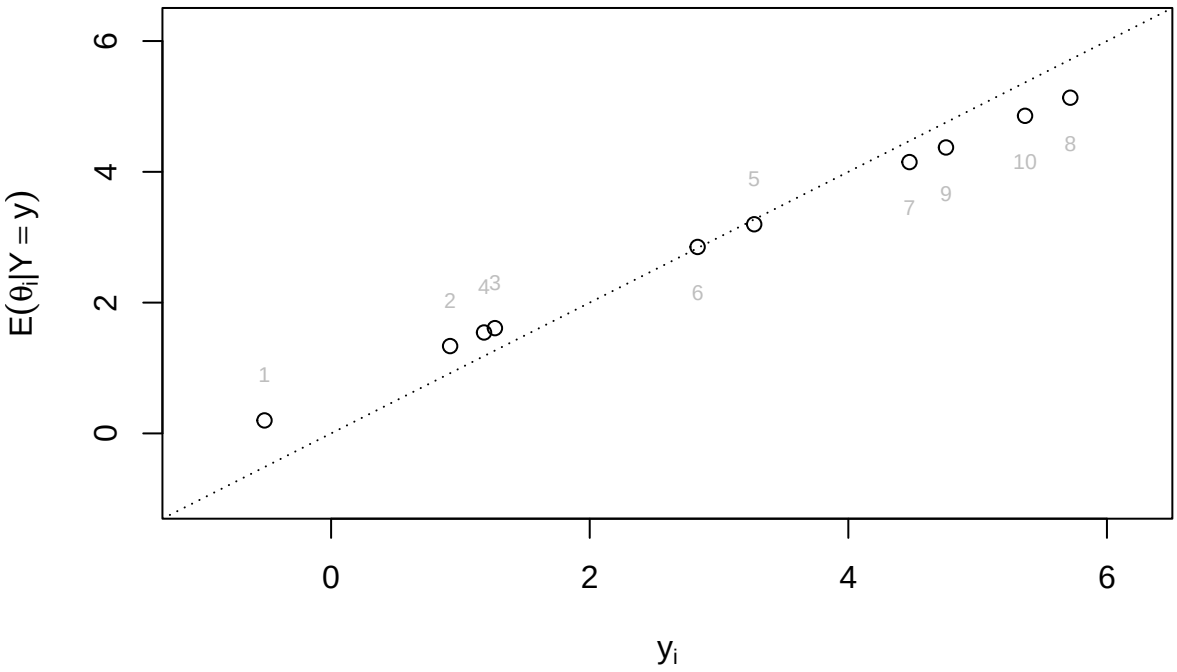
Then as a sanity check, note we actually have two routes to calculating $E(\theta_i|y)$, so its worth checking they agree:

```
post.mean <- a.hat + b.hat*(y.test-a.hat)

### alternatively, just use the MC draws for theta
post.mean.alt <- apply(ans.test$theta,2,mean)

### should agree to within numerical error
summary(abs(post.mean.alt - post.mean))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.00051 0.00078 0.00166 0.00285 0.00360 0.00952
```

To see the shrinkage effect in action, we plot the posterior means against the elements of y, with unit labels indicated, and with the identity line displayed for reference.

```
plot(y.test, post.mean,
     xlim=range(y.test)+0.5*c(-1,1),
     ylim=range(y.test)+0.5*c(-1,1),
     xlab=expression(y[i]), ylab= expression(E(paste(theta[i],"|",Y==y))))

ofst <- 0.7*c(rep(1,5),rep(-1,5))
text(y.test,  a.hat+b.hat*(y.test-a.hat) + ofst, as.character(1:10),
     cex=0.7, col="grey")
abline(c(0,1), lty=3)
```

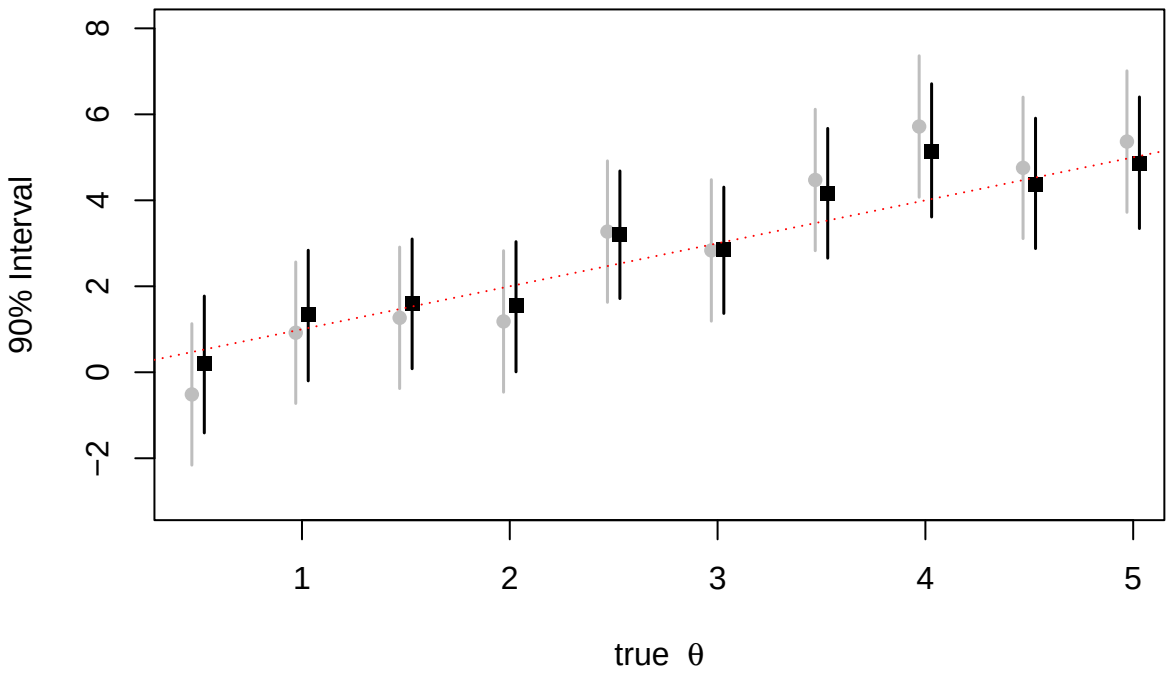


In looking at the above plot, we see the downward/upward shrinkage for units with large/small values of y_i. But because we simulated the data vector, we can go one step further and see if this shrinkage has helped. On aggregate, are the posterior means of $(\theta|Y = y)$ indeed closer to their targets than are the elements of Y themselves? Taking the square root of the average (of the ten) estimates as the typical error incurred, we see a typical error of 0.805 for y_i as an estimate of θ_i, compared to a typical error of 0.517 for $E(\theta_i|Y = y)$ as an estimate of θ_i. This 36% reduction in error is very promising. Though of course we shouldn't get too carried away by what happens for any one particular data vector.

Now let's take a look at interval estimation, both with, and without, shrinkage. For the former, 0.05 and 0.95 quantiles of the marginal posterior distribution of $(\theta_i|Y = y)$ form a 90% (equal-tailed) credible interval for θ_i. For the latter, treating our ten inference problems as unrelated suggests $y_i \pm 1.645$ as a 90% confidence interval for θ_i. (As a bit of a rabbit hole here, we could also motivate $y_i \pm 1.645$ as a 90% credible interval for θ_i if (i), we only observe Y_i, not Y, and (ii), the prior for θ_i is maximally wide, in a sense. Framed this way, we can focus in on the difference between using all of Y, or just Y_i, when estimating θ_i.)

```
### equal-tailed 90% credible intervals
int.90.Bayes <- apply(ans.test$theta, 2, quantile, probs=c(0.05, 0.95))

### frequentist 90% credible intervals
int.90.freq <- rbind(y.test - 1.645, y.test + 1.645)
```

```
jtr <- 0.03
plot(theta.tr-jtr, y.test, pch=16, col="grey", ylim=c(-3,8),
     xlab=expression(paste(true,phantom(0),theta)), ylab="90% Interval")
points(theta.tr+jtr, a.hat + b.hat*(y.test-a.hat), pch=15)

for (i in 1:10) {
  points(rep(theta.tr[i]+jtr,2), int.90.Bayes[,i], type="l",lwd=1.5)
  points(rep(theta.tr[i]-jtr,2), int.90.freq[,i], col="grey",type="l", lwd=1.5)
}

abline(c(0,1),lty=3, col="red")
```



For the (unshrunk) frequentist 90% confidence intervals, depicted in grey in the above plot, we know the number of the intervals that miss their target with be distributed as Binomial$(10, 0.1)$. As luck has it, for this realization of the data vector the number of misses is 1. Also as luck dictates, the number of misses we happen to see for the (shrunk) Bayesian intervals is 0.

From the figure, we see that the Bayesian credible intervals are narrower than their frequentist counterparts. This makes sense in that the latter can be (roughly) thought of as credible intervals for $(\theta_i|Y_i = y_i)$, whereas the former are, by construction, credible intervals for $(\theta_i|Y = y)$. By using more data, we have more knowledge. In fact, the ratio of interval widths (shrunk to unshrunk) ranges from 0.893 (for unit 6) to 0.968 (for unit 1). While we should resist the temptation to "read too much" into what happens for a particular data vector, the reduction in interval widths, plus the aforementioned reduction in estimation error, are teasingly appealing!

A modest extension: Datapoints of varying precision

We have started with a very simple sandbox in which to explore shrinkage, particularly in assuming that for each of I units, Y_i arises from its underlying mean θ_i via a known amount of normal variation, with this amount being the same for each unit. In practice, it is very easy to imagine situations where some datapoints are more precise than others, which we represent with $Y_i \sim N(\theta, \sigma_i^2)$, with σ_i known.

Under the hood we provide a function **posterior.B()** to handle this modest extension. In brief, whereas previously we could leverage importance sampling to draw Monte Carlo realizations from the posterior of (μ, τ, θ), now we have to fuss slightly more. With unequal σ_i we can obtain a closed-form expression for the posterior marginal density of τ, but this expression does not correspond to a standard distribution. Hence we evaluate the expression on a fine grid, in order to draw Monte Carlo samples of $(\tau|Y = y)$. After this we can draw from $(\mu|\tau, y)$ and $(\theta|\mu, \tau, y)$, both of which are normal distributions.

An obvious test of our coding is that output of **posterior.B()** applied to a problem with equal variances needs to match the output of **posterior.A()**, up to some numerical tolerance.

```
ans.check <- posterior.B(y=y.test, sig=rep(1,length(y.test)),
                         mn.mu=0, sd.mu=10, tau.upr=50,
                         lg.pri.tau=lg.pri.tau)
```

If all is working well, both functions are providing draws from the same (posterior) distribution. So frequentist hypothesis tests applied to the Monte Carlo output should not find evidence to the contrary.

```
### for instance, should not detect a difference in terms of the
### mean of the marginal posterior distribution of tau

t.test(ans.test$tau, ans.check$tau)[c("statistic","p.value")]
```

```
## $statistic
##     t
## -1.02
##
## $p.value
## [1] 0.309
```

```
### for instance, should not detect a difference in terms of the
### mean of the marginal (posterior) distribution of theta[3]

t.test(ans.test$theta[,3], ans.check$theta[,3])[c("statistic","p.value")]
```

```
## $statistic
##     t
## 0.525
##
## $p.value
## [1] 0.599
```

Application to meta-analysis

We can employ our hierarchical model in the task of *meta-analysis*. Here the task is to analyze a set of medical studies as a whole, rather than one-by-one. As an example, we take data from $I = 17$ randomized trials examining post-operative survival of patients with malignant gliomas. Specifically, each trial randomizes patients to either radiotherapy alone (the "control" group/arm) or radiotherapy plus adjuvant chemotherapy (the "experimental" or "treatment" or "active treatment" group/arm).

```
### For more background and references for these data:

require("metadat")
help(dat.fine1993)
```

```
## Loading required package: metadat
```

As notation, we take the number of patients in the j-th arm of the i-th trial to be m_{ij}, with $j = 0$ and $j = 1$ representing control and treatment respectively. And we observe $V_{ij} = v_{ij}$ out of the m_{ij} patients to experience the outcome event, which in our example is survival beyond 12 months post-surgery:

```
### wrangling the data into our notation

dat <- data.frame(
  m0=dat.fine1993$nci, v0=dat.fine1993$c2i,
  m1=dat.fine1993$nei, v1=dat.fine1993$e2i)

num.studies <- dim(dat)[1]

dat
```

```
##     m0 v0  m1  v1
## 1   22 12  19  11
## 2   35 12  34  18
## 3   68 15  72  21
## 4   20  5  22  14
## 5   32 13  70  42
## 6   94 33 183  80
## 7   50 18  26  13
## 8   55 30  61  37
## 9   25 12  36  23
## 10  35 14  45  19
## 11 208 76 246 106
## 12 141 46 386 170
```

```
## 13  32 17  59  34
## 14  15  3  45  18
## 15  18 14  14  13
## 16  19 10  26  12
## 17  75 40  74  42
```

The starting point for meta-analysis is to regard the i-th trial data as drawn from its own population, with its own *treatment effect* as the target parameter. We take this target θ_i to be the *log-odds-ratio* describing the association between being on active treatment (rather than control) and experiencing the outcome event (versus not). It is generally plausible that the treatment effects will be *similar*, but *not identical*, across trials. (Identical is generally quite a stretch. Inevitably, different trials of the same therapy will have somewhat different clinical protocols, and will be drawing patients from different geographic populations.) So completing the hierarchical model specification by presuming $\theta_i \sim N(\mu, \tau^2)$, with μ and τ unknown, fits the bill. We regard μ as a "typical" treatment effect across study populations, which indeed will be *a priori* unknown. And then τ describes the across-study variation in treatment effect, which is also unknown *a priori*.

A first-principles approach would be to consider the i-th trial data as a pair of binomial observations. To pursue this formulation, in addition to parameters θ describing the study-specific treatment effects, we would need parameters, say κ, to describe the trial-specific outcome rates in control populations. Then given κ and θ, all the counts V_{ij} could be viewed as mutually independent binomial observations of the form:

$$V_{0i} \sim \text{binomial}\{m_{0i}, \text{expit}(\kappa_i)\},$$
$$V_{1i} \sim \text{binomial}\{m_{1i}, \text{expit}(\kappa_i + \theta_i)\}.$$

It would be quite reasonable, and indeed recommended, to integrate these binomial distributions directly into the hierarchical model specification. For pedagogical purposes, however, we do something simpler, but still quite reasonable. It turns out that the information about θ_i contained in the binomial counts (V_{i0}, V_{i1}) is very well approximated by the information about θ_i contained in Y_i, presuming $Y_i \sim N(\theta_i, \sigma_i^2)$, where

$$Y_i = \text{logit}(V_{1i}/m_{1i}) - \text{logit}(V_{0i}/m_{0i}),$$

and the (presumed) known value of σ_i^2 is taken to be

$$\sigma_i^2 = \frac{1}{v_{0i}} + \frac{1}{m_{0i} - v_{0i}} + \frac{1}{v_{1i}} + \frac{1}{m_{1i} - v_{1i}}. \tag{5}$$

In applied statistics parlance, Y_i is the sample log-odds-ratio from trial i, while σ_i is the corresponding standard error. Note that we now have a version of our problem which exactly matches the "varying precision" version of our simple hierarchical model. And we know this model will produce an inference about θ_i (the population treatment effect underlying the

i-th clinical trial) which will depend on the summarized Y data from all the trials, not just the i-th trial.

As a sidebar for those interested, (5) arises from a "delta-method" approximation. Its form, as the sum of reciprocal cell counts from the 2×2 table cross-classifying the binary treatment and outcome statuses, is curious. Most particularly, the smallest of the four cell counts will be the weak link, making the largest (often by far) contribution to the estimated uncertainty of the study-specific sample log-odds-ratio as an estimate of its population counterpart.

We can easily wrangle together the needed data vector y and corresponding standard deviations σ:

```
### data wrangling, each trial has a y[i] and sig[i]

y.meta <- logit(dat$v1/dat$m1) - logit(dat$v0/dat$m0)

sig.meta <- sqrt(1/dat$v0 + 1/(dat$m0-dat$v0) + 1/dat$v1 + 1/(dat$m1-dat$v1))
```

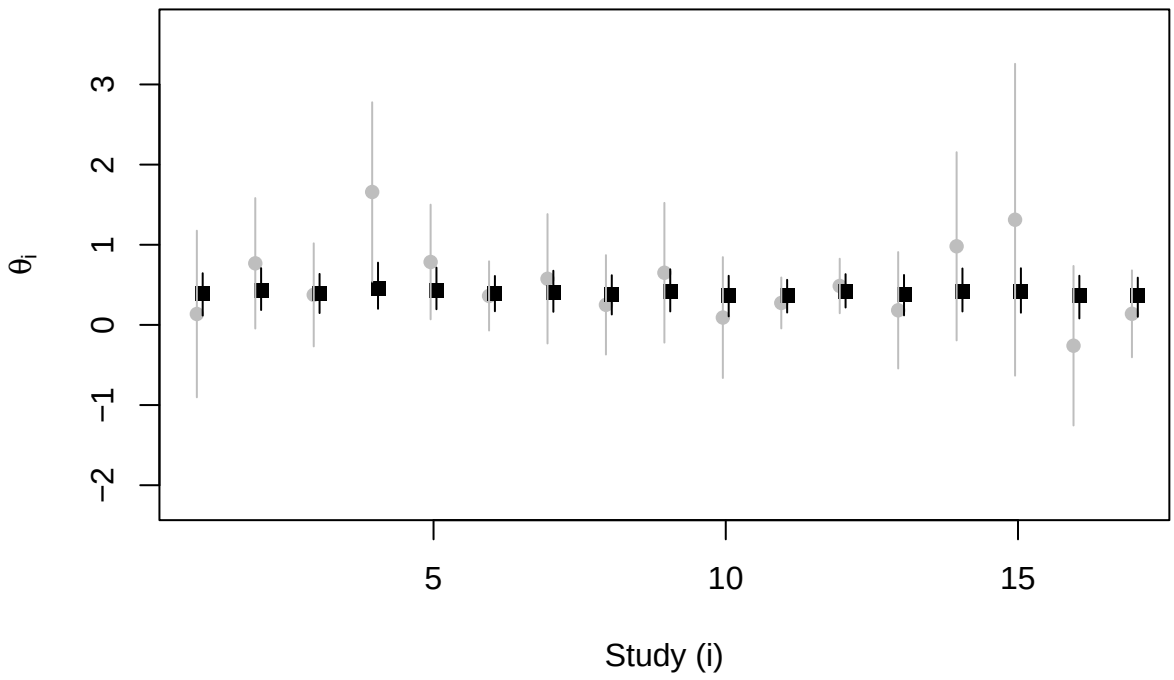
Now we are all set, as we have the inputs to produce the posterior distribution of (θ, μ, τ), which we obtain (again in the form a large Monte Carlo sample) as follows:

```
ans.meta <- posterior.B(y=y.meta, sig=sig.meta,
                        mn.mu=0, sd.mu=10, tau.upr=1,
                        lg.pri.tau=lg.pri.tau)
```

As in our earlier example, we can look at both point and interval estimation, both without, and with, shrinkage. To be clear, "with" involves the posterior mean and 90% equal-tailed credible interval for each θ_i based on the whole data vector Y. Whereas "without" takes y_i as the point estimate of θ_i and $y_i \pm 1.645\sigma_i$ as the 90% (frequentist) interval estimate, as befits using only Y_i when estimating θ_i.

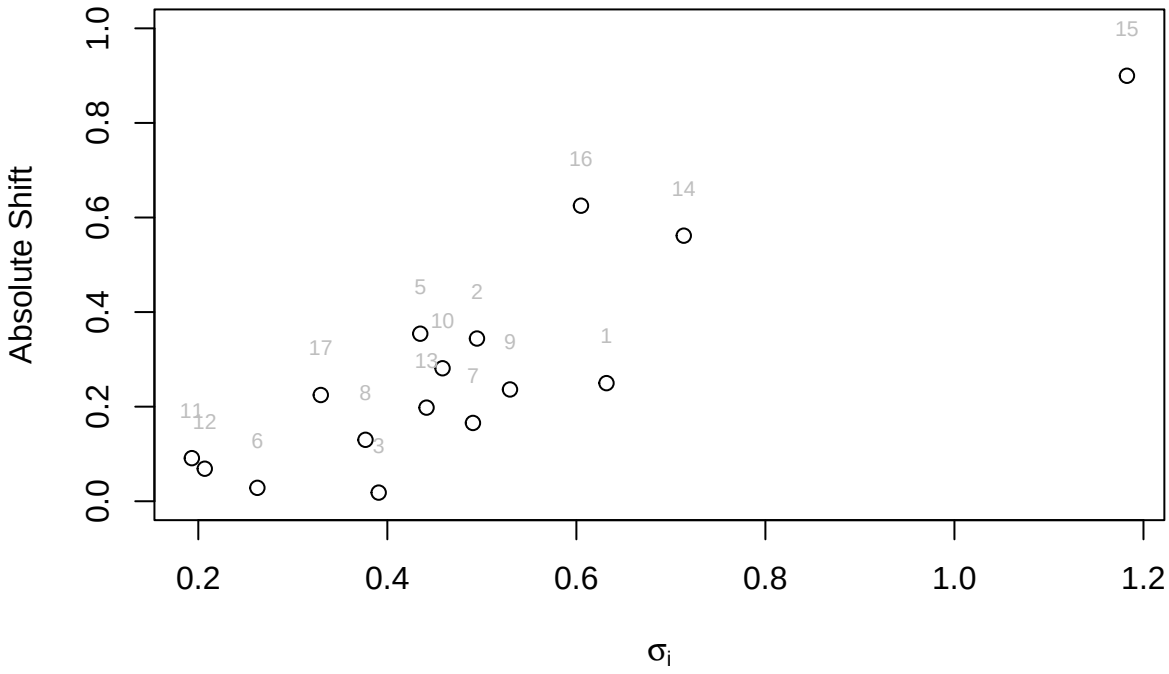
```
pst.mn <- apply(ans.meta$theta, 2, mean)
int.90 <- apply(ans.meta$theta, 2, quantile, probs=c(0.05, 0.95))
```

```
jtr <- 0.05
plot(  1:num.studies - jtr, y.meta,   pch=16, col="grey", ylim=c(-2.2,3.7),
       xlab="Study (i)", ylab=expression(theta[i]))
points(1:num.studies + jtr, pst.mn, pch=15)

for (i in 1:num.studies) {
  points(rep(i + jtr,2), int.90[,i], type="l")
  points(rep(i - jtr,2), y.meta[i]+1.645*sig.meta[i]*c(-1,1), col="grey",type="l")
}
```



The figure reinforces earlier messages. We see shrinkage (i), moving estimates toward the "middle of the pack", and (ii), reducing interval width. But we also see an additional feature at play now. Some units, need, and receive, considerably more help than others. Very intuitively, the units with less precise datapoints garner the greater interventions. That is, we tend to see a larger absolute shift, $|y_i - E(\theta_i|Y = y)|$, when σ_i is larger:

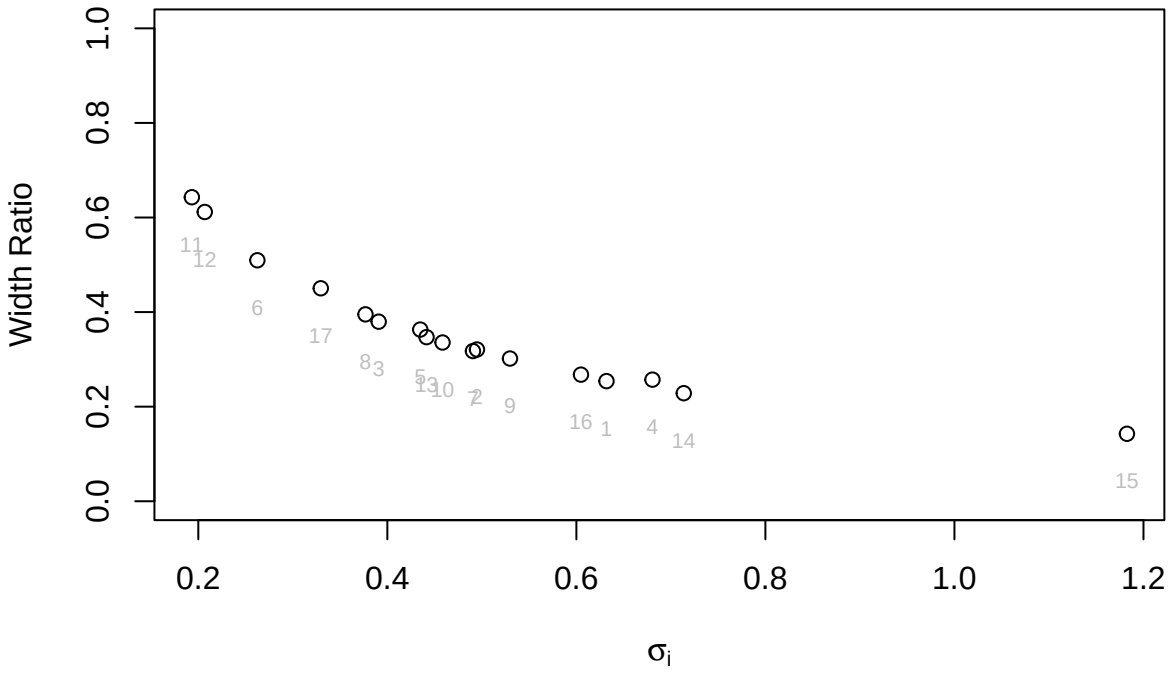
```
plot(sig.meta, abs(pst.mn-y.meta), ylim=c(0,1),
     xlab=expression(sigma[i]), ylab="Absolute Shift")

jtr <- 0.1
text(sig.meta, abs(pst.mn-y.meta) + jtr, as.character(1:num.studies),
     cex=0.7, col="grey")
```



Note though that this relationship is only moderately strong. Particularly, if the i-th unit has a "middle-ish" value of y, then it won't see much of a shift, regardless of its corresponding precision.

Turning to interval estimates, we find a stronger sense in which units with less precise datapoints receive more of an intervention. The shrunk interval width is a smaller fraction of the unshrunk interval width when σ_i is larger:

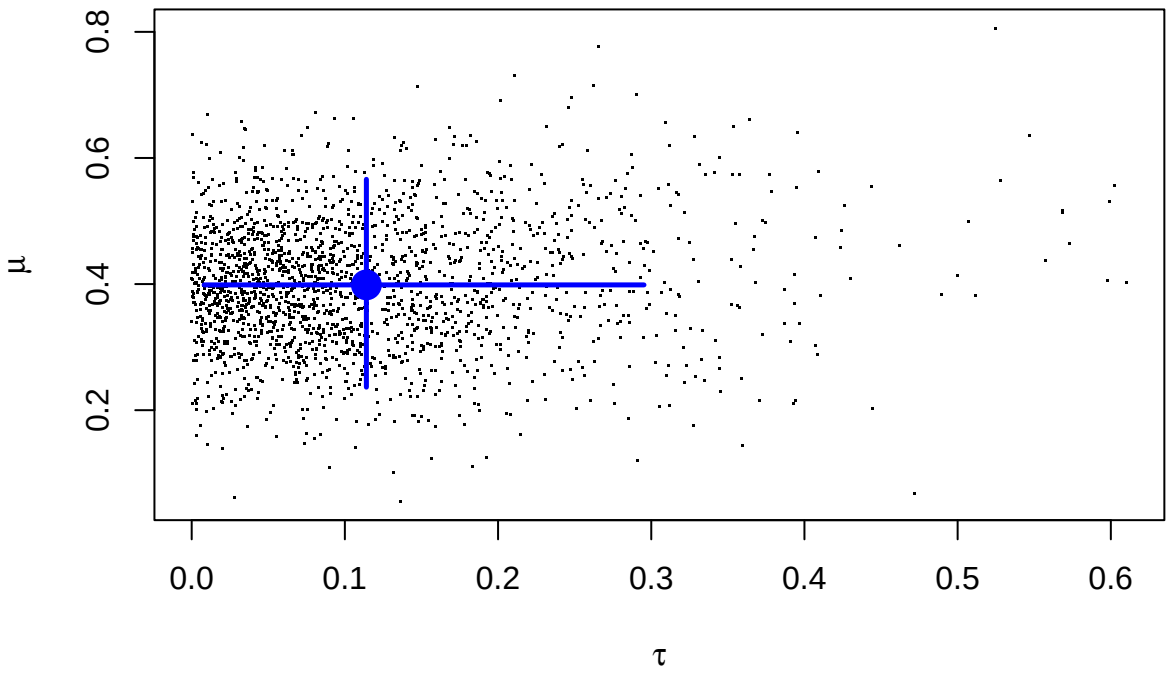
```
plot(sig.meta, (int.90[2,]-int.90[1,])/(2*1.645*sig.meta), ylim=c(0,1),
     xlab=expression(sigma[i]), ylab="Width Ratio")

jtr <- 0.1
text(sig.meta, (int.90[2,]-int.90[1,])/(2*1.645*sig.meta) - jtr,
     as.character(1:num.studies), cex=0.7, col="grey")
```



Inference on second-stage parameters

So far our discussion of borrowing of strength and shrinkage has been predicated on trying to learn about the first-stage (unit-specific) parameters $\theta = (\theta_1, \ldots \theta_I)$. In application to meta-analysis, however, parameters μ and τ^2 are generally of greater inferential interest. Foremost, since μ is regarded as the typical effect of treatment across different populations, inference about it is reported as a global summary of treatment efficacy. Secondarily, there is inherent interest in τ as the descriptor of treatment efficacy is across different populations.

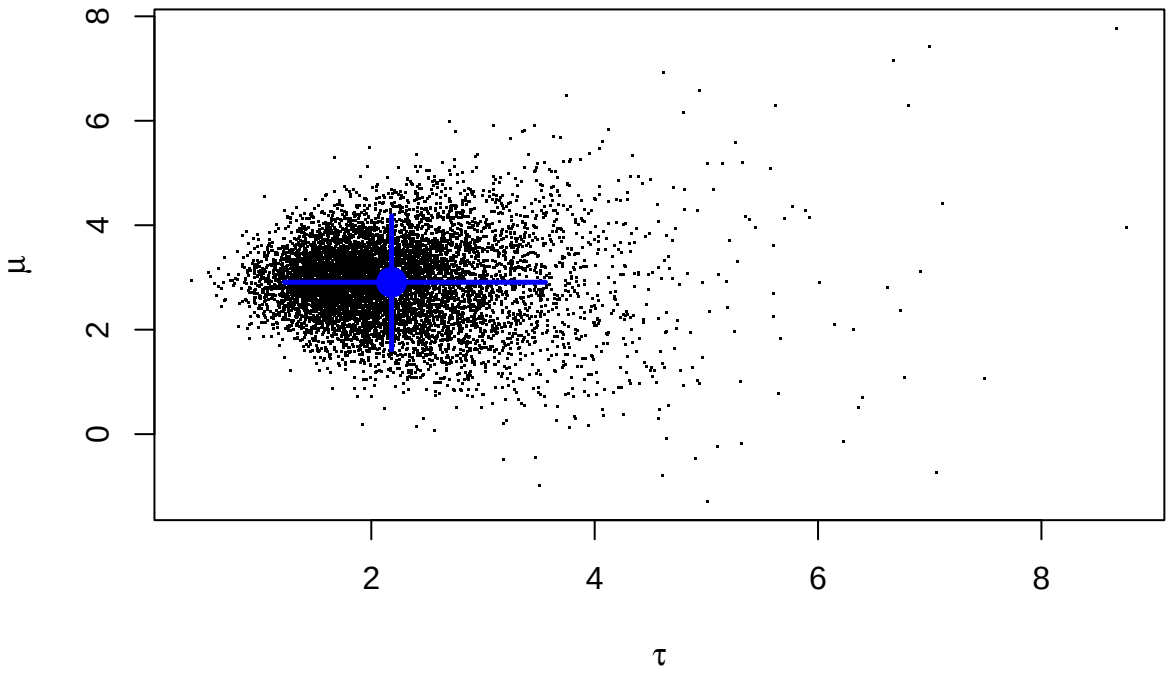
For the present glioma data, we can visualize the bivariate posterior distribution of (μ, τ), along with the posterior mean and equal-tailed 90% credible interval for each parameter:



With a point estimate for μ of $E(\mu|Y = y) = 0.4$, along with the 90% equal-tailed credible interval for μ running from 0.24 to 0.57, we have evidence that the treatment is beneficial in a typical population. Note that we can also report inferences on the more interpretable odds-ratio scale. For instance, we simply exponentiate the endpoints of the credible interval for μ, yielding $(1.27, 1.76)$ as the 90% equal-tailed credible interval for the typical odds ratio e^{μ}.

A feature of the bivariate posterior distribution of (μ, τ) depicted above is a weak dependence, whereby there is slightly more variation in μ when τ is larger. This is intuitive. Larger values of τ correspond to less commonality amongst the elements of θ. In turn this reduces the extent of borrowing-of-strength, such than more uncertainty about μ remains. For the glioma data, however, this dependence is indeed quite weak, since generally the posterior favors quite small values of τ.

To explore the posterior dependence between μ and τ in a situation where it is more evident, we return to the earlier synthetic data vector (with $I = 10$ elements) used to illustrate the simpler case of equally-precise datapoints (known that $\sigma_i = 1$ for all i). We examine the joint posterior distribution of (μ, τ) for these data (again with posterior means and 90% equal-tailed credible intervals given for reference):



In this case there is wider uncertainty about τ, hence the posterior association between τ and μ is more evident. Note also that these data effectively rule out values of τ very close to zero. The data tell us that assuming a single common value for the elements of θ would be inappropriate. As a sanity check here, we momentarily probe the implications if we *did* make this assumption. A 90% frequentist confidence interval for the common mean would be $\bar{y} \pm 1.645(1/10)^{1/2}$, which works out to be $(2.41, 3.45)$. This is considerably narrower than the 90% equal-tailed credible interval based on the posterior distribution of $(\mu|Y = y)$ depicted above, which happens to be $(1.62, 4.18)$. Thus the hierarchical model analysis is guarding against overconfidence that would arise from the pretense that all the datapoints arose from the same underlying mean. As a related point, by eyeballing the plot above we discern that the (overconfident) "fully pooled" interval estimate of $(2.41, 3.45)$ is roughly commensurate with what we can visualize for the posterior distribution of μ conditioned on τ being close to zero.

Wrap-up

Whether we are estimating first-stage or second-stage parameters, the hope is that the demonstrations above have whetted your appetite to learn more about Bayesian hierarchical models. The ideas of shrinkage and borrowing-of-strength are powerful, and can be employed in a multitude of statistical settings. And while there are non-Bayesian routes into combining data arising under similar but not identical circumstances, these can be "work" in terms of

establishing frameworks and principles. In contrast, shrinkage and borrowing-of-strength fall out naturally and effortlessly once a Bayesian hierarchical model has been specified. For a deeper dive on this, see Chapter 6 of *Bayesian Statistical Inference: Concepts, Hallmarks, and Operating Characteristics*.