STAT 8700 Homework 5

Brian Detweiler

Friday, September 30, 2016

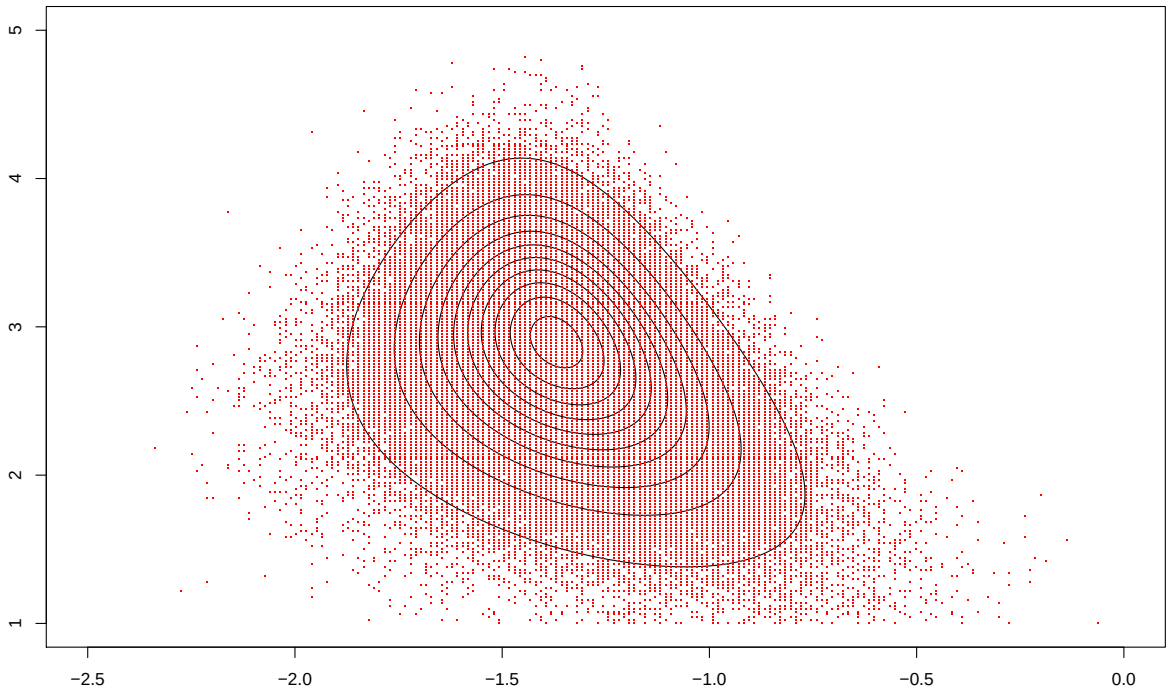
1. Consider the data provided in Table 3.3. We are only going to use the data listed in the top section, that is where Type of Street is residential, and Bike Route is Yes.

```
bikes <- data.frame('y' = c(16, 9, 10, 13, 19, 20, 18, 17, 35, 55),
                    'other' = c(58, 90, 48, 57, 103, 57, 86, 112, 273, 64))
bikes$N <- bikes$y + bikes$other
```

(a) Set up a model for the data so that, for $j = 1, \ldots, 10$, the observed number of bicycles at location j is binomial with unknown probability θ_j and sample size equal to the total number of vehicles (bicycles included) at that location. The parameter θ_j can be interpreted as the underlying or "true"" proportion of traffic at location j that is bicycles. Assign a beta population distribution (prior) for the parameters θ_j and a noninformative hyperprior distribution as in the rat tumor example of Section 5.3. Follow the rat tumor example, and obtain posterior simulations for the θ_j's.

```
log.post2 <- function(u, v, data) {


  success <- data$y
  N <- data$N
  alpha <- exp(u + v) / (1 + exp(u))
  beta <- exp(v) / (1 + exp(u))
  ldens <- 0

  for(i in 1:length(N)) {
    ldens <- (ldens
              + (lgamma(alpha + beta) + lgamma(alpha + success[i]) + lgamma(beta + N[i] - success[i]))
              - (lgamma(alpha) + lgamma(beta) + lgamma(alpha + beta + N[i])))
  }

  ldens - 5 / 2 * log(alpha + beta) + log(alpha) + log(beta)
}

contours <- seq(0.05, 0.95, 0.1)
u2 <- seq(-2.5, 0, length = 200)
v2 <- seq(1, 5, length = 200)
logdens2 <- outer(u2, v2, log.post2, bikes)
dens2 <- exp(logdens2 - max(logdens2))
contour(u2, v2, dens2, levels = contours, drawlabels = FALSE)
```

```r
nsim <- 100000
dens.u <- apply(dens2, 1, sum)
uindex <- sample(1:length(u2), nsim, replace = TRUE, prob = dens.u)
sim.u <- u2[uindex]
sim.v <- rep(NA, nsim)

for (i in (1:nsim)) {
  sim.v[i] <- sample(v2, 1, prob = dens2[uindex[i], ])
}

points(sim.u, sim.v, col="red", pch='.')
```

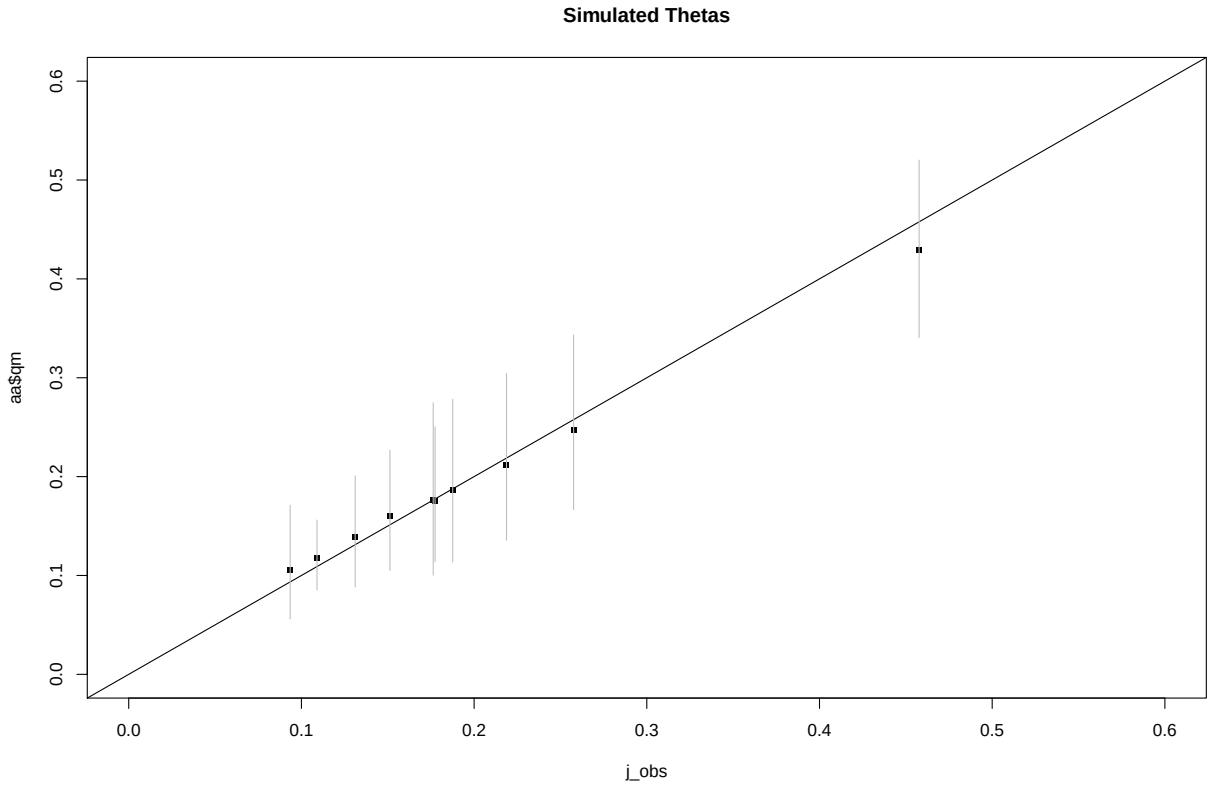



```r
sim.alpha <- exp(sim.u + sim.v) / (1 + exp(sim.u))
sim.beta <- exp(sim.v) / (1 + exp(sim.u))
theta.sim <- matrix(NA, nrow = nsim, ncol = length(bikes$N))

for (j in 1:length(bikes$N)) {
    theta.sim[,j] <- rbeta(nsim, sim.alpha + bikes$y[j], sim.beta + bikes$N[j] - bikes$y[j])
}
```

■

(b) As in the rat tumor example, compare the posterior distributions of the θ_j's to the observed proportion of bikes at each location.

```
aa <- data.frame(obs = bikes$y / bikes$N,
                 q025 = apply(theta.sim, 2, quantile, 0.025),
                 qm = apply(theta.sim, 2, quantile, 0.5),
                 q975 = apply(theta.sim, 2, quantile, 0.975))

j_obs = jitter(aa$obs, amount=.005)

plot(j_obs, aa$qm, xlim = c(0, 0.6), ylim = c(0, 0.6), pch = 15, cex = 0.75, main = 'Simulated Thetas')

abline(a = 0, b = 1)

for (i in 1:length(bikes$N)){
  lines(c(j_obs[i], j_obs[i]), c(aa$q025[i], aa$q975[i]), col ="grey")
}
```



■

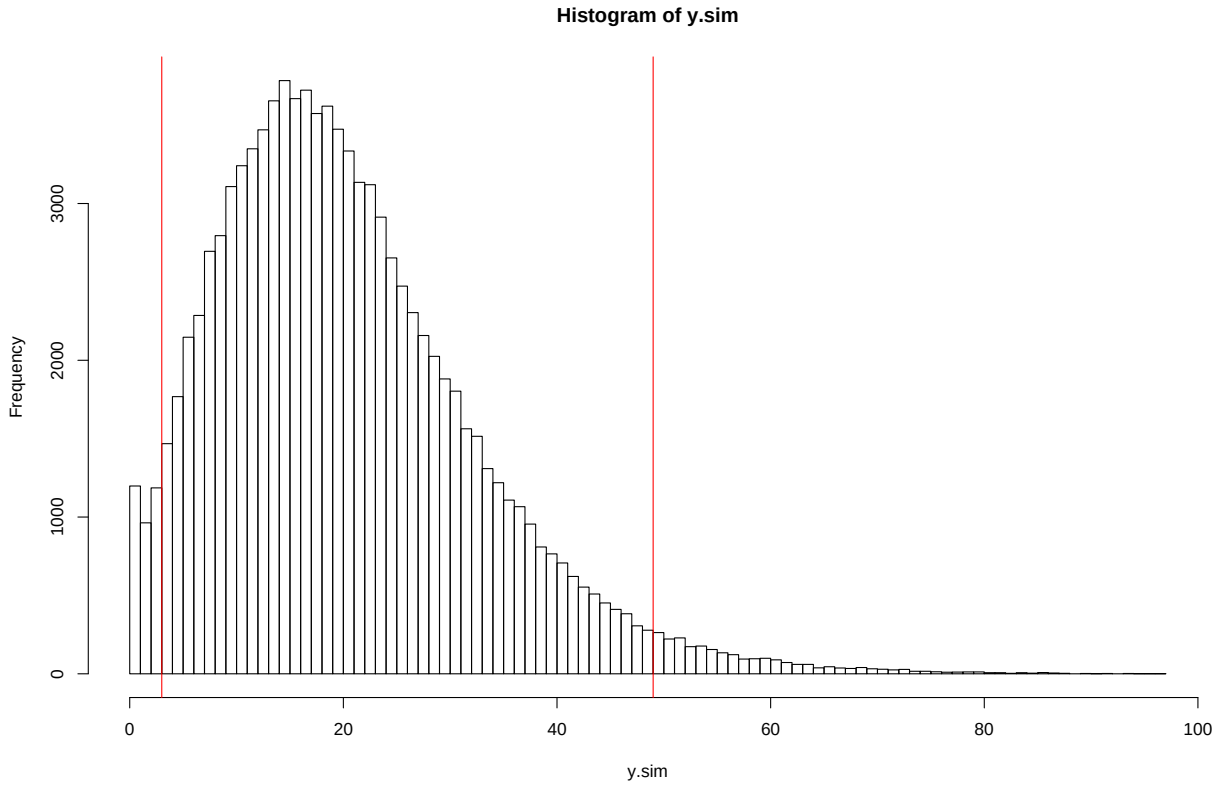
(c) Suppose we plan to visit an 11th location to observe the traffic. Use the prior distribution to simulate possible values for θ_{11}.

```
nsim <- 100000
theta11.sim <- rbeta(nsim, sim.alpha, sim.beta)
```

■

(d) Now suppose at the 11th location, 100 vehicles of all kinds go by in one hour of observation, simulate the number of bicycles that pass during that hour, and construct a 95% interval.

```r
# Use the mean of all the N_i's
N <- 100
y.sim <- rbinom(n = 100000, size = N, prob = theta11.sim)
y.sim.sorted <- sort(y.sim)
q025 <- y.sim.sorted[length(y.sim.sorted) * 0.025]
q975 <- y.sim.sorted[length(y.sim.sorted) * 0.975]
hist(y.sim, breaks = 100)
abline(v = q025, col='red')
abline(v = q975, col='red')
```

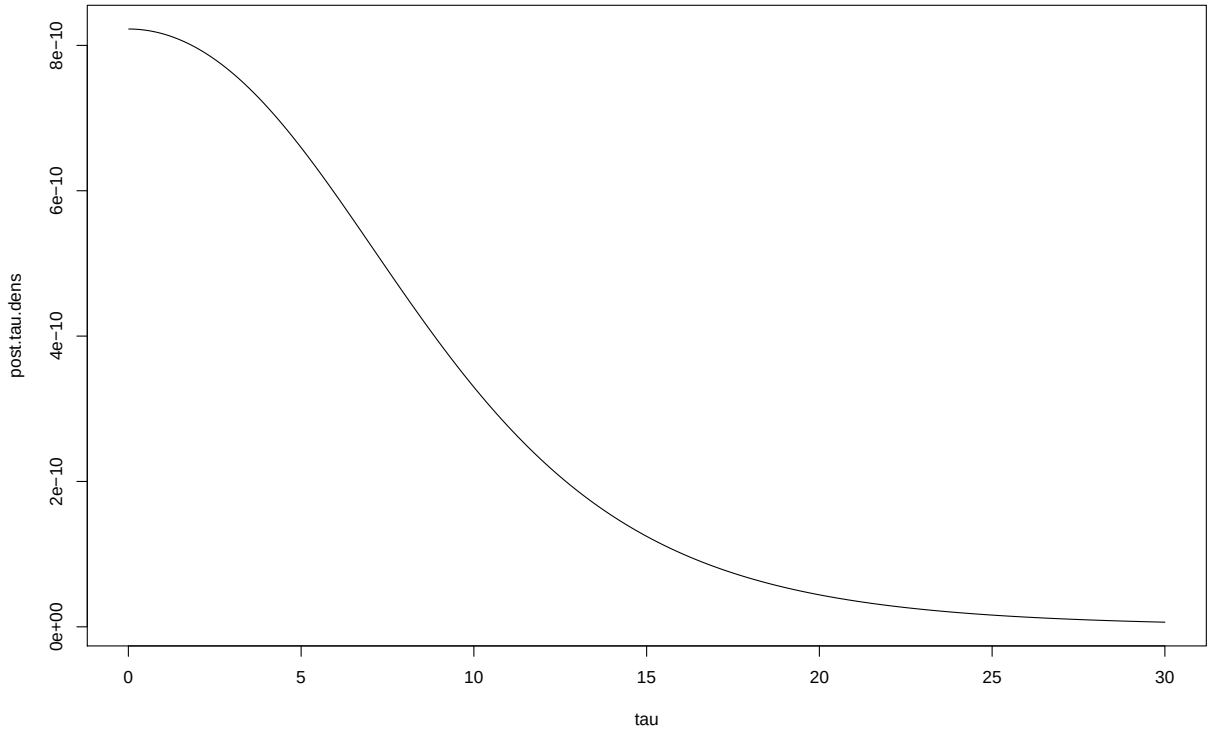


A 95% credible interval for y_{11} is (3, 49).

■

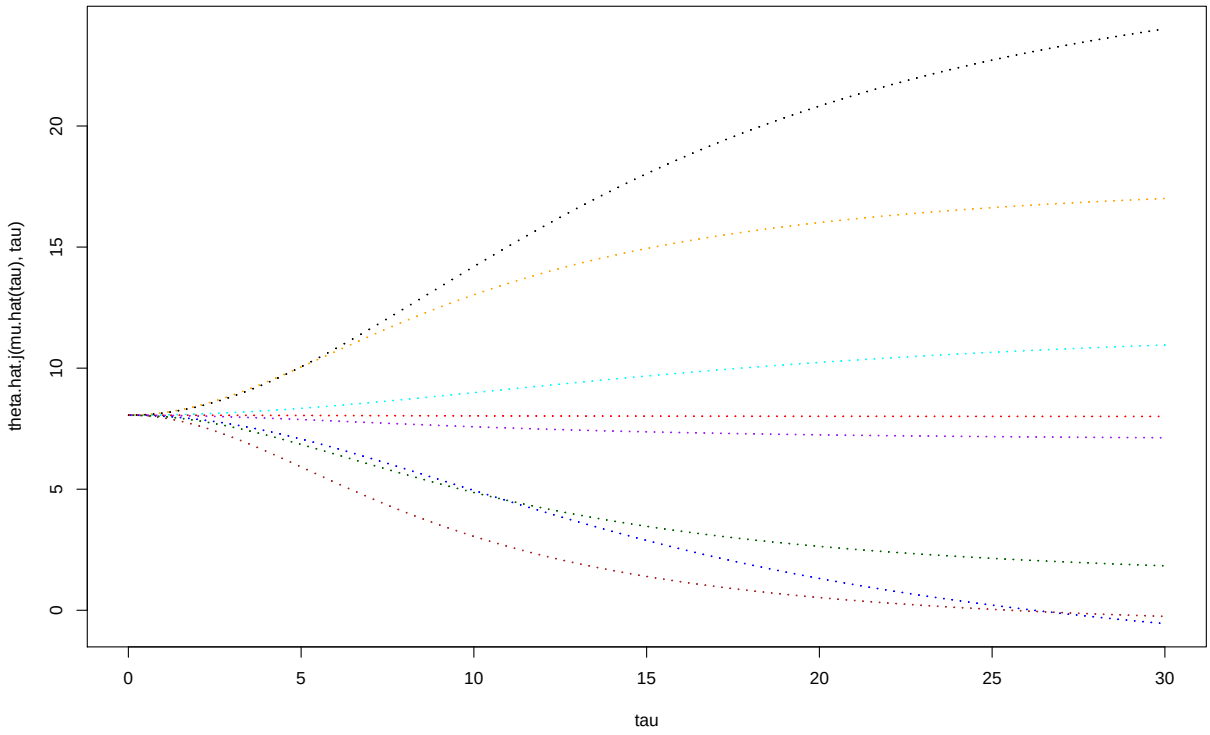
2. Consider the Schools example.

```
schooldata.y <- c(28, 8, -3, 7, -1, 1, 18, 12)
schooldata.sigmaj <- c(15, 10, 16, 11, 9, 11, 10, 18)

mu.hat <- function(tau) {
  sum ((1 / (schooldata.sigmaj^2 + tau^2)) * (schooldata.y)) /
    sum((1/(schooldata.sigmaj^2 + tau^2)))
}

V.mu.inv <- function(tau) {
  sum((1 / (schooldata.sigmaj^2 + tau^2)))
}

# p(tau | y)
post.tau <- function(tau) {
  V.mu.inv(tau)^(-(1 / 2)) *
    prod(((schooldata.sigmaj^2 + tau^2)^(-1 / 2)) *
           exp(-((schooldata.y - mu.hat(tau))^2) / (2 * (schooldata.sigmaj^2 + tau^2)))
         )
}

# Simulate a bunch of tau's using the post.tau() function
tau <- seq(0.01, 30, length = 1000)
post.tau.dens <- apply(as.array(tau), 1, FUN = "post.tau")

# The plot looks like a slide, so smaller taus are more likely
plot(tau, post.tau.dens, "l")
```

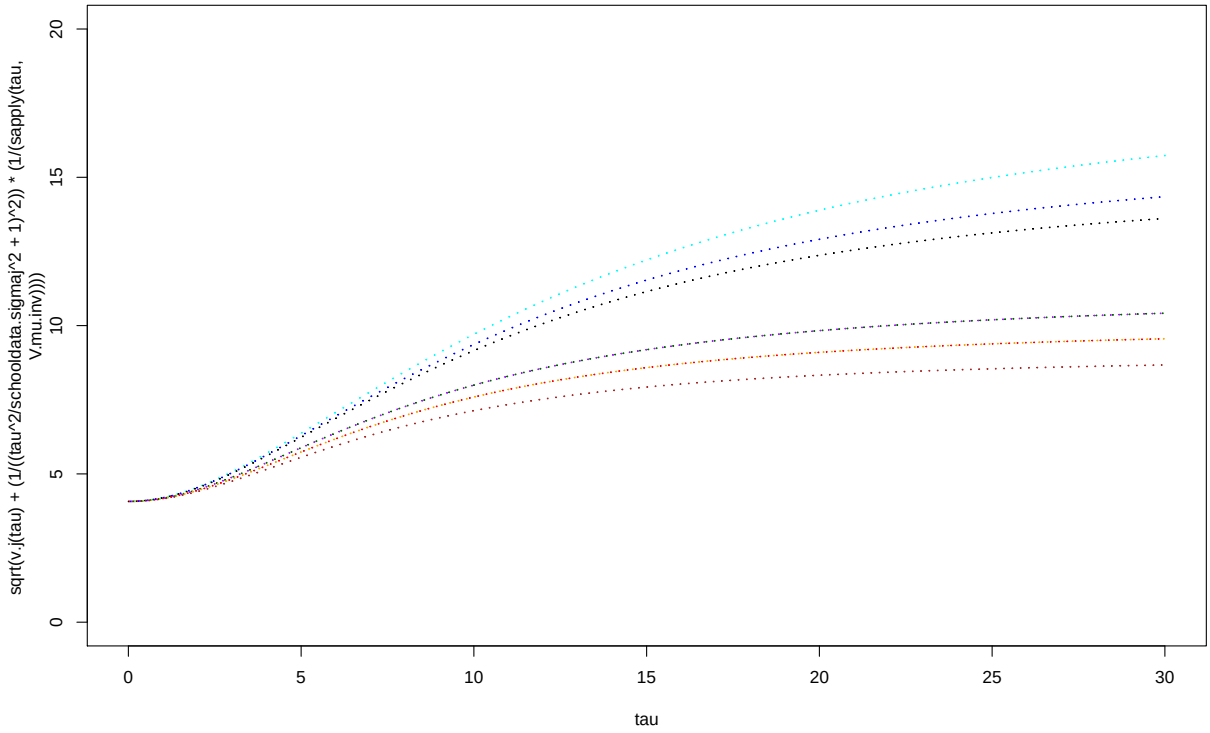



```
# Now take 200 samples with replacement from the taus, using the tau posterior density
sim.tau <- sample(tau, 200, prob = post.tau.dens, replace = TRUE)
# Now we can simulate mu.hat and V.mu.inv
sim.mu.hat <- apply(as.array(sim.tau), 1, FUN = "mu.hat")
sim.Vinv <- apply(as.array(sim.tau), 1, FUN="V.mu.inv")
sim.mu <- rnorm(200, sim.mu.hat, (sim.Vinv)^(-1/2))

v.j <- function(tau) {
  1 / ((1 / schooldata.sigmaj^2) + (1 / tau^2))
}

theta.hat.j <- function(mu, tau) {
  ((schooldata.y / schooldata.sigmaj^2) + (mu / (tau^2))) /
    ((1 / schooldata.sigmaj^2) + (1 / tau^2))
}

# Figure 5.6
plot(tau,
     theta.hat.j(mu.hat(tau), tau),
     cex = 0.1,
     col = c("black",
             "red",
             "blue",
             "purple",
             "brown",
             "darkgreen",
             "orange",
```

```
            "cyan"))
```




```
##### Figure 5.7
plot(tau,
     sqrt(v.j(tau) +
          (1 / ((tau^2 / schooldata.sigmaj^2 + 1)^2)) *
          (1 / (sapply(tau, V.mu.inv)))),
     cex = 0.1,
     ylim = c(0, 20),
     col = c("black",
             "red",
             "blue",
             "purple",
             "brown",
             "darkgreen",
             "orange",
             "cyan"))
```




```
sim.vj <- apply(as.array(sim.tau), 1, FUN = "v.j")
sim.thetahat.j <- matrix(NA, ncol = 8, nrow = length(sim.mu))

for (j in (1:8)) {
  for (i in (1:length(sim.mu))) {
    sim.thetahat.j[i, j] <- theta.hat.j(sim.mu[i], sim.tau[i])[j]
  }
}

sim.theta <- matrix(NA, ncol = 8, nrow = length(sim.mu))
for (j in (1:8)) {
  for (i in (1:length(sim.mu))) {
    sim.theta[i, j] <- rnorm(1, sim.thetahat.j[i, j], (sim.vj[i])^(1 / 2))
  }
}
```

(a) If τ is set to ∞ then each school would be estimated separately, and the posterior distribution for θ_j is simple. Write down the posterior for each of $\theta_1, \ldots, \theta_8$. What is the probability that θ_1 is greater than θ_7.

The posterior distribution for the extreme case of treating each experiment independently is

$$\theta_j|y \sim Normal(\overline{y}_j, \frac{\sigma^2}{n_j}), \text{ s.t. } \overline{y}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j}, \frac{\sigma^2}{n_j} = \sigma_j^2$$

$$\begin{aligned}
\theta_1|y &\sim Normal(28, 15) \\
\theta_2|y &\sim Normal(8, 10) \\
\theta_3|y &\sim Normal(-3, 16) \\
\theta_4|y &\sim Normal(7, 11) \\
\theta_5|y &\sim Normal(-1, 9) \\
\theta_6|y &\sim Normal(1, 11) \\
\theta_7|y &\sim Normal(18, 10) \\
\theta_8|y &\sim Normal(12, 18)
\end{aligned}$$

The probability that $\theta_1 > \theta_7$ is given by $P(\theta_1 - \theta_7 > 0)$, where $\theta_1 - \theta_7 \sim Normal(28 - 18, 15 + 18)$.

$$\begin{aligned}
P(\theta_1 - \theta_7 > 0) &= P(X - Y > 0), \text{ s.t. } X - Y \sim Normal(10, 33) \\
&= 0.6190666
\end{aligned}$$

■

(b) If τ is set to 0, then we are assuming that all schools share a common θ, write down the posterior for this common θ.

With a pooled estimate, we treat all observations as coming from the same θ.

$$\theta|y \sim Normal\left(\overline{y}, \frac{\sigma^2}{n}\right)$$

Where

$$\overline{y} = \frac{\sum_{j=1}^{J} n_j \overline{y}_j}{\sum_{j=1}^{J} n_j}, \text{ s.t. } n_j = \frac{\sigma^2}{\sigma_j^2}$$

Substituting in $\frac{\sigma^2}{\sigma_j^2}$ for n_j, we get

$$\begin{aligned}
\overline{y} &= \frac{\sum_{i=1}^{J} \frac{\sigma^2}{\sigma_j^2} \overline{y}_j}{\sum_{i=1}^{J} \frac{\sigma^2}{\sigma_j^2}} \\
&= \frac{5.5619949}{0.6776515} \\
&= 8.2077511
\end{aligned}$$

Now we can get $\frac{\sigma^2}{n}$ using a similar subsitution and the fact that $\sum_{j=1}^{J} n_j = n$,

$$\begin{aligned}
\overline{y} &= \frac{\sum_{i=1}^{J} \frac{\sigma^2}{\sigma_j^2} \overline{y}_j}{\sum_{i=1}^{J} n_j} \\
n \cdot \overline{y} &= \sum_{i=1}^{J} \frac{\sigma^2}{\sigma_j^2} \overline{y}_j \\
n \cdot 8.2077511 &= \sigma^2 \cdot 5.5619949 \\
\frac{\sigma^2}{n} &= 0.6776515
\end{aligned}$$

Now we have

$$\theta|y \sim Normal(8.2077511, 0.6776515)$$

■

(c) In R, if we have two vectors s and t and we would like to know what proportion of cases the entry in vector s is greater than the corresponding entry in vector t we can use the following command: `sum(s > t) / length(s)`. Use this to calcate the posterior probability in the hierarchical model that θ_1 is greater than θ_7.

```
sum(sim.theta[, 1] > sim.theta[, 7]) / length(sim.theta[, 1])
```

```
## [1] 0.585
```

■

(d) In **R**, if we have a matrix M then the command `table(apply(M, 1, which.max)) / nrow(M)` will calculate the the proportion of rows for which the value in each column is the biggest in the row. Use this to calculate the posterior probability that each school's coaching program is the best of the eight.

```
table(apply(sim.theta, 1, which.max)) / nrow(sim.theta)
```

```
##
##     1     2     3     4     5     6     7     8
## 0.430 0.060 0.035 0.065 0.035 0.065 0.205 0.105
```

■