# 8.1

(a) $var(y_{i\cdot j} \mid \mu, \tau^2)$ 이 $var(y_{i\cdot j} \mid \theta_j, \sigma^2)$ 보다 클 것이다.

btw-group sampling variance도 존재하기 때문이다.

(b) $cov(y_{i_1\cdot j}, y_{i_2\cdot j} \mid \theta_j, \sigma^2) = 0$ $\because$ $\theta_j, \sigma^2$ 가 fixed 일 때 indep

$cov(y_{i_1\cdot j}, y_{i_2\cdot j} \mid \mu, \tau^2) = positive$ $\because$ 서로 같은 $\theta_j$ 와 $\sigma^2$ 를 공유

(c) $var(y_{i\cdot j} \mid \theta_j, \sigma^2) = \sigma^2$

$var(\bar{y}_{\cdot j} \mid \theta_j, \sigma^2) = \frac{\sigma^2}{n}$

$cov(y_{i_1\cdot j}, y_{i_2\cdot j} \mid \theta_j, \sigma^2) = 0$

$$
\begin{aligned}
var(y_{i\cdot j} \mid \mu, \tau^2) &= var(E(y_{i\cdot j} \mid \theta_j, \sigma^2) \mid \mu, \tau^2) + E(var(y_{i\cdot j} \mid \theta_j, \sigma^2) \mid \mu, \tau^2) \\
&= var(\theta_j \mid \mu, \tau^2) + E(\sigma^2 \mid \mu, \tau^2) \\
&= \tau^2 + \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
var(\bar{y}_{\cdot j} \mid \mu, \tau^2) &= var(E(\bar{y}_{\cdot j} \mid \theta_j, \sigma^2) \mid \mu, \tau^2) + E(var(\bar{y}_{\cdot j} \mid \theta_j, \sigma^2) \mid \mu, \tau^2) \\
&= var(\theta_j \mid \mu, \tau^2) + E(\tfrac{\sigma^2}{n_j} \mid \mu, \tau^2) \\
&= \tau^2 + \sigma^2 / n_j
\end{aligned}
$$

$$
\begin{aligned}
cov(y_{i_1\cdot j}, y_{i_2\cdot j} \mid \mu, \tau^2) &= cov(E(y_{i_1\cdot j} \; y_{i_2\cdot j} \mid \theta_j, \sigma^2) \mid \mu, \tau^2) \\
&\quad + E(cov(y_{i_1\cdot j}, y_{i_2\cdot j} \mid \theta_j, \sigma^2) \mid \mu, \tau^2) \\
&= cov(\theta_j, \theta_j) + E(0) \\
&= \tau^2
\end{aligned}
$$

(d) $p(\mu \mid \underbrace{\theta_1, \theta_2, \cdots, \theta_m}_{\theta}, \sigma^2, \tau^2, \underbrace{y_1, \cdots, y_m}_{D})$

$$
\begin{aligned}
&= \frac{p(\mu, \theta, \sigma^2, \tau^2, D)}{\int p(\mu, \theta, \sigma^2, \tau^2, D)\, d\mu} \\
&= \frac{p(\mu)\, \cancel{p(\tau^2)}\, \cancel{p(\sigma^2)}\, \cancel{p(D \mid \theta, \sigma^2)}\, p(\theta \mid \mu, \tau^2)}{\int p(\mu)\, \cancel{p(\tau^2)}\, \cancel{p(\sigma^2)}\, \cancel{p(D \mid \theta, \sigma^2)}\, p(\theta \mid \mu, \tau^2)\, d\mu} \\
&= \frac{p(\mu)\, p(\theta \mid \mu, \tau^2)}{\int p(\mu)\, p(\theta \mid \mu, \tau^2)\, d\mu} = p(\mu \mid \theta, \tau^2)
\end{aligned}
$$

# ESC HW3

*최우현*

*2019년 11월 7일*

# a

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)

schools.list = lapply(1:8, function(i) {
  s.table = paste0('http://www.stat.washington.edu/people/pdhoff/Book/Data/hwdata/school', i,
'.dat') %>%
    url %>%
    read.table

  data.frame(
    school = i,
    hours = s.table[, 1] %>% as.numeric
  )
})
schools.raw = do.call(rbind, schools.list)
Y = schools.raw
```

## Setting priors

```
#mu
mu0<-7
g20<-5

#inverse tau square
eta0<-2
t20<-10

#inverse sigma square
nu0<-2
s20<-15
```

## Starting values

```
m = length(unique(Y[, 1]))
# Starting values - use sample mean and variance
n = sv = ybar = rep(NA, m)
head(Y)
```

```
##   school hours
## 1      1  2.11
## 2      1  9.75
## 3      1 13.88
## 4      1 11.30
## 5      1  8.93
## 6      1 15.66
```

```
for (j in 1:m) {
  Y_j = Y[Y[, 1] == j, 2]
  ybar[j] = mean(Y_j)
  sv[j] = var(Y_j)
  n[j] = length(Y_j)
}
```

8개의 school별로 hours를 구분: Y_j 각 group의 mean value: ybar 각 group의 variance: sv 각 group의 length: n

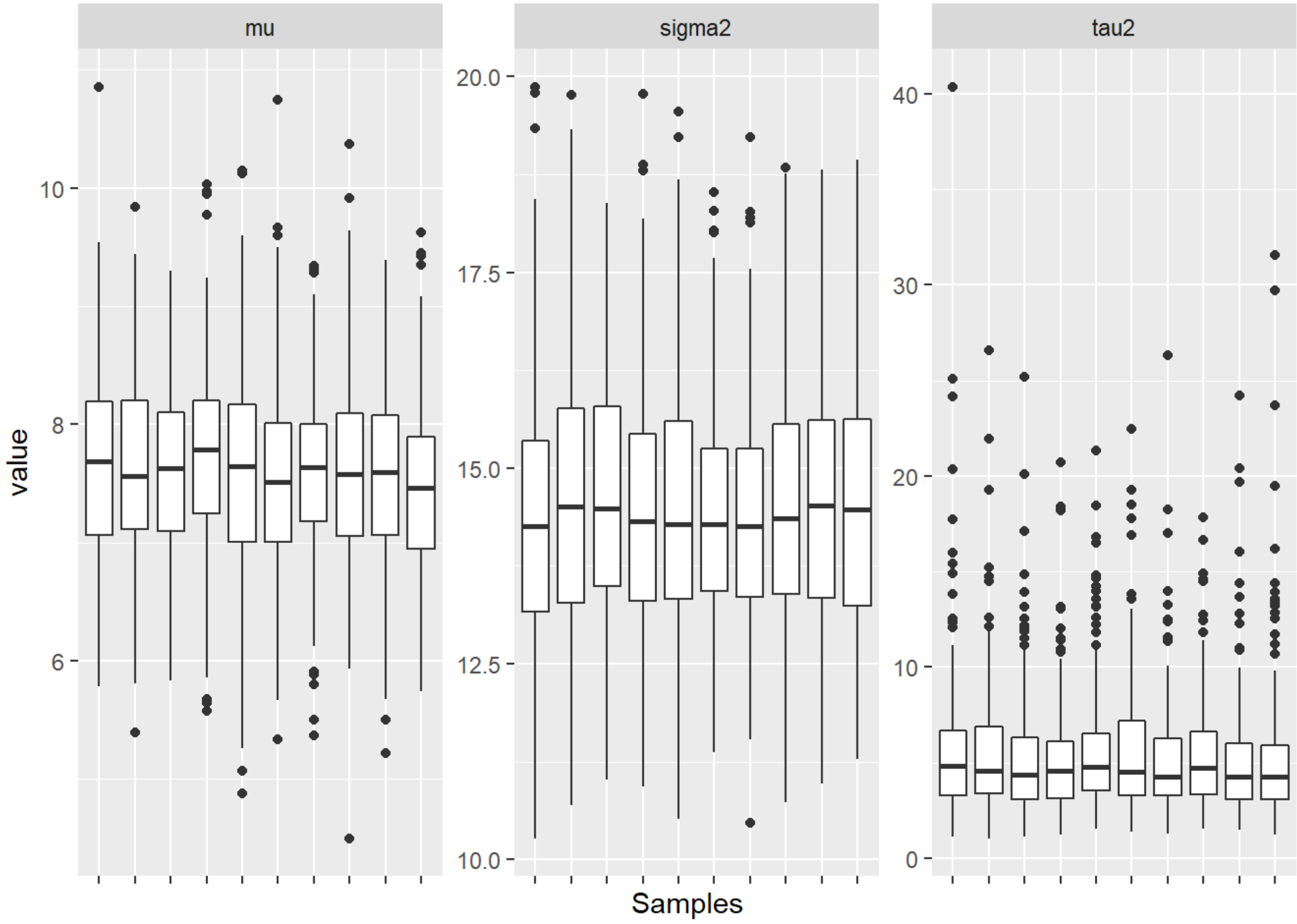
theta-means sigma2-variances mu-means tau2-variances

```
theta = ybar
sigma2 = mean(sv)
mu = mean(theta)
tau2 = var(theta)
```

## setup MCMC

```r
set.seed(2016131012)
S = 2000
THETA = matrix(nrow = S, ncol = m)
# Storing sigma, mu, theta together
SMT = matrix(nrow = S, ncol = 3)
colnames(SMT) = c('sigma2', 'mu', 'tau2')
for (s in 1:S) {
  # Sample thetas
  for (j in 1:m) {
    vtheta = 1 / (n[j] / sigma2 + 1 / tau2)
    etheta = vtheta * (ybar[j] * n[j] / sigma2 + mu / tau2)
    theta[j] = rnorm(1, etheta, sqrt(vtheta))
  }

  # Sample sigma2
  nun = nu0 + sum(n) # TODO: Could cache this
  ss = nu0 * s20
  # Pool variance
  for (j in 1:m) {
    ss = ss + sum((Y[Y[, 1] == j, 2] - theta[j])^2)
  }
  sigma2 = 1 / rgamma(1, nun / 2, ss / 2)

  # Sample mu
  vmu = 1 / (m / tau2 + 1 /g20)
  emu = vmu * (m * mean(theta) / tau2 + mu0 / g20)
  mu = rnorm(1, emu, sqrt(vmu))

  # Sample tau2
  etam = eta0 + m
  ss = eta0 * t20 + sum((theta - mu)^2)
  tau2 = 1 / rgamma(1, etam / 2, ss / 2)

  # Store params
  THETA[s, ] = theta
  SMT[s, ] = c(sigma2, mu, tau2)
}
###
```

Assess convergence with diagnostic boxplots:




Evaluate effective sample size:

```
# Tweak number of samples until all of the below are above 1000
library(coda)
effectiveSize(SMT[, 1])
```

```
## var1
## 2000
```

```
effectiveSize(SMT[, 2])
```

```
##     var1
## 1486.379
```

```
effectiveSize(SMT[, 3])
```

```
##     var1
## 1504.242
```

# b

Posterior means and confidence intervals

```
t(apply(SMT, MARGIN = 2, FUN = quantile, probs = c(0.025, 0.975)))
```

```
##              2.5%     97.5%
## sigma2 11.696551 17.959522
## mu      6.044540  9.103001
## tau2    1.894272 14.381404
```

```
t(apply(SMT, MARGIN = 2, FUN = mean))
```

```
##        sigma2       mu     tau2
## [1,] 14.48816 7.585376 5.433806
```

Comparing posterior to prior:

```
# For dinvgamma
library(MCMCpack)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

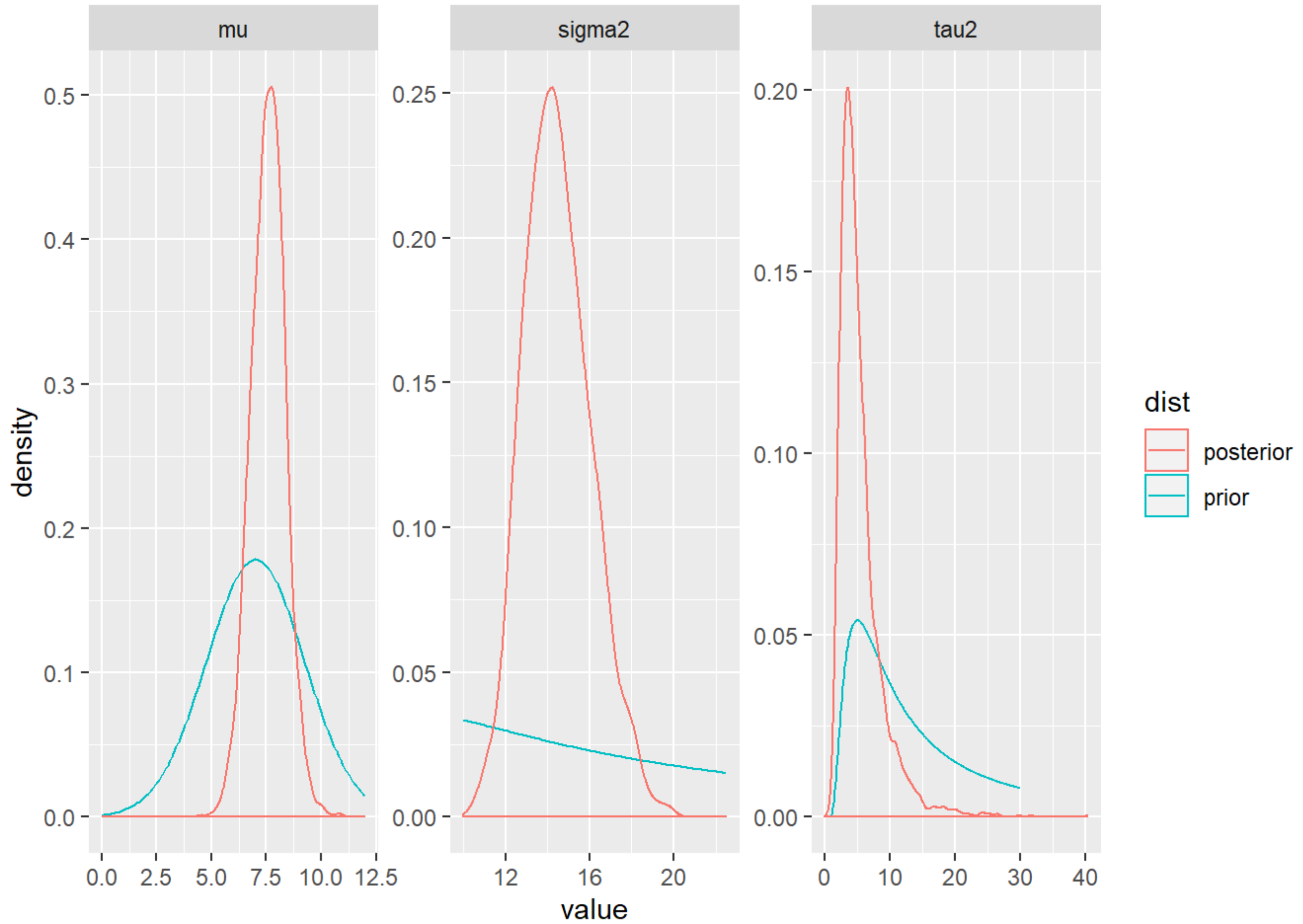
```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
```

```
## ## Copyright (C) 2003-2019 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
```

```
## ##
## ## Support provided by the U.S. National Science Foundation
```

```
## ## (Grants SES-0350646 and SES-0350613)
## ##
```

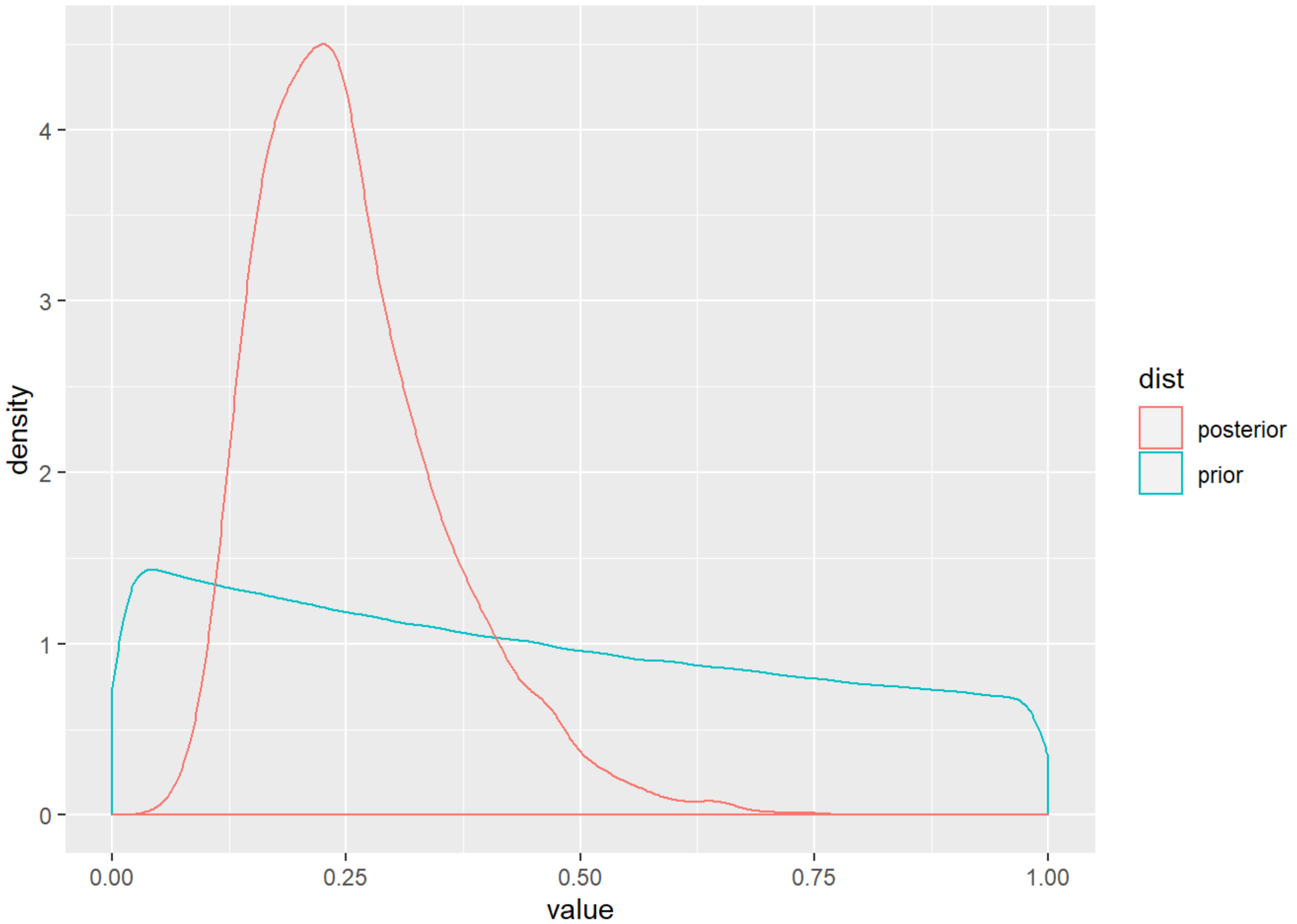
```
sigma2_prior = data.frame(
  value = seq(10, 22.5, by = 0.1),
  density = dinvgamma(seq(10, 22.5, by = 0.1), nu0 / 2, nu0 * s20 / 2),
  variable = 'sigma2'
)
tau2_prior = data.frame(
  value = seq(0, 30, by = 0.1),
  density = dinvgamma(seq(0, 30, by = 0.1), eta0 / 2, eta0 * t20 / 2),
  variable = 'tau2'
)
mu_prior = data.frame(
  value = seq(0, 12, by = 0.1),
  density = dnorm(seq(0, 12, by = 0.1), mu0, sqrt(g20)),
  variable = 'mu'
)
priors = rbind(sigma2_prior, tau2_prior, mu_prior)
priors$dist = 'prior'
smt.df$dist = 'posterior'
ggplot(priors, aes(x = value, y = density, color = dist)) +
  geom_line() +
  geom_density(data = smt.df, mapping = aes(x = value, y = ..density..)) +
  facet_wrap(~ variable, scales = 'free')
```




Our prior estimates for $\mu$ and $\tau^2$ were fairly estimate, but our estimate for $\sigma^2$ was very far off. After this analysis, we have estimates for $\mu$, the average amount of hours of schoolwork spent at a typical school, $\tau^2$, the variability between schools in the average hours of schoolwork, and $\sigma^2$, the variability among students' hours in each school.

# C

```
t20_prior = (1 / rgamma(1e6, eta0 / 2, eta0 * t20 / 2))
s20_prior = (1 / rgamma(1e6, nu0 / 2, nu0 * s20 / 2))
R_prior = data.frame(
  value = (t20_prior) / (t20_prior + s20_prior),
  dist = 'prior'
)
R_post = data.frame(
  value = SMT[, 'tau2'] / (SMT[, 'tau2'] + SMT[, 'sigma2']),
  dist = 'posterior'
)
ggplot(R_prior, aes(x = value, y = ..density.., color = dist)) +
  geom_density(data = R_prior) +
  geom_density(data = R_post)
```




```
mean(R_post$value)
```

```
## [1] 0.25744
```

```
mean(R_prior$value)
```

```
## [1] 0.43337
```

R이 1보다는 0에 가까운 형태이므로 그룹들 간의 variability보다 그룹 내에서의 variability가 더 크다고 할 수 있다.

# d

```
theta7_lt_6 = THETA[, 7] < THETA[, 6]
mean(theta7_lt_6)
```

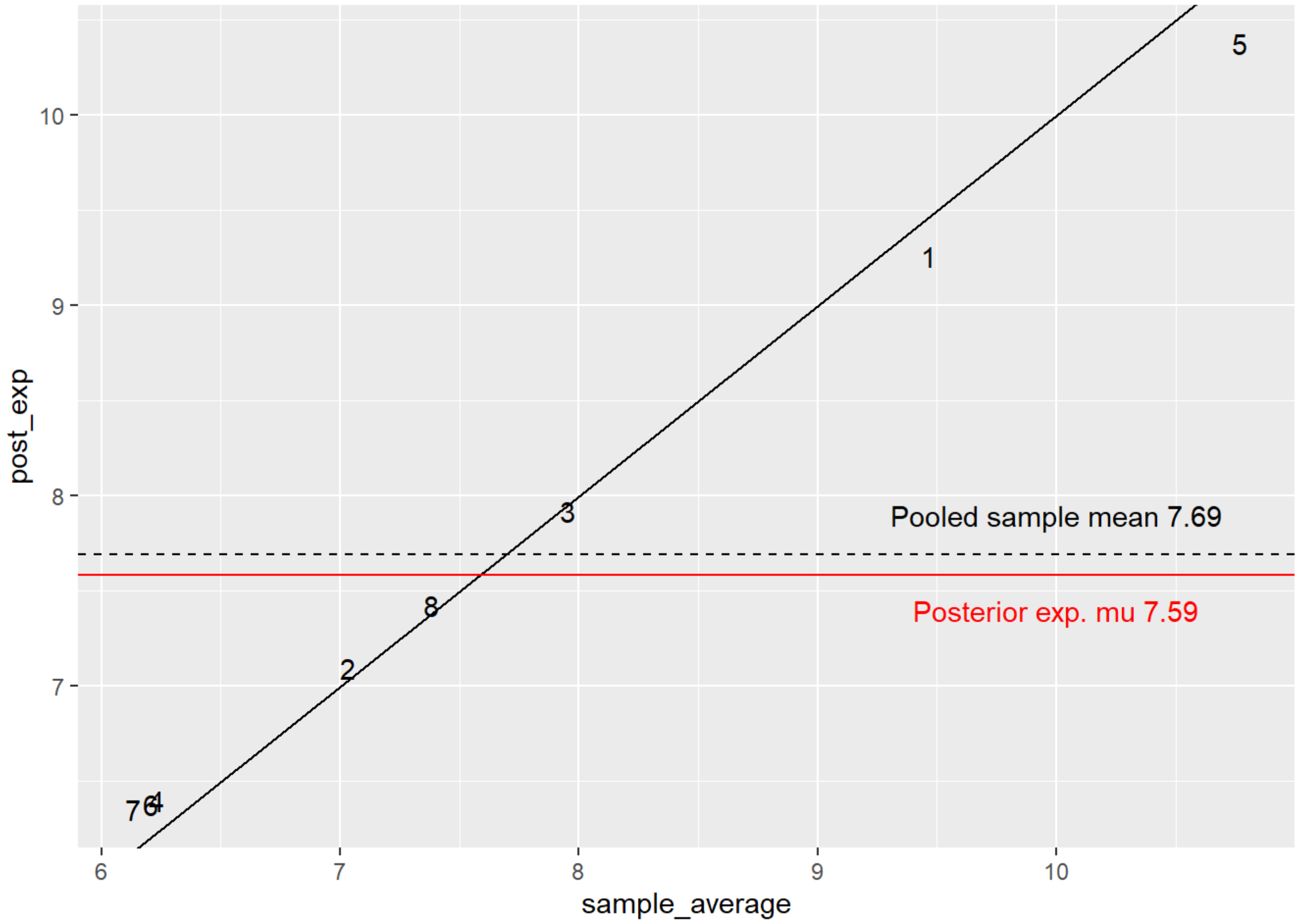
```
## [1] 0.5085
```

```
theta7_smallest = (THETA[, 7] < THETA[, -7]) %>%
  apply(MARGIN = 1, FUN = all)
mean(theta7_smallest)
```

```
## [1] 0.307
```

# e

```
relationship = data.frame(
  sample_average = ybar,
  post_exp = colMeans(THETA),
  school = 1:length(ybar)
)
ggplot(relationship, aes(x = sample_average, y = post_exp, label = school)) +
  geom_text() +
  geom_abline(slope = 1, intercept = 0) +
  geom_hline(yintercept = mean(schools.raw[, 'hours']), lty = 2) +
  annotate('text', x = 10, y = 7.9, label = paste0("Pooled sample mean ", round(mean(schools.ra
w[, 'hours']), 2))) +
  geom_hline(yintercept = mean(SMT[, 'mu']), color = 'red') +
  annotate('text', x = 10, y = 7.4, label = paste0("Posterior exp. mu ", round(mean(SMT[, 'mu'
]), 2)), color = 'red')
```


post_exp
sample_average
Pooled sample mean 7.69
Posterior exp. mu 7.59
5
1
3
8
2
76 4
10
9
8
7
6
7
8
9
10