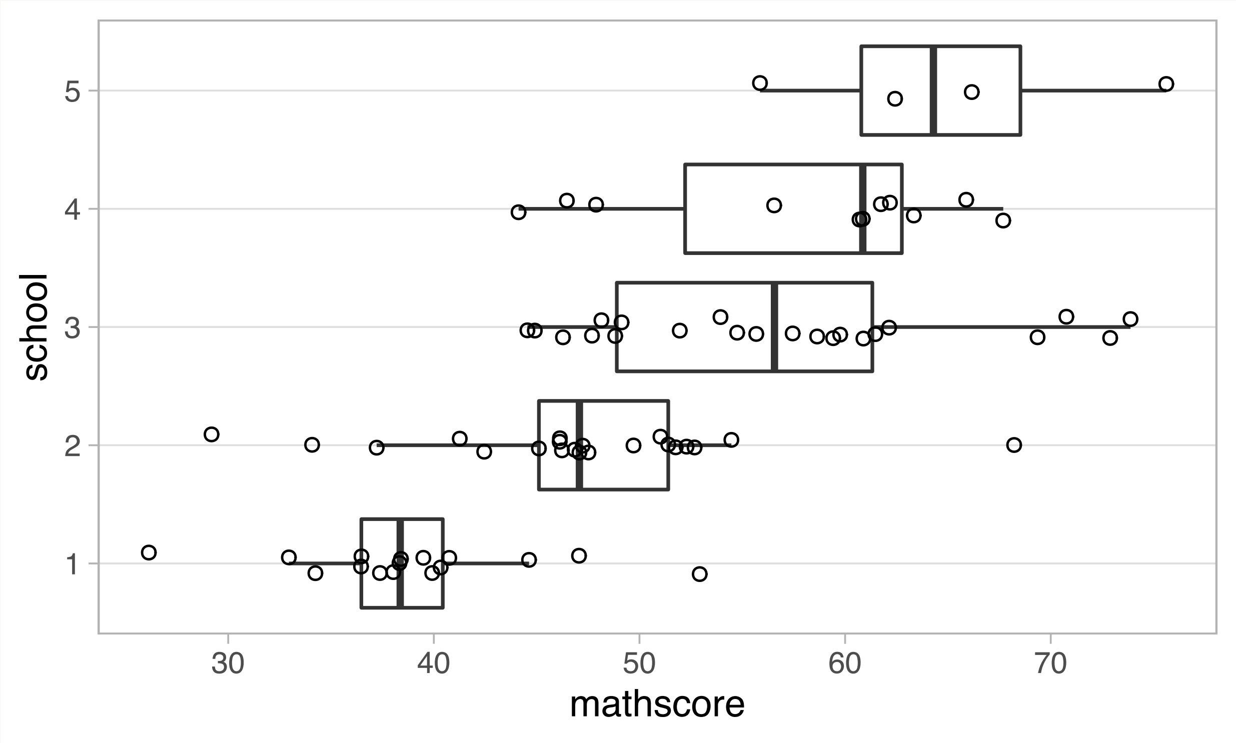
Hierarchical models

Stat 340: Bayesian Statistics

Example: ELS math scores

- 2002 Educational Longitudinal Study (ELS)
- Survey from schools across the United States
- Data are collected by sampling schools and then sampling students within each selected school
- We'll focus on 10th grade math scores from a sample of 10 schools
- Math tests contained items in arithmetic, algebra, geometry, data/probability, and advanced topics were divided into process categories of skill/knowledge, understanding/ comprehension, and problem solving

ELS math scores



Possible questions:

- What's the typical math score?
- To what extent do scores vary from school to school?
- For any single school, how much might scores vary from student to student?

Possible analysis strategies

Complete pooling (combined estimates)
Ignore schools and lump all students together

No pooling (separate groups)
Separately analyze each school and assume that one school's data doesn't contain valuable information about another school

Partial pooling (compromise estimates)
Acknowledge the grouping structure, so that even though schools differ in performance, they might share valuable information about each other and about the broader population of schools

What have we seen so far?

- Completely pooled model does not acknowledge differences between schools
- No pooled model acknowledges that some schools tend to score higher than others
- No pooled model ignores data on one school when learning about the typical score of another
- No pooled model cannot be generalized to schools outside our sample

Hierarchical model

Let's compromise between the the complete pooled and no pooled models

How? By using a *two-stage prior* specification

Hierarchical model specification for JAGS

```
modelString <-"model {

## sampling
for (i in 1:N){
   y[i] ~ dnorm(mu_j[school[i]], invsigma2)
}

## priors
for (j in 1:J){
   mu_j[j] ~ dnorm(mu, invtau2)
}

invsigma2 ~ dgamma(a_s, b_s)
sigma <- sqrt(pow(invsigma2, -1))

## hyperpriors
mu ~ dnorm(mu0, g0)
invtau2 ~ dgamma(a_t, b_t)
tau <- sqrt(pow(invtau2, -1))
}
"
```

Define the data and prior parameters

```
y <- sub_school$mathscore
school <- sub_school$school
N <- length(y)
J <- length(unique(school))
the_data <- list(y = y, school = school,
                 N = N, J = J,
                 mu0 = 50, g0 = .04,  # prior parameters
                 a_t = 1, b_t = .01,  # hyperparameters
                 a_s = 1, b_s = .01)  # hyperparameters
```

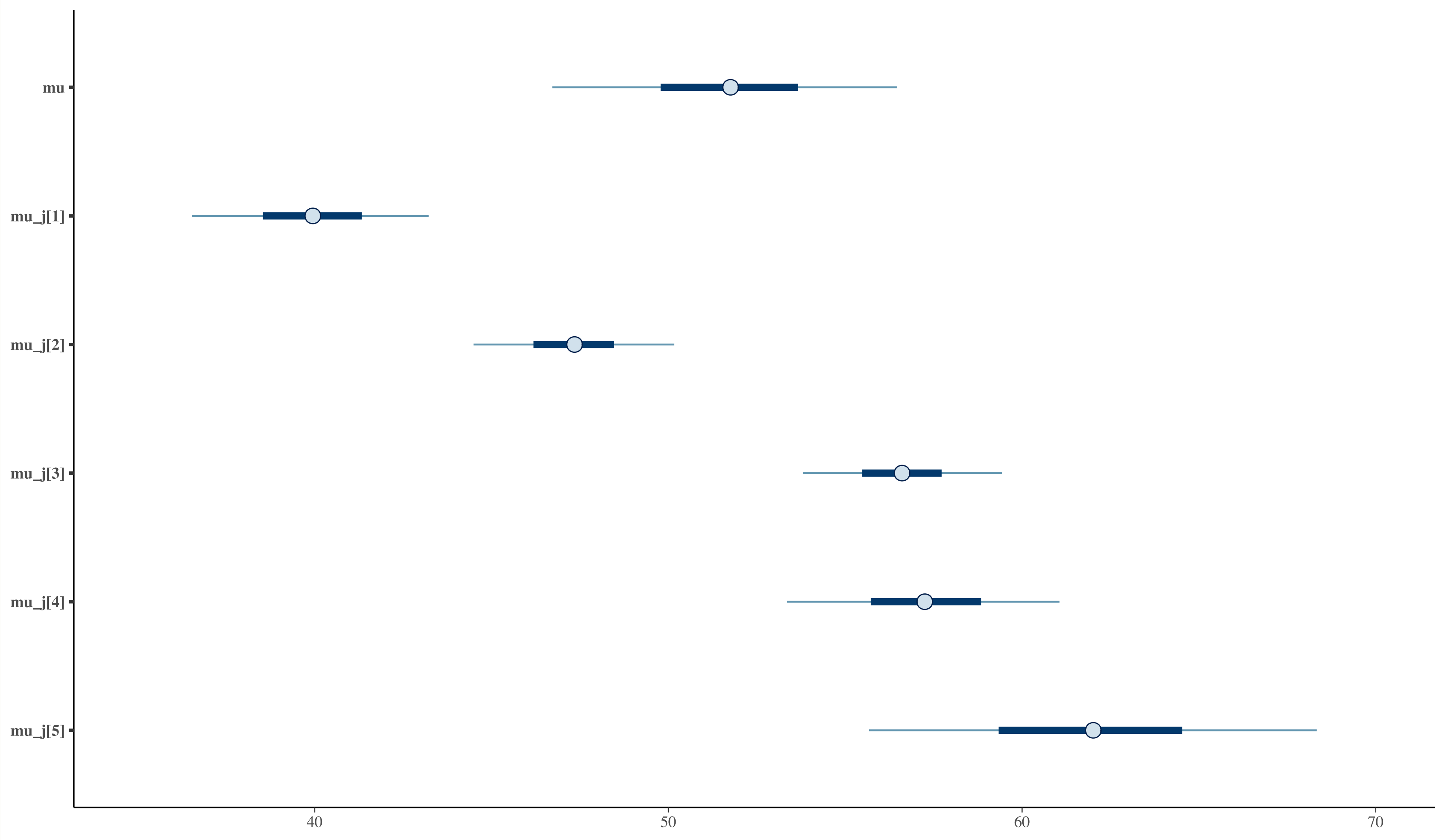
Run MCMC

```r
posterior <- run.jags(
  modelString,
  n.chains = 1,
  data = the_data,
  monitor = c("mu", "tau", "mu_j", "sigma"),
  adapt = 1000,
  burnin = 5000,
  sample = 5000,
  silent.jags = TRUE
)
```

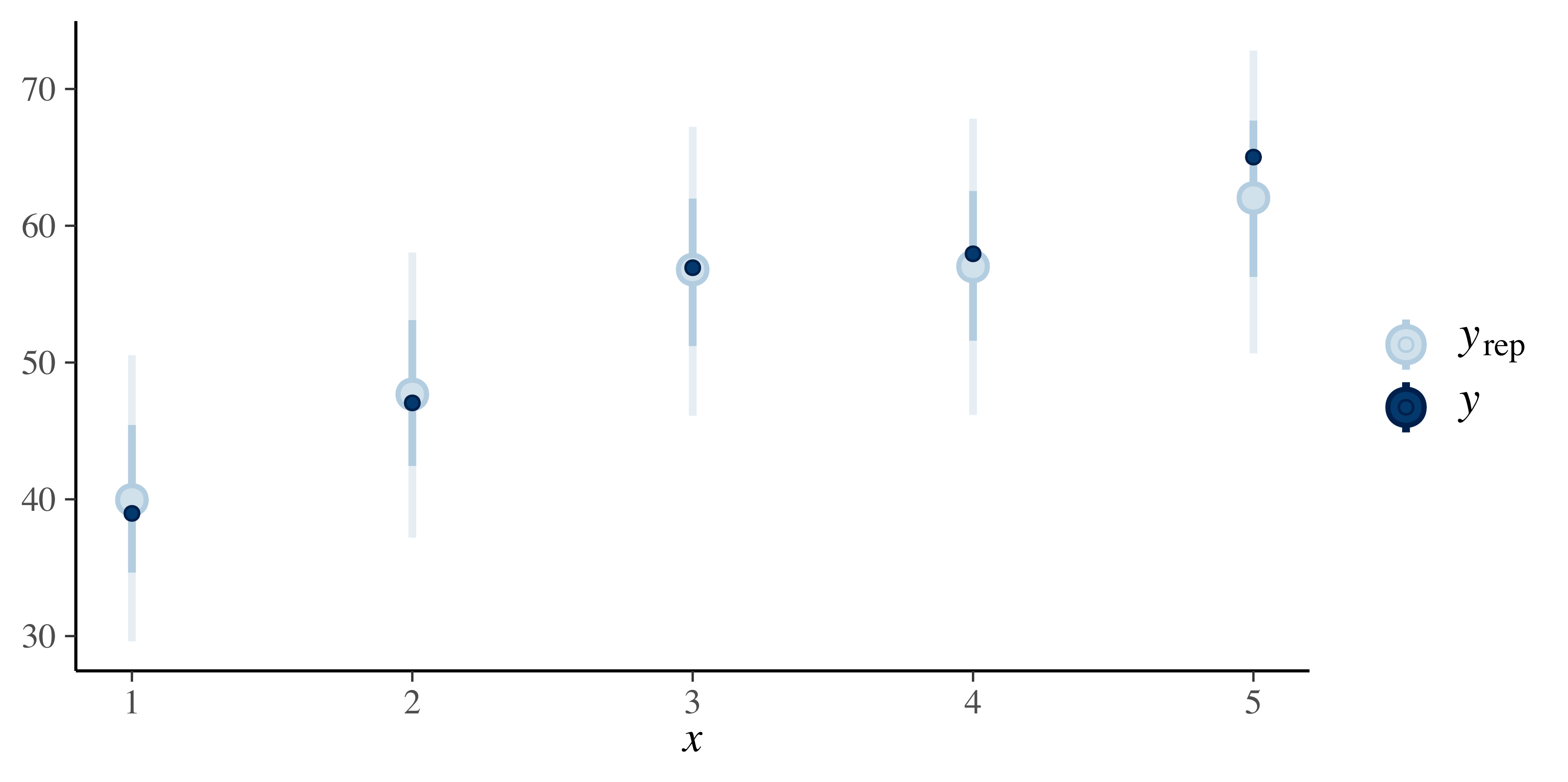
```r
print(posterior, digits = 3)
```

```
##
## JAGS model summary statistics from 5000 samples (adapt+burnin = 6000):
##
##           Lower95 Median Upper95 Mean    SD Mode  MCerr MC%ofSD SSeff    AC.10
## mu           45.8   51.8    57.6 51.7  3.01 51.8 0.0453     1.5  4411 -0.00241
## tau          3.71   7.68    14.3  8.3  3.12  6.9 0.0553     1.8  3189   0.0281
## mu_j[1]      35.9   39.9    43.9 39.9  2.04 40.3 0.0315     1.5  4184 -0.00417
## mu_j[2]      43.9   47.3    50.7 47.3  1.73 47.5  0.025     1.4  4783   0.0104
## mu_j[3]      53.3   56.6      60 56.6   1.7 56.6 0.0247     1.4  4763 -0.00338
## mu_j[4]      52.7   57.2    61.8 57.2  2.32 57.1 0.0339     1.5  4708  0.00782
## mu_j[5]      54.2     62    69.2   62  3.86 61.9 0.0624     1.6  3826   0.0131
## sigma        6.71   7.95    9.37    8 0.693  7.9 0.0105     1.5  4318 -0.00179
##
##           psrf
## mu          --
## tau         --
## mu_j[1]     --
## mu_j[2]     --
## mu_j[3]     --
## mu_j[4]     --
## mu_j[5]     --
## sigma       --
##
## Total time taken: 0.8 seconds
```

```r
mcmc_intervals(posterior$mcmc, regex_pars = "mu")
```

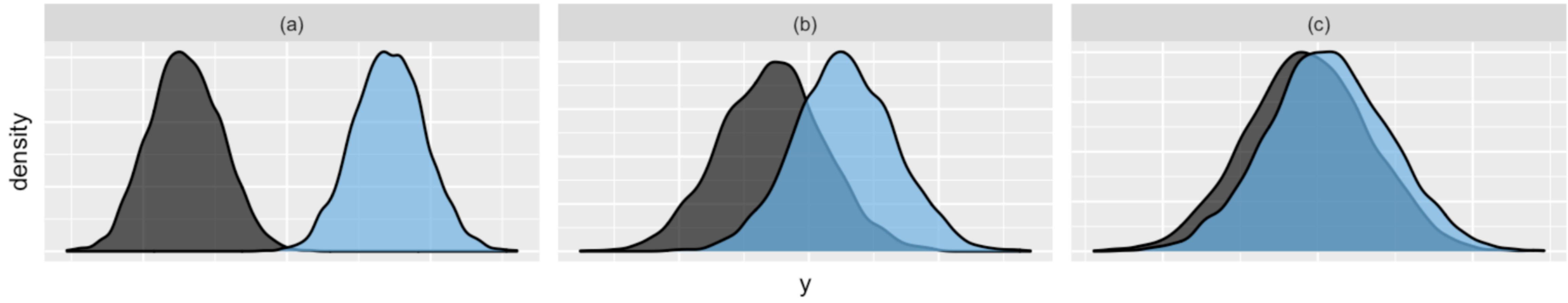


Hierarchical predictions vs. sample means



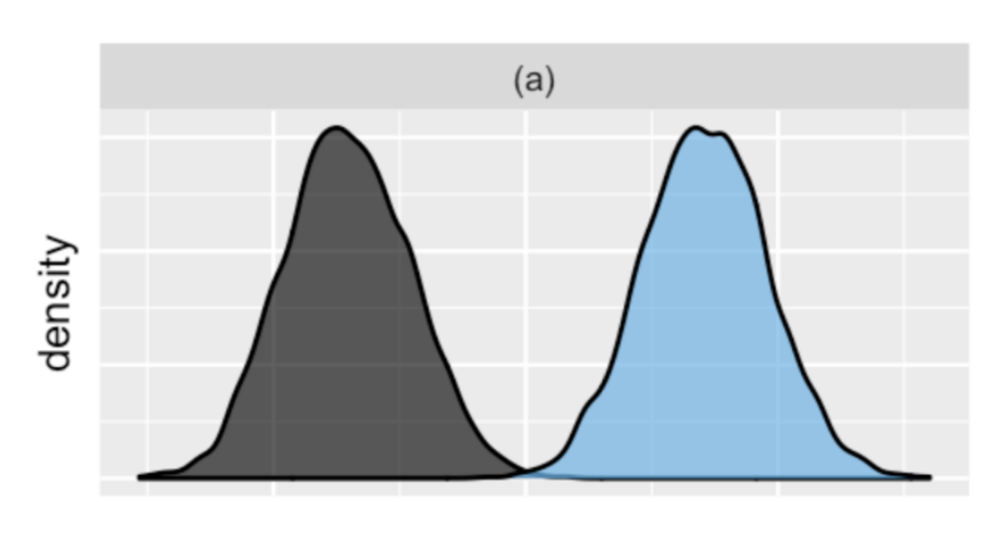
Comparing sources of variability

between > within | between ≈ within | between < within



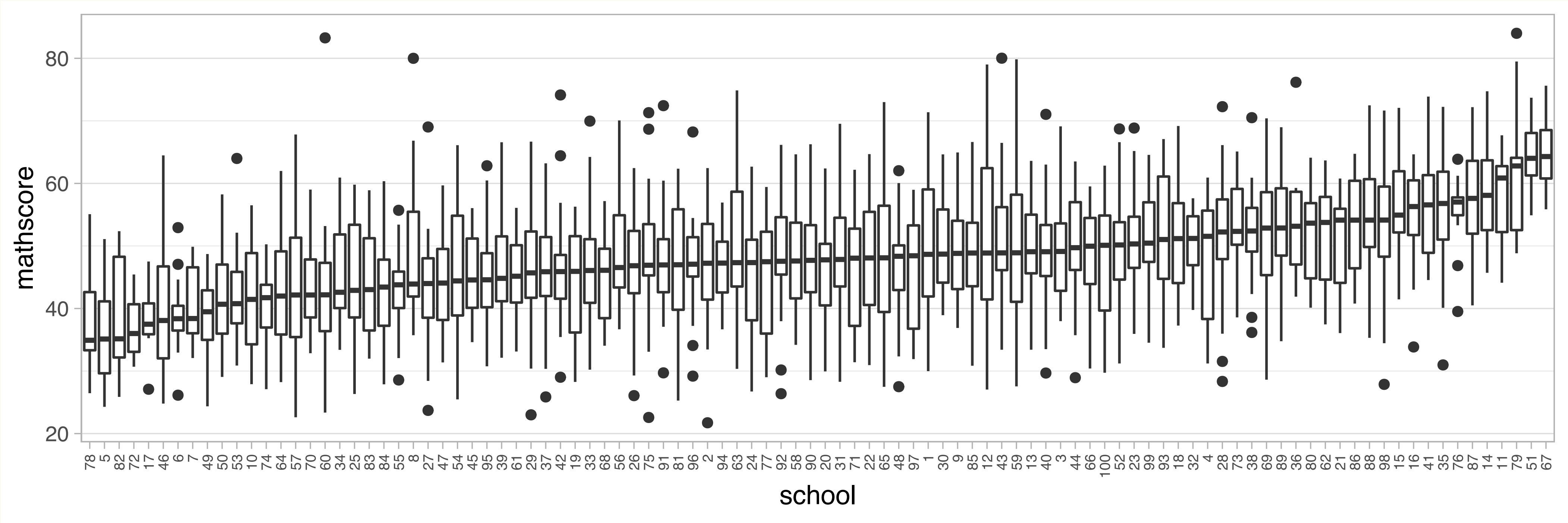
Within-group (intraclass) correlation

Suppose we're in the situation where between group variability is much larger than within group variability



- Two observations within the same group are more similar than two observations from different groups
- Observations within the same group are correlated (generally true, easier to see in this extreme case)

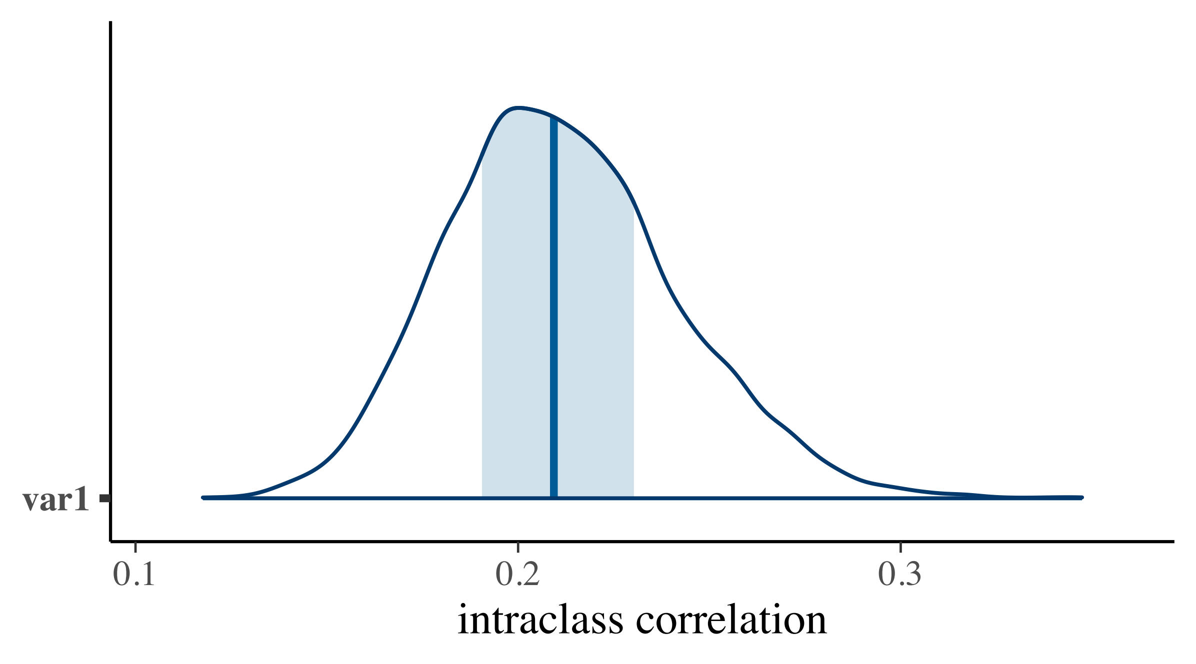
Full ELS data set



- 100 schools from urban settings in the full data set
- Sample sizes range from 4 to 32 students

ELS intraclass correlation

```
draws <- posterior_full$mcmc[[1]]
icc <- draws[,"tau"]^2 / (draws[,"tau"]^2 + draws[,"sigma"]^2)
```

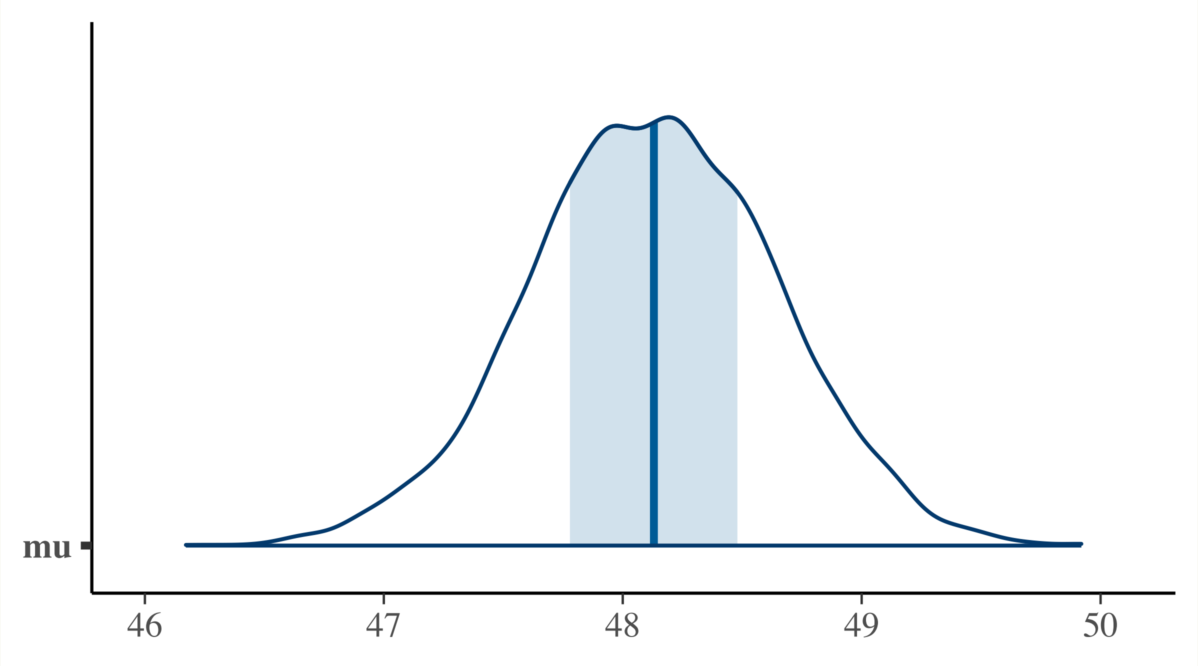


90% credible interval:

```
##    5%   95%
## 0.165 0.264
```

ELS global mean

Inference for the global parameters proceeds as always

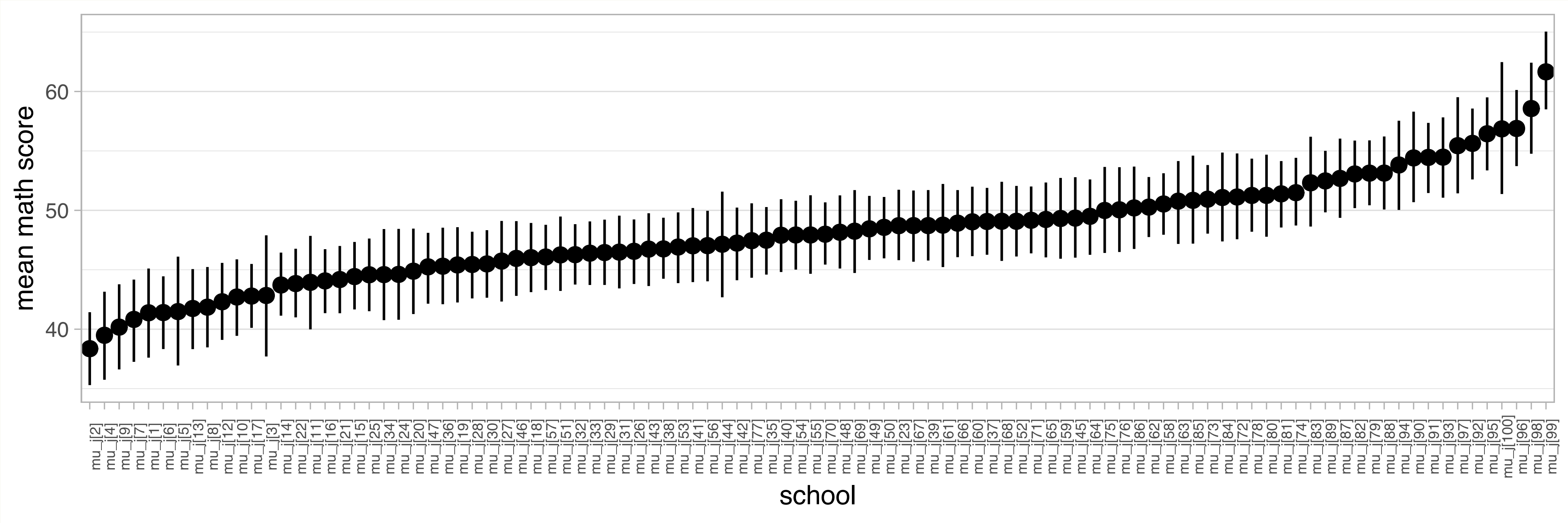


90% credible interval:

```
##     5%   95%
## 47.27 48.97
```

Inference for group-specific means

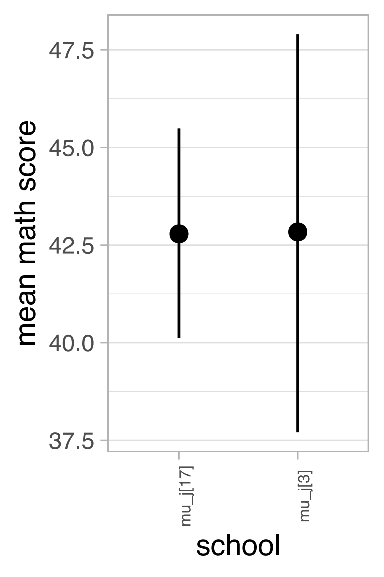
There are often *a lot* of parameters to manage for group-specific inference



Your turn

School 3 and 17 have roughly the same posterior mean, but substantially different credible interval widths

Discuss with a neighbor why you think this difference occurs.



02:00

Prediction for observed group

Suppose we want to make a prediction for school 13, then we need a posterior predictive distribution

```
# Work with a data frame
post_df <- as.data.frame(posterior_full$mcmc[[1]])

# Create the posterior predictive via simulation
pp_school13 <- post_df %>%
  select(mu_j = "mu_j[13]", sigma) %>%                # select relevant cols
  mutate(y_pred = rnorm(nrow(post_df), mu_j, sigma)) # simulate a new obs.

# Check it out
head(pp_school13, 3)
```

```
##          mu_j    sigma   y_pred
## 6001 40.65318 9.220115 17.68957
## 6002 42.60158 8.925333 50.84780
## 6003 40.14430 9.230251 37.15336
```

Prediction for unobserved group

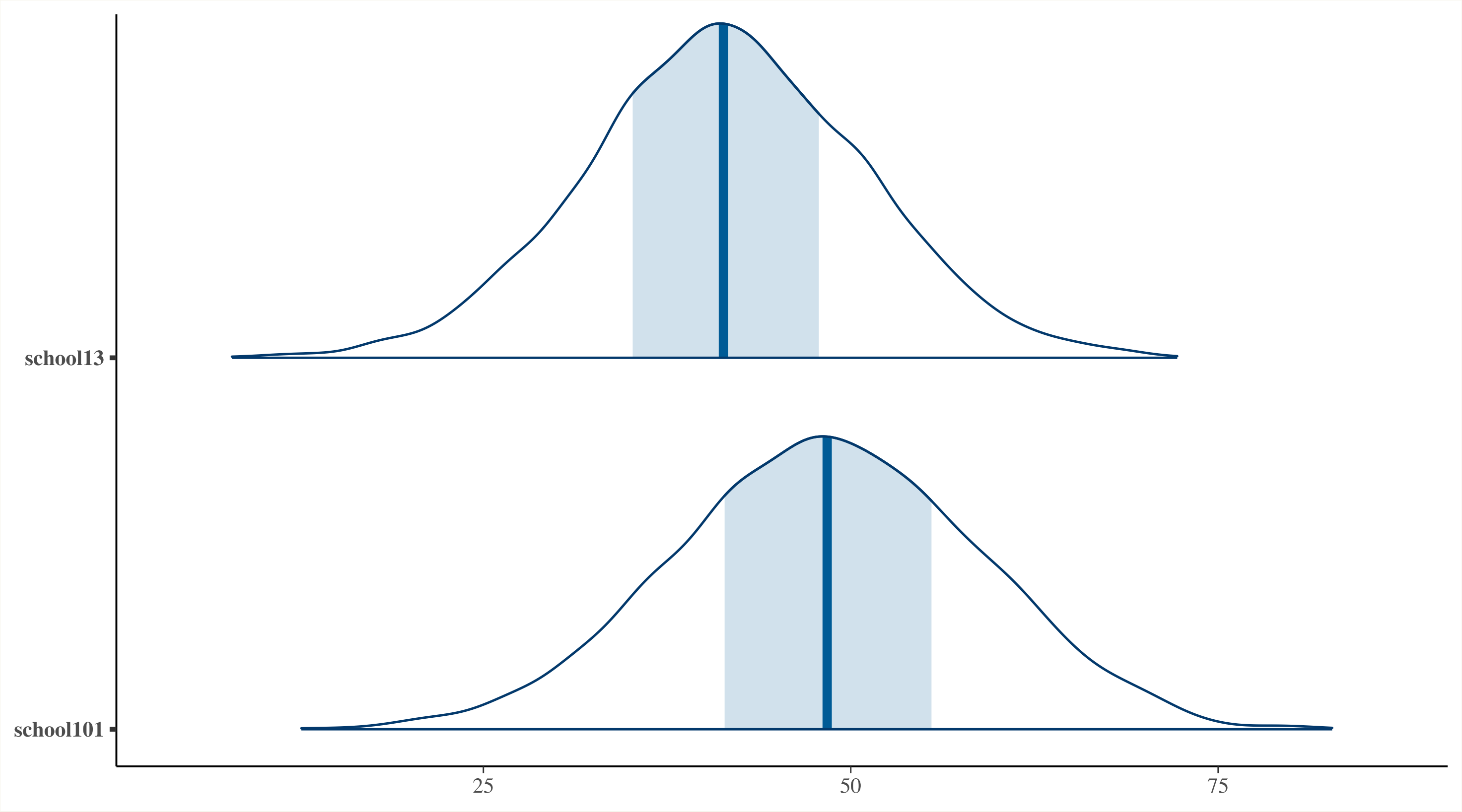
Suppose we want to make a prediction for a school we didn't observe, let's call it school 101

```
# Create the posterior predictive via simulation
pp_school101 <- post_df %>%
  select(mu, tau, sigma) %>%                         # select global params
  mutate(
    mu_j = rnorm(nrow(post_df), mu, tau),            # generate mu_j
    y_pred = rnorm(nrow(post_df), mu_j, sigma) # generate y
  )

# Check it out
head(pp_school101, 3)
```

```
##                mu      tau    sigma     mu_j   y_pred
## 6001 48.76941 4.738100 9.220115 41.14001 51.18014
## 6002 48.92055 4.394430 8.925333 47.45571 40.40386
## 6003 47.77900 4.468838 9.230251 42.37273 43.12030
```

How do the predictions compare?



Prior distributions for between group variance

For $J > 5$, common choices include

- improper uniform (either on the variance or log variance scale), but not allowed in JAGS
- proper uniform (with large upper bound) - `dunif`
- proper normal (with large variance $\rightarrow$ small precision) - `dnorm`

Gelman (2006). Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*. 1 (3) 515 - 534.

Prior distributions for between group variance

For $J < 5$, or in a situation where more prior information is helpful,

- Use half-t distribution
- Half-Cauchy with reasonably large scale parameter - `dhalfcauchy(sigma)`

Gelman (2006). Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*. 1 (3) 515 - 534.

Prior distributions for between group variance

- In general, Gelman (2006) does not recommend the $\mathrm{Gamma}(\epsilon, \epsilon)$ family of noninformative prior distributions because the inferences are often sensitive to the choice of ϵ.
- Different than advice in the textbook

Gelman (2006). Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*. 1 (3) 515 - 534.