Estimation of group effects

560 Hierarchical modeling

Peter Hoff

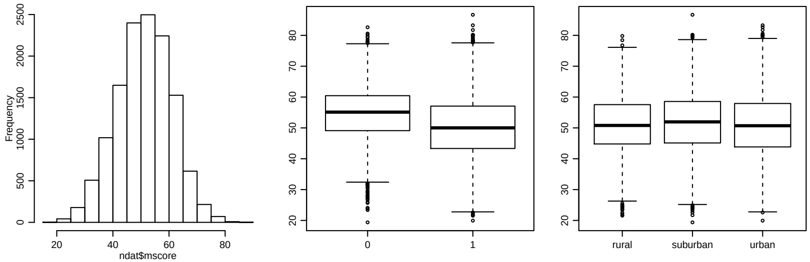
Statistics, University of Washington

Data for today

```
ndat[1:5,]

##   school enroll flp public urbanicity hwh   ses mscore
## 1   1011      5   3      1      urban   2 -0.23  52.11
## 2   1011      5   3      1      urban   0  0.69  57.65
## 3   1011      5   3      1      urban   4 -0.68  66.44
## 4   1011      5   3      1      urban   5 -0.89  44.68
## 5   1011      5   3      1      urban   3 -1.28  40.57

table(ndat$public)

##
##    0    1
## 3161 9813

table(ndat$urbanicity)

##
##    rural suburban    urban
##     2349     6114     4511
```

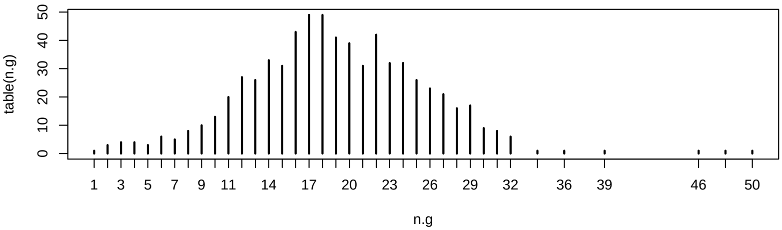
Data for today

```
par(mfrow=c(1,3),mar=c(3,3,2,1),mgp=c(1.75,.75,0))
hist(ndat$mscore,main="")
boxplot(ndat$mscore~ndat$public)
boxplot(ndat$mscore~ndat$urbanicity)
```



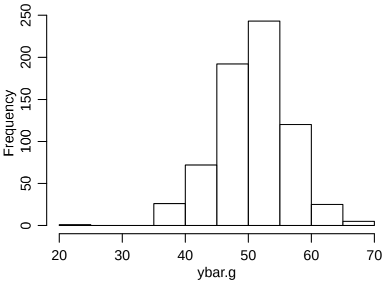
Data for today

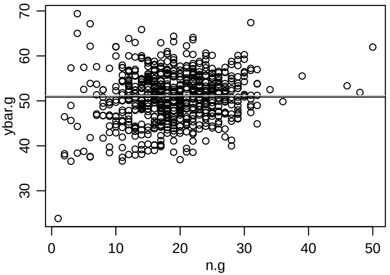
```
y<-ndat$mscore
g<-match(ndat$school , sort(unique(ndat$school)))

# school specific sample sizes
n.g<-c(table(g) )

plot(table(n.g))
```



Data for today

```
# school specific mscore means
ybar.g<-c(tapply(y,g,"mean"))

par(mfrow=c(1,2),mar=c(3,3,2,1),mgp=c(1.75,.75,0))
hist(ybar.g,main="")
plot(ybar.g~n.g)
abline(h=mean(ybar.g))
abline(h=mean(y),col="gray")
```





Testing for across-group differences

```
fit.ols<-lm(y~as.factor(g))
anova(fit.ols)

## Analysis of Variance Table
##
## Response: y
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(g)    683 342385  501.30  6.8118 < 2.2e-16 ***
## Residuals     12290 904450   73.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MLEs

```r
library(lme4)
fit.lme<-lmer(y~1+(1|g),REML=FALSE)
summary(fit.lme)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ 1 + (1 | g)
##
##      AIC      BIC   logLik deviance df.resid
##  93919.3  93941.7 -46956.6  93913.3    12971
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8112 -0.6534  0.0093  0.6732  4.6999
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  g        (Intercept) 23.63    4.861
##  Residual             73.71    8.585
## Number of obs: 12974, groups:  g, 684
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  50.9391     0.2026   251.4
```

Parameter estimates

```
VarCorr(fit.lme)

##  Groups   Name        Std.Dev.
##  g        (Intercept) 4.8615
##  Residual             8.5854

t2.mle<-as.numeric(VarCorr(fit.lme)$g)

t2.mle

## [1] 23.63411

sigma(fit.lme)

## [1] 8.585362

s2.mle<-sigma(fit.lme)^2

s2.mle

## [1] 73.70844

fixef(fit.lme)

## (Intercept)
##     50.9391

mu.mle<-fixef(fit.lme)
```

Group-specific estimates

What about estimates of $\mu_1, \ldots, \mu_m$?

Unbiased estimate:

$$\begin{aligned} \mathrm{E}[\bar{y}_j - \mu_j | \mu_j] &= \mathrm{E}[\bar{y}_j | \mu_j] - \mathrm{E}[\mu_j | \mu_j] \\ &= \mu_j - \mu_j = 0 \end{aligned}$$

$\bar{y}_j$ is an *unbiased estimator* of μ_j.

Expected squared error of unbiased estimate:

$$\begin{aligned} \mathrm{E}[(\bar{y}_j - \mu_j)^2 | \mu_j] &= \mathrm{Var}[\bar{y}_j | \mu_j] \\ &= \sigma^2 / n_j \end{aligned}$$

Standard error of unbiased estimate:

$$\mathrm{se}[\bar{y}_j | \mu_j] = \hat{\sigma} / \sqrt{n_j}$$

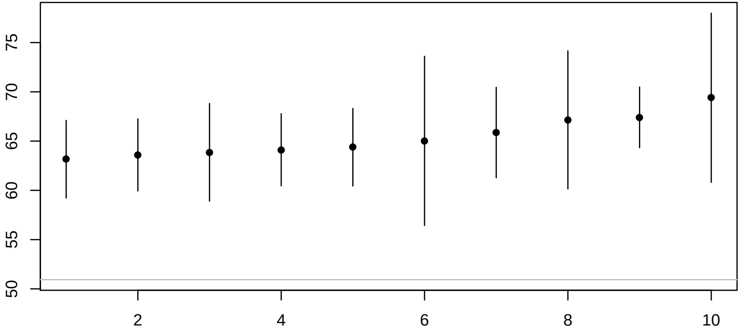
League tables

```
### top ten schools
topten<-order(ybar.g,decreasing=TRUE)[1:10]

topten

##  [1] 639 349 618 616 386 337 637  73 680 352

ybar.g[topten]

##      639      349      618      616      386      337      637       73
## 69.40250 67.40645 67.15500 65.86786 65.01750 64.37632 64.12091 63.86083
##      680      352
## 63.59818 63.16263

### top three schools
ybar.t3<-c(ybar.g[topten[1]] ,  ybar.g[topten[2]], ybar.g[topten[3]] )

ybar.t3

##      639      349      618
## 69.40250 67.40645 67.15500
```

Approximate confidence intervals

```
### sample sizes of top three
n.t3<-c(n.g[topten[1]] ,  n.g[topten[2]], n.g[topten[3]] )

n.t3

## [1]  4 31  6

### se of ybar for top three
se.t3<-sqrt(s2.mle/n.t3)

se.t3

## [1] 4.292681 1.541977 3.504959

### approximate 95 CIs
rbind(ybar.t3+2*se.t3, ybar.t3-2*se.t3)

##           639     349      618
## [1,] 77.98786 70.4904 74.16492
## [2,] 60.81714 64.3225 60.14508
```

More approximate confidence intervals



MSE and shrinkage estimates

MSE: The mean squared error of an estimator $\hat{\theta}$ in estimating θ is

$$\mathsf{MSE}(\hat{\theta}|\theta) = \mathsf{E}[(\hat{\theta} - \theta)^2|\theta]$$

Quiz: What is the MSE of $\bar{y}_j$ for estimating μ_j?

$$\begin{aligned} \mathsf{E}[(\bar{y}_j - \mu_j)^2|\mu_j] &= \mathsf{Var}[\bar{y}_j|\mu_j] \\ &= \sigma^2/n_j \end{aligned}$$

General result: The MSE of an unbiased estimator is its variance.

HW: What is the unconditional MSE of $\bar{y}_j$, treating μ_j as sampled?

$$MSE(\bar{y}_j) = \sigma^2/n_j$$

A shrinkage estimator

Suppose μ, σ^2, τ^2 are known. Can we find a better estimator than $\bar{y}_j$?

Intuition: If τ^2 is small and σ^2/n_j large, then

- $\bar{y}_j$ might be far from μ_j;
- μ_j should be close to μ.

This suggests the following "shrinkage estimator:"

$$\hat{\mu}_j = w_j \bar{y}_j + (1 - w_j)\mu \ , \text{ where } w_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}.$$

Quiz: Describe how $\hat{\mu}_j$ changes with

- n_j
- σ^2
- τ^2

Let $\mu = 0$ so $\hat{\mu}_j = w\bar{y}_j$.

$$MSE(\hat{\mu}_j|\mu_j) = \mathrm{E}[(w\bar{y}_j - \mu_j)^2|\mu_j]$$
$$MSE(\hat{\mu}_j) = \mathrm{E}[MSE(\hat{\mu}_j|\mu_j)]$$

Useful for calculations is the following identity:

$$\begin{aligned}(w\bar{y}_j - \mu_j)^2 &= (w(\bar{y}_j - \mu_j) - (1-w)\mu_j)^2 \\ &= w^2(\bar{y}_j - \mu_j)^2 - 2w(1-w)(\bar{y}_j - \mu_j)\mu_j + (1-w)^2\mu_j^2\end{aligned}$$

Unconditional MSE:

$$\mathrm{MSE}(\hat{\mu}_j) = \mathrm{E}[\mathrm{MSE}(\hat{\mu}_j|\mu_j)] = w^2\sigma^2/n_j + (1-w)^2\tau^2$$

MSE Comparison

$$\text{MSE}(\bar{y}_j) = \sigma^2/n_j$$
$$\text{MSE}(\hat{\mu}_j) = w^2\sigma^2/n_j + (1-w)^2\tau^2$$

Which is bigger?

Recall:

$$w = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} \in (0,1)$$

This implies $w^2 \in (0,1)$, so

$$w^2\sigma^2/n_j < \sigma^2/n_j$$

What about the other part of $MSE(\hat{\mu}_j)$?

Intuition: If τ^2 small $\Rightarrow$ other part is small, expect $\text{MSE}(\hat{\mu}_j) < \text{MSE}(\bar{y}_j)$.

Result: In fact,

$$\text{MSE}(\hat{\mu}_j) = \left(\frac{\tau^2}{\tau^2+\sigma^2/n}\right)\sigma^2/n < \sigma^2/n = \text{MSE}(\bar{y}_j)$$

for all $\mu, \sigma^2, \tau^2, n_j$.

Bias and variance

More generally, let

- $\hat{\theta}$ be an estimator of θ.
- $\mathrm{E}[\hat{\theta}|\theta] = \theta_0$.

If $\theta_0 = \theta$, then $\hat{\theta}$ is *unbiased*.

$$
\begin{aligned}
MSE(\hat{\theta}|\theta) &= \mathrm{E}[(\hat{\theta}-\theta)^2|\theta] \\
&= \mathrm{E}[([\hat{\theta}-\theta_0]+[\theta_0-\theta])^2|\theta] \\
&= \mathrm{E}[(\hat{\theta}-\theta_0)^2|\theta] + 2\times \mathrm{E}[(\hat{\theta}-\theta_0)(\theta_0-\theta)|\theta] + \mathrm{E}[(\theta_0-\theta)^2|\theta]
\end{aligned}
$$

- $\mathrm{E}[(\hat{\theta}-\theta_0)^2|\theta] = \mathrm{Var}[\hat{\theta}|\theta]$
- $\mathrm{E}[(\hat{\theta}-\theta_0)(\theta_0-\theta)|\theta] = 0$
- $\mathrm{E}[(\theta_0-\theta)^2|\theta] = (\theta_0-\theta)^2 = \mathrm{bias}(\hat{\theta}|\theta)^2$.

Bias-variance tradeoff

In general,

$$MSE(\hat{\theta}|\theta) = \mathrm{Var}[\hat{\theta}|\theta] + \mathrm{bias}(\hat{\theta}|\theta)^2$$

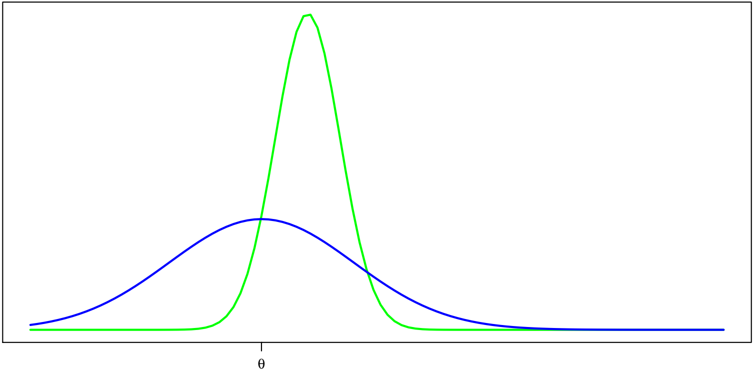
How well an estimator $\hat{\theta}$ does at estimating θ depends on *variance* and *bias*.

In general,

- estimators with low bias have have high variance;
- estimators with low variance have high bias.

Minimizing MSE requires balancing bias and variance.

Bias-variance tradeoff



Summary of bias and variance for the hierarchical model

If we are interested in how well we do across groups, we would compute

$$MSE(\hat{\mu}_j) = \mathrm{E}[MSE(\hat{\mu}_j|\mu_j)]$$

where the second expectation is with respect to $\mu_j \sim N(\mu, \tau^2)$.

Bias and variance of $\bar{y}_j$:

$$\begin{aligned} MSE(\bar{y}_j|\mu_j) &= \sigma^2/n_j \\ MSE(\bar{y}_j) = \mathrm{E}[MSE(\bar{y}_j|\mu_j)] &= \sigma^2/n_j \end{aligned}$$

Bias and variance of $\hat{\mu}_j$:

$$\begin{aligned} MSE(\hat{\mu}_j|\mu_j) &= w^2\sigma^2/n_j + (1-w)^2(\mu_j - \mu)^2 \\ MSE(\hat{\mu}_j) &= w^2\sigma^2/n_j + (1-w)^2\tau^2 \end{aligned}$$

You can show that the value of w that minimizes the unconditional MSE is

$$w_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}$$

Terminology: BLUPs, Bayes and shrinkage

The shrinkage estimators $\hat{\mu}_1, \dots, \hat{\mu}_m$ are also called

- Bayes estimators;
- BLUPs (best unbiased linear predictors)

Bayesian interpretation: If

- $\mu_j \sim N(\mu, \tau^2)$ represents your uncertainty about μ_j, and
- you observe $y_{1,j}, \dots, y_{n,j} \sim$ i.i.d. $N(\mu_j, \sigma^2)$, then

your optimal guess about μ_j is

$$\hat{\mu}_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu.$$

BLUPs

The $\hat{\mu}_j$'s are sometimes called the *best unbiased linear predictors (BLUPs)* .

This is confusing, as we have discussed how these estimators are biased:

$$\begin{aligned} \mathsf{E}[\hat{\mu}_j|\mu_j] &= \mathsf{E}[w\bar{y}_j + (1-w)\mu|\mu_j] \\ &= w\mu_j + (1-w)\mu \neq \mu_j \end{aligned}$$

$\hat{\mu}_j$ is *conditionally* biased.

The "U" in BLUP refers to bias only in an unconditional sense:

$$\begin{aligned} \mathsf{E}[\hat{\mu}_j] &= \mathsf{E}[\mathsf{E}[\hat{\mu}_j|\mu_j]] \\ &= \mathsf{E}[w\mu_j + (1-w)\mu] \\ &= w\mu + (1-w)\mu = \mu. \end{aligned}$$

Since $\mathsf{E}[\hat{\mu}_j] = \mathsf{E}[\mu_j] = \mu$ *unconditionally*, people might say $\hat{\mu}_j$ is "unbiased."

Understanding conditional and unconditional expectation

school	A	B	C	D	E	F	G	H	I	J
mean	μ_A	μ_B	μ_C	μ_D	μ_E	μ_F	μ_G	μ_H	μ_I	μ_J

Let $\mu = (\mu_A + \cdots \mu_J)/10$.

Study design:

- sample m schools at random from the population of schools.
- sample n students at random from each of the m schools.

What is the expectation of μ_1, $\bar{y}_1$, $\hat{\mu}_1$?

Expectation of μ_1 : Since each school A through J has equal probability of being selected as unit 1:

$$\begin{aligned} \mathrm{E}[\mu_1] &= \mu_A \times \Pr(\text{unit 1 =A}) + \cdots + \mu_J \times \Pr(\text{unit 1=J}) \\ &= \mu_A \tfrac{1}{10} + \cdots + \mu_J \tfrac{1}{10} = \mu \end{aligned}$$

Understanding conditional expectation

$$\mathrm{E}[\bar{y}_1 - \mu_1 | \text{unit } 1 = \text{D}\ ] = \mathrm{E}[\bar{y}_D - \mu_D] = \mu_D - \mu_D = 0$$

$$\begin{aligned}\mathrm{E}[\hat{\mu}_1 - \mu_1 | \text{unit } 1 = \text{D}] &= \mathrm{E}[w\bar{y}_D + (1-w)\mu - \mu_D] \\ &= w\mu_D + (1-w)\mu - \mu_D = (1-w)(\mu - \mu_D) \neq 0\end{aligned}$$

Conditionally on unit 1=D,

- $\bar{y}_1 = \bar{y}_D$ is unbiased for μ_D,
- $\hat{\mu}_1 = \hat{\mu}_D$ is biased for μ_D.

In English, if your first sampled school is school D, then

- $\bar{y}_1 = \bar{y}_D$ and $\bar{y}_D$ is unbiased for μ_D
- $\hat{\mu}_1 = \hat{\mu}_D$ and $\hat{\mu}_D$ is biased for μ_D.

Understanding unconditional expectation

Before you sample the schools, unit 1 is equally likely to be school A, B, . . . , J.

$$
\begin{aligned}
\mathrm{E}[\hat{\mu}_1 - \mu_1] &= \mathrm{E}[\hat{\mu}_A - \mu_A]\,\mathrm{Pr}(\text{unit }1{=}\mathrm{A}) + \cdots + \mathrm{E}[\hat{\mu}_J - \mu_J]\,\mathrm{Pr}(\text{unit }1{=}\mathrm{J}) \\
&= (1-w)(\mu - \mu_A) \times \tfrac{1}{10} + \cdots + (1-w)(\mu - \mu_J) \times \tfrac{1}{10} \\
&= (1-w)\mu - (1-w)(\mu_A + \cdots + \mu_J)\tfrac{1}{10} \\
&= (1-w)\mu - (1-w)\mu = 0.
\end{aligned}
$$

This unconditional expectation, and the "U" in BLUP, refers to averaging across the possibilities for the samples:

- $\hat{\mu}_j$ will be a biased estimator of the mean of whatever unit is picked jth.
- on average across studies, $\hat{\mu}_1, \ldots, \hat{\mu}_m$ will be unbiased.

Summary

In most applications I am familiar with, interest is more in the conditional expectations.

From this perspective, the shrinkage estimators $\hat{\mu}_1, \ldots, \hat{\mu}_m$

- are biased;
- have conditional MSE given by

$$w^2\sigma^2/n_j + (1-w)^2(\mu_j - \mu)^2,$$

- which is usually lower than the conditional MSE of $\bar{y}_j$.

Plug-in estimates

In practice, we replace μ, σ^2, τ^2 with estimates:

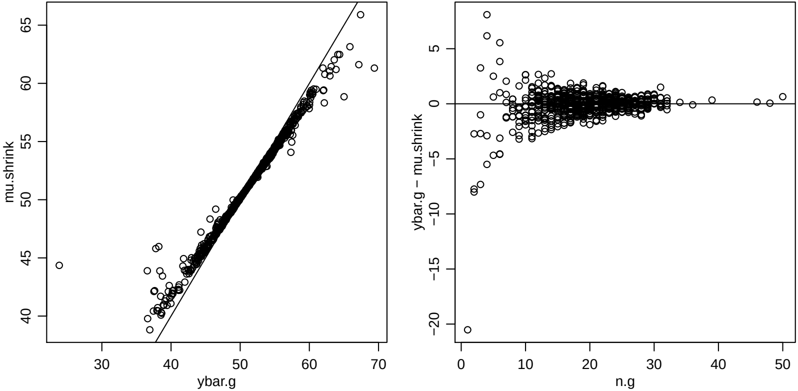
$$\hat{\mu}_j = w_j \bar{y}_j + (1 - w_j)\hat{\mu} \text{ , where } w_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}.$$

```
w.shrink<- (n.g/s2.mle) /(n.g/s2.mle + 1/t2.mle)

mu.shrink<-w.shrink*ybar.g + (1-w.shrink)*mu.mle

mu.mle

## (Intercept)
##     50.9391

cbind(ybar.g, n.g, mu.shrink)[1:8,]

##     ybar.g n.g mu.shrink
## 1 51.19300  30  51.16909
## 2 49.37133  15  49.64119
## 3 38.06833  12  40.72335
## 4 46.12172  29  46.58949
## 5 44.36308  13  45.63544
## 6 48.53091  22  48.82991
## 7 50.28111  18  50.37828
## 8 55.55792  24  55.02674
```

Shrinkage

```
cbind(ybar.g, n.g, mu.shrink)[topten,]

##       ybar.g n.g mu.shrink
## 639 69.40250   4  61.31365
## 349 67.40645  31  65.90120
## 618 67.15500   6  61.60894
## 616 65.86786  14  63.14810
## 386 65.01750   4  58.84972
## 337 64.37632  19  62.48167
## 637 64.12091  22  62.48426
## 73  63.86083  12  61.19530
## 680 63.59818  22  62.02644
## 352 63.16263  19  61.43912
```

Shrinkage



Shrinkage estimates from lme4

```
mu.shrink[1:10]

##        1        2        3        4        5        6        7        8
## 51.16909 49.64119 40.72335 46.58949 45.63544 48.82991 50.37828 55.02674
##        9       10
## 51.19648 48.70906

a.shrink<-ranef(fit.lme)[[1]][,1]

mu.mle+a.shrink[1:10]

##  [1] 51.16909 49.64119 40.72335 46.58949 45.63544 48.82991 50.37828
##  [8] 55.02674 51.19648 48.70906
```

In `lme4`, `ranef(fit.lme)[[k]][,l]` refers to the

- `l`th random effect for the
- `k`th grouping variable.

Confidence intervals for group means

CIs from unbiased estimators: How far away is $\bar{y}_j$ from μ_j?

$$\mathrm{E}[(\bar{y}_j - \mu_j)^2 | \mu_j] = \sigma^2 / n_j$$

An approximate 95% CI for μ_j is

$$\bar{y}_j \pm 2\sqrt{\hat{\sigma}^2 / n_j}$$

Confidence intervals for group means

CIs from shrinkage estimators: How far away is $\hat{\mu}_j$ from μ_j? We showed

$$\mathrm{E}[(\hat{\mu}_j - \mu_j)^2 | \mu_j] = w^2 \sigma^2 / n_j + (1 - w)^2 (\mu_j - \mu)^2$$

On average across groups, this squared distance is

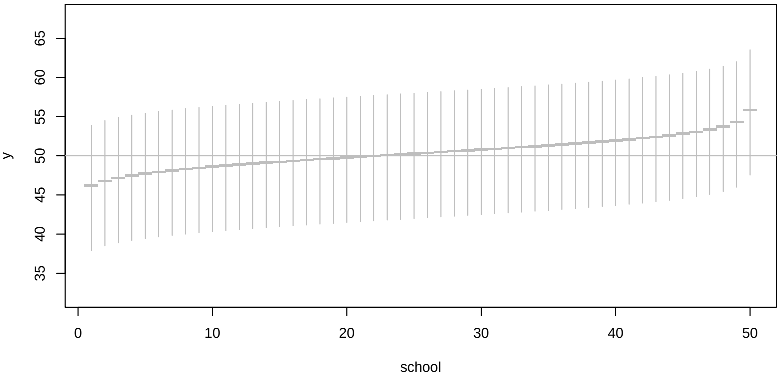
$$\begin{aligned} \mathrm{E}[(\hat{\mu}_j - \mu_j)^2] = w^2 \sigma^2 / n_j + (1 - w)^2 \tau^2 &= \left(\tfrac{\tau^2}{\tau^2 + \sigma^2 / n_j} \right) \sigma^2 / n_j \\ &= \frac{1}{1/\tau^2 + n_j / \sigma^2} \end{aligned}$$

An approximate 95% CI for μ_j is

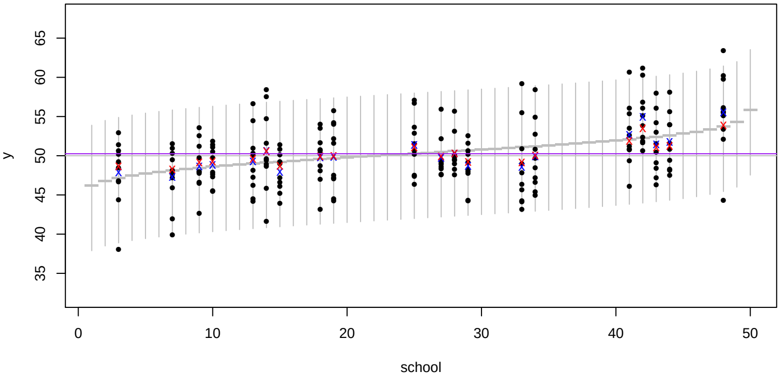
$$\hat{\mu}_j \pm 2 \sqrt{\tfrac{1}{1/\tau^2 + n_j / \sigma^2}}.$$

The 95% coverage is on average, across groups.

Simulation study

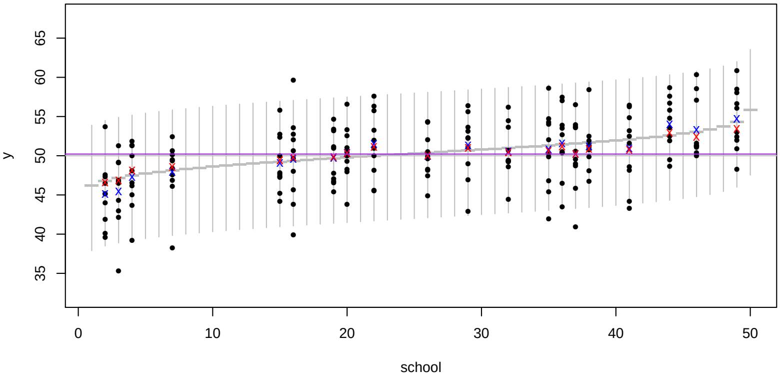


Simulation study



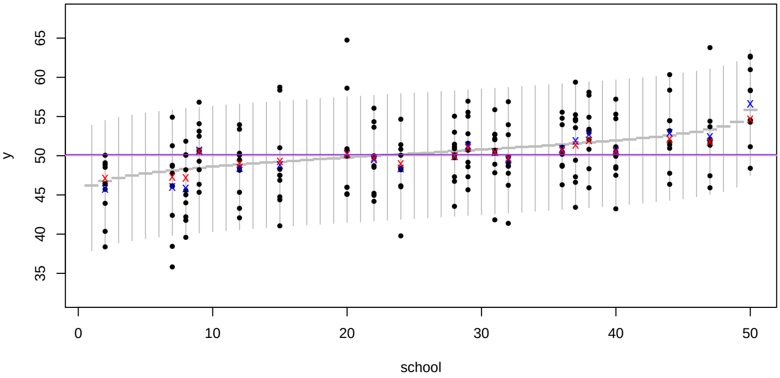

```
mean( (mu.shrink - mu.G[j.samp])^2 )

## [1] 1.036385

mean( (ybar - mu.G[j.samp])^2 )

## [1] 1.647306
```

Simulation study



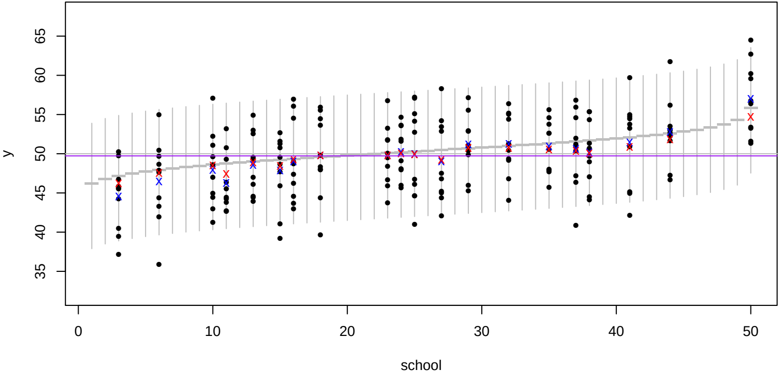

```
mean( (mu.shrink - mu.G[j.samp])^2 )

## [1] 0.4553131

mean( (ybar - mu.G[j.samp])^2 )

## [1] 0.7501884
```

Simulation study



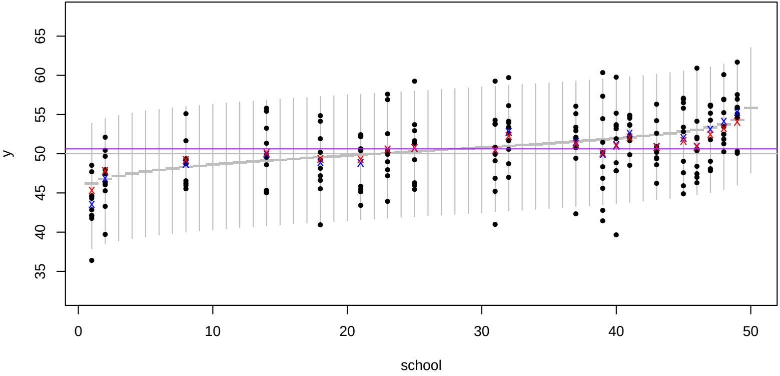

```
mean( (mu.shrink - mu.G[j.samp])^2 )

## [1] 0.7377844

mean( (ybar - mu.G[j.samp])^2 )

## [1] 1.266737
```

Simulation study




```r
mean( (mu.shrink - mu.G[j.samp])^2 )

## [1] 0.4162846

mean( (ybar - mu.G[j.samp])^2 )

## [1] 1.01239
```

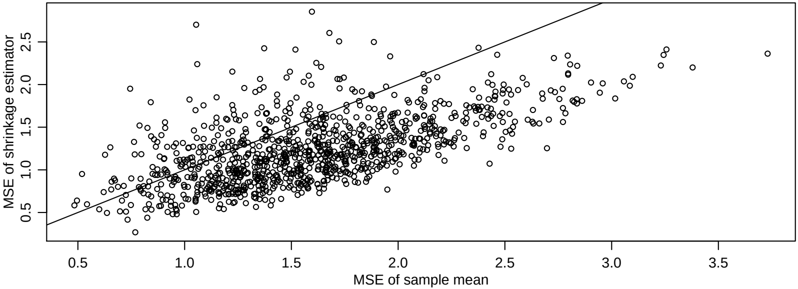
Simulation study




```
mean( (mu.shrink - mu.G[j.samp])^2 )

## [1] 0.9041727

mean( (ybar - mu.G[j.samp])^2 )

## [1] 1.247571
```

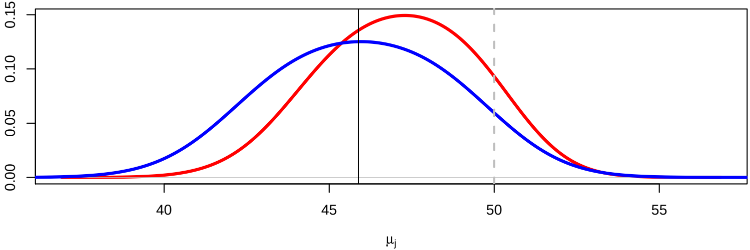
Simulation study

```
## Warning in optwrap(optimizer, devfun, getStart(start, rho$lower, rho$pp), :
convergence code 3 from bobyqa:  bobyqa -- a trust region step failed to reduce q
```



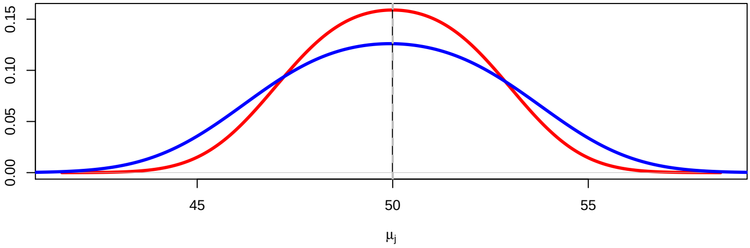

```
mean(MSE[,1])

## [1] 1.60045

mean(MSE[,2])

## [1] 1.24197

mean(MSE[,2]<MSE[,1])

## [1] 0.813
```

Inference for an underperforming school



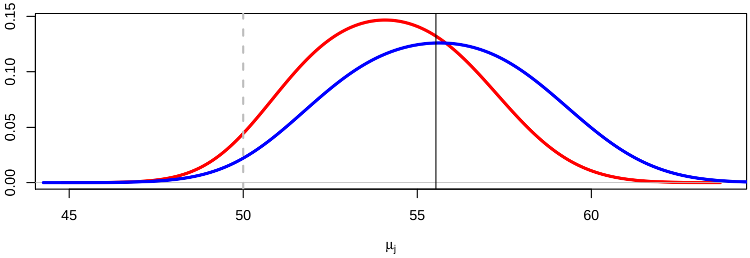

```
# MSE of ybar
mean( (UCI[[j]][,2] - mu.G[j] )^2 )

## [1] 1.748728

# MSE of shrinkage estimator
mean( (SCI[[j]][,2] - mu.G[j] )^2 )

## [1] 2.824414
```

Inference for a middling school



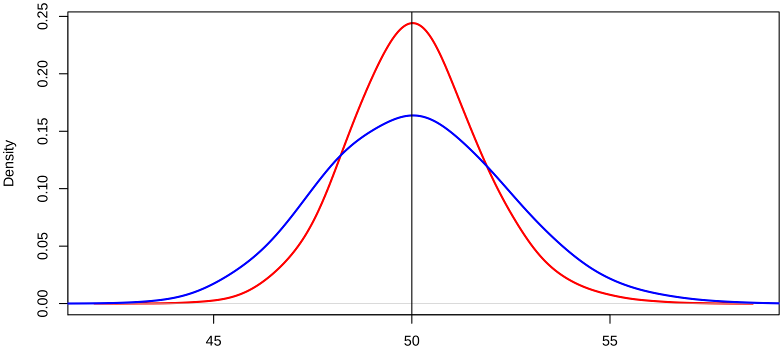

```r
# MSE of ybar
mean( (UCI[[j]][,2] - mu.G[j] )^2 )

## [1] 1.566342

# MSE of shrinkage estimator
mean( (SCI[[j]][,2] - mu.G[j] )^2 )

## [1] 0.7849105
```

Inference for an overperforming school

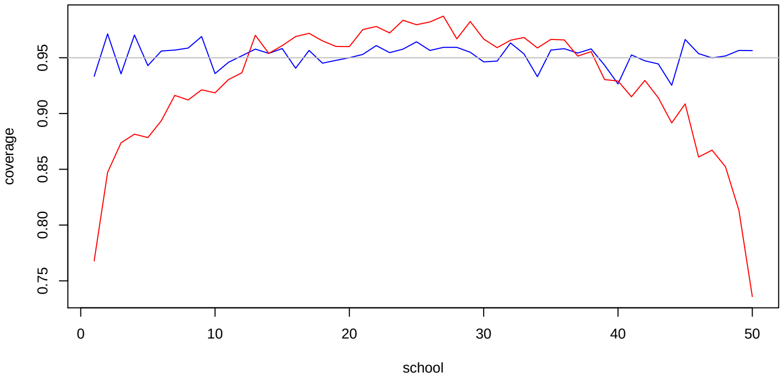



```r
# MSE of ybar
mean( (UCI[[j]][,2] - mu.G[j] )^2 )

## [1] 1.627534

# MSE of shrinkage estimator
mean( (SCI[[j]][,2] - mu.G[j] )^2 )

## [1] 3.129529
```

Unconditional unbiasedness of estimates



Confidence interval coverage



Summary

- coverage for schools with extreme values of μ_j is too low;
- coverage for schools with middling values of μ_j is too high.

Advice:

- Estimation and confidence interval construction are different tasks.
- Use a procedure that aligns with your data analysis goals.
- Be aware of the statistical properties of your analysis procedures.