Statistics 360/601 – Modern Bayesian Theory

Alexander Volfovsky

Lecture ? - Oct 30, 2018

A little bit about our data

https://nces.ed.gov/surveys/els2002/

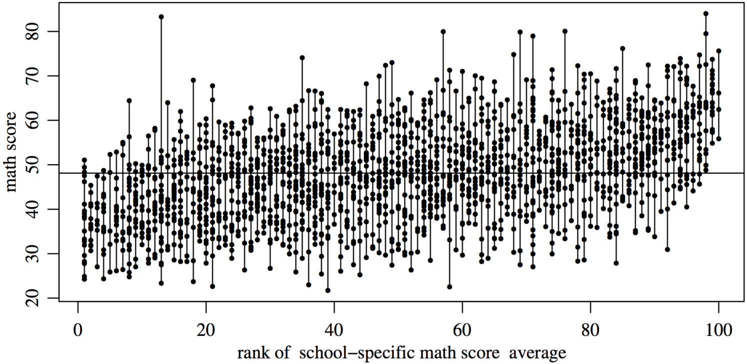
- Nationally representative, longitudinal study of 10th graders in 2002 and 12th graders in 2004
- Students followed throughout secondary and postsecondary years
- Surveys of students, their parents, math and English teachers, and school administrators
- Student assessments in math (10th & 12th grades) and English (10th grade)
- 2002 Focus: What are students' trajectories from the beginning of high school into postsecondary education, the workforce, and beyond?
- 2002 Focus: What are the different patterns of college access and persistence that occur in the years following high school completion

A little bit about our data

Table 1. Percentage of spring 2002 high school sophomores, by high school completion status and select student characteristics: 2006

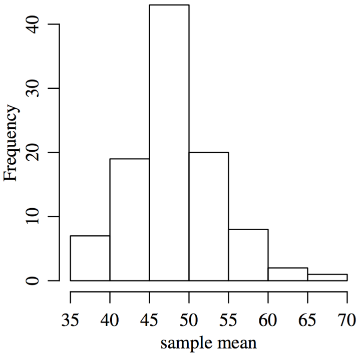
Characteristic	Received high school diploma	Received GED or other equivalency	Enrolled in high school or working toward equivalency	No diploma; not enrolled or working toward equivalency
Total	87.8	3.9	3.2	4.6
Sex				
Female	90.0	3.2	2.6	3.8
Male	85.7	4.7	3.8	5.4
Race/ethnicity[1]				
American Indian or Alaska Native	74.7	9.3	4.4	11.7!
Asian or Pacific Islander	92.6	2.1	2.6	2.4
Black or African American	82.2	4.6	6.0	5.7
Hispanic or Latino	80.9	4.3	4.3	10.0
White	91.1	3.5	2.2	2.9
More than one race	85.1	5.7	4.0	5.0
Family income				
$0–20,000	78.0	6.2	5.6	9.1
$20,001–50,000	85.3	4.5	3.9	6.0
$50,001–100,000	92.1	3.0	2.1	2.5
$100,001 or more	95.5	2.0	1.3	0.9
Parental education				
High school or less	80.4	4.6	5.1	9.1
Some college	88.1	4.6	3.2	3.7
Bachelor's degree	92.7	2.8	2.0	2.3
Graduate/professional degree	93.4	2.8	1.6	1.9
Native language[2]				
English	88.6	4.0	3.1	3.8
Non-English	82.8	3.2	3.7	9.5
School sector				
Public	87.0	4.1	3.4	4.9
Catholic	98.1	1.3	0.2	0.3
Other private	96.2	2.1	0.8	0.7
Educational expectation in 10th grade				
High school or less	62.5	10.1	8.8	17.3
Some college	79.0	7.2	5.6	7.6
Bachelor's degree	91.8	3.0	2.4	2.6
Graduate/professional degree	94.8	1.7	1.5	1.7

What are we studying?



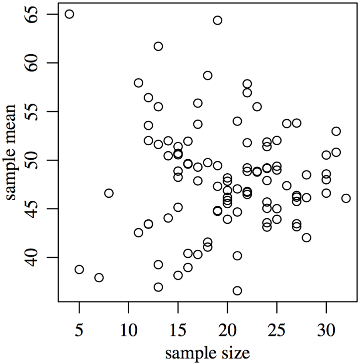
- Test scores for 10th graders in 100 urban high schools.
- Ordered according to school averages.

More about the data



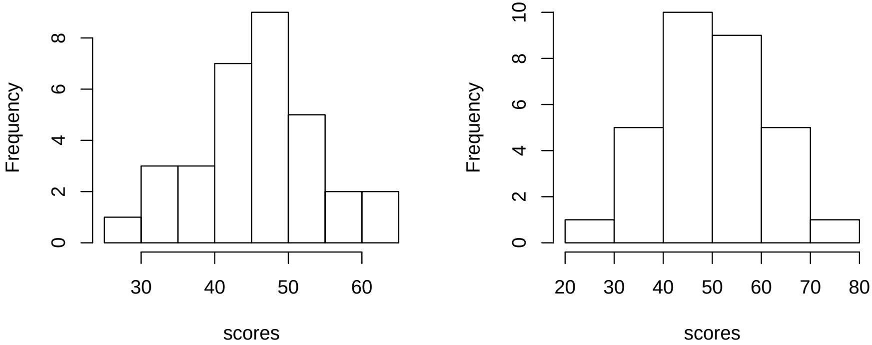
- Looks “smooth” (this is a histogram of averages).
- Normal model seems reasonable.

Are all data points created equal?



- We should trust averages based on more data points.
- Essentially: if $\theta_j = \theta$ then $E[\bar{Y}|\theta_j, \sigma^2] = \theta$ but $Var[\bar{Y}|theta_j, \sigma^2] = \sigma^2/n_j$

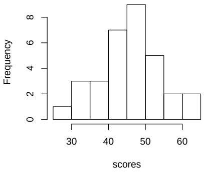
What happens within a school?



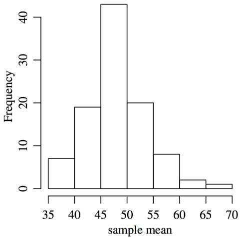
- Two of the larger schools ($n_{37} = 32$ and $n_1 = 31$)
- Also looks pretty normal.

Sampling model?

- Student i within school j: $Y_{ij}|\theta_j, \sigma^2 \sim$ *normal*(θ_j, σ^2)



- Mean of school j: $\theta_j|\mu, \tau^2 \sim$ *normal*(μ, τ^2)



Prior setup

- Need to specify prior parameters for $p(\mu), p(\sigma^2), p(\tau^2)$.
- Within school variance: $\sigma_0^2 = 100$ because the test is designed to have this variance. $\nu_0 = 1$ for a weakly concentrated prior.
- Between school variance: $\tau_0^2 = 100$ and $\eta_0 = 1$ for the same reason.
- School means: $\mu_0 = 50$ because the test is designed to have this mean, $\gamma_0^2 = 25$ lets the actual school mean move a bit but not too much.

Gibbs sampler

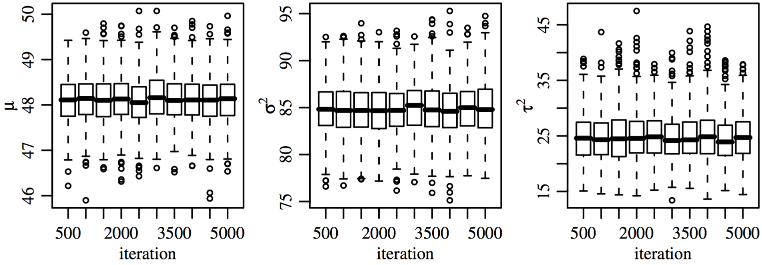
1. Sample $\mu^{(s+1)} \sim \; p(\mu|\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \tau^{2(s)})$
2. Sample $\tau^{2(s+1)} \sim \; p(\tau^2|\theta_1^{(s)}, \ldots, \theta_m^{(s)}, \mu^{(s+1)})$
3. Sample $\sigma^{2(s+1)} \sim \; p(\sigma^2|\theta_1^{(s)}, \ldots, \theta_m^{(s)}, y_1, \ldots, y_m)$
4. For each $1 \leq j \leq m$ sample $\theta_j^{(s+1)} \sim \; p(\theta_j|\mu^{(s+1)}, \tau^{2(s+1)}, \sigma^{2(s+1)}, y_j)$

Possible implementation orders:

- Do 1, 2, 3, 4
- Do 2, 1, 3, 4 (with obvious changes)
- ...
- $(i_1, i_2, i_3, i_4) \sim \pi$ where π

MCMC diagnostics

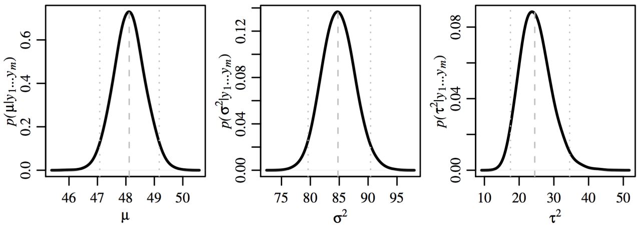
1. Stationarity plots



2. Lag-t autocorrelation
   Lag-1 for $\mu, sigma^2, \tau^2$ are 0.15, 0.053 and 0.312.
3. Effective sample sizes
   For $\mu, sigma^2, \tau^2$: 3706, 4499, and 2503.
4. Monte Carlo standard errors
   For $\mu, sigma^2, \tau^2$: 0.009, 0.04, 0.09

Posterior summaries

Marginal posterior distributions



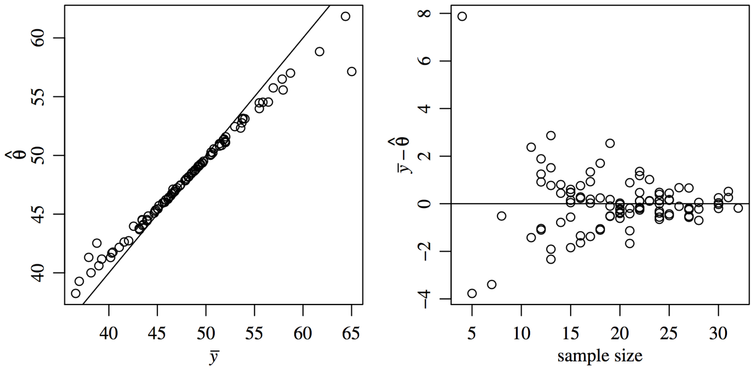
- Posterior means for $\mu, sigma, \tau$ are 48.12, 9.21, 4.97.
- 95% of scores within a classroom are within $\approx 4 \times 9.21 \approx 37$ points of each other.
- 95% of the average classroom scores are within $\approx 4 \times 4.97 \approx 20$ points of each other.

Why hierarchical models?

- Shrinkage!
- Conditioning on μ, τ^2, σ^2 and the data we have

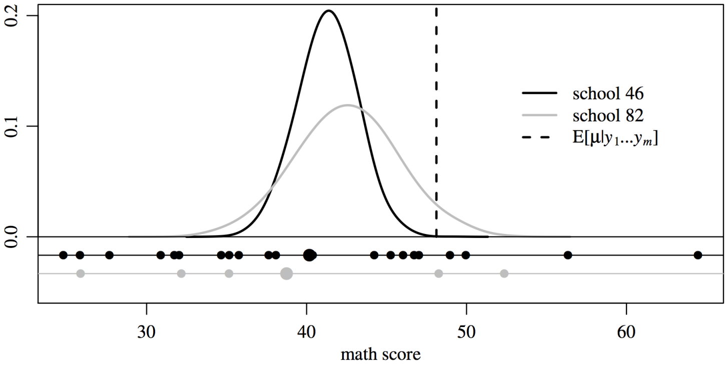
$$E[\theta_j | y_j, \mu, \tau^2, \sigma^2] = \frac{\bar{y}_j n_j / \sigma^2 + \mu / \tau^2}{n_j / \sigma^2 + 1 / \tau^2}$$

- n_j small then $E[\theta_j | \ldots]$ is pulled away from $\bar{y}_j$ towards to μ.

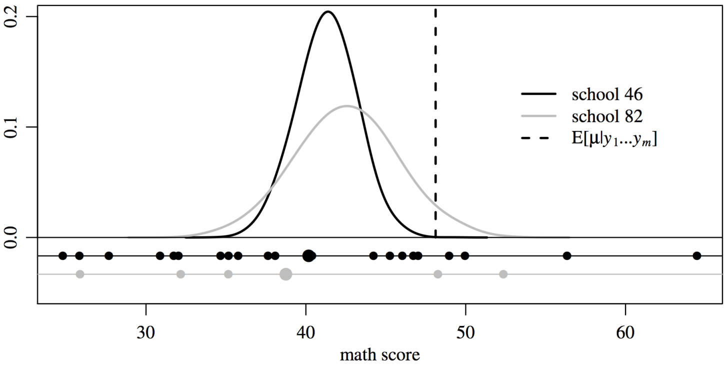


Does this change inference?

- It can — we are interested in ranking the schools based on their students' performances.
- If we give everyone the test in each school it makes sense to compare posterior expectations $E[\theta_j|y_1, \ldots, y_m]$.
- Almost the same inference as from $\bar{y}_j$...



Does this change inference?



- $\bar{y}_{46} = 40.18,\ n_{46} = 21,\ E[\theta_{46}| \dots] = 41.31$
- $\bar{y}_{82} = 38.76,\ n_{82} = 5,\ E[\theta_{82}| \dots] = 42.53$
- Removing lowest scores: $\tilde{y}_{46} = 40.9,\ \tilde{y}_{82} = 41.99$

So many parameters...

- Sometimes setting hyperparameters is hard.
- Consider our sampling model:

$$p(y_{ij}|\theta_j, \sigma^2)p(\theta_j|\mu, \tau^2)$$

- The marginal of the data is

$$\int p(y_{ij}|\theta_j, \sigma^2)p(\theta_j|\mu, \tau^2)d\theta_j$$

- We can estimate μ, τ^2 from that — this the beginning of an Empirical Bayes procedure.
- EB provides inference for parameters of interest θ_j but ignores uncertainty about hyperparameters like τ^2.
- Empirical Bayes estimators: Based on the posterior $p(\theta_j|y_j, \hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2)$.