Lecture 7: Hierarchical Modeling

Professor Alexander Franks

2020-11-29

Announcements

- Fill out ESCI evaluations online!
    - What did you like? Would you like to see more courses like this?
    - What can be improved?

Comparing Multiple Related Groups

- Hierarchy of nested populations
- Models which account for this are called *hiearchical* or *multi-level* models

Some examples:

- Patient outcomes within several different hospitals
- People within counties in the United States (e.g. Asthma mortality example)
- Athlete performance in sports
- Genes within a group of animals

Eight schools example

- A study was performed for the Educational Testing Service (ETS) to evaluate the effects of coaching programs on SAT preparation
- Each of eight different schools used a short-term SAT prep coaching program
- Compute the average SAT score in those who did take the program minus those that did not participate in the program
- We observe the average difference varies by school. What accounts for these differences?

Eight schools example

- Interested in "real" differences due to training
- Want to reduce effect of chance variability
- How do we estimate the effect of the program in each of the schools?
- Two extremes:
  - Estimate the effect of the program in every school independently
  - A separate prior distribution for each school effect
  - Or assume the effect is the same in every school
  - Combine all the data
  - A compromise between the above 2 options?

Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- y_j is the observed effects of the program in school j
    - Based on a sample of test scores from those in the program and those not in the program
- θ_j are the true *unknown* effects of the program in school j
- Variances, σ_j^2, are *known*
    - Determined by the number of students in the sample

Eight Schools Example

```
J <- 8
y = c(28,  8, -3,  7, -1,  1, 18, 12)
sigma <- c(15, 10, 16, 11,  9, 11, 10, 18)
```

- Assuming the effect of the program on each school is identical.
- What are the chances of seeing a value as large as 28?
- As small as -3?

Eight Schools Example

If effects are actual equal, what is it?

```
## Compute the precision frome each school
prec <- 1/sigma^2

## global estimate is a weighted vareage
mu_global <- sum(prec * y / sum(prec))
mu_global
```

```
## [1] 7.685617
```

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- What are the chances of seeing a value as large as 28?
- As small as -3?

```
# 1000 datasets with mean mu_global but different sigmas

## Chance of seeing a value greater than 28
mean(sapply(1:1000, function(x)
  max(rnorm(J, mean=mu_global, sd=sigma))) >= 28)
```

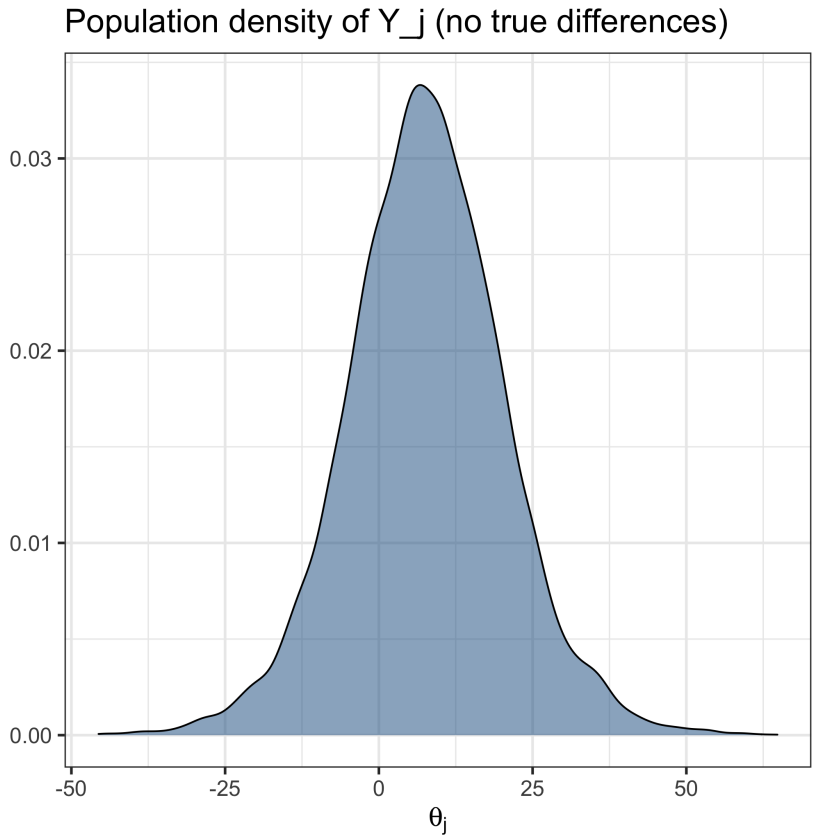
```
## [1] 0.378
```

```
## Chance of seeing a value less than -3
mean(sapply(1:1000, function(x)
  min(rnorm(J, mean=mu_global, sd=sigma))) <= -3)
```

```
## [1] 0.81
```

Eight Schools Example

Density of $Y_j \sim N(\theta_{\text{pooled}}, \sigma_j^2)$



Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- θ_j are the true unknown effects of the program in school j
- y_j is the observed effects of the program in school j
    - Based on a sample of test scores from those in the program and those not in the program
    - Number of people in the sample determine the magnitude of σ_j^2

Eight Schools Example

- How do we estimate θ_j?
  - Independent: $\hat{\theta}_j^{(MLE)} = y_j$ is the MLE
  - Identical effects: $\hat{\theta}_j^{(pool)} = \frac{\sum_i \frac{1}{\sigma_i^2} y_i}{\sum \frac{1}{\sigma_i^2}}$
    - Same effect for all schools: estimate using a weighted average of the observed effects
- Compromise: $\hat{\theta}_j^{(shrink)} = w\theta_j^{(MLE)} + (1-w)\theta_j^{(pool)}$

Eight Schools

```
theta_j_mle <- y
theta_j_mle
```

```
## [1] 28  8 -3  7 -1  1 18 12
```

```
theta_j_pooled <- rep(sum(1/sigma^2 * y) / sum(1/sigma^2), J)
theta_j_pooled
```

```
## [1] 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685
```

Is there a middle ground between two extremes?

Eight schools example

A hierarchical model:

$$\theta_i \sim N(\mu, \tau^2)$$
$$y_j \sim N(\theta_j, \sigma_j^2)$$

- Add a *shared* normal prior distribution to θ_j
- Assume the global mean, μ is also unknown
- How do we choose prior for μ?

$\mu \sim N(\mu_0, \tau_0^2)$? or $p(\mu) \propto 1$?

- Need to estimate all of $(\mu, \theta_1, \ldots, \theta_8)$ with MCMC
- τ^2 determines how much weight weight we put on the independent estimate vs the pooled estimate

Intuition behind shrinkage

- $Y_j = \theta_j + \epsilon_j$ and for simplicity assume that the variance of ϵ, σ^2 for all j
    - θ_j represents true differences between schools (signal)
    - ϵ_j is sampling variability (noise, chance variation)
- $Var(Y_j) = Var(\theta_j) + Var(\epsilon_j) = \tau^2 + \sigma^2$
    - The variance of the observed outcomes is the sum of signal variance, τ^2, and the sampling variance σ_j^2
- Consequence: the observed outcomes always have higher variance across groups than the signal
    - $\text{Var}(Y_j) > \text{Var}(\theta_j)$
- Solution: reduce the variance by shrinking them!
    - Want the variance of the shrunken estimates to be close to τ^2

Eight Schools examples

Comments:

- The global average, μ, is a parameter so also has uncertainty
- How dow we determine how much to shrink, e.g. how do we determine τ^2?
- Is the training program effective in school j?
  - What is $P(\theta_j > 0 \mid y)$?
- On avearge (over all schools) is the training program effective?
  - What is $P(\mu > 0 \mid y)$?

Eight schools example

- If τ^2 is large, the prior for θ_j is not very strong
  - If $\tau^2 \to \infty$ equivalent to the no pooling model
- If τ^2 is small, we assume a priori that θ_j are very close
  - if $\tau^2 \to 0$ equivalent to the complete pooling model, $\theta_j = \mu$

Estimating parameters

- MH: Need to generate a proposal from a 9-dimensional posterior distribution
    - Eight parameters for θ_j and one for μ
- Gibbs: sample each of the 9 parameters from the complete conditionals
    - Sample $p(\theta_j \mid \theta_{-j}, \mu)$
    - Sample $p(\mu \mid \theta_1, \ldots \theta_8, \mu)$
- Stan

Eight Schools Estimation

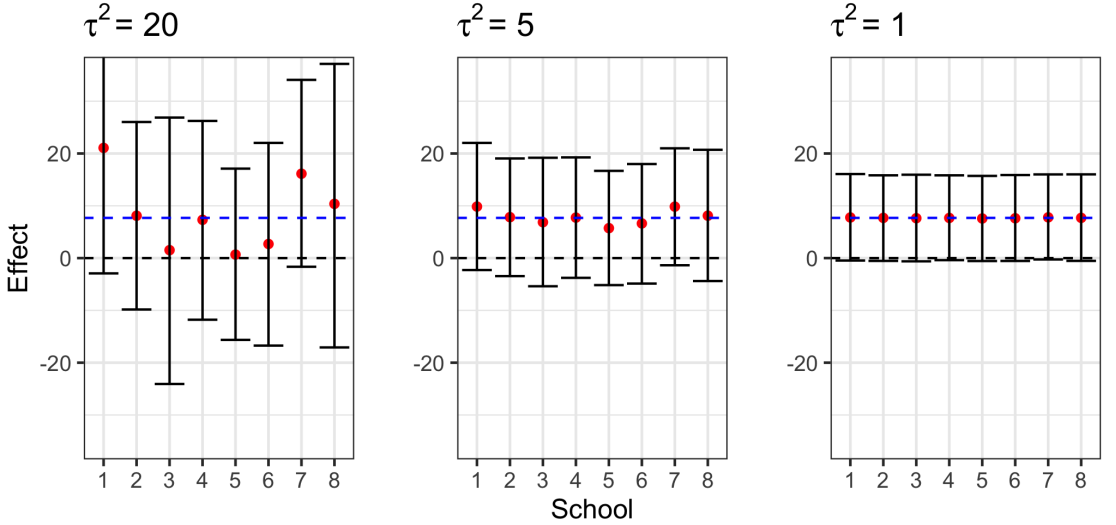
```
J <- 8
y = c(28,  8, -3,  7, -1,  1, 18, 12)
sigma <- c(15, 10, 16, 11,  9, 11, 10, 18)

## fix the variance of the prior to a number
tau <- 20
```

Eight Schools in Stan

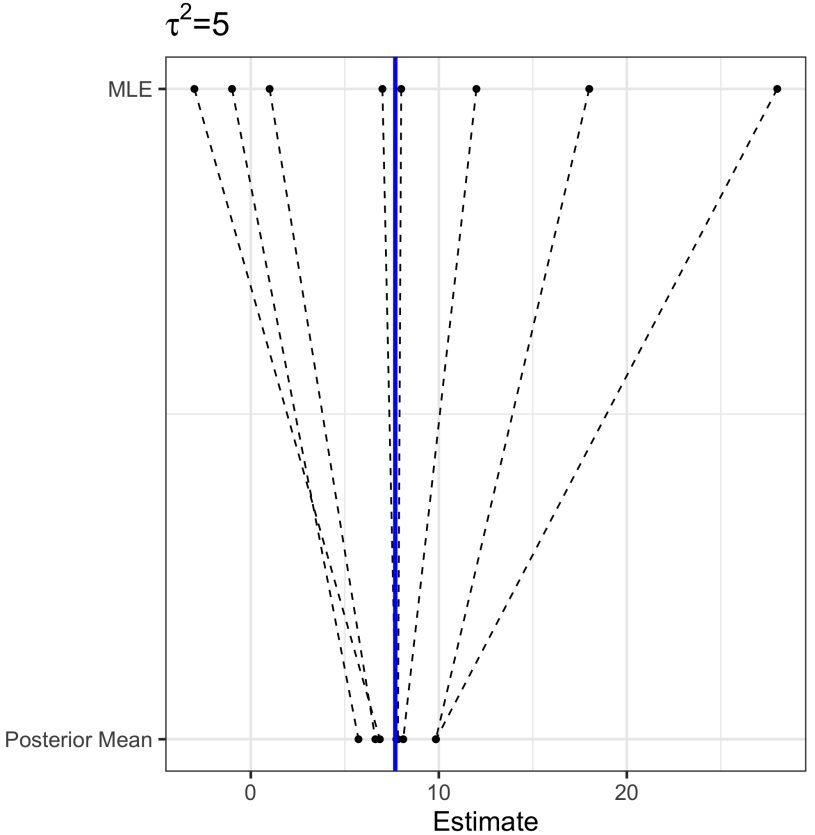
```
// saved as 8schools.stan
data {
  int<lower=0> J;         // number of schools
  real y[J];              // estimated treatment effects
  real<lower=0> sigma[J]; // standard error of effect estimates
  real<lower=0> tau;      // shrinkage standard deviation
}
parameters {
  real mu;         // population treatment effect
  vector[J] eta;  // unscaled deviation from mu by school
}
transformed parameters {
  vector[J] theta = mu + tau * eta;  // school treatment effects
}
model {
  target += normal_lpdf(eta | 0, 1);       // prior log-density
  target += normal_lpdf(y | theta, sigma); // log-likelihood
}
```

Eight Schools example

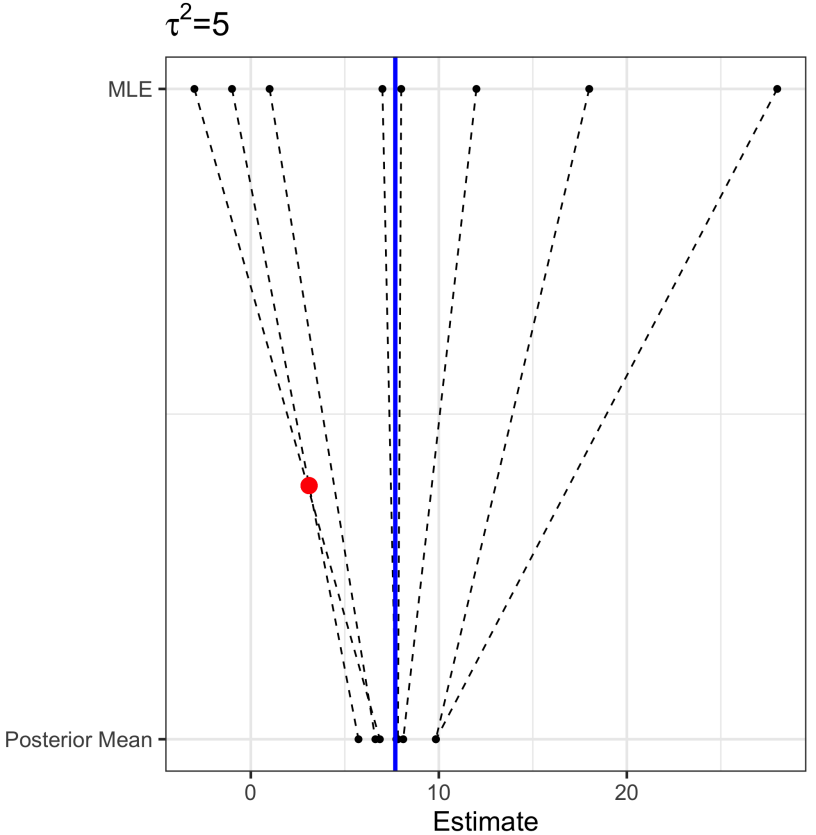
```
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```



MLE vs Posterior Mean



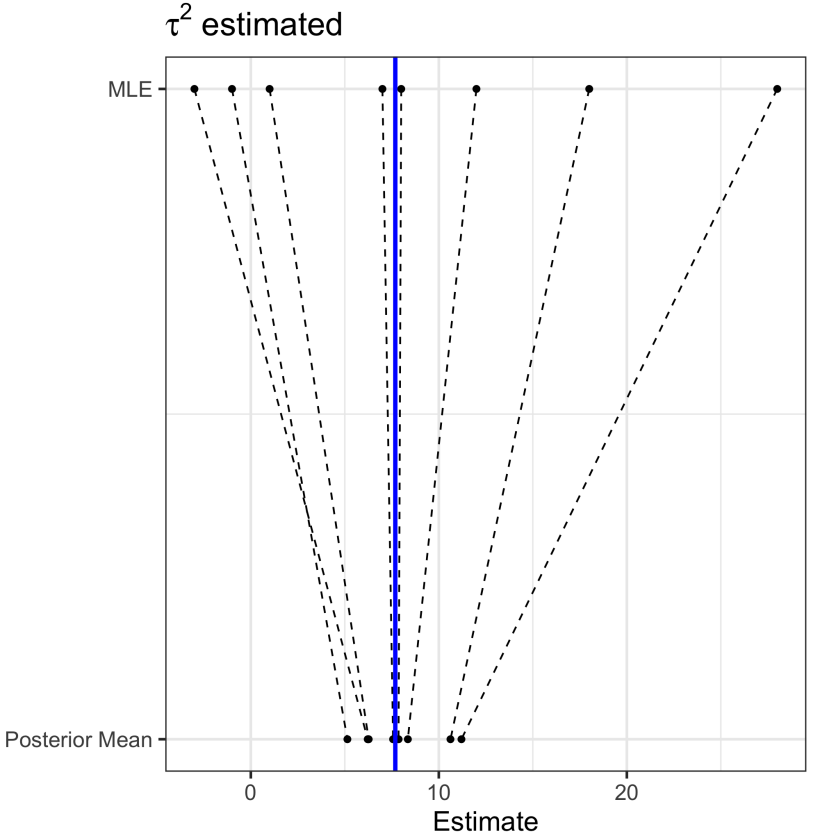
MLE vs Posterior Mean



Eight Schools in Stan

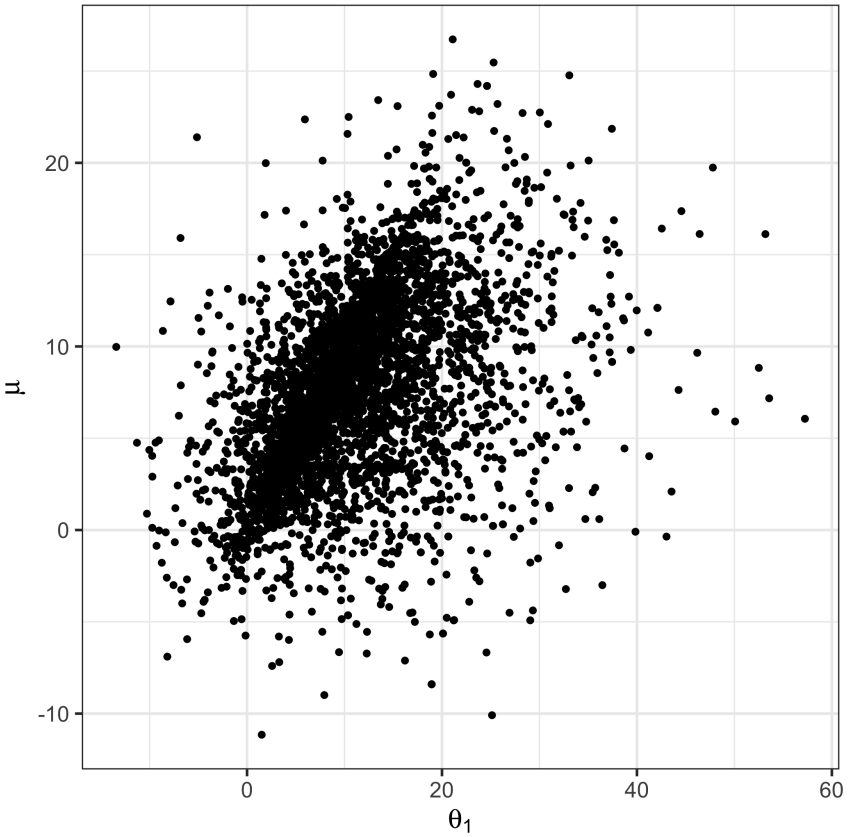
```
// saved as 8schools.stan
data {
  int<lower=0> J;         // number of schools
  real y[J];              // estimated treatment effects
  real<lower=0> sigma[J]; // standard error of effect estimates
}
parameters {
  real mu;             // population treatment effect
  real<lower=0> tau;  // standard deviation in treatment effects #
  vector[J] eta;       // unscaled deviation from mu by school
}
transformed parameters {
  vector[J] theta = mu + tau * eta; // school treatment effects
}
model {
  target += -2log(tau);        // prior for tau
  target += normal_lpdf(eta | 0, 1);       // normal model for the
  target += normal_lpdf(y | theta, sigma); // log-likelihood
}
```

MLE vs Posterior Mean

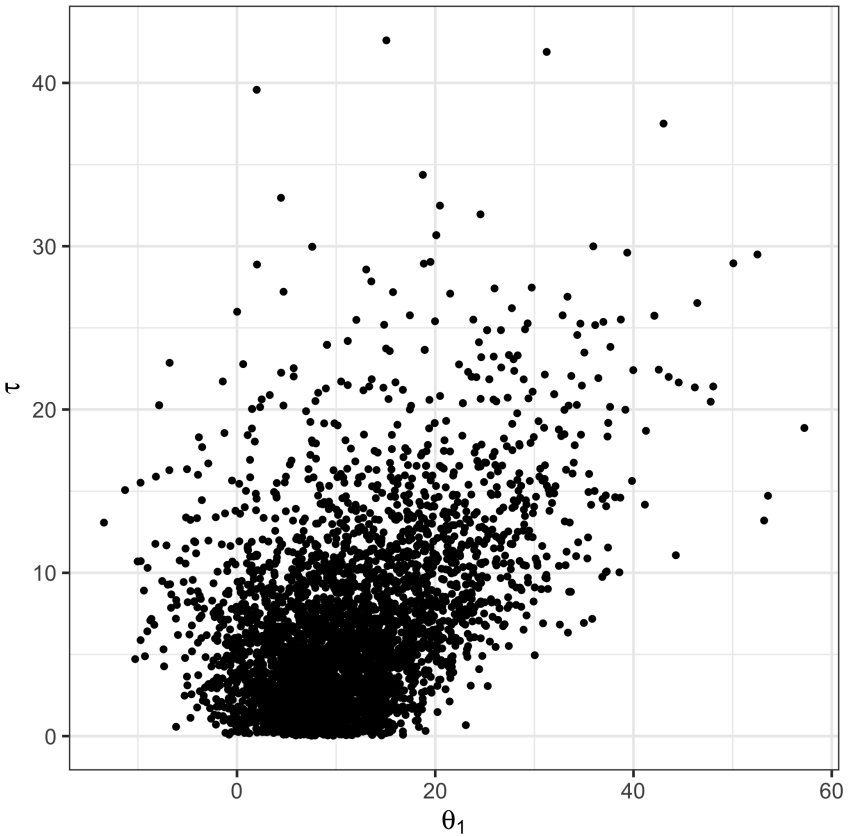


Posterior mean of $\tau^2 = 6.4633808$

Eight Schools Scatter Plot: θ_1 vs μ



Eight Schools Scatter Plot: θ_1 vs τ

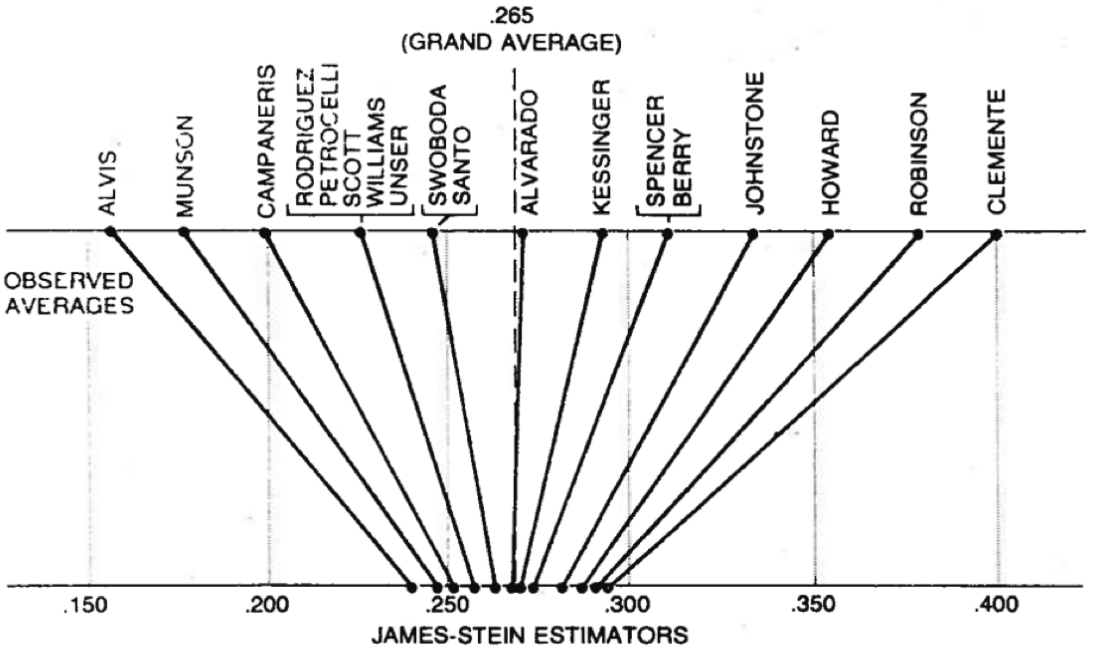


Hierarchical modeling in sports

1. 1970 Batting Averages for 18 Major League Players and Transformed Values X_i, θ_i

i	Player	Y_i = batting average for first 45 at bats (1)	p_i = batting average for remainder of season (2)	At bats for remainder of season (3)	X_i (4)	θ_i (5)
1	Clemente (Pitts, NL)	.400	.346	367	−1.35	−2.10
2	F. Robinson (Balt, AL)	.378	.298	426	−1.66	−2.79
3	F. Howard (Wash, AL)	.356	.276	521	−1.97	−3.11
4	Johnstone (Cal, AL)	.333	.222	275	−2.28	−3.96
5	Berry (Chi, AL)	.311	.273	418	−2.60	−3.17
6	Spencer (Cal, AL)	.311	.270	466	−2.60	−3.20
7	Kessinger (Chi, NL)	.289	.263	586	−2.92	−3.32
8	L. Alvarado (Bos, AL)	.267	.210	138	−3.26	−4.15
9	Santo (Chi, NL)	.244	.269	510	−3.60	−3.23
10	Swoboda (NY, NL)	.244	.230	200	−3.60	−3.83
11	Unser (Wash, AL)	.222	.264	277	−3.95	−3.30
12	Williams (Chi, AL)	.222	.256	270	−3.95	−3.43
13	Scott (Bos, AL)	.222	.303	435	−3.95	−2.71
14	Petrocelli (Bos, AL)	.222	.264	538	−3.95	−3.30
15	E. Rodriguez (KC, AL)	.222	.226	186	−3.95	−3.89
16	Campaneris (Oak, AL)	.200	.285	558	−4.32	−2.98
17	Munson (NY, AL)	.178	.316	408	−4.70	−2.53
18	Alvis (Mil, NL)	.156	.200	70	−5.10	−4.32

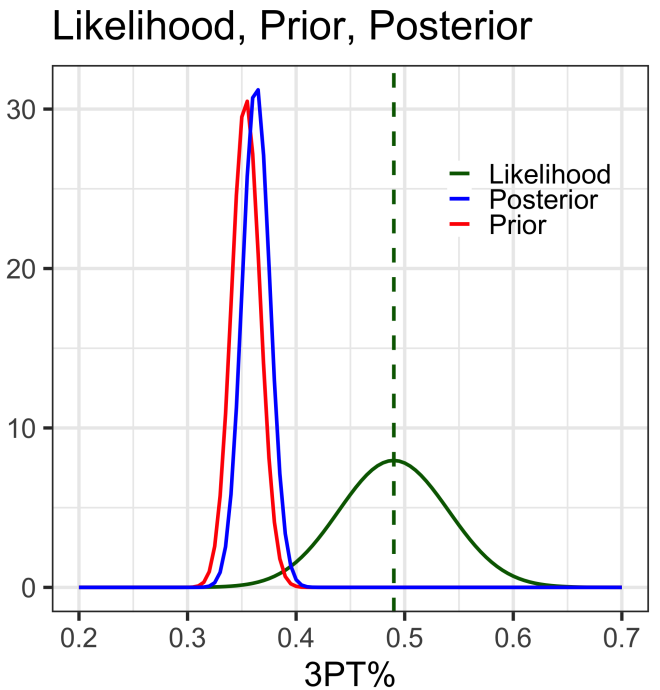
Hierarchical modeling in sports



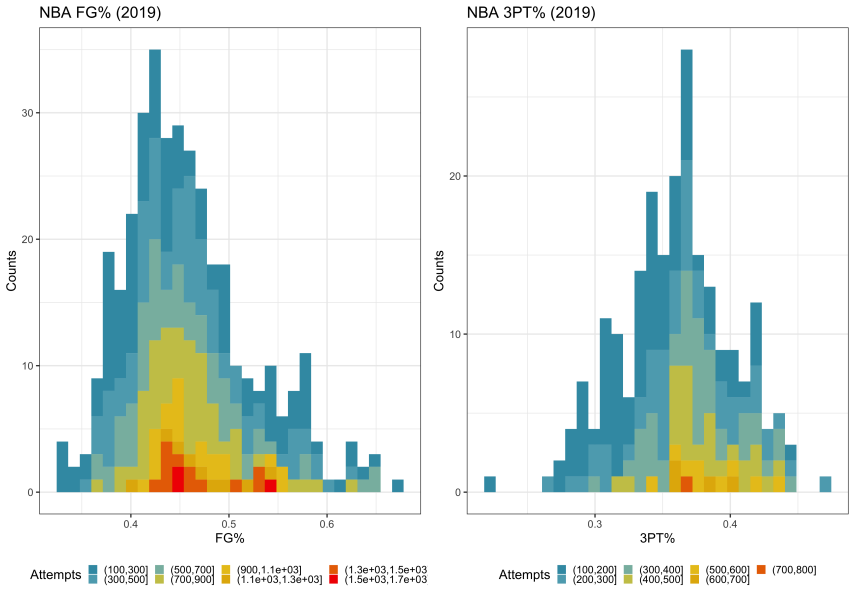
JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Basketball Example

After 100 shots Robert Covington's 3PT% was 0.49



Basketball Example



Eight Schools and Independence

- The effects θ_j in each school are *not* independent
- E.g. if we know θ_1 is really large, it means θ_j is also more likely to be large
- However, before we see data the distributions of θ_j's, are indistinguishable
  - Don't know which θ's will be large and which small
  - This kind of dependence is known as *exchangeability*

Exchangeability

An exchangeable sequence of random variables is a finite (or infinite) sequence $X_1, X_2, \ldots X_n$ of random variables such that for any finite permutation π of the indices 1, 2, 3, ..., , the joint probability distribution of the permuted sequence

$$X_{\pi(1)}, X_{\pi(2)}, X_{\pi(3)}, \ldots$$

is the same as the joint probability distribution of the original sequence.

Exchangeability

- Consider a set of J experiments in which experiment j as data y_j and parameter θ_j
- The J experiments are related, but no information in the indices that distinguish θ_j
- Assume an exchangeable prior distribution: $p(\theta_1, \ldots \theta_J) = p(\theta_{\pi_1}, \ldots \theta_{\pi_J})$ where π is any permutation of the indices $1 \ldots J$
- Equivalent to a prior assumption about symmetry among the parameters $(\theta_1, \ldots \theta_J)$

Exchangeability

- Example: All i.i.d random variables are exchangeable
- Example: Non-independent random variables can also be exchangeable
- I flip a coin 5 times and observe 2 heads. What is the distribution over the order in which they were flipped?
  - Let X_i be a 1 if the ith outcome of the flip was heads and zero otherwise
  - X_i are not independent because the number of observed heads is fixed
  - $p(X_1, X_2, \ldots X_5)$ is exchangeable

Exchangeability

- Example: multivariate normal with common mean μ and equi-correlation

$$Cov(Y) = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

Non-exchangeable random variables

Examples of *not* exchangeable variables include

- Time series data (closer in time = more correlated)
- Spatial data (closer in space = more correlated)
- Typically, ignorance (about indices) implies exchangeability
- When exchangeability doesn't hold, can often assume conditional exchangeability

Mixture of i.i.d random variables

- i.i.d random variables are exchangeable:
  $p(\theta_1, \ldots, \theta_n) = \prod_{j=1}^{J} p(\theta_j \mid \phi)$
- Mixtures of i.i.d random variables:

$$p(\theta_1, \ldots, \theta_n) = \int \left\{ \prod_{j=1}^{n} p\left(\theta_j | \phi\right) \right\} p(\phi) d\phi$$

  - If ϕ were known, θ_j would be i.i.d.
  - Since ϕ is not known, integrate over uncertainty
  - θ_j are a mixture of i.i.d random variables
  - Dependent but exchangeable

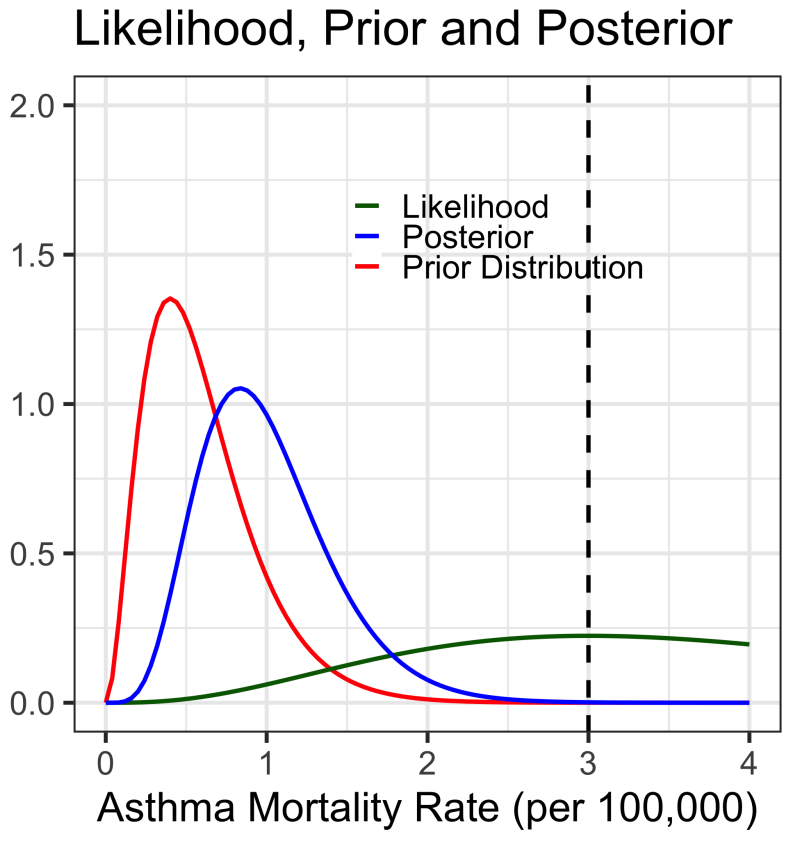
de Finetti's theorem

- **Theorem:** As $J \to \infty$ all exchangeable distributions can be expressed as a mixture of independent and identical distributions
- For a finite J, no guarantee that the variables can be represented as a mixture of i.i.d random variables
- But, if variables are exchangeable usually can be modeled as *approximately* a mixture of i.i.d
- Mixture's of i.i.d random variables $\to$ a hierarchical model!
    - For this reason de Finetti's theorem is often cited
- Good discussion in 2.7-2.8 of Hoff book

Poisson Example, 1 county

- In a particular county 3 people out of a population of 100,000 died of asthma
- Assume a Poisson sampling model with rate λ (units are rate of deaths per 100,000 people)
- Want to know the true mortality rate, λ
- We discussed how to specify a prior distribution for λ

Asthma Mortality



Poisson Example, many counties

- Assume we measure asthma mortality, y_j in n counties
- $Y_j \sim \text{Pois}(\lambda_j)$
- Does it make sense to model λ_j exchangeable?
- If we don't know how j relates to geographic location
  - $\lambda_j \sim \text{Gam}(a, b)$ and $p(a, b)$
- If we do know which geographic information relates to each j shouldn't model θ_j as exchangeabale

Poisson Example, many counties, many states

- Assume we measure asthma mortality, y_{ij} in county j in state i
- $Y_{ij} \sim \text{Pois}(\lambda_{ij})$
- Does it make sense to model λ_{ij} as exchangeable?
- Model: $\lambda_{ij} \sim \text{Gam}(a_i, b_i)$
- Can model county rates *within* state as exchangeable
- Can model parameters for state averages, (a_i, b_i), as exchangeable