S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Lecture 4.3
Randomized Experiments

Schedule Updates

We are slightly rearranging the final assignments

Final Project EDA	Due Tues, Nov 5
Project B (Group Assignment)	Due Fri, Nov 15
Final Project Presentations	December 12

(Multiple blocks in the 9am-4pm range.)

Technical report due day of final project presentation

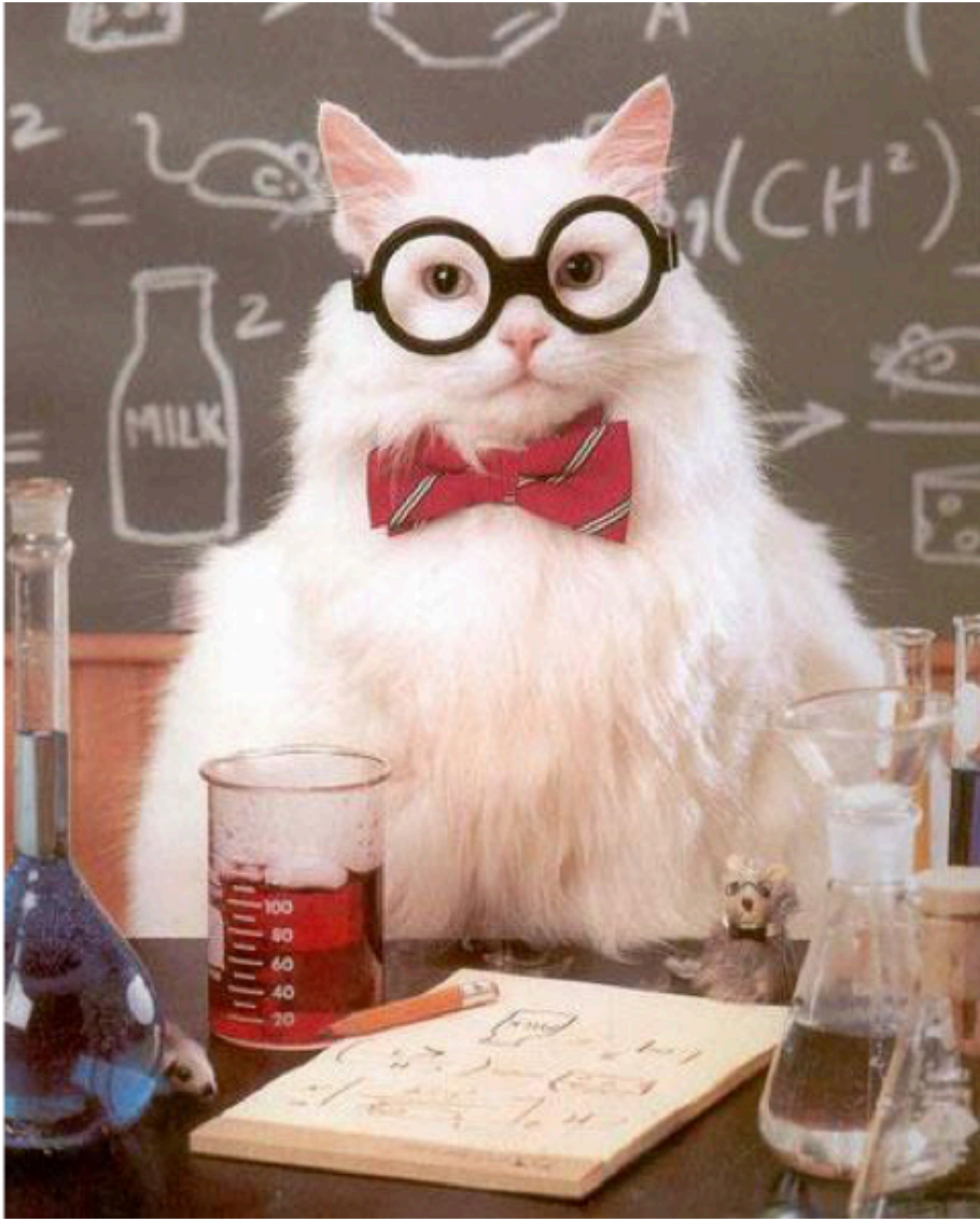
Goals for today

Talk about randomized trials and how you can analyze them with MLMs—the de facto standard in Education

In particular

- ★ Cluster randomized experiments (each school gets treated or not)
  - Power and cost effectiveness tradeoffs
- ★ Multisite randomized experiments (bunch of schools, where a fraction of the students in each get treated)
  - Correctly handling some potential sources of bias
  - Detecting cross site impact variation

Randomized Experiments



Randomized Experiments

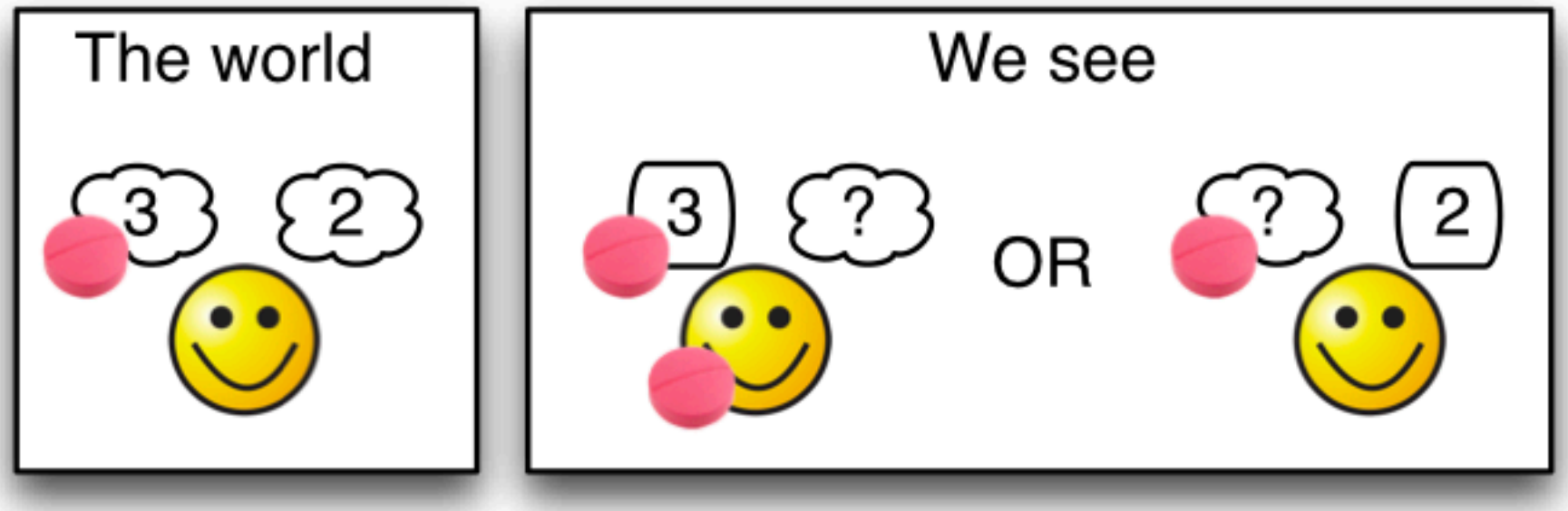
The experimenter obtains a collection of units

Then she randomizes units into treatment and control

This makes, for *typical* randomizations, the treatment group and the control group more or less the same.

Now, if we see any differences in outcome, we can ascribe this to the treatment itself.

The Neyman-Rubin Potential Outcomes Framework



Assume treatment assignment for any unit has no impact on any other unit (SUTVA).

Each unit then has two outcomes:

$y_i(1)$: what happens when you treat it

$y_i(0)$: what happens when you do not

The *treatment effect* for unit i is then $\tau_i = y_i(1) - y_i(0)$

We observe *either* $y_i(1)$ or $y_i(0)$ depending on whether we treat unit i or not.[6]

Assignment Mechanisms: The source of randomness

The Potential Outcomes are fixed:
The randomness comes from how treatment is assigned.

For each unit we have:

T_i, the treatment assignment. 1 indicates treatment, 0 not.

Y_i, the *observed* outcome.

$$Y_i = T_i y_i(1) + (1 - T_i) y_i(0)$$

How treatment is assigned is the *Assignment Mechanism.*

Comments

For the whole course we have thought of *randomness* as perturbances to predictions or sampling from some larger population (the random school effects, random student effects).

The potential outcomes view is an alternate view. Randomness is due to assignment. The units can be considered fixed.

We can informally merge the views. The **key idea is that any individual has two outcomes, only one of which we see.**

With multilevel data, two forms of randomized experiment

Consider you have J=60 schools (with m students nested in each school) and a novel attitude-adjusting treatment. You have two primary options:

Cluster-Randomized Trial

- ★ Randomize the schools into treatment and control.
- ★ This is a single experiment on schools!

Multi-Site Trial

- ★ You randomize some proportion of the students in each school into treatment.
- ★ This is 60 mini-experiments!

Notation Table

Y: Average treatment impact

T_i, T_{ij}: Treatment indicator. If double-subscript we know it is for individual i in unit j. Single subscript means we have a single experiment (no multilevel)

β: Average treatment impact

τ: Variation in average treatment impacts across sites

Cluster-Randomized Trials

For further discussion see

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. Psychological Methods.

You must take clustering into account

"Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception"

-Cornfield, 1978, p. 101

(This is due to the same reasons we have been talking about all term.)

What are we estimating?

Each site has an average treatment effect

$$B_j = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij}(1) - Y_{ij}(0)$$

Average treatment effect of *sites*:

$$\beta_{sites} = \frac{1}{J} \sum_{j=1}^{J} B_j$$

Average treatment effect of *people*:

$$\beta_{peeps} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} Y_{ij}(1) - Y_{ij}(0) = \frac{1}{N} \sum_{j=1}^{J} N_j B_j$$

Standard (Old fashioned) Analysis: ANOVA

Standard analysis of cluster-randomized experiments is mixed, two-factor nested ANOVA

ANOVA is a weird and technical world.

You have to worry about balance and equal group sizes and lots of stuff.

I tend to avoid it. Use regression methods instead (such as MLM).

A simple model

lmer call?

$$Y_{ij} = \beta_j + r_{ij}$$

$$\beta_j = \gamma_0 + \gamma_1 T_j + u_j$$

$$u_j \sim N(0, \sigma_u^2)$$

Treatment, T_j, is a level-2 variable.

We cannot separate variation in treatment impact and variation in clusters.

We might reasonably worry about the second-level variance term: we could let it vary by treatment status with extended model with different variances for the two groups.

Cluster-randomization: how many units?

If the clusters are radically different, and the units within cluster very much the same, then we really have only J units in our experiment.

If the clusters are the same, and individuals within cluster basically independent, then we have closer to nJ units.

Moral: Cluster-randomization can be a *much smaller experiment than you think*.

(But design can save you in some cases.)

Rough estimate of uncertainty

For equal-sized clusters, random intercept model, etc., we have

$$\hat{\gamma}_1 = \bar{Y}_T - \bar{Y}_C$$

and

$$Var(\hat{\gamma}_1) = \frac{1}{J}\left(4\sigma_u^2 + \frac{\sigma^2}{n}\right)$$

with σ_u being the standard deviation of random intercepts and σ being the residual standard deviation.

J = number of clusters
n = number of students in each cluster
n_J = total number of students

Cost of data collection

$$Cost = J(C_1 n + C_2)$$

Incremental cost per student in cluster

Fixed cost per cluster

Now we have variance and cost.

For fixed cost we can optimize what *n* and *J* should be.

We get

$$n_{opt} = \frac{\sigma}{\sigma_u}\sqrt{\frac{C_2}{C_1}}$$

So plug in and get a cost per cluster and go from there.

Optimal sample sizes and corresponding variances as function of ICC and cost

Intracluster correlation (ρ)	Cluster/ person cost ratio (C_2)	n(optimal)	J	Var ($\hat{\gamma}_1$)
.01	2	14	31	.0103*
.01	10	31	12	.0138*
.01	50	70	4	.0232
.05	2	6	61	.0133*
.05	10	14	21	.0226
.05	50	31	6	.0522
.10	2	4	80	.0156*
.10	10	9	26	.0304
.10	50	21	7	.0811
.20	2	3	104	.0186*
.20	10	6	31	.0426
.20	50	14	8	.1317
.50	2	1	146	.0233
.50	10	3	38	.0693
.50	50	7	9	.2606

Our SE for our estimated ATE

* = reasonable precision.

Comments

You can do power and optimal cost/design calculations for specific scenarios but they might not generalize.

Could instead *simulate* to model more complex and authentic scenarios.

That being said, general findings of simple models will generally carry over unless one has many predictive covariates or very atypical balances of units not taken into account by simple scenario.

Multi-Site Trials

Reading:

Bloom, H.S., Raudenbush, S.W., Weiss, M., & Porter, K. Using Multi-site Experiments to Study Cross-site Variation in Treatment Effects: A Hybrid Approach with Fixed Intercepts and a Random Treatment Coefficient

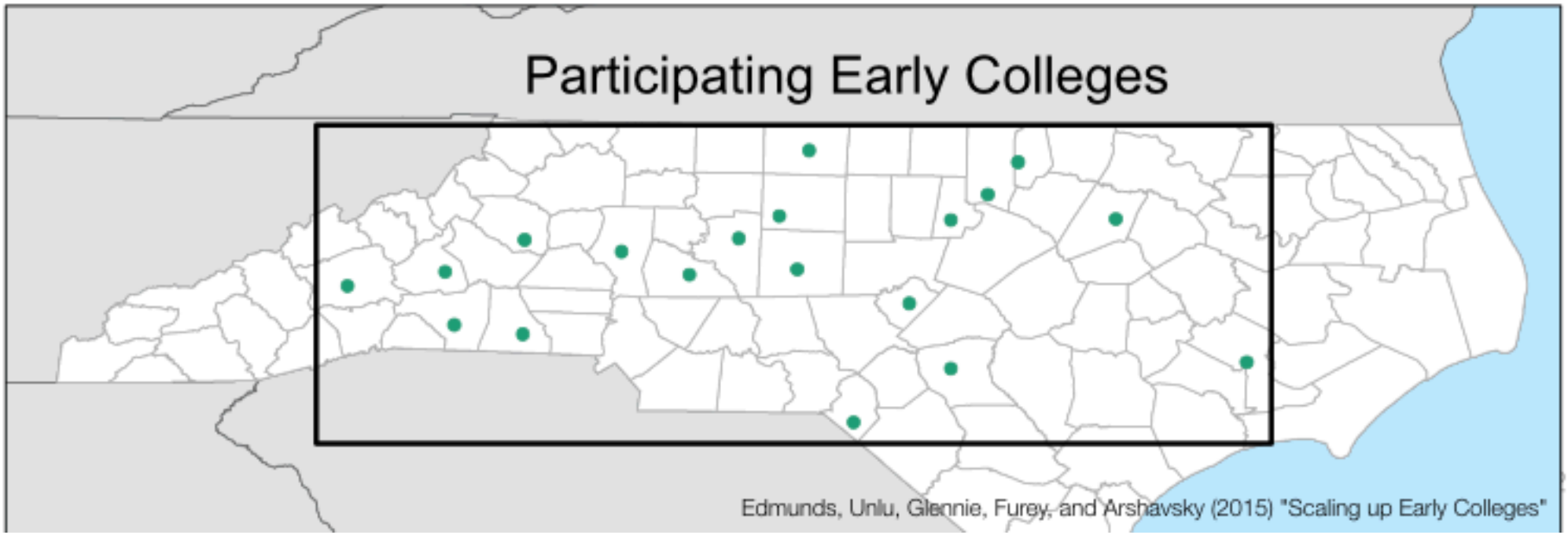
Some slides adapted from Raudenbush slides connected to his HLM program

Running Example: ECHS Study

ECHS are autonomous small schools managed by local school districts in partnership with two/four-year college

Our study is an effective RCT of ECHS in North Carolina

- **44 lotteries**: **19 schools** × (up to) **6 years**
- First cohort started 2005
- **Goal:** Improve HS graduation and college going rates



Many small experiments!

Each site is a mini-experiment.

Just like the many worlds framework, we might borrow information across worlds to get better understanding of individual worlds.

Our Research Questions

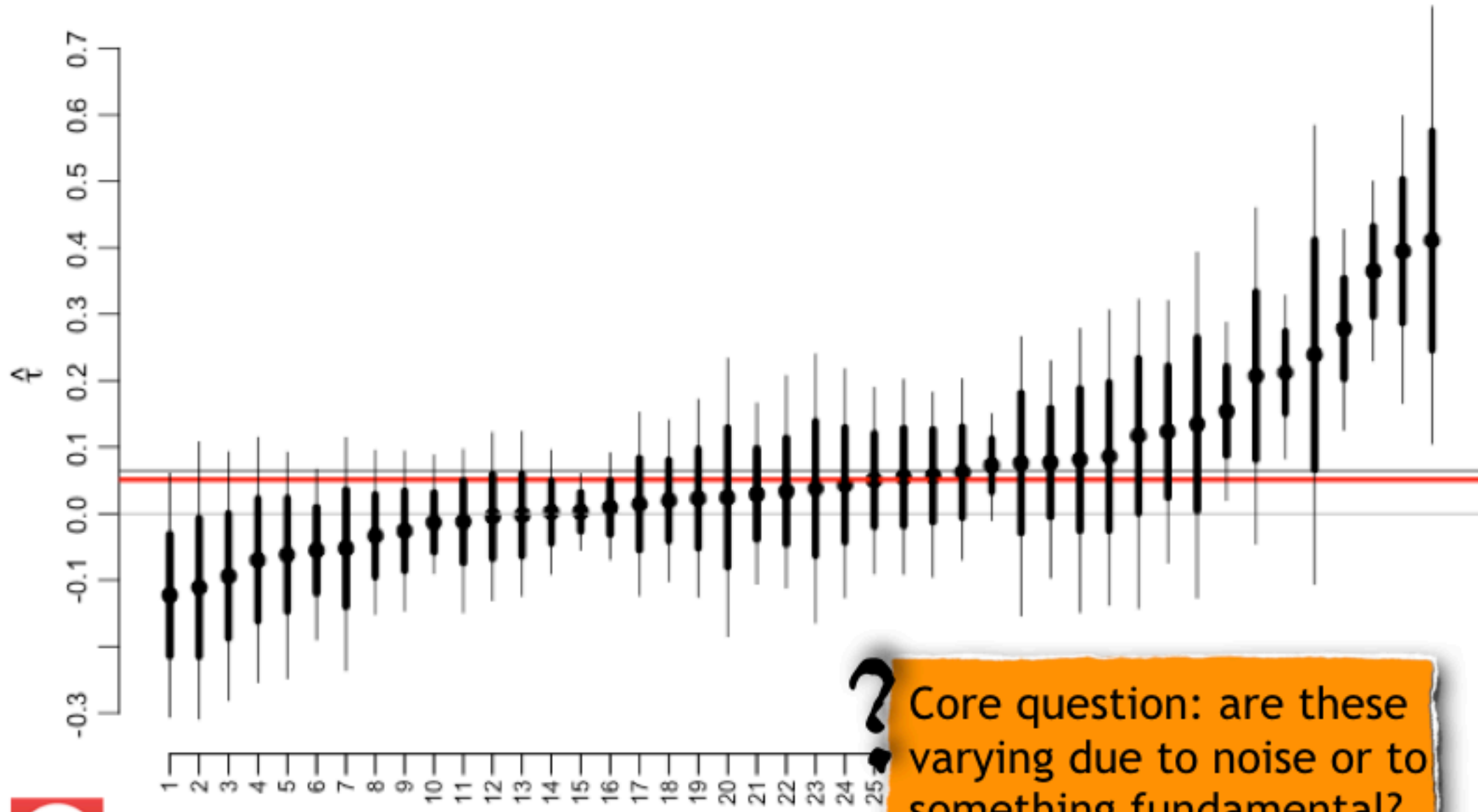
- ★ RQ1: What is the average impact across sites?
- ★ RQ2: What is the variation in impacts across sites?
- ★ RQ3: What is the impact for a specific site? (Harder.)

Potential Opportunities and Features of Multi-Site Trials

- ★ The different mini-experiments are different! This potentially helps us.
- ★ I.e., can we use that variation to assess generalizability of impact?
- ★ Heterogeneity makes seeing things hard: variation in multisite trials does cause trouble.
- ★ Often multiple sites, but students *still* nested in classrooms. What is "sample size" then?
- ★ Compare to cluster-randomization: here we get treatment info for each site. Cluster-randomization gives *no* treatment information for *any* site.

Raw site x cohort-level estimates for ECHS

(for 9th grade on track as outcome, ATE = 5.9pp)



Is there evidence of cross cohort variation?

Core question: are these varying due to noise or to something fundamental?

We might imagine using the usual model

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + r_{ij} \qquad r_{ij} \sim (0, \sigma^2)$$

Level 2:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix} \right]$$

The "random slopes" are the treatment effects (T_{ij} coded as 0, 1)

With randomized experiments, covariates help precision, but can be skipped since the randomization ensures balance (on average)

What are we estimating?

Each site has its own average treatment effect: β_{1j}

Average treatment effect across *sites*:

$$\gamma_{sites} = \frac{1}{J}\sum_{j=1}^{J}\beta_{1j}$$

Average treatment effect of *people*:

$$\gamma_{peeps} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\beta_{1j} = \frac{1}{J}\sum_{j=1}^{J}N_j\beta_{1j}$$

Questions still remain...

Are the sites representative of a larger whole?

- ★ Sample average treatment effect vs. population average treatment effect.
- ★ MLM models assume i.i.d. draws from superpopulations.
- ★ Randomized experiments are on a specific (usually nonrandom) sample.

Are the N_j the number of sampled people?
Should we weight by the *total size* of the site?

Do we Satisfy the Six Assumptions?

1. $r_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ ← Student residuals (beyond the school-level effects)

2. $Cov(r_{ij}, u_{qj}) = 0$ ← Student residuals and school random effects

3. $u_j = (u_{0j}, \ldots, u_{Qj})' \overset{i.i.d.}{\sim} N(0, \Sigma)$

4. $Cov(W_{sj}, u_{qj}) = 0$ ← Level 2 covariates

5. $Cov(X_{qij}, r_{ij}) = 0$ ← Level 1 covariates. (For us, the randomized treatment)

6. $Cov(X_{qij}, u_{q'j}) = 0 \qquad Cov(W_{sj}, r_{ij}) = 0$

Imbalance

Some forms of imbalance:

- ★ Unequal N_j per site
  If big sites have larger effects, we will estimate larger average effects
- ★ Unequal proportion of units treated per site
  If sites with higher proportion of treated units have systematically different average effects, we will again violate our modeling assumptions

Problem with imbalance:

- ★ Imbalance can cause our final estimate to be a weighted average of site level effects where the weights are not what we want.

ECHS is a canonical scenario for modern times

The Assignment Mechanism:

- ★ ECHS schools offer a fixed number of slots.
- ★ People apply to the slots.
- ★ A *lottery* allocates who gets in.
- ★ Treatment=Get access to charter school. Control is no dice.

Food for thought:

- ★ Higher quality ECHS schools could have more applicants (and thus a lower proportion treated).
- ★ Larger schools might have more slots, people applying, etc.

Multi-site trials with fixed effects models

The *fixed effect* (dummy variables for site) model gives:

Tx impact

Fixed effect

$$Y_{ij} = \gamma_1 T_{ij} + \alpha_j + r_{ij}$$

This gives the following estimate:

$$\hat{\gamma}_1 = \frac{1}{\sum_{j=1}^{J} N_j \bar{T}_j (1-\bar{T}_j)} \sum_{j=1}^{J} N_j \bar{T}_j (1-\bar{T}_j) \hat{\beta}_{1j}$$

This is just a weighted average of effects.

with $\hat{\beta}_{1j} = \bar{Y}_{Tj} - \bar{Y}_{Cj}$

(these are the individual unpooled site estimates of treatment effect for the sites).

A modern multilevel modeling approach: Fixed Intercept, Random Coefficient (FIRC)

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \text{ fixed effect}$$

Fixed effect means we do not shrink our intercepts. Each intercept will be the mean of the control units in that site.

$$\beta_{1j} = \gamma_{00} + u_j$$

$$u_j \sim N(0, \tau)$$

The gamma$_{00}$ is the average treatment impact across our sites.

$$\epsilon_{ij} \sim \begin{cases} N(0, \sigma_1^2) & \text{if } Z_{ij} = 1, \\ N(0, \sigma_0^2) & \text{if } Z_{ij} = 0 \end{cases}$$

Residuals are different for treatment and control groups

Issues: We have wiped out all information on the baseline (control) outcome, and how it connects to treatment impact.

See Bloom reading

Models can be targeted to specific tasks

The prior model is designed to answer some questions. To do so, we lose the ability to answer others.

Wins:

- ★ This model allows us to account for treatment effect heterogeneity
- ★ We get a good estimate of that heterogeneity (the τ term)

Losses:

- ★ We cannot estimate other parameters such as baseline site-level variation, average site outcomes, or correlation of site level outcome and treatment impact.

An alternate (preferred) method for detecting variation: Q-Statistics

$$Q = \sum_{j=1}^{J} \frac{(\hat{\beta}_j - \bar{\beta})^2}{\widehat{SE}_j^2}$$

The $\hat{\beta}_j$ and $\widehat{SE}_j$ come from, e.g., separate OLS or fixed effect model with interactions.

We then compare this to χ^2_{J-1}

$$\bar{\beta} = \frac{1}{M} \sum_{j=1}^{J} \frac{1}{\widehat{SE}_j^2} \hat{\beta}_j \text{ with}$$

$$M = \sum_{j=1}^{J} \frac{1}{\widehat{SE}_j^2}$$

Advantages:

- ★ No multi-level model needed
- ★ Simple to implement

These come from meta-analysis via Weiss et al. (2016)

Systematic cross site variation

With multi-level models, we can naturally include site-level covariates:

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j} T_{ij} + \epsilon_{ij} \\
\beta_{0j} &= \text{ fixed effect} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} W_j + u_j \\
u_j &\sim N(0, \tau)
\end{aligned}
$$

This is a site by covariate interaction. (Systematic variation.)

Now we could test

$$H_0^{(sys)} : \gamma_{11} = 0$$

or, even better

$$H_0^{(comb)} : \tau = 0 \text{ and } \gamma_{11} = 0$$

Fitting (modified) FIRC to ECHS

```
> lmer(formula = all_p_9 ~ 1 + treatment +
       (0 + treatment | echs_cohort),
        data = sdat)

> display( M1 )
lmer(formula = all_p_9 ~ 0 + treatment + echs_cohort + (0 + treatment |
    echs_cohort), data = sdat, REML = FALSE)
                coef.est coef.se
treatment       0.04     0.02
echs_cohortC1   0.89     0.03
echs_cohortC10  0.93     0.04
echs_cohortC11  0.82     0.03
...
echs_cohortC8   0.98     0.04
echs_cohortC9   0.98     0.04

Error terms:
 Groups       Name       Std.Dev.
 echs_cohort  treatment  0.09
 Residual                0.26
---
number of obs: 3820, groups: echs_cohort, 44
AIC = 811.7, DIC = 717.7
deviance = 717.7
```

Average treatment effect
4 percentage points on "on track" status

But different sites varied in their effectiveness by 9pp up or down

Variation in average impacts across sites:
0.04 +/- 2 * 0.09 = -0.14 to 0.22 percentage points
(middle 95% prediction interval)

Testing for Variation

```
> lmer(formula = all_p_9 ~ 1 + treatment +
        (0 + treatment | echs_cohort),
         data = sdat)

> display( M1 )
             coef.est coef.se
(Intercept)  0.88      0.01
treatment    0.07      0.01

Error terms:
 Groups       Name        Std.Dev.
 echs_cohort  treatment   0.07
 Residual                 0.27
---
number of obs: 3820, groups: echs_cohort, 44
AIC = 1023.2, DIC = 984.8
deviance = 1000.0
```

Average treatment effect
7 percentage points on "on track" status

But different sites varied in their effectiveness by 7pp up or down

Cross site impact variation:
0.07 +/- 2 * 0.07 = -0.07 to 0.21 percentage points
(middle 95% prediction interval)

Alternate: Random intercept, random slope model

```
> M2 = lmer( all_p_9 ~ 1 + treatment + (1+treatment|
echs_cohort), data=sdat, REML=FALSE )
> display( M2 )

coef.est coef.se
(Intercept) 0.90     0.02
treatment   0.04     0.02

Error terms:
  Groups       Name        Std.Dev. Corr
echs_cohort (Intercept) 0.10
treatment   0.11     -0.80
Residual                0.26
---
  number of obs: 3820, groups: echs_cohort, 44
AIC = 811.9, DIC = 799.9
deviance = 799.9
```

lr test for treatment variation

```
> M2B = lmer( all_p_9 ~ 1 + treatment + (1|echs_cohort),
                          data=sdat, REML=FALSE )
> anova( M2, M2B )
Data: sdat
Models:
  M2B: all_p_9 ~ 1 + treatment + (1 | echs_cohort)
M2: all_p_9 ~ 1 + treatment + (1 + treatment | echs_cohort)
Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
M2B  4 899.43 924.42 -445.71   891.43
M2   6 811.90 849.39 -399.95   799.90 91.524      2  < 2.2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Liklihood ratio test highly significant:
different cohorts had different average impacts.

Q statistic test for variation

```
# Custom function I wrote as part of my research.
> estimate.Q.confint(sdat$all_p_9, sdat$treatment,
sdat$echs_cohort )
$reject
[1] TRUE

$p.value
[1] 5.314549e-20
```

P-value is tiny tiny tiny!

```
$Q
[1] 187.7686

$CI_low
[1] 0.09

$CI_high
[1] 0.13
```

This is from test inversion: this confidence interval is all values of tau that we could not reject.

We see significant variation, again. This bolsters our belief from the multilevel modeling estimates.

Some comments and admissions

This analysis is approximate, given current course knowledge.

To do FIRC right, we need different variances on the treatment and control unit residuals

Need later material, when we talk about error structure, to do this.

Code available upon request.

With FIRC, lrtest compares a linear model with multilevel model (no random effects if no treatment variation)

- ★ This introduces some coding trickiness; more code available upon request.

Recap

Recap

http://cs179.org/lec43

Randomized experiments with multilevel modeling have some wrinkles to track.

Using MLM is generally safe, but you might not be estimating the exact average treatment effect you think.

If clusters are same sizes, and proportion treated is constant, these issues generally go away.

Multisite trials allow for assessing *variation* in treatment, which is an interesting policy-related quantity.