# Lecture 1

# Lecture 1 Outline

- Introduction
- Housekeeping for the course (for the record)
- Statistical background for MLMs
  - Main ideas: multi-levels and decomposition as am approach to analysis
  - Notation
  - Longitudinal data as special case of MLMs
  - What creates the clustering in practice?
  - A simple example
  - Key MLM components
  - Accounting for within-cluster associations

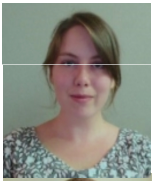
# Introduction

- This course describes statistical methods for the analysis of “multi-level” data
- Develops skills to implement and interpret random effects, variance component models that reflect the multi-level structure for both predictor and outcome variables.
- See course schedule for specific topics covered

- The emphasis for student evaluations is on analysis, interpretation, and communication of results

# Teaching Team

Instructor: Scott L Zeger, John C. Malone Professor - sz@jhu.edu; Office hours: Friday, 12-12:50pm ZegerZoomRoom (ZZR- https://jhjhm.zoom.us/j/9624766044 ) or by appointment



Teaching assistants: Office hours Tuesday, 12-12:50 at TBN

- Lacey Etzkorn, 4th year PhD Biostatistics, Lead TA - letzkor1@jhmi.edu

- Emily Scott MS, 2nd year PhD Biostatistics, TA - escott29@jhu.edu

## Commitments Required to Participate in this Course
## Teachers □ Learners

1. The teachers commit to support student learning with our intention, time and materials. We very much want you to learn MLM.

2. The learners commit to do their best to master the course content using the course meetings and materials.

3. The course materials (handouts, recordings, quizzes, exams,…) have been created by the instructors to support your learning. They are the property of JHU and its faculty. We use them multiple times. Learners have the right to use the materials during their participation in the course **but agree not to share them with others**.

# Course Information

- Basic class structure:
  - Zoom Lectures/discussions with Instructor M/W 10:30 to 11:50
  - Zoom Lab working sessions led by TAs, Wednesday, 9 - 10:20
    ~~TBD~~ Zoom Link
- Utilize CoursePlus for all course materials
  - Links for all course materials under “Class Materials and Resources”, “Class Sessions”
  - Zoom recordings of each lecture will be posted; pre-recordings should be used prior to lectures when asked.
  - No recordings of the Wed lab sessions
- Discussion Forum
  - Students can register for participation in this forum where you may post questions/answers throughout the course
  - We prefer postings to this bulletin board as many people have the same questions!

# Student Evaluations

- 5 structured lab sessions (Lab0 through Lab4)
  - Work in groups to complete a guided exercise
  - Multiple choice quiz will follow labs 1 through 4
- 2 homework assignments
  - Data analysis driven, you write a short report summarizing your findings AND, answer quiz questions after each lab
  - Students are encouraged to work together on the data analysis
  - May submit the assignment in teams of up to 4 students
  - One lab session will be dedicated to addressing questions on each homework assignment.
- Final exam
  - Take home data analysis
  - Students may work together on the data analysis

# Student Evaluations

- 5 structured lab sessions (Lab0 through Lab4)
  - Work in groups to complete a guided exercise
  - Multiple choice quiz will following labs 1 through 4
- 2 homework assignments
  - Same structure as LDA: data analysis driven, you write a short report summarizing your findings AND, three short answer questions
  - Students are encouraged to work together on the data analysis
  - May submit the assignment in teams of up to 4 students
  - One lab session will be dedicated to addressing questions on each homework assignment.
- Final exam (3rd homework assignment)
  - Take home data analysis
  - Students may work together on the data analysis
  - Each student must write up their own report

# Course Information

**Important Dates**

| | | |
|---|---|---|
| Lab 1 quiz | April 2 | - 10% |
| Lab 2 quiz | April 9 | - 10% |
| Homework 1 | April 13 | - 20% |
| Lab 3 quiz | April 23 | - 10% |
| Homework 2 | May 4 | - 20% |
| Lab 4 quiz | May 7 | - 10% |
| Final Exam | May 18 | - 20% |

(Homework 3)

fixed

# Course Information

- Texts:
  - Recommended:
  - Multilevel and Longitudinal Modeling using Stata (College Station, TX: Stata Press.). Sophia Rabe-Hesketh and Anders Skrondal (Second Edition).
  - 3rd Edition
- Supporting Stata and R
- SAS code available at: http://www.ats.ucla.edu/stat/sas/topics/MLM.htm

# Statistical Background on MLMs

- Main ideas
  - multi-level causal models
  - Decomposition as an analytic strategy
- Notation
- Longitudinal data as special case of MLMs
- What creates the clustering in practice?
- A simple example
- Key MLM components
- Accounting for within-cluster associations

# A rose is a rose?

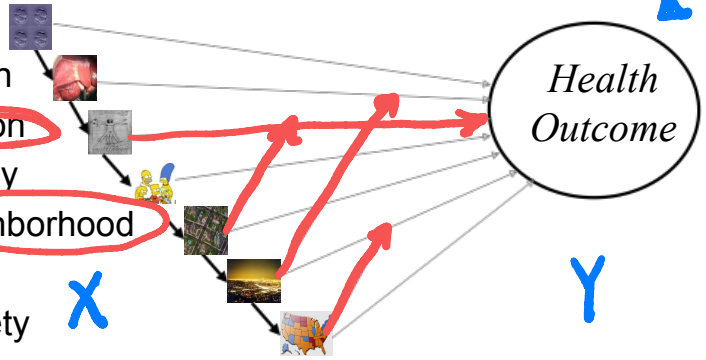
- *Multi-level* model
- *Random effects* model
- *Mixed* model
- *Random coefficient* model
- *Hierarchical* model
- *Growth* models
- *Meta-analysis* (special case)

Multiple (social sciences)

Many names for similar models, analyses, and goals.

# Multi-level Models – Main Idea

- Biological, psychological and social processes that influence health occur at many **levels**:
  - Cell
  - Organ
  - Person
  - Family
  - Neighborhood
  - City
  - Society




- An analysis of health outcomes should consider:
  - Each of these levels
  - And potential interactions between levels

# COVID-19 example

RAAS

Molecule : Spike protein ; ACE-II

Cell: COVID-19 , any blood

Organ: LUNG, HEART

Person: COVID-19; bmi, mask-wearing behavior

Family: COVID-19 family, mask-wearing,

Neighborhood:

City: Belt + 25% COVID-cases

State:

Nation:

Species

:

Causes

Covid-19 death

Y

# Example: Alcohol Abuse

| *Level:* | |
|---|---|
| 1. | Cell: Neurochemistry |
| 2. | Organ: Ability to metabolize ethanol |
| 3 | Person: Genetic susceptibility to addiction |
| 4 | Family: Household environment |
| 5 | Neighborhood: Availability of bars |
| 6 | Society: Regulations; organizations; social norms |

# Example: Alcohol Abuse; Interactions between Levels

| *Level:* | |
|---|---|
| 5 | Availability of bars *and* |
| 6 | State laws about drunk driving |
| 3 | Genetic predisposition to addiction *and* |
| 2 | Person's ability to metabolize ethanol |
| 3 | Genetic predisposition to addiction *and* |
| 4 | Household environment |
| 6 | State regulations about intoxication *and* |
| 3 | Job requirements |

# Notation:

highest level

lowest level

Person: *sijk*

Outcome: *Ysijk*

Predictors: *Xsijk*

$X_{sijk} \rightarrow Y_{sijk}$

Population

**State:** $s=1,\ldots,S=50$

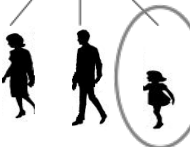
**Neighborhood:**

$i=1,\ldots I_s$

**Family:**

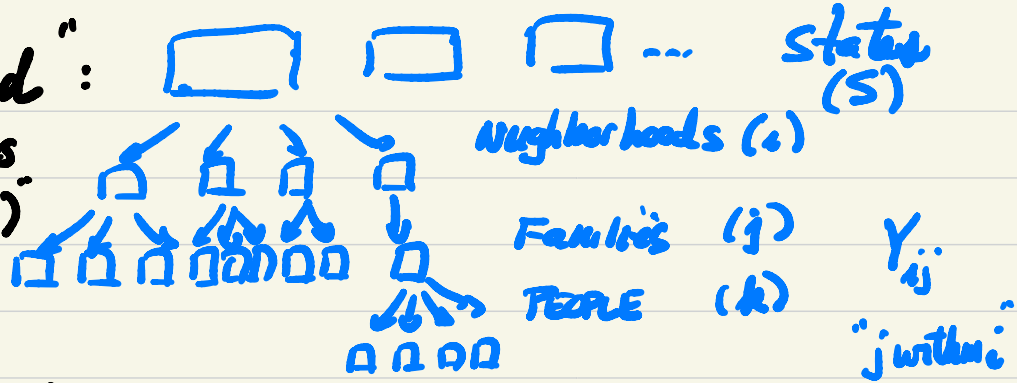
$j=1,\ldots J_{si}$

**Person:** $k=1,\ldots,K_{sij}$

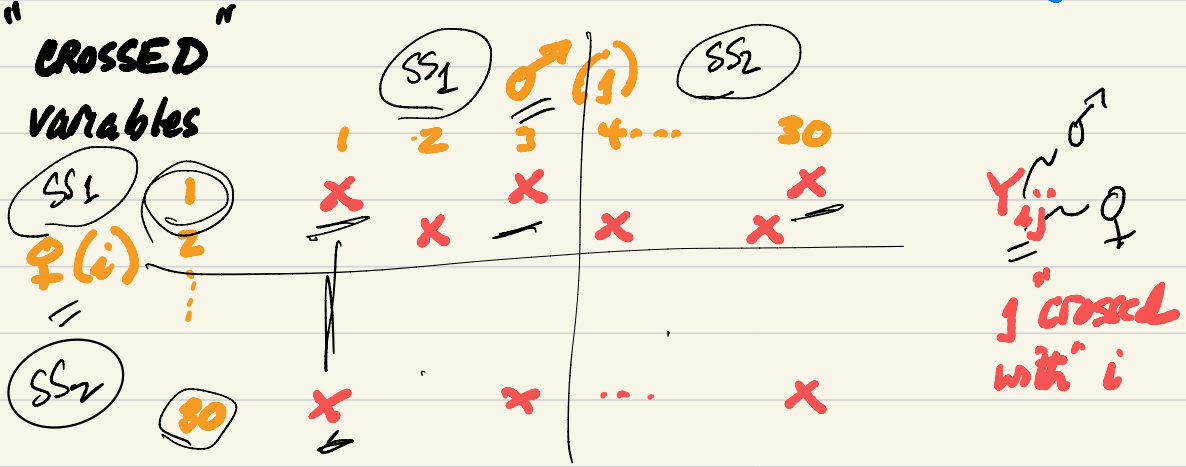


$(y1223, x1223)$

"Nested" variables "(TREE)":




"CROSSED" variables




We will use similar models for both cases, even though they represent different computational problems

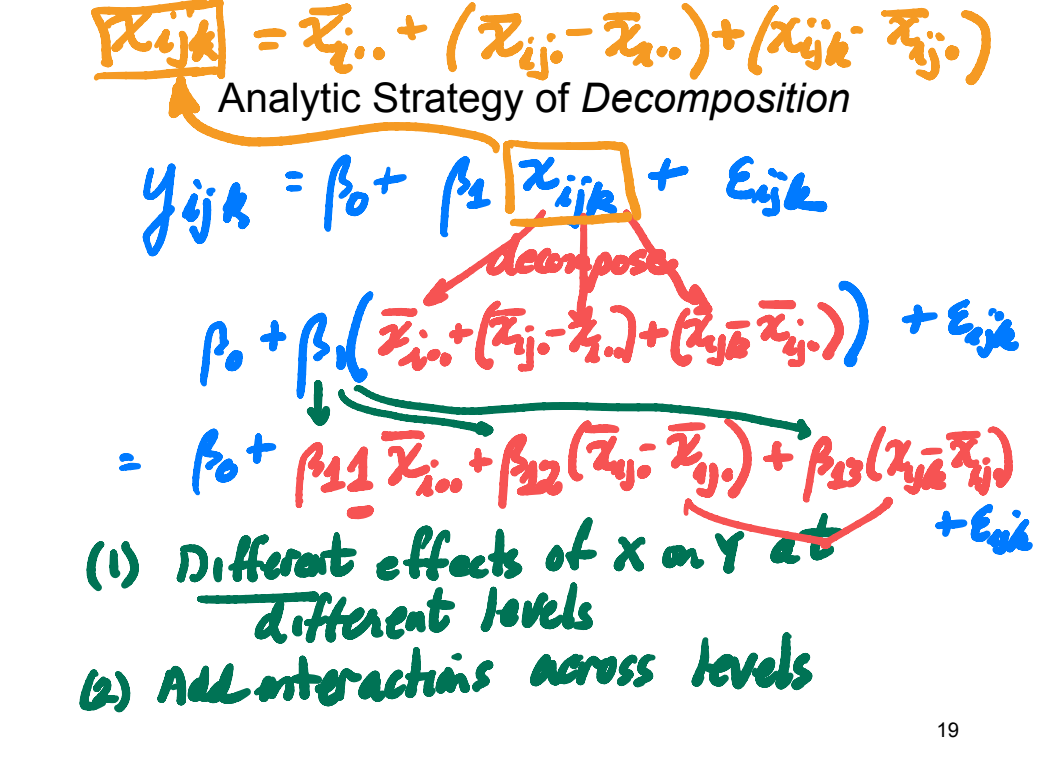
# Notation (cont.)

- $(y_{sijk}, x_{sijk})$ are (response, predictors) for
  - person $k = 1, ..., K_{sij}$ in
  - family $j = 1, ..., J_{si}$ in
  - neighborhood $i = 1, ..., I_s$ in
  - state $s = 1, ..., S$
- $\mu_{sijk} = \mathsf{E}(y_{sijk}|x_{sijk})$

Goal is to construct a regression model for the mean of ysijk

# Describing the Hierarchy

- Referred to as “levels”
  - Lowest level
    - i.e. the most nested unit
    - Level 1
  - Then count “levels” up from there!


$x_{ijk} = \bar{x}_{i..} + (\bar{x}_{ij.} - \bar{x}_{i..}) + (x_{ijk} - \bar{x}_{ij.})$
Analytic Strategy of *Decomposition*
$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \varepsilon_{ijk}$
decompose
$\beta_0 + \beta_1 (\bar{x}_{i..} + (\bar{x}_{ij.} - \bar{x}_{i..}) + (x_{ijk} - \bar{x}_{ij.})) + \varepsilon_{ijk}$
$= \beta_0 + \beta_{11} \bar{x}_{i..} + \beta_{12} (\bar{x}_{ij.} - \bar{x}_{ij.}) + \beta_{13} (x_{ijk} - \bar{x}_{ij.}) + \varepsilon_{ijk}$
(1) Different effects of X on Y at different levels
(2) Add interactions across levels

# Air pollution examples of decomposition

$X_{\text{location}(i),\,\text{time}(j)} \rightarrow \lambda_{ij} \rightarrow Y_{ij}$

eg. $PM_{2.5}$ — mortality Rate — observed deaths in $N_{ij}$ people

$$\mu_{ij} = N_{ij}\lambda_{ij} = N_{ij}e^{x_{ij}\beta}$$

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$X_{ij} = \underbrace{\bar{x}_{i\cdot}}_{x_{ij1}} + \underbrace{\bar{x}_{\cdot j}}_{x_{ij2}} + \underbrace{(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j})}_{x_{ij3}}$$

$x_{ij1}$: space variation ($i$)

$x_{ij2}$: time variation ($j$)

$x_{ij3}$: space × time variation ($i$ and $j$)

By assumption

$$\mu_{ij} = e^{x_{ij1}\beta + x_{ij2}\beta + x_{ij3}\beta}$$

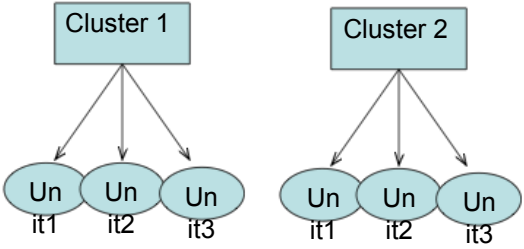
Test this by fitting

$$\mu_{ij} = e^{x_{ij1}\beta_1 + x_{ij2}\beta_2 + x_{ij3}\beta_3}$$

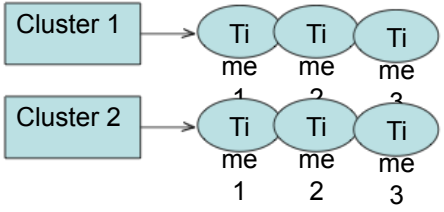
Ask whether $\beta_1 = \beta_2 = \beta_3 = \beta$ ?

If not, perhaps some estimates are less confounded by unmeasured confounders than others ??

# Longitudinal data
# as special case of multi-level data




GENERAL: Multi-level model




Longitudinal Data:
- Clusters are subjects
- Units within subject represent time (ordered)

# What creates the clustering in practice?

- Naturally occurring hierarchy in the target population
  - Development toxicity studies: mother is given treatment/intervention, offspring are measured for outcomes
  - Clinical outcome studies where patients are naturally nested within hospital/clinic
  - Family studies: Family SES and person level income and health outcomes

# What creates the clustering in practice?

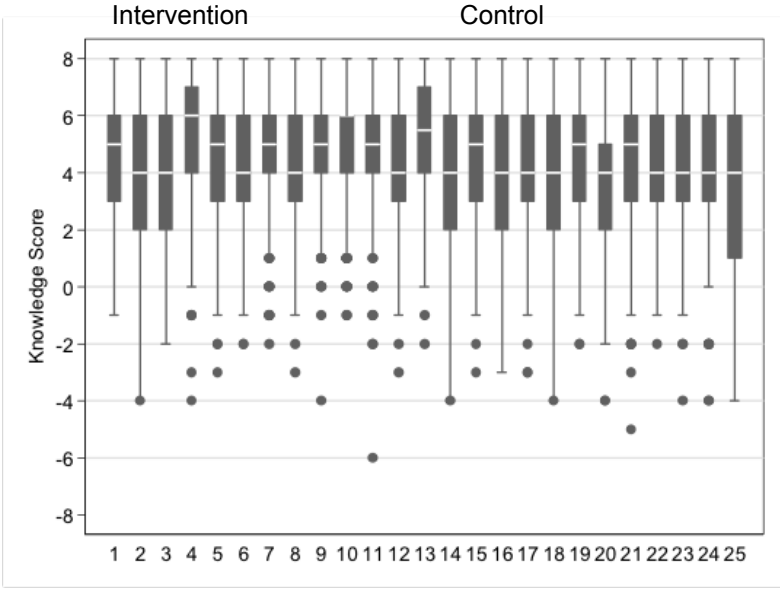
- Study design
  - Cluster randomized trial: intervention is randomized to schools, outcomes are measured at the student level within each school
  - Multi-stage sampling design:
    - NHANES
      - First stage: Primary Sampling Unit (PSU) based on US counties
      - Second stage: Census blocks sampled within PSU
      - Third stage: Households are sampled within Census blocks
      - Fourth stage: Eligible persons are selected within households
- Or Both!
  - Multi-center longitudinal clinical trials:
    - Data are clustered by center (natural structure) and then subjects sequentially followed over time (study design)

# A “simple” example

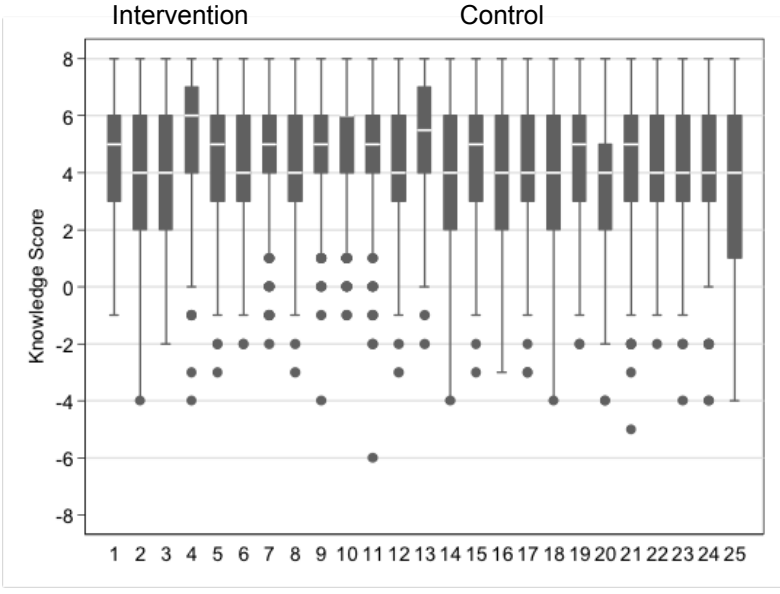
- Cluster-randomized trial of sex education
  - Wight et al (2002)
  - 25 secondary schools randomly assigned to control group or SHARE
    - Sexual Health and Relationships: Safe Happy and Responsible
  - Teachers in intervention schools
    - 5 days of training
    - delivered 10 sessions in 3rd and 4th year of secondary school (ages 13-15) to two successive cohorts.
  - 8,430 total students recruited for the 25 schools
  - 5,854 students successfully followed for 2 years
  - Outcome: knowledge score at follow-up

5854/25 ≈ 234/school

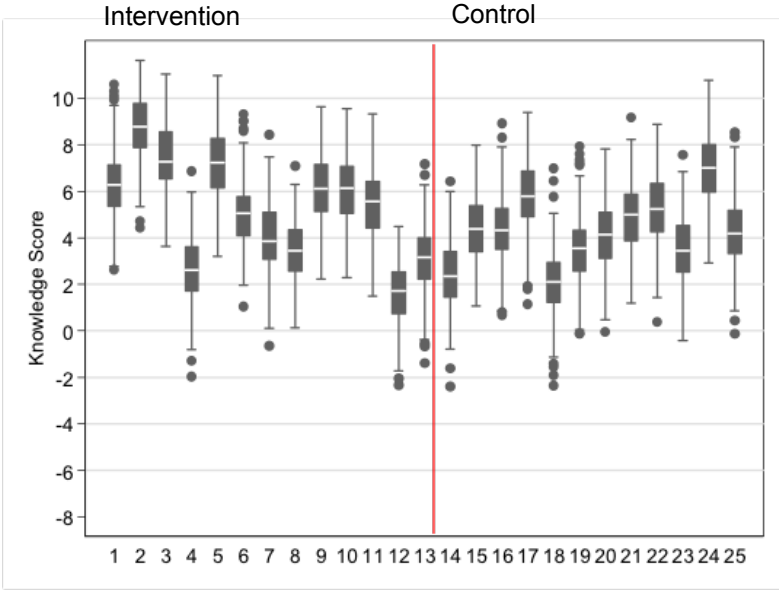
# Data


Intervention
Control
Knowledge Score
8
6
4
2
0
-2
-4
-6
-8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

# Observed “Non-Clustered” Data



# Simulated Clustered Data



# Within-Cluster Correlation

- Correlation of two observations from same cluster =

$$\frac{\text{Tot Var - Var Within}}{\text{Tot Var}}$$

- Non-Clustered = (5.5-5.4) / 5.5 = 0.02
- Clustered = (5.5-2.0) / 5.5 = 0.64

# “Quiz”: Most Important Assumptions of Regression Analysis?

A. Data follow normal distribution

B. All the key covariates are included in the model

C. Xs are fixed and known

D. Responses are independent

# “Quiz”: Most Important Assumptions of Regression Analysis?

A. Data follow normal distribution

B. All the key covariates are included in the model

C. Xs are fixed and known

D. Responses are independent

# Regression with Correlated Data

Must take account of correlation to:

- Obtain valid inferences
  - standard errors
  - confidence intervals
- Make efficient inferences

# Impact of Ignoring the Clustering!

| Model | Observed Data (ICC = 0.02) | | | Simulated Data (ICC = 0.64) | | |
|---|---|---|---|---|---|---|
| | Trt (se) | Z | p | Trt (se) | Z | p |
| No clustering | 0.62 (0.063) | 9.74 | < 0.001 | 0.60 (0.059) | 10.24 | < 0.001 |
| Clustered | 0.54 (0.16) | 3.30 | 0.001 | 0.90 (0.73) | 1.23 | 0.217 |

$$\frac{SE_{wrong}}{SE_{right}} = \frac{3.3}{9.7} \approx \frac{1}{3} \approx \sqrt{\frac{1}{9}}$$

# Why do we see a increase in the standard error for the treatment effect when accounting for the clustering?

- Consider estimation of the mean knowledge score within a school:
  - Sample mean: $\overline{Y}_i = \sum_{j=1}^{n_i} Y_{ij} / n_i$
  - Variance of sample mean:

$$Var(\overline{Y}_i) = \frac{1}{{n_i}^2} \mathrm{Var}\left(\sum_{j=1}^{n_i} Y_{ij}\right)$$

$$= \frac{1}{{n_i}^2} \left[ \sum_{j=1}^{n_i} Var(Y_{ij}) + 2 \times \frac{n_i(n_i-1)}{2} Cov(Y_{ij}, Y_{ik}) \right]$$

$$= \frac{1}{n_i} \left[\sigma^2 + (n_i - 1)\rho\right]$$

$\sigma^2] = \frac{\sigma^2}{n} [1 + (n_i - 1)\rho]$

$= \frac{\sigma^2}{n} [1 + 234\rho] = \frac{\sigma^2}{n} 5.6$

# Accounting for within cluster correlations

- Marginal model approach

  Model mean response and covariances among responses from same clusters

- Conditional model approach

  Model mean response given cluster-specific coefficients and distribution of the coef's among clusters

# Key Components of Multi-level Models

- Specification of predictor variables from multiple levels **(Fixed Effects)**
  - Variables to include

- Specification of correlation among responses from same clusters **(Random Effects)**

- Choices must be driven by scientific understanding, the research question and empirical evidence.

# Key Points

- "Multi-level" Models:
  - Have covariates from many levels and potentially interactions of covariates measured at different levels
  - Acknowledge correlation among observations from within a level (cluster)
- Random effect MLMs condition on unobserved "latent variables" to account for the correlation
- Assumptions about the latent variables determine the nature of the within cluster correlations