

## Introduction

Peter Hoff  
Duke STA 610

Multilevel data  
○○○

Subpopulation inferences  
○○○

Population inferences  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

Cross-level inferences  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

Multilevel data

Subpopulation inferences

Population inferences

Cross-level inferences

## Multilevel data

**Multilevel data:** Data for which there are

- multiple nested levels of sampling, and/or
- multiple nested sources of variability.

Such data are also often called *hierarchical data* or *clustered data*.

### Examples:

**Educational testing:** students nested within classes;

**Small area estimation:** households nested within counties;

**Agricultural experiments:** subplots nested within whole plots;

**Clinical trials:** measurements nested within patients, patients within hospitals.

## Terminology

*observational unit*: an object or condition for which data are measured.

*macro-level unit*: a unit within which other units are nested.

*micro-level unit*: a unit nested within another unit.

### Synonyms:

- macro-level unit, top-level unit, clusters, groups;
- micro-level unit, bottom-level unit, units.

If there are only two levels, we will say *units* are nested within *groups*.

**Notation:**  $y_{i,j}$  = measurement of *i*th unit in *j*th group.

### Populations:

- The population: all possible units from all possible groups;
- A subpopulation: all possible units from a single group group.

## Types of multilevel inference

**Subpopulation inferences:** Group-specific features are of primary interest.

- What is the mean within each group, based on a sample from each group?
- What is the treatment effect for each group?
- Do the groups differ? If so, how do they differ?

**Population inferences:** Across-group averages are of primary interest.

- What is the population mean, based on cluster sample?
- What is the population treatment effect?

**Cross-level inferences:** Both types of features are important.

- What is the average treatment effect, adjusting for group differences?

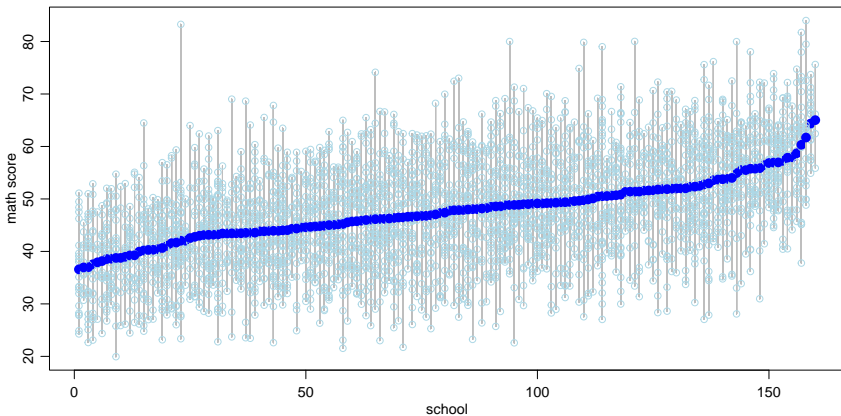
Multilevel data  
○○○

Subpopulation inferences  
●○○

Population inferences  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

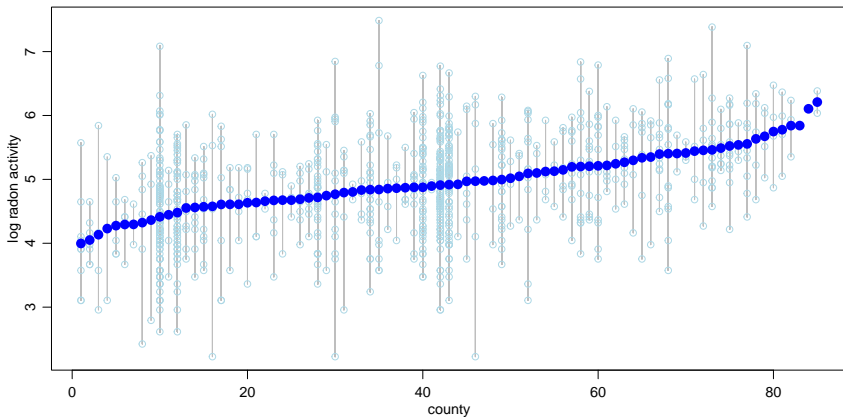
Cross-level inferences  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

## Example: Educational testing data



**Exercise:** Identify the populations and subpopulations.

## Example: Environmental monitoring data



**Exercise:** Identify the populations and subpopulations.

## Group-specific inferences

**Targets of inference:** Subpopulation means  $\theta_1, \dots, \theta_p$ .

**Data:** Subpopulation samples  $\{y_{1,1}, \dots, y_{1,n_1}\}, \dots, \{y_{1,p}, \dots, y_{1,n_p}\}$ .

**Statistical methods:**

- Variance tests and estimation: What is  $\text{Var}[\theta_1, \dots, \theta_p]$ ? Is it zero?
- Estimates of  $\theta_j$ :  $\hat{\theta}_j = \bar{y}_{\cdot j}$  or  $\hat{\theta}_j = w\bar{y}_{\cdot j} + (1 - w)\bar{y}_{\cdot}$ ;
- Confidence intervals:  $\Pr(\theta_j \in C(\mathbf{y}) | \theta_j) = 1 - \alpha$ , or  $\Pr(\theta^* \in C(\mathbf{y})) = 1 - \alpha$ ;



## Cluster sampling

**Survey design:** Consider the costs of obtaining soil samples from

- 100 randomly sampled locations in a city, versus
- 10 randomly sampled locations from 10 randomly sampled neighborhoods.

**Cluster sampling:**

The second sampling scheme is called *cluster sampling* or *two-stage sampling*.

Cluster sampling

- is often cheaper per sampled unit;
- often gives less reliable estimates of population means.

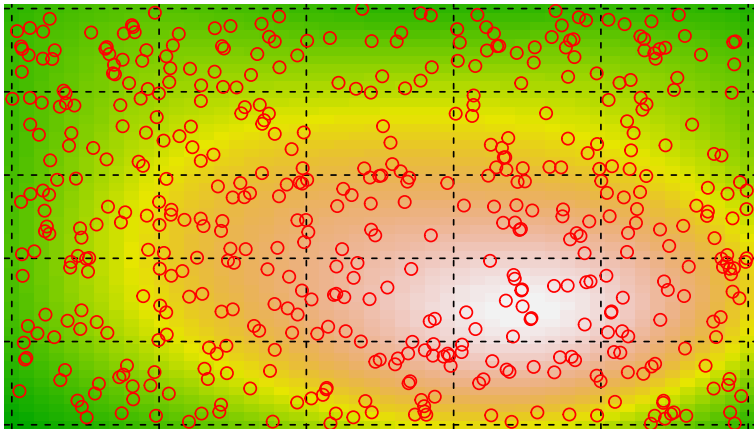
## Estimation of a population mean

**Task:** Estimate the population mean  $\mu$  from sample data.

### Questions:

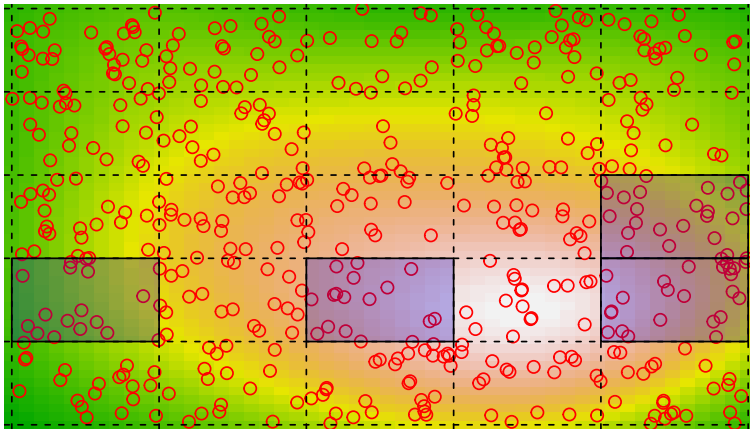
- How do cluster sampling and SRS compare?
- How do you infer  $\mu$  from cluster sample data?

## Two-stage sampling



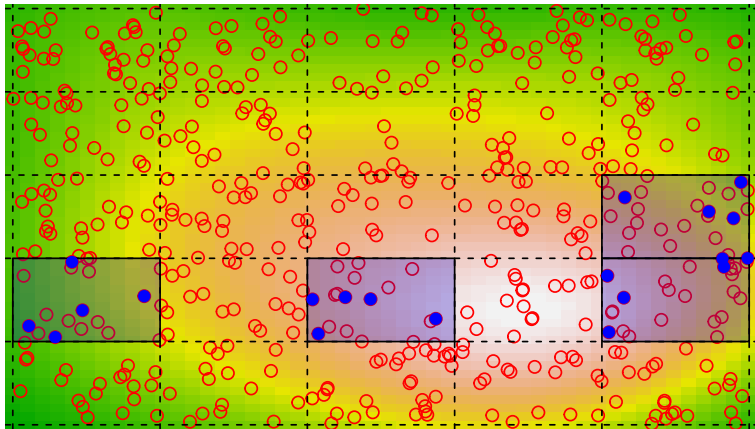
$$\mu=2.0494009$$

## Two-stage sampling



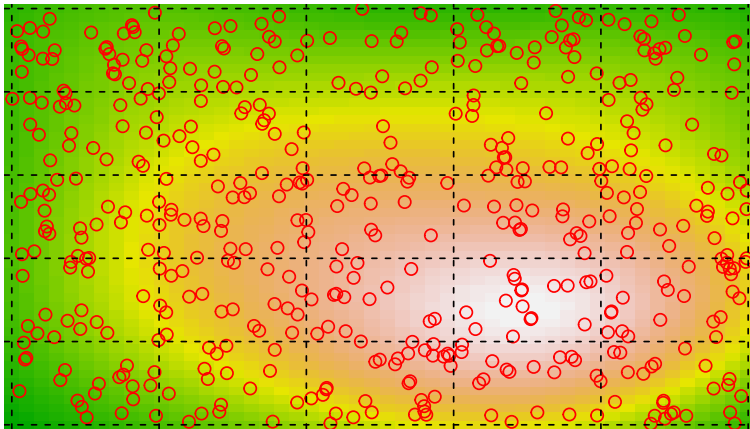
$$\mu=2.0494009$$

## Two-stage sampling



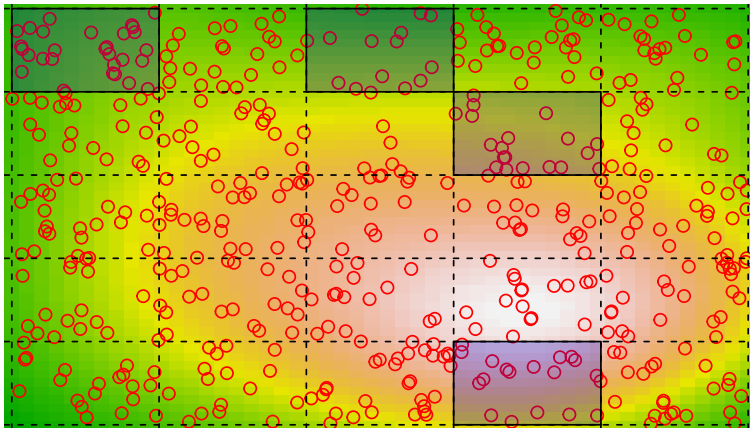
$$\mu=2.0494009 \quad , \quad \bar{y}=2.3547727$$

## Two-stage sampling



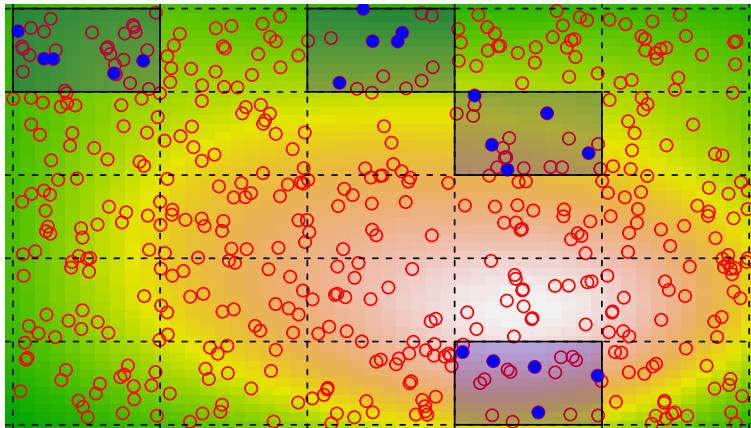
$$\mu=2.0494009$$

## Two-stage sampling



$$\mu=2.0494009$$

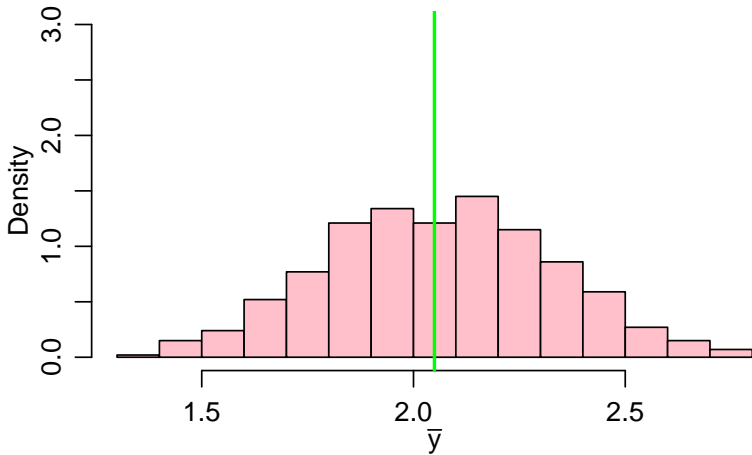
## Two-stage sampling



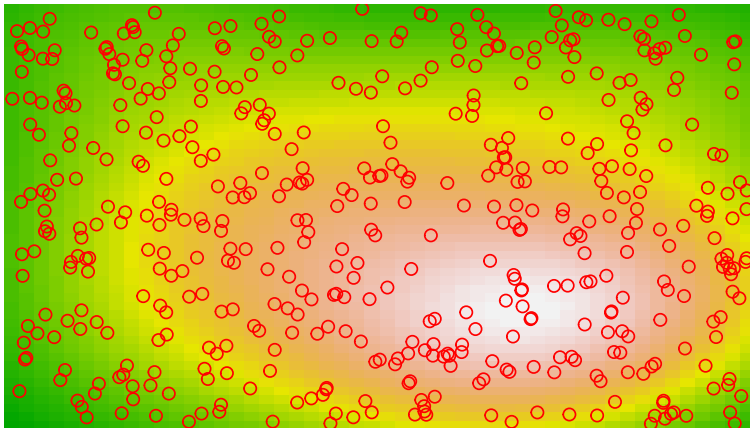
$$\mu=2.0494009 \quad , \quad \bar{y}=1.896463$$



## Variability of sample mean

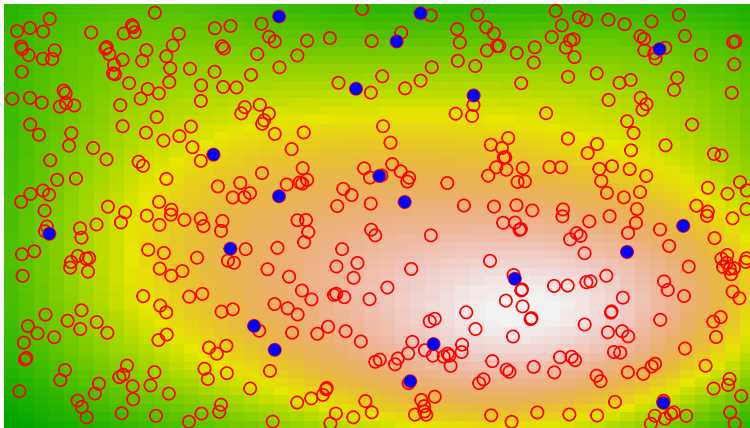


## Comparison to SRS



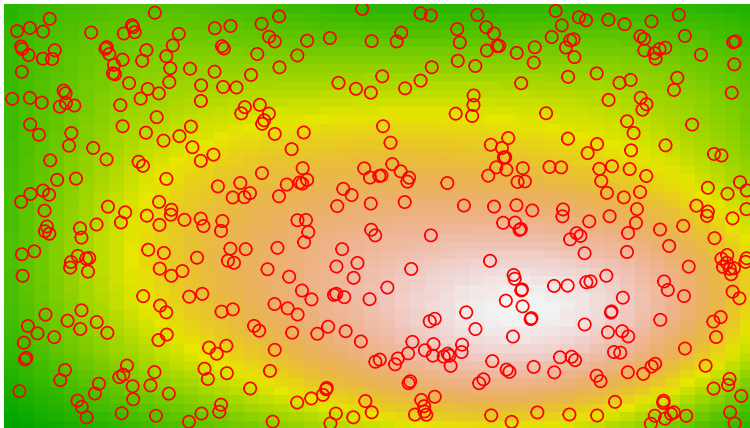
$$\mu=2.0494009$$

## Comparison to SRS



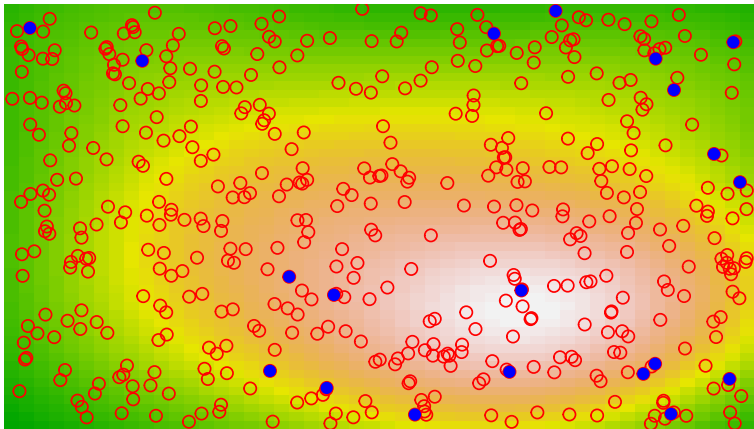
$$\mu=2.0494009 \quad , \quad \bar{y}=2.1696295$$

## Comparison to SRS



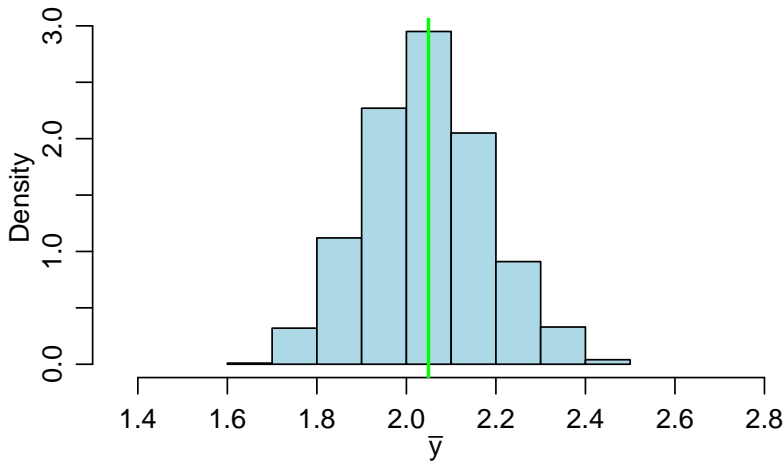
$$\mu=2.0494009$$

## Comparison to SRS

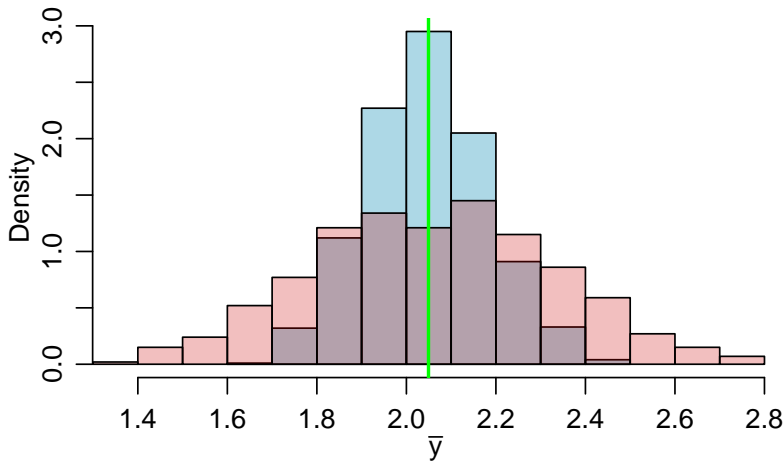


$$\mu=2.0494009 \quad , \quad \bar{y}=1.9804926$$

## Variability of sample mean



## Comparison of sampling variability



## Heterogeneity, homogeneity and dependence

As we will show mathematically,

across-group heterogeneity  $\Leftrightarrow$  within-group homogeneity  
 $\Leftrightarrow$  within-group correlation or dependence

Across-group heterogeneity increases the variance of the sample mean, and so

$$\text{Var}[\bar{y}_{tss}] \geq \text{Var}[\bar{y}_{srs}]$$

if the total samples sizes are the same.



## Ignoring across-group heterogeneity

**Task:** Construct a 95% CI for the population mean.

***t*-interval for SRS:**

If  $y_1, \dots, y_n$  is an iid sample with  $E[y_i] = \mu$  and  $\text{Var}[y_i] = \sigma^2$ ,

$$E[\bar{y}] = \mu, \text{Var}[\bar{y}] = \sigma^2/n.$$

By the central limit theorem,

$$\bar{y} \dot{\sim} N(\mu, \sigma^2/n), \quad \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0, 1).$$

As  $\sigma^2$  is generally unknown, we use

$$\frac{\bar{y} - \mu}{s/\sqrt{n}} \dot{\sim} t_{n-1}, \quad \text{where } s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2.$$

From this, we have

$$\bar{y} \pm t_{n-1, .975} \times s/\sqrt{n} \text{ is a 95\% CI for } \mu.$$

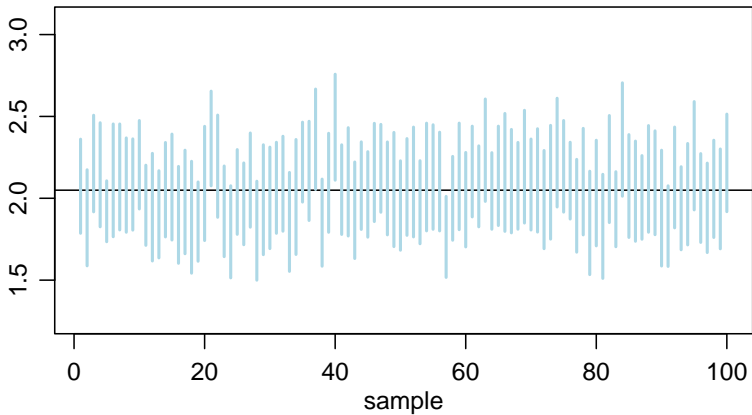
Multilevel data  
○○○

Subpopulation inferences  
○○○

Population inferences  
○○○○○○○○○○○○○○○○○○○○●○○○

Cross-level inferences  
○○○○○○○○○○○○○○○○○○

## Ignoring across-group heterogeneity



## Ignoring across-group heterogeneity

$$\bar{y} \pm t_{n-1, .975} \times s/\sqrt{n}$$

What if we apply the formula to data from a cluster sample?

If  $y_1, \dots, y_n$  are from a SRS, then

$$\text{Var}[\bar{y}] = \sigma^2/n = E[s^2/n].$$

$s/\sqrt{n}$  provides a good estimate of the sd of  $\bar{y}$ .

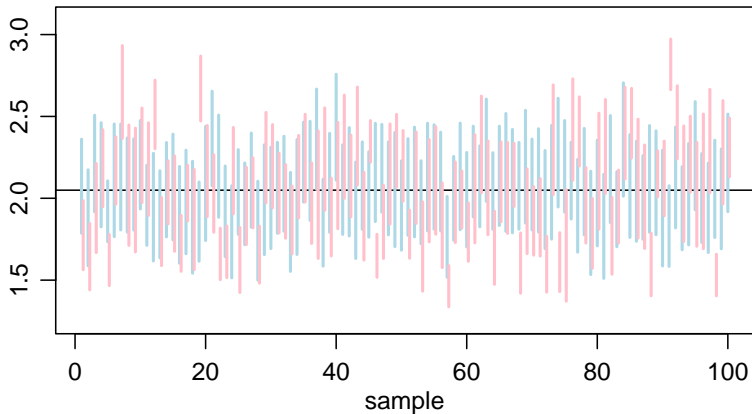
If  $y_1, \dots, y_n$  are from a cluster sample, then generally

$$\text{Var}[\bar{y}] > \sigma^2/n \approx E[s^2/n].$$

$s/\sqrt{n}$  is generally an underestimate of the sd of  $\bar{y}$ .

How will the resulting confidence interval behave if  $\text{sd}(\bar{y}) > s/\sqrt{n}$ ?

## Ignoring cross-group heterogeneity



## Ignoring across-group heterogeneity

### Summary:

- Across-group heterogeneity = within-group similarity.
- Within-group similarity leads to positively correlated cluster sample data.
- The variance of the sample mean from (positively) correlated data is higher than that of the mean of uncorrelated data.
- Statistical inference ignoring such correlation will be inaccurate.

### Remedy: We will develop techniques to

- evaluate within- and across-group heterogeneity;
- provide accurate statistical inference based on cluster samples.

## Estimation of a treatment effect

Suppose

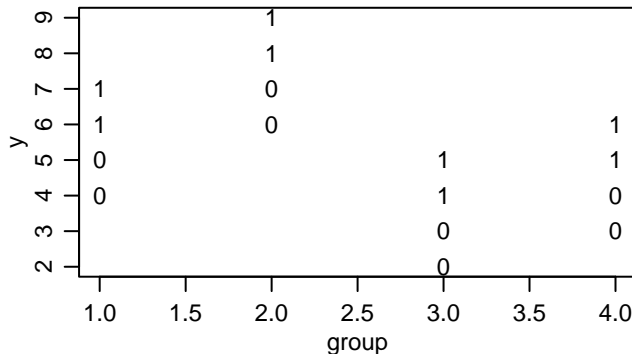
- $x \in \{0, 1\}$
- $\mu_1 = E[y|x = 1]$
- $\mu_0 = E[y|x = 0]$

**Task:** Estimate the difference  $\delta = \mu_1 - \mu_0$  based on cluster sample data.

**Data:** For each group  $j$ , we have  $(y_{1,j}, x_{1,j}), \dots, (y_{n,j}, x_{n,j})$ .

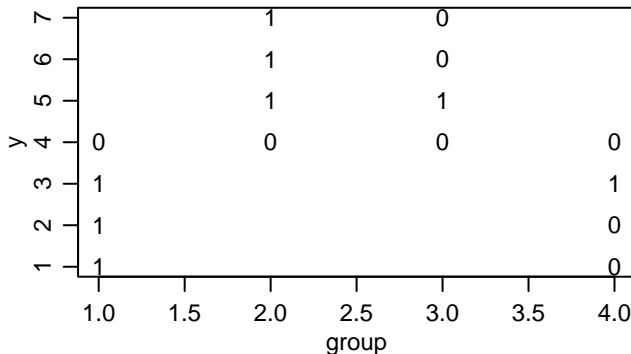
**Question:** What could go wrong by ignoring the multilevel nature of the data?

## Overconservative analysis



- Overlap across groups, no overlap within groups.
- Across-group variation is *large* compared to the treatment effect.
- Ignoring group differences can lead to *overconservative analysis*.

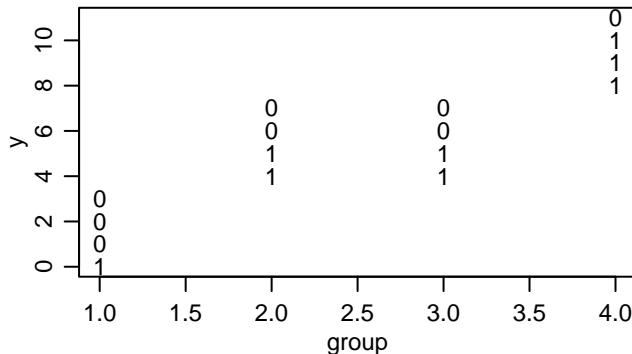
## Underconservative analysis



- The population mean difference is zero.
- The sample mean difference based on pairs of two groups is not zero.
- Ignoring group differences can lead to *underconservative analysis*.



## Effect reversal



- $\mu_1 - \mu_0 > 0$  in population,  $\mu_{1,j} - \mu_{0,j} < 0$  in every group.
- Within-group effects may be different from population effects.
- This is sometimes called *Simpson's paradox*.

## Consequences of across-group heterogeneity

### Summary:

- Across-group heterogeneity can lead to *over or under* conservative analysis.
- Population-level effects may be different from group-level effects.
- Data analysis ignoring groups can be inaccurate in *unpredictable* ways.

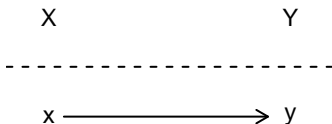
### Remedy: We will develop techniques to

- differentiate between macro and micro level effects;
- appropriately control for within and between-group heterogeneity.

## Macro and micro effects

$X$ ,  $x$  are macro and micro level explanatory variables

$Y$ ,  $y$  are macro and micro level outcome variables



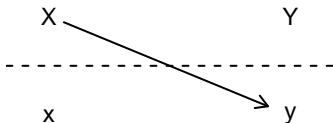
What are the effects of SES ( $x$ ) on political opinion ( $y$ )?

(a *micro-micro effect*)

## Macro, micro and cross-level effects

$X, x$  are macro and micro level explanatory variables

$Y, y$  are macro and micro level outcome variables



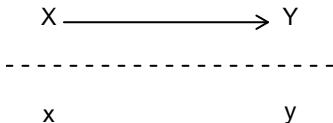
What are the effects of State GDP ( $X$ ) on political opinion ( $y$ ) ?

(a *macro-micro effect*)

## Macro, micro and cross-level effects

$X, x$  are macro and micro level explanatory variables

$Y, y$  are macro and micro level outcome variables

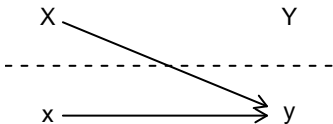


What are the effects of State GDP ( $X$ ) on statewide political opinion ( $Y$ )?  
(a *macro-macro effect*)

## Macro, micro and cross-level effects

$X$ ,  $x$  are macro and micro level explanatory variables

$Y$ ,  $y$  are macro and micro level outcome variables



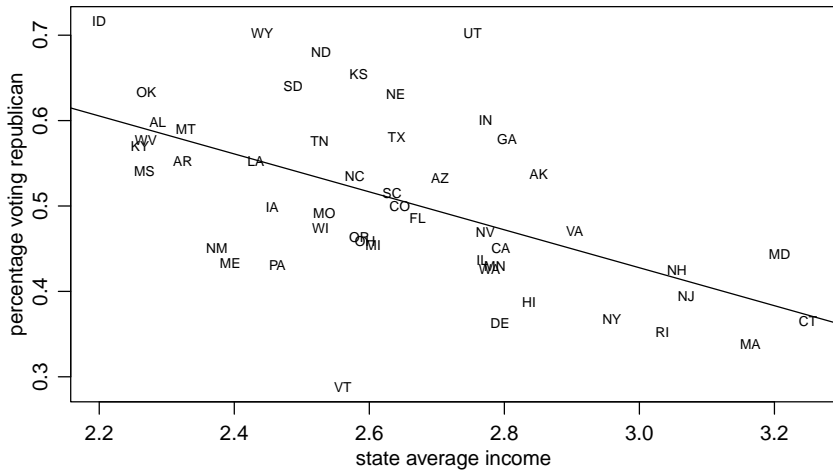
What are the effects of State GDP ( $X$ ) and SES ( $x$ ) on political opinion ( $y$ )?  
(*multilevel effects*)

## Example: Income and voting patterns

Exit poll data from 2004 presidential election

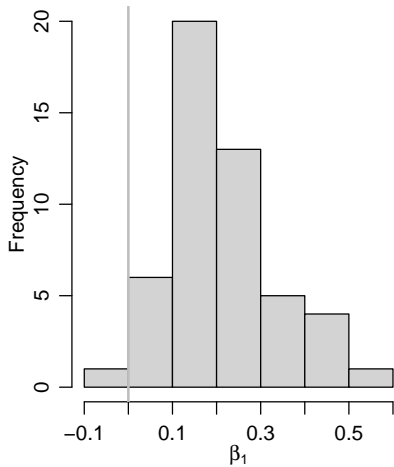
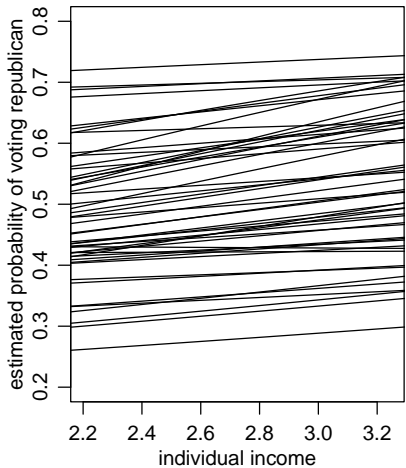
- $j \in \{1, \dots, 50\}$  indexes the states,
- $y_{i,j}$  is the voting variable for person  $i$  in state  $j$ ,
- $x_{i,j}$  is a measure of income for person  $(i, j)$ .

## Macro effects





## Micro effects



## Joint estimation of effects

In general we may be interested in understanding all of the following:

- macro level effects,
- micro level effects,
- macro effects on micro variables,
- heterogeneity of micro effects across groups.

Inference for these items can be made with LME and GLME models:

$$\begin{aligned} y_{i,j} &\sim a_j + b_j x_{i,j} + \epsilon_{i,j} \\ &= (\alpha_0 + \alpha_1 w_j + z_j) + (\beta_0 + \beta_1 w_j + e_j) x_{i,j} + \epsilon_{i,j}. \end{aligned}$$