Thesis Proposal

Sunthud Pornprasertmanit
W. Joel Schneider

Sample size estimation for Two-Group Cluster Randomized Designs

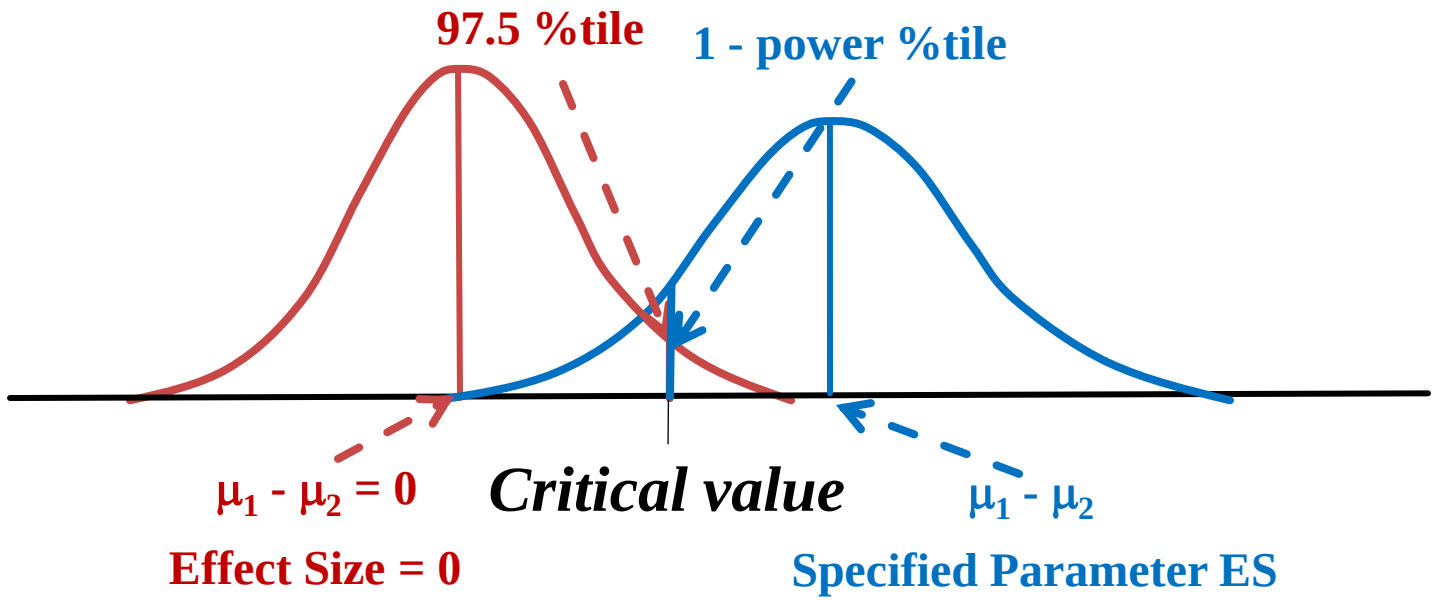
Outline

- Sample Size Estimation Approach
  - Power
  - Accuracy in Parameter Estimation
- Illustration in independent *t*-test
- Cluster Randomized Design (CRD)
- The Proposed Program

Two Approaches of Sample Size Estimation

- Power analysis
  - The probability of a significant result when there is a real effect in the population
- Width of Confidence Interval of Effect Size (*CI* of *ES*)
  - The accuracy of effect size estimation

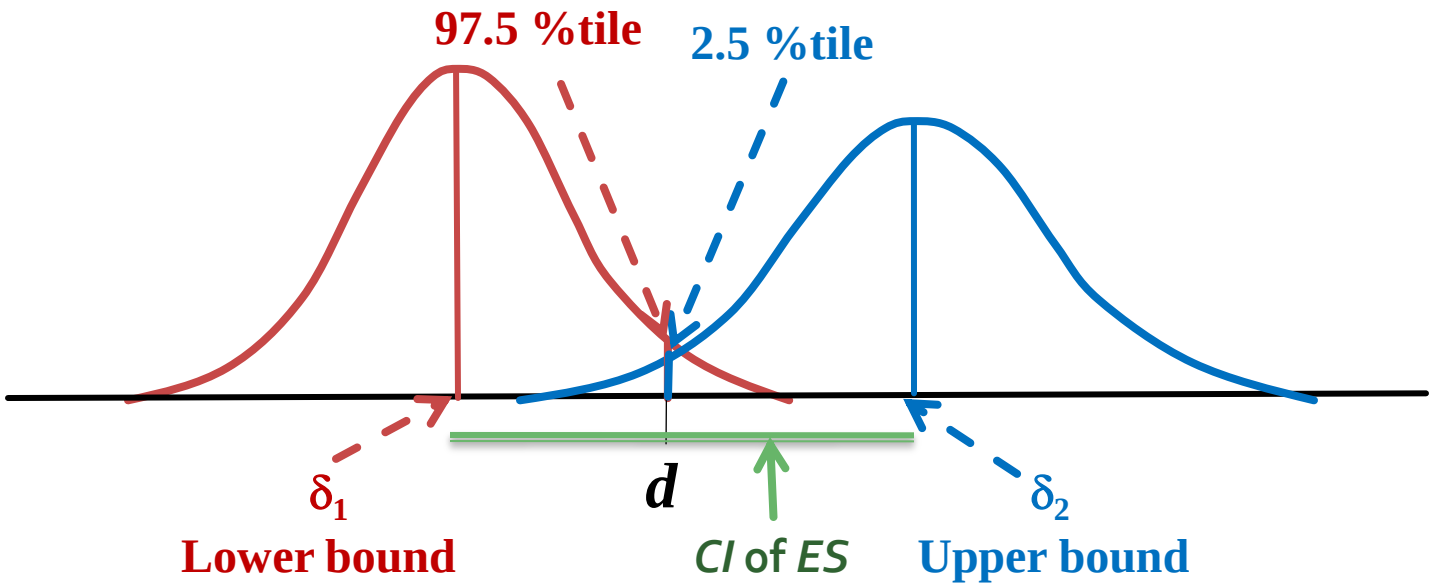
Power of Independent *t*-tests



- More *n* → Less *SE* → More power

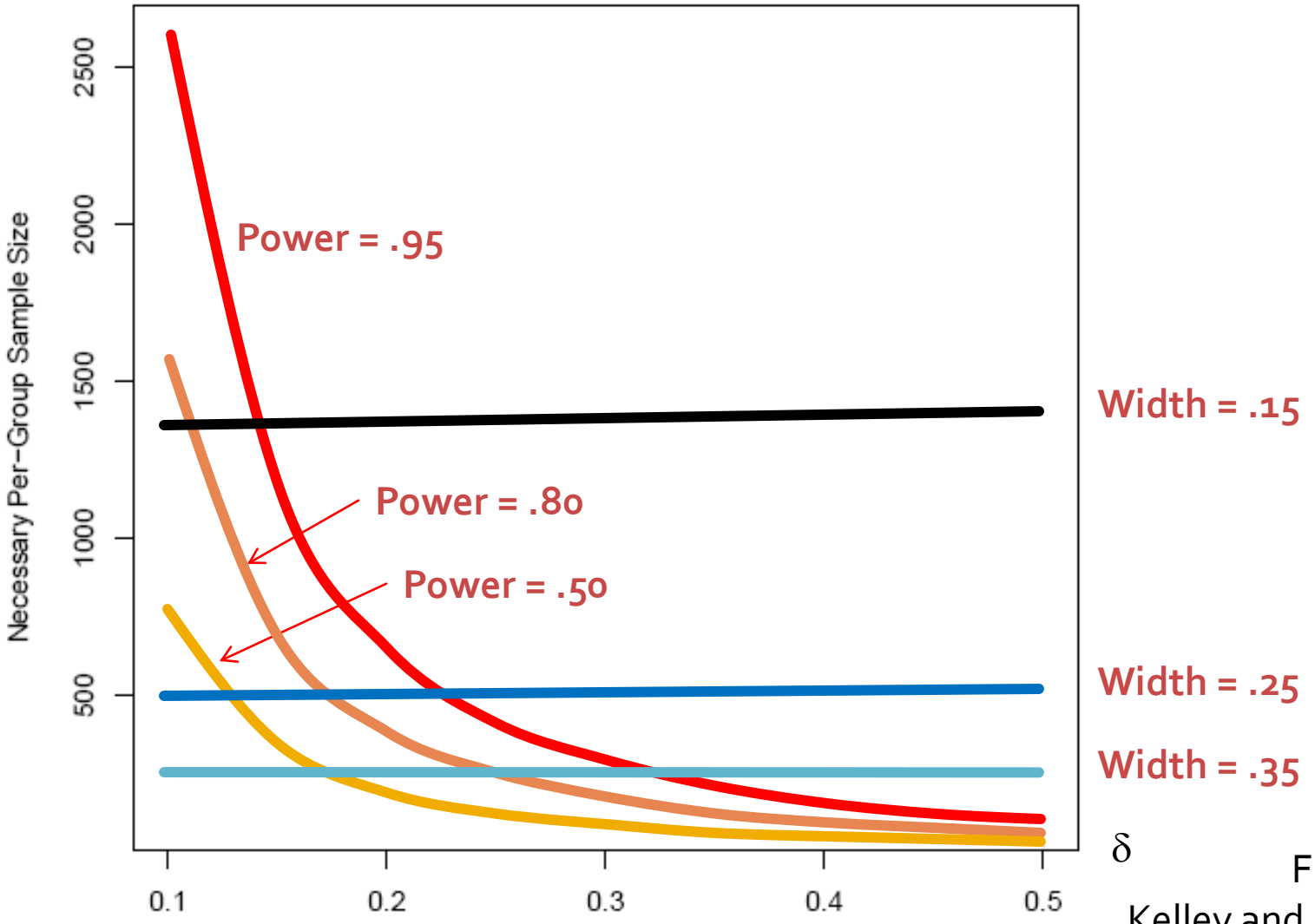
Width of CI of ES

- 95 % CI of a difference between independent means



- More *n* → Less *SE* → Less Width of *CI* of *ES*

Comparing two approaches

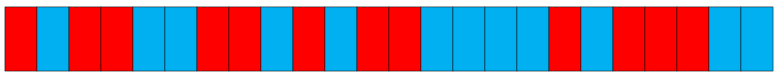


From
Kelley and Rausch (2006)

Cluster Randomized Design

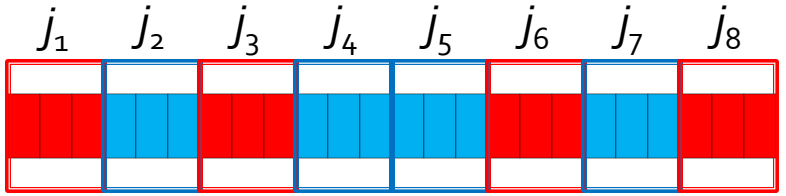
- CRD is the analysis of group differences when groups are randomly assigned to different conditions

Independent t-test



All sample size = 24

Two-group CRD



All sample size = 24

$J = 8$

$n = 3$

Cluster Randomized Design

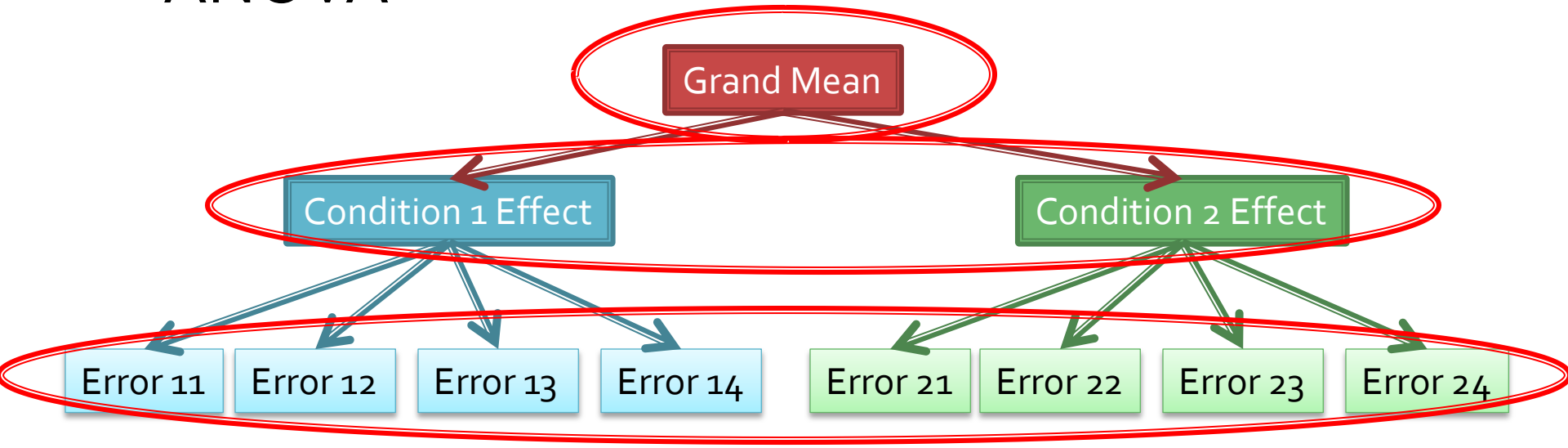
- Using Independent *t*-test
  - Independence of error terms assumption has been violated
    - Similar experience within clusters
  - Inflate type I error
- CRD accounts for interdependence

Basic Concepts in CRD

- Two types of errors in CRD
  - Group-level error variance
  - Individual-level error variance
- Intraclass correlation (ICC)
- Effect Size in CRD
- *CI* of *ES*
- Covariate Effect in CRD

Error terms

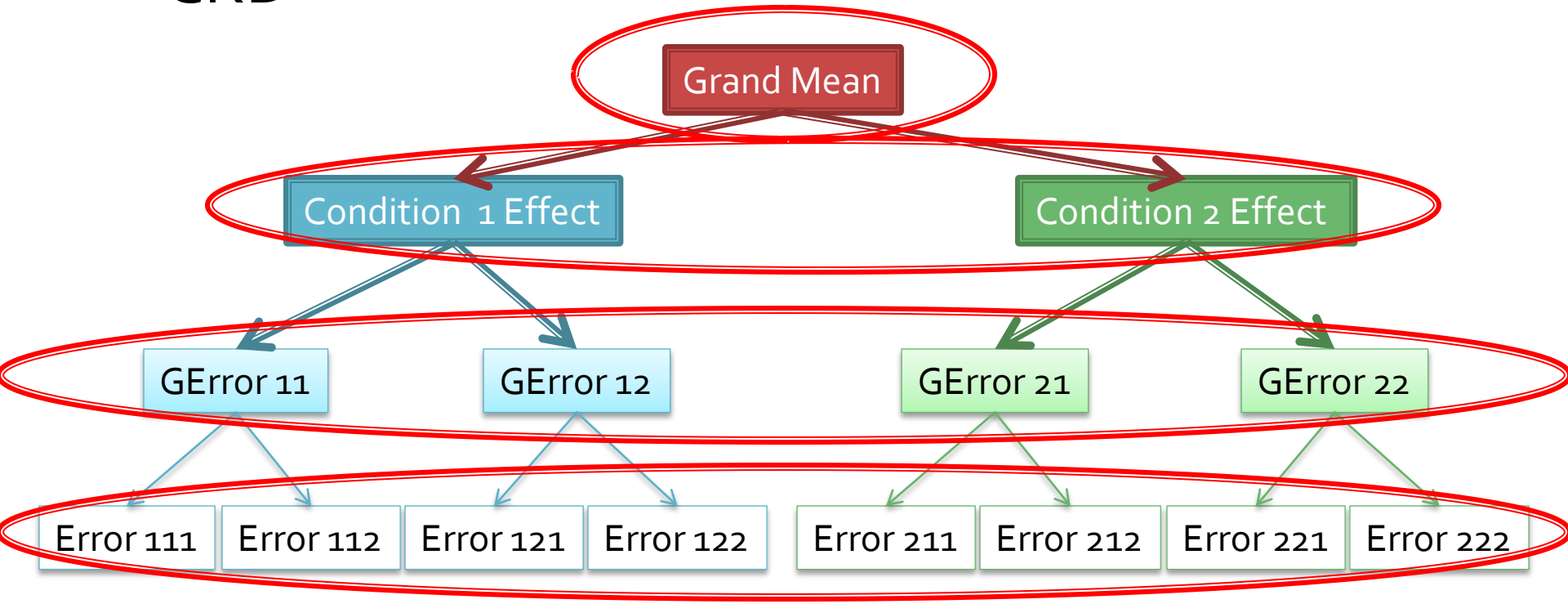
- ANOVA



$$Y_{ki} = \bar{Y}.. + \alpha_k + e_{ki}$$

Error terms

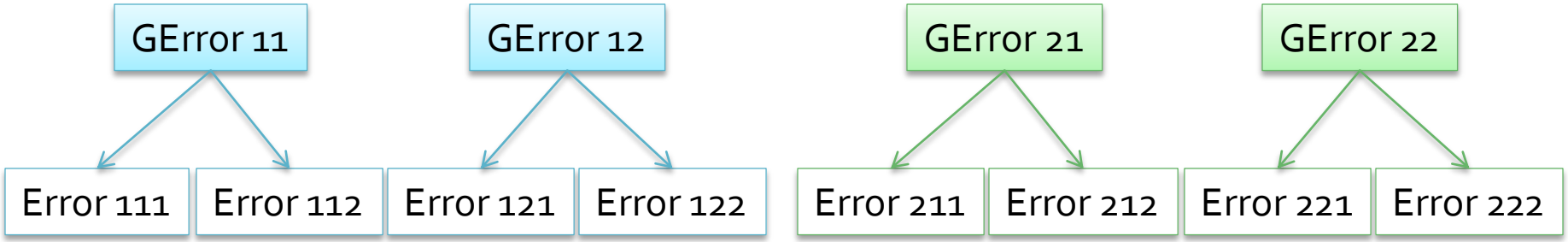
- CRD



$$Y_{kij} = \bar{Y}.. + \alpha_k + u_{ki} + e_{kij}$$

Error terms

- Group error → common experience in a group
- Individual error → unique experience of each individual



$$Y = \overline{Y}.. + \alpha_k + u_{ki} + e_{kij}$$

Error terms

- CRD

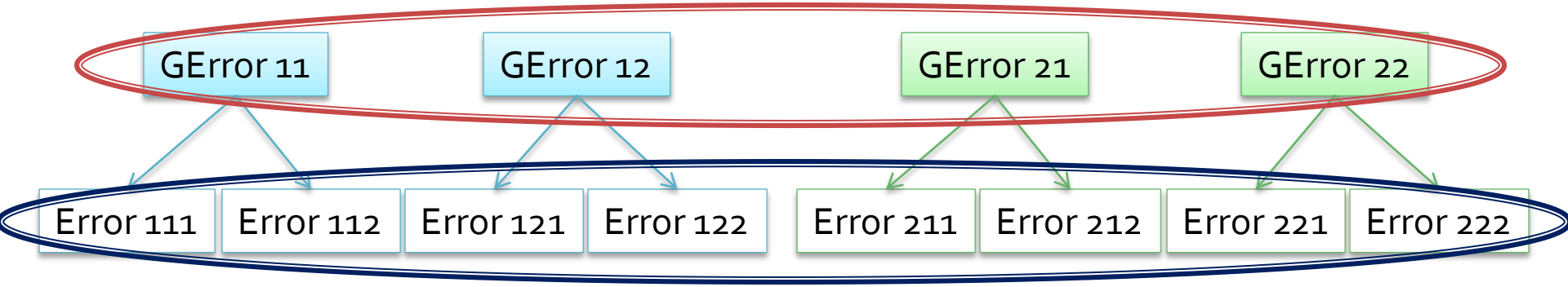
$$Y = \bar{Y}.. + \alpha_k + u_{ki} + e_{kij}$$

Intraclass Correlation

$$\rho = \frac{\tau}{\tau + \sigma}$$

Group Error Variance

$$Var(u_{ki}) = \tau$$



Individual Error Variance

$$Var(e_{kij}) = \sigma$$

Effect Size

- Effect Size Definition

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

- In single level design, σ is pooled *SD* or $\sqrt{MS_{error}}$
- In CRD, three types of pooled *SD*
  - Group or $\sqrt{\tau}$
  - Individual or $\sqrt{\sigma}$
  - Total or $\sqrt{\tau + \sigma}$

Effect Size

- Hedges (2007) guideline
- In this study, use only individual pooled *SD*
- Assume $\sigma = 1$ → Effect Size = Condition Difference

CI of *ES*

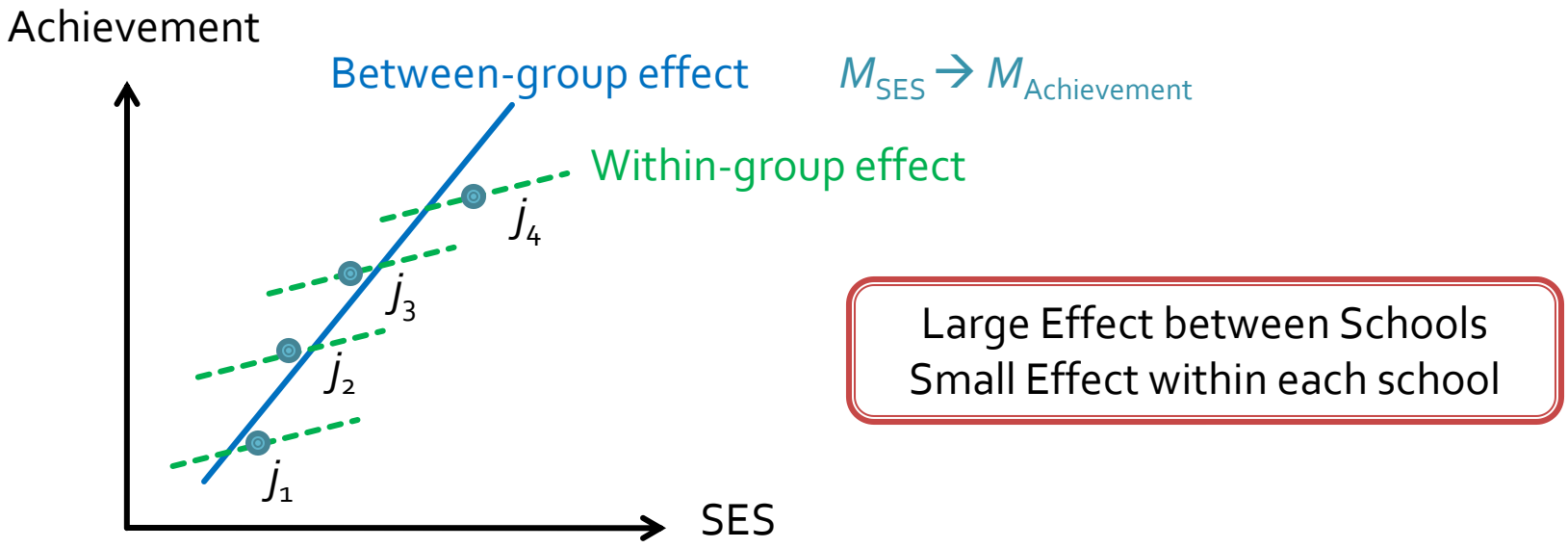
- Formula by Hedges (2007)
- Nonlinear constraint in Mplus
  - Duplicate *Y*
  - Regress the original *Y* on the treatment variable (and the covariate)
  - Make the nonlinear constraint:

$$\text{Effect Size} = \frac{\gamma_{Y \text{ on } X}}{\sqrt{\sigma_{\text{Duplicated}Y}}}$$

  - Mplus will find *CI* of *ES* based on ML estimation

Covariate Effect

- Different amount of error variance explained across levels
  - e.g., SES → Achievement [4 Schools]



Covariate Effect

- For an individual-level covariate
  - Cluster-level R^2 and Individual-level R^2
- For a cluster-level covariate
  - Cluster-level R^2

Finding Sample Size

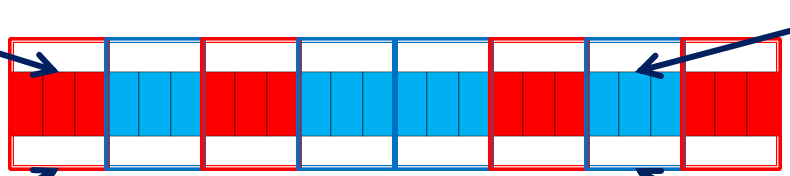
- Different Combination of three factors can yield the same power or width of *CI*
  - Number of Clusters (*J*)
  - Cluster size (*n*)
  - Proportion of treatment clusters (*p*)
- Different Combination also yield same costs

Finding Sample Size

- Four costs

Treatment Individual Cost (TIC)

Control Individual Cost (CIC)



Treatment Group Cost (TGC)

Control Group Cost (CGC)

Each Treatment Group Cost = **TGC** + (*n* x **TIC**)

Each Control Group Cost = **CGC** + (*n* x **CIC**)

Number of Treatment Groups = pJ

Number of Control Groups = $(1-p)J$

$$\text{Total Cost} = pJ(\mathbf{TGC} + (n \times \mathbf{TIC})) + (1-p)J(\mathbf{CGC} + (n \times \mathbf{CIC}))$$

Finding Sample Size

- Three criteria
  - Minimize number of overall individuals by specified power/width
    - Find various *n*, *J*, *p* for given power/width → Find lowest *nJ*
  - Minimize cost by specified power/width
    - Find various *n*, *J*, *p* for given power/width → Find lowest cost
  - Maximize power/ Minimize width by specified cost
    - Find various *n*, *J*, *p* for given cost → Find highest power/width

Logic

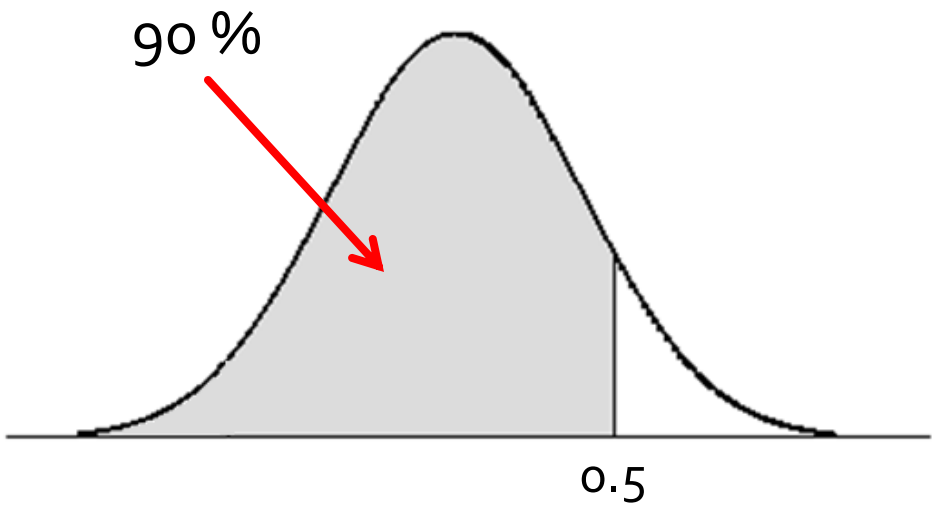
- Find starting values by normal approximation
- Find more accurate result by a priori Monte Carlo Simulation

Features of the Program

- Finding Sample Sizes for three criteria
- Accounted for a covariate
- Post Hoc / A priori
- May specify *n*, *J*, or *p*, in advance
- Report either the starting value or the Monte Carlo result
- Degree of certainty in *CI* of *ES*

Degree of Certainty

- Width is also prone to sampling error
- Degree of Certainty = The proportion of width less than a specified level
  - Desired width = 0.5 with 90% Degree of Certainty



Future Plan

- Test the program accuracy
- Check for difference between
  - Starting values
  - A priori Monte Carlo results
- Writing user's manual
- (May make a video clip for explaining how to use the program)

Testing Program Accuracy

- Test backward
- The proposed program will find the estimated sample sizes combination
  - "PINT" find the variance of treatment effect
- Not exactly equal but close enough

Program Illustration