Thesis Progress

Sunthud Pornprasertmanit
W. Joel Schneider

Sample size estimation for Two-Group Cluster Randomized Design

Introduction

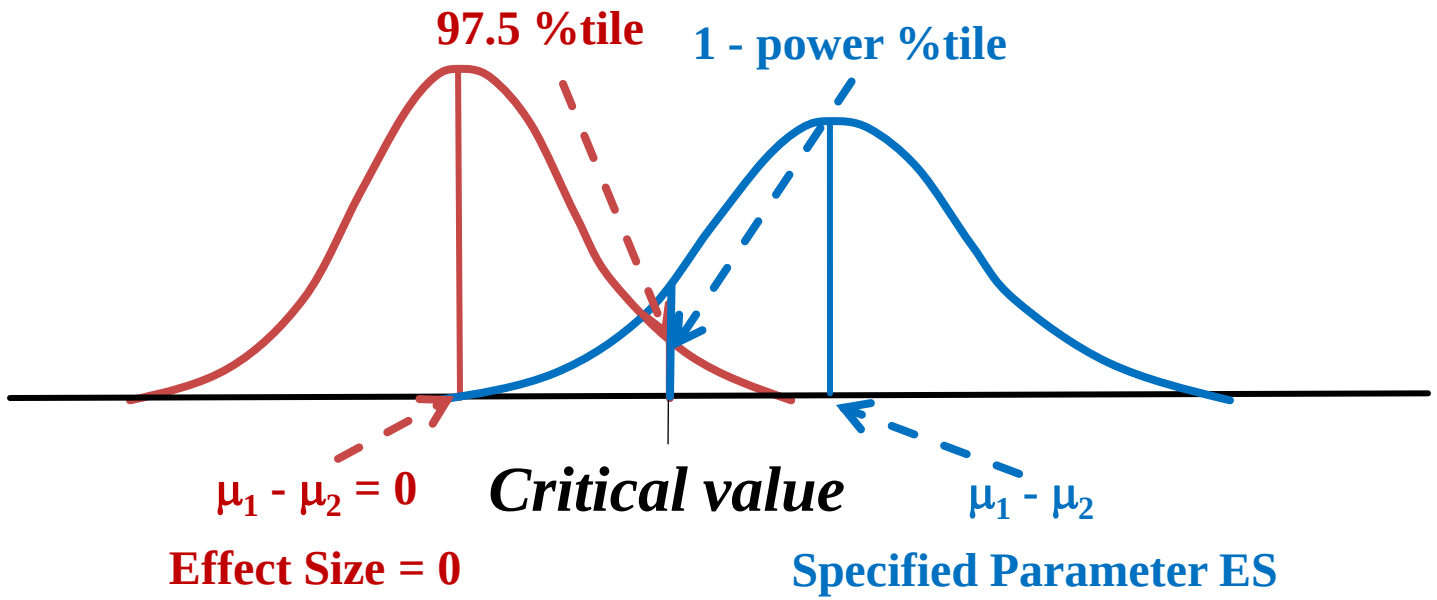
- Two approaches of sample size estimation
- Advantages of CRD over ANOVA
- Basic Concepts for CRD
- Two-Group CRD Formula
- Sample Size Estimation in CRD

Two Approaches of Sample Size Estimation

- Power analysis
  - The probability of significant result from real effect in population
- Width of *CI* of *ES*
  - The accuracy of effect size estimation

Power Analysis

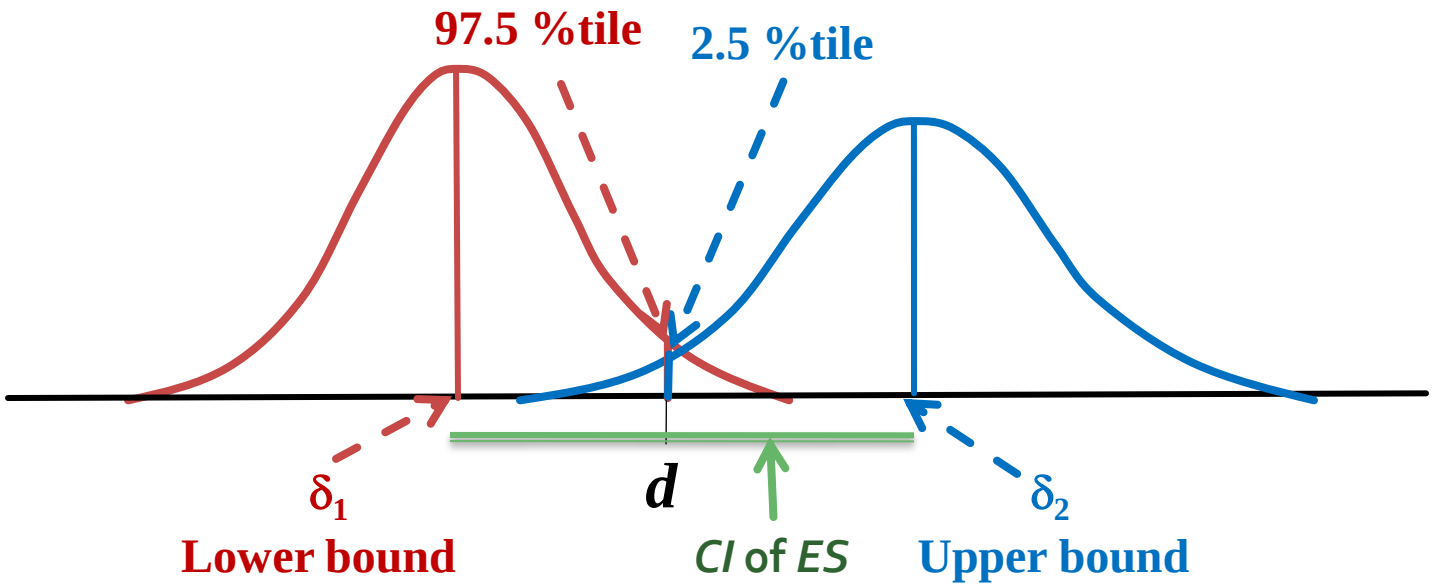
- Example → Independent *t*-test
  - Power of difference between two independent means



  - More *n* → Less *SE* → More power

Width of CI of ES

- Example → Independent *t*-test
  - 95 % CI of a difference between independent means

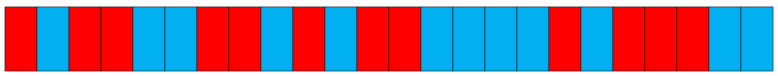


  - More *n* → Less *SE* → Less Width of *CI* of *ES*

Cluster Randomized Design

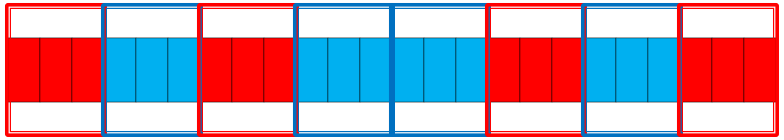
- CRD is the analysis of group differences when groups are randomly assigned to different conditions

Independent t-test



All sample size = 24

Two-group CRD



All sample size = 24

$J = 8$

$n = 3$

Cluster Randomized Design

- Characteristics of CRD data
  - Similarity within group
  - The errors within group are correlated
  - Inflated variability of random error

Cluster Randomized Design

- Find error variance in each design
  - Variance-covariance matrix

ANOVA data

σ_e	0	0	0	0	0
0	σ_e	0	0	0	0
0	0	σ_e	0	0	0
0	0	0	σ_e	0	0
0	0	0	0	σ_e	0
0	0	0	0	0	σ_e

$\mathrm{Var}(M_e) = \sigma_e$

CRD data

σ_e	τ	τ	0	0	0
τ	σ_e	τ	0	0	0
τ	τ	σ_e	0	0	0
0	0	0	σ_e	τ	τ
0	0	0	τ	σ_e	τ
0	0	0	τ	τ	σ_e

$\mathrm{Var}(M_e) > \sigma_e$

Cluster Randomized Design

- What happened when H_o is true and using ANOVA

<u>ANOVA data</u>

Independent error terms

$$\text{Var}(M_e) = \sigma_e$$

$$F = \frac{\sigma_{M_e}}{\sigma_e} = \frac{\sigma_e}{\sigma_e} = 1$$

Accurate type I error

<u>CRD data</u>

Correlated error terms

$$\text{Var}(M_e) > \sigma_e$$

$$F = \frac{\sigma_{M_e}}{\sigma_e} = \frac{> \sigma_e}{\sigma_e} \text{ then } F > 1$$

Inflated type I error

Cluster Randomized Design

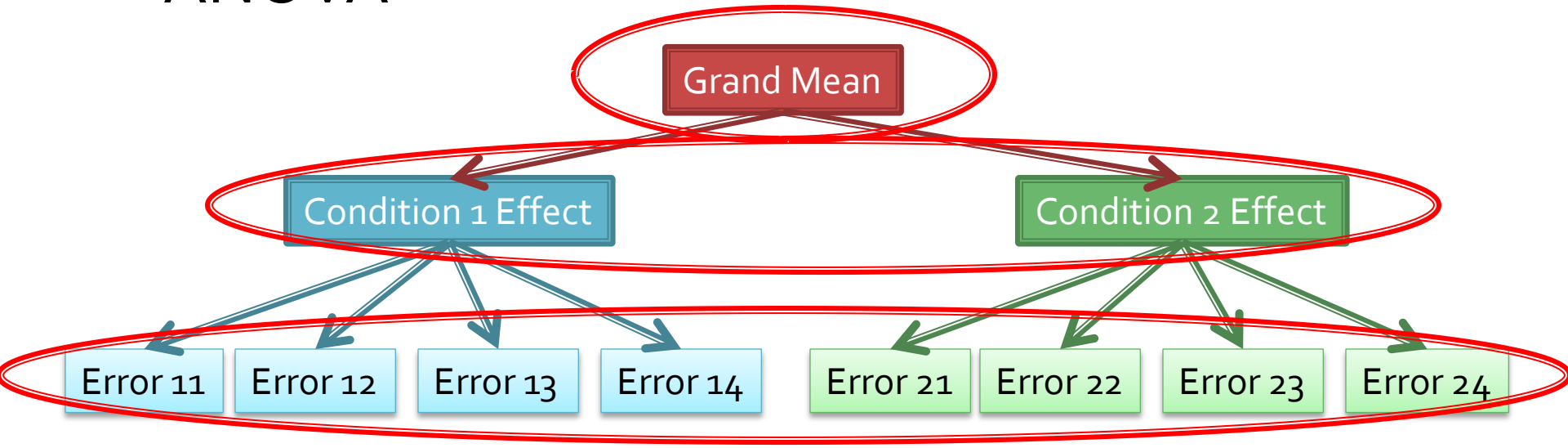
- CRD is accounted for inflated type I error
  - When groups are randomly assigned to different conditions
  - Subset of multilevel analysis

Basic Concepts in CRD

- Two types of errors in CRD
- Group-level error variance
- Individual-level error variance
- Intraclass correlation
- Effect Size in CRD

Error terms

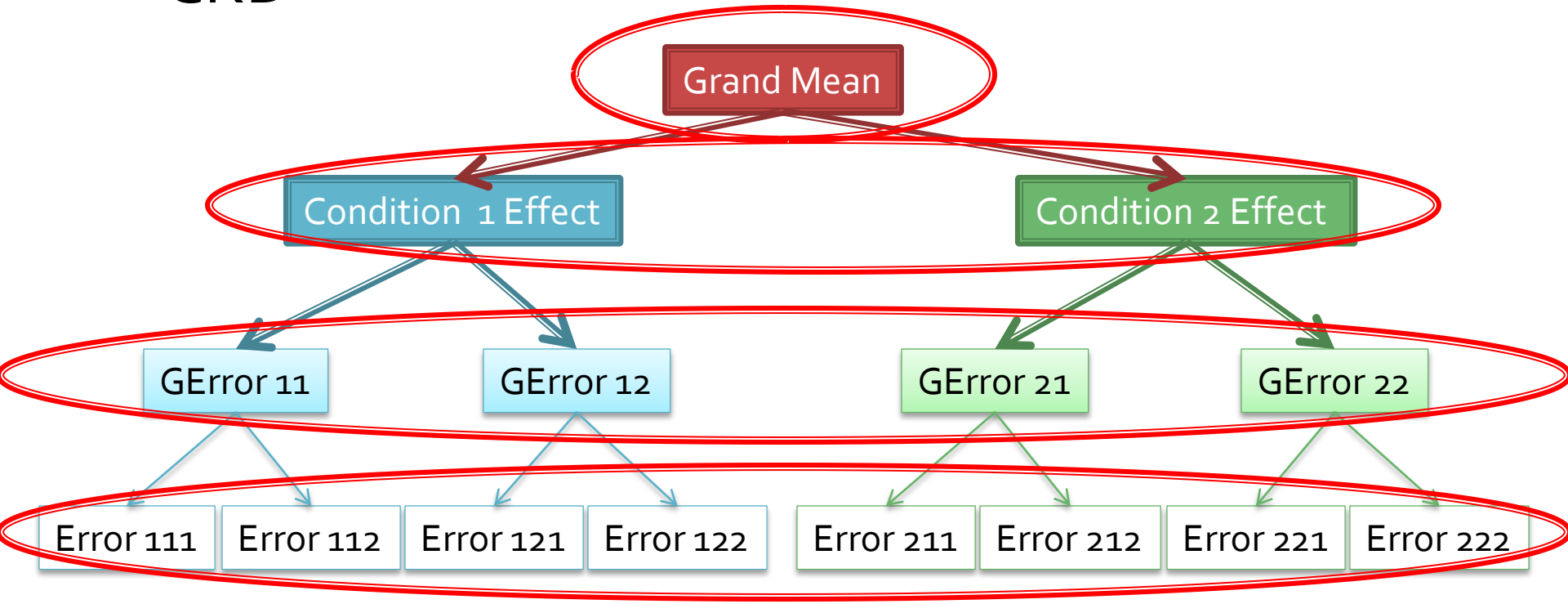
- ANOVA



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

Error terms

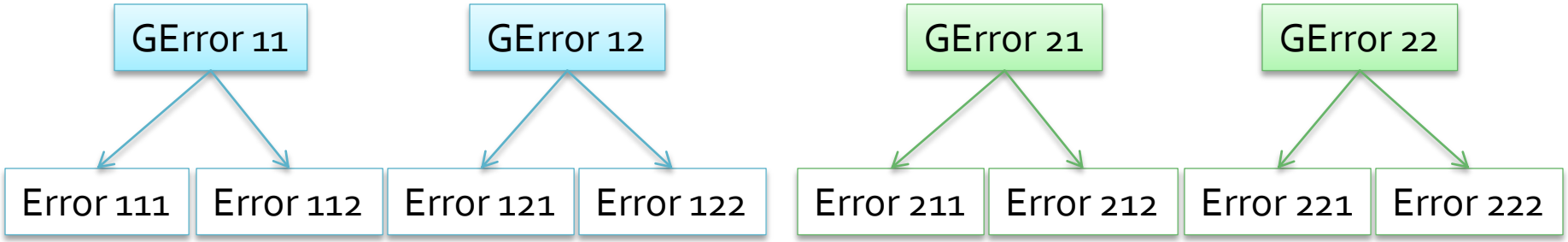
- CRD



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

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

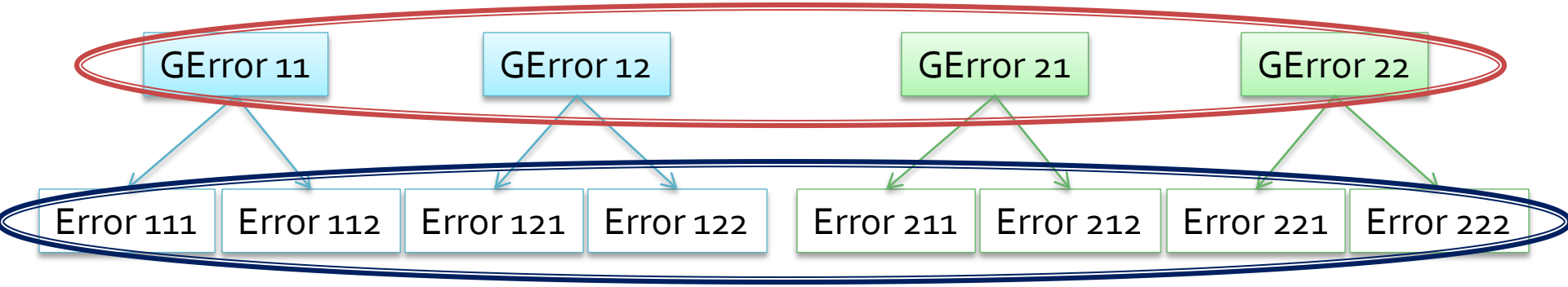
$$Y = \bar{Y}.. + \alpha_k + u_{ki} + e_{kij}$$

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) proposed
  - In group-individual levels → use individual
    - School-Students; Organization-Incumbents
  - In individual-measurement → use group
    - Applicants-GRE scores; Individuals-Social Supports
- In this study, use only individual pooled *SD*
- Assume $\sigma = 1$ → Effect Size = Group Diff

Two-Group CRD

- Equation

$$Y = M_0 + dX + u_j + e_{ij}$$

- Test group difference (d)

$$Var(d) = \frac{\sigma/n + \tau}{Jp(1-p)}$$

Finding Sample Size

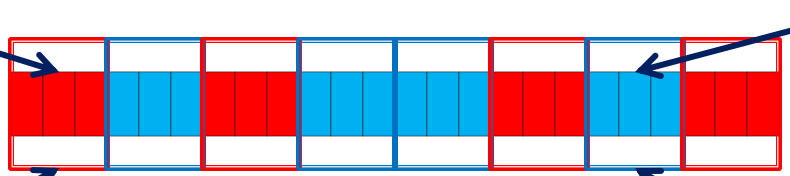
- Different Combination of three factors can yield the same power or width of CI
  - Number of Clusters (*J*)
  - Cluster size (*n*)
  - Proportion of treatment clusters (*p*)
- Different Combination also yield same costs

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 = $(1-p)J$

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

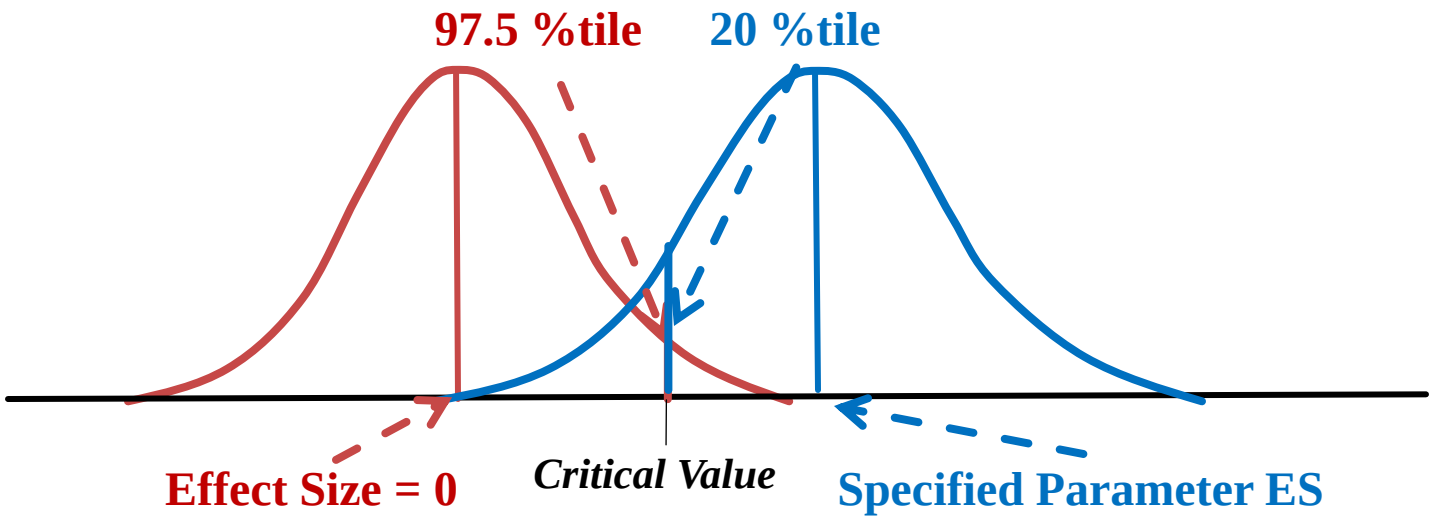
Finding Sample Size

- 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

Finding Sample Size: Criterion 1 and 2

1. Find Starting Value - Normal Dist
    1) Find combination of *n*, *J*, *p* for given power/width
    2) Find lowest *nJ* or cost
2. A Priori Monte Carlo Simulation by Mplus
    1) Adjust *n*,*J*, *p* for given power/width
    2) Find lowest *nJ* or cost
3. Summarize data by Mplus

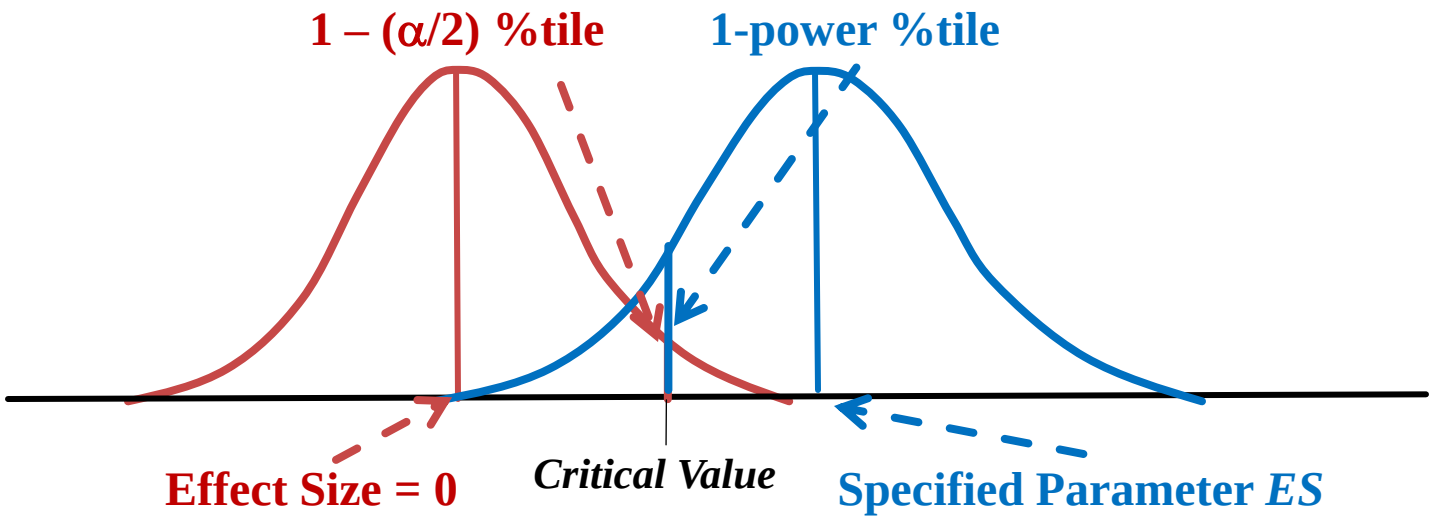
Criterion 1 and 2: Power Analysis



- Assume large sample theory

$$z = \frac{d}{\sqrt{Var(d)}}$$

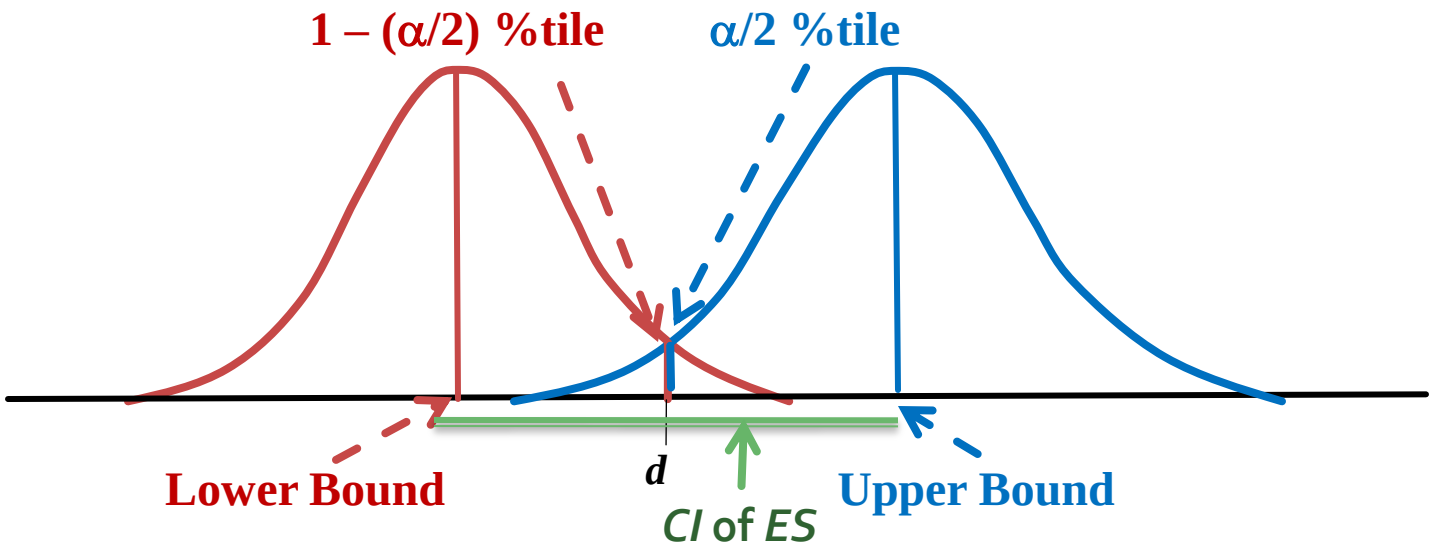
Criterion 1 and 2: Power Analysis



$$z_{1-\alpha/2} = \frac{Critical\ Value - 0}{\sqrt{Var(d)}} \qquad z_{1-power} = \frac{Critical\ Value - d}{\sqrt{Var(d)}}$$

$$Var(d) = \left(\frac{d}{z_{1-\alpha/2} - z_{1-power}} \right)^2$$

Criterion 1 and 2: CI of ES



$$z_{1-\alpha/2} = \frac{d - Lower\ bound}{\sqrt{Var(d)}} \qquad z_{\alpha/2} = \frac{d - Upper\ bound}{\sqrt{Var(d)}}$$

$$Var(d) = \left(\frac{Upper\ bound - Lower\ bound}{z_{1-\alpha/2} - z_{\alpha/2}} \right)^2 = \left(\frac{width}{2z_{1-\alpha/2}} \right)^2$$

Criterion 1 and 2 : Desired Variance Known

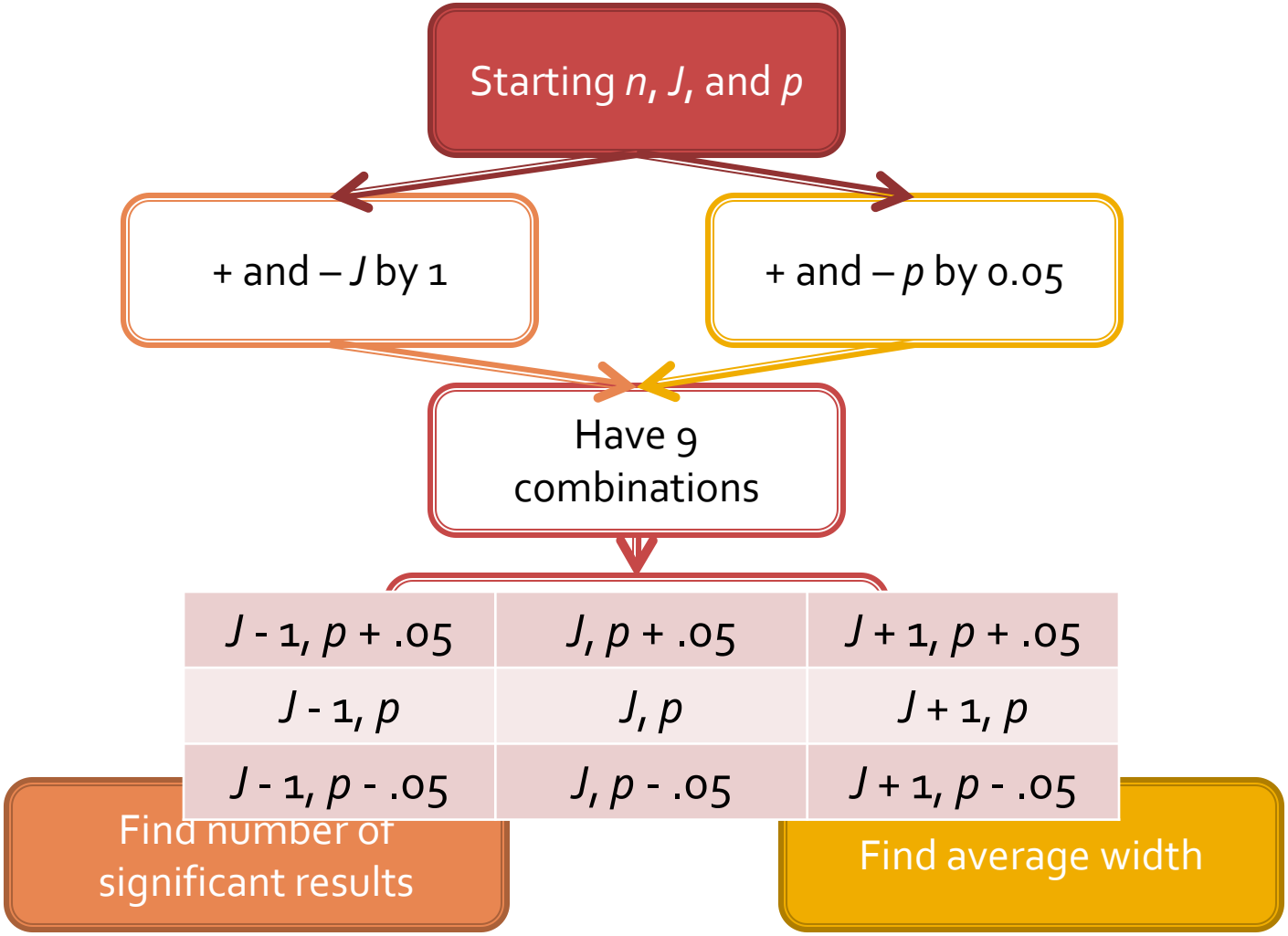
- Since *Var*(*d*) is known, we solve for various *n*, *J*, *p* by

$$Var(d) = \frac{\sigma/n + \tau}{Jp(1-p)}$$ when $\sigma = 1$; $\tau = \rho/(1 - \rho)$

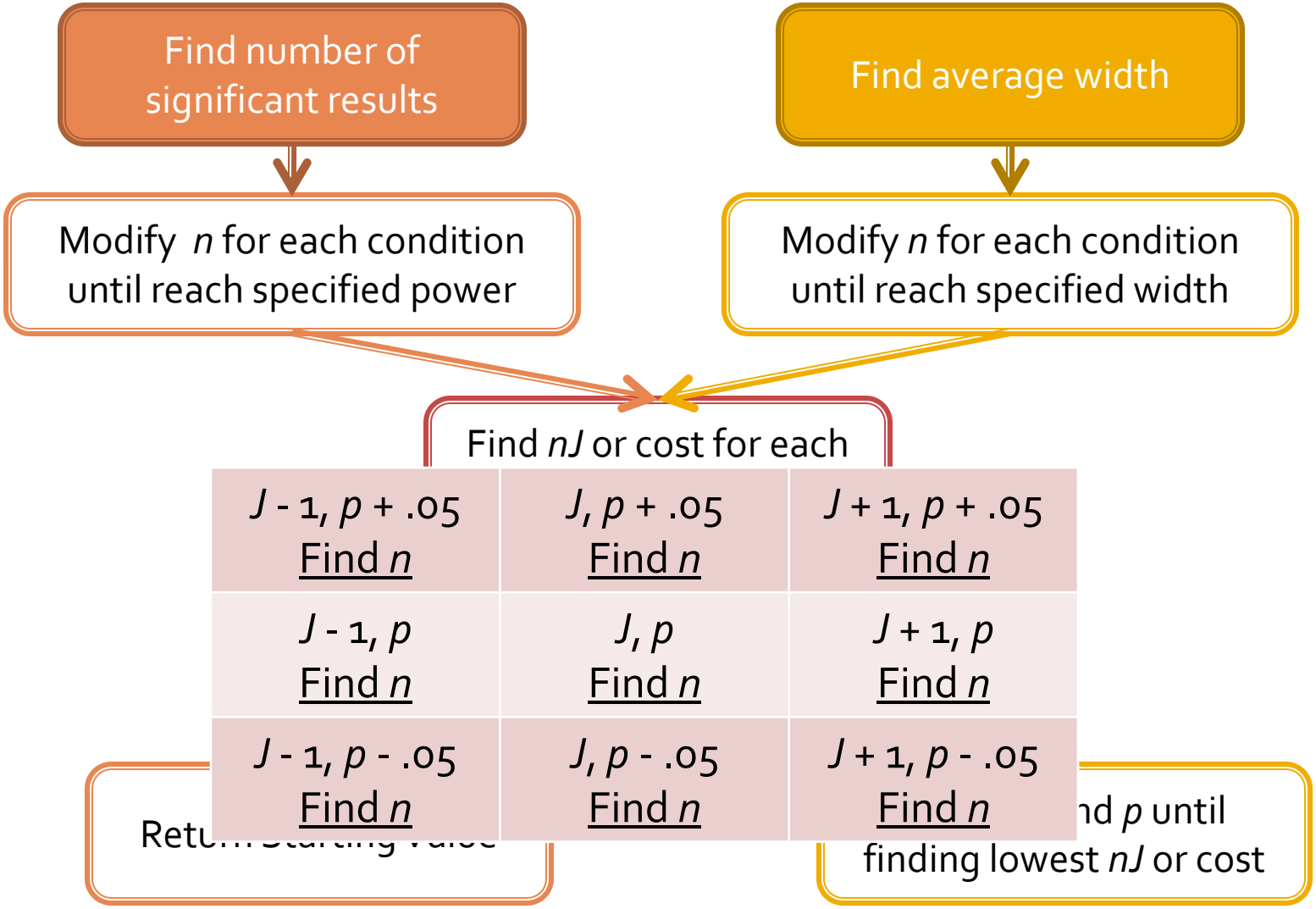
- Find the combination of *n*, *J*, *p* which
  - Criterion 1: lowest *nJ*
  - Criterion 2: lowest total cost from

Total Cost = pJ(**TGC** + (n x **TIC**)) + $(1-p)J$(**CGC** + (n x **CIC**)

Criterion 1 and 2: A Priori Monte Carlo Simulation



Criterion 1 and 2: A Priori Monte Carlo Simulation



Finding Sample Size: Criterion 3

- Since total cost is determined, we solve for various *n*, *J*, *p* by

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

- Find the combination of *n*, *J*, *p* which have highest power or lowest width
- Confirm result of power and width by running Mplus

Other Features

- Covariate
  - Intraclass correlation of covariate
  - Group effect and individual effect
- Degree of certainty in *CI* of *ES*

Program Illustration