Sample Size and Power Calculations

IPA/JPAL/CMF Training

Limuru, Kenya
28 July 2010

Owen Ozier
Department of Economics
University of California at Berkeley

Slides revised 14 September 2010

Thanks and Introduction

Thanks to everyone from JPAL/IPA who made this happen!

Thanks and Introduction

Thanks to everyone from JPAL/IPA who made this happen!

My background: randomized evaluations in Busia, Kenya.

Motivation

Motivation

Program evaluation:
bringing the scientific method to social science

Motivation

Program evaluation:
bringing the scientific method to social science

First steps:

Propose a hypothesis

Motivation

Program evaluation:
bringing the scientific method to social science

First steps:

Propose a hypothesis

Design an experiment to test the hypothesis

Motivation

Program evaluation:
bringing the scientific method to social science

First steps:

Propose a hypothesis

Design an experiment to test the hypothesis

This involves gathering data...

Motivation

Program evaluation:
bringing the scientific method to social science

First steps:

Propose a hypothesis

Design an experiment to test the hypothesis

This involves gathering data...
...but how much data will we need?

Usually a lot

How this will work

"Numerical data should be kept for eternity; it's great stuff."
- Glenn Stevens, Boston University

Outline

1. Motivation

Outline

Outline

Outline

1. Motivation
2. Probability basics
   - Coin tossing
3. Power calculation
   - Terminology/Concepts
   - The Basic Calculation
   - Clusters
   - Covariates
   - Details
4. Exercises

A hypothesis and a kind of test

A hypothesis and a kind of test

- "Null" Hypothesis: the coin is fair
  50% chance of heads, 50% chance of tails.

A hypothesis and a kind of test

- "Null" Hypothesis: the coin is fair
  50% chance of heads, 50% chance of tails.
- Structure of the data:
  Toss the coin a number of times, count heads.

A hypothesis and a kind of test

- "Null" Hypothesis: the coin is fair
  50% chance of heads, 50% chance of tails.
- Structure of the data:
  Toss the coin a number of times, count heads.
- The test:
  "Accept" hypothesis if within some distance of the mean under the null;
  "Reject" otherwise.

A hypothesis and a kind of test

- "Null" Hypothesis: the coin is fair
  50% chance of heads, 50% chance of tails.
- Structure of the data:
  Toss the coin a number of times, count heads.
- The test:
  "Accept" hypothesis if within some distance of the mean under the null;
  "Reject" otherwise.
- If we only had 4 tosses of the coin, what distance cutoffs could we use?

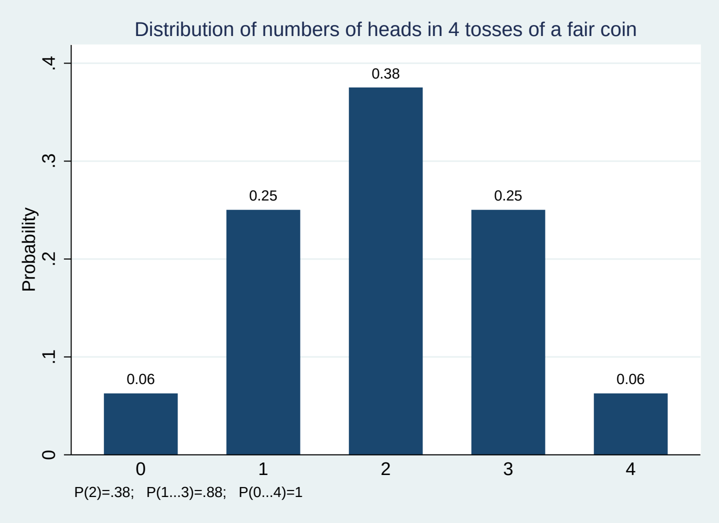
A hypothesis and a kind of test

- "Null" Hypothesis: the coin is fair
  50% chance of heads, 50% chance of tails.
- Structure of the data:
  Toss the coin a number of times, count heads.
- The test:
  "Accept" hypothesis if within some distance of the mean under the null;
  "Reject" otherwise.
- If we only had 4 tosses of the coin, what distance cutoffs could we use?
  Could accept (A) never,
  (B) when exactly the mean (2 heads),
  (C) when within 1 (1, 2, or 3 heads),
  or (D) always.

A hypothesis and a kind of test

- "Null" Hypothesis: the coin is fair
  50% chance of heads, 50% chance of tails.
- Structure of the data:
  Toss the coin a number of times, count heads.
- The test:
  "Accept" hypothesis if within some distance of the mean under the null;
  "Reject" otherwise.
- If we only had 4 tosses of the coin, what distance cutoffs could we use?
  Could accept (A) never,
  (B) when exactly the mean (2 heads),
  (C) when within 1 (1, 2, or 3 heads),
  or (D) always.
- We don't want to reject the null when it is true, though;
  How much accidental rejection would each possible cutoff give us?

Distribution of possible results



Not enough data.

Not enough data.

- There is no way* to create such a test with four coin tosses so that the chance of accidental rejection under the “null” hypothesis (sometimes written H_0) is less than 5%, a standard in social science.

Not enough data.

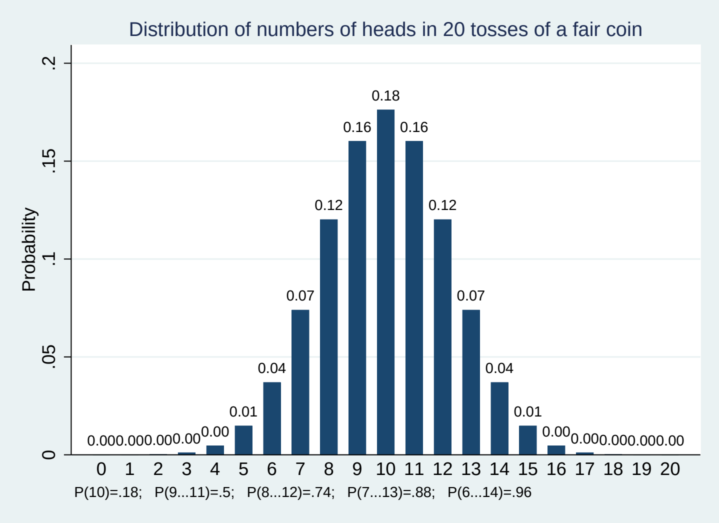
- There is no way* to create such a test with four coin tosses so that the chance of accidental rejection under the "null" hypothesis (sometimes written H_0) is less than 5%, a standard in social science.

    * *(Except the "never reject, no matter what" rule. Not very useful.)*

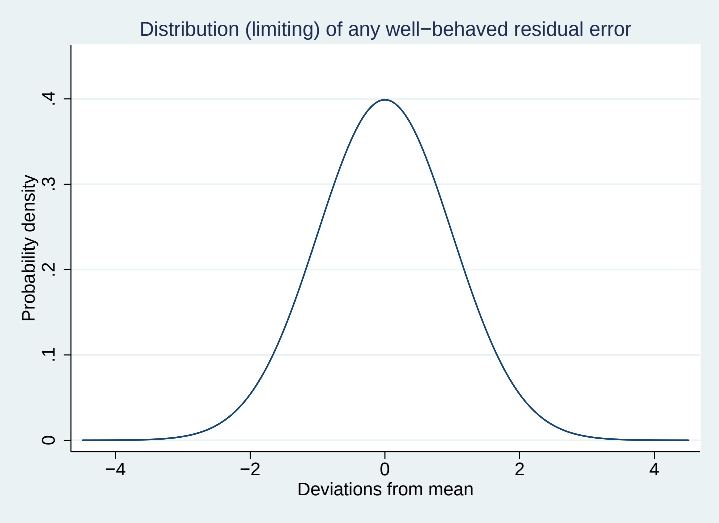
Not enough data.

- There is no way* to create such a test with four coin tosses so that the chance of accidental rejection under the "null" hypothesis (sometimes written H_0) is less than 5%, a standard in social science.
  * *(Except the "never reject, no matter what" rule. Not very useful.)*
- What about 20 coin tosses?

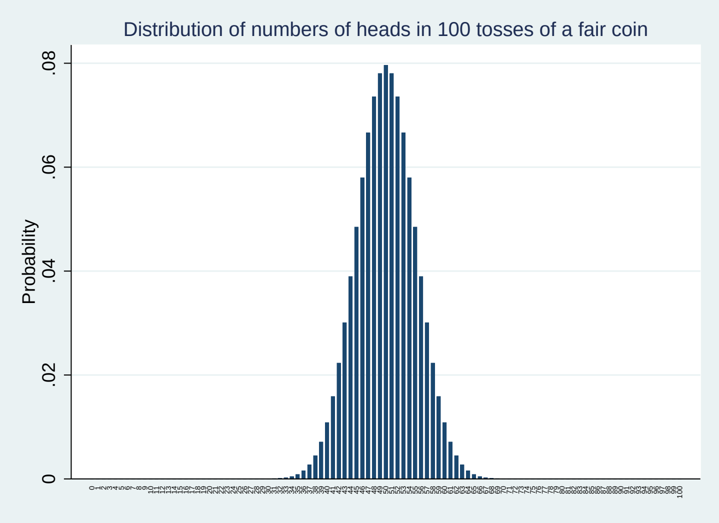
Distribution of possible results



The normal distribution



As sample size increases more:



Types of error

Test result

	"Reject Null," Find an effect!	"Accept Null," Conclude no effect.
Truth: There is an effect	Great!	"Type II Error" (low power)
Truth: There is NO effect	"Type I Error" (test size)	Great!

Types of error

	Test result	
	"Reject Null," Find an effect!	"Accept Null," Conclude no effect.
Truth: There is an effect	Great!	"Type II Error" (low power)
Truth: There is NO effect	"Type I Error" (test size)	Great!

The probability of Type I error is what we just discussed: the "size" of the test. By convention, we are usually interested in tests of "size" 0.05.

Types of error

Test result

	"Reject Null," Find an effect!	"Accept Null," Conclude no effect.
Truth: There is an effect	Great!	"Type II Error" (low power)
Truth: There is NO effect	"Type I Error" (test size)	Great!

The probability of Type I error is what we just discussed: the "size" of the test. By convention, we are usually interested in tests of "size" 0.05.

The probability of Type II error is also very important;
If P(failure to detect an effect) $= 1 - \kappa$, then the power of the test is κ.

Types of error

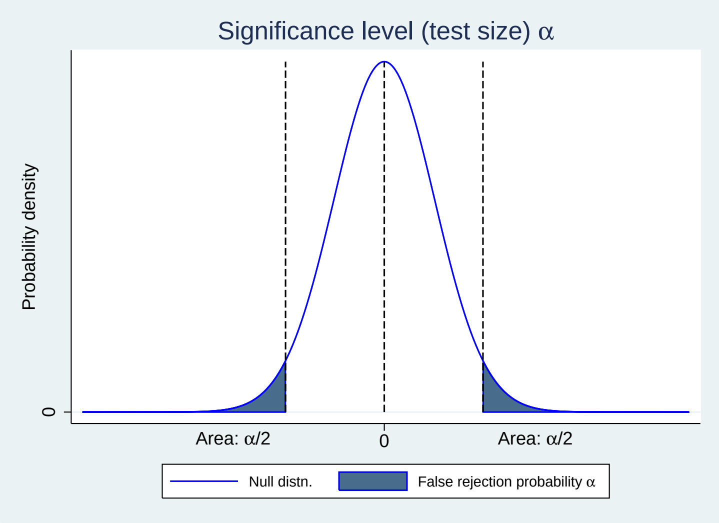
Test result

	"Reject Null," Find an effect!	"Accept Null," Conclude no effect.
Truth: There is an effect	Great!	"Type II Error" (low power)
Truth: There is NO effect	"Type I Error" (test size)	Great!

The probability of Type I error is what we just discussed: the "size" of the test. By convention, we are usually interested in tests of "size" 0.05.

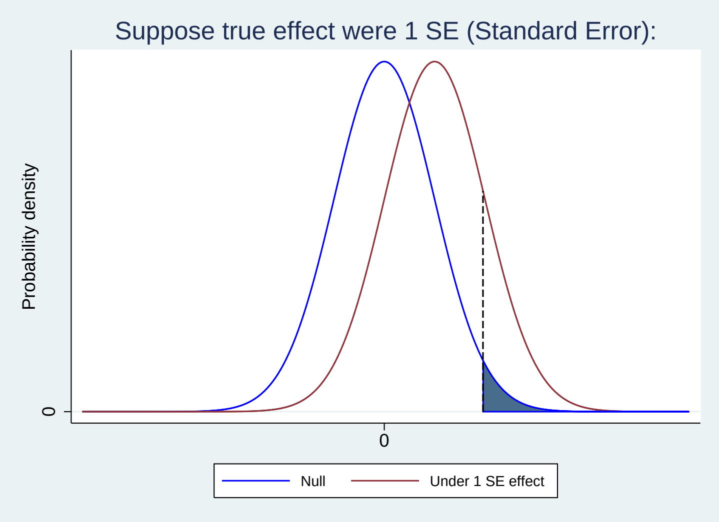
The probability of Type II error is also very important;
If P(failure to detect an effect) $= 1 - \kappa$, then the power of the test is κ.
Power depends on anticipated effect size; typical desired power is 80% or higher.

Rejecting H_0 in critical region



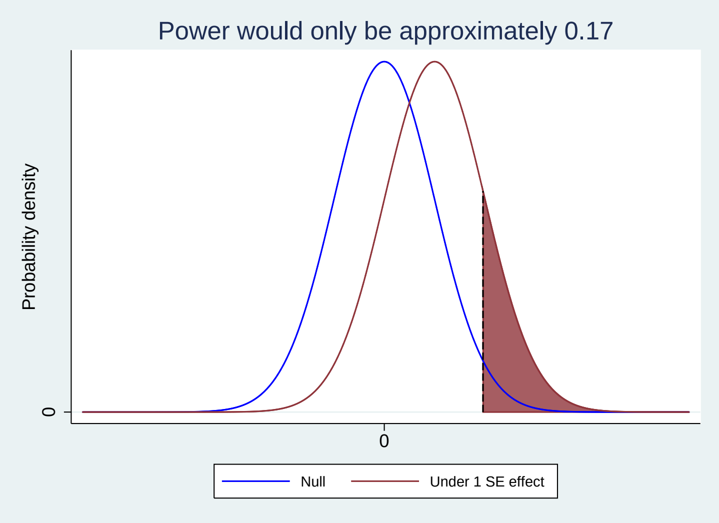
Note: two-sided test.

Under an alternative:



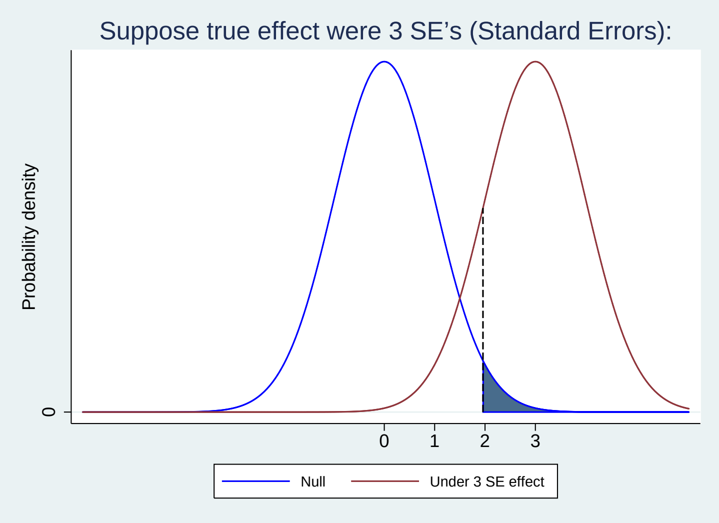
Note: two-sided test.

Under an alternative:



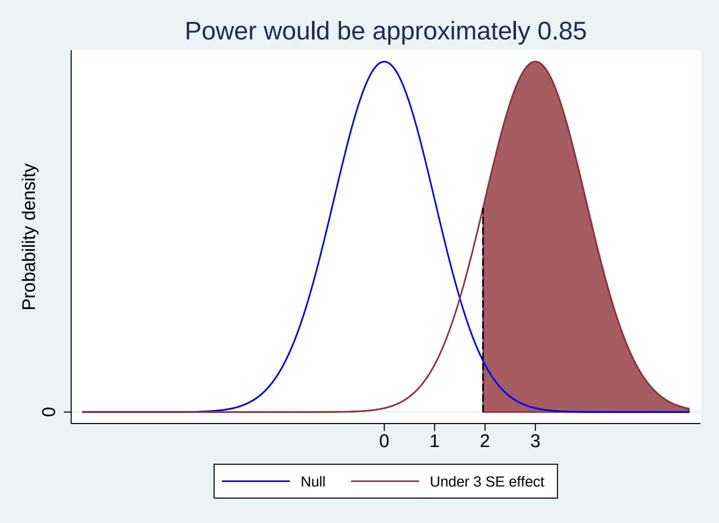
Note: two-sided test.

Under an alternative:



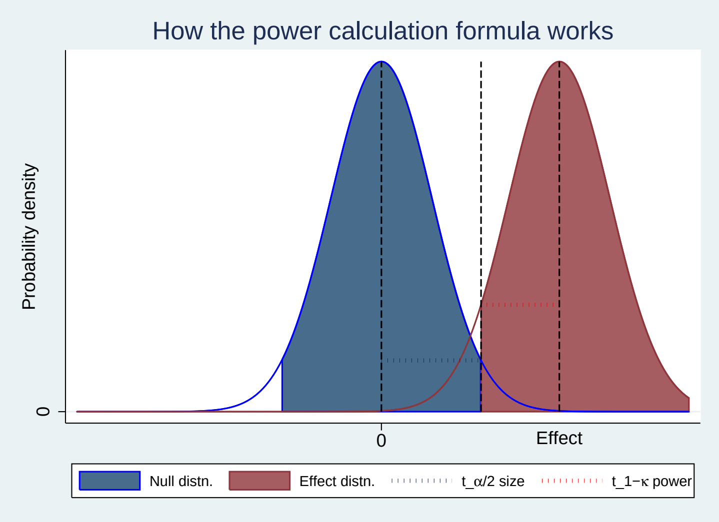
Note: this example follows slides by Marc Shotland (JPAL).

Under an alternative:



Note: this example follows slides by Marc Shotland (JPAL).

Power calculation, visually



Note: see the related figure in the *Toolkit* paper.

The formula: for power κ and size α,

$Effect > (t_{1-\kappa} + t_{\alpha/2})SE(\hat{\beta})$ Notation: $t_{1-p} = p^{th}$ percentile of the t dist'n.

The formula: for power κ and size α,

$Effect > (t_{1-\kappa} + t_{\alpha/2}) SE(\hat{\beta})$ Notation: $t_{1-p} = p^{th}$ percentile of the t dist'n.

$$MDE = (t_{1-\kappa} + t_{\alpha/2}) \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}}$$

The formula: for power κ and size α,

$Effect > (t_{1-\kappa} + t_{\alpha/2})SE(\hat{\beta})$ Notation: $t_{1-p} = p^{th}$ percentile of the t dist'n.

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \approx (z_{1-\kappa} + z_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

In practice (Stata): **sampsi**

Note: Stata uses the normal rather than t distribution (avoiding the D.O.F. issue).

The formula: for power κ and size α,

$Effect > (t_{1-\kappa} + t_{\alpha/2})SE(\hat{\beta})$ Notation: $t_{1-p} = p^{th}$ percentile of the t dist'n.

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \approx (z_{1-\kappa} + z_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

In practice (Stata): **sampsi**

Note: Stata uses the normal rather than t distribution (avoiding the D.O.F. issue).

Where do these numbers come from, σ^2 and the effect size? Two basic options:

The formula: for power κ and size α,

$Effect > (t_{1-\kappa} + t_{\alpha/2})SE(\hat{\beta})$ Notation: $t_{1-p} = p^{th}$ percentile of the t dist'n.

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \quad \approx \quad (z_{1-\kappa} + z_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

In practice (Stata): **sampsi**

Note: Stata uses the normal rather than t distribution (avoiding the D.O.F. issue).

Where do these numbers come from, σ^2 and the effect size? Two basic options:

- Consider standardized effect sizes in terms of standard deviations

The formula: for power κ and size α,

$Effect > (t_{1-\kappa} + t_{\alpha/2})SE(\hat{\beta})$ Notation: $t_{1-p} = p^{th}$ percentile of the t dist'n.

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \approx (z_{1-\kappa} + z_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

In practice (Stata): **sampsi**

Note: Stata uses the normal rather than t distribution (avoiding the D.O.F. issue).

Where do these numbers come from, σ^2 and the effect size? Two basic options:

- Consider standardized effect sizes in terms of standard deviations
- Draw on existing data: What is available that could inform your project?

What if treatment is assigned by groups?

We have been thinking here of randomizing at the individual level.
But in practice, we often randomize larger units.

What if treatment is assigned by groups?

We have been thinking here of randomizing at the individual level.
But in practice, we often randomize larger units.
Examples:

- **Entire schools** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual pupil
- **Classes within a school** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual pupil
- **Households** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual family member
- **Sub-district locations** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual road
- **Microfinance branch offices** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual borrower

What if treatment is assigned by groups?

We have been thinking here of randomizing at the individual level.
But in practice, we often randomize larger units.
Examples:

- **Entire schools** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual pupil
- **Classes within a school** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual pupil
- **Households** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual family member
- **Sub-district locations** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual road
- **Microfinance branch offices** are assigned to treatment or comparison groups, and we observe outcomes at the level of the individual borrower

What does this do?
It depends on how much variation is explained by the group each individual is in.

Intuitive example

You want to know how close the upcoming Kenyan referendum will be.

Intuitive example

You want to know how close the upcoming Kenyan referendum will be.

Method 1: Randomly select 50 people from entire Kenyan population

Intuitive example

You want to know how close the upcoming Kenyan referendum will be.

Method 1: Randomly select 50 people from entire Kenyan population

Method 2: Randomly select 5 families, and ask ten members of each extended family their opinion

The formula

Scale the effective standard error by:

$$DesignEffect = \sqrt{1 + (n_{groupsize} - 1)\rho}$$

ρ ("rho") is the intra-class correlation.

The formula

Scale the effective standard error by:

$$DesignEffect = \sqrt{1 + (n_{groupsize} - 1)\rho}$$

ρ ("rho") is the intra-class correlation.
In practice (Stata): **loneway** and **sampclus**

The formula

Scale the effective standard error by:

$$DesignEffect = \sqrt{1 + (n_{groupsize} - 1)\rho}$$

ρ ("rho") is the intra-class correlation.
In practice (Stata): **loneway** and **sampclus**

Recall earlier formula:

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

The formula

Scale the effective standard error by:

$$DesignEffect = \sqrt{1 + (n_{groupsize} - 1)\rho}$$

ρ ("rho") is the intra-class correlation.
In practice (Stata): **loneway** and **sampclus**

Recall earlier formula:

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

But where does this ρ number come from? Two basic options:

The formula

Scale the effective standard error by:

$$DesignEffect = \sqrt{1 + (n_{groupsize} - 1)\rho}$$

ρ ("rho") is the intra-class correlation.
In practice (Stata): **loneway** and **sampclus**

Recall earlier formula:

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

But where does this ρ number come from? Two basic options:

- Consider what might be reasonable assumptions

The formula

Scale the effective standard error by:

$$DesignEffect = \sqrt{1 + (n_{groupsize} - 1)\rho}$$

ρ ("rho") is the intra-class correlation.
In practice (Stata): **loneway** and **sampclus**

Recall earlier formula:

$$MDE = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

But where does this ρ number come from? Two basic options:

- Consider what might be reasonable assumptions
- Draw on existing data (again):
  What is available that could inform your project?

Intra-class correlations we have known

Data source	**ICC (ρ)**
Madagascar Math + Language	0.5
Busia, Kenya Math + Language	0.22
Udaipur, India Math + Language	0.23
Mumbai, India Math + Language	0.29
Vadodara, India Math + Language	0.28
Busia, Kenya Math	0.62

Source: Marc Shotland (JPAL) slides

Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.

Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.
- Baseline survey data on the outcome of interest is especially useful when the within-individual correlation of outcomes (in the absence of treatment) is high. Consider academic test scores, for example.

Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.
- Baseline survey data on the outcome of interest is especially useful when the within-individual correlation of outcomes (in the absence of treatment) is high. Consider academic test scores, for example.
- However, it may risk contaminating an otherwise clean randomized design. *(randomizing was supposed to avoid omitted variable biases in OLS!)*

Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.
- Baseline survey data on the outcome of interest is especially useful when the within-individual correlation of outcomes (in the absence of treatment) is high. Consider academic test scores, for example.
- However, it may risk contaminating an otherwise clean randomized design. *(randomizing was supposed to avoid omitted variable biases in OLS!)*
- Stratifying is a good option:

Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.
- Baseline survey data on the outcome of interest is especially useful when the within-individual correlation of outcomes (in the absence of treatment) is high. Consider academic test scores, for example.
- However, it may risk contaminating an otherwise clean randomized design. *(randomizing was supposed to avoid omitted variable biases in OLS!)*
- Stratifying is a good option:
    - On baseline outcome values

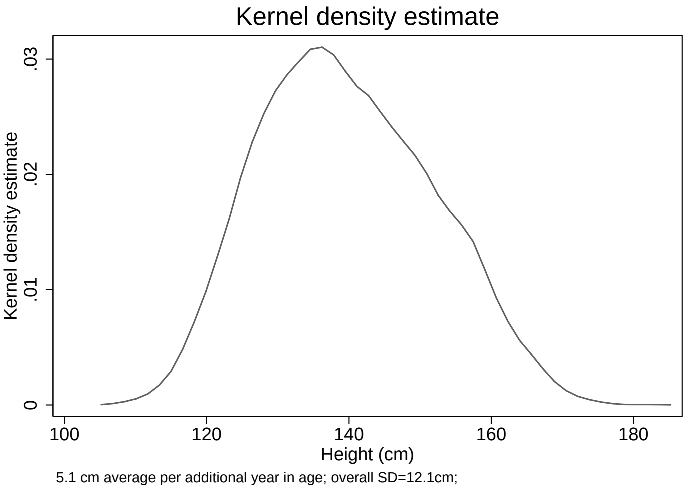
Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.
- Baseline survey data on the outcome of interest is especially useful when the within-individual correlation of outcomes (in the absence of treatment) is high. Consider academic test scores, for example.
- However, it may risk contaminating an otherwise clean randomized design. *(randomizing was supposed to avoid omitted variable biases in OLS!)*
- Stratifying is a good option:
  - On baseline outcome values
  - On fixed observables that have high predictive power (parents' education, age, ...)

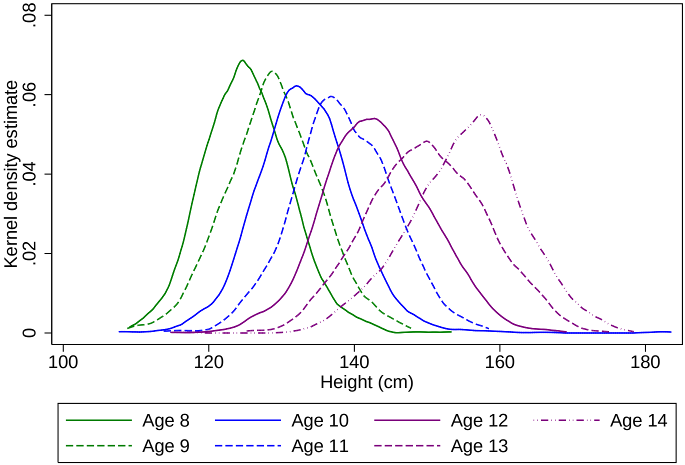
Baseline and other covariate data

- Controlling for covariates reduces the SE, so increases power.
- Baseline survey data on the outcome of interest is especially useful when the within-individual correlation of outcomes (in the absence of treatment) is high. Consider academic test scores, for example.
- However, it may risk contaminating an otherwise clean randomized design. *(randomizing was supposed to avoid omitted variable biases in OLS!)*
- Stratifying is a good option:
    - On baseline outcome values
    - On fixed observables that have high predictive power (parents' education, age, ...)
    - On observables that delineate subpopulations you may want to test within (you might want to do another power calculation on this subsample)

Power calculation, visually



Power calculation, visually



5.1 cm average per additional year in age; overall SD=12.1cm; within SD=7.21cm

Some things to watch out for

- Imperfect compliance with treatment:

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)
- How randomization is done:

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)
- How randomization is done:
  - Alternative test based on statistical work of Fisher in early 20th century

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)
- How randomization is done:
  - Alternative test based on statistical work of Fisher in early 20^{th} century
  - Issues with 1,2,3,1,2,3... if alphabetical, not a problem if order is random

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)
- How randomization is done:
  - Alternative test based on statistical work of Fisher in early 20^{th} century
  - Issues with 1,2,3,1,2,3... if alphabetical, not a problem if order is random
  - (recently: McKenzie and Bruhn)

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)
- How randomization is done:
  - Alternative test based on statistical work of Fisher in early 20^{th} century
  - Issues with 1,2,3,1,2,3... if alphabetical, not a problem if order is random
  - (recently: McKenzie and Bruhn)
- Much more: check out Duflo, Glennerster, and Kremer:
  Using randomization in development economics research: a toolkit

Some things to watch out for

- Imperfect compliance with treatment:
  - The most straightforward approach is to decrease the effect size.
- Multiple treatments:
  - Optimal treatment structure depends on several things;
    May be better to have control group larger than any of the treatment groups.
  - To get started, calculate power for the smallest effect you hope to test.
  - Multiple test correction may be needed later. (recently: Michael Anderson)
- How randomization is done:
  - Alternative test based on statistical work of Fisher in early 20^{th} century
  - Issues with 1,2,3,1,2,3... if alphabetical, not a problem if order is random
  - (recently: McKenzie and Bruhn)
- Much more: check out Duflo, Glennerster, and Kremer:
  Using randomization in development economics research: a toolkit
- *"A first comment is that, despite all the precision of these formulas, power calculations involve substantial guess work in practice."*

Exercises

First make sure you have the files:

- Make sure **sampclus** is installed; either by using **findit sampclus** or by copying the two files into **C:/ado/plus/s/** (typically).
- Various **.do** and **.dta** files

- type-I-error-reject-null-when-true.do
- type-II-error-fail-to-reject-when-alt-is-true.do
- sampsi-syntax.do
- icc-example.do
- SampsiExerciseB.do - here we need to adjust the directory in the .do file before running.

Exercises

First make sure you have the files:

- Make sure **sampclus** is installed; either by using **findit sampclus** or by copying the two files into **C:/ado/plus/s/** (typically).
- Various **.do** and **.dta** files

- type-I-error-reject-null-when-true.do
- type-II-error-fail-to-reject-when-alt-is-true.do
- sampsi-syntax.do
- icc-example.do
- SampsiExerciseB.do -
  here we need to adjust the directory in the .do file before running.

Exercises

First make sure you have the files:

- Make sure **sampclus** is installed; either by using **findit sampclus** or by copying the two files into **C:/ado/plus/s/** (typically).
- Various **.do** and **.dta** files

- type-I-error-reject-null-when-true.do
- type-II-error-fail-to-reject-when-alt-is-true.do
- sampsi-syntax.do
- icc-example.do
- SampsiExerciseB.do - here we need to adjust the directory in the .do file before running.

Exercises

First make sure you have the files:

- Make sure **sampclus** is installed; either by using **findit sampclus** or by copying the two files into **C:/ado/plus/s/** (typically).
- Various **.do** and **.dta** files

- type-I-error-reject-null-when-true.do
- type-II-error-fail-to-reject-when-alt-is-true.do
- sampsi-syntax.do
- icc-example.do
- SampsiExerciseB.do -
  here we need to adjust the directory in the .do file before running.

Exercises

First make sure you have the files:

- Make sure **sampclus** is installed; either by using **findit sampclus** or by copying the two files into **C:/ado/plus/s/** (typically).
- Various **.do** and **.dta** files

- type-I-error-reject-null-when-true.do
- type-II-error-fail-to-reject-when-alt-is-true.do
- sampsi-syntax.do
- icc-example.do
- SampsiExerciseB.do -
  here we need to adjust the directory in the .do file before running.

Exercises

First make sure you have the files:

- Make sure **sampclus** is installed; either by using **findit sampclus** or by copying the two files into **C:/ado/plus/s/** (typically).
- Various **.do** and **.dta** files

- type-I-error-reject-null-when-true.do
- type-II-error-fail-to-reject-when-alt-is-true.do
- sampsi-syntax.do
- icc-example.do
- SampsiExerciseB.do -
  here we need to adjust the directory in the .do file before running.