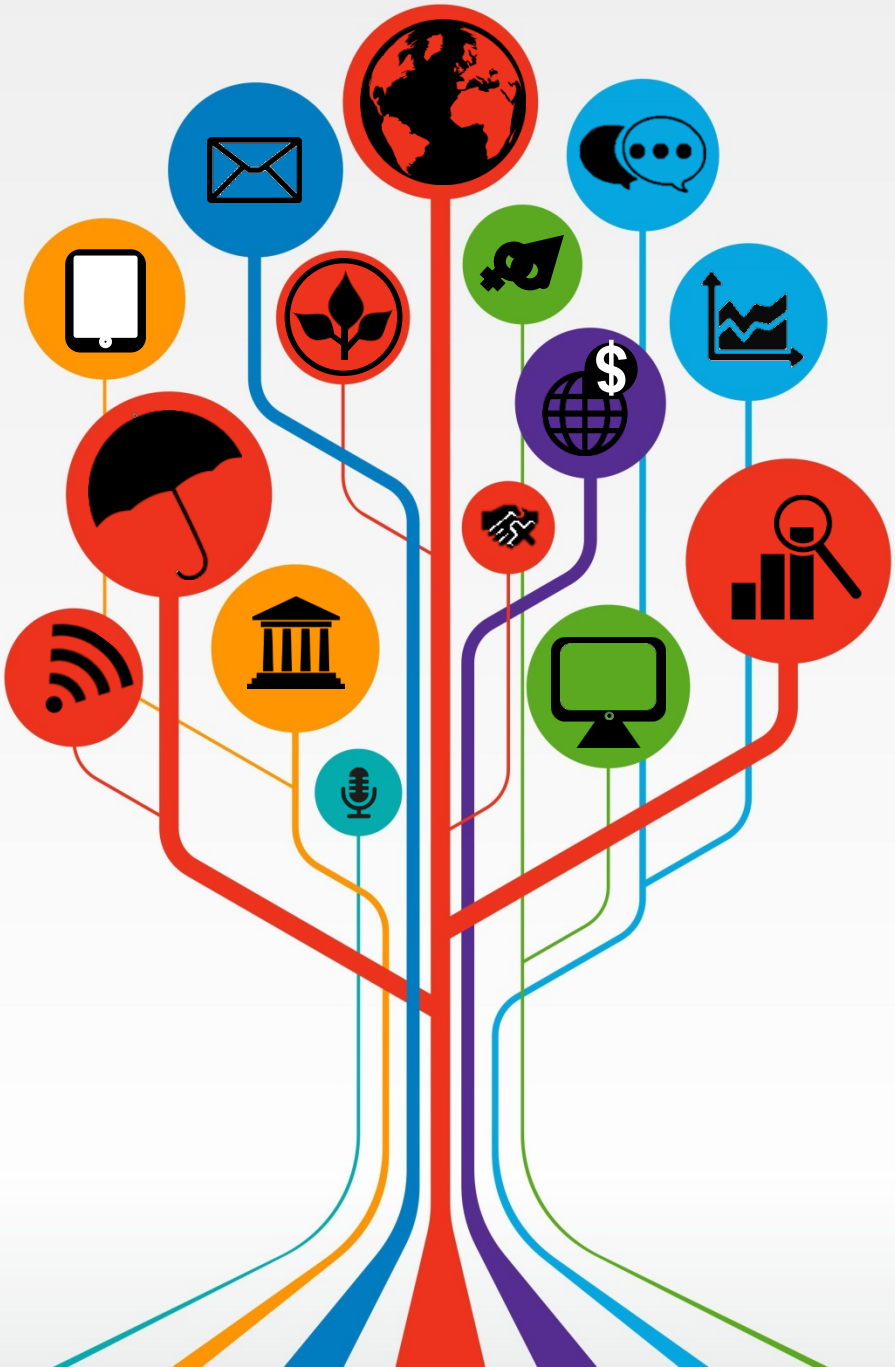
FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC



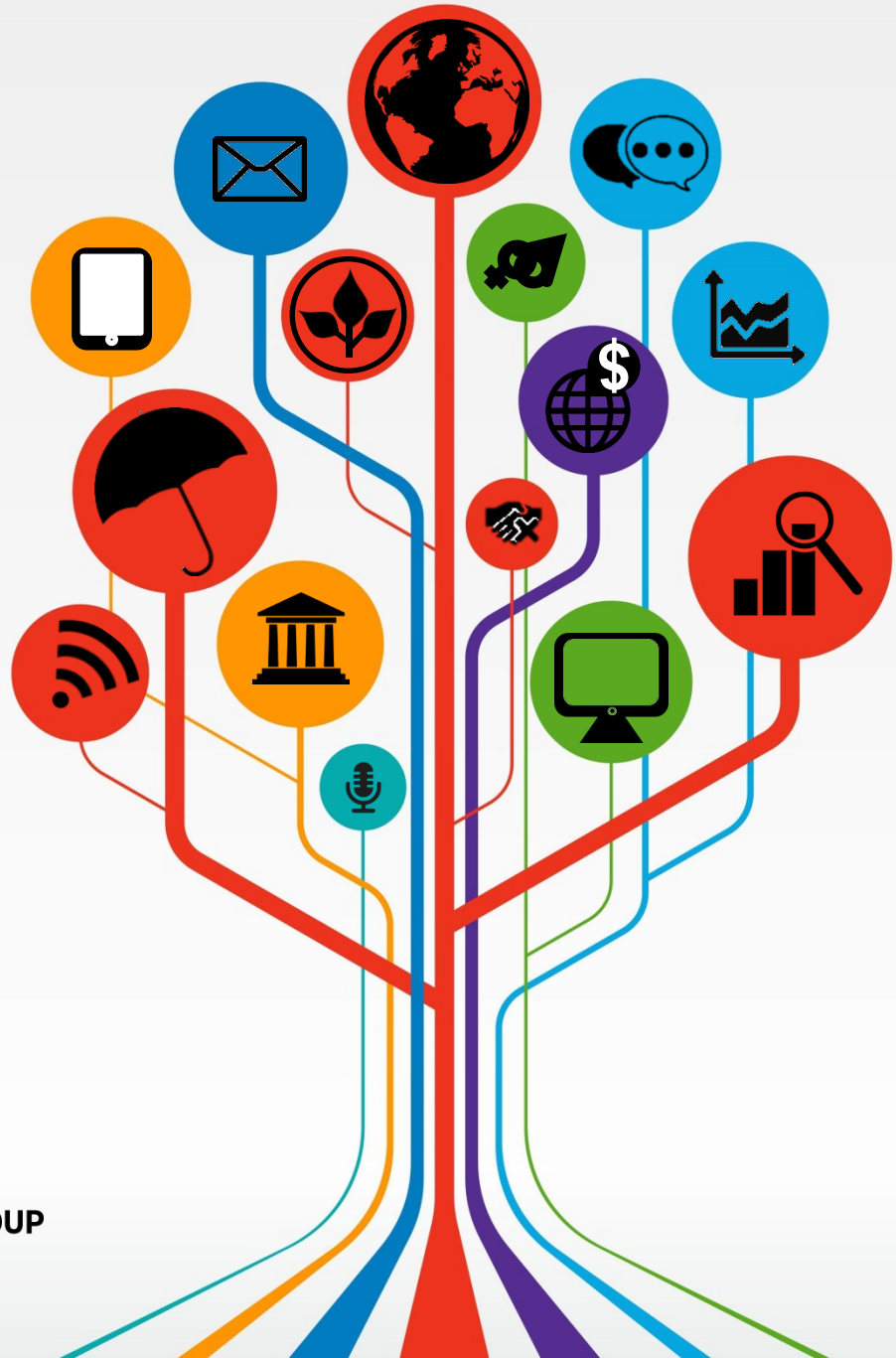
i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GR

Sampling: Track 2

Maria Jones & Roshni Khincha

21 June 2018



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

outline

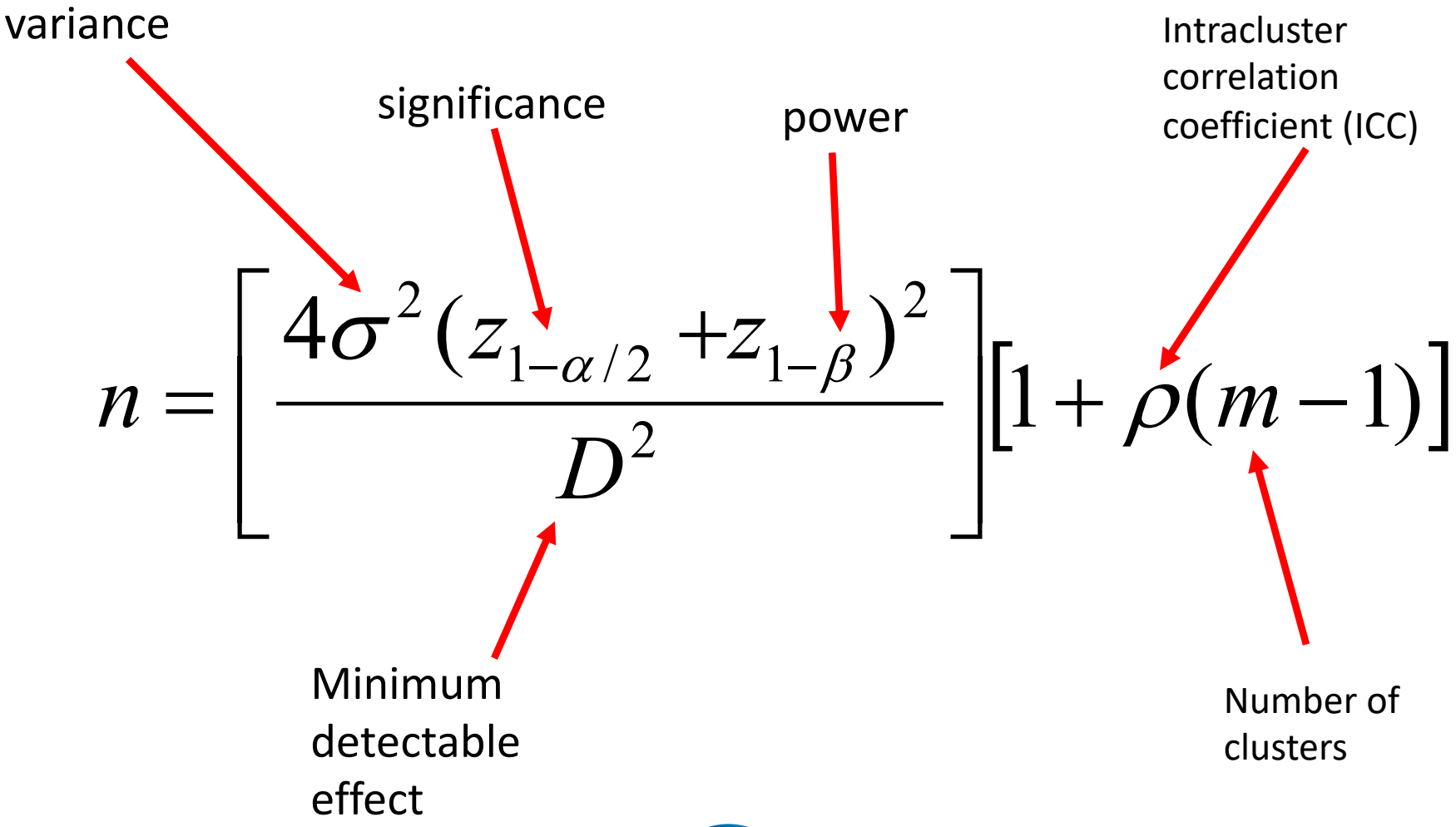
1. Sample size calculations: key parameters
   - Choose your own adventure!
   - Quick quizzes, we'll spend more time where there are knowledge gaps

2. Sampling in challenging contexts: 2 case studies
   - Market and trader survey
   - Farm survey in an irrigation scheme

3. Programming power calculations: Stata options
   - For your reference as an appendix

Sample Size Calculations -
Key Parameters to Understand

sample size equation



types of power calcs

- Three options
  - Compute sample size given power and MDES
  - Compute power given sample size and effect size
  - Compute MDES given power and sample size

- For IE, typically assume power of 80% and solve for either sample size or MDES
  - Most often, take sample size as given (based on IE design / population / budget) and solve for *MDES*
  - If MDES too large, useful to reverse – put in largest acceptable MDES and solve for sample size

power and confidence

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] \left[1 + \rho(m-1) \right]$$

significance and confidence

- type I error: false positive
  - detect an effect when no effect is present
    - e.g. result indicates a treatment has an effect when in truth it has no effect
- significance level: α (alpha)
  - probability of a type I error
- confidence level : (1 – α)
  - probability that we do not find a statistically significant effect if the treatment effect is zero
- for power calcs assume 95% confidence

power

- type II error: false negative
  - fail to detect an effect when an effect is present
  - e.g. a result that indicates that a treatment has no effect when in truth it has an effect
- β (beta) is the likelihood of making a type II error
- power: (1 – β)
  - probability of correctly rejecting H_0
- for power calcs assume 80% power

effect size

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] \left[1 + \rho(m-1) \right]$$

Effect size

- What is minimum detectable effect?

- How should you determine a reasonable MDES?

MDES

- D: the smallest effect size that, if it were any smaller, the intervention would not be worth the effort
  - a.k.a **Minimum Detectable Effect Size** (MDES)
- The smaller the effect you want to be able to detect, the larger the sample you will need
  - larger sample → more precise measuring device
- Very common to solve for MDES when doing IE power calculations

QUIZ

Indicators	Core Sector Indicator	Core GAFSP Indicator	Unit	Base-line	Cumulative Target Value				
					PY1	PY2	PY3	PY4	PY5
Project Development Objective – Outcomes:									
1. Improved technologies (crop and livestock) released for project area farmers	☒	☐	No.	0	2	11	25	29	29
2. Increased productivity of	☐	☐	% over BL						
• Crops[5]				BL7			15%		30%
• Livestock[6]				BL 8			50%		75%

Baseline values: paddy: 2.6 tons/ha; wheat: 1.9 tons/ha; maize: 2.2 tons/ha; potato: 12.9 tons/ha

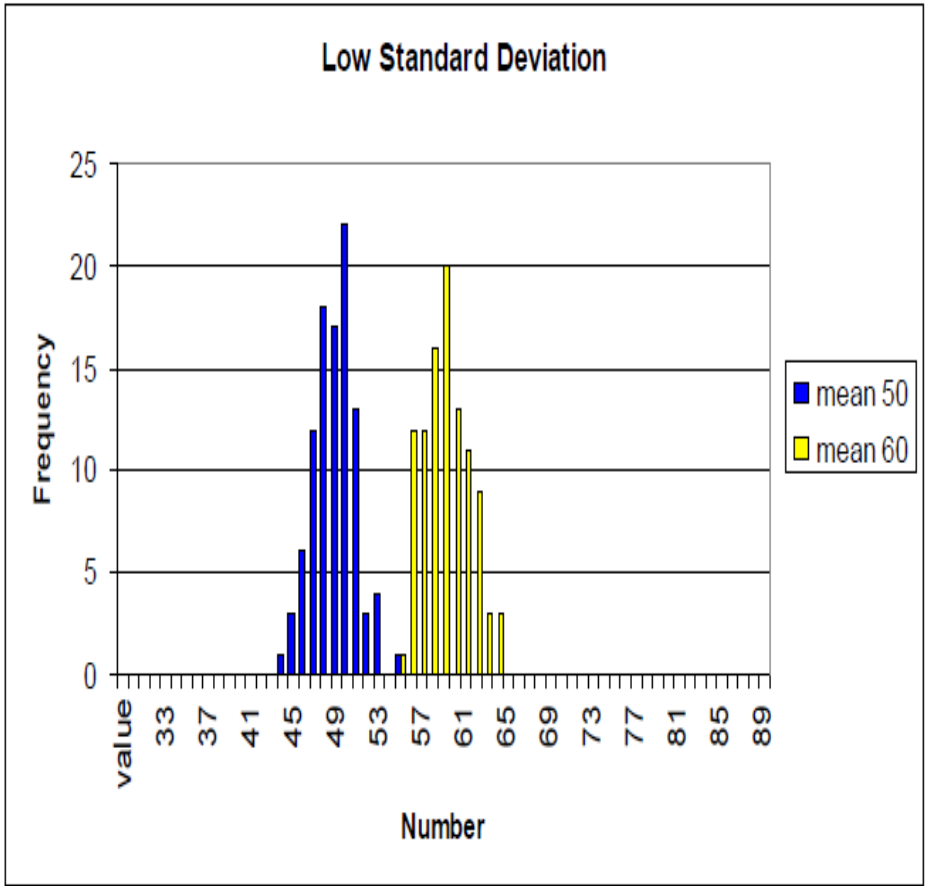
The impact evaluation focuses on improving crop productivity through farmer field schools. Based on the above results framework, how would you think about deciding a reasonable MDES?

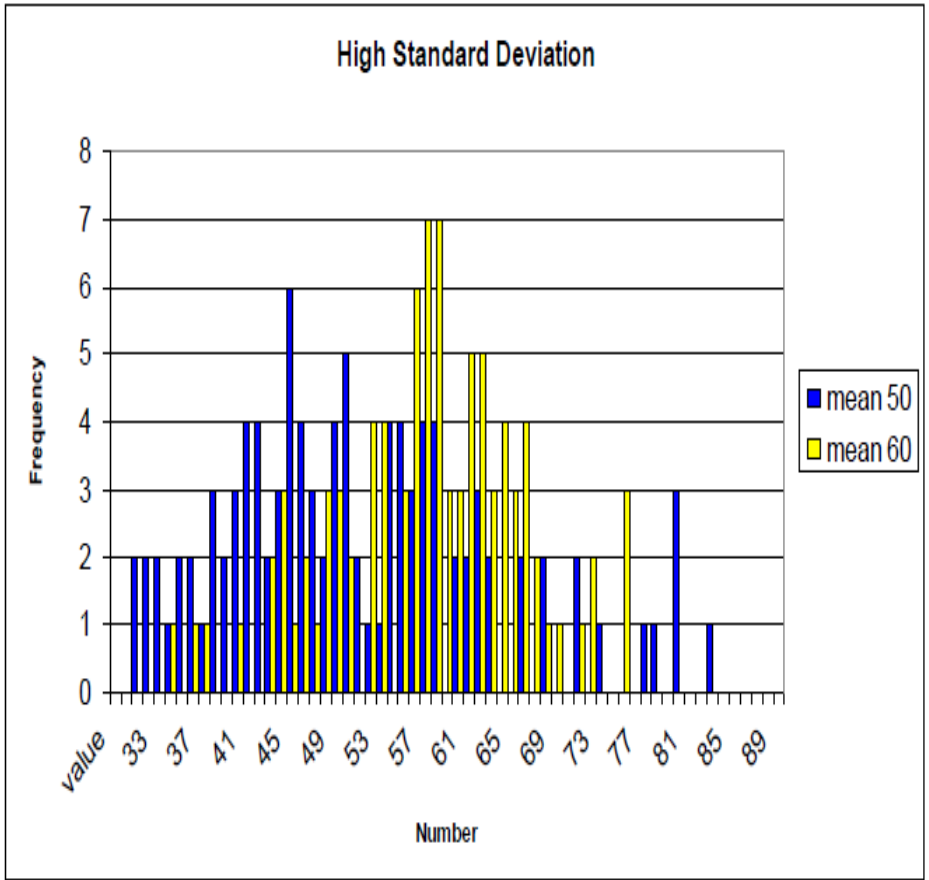
variance of outcomes

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] \left[1 + \rho(m-1) \right]$$

QUIZ

An intervention increases employment by 10% for treatment group on average in two different populations. Would you expect a difference in sample size needed to detect the effect in the two populations?





variance of outcomes

- σ = variance of the outcome of interest for the study population

- More underlying variance (heterogeneity)
  - → more difficult to detect difference
  - → need larger sample size

- **Tricky:** How do we know about heterogeneity *before* we decide our sample size and collect our data?
  - Ideal: pre-existing data … but often non-existent
  - Can use pre-existing data from a *similar* population
    - Example: LSMS, data routinely collected by govt, satellite imagery
  - Common sense

clustering (aka “design effect”)

$$n = \left[\frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{D^2} \right] \left[1 + \rho(m-1) \right]$$

QUIZ

Which sampling strategy is likely to give you more statistical power?

A. 400 classrooms, 5 students per classroom = 2,000 students
B. 50 classrooms, 40 students per classroom = 2,000 students
C. Both should give you similar statistical power
D. Don’t know

clustering

- Unit for sample size calculation depends on both:
  - Level of intervention AND
  - Level of measured impacts
- Example: intervention at village level, interested in impacts at HH level
  - Randomly assign villages to treatment / control
  - Sample household within villages

clustering

- Level of intervention (“cluster”) most important for sample size calculation

- If few clusters, precision will be limited, regardless of number of HHs sampled

QUIZ

Which sampling strategy is likely to give you more statistical power?

A. 100 villages, 5 HHs per village = 2,000 HHs
B. 100 villages, 50 HHs per village = 2,000 HHs
C. Both should give you similar statistical power
D. Don’t know

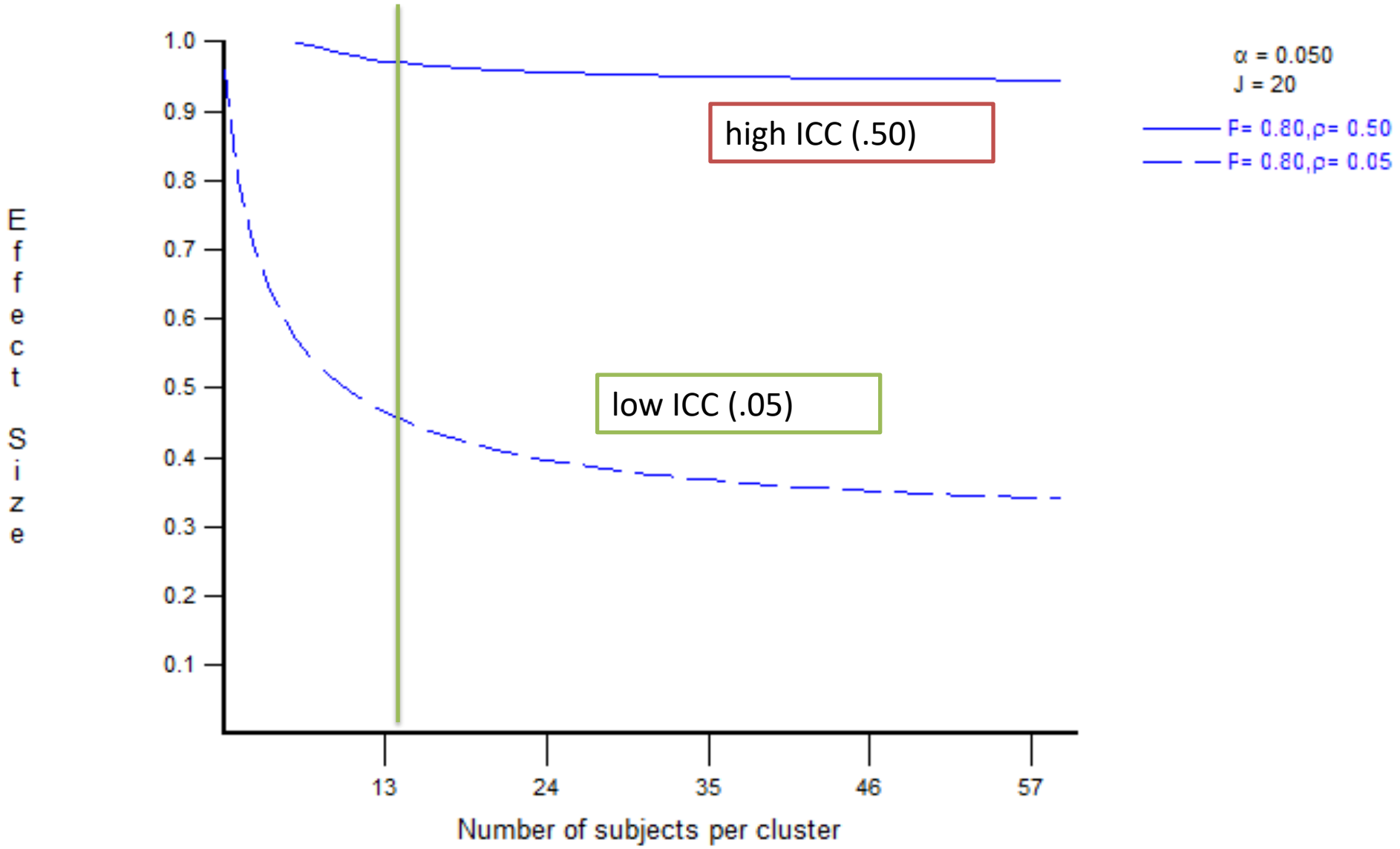
clustering

- **Intracluster correlation (ICC):** similarity of units within clusters

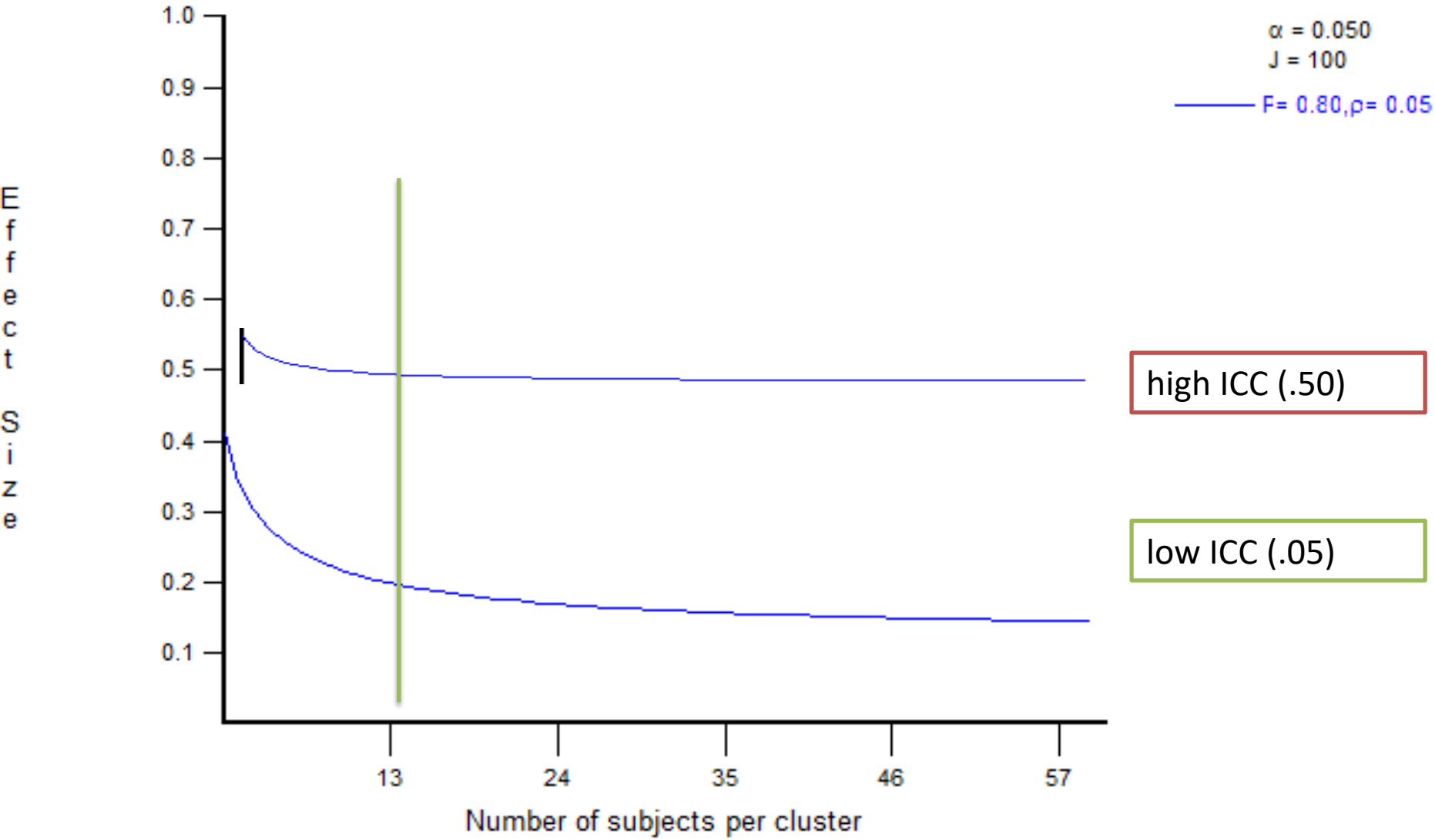
- Is the variation in outcome of interest coming mostly from differences *within* villages (low ICC), or *between* villages (high ICC)?
  - If HHs in village A are similar to each other, but different from HHs in village B, high ICC
  - If HHs in village A are similar to HHs in village B, low ICC

- If ICC = 0, no design effect

20 clusters


α = 0.050
J = 20
F= 0.80,ρ= 0.50
F= 0.80,ρ= 0.05
high ICC (.50)
low ICC (.05)
Effect Size
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
13
24
35
46
57
Number of subjects per cluster

100 clusters



clustering

Takeaway

High *intra-cluster correlation* (HHs in same cluster similar)

lower marginal value per extra sampled unit in the cluster

More clusters needed

Rule of thumb: <u>at least</u> 40 clusters per treatment arm

QUIZ

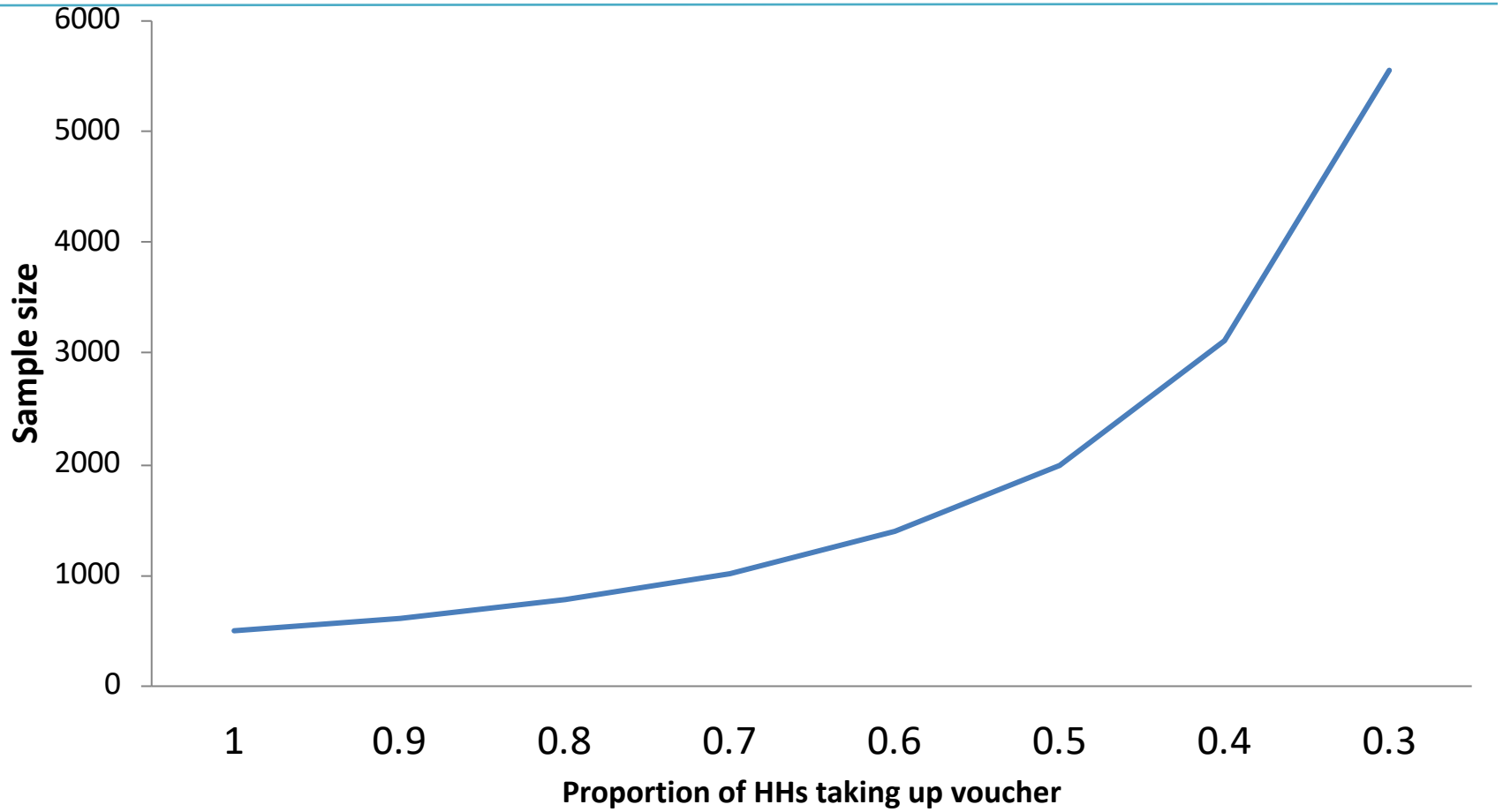
- You do power calculations and decide you need a sample of 1,000 HHs for an impact evaluation. The project starts. 6 months in, a monitoring survey shows that take-up of the intervention is 50%.
- What effect will this have on statistical power, given the sample size of 1,000 HHs?
- What could you do to improve power, if increasing the number of HHs is not feasible?

take-up

- Low take-up (rate) for intervention lowers precision
  - Effectively decreases sample size / increases minimum detectable effect
  - Can only detect an effect if it is really large

- Unfortunately, to account for take-up rate of 50%, have to increase sample size by factor of 4

Take up vs. sample size



Other factors

- Attrition
  - Effectively the same problem as take-up
  - Especially serious if cluster-level
- Compliance
  - If contamination (control also adopts) it will be more difficult to discern a meaningful difference
- Data quality
  - Missing observations: approximately = attrition
  - Measurement error → increased variance, less precision

conclusions

The smaller effects that we want to detect	**The larger the sample size has to be**
The more underlying heterogeneity (variance)	
The higher the level of clustering	
The lower take up	
The lower data quality	

Now you know how many people to sample ….

How do you identify them?

Sampling in practice

- Best case scenario: complete sampling frame already exists
- Most often this is not the case as impact evaluation samples are quite specific
- First step is usually to conduct a listing, then sample
- However, that may not be entirely straightforward, as the two case studies show

Case Study 1 -

Market Listing & Trader Survey

Context

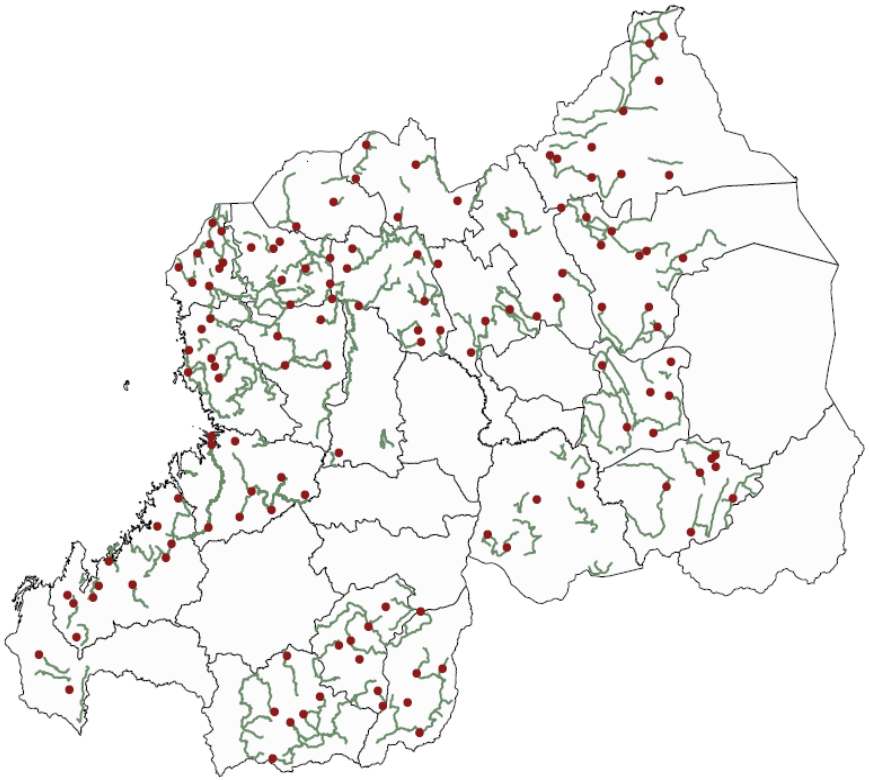
Rural Feeder Roads:

Does improved connectivity change lives?

Market survey - setup

- Understand how market structure and composition has changed over time through
  - a visual listing of all traders present in the market
  - conducting a short trader survey for a sub-sample of traders listed in each market

Market sample



Sampling for trader survey

- Based on power calculations and field practicalities, research team designed a sampling strategy in which the number of traders to survey per market depends on total market size

Market size	Sample size
<=30	100%
31-100	50%
101-400	33.3%
401-500	25%
501-600	20%
601-700	16.67%
701-800	14.28%
801-900	12.5%
901-1000	11.11%
...so on	

Sampling for trader survey

- The trader survey has to be conducted on the same day as the listing
  - for quicker completion as markets do not meet everyday
  - to avoid attrition / confusion as traders are identified by location, clothing, and type of goods, which will change between market days

How can the sample be dynamically selected as soon as listing is complete?

Selecting the sample – method 1

- Allow the enumerators to select the traders to interview

- Pros:
  - Easiest for the enumerator
- Cons:
  - Enumerators are likely to chose the traders they can find easily
  - No guarantee of representation of all types of traders

Selecting the sample – method 2

- Provide a walking skip pattern based on market size for enumerators to follow

Market size	Sample size	Skip pattern to follow
<=30	100%	every **trader** will be interviewed
31-100	50%	every **2nd trader** will be interviewed
101-400	33.3%	every **3rd trader** will be interviewed
401-500	25%	every **4th trader** will be interviewed
501-600	20%	every **5h trader** will be interviewed
601-700	16.67%	every **6th trader** will be interviewed
701-800	14.28%	every **7th trader** will be interviewed
801-900	12.5%	every **8th trader** will be interviewed
901-1000	11.11%	every **9th trader** will be interviewed
...so on		

Selecting the sample – method 2

- Pros
  - There is some form of randomness
  - All trader types are likely to be represented
- Cons:
  - Enumerators have to do a lot of mental math!
  - Very hard to verify whether sampling pattern was followed

Selecting the sample – method 3

- Rely on technology - Program the survey form to dynamically pick the traders to survey

- Pros
  - Enumerators just have to locate the stall listed on the tablet screen
  - All trader types are likely to be represented
- Cons:
  - Programming of the randomization might take time
  - Randomization is not replicable if done on SurveyCTO

What would you do?

What was actually done?

- Weighing the pros and cons of each available method, we left the selection to technology (method 3)
- Overestimated the required sample in each market to ensure required number of trader surveys were reached

Case Study 2 -

Irrigation Scheme Farmer Survey

Irrigation impact evaluation

sampling

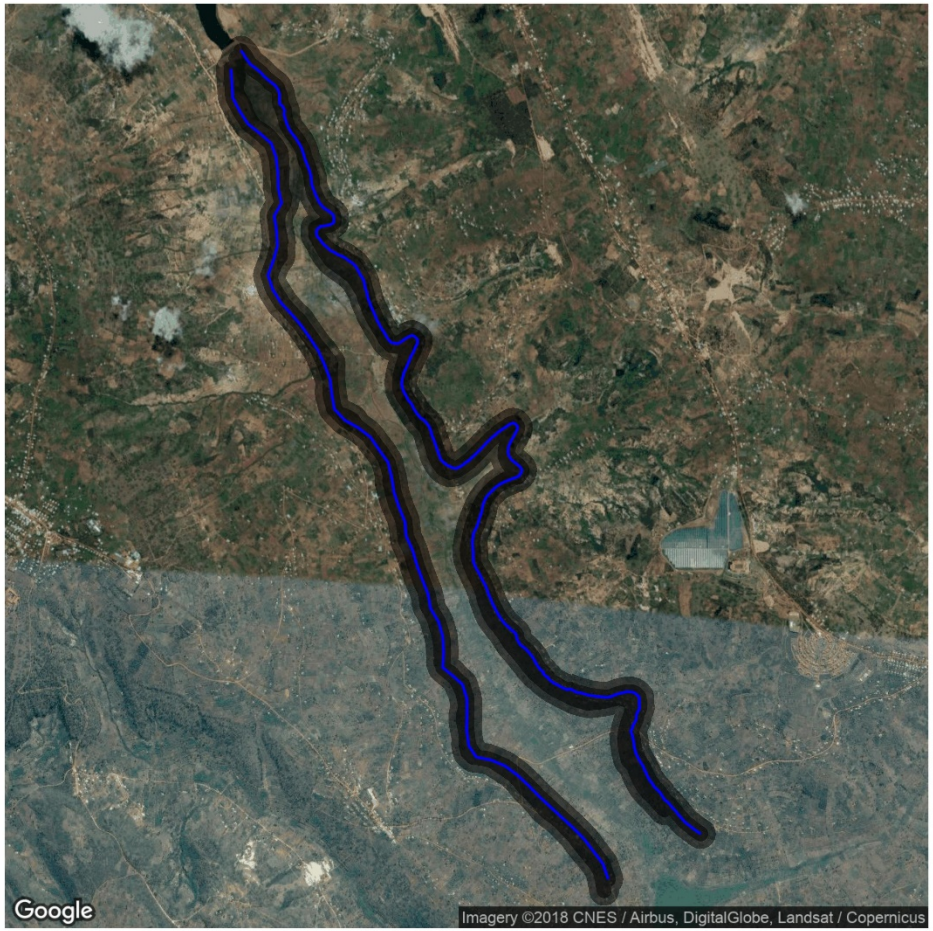
- Spatial regression discontinuity design
  - Compare plots just below irrigation canal to those just above
    - Therefore need to sample plots close to irrigation canal

- How to do that?
  - Listing HHs in the neighboring villages?
    - Plots aren't necessarily close to villages
    - People won't accurately be able to say whether the plot is within 50m of the canal

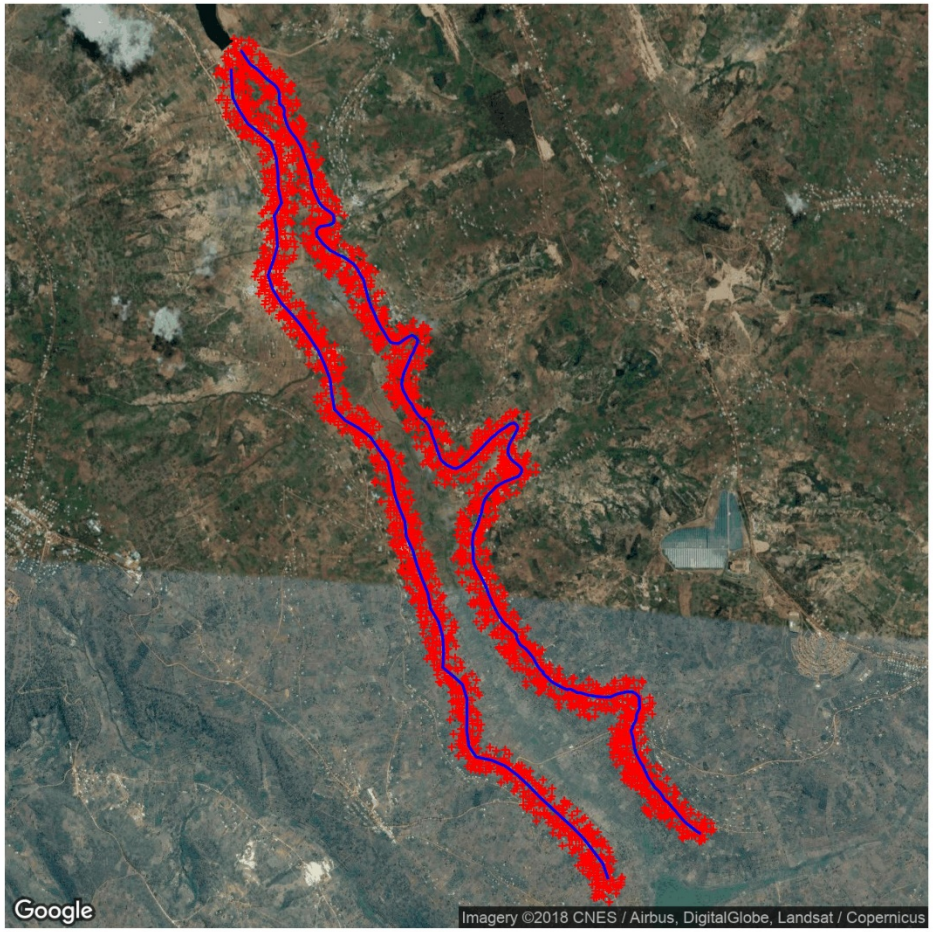
What did we do?

- Dropped uniform grid of points across full site at 2m resolution
- Randomly sampled points, excluding any point within 10m of a point selected
- Enumerators equipped with GPS units visited each sampled point to identify whether the point is agricultural land, and if so find out who cultivates it

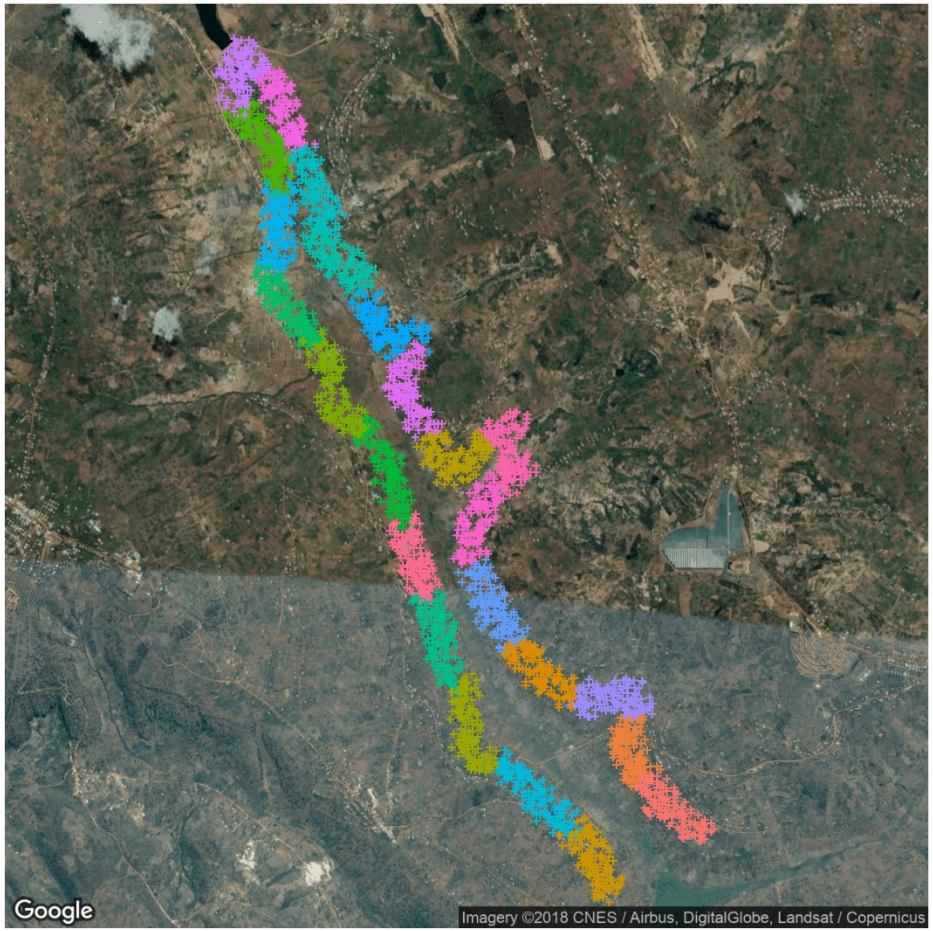
Draw 50m buffer



Drop 3000 points



Assign 24 enumerators to points



outcome

- From 3000 points, 2932 successfully visited and description recorded
- 1,058 distinct village name + cultivator combinations
- Contact village leaders to verify names and remove duplicate households (e.g. husband and wife) → 810 households
- 670 households successfully interviewed (more duplicates discovered during interviews, some names not recognized)
- Once plots were mapped, 76% have sample point within boundaries

APPENDIX

Programming Power Calculations - Summary of Stata Options

data you need to have on hand

- Mean and variance for outcome variable for population of interest
  - Assume mean and SD same for tmt and control if randomized

- Sample size (assuming you are calculating MDES (δ))
  - If individual randomization, number of people/units (*n*)
  - If clustered, number of clusters *(k),* number of units per cluster *(m),* intracluster correlation (ICC, ρ) and ideally, variation in cluster size (e.g. min-max cluster size)

- The following standard conventions
  - Significance level (α) = 0.05
  - Power = 0.80 (i.e. probability of type II error (β) = 0.20

ideally, you also know

- Baseline correlation of outcome with covariates
  - Covariates (individual and/or cluster level) reduce residual variance of the outcome variable, reducing required sample size
    - Reducing individual level residual variance is akin to increasing # obs per cluster (bigger effect if ICC low)
    - Reducing cluster level residual variance is akin to increasing # of clusters (bigger effect if ICC and *m* high)
  - If you have baseline data, this is easy to obtain
    - Including baseline autocorrelation will improve power (keep only time invariant portion of variance)
- Number of follow-up surveys
- Autocorrelation of outcome between FUP rounds
- Take-up for the intervention

But… what if I don't have this data?

- You will basically never have all of this for your exact population of interest when you first do power calculations

- So, use the best available data to estimate values for each parameter. Sources to consider
    - high-quality nationally representative survey (e.g. LSMS)
    - Data from DIME IE in same country (or region, if pressed)
    - Review the literature – especially published papers on the sector and country. Will certainly include effect sizes, possibly also useful descriptives (e.g. that you could use for assumptions about means)

- If you can't come up with a specific value you feel very confident in, run a few different power calculations with alternate assumptions to generate bounds

Commonly-used Stata options

- power
- sampsi
- sampclus
- clsampsi
- clustersampsi

power calcs in stata quick reference

Which Stata package should I use?	Clustering	Multiple survey rounds	Different size tmt and control groups	Using old version of stata (12 or lower)	Directly calculate MDES
power	YES (only as of Stata 15)	NO	YES	NO	YES
sampsi	NO	YES	YES	YES	NO
clsampsi	YES	NO	YES	YES	NO
clustersampsi	YES	NO	NO	YES	YES

power

- Stata's newest updated to power calculations
  - Introduced with Stata13, replaces *sampsi*
  - As of Stata15, allows for clustered sample designs

- Pros
  - **Better output: more info, graph option. Automatically saved.**
  - More flexible in terms of input/output choices
  - Can compute sample size of control group given treatment group size (or vice versa)
  - Directly calculate MDES
  - Allows for treatment and control groups of different sizes

- Cons
  - No straightforward way to control for repeated measures

power

- Options
  - *cluster*
    - *Allows for clustered sampling designs*
  - *power onemean*
    - assume means same in tmt & control (e.g. randomization)
  - *n* sample size
    - *n1()* control group size, *n2()* treatment group size
  - *nratio* ratio of n1/n2
    - default is 1
    - not necessary to specify if you list n1 and n2
  - *power, table* outputs results in table format
  - *power, saving(filename, [replace])* saves results in .dta format

sampsi

- No longer official stata package (replaced by power), though it continues to work

- Pros
  - Works with Stata13 or less
  - Allows repeated measures (multiple follow-ups)

- Cons
  - Does not allow clustering
  - Have to impute MDES
  - Defaults to 90% power (not really a con, but be aware)

sampsi options

- *onesample:* use if randomized (assume means the same between treatment and control)
- Sample size
  - *n1(#)* size of treatment group
  - *n2(#)* size of control group
  - *ratio()* n1/n2, default is 1
- Repeated measures
  - *pre* number of baseline measurements
  - *post* number of follow-up measurements
  - *r0(#)* correlation between baseline measures (default *r0 = r1)*
  - *r1(#)* correlation between follow-up measures
  - *r01(#)* correlation between baseline and follow-up
- *method(post change anova* or *all),* default is all

sampsi

- Default is to compute sample size
- To compute power: specify n1 or n2
- To compare means (not proportions), must specify *sd1(#)* or *sd2(#)*
- For repeated measures, sd1(#) or sd2(#) must be specified

sampsi example syntax

- Simple case: one-sample comparison of mean to hypothesized value.
  - Take sample size as given to compute power:
    - *sampsi # (baseline mean) # (hypothesized mean), sd (postulated sd) n (sample size) onesample*
    - *sampsi 0 2.5, sd(4) n(25) onesample*

- More complex: repeated measures
  - Need to know sample sizes, mean, sd, expected correlation
  - *sampsi # (BL mean) # (hypothesized mean), n1 (control sample size) n2 (tmt sample size) sd1 (control sd) sd2 (tmt sd) method(change) pre (number of BL measures) post (number of FUP measures) r1 (correlation btw FUP measures)*
  - *sampsi 485 500, n1(15) n2(15) sd1(20.2) sd2(19.5) method(change) pre(1) post(3) r1(.7)*

clsampsi

- Pros
  - Allows for clustering
- Cons
  - Have to impute MDES
  - Does not allow for repeated measures
  - Does not allow for baseline correlation

clsampsi options

- m(#) cluster size in treatment and control assuming equal cluster size in tmt & control
  - alternative *m1(#)* and *m2(#)*
- k(#) number of clusters in tmt and control assuming equal number in tmt & control
  - Alternative *k1(#)* and *k2(#)*
- *sd(#)* standard deviation assuming same sd in tmt & control
  - Alternative *sd1(#)* and *sd2(#)*
- *rho(#)* ICC assuming same in tmt & control
  - Alternatively rho1 and rho2
- *sampsi* determines power of means (or proportion) comparison using the standard sampsi command

clsampsi less common options

- *varm(#)* cluster size variation assuming same in tmt & ctl
  - only affects power if larger than *m(#)* and *rho(#)*>0
  - Calculate the effect of cluster-size variation (varm1()) on the required sample size
    - *clsampsi 3 2.3, sd1(2) sd2(1.55) m1(6) m2(8) varm1(100) rho1(0.2)*

clustersampsi

- Pros
  - Allows for clustering
  - Allows for baseline correlations
  - Directly calculates MDES

- Cons
  - Doesn't allow for different sized treatment / control groups
  - Doesn't allow for repeated measures

clustersampsi options

- *detectabledifference* calculate MDES
  - Alternative options: *power, samplesize*
  - to use *detectabledifference* must specify *m*, *k, mu1*
- *rho(#)* ICC
- *k(#)* number of clusters in each arm
- *m(#)* average cluster size
- *size_cv(#)* coefficient of variation of cluster sizes (default is 0). Can be any number greater than 1.
- *mu1* mean for tmt (*mu2* = mean for control)
- *sd1* mean for tmt (*sd2* = mean for control)
- *base_correl* correlation btw baseline measurements (or other predictive covariates) and outcome

clustersampsi example

- Detectable difference for fixed sample size: compute the difference detectable with 20 clusters per arm each of size 10 between two means where the baseline mean is 300 and ICC is 0.05.
  - *clustersampsi, detectabledifference mu1(300) m(10) k(20) rho(0.05)*

sampclus

- Add-on to sampsi that allows for clustering
- Must be preceded by sampsi

sampsi 200 185, alpha(.01) power(.8) sd(30)
sampclus, obsclus(10) rho(.2)
sampclus, obsclus(10) rho(.1)

- Corrects sample size and computes number of clusters from a t-test
- Adjusts this sample size calculation for 10 observations per cluster and an ICC of 0.2
- Repeats for an intraclass correlation of 0.1

example of reporting power calcs – simple individual-level randomization

Parameter	Values	Definition	Source of parameter
α	0.05	Significance level	Assumption
β	0.8	Desired power of the test	Assumption
Tail	2	One-tailed or two-tailed test	Detect either an increase or a decrease in yields
μ	12761	Pooled mean of outcome variable (yield in RWF/ha)	Baseline survey in the three ongoing sites
σ_y	24233	Pooled standard deviation of outcome variable (yield in RWF/ha)	Baseline survey in the three ongoing sites
P	0.52	The proportion of the study sample randomly assigned to treatment	Actual treatment/control ratio in two of the ongoing sites; expectation is ~.5
N	690	The size of the study sample	Actual sample size in two of the ongoing sites (expected to double with new sites)
stata code	*sampsi 12761 15930, sd1(24233) method(change) n1(359) n2(331)*	stata package: sampsi	
D	0.13	Minimum detectable effect (in standard deviations)	Calculation

Example of reporting power calcs – clustered randomization

Parameter	Test 1: Low ICC	Definition	Source of parameter - comments
α	0.05	Significance level	Assumption
β	0.8	Desired power of the test	Assumption
Tail	2	One-tailed or two-tailed test	Detect either increase or decrease in yields
μ	12761	Pooled mean of outcome variable (yield in RWF/ha)	Baseline survey in the three ongoing sites
σ_y	24233	Pooled SD of outcome variable (yield in RWF/ha)	Baseline survey in the three ongoing sites
K	152	Number of clusters	Actual sample size in the three ongoing sites (expected to ~double with 3 new sites)
N	7	Number of observations per cluster	
ρ	0.75	Correlation between baseline and follow-up measurements	Assumption
icc	0.05	Intracluster correlation	Based on icc for agricultural yield data for two of the three sites, from a 2013 survey run by the same research team.
stata code	*clustersampsi, detectabledifference mu1(12761) sd1(24233) m(7) k(152) rho(0.05) base_correl(0.75)*	stata package: clustersampsi	*** note that estimates are conservative. clustersampsi doesn't allow for imbalanced treatment/control, so assumes comparing 76 tmt v. 76 ctl. Also only allows for 1 follow-up (unlike sampsi, which has post option).*
D	0.12	Minimum detectable effect (in standard deviations)	Calculation