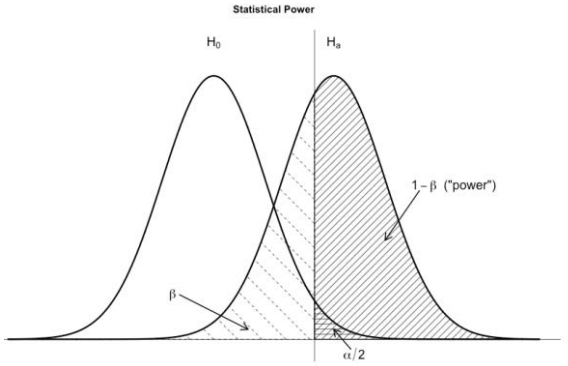
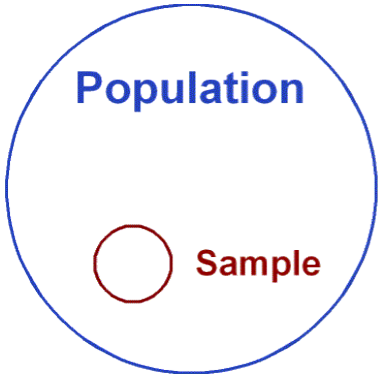


Statistical Power Sampling Design and sample Size Determination



Deo-Gracias HOUNDOLO
Impact Evaluation Specialist
dhoundolo@3ieimpact.org

Outline

1. Sampling basics
2. What do evaluators do?
3. Statistical Power?
4. Sample Design: SRS and TSS
5. How to determine required sample size
6. Exercise 1: Simple random sample case
7. Exercise 2: Two stage random sample case
8. Take Away: Why power calculation?
9. Things to know
10. Final words

Sampling basics

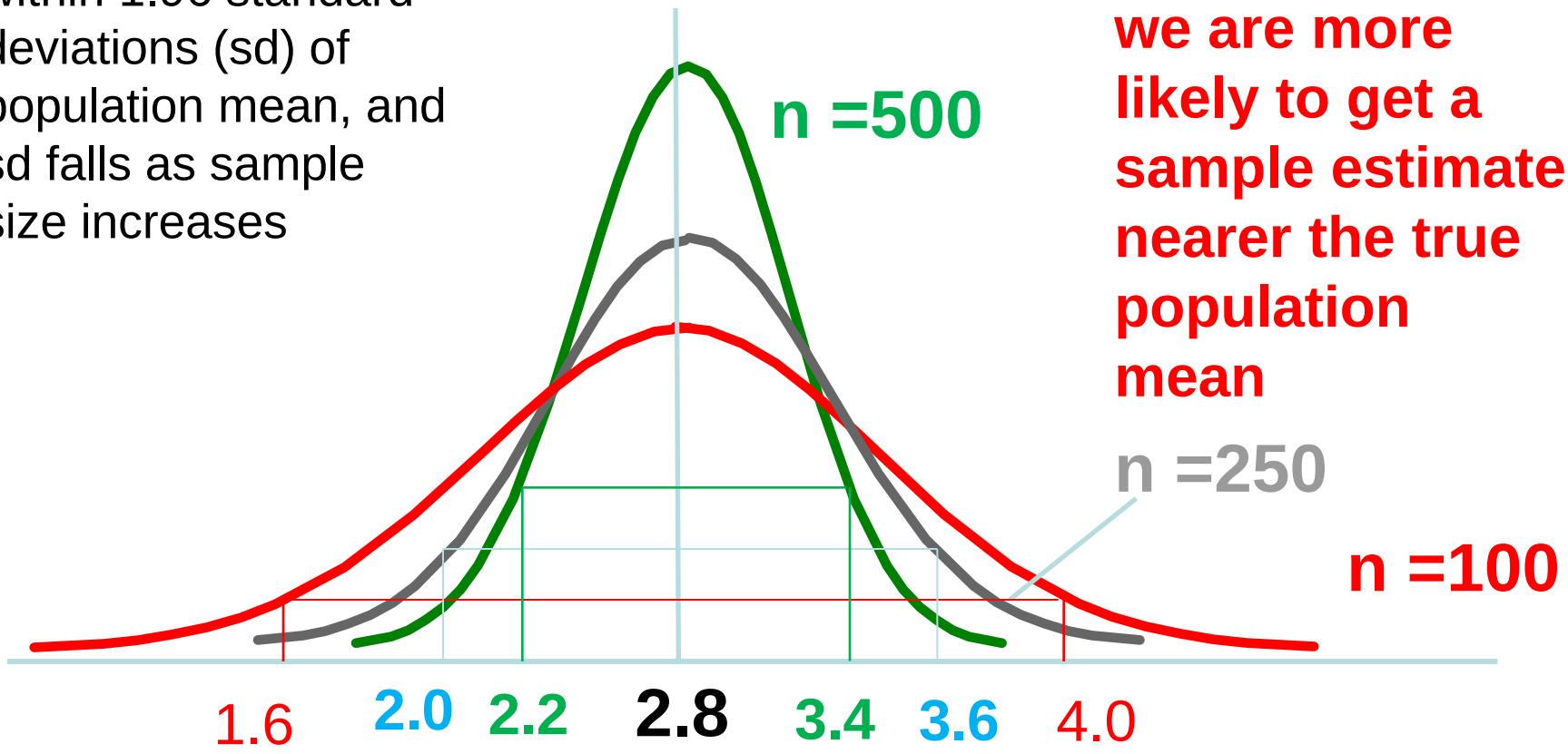
Population mean: The true value of a parameter, i.e. the average weight for age of all children aged under in the region of interest.

Sample mean: the average weight for age in a sample drawn from the population.

The larger the sample the more likely it is that the sample mean is close to the population mean (provided our sample is a random sample)

Distribution of sample means

95% of estimates fall within 1.96 standard deviations (sd) of population mean, and sd falls as sample size increases



So as sample size increases we are more likely to get a sample estimate nearer the true population mean

Some sampling basics

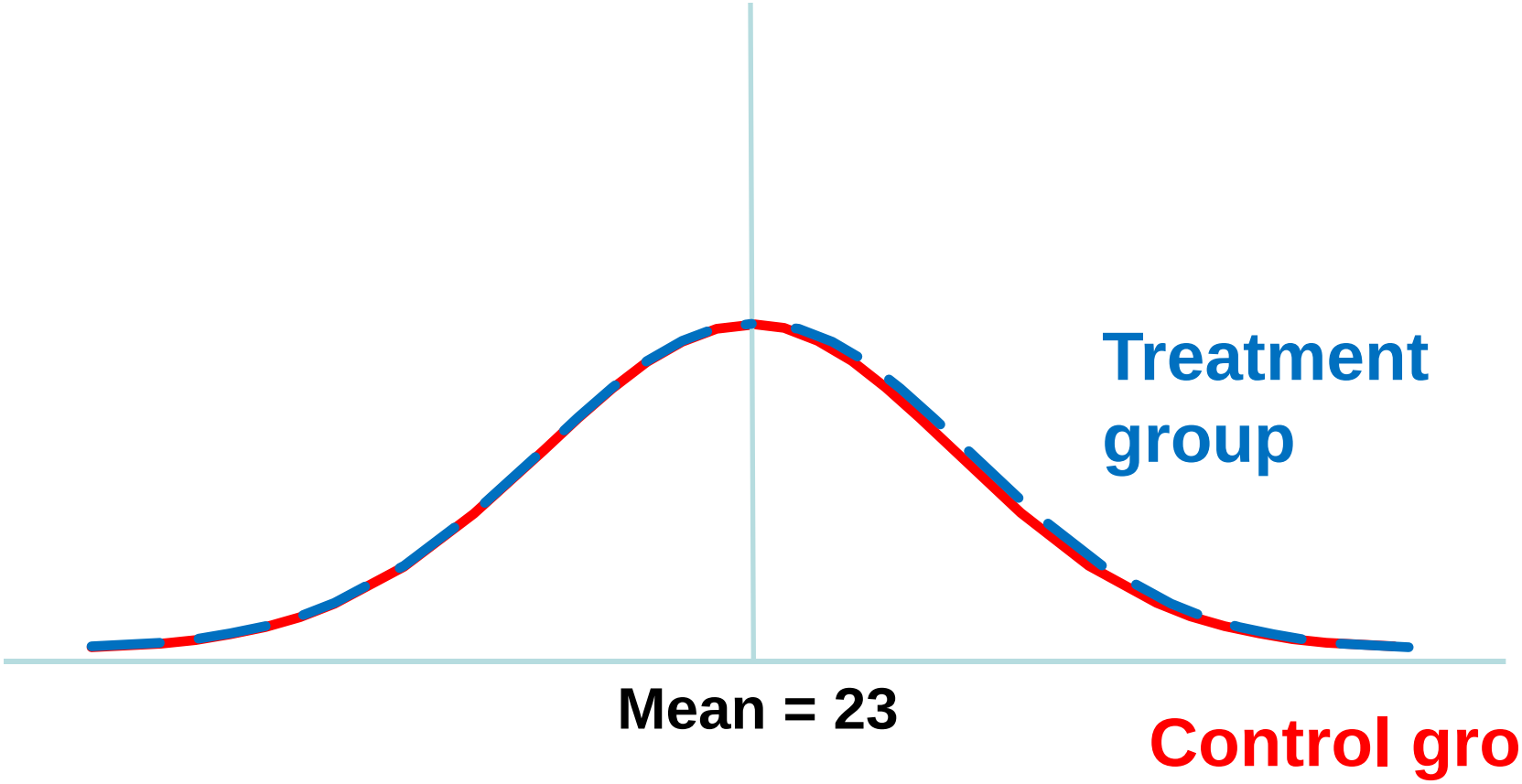
Table 1 Average characteristics by different sample sizes (n)

	Rural (%)		Years of education		Number of household members	
	Treatment	Control	Treatment	Control	Treatment	Control
n=2	100	0	12.0	9.0	9.0	5.0
n=20	70	80	6.4	5.8	6.4	6.7
n=50	72	60	5.8	5.3	6.4	6.5
n=200	65	61	6.0	5.0	6.7	6.5
n=2,000	66	64	5.2	5.4	6.5	6.5

The larger the sample the more likely it is that treatment and control are comparable

What do evaluators do before treatment?

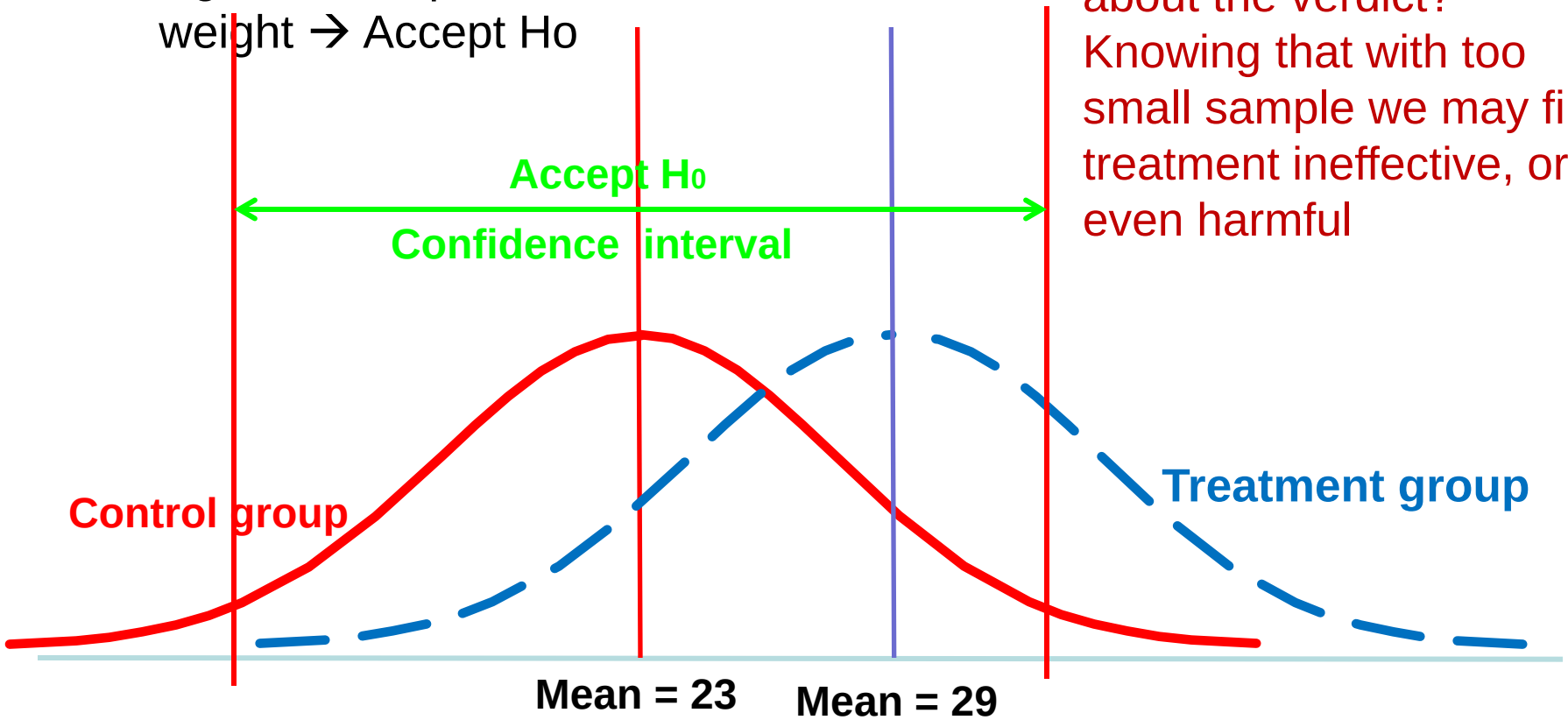
Distribution of students' weights in treatment and control groups before a FNS program treatment



What do evaluators do after treatment?

There is insufficient evidence to conclude that the treatment has a significant impact on students' weight → Accept Ho

How confident are we about the verdict? Knowing that with too small sample we may find treatment ineffective, or even harmful



Hypothesis testing in impact evaluation

Formally, impact evaluators ***tests the null hypothesis***…

- H_0 : impact = 0

The null hypothesis is that the program does not have an impact

…***against the alternative hypothesis***:

- H_a : impact ≠ 0

The alternative hypothesis is that the program has an impact

Errors in hypothesis testing

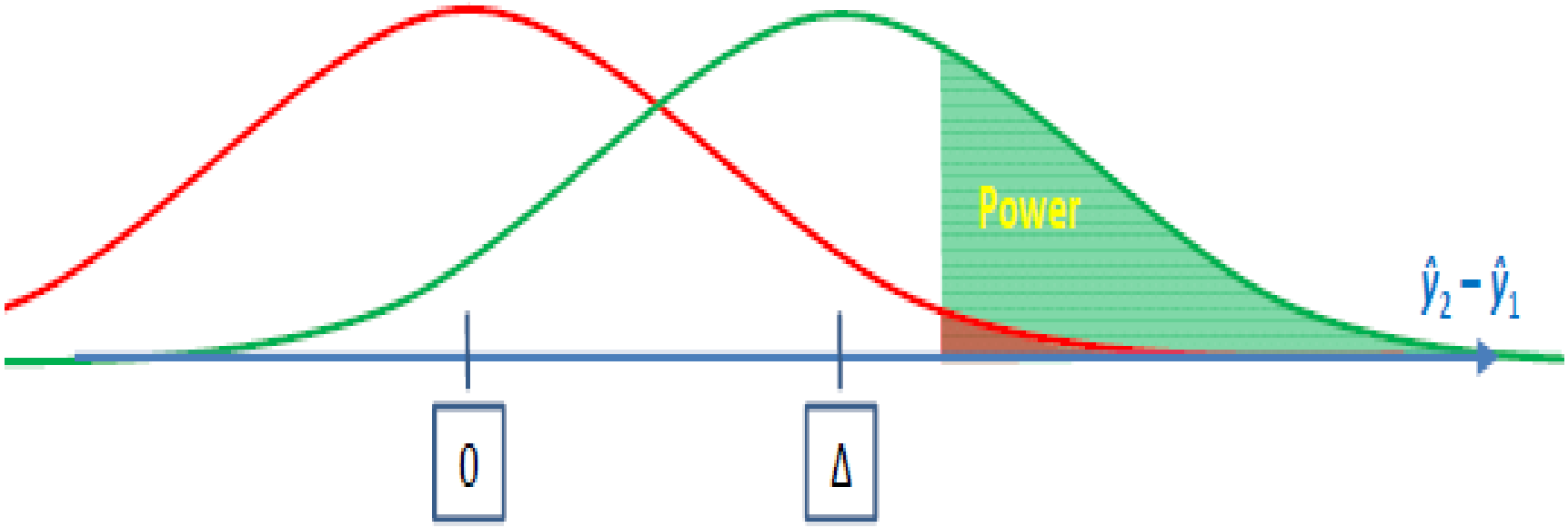
Evaluator	H_0 true	H_0 false
Accept H_0	No error $1-\alpha = 95\%$	Type II error $\beta = 20\%$ in Social Science
Reject H_0	Type I error = 5% in Social Science	No error => Power $1-\beta = 80\%$

What is power?

The *power (or statistical power)* of an impact evaluation design is the likelihood that it will detect a difference between the treatment and comparison groups, when in fact one exists.

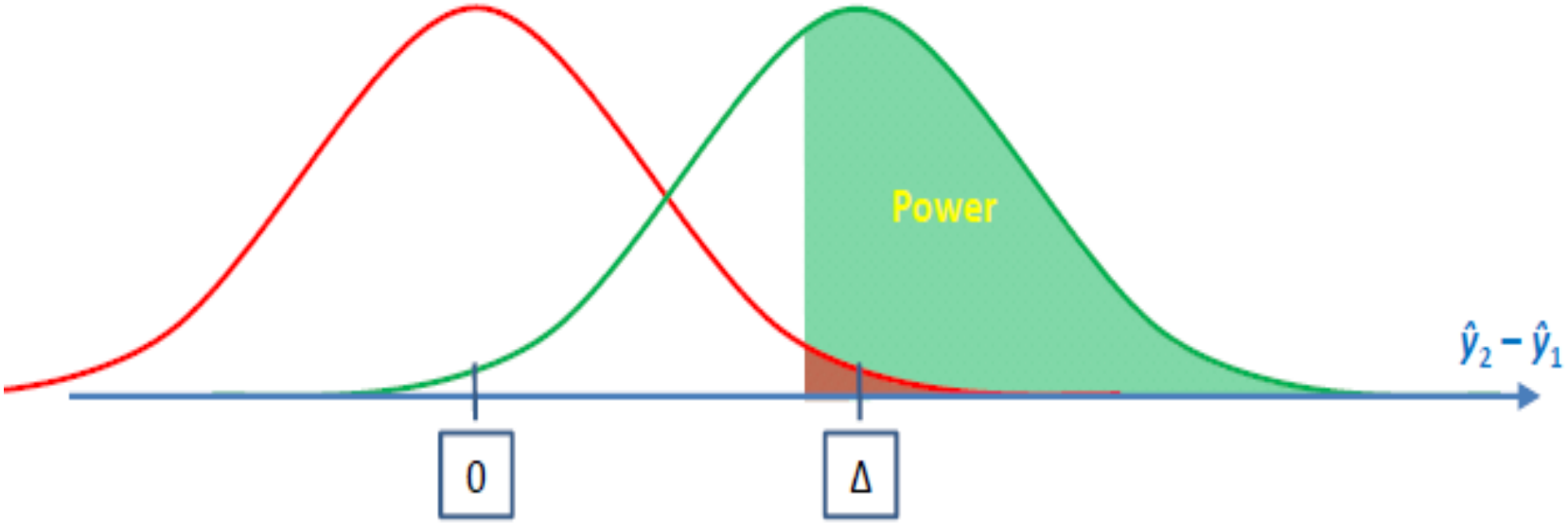
Power calculation indicate the smallest sample size required for an evaluation design to detect a meaningful difference (Minimum Detectable Effect) in outcomes between the treatment and comparison groups.

What affect the power of a design?



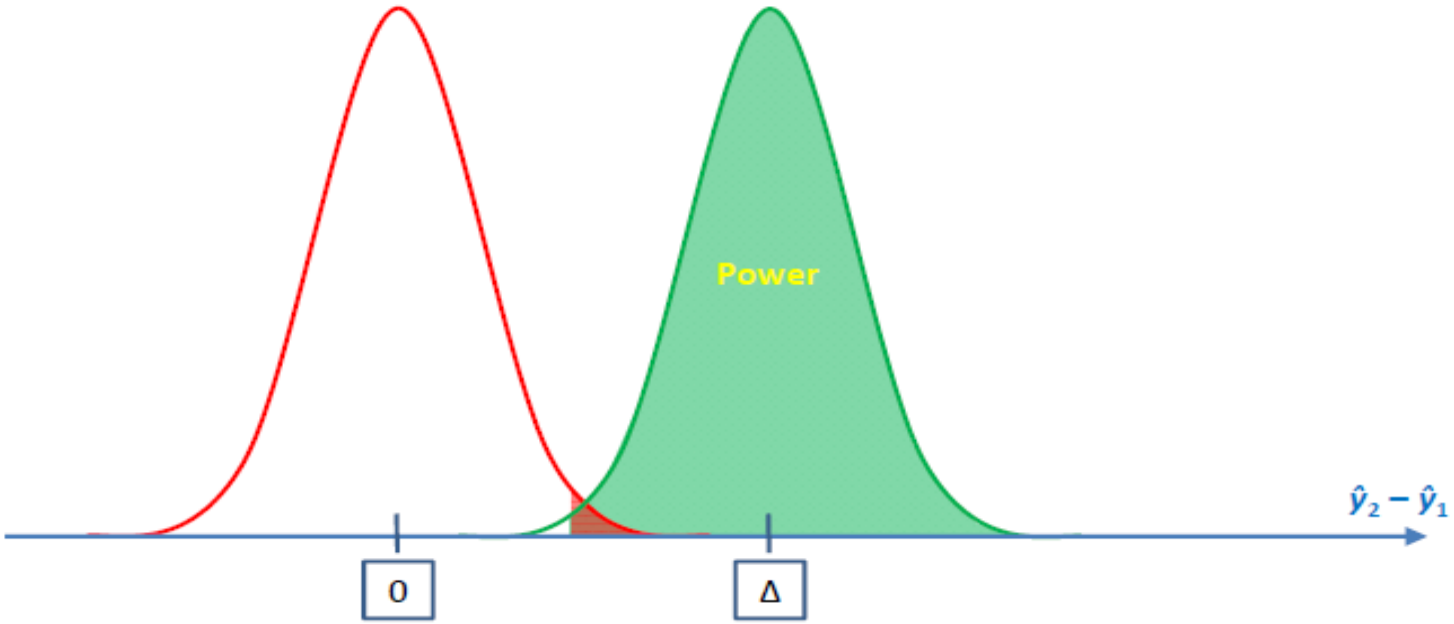
Depiction of power (in green)

What affect the power of a design?



Depiction of power (in green) with increased sample size vis-à-vis previous chart

What affect the power of a design?



Depiction of power (in green) with larger sample size

Several parameters affects power

Bu what matters in the end is an

effective sample size

Simple Random Sampling

SRS may be an option in certain cases, but it may not be practical if:

– if we need estimations for subgroups of the population
– especially if some of the subgroups are small
– we don't dispose of an adequate sample frame
– a Simple Random Sample would be too scattered in the territory

We then resort to other techniques

– Stratification
– Sampling in stages

Stratification

HOW

- We divide the population into subgroups, called strata
- We take a separate sample in each stratum

WHEN

- Stratification may be needed if:

– We want to reduce the standard error, by gaining control of the composition of the sample

– We want to assure the representativity of certain groups

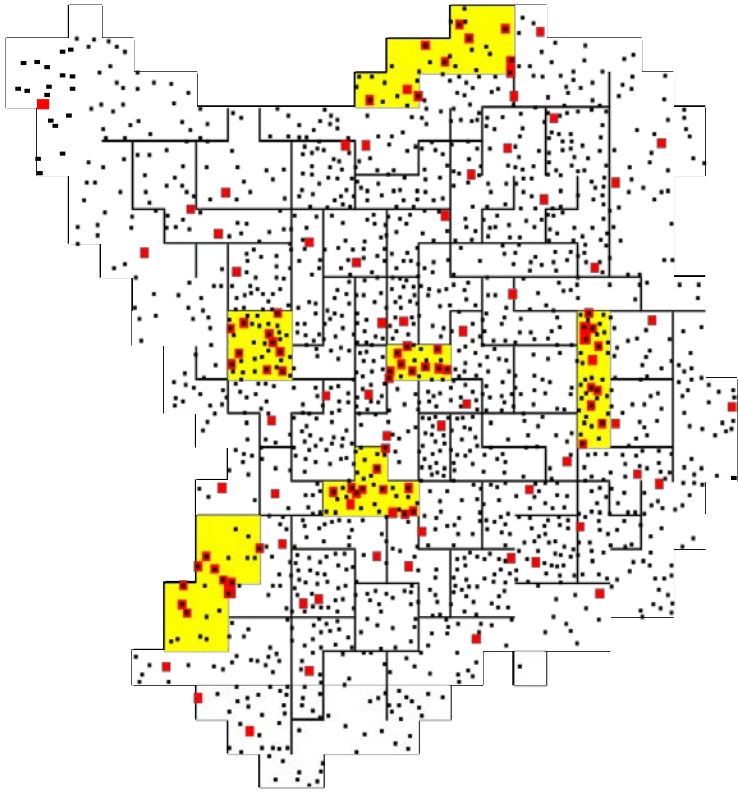
Two-stage sampling

Instead of taking a SRS

- We divide the territory into small areas, called Primary Sampling Units (PSUs).

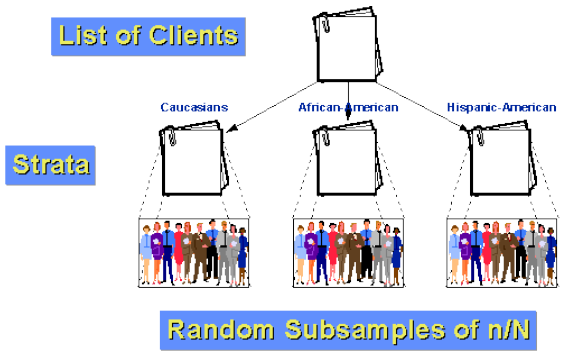
  – In the first stage, we choose PSUs.

  – In the second stage, we select households in the chosen PSUs



Two-stage sampling

- Solves the problems of SRS
  - Reduces transportation costs
  - Reduces sample frame problems
- The sample can be made self-weighted if
  - We choose PSUs with Probability Proportional to Size (PPS), and then
  - We take a fixed number of households in each PSU
- The price to pay is cluster effect



Cluster-Randomization

- Randomization addresses the problem of selection bias by the random allocation of the treatment

- Randomization may not be at the same level as the unit of observation
  - Randomize across schools but measure individual learning outcomes
  - Randomize across sub-districts but measure village-level outcomes

- You need to randomize across a 'reasonable number' of units

Why Cluster-Randomization?

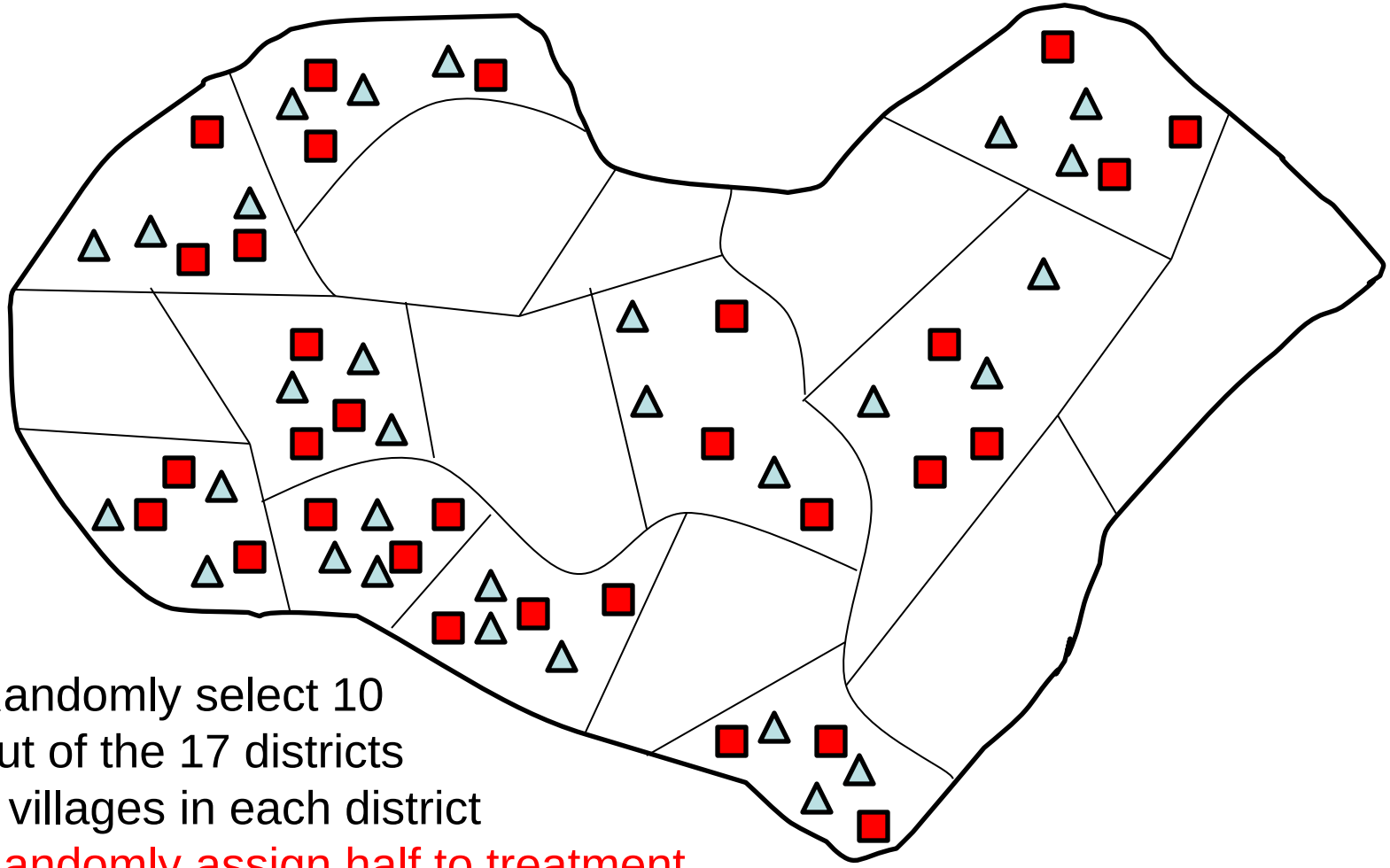
1. Ethics

- Not fair to provide one individual in a village with benefits and another individual not.

2. Spillovers

- Spillovers → Non-beneficiaries also benefit

Cluster-Randomization in Practice



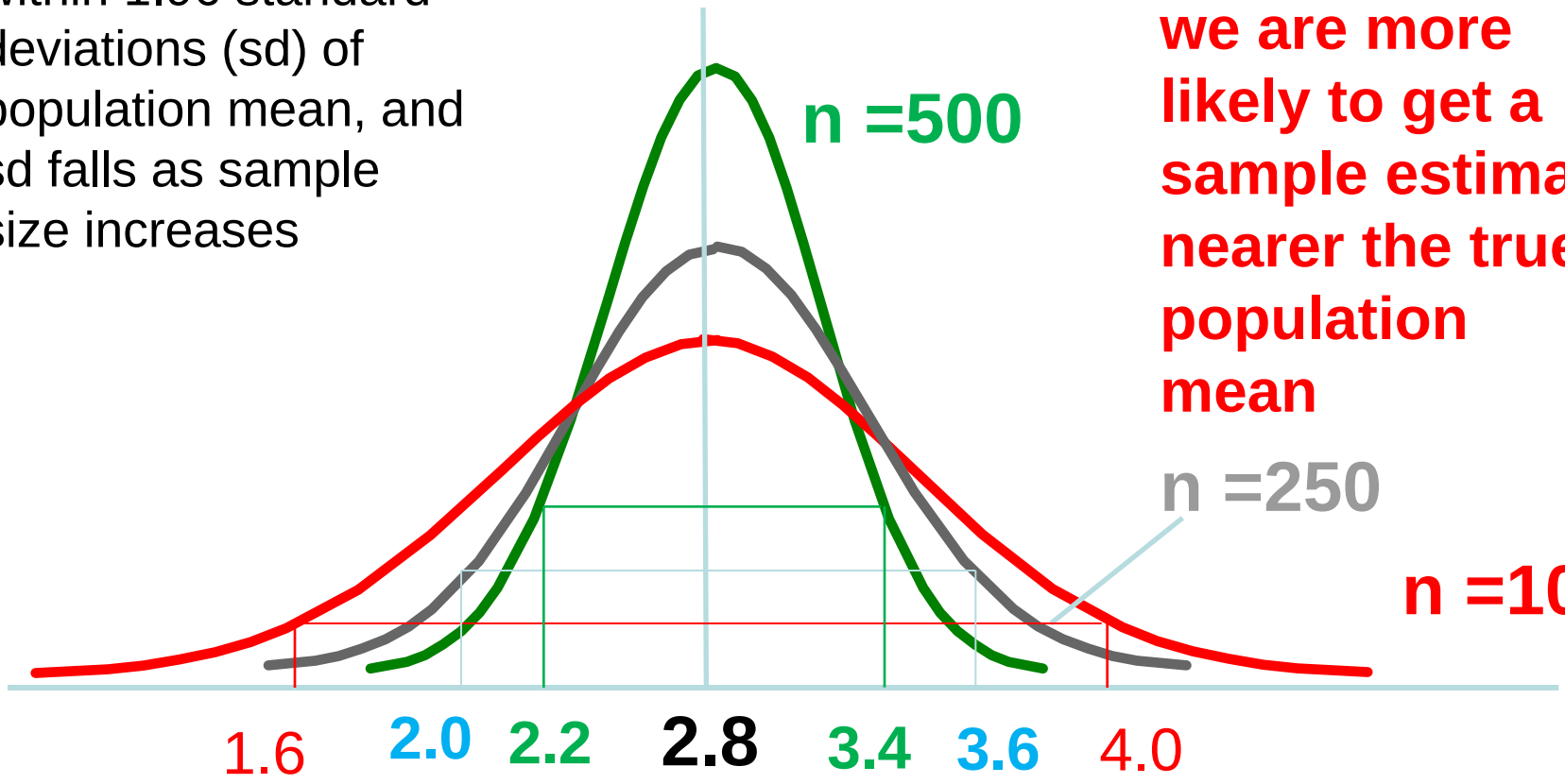
Randomly select 10
out of the 17 districts
6 villages in each district
Randomly assign half to treatment
10 hhs/village, n= 600

Sample Size Calculations

How big is big enough?

Distribution of sample means

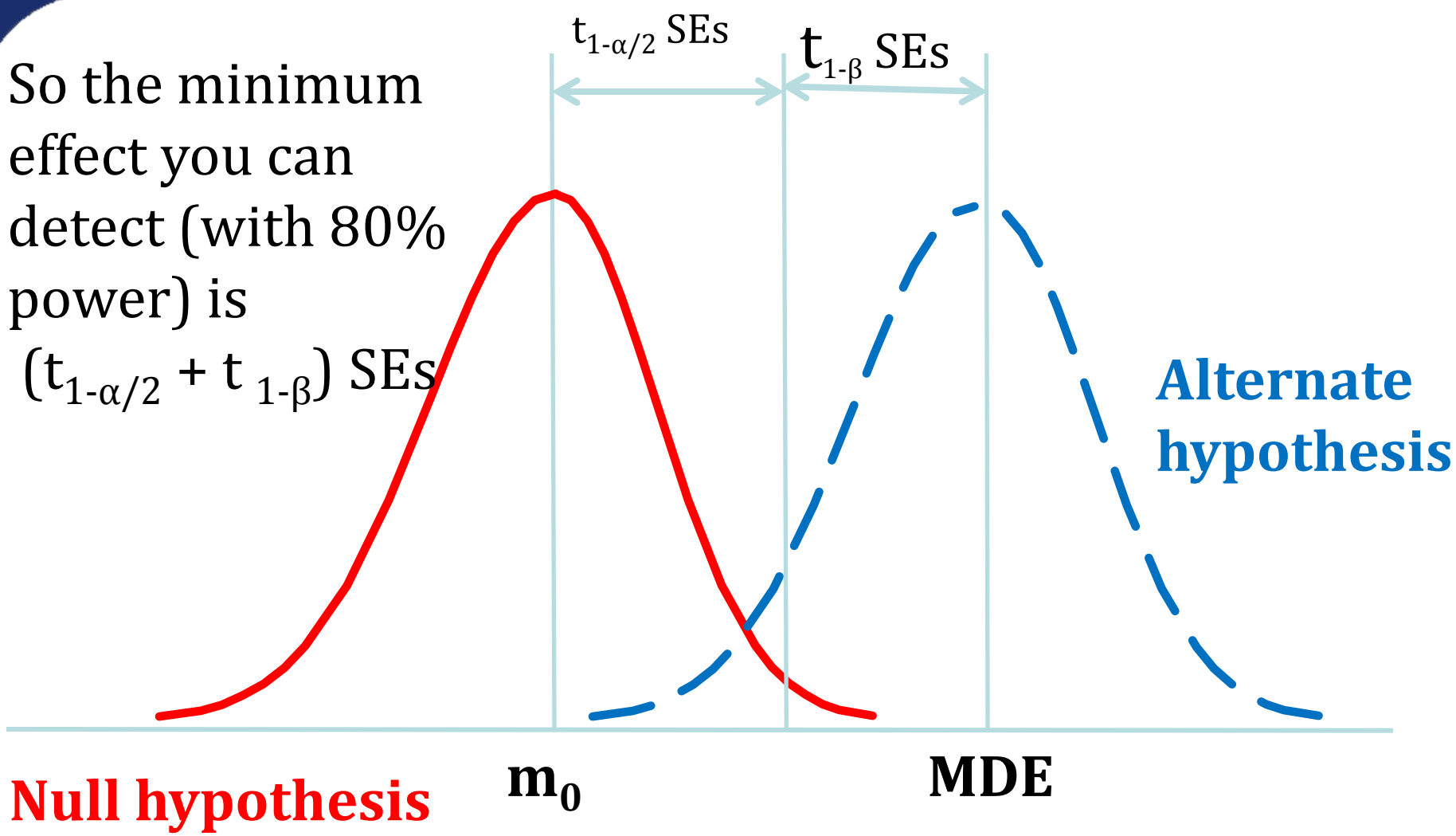
95% of estimates fall within 1.96 standard deviations (sd) of population mean, and sd falls as sample size increases



So as sample size increases we are more likely to get a sample estimate nearer the true population mean

How far apart do the distributions need to be?

So the minimum effect you can detect (with 80% power) is $(t_{1-\alpha/2} + t_{1-\beta})$ SEs



How to determine sample size (1)?

$$MDE = (t_{1-\alpha/2} + t_{1-\beta})e$$

MDE: Minimum detectable effect

$\alpha/2$: Rate of Type I errors (false positives)
(typically $\alpha/2$ = 2.5%)

β: Rate of Type II errors (false negatives)
(typically β = 10 - 20% ←→ Power = 90 – 80%)

e: Standard error of the estimated effect

How to determine sample size (2)?

For treatment and control groups of the same size, selected with SRS, and the same variance

$$e = \sqrt{\frac{2\sigma^2}{n}}$$

e: Standard error

n: Sample size of each group

σ^2: Variance of the outcome

- for a prevalence, σ^2=P(1-P)

$$n = \left(\frac{t_{1-\alpha/2} + t_{1-\beta}}{MDE}\right)^2 2\sigma^2$$

Time For Fun Part 1☺

Determine on your own required sample size

Exercise 1: Power calculation using Simple Random Sample

i. 42% youth unemployment rate (national survey Jan 2014).
ii. Youth wage voucher programme to reduce unemployment to 20% in 2 years
iii. 4000 youngsters are eligible
iv. Minister decides 420 equally distributed in T and C groups.
v. You lead a 3ie impact evaluation team and a journalist asks you:

Do you think that a sample size of 420 is enough for the evaluation?

t(1- β)=0.84 (if β=0.2) and

t(1-α/2)=1.96 (if α =0.05)

σ^2=P(1-P)

$$n = \left(\frac{t_{1-\alpha/2} + t_{1-\beta}}{MDE}\right)^2 2\sigma^2$$

RESULTS

- IF power is 80%, β=0.2 and α =0.05 then

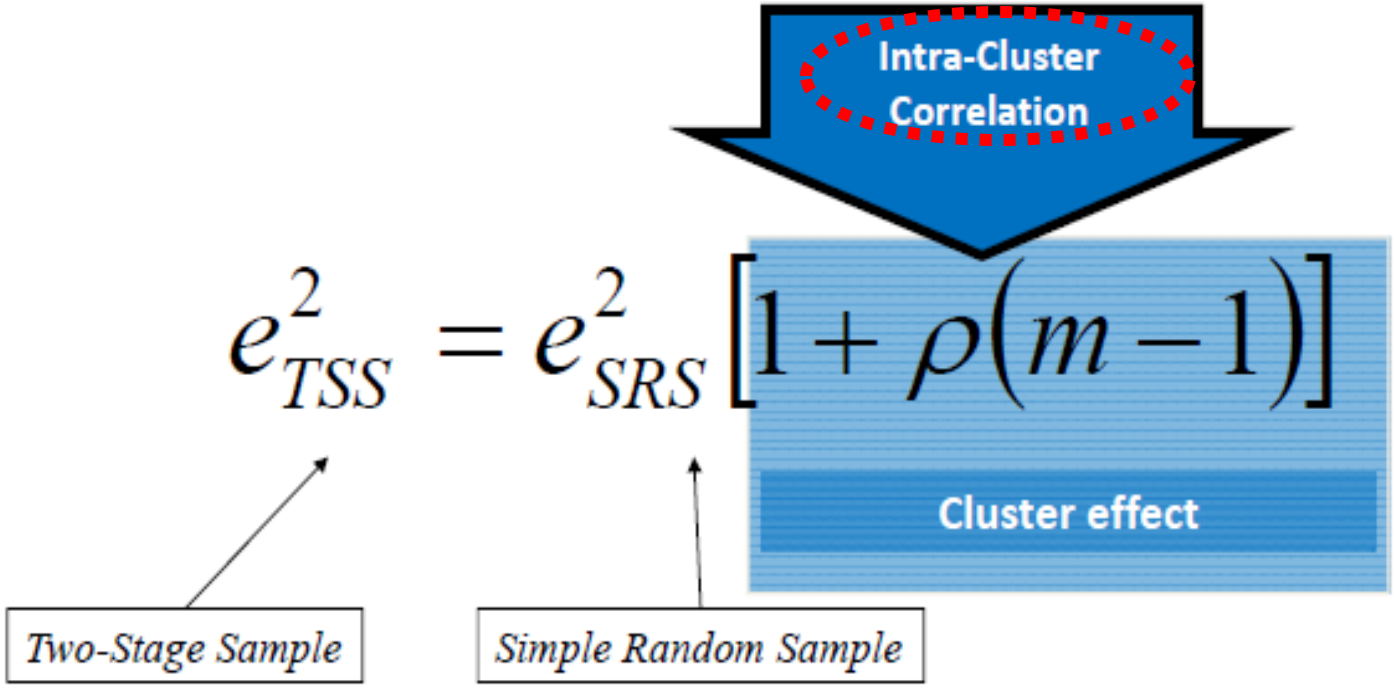
n =(((1.96+0.84)/0.22)^2)*(2*(0.42*(1-0.42)))=79 participants in each group ➔ 158 participants in total.

- Assuming 5% attrition rate, we may plan to sample 83 participants by group → 166 participants in total

YES 420 participants is enough if not far beyond what is required.

How to determine sample size using 2 stage sampling?

Standard error grows if, instead of taking a Simple Random Sample of n households, we take a two-stage sample, with k PSUs and m households per PSU ($n=k \cdot m$)



Rho: Intra-cluster correlation

- We want variation *within* clusters
- So a lower value of ρ is better
- If there is no variation it is as if each cluster is just one observation
- You need to use existing data to get a value of ρ, which will usually be in the range 0.15-0.25

Cluster Effect

For a total sample size of 12,000 households

Number of PSUs	HHs per PSU	Intra-Cluster Correlation				
		0.01	0.02	0.05	0.10	0.20
3,000	4	1.03	1.06	1.15	1.30	1.60
2,000	6	1.05	1.10	1.25	1.50	2.00
1,500	8	1.07	1.14	1.35	1.70	2.40
1,000	12	1.11	1.22	1.55	2.10	3.20
800	15	1.14	1.28	1.70	2.40	3.80
600	20	1.19	1.38	1.95	2.90	4.80
400	30	1.29	1.58	2.45	3.90	6.80
300	40	1.39	1.78	2.95	4.90	8.80
200	60	1.59	2.18	3.95	6.90	12.80
150	80	1.79	2.58	4.95	8.90	16.80
100	120	2.19	3.38	6.95	12.90	24.80

- **Number of clusters drives power,**

not

- **no. of observations in a cluster**

Time For Fun Part 2 ☺

Determine on your own required sample size required

Exercise 2: Power calculation using Two Stage Random Sampling

i. 42% youth unemployment rate (national survey Jan 2014).
ii. Youth wage voucher programme to reduce unemployment to 20% in 2 years
iii. 4000 youngsters are eligible
iv. Minister decides 15 youngsters in 28 communities (420 youngsters) equally distributed in T and C communities.
v. You lead a 3ie impact evaluation team and a journalist asks you:

Do you think that a sample size of 420, 15 youngsters in 28 communities, is enough for the evaluation?

$t(1-\beta)=0.84$ (if $\beta=0.2$) ;

$t(1-\alpha/2)=1.96$ (if $\alpha=0.05$) and $\rho = 0.2$

Simple Random Sample

RESULTS

$$J = 1 + \frac{(Z_1 + Z_2)^2\left[\frac{(\mu_0+\mu_1)}{n} + k^2(\mu_0^2 + \mu_1^2)\right]}{(\mu_0 - \mu_1)^2}$$

Parameter	Value	Definition
α	0.05	Significance Level
β	0.8	Desired power of the test
Tail	2	One-tailed or two-tailed test
z_1	1.96	Z-value corresponding to the desired significance level of the test
z_2	0.84	Z-value corresponding to the desired power of the design
a	7.85	(z1+z2)^2*
R2	0	The coefficient of variation of true proportions between clusters within each group
m	15	Number of individuals in each cluster
μ_0	0.58	True (population) rate in the absence of the intervention
μ_1	0.8	True (population) rate in the presence of the intervention
k	16	Number of clusters in each group

n=m*k=15*16= 240 participants

5% attrition out of 15 → 1extra/community → 245 in total

YES a sample size of 420 is enough for the evaluation.

Account for attrition as a result of:

- Households which can't be located
- Or aren't in
- Or refuse
- Or return unusable data
- Or don't comply with treatment

Why power calculation?

- Not acceptable to conduct a study that would not be stringent enough to detect a real effect due to a lack of statistical power.
- Not acceptable to conduct a study by recruiting 1000s of participants when sufficient data could be obtained with 100s of participants instead.
- Recruiting more participants than required would also be a waste of both resources and time

Things to know

1. The smaller MDE → The larger sample size required

2. Better have large number of clusters than large number of households within clusters

3. The lower the take-up the lower the power

4. The size of study population has very little to do with the sample size required for an evaluation

My final words

If a study does not detect a statistically significant effect of an intervention, it does not necessarily mean that the study is under-powered. It may be because the intervention fails to deliver according to plan (implementation failure) or it is just not the right intervention for the problem at hand. Do not blame power whenever there is no statistically significant result…

Deo-Gracias HOUNDOLO
Impact Evaluation Specialist
dhoundolo@3ieimpact.org

Errors in testing hypothesis

Two Potential Errors when testing hypothesis in Impact Evaluations:

- An error would occur when an evaluator concludes that a program *had no impact*, when in fact the ***program does have the expected impact***

  →Type I.

- Another error would occur when an evaluator concludes that a program *had an impact*, when in reality the ***program does not have the expected impact***

  → Type II.

SRS vs. TSS

Two Stage random sampling

- 245 participants required to detect 22% drop of umemployment with 80% power.

Simple random sampling

- 166 participants required to detect 22% drop of umemployment with 80% power.

The horrifying truth about hypothesis testing

- Confidence intervals are needed because our data are a sample
- If the ‘null hypothesis’ is correct (null = no programme impact) then we will correctly agree with the null 95% of the time (we are wrong 5% of the time)
- But if the null hypothesis is wrong (the programme works) then we probably incorrectly conclude the programme doesn’t work 40-60% of the time!!!

Implications

- An RCT is no better than tossing a coin at determining if a successful programme is working so
- Power, power, power
- A theory-based approach can lead us to think correct or false negative
- We also need replicate ‘unsuccessful’ programmes
- And we really REALLY need to do SRs (we will see why shortly)

Sample Size Calculations

Larger sample → more likely that treatment and control are comparable

	Years of education	
	Treatment	Control
n=2	12.0	9.0
n=20	6.4	5.8
n=50	5.8	5.3

Why does sample size matter?

- Minimizing error associated to estimation of

$$MeanT \text{ and } MeanC$$

- Use rational behind Law of large number and Central Limit Theorem

- $$Var(\text{estimated mean } Y) = Var\ (\text{mean } Y) / n$$

Rules of thumb for power calculation

1. Even though power calculation is a technical task, it is also true that there are **a few rules of thumb** that are applied and can always serve as guidance.
2. When power increases, the probability to find a true impact of the intervention (if it exists) increases. In social science, researchers aim to have at least 80% power which means allowing 20% chance of committing a type II error.
3. The larger your sample size, the smaller the standard error and therefore the higher your power.
4. The smaller the Minimum Detectable Effect, the larger the sample size needs to be.
5. For any given number of clusters, the larger the intra-cluster correlation, the lower the power.
6. For any given number of unit of observation per cluster, the larger the number of clusters the higher the power.
7. Increasing the units of observation per clusters will generally not improve power as much as would increase the number of clusters (unless ICC is 0).
8. Intra-cluster correlation increases when observations within clusters are getting more and more identical relative to other clusters, which lowers the number of independent observations and, effectively, the sample size.
9. Baseline covariates are used in model specification to increase the statistical power of the study because they reduce the standard error of outcome and therefore increase the likelihood to reduce the minimum effect that the design can detect.

Common pitfalls for power calculation

Sample size should be determined for all main outcome variables before final decision on study sample size is made. It is not appropriate for instance to run power calculation only for school attendance when, for instance, learning outcomes are also a main outcome of interest.

Minimum Detectable Effect of an intervention is highly a function of the impact trajectory of the intervention over time. and therefore it is necessary and essential to take into account the expected timeline of the intervention to evaluate before deciding the magnitude of the Minimum Detectable Effect.

Power calculations must account for intra-cluster correlation in case of cluster sampling, as it does affect power. That is, not all samples of the same size have equal power).

If a study does not detect a statistically significant effect of an intervention, it does not necessarily mean that the study is under-powered. It may be because the intervention fails to deliver according to plan (implementation failure) or it is just not the right intervention for the problem at hand. Do not blame power whenever there is no statistically significant result.

Common pitfalls for power calculation

Attrition is a major threat to evaluation because it decreases the sample size with full information and therefore reduces power. There is no genuine way to rectify sample size after attrition occurs. To minimize attrition, be sure to collect enough data to be able to track participants. To avoid the effects of expected attrition, it is necessary to over-sample or take all-necessary measures that do not compromise the intervention to avoid or limit attrition.

Spillover and contamination are other ghosts that bias estimates and therefore affect power. Spillover makes control group affected by intervention through different mechanism, while contamination makes treatment or control groups affected by similar intervention during the study and therefore bias attributable effect estimates for the intervention studied. Hence study design and implementation should be guarded against.

Power calculation is run to decide on the sample size required for an evaluation study. It is an *ex-ante* activity and not an *ex-post* decision. When run *ex-post*, it can check actual power but the purpose is completely different from that of power calculation. In *ex-post* power check, the objective is to determine the power of the study, given the actual sample size used for analysis but using the same values for all other parameters used while running *ex-ante* power calculation. ,.

Using a randomized control trial as the identification strategy does not alone guarantee that power will be sufficient. .

Power calculation formulae or programming are not the same for continuous versus binary outcome measures. It is a mistake to use the same formulae in each case. Even when using software packages,, it is critical to specify the nature (continuous or binary) of the outcome variable of interest.

The notion of power

H_0 : The hypothesis is that the program does not have an impact

What we know

What researchers conclude?

In reality what happens in the eligible population

Null hypothesis H_0
No impact
H_0: $MoyT \_ MoyC = \Delta 1$

Alternative hypothesis H_A
There is impact
H_A: $MoyT \_ MoyC = \Delta 2$

No impact
Accept H_0 if
$MeanT \_ MeanC \leq t$

P (correctly accept H_0) = 1-α/2

P (wrong acceptance of Ho:Type II error) = β

There is impact
Reject H_0 if
$MeanT \_ MeanC > MDE$

P (wrong rejection of Ho: Type I error) = α/2

P (correctly reject H_0) = 1- β

$MeanT \_ MeanC = \Delta$

These means are estimated with errors

Significance

Power

Visualization of statistical power and other key parameters in hypothesis testing