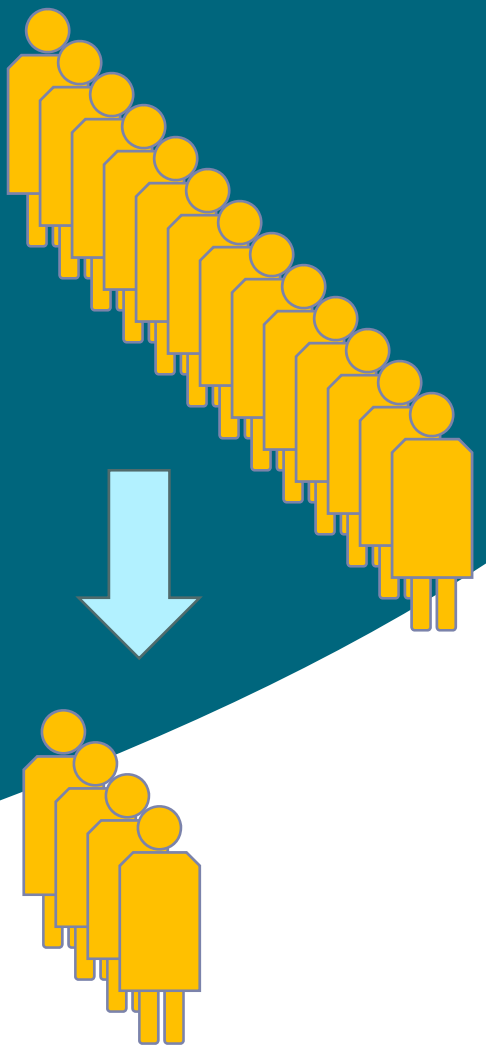
What sample size do I need for my evaluation?

An overview of basic concepts for novices with examples from health care evaluations



AEA Conference 2018

Cleveland, OH

Nov. 3, 2018

Eva Bazant, DrPH, Research Team Lead

Mark Kabue, DrPH, Senior Research Advisor

Johns Hopkins University Affiliate

jhpiego

Saving lives. Improving health. Transforming futures.

Objectives

Understand the…

- connection between sample size, study power, study hypothesis and outcomes
- difference between type I and II errors
- inputs for sample size and power calculation
- available examples from the literature
- example from Jhpiego's work
- sample size computation resources available

jhpiego
Saving lives. Improving health. Transforming futures.

Research Starts with a Question

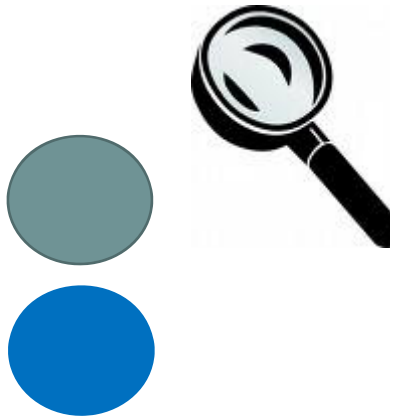
Example: Does the proportion of people who develop Outcome X differ between those in Group A who were exposed to the innovative program and those exposed to the standard program?

jhpiego
Saving lives. Improving health. Transforming futures.

Four Possible Answers to The Study Question

The two groups **do not differ** (with respect to outcome) and we **correctly conclude** that they do not differ

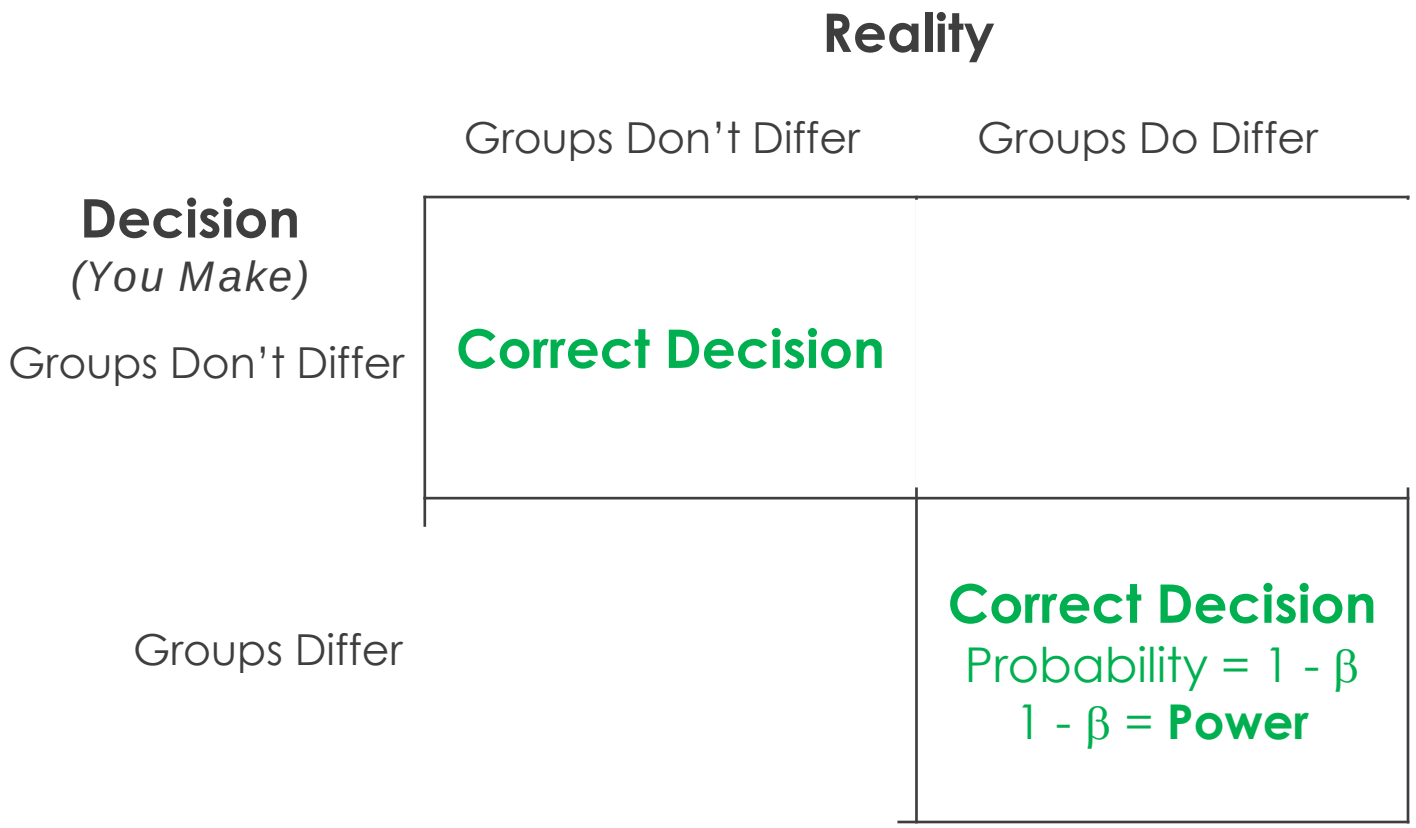
The two groups **differ** and we **correctly conclude** that they differ



Four Possible Answers to The Study Question

1. The groups **do not differ** (with respect to outcome) and we **correctly conclude** that they do not differ
2. The groups **do NOT differ** but we **incorrectly conclude** that they differ **Type I Error**
3. The groups **differ** and we **correctly conclude** that they differ
4. The groups **differ** but we **incorrectly conclude** that they do NOT differ **Type II Error**

Probabilities of Answers



Gordis. *Epidemiology*. 5th Ed.

Probabilities of Answers

Reality

Decision *(You Make)*	Groups Don't Differ	Groups Differ
Groups Don't Differ	**Correct Decision**	**Type II Error** False *Negative* Probability = β
Groups Differ	**Type I Error** False *Positive* Probability = α α = P-Value	**Correct Decision** Probability = 1 - β 1 - β = **Power**

Gordis. Epidemiology. 5th Ed.

Type 1 Error (α, P-Value)

"alpha" α = Probability that we detect -- by chance alone -- a difference that *does not really exist*

- Default is often 0.05

Importance: false positive findings

- Example: falsely stating that a treatment reduces mortality when it does not
- **Question: When there is a smaller α, does this need a smaller or larger sample size?**

Type 2 Error (β)

β ("beta") = The probability of not detecting a difference *that exists*

- 1 - β = Power; **High power = low β**
  - **Default is 80% power:** The default probability for **not** detecting a difference *that exists* is 20% (80% = 1 - β)
    - language: 'not detecting' = 'failing to detect' = failing to observe
  - **Does higher power need a smaller or larger sample size?**

Importance – False negative findings

- Failing to observe a true effect of a toxin on increased risk of cancer

Research Question leads to Hypothesis Testing

Hypothesis Testing is the basis of doing sample size and power calculations

- ***Start with a 'Null' hypothesis:***
  - H_0: Incidence of Outcome in exposed = Incidence Outcome in unexposed

- ***Alternative hypothesis (what we expect):***
  - H_1, Incidence Outcome in exposed ≠ Incidence Outcome in unexposed

- **Power = Probability of rejecting H_0 when H_1 is true**

Towards Sample Size - Inputs

Quantify study question (hypothesis) – Example:

- **Hypothesis:** The rate of HIV incidence will be lower for those who undergo medical male circumcision compared to those who do not.
- **Quantify the expected difference**: From literature review or based on expert opinion, we expect a *50% lower* HIV sero-incidence in men in western Kenya who receive medical male circumcision compared to those who didn't
- **We use this information to determine the sample size** that will be needed to detect this difference with 80% (or higher) power and a significance level of 0.05 (or lower)

Towards Sample Size - Inputs

- The technical, program and evaluation staff can come together to specify:

  - the **primary outcome, and other outcomes**.
  - a meaningful difference or **change** from the field experience or past similar studies…. “of program value” or “program effect”

Five Inputs. Count 'Em: 5.

1. Expected rate of outcome among exposed
2. Expected rate of outcome among unexposed
    - Or the baseline rate and expected difference
3. Number of controls per case (exposed) (can be same)
4. Desired power level 1 - β, standard = 0.80
5. Level of statistical significance desired.
    - α, standard = 0.05;

Also called parameters

Inputs (continued)

1. Expected rate of outcome among exposed
2. Expected rate of outcome among unexposed
   - Or this can be the change between two time points
   - Note: this presupposes that a) there is 1 primary indicator and b) there is some baseline level established that we can work with
   - We may want to choose the indicator that will need the largest sample size, if we have multiple indicators of interest

Inputs (continued)

3. Desired power level
   - 1 - β, standard (default)= 0.80
   - In some cases, we may want higher power (for example, often in clinical drug trials)

4. Number of controls for each case (exposed)
   - Do we expect the same number of participants in the control group as in the intervention group?

Inputs (continued)

5. Level of statistical significance desired
   - α, standard = 0.05; One-sided or two-sided?
   - Usually, two-sided test is chosen (noting that an outcome can increase or decrease, go in either direction).
   - But if we are confident that the outcome will go in one direction, we can use a one-sided test.)

Quick review: which are the five initial inputs to sample size?

1. ___
2. ___
3. ___
4. ___
5. ___

Caution: Some studies find no effect because they were underpowered

"…most of the 2000 randomized clinical trials underway worldwide are of 'little to no scientific value', based… on the fact that these studies are of **inadequate sample size** to detect effects reliably."

C.H. Hennekens. 1987. *Epidemiology in Medicine*. Boston: Little Brown and Co.
Reference to:
Peto R. 1982. *Statistics of Cancer Trials.* In KE Halnan (ed.) Treatment of Cancer. London: Chapman and Hall.

How to estimate the amount of change or difference to expect?

1. Literature review
2. Expert opinion
3. Pilot studies
4. Desired effect or minimal biologically or clinically meaningful difference to detect in the outcome
5. Existing data: Surveillance data, registry data, hospital data, chart review

"effect size" ==== "program effect"

Solving for Power or MDE

- Solve for Power…when we already know the sample size

- Solve for "**Minimal detectable difference or effect**" on the primary outcome – if you already know the maximum sample and the power used is standard.

Equations:
We don't usually solve them by hand

$$n = [Z_{1-\alpha/2}\,(2pq)^{1/2} + Z_{\beta}(p_1q_1 + p_0q_0)^{1/2}]^2 / (p_1 - p_0)^2$$

Where p = mean proportion of exposure = $(p_1 + p_0)/2$

p_o = proportion among the control participants

p_1 = proportion among the exposed intervention participants

$q = 1-p$, $q_0 = 1-p_0$, $q_1 = 1-p_1$

$$\frac{[(P_1 - P_0) - 1.96(p(1-p)(N_1^{-1} + N_0^{-1}))^{1/2}]}{N_1^{-1}p_{01}(1-p_1) + N_0^{-1}\,p_0(1-p_0))^{1/2}}$$

$$\frac{(1.96 + 0.841)^2(k+1)}{k4\lambda(R^{1/2}-1)^2}$$

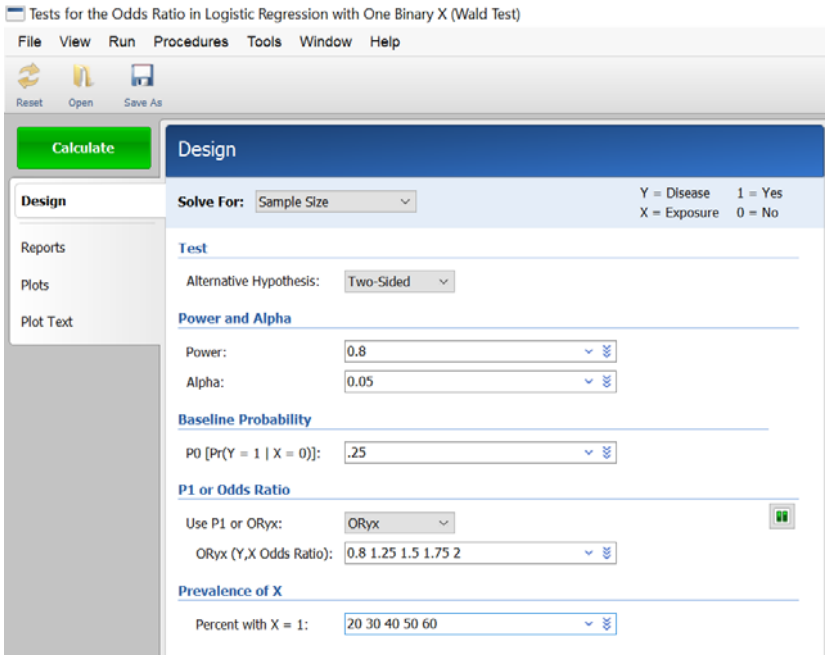
$$R = \lambda_1 / \lambda_0$$

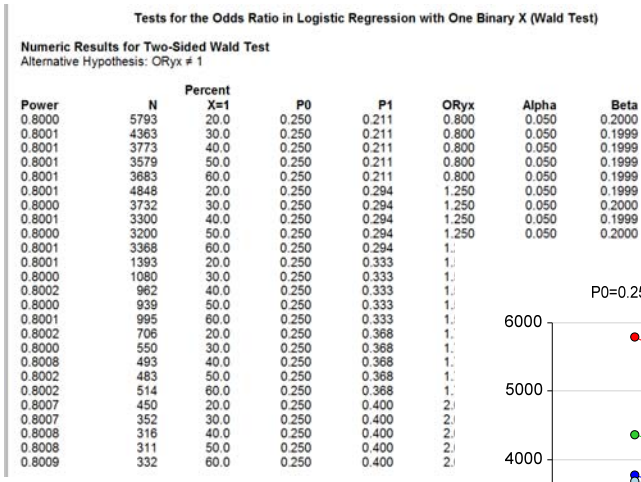
$$\frac{[1.96\,B_1^{1/2} + 0.841(B_1 + B_2 log\,(R))^{1/2}]^2}{[B_1\,log(R) + B_2 log(R)^2/2]^2}$$

R = Relative Hazard

Specialty software can help

- Stata, SAS, free R software, and others
- **PASS:** Expensive, user friendly, menu driven, "good looking" output. **Can free download 'trial version'**





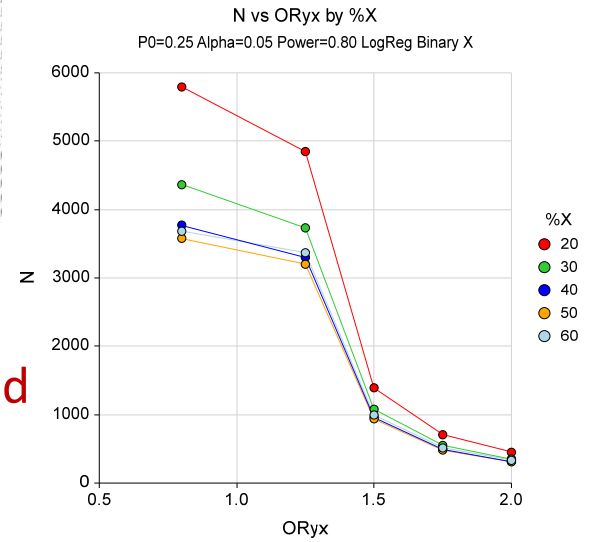
Tests for the Odds Ratio in Logistic Regression with One Binary X (Wald Test)

Numeric Results for Two-Sided Wald Test

Alternative Hypothesis: ORyx ≠ 1

Power	N	Percent X=1	P0	P1	ORyx	Alpha	Beta
0.8000	5793	20.0	0.250	0.211	0.800	0.050	0.2000
0.8001	4363	30.0	0.250	0.211	0.800	0.050	0.1999
0.8001	3773	40.0	0.250	0.211	0.800	0.050	0.1999
0.8001	3579	50.0	0.250	0.211	0.800	0.050	0.1999
0.8001	3683	60.0	0.250	0.211	0.800	0.050	0.1999
0.8001	4848	20.0	0.250	0.294	1.250	0.050	0.1999
0.8000	3732	30.0	0.250	0.294	1.250	0.050	0.2000
0.8001	3300	40.0	0.250	0.294	1.250	0.050	0.1999
0.8000	3200	50.0	0.250	0.294	1.250	0.050	0.2000
0.8001	3368	60.0	0.250	0.294	1.		
0.8001	1393	20.0	0.250	0.333	1.		
0.8000	1080	30.0	0.250	0.333	1.		
0.8002	962	40.0	0.250	0.333	1.		
0.8000	939	50.0	0.250	0.333	1.		
0.8001	995	60.0	0.250	0.333	1.		
0.8002	706	20.0	0.250	0.368	1.		
0.8000	550	30.0	0.250	0.368	1.		
0.8008	493	40.0	0.250	0.368	1.		
0.8002	483	50.0	0.250	0.368	1.		
0.8002	514	60.0	0.250	0.368	1.		
0.8007	450	20.0	0.250	0.400	2.		
0.8007	352	30.0	0.250	0.400	2.		
0.8008	316	40.0	0.250	0.400	2.		
0.8008	311	50.0	0.250	0.400	2.		
0.8009	332	60.0	0.250	0.400	2.		

All have manuals and YouTube videos! ☺



Example in the Literature:
A Cluster Randomized Trial Comparing Sand to Wood Chip Surfaces … Arm Fractures in Children

Howard AW, et al. *PLoS Med* 2011

- **Background:** Playground injuries are common. Surface area is affects severity. **Research Question:** Is there a difference in playground arm fracture rates in school playgrounds with wood fiber versus granite sand surfacing?

Example: Sample Size Estimation

Sample size. Based on retrospectively collected data from 1999–2001, a baseline arm fracture rate of 40 per 100,000 student-months was estimated. A clinically significant difference would be a halving of this rate to 20 per 100,000 student-months. Estimating 410 students per school provided 820 student years (y) of data over the 2-y study. Hayes' method of sample size estimation for cluster randomization was used [19]. Setting alpha = 0.05 and power at 80% and k (coefficient of variation between clusters) at 0.2, we estimated that 17 clusters per arm or 34 schools in total would be required.

5 initial inputs to sample size calculation:

Example: Sample Size Estimation

Sample size. Based on retrospectively collected data from 1999–2001, a baseline arm fracture rate of 40 per 100,000 student-months was estimated. A clinically significant difference would be a halving of this rate to 20 per 100,000 student-months. Estimating 410 students per school provided 820 student years (y) of data over the 2-y study. Hayes' method of sample size estimation for cluster randomization was used [19]. Setting alpha = 0.05 and power at 80% and *k* (coefficient of variation between clusters) at 0.2, we estimated that 17 clusters per arm or 34 schools in total would be required.

5 initial inputs to sample size calculation:

1. **What is the primary outcome? AND Baseline rate?**
2. **What is the expected difference**
3. **What is the alpha? (one or two-sided) p = ?**
4. **What is desired power ?**
5. **Number of control subjects for each intervention subject?**

Take Away Messages

The best estimates of sample size are the most accurate and most feasible

- Expected exposure and outcome rates
- Expected change or effect size or difference and direction
- Can I really get all of those people in my study?

Take Away Messages (cont'd)

- Sample size and power calculations are done according to the research question and study design;
- The process is *iterative.* **Estimate a range of plausible values.**

- **Technical, program and M&E** colleagues should be involved.
- Know the inputs
- Involve a statistician and have a software

PRACTICAL EXERCISES

Stata software --- One Sample, using proportion to compute sample size (Point Prevalence – Scenario 1)

- power oneproportion .5 .55

- Estimated sample size for a one-sample proportion test
- Score z test
- Ho: p = p0 versus Ha: p != p0

- Study parameters:
- alpha = 0.0500
- power = 0.8000
- delta = 0.1000
- p0 = 0.5000 (Estimated Baseline prevalence of 50%)
- pa = 0.6000 (Estimated Endline prevalence of 55%)

- Estimated sample size: **N = 783**

jhpiego
Saving lives. Improving health. Transforming futures.

Stata software --- One Sample, using proportion to compute sample size (Point Prevalence – Scenario 2)

- power oneproportion .5 .6

- Estimated sample size for a one-sample proportion test
- Score z test
- Ho: p = p0 versus Ha: p != p0

- Study parameters:
- alpha = 0.0500
- power = 0.8000
- delta = 0.1000
- p0 = 0.5000 (Estimated Baseline prevalence of 50%)
- pa = 0.6000 (Estimated Endline prevalence of 60%)

- Estimated sample size: **N = 194**

jhpiego
Saving lives. Improving health. Transforming futures.

Stata software --- One Sample, using proportion to compute sample size (Point Prevalence – Scenario 3)

- power oneproportion .5 .65

- Estimated sample size for a one-sample proportion test
- Score z test
- Ho: p = p0 versus Ha: p != p0

- Study parameters:
- alpha = 0.0500
- power = 0.8000
- delta = 0.1000
- p0 = 0.5000 (Estimated Baseline prevalence of 50%)
- pa = 0.6000 (Estimated Endline prevalence of 65%)

- Estimated sample size: **N = 85**

jhpiego
Saving lives. Improving health. Transforming futures.

Stata software --- One Sample, using proportion to compute sample size (Point Prevalence – Scenario 4a)

- power oneproportion .5 .7

- Estimated sample size for a one-sample proportion test
- Score z test
- Ho: p = p0 versus Ha: p != p0

- Study parameters:
- alpha = 0.0500
- power = 0.8000
- delta = 0.1000
- p0 = 0.5000 (Estimated Baseline prevalence of 50%)
- pa = 0.6000 (Estimated Endline prevalence of 70%)

- Estimated sample size: **N = 47 [Two-sided test]**

jhpiego
Saving lives. Improving health. Transforming futures.

Stata software --- One Sample, using proportion to compute sample size (Point Prevalence – Scenario 4b)

- power oneproportion .5 .7

- Estimated sample size for a one-sample proportion test
- Score z test
- Ho: p = p0 versus Ha: p != p0

- Study parameters:
- alpha = 0.0500
- power = 0.8000
- delta = 0.1000
- p0 = 0.5000 (Estimated Baseline prevalence of 50%)
- pa = 0.6000 (Estimated Endline prevalence of 70%)

- Estimated sample size: **N = 37 [One sided test]**

jhpiego
Saving lives. Improving health. Transforming futures.

Using Stata software --- Demonstration for Cluster randomized control trial sample size

Open Stata

- Download the free "`clustersampsi`" command

- For real-life scenario we faced, see 'Extra slides"

jhpiego
Saving lives. Improving health. Transforming futures.

Stata Output: Scenario 1

- clustersampsi, samplesize mu1(75) mu2(85) sd1(15) sd2(15) k(10) ho(0.15)
- Sample size calculation to determine number of observations required per cluster, for a two sample comparison of means (using normal approximations).
- mean 1: 75.00
- mean 2: 85.00
- standard deviation 1: 15.00
- standard deviation 2: 15.00
- significance level: 0.05
- power: 0.80
- baseline measures adjustment (correlation): 0.00
- number of clusters available: 10
- intra cluster correlation (ICC): 0.1500
- coefficient of variation (of cluster sizes): 0.00
- clustersampsi estimated parameters:
- Firstly, assuming individual randomisation:
- sample size per arm: 36
- Then, allowing for cluster randomisation:
- average cluster size required: 9
- sample size per arm: 90
- TOTAL SAMPLE SIZE 180

Stata Output: Scenario 2

- clustersampsi, samplesize mu1(75) mu2(85) sd1(15) sd2(15) k(8) rho(0.15)
- Sample size calculation to determine number of observations required per cluster, for a two sample comparison of means (using normal approximations).
- mean 1: 75.00
- mean 2: 85.00
- standard deviation 1: 15.00
- standard deviation 2: 15.00
- significance level: 0.05
- power: 0.80
- baseline measures adjustment (correlation): 0.00
- number of clusters available: 8
- Cluster trials with few clusters per arm (say less than 10)
- might be infeasible due to small number of randomisation units.
- intra cluster correlation (ICC): 0.1500
- coefficient of variation (of cluster sizes): 0.00
- clustersampsi estimated parameters:
- Firstly, assuming individual randomisation:
- sample size per arm: 36
- Then, allowing for cluster randomisation:
- average cluster size required: 18
- sample size per arm: 144
- sample size TOTAL: 288

Stata Output: Scenario 3

- clustersampsi, samplesize mu1(75) mu2(85) sd1(15) sd2(15) k(10) rho(0.13)
- Sample size calculation to determine number of observations required per cluster, for a two sample comparison of means (using normal approximations).
- For the user specified parameters:
- mean 1: 75.00
- mean 2: 85.00
- standard deviation 1: 15.00
- standard deviation 2: 15.00
- significance level: 0.05
- power: 0.80
- baseline measures adjustment (correlation): 0.00
- number of clusters available: 10
- intra cluster correlation (ICC): 0.1300
- coefficient of variation (of cluster sizes): 0.00
- clustersampsi estimated parameters:
- Firstly, assuming individual randomisation:
- sample size per arm: 36
- Then, allowing for cluster randomisation:
- average cluster size required: 7
- sample size per arm: 70

TOTAL SAMPLE SIZE 140

Stata Output: Scenario 4

- clustersampsi, samplesize mu1(75) mu2(85) sd1(15) sd2(15) k(8) rho(0.13)
- Sample size calculation to determine number of observations required per cluster, for a two sample comparison of means (using normal approximations).
- For the user specified parameters:
- mean 1: 75.00
- mean 2: 85.00
- standard deviation 1: 15.00
- standard deviation 2: 15.00
- significance level: 0.05
- power: 0.80
- baseline measures adjustment (correlation): 0.00
- number of clusters available: 8
- Cluster trials with few clusters per arm (say less than 10)
- might be infeasible due to small number of randomisation units.
- intra cluster correlation (ICC): 0.1300
- coefficient of variation (of cluster sizes): 0.00
- clustersampsi estimated parameters:
- Firstly, assuming individual randomization, sample size per arm: 36
- Then, allowing for cluster randomisation:
- average cluster size required: 13
- sample size per arm: 104

TOTAL SAMPLE SIZE 208

Summary of Scenarios of Sample Size

	Mean Score (a) of 75% and 85%. (Difference of 10%)
Clusters per arm, n=8	
ICC of 0.13	SD of 15 for each group. n=**104** per arm.
ICC of 0.15	SD of 15 for each group. **n=144** per arm.
Clusters per arm, n=10	
ICC of 0.13	SD of 15 for each group. n=**70** per arm.
ICC of 0.15	SD of 15 for each group. **n=90** per arm.

jhpiego
Saving lives. Improving health. Transforming futures.

Acknowledgements

- Prof. Supriya Mehta, University of Illinois at Chicago (UIC), for the first iteration of the sample size lecture

- Dr. Gayane Yenoykan, Johns Hopkins School of Public Health, for assistance with clustersampsi scenarios

Contact us at....

Dr. Eva Bazant

Eva.Bazant@Jhpiego.org

Dr. Mark Kabue

Mark.Kabue@Jhpiego.org

jhpiego

Saving lives. Improving health. Transforming futures.

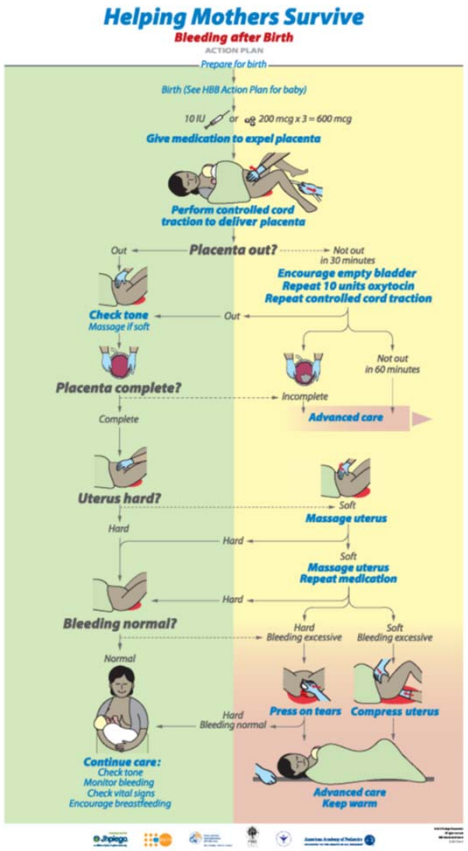
jhpiego
Saving lives. Improving health.
Transforming futures.

EXTRA SLIDES

Example: Additionally…

- What information was given in the Howard article abstract related to doing the study in clusters?
- Additional inputs are **intra-class correlation (ICC)** and sometimes, coefficient of variation (CoV) (a measure of heterogeneity, for example if we had a sample of hospitals and health centers)

Case Study (1)



- A Jhpiego colleague is a nurse/midwife & trainer and developing a new training module for health care providers. She wanted to supplement the training module with video clips of clinical scenarios. She wants to evaluate the training + video and is interested in provider learning.
- Colleague would like to compare health care providers who go through the training module and receive extra video supplementation to those who do not.
- There are 2 groups of health providers.
- Evaluation is a pre and post intervention.

Case Study (2)

- Discussion centered around the primary outcome of ‘learning’. How would this be measureable?
- There is a new evaluation tool being developed, a simulated skills demonstration called an objective structured clinical examination (OSCE). It is for providers going through a new training module on caring for women with normal birth in low-resource settings
- Is the outcome continuous (mean percent correct on exam) or is it binary (pass, yes/no)?

jhpiego
Saving lives. Improving health. Transforming futures.

Case Study (3)

- Hypothesis: Helping Mothers Survive Training with a live trainer and standardized videos supplementing usual materials will result in 10% higher health worker OSCE competency scores than demonstration by live trainer without video.
- The comparison between providers in two groups will be made at post-test. Looking at post-tests from past study OSCEs, at post-test, the scores are usually around 80% or higher.
- Providers will be invited from several hospitals in a country in East Africa. In past studies using other OSCEs, looking at the datasets, we can expect an intra-class correlation (ICC) of 0.10 to 0.15. "rho" (greek letter used in biostatistics)
- NOTE: ICC acts as a multiplier requiring more sample – and reaching a higher sample will be more expensive.

jhpiego
Saving lives. Improving health. Transforming futures.

Case Study (4)

- We expect providers in one group will have a post-training score of 75% and another group the score of 85% (10% point difference) with a standard deviation (SD) of 15% points (measure of variability, must be specified along with the mean).
- We expect providers to come from 8 to 10 health facilities or clusters (a number used our control).
- There is a budget for ~200 providers total.
- We will attempt to sample providers randomly from those eligible.