Sample size and power supplementary materials and R exercise

Drew Cameron

Intervention Trial Design PH226C
27 September 2019

Slides 34 – 53
(from Jack’s full presentation)

Example #1: Mobile app RCT (individual)

Original Paper

A Mobile App for the Self-Management of Type 1 Diabetes Among Adolescents: A Randomized Controlled Trial

Shivani Goyal[1,2*], BEng, MSc, PhD; Caitlin A Nunn[3*], MSc; Michael Rotondi[4], PhD; Amy B Couperthwaite[4], MSc; Sally Reiser[5], RD; Angelo Simone[5], MD; Debra K Katzman[6,7], MD, FRCP(C); Joseph A Cafazzo[1,2,8], PhD, PEng; Mark R Palmert[3,6,9], MD, PhD

[1]Centre for Global eHealth Innovation, Techna Institute, University Health Network, Toronto, ON, Canada

[2]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada

[3]Division of Endocrinology, The Hospital for Sick Children, Toronto, ON, Canada

[4]School of Kinesiology & Health Science, York University, Toronto, ON, Canada

Given parameters

“Sample size was determined based on a nominal 2-sided type 1 error rate of 5% and 80% power. Estimates of standard deviation in HbA1c ranging from 0.50 to 0.75 were used to determine the minimum number of participants required to detect a clinically relevant (>=0.5) change in HbA1c levels.” (Goyal et al, 2017)

Parameter	Value
Type 1 error rate (alpha)	
Power (1-beta)	
MDE (“clinically relevant” change in HbA1c levels)	
Standard deviation	
Baseline HbA1c (from eligibility criteria)	

Given parameters

“Sample size was determined based on a nominal 2-sided type 1 error rate of 5% and 80% power. Estimates of standard deviation in HbA1c ranging from 0.50 to 0.75 were used to determine the minimum number of participants required to detect a clinically relevant (>=0.5) change in HbA1c levels.” (Goyal et al, 2017)

Parameter	Value
Type 1 error rate (alpha)	0.05
Power (1-beta)	0.80
MDE (“clinically relevant” change in HbA1c levels)	>=0.5
Standard deviation	0.50-0.75
Baseline HbA1c (from eligibility criteria)	8.0-10.5

http://www.sample-size.net/sample-size-means/

Sample size – Means

Compare the mean of a continuous measurement in two samples

The sample sizes are calculated in two different ways: first using the T statistic (with a non-centrality parameter), then using the Z statistic. The Z statistic approximates the T statistic, but provides sample sizes that are slightly too small. (We provide the Z statistic calculation to allow comparison with other calculators which use the Z approximation.)

Instructions: Enter parameters in the red cells. Answers will appear in blue below.

α (two-tailed) =	0.05	*Threshold probability for rejecting the null hypothesis. Type I error rate.*
β =	0.2	*Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.*
q_1 =	0.5	*Proportion of subjects that are in Group 1 (exposed)*
q_0 =	0.500	*Proportion of subjects that are in Group 0 (unexposed); 1-q_1*
E =	0.5	*Effect size*
S =	0.75	*Standard deviation of the outcome in the population*

Calculate

Let's make the most conservative assumption here

Slide created by Yoshika Crider

http://www.sample-size.net/sample-size-means/

Sample size – Means

Compare the mean of a continuous measurement in two samples

The sample sizes are calculated in two different ways: first using the T statistic (with a non-centrality parameter), then using the Z statistic. The Z statistic approximates the T statistic, but provides sample sizes that are slightly too small. (We provide the Z statistic calculation to allow comparison with other calculators which use the Z approximation.)

Instructions: Enter parameters in the red cells. Answers will appear in blue below.

α (two-tailed) =	0.05	*Threshold probability for rejecting the null hypothesis. Type I error rate.*
β =	0.2	*Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.*
q_1 =	0.5	*Proportion of subjects that are in Group 1 (exposed)*
q_0 =	0.500	*Proportion of subjects that are in Group 0 (unexposed); 1-q_1*
E =	0.5	*Effect size*
S =	0.75	*Standard deviation of the outcome in the population*

Calculate

1. Calculation using the T statistic and non-centrality parameter:

N_1: 37
N_0: 37
Total: 74

2. Normal approximation using the Z statistic instead of the T statistic:

$A = (1/q_1 + 1/q_0) = 4.00000$

$B = (Z_\alpha + Z_\beta)^2 = 7.84887$

Total group size = $N = AB/(E/S)^2 = 70.640$

N_1: 36
N_0: 35
Total: 71

This formula uses the Z statistic to approximate the T statistic. As a result it slightly underestimates the sample size. We provide this approximation to allow comparison to other calculators that use the Z statistic.

But their final sample size was 46 per arm, 92 total participants.... Why?

http://www.sample-size.net/sample-size-means/

Sample size – Means

Compare the mean of a continuous measurement in two samples

The sample sizes are calculated in two different ways: first using the T statistic (with a non-centrality parameter), then using the Z statistic. The Z statistic approximates the T statistic, but provides sample sizes that are slightly too small. (We provide the Z statistic calculation to allow comparison with other calculators which use the Z approximation.)

Instructions: Enter parameters in the red cells. Answers will appear in blue below.

α (two-tailed) =	0.05	*Threshold probability for rejecting the null hypothesis. Type I error rate.*
β =	0.2	*Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.*
q_1 =	0.5	*Proportion of subjects that are in Group 1 (exposed)*
q_0 =	0.500	*Proportion of subjects that are in Group 0 (unexposed); 1-q_1*
E =	0.5	*Effect size*
S =	0.75	*Standard deviation of the outcome in the population*

Calculate

1. Calculation using the T statistic and non-centrality parameter:

N_1: 37
N_0: 37
Total: 74

2. Normal approximation using the Z statistic instead of the T statistic:

$A = (1/q_1 + 1/q_0) = 4.00000$

$B = (Z_\alpha + Z_\beta)^2 = 7.84887$

Total group size = $N = AB/(E/S)^2 = 70.640$

N_1: 36
N_0: 35
Total: 71

This formula uses the Z statistic to approximate the T statistic. As a result it slightly underestimates the sample size. We provide this approximation to allow comparison to other calculators that use the Z statistic.

Why?

Buffered for up to 25% loss to follow up (37*1.25=46)

Let’s try this in Stata…

(We get more flexibility to test out different assumptions.)

What are we telling Stata?

```
power twomeans 8 8.5, sd(0.75)
```

power defaults: alpha = 0.05, power = 0.80, two-tailed

Parameter	Value
Type 1 error rate (alpha)	0.05
Power (1-beta)	0.80
MDE (“clinically relevant” change in HbA1c levels)	>=0.5
Standard deviation	0.50-0.75
Baseline HbA1c (from eligibility criteria)	8.0-10.5

Real time in Stata…

```
SET BASED ON BASELINE'S LOW-END MEAN:  power twomeans ??  ?? , sd(??)

OR

SET BASED ON BASELINE'S HIGH-END MEAN: power twomeans ??  ?? , sd(??)
```

What's the difference?

```
. power twomeans 8 8.5, sd(0.75)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.5000
           m1 =    8.0000
           m2 =    8.5000
           sd =    0.7500

Estimated sample sizes:

            N =        74
  N per group =        37
```

```
. power twomeans 10 10.5, sd(0.75)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.5000
           m1 =   10.0000
           m2 =   10.5000
           sd =    0.7500

Estimated sample sizes:

            N =        74
  N per group =        37
```

Alternate code: `power twomeans 10, sd(0.75) diff(0.5)`

What's the difference?

```
. power twomeans 8 8.5, sd(0.75)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.5000
           m1 =    8.0000
           m2 =    8.5000
           sd =    0.7500

Estimated sample sizes:

            N =        74
  N per group =        37
```

```
. power twomeans 10 10.5, sd(0.75)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.5000
           m1 =   10.0000
           m2 =   10.5000
           sd =    0.7500

Estimated sample sizes:

            N =        74
  N per group =        37
```

Alternate code: `power twomeans 10, sd(0.75) diff(0.5)`

What if we assume a smaller sd?

Real time in Stata…

```
SET USING BASELINE'S HIGH-END WITH LESS-CONSERVATIVE SD:

power twomeans 10  10.5 , sd(??)
```

What if we assume a smaller sd?

```
. power twomeans 10 10.5, sd(0.5)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.5000
           m1 =   10.0000
           m2 =   10.5000
           sd =    0.5000

Estimated sample sizes:

            N =        34
  N per group =        17
```

Real time in Stata…
Graphing MDE & sample size tradeoffs

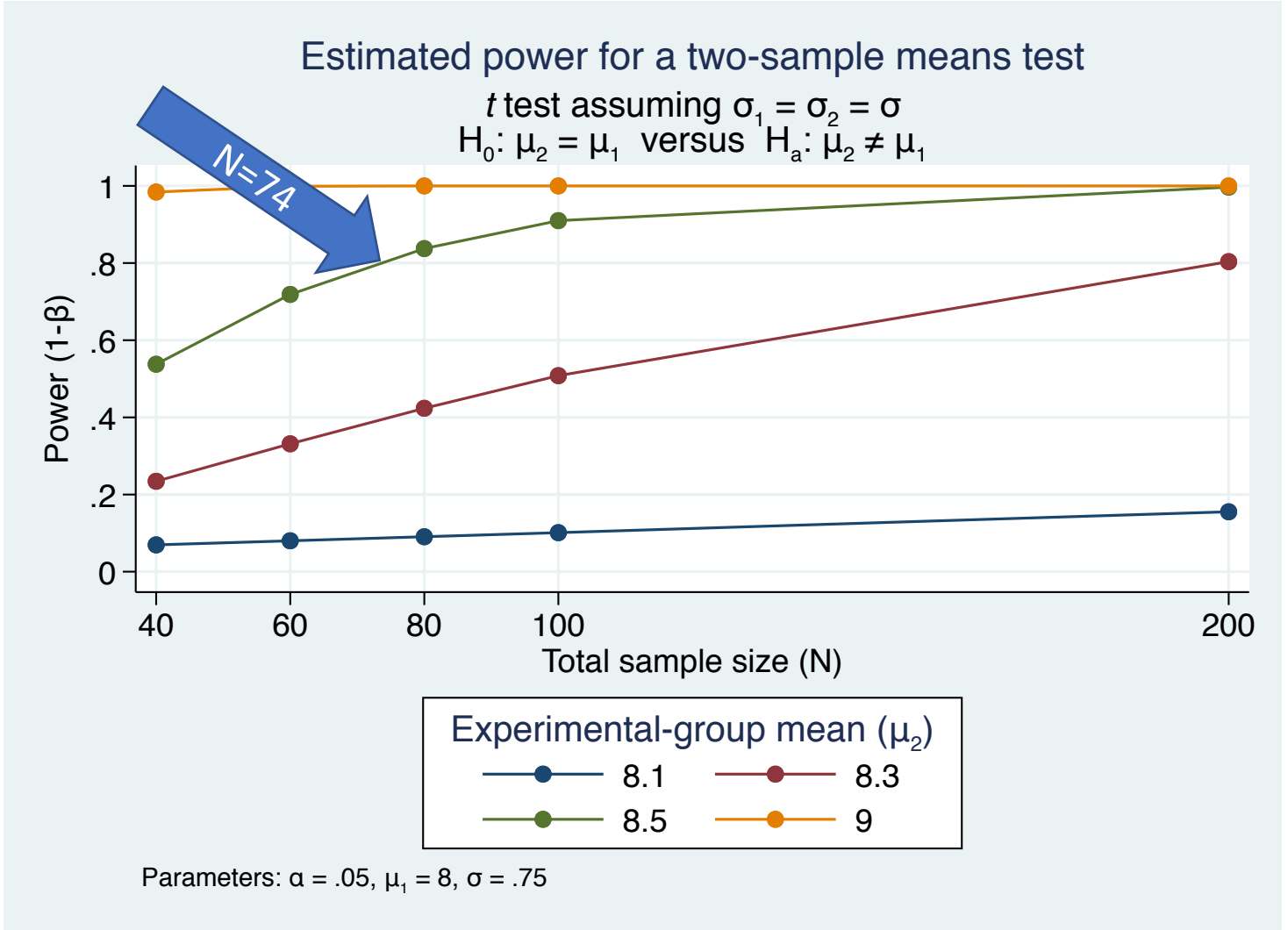
```
power twomeans 8 (?? ?? ?? ??), n(40 60 80 100 200) sd(.75) graph
```

Real time in Stata…
Graphing MDE & sample size tradeoffs

```
power twomeans 8 (8.1 8.3 8.5 9), n(40 60 80 100 200) sd(.75) graph
```

Real time in Stata…
Graphing MDE & sample size tradeoffs

```
power twomeans 8 (8.1 8.3 8.5 9), n(40 60 80 100 200) sd(.75) graph
```

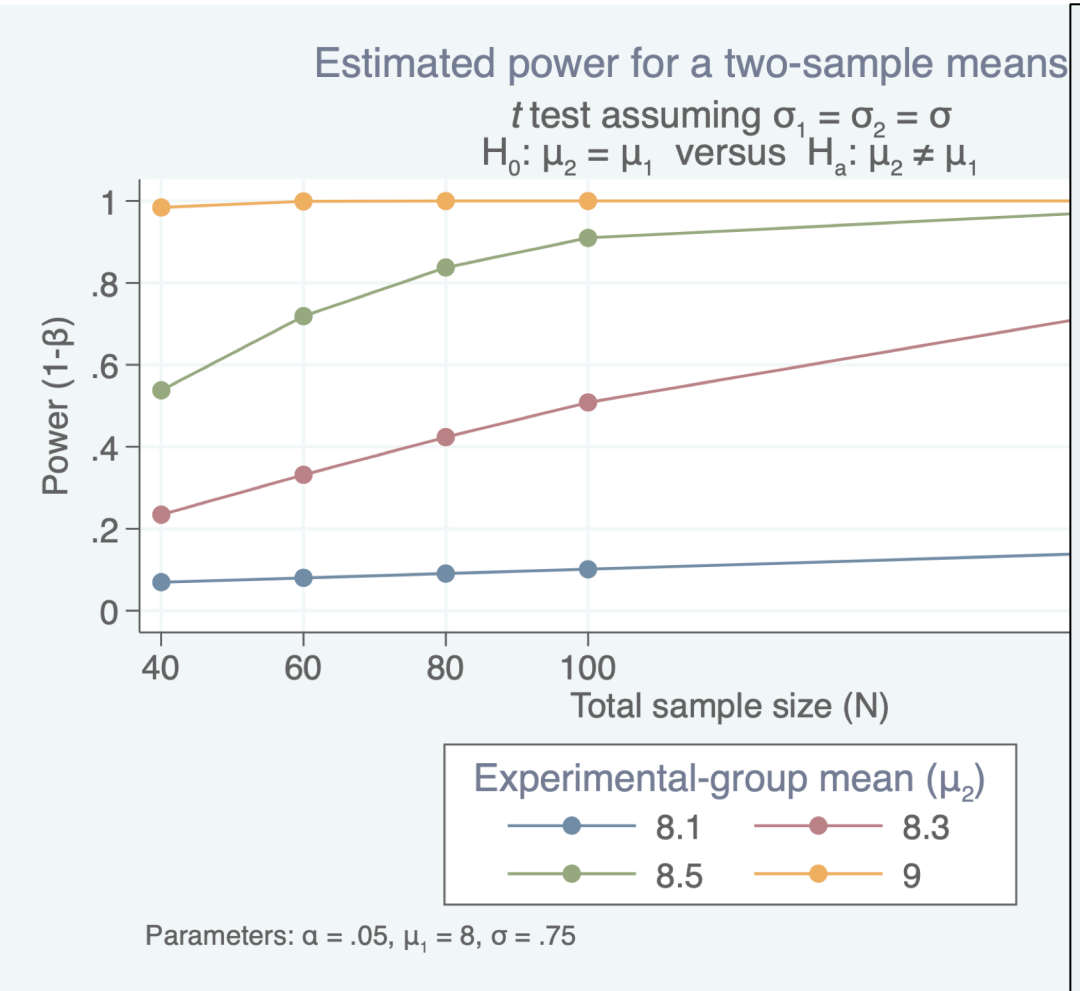


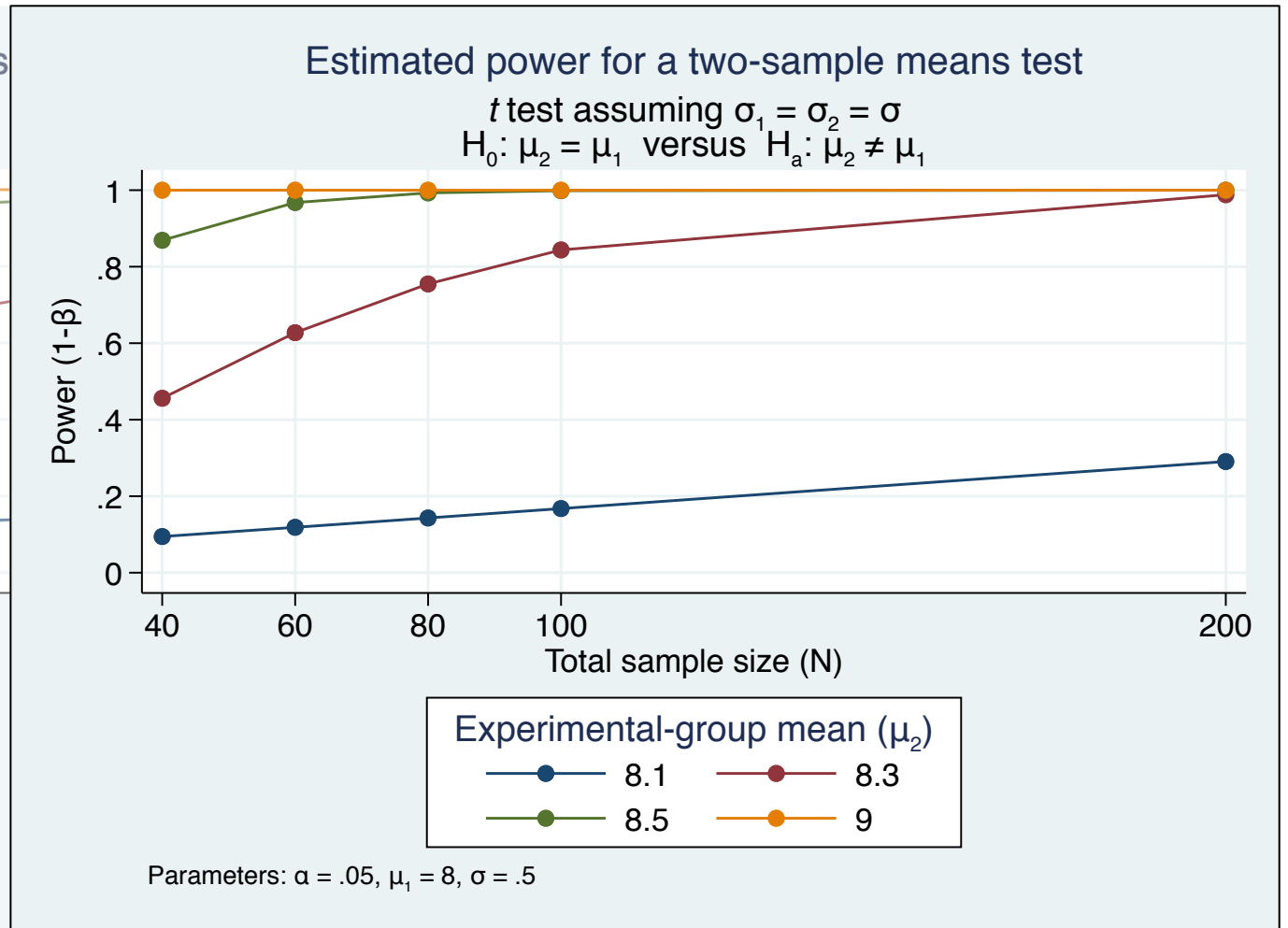
Real time in Stata…
Graphing MDE & sample size tradeoffs

```
power twomeans 8 (8.1 8.3 8.5 9), n(40 60 80 100 200) sd(??) graph
```

Real time in Stata…
Graphing MDE & sample size tradeoffs

```
power twomeans 8 (8.1 8.3 8.5 9), n(40 60 80 100 200) sd(.5) graph
```





Real time in Stata…
MDE & sample size tradeoffs in Table form

```
power twomeans 8 (8.1 8.3 8.5 9), n(40 60 80 100 200) sd(.75) table
```

alpha	power	N	N1	N2	delta	m1	m2	sd
.05	.06957	40	20	20	.1	8	8.1	.75
.05	.08004	60	30	30	.1	8	8.1	.75
.05	.09059	80	40	40	.1	8	8.1	.75
.05	.1012	100	50	50	.1	8	8.1	.75
.05	.1553	200	100	100	.1	8	8.1	.75
.05	.2343	40	20	20	.3	8	8.3	.75
.05	.3315	60	30	30	.3	8	8.3	.75
.05	.4235	80	40	40	.3	8	8.3	.75
.05	.5082	100	50	50	.3	8	8.3	.75
.05	.8036	200	100	100	.3	8	8.3	.75
.05	.5378	40	20	20	.5	8	8.5	.75
.05	.7187	60	30	30	.5	8	8.5	.75
.05	.8376	80	40	40	.5	8	8.5	.75
.05	.91	100	50	50	.5	8	8.5	.75
.05	.9968	200	100	100	.5	8	8.5	.75
.05	.9841	40	20	20	1	8	9	.75
.05	.9991	60	30	30	1	8	9	.75
.05	1	80	40	40	1	8	9	.75
.05	1	100	50	50	1	8	9	.75
.05	1	200	100	100	1	8	9	.75

N=74

Be sure to check out:
Slides 80-94

(from Jack's expanded lecture slides – these were not included in the class lecture)

Additional sample size resources

- www.power-calculator.org
  - RCT, cluster RCT calculations for continuous and binary outcomes (can be super buggy and slow though!)
- https://jadebc.shinyapps.io/samplesize/
  - Individually randomized trial calculations for continuous and binary outcomes
  - Shows curves to visualize trade-offs in parameters
- http://www.sample-size.net/
  - Individual and clustered design options
  - No visualization, just table output
- https://ssc.researchmethodsresources.nih.gov/ssc/
  - Group trial calculations
  - Lots of parameters that can get a little complicated
- **STATA** – take note of the defaults (e.g. power = 0.90 for *sampsi*, 0.80 for *power*)
  - Can use *sampsi* or *power*
  - *https://www.stata.com/features/power-and-sample-size/*
- Djimeu and Hondoulo (2014; 2016)
  - Draft article: https://www.3ieimpact.org/file/8081/download?token=CoEfkFc4
  - Spreadsheet tool: https://www.3ieimpact.org/sites/default/files/2017-11/3ie-sample-size-minimum-detectable-effect-calculator.xlsx
  - Landing page: https://www.3ieimpact.org/evidence-hub/publications/working-papers/power-calculation-causal-inference-social-science-sample
  - Note there is a published version we put in bCourses, but it is gated online so you'll need library access to find it online!

Slide created by Yoshika Crider w. amendments from Drew Cameron

Power calculations

(Drew's dissertation as an example)

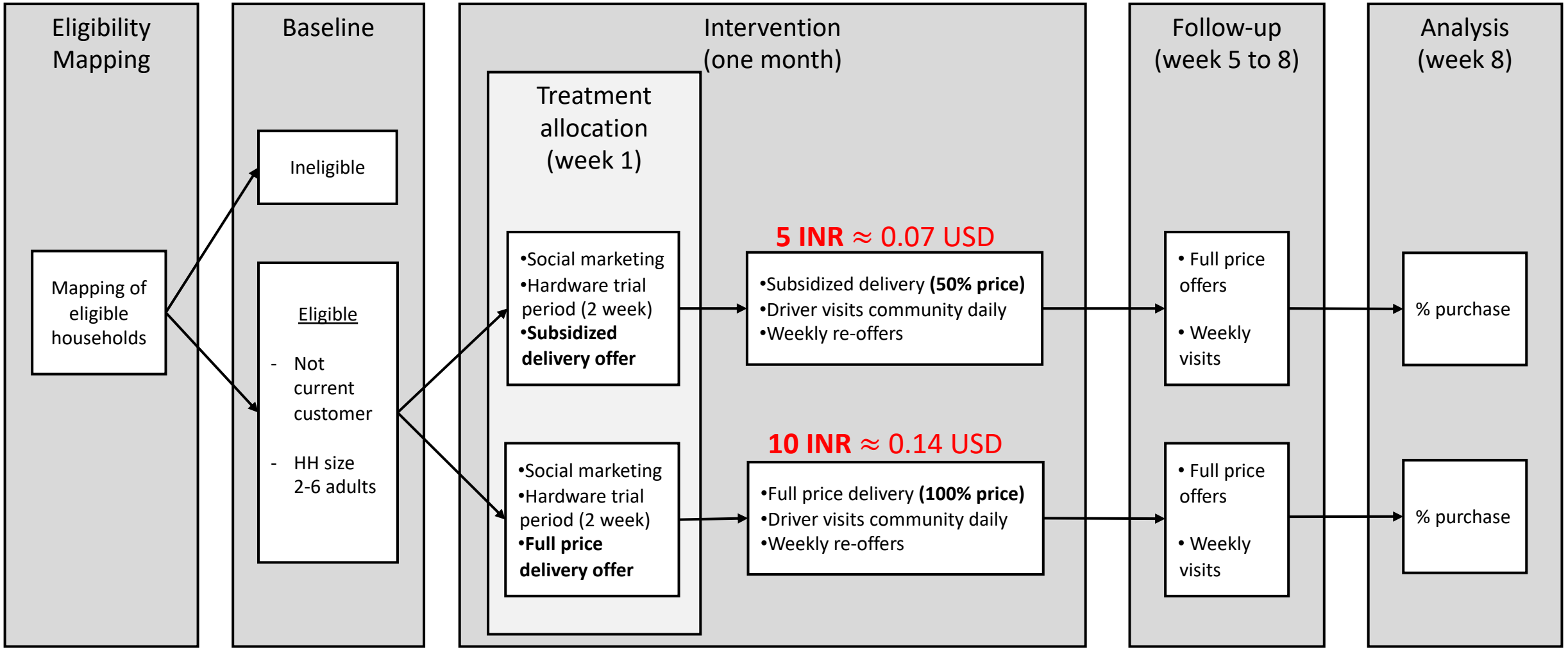
Research question

- What is the impact of subsidies for daily potable water delivery on future demand?

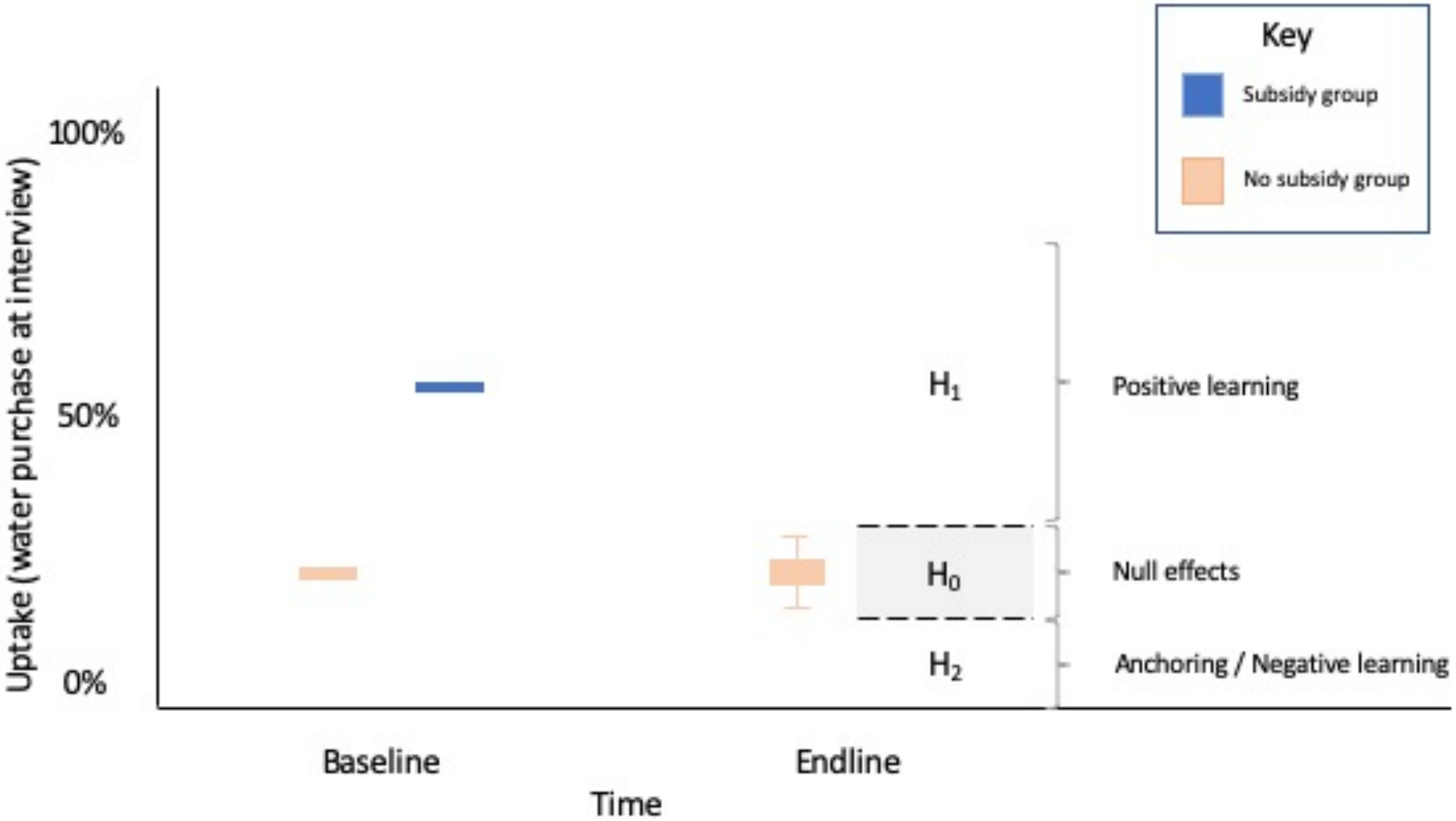
  - H_0 – There is no difference in demand between the two groups after two months
  - H_1 – The treatment group has greater demand after two months
  - H_2 – The treatment group has less demand after two months

  - But how did I decide the MDE?

Study design



Proposed analysis



Power calculations (two proportions test)

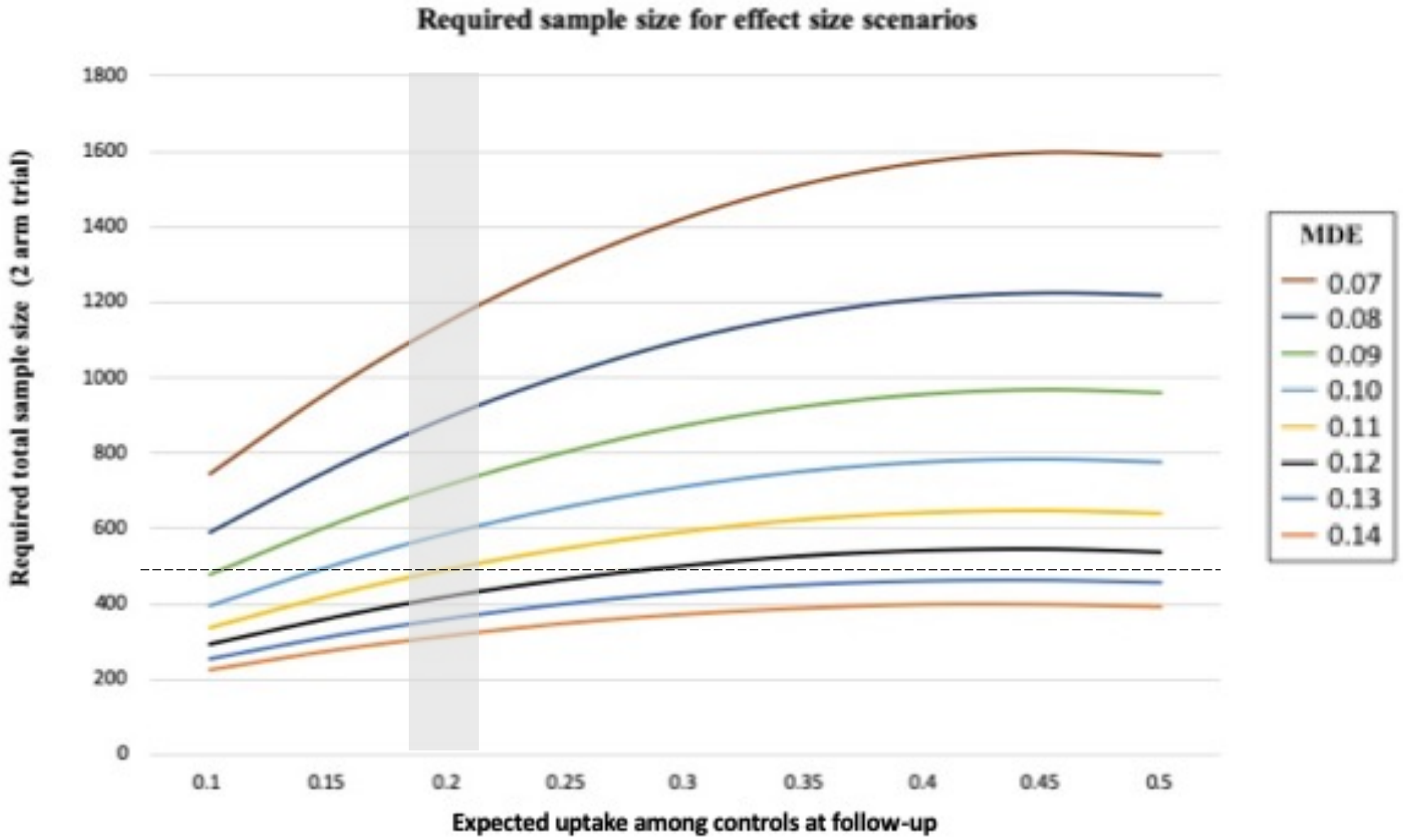
$$n = \left\{ \frac{P}{T\delta^2} + \frac{-P+1}{-T+1}(-t_1 - t_2)^2 \right\}$$

Two arms

- Two-tailed test
- $\alpha = 0.05; \beta = 0.80$, thus $t_1 = 2.04; t_2 = 0.85$
- Population proportion w/out intervention, $P = 0.20$ (Deliare et al. 2017)
- Proportion of sample randomly assigned treatment, $T = 0.5$ (Hasselblad 2016)
- MDE 10 percentage-point difference in uptake, $\delta = 0.1$ (Dupas 2014; Fischer 2018)
- **538 households**

- 5% attrition, 1.05(805) = 830 (Dupas 2014; Wright 2016; Fischer 2018; Burt 2017)
- **564 total households → 269 per arm**
- With covariates, where: $R^2 = 0.1; 0.3; 0.5$ → $\text{MDE} = 0.09; 0.08; 0.07$, respectively

Power sensitivity cont'd



47

Power sensitivity

$$n = \left\{ \frac{P}{T\delta^2} + \frac{-P+1}{-T+1}(-t_1 - t_2)^2(-R^2+1) \right\}$$

MDE given proportion of outcome variance explained by level-1 covariates

n	R^2	*MDE*
values specific to 507 households in 2 arms before adjusting for 5% attrition	0.00	0.097
	0.05	0.094
	0.10	0.092
	0.15	**0.089**
	0.20	0.087
	0.25	0.084
	0.30	0.081
	0.35	**0.078**
	0.40	0.075
	0.45	0.072
	0.50	**0.069**

Power (ex-post or after the study was done)

We selected an individually randomized RCT.
After a the study was over, we had recruited 526 households at baseline, dropping to 503 at endline (4.3% attrition over two-months), had randomized 52.9% into the treatment group, and had 13% purchase in the control group at follow up…

$two - tailed\ test$
$\alpha = 0.05$
$\beta = 0.8$
$\rho = 0.130$ ←**We had far less take-up in the control group than we were worried about!**
$\tau = 0.5285$ ← Balance was pretty good
$n = 503$ ← Attrition was lower than expected (~4.5% instead of 5%)
$\boldsymbol{\delta = 0.084}$

$R^2 = 0.323$ ← Individual-level controls also improved our estimate substantially!
$\boldsymbol{\delta = 0.069}$

Power sensitivity cont'd (what if I could afford a cluster RCT?)

$$J = 1 + \frac{(z_1 + z_2)^2[\frac{\mu_0(1-\mu_0)}{n} + \frac{\mu_0(1-\mu_1)}{n} + k^2(\mu_0^2 + \mu_1^2)]}{(\mu_0 - \mu_1)^2}$$

- Two Arms
  - 2 tailed test
  - $\alpha = 0.05$; $\beta = 0.80$, thus $z_1 = 1.96$; $z_2 = 0.84$
  - $n = 4$ number of households per cluster
  - $\mu_0 = 0.2$; $\mu_1 = 0.3$ 10pp difference in uptake between groups
  - $k = 0$ Intra-cluster correlation
  - $\boldsymbol{J = 74}$ **number of clusters per arm**

  - 296 households per arm = 592 households
  - **Attrition = 1.05(592) = 622 households**

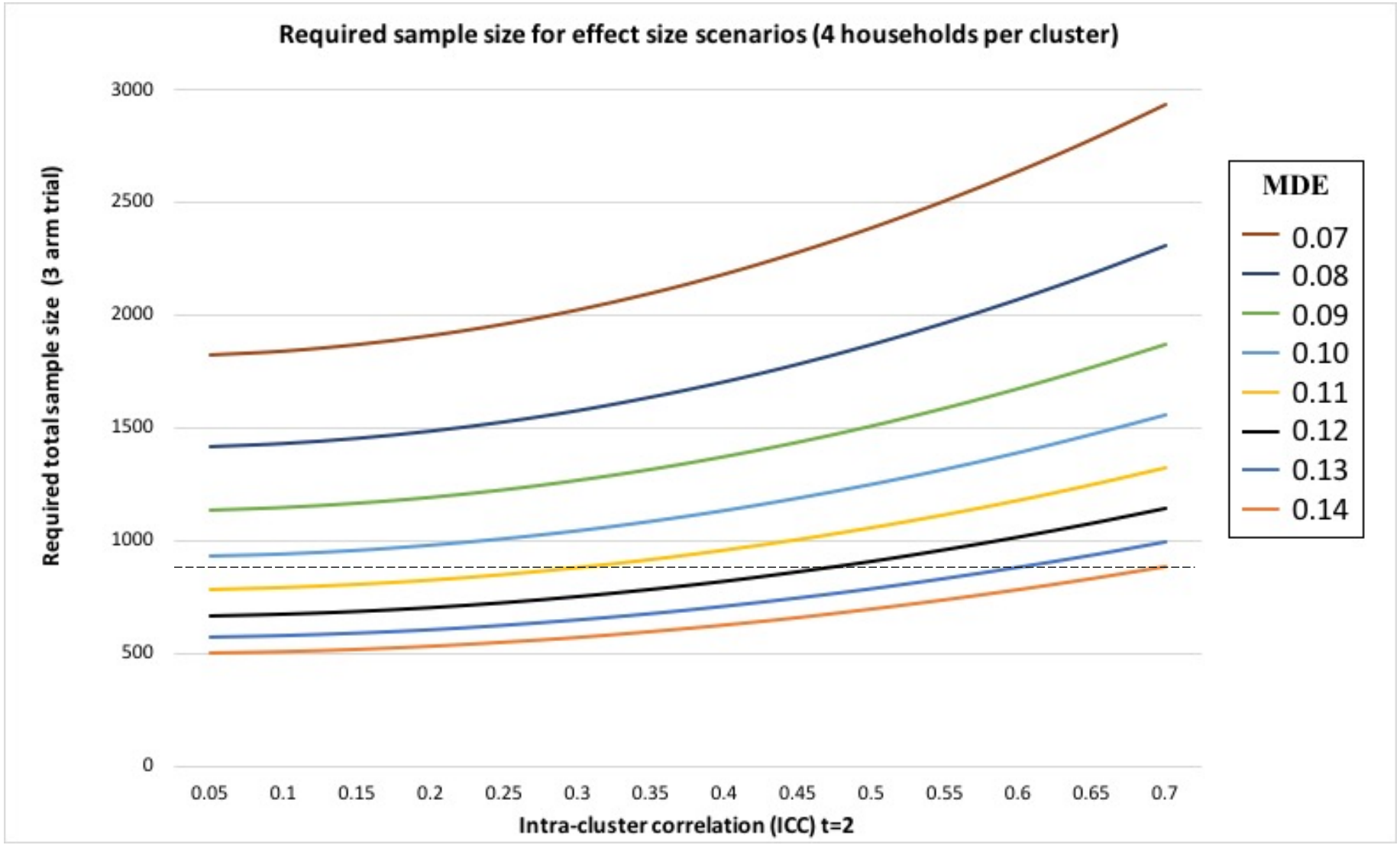
Power sensitivity cont'd (cluster RCT)

Sensitivity by number of households per cluster

Households per cluster	*Number of clusters per arm*	*Clusters (2 arms)***	*Total households needed (2 arms)***	*Total clusters needed (3 arms)***	*Total households needed (3 arms)***
1	291	612	612	*917*	*917*
2	146	307	614	*460*	*920*
3	98	206	618	*309*	*927*
4	**74**	**156**	**624**	***234***	***936***
5	59	124	620	*186*	*930*
6	49	103	618	*155*	*930*

Notes: **Total households needed (2 and 3 arms) includes 5% attrition rate; Estimates assume a .10 MDE between any two arms

Power sensitivity cont'd



Appendix: Power calculations (ICC scenarios for cluster RCT)