Berkeley School of Public Health

Propensity score matching

PHW250 B - Andrew Mertens

Epidemiologic analysis topics (already covered)

- Overview of epidemiologic analyses
- Types of variables
- Types of analyses
  - Univariable analyses
  - Bivariable analyses
  - Multivariable analyses for
    - Linear regression
    - Logistic regression
    - Log-linear regression
- Statistical modeling for other types of data
  - Longitudinal data
  - Repeated measures data
  - Survival data
  - Matched data

Background

- Traditional approaches to handling confounding in the analysis phase:
  - Stratification
  - Matching
  - Multivariable regression models
- Modern methods from the causal inference literature
  - Propensity score matching
  - Inverse probability of treatment weighting
  - G-computation
  - Double robust estimation
  - Instrumental variables
  - Mendelian randomization

Multivariable models are biased in the presence of time-dependent confounding

Not covered in this course.

Learning objectives related to propensity scores

- Define a propensity score
- Explain how propensity scores attempt to approximate counterfactuals
- In words, explain how propensity scores are estimated
- Explain the purpose of propensity score matching
- Describe the process of propensity score matching
- Make a causal interpretation of the results of a propensity score matched analysis

Definition of propensity scores

- **Propensity score:** the predicted probability ("propensity") of exposure in a particular individual based on a set of characteristics

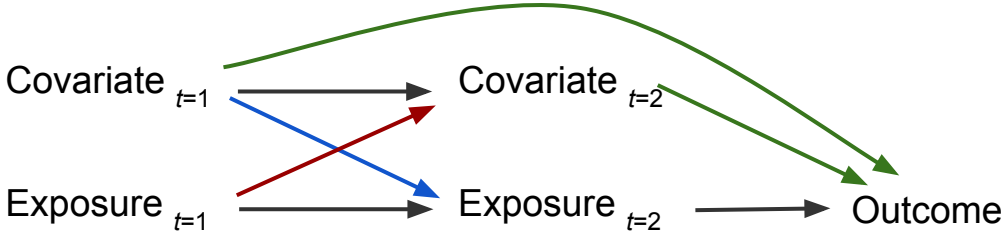
(Szklo & Nieto, 3rd) Ed.

Why use propensity scores?

- **Common use 1:** study with time-dependent confounding
  - Adjusting for a time-dependent confounder changes the quantity that we estimate because it is on the causal path from the exposure to the outcome.
  - If we don't adjust for a time-dependent confounder, our estimate will be confounded.
- **Common use 2:** You want to mimic randomization in an observational study (often one that evaluates an existing intervention).
  - Use propensity scores to identify who can be enrolled as a comparison group, attempting to mimic randomization by finding untreated individuals who are ideally exchangeable with those who were treated.

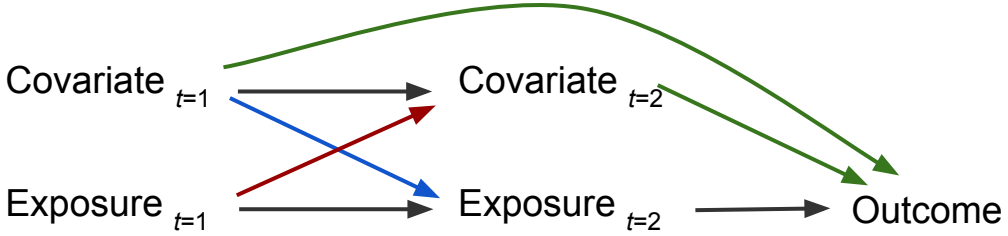
Time dependent confounding

- A time-dependent confounder is a variable that:
  - Is affected by prior exposure
  - Predicts subsequent exposure
  - Associated with / causes the outcome
- In the DAG below we use subscripts to indicate the time (t) of measurement.



Time dependent confounding

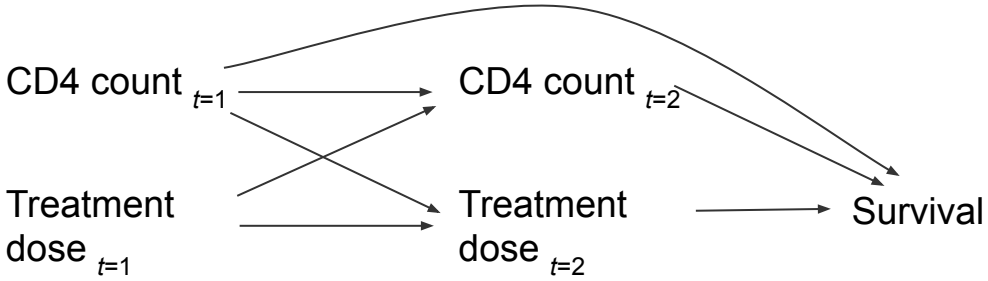
- A time-dependent confounder is a variable that:
  - Is affected by prior exposure
  - Predicts subsequent exposure
  - Associated with / causes the outcome
- In the DAG below we use subscripts to indicate the time (t) of measurement.



- Adjusting for a time-dependent confounder changes the quantity that we estimate because it is on the causal path from the exposure to the outcome.
- If we don't adjust for a time-dependent confounder, our estimate will be confounded.

Example of time dependent confounding

- The HIV virus reduces the number of CD4 lymphocyte cells (T cells) in the body when untreated.
- A patient's CD4 count affects a physician's choice of treatment dose for an HIV patient, and the treatment dose can in turn affect later CD4 count.



Why use propensity scores?

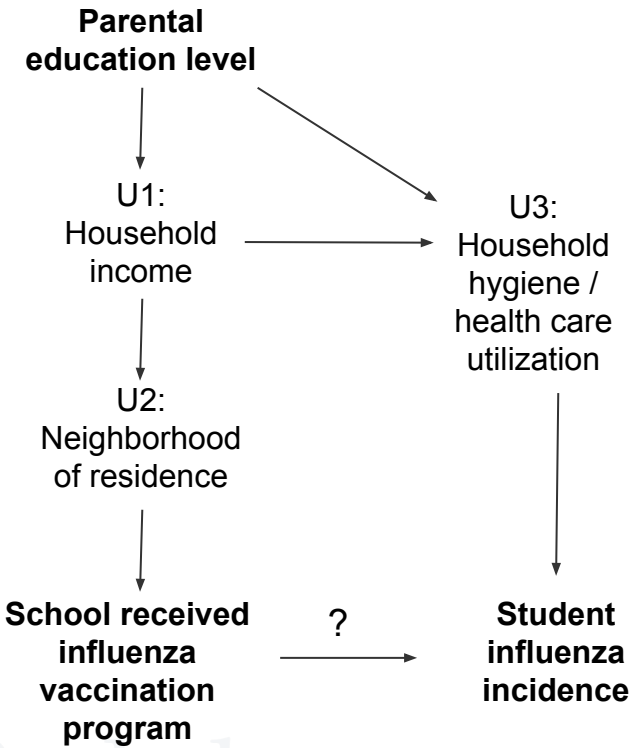
- **Common use 1:** study with time-dependent confounding
  - Adjusting for a time-dependent confounder changes the quantity that we estimate because it is on the causal path from the exposure to the outcome.
  - If we don't adjust for a time-dependent confounder, our estimate will be confounded.
- **Common use 2:** You want to mimic randomization in an observational study (often one that evaluates an existing intervention).
  - Use propensity scores to identify who can be enrolled as a comparison group, attempting to mimic randomization by finding untreated individuals who are ideally exchangeable with those who were treated.

Propensity score matching to evaluate an existing intervention

- Example: a school-located influenza vaccination program was deployed in 30 schools in the San Francisco Bay Area in 2014.
  - Example is loosely based on an evaluation of the Shoo the Flu Program
- You want to rigorously evaluate whether the program reduced influenza among students and their household members.
- However, you must use an observational design because the program did not want to randomize schools to the program vs. control.

Propensity score matching to evaluate an existing intervention

- **Step 1:** Obtain publicly available pre-intervention data from 2013 from the California Department of Education on all schools in the San Francisco Bay Area.
- Variables include:
  - Average class size
  - Student race
  - Parent education level
  - Standardized test scores
  - % of students receiving free and reduced price lunch
  - % of English language learners
- Goal is to obtain data on variables that are potential confounders.



Propensity score matching to evaluate an existing intervention

- **Step 2:** Estimate the probability of being treated or exposed conditional on a set of characteristics using a regression model (typically logistic regression)
  - Y = influenza incidence
  - A = child attends school with influenza vaccination program
  - W = confounders

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$

$$\ln\left(\frac{\Pr(A|W=w)}{1-\Pr(A|W=w)}\right) = -1 + 2 \cdot W_1 + 3 \cdot W_2$$

$$\Pr(A|W=w) = \frac{1}{1+e^{-(-1+2\cdot W_1+3\cdot W_2)}}$$

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\Pr(A|W = w) = \frac{1}{1 + e^{-(-1+2\cdot W_1+3\cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*0 + 3*0))) = 0.27

Propensity score matching to evaluate an existing intervention

- **Step 3:** Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each school

$$\Pr(A|W = w) = \frac{1}{1 + e^{-(-1+2\cdot W_1+3\cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*0 + 3*0))) = 0.27

88% probability that id 2 was treated

89% probability that id 3 was treated

Propensity score matching to evaluate an existing intervention

- **Step 4:** Match each treated school with an untreated school with a similar probability of being treated conditional on covariates.

$$\Pr(A|W = w) = \frac{1}{1 + e^{-(-1+2\cdot W_1+3\cdot W_2)}}$$

id	A	W_1	W_2	Pr(A|W=w) (Propensity scores)
1	1	1	0	1 / (1 + exp(-(-1 + 2*1 + 3*0))) = 0.29
2	0	0	1	1 / (1 + exp(-(-1 + 2*0 + 3*1))) = 0.88
3	1	1	1	1 / (1 + exp(-(-1 + 2*1 + 3*1))) = 0.89
4	0	0	0	1 / (1 + exp(-(-1 + 2*0 + 3*0))) = 0.27

Untreated id

Treated id

Match ids with similar propensity scores

Summary of steps

1. Obtain pre-intervention data on treated and untreated individuals or clusters.
2. Estimate the probability of being treated or exposed conditional on a set of characteristics using a regression model (typically logistic regression)
3. Use the regression model fit to estimate the probability of being treated or exposed conditional on a set of characteristics for each individual or cluster.
4. Match each treated individual or cluster with an untreated individual or cluster with a similar probability of being treated conditional on covariates.
    - Usually this means identifying a subset of untreated individuals / clusters with similar probabilities of treatment to those of the treated individuals. These individuals are enrolled as the control group.
    - Individuals who are not matched are dropped from the analysis. This means that we are essentially estimating our measure of association on a population that is similar to those who received treatment.

Analysis and interpretation of propensity score matched data

- The analysis of a propensity score matched study must account for the matching.
- Interpretation: Estimate of the average association between treatment and the outcome for the treated population (if all A=1 matched to an A=0)
- **Statistical notation:**
  - RD: $E[Y|A=1, W] - E[Y|A=0, W]$
  - RD in propensity matched study: $E_w\{E[Y|A=1, W]|A=1\} - E_w\{E[Y|A=0, W]|A=1\}$
- **Counterfactual notation:**
  - RD: $E[Y_1] - E[Y_0]$
  - RD in propensity matched study: $E[Y_1|A=1] - E[Y_0|A=1]$

Summary of key points

- Propensity score: the predicted probability (“propensity”) of exposure in a particular individual based on a set of characteristics
- Common uses:
  - Study with time-dependent confounding
  - You want to mimic randomization in an observational study (often one that evaluates an existing intervention).
- Analysis of propensity score matched data must account for that matching.
- The interpretation of measures of association is different in a propensity score matched study.

Berkeley School of Public Health

IPTW & G-Computation

PHW250 B - Andrew Mertens

Learning objectives related to these topics

- For inverse probability of treatment weights / G-computation
  - Explain its purpose
  - Explain its process
  - Explain how it attempts to approximate counterfactuals
  - Explain how to interpret its results
  - Describe limitations
- Explain how IPTW is related to standardization
- Compare and contrast IPTW and G-computation approaches

Inverse Probability of Treatment Weighting

IPTW: Summary of steps

1. Estimate the **propensity score** — the probability of being treated or exposed conditional on a set of characteristics (typically use logistic regression)
2. Define weights based on the inverse propensity score (i.e. the inverse probability of treatment)
    - Weight for exposed individuals = 1 / PS
    - Weight for unexposed individuals = 1 / (1-PS)
    - This is the simplest form of weights. Many variations exist.
3. Apply weights to the data, creating a “pseudo-population” to approximate two counterfactual populations
4. Calculate the mean of the outcome times the weight
    - If the backdoor criterion holds conditional on W, then the inverse probability weighted measure of association should obtain the causal effect of the exposure or treatment

Example of IPTW

- **Example**: an observational study whether influenza vaccination reduces the risk of visiting a doctor's office for influenza-like symptoms

	Doctor visit	No doctor visit
Influenza vaccine	128	282
No influenza vaccine	64	26

Crude RR = 0.44

Example of IPTW

- **Example**: an observational study whether influenza vaccination reduces the risk of visiting a doctor's office for influenza-like symptoms

	Doctor visit	No doctor visit
Influenza vaccine	128	282
No influenza vaccine	64	26

Crude RR = 0.44

Confounding is present!

Age <65 years		
	Doctor visit	**No doctor visit**
Influenza vaccine	108	252
No influenza vaccine	24	16

RR if Age <65 years = 0.5

Age ≥ 65 years		
	Doctor visit	**No doctor visit**
Influenza vaccine	20	30
No influenza vaccine	40	10

RR if Age ≥ 65 years = 0.5

IPTW Step 1: estimate propensity score

	Influenza vaccine	No influenza vaccine
Age <65 years	360	40
Age ≥ 65 years	50	50

Step 1: Estimate the propensity score — the probability of being treated or exposed conditional on a set of characteristics

Propensity score if age < 65 years: 360 / (360 + 40) = 0.90
Propensity score if age ≥ 65 years: 50 / (50 + 50) = 0.50

Notice that the distribution of age is different within the exposed and unexposed groups.

IPTW Step 2: Define weights

	Influenza vaccine	No influenza vaccine
Age <65 years	360	40
Age ≥ 65 years	50	50

Step 2: Define weights based on the inverse propensity score (i.e. the inverse probability of treatment)

Propensity score if age < 65 years: 360 / (360 + 40) = 0.90
Propensity score if age ≥ 65 years: 50 / (50 + 50) = 0.50

- **Weight for age <65 years**
  - Exposed = 1 / 0.90 = 1.11
  - Unexposed = 1 / (1-0.90) = 10
- **Weight for age ≥ 65 years**
  - Exposed = 1 / 0.50 = 2
  - Unexposed = 1 / (1-0.50) = 2

Notice that the distribution of age is different within the exposed and unexposed groups.

IPTW Step 3: Apply weights to the data

	Influenza vaccine	No influenza vaccine
Age <65 years	360 * **1.11 = 400**	40 * **10 = 400**
Age ≥ 65 years	50 * **2 = 100**	50 * **2 = 100**

Step 3: Apply weights to the data, creating a "pseudo-population" to approximate two counterfactual populations

Propensity score if age < 65 years: 360 / (360 + 40) = 0.90
Propensity score if age ≥ 65 years: 50 / (50 + 50) = 0.50

- **Weight for age <65 years**
  - Exposed = 1 / 0.90 = 1.11
  - Unexposed = 1 / (1-0.90) = 10
- **Weight for age ≥ 65 years**
  - Exposed = 1 / 0.50 = 2
  - Unexposed = 1 / (1-0.50) = 2

After weighting, the distribution of age is the same within the exposed and unexposed groups.

Weighting creates a pseudo population in which exposure is independent of confounders— this mimics the exposure and confounder distribution we would see in a trial.

IPTW Step 3: Apply weights to the data

Age <65 years (Weighted)		
	Doctor visit	**No doctor visit**
Influenza vaccine	108 * **1.11 = 120**	252 * **1.11 = 280**
No influenza vaccine	24 * **10 = 240**	16 * **10 = 160**

Age ≥ 65 years (Weighted)		
	Doctor visit	**No doctor visit**
Influenza vaccine	20 * **2 = 40**	30 * **2 = 60**
No influenza vaccine	40 * **2 = 80**	10 * **2 = 20**

Weight for age <65 years

- Exposed = 1 / 0.90 = 1.11
- Unexposed = 1 / (1-0.90) = 10

Weight for age ≥ 65 years

- Exposed = 1 / 0.50 = 2
- Unexposed = 1 / (1-0.50) = 2

IPTW Step 4: Calculate the measure of association

Pooled, weighted data		
	Doctor visit	**No doctor visit**
Influenza vaccine	**160**	**340**
No influenza vaccine	**320**	**180**

Recall:

- Crude RR = 0.44
- RR if age < 65 = 0.5
- RR if age ≥ 65 = 0.5

Step 4: Calculate the mean of the outcome times the weight

IPTW RR = R(e) / R(u) = 0.32 / 0.64 = (160 / (160 + 340)) / (320 / (320 + 180)) = 0.5

The weighting in IPTW has achieved the same RR that we would obtain by calculating a Mantel-Haenszel RR controlling for age.

IPTW Step 4: Calculate the measure of association

Age <65 years (Weighted)		
	Doctor visit	No doctor visit
Influenza vaccine	120	280
No influenza vaccine	240	160

RR = 0.5

Age ≥ 65 years (Weighted)		
	Doctor visit	No doctor visit
Influenza vaccine	40	60
No influenza vaccine	80	20

RR = 0.5

Creating a pseudo-population

- Weighting creates a pseudo-population.
  - If the weight for subject *i* = 4, then there are 4 copies of subject *i* in the pseudo-population.
- After weighting, the number of participants is balanced between exposed and unexposed within confounder strata.
  - This mimics what we would achieve in a randomized trial: covariate balance between the treatment and control group.

Interpreting IPTW estimates

- Ratio or difference in the average outcome **if everyone had treatment *X*=*x***, for any *x*, compared to if **no one had the treatment *X*=*x***.
  - You can choose whether to estimate a relative or additive scale measure of association.

- For our example: the average risk of a medical visit for influenza-like illness if everyone was vaccinated for influenza compared to if no one was vaccinated for influenza

  - **RR = 0.32 / 0.64 = 0.50**
  - **RD = 0.32 - 0.64 = -0.32**

- Several assumptions are required to make a causal interpretation of measures of association estimated using IPTW.

G-computation

G-computation: Summary of steps

1. Estimate the association between the exposure and outcome adjusting for confounders
    a. Could use linear, log-linear, or logistic regression
2. Use the coefficients from the model to obtain a counterfactual value for each individual using their particular values of each confounder in two scenarios:
    a. Setting exposure variable to “exposed”
    b. Setting exposure variable to “unexposed”
3. Estimate the measure of disease in the population set to “exposed” and in the population set to “unexposed”
4. Estimate the measure of association in this “counterfactual population”

Example of G-computation

- **Example**: an observational study whether influenza vaccination reduces the risk of visiting a doctor's office for influenza-like symptoms
- **Step 1: Estimate the association between the exposure and outcome adjusting for confounders**
  - Y = doctor's office visit for influenza-like symptoms (binary: yes/no)
  - X = influenza vaccination status (binary: yes/no)
  - W = age
  - Could use log-linear regression in R:

    glm.fit = glm(Y~X+W, data = d, family=poisson(link="log"))

$$\ln(E(Y|X,W) = \beta_0 + \beta_1 X + \beta_2 W$$
$$\ln(E(Y|X,W) = -0.223 - 0.693X - 0.288W$$

Example of G-computation

- **Step 2: Use the coefficients from the model to obtain a counterfactual value for each individual using their particular values of each confounder in two scenarios:**

$$\ln(E(Y|X,W) = -0.223 - 0.693X - 0.288W$$

Subset of 4 rows:

id	Y	X	W	Counterfactual risk if X=1	Counterfactual risk if X=0
1	1	1	1	exp(-0.223 - 0.693(1) - 0.288(1)) = exp(-1.204) = 0.3	exp(-0.223 - 0.693(0) - 0.288(1)) = exp(-0.511) = 0.6
2	0	1	1	exp(-1.204) = 0.3	exp(-0.511) = 0.6
3	1	0	0	exp(-0.916) = 0.4	exp(-0.223) = 0.8
4	0	0	0	exp(-0.916) = 0.4	exp(-0.223) = 0.8

Example of G-computation

- **Step 3: Estimate the measure of disease in the population set to "exposed" and in the population set to "unexposed"**

Subset of 4 rows:

id	Y	X	W	Counterfactual outcome if X=1	Counterfactual outcome if X=0
1	1	1	1	0.3	0.6
2	0	1	1	0.3	0.6
3	1	0	0	0.4	0.8
4	0	0	0	0.4	0.8
				...	**...**
Mean across all rows:				**0.32**	**0.64**

Counterfactual risk setting X=1 for everyone in the population = 0.32

Counterfactual risk setting X=0 for everyone in the population = 0.64

Example of G-computation

- **Step 4: Estimate the measure of association in this "counterfactual population"**

Subset of 4 rows:

id	Y	X	W	Counterfactual outcome if X=1	Counterfactual outcome if X=0
1	1	1	1	0.3	0.6
2	0	1	1	0.3	0.6
3	1	0	0	0.4	0.8
4	0	0	0	0.4	0.8
				...	...
Mean across all rows:				**0.32**	**0.64**

Counterfactual risk setting X=1 for everyone in the population = 0.32

Counterfactual risk setting X=0 for everyone in the population = 0.64

RR = 0.32 / 0.64 = 0.50
RD = 0.32 - 0.64 = -0.32

We obtained the same RR using IPTW and G-computation as we would have obtained by calculating a Mantel-Haenszel RR controlling for age.

Interpreting G-computation estimates

- Ratio or difference in the average outcome **if everyone had treatment *X*=*x***, for any *x*, compared to if **no one had the treatment *X*=*x***.

- For our example: the average risk of a medical visit for influenza-like illness if everyone was vaccinated for influenza compared to if no one was vaccinated for influenza

  - **RR = 0.32 / 0.64 = 0.50**
  - **RD = 0.32 - 0.64 = -0.32**

- Several assumptions are required to make a causal interpretation of measures of association estimated using G-computation.

Flexible parameter definition under IPTW and G-computation

- Typically IPTW and G-computation are used to compare the risk when everyone is exposed compared to when no one is exposed.
- However, sometimes this is not a realistic contrast.
  - E.g., it would be highly unlikely for no one to receive the influenza vaccine in most populations in the U.S.
- IPTW and G-computation can be used to estimate other counterfactual contrasts
- Alternative parameter:
  - What is the difference or ratio of risk comparing:
    - Scenario with current level of vaccination coverage (50%)
    - Scenario with 80% vaccination coverage
- More on this in the unit on population intervention models

Flexible parameter definition under IPTW and G-computation

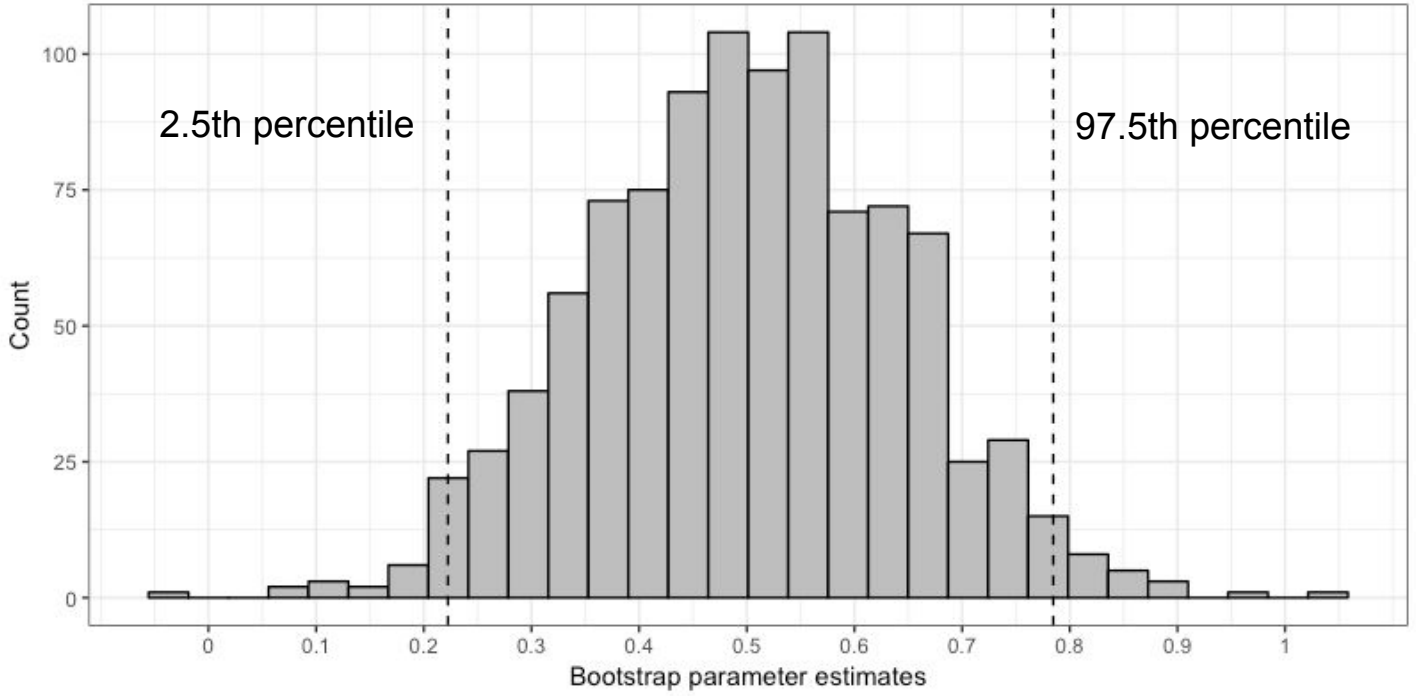
- Flexible parameter definition is possible because both methods separate the step that adjusts for confounding from the step that estimates the parameter.
  - IPTW adjusts for confounding when estimating the propensity score.
  - G-computation adjusts for confounding when estimating the outcome conditional on exposure and confounders.
- This separation allows for estimation of novel parameters.
- It also allows for estimation methods in the step that adjusts for confounding other than traditional regression models (e.g., machine learning based estimation).

How do we obtain confidence intervals for IPTW and G-computation?

- There is no formula for calculating standard errors of parameter estimates from IPTW or G-computation.
- We must use bootstrapping to obtain standard errors, confidence intervals, and p-values.
- What is bootstrapping?

    1. Take a random sample with replacement from your dataset.
    2. Estimate the parameter in the random sample.
    3. Repeat steps 1-2 many times (1000 - 10,000 times) to obtain a distribution of bootstrapped parameter estimates.
    4. Obtain the standard error by applying the normal distribution to the parameter estimates or using quantiles of the distribution of the parameter estimates.

Example of bootstrap distribution



Lower bound = 0.22

Upper bound = 0.78

Limitations of IPTW and G-computation

- "Garbage in, garbage out"
  - Sophisticated methods can't always make up for bias in the study design or due to unmeasured confounding.
- The model used to adjust for confounding must meet the backdoor criterion in order to make causal inferences. (in addition to other assumptions)
  - In IPTW, the model used to estimate the propensity score must adjust for confounders that meet the backdoor criterion.
  - In G-computation, if the model used to estimate the outcome conditional on the exposure and covariates must adjust for confounders that meet the backdoor criterion.
  - These slides show an example with a single confounder — in practice we almost always must adjust for multiple confounders.

Comparison of IPTW to standardization

IPTW Steps

1. Estimate the propensity score
2. Apply **weights** to the data
3. Calculate the mean of the outcome times the **weights**

Standardization Steps

1. Obtain population counts stratified by a confounder in the reference population
2. Apply **population counts** to the data
3. Calculate the mean of the outcome times the **population counts**

Double robust estimation

$$P_O(Y, X, W) = \underbrace{P(Y|X, W)P(W)}_{\text{Targeted by G-computation}}\underbrace{P(X|W)}_{\text{Targeted by IPTW}}$$

- To obtain an unbiased measure of association, we must use the correct model for the outcome under G-computation and the correct model for the treatment under IPTW.
- What if we get one of them wrong?
- Double robust estimation provides an unbiased estimate if either model is correct.
  - Example: Targeted maximum likelihood estimation

Summary of key points

- New methods from the causal inference literature: propensity score matching, IPTW, G-computation
- Using these methods alone isn't enough to make causal inferences — other assumptions must be met.
- Don't necessarily have to be used for causal inference.
  - IPTW and G-computation can be used to estimate novel alternative parameters
- These methods are superior to regression based methods for controlling for time-dependent confounding.
- Garbage in, garbage out
  - We cannot make up for data that is biased due to flawed study design or data collection by simply using these methods.
  - If we use incorrect models to estimate treatment in IPTW or the outcome in G-computation, our results may be biased.

Sample size & power

This lecture was created by Dr. Ben Arnold @ UC Berkeley

Current use: PH250G

Goals for this session

1. Familiarize you with the basic issues that underlie sample size calculations

2. Walk through the steps of real calculations, including one with clustering

3. Point out some additional wrinkles that you may

Slide created by Ben Arnold

Presentation overview

- **Big-picture concepts & motivation**
- Example 1: Understanding power & marine water exposure / gastrointestinal illness case study
- Example 2: Calculate sample size for an individually randomized trial
- Additional issues and recap

Motivation: need a sufficient sample size

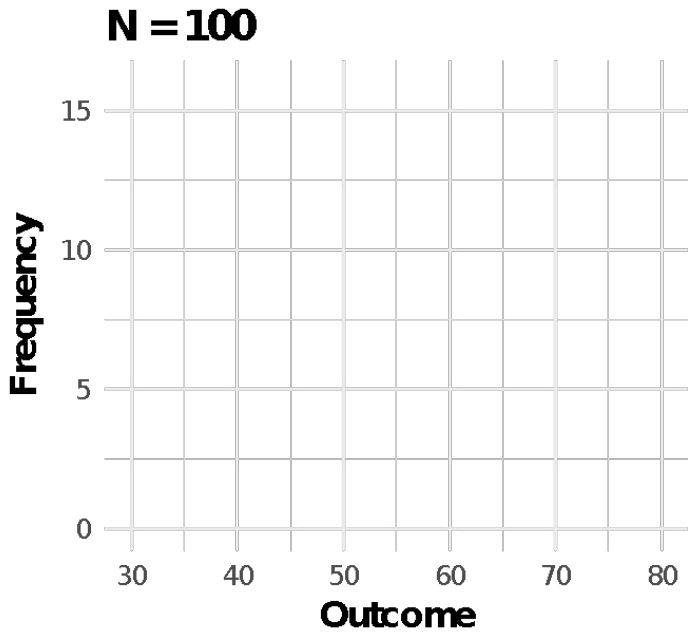
- In the planning stage of a study, a key activity for epidemiologists and biostatisticians is to determine how many participants to enroll.

- A study needs to be large enough to distinguish a true difference from random variation

Slide created by Ben Arnold

Heuristic, hypothetical example: Do the populations differ?

With a smaller sample size, the difference between populations is less clear



Population

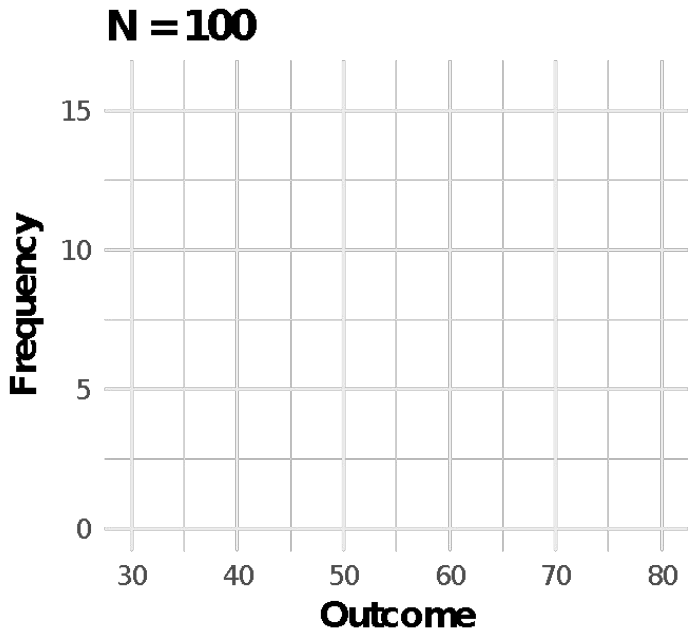
Mean = 50

Mean = 60

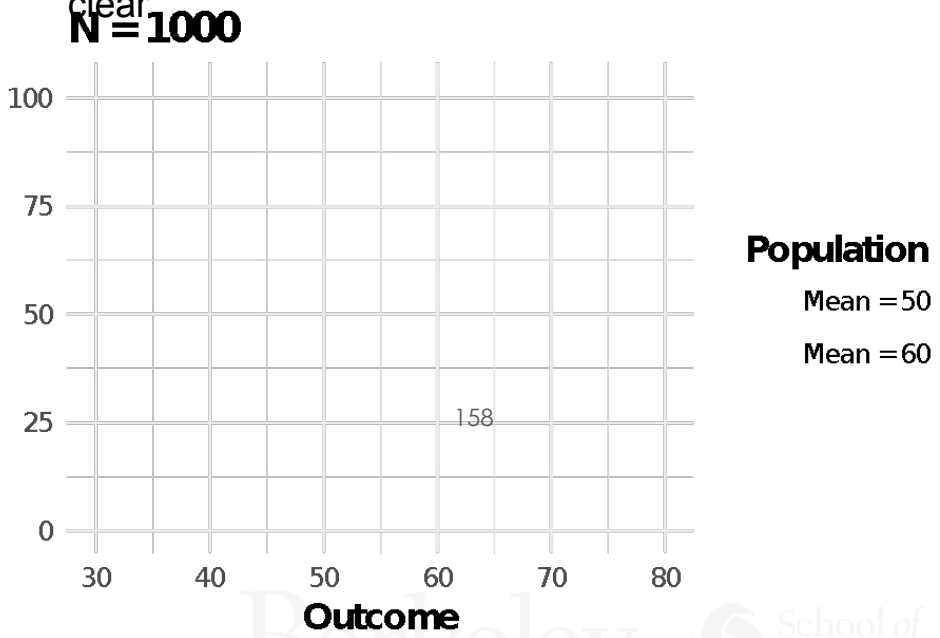
Slide created by Ben Arnold

Heuristic, hypothetical example: Do the populations differ?

With a smaller sample size, the difference between populations is less clear



With a larger sample size, the difference between populations is much more clear



Slide created by Ben Arnold

Motivation: …but samples should not be too big

- Samples that are too large are a problem:
  - Waste time and resources
  - Impose undue burden on the study population
  - At some point study logistics begin to threaten internal validity

Slide created by Ben Arnold

Sample Size and Minimum Detectable Effects

- When designing a study, we're typically interested in one of these two related questions:

1. Given a desired minimum detectable effect (MDE), how many participants does the study need to enroll? (Example 1)

2. Given a fixed sample size and design, what is the MDE that the study can detect? (Example 2)

- Tied up in both of these questions is a decision of how much power we want the design to have

Slide created by Ben Arnold

Presentation overview

- Big-picture concepts & motivation
- **Example 1: Understanding power & marine water exposure / gastrointestinal illness case study**
- Example 2: Calculate sample size for an individually randomized trial
- Additional issues and recap

Example 1:
Does swimming in marine water increase the risk of gastrointestinal illness?

WATER RESEARCH 46 (2012) 2176–2186

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/watres

Using rapid indicators for *Enterococcus* to assess the risk of illness after exposure to urban runoff contaminated marine water

John M. Colford Jr.[a,*], *Kenneth C. Schiff*[b], *John F. Griffith*[b], *Vince Yau*[a,1], *Benjamin F. Arnold*[a], *Catherine C. Wright*[a], *Joshua S. Gruber*[a], *Timothy J. Wade*[c], *Susan Burns*[d], *Jacqueline Hayes*[d], *Charles McGee*[e], *Mark Gold*[f], *Yiping Cao*[b], *Rachel T. Noble*[g], *Richard Haugland*[h], *Stephen B. Weisberg*[b]

Slide created by Ben Arnold

General setup using the recreational water example

- Scientific Question: Does swimming in marine water increase the risk of gastrointestinal illness?

- Swimmers ($A = 1$) and non-swimmers ($A = 0$) are enrolled at the beach, and followed for 2 weeks for incident GI illness (Y).

- One comparison of interest, θ, is the risk difference between groups:

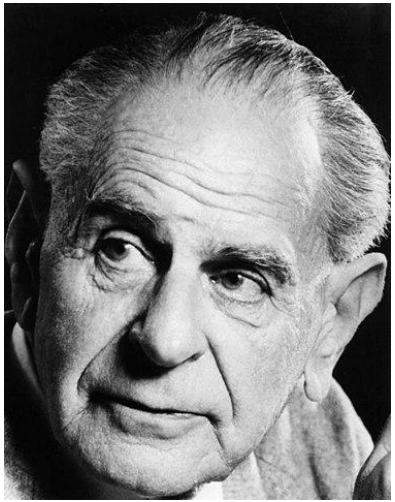
$$\theta = \text{risk among swimmers} - \text{risk among non-swimmers}$$
$$= p_2 - p_1 = Pr(Y=1 \mid A=1) - Pr(Y=1 \mid A=0)$$

163

- Sample size question: How large does the study need to be to credibly detect an increase of 1 percentage point in the incidence proportion from a base of 3.4% to 4.4% ?

Slide created by Ben Arnold

Statistical Power



Sir Karl Popper (1902-1994)

https://en.wikipedia.org/wiki/Karl_Popper

- It is common to use statistical power to help determine the size of a study.
- Statistical power is defined as the probability of rejecting the null hypothesis when it is false.
- Implicit in its definition is that an investigator must specify a null and alternative hypothesis.
  - i.e., power is defined in terms of a specific hypothesis test
- Grounded in Popper's epistemology of falsification: scientific knowledge progresses through testing falsifiable statements.

1 : specify the null and alternative hypotheses

From the recreational water example:

- The null hypothesis, H_0, is that $\theta_0 = p_2 - p_1 = 0$
  (there is no effect)

- An alternative hypothesis, H_a, is that $\theta_a = p_2 - p_1 = 0.044 - 0.034 = 0.01$
  (water exposure increases risk by 1 percentage point)

Slide created by Ben Arnold

2: pick a test statistic

- Given a hypothesis, the next step is to choose a test statistic

- One example of a (commonly used) test statistic is a Wald statistic, which is the difference between groups divided by its standard

$$X = \frac{\hat{\theta}_0}{SE(\hat{\theta}_0)} \sim N(0,1)$$

- A Wald statistic, X, will have a standard

Slide created by Ben Arnold

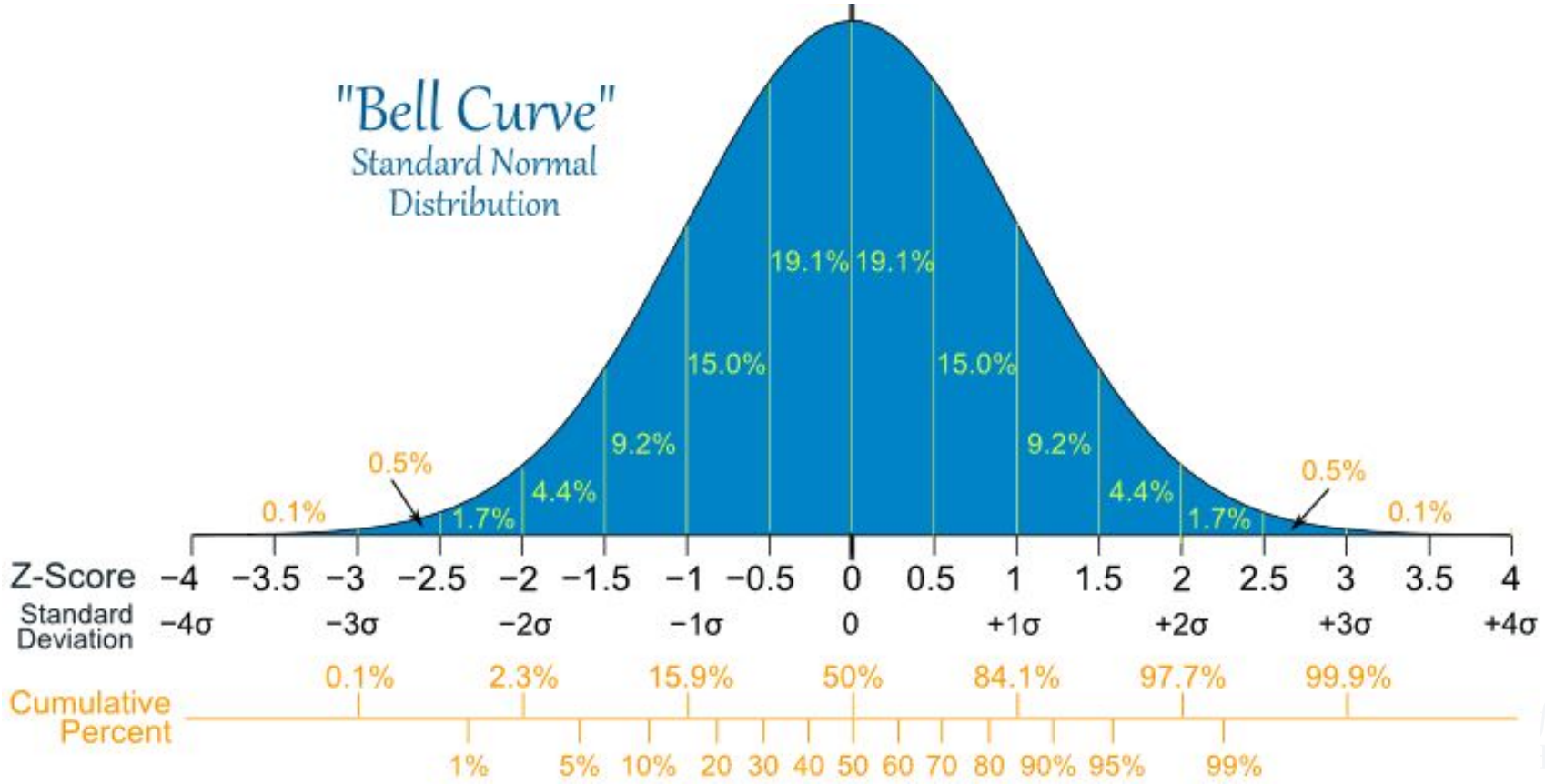
3 : specify critical values based on Type I & Type II errors

Decision made (based on test statistic)	H_0 is true	H_a is true
Accept H_0	correct decision	Type II error (β)
Reject H_0	Type I error (α)	correct decision

- Type I error: incorrectly reject the null when there is truly no difference (false positive)
- Type II error: incorrectly accept the null when there is, in fact, a difference (false

Slide created by Ben Arnold

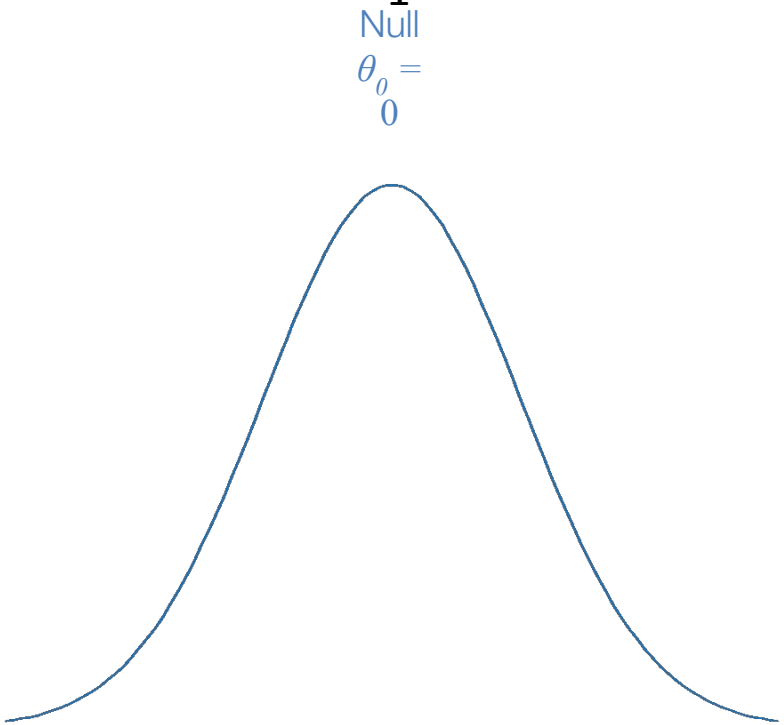
Reminder – Standard normal distribution



What is power?

Start with the null distribution of X (2 slides ago)

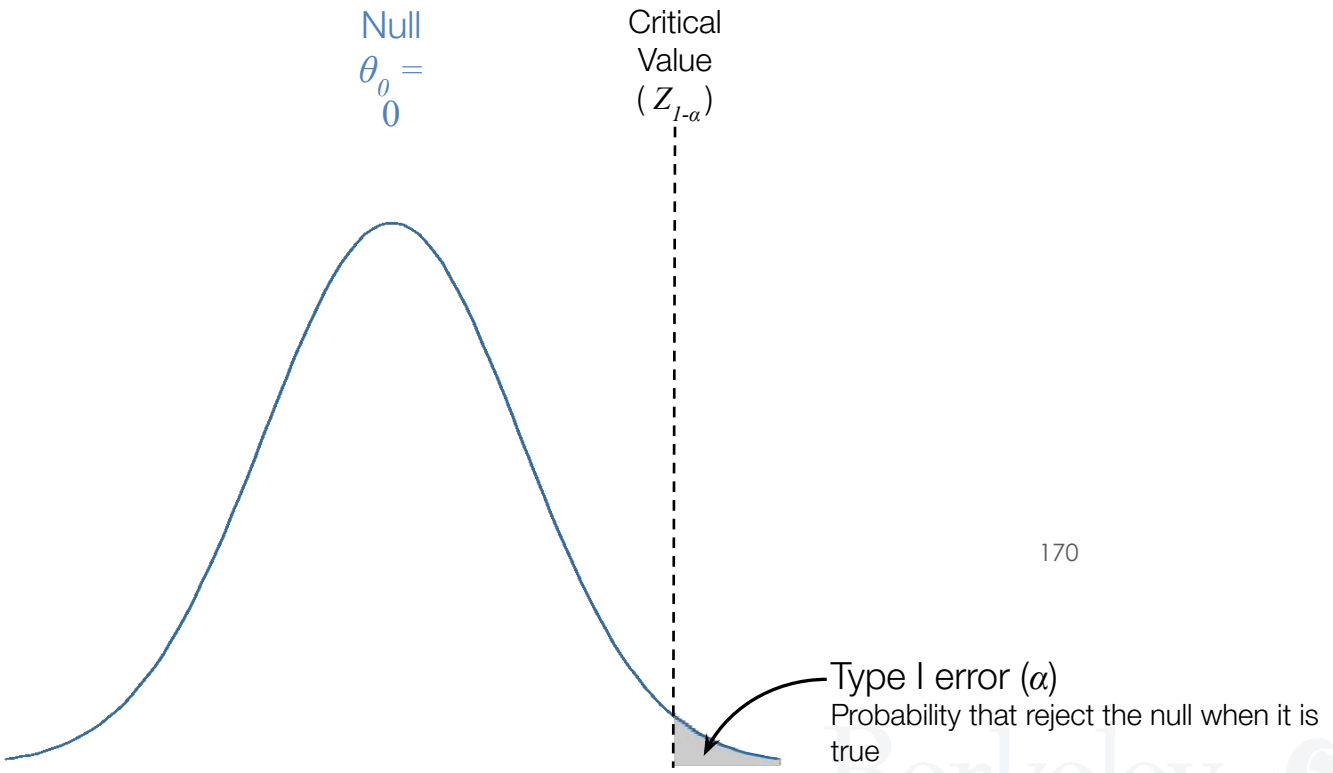
It is normally distributed with mean 0 and standard deviation 1



Slide created by Ben Arnold

Type I error (α) probability is the area under the null distribution of X to the right of the **critical value**

If α = 0.05, then $Z_{1-\alpha}$ is the 95th percentile of the standard normal distribution.



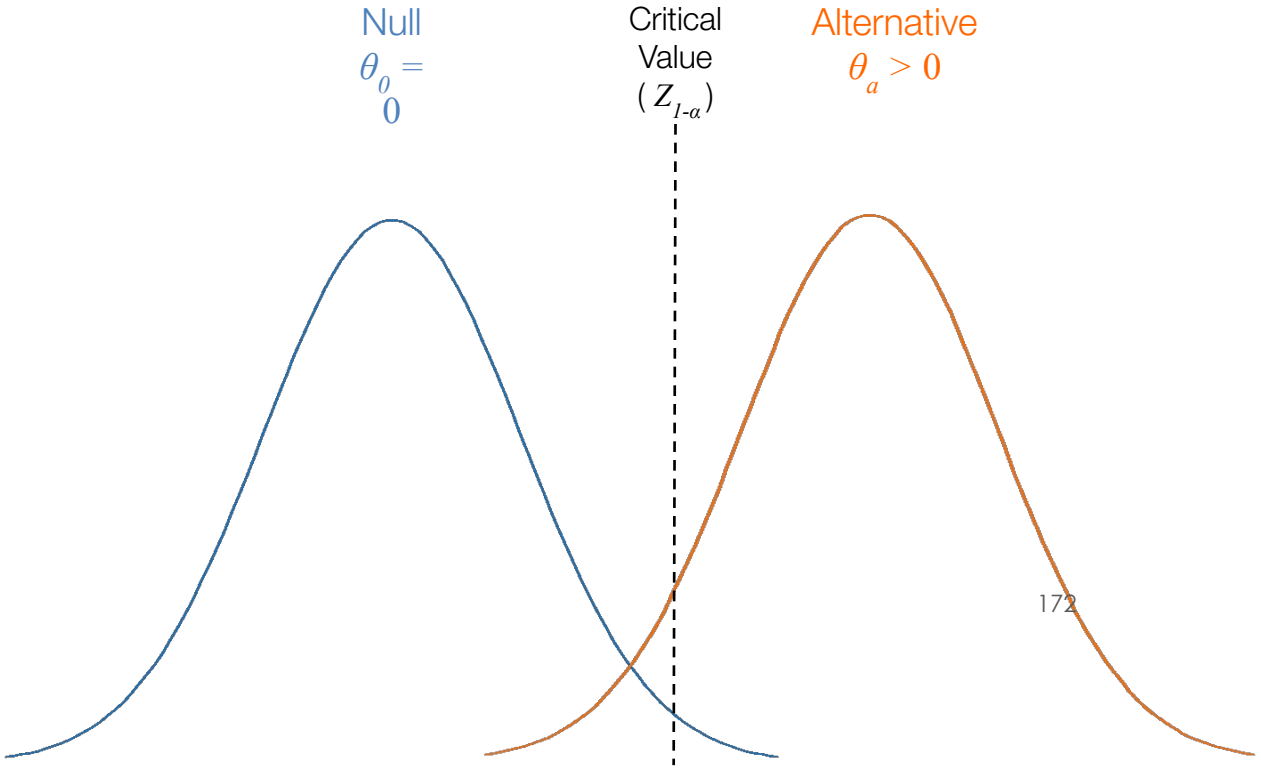
Slide created by Ben Arnold

QUESTIONS

- Can we just set alpha = zero to make sure our result/finding is not due to random error?
  - No – Some random error (i.e, some sampling error) is unavoidable

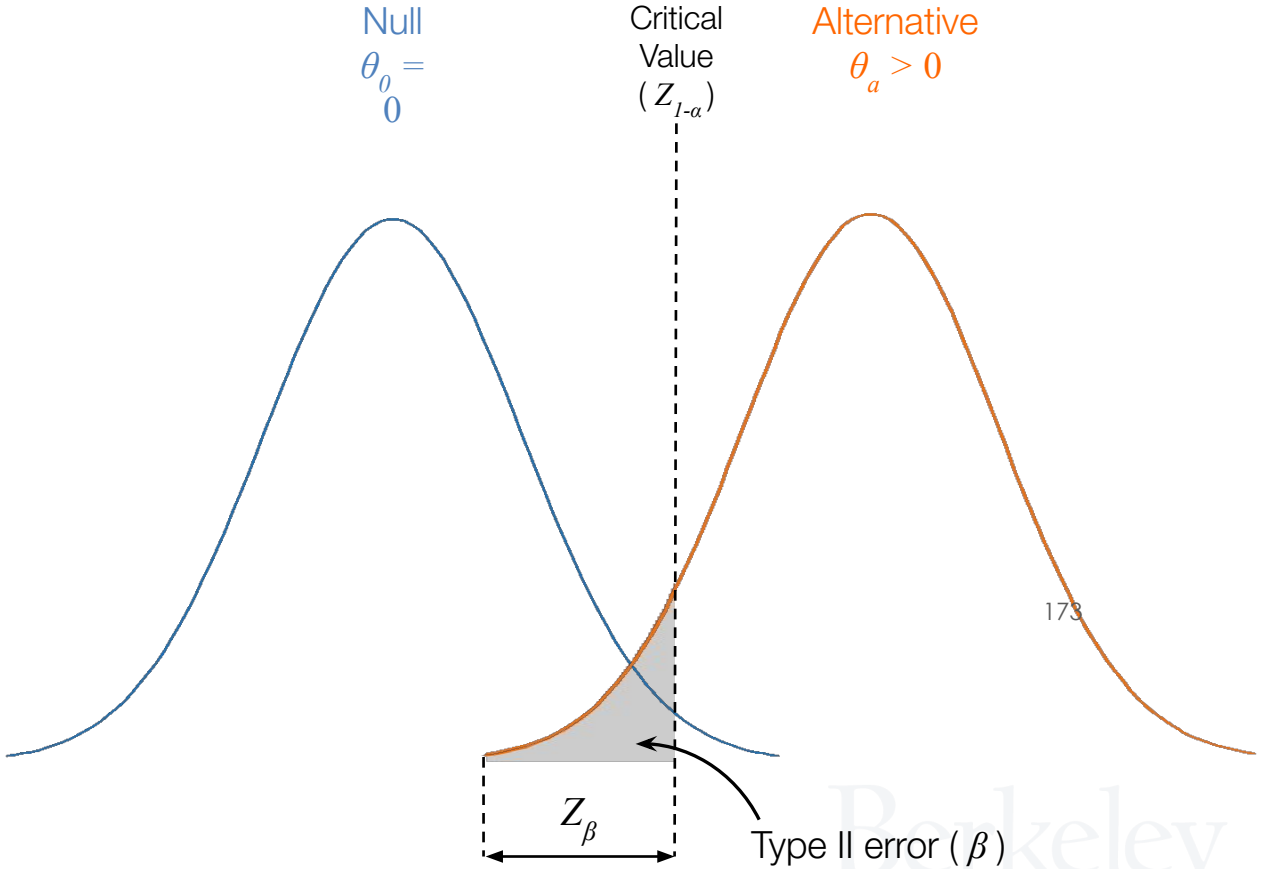
- What if we increase our sample size until it includes all sampling units?
  - This is a census, not a sample
    - In which case don't need to estimate parameters we

The null and alternative distributions for X



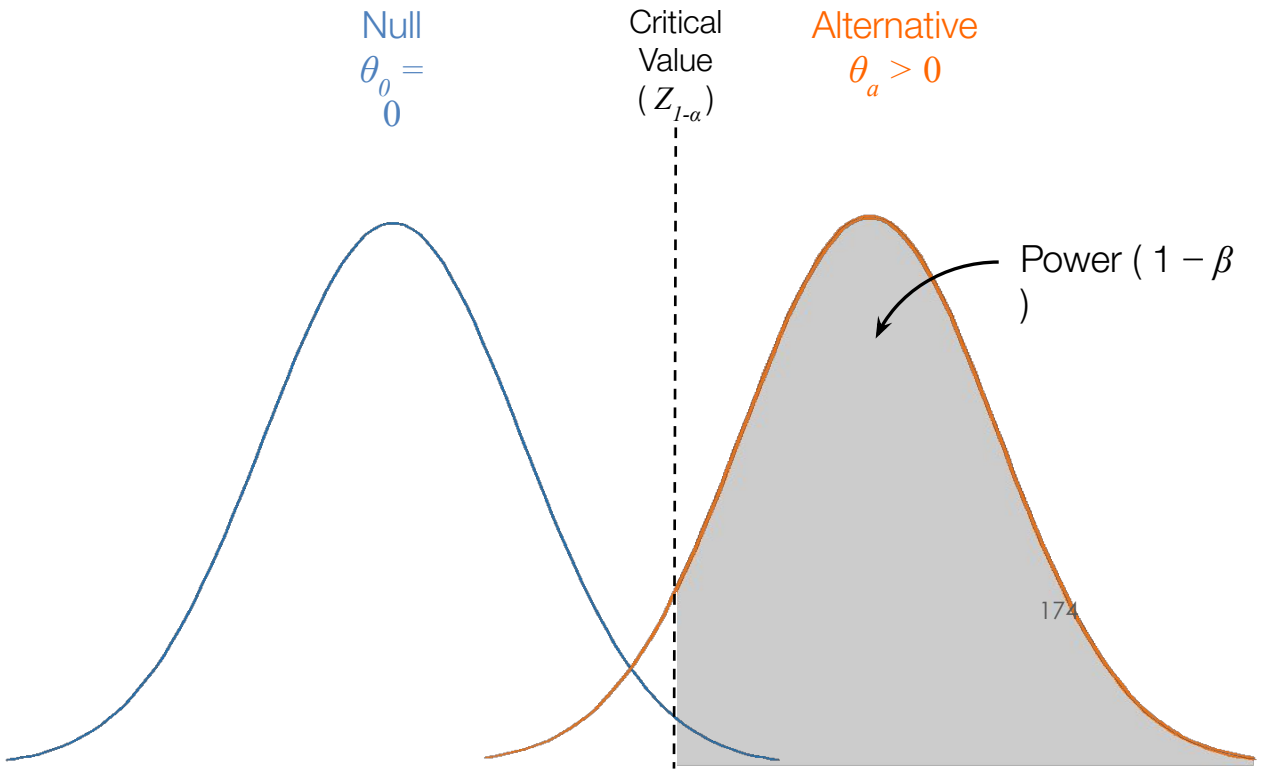
Slide created by Ben Arnold

Type II error probability is the area under the **alternative distribution** to the left of the critical value



Slide created by Ben Arnold

Power ($1-\beta$) is the area under the alternative distribution to the right of the critical value.



Slide created by Ben Arnold

How to increase power?

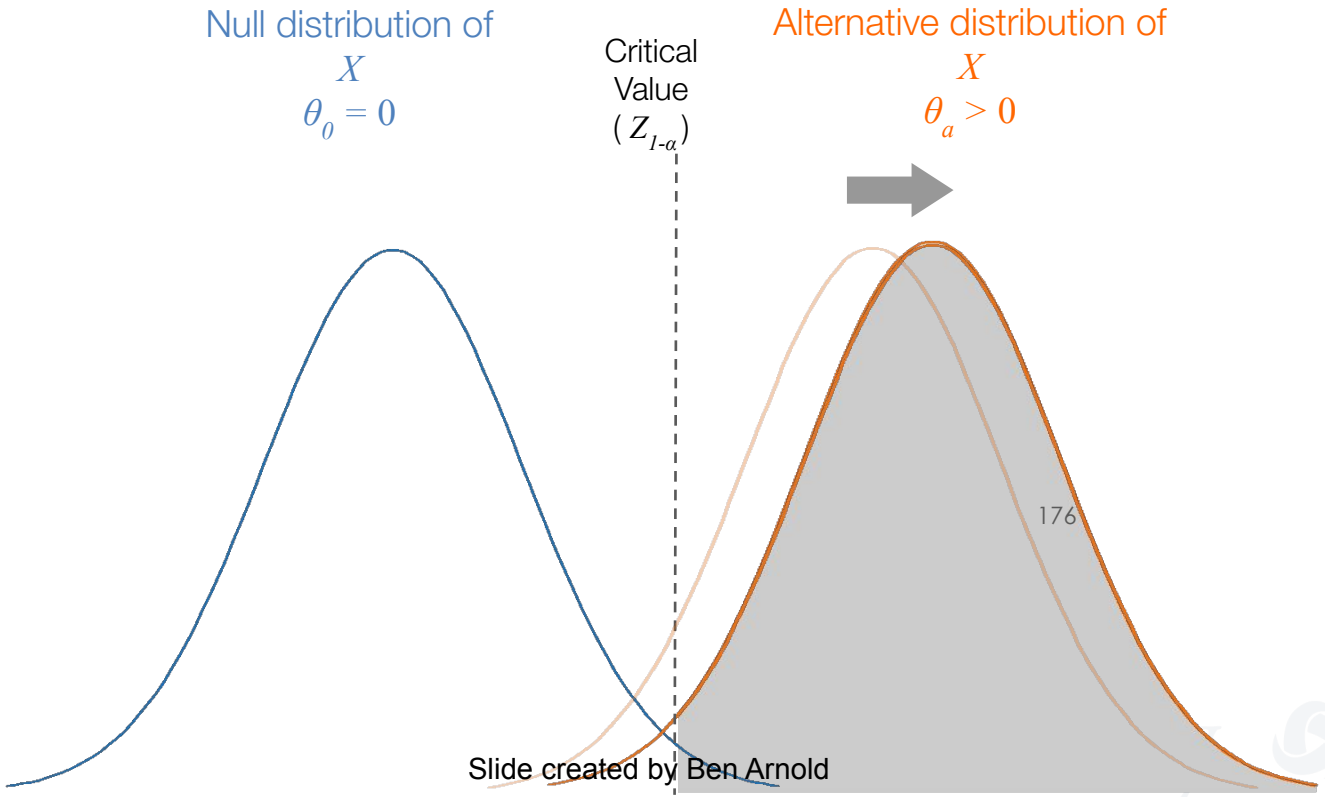
- To increase power, need to shift the alternative distribution further away from the null distribution.

$$X = \frac{\hat{\theta}_a}{SE(\hat{\theta}_a)}$$

- How to increase this test statistic?

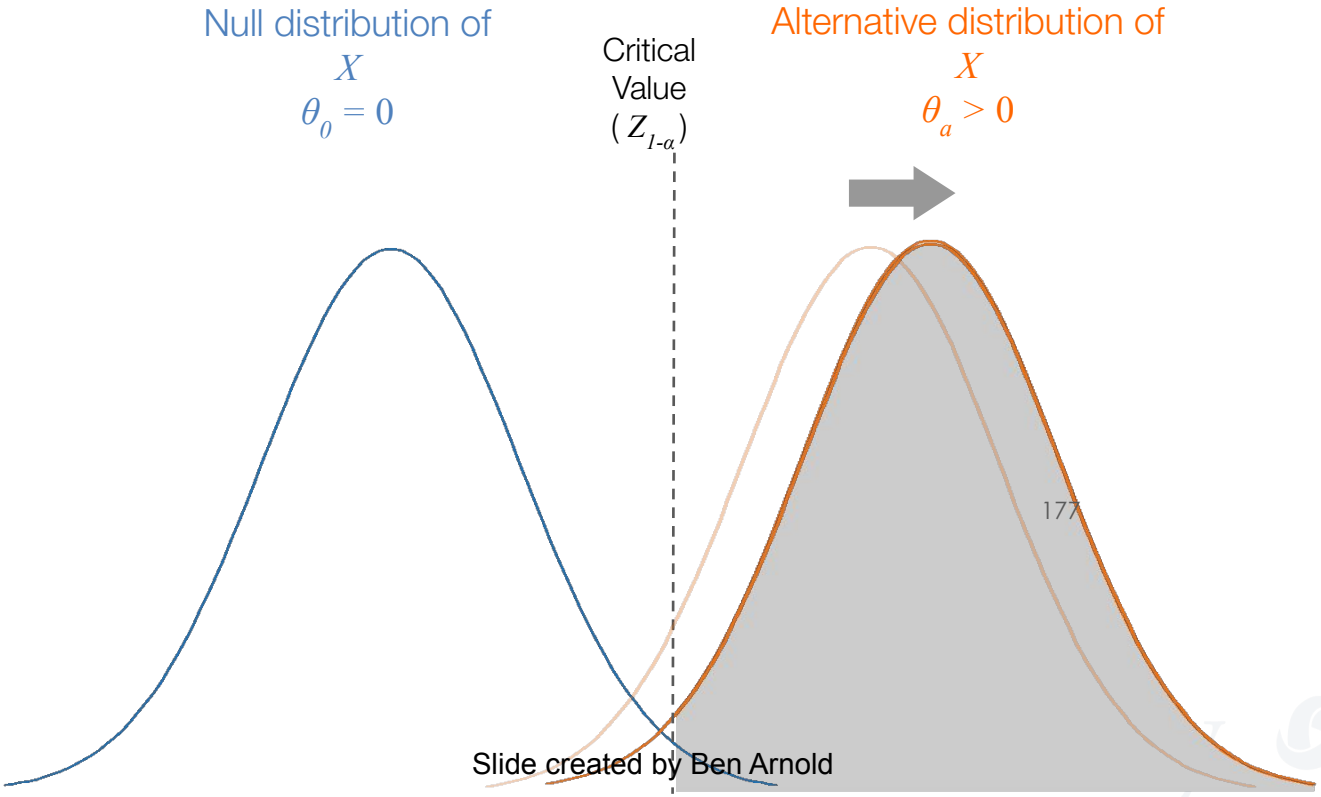
Slide created by Ben Arnold

Increase X and the area beyond the critical value (power) by increasing the effect size, θ_a



Slide created by Ben Arnold

Increase X and the area beyond the critical value (power) by increasing the sample size. This shrinks $SE(\hat{\theta}_a)$ because: $SE(\hat{\theta}_a) = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$



Slide created by Ben Arnold

Example calculation from the recreational water study

1. Specify hypotheses
   - $H_0 : \theta_0 = 0$; $H_a : \theta_a > 0$

2. Pick critical values
   - Power (aka, Type 2 error rate): 80% = $Z_{0.80} = 0.84$
   - Precision (aka, Type 1 error rate): 5% for a two-sided test = $Z_{1-\alpha/2} = Z_{0.975} = 1.96$

3. Specify parameters required in the calculation
   - $p_1 = 0.034$ cumulative incidence among non-swimmers from a nearby study (Colford 2007)
   - $p_2 = 0.044$ cumulative incidence among swimmers (for a 1% increase)

4. Calculate sample size (by hand or with software – next slides)

Slide created by Ben Arnold

Sample size equation for a risk ratio of proportions

(algebra challenge: derive the equation in panel 2 of Schulz and Grimes 2005)

$$m = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 \left[p_1(1-p_1) + p_2(1-p_2)\right]}{(p_1 - p_2)^2}$$

- m = sample size in each group (swimmers, non-swimmers; assumed equal size)
- p_1 = cumulative incidence among swimmers
- p_2 = cumulative incidence among non-swimmers
- Z_x = the x'th percentage point of the standard normal distribution

Slide created by Ben Arnold

Recreational water example, sample size arithmetic

$$
\begin{aligned}
m &= \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 \left[p_1(1-p_1) + p_2(1-p_2)\right]}{(p_1 - p_2)^2} \\
&= \frac{(0.84 + 1.96)^2 \left[0.044(1-0.044) + 0.034(1-0.034)\right]}{(0.044 - 0.034)^2} \\
&= 5,873
\end{aligned}
$$

180

(note: $m = 5,879$ using the equation, but without rounding on the critical values for Z)

"Most hand [sample size] calculations diabolically strain human limits, even for the easiest formula"

(Schulz & Grimes 2005)

Recommend that you practice writing functions in R that use these formulas.
See example in the video on sample size for cluster randomized trials.

Presentation overview

- Big-picture concepts & motivation
- Example 1: Understanding power & marine water exposure / gastrointestinal illness case study
- **Example 2: Calculate sample size for an individually randomized trial**
- Additional issues and recap

Example #2: 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

Slide created by Yoshika Crider

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)	

Slide created by Yoshika Crider

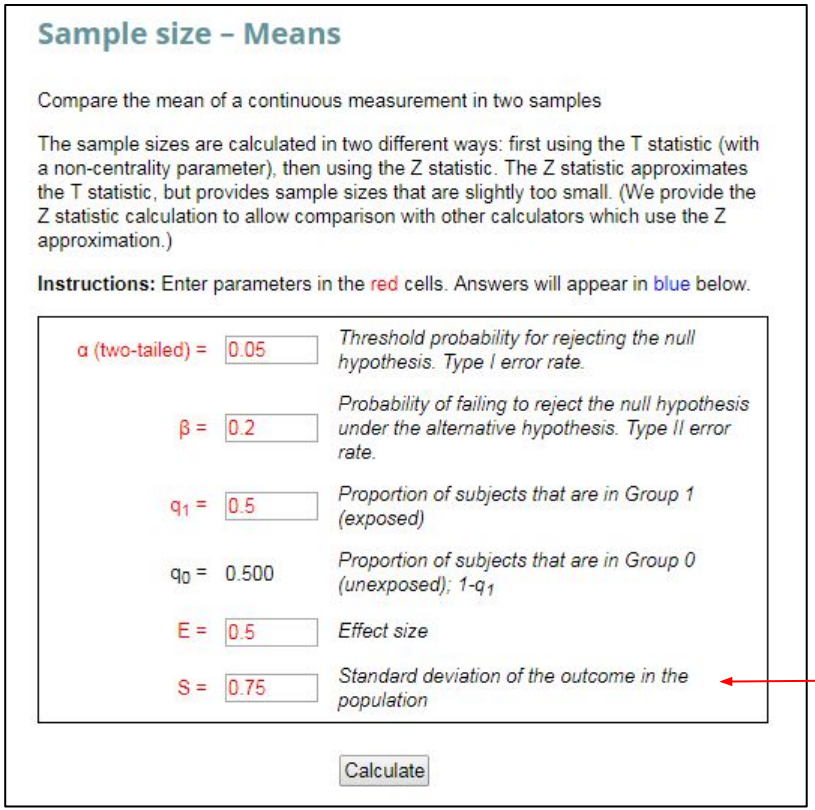
Given parameters

"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

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

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)

Additional issues you may need to consider

- Loss to follow-up
  - Increase your sample size by your anticipated attrition rate (e.g., 10%)
- Study design for multiple primary outcomes
  - Repeat the exercise and find the limiting outcome
- Study design for planned subgroup analyses
  - Size the study around subgroups (may require sampling based on subgroup)

- Binary outcomes with repeated measures
  - Similar, but slightly modified equations (Leon 2004)
- Sample size / power calculations for complicated designs
  - Consider the use of Monte Carlo simulation (Feiveson 2002, Arnold 2011)

Slide created by Ben Arnold

Summary of Key Points

- Sample size calculations require that you specify:
  1. basic design (fraction treated/exposed), parameter of interest, and hypothesis
  2. outcome variability (get this automatically for binomial outcomes)
  3. minimum detectable effect
  4. desired level of power and Type I error (alpha)

- Sample size, MDE, and power equations are all just re-arranged versions of the same

Slide created by Ben Arnold

Additional comments

- These calculations are important, but ultimately they rely on a lot of guesswork. In practice, there is a lot of interplay between the science, the budget, and the sample size calculations. This is the "sample size samba" (Schulz & Grimes 2005)

- These skills are extremely useful to have in your study design tool box

- If you are interested in the derivation that underlies the basic sample size and power equations, Steve Selvin's book *Statistical Analysis of Epidemiologic Data* has a nice introduction (Ch 3)

Slide created by Ben Arnold

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

Slide created by Yoshika Crider

References

- Arnold, B.; Hogan, D.; Colford, J. & Hubbard, A. Simulation methods to estimate design power: an overview for applied research. BMC Med Res Methodol, 2011, 11, 94
- Arnold, B. F.; Null, C.; Luby, S. P.; et al. Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits Study design and rationale. BMJ Open, 2013, 3, e003476
- Colford JM, Schiff KC, Griffith JF, Yau V, Arnold BF, Wright CC, et al. Using rapid indicators for Enterococcus to assess the risk of illness after exposure to urban runoff contaminated marine water. Water Res. 2012;46: 2176–2186.
- Colford, J. M.; Wade, T. J.; Schiff, K. C.; et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology, 2007, 18, 27-35
- Campbell, M. J.; Donner, A. & Klar, N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med, 2007, 26, 2-19
- Duflo, E.; Glennerster, R. & Kremer, M. Using Randomization in Development Economics Research: A Toolkit. 61 Handbook of Development Economics, 2007, Volume 4, 3895-3962
- Feiveson, A. H. Power by simulation. Stata Journal, 2002, 2, 107-124
- Leon, A. C. Sample-size requirements for comparisons of two groups on repeated observations of a binary outcome Eval Health Prof, 2004, 27, 34-44
- Murray, D. M.; Varnell, S. P. & Blitstein, J. L. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health, 2004, 94, 423-432

- Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44: 1051–1067.
- Schulz, K. F. & Grimes, D. A. Sample size calculations in randomised trials: mandatory and mystical. Lancet, 2005, 365, 1348-1353
- Selvin, S. Statistical Analysis of Epidemiologic Data (2nd ed). Oxford University Press, 1996
- Victora, C. G.; Adair, L.; Fall, C.; Hallal, P. C.; Martorell, R.; Richter, L.; Sachdev, H. S. Maternal and child undernutrition: consequences for adult health and human capital. Lancet, 2008, 371, 340-357
- Victora, C. G.; de Onis, M.; Hallal, P. C.; Blossner, M. & Shrimpton, R. Worldwide timing of growth faltering: revisiting implications for interventions. Pediatrics, 2010, 125, e473-e480

Slide created by Ben Arnold

Sample size and power for cluster randomized trials

Jack Colford

Adapted from slides by Ben Arnold

PH250G

Presentation overview

- **Big-picture concepts related to clustering**
- Example: Clustering-related considerations & nutritional intervention / child growth case study
- Additional issues and recap

Clustered designs : common in epidemiologic studies

- Often it makes the most sense (scientifically and/or logistically) to deliver an intervention or program to a group of individuals (Murray 2004)

- Changes the physical or social environment (handwashing behavior change)
  - Cannot be delivered to individuals (centralized water treatment)
  - Investigators wish to capture group-level dynamics (deworming campaigns)

- Outcomes are often measured at the individual level, where individuals are grouped into clusters.

- Clusters can be defined by space (e.g., village membership) or by any shared characteristic that connects group members within a cluster

Slide created by Ben Arnold

Clustered designs : Independence assumptions

- A defining characteristic of clustered designs is that in the analysis, repeated observations within the cluster are not assumed to be independent. They are assumed to be correlated.

- However, investigators typically design studies so that clusters are independent (i.e., no interference between individuals in different clusters)

- The result of correlated outcomes within each cluster is that each measurement tends to

Clustered designs : Within-cluster correlation

- When observations within a cluster are correlated, a common way to summarize the correlation is with the intraclass correlation coefficient (ICC):

$$ICC = \rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

where τ^2 is the between-cluster variance and σ^2 is the within-cluster variance.

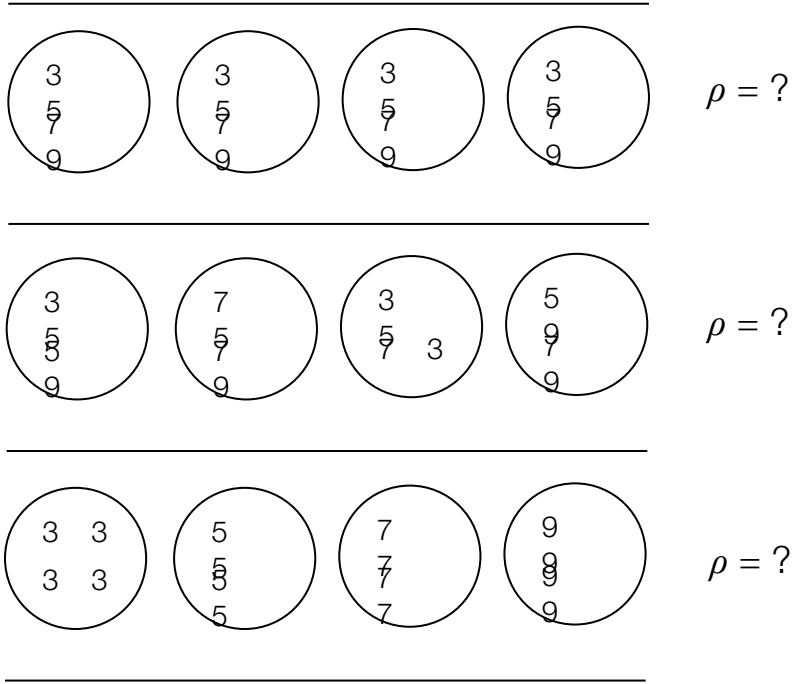
- The ICC is the fraction of the total variance ($\tau^2 + \sigma^2$) that is explained by the between-cluster variance (τ^2).

- If cluster membership explains a lot of the variability in the outcome, then:
  - the ICC will be larger
  - outcomes within each cluster will be more correlated than if cluster membership had no effect on the outcome

Slide created by Ben Arnold

A picture of difference ICCs (ρ)

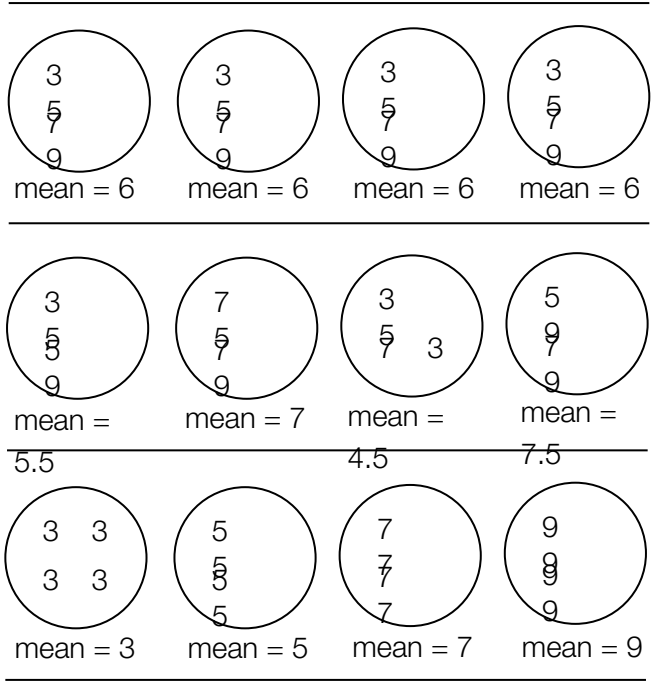
Clusters (circles) with individual level outcomes



Slide created by Ben Arnold

A picture of difference ICCs (ρ)

Clusters (circles) with individual level outcomes



$\rho = 0$

No between cluster variation
No within cluster correlation
Cluster membership is not informative

$\rho = 0.18$

Some between cluster variation
Some within cluster correlation
Cluster membership is a bit informative

$\rho = 1$

High between cluster variation
Perfect within-cluster correlation
Cluster membership tells you everything

Slide created by Ben Arnold

Clustered designs : The Design Effect

- The ICC influences sample size calculations through the design effect (*Deff*):

$$D_{eff} = 1 + (m - 1)\rho$$

where m is the average number of observations per cluster

- The design effect is the ratio between the variances of a design with group-level correlation versus a design where all units are independent.
  - $\rho = 0 \Rightarrow$ have as much power as an individually randomized trial
  - $\rho = 1 \Rightarrow$ effective sample size is the number of clusters

Slide created by Ben Arnold

clustered designs with a continuous outcome (example from Duflo 2007)

$$k = \frac{\sigma^2(Z_{1-\beta}+Z_{1-\alpha})^2}{d^2P(1-P)m} \times [1 + (m-1)\rho]$$

Design effect

$$MDE = d = (Z_{1-\beta} + Z_{1-\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{k}} \times \sqrt{\rho + \frac{1-\rho}{m}}$$

- σ^2 : variability of the outcome, Y (here: assumed equal in both groups)
- P : proportion of units allocated to treatment (optimal allocation is 50%)
- d : minimum detectable effect between groups
- k : total number of clusters enrolled in the study
- m : average number of individuals per cluster
- ρ : cluster level ICC

Another good reference is Rutterford et al. 2015 IJE (see references at end of slides)

Clustered designs : Practical considerations

- In clustered designs, investigators must choose the number of clusters per treatment arm (*k*) and the number of individuals (*m*) to measure within each cluster.

- As a general rule of thumb, gains in power are small for m > 1/ρ (Campbell 2007)

- Designs with more clusters and fewer observations per cluster are usually optimal from a statistical perspective
  - … but are often sub-optimal from a cost / logistical perspective

- Statistical power for clustered designs is usually driven by the number of clusters per arm and by

Slide created by Ben Arnold

Example: length-for-age in WASH Benefits Bangladesh

BMJ open

Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits study design and rationale

Benjamin F Arnold,[1] Clair Null,[2,3] Stephen P Luby,[4,5] Leanne Unicomb,[4] Christine P Stewart,[6] Kathryn G Dewey,[6] Tahmeed Ahmed,[7,8] Sania Ashraf,[4] Garret Christensen,[3,9] Thomas Clasen,[2] Holly N Dentz,[2,3] Lia C H Fernald,[1] Rashidul Haque,[4,10] Alan E Hubbard,[1] Patricia Kariger,[1] Elli Leontsini,[11] Audrie Lin,[1] Sammy M Njenga,[12] Amy J Pickering,[13] Pavani K Ram,[14] Fahmida Tofail,[7] Peter J Winch,[11] John M Colford Jr[1]

To cite: Arnold BF, Null C, Luby SP, *et al*. Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits study design and rationale. *BMJ Open* 2013;**3**:e003476. doi:10.1136/bmjopen-2013-003476

ABSTRACT

Introduction: Enteric infections are common during the first years of life in low-income countries and contribute to growth faltering with long-term impairment of health and development. Water quality, sanitation, handwashing and nutritional interventions can independently reduce enteric infections and growth faltering. There is little evidence that directly compares the effects of these individual and combined interventions on diarrhoea and growth when delivered to infants and young children. The objective of the WASH Benefits study is to help fill this knowledge gap.

boards at the University of California, Berkeley, Stanford University, the International Centre for Diarrheal Disease Research, Bangladesh, the Kenya Medical Research Institute, and Innovations for Poverty Action. Independent data safety monitoring boards in each country oversee the trials. This study is funded by a grant from the Bill & Melinda Gates Foundation to the University of California, Berkeley.

Registration: Trial registration identifiers (http://www.clinicaltrials.gov): NCT01590095 (Bangladesh), NCT01704105 (Kenya).

Slide created by Ben Arnold

Presentation overview

- Big-picture concepts related to clustering
- **Example: Clustering-related considerations & nutritional intervention / child growth case study**
- Additional issues and recap

Nutritional Intervention in WASH Benefits

Behavior Change

Exclusive breastfeeding through 6 months
Continued breastfeeding until 24 months
Encourage micronutrient dense food

www.aliveandthrive.org

+

Nutritional Supplement

Nut-based daily supplement 6 – 24 months
118 kcal/day + fatty acids + micronutrients

206

Slide created by Ben Arnold

Nutritional interventions and length-for-age

- A primary outcome in the Bangladesh trial was a child's length-for-age Z-score (LAZ) measured 2 years after intervention
  - A child's length (height) is mapped to international standards based on age and sex. The score is continuous; LAZ < −2 indicates the child is stunted.

- **Stunting** is low height-for-age (ie, low LAZ) –> often due to poor nutrition
  - **Wasting** is low weight-for-height –> often due to periods of insufficient food intake

- Stunting by age 24 months is associated with life-long deficits in health and human capital (Victora 2008)

- A testable hypothesis: The nutritional intervention would improve LAZ

Slide created by Ben Arnold

Sample size steps in this example

1. Specify the parameter of interest and H_0 , H_a
2. Obtain measures of outcome variability and the ICC
3. Calculate preliminary sample size and MDE estimates
4. Check the budget and logistics

   (iterate steps 3 + 4 multiple times, potentially with multiple outcomes)
5. Settle on a final sample size / design

Slide created by Ben Arnold

Step 1: Identify the parameter of interest & hypotheses

- We were interested in estimating the difference in LAZ between the nutrition arm and the comparison arm.

- Our parameter of interest, θ, was the mean difference between groups:

$$\theta = E(Y \mid A = 1) - E(Y \mid A = 0)$$

where Y is LAZ and A is a dichotomous indicator equal to 1 if a child received the nutrition intervention and 0 if a child was in the control arm.

- Our null hypothesis, H_0, was that $\theta = 0$ (no effect).

- Our alternative hypothesis, H_a, was that $\theta > 0$ (intervention beneficial).

Slide created by Ben Arnold

Step 2: Obtain measures of outcome variability and the ICC

- We had LAZ measurements from 982 children < 3 years old from rural Bangladesh that were part of an existing cohort. (Huda 2012)

- To estimate the variability of LAZ you can calculate the standard deviation in R using summarise function in the dplyr package

210

Slide created by Ben Arnold

Step 2: Obtain measures of outcome variability and the ICC

- and estimate the ICC using the ICC package in R.

```
install.packages("ICC", dependencies = TRUE)
library(ICC)
ICCest(x = clusterid, y = laz, data = anthro)
$ICC[1] 0.008
```


Slide created by Ben Arnold

Steps 3, 4 & 5: Calculate preliminary sample size and MDE estimates, check budget/logistics, iterate

- In the case of this study, we repeated MDE calculations for a range of designs and settled on the following:
  - A single post-treatment measurement of LAZ after 2 years of intervention
  - 7 newborns per cluster (fixed by demographics)

Slide created by Ben Arnold

Step 6 : Given a design, calculate the minimum detectable effect size

$$
\begin{aligned}
MDE &= (Z_{1-\beta} + Z_{1-\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{k}} \times \sqrt{\rho + \frac{1-\rho}{m}} \\
&= (0.84 + 1.64)\sqrt{\frac{1}{0.33 \times (1-0.33)}}\sqrt{\frac{1.24^2}{270}} \times \sqrt{0.008 + \frac{1-0.008}{7}} \\
&= 0.154
\end{aligned}
$$

- Note: $k = 270$ and $P = 0.33$ because we are comparing 1 treatment arm (90 clusters) to a double-sized control arm (180 clusters)

- The design will be able to detect a difference of $+0.15$ LAZ with 80% power and a one-sided $\alpha = 0.05$ for the comparison of any treatment arm to the control arm. At age 24 months, this is equivalent to ≈ 0.45 cm.

Slide created by Ben Arnold

Very easy to implement in R!

```
#-------------------------------------
# mde function
# k:     total number of independent units (clusters)
# m:     average number of repeated measures per unit
# sd:    standard deviation of the outcome in the population
# rho:   intra-class correlation of the outcome within units
# P:     proportion of clusters allocated to treatment (optimal = 0.5)
# alpha: Type-II error rate (for a one-sided test, double the alpha)
# power: 1 - Type II error rate
#-------------------------------------

mde <- function(k,m,sd,rho,P=0.5,alpha=0.05,power=0.8) {

  Za <- qnorm(1-(alpha/2))

  Zb <- qnorm(0.8)

  return( (Za+Zb) * sqrt(1/(P*(1-P)))*sqrt(sd^2/k) * sqrt(rho+(1-rho)/m) )

}

# one-sided hypothesis test, consistent with the example (double alpha from 0.05 to 0.1)

mde(k=270,m=7,sd=1.24,rho=0.008,P=0.33,alpha=0.1,power=0.8)

[1] 0.1544049

# two-sided hypothesis test

mde(k=270,m=7,sd=1.24,rho=0.008,P=0.33,alpha=0.05,power=0.8)

[1] 0.1739726
```


Slide created by Ben Arnold

Example of an intermediate step (alternate scenarios):

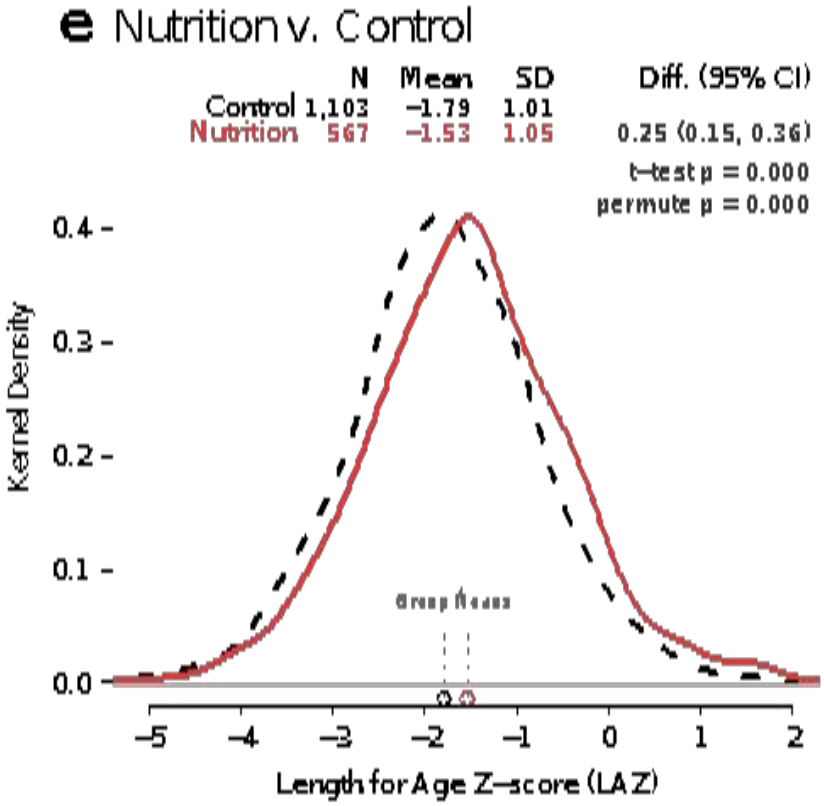
- Considering alternate designs with a fixed total N (1,890), but different allocation:

Children per cluster (m)	Total clusters (k)	MDE (d)
7	270	0.154
10	189	0.156
14	135	0.158
21	90	0.162
30	63	0.167

- Not much increase in MDE from enrolling fewer, larger clusters so this would be a preferred strategy if logistics permit (in this real example, we were limited by demographics and timeline)
- Note: since ρ = 0.008, our rule of thumb suggests we could enroll up to 1/0.008 = 125 children per cluster without a large loss in power

Slide created by Ben Arnold

… and in the end: WASH Benefits Nutrition results



Luby et al. 2018

- Assumptions were slightly conservative
- SD was 1.01 rather than 1.24
- Actual MDE was about 0.11, slightly smaller than the designed MDE of 0.15

- Nutrition intervention improved LAZ by +0.25 [216]
  - Still very far from removing all growth faltering (mean LAZ = -1.53)

Slide created by Ben Arnold

Presentation overview

- Big-picture concepts related to clustering
- Clustering-related considerations & nutritional intervention / child growth case study
- **Additional issues and recap**

Summary of Key Points (1)

- Sample size calculations require that you specify:
  1. basic design (fraction treated/exposed), parameter of interest, and hypothesis
  2. outcome variability (get this automatically for binomial outcomes)
  3. minimum detectable effect
  4. desired level of power and Type I error (alpha)

- Additionally, for clustered designs you also

Slide created by Ben Arnold

Summary of Key Points (2)

- The power of clustered designs is most sensitive to the number of clusters per arm and the ICC.
- Even if the ICC is small, there can be large design effects if cluster size is large
- Use estimates of variability and the ICC from similar populations to your study.
  - If estimates are not available to you, calculate sample size and the MDE over a range of plausible values for these parameters.

Slide created by Ben Arnold

References

- Arnold, B.; Hogan, D.; Colford, J. & Hubbard, A. Simulation methods to estimate design power: an overview for applied research. BMC Med Res Methodol, 2011, 11, 94
- Arnold, B. F.; Null, C.; Luby, S. P.; et al. Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits Study design and rationale. BMJ Open, 2013, 3, e003476
- Colford JM, Schiff KC, Griffith JF, Yau V, Arnold BF, Wright CC, et al. Using rapid indicators for Enterococcus to assess the risk of illness after exposure to urban runoff contaminated marine water. Water Res. 2012;46: 2176–2186.
- Colford, J. M.; Wade, T. J.; Schiff, K. C.; et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology, 2007, 18, 27-35
- Campbell, M. J.; Donner, A. & Klar, N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med, 2007, 26, 2-19
- Duflo, E.; Glennerster, R. & Kremer, M. Using Randomization in Development Economics Research: A Toolkit. 61 Handbook of Development Economics, 2007, Volume 4, 3895-3962
- Feiveson, A. H. Power by simulation. Stata Journal, 2002, 2, 107-124
- Leon, A. C. Sample-size requirements for comparisons of two groups on repeated observations of a binary outcome Eval Health Prof, 2004, 27, 34-44
- Murray, D. M.; Varnell, S. P. & Blitstein, J. L. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health, 2004, 94, 423-432

- Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44: 1051–1067.
- Schulz, K. F. & Grimes, D. A. Sample size calculations in randomised trials: mandatory and mystical. Lancet, 2005, 365, 1348-1353
- Selvin, S. Statistical Analysis of Epidemiologic Data (2nd ed). Oxford University Press, 1996
- Victora, C. G.; Adair, L.; Fall, C.; Hallal, P. C.; Martorell, R.; Richter, L.; Sachdev, H. S. Maternal and child undernutrition: consequences for adult health and human capital. Lancet, 2008, 371, 340-357
- Victora, C. G.; de Onis, M.; Hallal, P. C.; Blossner, M. & Shrimpton, R. Worldwide timing of growth faltering: revisiting implications for interventions. Pediatrics, 2010, 125, e473-e480

Slide created by Ben Arnold