




Representative vs. Non-Representative samples

Briana Mezuk
MCUAAAR AnC Mini Training
1/11/2023





Goals for today:

- Define key terms relevant to sampling theory
- Provide a framework for understanding the utility of non-representative samples
- Set up the full-training session on mapping and modeling population (Census) data

Definitions

Population: The entire group that you want to make inferences about

Examples

- All undergraduate students at the University of Michigan
- All physicians who work at Michigan Medicine
- All residents of the state of Michigan
- All African Americans
- All ED admissions for opioid overdose
- All nations

Sample: Some subset of the population that you will collect data from in order to draw inferences about the population

Examples

- Undergraduate students who enrolled at UM in 2021 at the Ann Arbor campus
- Physicians who see >100 patients each week at Michigan Medicine
- Residents who speak English, have a working telephone number, and who have lived in MI for at least 6 months
- African Americans recruited for the National Survey of American Life
- ED admissions to hospitals in counties with at least 100,000 people
- Nations with reliable mortality records

Definitions

NOTE: The SIZE of the sample is not in any way a measure of its “representativeness”!

Representative sample

- A sample that is accurately representative of the characteristics of the population it is seeking to represent
 - Which characteristics are you trying to represent?
Not all meaningful characteristics are known/observed.
- Random sample: All individuals in the population have **equal probability** of being selected, regardless of their characteristics.
- Probability sample: All individuals have a probability of being selected, but that **probability varies based on their characteristics**.
 - Often paired with stratification and clustering to create subsets
 - Examples: NSAL

Non-representative sample

- Not necessarily a “Convenience” sample
 - Convenience = CRaP Sampling
 - Cheap
 - Readily available
 - Presto!
- Purposeful sampling: Investigator uses knowledge about the topic to inform sample
- Respondent-driven sampling: Modification of “snowball” sampling where respondents can refer others in their network with specific characteristics into the study

Pros and Cons of Representative samples

Pros

1. [Assuming a large enough sample size] Captures the heterogeneity (especially in exposures) represented in the population
2. [Assuming minimal selection and attrition bias] May have fewer threats to external validity (i.e., more generalizable)
3. Most appropriate sampling approach for estimating prevalence of health conditions in the population and for generating statistics important for understanding healthcare needs, service planning, etc.

Cons

1. Too expensive/impracticable
2. Unless the N is very large, it will have few cases of rare outcomes/exposures
3. Unless all subgroups are collected with sufficient sample sizes, will necessitate imprecise comparisons (e.g., White vs. non-White, Rural vs. Urban) or evaluations of moderation

Pros and cons of non-representative samples

Con #1

Non-representative samples are too homogenous (especially in terms of exposures)

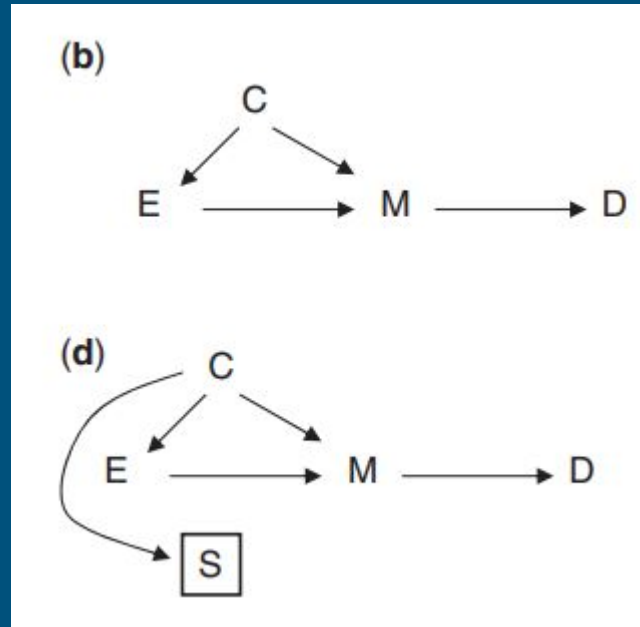
- **Response:** Homogenous samples can enhance the statistical power/statistical efficiency to detect main effects, especially in under-represented subgroups or for rare exposures
 - If you were interested in understanding risk factors for LBW among women who gave birth after age 40, a representative sample would not be the most efficient
- **Response:** Homogenous samples can reduce the risk that interactions/moderators are simply statistical artifacts due to outliers
 - You want \sim N's across your moderator strata
- **Response:** "Generalizability" is study-question dependent!
 - Do you really need a representative sample to test the hypothesis that tobacco causes lung cancer?

Pros and cons of non-representative samples

Con #2

If exposure is associated with probability of selection/some feature you are using to sample on, your exposure-outcome relationship may be biased

Ex. If SES at birth is associated with both tobacco use and likelihood of going to medical school, then using a sample of physicians to test the relationship between tobacco and lung cancer may generate a biased estimate (Collider bias)



Representative sample:
E→D through M, with confounder C

Non-Representative sample:
E→D through M, with confounder C that also impacts selection into the sample

Pros and cons of non-representative samples

Con #2

If exposure is associated with probability of selection/some feature you are using to sample on, your exposure-outcome relationship may be biased

Ex. If SES at birth is associated with both tobacco use and likelihood of going to medical school, then using a sample of physicians to test the relationship between tobacco and lung cancer may generate a biased estimate (Collider bias)

- **Response:** Unless the links between the exposure and selection factor are very strong, the amount of bias that will be introduced is small
 - Simulations estimate this bias is typically on the order of 10%.
- **Response:** Exposures like tobacco use, SES, etc. are confounders even in representative samples and failing to adequately control for them can also result in biased estimates
 - Therefore, restricting to a specific group (e.g., non-smokers, high SES, etc.) allows for better control of these confounders

Pros and cons of non-representative samples

Con #3

If a mediator is associated with selection into the sample, the exposure-outcome relationship may be biased

Response: This will most likely under-estimate (downwardly bias) the exposure-outcome relationship

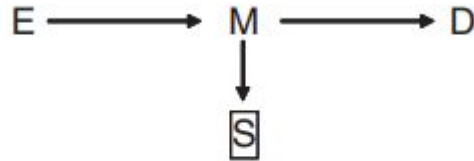


Figure 2 Selection of cohort participants (S) is affected by the mediator (M) of the exposure (E)-outcome (D) association

Summary

- With high internal validity, the valid assessment of the causal relationship may be widely generalizable, and does not require that the participants be representative of those to whom the new evidence will be applied.
 - We are ALWAYS applying findings generated from samples that are not representative of the current population, even if those are samples of past individuals to current/future individuals
- The purpose of the study - the scientific question to be answered - needs to drive all design decisions, including the benefit/drawback of representative vs. non-representative samples.
 - Studies seeking to understand causal relationships rarely require representative samples to generate inferences that are internally-valid.

Further reading



Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2013; all rights reserved.

International Journal of Epidemiology 2013;**42**:1012–1014
doi:10.1093/ije/dys223

POINT COUNTERPOINT

Why representativeness should be avoided

Kenneth J Rothman,^{1,2} John EJ Gallacher³ and Elizabeth E Hatch¹

¹Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA, ²RTI Health Solutions, RTI International, Research Triangle Park, NC, USA and ³Institute of Primary Care and Public Health, Cardiff University, Cardiff, UK

Accepted 21 November 2012

Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2013; all rights reserved.

International Journal of Epidemiology 2013;**42**:1018–1022
doi:10.1093/ije/dyt103

Commentary: Representativeness is usually not necessary and often should be avoided

Lorenzo Richiardi,^{1*} Costanza Pizzi^{1,2} and Neil Pearce^{3,4}

¹Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Turin, Italy, ²Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, London, UK, ³Departments of Medical Statistics and Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK and ⁴Centre for Public Health Research, Massey University, Wellington, New Zealand

*Corresponding author. Cancer Epidemiology Unit, Via Santena 7, 10126 Torino, Italy. E-mail: Lorenzo.richiardi@unito.it

Accepted 6 February 2013