UNIVERSITY OF MICHIGAN

# Inference for Non-Probability Samples

Brady T. West

Research Associate Professor, Survey Research Center, Institute for Social Research

# Lecture Overview

- **Problem**: **<u>Non-probability samples</u> do not let us rely on sampling theory** for making population inferences based on expected sampling distributions

**Two Approaches:**

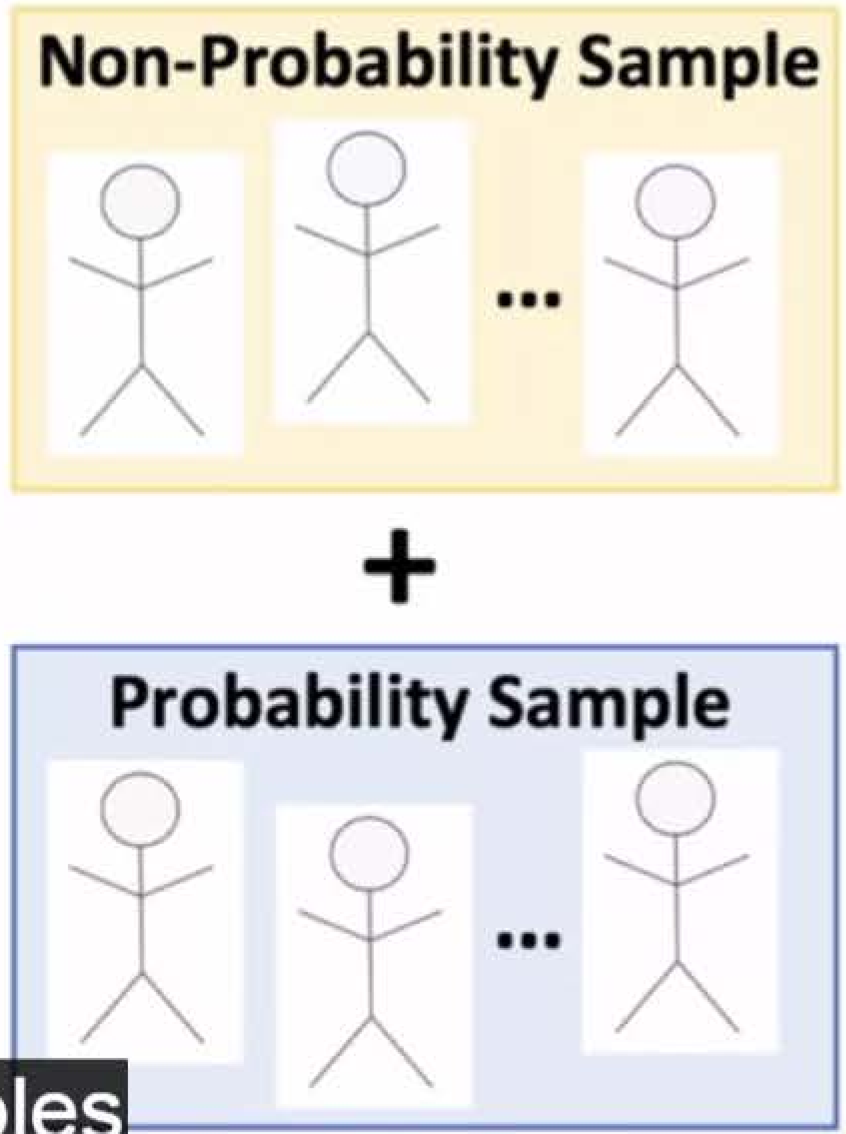
I. Quasi-Randomization (or pseudo-randomization)

II. Population Modelling

We're going to introduce each of these different two approaches in this lecture.

# Approach 1: "Quasi-Randomization"

**Big Idea**: Combine data from non-probability sample with data from probability sample that collected same types of measures




but it's important that both samples

# Approach 1: "Quasi-Randomization"

**Example:**

__ years-old?

White/Black/Asian/...

if we measure blood pressure, age, and race/ethnicity on a sample of **volunteers**,

→ combine with prior data from a probability sample (e.g., NHANES) that collected the same three measures

age, and race ethnicity on a sample of volunteers.

# Approach 1: "Quasi-Randomization"

- **Stack** the two data sets; non-probability sample may have other response variables we are really interested in
- **Code** NPSAMPLE = 1 if member of non-probability sample
  NPSAMPLE = 0 if member of probability sample

| NPSAMPLE | BloodPressure | Age | Race/Ethnicity | Response1 | Response2 |
|---|---|---|---|---|---|
| 0 | 100 | 52 | White | 83 | Yes |
| 0 | 120 | 45 | Asian | 92 | No |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1 | 130 | 64 | Black | 91 | No |
| 1 | 110 | 38 | White | 79 | No |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# Approach 1: "Quasi-Randomization"

Fit **logistic regression model**

→ predicting NPSAMPLE with common variables
weighting non-probability cases by 1 and
weighting probability cases by their survey weights

More on logistic regression later!

The weights that we've been talking about in

6:52 / 17:23

# Approach 1: "Quasi-Randomization"

**Big Idea:**

1. **Can predict probability of being in non-probability sample**, within whatever population is represented by probability sample!
2. **Invert predicted probabilities** for non-probability sample, **treat as survey weights** in standard weighted survey analysis

$$Survey\ Weight = \frac{1}{Predicted\ Probability}$$

we treat those, the inverse of those probabilities

# Approach 1: "Quasi-Randomization"

**Issue:** How to estimate sampling variance?

Not entirely clear …

Some kind of **replication method** is recommended (e.g. computing weighted estimates based on **bootstrap samples** or **jackknife samples** of the original units)

so-called replication method is needed to estimate that sampling variance.

# Approach 1: "Quasi-Randomization"

For a deep (and technical) dive into this approach, see the following article:

Elliott, M.R. and Valliant, R. (2017).
Inference for Non-Probability Samples.
*Statistical Science*, 32(2), 249-264.

For a deep and fairly technical dive into more about this approach,

# Approach 2: Population Modeling

**Big Idea:**

1. **Use predictive modeling** to **predict aggregate sample quantities (usually totals) on key variables** of interest for population units **not** included in the non-probability sample
2. **Compute estimates of interest using estimated totals**

$$\text{e.g } Weighted\ Mean = \frac{Predicted\ Total\ Estimate}{Estimated\ Population\ Size}$$

**Note**: Don't need probability sample with same measures

don't need a probability sample that collected the same measures.

# Approach 2: Population Modeling

More about good models later!

- **Need good regression models** to predict key variables using other auxiliary information available at aggregate level (e.g., totals for overall population)

- **Standard errors** can be based on fitted regression models, or using similar replication methods!

See Elliott and Valliant article for more details

that Elliot and Valliant article for more details on this type of approach.

# Summary

**Inferential methods for non-probability samples <u>need to</u>:**

- **Leverage other auxiliary information** (reference probability samples or regression models)
- **Predict values** for population cases not included in probability sample (or at least probability of being included in non-probability sample!)

In absence of this information …
we will have a **hard time** making good population inferences!

it becomes a really difficult problem making good population inference, okay.

16:41 / 17:23

Suppose that you have collected data from a non-probability sample, and you've also identified a reference probability sample. The non-probability sample measured height, weight, and years of education. The probability sample measured age, gender, and income. You wish to make inference about the mean years of education in the population. Which approach could you use?

- The quasi-randomization approach: just stack the two data sets and estimate the probability of being in the non-probability sample as a function of all the variables.
- The superpopulation modeling approach: fit a regression model predicting years of education for the cases that were not in the non-probability sample.
- Simply estimate the mean age and the sampling variance for the estimated mean using the non-probability sample.
- Nothing: we don't have any common variables in the two samples to use for estimation.

**Correct**

For any of these estimation techniques for non-probability samples, we need to have common variables in the two data sets. Simple estimation of the mean based on the non-probability sample may lead to a biased estimate, and we can't estimate sampling variance from the non-probability sample.