Inference for Non-Probability Samples

Brady T. West

Lecture Overview

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

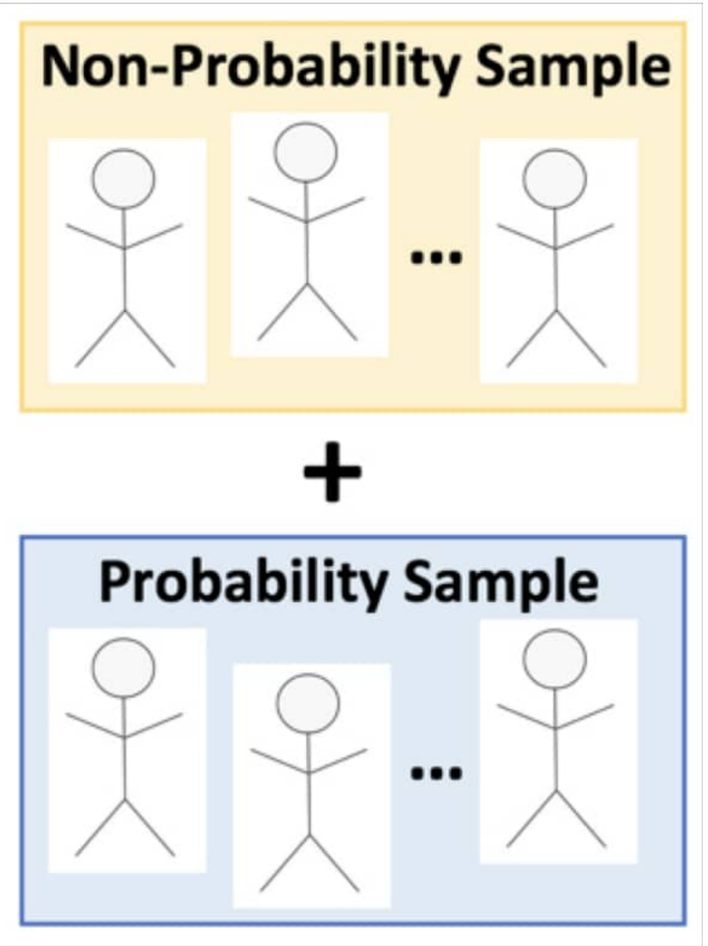
Two Approaches:

I. Quasi-Randomization (or pseudo-randomization)

II. Population Modelling

Approach 1: "Quasi-Randomization"

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



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

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	Blood Pressure	Age	Race/Ethnicity	Response 1	Response 2
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!

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

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

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)

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.

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**

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

Note: Don't need probability sample with same measures

Approach 2: Population Modeling

- **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

Summary

Inferential methods for non-probability samples need to:

- **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!