Non-Probability Sampling, Part 1

Brady T. West

Lecture Overview

- What defines a non-probability sample?

- **Common examples** of non-probability samples

- Intro to **two common population inference methods**

- Example of non-probability sampling: **Twitter data**

Logo from Twitter

What Are Non-Probability Samples?

- **Features of Non-probability samples:**
  - Probabilities of selection **can't be determined** for sampled units
  - **No random selection** of individual units
  - Sample divided into groups (strata) or clusters, but **clusters not randomly sampled** in earlier stage
  - Data collection often very **cheap** ¢ relative to probability sampling

Examples of Non-Probability Samples

- Studies of **volunteers**

Do you suffer from XXX?

Learn more about our clinical research study for an investigational drug
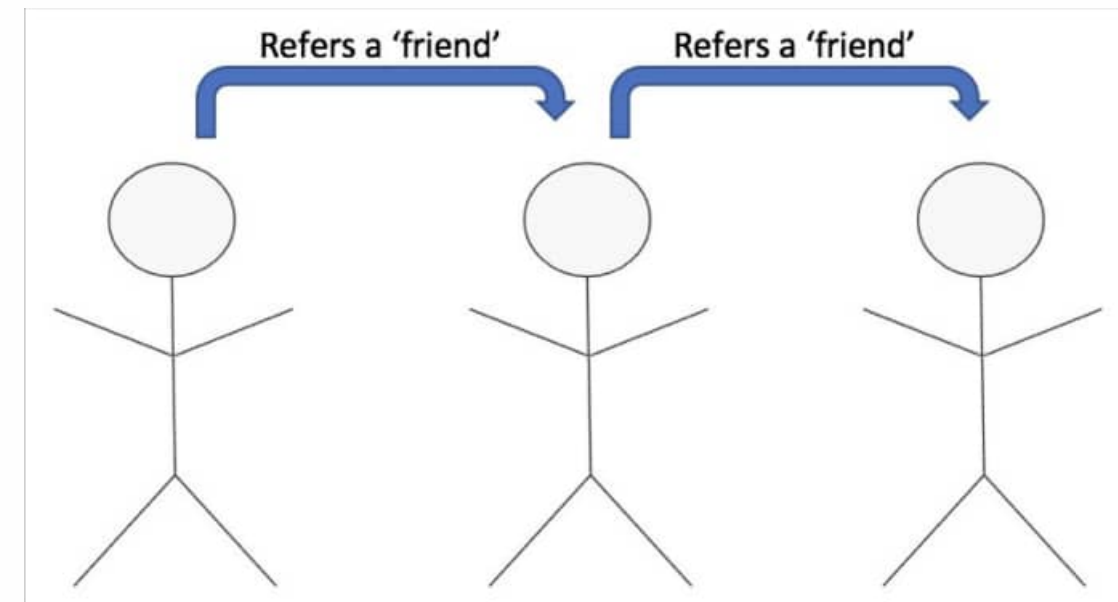
ABC Clinical Trials: (XXX) XXX-XXXX

Examples of Non-Probability Samples

- Studies of **volunteers** (e.g., clinical trials)

- **Opt-in** / Intercept web surveys

Examples of Non-Probability Samples

- Studies of **volunteers** (e.g., clinical trials)

- **Opt-in** / Intercept web surveys

- **Snowball** samples

Examples of Non-Probability Samples

- Studies of **volunteers** (e.g., clinical trials)

- **Opt-in** / Intercept web surveys

- **Snowball** samples
  (e.g., word-of-mouth data collection)

- **Convenience** samples

All Students in Psych 101

Examples of Non-Probability Samples

- Studies of **volunteers** (e.g., clinical trials)

- **Opt-in** / Intercept web surveys

- **Snowball** samples
  (e.g., word-of-mouth data collection)

- **Convenience** samples
  (e.g., all students in Psych 101)

- **Quota** samples

Recruit 1,000 males and 1,000 females in any way

Examples of Non-Probability Samples

Common Feature:
Probabilities of selection cannot be determined a priori!

So What Is The Problem?

- Non-probability sample → **no statistical basis for making inference** about larger population from which sample selected

- **Knowing probabilities of selection** (in addition to population strata and randomly sampled clusters) → **can estimate features of sampling distribution** *if were to take many random samples using same design*

So What Is The Problem?

- Sampled units **not selected at random** →
  strong risk of **sampling bias**
  (e.g., people actually interested in visiting particular web site)

So What Is The Problem?

- Sampled units **not selected at random** →
  strong risk of **sampling bias**
  (e.g., people actually interested in visiting particular web site)
- Sampled units **not generally representative**
  of larger target population of interest

Not Representative

Target Population

So What Is The Problem?

- Sampled units **not selected at random** →
strong risk of **sampling bias**
(e.g., people actually interested in visiting particular web site)

- Sampled units **not generally representative**
of larger target population of interest

- **"Big data"** (e.g., information from millions of tweets 🐦)
often from non-probability samples ~ be careful!

Logo from Twitter

So What Can We Do?

- Many data sets arise from non-probability samples **… can we say anything about a larger population?**

- **Two possible approaches**:
    - Pseudo-Randomization
    - Calibration

For technical "deep dive" into estimation approaches: Elliott and Valliant (2017, Statistical Science)