Sampling: Course Introduction

```
$ echo "Data Science Institute"
```

Learning Goal

- Explain the structure of this course
- Go through an overview of what to expect in the rest of this course (high-level)
- Sampling Intro!
- But first...

Hello!

- Meet the technical facilitator and learning support!
- Feel free to introduce yourself in the chat!

So...

Why are we here?

Study Design

- What comes before and after writing code to analyze our data?
  - Where do we get the data?
  - How do we get the data?
  - Why are we collecting data in specific ways and quantities?
  - How do we communicate our results?
  - Who are our results for?

Case Study: Why does sampling matter?

- 2016 US Presidential Election
- Donald Trump (Republican Nominee) vs. Hilary Clinton (Democratic Nominee)
- Polarization!
- Strongly negative campaigning from both parties
    - “Of nearly 70,000 political television ads that ran in recent days, less than one in 10 were primarily positive, a CNN analysis of data from Kantar Media/CMAG shows.”

Election Predictions

- Statisticians and newsrooms use data to predict which candidate will become president
- These data are largely from pre-vote surveys and interviews with prospective voters, with some projection based on historical trends (e.g. an area has always voted Democrat)

Election Results

- Trump overperformed expectations, winning the electoral college, and was elected President
- The predictions were wrong! Almost all of them!
- But why?

What do you think are some potential reasons that the predictions about the 2016 US Presidential election were so wrong?

Where does sampling come in?

- Timing
- Medium
- Dishonest responses
- Nonresponse bias
- Non-representative sampling

In a well-designed survey, we need to consider how each of these factors may impact our conclusions.

The Basics (Sampling Terminology)

Sampling

"Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population."

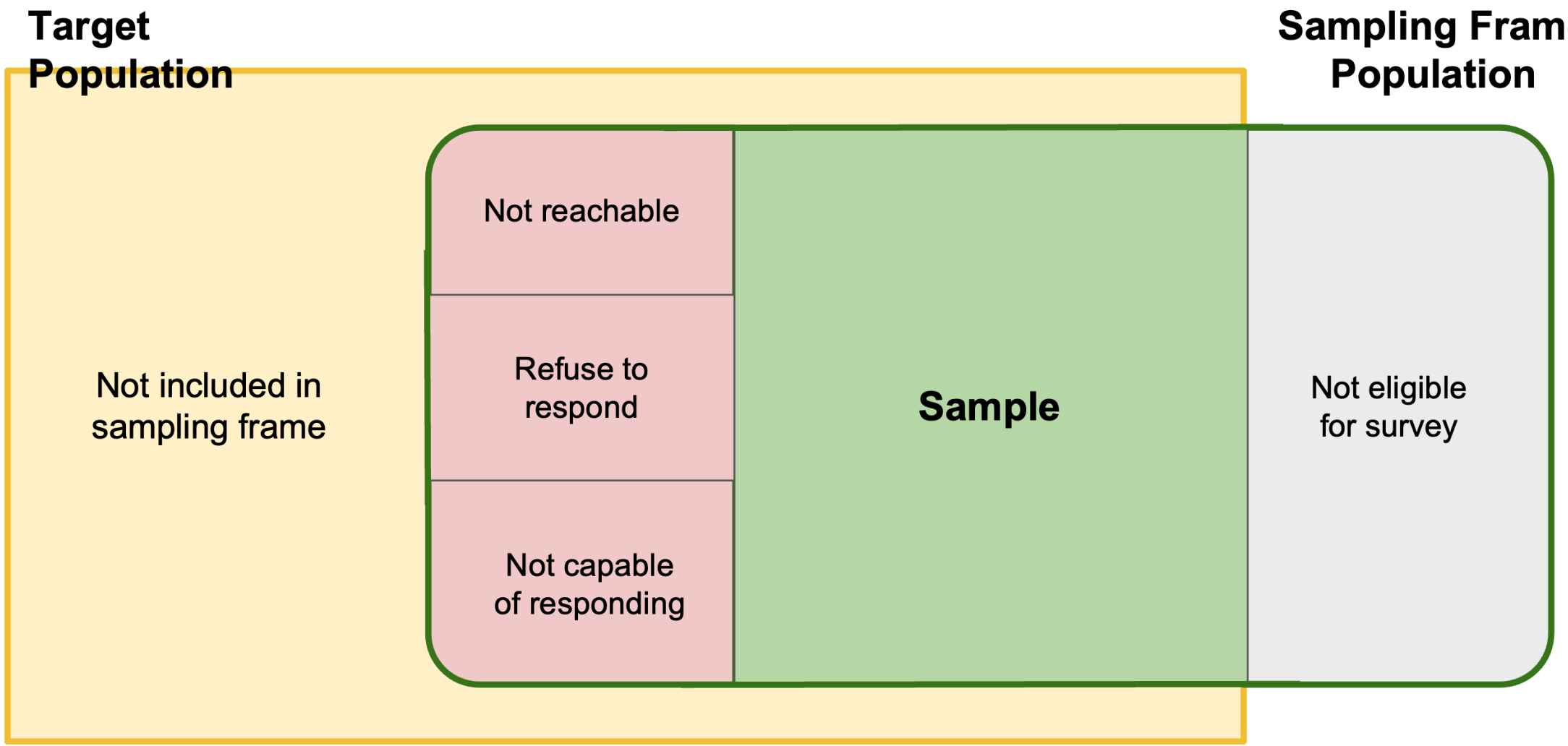
Thompson, 'Sampling: Third Edition' (2012)

What is a population?

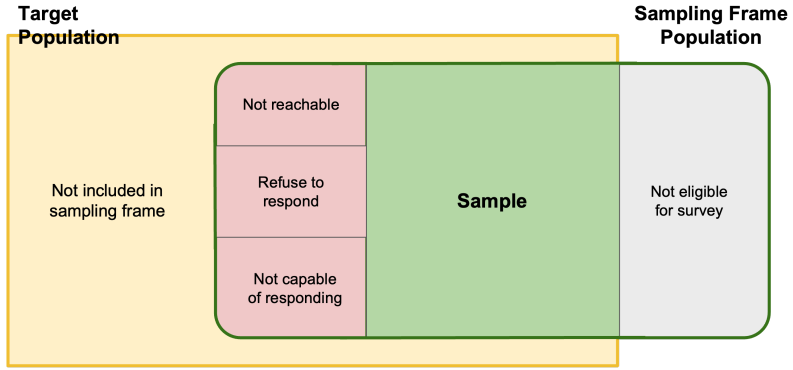
- Real or hypothetical set of units with characteristics of attributes that can be modeled by random variables
- Can be infinite or finite
- Inclusion and exclusion criteria must be clearly defined

Population levels

- **Target population** = all units covered by our study objective
- **Frame population** = all units covered by our sampling frame (the parts of our population we take our sample from)
- **Sampled population** = what we actually collected responses from


Target Population
Sampling Frame Population
Not included in sampling frame
Not reachable
Refuse to respond
Not capable of responding
Sample
Not eligible for survey

- We want the characteristics of the sample (green) to match the characteristics of the target population (yellow).
- For this to be true, all of the following need to match the characteristics of the target population (yellow) as well:
    - Individuals in the sampling frame (green)
    - Individuals excluded from the sampling frame (yellow)
    - Non-respondents (pink)
- Individuals not eligible for the survey (grey) must be well defined, distinct from the target population, and properly excluded from analysis



Why sampling?

- Sampling lets us collect data about individuals (sampling units) from only a subset of our target population, called our sample, and make inferences about the whole population
- This is more cost- and time-efficient than sampling an entire population; usually easier and more realistic
- How do we collect samples?

Surveys

- “A survey is an investigation about the characteristics of a given population by means of collecting data from a sample of that population and estimating their characteristics through the systematic use of statistical methodology.” (OECD)

Surveys

- “The term survey covers any activity that collects or acquires statistical data. Included are censuses, sample surveys, the collection of data from administrative records and derived statistical activities” (Statistics Canada)

Activity: Exploring Samples (Together)

A pilot survey for The Canadian Longitudinal Study on Aging (CLSA) was conducted in the province of Ontario. The survey intended to cover the general population of the province with age 45–80 (inclusive). Survey questionnaires were sent to randomly selected individuals through regular mail. Individuals and their mailing addresses were selected and obtained from the Provincial Health Records.

What are the target population, sampling units, frame population, and sampled population?

Activity: Exploring Samples (Try it!)

A teacher wanted to find the average number of hours each week spent on watching TV by 4 and 5 year old children in Waterloo. She conducted a survey using the list of 123 kindergartens administered by the Waterloo Region District School Board. She first randomly selected ten kindergartens from the list. Within each kindergarten, she was able to obtain a complete list of all 4 and 5 year old children, with contact information for their guardians. She then randomly selected 50 children from the list and mailed the survey to their guardians. The sample data were compiled from completed and returned surveys.

What are the target population, sampling units, frame population, and sampled population?

What will we cover in this course?

Coming up...

- Probability 101
- Populations and Samples
- Types of Samples
- Simple Random Samples, Stratified Samples, Cluster Samples, Resampling
- Errors in Sampling
- Survey Quality and Design
- Sampling Ethics and Privacy

Assignments

- 2 Assignments
- In the Github repo!

Learning Outcomes

- Ability to implement simple probability samples.
- Ability to understand more complicated sampling procedures and the tradeoffs involved.
- Ability to identify and understand sources of error or inaccuracies in data as a result of sampling strategies.
- Development of intuition around survey quality.

Be able to choose and justify a sampling approach based on your data science objectives