Data Collection

Lecture 2

Xiaofei Shi

Learning objectives

- Understand the common goals of different datasets
- Understand the common errors in using datasets
- Understand basic sampling terminology and sampling weights

When statistics are not based on strictly accurate calculations, they mislead instead of guide. The mind easily lets itself be taken in by the false appearance of exactitude which statistics retain in their mistakes, and confidently adopts errors clothed in the form of mathematical truth.

-- Alexis de Tocqueville, Democracy in America

COVID-19 Data

Recall the beginning of the pandemic before the lockdown

- Who was advised to get tested?
- What was being reported?
- What decisions were made based on the data?

COVID-19 Data

Recall the beginning of the pandemic before the lockdown

- Who was advised to get tested?
- What was being reported?
- What decisions were made based on the data?

Data Collection is Costly

- Measurements can be costly
- Gathering subjects can be costly
  - e.g. bipolar patient with sleeping disorders
- Operational logistics

Common Goals for Data Collections

- Prediction
- Monitor and pattern detection
- Estimation
- Causality

Prediction focused dataset

Minimum requirement:

- Features (X, independent variable)
- Labels/Response (Y, dependent variable)

Example:

- Advertisement targeting

Common error:

- Building a causal model from a prediction focused dataset

Monitoring focused datasets

Subgoals:

- Decision making or anomaly detection (e.g. economic depression)
- Motivate new hypotheses (e.g. subpopulations in your user base)

Requirements:

- Features/Outcomes
- (post-analysis validation)

Common error:

- These are not always representative of the population

Estimation focused datasets

Requirements:

- Outcome
- Sampling weights

Example:

- National Health and Nutrition Examination Survey

Common error:

- Ignoring the sampling weights
- These are not experiments

Causality focused datasets

Requirements:

- Treatment
- Outcome

Example:

- Clinical trials

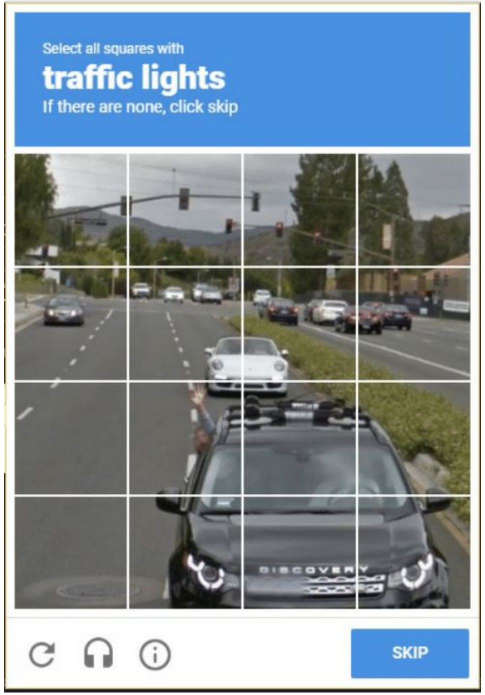
Common error:

- These do not need to be representative, e.g. college studies
- Using features to subdivide the sample until you find significance

Practice - there can be multiple answers

What are the goals of the following datasets?

- Test scores or GPA datasets
- Followers for a Instagram influencer
- CAPCHA dataset
- Ethnicity during an interview or application



Select all squares with
traffic lights
If there are none, click skip

An interesting traffic light

How to get a representative sample?

- Is representativeness the only goal? Why do we care?
    - How would you describe this mathematically?
- How do we do this?

Sampling Terminology

- Observation unit/element:

  An object on which a measurement is taken.
- Target population:

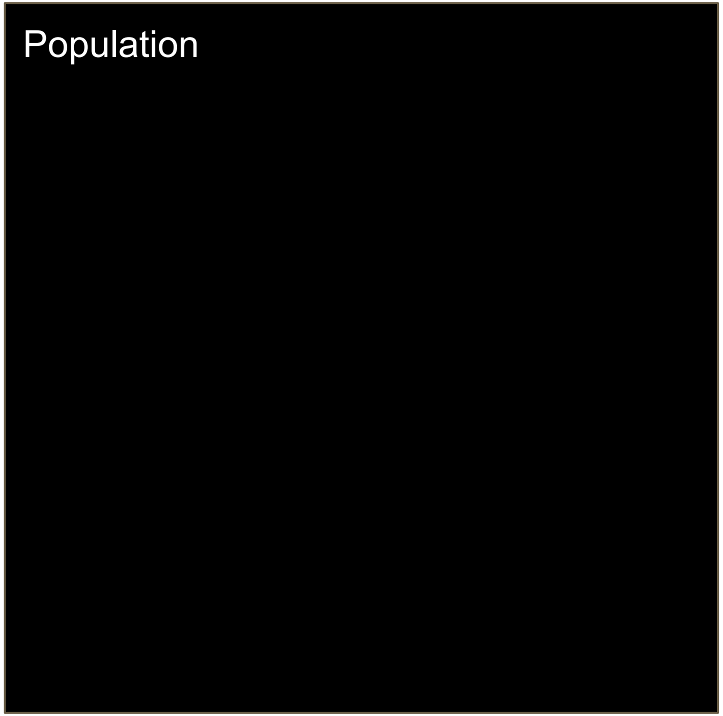
  The complete collection of observations we want to study.
- Sample:

  A subset of population.
- Sampled population:

  The population from which the sample was taken.
- Sampling unit:

  A unit that can be selected for a sample.
- Sampling frame:

  A list,map,or other specification of sampling units in the



Population

Sampling Terminology

- Observation unit/element:

  An object on which a measurement is taken.
- Target population:

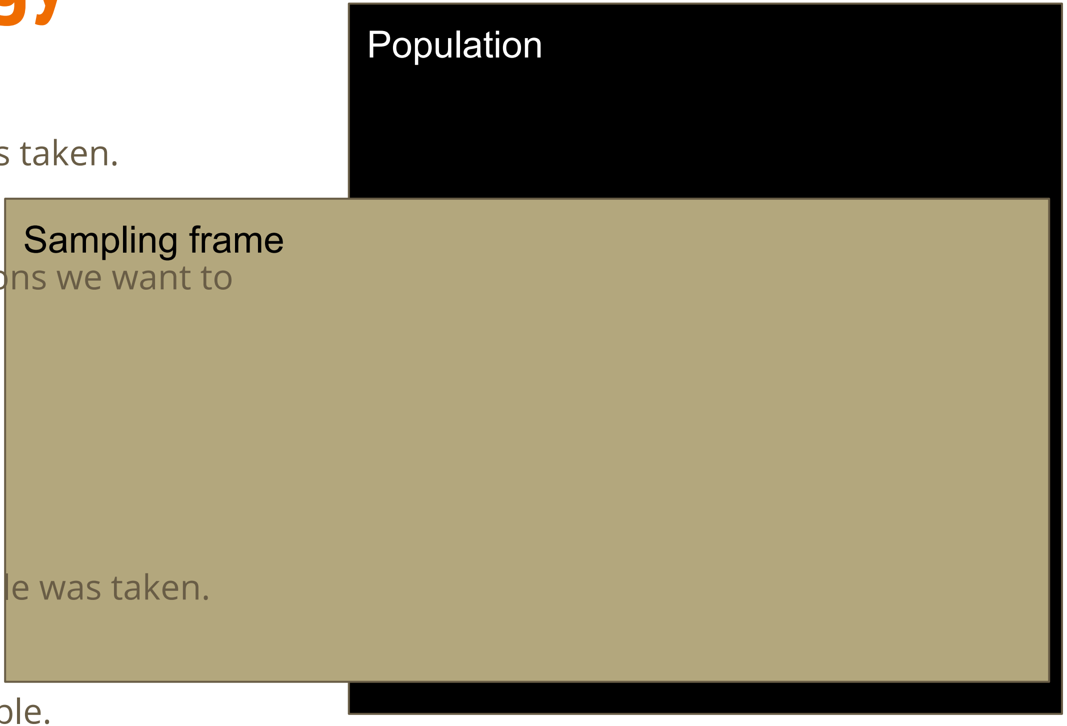
  The complete collection of observations we want to study.
- Sample:

  A subset of population.
- Sampled population:

  The population from which the sample was taken.
- Sampling unit:

  A unit that can be selected for a sample.
- Sampling frame:

  A list,map,or other specification of sampling units in the



Sampling Terminology

- Observation unit/element:

  An object on which a measurement is taken.
- Target population:

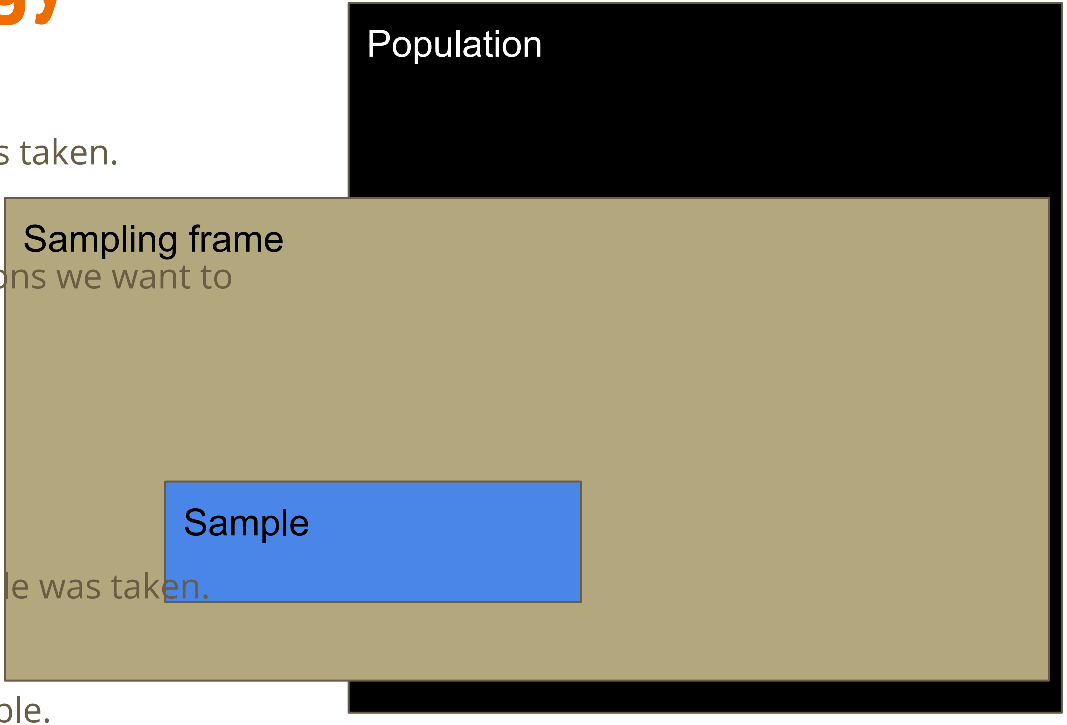
  The complete collection of observations we want to study.
- Sample:

  A subset of population.
- Sampled population:

  The population from which the sample was taken.
- Sampling unit:

  A unit that can be selected for a sample.
- Sampling frame:

  A list,map,or other specification of sampling units in the



Population

Sampling frame

Sample

Bias in Sampling

- In general, we hope that the samples we collect can represent the target population well.

- What could possibly go wrong?

Bias in Sampling

- Selection bias:

- Measurement error:

Sampling methods - multi-stage sampling

Simple random sample: sampling uniformly without replacement within the sampling frame

Cluster sampling: SRS across the clusters, then measure every unit within the sampled cluster

Stratified sampling: for every cluster, perform a SRS (size can vary)

Multi-stage sampling: mix of the sampling methods above

Questionnaire Design

- Always test your questions before taking the survey.
- Keep it simple and clear.
- Use specific questions instead of general ones, if possible.
- Relate your questions to the concept of interest.
- Decide whether to use open or closed questions.
- Report the actual question asked.
- Avoid questions that prompt or motivate the respondent to say what you would like to hear.
- Consider the social desirability of responses to questions, and write questions that elicit honest responses.
- Avoid double negatives.
- Use forced-choice, rather than agree/disagree questions.
- Ask only one concept per question.
- Pay attention to question order effects.

Sampling methods - SRS

Simple random sample: sampling uniformly without replacement within the sampling frame

Sampling methods - cluster sampling

Simple random sample: sampling uniformly without replacement within the sampling frame

Cluster sampling: SRS across the clusters, then measure every unit within the sampled cluster

Sampling methods - stratified sampling

Simple random sample: sampling uniformly without replacement within the sampling frame

Cluster sampling: SRS across the clusters, then measure every unit within the sampled cluster

Stratified sampling: for every cluster, perform a SRS (size can vary)

Sampling methods - quota sampling

Simple random sample: sampling uniformly without replacement within the sampling frame

Cluster sampling: SRS across the clusters, then measure every unit within the sampled cluster

Stratified sampling: for every cluster, perform a SRS (size can vary)

Multi-stage sampling: mix of the sampling methods above

Quota sampling (bad): for every cluster, get a convenient sample

Common Problem - Fitting models when you have the population

- NYC Public School admissions
- Class surveys

Why do we care about the sampling method?

$$\mu = \frac{1}{N} Y_i$$

The source of randomness comes from sampling

$$\mu = \frac{1}{N} Y_i$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{N} 1[i \in sample] \cdot Y_i$$

Sampling weights are inverse of their sampling probabilities

$$\mu = \frac{1}{N} Y_i$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{N} 1[i \in sample] \cdot Y_i$$

$$E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^{N} \underbrace{P(i \in sample)}_{\frac{n}{N}} Y_i = \frac{1}{N} \sum_{i} Y_i = \mu$$

More generally

$$\bar{Y} = \underbrace{\frac{1}{n}}_{weight} \sum_i 1[i \in sample] \cdot Y_i$$

$$\bar{Y} = \sum_i 1[i \in sample] \cdot weight_i \cdot Y_i$$

$$E(\bar{Y}) = \frac{1}{N} \sum_i Y_i$$

$$\Rightarrow weight_i = \frac{1}{N * P(i \in sample)}$$

Summary

- Different purposes for different datasets
- Sampling terminology
- Sampling methods

More details can be found in

Sampling: Design and Analysis by Sharon L. Lohr