SURVEY ANALYSIS 101: BASICS & TERMINOLOGY

ISAAC CORMIER, B.SC.
2025/07/25

MODEL-BASED VERSUS DESIGN-BASED ANALYSIS

Model-based statistics assume that:

- The population of interest is **random**, where the 'true' set of individuals and associated values are **generally unknown.**
- There is an underlying **probability model** (e.g., normal distribution, logistic distribution) describing how data is generated from the population.
- The values in your sample are **fixed** and can be used to estimate the unknown population parameters **based on your chosen model.**

Design-based statistics assume that:

- The population of interest is **fixed** and finite, with a **known** set of individuals and associated values.
- There are underlying **sampling probabilities** that determine how likely everyone is to be sampled from the population and how well their data **represents** the total population.
- The values in your sample are **random** and depend on the **sampling design** used to draw your sample from the population.

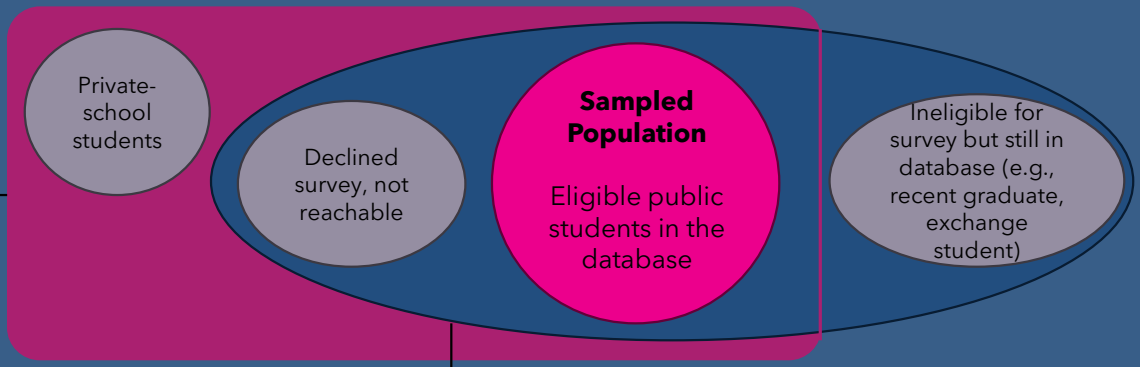
Lohr (2019), Lumley (2010)

THE PROBABILITY SAMPLE

- Unlike model-based samples, **probability samples** are used when the goal of your survey is to make generalizable inferences about a **target population.**
- Your sample is drawn from the population based on your **sampling frame.**
- Every individual or **unit** in the **sampled population** has a **known, non-zero probability** of being selected, and a random process is used to choose the specific units to be included.

Target Population

All Nova Scotia students currently enrolled in high school (10-12)



Sampling Frame

Ministry of Education's database of students enrolled in public high schools

Lohr (2019)

TYPES OF PROBABILITY SAMPLES

- **Simple Random Sample:** Every possible subset of units in the sampling frame has **an equal probability** of being selected.

  E.g., 500 random students are selected from list of **all** eligible public-school students.

- **Stratified Random Sample**: The sampling frame is divided into subgroups called **strata**, and a random sample is taken from each.

  E.g., NS public schools are divided into 20 rural regions and 4 urban high-density regions, and a random sample of 15 students is taken from each rural region, while 50 students are taken from each urban region (n=500).

- **Cluster Sample:** Units are aggregated into larger groups called **clusters**, and a random sample of the clusters called **Primary Sampling Units (PSUs)** are collected.
  - The elements within each cluster are called **Secondary Sampling Units (SSUs).**

  E.g., Schools are defined as clusters, and a random selection of 100 schools are drawn, with a random selection of 5 students from each cluster (n=500).

Lohr (2019), Lumley (2020)

TYPES OF PROBABILITY SAMPLES

- **Systematic Sampling:** Each unit in the population has an equal probability of selection, but rather than being randomly chosen, each unit is systematically selected from a population list based on a **sampling interval.**

E.g., A sampling interval of 2 is calculated based on a population total n=1000, and a starting point of 2 is randomly selected between 1 and the sampling interval. Every 2nd student is then sampled (n=500).

Systematic sampling is more efficient than **simple random sampling** as long as the population list is truly randomized and the order does not introduce bias.

Stratification increases precision and ensures representation of naturally occurring groups.

Clustering makes surveying easier by sampling groups instead of sampling individuals.

Lohr (2019), Lumley (2020)

SIMPLE SAMPLING WEIGHTS

- The probability of a given unit from the population being included in a sample is called the **inclusion probability** (π_i).
- Every unit in a sample has a **sampling weight** (w_i) that defines the number of population units represented by that given unit:

$$w_i = \frac{1}{\pi_i}$$

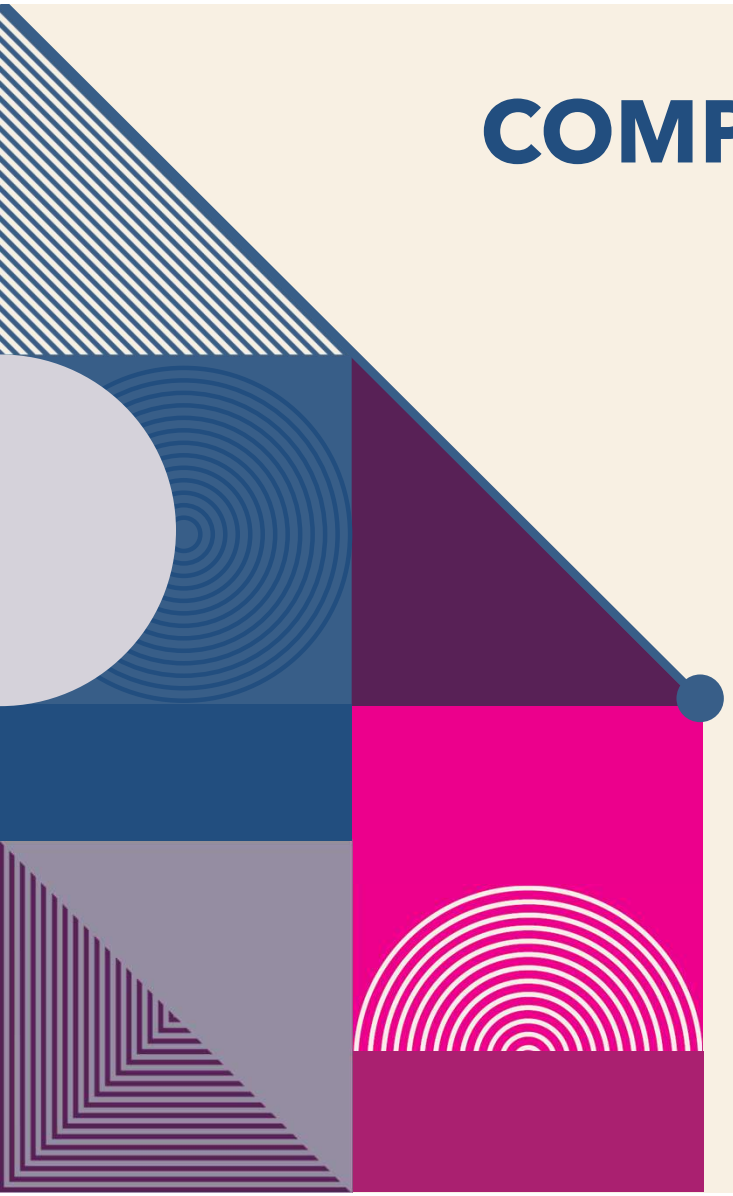
- In a simple random sample, the probability of being included is simply the **sample size divided by the population size.**
- E.g., n = 500 students; N = 1000 students in all of Nova Scotia

$$\pi_i = \frac{n}{N} = \frac{500}{1000} = 0.5 \qquad w_i = \frac{1}{\pi_i} = \frac{1}{0.5} = 2$$

Every 1 student in the sample represents 2 students from the population.

Lohr (2019)

COMPLEX SAMPLING WEIGHTS



- For more complex survey designs, sampling weights must be adjusted to account for:
  - Multi-stage sampling (e.g., stratification, clustering)
  - Likelihood of response (i.e., **nonresponse adjustment**)
  - Sampling frame corrections (e.g., removing duplicates, adjusting for biases in sampling)
  - **Oversampling** or **under-sampling** from certain subgroups to improve precision and ensure adequate data for analysis.
- Complex survey datasets will typically contain a **final weight variable** that has already been adjusted based on the specific survey design.
- You **cannot** perform regular model-based statistics on design-based data because they **do not account for sampling weights** and will **lead to bias estimates.**

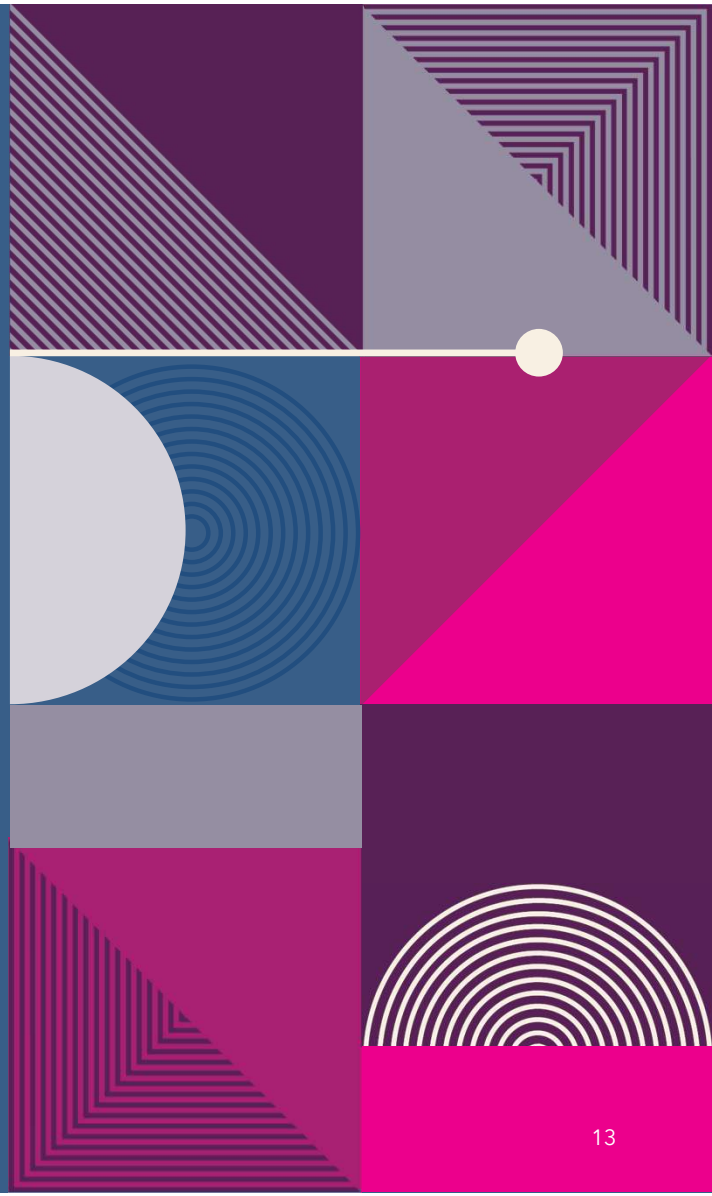
Lohr (2019), Lumley (2020)

THE 2018 COLOMBIA VACS DESIGN

- **Target Population:** All non-institutionalized civilians of Colombia ages 13 to 24 living in urban and rural areas of Colombia, including the District Capital (DC) Bogotá.
- **Sampling Frames:** Based on the 2005 Population and Household Census of Colombia.

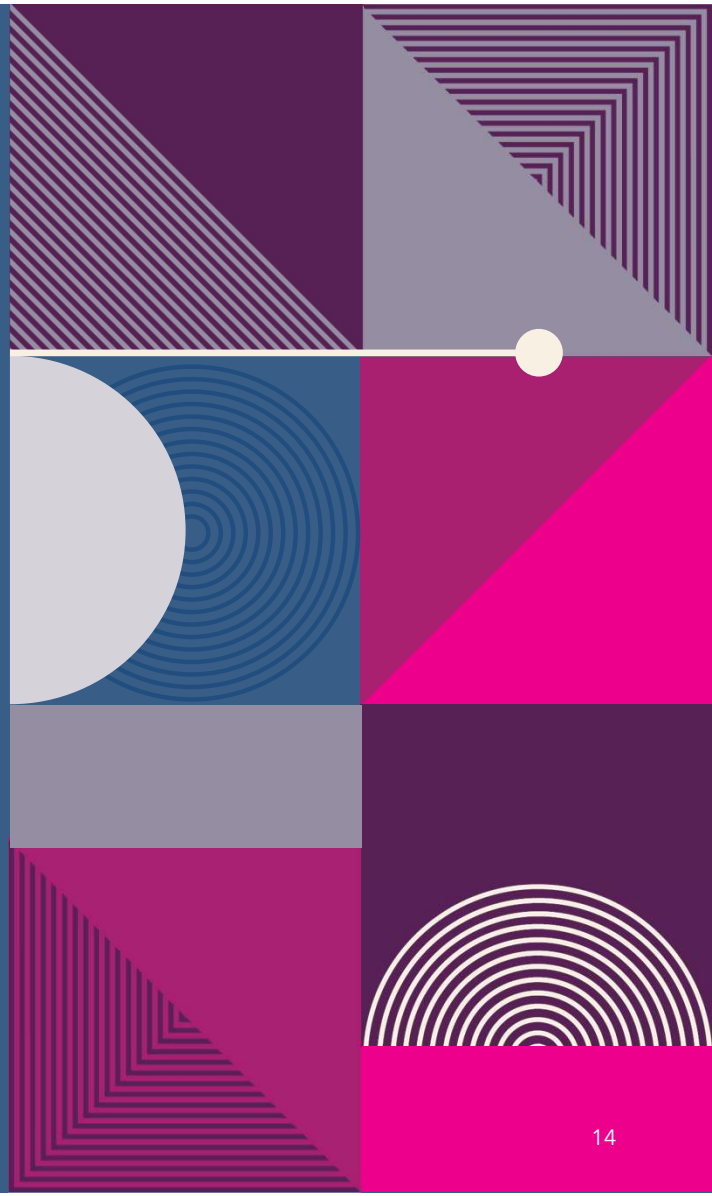
  National: 1,222 municpalities and 32 departments of Colombia and DC Bogotá.

  Priority: 170 municpalities historically exposed to conflict.
- **Sampled Population:** All eligible youth aged 13 to 24 residing in households within selected Enumeration Areas (EAs) for the given sampling frame.
  - Institutionalized individuals, EAs with less than 50 households or inaccessible/unsafe areas were excluded.



THE 2018 COLOMBIA VACS DESIGN

- **Stratification:** The sampling population was stratified in two levels:
  - By sampling frame (National or Priority)
  - By municipality/department
- **Clustering:** Sampling was done by EAs, which group multiple households by geography. Each EA was assigned randomly to be either male-only or female only.
  - **Primary Sampling Unit:** EAs randomly but independently selected from each sampling frame based on the assigned gender for that EA.
  - **Secondary Sampling Unit:** Households, with 24 households selected within each EA using systematic sampling.
- **Observation Unit:** 1 single, eligible respondent, randomly selected within each household based on the EA's assigned gender.



THE 2018 COLOMBIA VACS DESIGN

- **Weighting Procedure:** Base weights were calculated based on the selection probabilities of SSUs, gender, and eligibility, and were adjusted for individual and household non-response as well as post-stratification formed by gender and department.
- **Oversampling:** 20% oversampling was done from specific areas of the country's four main cities (Bogotá, Cali, Medellín, and Barranquilla.)