# 2nd Module : Data Collection

## Lesson 1: Sampling method, Bios, + Data Tables

April 28, 2023

### Data Collection

↳ collecting / gathering info on people / things

→ multi-billionaire industry - lots of $

→ Mark Zuckerberg
→ Steve Jobs
→ Bill Gates

they sell your info to companies who will know how to better manipulate you into buying stuff → marketing

What type of data are they collecting on you?

examples
→ amount of time spent on a website
→ age
→ height
→ # of friends
→ how $ you make

→ voting intentions
→ what internet sites visited
→ gender
→ interests
→ relationship status
→ beliefs
→ habits
→ location
→ who your friends are
→ what profession

**quantitative data** v.s. **qualitative data**

↳ info that can be expressed as a #

↳ info that can NOT be expressed as a #. ↳ uses words.

Who : (collecting Data on)

Recall : Definitions :

Population : the set of all individuals / things we want to understand

300 students

e.v. { Ashmi, Leo, Devonte ... Jada }

Sample : a subset / subgroup / smaller group of the population that is ideally representative of the population.

ex. { Ms. Short's class }

ex. to find out if students of JLAEC like hybrid learning

What (type of info collected)

Recall : Definitions

**Distribution** → a set/group of data values (elements that measure things)

ex. a set of grade = x

{ 87%, 70%, 90% }

**Variable** → a letter/symbol that rep. the elements of a dist. that vary/change.

**Variable Types**

- Quantitative
  - Discrete (gaps)
    - usually $\mathbb{Z}$ (integers) (+ and − whole #)
    - ex. # of facebook friends
      x = { 299, 300, 900, 150, 151 }
      nothing in between
    - s = ex. shoe size, where $s \in \mathbb{R}$
      { 4, 4.5, 5, 5.5, 6, 6.5 ... }
      nothing in between
  - Continuous (no gaps)
    - usually $\mathbb{R}$
    - ex. { amount of coffee beans bought in grams }
      { 3.3, 3.79, 4.59 }
      infinite possibilities between 3 and 3.7
    - ex. length of foot
      { 4, 4.1, 4.2 }
      infinite
- Qualitative
  - y = ex. { liberal, conservative }

Think - Pair - Share

5 mins, you do pg 2 and 3 of handout 7.

Check Answers w/ Partner!

### 1.1.1 Example

From among the following data, identify which are quantitative data and which are qualitative data:

- A survey asks participants what their political preferences are from among the following: Liberal, Conservative, New Democrat, Green, Other. qualitative
- A geological survey is collecting data on temperatures below certain rock formations. quantitative
- A census is being carried out asking for participants' age quant
- A census is being carried out asking for participants' city of residence. qual {Montreal, T.O.

### 1.1.2 Practice

For each of the following, determine whether it would consist in quantitative or qualitative data:

(a) Temperature quant.
(b) Time quant
(c) Ice Cream flavor preference qual. {vanilla
(d) Shoe Size quant
(e) Weight quant
(f) Length
(h) Country of residence qual
(i) Favorite Hobbies
(j) Mass of an object
(k) Names
(l) Color of an object
(m) Volume of transportation crates

### 1.2.1 Example

From among the following set of quantitative, identify which are discrete and which are continuous:

- Number of students in a school discrete quant
- Shoe Size discrete
- Temperature continuous quantitative
- Weight cont

### 1.2.2 Practice

Classify the following data as either *discrete* or *continuous*

(a) Time cont

(b) Volume of Air cont

(c) Grade on a math exam discrete

(d) Number of shoppers in a shopping mall discrete

~~80.4~~ %
80.5 %
80 %, 81 %

How (are we collecting Data)

Recall: 3 statistical tools

i. **Sample survey:** a questionnaire (survey) that you give to a sample of ppl of the population
→ easy → less accurate esp. if sample is not representative

ii. **census:** a questionnaire that you give to the entire population
→ long → more accurate

e.x. collect data on whether students of JLAEC like hybrid learning.

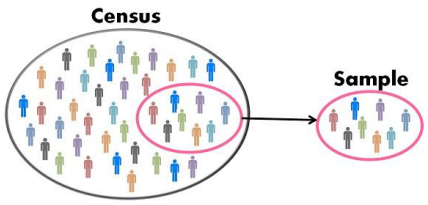
iii. **study:** indepth survey/study/questionnaire performed by experts

e.x collect data on how hybrid learning has affected learning.

**Census v.s. Sample v.s. Study**

**Census:** involves the entire population (for example, you use a census to collect data on the address of residency for every Canadian)

**Sample:** involves only a subset (sample) of the population (for example, you wish to estimate the average height of all Canadians by sampling 1000 Canadians that are representative of the entire population).




**Study:** A sample survey that relies on experts (ex. Scientific trials involving testing a vaccine; nutrition studies on the effects of sugar consumption on the central nervous system, etc.)

Ref: https://healthsci.mcmaster.ca/news-events/news/news-releases/2020

You do 1.3 and 1.3.1 on 4 and 5

### 1.3 Types of Surveys (Census, Sample Survey and Study)

How do we collect data? There are three major types of surveys for data collection. Consider the following research goals and ask yourselves how we should go about collecting the relevant data:

**(a) Determining the address of residence of every Canadian**

census

**(b) Determining the average height of Canadians**

sample survey

**(c) Determining the link between a vaccine and a specific potential side effect**

study

#### 1.3.1 Practice

For each of the following situations, state whether we should conduct a census, a sample survey or a study. *Justify your answer.*

(a) The Municipal government wants to determine exact the number of people working for the city in Montreal.

census

(b) A journalist wants to find out peoples' opinions concerning the religious symbols ban in Quebec.

sample survey

(c) A company manufacturing bolts wants to to ensure their quality.

sample / census

(d) The personnel manager in a multinational corporation wants to know the number of employees who have elected to take the supplemental health policy.

census

(e) A Canadian Health agency wishes to evaluate the number of people affected by work-related stress in Vancouver

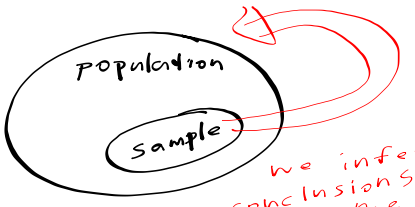
study

# Understanding Bias in Data Collection

(Bad) ⇒ ~~Stereotype~~

Who: population (goal)

How: sample survey




Bias happens when what we infer / conclude / say about a population (sample survey) based on a statistical tool is likely different from reality.

we infer conclusions about the population.

nota bene:

If our sample is not representative, then there's a bias, and what we conclude is (statistically) invalid.

ex. I want to estimate average height of male canadians.

population

P

sample: toddlers aged 2 to 4

∴ men's ave height is 3'6"

ave. 3'6"

✗ wrong cuz sample wasn't representative

sample: males of MTL whose average height is 5'10"

still not representative ∴ biased

∴ 5'10" is invalid

4 Methods to Ensure Sample is Representative of ... the Population
(in notes: refer to handout 2 of lesson 1)
(write on memory aid)

## 2 Data Collection — Sampling Methods: Random, Systematic, Cluster, Stratified Sampling.

A way to avoid bias is to select your sample from the target population in a way that leads to a representative sample. We will look at four ways of doing so.

### 2.1 Random Sampling

pg 4 of textbook

A **random sample** selects participants randomly from a target population. Each member of the the target population is just as likely as any other to be selected.

sample

the population

**Example** A teacher wants to select 4 students to represent the class. The teacher places all the names in a hat and draws four names.

**Q:** Explain why this is a random sample.

- names picked by chance
- every student has the same chance of being selected

v.s. the teacher points at 4 students

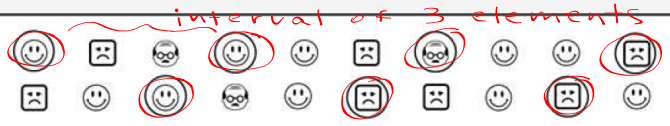
(implicit bias) unconscious bias could affect chance

(2)

## 2.2 Systematic Sampling

pg 4 of textbook

A **systematic sample** involves selecting the sample using a skip or sample interval (e.g. selecting every 4th person). Often the target population is put in some order/list first.




**Example** A teacher wants to select 4 students to represent the class. The teacher takes out the alphabetical class list and selected every 8th student on the list.

**Q:** Explain why this is a stratified sample.

Definitions:

(pl. clusters)

* a cluster of the population is a subgroup / subset that is heterogeneous.

(diverse / different elements)

∴ representative, so pick 1 cluster or more for sample.

(pl. strata)

* a stratum of the pop. is a subgroup / subset that is homogeneous.

(similar / same elements)

{all students who are female}

∴ not representative, so pick a % of all strata for sample

(same % as population %)

3

## 2.3 Cluster Sampling

A **cluster sample** involves breaking the target population into smaller groups where each groups is representative of the entire target population as a whole. Then you select *everyone* from one or more of these clusters (groups) randomly to be part of the sample.

heterogeneous


groups

**Example** At a local adult education centre, five math classes are randomly selected out of 20 and all of the students from each class are interviewed.

**Q:** Explain why this is a ~~stratified~~ cluster sample.

4

## 2.4 Stratified Sampling

→ strata / homogeneous groups

A **stratified sample** involves breaking the target population into smaller groups according to some particular quality. Then you select individuals from each group randomly (respecting the percentages of the population as a whole)


stratum 1 } the population

**Example** At a local adult education centre, they are trying to determine whether people with certain hair colors spend more time getting ready in the morning than others. There are 10 000 students in all and a sample of 100 is selected as follows:

| | Population | Sample |
|---|---|---|
| Brown/Black Hair | 6000 | 60 |
| Blonde Hair | 3800 | 38 |
| Red Hair | 200 | 2 |
| **Total** | 10000 | 100 |

$$\% = \frac{\text{part}}{\text{total}} \times 100\%$$

**Q:** Explain why this is a stratified sample.

↳ check %'s in pop. and sample

$$\%_{bb} = \frac{6000}{10000} \times 100\% = 60\%$$

$$\%_{bb\ pop} = 60\%$$

$$\%_{bb\ sample} = \frac{60}{100} \times 100\% = 60\%$$

$$\%_{bb\ sample} = 60\%$$

must be same %

$$\%_{bh\ pop} = \frac{3800}{10000} \times 100\% = 38\%$$

$$\%_{bh\ sample} = \frac{38}{100} \times 100\% = 38\%$$

$$\%_{RH\ pop} = \frac{200}{10000} \times 10\% = 2\%$$

$$\frac{2}{100} \times 100\% = 2\%$$

$\%_{RH\ sample} = 2\%$

all same % ∴ stratified!

# Determining the # of people to include in stratified sample

| | # in Population | # in sample |
|---|---|---|
| francophone | 2400 | 120 |
| anglophone | 1200 | 60 |
| allophone | 400 | 20 |
| total | 4000 | 200 |

given:
sample size = 200 students

1) Find % of population:

$\% = \frac{\text{Part}}{\text{total}} \times 100\%$

$\%_f = \frac{2400}{4000} \times 100\%$

$\%_f = 60\%$

$\%_a = \frac{1200}{4000} \times 100\%$

$= 30\%$

$\%_{all} = 10\%$

2) Find # of elements in sample

60% of 200

$60\% \times 200$

$\frac{60}{100} \times 200$

$0.6 \times 200 = 120$

30% of 200

$0.3 \times 200 = 60$

10% of 200

$\frac{10}{100} \times 200 =$

$0.1 \times 200 = 20$

### 2.5 Practice

You do pg 11 of handout 3

Identify the type of sampling method used in each of the following situations:

(a) An administrator at an adult education centre randomly selects seven classes and surveys each student in those classes.

groups

— cluster

(b) An avid rare book collector needs to select which books to display. She enters her temperature controlled storage where the books are stored on shelves. She picks a random book to start and selects every 5th book.

skip interval — systematic

(c) Publique Sant'e randomly selects your school out of hundreds of schools for testing and all the students are tested.

groups

— cluster

(d) Every fourth person entering the public theater is searched in order to discourage people from sneaking in food.

– systematic

(e) A pizzeria serves vegetarian and meat lovers pizza. 30% of the customers order vegetarian and the rest order meat lovers. We sample 3 customers who purchase vegetarian pizza and 7 customers who purchase meat lovers pizza to get their opinions on the pizza shop.

– stratified

(f) A researcher for an airline interviews all of the passengers on five randomly selected flights.

– clusted

hetergroups

Think – Pair – Share

% are same

$\frac{3}{10} = 30\%$ $\frac{7}{7+3} = 70\%$

### 2.6 Example: Determining Missing Values in a A Stratified Sample

Determine the missing values in following stratified sample of 50 ice cream lovers:

| | Population | Sample |
|---|---|---|
| Chocolate Ice Cream Lovers | 220 | ? |
| Vanilla Ice Cream Lovers | 80 | ? |
| Caramel Ice Cream Lovers | 140 | 14 |
| Peppermint Ice Cream Lovers | 60 | ? |

Using / Creating Different (4) Data Tables

(easier to present survey results)

Condensed Data Table for Discrete Quantitative and Qualitative Variables

(refer to handout 3)

### 3.1 Condensed Data Tables

The following data show the survey results of the number of children in families. (brut)

a distribution - raw data

$x = \{3,5,2,4,3,1,4,1,2,5,3,6,4,3,5,4,4\}$

Let's create a condensed table (frequency table) to visualize this data

x variable

| No of Children | Tally | Frequency |
|---|---|---|
| 1 | II | 2 |
| 2 | II | 2 |
| 3 | III | 4 |
| 4 | IIII | 5 |
| 5 | III | 3 |
| 6 | I | 1 |
| Total | 17 | 17 |

Condensed Data Tables are often called **Tally Tables** (if they include the tally column only) or **Frequency Tables** (1)

Definition: Frequency - the # of times a data value/element appears in a distribution.

### 3.1.1 Condensed Data Tables: Relative Frequency

Another useful visualization would be to represent the *percentage* of each frequency row. This would give us a quick indication of just how frequent a particular value is! Let's visualize the same data from the previous example, but this time we will also add a *relative frequency column*:

The following data show the survey results of the number of children in families.

1, 1, 2, 2, 3, 3, 3, 3
4, 4, 4, 4, 4, 5, 5, 5, 6

Let's create a condensed table (frequency table) to visualize this data

| No of Children | Tally | Frequency | Rel. Frequency |
|---|---|---|---|
| 1 | | 2 | 11.8% |
| 2 | | 2 | 11.8% |
| 3 | | 4 | 23.5% |
| 4 | | 5 | 29.4% |
| 5 | | 3 | 17.6% |
| 6 | | 1 | 5.9% |
| Total | | 17 | 100% |

Relative Frequency → Put titles on memory aid.

$R.F. = \frac{\text{Freq}}{\text{total freq}} \times 100\%$

$R.F._1 = \frac{2}{17} \times 100\%$

$R.F._1 = 11.8\%$

You do pg 13 of handout 3

Condensed Data Tables that include *relative frequency columns* are usually called **relative frequency tables**

Definition: Relative Frequency: how often/frequent (%) a data value appears in relation to the whole distribution.

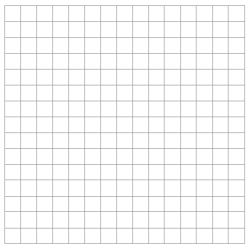
$$\% = \frac{\text{part}}{\text{whole}} \times 100\%$$

### 3.1.2 Practice

The following data show the survey results of the number of pets owned per household.

3, 0, 2, 4, 3, 1, 0, 1, 2, 1, 3, 0, 4, 3, 2, 1, 4, 0, 1, 1, 2

Create a full condensed table (including relative frequency) to visualize this data

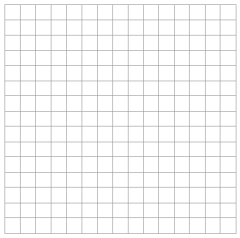


### 3.1.3 Practice

The following data show the survey results of the students' favorite color.

black, blue, orange, orange, red, purple, blue, green, blue, orange, red, red, brown, white, blue, blue, green, blue, green

Create a full condensed table (including relative frequency) to visualize this data



### 3.1.2 Practice

The following data show the survey results of the number of pets owned per household.

3, 0, 2, 4, 3, 1, 0, 1, 2, 1, 3, 0, 4, 3, 2, 1, 4, 0, 1, 1, 2

Create a full condensed table (including relative frequency) to visualize this data

Relative Frequencies:

① $\frac{4}{21} \times 100 = 19.0\%$ or simply $19\%$

② $\frac{6}{21} \times 100 = 28.6\%$

③ same as ①

④ same as ①

⑤ $\frac{3}{21} \times 100 = 14.3\%$

| No. of Pets | Tally | Frequency | Rel. Frequency | |
|---|---|---|---|---|
| 0 | IIII | 4 | 19.0% | ① |
| 1 | MLI | 6 | 28.6% | ② |
| 2 | IIII | 4 | 19.0% | ③ |
| 3 | IIII | 4 | 19.0% | ④ |
| 4 | III | 3 | 14.3% | ⑤ |
| Total | 21 | 21 | 100% | |

Homework:

Read and do pages 8 - 13

Do pages 14 -15

Read and do page 3 - 4

Read page 24 and do page 25

### 3.1.3 Practice

The following data show the survey results of the students' favorite color.

black, blue, orange, orange, red, purple, blue, green, blue, orange, red, red, brown, white, blue, blue, green, blue, green

Create a full condensed table (including relative frequency) to visualize this data

① $\frac{1}{19} \times 100 = 5.3\%$

② $\frac{6}{19} \times 100 = 31.6\%$

③ $\frac{3}{19} \times 100 = 15.8\%$

④ same as ③

⑤ same as ①

⑥ same as ②

⑦ same as ①

⑧ same as ①

| Favorite Color | Tally | Frequency | Relative Frequency | |
|---|---|---|---|---|
| Black | I | 1 | 5.3% | ① |
| Blue | MLI | 6 | 31.6% | ② |
| Orange | III | 3 | 15.8% | ③ |
| Red | III | 3 | 15.8% | ④ |
| Purple | I | 1 | 5.3% | ⑤ |
| Green | III | 3 | 15.8% | ⑥ |
| Brown | I | 1 | 5.3% | ⑦ |
| White | I | 1 | 5.3% | ⑧ |
| TOTAL | 19 | 19 | 100% | |