Introduction to Statistics

Types of data, levels of measurement and sampling methods

Instructor: Qasim Ali

Let’s Introduce ourselves

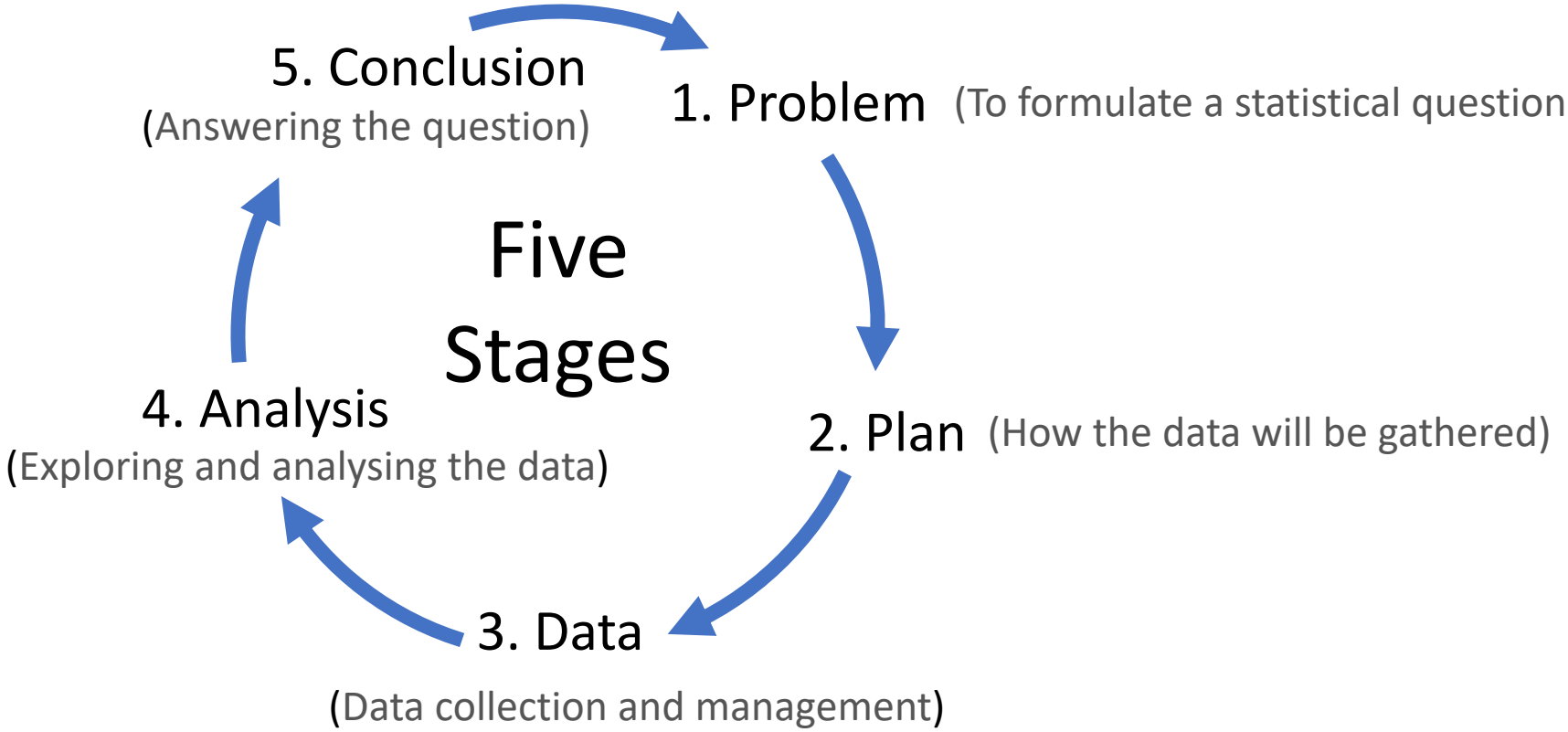
UNIVERSITY OF
GOTHENBURG

CONESTOGA
Connect Life and Learning

HUMBER

Statistics

- **Statistics** is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

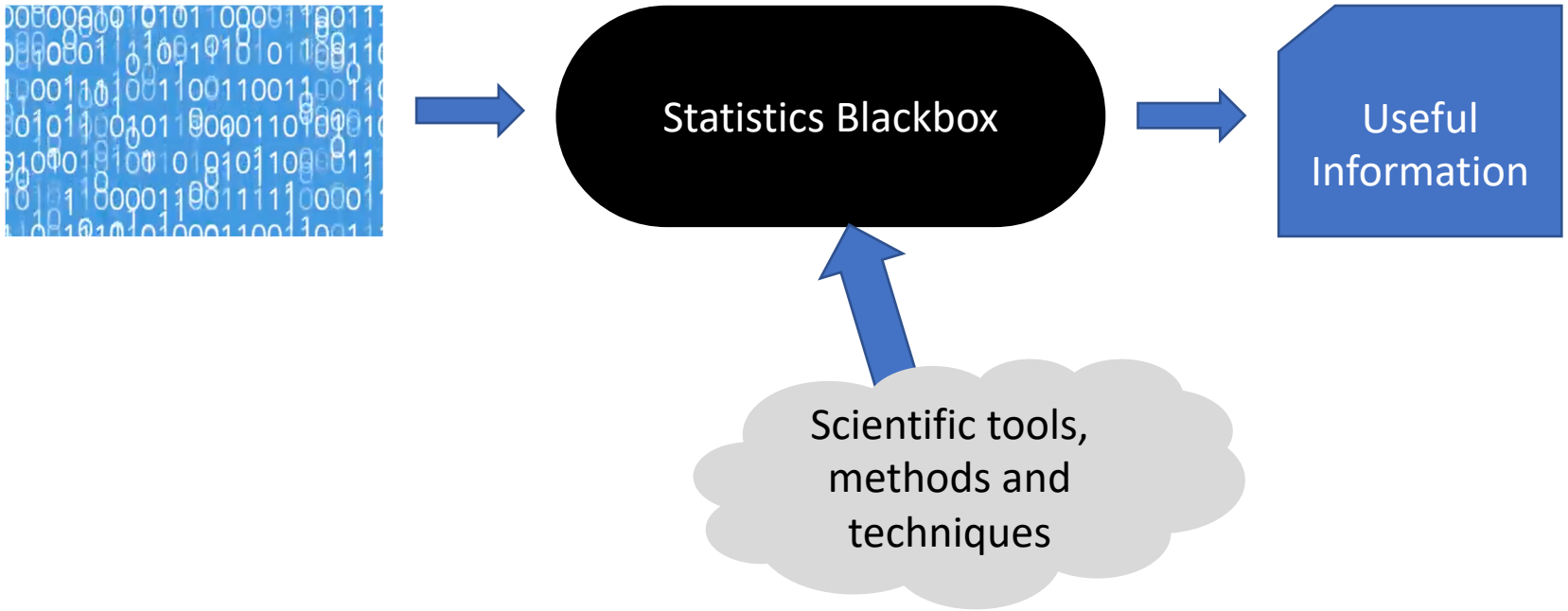


Statistics

Why statistics matter in your life?

- **Statistics** help us quantify uncertainties by telling which numbers are important and which are useless.

- **Statistics** can help us discern if the results are providing us a true illustration of a situation or if the results are presenting us with a biased view.



Population vs Sample

- **Population:** A complete collection of all elements (scores, people, measurements, and so on) to be studied.

  Numerical measurement describing some characteristic of a ***population*** is called as **parameter**

Example: For the population of 21,500 Humber College female students the average height is 64.2 in. The average height is a parameter of this population.

- **Sample:** A sub-collection of elements drawn from complete collection of all elements (population).

  Numerical measurement describing some characteristic of a ***sampling*** is called as statistic

Example: For a group of Humber College female students looking for a work in a fashion agency, the average height is 69.3 in.

Population vs Sample

Practice Problems

In the following examples, determine whether the given value is a **statistic** *or a* **parameter***.*

- For a random sample of 200 Humber College students the average age is 22.7 years.

  Answer: Statistic

- According to the latest census, 2.83% of families in BC have an annual income less than $10,000.

  Answer: Parameter

- 93% of 2100 people questioned in a survey think that legal services in Ontario (and, generally speaking, anywhere in Canada) should be made more affordable.

  Answer: Statistic

Data

- **Data** are observations (such as measurements, genders, and survey responses) that have been collected.

Qualitative Data

- A categorial data that describe attributes of an observation
- Data is only discrete, e.g. name, colors, country, city etc.

Examples

- Females have brown, black, blonde, and red hair
- The cake is orange, blue, and black in color

Quantitative Data

- A numeric data that describe counts or measurements
- Data can be discrete or continuous, e.g. length, weight, price etc.

Examples

- There are four cakes and three muffins kept in the basket
- One glass of fizzy drink has 97.5 calories

Discrete Data

- Shoe sizes:

  6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 11, 12, 13

- Total quantity of items sold

  1 item, or 2 items or 5 items or 10 items

1 QTY Gift Card	$25.00
1 QTY Sport Shirt	$25.00
SUBTOTAL	$140.00
GST 5%	$4.50
PST 7%	$6.30
TOTAL	$150.80
DOWN PAYMENT	$150.80
CHANGE DUE	$0.00
CASH TEND	$150.80
STATUS	Success
Type	Sale
Transaction#	23
Transaction Date	28/07/2017 12:14:18
Payment#	PYMT-27
Payment Date	28/07/2017 12:14:17
Clerk	PDEMO

- Your Birthday

  Jan. 19, April 24, July 7

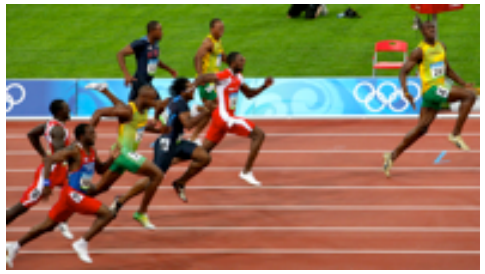
Continuous Data

- Height of Jurassic world animals

  between 5.5 feets to 30 feets

- Time to complete 200m race

  Any value between 40 secs to 1 minutes



- Price of products

  it can $4.86 or $132,579.05

My Restaurant

From: Friday, Sep 7, 2018, 2:39 PM
To: Wednesday, May 29, 2019, 8:28 AM

Menu Item List Report

Appetizers

Caesar Salad	
Cost: $0.00	Price: $5.50
Freedom fries	
Cost: $2.00	Price: $8.00
Fries	
Cost: $0.75	Price: $2.00
Ham Delights	
Cost: $1.50	Price: $5.75
House Salad	
Cost: $1.00	Price: $4.00
Loaded Nachos	
Cost: $2.00	Price: $5.00

Data

Examples

In the following examples, please identify the type of data.

- The diagnosis of a patient admitted to a hospital — **Answer:** Qualitative
- The number of students in statistics class — **Answer:** Quantitative discrete
- Eye color of a fashion show model — **Answer:** Qualitative
- The outside temperature in Etobicoke — **Answer:** Quantitative continuous
- The distance between two subway stations (in meters) — **Answer:** Quantitative continuous
- The number of cases of Covid-19 in a neighborhood — **Answer:** Quantitative discrete

Levels of Measurement of Data

Level	Summary	Example
Nominal	Categories only. Data cannot be arranged in an ordering scheme.	Student cars: Corvettes, Ferraris, Porsches } Categories or names only.
Ordinal	Categories are ordered, but differences cannot be determined or they are meaningless.	Student cars: compact, mid-size, full-size } An order is determined by "compact, mid-size, full-size".
Interval	Differences between values can be found, but there is no inherent starting point. Ratios are meaningless.	Campus temperatures: $15^\circ C$, $20^\circ C$, $30^\circ C$ } 90°C is not twice as hot as 45°C.
Ratio	Like interval, but with an inherent starting point. Ratios are meaningful.	Weights of university football players: 150 lb, 195 lb, 300 lb } 300 lb is twice 150 lb.

Nominal data does not provide any quantitative value. It is sometimes referred to as labelled or named data

Data at the ordinal level of measurement are quantitative or qualitative.

Data at the interval level of measurement are quantitative.

Data at the interval level of measurement are quantitative in nature.

Levels of Measurement of Data

Examples

In the following problems, determine which of the four levels of measurement (nominal, ordinal, interval or ratio) is most appropriate

- Ratings of superior, above average, average, below average, or poor for a TV fashion show

  Answer: Ordinal

- Blood alcohol content

  Answer: Ratio

- Fashion styles presented in a fashion catalogue

  Answer: Nominal

- Real estate prices in Toronto in April 2020

  Answer: Ratio

- Temperature (in degrees Celsius) of Covid-19 patients

  Answer: Interval

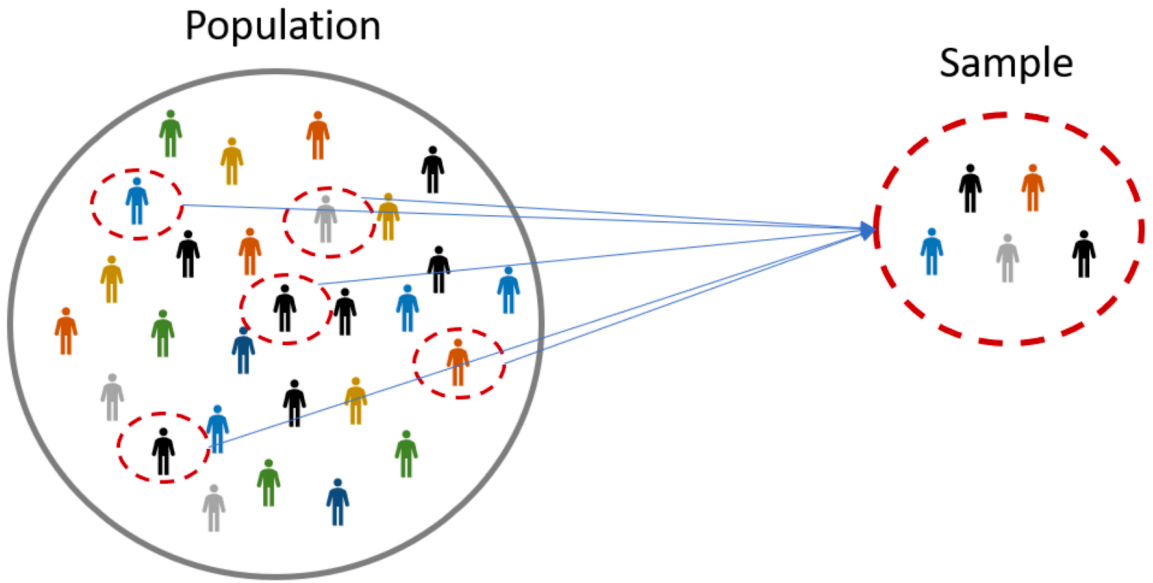
Sampling Methods

- Sampling is the process of selecting observations (a sample) to provide an adequate description and inferences of the population.

- Sample
  - It is a unit that is selected from populati
  - Represents the whole population
  - Purpose to draw the inference

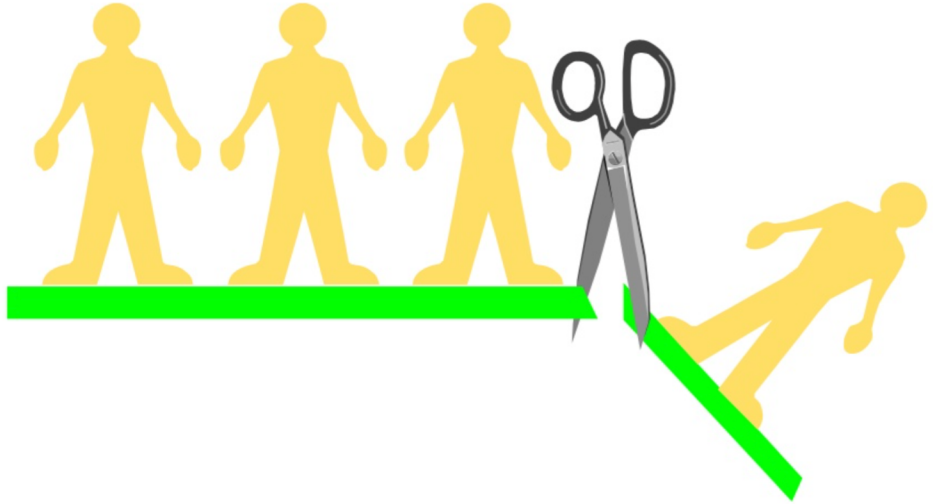
- Why Sample???

- Sampling Frame Listing of population from which a sample is chosen

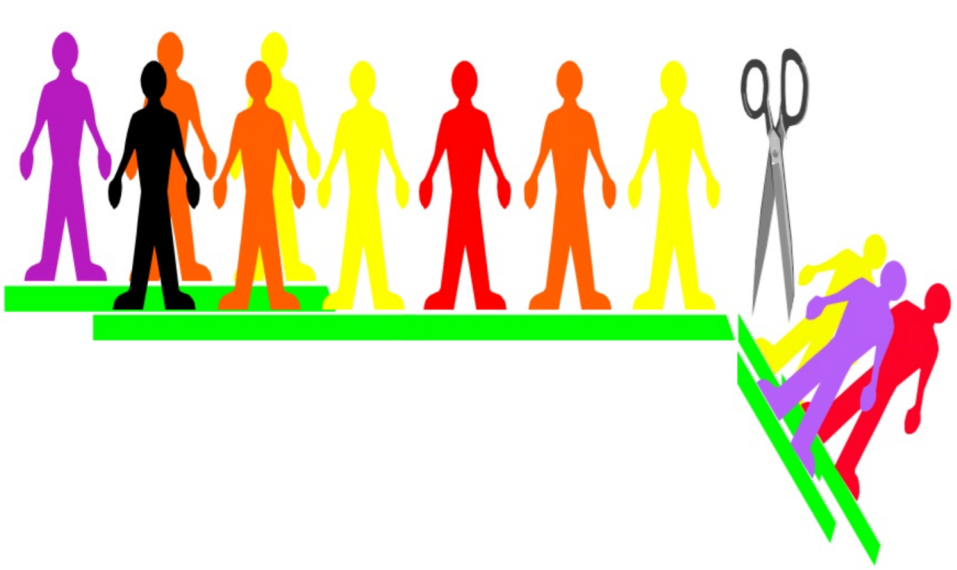


Sampling Methods

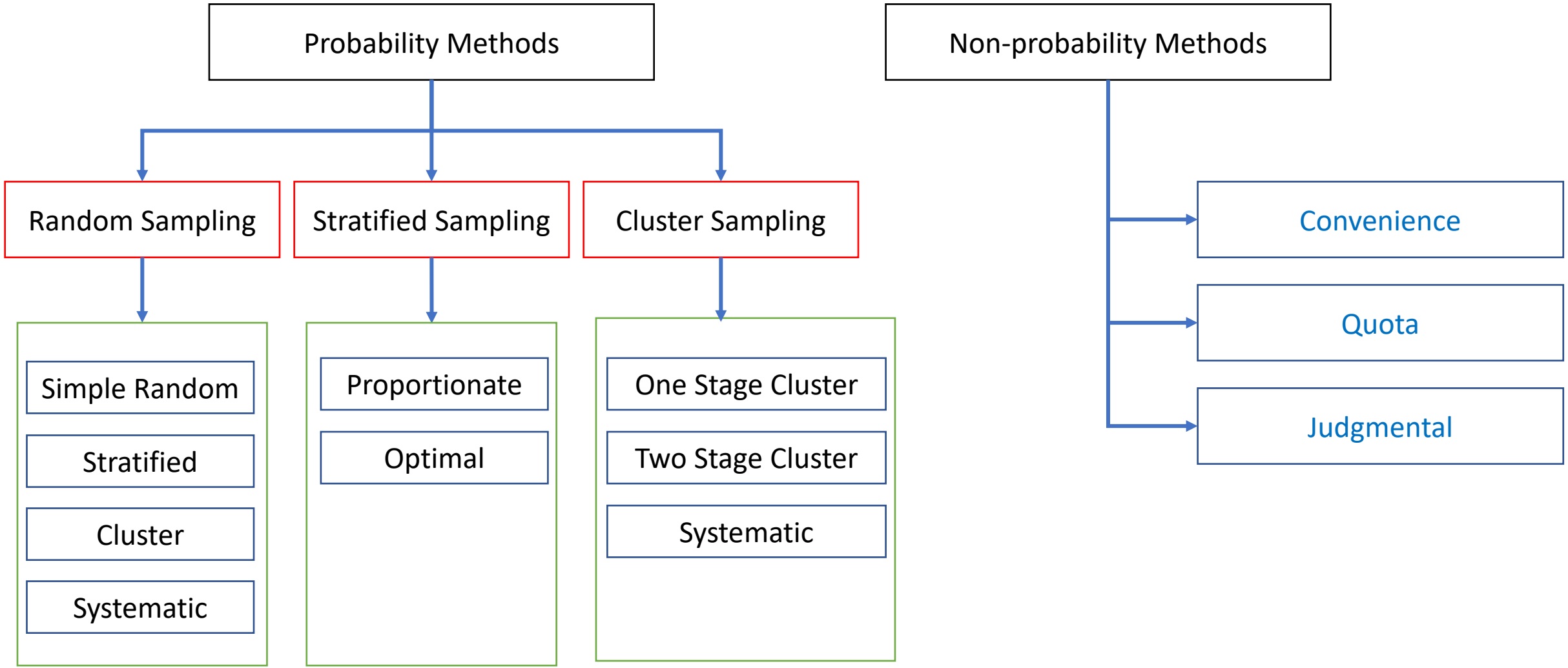
- Homogenous Sample



- Heterogenous Sample



Sampling Methods

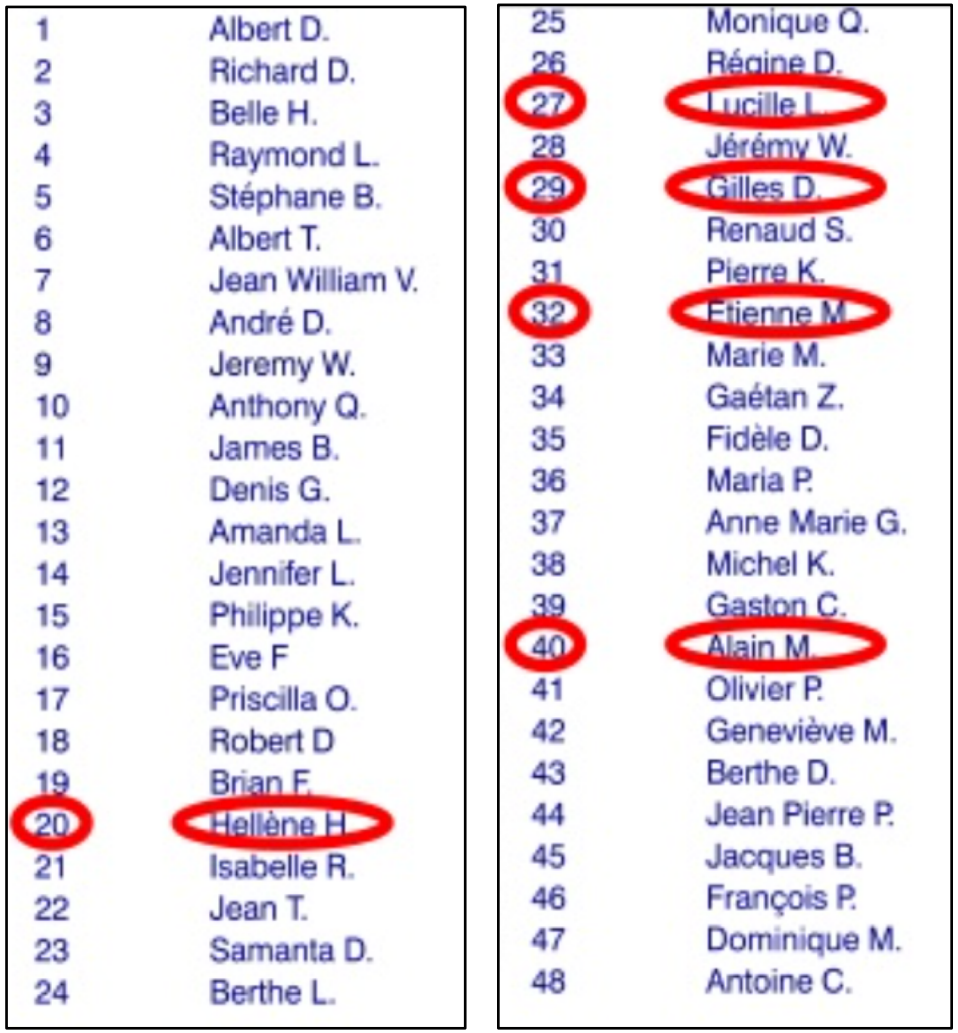


Random Sampling

In a **random sample**, members of the population are selected in such a way that each individual has an *equal chance* of being selected.

Simple random sample

- All subsets of the frame are given an equal probability
- Random number generator



1	Albert D.
2	Richard D.
3	Belle H.
4	Raymond L.
5	Stéphane B.
6	Albert T.
7	Jean William V.
8	André D.
9	Jeremy W.
10	Anthony Q.
11	James B.
12	Denis G.
13	Amanda L.
14	Jennifer L.
15	Philippe K.
16	Eve F
17	Priscilla O.
18	Robert D
19	Brian F.
20	Hellène H.
21	Isabelle R.
22	Jean T.
23	Samanta D.
24	Berthe L.
25	Monique Q.
26	Régine D.
27	Lucille L.
28	Jérémy W.
29	Gilles D.
30	Renaud S.
31	Pierre K.
32	Etienne M.
33	Marie M.
34	Gaétan Z.
35	Fidèle D.
36	Maria P.
37	Anne Marie G.
38	Michel K.
39	Gaston C.
40	Alain M.
41	Olivier P.
42	Geneviève M.
43	Berthe D.
44	Jean Pierre P.
45	Jacques B.
46	François P.
47	Dominique M.
48	Antoine C.

Random Sampling

Advantages

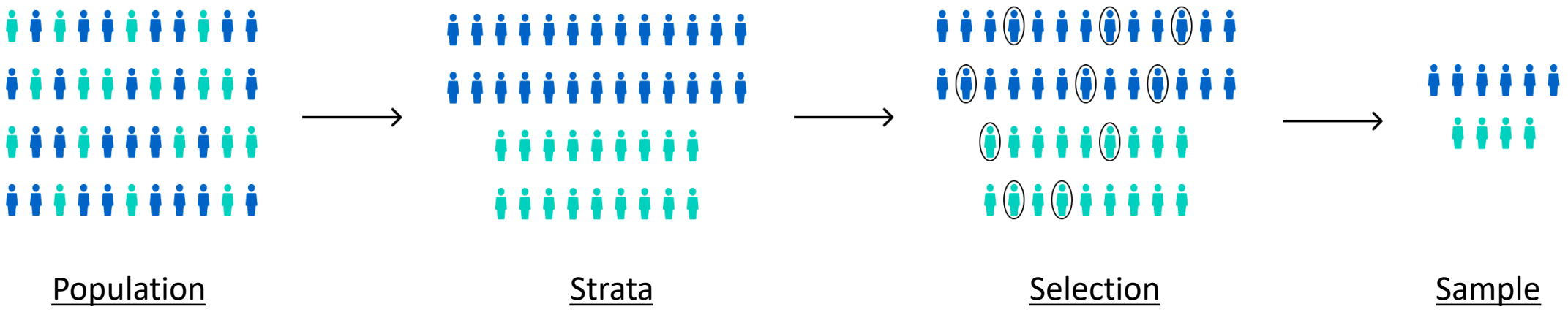
- Minimal knowledge of population used or needed
- Easy to analyze data
- Large subsets are required to reduce errors

Disadvantages

- Low frequency of use
- Can be performed without expertise
- Random error is high

Stratified Sampling

- Population is divided into two or more subgroups called as strata that share same characteristics (e.g. gender, ethnicity, education)
- Subsamples are selected from each strata



Stratified Sampling

Advantages

- Assures representation of all groups in the sample population
- Characteristics of each stratum can be estimated and compared with others

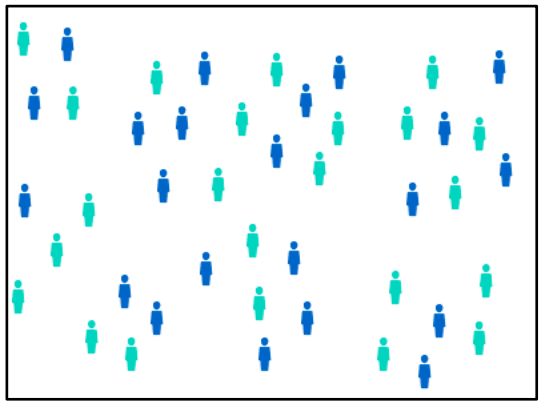
Disadvantages

- Accurate information is required on proportions of each stratum
- Needs deterministic work and therefore costly to prepare the stratified lists

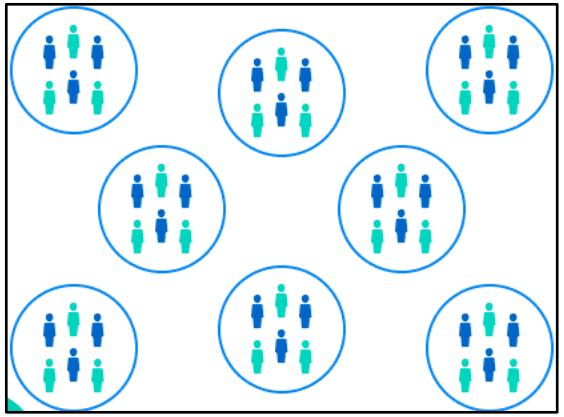
Cluster Sampling

In **cluster sampling**, we first divide the population area into sections (or clusters), then randomly select a few of those sections, and then choose *all* the members from those selected sections.

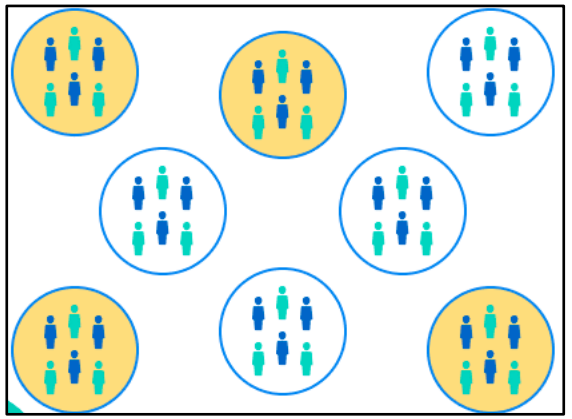
- The population is divided into subgroups (clusters)
- Randomly select few of those subgroups
- Choose all the members from those subgroups



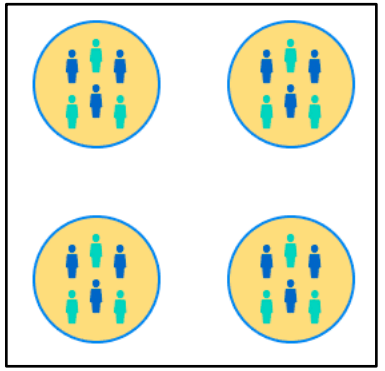
Population



Clusters



Selection



Sample

Cluster Sampling

Advantages

- Can estimate characteristics of both cluster and population

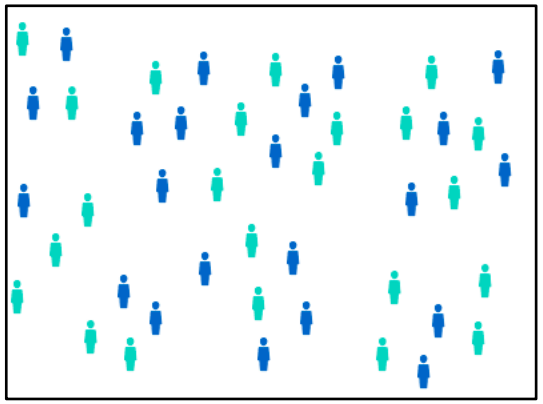
Disadvantages

- Method is costly to reach an element of a sample
- Each stage of cluster generates error. If there are more stages of cluster, there will be more error.

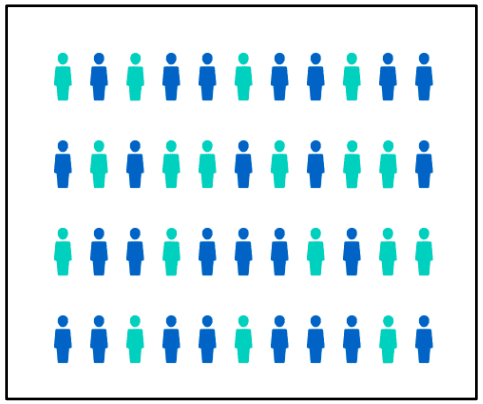
Systematic Sampling

In **systematic sampling**, we select some starting point and then select every *k*th (such as every 50th) element in the population.

- Order all population in a sequence
- Define a sampling rules (e.g. interval)
- Select kth number in the sequence

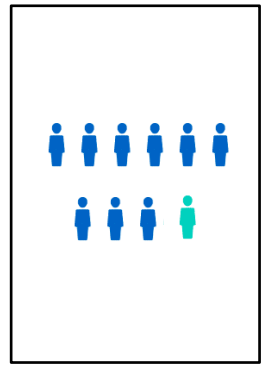


Population



Sequenced

Selection



Sample

Systematic Sampling

Advantages

- Moderate cost, moderate usage
- Simple to draw sample
- Easy to verify

Disadvantages

- Periodic ordering required
- Error is high

Convenience Sampling

Choosing respondents that are readily available

Advantages

- Very low cost
- Extensively used and understood

Disadvantages

- Variability and biased cannot be measured or controlled
- Projecting data beyond sample not justified
- Restriction of generalization

Quota Sampling

Segment the population into mutually exclusive subgroups (like strata)

Advantages

- Good for limited research budget
- Extensively used and understood

Disadvantages

- Variability and biased cannot be measured or controlled
- Projecting data beyond sample is not justifiable
- More time consuming than convenience sampling

Sampling Error

Errors arise due to the sampling methods and sampling surveys are known as Sampling Error.

This means that the sample does not represent the total population of given data.

Types of Sampling Errors

Biased errors: Due to selection of sampling techniques, e.g. size of the sample

Unbiased errors/ Random sampling errors: Selection of the members of the population included or not

Sampling Error

How to reduce sampling error?

- Specific problem selection
- Systematic documentation of related research
- Effective enumeration
- Effective pre testing
- Controlling methodological bias
- Selection of appropriate sampling techniques

Population vs Sample

Practice Problems

- *In the following examples, identify which of these types of sampling is used: random, stratified, systematic, cluster, or convenience.*

- When she wrote *Marriage and Divorce: Legal and Psychological Issues*, the author based her conclusions on 4500 responses from 100,000 questionnaires distributed to her Facebook followers.

  Answer: Convenience sampling

- A psychologist at the University of Guelph-Humber surveys all students from each of 20 randomly selected classes.

  Answer: Cluster sampling

- Toronto Tech, an electronic component manufacturer, usually selects every 100th electronic component from assembly line and conducts a thorough test of quality.

  Answer: Systematic sampling

Design of Experiments

In an **observational study**, we observe and measure specific characteristics, but we don't attempt to manipulate or modify the subjects being studied.

In an **experiment**, we apply some *treatment* and then proceed to observe its effects on the subjects.

Example:

- A survey of citizens to determine what percentage of the population supports the tax policy of the federal government.

- A medical treatment given to a group of patients in order to determine the effectiveness of the method.

Summary

- Statistics
- Population vs Samples
- Types of Data
- Level of Measurements
- Sampling Methods
- Sampling and Non-sampling Errors
- Design of Experiments