Lingnan University 嶺南大學 香港 Hong Kong

SDS School of Data Science 數據科學學院

CDS 533
Statistics for Data Science

Instructor: Lisha Yu

Division of Artificial Intelligence

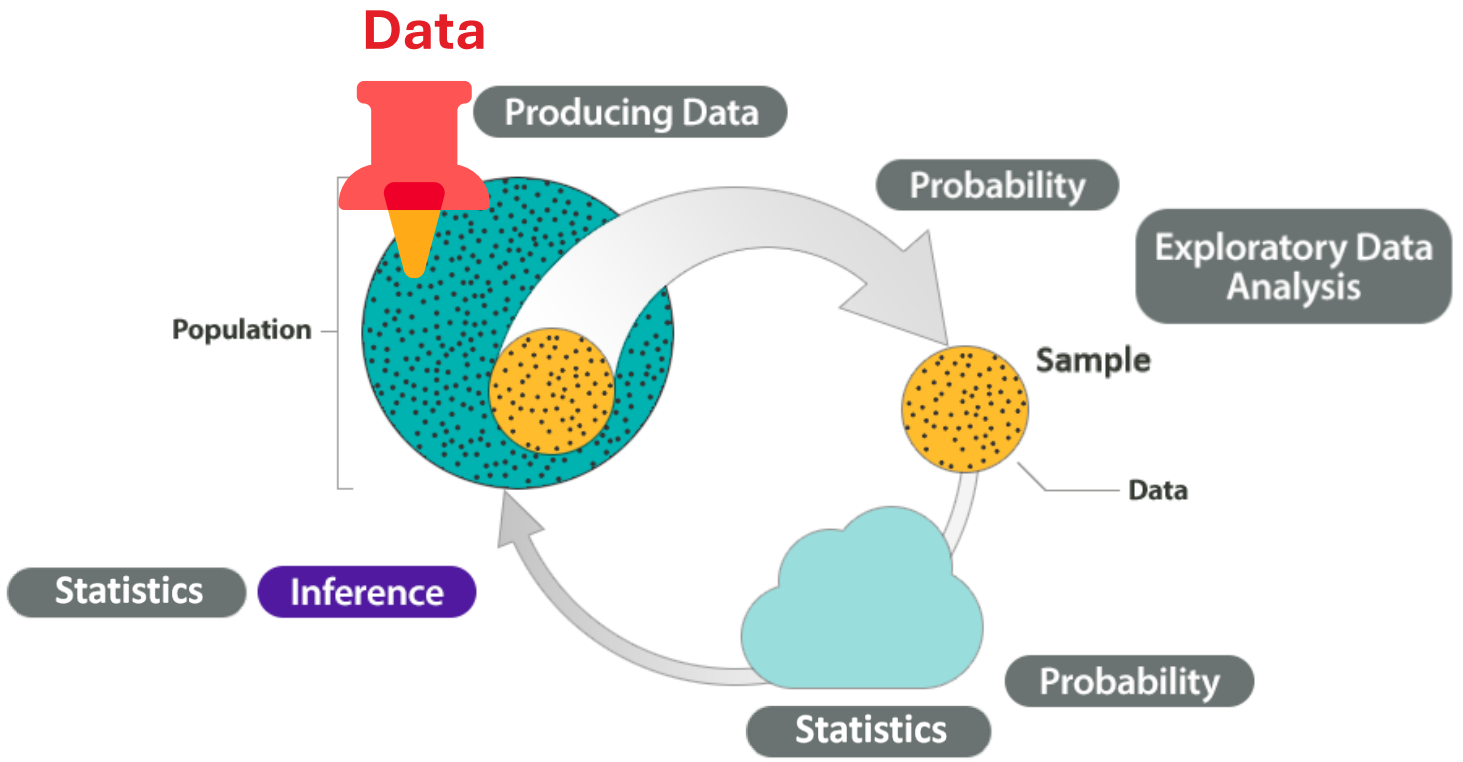
School of Data Science

Lingnan University

Fall 2024

Data: The Heart of Statistics

Big Picture of Statistics



What is Data?

Look around you, there is data everywhere.

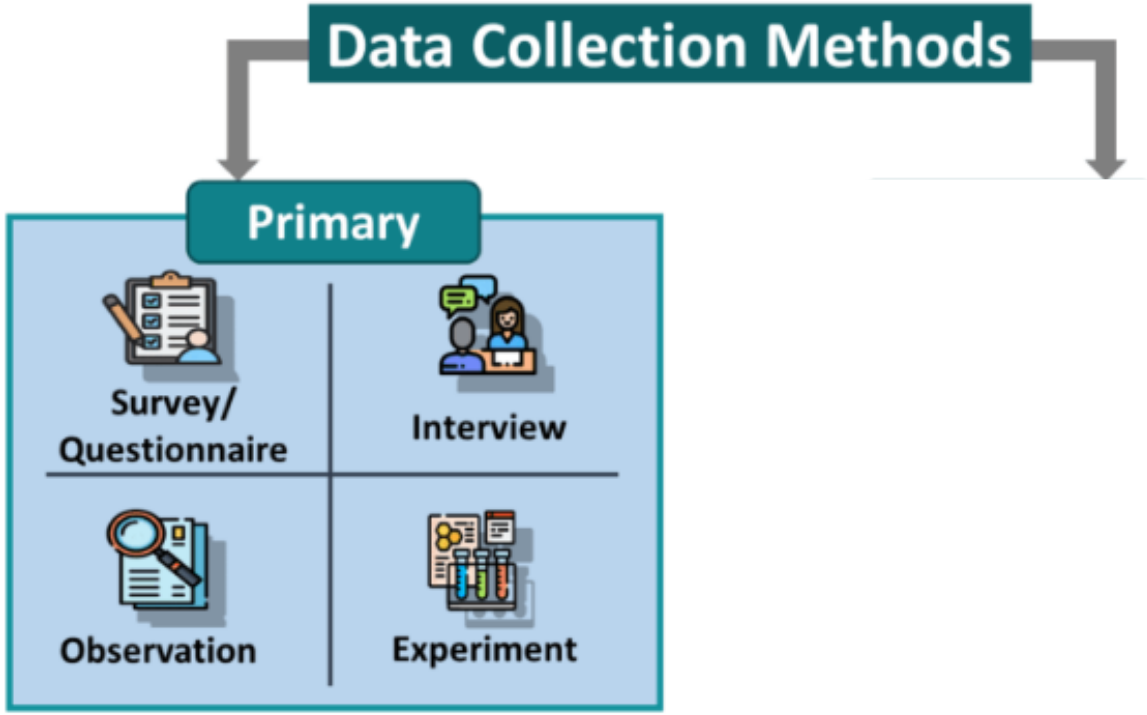
Data refers to facts and statistics collected together for reference or analysis.

(Not limit to)

Sources of Data

Depending on the objective of the data collection, the source of data is said to be

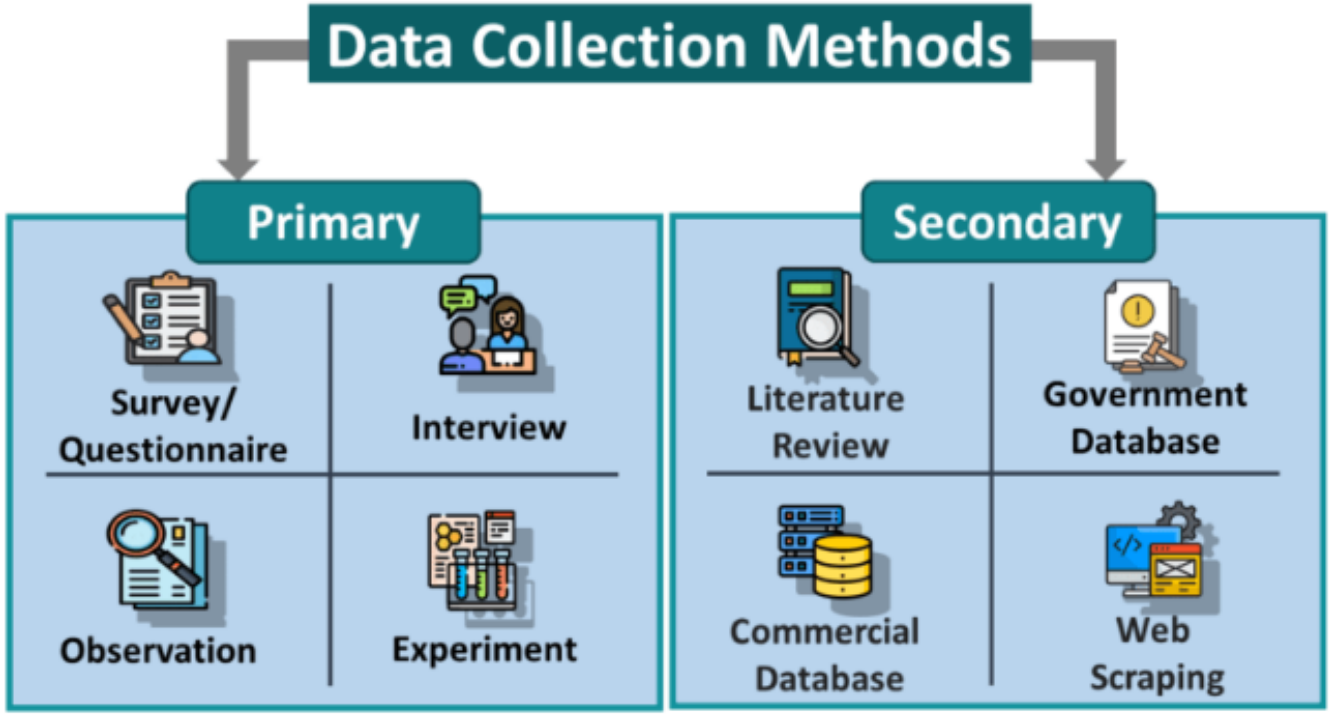
- **Primary data**
  - Information collected directly from source
  - Data collected for very 1st time
  - Data is original and specific to your research



Sources of Data

Depending on the objective of the data collection, the source of data is said to be

- **Secondary data**
  - Data collected by someone else (Often the primary data)
  - Data collected in the past
  - Data is NOT original or 1st hand data



Sources of Data

Questions:

A marketing agency utilizes sales reports and customer feedback from a retail store to analyze market trends for a new product launch.

A research team conducts surveys and interviews with patients to gather information about the effectiveness of a new drug.

Sources of Data

Questions:

A marketing agency utilizes sales reports and customer feedback from a retail store to analyze market trends for a new product launch.

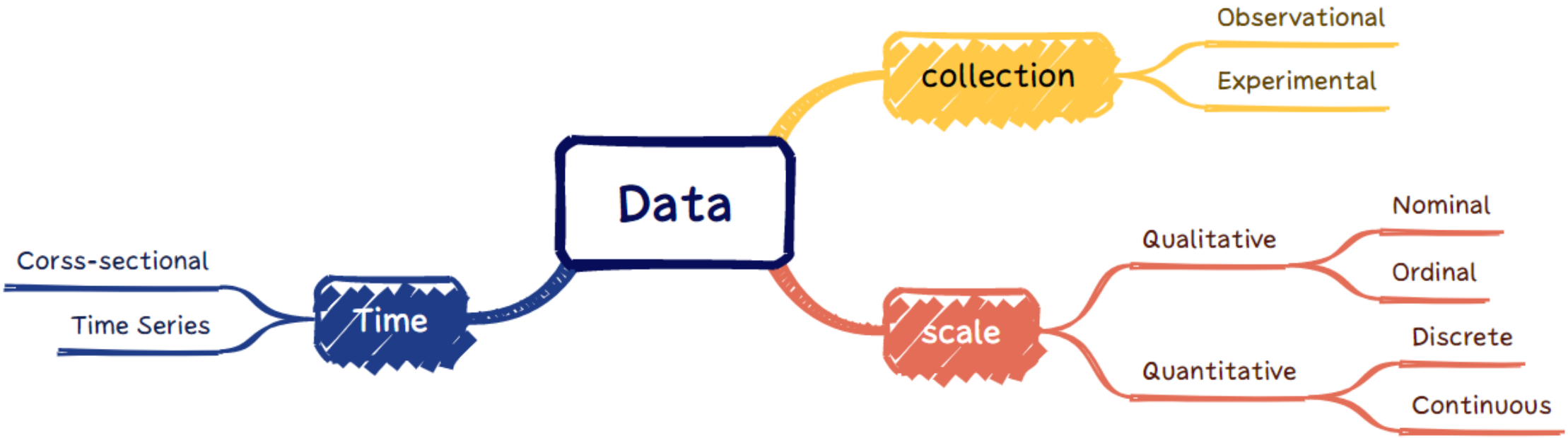
ANSWER: Secondary data

A research team conducts surveys and interviews with patients to gather information about the effectiveness of a new drug.

ANSWER: Primary data

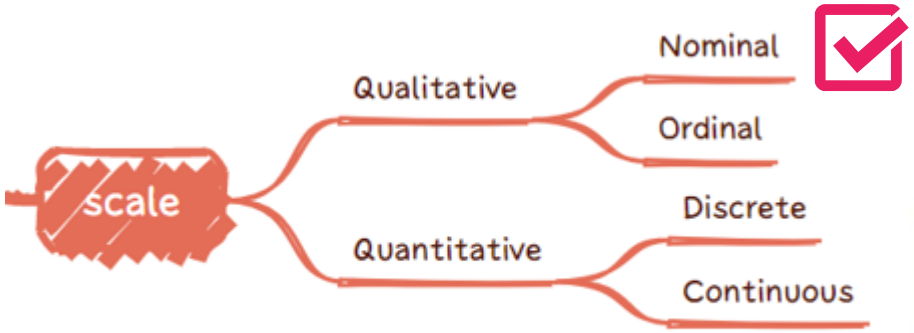
Categories of Data

Statistical way...



Categories of Data

Qualitative data represents groupings



Nominal Data

This type of data is qualitative in nature which has **no inherent mathematical significance**. It is sort of a fixed value under which a unit of observation is assigned or "**categorized**".

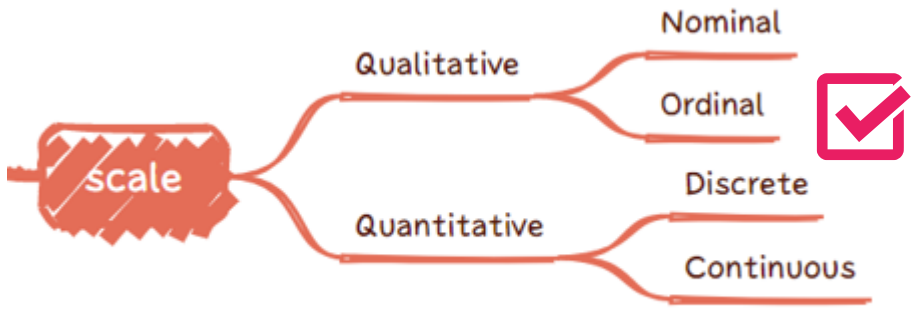
What is your gender?
- M – Male
- F – Female

What is your hair color?
- 1 – Brown
- 2 – Black
- 3 – Blonde
- 4 – Gray
- 5 – Other

Where do you live?
- A – North of the equator
- B – South of the equator
- C – Neither: In the international space station

Categories of Data

Qualitative data represents groupings



Ordinal Data

This type of data is the **combination of numerical and categorical data**, i.e. categorical data having some mathematical significance.

How do you feel today?

- 1 – Very Unhappy
- 2 – Unhappy
- 3 – OK
- 4 – Happy
- 5 – Very Happy

How satisfied are you with our service?

- 1 – Very Unsatisfied
- 2 – Somewhat Unsatisfied
- 3 – Neutral
- 4 – Somewhat Satisfied
- 5 – Very Satisfied

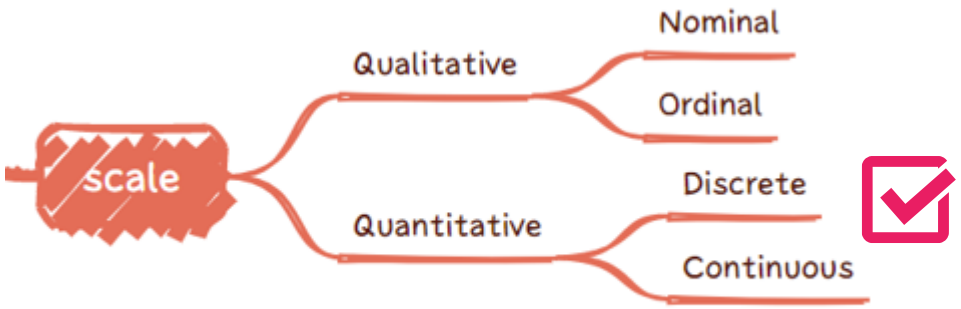
Categories of Data

Qualitative data represents groupings

Type of variable	What does the data represent?	Examples
Binary variables (aka dichotomous variables)	Yes or no outcomes.	• Heads/tails in a coin flip • Win/lose in a football game
Nominal variables	Groups with no rank or order between them.	• Species names • Colors • Brands
Ordinal variables	Groups that are ranked in a specific order.	• Finishing place in a race • Rating scale responses in a survey, such as Likert scales*

Categories of Data

Quantitative data represents amounts



Discrete Data

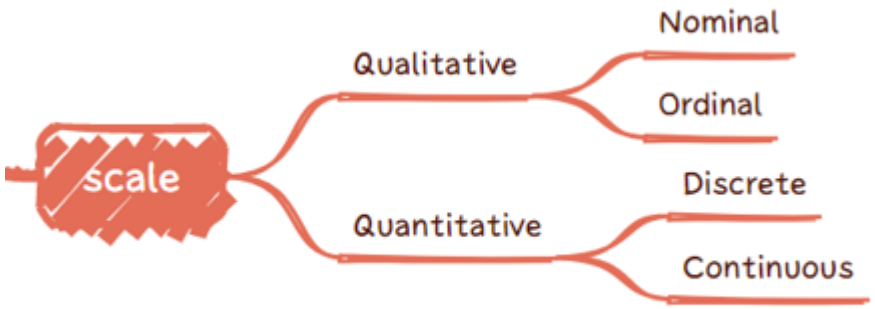
Discrete data is the information that often **counts of some event**, i.e. can only take specific values. These are often *integer-based*, but not necessarily.

eg:

- Number of times a coin was flipped
- Shoe sizes of people

Categories of Data

Quantitative data represents amounts



Continuous Data

Continuous Data is the information that has the **possibility of having infinite values**, i.e. can take any value within a range.

eg:

- How many centimeters of rain fell on a given day
- Distance, volume

Categories of Data

Quantitative data represents amounts

Type of variable	What does the data represent?	Examples
Discrete variables (aka integer variables)	Counts of individual items or values.	• Number of students in a class • Number of different tree species in a forest
Continuous variables (aka ratio variables)	Measurements of continuous or non-finite values.	• Distance • Volume • Age

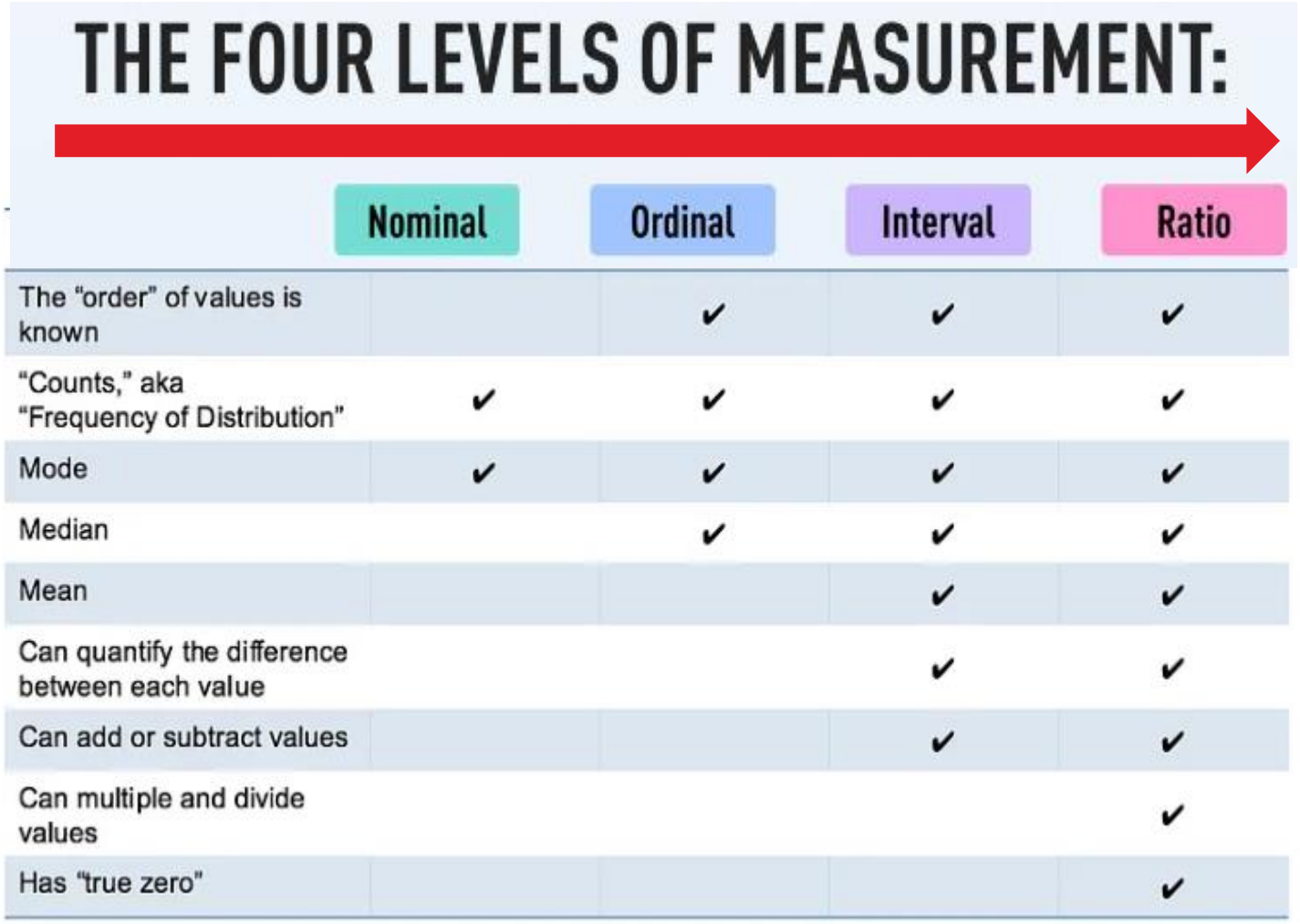
Level of Measurement

When we conducting statistical analysis…

Data is often represented in a rectangular array where each column is a variable and each row is an observational unit or sampling unit.

It is important, as it determines the type of statistical analysis you can carry out.

Level of Measurement

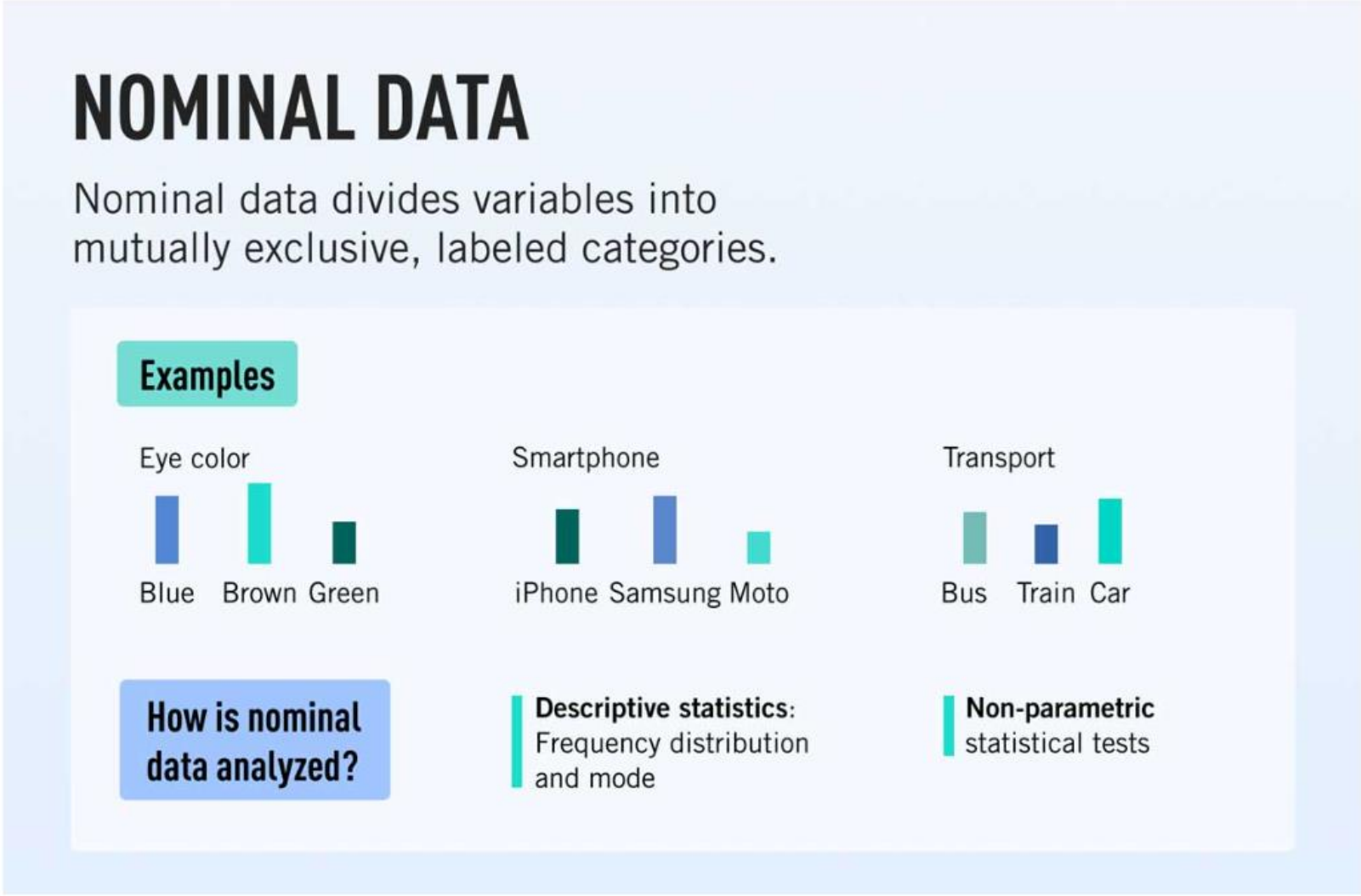


THE FOUR LEVELS OF MEASUREMENT:

	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

- **Nominal variables** are used to "name" a series of values.
- **Ordinal scales** provide good information about the order of choices.
- **Interval scales** give us the order of values + the ability to quantify the difference between each one.
- **Ratio scales** give us the ultimate–order, interval values, plus the ability to calculate ratios since a "true zero" can be defined.

Level of Measurement



Level of Measurement

ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades

Education level

Seniority level

How is ordinal data analyzed?

Descriptive statistics: Frequency distribution, mode, median, and range

Non-parametric statistical tests

Level of Measurement

INTERVAL DATA

Interval data is measured along a numerical scale that has equal intervals between adjacent values.

Examples

Temperature

IQ score

Income ranges

How is interval data analyzed?

- **Descriptive statistics**: Frequency distribution; mode, median, and mean; range, standard deviation, and variance
- **Parametric statistical tests** (e.g. t-test, linear regression)

Level of Measurement

RATIO DATA

Ratio data is measured along a numerical scale that has equal distances between adjacent values, and a true zero.

Examples

Weight in KG

...50 70 90...

Number of staff

...10 30 50...

Income in USD

...20k 40k 60k...

How is ratio data analyzed?

- **Descriptive statistics**: Frequency distribution; mode, median, and mean; range, standard deviation, variance, and coefficient of variation
- **Parametric statistical tests** (e.g. ANOVA, linear regression)

Data Collection Method

There are many methods to collect data:

- Census
- Sample survey
- Administrative data

- Crowdsourcing
- Web scraping
- Remote sensing
- Statistical registers
- Open data
- Big data

Type of Data

Data can be passive and active.

Passive data
e.g., web browsing history, sensor data, GPS, ect.

Active data
e.g., survey response, social media post, form submissions, ect.

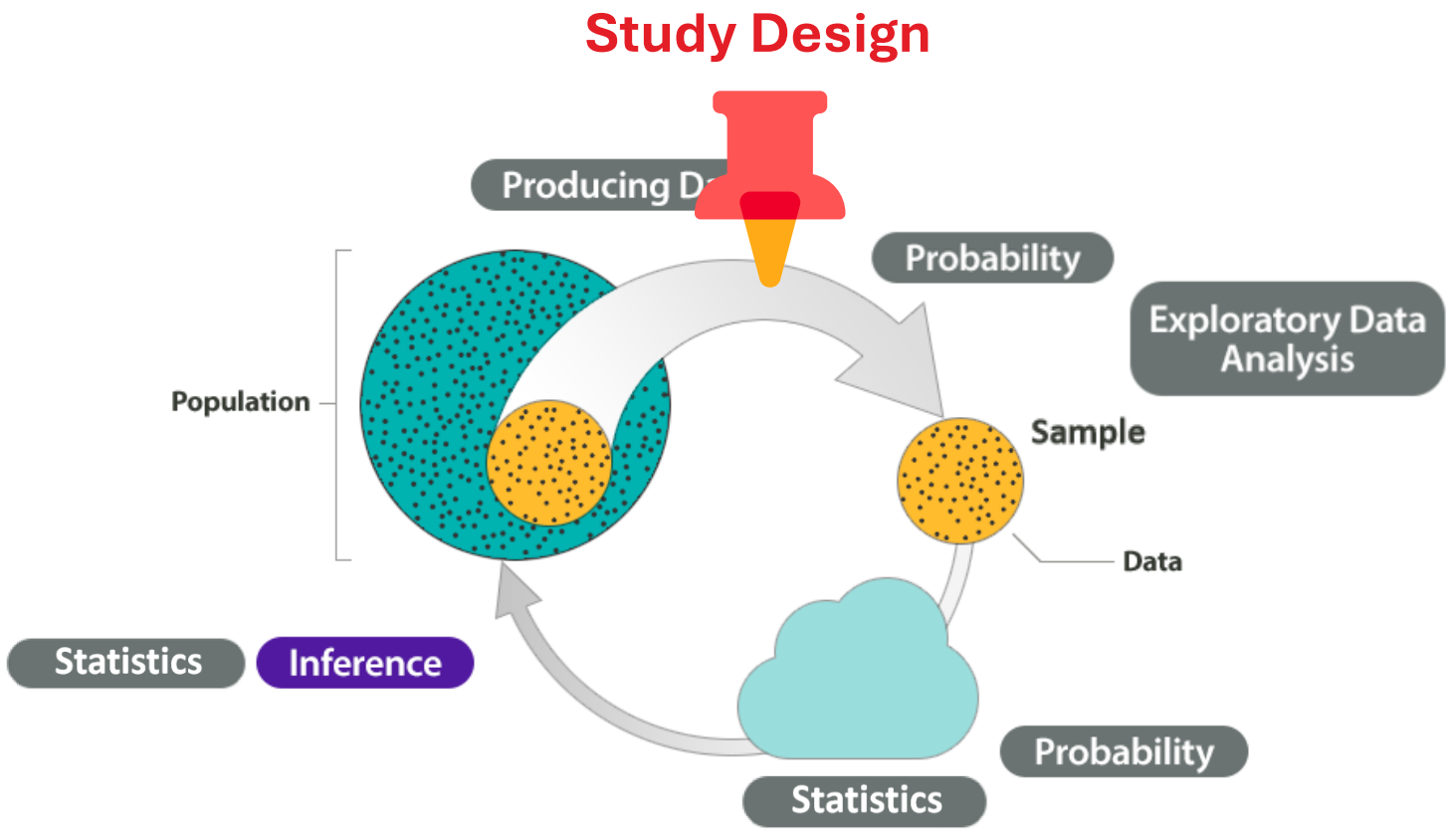
Data Quality

Generally speaking, statistical data is evaluated in terms of its "fitness for use" - that is, the extent to which the statistical information can be relied upon to fulfill the information needs.

Quality dimensions:

- Relevance (e.g., relate to question or data gap)
- Accessibility (e.g, easy to access or affordable)
- Accuracy (e.g., cover the required population and period of reference)
- Timeliness (e.g., willing to accept lower accuracy to get the data faster)
- Interpretability (e.g., useful, reliable, complete)
- Coherence (e.g., information consistent over time)

Big Picture of Statistics



Data Gathering

To begin, the following questions should be addressed:

- Why is this being conducted?
- Whom will the collected information be about?
- What do I need to know?
- How will the information be used?
- How accurate and timely does the information have to be?

Data is everywhere …
What is your research question

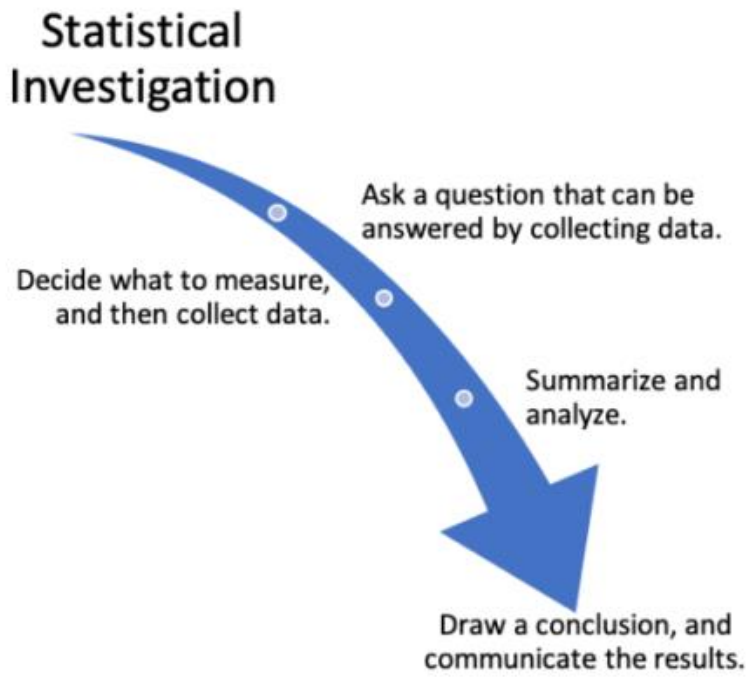
Formulate a relevant research question

The Role of Statistics

- Design: Planning and carring out research study
- Description: Summarizing and exploring data
- Inference: Making predictions and generalization

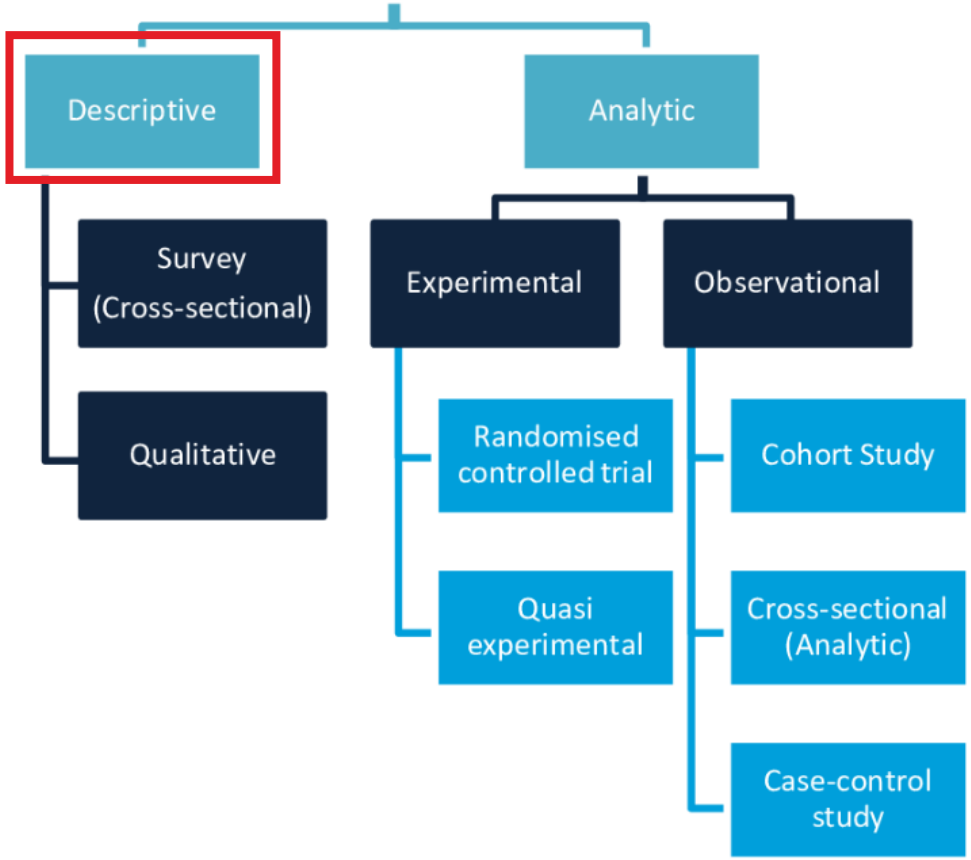
Study Design

In statistical studies, the type of study design used and the details of the design are important in determining what kind of conclusions we may draw from the results.



Study Design

In statistical studies, the type of study design used and the details of the design are important in determining what kind of conclusions we may draw from the results.



Sample Surveys

Sample surveys are a special type of data collection that usually aim to discover the opinions of people on certain topics.

In a sample survey, a sample of individuals is selected from a larger population of individuals.

The idea of **sampling** is to study a small part of the population in order to gain information about the population as a whole.

Conclusions drawn from a sample are valid only when the sample is drawn in a well-defined way.

Survey Design

To design a survey, many decisions have to be made about the following issues.

- Survey objectives
- Target population
- Data requirements
- Type of collection
- Minimizing error
- Sample size
- Analysis plan
- Questionnaire design
- Data collection methods
- Data processing plan
- Quality control
- Analysis and dissemination of results

Survey Bias

Bias is defined as a "deviation of results or inferences from the truth, or processes leading to such a deviation" and it occurs in every survey.

- **Selection:** How was the survey sample selected? How many participants completed the survey? Was the sample broad enough to capture the most valuable insights?

- **Response:** How are participants swayed by leading factors from the interviewer? Such as the questions asked, their format, and the respondent's desires to be socially accepted?

- **Interviewer:** Is the interviewer unconsciously sending signals to participants that could alter their answers? Are the interviewers biased? Are the survey questions tailored towards specific outcomes?

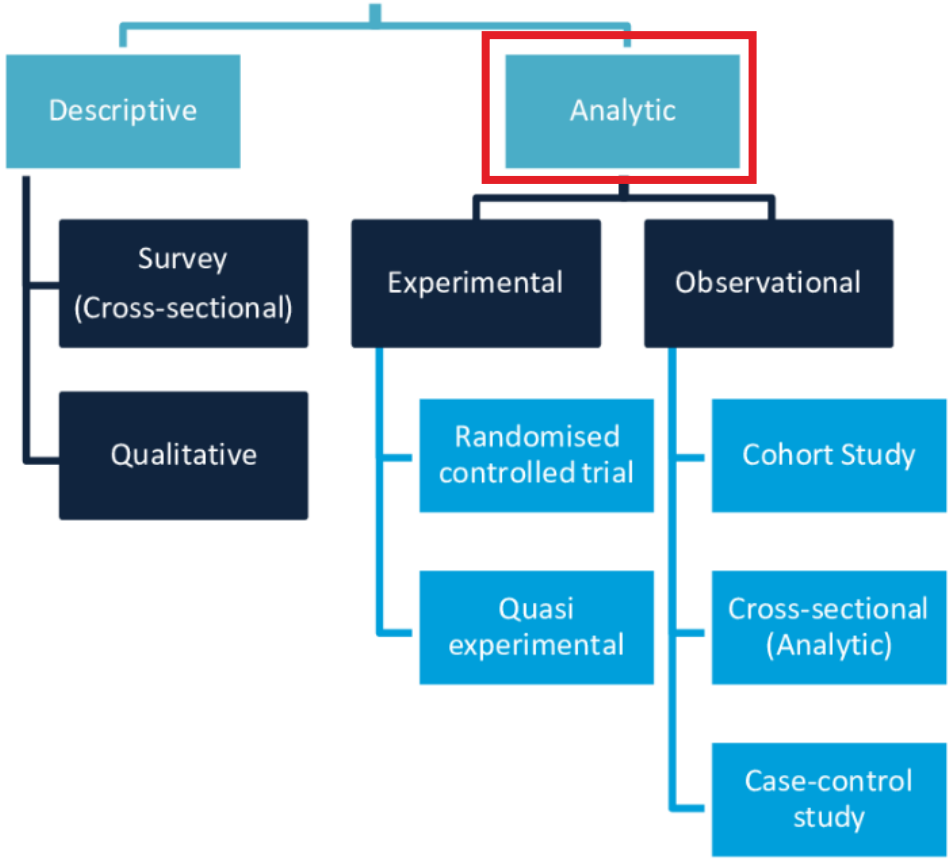
Survey Example

Doctor of Artificial Intelligence

https://forms.office.com/Pages/ResponsePage.aspx?id=nReQ_geCskyINAzifuAWLbxijFLlNFFOrsXVJM7sZEBUNFZEM0JHVVRXRUxOQ0VRQ0ZHNldPUUQ4MS4u

Study Design

In statistical studies, the type of study design used and the details of the design are important in determining what kind of conclusions we may draw from the results.



Study Design

Statistical methods are driven by te data that we collect. We typically obtain data from two distinct soures:

- Observational studies
  - observing and mesuring specific charactertistics without attempting to modify the subject being studied
  - Focus on the association

➢ *Example: Company officials wished to study the relation between the age of an employee and the number of days of illness in a year.*

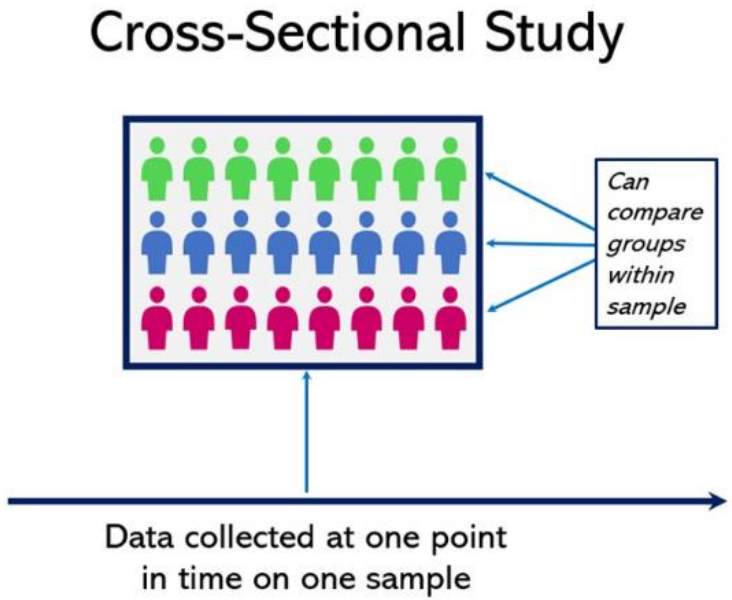
Explanatory variable not controlled → age is observed

Establish associations but no cause-and-effect: a positive relation between age and number of days of illness may not imply that number of days of illness is the direct result of age → younger employees work indoors while older employees usually work outdoors, and therefore work location is more responsible for the number of days of illness instead of age

Observational Studies

- Cross sectional study

Data are observed, measured, and collected at one point in time.



- Longitudinal study
  - Retrospective
  - Prospective

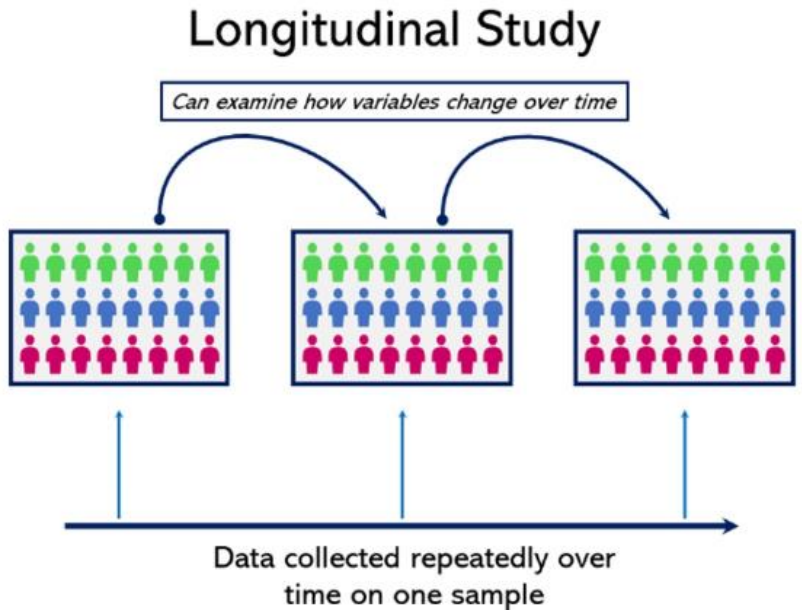
Observational Studies (Longitudinal)

Retrospective (or case control) study

Data are collected from the past by going back in time (examine records, interviews,...)

Prospective (or longitudinal or cohort) study

Data are collected in the future form groups sharing common factors (call cohorts).



Study Design

Statistical methods are driven by te data that we collect. We typically obtain data from two distinct soures:

- Experimental studies
  - apply some treatment and then observe its effects on the subjects; (subjects in experiments are called experimental units)
  - Focus on cause-and-effect

➢ *Example: Effect of Vitamin C on prevention of colds in 800 children. Half of the children were selected at random and received Vit C (treatment group) the remaining children received a placebo (control group)*

Qualitative explanatory factor with two levels and children as experimental units

Experimental Study

Randomized controlled trial (RCT)

1. Eligible people are randomly assigned to one of two or more groups.
2. One group receives the intervention (such as a new drug) while the control group receives nothing or an inactive placebo.
3. The researchers then study what happens to people in each group.
4. Any difference in outcomes can then be linked to the intervention.

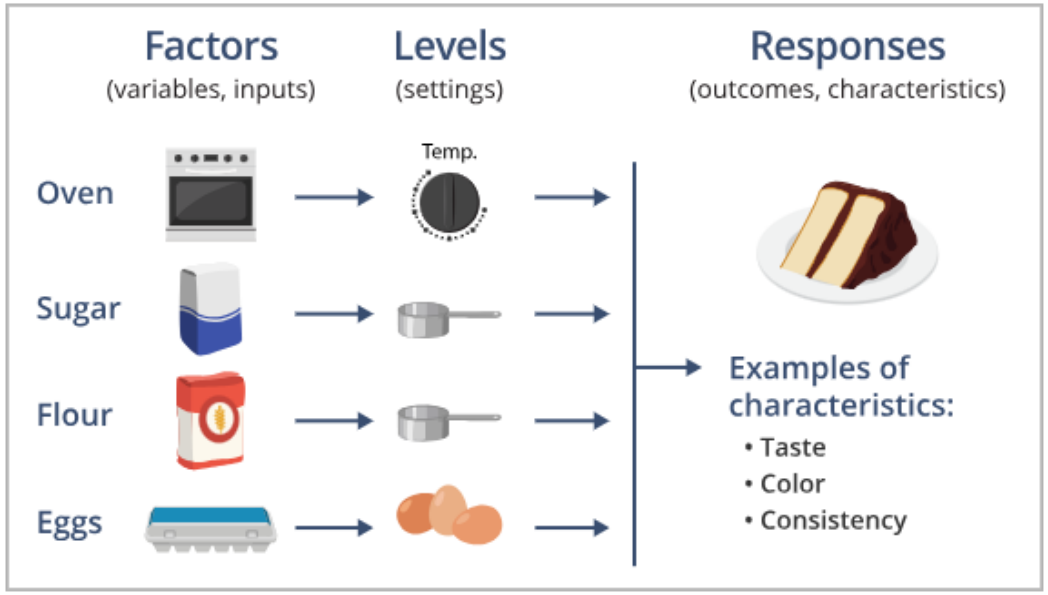
Study Design

Design of Experiment

7-step procedure in DOE

1. State Objective → 2. Choose Response
3. Choose Factors, Levels, Region → 4. Choose Design
5. Perform Experiment → 6. Statistical Data Analysis
7. Draw Conclusions and Make Recommendations

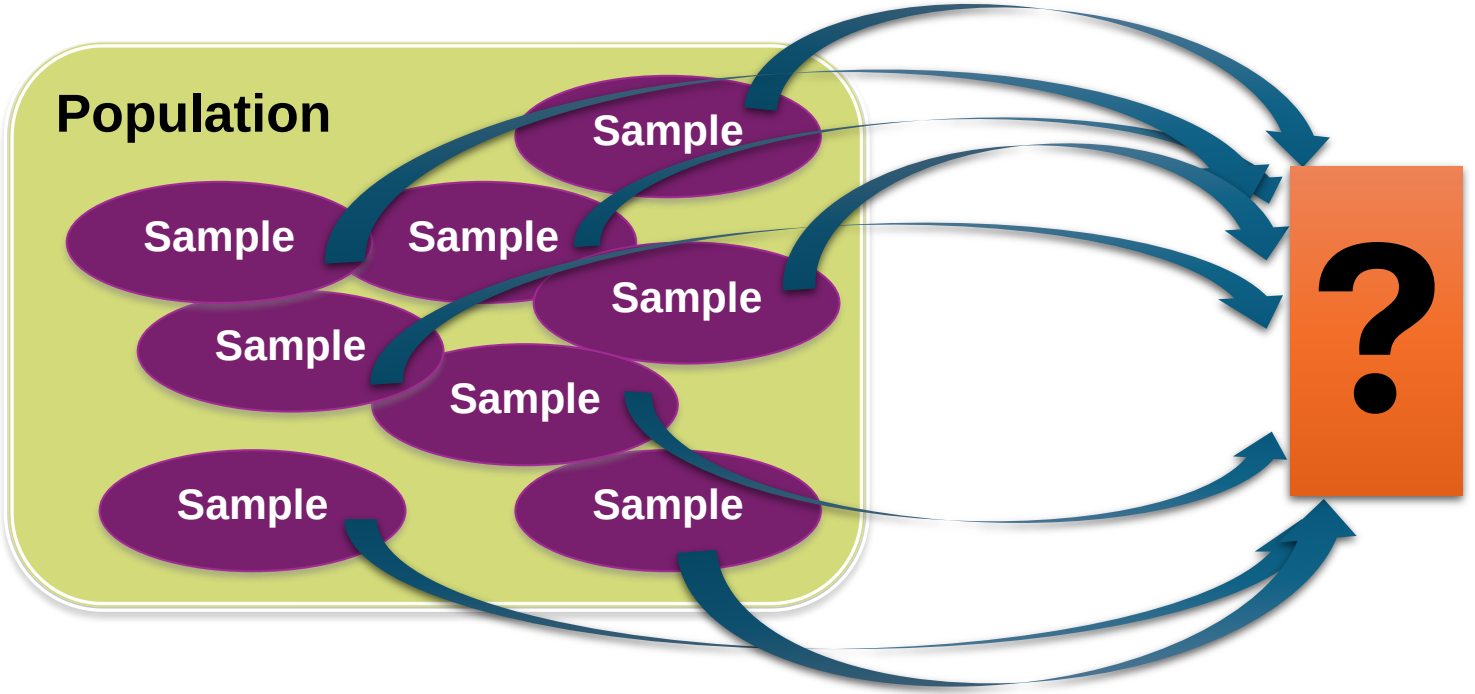
- Requiring inter-discipline collaboration: steps 1, 2, 3, 7
- Statistician: steps 4, 6
- Engineer: step 5



Further readings: https://www.moresteam.com/toolbox/design-of-experiments

Bias and Variability

What would happen if we took many samples?



Measurements: precision versus accuracy

Precision of a variable (Variability): the degree to which a variable has nearly the same value when measured several times. It is a function of random error (chance) and is assessed as the reproducibility of repeated measurements. [**Spread**]

Example: *weigh the same person 3 times on an electronic balance and obtain slightly different measurements – 67.5 kg, 67.4 kg and 67.6 kg*

Variability may be due to operator, instrument and subject

Minimize random error and improve precision

- Operating manuals, training the operator, refining / automating instruments
- Repeat the measurement and average over a larger number of observations (but! added cost, practical difficulties)

Measurements: precision versus accuracy

Accuracy of a variable (Bias): the degree to which a variable actually represents what it is supposed to represent. It is a function of systematic error (bias) which is often difficult to detect and has important influence on the validity of the result. [**Center**]

Example: *incorrect calibration of an instrument*

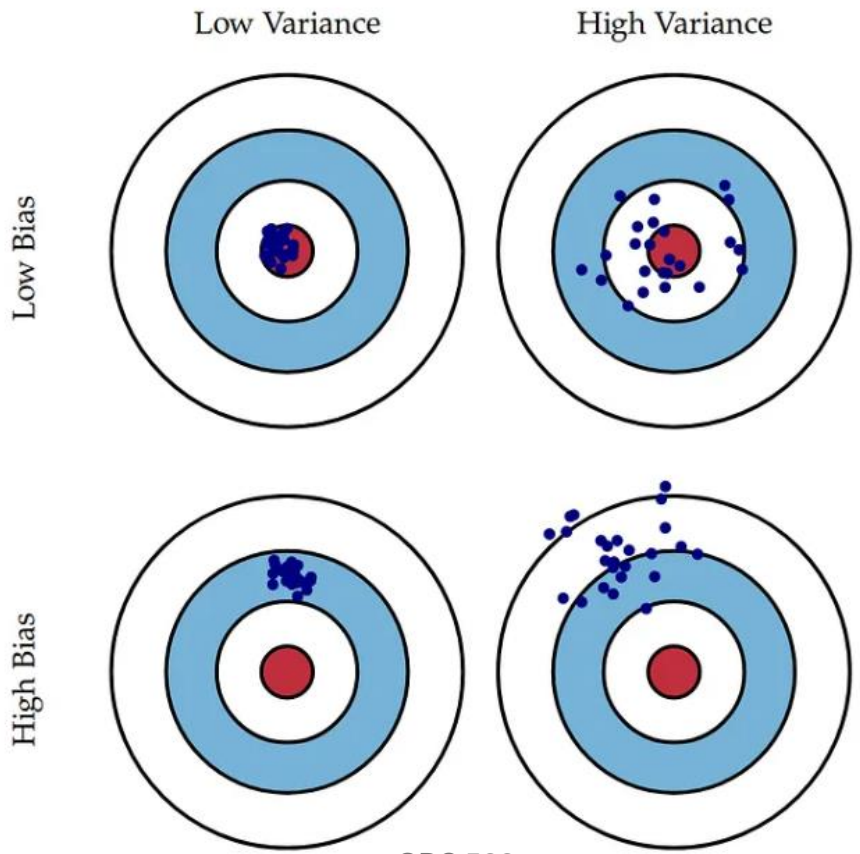
Improve accuracy and minimize bias

- Operating manuals, training the operator, refining / automating instruments
- Periodic calibration using a gold standard
- Blinding: double–blind study: the experimental subject and the evaluator *have no information on which treatment that they receive or give*, any inaccuracy in measuring the outcome will be the same in the 2 groups

Measurements: precision versus accuracy

Bias and variance in shooting arrows at a target. Bias means that the archer systematically misses in the same direction. Variance means that the arrows are scattered.

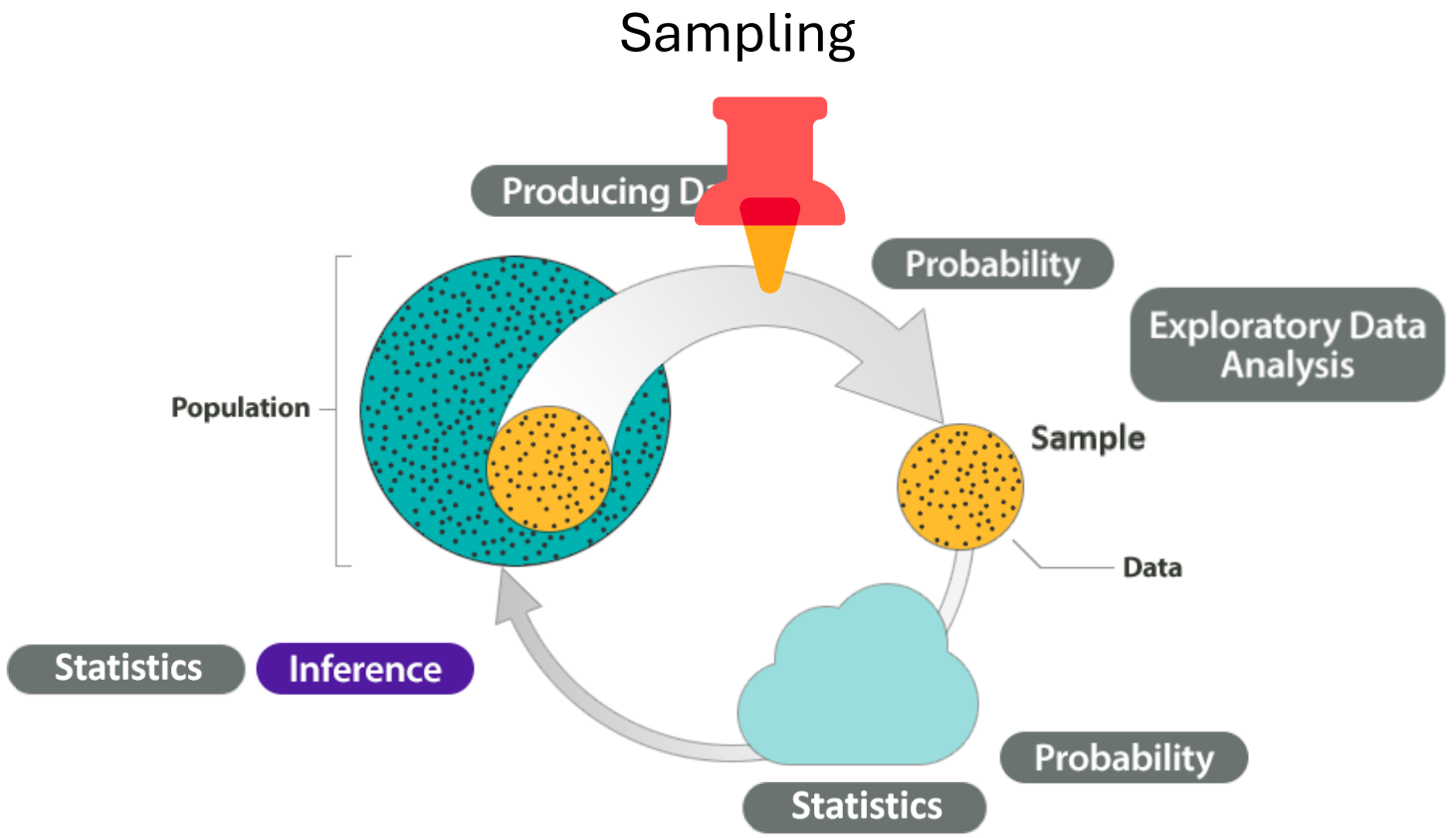
- Tradeoff



Review – Type of Data
– Study Design

www.kahoot.it

Big Picture of Statistics



Sampling

Scenario: The university wants to assess student satisfaction with campus facilities.

How will you collect the data?

Sampling

Sampling process of selecting a small number of elements (samples) from a larger defined target group of elements (Population) such the information gathered from the samples will allow judgments to the population.

Sampling

Scenario: The university wants to assess student satisfaction with campus facilities.

How will you collect the data?

Poor Data Collection Method:
The university decides to distribute an online survey only to students who frequently visit the campus coffee shop.

Sampling

If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical tutoring can salvage them.

Method used to collect sample data influences the quality of the staitstical analysis.

What other data collection problems can you think of?

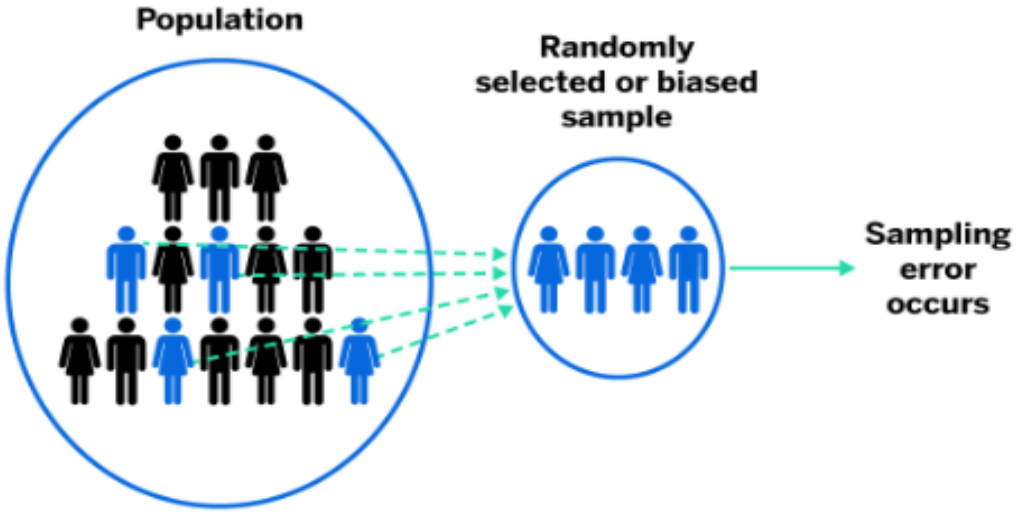
Sampling Distortion

- Sampling Error
- Sampling Bias

Sampling Error

Sampling error is the error caused by observing a sample instead of the whole population.

- difference between $\bar{x}$ and μ
- inevitable gap

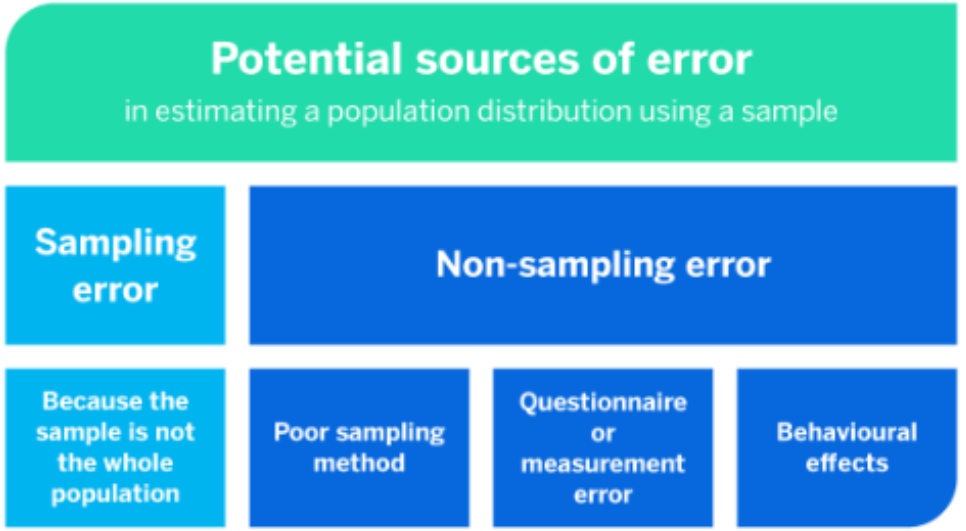


Sampling Error

Non-sampling error refers to all sources of error that are unrelated to sampling.

- when there are problems with the sampling method, or the way the survey is designed or carried out.

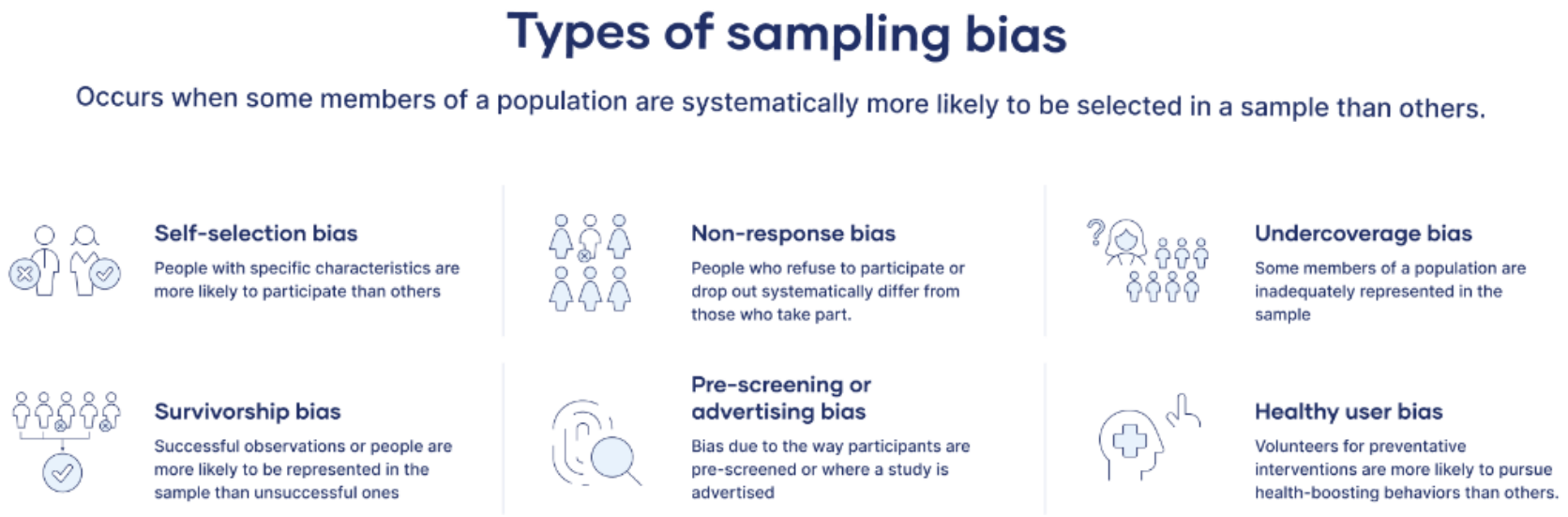
Table summarising types of error.



(part)

Sampling Bias

Sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower sampling probability than others.



Discussion

A health organization wants to study the dietary habits of teenagers. They decide to survey students at a local high school known for its strong athletic program.

What types of sampling distortion are present in this scenario, and how might it influence the conclusions about the dietary habits of all teenagers?

- Sampling Bias
- The dietary habits reported may reflect the nutritional needs and preferences of athletes rather than those of all teenagers, that could lead to misguided policy recommendations or health initiatives aimed at the broader teenage population.

Discussion

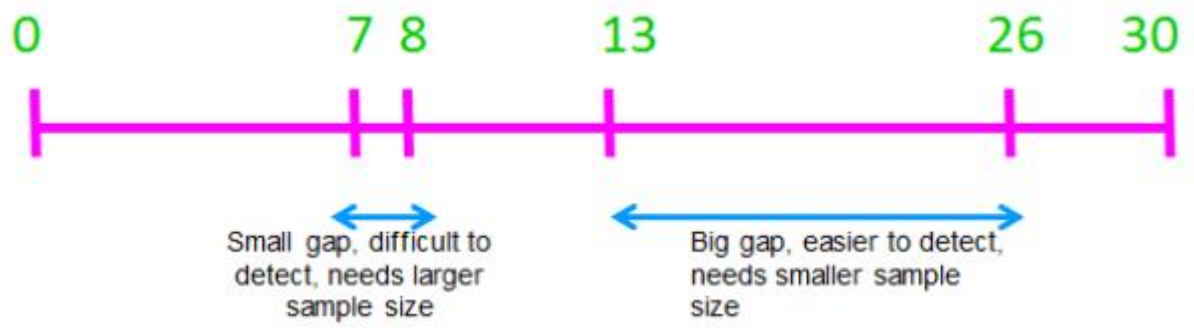
A researcher is studying the average test scores of high school students in a district. They randomly select 30 students from one school, which happens to have a particularly high-performing academic reputation.

What sampling sampling distortion could occur in this scenario, and how might it impact the estimation of the average test scores for all high school students in the district?

- Sampling Error
- The average test score calculated may not accurately reflect the performance of all high school students in the district, may potentially influencing funding and resource allocation decisions.

Characterics of Good Sample

- **Representativeness**
- Accessible
- Cost effective
- **Of the right size**
- Generalizability (lack of bias)
- Obtained with minimum sampling error
- It shoud be suitable for anlysis as per the study design



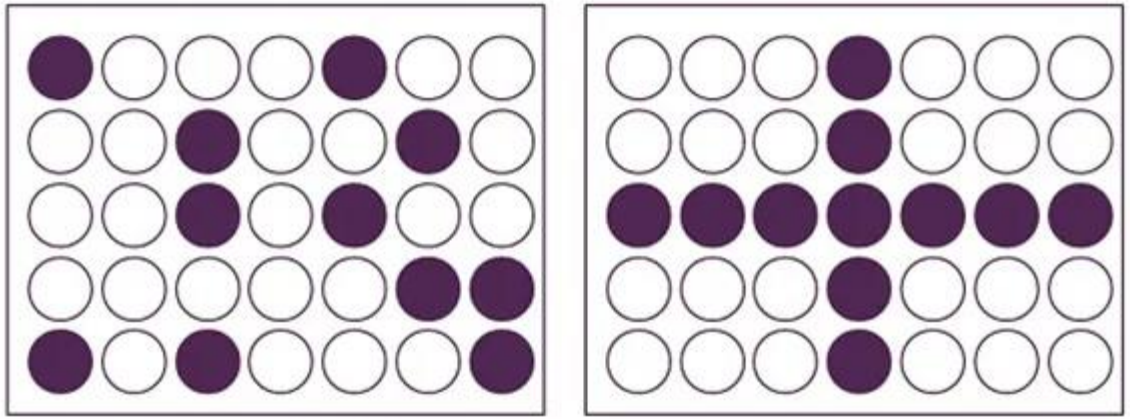
How can we select representative samples?

Sampling Method

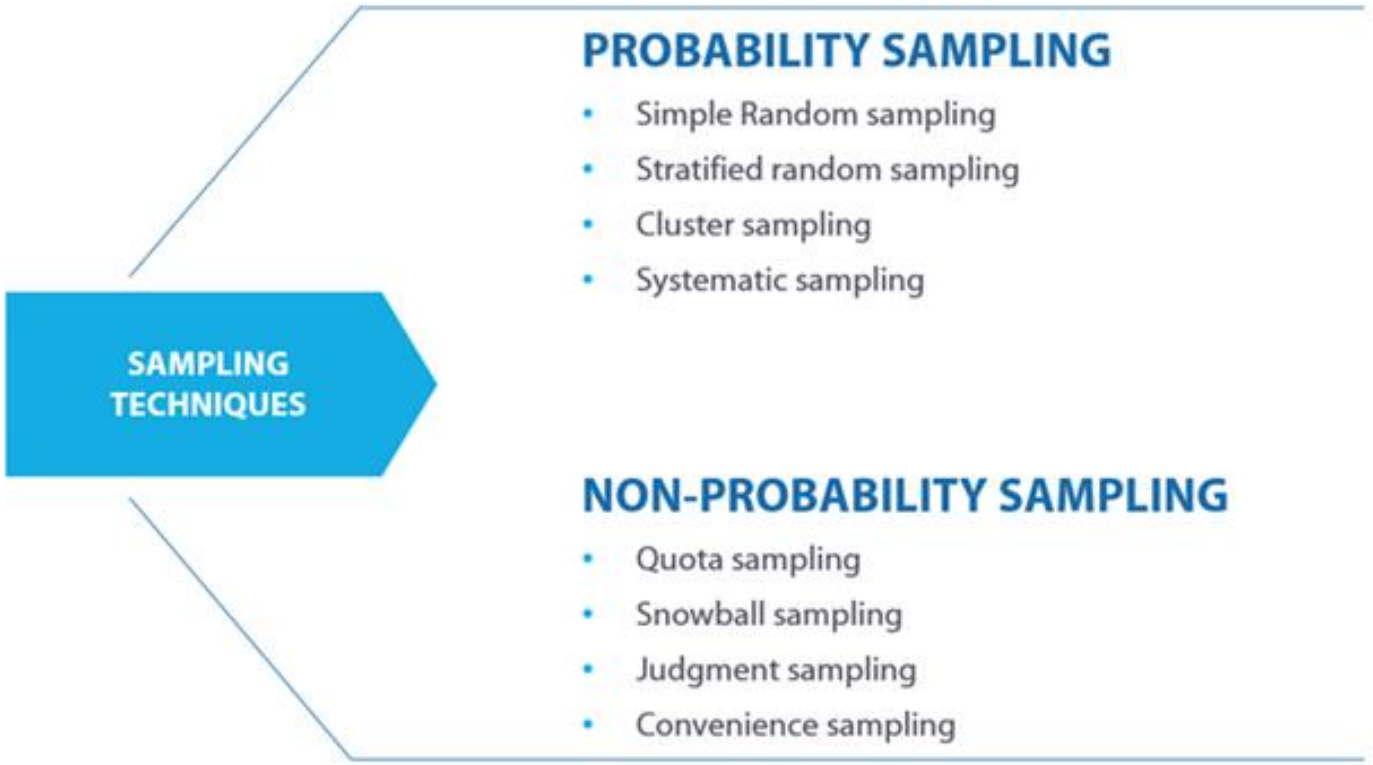
There are two primary types of sampling methods:

- ***Probability sampling*** involves random selection, allowing you to make strong statistical inferences about the whole group. [quatitative reseach]

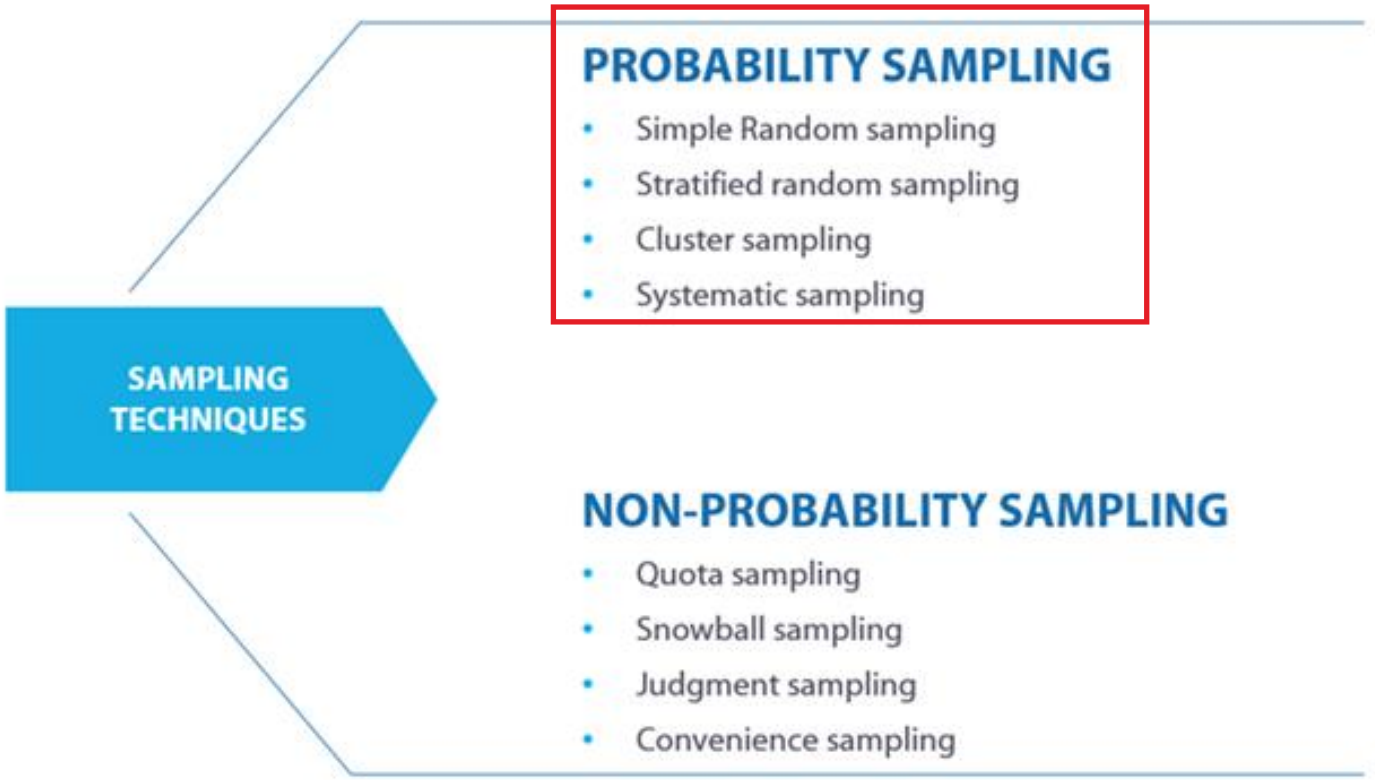
- ***Non-probability sampling*** involves non-random selection based on convenience or other criteria, allowing you to easily collect data. [exploratory and qualitative research]



Sampling Techniques

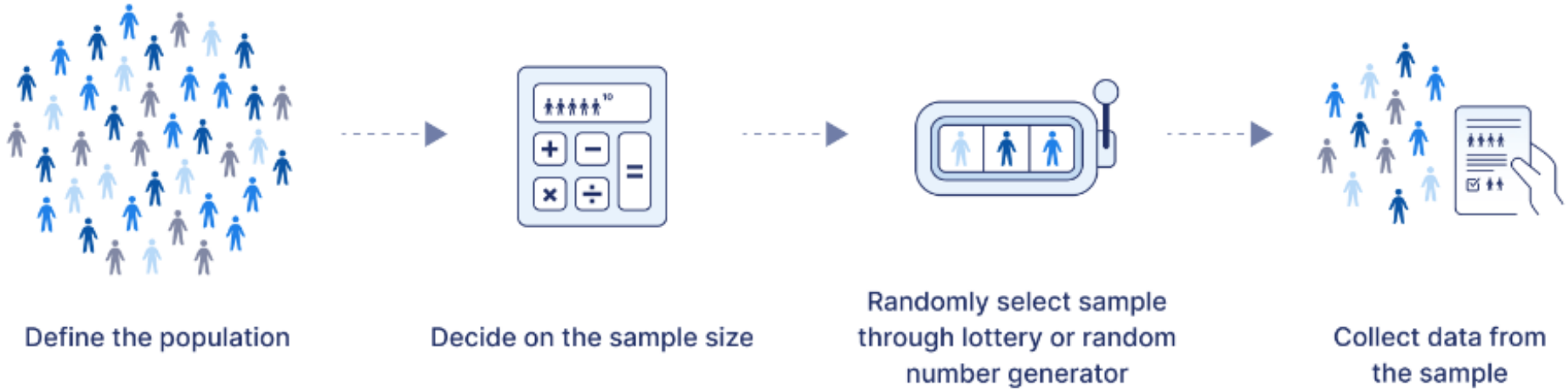


Sampling Techniques



Simple Random Sample

Simple random sample is a randomly selected subset of a population.



- Most straightforward
- Lower risk for research biases like sampling bias and selection bias

Simple Random Sample

Example

The American Community Survey (ACS) uses simple random sampling. Officials from the United States Census Bureau follow a random selection of individual inhabitants of the United States for a year, asking detailed questions about their lives in order to draw conclusions about the whole population of the US.

- High internal validity: randomization is the best method to reduce the impact of potential confounding variables.
- With a large enough sample size, a simple random sample has high external validity: it represents the characteristics of the larger population.

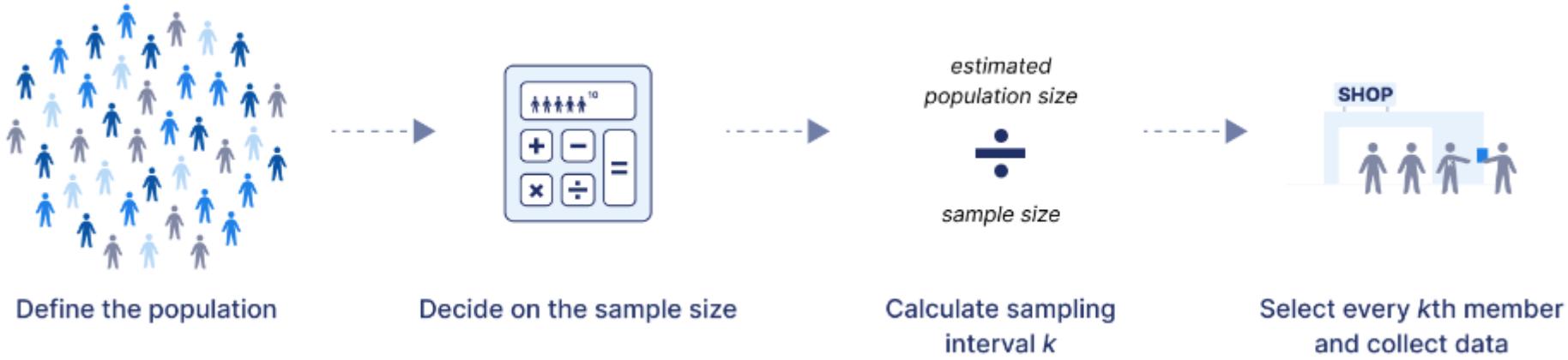
Simple Random Sample

- Simple random sampling can be **challenging** to implement in practice. To use this method, there are some prerequisites, you
  - have a complete list of every member of the population
  - can contact or access each member of the population if they are selected
  - have the time and resources to collect data from the necessary sample size

Simple random sampling works best if you have a lot of time and resources to conduct your study, or if you are studying a limited population that can easily be sampled.

Systematic Sampling

Systematic sampling is a method in which researchers select members of the population at a regular interval (or k) determined in advance.



- Applied with/without a list of the entire population
- It's essential to consider the order in which your population is listed to ensure that your sample is valid.

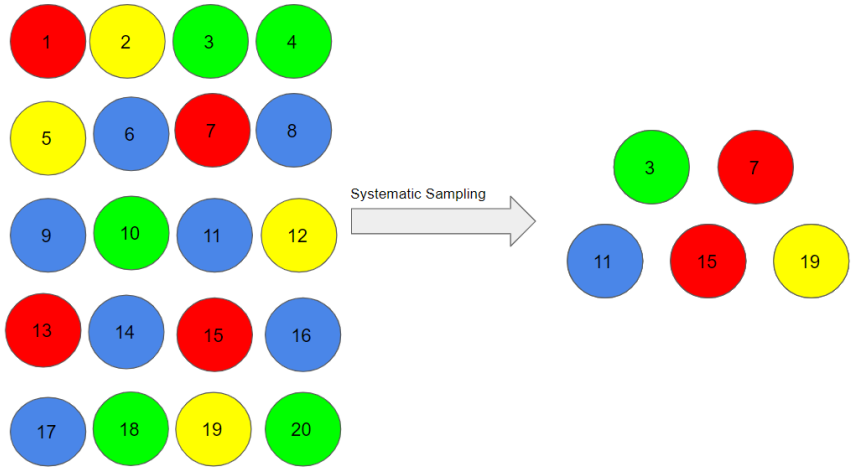
Systematic Sampling

Example: Systematic sampling

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

Example: Alternating list

Your population list alternates between men (on the even numbers) and women (on the odd numbers). You choose to sample every tenth individual, which will therefore result in only men being included in your sample. This would obviously be unrepresentative of the population.



You should not use systematic sampling if your population is ordered cyclically or periodically, as your resulting sample cannot be guaranteed to be representative.

Stratified Sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways (e.g., gender identity, age range, income bracket, job role).

- Rely on stratified sampling when a population's characteristics are diverse and they want to ensure that every characteristic is properly represented in the sample.

Stratified Sampling

> **Example: Stratified sampling**
>
> The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

- To use stratified sampling, you need to be able to divide your population into mutually exclusive and exhaustive subgroups.
  - Every member of the population can be ONLY classified into one subgroup.
- Best choice when you believe that subgroups will have different mean values for the variable(s) you're studying.

Stratified Sampling

It has several potential advantages:

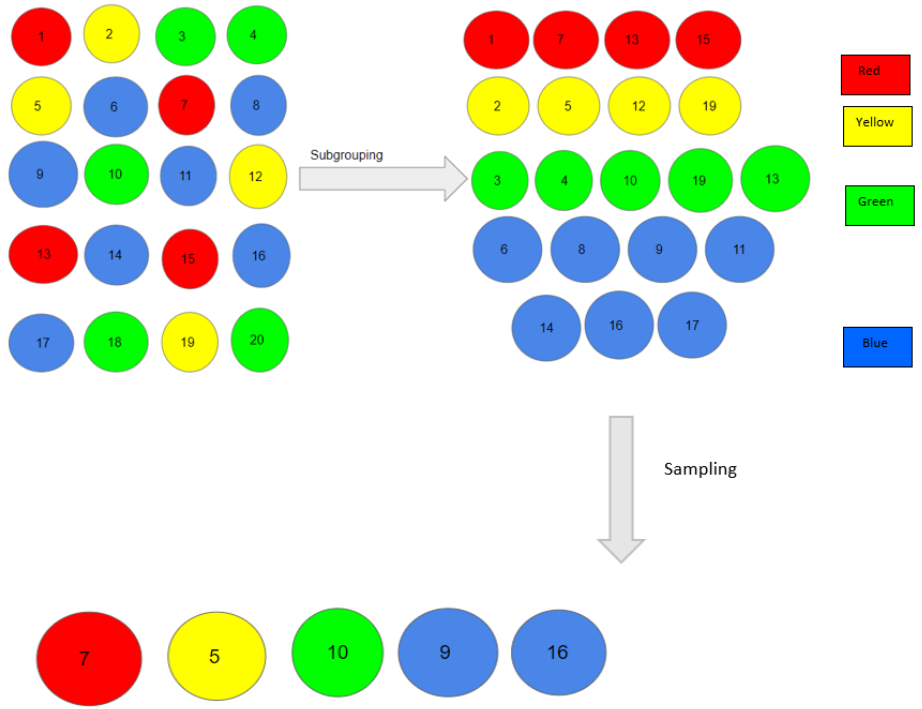
- Ensuring the diversity of your sample
- Ensuring similar variance
- Lowering the overall variance in the population
- Allowing for a variety of data collection methods

Research example

You are interested in how having a doctoral degree affects the wage gap between gender identities among graduates of a certain university.

Because only a small proportion of this university's graduates have obtained a doctoral degree, using a simple random sample would likely give you a sample size too small to properly compare the differences between men, women, and those who do not identify as men or women with a doctoral degree versus those without one.

Therefore, you decide to use a stratified sample, relying on a list provided by the university of all its graduates within the last ten years.



Cluster Sampling

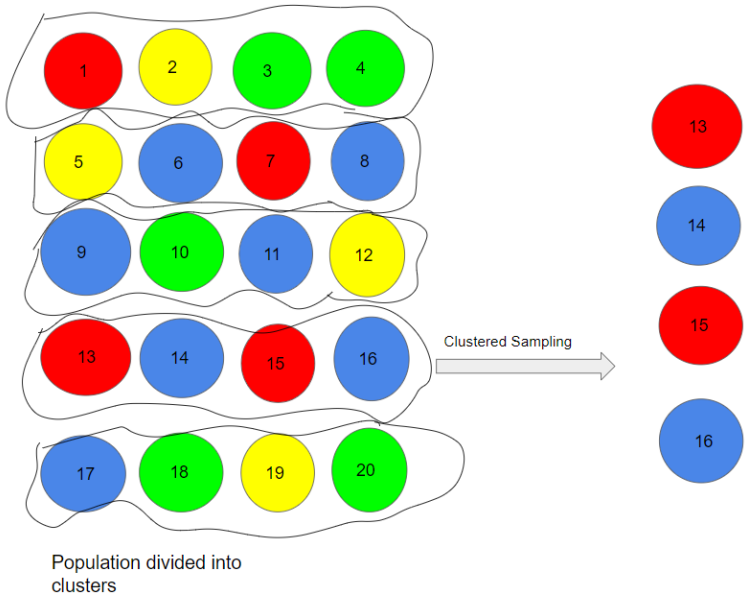
Cluster sampling divides a population into smaller groups known as clusters (each cluster should have similar characteristics to the whole sample). They then randomly select among these clusters to form a sample.

- It is often used to study large populations, particularly those that are widely geographically dispersed.
- Researchers usually use pre-existing units such as schools or cities as their clusters.

Cluster Sampling

Example: Cluster sampling

The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.



- This method is good for dealing with large and dispersed populations.
- There is more risk of error in the sample, as there could be substantial differences between clusters.
- It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Multistage Sampling

Multistage sampling, or multistage cluster sampling, you draw a sample from a population using smaller and smaller groups (units) at each stage. It's often used to collect data from a large, geographically spread group of people in national surveys.

Multistage Sampling

Research example

Your population is all students aged 13–19 registered at schools in your state.

If you're unable to access a complete sampling frame, you can't use single-stage probability sampling from the whole population. In addition, collecting data from a sample of individuals across the state would be very difficult, costly, and time-intensive.

Instead, you decide to use a multistage sampling method to collect a representative sample of participants.

A combination of different sampling methods.

Multistage Sampling

Multistage sampling

In the **first stage**, you make a list of school districts within the state. You select 15 school districts as your PSUs.

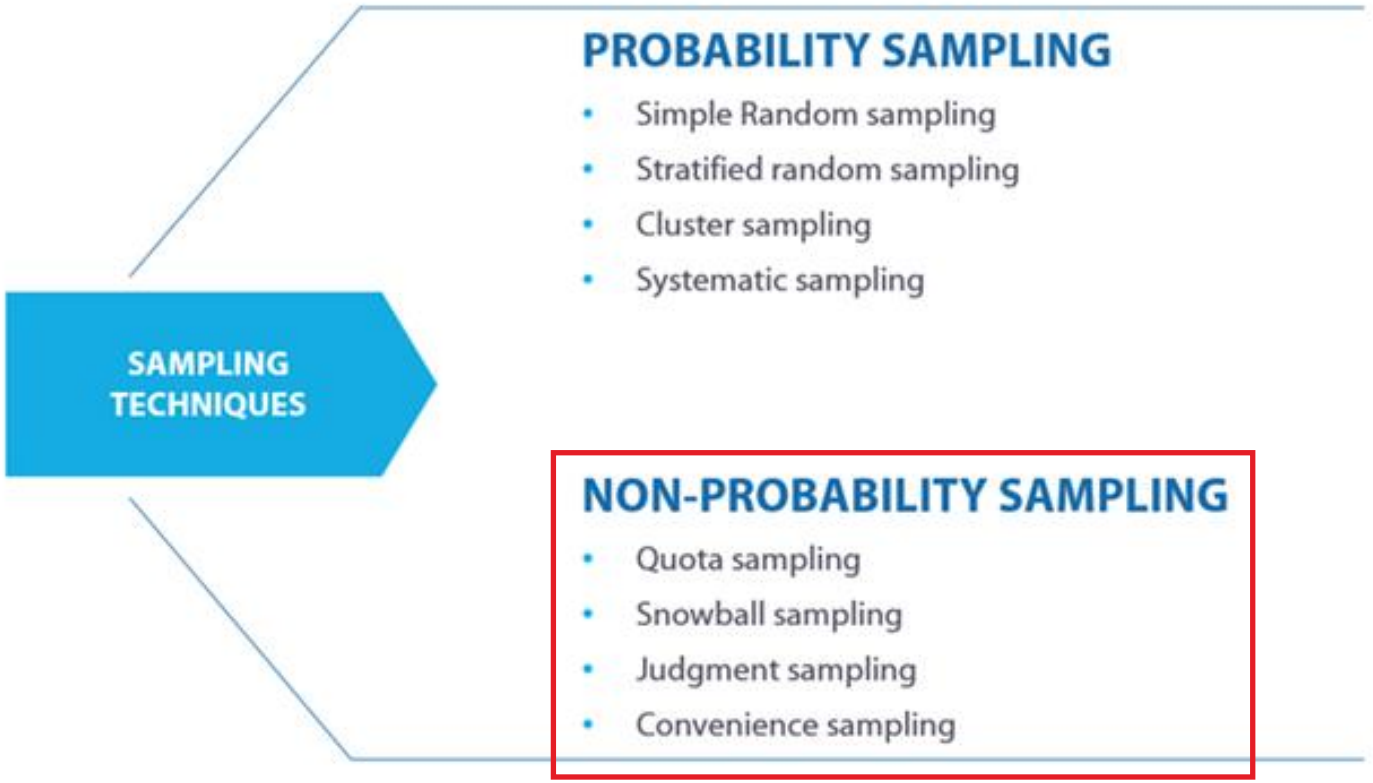
In the **second stage**, you list all schools within those school districts. You select 10 schools from each district as your SSUs.

In the **third stage**, you obtain a list of all students within those schools. You select 50 students from every school as your USUs, and collect data from those students.

Advantage:

- No need to start with a sampling frame of target population.
- It's relatively inexpensive and effective when have a large or geographically dispersed population.
- It's flexible— can vary sampling methods between stages based on what's appropriate or feasible.

Sampling Techniques



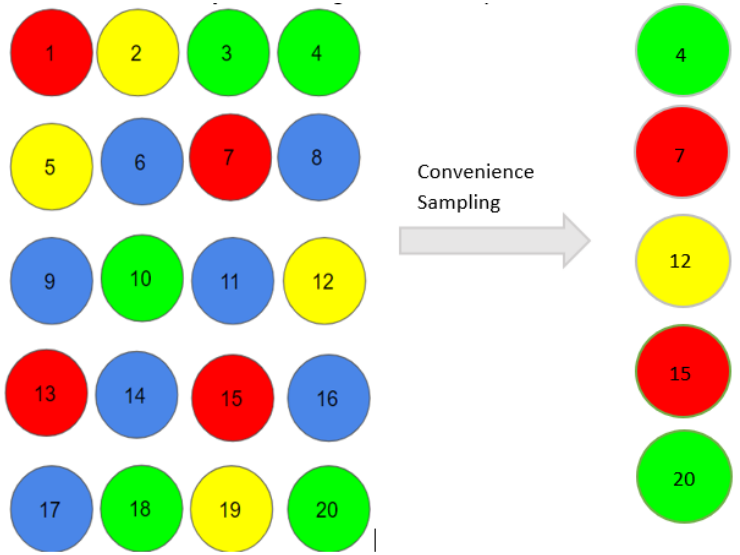
Convenience Sampling

Convenience sample simply includes the individuals who happen to be most accessible to the researcher.

- Easier and cheaper to access
- Ways to collect: online, social media post, in-person, crowdcourcing, pre-existing groups, ect..
- Higher risk of sampling bias

Convenience Sampling

> **Example: Convenience sampling**
>
> You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.



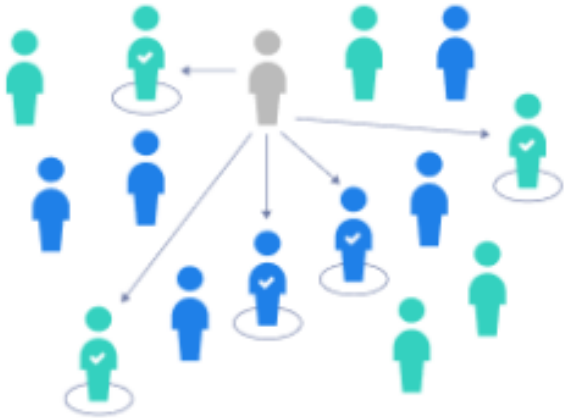
Convenience sampling could be a good fit if:

- want to get an idea of people's attitudes and opinions
- want to run a test pilot for survey
- want to generate hypotheses that can be tested in greater depth in future research

Be aware of research bias, such as selection bias and sampling bias.

Purposive Sampling

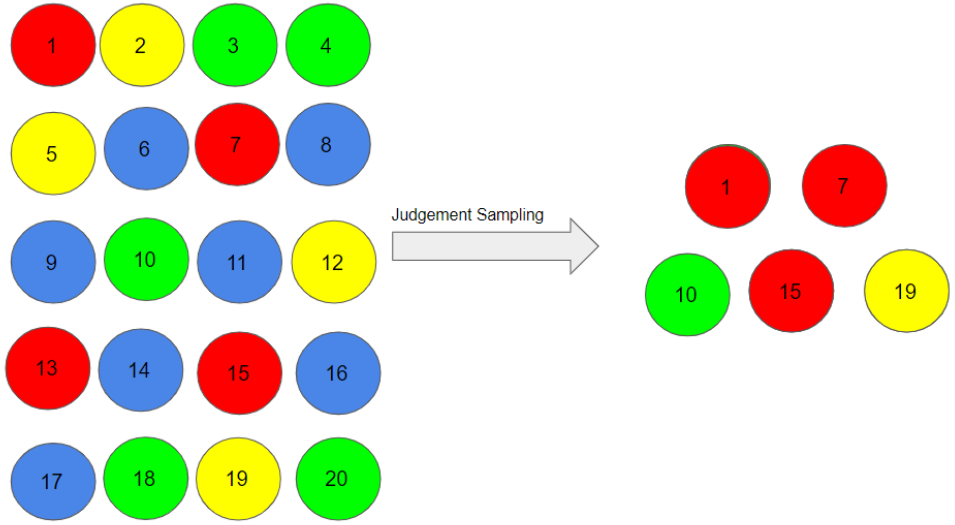
Purposive sample, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.



- wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific.
- Make sure to describe inclusion and exclusion criteria and beware of observer bias affecting arguments.

Purposive Sampling

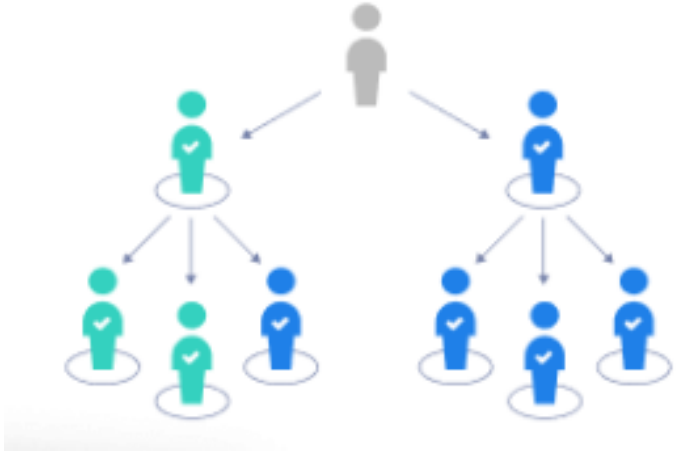
> **Example: Purposive sampling**
>
> You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.



Purposive sampling is best used when want to focus in depth on relatively small samples. Perhaps would like to access a particular subset of the population that shares certain characteristics when you have a lot of background information about the topic. It is at high risk for research biases like observer bias.

Snowball Sampling

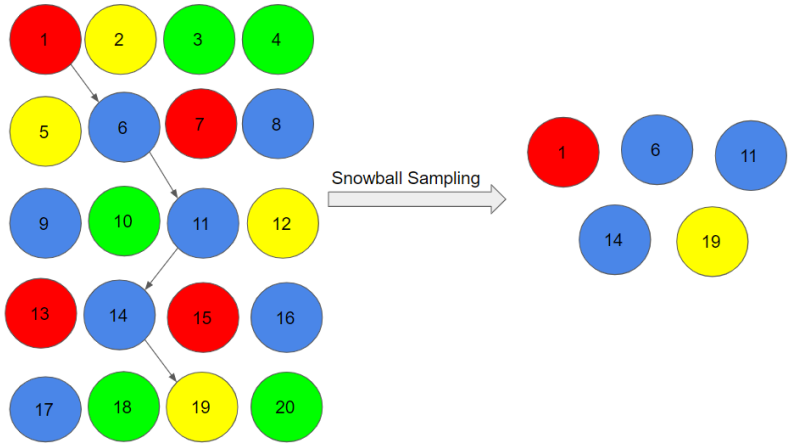
Snowball sampling is a sampling method where new units are recruited by other units to form part of the sample.



- useful way to conduct research about people with specific traits who might otherwise be difficult to identify (e.g., people with a rare disease)
- downside here is representativeness due to the reliance on participants recruiting others, which leads to sampling bias.

Snowball Sampling

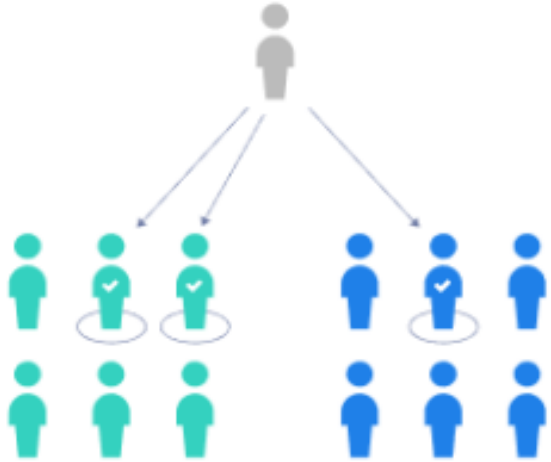
> **Example: Snowball sampling**
>
> You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.



- widely employed method in qualitative research, specifically when studying hard-to-reach populations. These may include:
  - Populations that are small relative to the general population
  - Geographically dispersed populations
  - Populations possessing a social stigma or particular shared points of interest
  - used to study sensitive topics, or topics that people may prefer not to discuss publicly

Quota Sampling

Quota sampling is a sampling method that relies on the non-random selection of a predetermined number or proportion of units.

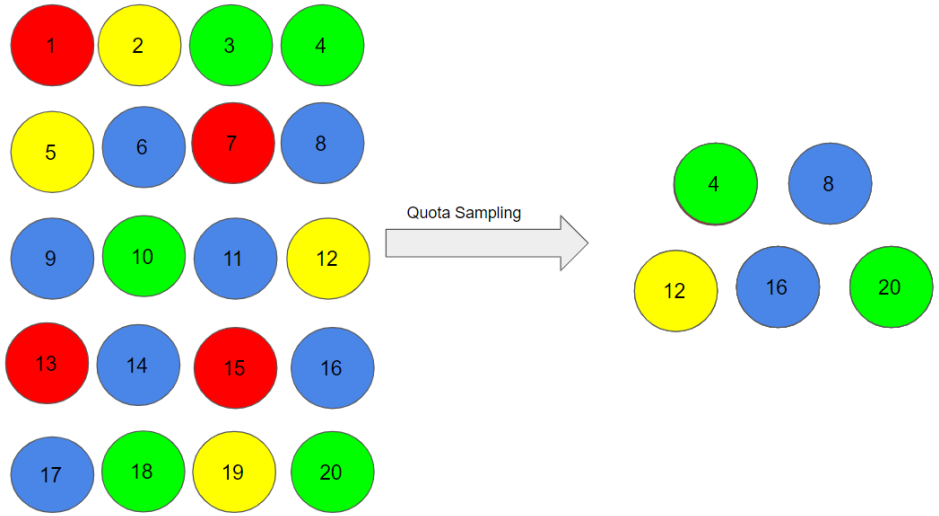


- It first divide the population into mutually exclusive subgroups (called strata) and then recruit sample units until you reach the quota.
- The aim of quota sampling is to control what or who makes up the sample.
- To gain insight about a characteristic of a particular subgroup or investigate relationships between different subgroups.

Quota Sampling

Example: Quota sampling

You want to gauge consumer interest in a new produce delivery service in Boston, focused on dietary preferences. You divide the population into meat eaters, vegetarians, and vegans, drawing a sample of 1000 people. Since the company wants to cater to all consumers, you set a quota of 200 people for each dietary group. In this way, all dietary preferences are equally represented in your research, and you can easily compare these groups.You continue recruiting until you reach the quota of 200 participants for each subgroup.



- It is most commonly used in research studies where there is no sampling frame available, since it can help researchers obtain a sample that is as representative (a broad picture) as possible of the population being studied.
- It cannot be generalized to the wider population and is at high risk for bias.

Differences

Difference between non-probability sampling and probability sampling:

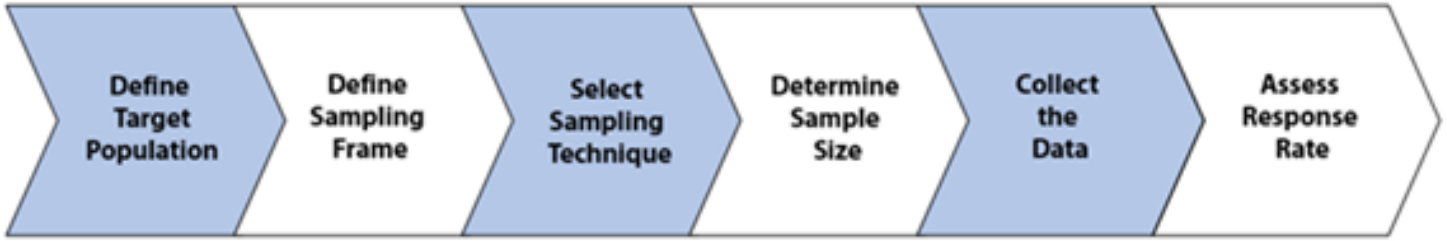
Non-probability sampling	Probability sampling
Sample selection based on the subjective judgment of the researcher.	The sample is selected at random.
Not everyone has an equal chance to participate.	Everyone in the population has an equal chance of getting selected.
The researcher does not consider sampling bias.	Used when sampling bias has to be reduced.
Useful when the population has similar traits.	Useful when the population is diverse.
The sample does not accurately represent the population.	Used to create an accurate sample.
Finding respondents is easy.	Finding the right respondents is not easy.

Sampling Methods to use

For any research, it is essential to choose a sampling method accurately to meet the goals of the study. The effectiveness of sampling relies on various factors. Here are some steps expert researchers follow to decide the best sampling method.

- Jot down the research goals. Generally, it must be a combination of cost, precision, or accuracy.
- Identify the effective sampling techniques that might potentially achieve the research goals.
- Test each of these methods and examine whether they help achieve the goal and select the method that works best for the research.
- Many statistical methods assume random sampling; however, it is often impractical to obtain random samples.
- Inferences to populations from nonrandom samples can be justified, but this depends on background information sufficient to determine that a sample is representative.
- Biases in sampling procedures can mislead.
- Conclusions from a study should be consistent with how the data was sampled.

Sampling Example

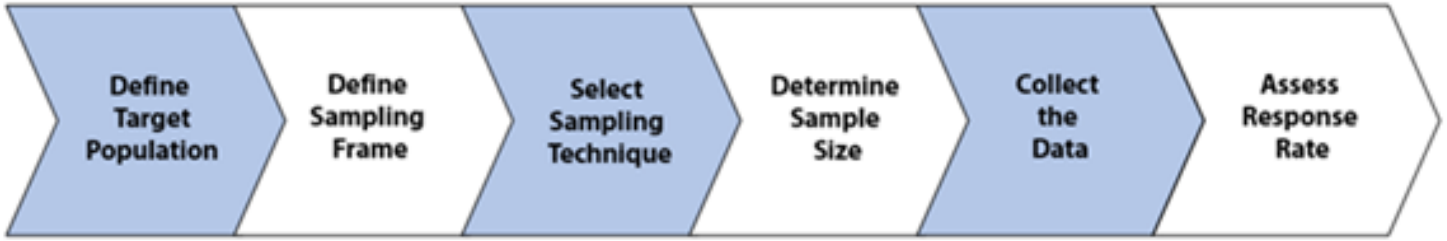


Sampling Process Flow

Let's take an interesting case study and apply these steps to perform sampling. Recently conducted General Elections in India a few months back. The public opinion polls every news channel was running at the time:

Were these results concluded by considering the views of all 900 million voters of the country or a fraction of these voters?

Sampling Example

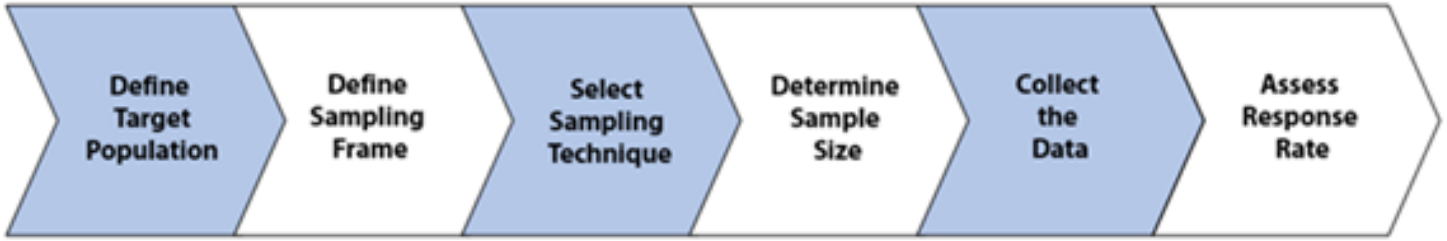


Sampling Process Flow

Step 1
The first stage in the sampling process is to clearly define the target population.
So, to carry out opinion polls, polling agencies consider only the people who are above 18 years of age and are eligible to vote in the population.

Step 2
Sampling Frame – It is a list of items or people forming a population from which the sample is taken.
So, the sampling frame would be the list of all the people whose names appear on the voter list of a constituency.

Sampling Example



Sampling Process Flow

Step 3

Generally, probability sampling methods are used because every vote has equal value and any person can be included in the sample irrespective of his caste, community, or religion. Different samples are taken from different regions all over the country.

Step 4

Sample Size – It is the number of individuals or items to be taken in a sample that would be enough to make inferences about the population with the desired level of accuracy and precision.Larger the sample size, more accurate our inference about the population would be.

For the polls, agencies try to get as many people as possible of diverse backgrounds to be included in the sample as it would help in predicting the number of seats a political party can win.

Review – Sampling methods

www.kahoot.it

Lab Time