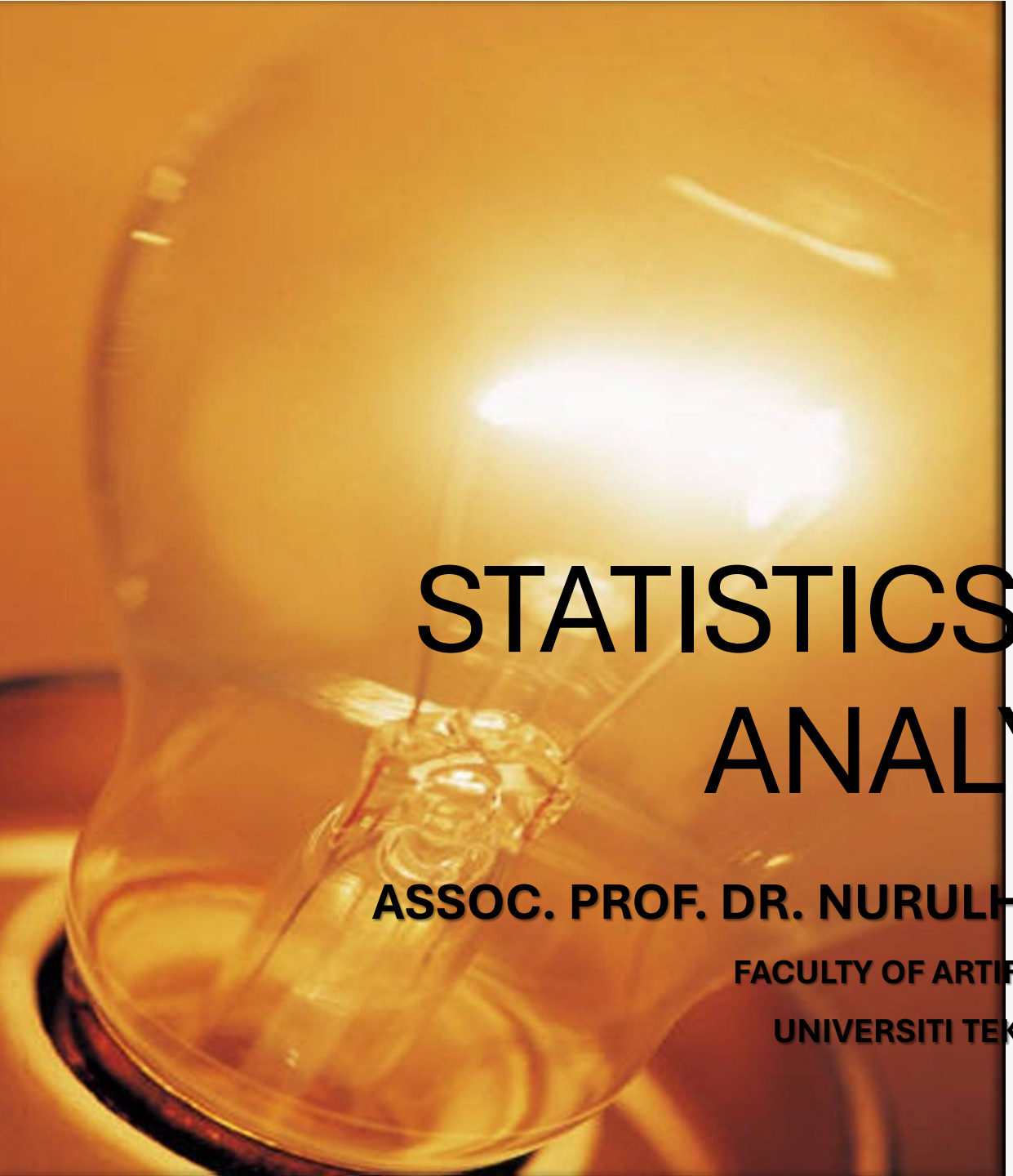


STATISTICS AND DATA ANALYTICS

ASSOC. PROF. DR. NURULHUDA FIRDAUS BT MOHD AZMI

FACULTY OF ARTIFICIAL INTELLIGENCE,

UNIVERSITI TEKNOLOGI MALAYSIA


DATA ANALYTICS

Introduction to Data Analytics

Definition: The process of analyzing raw data to find insights and trends.

Importance: Helps in making informed decisions and predictions.

Types of Analytics:

1. Descriptive (What happened?)
2. Diagnostic (Why did it happen?)
3. Predictive (What will happen?)
4. Prescriptive (What should be done?)

Key Goals: Identify patterns, trends, and relationships in data.

Basics of Data Analysis

1. Formatting and Organization of Data:
   - Structured - Highly organized and formatted, making it easy to search and analyze (e.g., databases, spreadsheets)
   - Unstructured - Lacks a predefined format, requiring specialized tools for processing (e.g., text, images, audios, videos).
2. Data Type:
   - Quantitative Data
   - Qualitative Data
3. Data Hierarchy (from lowest to highest): Nominal → Ordinal → Ratio/Interval
4. Analysis does not mean using computer software package
5. Analysis is looking at the data considering the questions you need to answer:
   - For example: How would you analyze patient outcomes to determine whether a healthcare program is effectively improving patient care and meeting its health objectives?"

Question:
Is my healthcare program improving patient outcomes?
Analysis:
Compare program targets (e.g., patient recovery rates, hospital re-admission rates) with actual performance to assess progress.

Interpretation:
Why have you or have you not achieved the healthcare targets? What does this mean for patient care and treatment effectiveness?
May require additional patient data and healthcare metrics for deeper insights.

BASIC UNDERSTANDING ABOUT STATISTICAL DATA ANALYSIS

NEED TO KNOW....

- How to establish statistical data analysis?
- What are the statistical branches?
- Does mathematics and statistics same?
- Does data and variable reflect the same meaning?
- Do data hierarchy important?
- Which do I need to choose, population or sample?
- How to determine the sampling size is enough?

STATISTICS

1. Statistical analysis is the science of collecting data and uncovering patterns and trends.

2. Conducting statistical data analysis is rely on statistical assumption.

3. Statistics comes from the branch of Mathematics

8

STATISTICS IN HEALTHCARE DOMAIN

- In healthcare, statistics helps in **clinical decision-making, public health policy, and medical research**.

- Importance of Statistics in Healthcare:
  - **Evidence-Based Practice** – Supports clinical guidelines and treatment plans.
  - **Epidemiology** – Tracks disease patterns and outbreaks.
  - **Clinical Trials** – Evaluates the effectiveness of new drugs and treatments.
  - **Pharmacovigilance** – Monitors drug safety and adverse effects.

9

MATHEMATICS VERSUS STATISTICS?

MATHEMATICS:

- When the problem is solved correctly, the result can be reported as 100% certainty.
- Mathematic problem: Mary and Jane is asked to solve the value of x given 3x+5=11

STATISTICS:

- When the problem is solved, the result do not have 100% certainty. We might say, for example, we are 95% confidence that the average.....
- Mary and Jane are asked to **estimate the average recovery time for patients after a minor surgery** at a local hospital.
- To do this, they collect data on the **number of days patients take to fully recover** after undergoing a **laparoscopic appendix removal**.
- Since recovery time can vary based on **age, pre-existing conditions, and lifestyle factors**, their analysis will help in understanding the **average expected recovery period** for different patient groups.

THE PROCESS OF STATISTICS

Identify the research objectives.

- Determine the question(s) to answer
- Identify the population that is to be studied.

Collect the data needed.

- Look for sample.

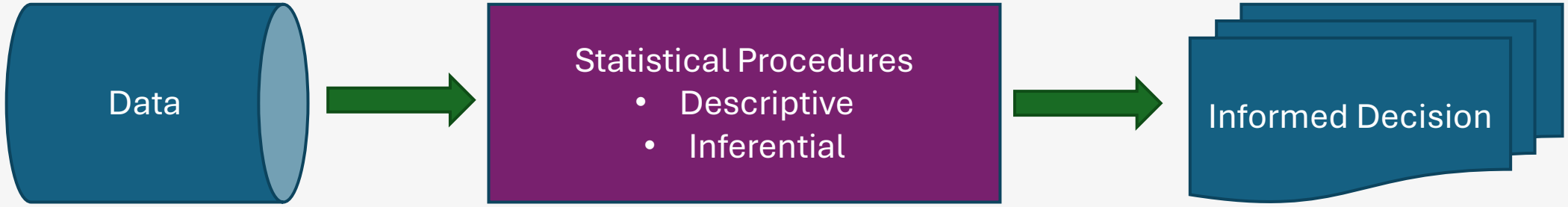
Describe the data.

- Data editing, pre-processing.
- Exploring and summarize the data.

Perform Inference.

- Apply the appropriate techniques to extend the results obtained from the sample to the population and report a level of reliability of the results.

THE BRANCH OF STATISTICS

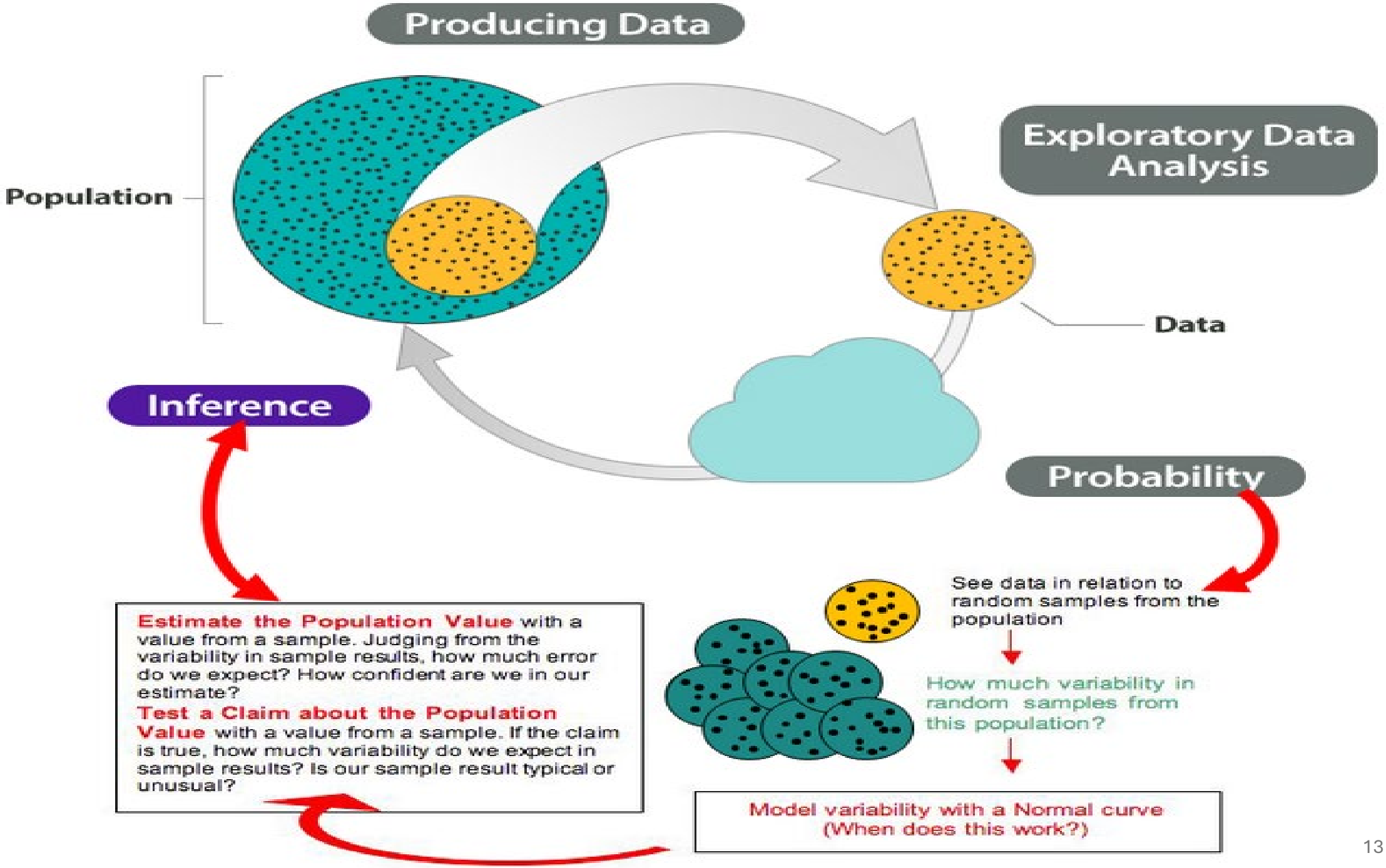


Descriptive Statistics:

- Consists of exploring and summarizing data.
- Describe about the data either through visual (chart or graph) and/or numerical measures.

Inferential Statistics:

- Uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.
- Making estimation & hypothesis testing to help in decision making process.
- Include a level of a confidence in the result since sample cannot tells everything about a population.


Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability
Estimate the Population Value with a value from a sample. Judging from the variability in sample results, how much error do we expect? How confident are we in our estimate?
Test a Claim about the Population Value with a value from a sample. If the claim is true, how much variability do we expect in sample results? Is our sample result typical or unusual?
See data in relation to random samples from the population
How much variability in random samples from this population?
Model variability with a Normal curve
(When does this work?)

DATA VERSUS VARIABLE

VARIABLE

- A characteristic of an item or an individual that will be analyzed using statistics.
- Should be place as columns in the record sheet.
- A hospital is conducting a study on **patients with hypertension** to analyze factors that may influence their blood pressure levels. The research team collects various data points from patients during their routine check-ups.
- **Individuals in the study**: Patients with hypertension
- **Variables**:
    - **Systolic and Diastolic Blood Pressure Levels** (measured in mmHg) – *Numerical (Continuous)*
    - **Patient's BMI Category** (Underweight, Normal, Overweight, Obese) – *Categorical (Ordinal)*
    - **Medication Usage** (Yes/No) – *Categorical (Nominal)*
    - **Daily Sodium Intake** (mg/day) – *Numerical (Continuous)*

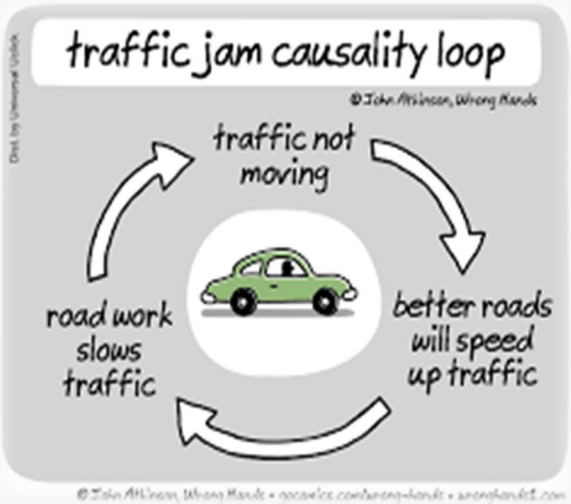
DATA

- Observations that is recorded.
- Should be place as rows in the record sheet.
- Varies (same weight? eat same amount of food everyday?)
    - variability in data may help to explain different result obtained
- Misused of data occurs when the data are incorrectly obtained (where the data comes from) or incorrectly analyzed/presented

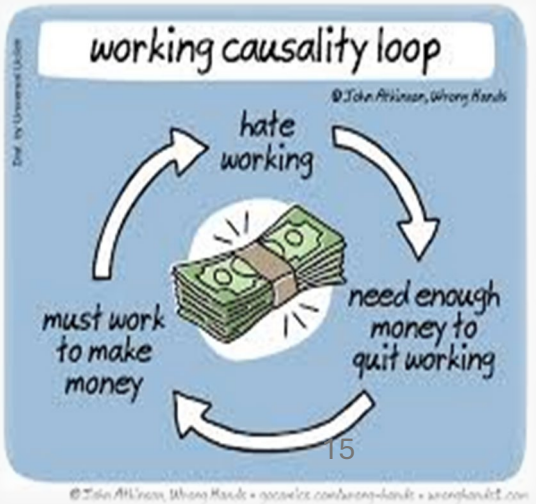
To understand causality:

- Independent variable (IV) – represent inputs or causes

- Dependent variable (DV) – the output or outcome whose variation is being studied

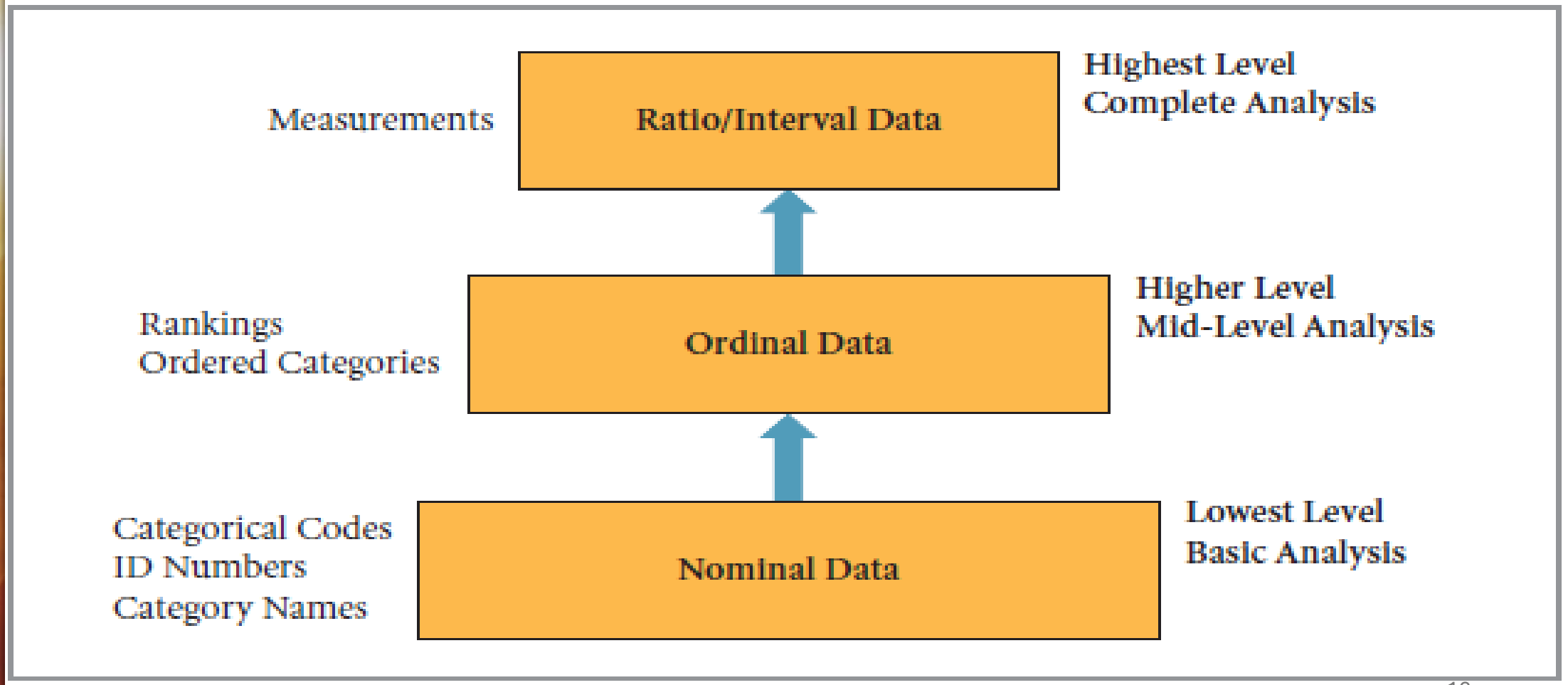


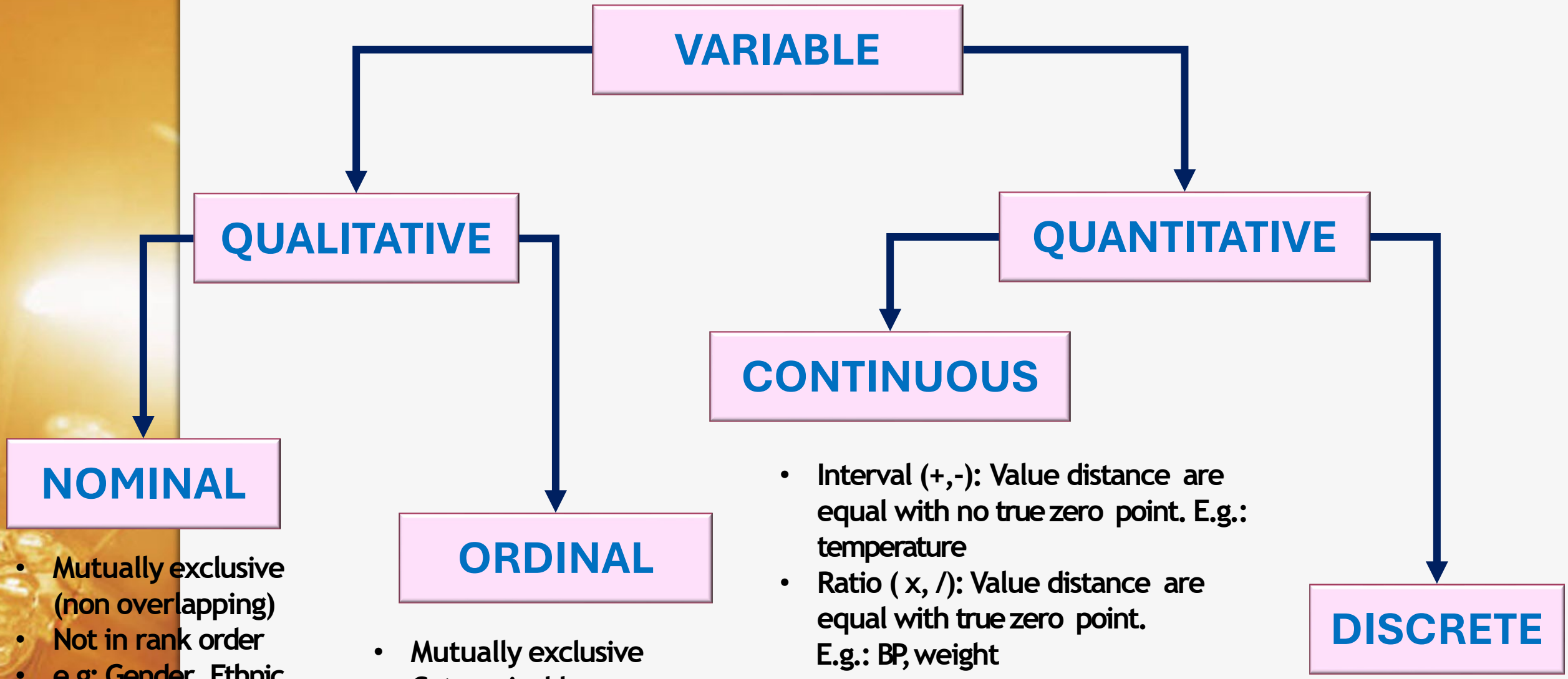
Does Correlation same with Causation?



DO DATA HIERARCHY IMPORTANT?

To analyze the data, it is dependent on the level and type of data you have available





** Where does Interval and Ratio data categorize?

DATA-HIERARCHY ISSUES

- A **hospital** recently conducted a survey to determine the **blood type distribution** among its patients. The data were coded as follows:
  - **1 = Type A**
  - **2 = Type B**
  - **3 = Type AB**
  - **4 = Type O**
- 17 responses from patient collected were: response = {1,1,4,2,1,2,2,2,3,1,1,1,3,2,2,1,2}
- If the **mean** is calculated, then the **sample mean, $\bar{x}$ = 1.82**
- This result is **nonsensical**, as it suggests a blood type that does not exist between **Type A and Type B**. Since blood type is **categorical (nominal data)**, calculating a mean is **meaningless** and inappropriate. Instead, we should use **mode** (most frequent category) or **proportions** to summarize the data.

DATA-HIERARCHY ISSUES

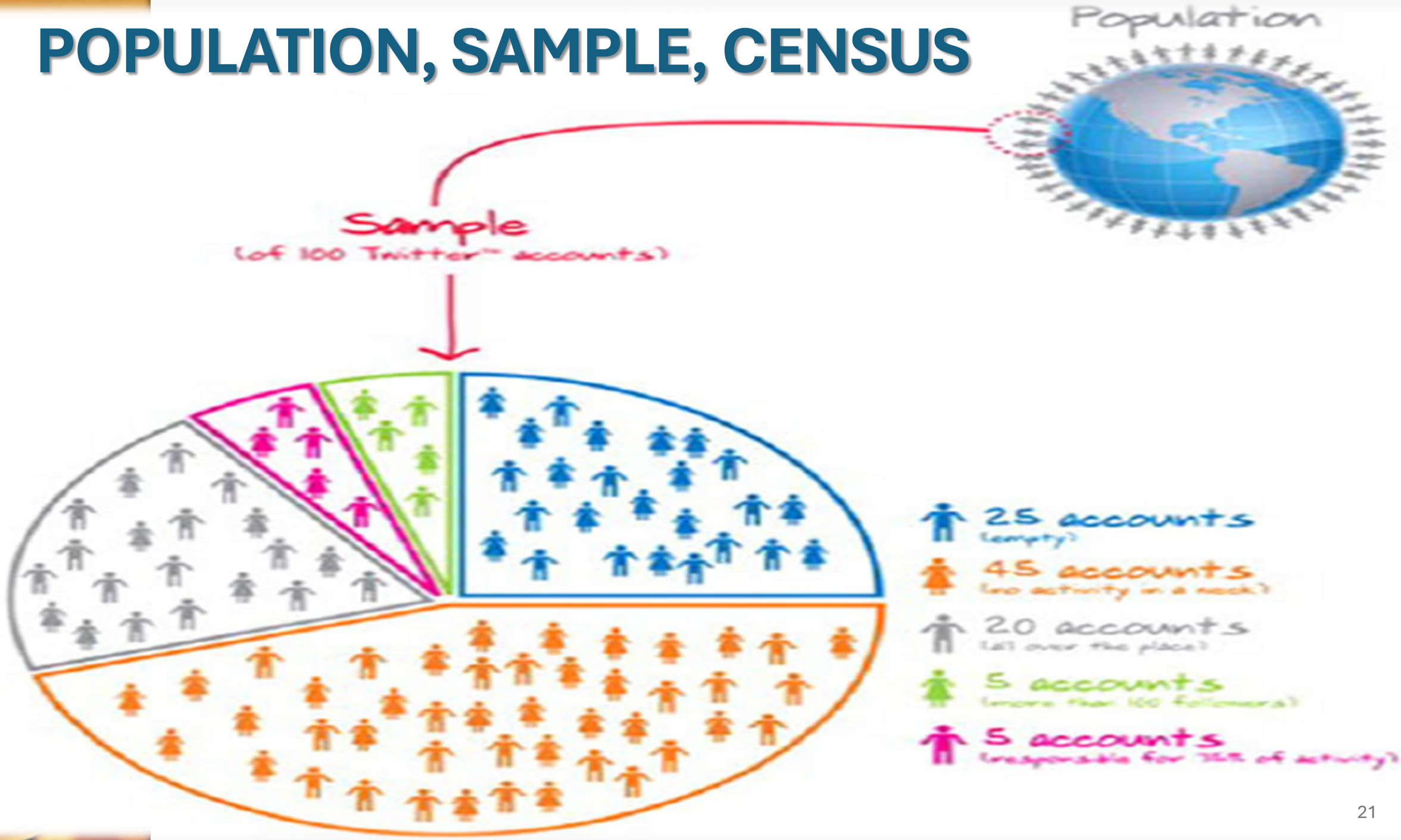
- Common mistakes to compute means on ordinal-level data.
- In a **hospital patient satisfaction survey**, a **5-point scale** is often used to measure **patients' pain levels** after a medical procedure. The scale is defined as follows:
  - **1 = No Pain**
  - **2 = Mild Pain**
  - **3 = Moderate Pain**
  - **4 = Severe Pain**
  - **5 = Extreme Pain**
- The responses from **10 patients** are recorded as: response = {2,2,1,3,3,1,5,2,1,3}
- If the **mean** is calculated, then the **sample mean pain score, $\bar{x}$ = 2.3**
- This result is **misleading**, as it suggests a **precise pain level between "Mild Pain" and "Moderate Pain"**, when pain **is subjective and ordinal**. Since **pain levels are ranked and do not have equal intervals**, calculating a mean is **inappropriate**. Instead, we should use the **median or mode** to summarize the responses accurately.

DATA COLLECTION

Methods of Data Collection

- **Primary Data** – Collected directly from patients (e.g., surveys, laboratory tests).
- **Secondary Data** – Retrieved from records (e.g., hospital databases, clinical reports).

POPULATION, SAMPLE, CENSUS



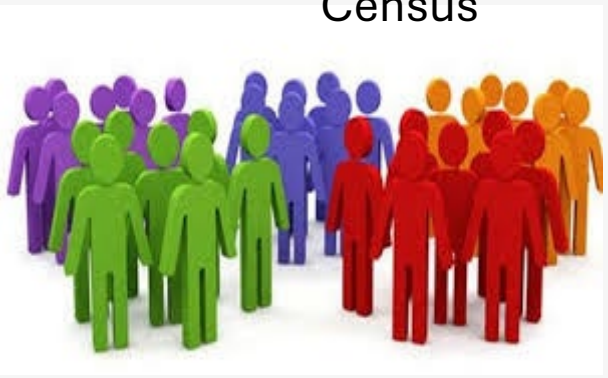
THE DIFFERENCES

- **Population** – the entire group to be studied.
- **Sample** – a subset of the population that is being studied.
- **Individual** – a person or object that is a member of the population being studied.
- **Census** – a list of all individuals in a population along with the characteristics of everyone.

Population

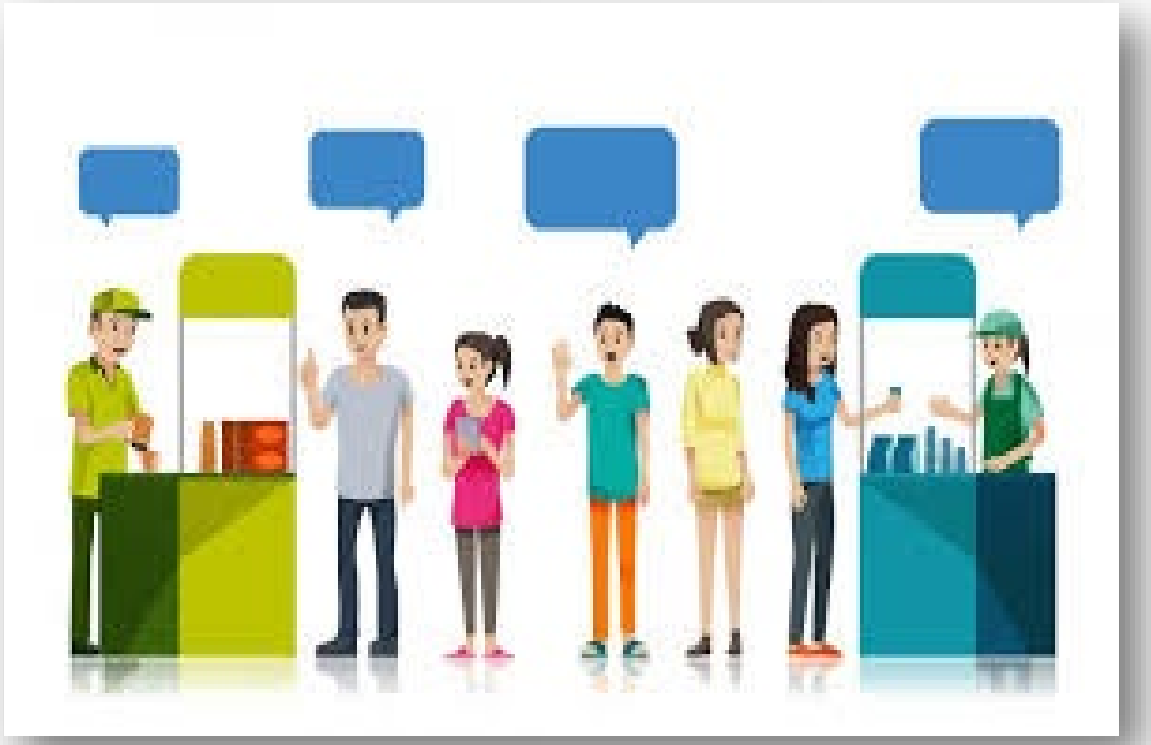
Sample

Census



Individual

NON-PROBABILITY (NON-STATISTICAL) SAMPLING



SAMPLING

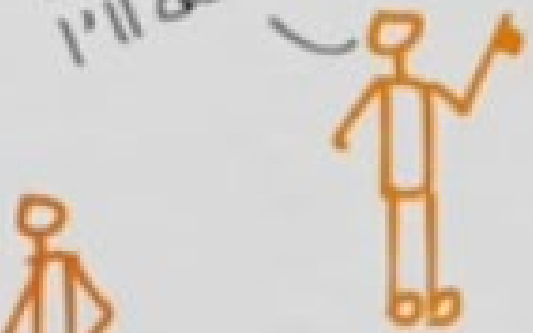
NON-PROBABILITY SAMPLING

Purposive Sampling

Convenience (Accidental)

4pm-7pm

Voluntary Sampling

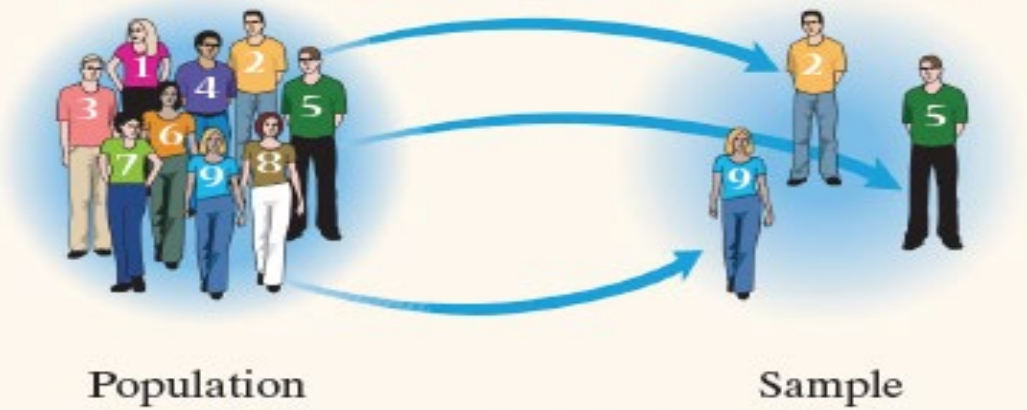


Snowball Sampling

PROBABILITY (STATISTICAL) SAMPLING

Simple Random Sampling

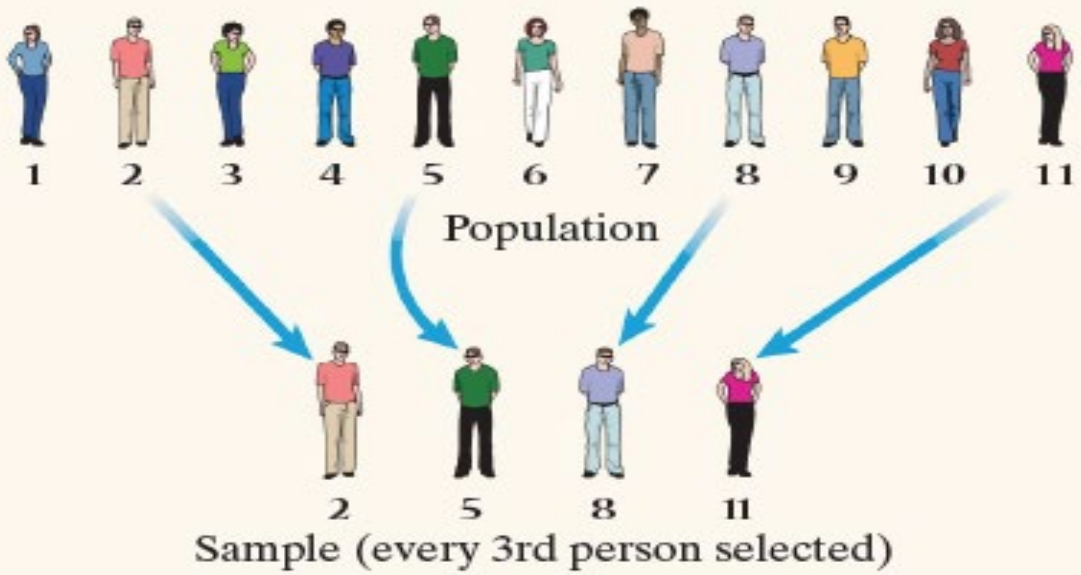
https://youtu.be/yx5KZi5QArQ



Stratified Sampling

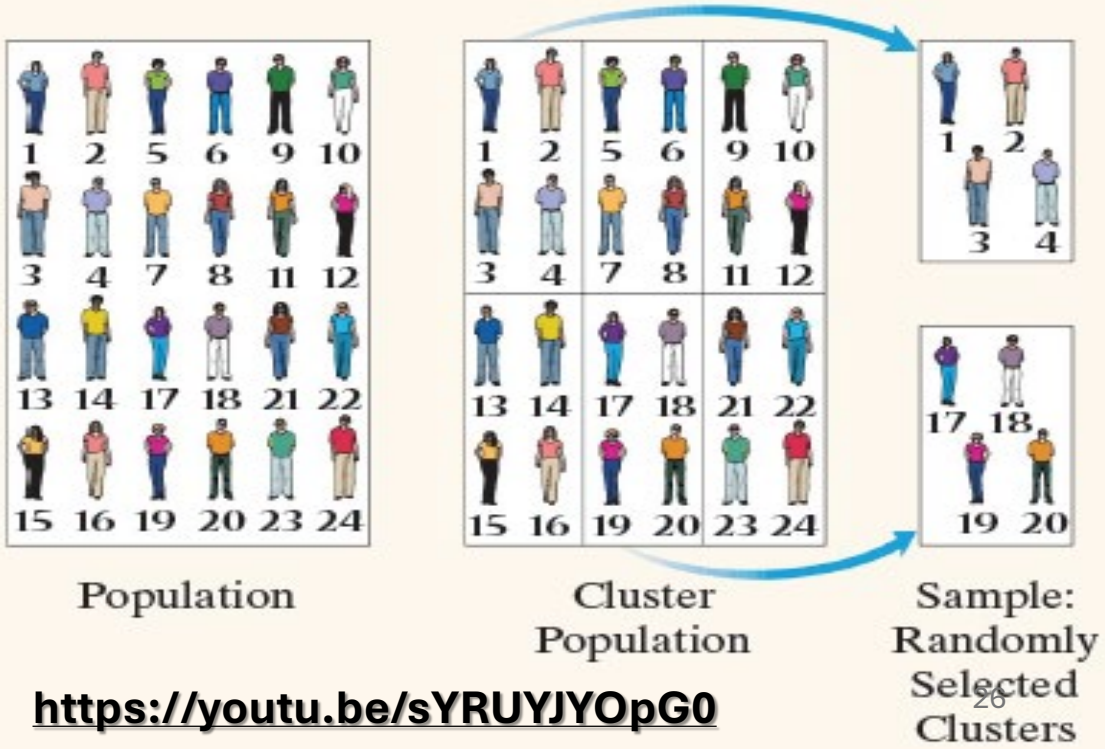
https://youtu.be/sYRUYJYOpG0

Systematic Sampling



https://youtu.be/QFoisfSZs8I

Cluster Sampling



https://youtu.be/sYRUYJYOpG0


DECIDING ON
SAMPLE SIZE

SAMPLE SIZE CONSIDERATION: KEY FACTORS & BEST PRACTICES

1. **Determining the Sample Size**
   - Researchers must calculate the **appropriate sample size** to ensure representativeness.
   - Key factors include **confidence level, margin of error, population size, variability**.
2. **Balancing Reliability & Cost**
   - **Larger samples** increase accuracy but raise costs and require more time.
   - Find the **minimum viable sample size** that maintains statistical power.
3. **Confidence Level & Margin of Error**
   - **Higher confidence levels** (95% or 99%) require larger samples to reduce uncertainty.
   - **Smaller margin of error** (e.g., ±3%) needs a larger sample for precision.
4. **Impact of Population Size**
   - **Small populations** need a larger proportion sampled.
   - **Large populations** require a well-calculated sample using formulas or online calculators.
5. **Adjusting for Non-Response & Variability**
   - Anticipate non-response rates by **oversampling**.
   - **High population variability** requires a larger sample to detect differences.

SAMPLE SIZE CALCULATION

- To estimate the sample size, the decision maker need to specify their confidence level and their desired margin error, *e*
- For population mean:

$$n = \frac{z^2\sigma^2}{e^2} = \left(\frac{z\,\sigma}{e}\right)^2$$

- For population proportion:

$$n = \frac{z^2 P(1-P)}{e^2}$$

Where,

n = sample size
z = critical value for specified confidence level
e = desired margin error
σ = population standard deviation (can be known or unknown)
P = population proportion

SAMPLE SIZE ONLINE CALCULATOR

- There are many sample size calculators online. Different formulas are used depending on whether you have subgroups or how rigorous your study should be (e.g., in clinical research).

- As a rule of thumb, a **minimum of 30 units** or more per subgroup is necessary.

FURTHER JOURNEY IN STATISTICAL DATA ANALYSIS…

- Once you've collected all your data, you can inspect them and calculate **descriptive statistics** that summarize them.
- A number that describes a sample is called a **statistic**, while a number describing a population is called a **parameter**. Using **inferential statistics**, you can make conclusions about population parameters based on sample statistics.


YOUR DATA IS READY....
LET'S BEGIN OUR STATISTICAL DATA ANALYSIS