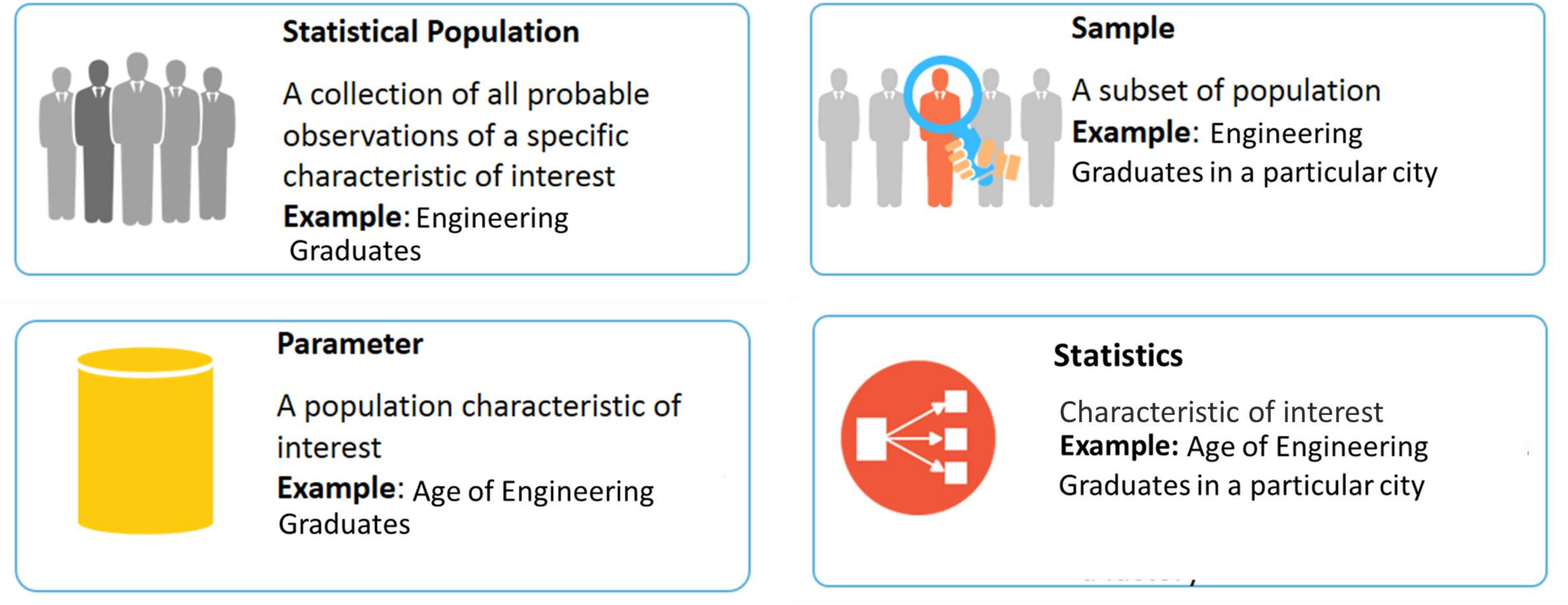
Statistics

Why Statistics?

- Data are everywhere
- Statistical techniques are used to make many decisions that affect our lives
- No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions effectively

Population and Sample

- A population is the set of all measurements of interest to the study.
- A sample is a selected subset of measurements of a population to represent the population.



Statistical Population

A collection of all probable observations of a specific characteristic of interest

Example: Engineering Graduates

Sample

A subset of population

Example: Engineering Graduates in a particular city

Parameter

A population characteristic of interest

Example: Age of Engineering Graduates

Statistics

Characteristic of interest

Example: Age of Engineering Graduates in a particular city

Population and Sample

- **Market Share of a Product**
  - For example you need to estimate the market share of a detergent product specifically, say, Tide
  - Population here is the entire population
  - Sample is the a set of Supermarkets/shops
  - Market Share is calculated on the sample, not the population

Sources of Data

- **Primary Data**
  - Surveys
    - Mail: Lowest rate of response, usually the lowest cost
    - Web: Faster response and inexpensive
    - Telephone: Fastest response
    - Personal Interview: Usually focus groups. Most costly. Interviewer effects can be seen
- **Secondary Data**
  - This is the data that has been compiled or published elsewhere
  - Example: Census Data
  - Advantages: It can be gathered quickly and inexpensively
  - Disadvantages: May be outdated. May not be accurate

Errors

- **Response Errors**
  - Subject lies
  - Subject makes a mistake
  - Interviewer makes a mistake
  - Interviewer effects
- **Non Response Errors**
  - If the rate of response is low, then the sample is not representative
  - Might get a biased view of the population

Which is better?

Sample 1

- N = 2000
- Response rate = 90%

Sample 2

- N = 1,000,000
- Response rate = 20%

Which is better?

- Small but representative sample can be useful in making inferences
- A large sample which is unrepresentative, which makes them biased, is useless. There is no way to correct for it
- Therefore, sample 1 is better than sample 2

Types of Data

Types of Data

Categorical Data

- This refers to data that can be classified into separate groups.
- It is also called qualitative data.
- This data represents characteristics.
- For example, gender of a person can be male or female. It can also have numerical values like 1 for male and 0 for female.
- Categorical data can be further classified as nominal or ordinal.

Types of Data

Numerical Data

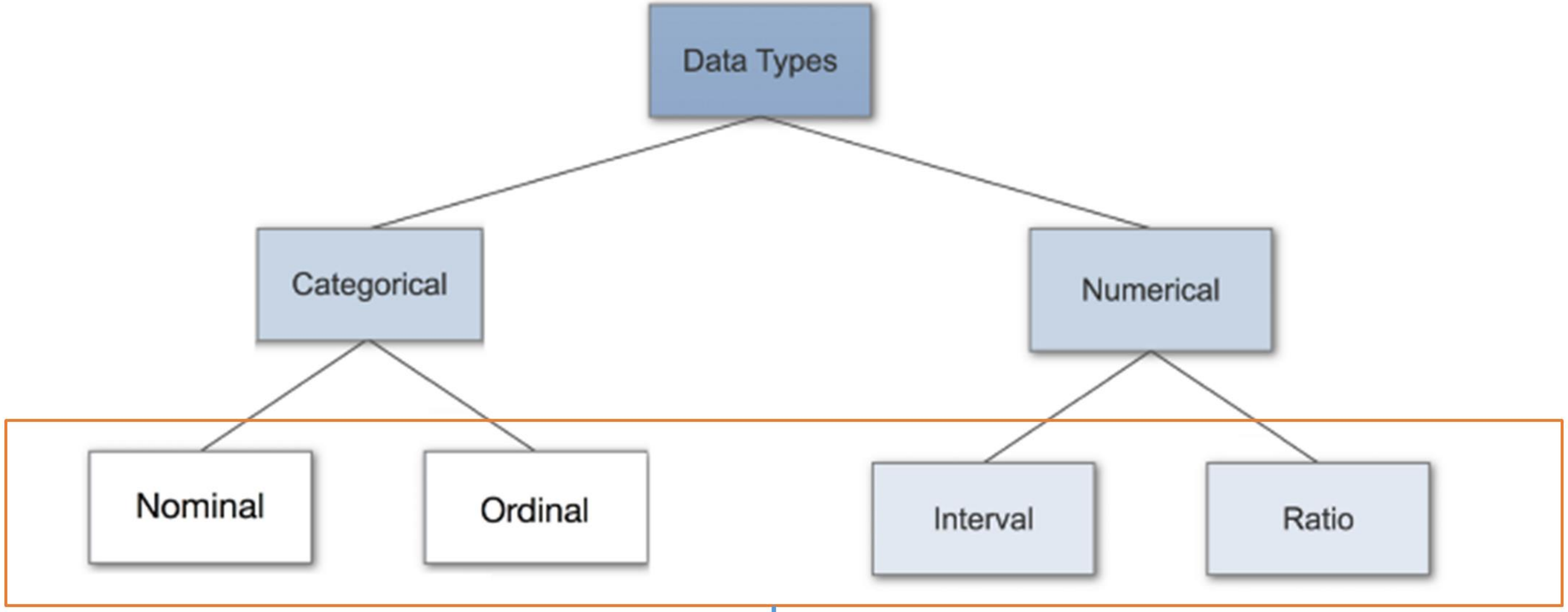
- Data that can be measured is called numerical data.
- It is also called quantitative data.
- Discrete Data:
  - If the values can be clearly separated from each other, then it is discrete data.
  - **Example:** Number of children
- Continuous data
  - **Example:** height of a person

Types of Data

Numerical Data

- One simple way to check if the data is continuous or discrete is to check whether if we can add more decimal points to the data
    - You might say you are 5’11’’ tall. But in actuality you may be 5’11.23432” tall
    - If you say you have 2 children, you cannot have 2.234545 children

Types of Data



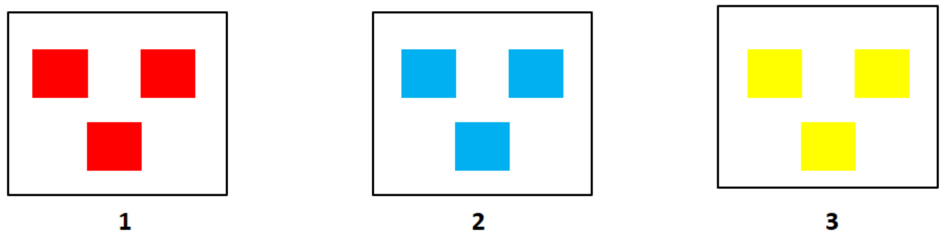
These are also referred to as scales of measurement.

Scales of Measurement

Scales of Measurement - Nominal

- **All we can say is that one is different from each other**
  - Gender: Male, Female, Transgender
  - Eye color: Blue, Green, Brown, Hazel
  - Type of house: Bungalow, Duplex, Ranch
  - Type of pet: Dog, Cat, Rodent, Fish, Bird

Example: Numbers are used to denote different colored items in a data set.



In this measurement:

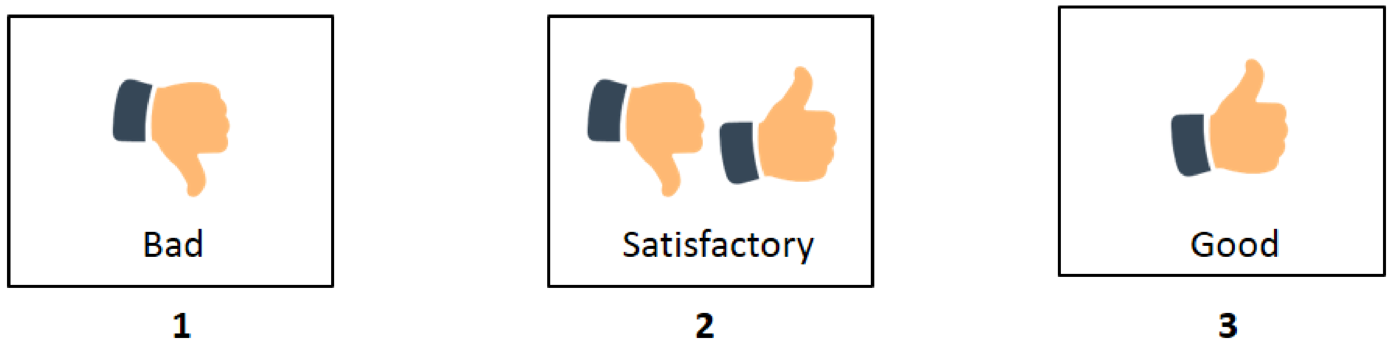
- Numbers are used to label qualitative items (not quantitative data) in classes or groups.
- The main purpose is to categorize items.
- Values cannot be altered in a numerical manner.
- Arithmetic operations are not possible.

Scales of Measurement - Ordinal

- **Ordinal scale of measurement refers to ordered series of relationships or rank order.**
- The ordinal scale contains data that can be placed in order.
- Ordinal scales do not represent a measurable quantity. It is **difficult to measure the interval between the values**.

  - High school class rankings: 1st, 2nd, 3rd, etc
  - Social economic class: working, middle, upper
  - The Likert Scale: agree, strongly agree, disagree

Numbers are used to denote customer preferences for a product.

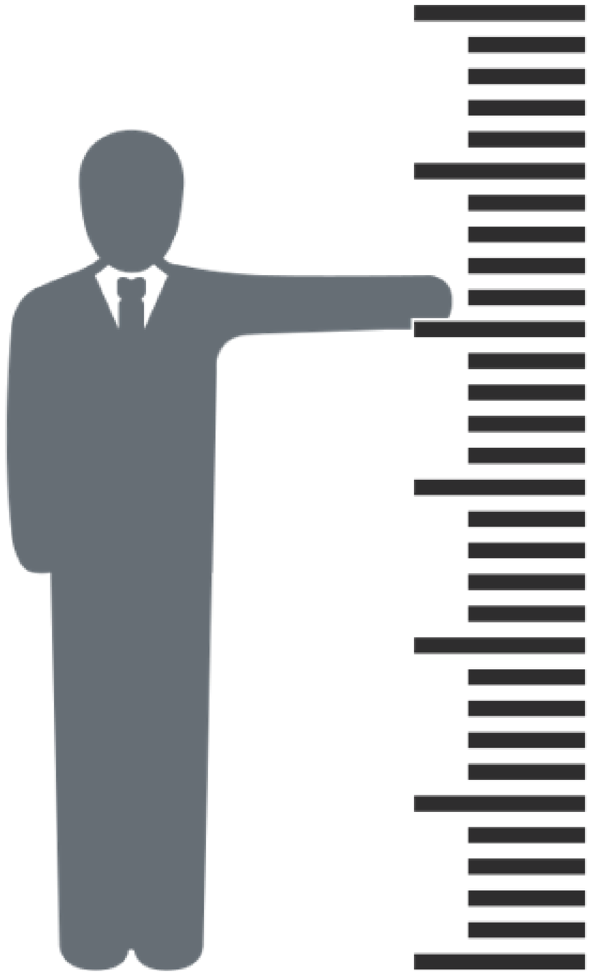


Scales of Measurement - Interval

- **An interval scale has measurements where the difference between the values are equal.**
- The Fahrenheit scale - the difference between 100 degrees and 80 degrees is the same difference as between 50 degrees and 70 degrees.
- Interval variables do not have a **meaningful zero-point**.
- For example, zero degrees does not mean that there is no temperature at all.
- If the zero becomes meaningful, then it is termed ratio scale.

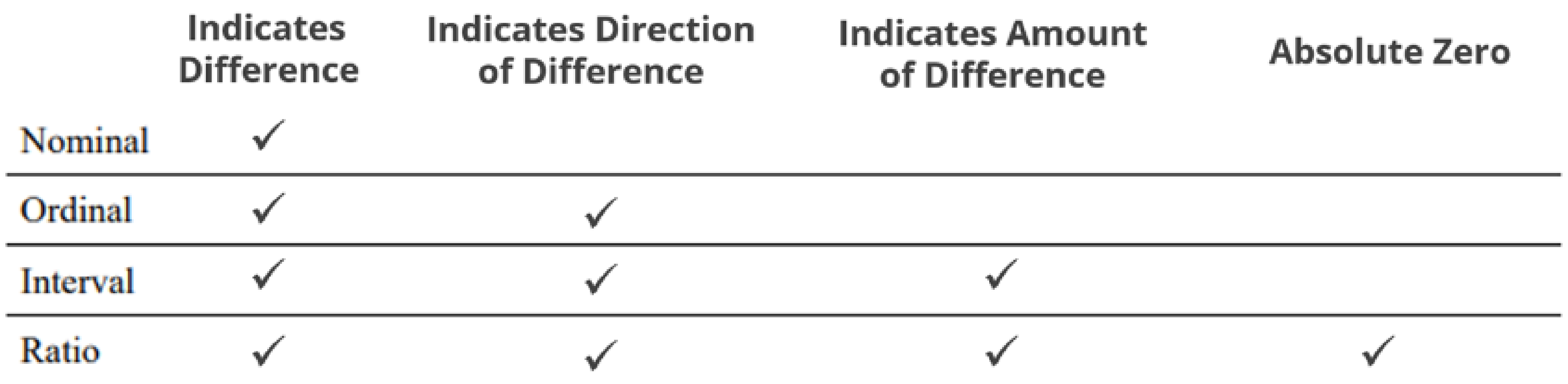
Scales of Measurement - Ratio

- In ratio scale, zero is meaningful.
- In this scale, no numbers below zero exist, i.e., it has absolute zero.
- Arithmetic operations can be performed on a ratio scale.
- If the length of a piece of cloth is measured in inches, then the measurement cannot become zero or less than that. A negative length is not possible.



Scales of Measurement

- The differences between the four scales of measurement can be easily understood from the table:



	Indicates Difference	Indicates Direction of Difference	Indicates Amount of Difference	Absolute Zero
Nominal	✓			
Ordinal	✓	✓		
Interval	✓	✓	✓	
Ratio	✓	✓	✓	✓

- It is clear from the table that ratio scale satisfies all the four properties of scales of measurements