Hodeidah University Mathematics Department

Lectures in advanced biostatistics for medical students

5-Level

Prof. Dr. Salem Saleh

2024

Lecture 1, Review of Principle Concepts in Biostatistics (1)

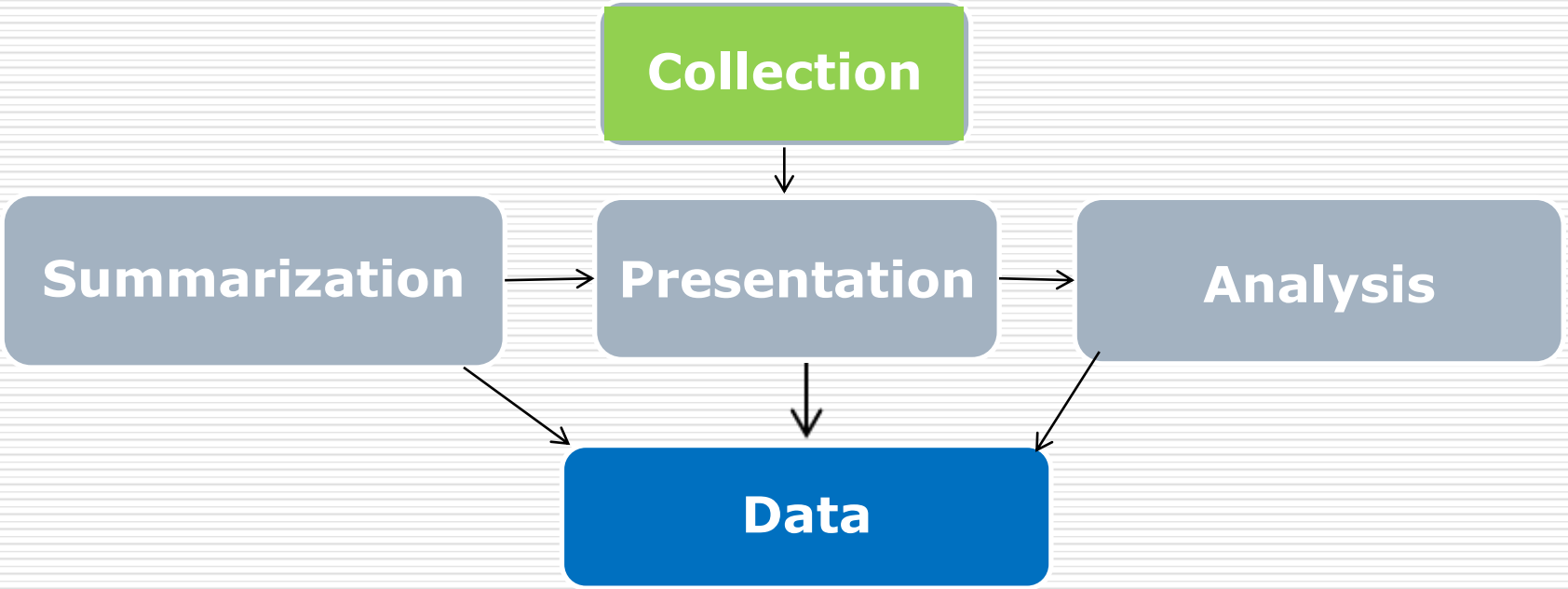
- ✓ What is statistics- Biostatistics, their sections and fields?
- ✓ Data and Variables with their types.
- ✓ Collection Data Methods.
- ✓ Measurement Levels
- ✓ Population and Sample with their types.
- ✓ Some statistical data analysis software.

Statistics and Biostatistics

- Statistics: is the science that interest with collection, summarization, presentation, analysis, and illustration the data for obtain acceptable scientific results, to make a suitable decisions.
- Biostatistics: Biostatistics is a branch of statistics that applies statistical methods to biological, health, and medical researches.

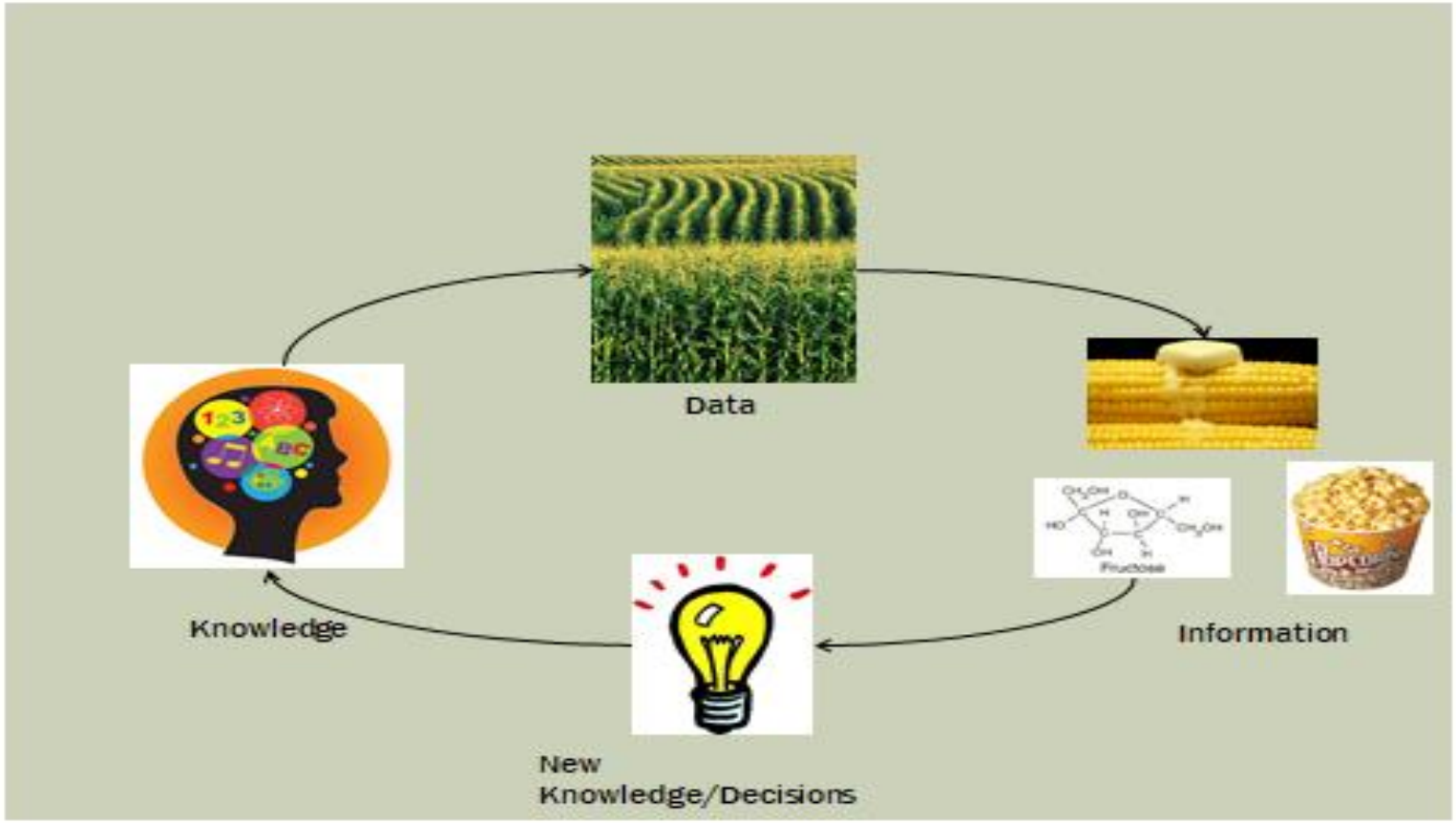
Basics of Statistics

Statistics: is the science concerned with:



In order to obtain acceptable scientific results,
to make appropriate decisions.

Knowledge, Data, Information, Analyze, Decisions

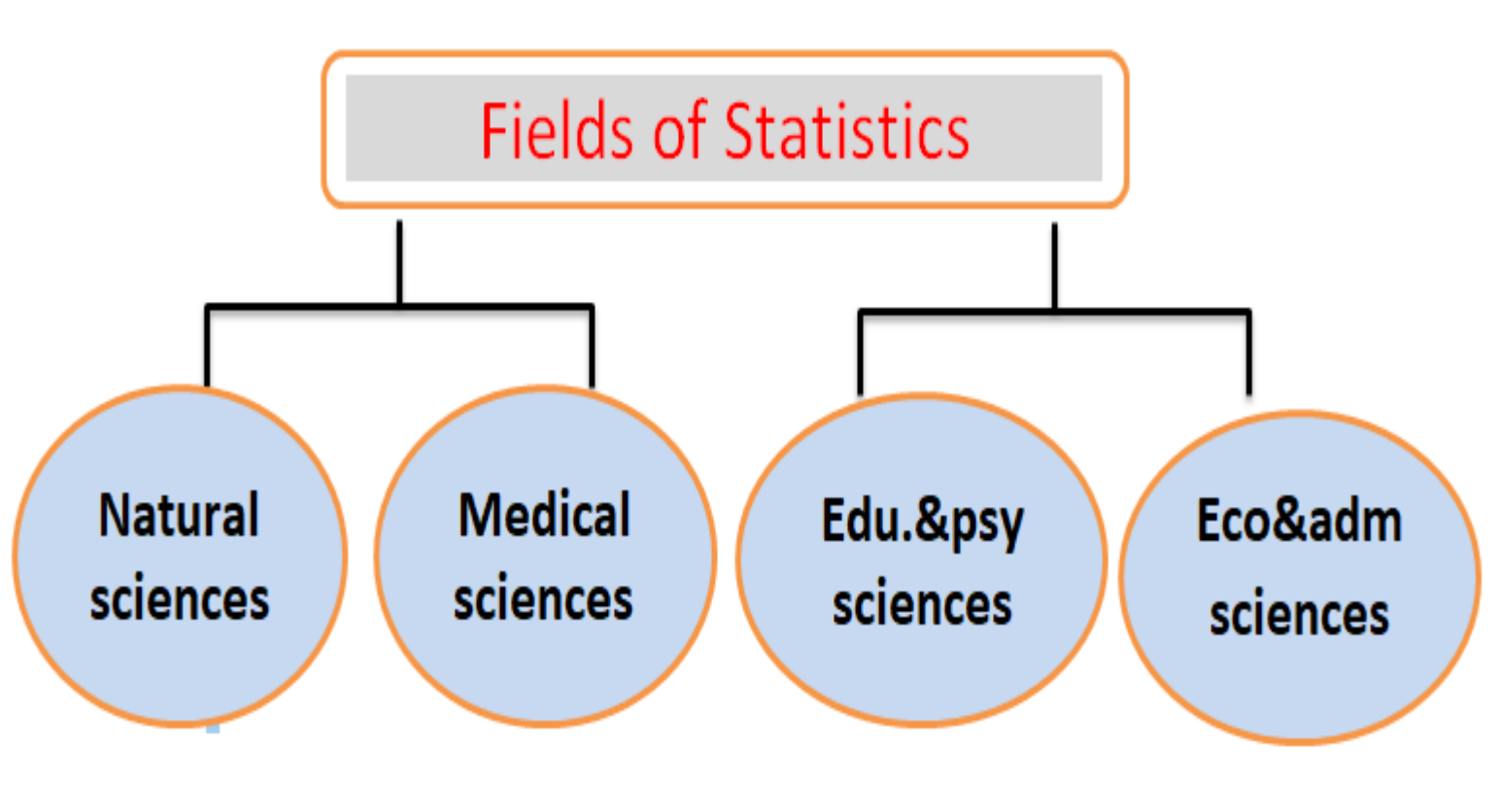


importance of statistics

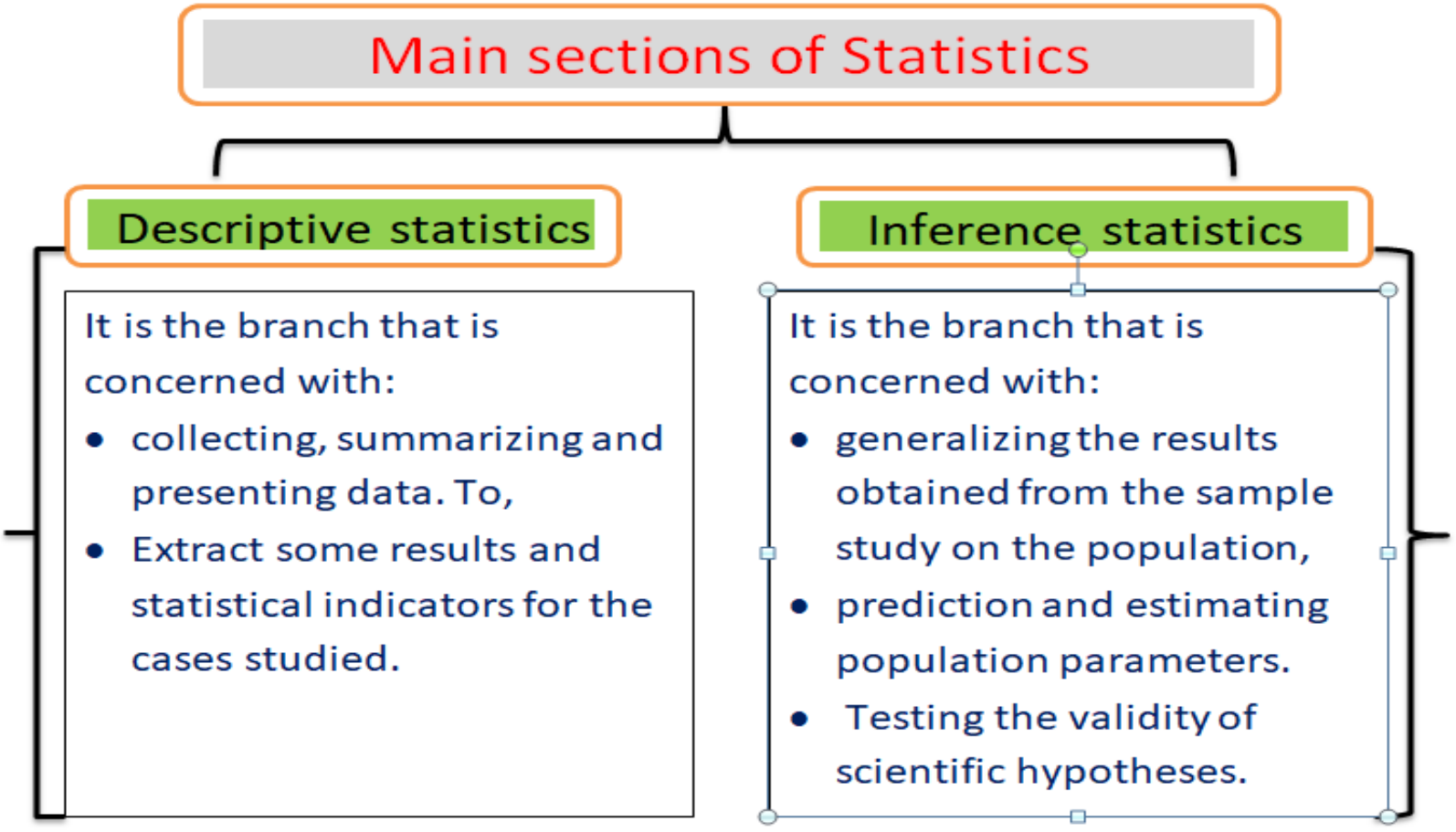
The importance of statistics is can be summarized in the following points:

- ✓ Statistics is the main tool in scientific research.
- ✓ It is considered one of the important tools for drawing strategic policies for development.
- ✓ It is used in various areas of life on the personal and institutional level in managing life affairs.

Fields of statistics



Main sections of Statistics



Variables and Data

Variable:
is a characteristic or attribute that can have different values or attributes, such as age, gender, level education, Marital status, Blood groups.

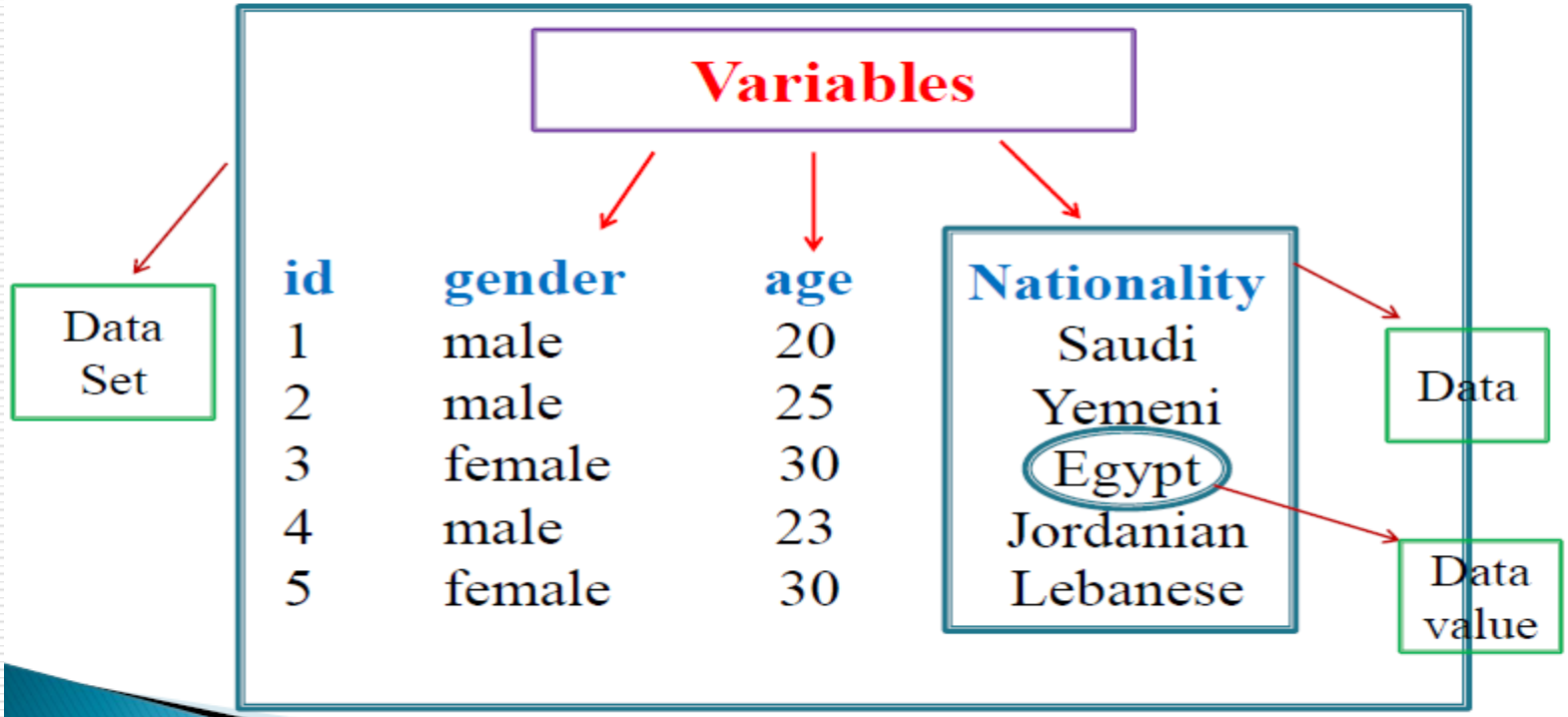
- The variables whose values are determined by chance they are called random variables.

Data:
is the values that a variable can take it.

- A collection of data values forms a data set.

Variables and Data

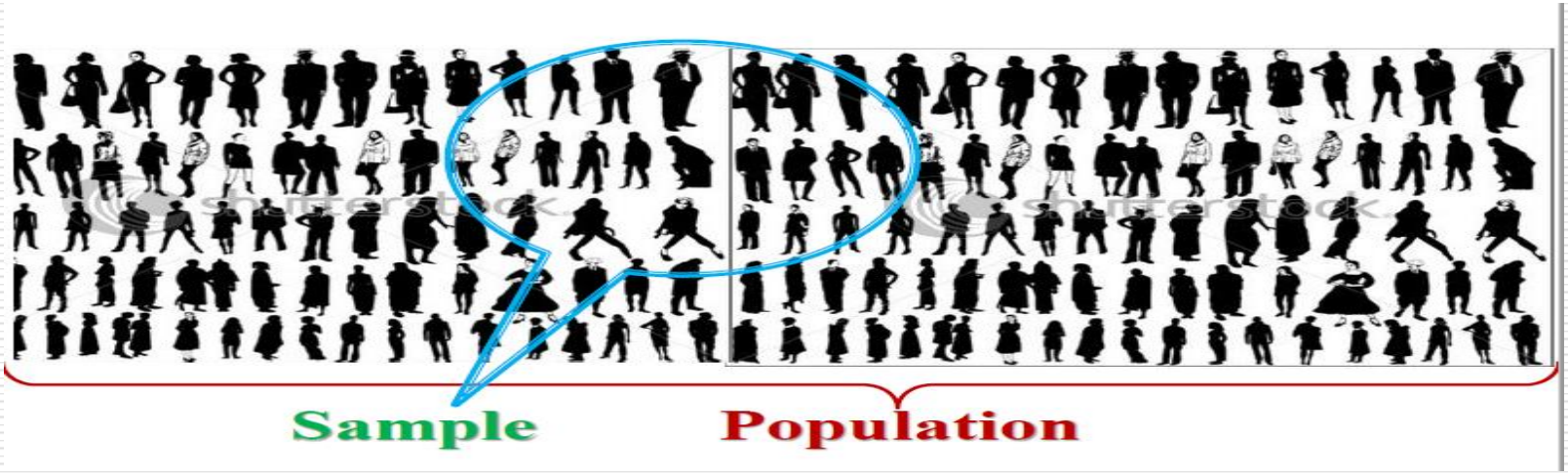
For example: A data set in table



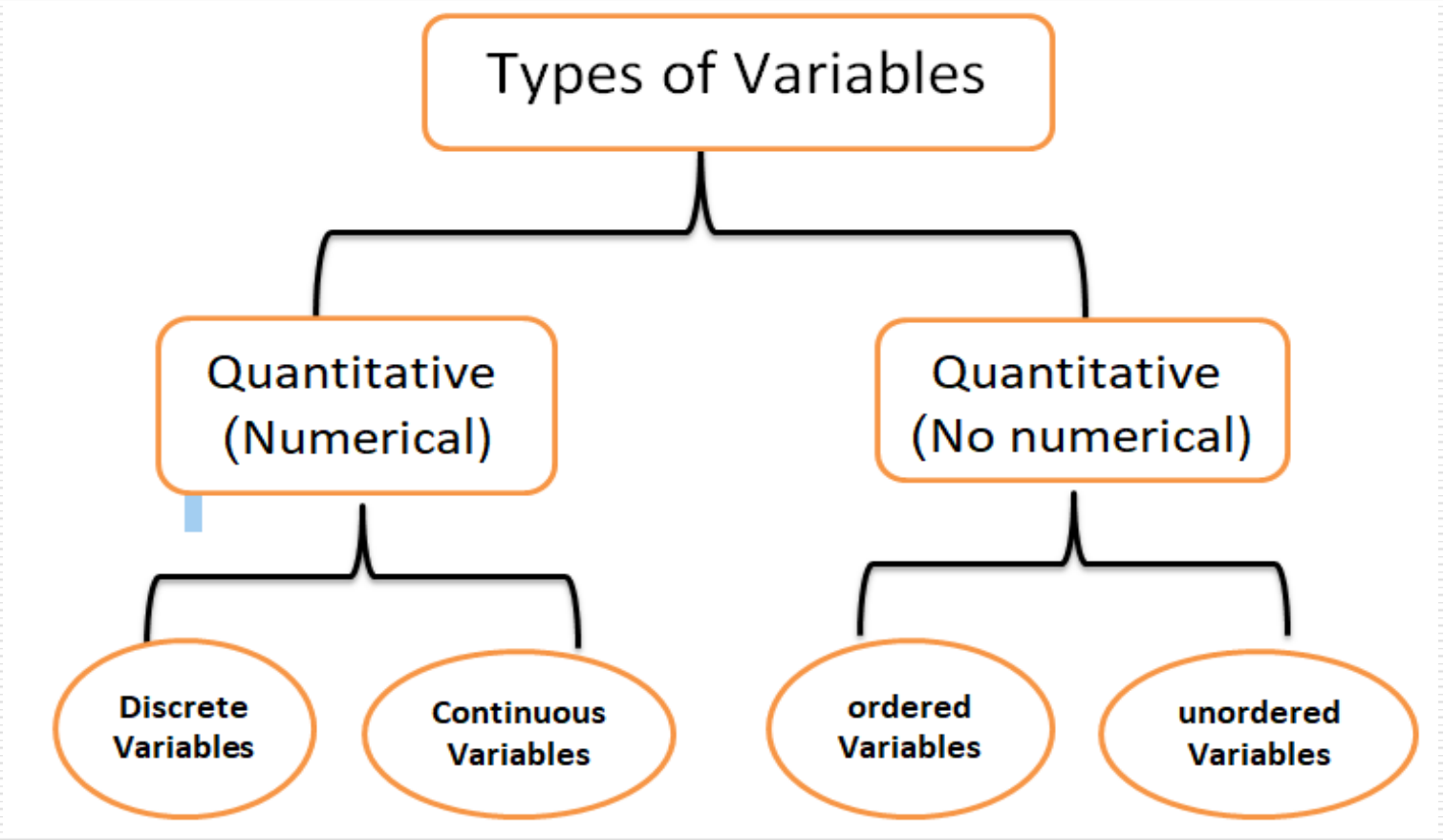
Some principle concepts

- A population: consists of all subjects (human or elements or otherwise) that are studied.

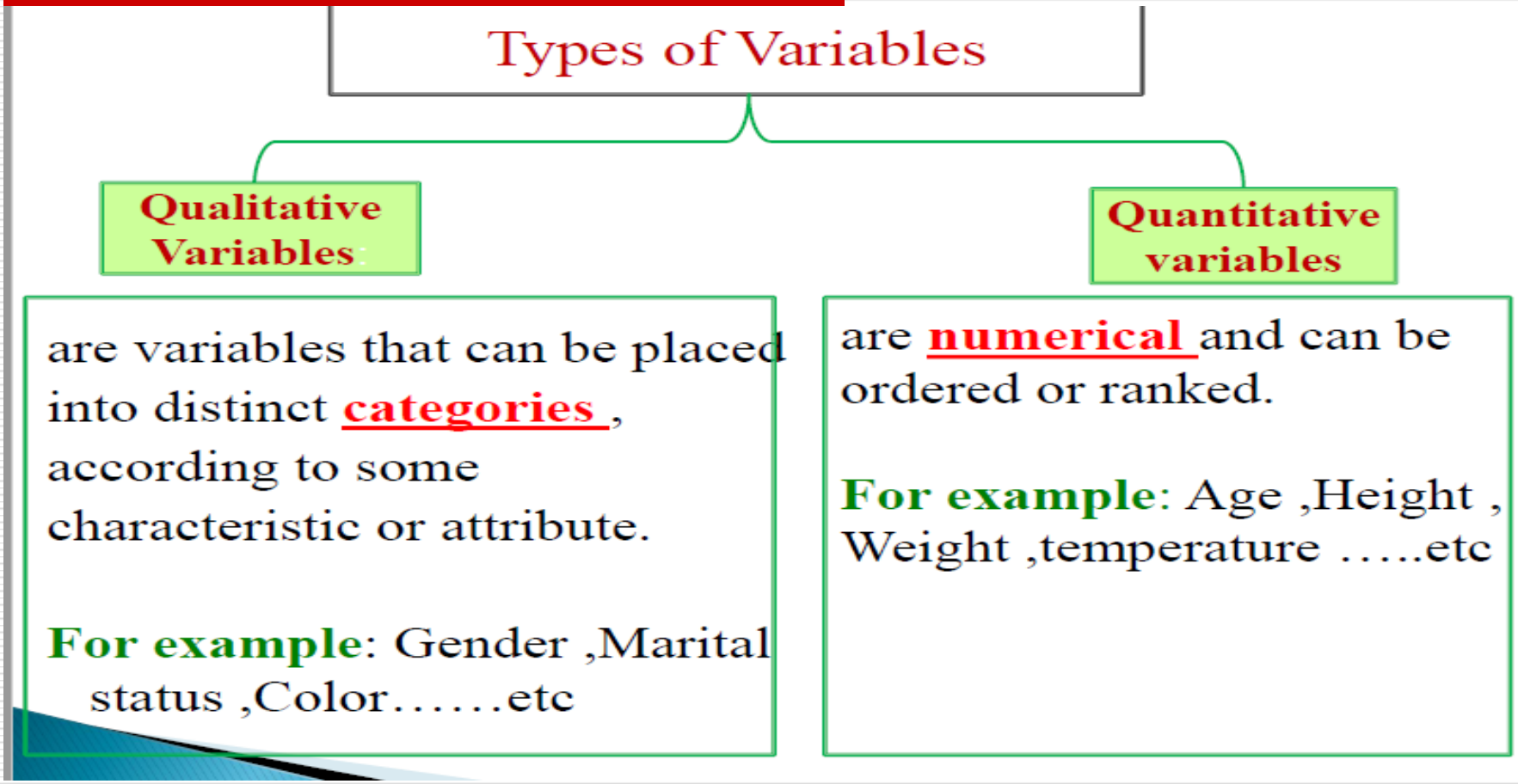
- A sample: is a subset of the population or (is a group selected from a population).



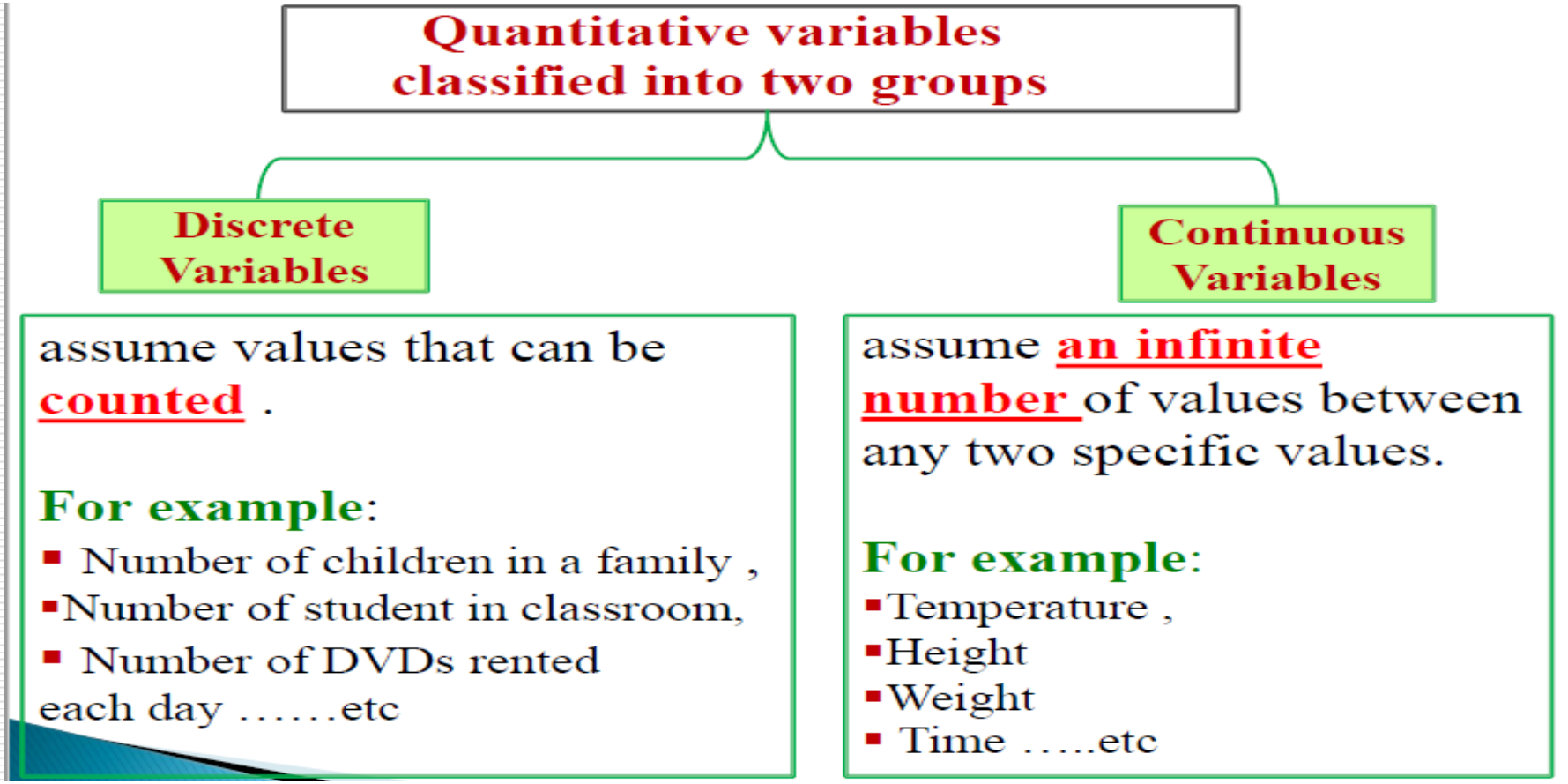
Some basic concepts



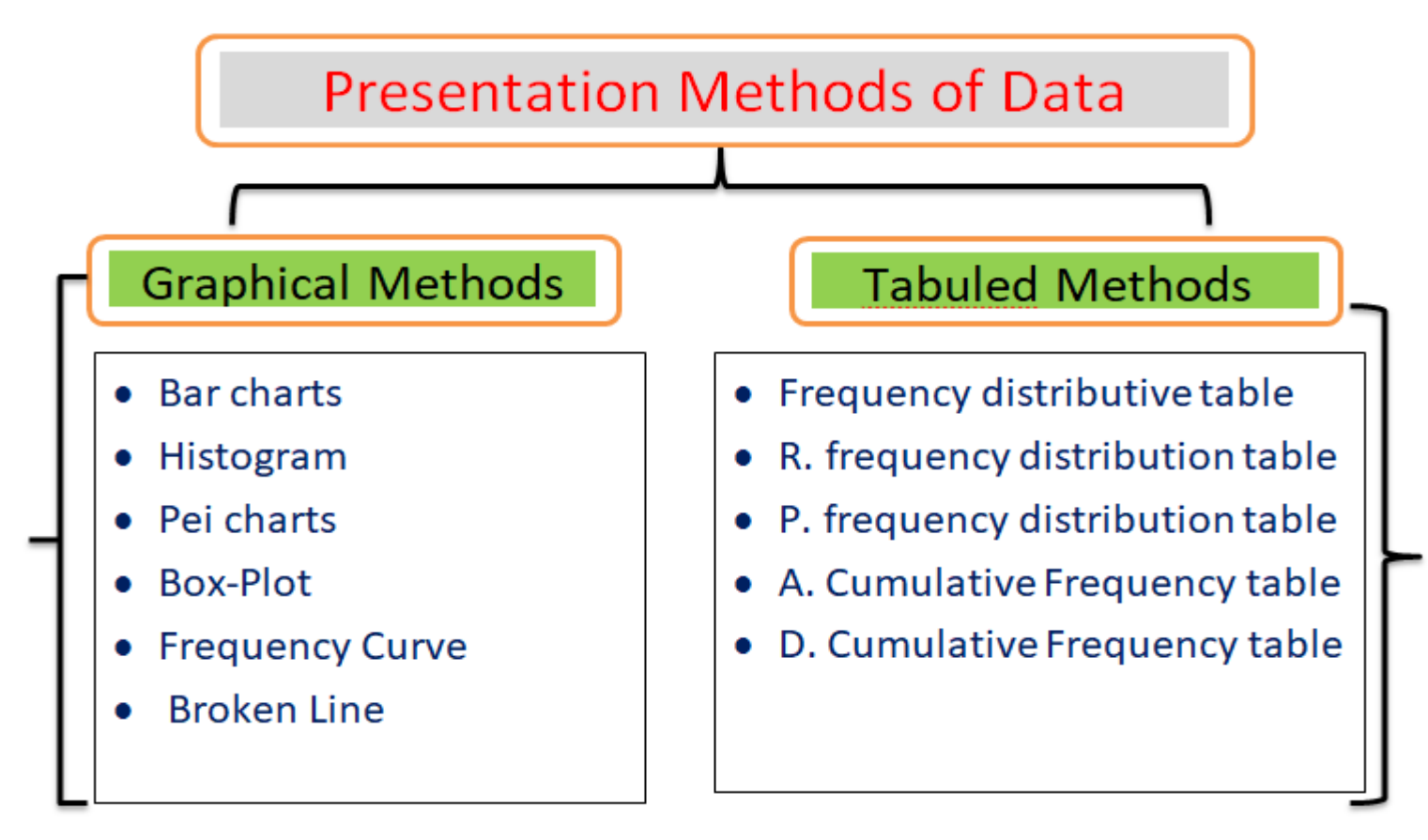
Some basic concepts



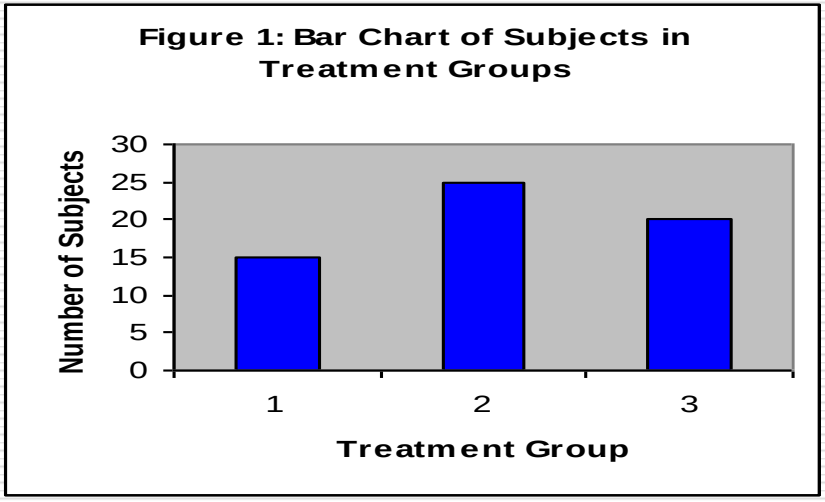
Some basic concepts

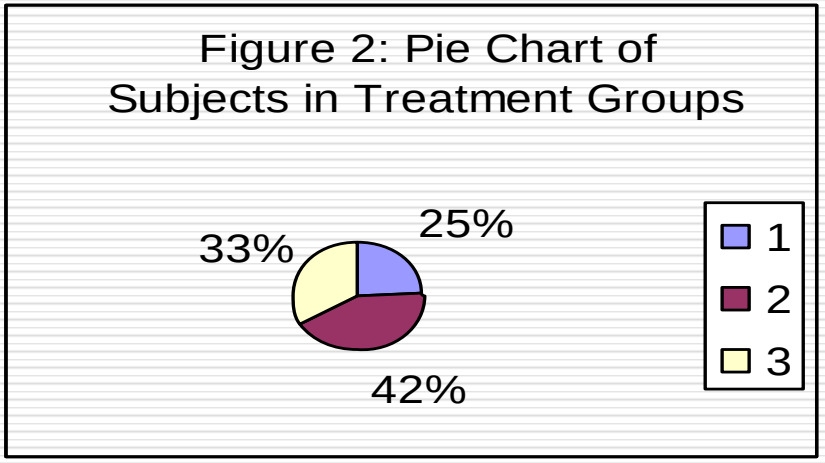


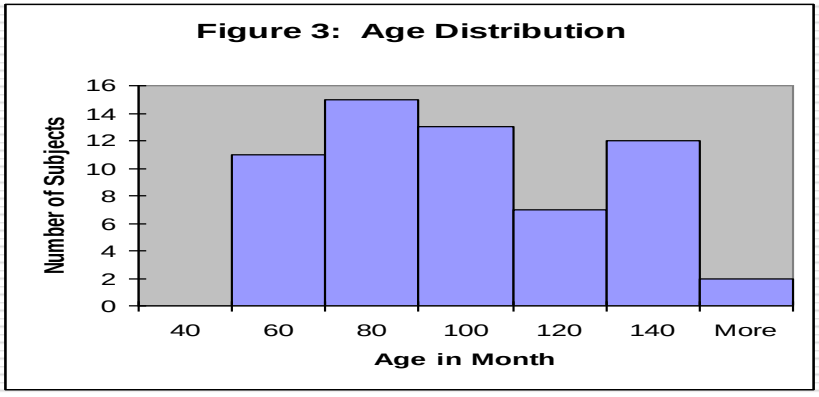
Presentation Methods of Data

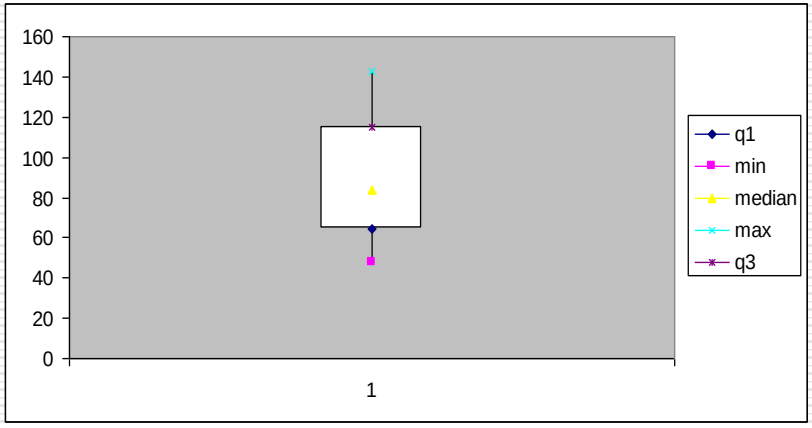


For examples of graphical Methods

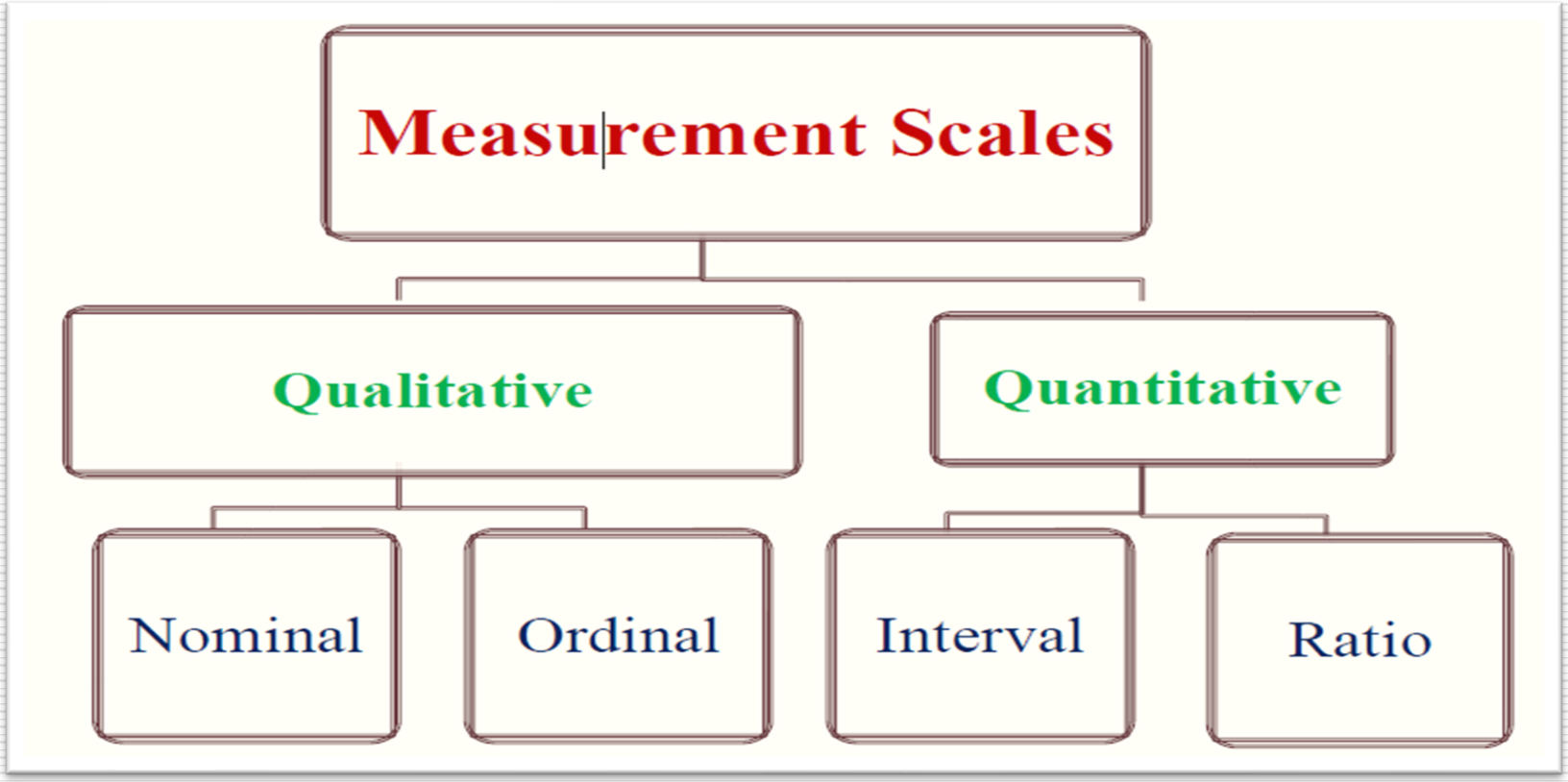




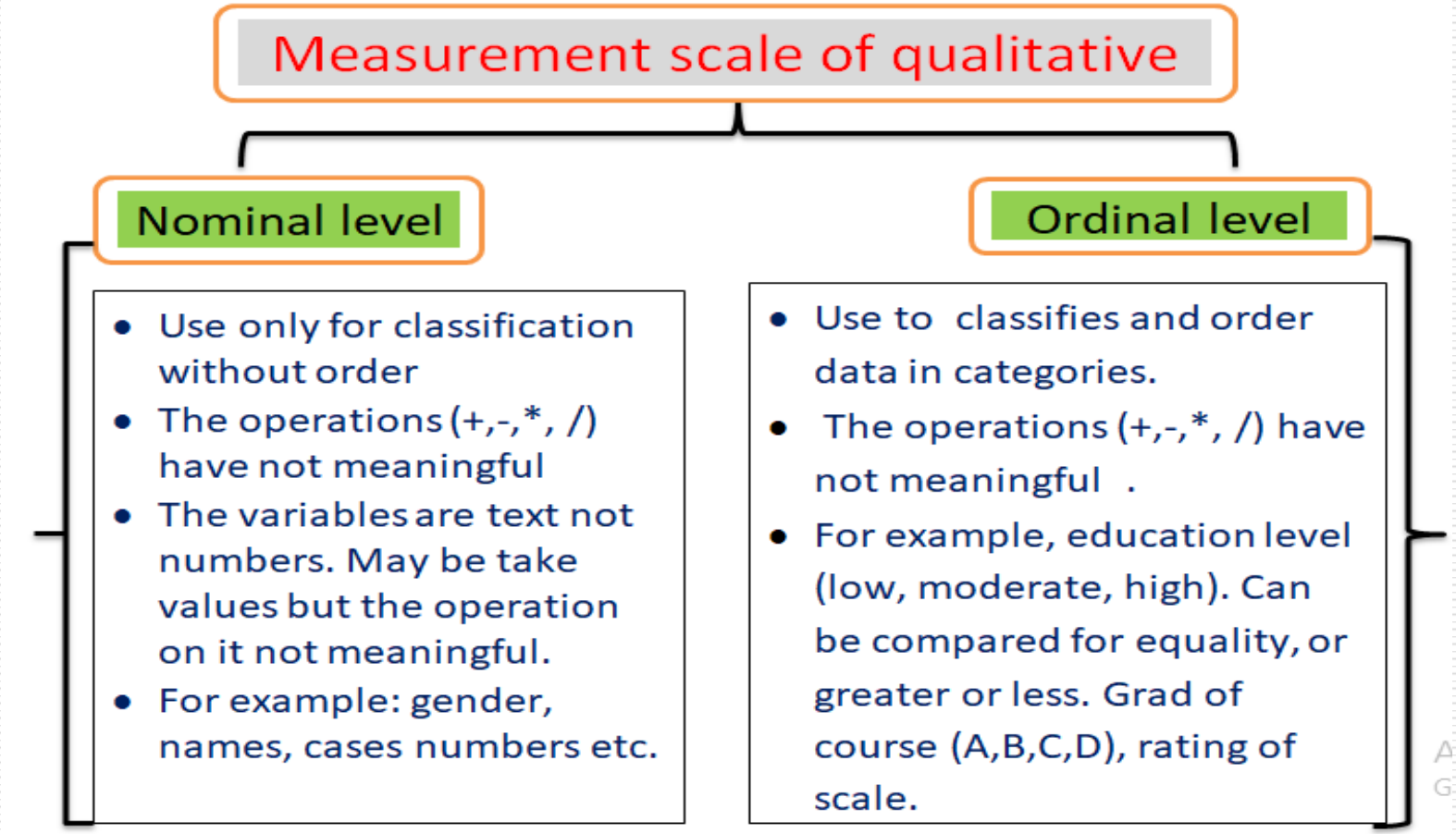




Measurement System



Measurement System



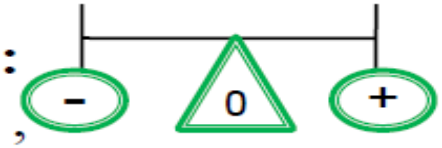
Measurement System

Measurement Scale of Quantitative

Interval level

ranks data and precise differences between units of measure do exist ,however there is **no meaningful zero**.

For example:
Temperature ,
IQ test…etc

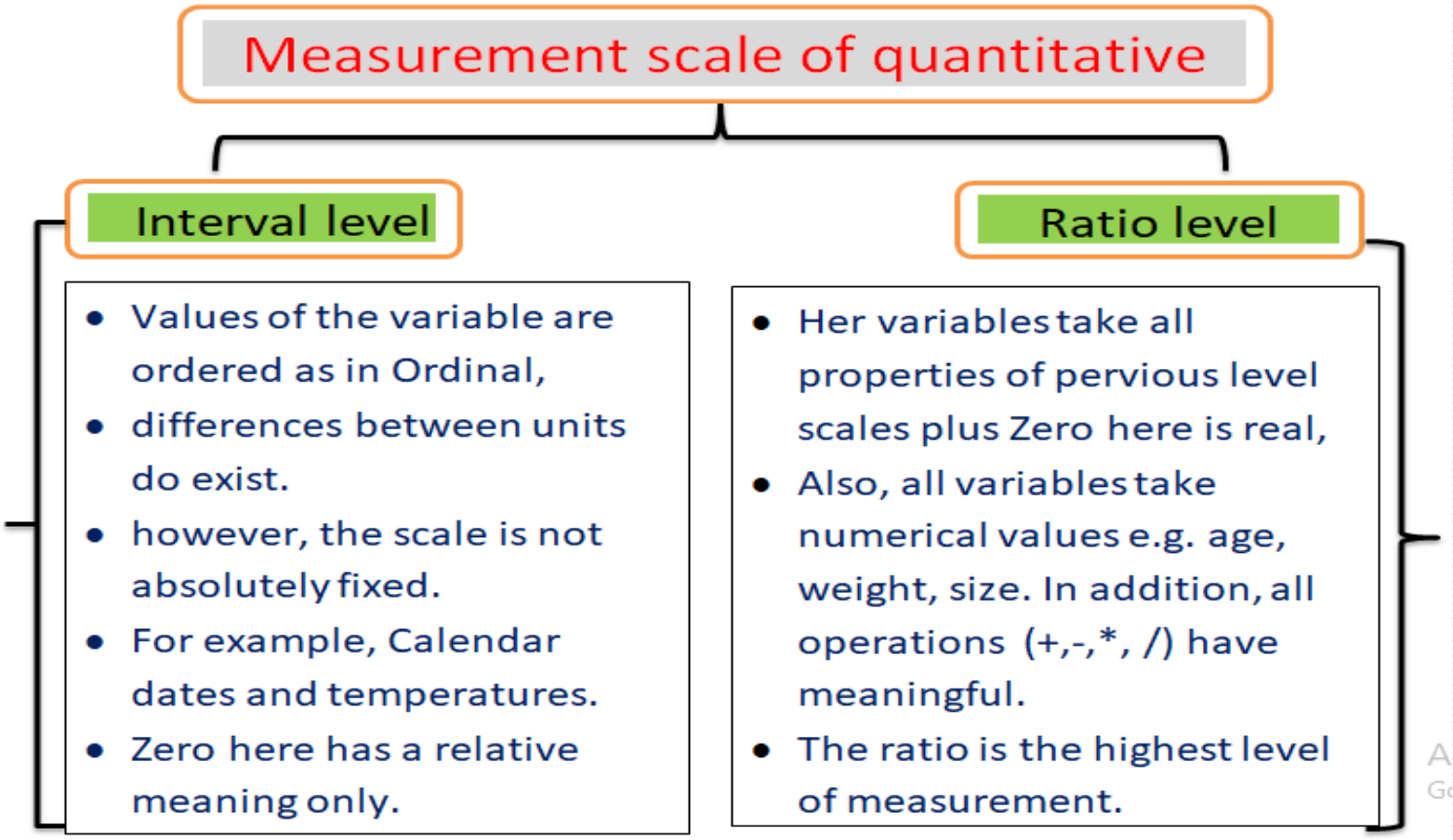


Ratio level

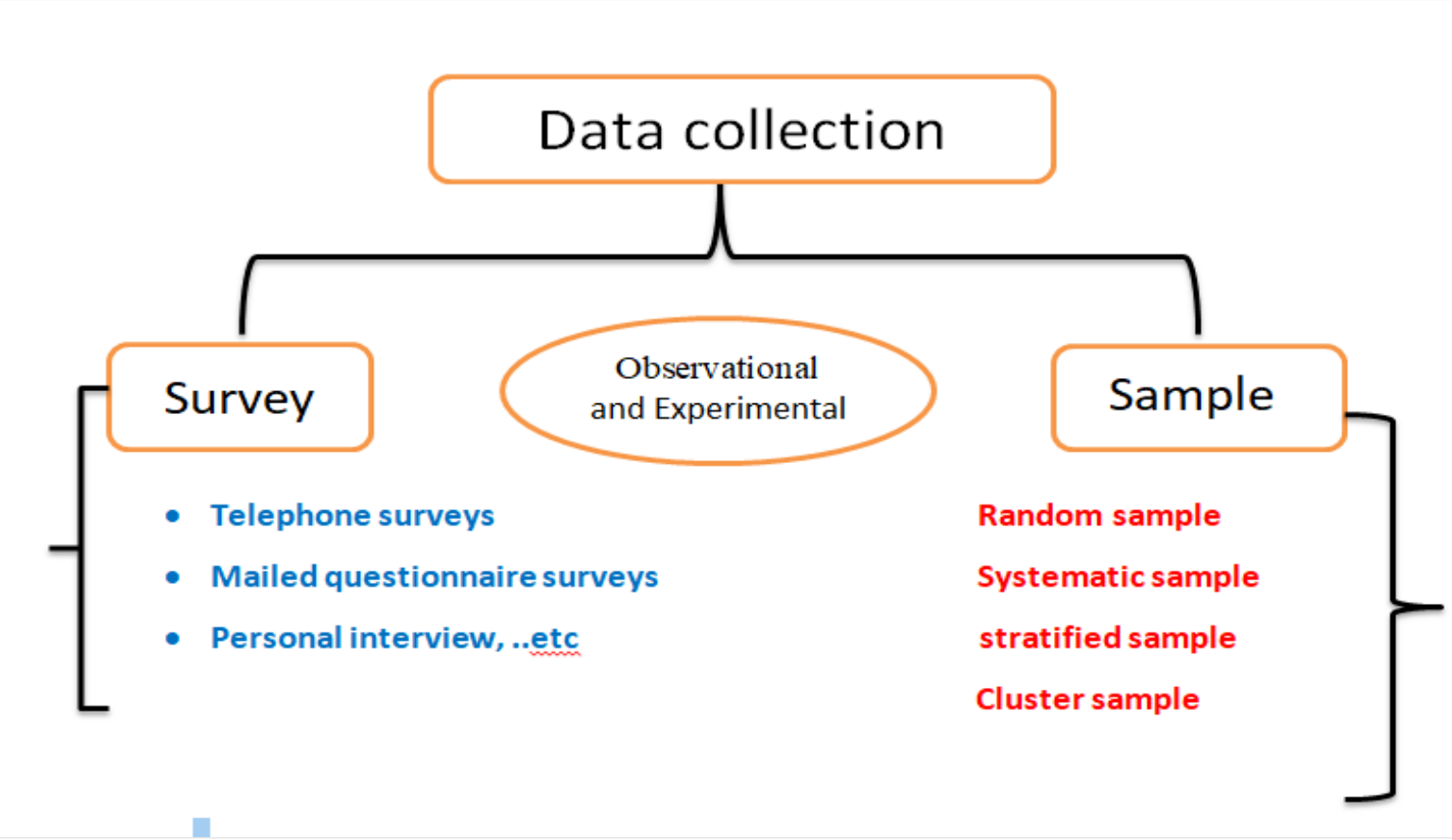
possesses all the characteristics of interval and there exist **a true zero**.

For example:
Height , Weight, Time,
Salary , Age …etc

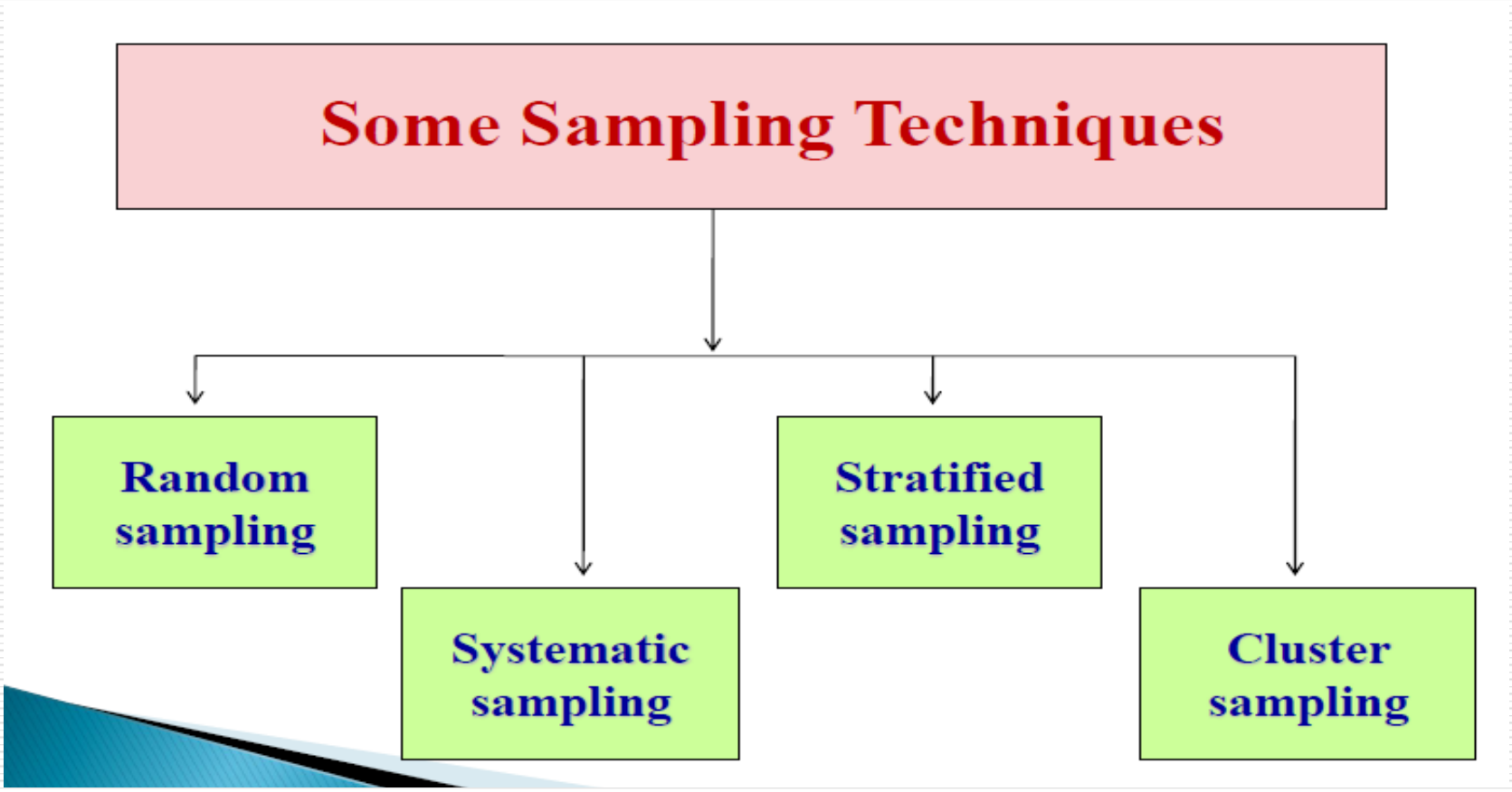
Measurement System



Methods of Data collection



Some Types of Samples



- **Random sample:**

are selected by using chance methods or random numbers.
For example

Table 1–3 Random Numbers

79	41	71	93	60	35	04	67	96	04	79	10	86
26	52	53	13	43	50	92	09	87	21	83	75	17
18	13	41	30	56	20	37	74	49	56	45	46	83
19	82	02	69	34	27	77	34	24	93	16	77	00
14	57	44	30	93	76	32	13	55	29	49	30	77
29	12	18	50	06	33	15	79	50	28	50	45	45
01	27	92	67	93	31	97	55	29	21	64	27	29
55	75	65	68	65	73	07	95	66	43	43	92	16
84	95	95	96	62	30	91	64	74	83	47	89	71
62	62	21	37	82	62	19	44	08	64	34	50	11
66	57	28	69	13	99	74	31	58	19	47	66	89
48	13	69	97	29	01	75	58	05	40	40	18	29
94	31	73	19	75	76	33	18	05	53	04	51	41
00	06	53	98	01	55	08	38	49	42	10	44	38
46	16	44	27	80	15	28	01	64	27	89	03	27
77	49	85	95	62	93	25	39	63	74	54	82	85
81	96	43	27	39	53	85	61	12	90	67	96	02
40	46	15	73	23	75	96	68	13	99	49	64	11

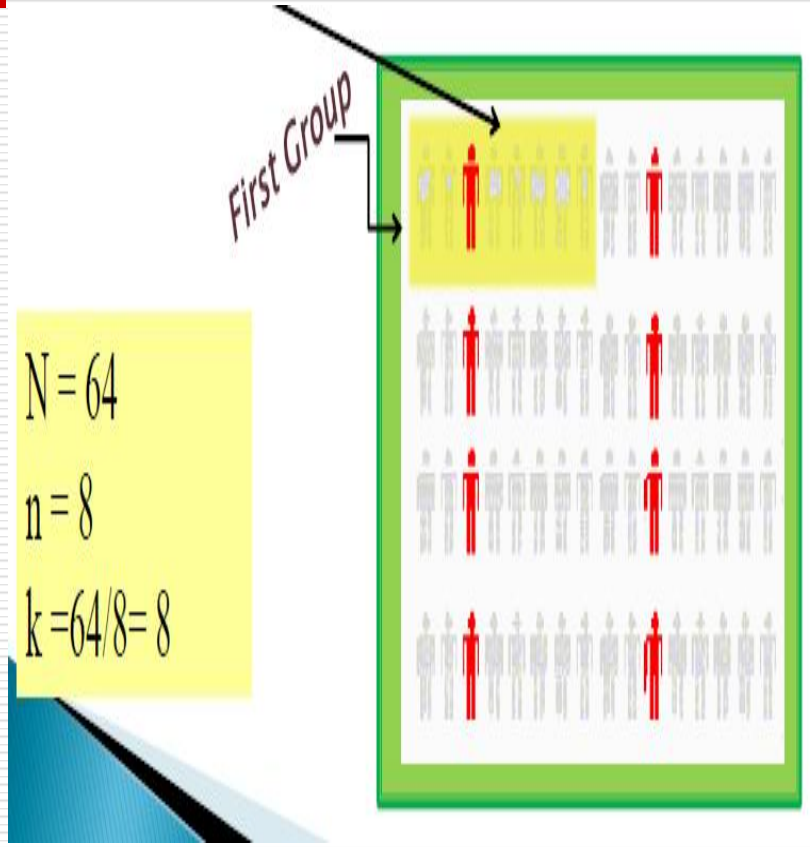
Q: Select random sample of 15 subjects out of 85 subjects:

A: 12, 27, 75, 62, 57, 13, 31, 06, 16, 49, 46, 71, 53, 41, 02

- **Systematic sample**

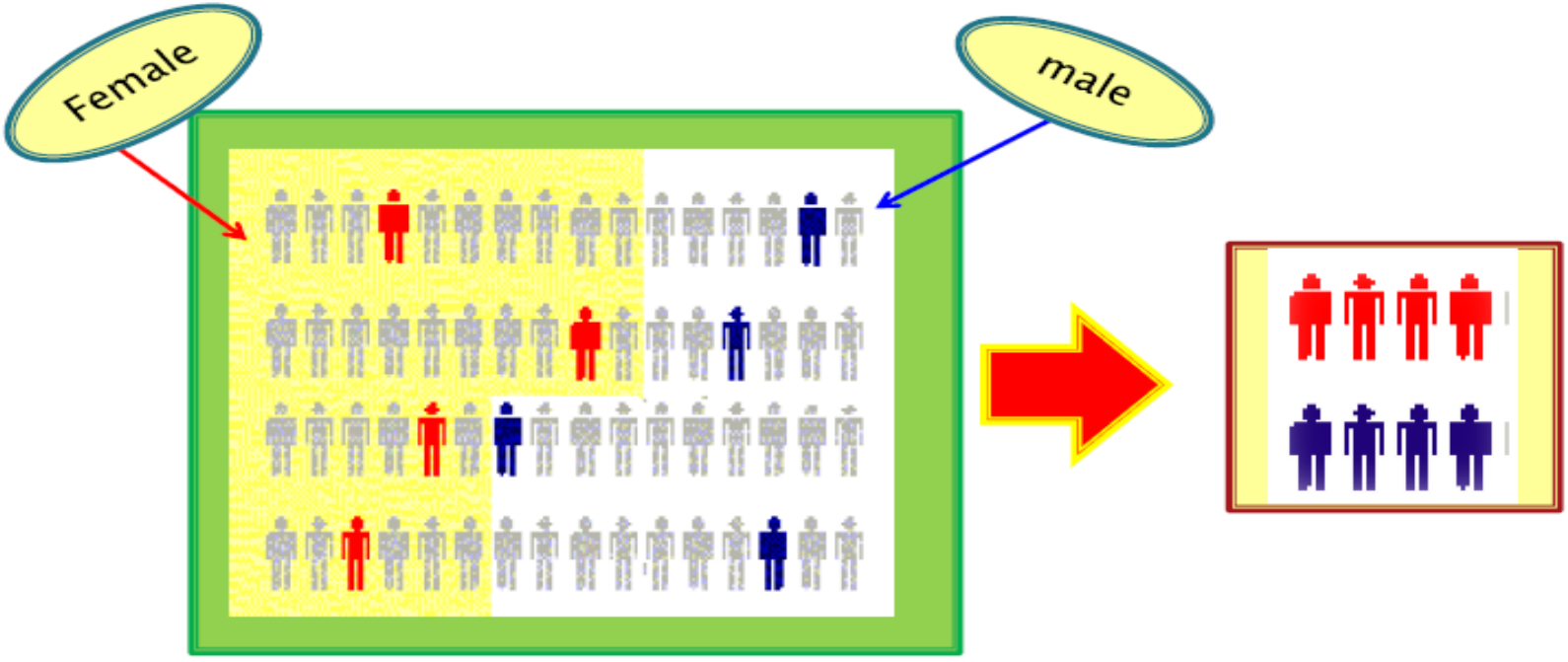
are obtained by numbering each element in the population and then selecting the k^{th} value. For example,

- Determine the sample size: n
- Determine the systematic period k, where k= N/n
- Randomly Select any number in k
- Add k to this selected number to obtain the next value.



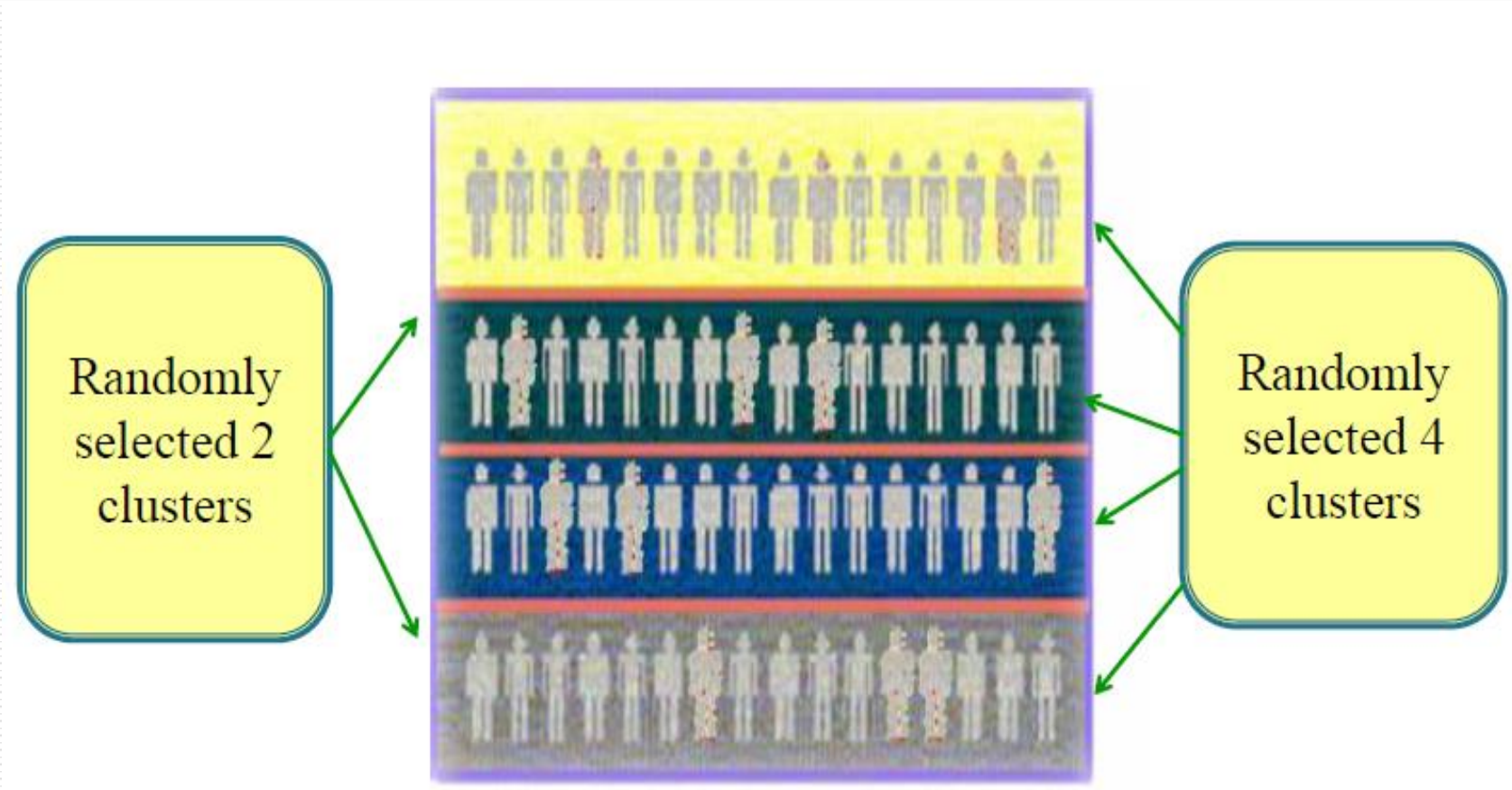
- **Stratified sample:** are selected by dividing the population into groups (strata) according to some characteristic and then taking samples from each group. For example,

A researcher select a random sample from each gender to check their blood pressure



- **Cluster sample**

  are selected by dividing the population into groups and then taking samples of the groups. For example,



Functions of Biostatistics:

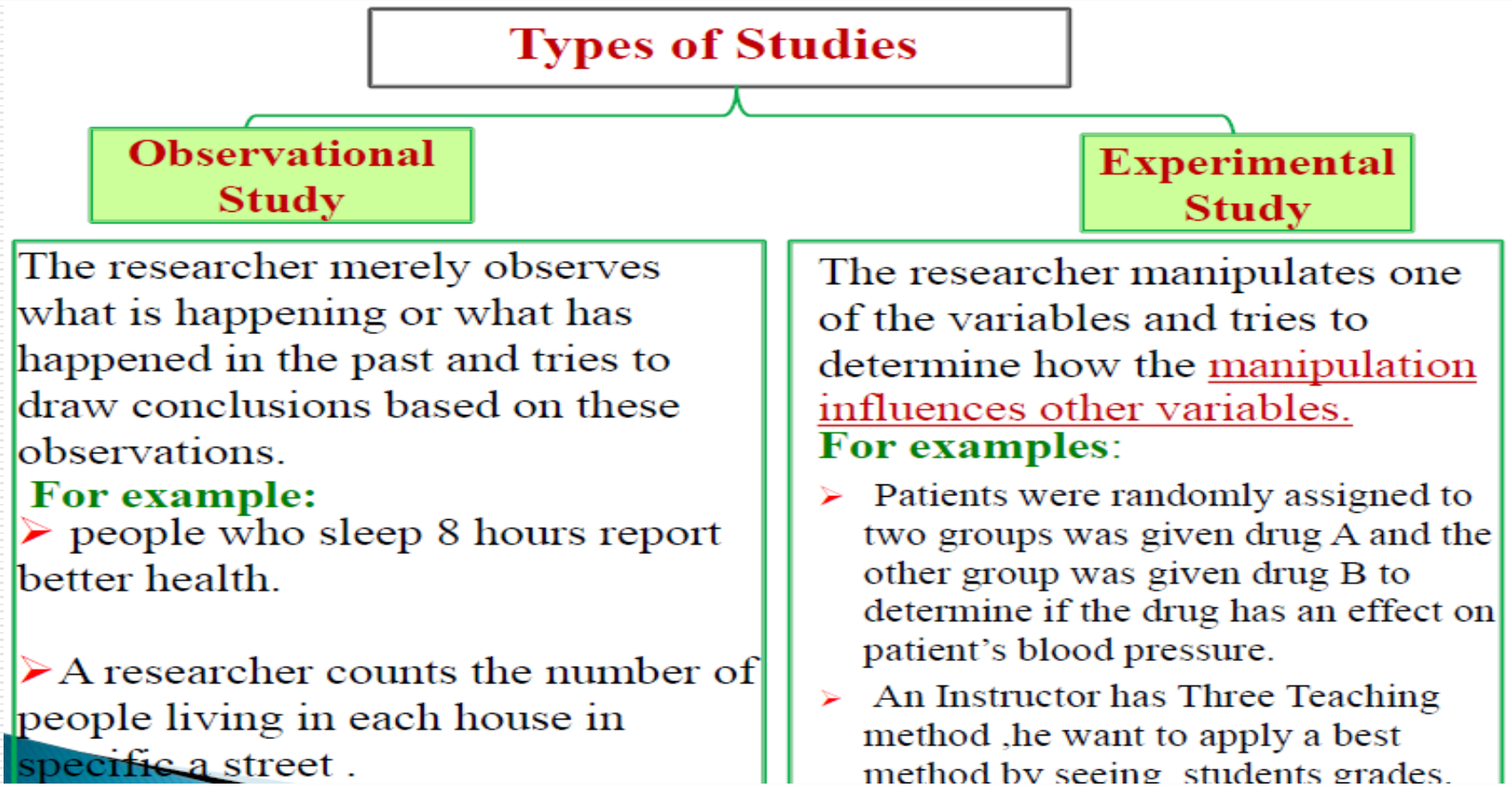
Functions of Statistics:

- It presents facts in a definite form.
- It simplifies mass of figures.
- It facilitates comparison.
- It helps in formulating and testing of hypothesis.
- It helps in prediction.
- It helps in the formulation of suitable policies.

Role of statisticians

- To guide the design of an experiment or survey prior to data collection.
- To analyze data using proper statistical procedures and techniques.
- To present and interpret the results to researchers and other decision makers

Some Types of studies:



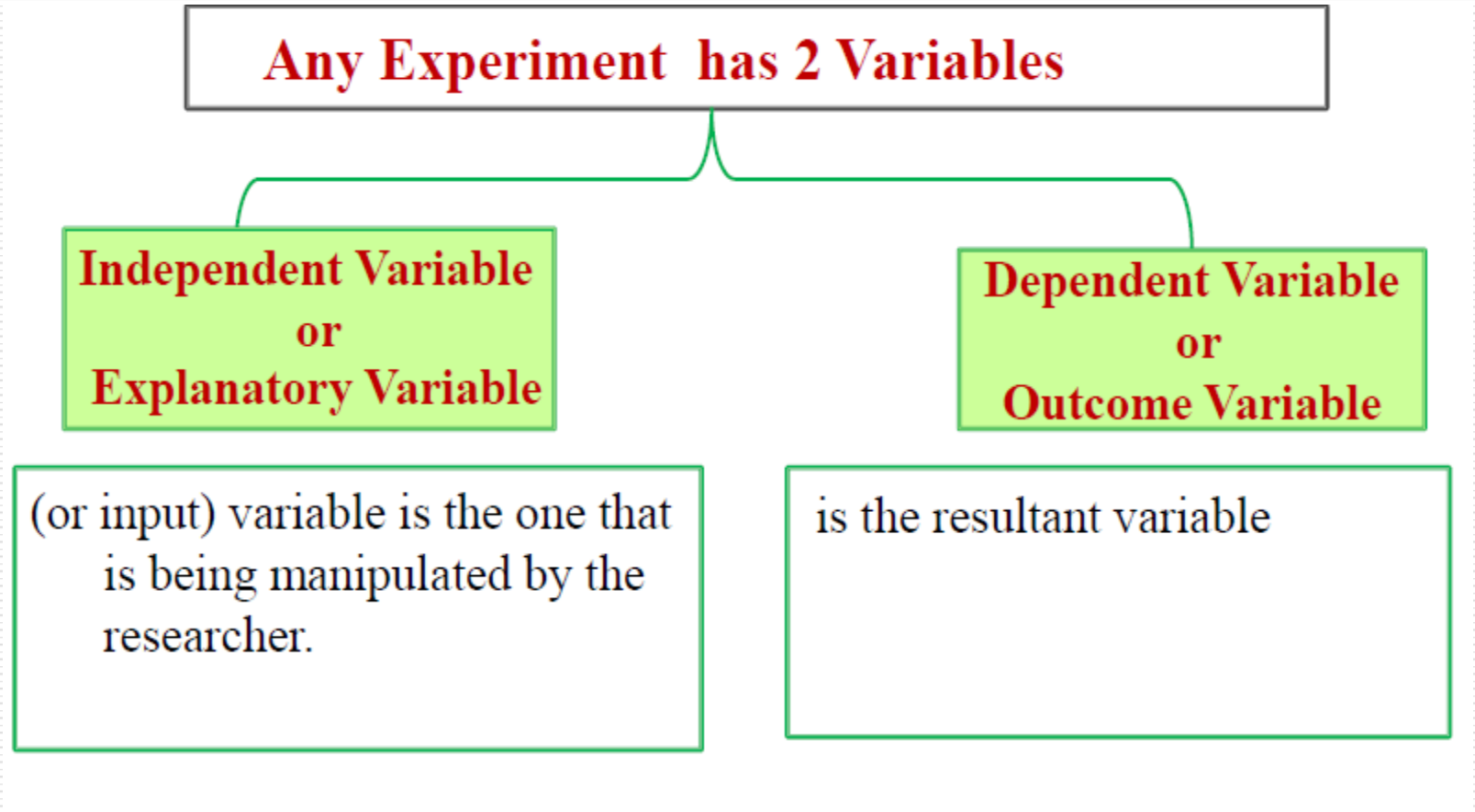
Variables of study:

For example

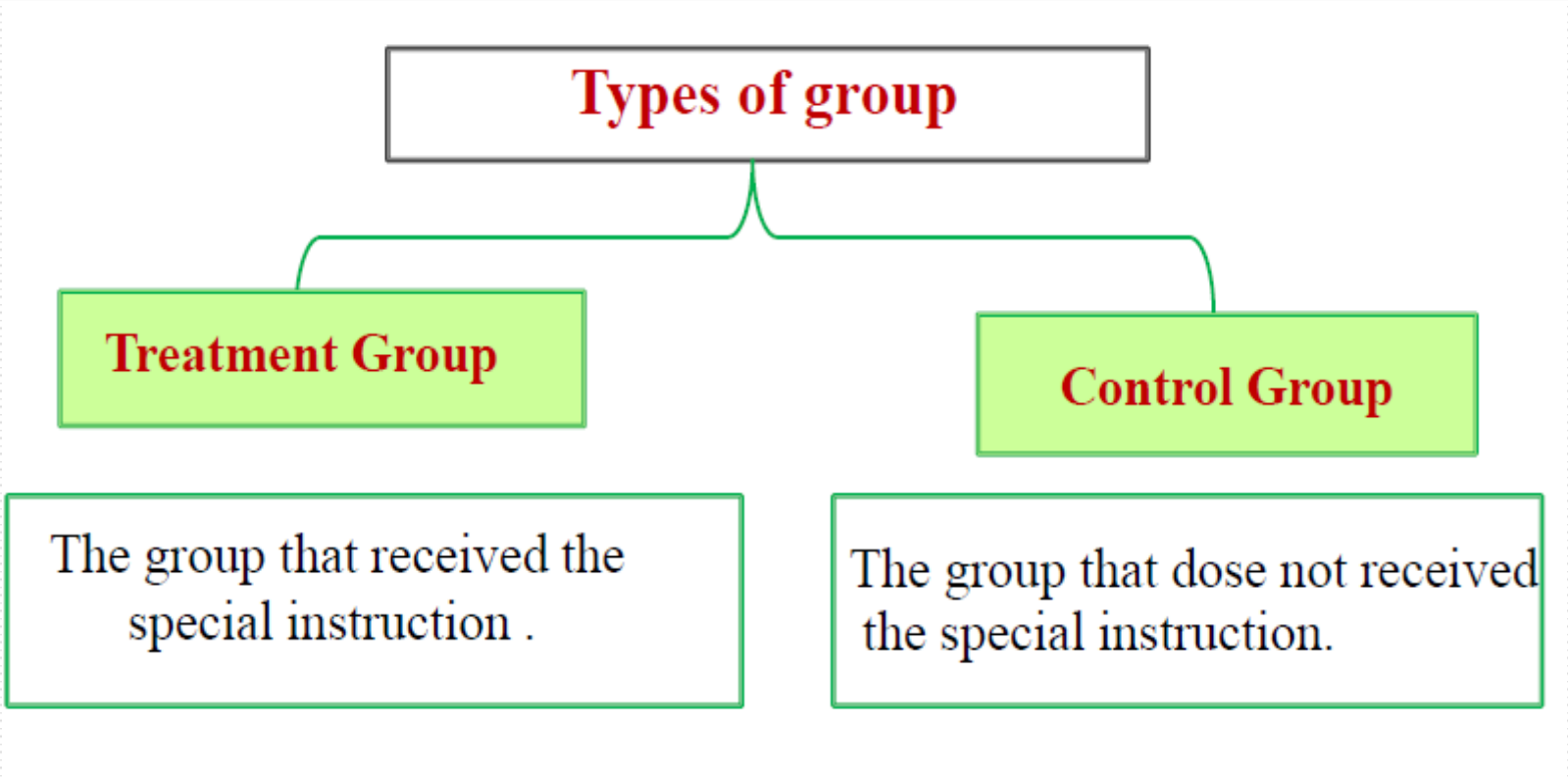
Independent	temperature of water	exercise
Dependent	time to cook an egg	health

Note :Statistical studies usually include one or more independent variables and one dependent variable.

Variables of study:



Experimental study:



Some Statistical Data Analysis Software

The Best Statistical Software tools:

- IBM SPSS Statistics
- SAS/STAT
- Stata
- Minitab
- Graph Pad Prism
- SmartPLS

Lecture 2. Linear Correlation

Introduction:

Today, we'll explore the concept of linear correlation, an essential statistical tool used to measure the strength and direction of the linear relationship between two continuous variables. In biostatistics, understanding the relationship between variables like blood pressure and age, or cholesterol levels and body weight, can provide valuable insights into health outcomes.

1. What is Linear Correlation?

Linear correlation quantifies how much two variables change together. When two variables tend to increase or decrease together, they are said to have a positive correlation. If one variable tends to increase when the other decreases, the correlation is negative.

Types of correlation

- **Positive Correlation:** As one variable increases, the other variable also increases.
- **Negative Correlation:** As one variable increases, the other variable decreases.
- **No Correlation:** There's no apparent relationship between the variables.

Correlation Coefficient

The strength and direction of a linear correlation are expressed by the **correlation coefficient** (often denoted by r).

- r ranges from -1 to 1.
  - $r = 1$ indicates a perfect positive correlation.
  - $r = -1$ indicates a perfect negative correlation.
  - $r = 0$ indicates no linear correlation.

Formula for Pearson Correlation Coefficient (r):

The Pearson correlation coefficient is calculated using the following formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- n is the number of data points.
- x and y are the individual data points for the two variables

Examples of Correlation

3. Example 1: Correlation between Age and Blood Pressure

Imagine we want to investigate the relationship between age and systolic blood pressure. We collect data from a sample of patients and plot age on the x-axis and systolic blood pressure on the y-axis.

- If we find $r = 0.85$, this indicates a strong positive correlation. As age increases, systolic blood pressure also tends to increase.

Examples of Correlation

4. Example 2: Correlation between Exercise and Cholesterol Levels

Consider another example where we examine the relationship between the amount of exercise (measured in hours per week) and cholesterol levels. If $r = -0.70$, this suggests a strong negative correlation—more exercise is associated with lower cholesterol levels.

Interpretation in Biostatistics

5. Interpretation in Biostatistics:

In biostatistics, linear correlation helps to identify potential relationships between biological variables. However, it's crucial to remember:

- **Correlation does not imply causation.** Even if two variables are strongly correlated, it doesn't mean one causes the other.
- **Outliers can affect correlation.** A few extreme values can significantly influence the correlation coefficient.
- **Only measures linear relationships.** If the relationship between the variables is nonlinear, the correlation coefficient may be misleading.

Application in Research

Biostatisticians often use linear correlation in epidemiological studies to explore associations between risk factors and health outcomes, or in clinical trials to investigate the effect of treatments on health indicators.

For example, researchers may study the correlation between a new drug dosage and the reduction in tumor size among patients. A strong negative correlation could indicate that higher doses are associated with smaller tumors, leading to further investigation.

Example: Correlation Between Age and Blood Pressure

Suppose we have data for five individuals on their age and corresponding systolic blood pressure. The data is as follows:

Individual	Age (X)	Blood Pressure (Y)
1	25	120
2	30	122
3	35	125
4	40	130
5	45	135

Steps to Calculate the Correlation Coefficient (r)

1. Calculate the sums:

$$\Sigma X = 25 + 30 + 35 + 40 + 45 = 175$$

$$\Sigma Y = 120 + 122 + 125 + 130 + 135 = 632$$

2. Calculate the sum of squares for each variable:

$$\Sigma X^2 = 25^2 + 30^2 + 35^2 + 40^2 + 45^2 = 625 + 900 + 1225 + 1600 +$$

$$\Sigma Y^2 = 120^2 + 122^2 + 125^2 + 130^2 + 135^2 = 14400 + 14884 + 1562$$

3. Calculate the sum of the products of X and Y:

$$\Sigma XY = (25 \times 120) + (30 \times 122) + (35 \times 125) + (40 \times 130) + (45 :$$

Steps to Calculate the Correlation Coefficient (r)

4. **Use the formula to calculate r:**

The Pearson correlation coefficient r is calculated as:

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

Where:

- $n = 5$ (number of individuals)
- $\Sigma X = 175$
- $\Sigma Y = 632$
- $\Sigma X^2 = 6375$
- $\Sigma Y^2 = 80034$
- $\Sigma XY = 22310$

Plug in the values we get:

$$r = \frac{5(22310) - (175)(632)}{\sqrt{[5(6375) - 175^2][5(80034) - 632^2]}}$$

Simplifying:

$$r = \frac{111550 - 110600}{\sqrt{[31875 - 30625][400170 - 399424]}}$$

$$r = \frac{950}{\sqrt{[1250][746]}}$$

$$r = \frac{950}{\sqrt{932500}}$$

$$r = \frac{950}{965.13}$$

$$r \approx 0.9844$$

Interpretation

The correlation coefficient $r \approx 0.9844$ indicates a very strong positive linear relationship between age and blood pressure. This suggests that as age increases, systolic blood pressure tends to increase as well.

This example demonstrates how to calculate and interpret the correlation coefficient between age and blood pressure, providing insight into the strength and direction of their relationship.

Conclusion

Linear correlation is a fundamental concept in biostatistics that helps to understand the relationships between variables. By interpreting the correlation coefficient, researchers can gain insights into how two variables interact, providing a foundation for further analysis and hypothesis testing.

Always remember to carefully interpret the correlation within the context of your study and consider additional statistical analyses to confirm your findings.

Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (often denoted as ρ or r_s) is a non-parametric measure of the strength and direction of association between two ranked variables. Unlike Pearson's correlation, which assumes that the variables are linearly related and normally distributed, Spearman's correlation does not require these assumptions. This makes it particularly useful in medical research where the data may not meet the stringent requirements of parametric tests.

Key Features

- **Non-parametric**: Spearman's correlation is suitable for both continuous and ordinal data.
- **Rank-based**: It evaluates the relationship based on the ranks of the data rather than the raw data values.
- **Range**: The coefficient ranges from -1 to 1:
  - $r_s = 1$: Perfect positive correlation.
  - $r_s = -1$: Perfect negative correlation.
  - $r_s = 0$: No correlation.

Steps to Calculate Spearman's Rank Correlation

1. **Assign Ranks**: Rank the data for each variable. If there are tied ranks (i.e., the same value occurs more than once), assign each value the average of the ranks they would have received.
2. **Calculate the Difference**: For each pair of observations, subtract the rank of one variable from the rank of the other variable. Denote this difference as d_i.
3. **Square the Differences**: Square each difference (d_i^2).
4. **Sum the Squared Differences**: Calculate the sum of the squared differences ($\sum d_i^2$).
5. **Apply the Formula**:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n is the number of pairs of observations.

Example;

Suppose you want to investigate the correlation between two variables: the severity of a symptom and the age of patients. You have data for five patients:

Patient	Symptom Severity (Rank)	Age (Rank)	Difference (d)	d^2
A	2	5	-3	9
B	4	3	1	1
C	3	4	-1	1
D	1	1	0	0
E	5	2	3	9

Now, sum the squared differences:

$$\sum d_i^2 = 9 + 1 + 1 + 0 + 9 = 20$$

Example

Apply the Spearman's formula:

$$r_s = 1 - \frac{6 \times 20}{5(5^2 - 1)} = 1 - \frac{120}{120} = 0$$

In this case, $r_s = 0$ indicates that there is no apparent correlation between symptom severity and age in this small sample.

Applications in Medicine

Spearman's rank correlation is widely used in medical research for scenarios like:

- **Assessing the relationship between patient satisfaction scores and the number of visits to a healthcare provider.**
- **Determining the association between the rank of a particular diagnostic test result and the severity of a disease.**
- **Evaluating correlations between different ordinal scales, such as pain levels and response to**

Thank you for all..