



الجامعة التقنية الشمالية  
كلية التقنيات الصحية والطبية الدور  
قسم تقنيات العلاج الطبيعي

# احصاء حيوي





# Lecture Notes

For Collage of Health and Medical Techniques -Aldouy Students  
**Second stage**



## *Biostatistics*

Dr. Ghassan Jasim Hadi  
Northern Technical University  
September 2024

اسم المادة	باللغة العربية	احصاء حيوي	عدد الساعات الاسبوعية			
	باللغة الانكليزية	Biostatistics	نظري	عملي	المجموع	عدد الوحدات
لغة التدريس للمادة	الانكليزية		2	2	4	3

# مفردات المادة

العنوان	الاسبوع
Introduction , Objectives Of Statistics, The Major Objectives Of Statistics, Classification Of Statistics, Stages Of Statistical Method in Scientific Research.	الاول والثاني
Sources of Data Collection.	الثالث
Samples: Introduction, Types of Samples.	الرابع
Types of data: Introduction, constant data, variables , Types of variables.	الخامس والسادس
Data display: Introduction, Display data numerically as Simple display or Raw data, Ordered Display or Array, Data display frequency table.	السابع
Display data graphically.	الثامن والتاسع
Measures of central tendency .	العاشر والحادي والثاني عشر
Measures of Dispersion.	الثالث عشر
Measures of Skewness and Kurtosis.	الرابع عشر
General Review.	الخامس عشر

# الاهداف العامة

## اهداف الهادة :

➤ الهدف العام: تعريف الطالب على التطبيقات الاحصائية وكذلك التعرف على الهندس الطبي واجرائاته.

➤ الهدف الخاص:

تعريف الطالب على اهمية الاحصاء في البحث العلمي 1.

اعداد خطط للهندس الطبي 2.

التعرف على طرق اخذ العينات من مجتمع الدراسة 3.

كيفية تطبيق القوانين في عرض بيانات الدراسة 4.



## References

1. Bland, M. (2000). An Introduction to Medical Statistics, 3<sup>rd</sup> ed. University Press, Oxford.
2. Altman, D.G. (1991). Practical Statistics for Medical Research, Chapman and Hall, London.
3. Armitage, P. and Berry, G. (1987). Statistical Methods in Medical Research, 2<sup>nd</sup> ed. Blackwell, Oxford.
4. Michael, J. (1999). Medical Statistics: A commonsense

# **First Lecture:**

## **Introduction to Statistics**

# First Lecture:

## Introduction to Statistics

### The importance of statistics in scientific research

Scientific research is the best way to obtain knowledge, understand phenomena, and use the statistical method in research. Scientific means providing data and information about the phenomena to be studied in that research, and this means that the possibility of applying . The statistical method in scientific research is related to the possibility of expressing the studied phenomenon quantitatively. The method is distinguished. Statistic is that it provides the researcher with an objective scientific method so that there is no interference or bias in the research results and this advantage. The statistical method has been used by researchers in the fields of pure sciences or humanities and others, and as a result of using Statistics in these fields has appeared names for statistics coupled with the name of another science and this does not change the goal of the science of statistics.



# Pre-Test

- 1-What do you know about Statistics?
- 2- What do you know about Data?

# Definitions

## Terms Statistical :

The word “statistics” is derived from the Latin word “status” (case) that solves facts and information for the collection of private information You can benefit from it in the matter of taxes, determining the labor force, and so on. There are many definitions of statistics and they vary in terms of content and comprehensiveness according to the development of the science and the expected benefits. Here are some of the definitions:

- 1 - Statistics are categorized facts from information about an individual and a country**
- 2 - Statistics is the science concerned with collecting, analyzing, and interpreting numerical data.**
- 3 - Statistics is the science of estimates and probability.**

# Definitions

**Statistics:** is defined as the scientific method that is concerned with collecting data and facts about a particular phenomenon or group of phenomena, organizing and classifying these data and facts in a way that can be analyzed and interpreted more easily and to reach the appropriate decision-making.

هو العلم الذي يهتم **Statistics** علم الاحصاء، بالطرق العلمية لجمع وتنظيم وتلخيص البيانات وعرضها وتحليلها بأساليب علمية للحصول منها على المعلومات اللازمة لاتخاذ القرارات المناسبة. تحليل البيانات اى المعالجة الاحصائية للبيانات قد تتم بطريقة او بطريقة **Manual Processing** يدوية باستخدام **Electronic Processing** الكترونية الكمبيوتر او بهزيج منهما.



# Definitions

**Definition: When different statistical methods are applied to biological, medical, and public health data they constitute the discipline of Biostatistics.**

# Definitions

**Definition:** When different statistical methods are applied to biological, medical, and public health data they constitute the discipline of Biostatistics.

**Biostatistics:** is the branch of applied statistics directed toward applications in the health sciences and biology.

**Biostatistics:** The tools of statistics are employed in many fields -business, education, psychology, agriculture, and economics, to mention only a few. When the data being analyzed are derived from public health data, biological sciences, and medicine, we use the term biostatistics to distinguish this particular application of statistical tools and concepts.



# Classification of Biostatistics

**Descriptive statistics** الإحصاء الوصفي: A statistical method that is concerned with the collection, organization, summarization, and analysis of data from a sample of the population.

**Inferential statistics** الإحصاء الاستدلالي: A statistical method that is concerned with concluding/inferences about a particular population by selecting and measuring a random sample from the population.

# Biostatistics



```
graph TD; A[Biostatistics] --> B[Descriptive statistics]; A --> C[Inferential statistics]; B --> D[Collecting, Organizing, Summarizing, Present of Data]; C --> E[Making inferences, Hypothesis testing, determining relationship, making the prediction];
```

## Descriptive statistics

Collecting, Organizing,  
Summarizing, Present of  
Data

## Inferential statistics

Making inferences,  
Hypothesis testing,  
determining relationship,  
making the prediction

## Definition of Some basic terms

---



**Population:** is the complete set of possible measurements for which inferences are to be made.



**Census:** a complete enumeration of the population. But in most real problems it cannot be realized, hence we take a sample.



**Sample:** A sample from a population is the set of measurements that are collected in the course of an investigation.



**Parameter:** Characteristic or measure obtained from a population.



**Statistic:** A statistic refers to a numerical quantity computed from sample data (e.g. the mean, the median, the maximum...).



**Data:** Refers to a collection of facts, values, observations, or measurements that the variables can assume.

# Definition of Some basic terms



**Statistics**: is a branch of mathematics dealing with data collection, organization, analysis, interpretation, and presentation.



**Sampling**: The process or method of sample selection from the population.



**Sample Size**: The number of elements or observations to be included in the sample.



**Variable**: It is an item of interest that can take on many different numerical values.

# Stages of Statistical Enquiry

- Here is a brief detail about the different stages of a statistical inquiry:

- **Collecting Data**

The collection of statistical data is one of the most important aspects of a statistical inquiry. In this stage, you collect relevant data from multiple sources – both primary and secondary in nature. The source is primary if the data (either published or unpublished) is originally collected by an investigator or an [agency](#).

On the other hand, the source is secondary if the data (published or unpublished) is taken from an agency or a person who have already used the data for their statistical requirements. It is also important to note that the difference between primary and secondary data is a matter of degree alone.

- **Organizing and Presenting Numerical Data**

While conducting a statistical inquiry, the second important stage is the collection and presentation of numerical data. When you collect data, the secondary source usually provides it in an organized form. However, data from the primary source is “raw” and unorganized.

Therefore, you need to edit, classify, and tabulate the data in order to organize it. Editing data involves the removal of omissions, inaccuracies, and inconsistencies present in the data.

Further, classifying data involves bringing together the data items which have common characteristics. Subsequently, you put the data in a tabular format and present it well. The presentation is either in the form of a chart, diagram, graph, etc.



# Stages of Statistical Enquiry

- **Analyzing the Numerical Data**

Once the data is collected, organized, and presented, it is important to analyze the numerical data to get a better understanding of the subject matter. You can use some popular measures to analyze numerical data like:

- Averages or measures of the central tendency
  - Dispersion
  - Correlation
  - Skewness
  - Regression
  - Association and Attributes
  - Interpolation and Extrapolation, etc.
  - Further, to simplify the data, you can use probability and distribution, sampling, index numbers, variance analysis, and time series.

- **Interpreting the Numerical Data**

Once you have analyzed the numerical data, you must draw conclusions and inferences from it. This is the interpretation of numerical data. It is a sensitive and difficult task requiring a high degree of skill, experience, common sense, and also a sense of balanced judgment of the investigator.

Further, if the investigator misinterprets the data, then he might draw conclusions which lead to a waste of time and resources. This can eventually defeat the purpose of the statistical inquiry.

# Data

**Statistical data:** data that refers to numerical descriptions of things. These descriptions may take the form of counts or measurements. Thus statistics of malaria cases in one of the malaria detection and treatment posts of Ethiopia include fever cases, number of positives obtained, sex and age distribution of positive cases, etc.

# Data collection methods



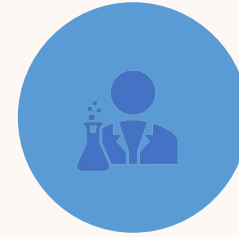
SURVEYS,



INTERVIEWS,



OBSERVATIONS,



EXPERIMENTS,



EXTERNAL DATA

# Data

## Choosing a Method of Data Collection

- Decision-makers need information that is relevant, timely, accurate and usable. The cost of obtaining, processing and analyzing these data is high. The challenge is to find ways, which lead to information that is cost-effective, relevant, timely, and important for immediate use. Some methods pay attention to timeliness and reduction in cost. Others pay attention to accuracy and the strength of the method in using scientific approaches.
- The statistical data may be classified under two categories, depending upon the sources.

1) Primary data      2) Secondary data



# Data

## 1. Primary Data:

- Data generated for the first time primarily/originally for the study in question.
- It needs the involvement of the researcher himself. Census and sample surveys are sources of primary types of data.

## 2. secondary Data:

Obtained from other pre-existing/ priorly collected sources In this case data were obtained from already collected sources

like newspapers, magazines, DHS, hospital records, and existing

data like:

- Mortality reports
- Morbidity reports
- Epidemic reports
- Reports of laboratory utilization (including laboratory test results)



# Data

**The selection of the method of data collection is also based on practical considerations, such as:**

- 1) The need for personnel, skills, equipment, etc. in relation to what is available and the urgency with which results are needed.**
- 2) The acceptability of the procedures to the subjects - the absence of inconvenience, unpleasantness, or untoward consequences.**
- 3) The probability that the method will provide good coverage, i.e. will supply the required information about all or almost all members of the population or sample. If many people do not know the answer to the question, the question is not appropriate.**

# Characteristics of statistical data

## Relate to an aggregate of facts

Single observation cannot be called a Statistics

Production of food grain in a particular year

percentage marks of a single student can not be Statistics

Yearly production of food grain for a few years, percentage marks of all students in a class can constitute Statistical Data

## Affected by multiple causes

If we consider the production of rice in a Maharashtra for few years, obviously the figures will differ every year.

This is because the Statistical data are affected to a marked extent by multiple causes.

## Numerically expressed

Statistical data should be expressed in terms of numbers.

Production of sugar is excellent in Maharashtra, Performance of students at the HSC Examination has improved over some time can not constitute Statistical Data

## Estimated by a reasonable standard of accuracy

Estimation is always crude expression without actual measurement.

## Collected in systematic manner & Collected for predetermined purpose

The purpose should be well-defined and clear, otherwise, some unnecessary information may be collected or necessary information may be ignored.

## Placed in relation to each other

Statistical data are always placed in relation to each other ie they are comparable.

# Post Test

1

Explain the meaning of statistics, population, Biostatistics, and sample.

2

What are the important Characteristics of statistical data?

3

Is there a deferent between statistics and biostatistics?

# sampling

- Pretest
- What is Sampling?
- What are the types of Sampling?

# sampling

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.



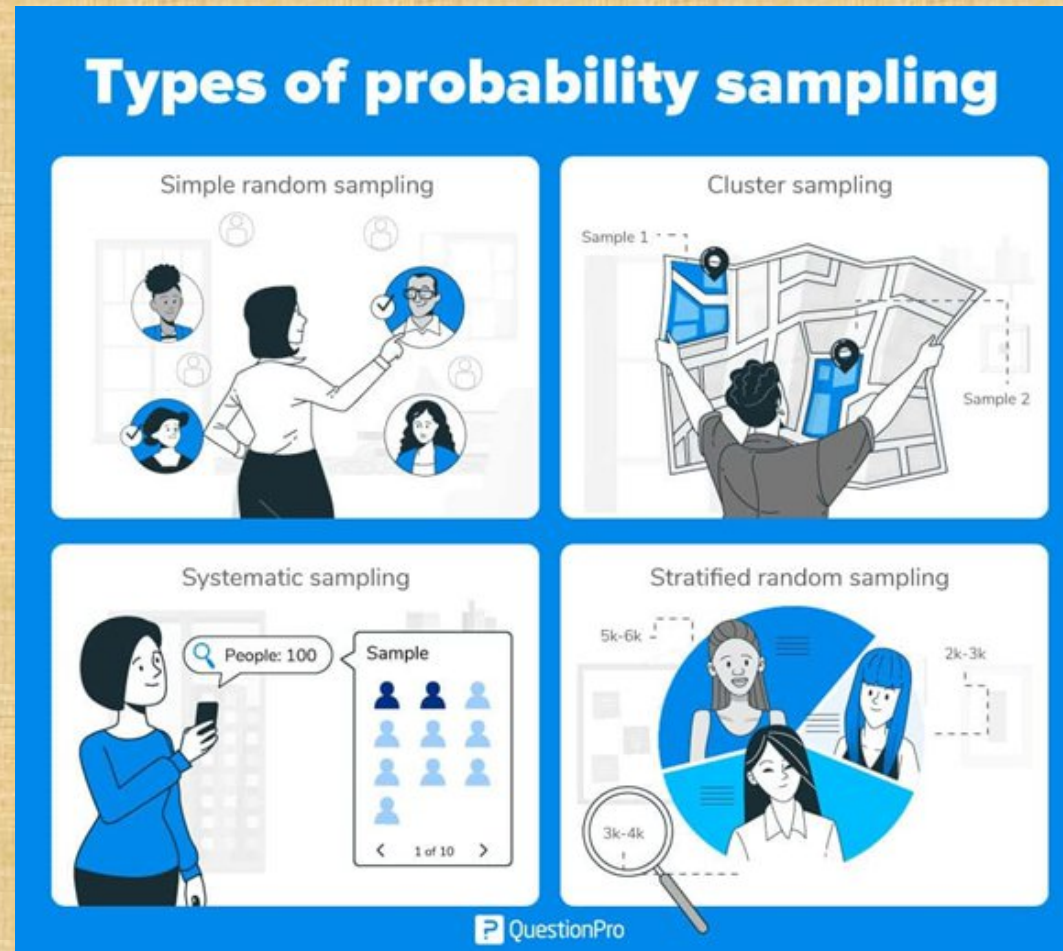
# sampling

- **Types of sampling: sampling methods**

- Sampling in market action research is of two types – probability sampling and non-probability sampling. Let's take a closer look at these two methods of sampling.
- **Probability sampling:** Probability sampling is a sampling technique where a researcher selects a few criteria and chooses members of a population randomly. All the members have an equal opportunity to participate in the sample with this selection parameter.
- **Non-probability sampling:** The non-probability method involves a collection of feedback based on a researcher or statistician's sample selection capabilities and not on a fixed selection process.

# sampling

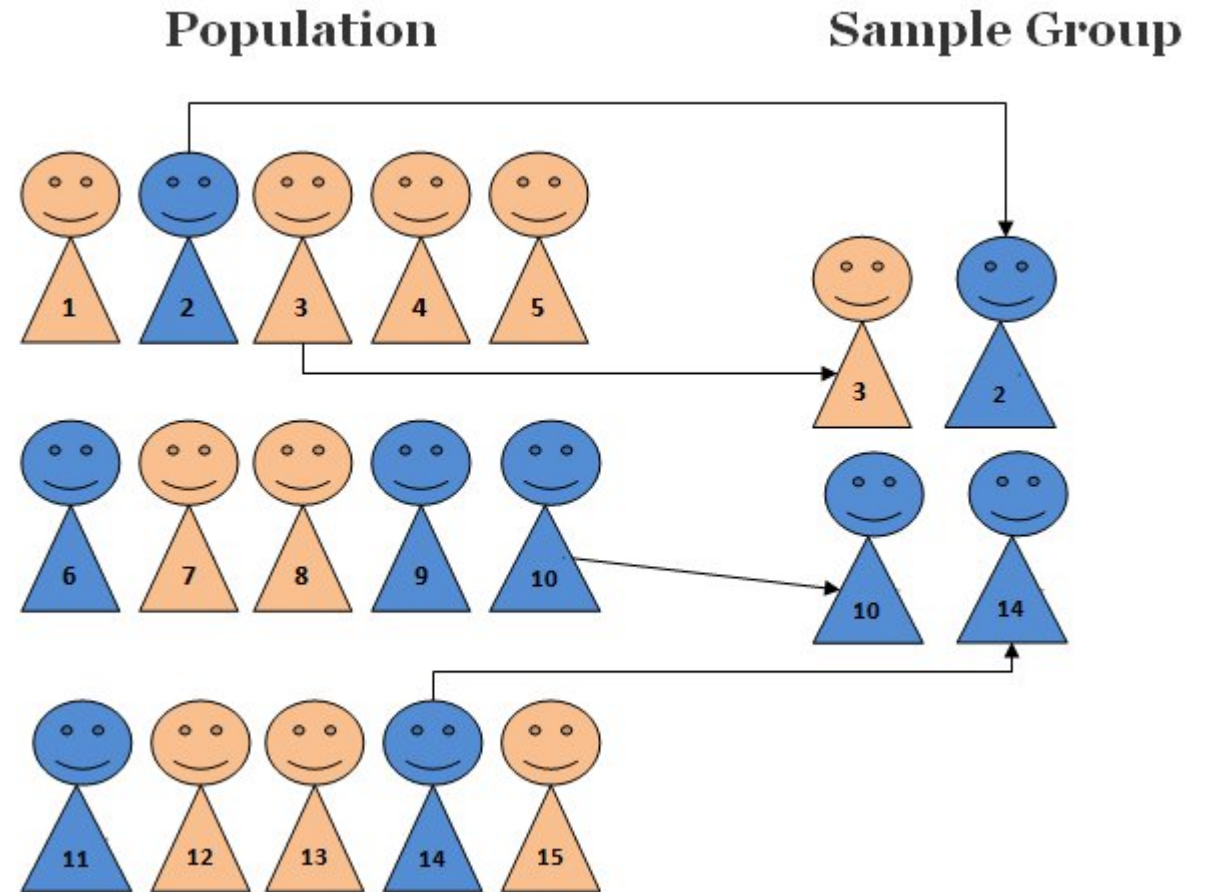
- There are four types of probability sampling techniques:



# sampling

There are four types of probability sampling techniques:

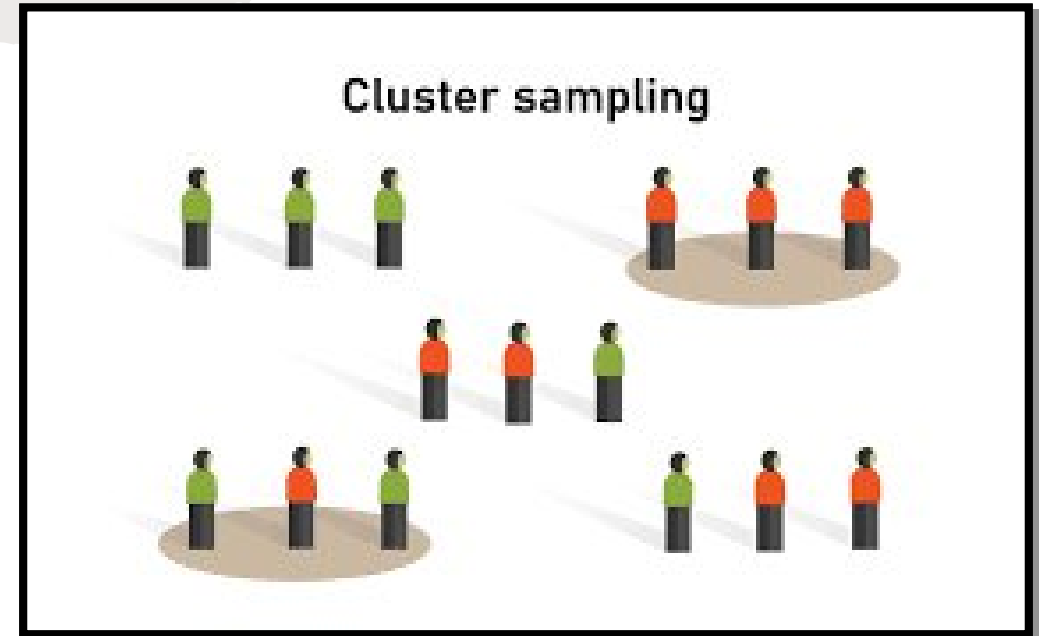
1. **Simple random sampling:** One of the best probability sampling techniques that helps in saving time and resources is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.



# sampling

There are four types of probability sampling techniques:

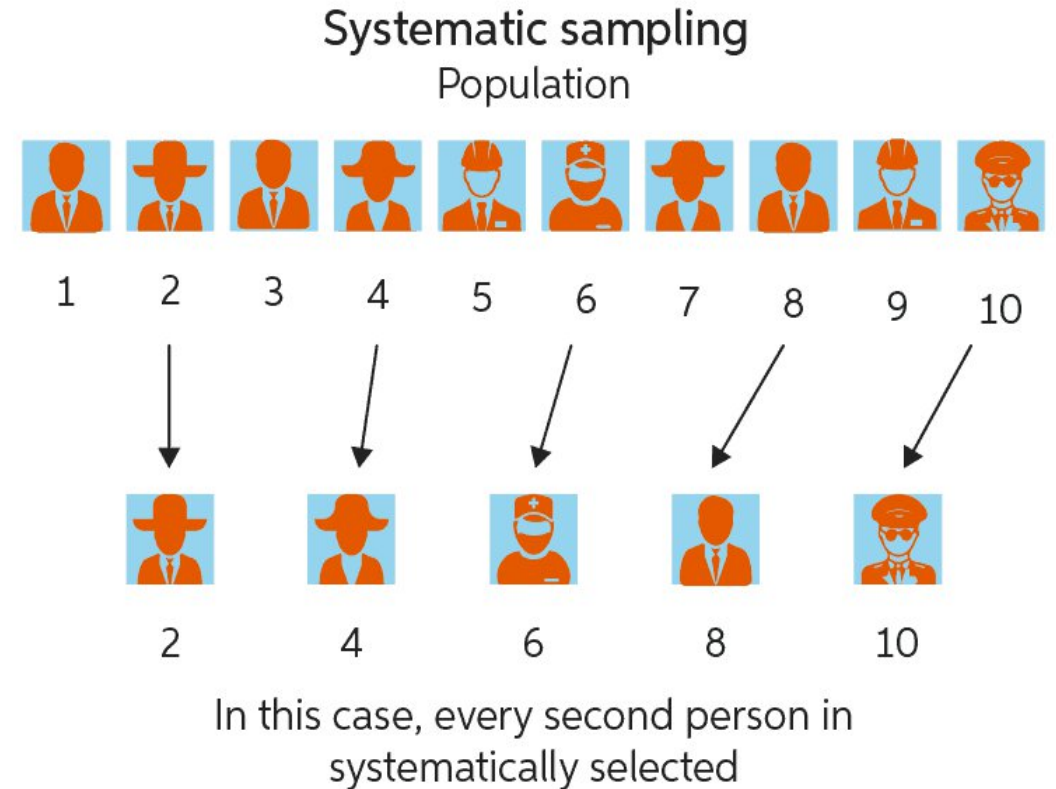
- 2. **Cluster sampling:** Cluster sampling is a method where the researchers divide the entire population into sections or clusters representing a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inferences from the feedback.
- Cluster sampling also involves dividing the population into subgroups, but each subgroup should have characteristics similar to those of the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.



# sampling

There are four types of probability sampling techniques:

3. **Systematic sampling:** Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires selecting a starting point for the sample and determining the sample size that can be repeated at regular intervals. This type of sampling method has a predefined range; hence, this sampling technique is the least time-consuming.



# sampling

There are four types of probability sampling techniques:

4. **Stratified random sampling:** Stratified random sampling is a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population. While sampling, these groups can be organized, and a sample from each group can be drawn separately.





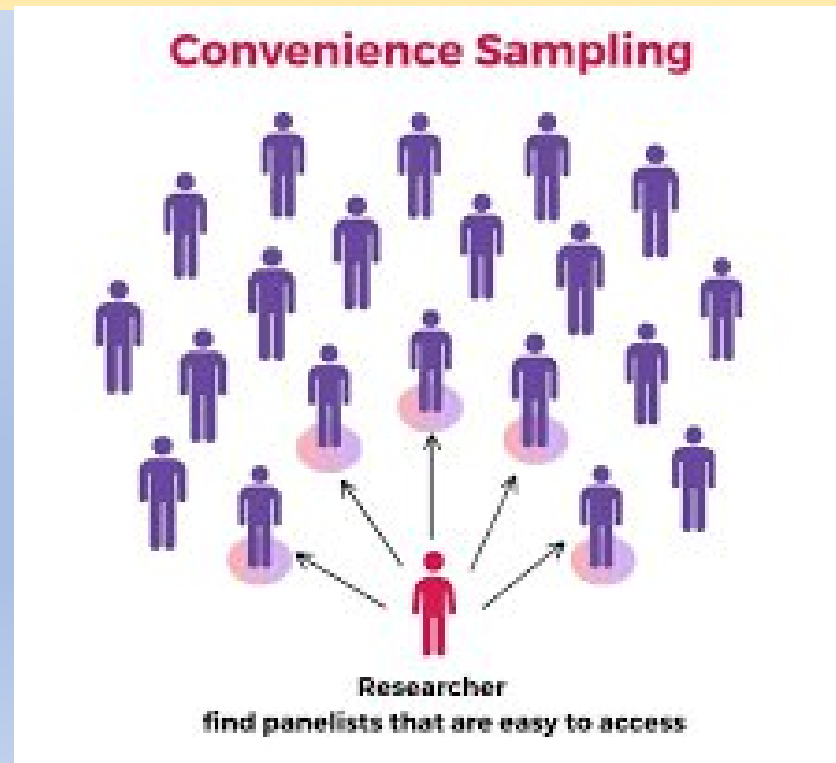
# sampling

- Types of non-probability sampling with examples

The non-probability method is a sampling method that involves a collection of feedback based on a researcher or statistician's sample selection capabilities and not on a fixed selection process. In most situations, the output of a survey conducted with a non-probable sample leads to skewed results, which may not represent the desired target population. However, there are situations, such as the preliminary stages of research or cost constraints for conducting research, where non-probability sampling will be much more useful than the other type.

# sampling

- Types of non-probability sampling with examples
- **Convenience sampling:** This method depends on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling because of the researcher's ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it's purely done based on proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations with resource limitations, such as the initial stages of research, convenience sampling is used.



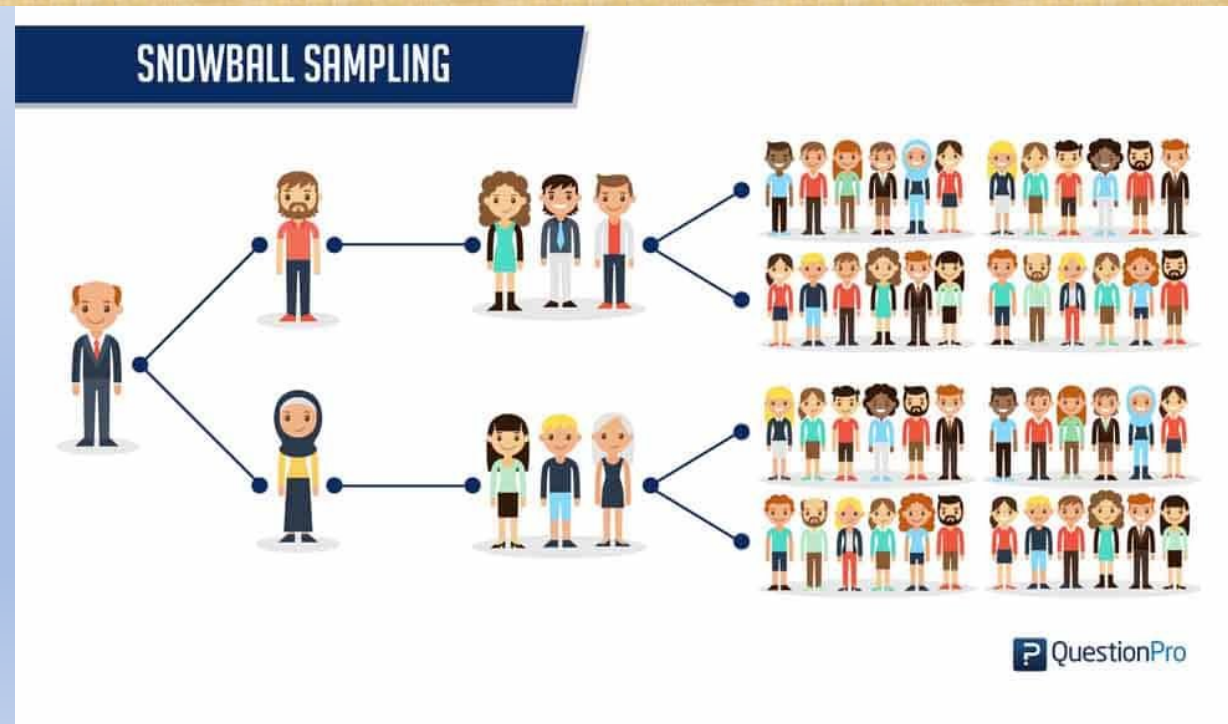
# sampling

- Types of non-probability sampling with examples
- **Judgmental or purposive sampling:** Judgmental or purposive samples are formed at the researcher's discretion. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master's degree. The selection criteria will be: "Are you interested in doing your masters in ...?" and those who respond with a "No" are excluded from the sample.



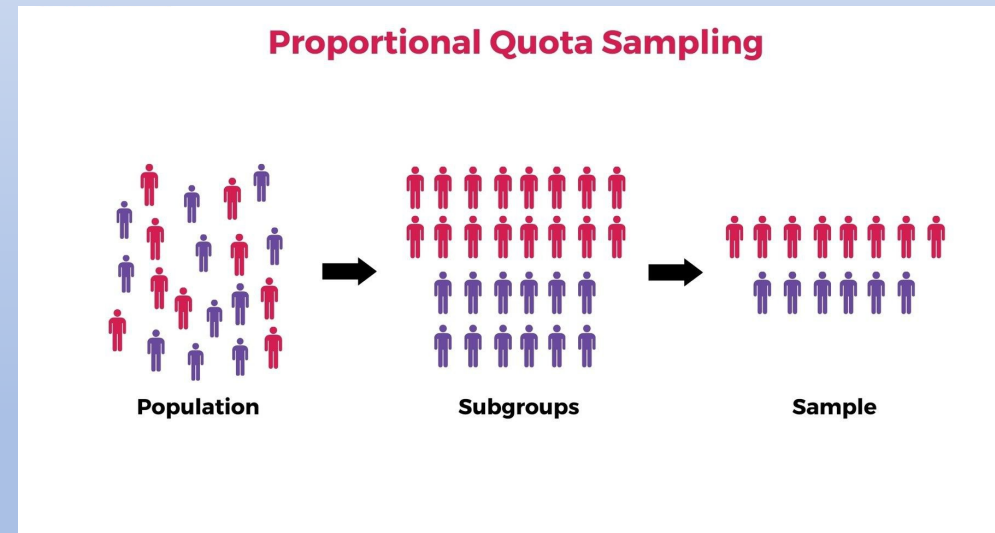
# sampling

**Snowball sampling:** Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, surveying shelterless people or illegal immigrants will be extremely challenging. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method when the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.



# sampling

- Types of non-probability sampling with examples
- **Quota sampling:** In Quota sampling, members in this sampling technique selection happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.



# sampling

- **How do you decide on the type of sampling to use?**
- For any research, it is essential to choose a sampling method accurately to meet the goals of your study. The effectiveness of your 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 your goal.
- Select the method that works best for the research.

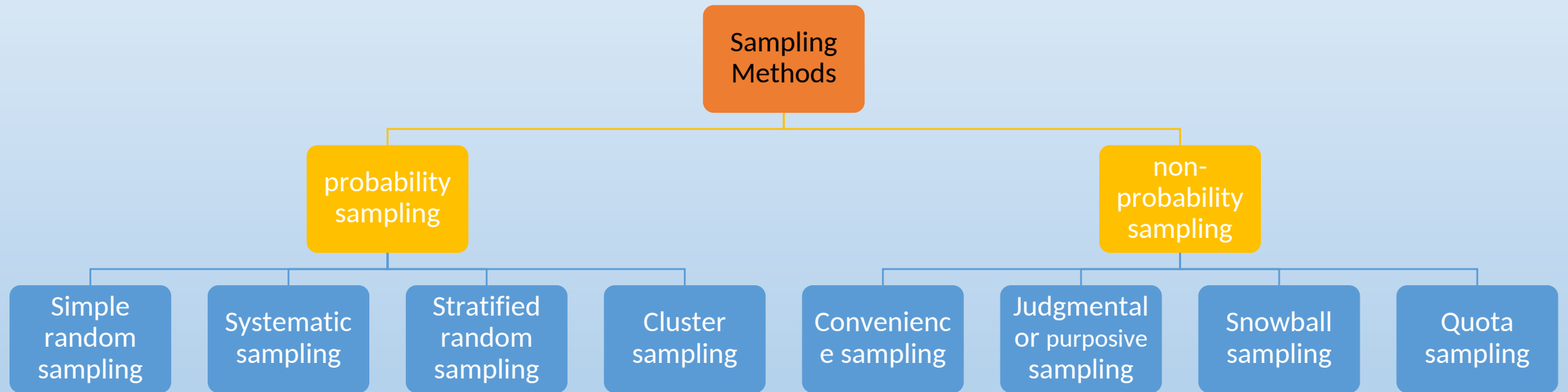
# Sampling

# Comparison

	<b>Probability Sampling Methods</b>	<b>Non-Probability Sampling Methods</b>
Definition	Probability Sampling is a sampling technique in which samples from a larger population are chosen using a method based on the theory of probability.	Non-probability sampling is a sampling technique in which the researcher selects samples based on the researcher's
Alternatively	Random sampling method.	Non-random sampling
Population	The population is selected randomly.	The population is
Nature	The research is conclusive.	The research is
Sample	Since there is a method for deciding the sample, the population demographics are conclusively represented.	Since the sampling method is arbitrary, the population demographics representation is almost
Time Taken	Takes longer to conduct since the research design defines the selection parameters before the market research study begins.	This type of sampling method is quick since neither the sample nor the selection criteria of
Results	This type of sampling is entirely unbiased; hence, the results are also conclusive.	This type of sampling is entirely biased, and hence the results are biased, too, rendering the
Hypothesis	In probability sampling, there is an underlying hypothesis before the study begins, and this method aims to	In non-probability sampling, the hypothesis is derived after



# sampling



# sampling

## Sampling Bias and How to Avoid

It

**Sampling bias** occurs when some members of a population are systematically more likely to be selected in a sample than others. It is also called ascertainment bias in medical fields.

# sampling

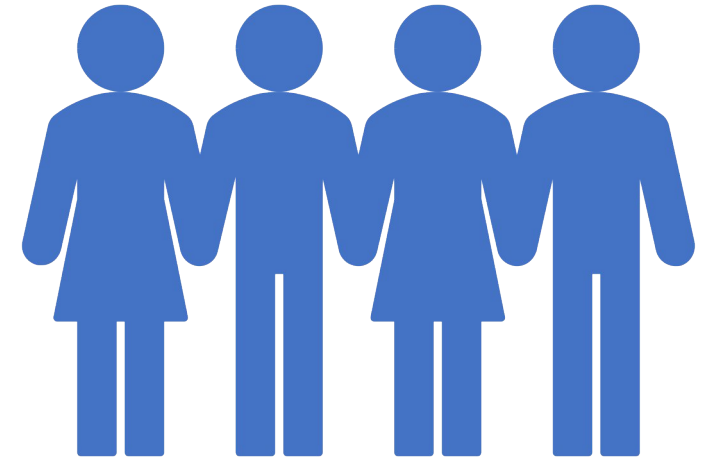
## Types of sampling bias

Type	Explanation	Example
Self-selection bias	People with specific characteristics are more likely to agree to take part in a study than others.	People who are more thrill-seeking are likely to take part in pain research studies. This may skew the data.
Nonresponse bias	People who refuse to participate or drop out from a study systematically differ from those who take part.	In a study on stress and workload, employees with high workloads are less likely to participate. The resulting sample may not vary greatly in terms of workload.
Undercoverage bias	Some members of a population are inadequately represented in the sample.	Administering general national surveys online may miss groups with limited internet access, such as the elderly and lower-income households.
Survivorship bias	Successful observations, people and objects are more likely to be represented in the sample than unsuccessful ones.	In scientific journals, there is strong publication bias towards positive results. Successful research outcomes are published far more often than <a href="#">null</a> findings.
Pre-screening or advertising bias	The way participants are pre-screened or where a study is advertised may bias a sample.	When seeking volunteers to test a novel sleep intervention, you may end up with a sample that is more motivated to improve their sleep habits than the rest of the population. As a result, they may have been likely to improve their sleep habits regardless of the effects of your intervention.
Healthy user bias	Volunteers for preventative interventions are more likely to pursue health-boosting behaviors and activities than other members of the population.	A sample in a preventative intervention has a better diet, higher physical activity levels, abstains from alcohol, and avoids smoking more than most of the population. The experimental findings may be a result of the treatment interacting with these characteristics of the sample, rather than just the treatment itself.

# sampling

## How to avoid or correct sampling bias

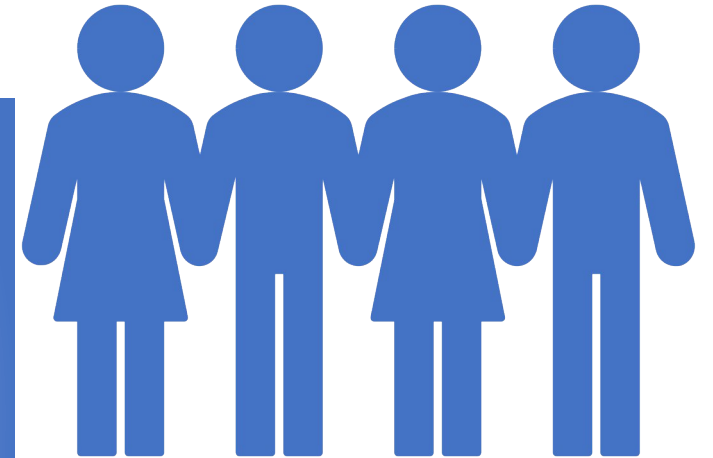
- Using careful research design and sampling procedures can help you avoid sampling bias.
- Define a target population and a sampling frame (the list of individuals that the sample will be drawn from). Match the sampling frame to the target population as much as possible to reduce the risk of sampling bias.
- Make online surveys as short and accessible as possible.
- Follow up on non-responders.
- Avoid convenience sampling.



# sampling

## Post tests

1. What is the difference between Probability sampling and Non-probability sampling?
2. A sample is the term that refers to the group surveyed any time the survey is not administered to all members of the population.(T or F)
3. At a computer facility, every 200th computer chip is inspected for defects. This is an example of what type of sampling method?
  - (a) random sampling
  - (b) systematic sampling
  - (c) stratified sampling
  - (d) cluster sampling



# variable

- Pretest
- What is a variable?
- What is a nominal variable?

# variable

A variable is a characteristic or attribute that can assume different values in different persons, places, or things.

- Example:
- Age,
- Diastolic blood pressure,
- Heart rate,
- The height of adult males,
- The weights of preschool children,
- Gender of Biostatistics students,
- Marital status of instructors at the University of Tikrit,
- Ethnic group of patients



Some examples  
of variables  
include:

Diastolic blood  
pressure,

heart rate,  
height,

The weights,

Stage of bladder  
cancer patients,

# Types of variables

- **Qualitative variable:** a variable or characteristic that cannot be measured in quantitative form but can only be identified by name or categories, for instance, place of birth, ethnic group, type of drug, stages of breast cancer (I, II, III, or IV), degree of pain (minimal, moderate, severe or unbearable).
- **Quantitative variable:** A quantitative variable can be measured and expressed numerically and can be of two types (discrete or continuous). The values of a discrete variable are usually whole numbers, such as the number of episodes of diarrhea in the first five years of life. A continuous variable is a measurement on a constant scale. Examples include weight, height, blood pressure, age, etc.

# Types of variables

*Example of  
Quantitative(Numerical)  
variable:*

*survival time*

*systolic blood  
pressure*

*number of  
children in a  
family*

*height, age,  
body mass  
index.*

# *Numerical variable*

## 1. Discrete Variables

- Have a set of possible values that is either finite or countably infinite.
- The values of a discrete variable are usually whole numbers.
- Numerical discrete data occur when the observations are integers that correspond with a count of some sort

## 2. continuous variables

- A continuous variable has a set of possible values including all values in an interval of the real line.
- No gaps between possible values.
- Each observation theoretically falls somewhere along the continuum.

# Examples of discrete variables

- ❖ Number of pregnancies,
- ❖ The number of bacteria colonies on a plate,
- ❖ The number of cells within a prescribed area upon microscopic examination,
- ❖ The number of heartbeats within a specified time interval,
- ❖ A mother's history of several births ( parity) and pregnancies (gravidity),
- ❖ The number of episodes of illness a patient experiences during some period, etc..

# Examples of Continuous variables

- ❖ Body mass index
- ❖ Height
- ❖ Blood pressure
- ❖ Serum cholesterol level
- ❖ Weigh,
- ❖ Age etc...

# Variables based on Scales(Level) of measurement:

## 1- Nominal:

Only "naming" and classifying observations is possible. When numbers are assigned to categories, it is only for coding purposes and it does not provide a sense of size

### Example:

- ❖ Sex of a person (M, F)
- ❖ eye color (e.g. brown, blue)
- ❖ religion (Muslim, Christian)
- ❖ place of residence (urban, rural) etc



# Variables based on Scales(Level) of measurement:

## 2. Ordinal:

- **Categorization and ranking (ordering) observations are possible We can talk of greater than or less than and it conveys meaning to the value but;**
- **It is impossible to express the real difference between measurements in numerical terms**

### Example:

- Socio-economic status (very low, low, medium, high, very high)**
- severity(mild, moderate, severe)**
- blood pressure (very low, low, high, very high) etc.**

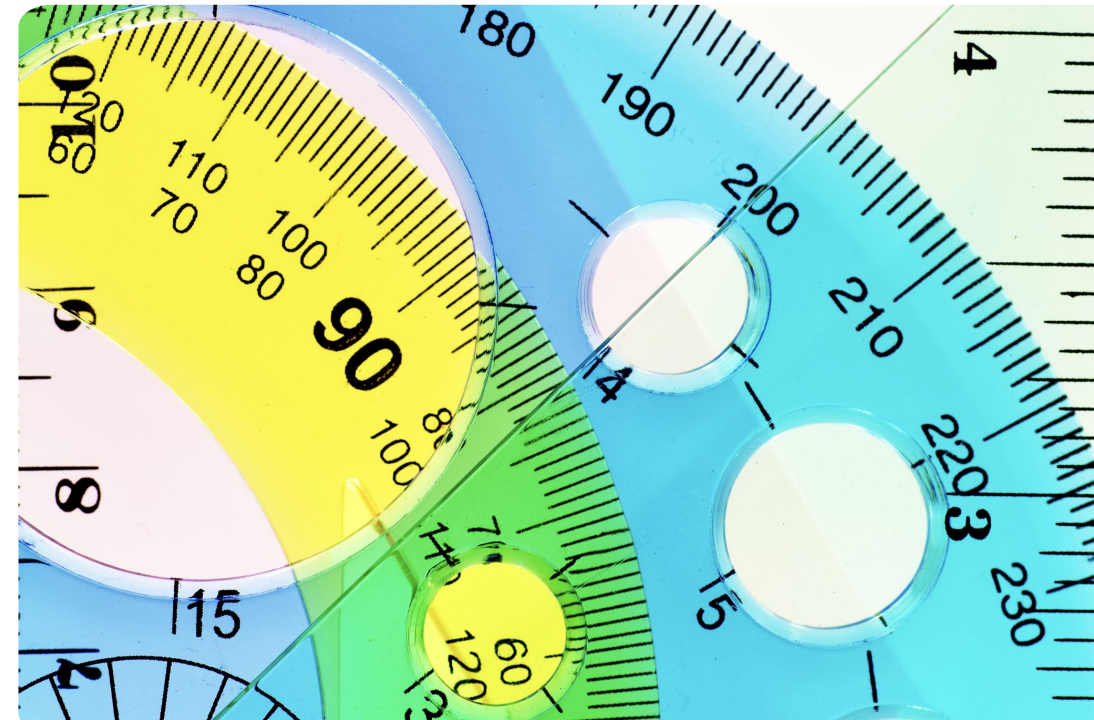
# Variables based on Scales(Level) of measurement:

## 3. Interval:

- Possible to categorize, rank, and tell the real distance between any two measurements
- Zero is not absolute

### Example:

- Body temperature in degrees F. and Celsius (measured in degrees).
- It is a meaningful difference



# Variables based on Scales(Level) of measurement:

## 4. Ratio:

- The highest level of measurement scale, characterized by the fact that equality of ratios, as well as equality of intervals, can be determined
- There is a true zero point. i.e. zero is absolute

### Example:

- volume
- height
- weight
- length
- time until death, etc...

# Variables

## Qualitative variable

### Nominal

e.g. Gender, ethnic

### Ordinal

## Quantitative variable

### Interval

e.g. Temp. F, C

### Ratio:

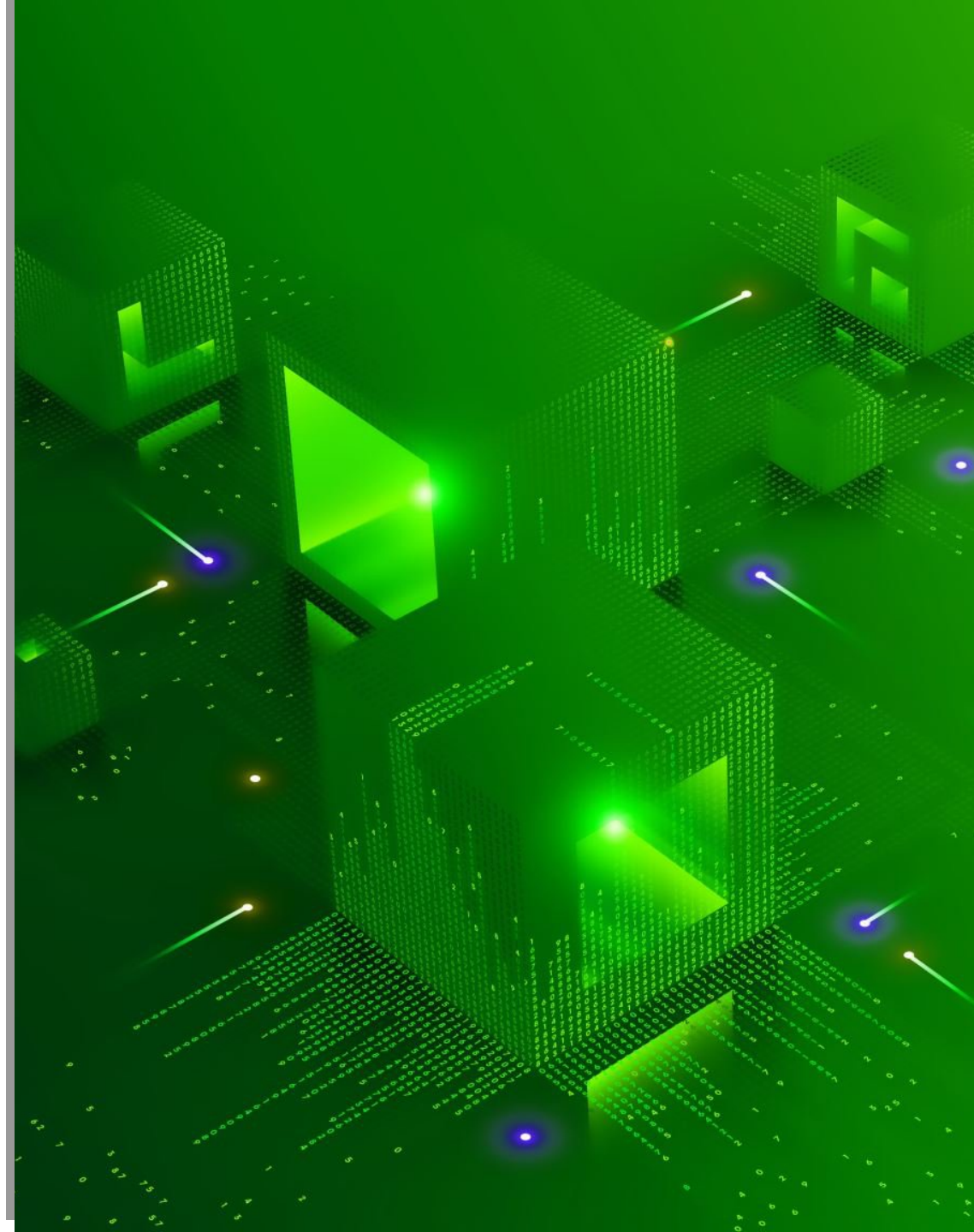
Operation	Nominal	Ordinal	Interval	Ratio
Equality	✓	✓	✓	✓
Order		✓	✓	✓
Add / subtract			✓	✓
Multiply / divide				✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Arithmetic mean			✓	✓
Geometric mean				✓

# variable

- Post Test
- What are the differences between discrete and continuous variables? Give some Examples
- What are the types of variables based on scale measurement?

## Methods of data organization and presentation

- The data collected in a survey is called **raw data**. In most cases, useful information is not immediately evident from the mass of unsorted data. Collected data need to be organized in such a way as to condense the information they contain in a way that will show patterns of variation clearly. Precise methods of analysis can be decided upon only when the characteristics of the data are understood. For the primary objective of this different techniques of data organization and presentation like **order array**, **tables**, and diagrams are used.





## Methods of data organization and presentation

### **Array (ordered array)**

- is a serial arrangement of numerical data in an ascending or descending order. This will enable us to know the range over which the items are spread and will also get an idea of their general distribution. The ordered array is an appropriate way of presentation when the data are small in size (usually less than 20).

## Methods of data organization and presentation

### Array (ordered array)

- EXAMPLE: Here are the ages of ten persons 10, 35, 47, 7, 13, 65, 55, 18, 6, 70
- **Ordered Array: 6,7,10,13,18,35,47,55,65,70**

## Methods of data organization and presentation

### Frequency Distributions

- For data to be more easily appreciated and to draw quick comparisons, it is often useful to arrange the data in the form of a table or one of several different graphical forms.
- When analyzing voluminous data collected from say, a health center's records, it is quite useful to put them into compact tables. Quite often, the presentation of data in a meaningful way is done by preparing a frequency distribution. If this is not done the raw data will not present any meaning and any pattern in them (if any) may not be detected.

## Methods of data organization and presentation

### Array (ordered array)

- EXAMPLE:
- A study in which 400 persons were asked how many full-length movies they had seen on television during the preceding week. The following gives the distribution of the data collected.

## Methods of data organization and presentation

Number of movies	Number of persons	Relative frequency (%)
0	72	18.0
1	106	26.5
2	153	38.3
3	40	10.0
4	18	4.5
5	7	1.8
6	3	0.8
7	0	0.0
8	1	0.3
Total	400	100.0

$$\text{Relative Frequency} = \frac{\text{Frequency of Classes}}{\text{Total Frequency}} * 100$$

In the above distribution Number of movies represents the **variable** under consideration, the Number of persons represents the **frequency**, and the whole distribution is called **frequency distribution**, particularly **simple frequency distribution**.

## Methods of data organization and presentation

### A categorical distribution

non-numerical information can also be represented in a frequency distribution. Seniors of a high school were interviewed on their plans after completing high school. The following data give plans of 548 seniors of a high school.

Seniors' plan	Number of seniors
Plan to attend college	240
May attend college	146
Plan to or may attend a vocational school	57
Will not attend any school	105
Total	548

# Methods of data organization and presentation

- Consider the problem of a social scientist who wants to study the age of persons arrested in a country. In connection with large sets of data, a good overall picture and sufficient information can often be conveyed by grouping the data into several class intervals as shown below.

<u>Age (years)</u>	<u>Number of persons(Frequency)</u>	<u>Relative Frequency %</u>
• Under 18	1,748	
• 18 - 24	3,325	
• 25 - 34	3,149	
• 35 - 44	1,323	
• 45 - 54	512	
• 55 and over	335	
• Total	10,392	

- This kind of frequency distribution is called **grouped frequency distribution**. Frequency distributions present data in a relatively compact form, give a good overall picture and contain information that is adequate for many purposes, but there are usually some things that can be determined only from the original data. For instance, the above-grouped frequency distribution cannot tell how many of the arrested persons are 19 years old, or how many are over 62.



## Methods of data organization and presentation

---

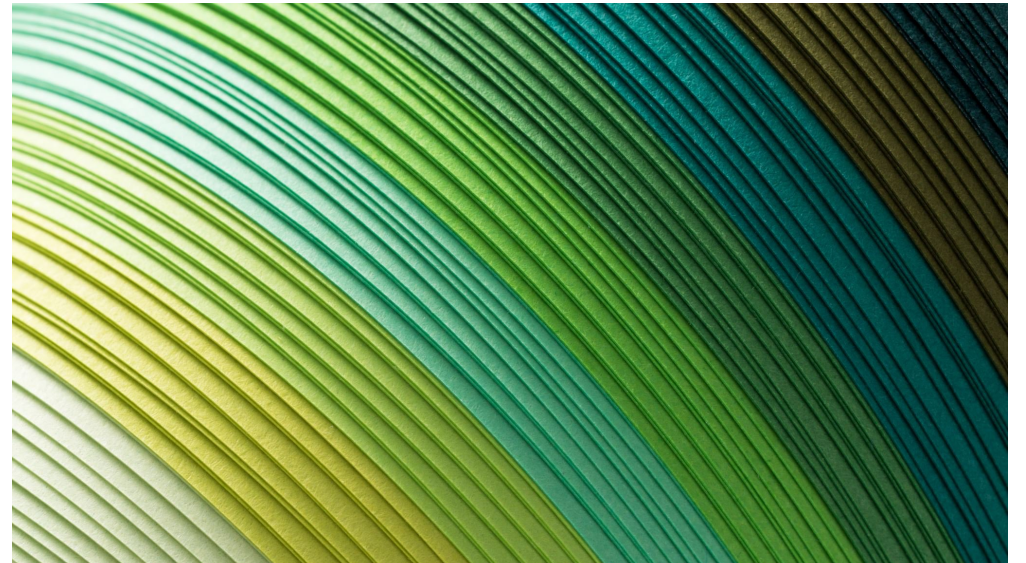
The construction of **grouped frequency distribution** consists essentially of four steps:

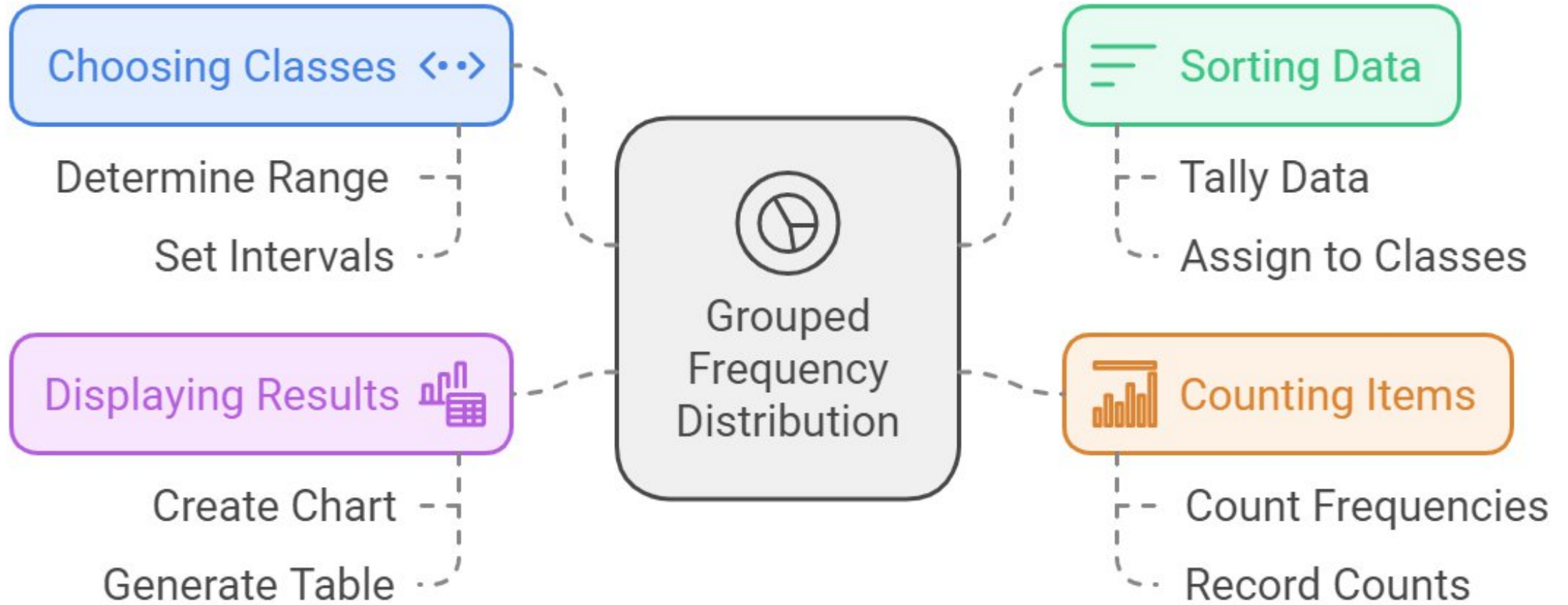
---

(1) Choosing the classes, (2) sorting (or tallying) the data into these classes, (3) counting the number of items in each class, and (4) displaying the results in the form of a chart or table

---

Choosing suitable classification involves choosing the **number of classes** and the **range of values each class** should cover, namely, from where to where each class should go. Both of these choices are arbitrary to some extent, but they depend on the nature of the data and its accuracy, and on the purpose the distribution is to serve.





## Methods of data organization and presentation

The following are some rules that are generally observed:

1) We seldom use fewer than 6 or more than 20 classes, and 15 generally is a good number, the exact number we use in a given situation depends mainly on the number of measurements or observations we have to group

A guide on the determination of the number of classes (k) can be the Sturge's Formula, given by:  
 $K = 1 + 3.322 \times \log(n)$ , where n is the number of observations

**Determine the overall range R :**

$$R = XL - XS$$

Where XS is the smallest value and XL is the biggest value.

And the length or width of the class interval (w) can be calculated by:

$$W = (\text{Maximum value} - \text{Minimum value}) / K = \text{Range} / K$$

## Methods of data organization and presentation

**The following are some rules that are generally observed:**

2) We always make sure that each item (measurement or observation) goes into one and only one **class**, i.e. classes should be mutually exclusive. To this end we must make sure that the smallest and largest values fall within the classification, that none of the values can fall into possible gaps between successive classes, and that the classes do not overlap, namely, that successive classes have no values in common.

Note that the Sturges rule should not be regarded as final, but should be considered as a guide only. The number of classes specified by the rule should be increased or decreased for convenient or clear presentation.

**The following are some rules that are generally observed:**

### **3) Determination of class limits:**

(i) Class limits should be definite and clearly stated. In other words, open-ended classes should be avoided since they make it difficult, or even impossible, to calculate certain further descriptions that may be of interest. These are classes like less than 10, greater than 65, and so on.

(ii) The starting point, i.e., the lower limit of the first class be determined in such a manner that the frequency of each class gets concentrated near the middle of the class interval. This is necessary because in the interpretation of a frequency table and subsequent calculations based on it, the mid-point of each class is taken to represent the value of all items included in the frequency of that class.

It is important to watch whether they are given to the nearest inch or the nearest tenth of an inch, whether they are given to the nearest ounce or the nearest hundredth of an ounce, and so forth. For instance, to group the weights of certain animals, we could use the first of the following three classifications if the weights are given to the nearest kilogram, the second if the weights are given to the nearest tenth of a kilogram, and the third if the weights are given to the nearest hundredth of a kilogram:



## Methods of data organization and presentation

The following are some rules that are generally observed:

It is important to watch whether they are given to the nearest inch or the nearest tenth of an inch, whether they are given to the nearest ounce or the nearest hundredth of an ounce, and so forth. For instance, to group the weights of certain animals, we could use the first of the following three classifications if the weights are given to the nearest kilogram, the second if the weights are given to the nearest tenth of a kilogram, and the third if the weights are given to the nearest hundredth of a kilogram:

Weight (kg)	Weight (kg)	Weight (kg)
10 - 14	10.0 - 14.9	10.00 - 14.99
15 - 19	15.0 - 19.9	15.00 - 19.99
20 - 24	20.0 - 24.9	20.00 - 24.99
25 - 29	25.0 - 29.9	25.00 - 29.99
30 - 34	30.0 - 34.9	30.00 - 34.99

# Methods of data organization and presentation

23	24	18	14	20	24	24	26	23	21
16	15	19	20	22	14	13	20	19	27
29	22	38	28	34	32	23	19	21	31
16	28	19	18	12	27	15	21	25	16
30	17	22	29	29	18	25	20	16	11
17	12	15	24	25	21	22	17	18	15
21	20	23	18	17	15	16	26	23	22
11	16	18	20	23	19	17	15	20	10

- Example: Construct a grouped frequency distribution of the following data on the amount of time (in hours) that 80 college students devoted to leisure activities during a typical school week:



## Methods of data organization and presentation

Using the above formula,  $K = 1 + 3.322 * \log (80) = 7.32 = 7$  classes

Maximum value = 38 and Minimum value = 10  $\rightarrow$  Range = 38 - 10 = 28 and  $W = 28/7 = 4$

Using a width of 4, we can construct a grouped frequency distribution for the above data as:

Time spent (hours) Classes	Tally	Frequency	Cumulative freq.	Relative Frequency
10-13		6	6	7.5
14-17		19	25	23.75
18-21		23	48	28.75
22-25		18	66	22.5
26-29		9	75	11.25
30-33		3	78	3.75
34-37		1	79	1.25
38-41		1	80	1.25

## Methods of data organization and presentation

**Cumulative and Relative Frequencies:** When frequencies of two or more classes are added up, such total frequencies are called Cumulative Frequencies. These frequencies help to find the total number of items whose values are less than or greater than some value. On the other hand, relative frequencies express the frequency of each value or class as a percentage of the total frequency.

**Note.** In the construction of cumulative frequency distribution, if we start the cumulation from the lowest size of the variable to the highest size, the resulting frequency distribution is called 'Less than cumulative frequency distribution', and if the cumulation is from the highest to the lowest value the resulting frequency distribution is called 'more than cumulative frequency distribution.' The most common cumulative frequency is the less than cumulative frequency.

### Mid-Point of a class interval and the determination of Class Boundaries

Mid-point or class mark ( $X_c$ ) of an interval is the value of the interval which lies mid-way between the lower true limit (LTL) and the upper true limit (UTL) of a class. It is calculated as:

$$x_c = \frac{\text{upper class limit} - \text{Lower class Limit}}{2}$$

True limits (or class boundaries) are those limits, that are determined mathematically to make an interval of a continuous variable continuous in both directions, and no gap exists between classes. The true limits are what the tabulated limits would correspond with if one could measure exactly.

## Methods of data organization and presentation

Example: Frequency distribution of weights (in Ounces) of Malignant Tumors Removed from the Abdomen of 57 subjects

Weight	Class boundaries	Xc	Freq.	Cum. freq.	Relative freq (%)
10-19	9.5 - 19.5	14.5	5	5	0.0877
20-29	19.5 - 29.5	24.5	19	24	0.3333
30-39	29.5 - 39.5	34.5	10	34	0.1754
40-49	39.5 - 49.5	44.5	13	47	0.2281
50-59	49.5 - 59.5	54.5	4	51	0.0702
60-69	59.5 - 69.5	64.5	4	55	0.0702
70-79	69.5 - 79.5	74.5	2	57	0.0352
Total			57		1.0000

Note: The width of a class is found from the true class limit by subtracting the true lower limit from the upper true limit of any particular class.

For example, the width of the above distribution is (let's take the fourth class)  $w = 49.5 - 39.5 = 10$ .

## Methods of data organization and presentation

### Posttest

EXA: The pulse rate of 50 adult individuals is listed below. Create a frequency table summarizing the number?

60	80	50	45	49	48	35	55	91	50
65	74	54	40	42	80	90	65	67	45
80	71	61	38	55	84	88	79	81	48
74	40	68	52	62	76	81	82	64	38
62	45	59	42	60	66	70	60	68	29

# Measures of Dispersion

Understanding Variability in Data

# Introduction to Measures of Dispersion

- **Definition**: Measures of Dispersion describe the spread or variability within a data set.
- **Purpose**: Helps understand how data points differ from the central tendency (mean, median, etc.).
- **Why it Matters**: Knowing the spread of data aids in comparing data sets, making informed decisions, and identifying outliers.

# Types of Measures of Dispersion

- 1. Range
- 2. Variance
- 3. Standard Deviation
- 4. Interquartile Range (IQR)
- 5. Mean Absolute Deviation



# Range (Simplest Measure)

- Definition: Difference between the maximum and minimum values in a data set.
- Formula:  $\text{Range} = \text{Maximum} - \text{Minimum}$
- Example: Data set: 4, 8, 15, 16, 23, 42
- - Maximum = 42, Minimum = 4
- - Range =  $42 - 4 = 38$
- Advantages: Easy to calculate.
- Disadvantages: Sensitive to outliers; doesn't give info on distribution.

# Variance and Standard Deviation

- Variance:
  - - Definition: Average squared difference from the mean, giving a measure of data spread.
  - - Formula (Sample):  $s^2 = \Sigma(x - \bar{x})^2 / (n - 1)$
- Standard Deviation:
  - - Definition: The square root of variance, represents spread in original units.
  - - Formula:  $s = \sqrt{s^2}$

# Example: Standard Deviation from Frequency Table

- Given frequency table of test scores:

Score (x)	Frequency (f)
10	3
20	7
30	4
40	6
50	5

- Calculations:
- $\Sigma(x * f) = (10 * 3) + (20 * 7) + (30 * 4) + (40 * 6) + (50 * 5) = 780$
- $\Sigma f = 3 + 7 + 4 + 6 + 5 = 25$
- $\bar{x} = 780 / 25 = 31.2$

Step 2: Calculate Each  $(x - \bar{x})^2 * f$

For each score, calculate the squared deviation from the mean, multiply by the frequency, and then sum these values:

Score (x)	Frequency (f)	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 \cdot f$
10	3	-21.2	449.44	1348.32
20	7	-11.2	125.44	878.08
30	4	-1.2	1.44	5.76
40	6	8.8	77.44	464.64
50	5	18.8	353.44	1767.2

$$\Sigma((x - \bar{x})^2 * f) = 4464$$

Step 3: Calculate the Variance and Standard Deviation

The variance ( $s^2$ ) for the sample is:

$$s^2 = \Sigma((x - \bar{x})^2 * f) / (\Sigma f - 1) = 4464 / 24 = 186$$

The standard deviation ( $s$ ) is the square root of the variance:

$$s = \sqrt{186} \approx 13.64$$

### **Summary**

- Mean ( $\bar{x}$ ): 31.2
- Variance ( $s^2$ ): 186
- Standard Deviation ( $s$ ): 13.64

# Pot Test Problem

Consider the following frequency table with age ranges:

Age Range (x)	Frequency (f)
10 - 19	5
20 - 29	8
30 - 39	12
40 - 49	10
50 - 59	4
60 - 69	6

**Calculate the Standard Deviation.**

## Solution Steps

1. **Calculate the Midpoint** for each age range:

- For example, for 10-19, the midpoint is  $(10 + 19)/2 = 14.5$ .
- Midpoints for all ranges: 14.5, 24.5, 34.5, 44.5, 54.5, 64.5.

2. **Calculate the Mean:**

- Use the formula  $\bar{x} = \frac{\sum(x \cdot f)}{\sum f}$ , where  $x$  is the midpoint.

3. **Calculate Each  $(x - \bar{x})^2 \cdot f$ :**

- Subtract the mean from each midpoint, square the result, multiply by the frequency, and then sum all values.

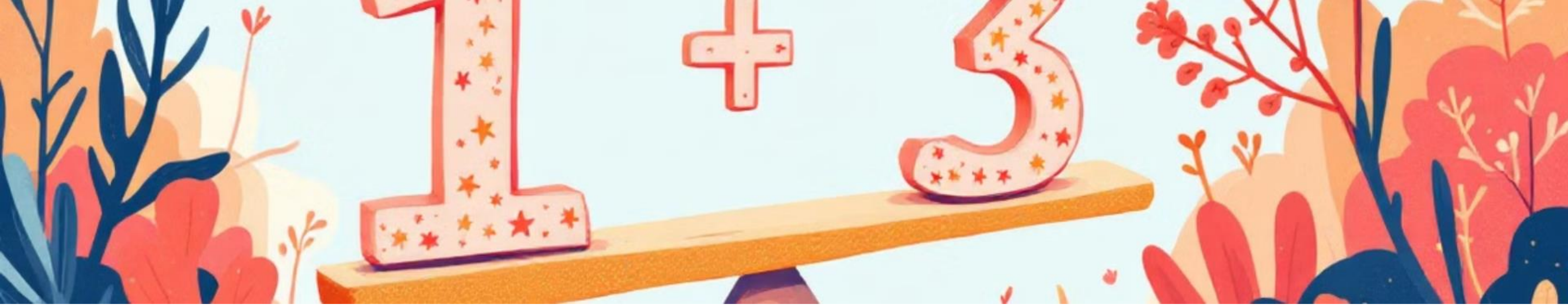
# Measures of Central Tendency

Measures of central tendency are essential statistical tools. They help summarize data sets by identifying a single value that represents the center or typical value. These measures include the mean, median, and mode.



م.د. غسان جاسم هادي





# Mean

## Definition

The mean is the average of all values in a data set. It's calculated by summing all values and dividing by the count of values.

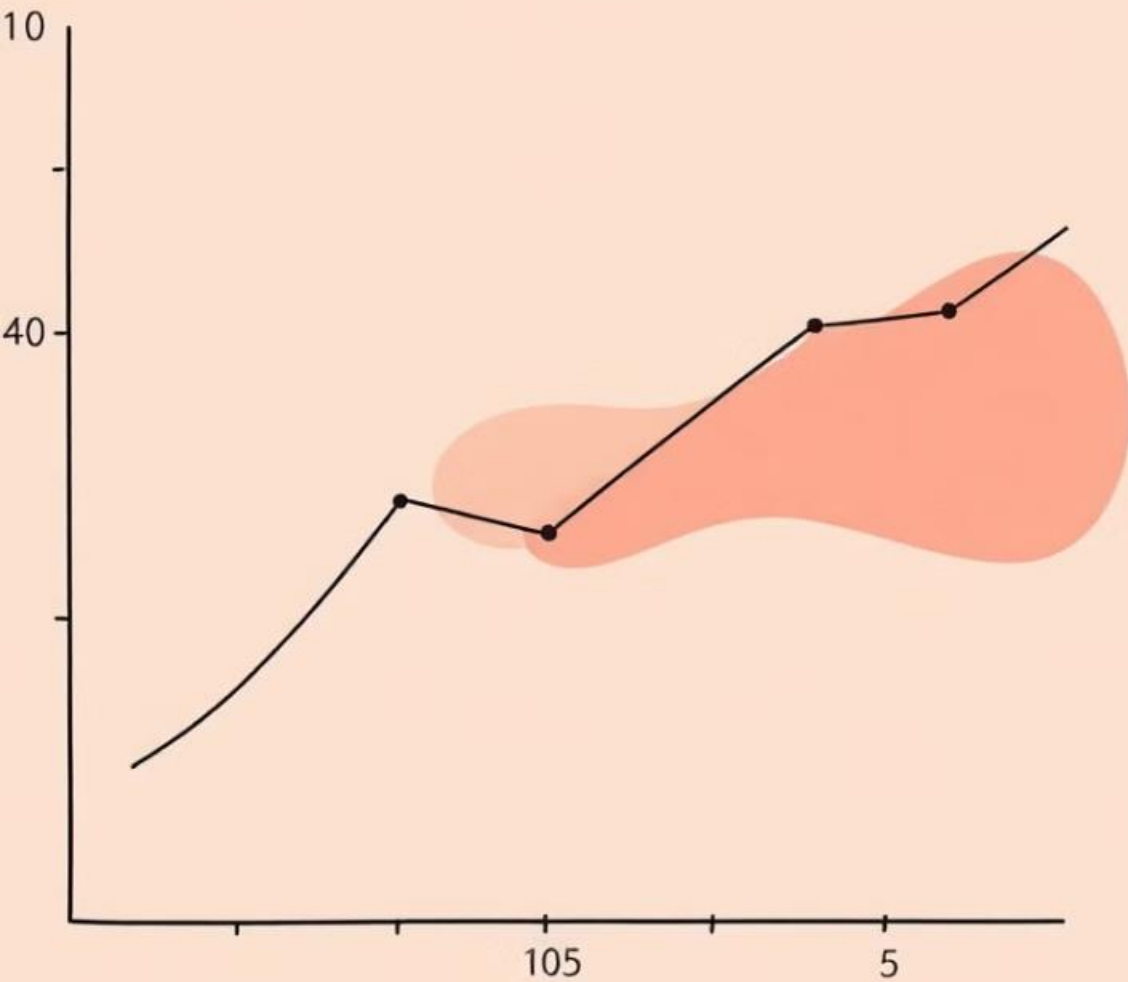
## Sensitivity

The mean is sensitive to extreme values or outliers. This can skew the representation of the data's center.

## Usage

It's widely used in various fields, from economics to science. The mean provides a quick summary of data.

# Median



1

## Definition

The median is the middle value when a data set is ordered from least to greatest.

2

## Robustness

It's less affected by outliers, making it useful for skewed distributions.

3

## Application

Often used in income and housing price statistics to avoid skew from extreme values.

# Mode



## Frequency

The mode is the value that appears most frequently in a data set.



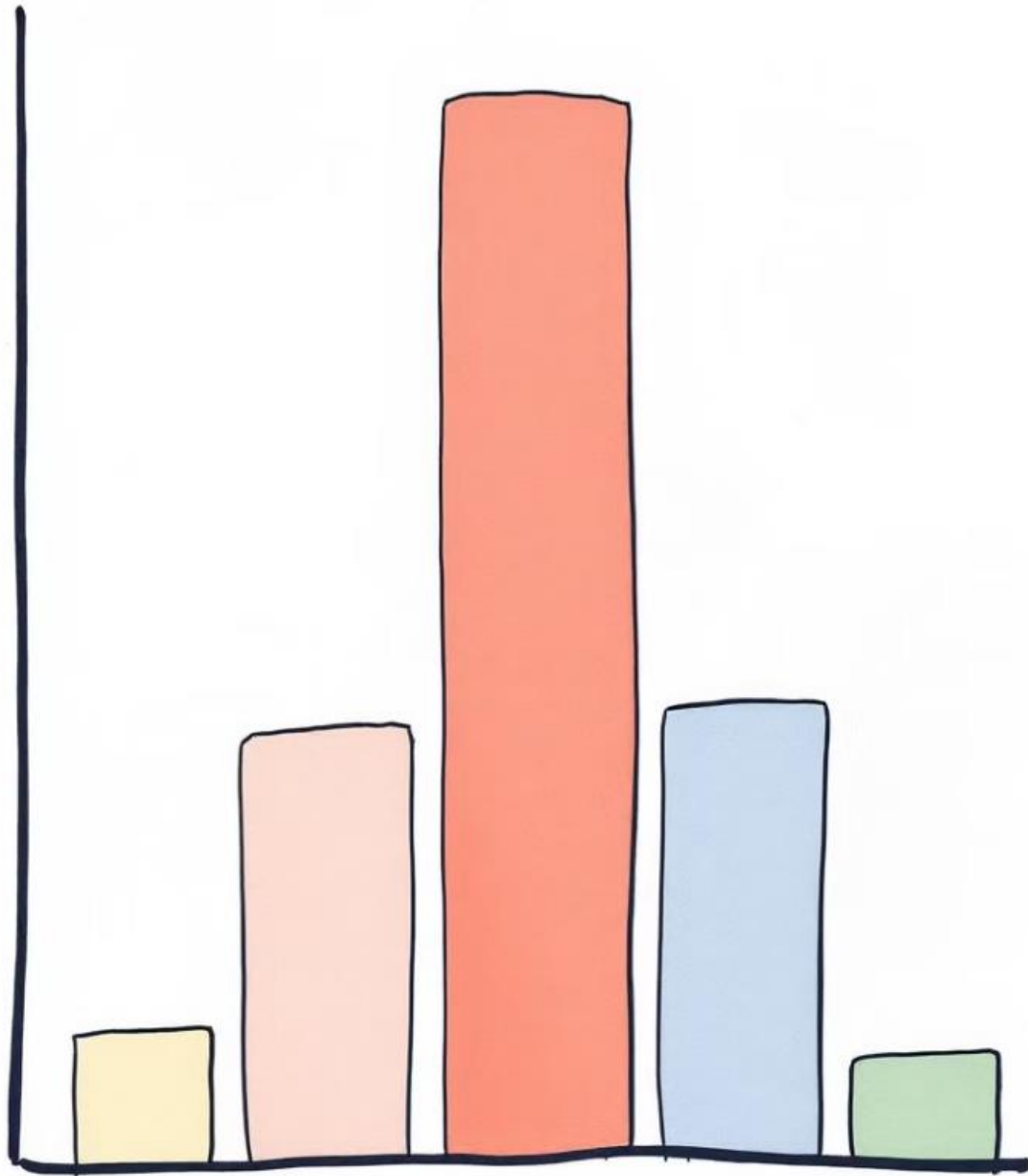
## Multiple Modes

A data set can have more than one mode if multiple values occur with equal frequency.



## Categorical Data

The mode is particularly useful for non-numeric or categorical data.



# Calculating the Mean

1

## Sum Values

Add up all the values in your data set.

2

## Count Values

Determine the total number of values in your data set.

3

## Divide

Divide the sum by the count to get the mean.



# Calculating the Mean

$$\bullet \bar{X} = \frac{1}{n} \sum X_i$$

**EXA: Given the following data: 5,10,15,20,25**

**To calculate the mean:**

$$(5+10+15+20+25)/5=75/5=15$$

**Thus, the mean is 15.**



# Calculating the Weighted Mean:

A **weighted mean** is a kind of average. Instead of each data point contributing equally to the final mean, some data points contribute more “weight” than others. If all the weights are equal, then the weighted mean equals the arithmetic mean (the regular “average” you’re used to). Weighted means are very common in statistics, especially when studying populations.

$$\bullet \bar{X} = \frac{\sum w_i * x_i}{\sum w_i}$$

- $\Sigma$  = **summation** (in other words...add them up!).
- $w$  = the weights.
- $x$  = the value.

# Calculating the Weighted Mean:

Scenario: Suppose a researcher is studying the average number of hours per week patients with different chronic conditions engage in physical exercise

Condition	Average Hours per Week	Number of Patients
Diabetes	4	30
Hypertension	5	50
Heart Disease	3	20

Here, the number of patients in each group serves as the weight, since the groups have different sizes.

# Calculating the Weighted Mean:

Solution: To calculate the weighted mean, use the formula:

$$\bar{X} = \frac{\sum w_i * x_i}{\sum w_i}$$

Where  $x_i$  is the average hours per week for each group,

$w_i$  is the number of patients in each group,

$\sum(x \cdot w)$  is the sum of the products of the averages and weights (number of patients),

$\sum w$  is the sum of the weights (total number of patients).



**Step 1: Multiply each group's average by its number of patients:**

- For Diabetes:  $4 \times 30 = 120$
- For Hypertension:  $5 \times 50 = 250$
- For Heart Disease:  $3 \times 20 = 60$

**Step 2: Find the sum of these products:**

$$120 + 250 + 60 = 430$$

**Step 3: Find the total number of patients (sum of the weights):**

$$30 + 50 + 20 = 100$$

**Step 4: Calculate the weighted mean:**

$$\text{Weighted Mean} = 430 / 100 = 4.3$$

$$\text{Weighted Mean} = 4.3$$



# Mean for grouped Data

## Example: Mean for Grouped Data in Biostatistics

In biostatistics, calculating the mean for grouped data involves estimating the central value of a distribution where data points are grouped into intervals or classes. The formula for the mean of grouped data is:

$$\text{Mean}(\bar{x}) = \frac{\sum(f \cdot x_m)}{\sum f}$$

Where:

- $f$  is the frequency of each class,
- $x_m$  is the midpoint of each class (class mark),
- $\sum f$  is the total number of observations (sum of frequencies),
- $\sum(f \cdot x_m)$  is the sum of the products of class frequencies and midpoints.

# Mean for grouped Data

## Example Scenario:

Suppose a study records the blood pressure levels of patients, grouped into intervals, and the frequencies of patients within each interval are provided as follows:

Blood Pressure (mmHg)	Frequency (f)
110-119	3
120-129	5
130-139	12
140-149	8
150-159	2

# Mean for grouped Data

## Step 3: Calculate the Mean

To find the mean, sum up the products of the frequencies and midpoints and divide by the total number of observations:

$$\Sigma(f \cdot x_m) = 343.5 + 622.5 + 1614.0 + 1156.0 + 309.0 = 4045.0$$

The total frequency ( $\Sigma f$ ) is:

$$\Sigma f = 3 + 5 + 12 + 8 + 2 = 30$$

Now, calculate the mean:

$$\text{Mean}(\bar{x}) = 4045.0 / 30 = 134.83 \text{ mmHg}$$

## Conclusion:

The mean blood pressure of the patients, based on the grouped data, is **134.83 mmHg**.





# Calculating the Median

- 1 Order Data**  
Arrange all values in the data set from lowest to highest.
- 2 Find Middle**  
For odd-numbered sets, select the middle value. For even-numbered sets, average the two middle values.
- 3 Identify Median**  
The selected or calculated middle value is the median.



## Calculating the Median

Example: Consider the following data, which consists of white blood counts taken on admission of all patients entering a small hospital on a given day. Compute the median white-blood count ( $\times 10^3$ ). 7,


35,5,9,8,3,10,12,8

Solution: First, order the sample as follows. 3,5,7,8,8,9,10,12,35.

Since  $n$  is odd, the sample median is given by the 5th,  $((9+1)/2)$ th,

largest point, which is equal to 8.

1 FOR NUMBER

1 \*  ★ most

2 \*

3 +

4 \*

5 \*

# Calculating the Mode

## 1 Frequency Count

Count how many times each value appears in the data set.


## 2 Identify Highest Frequency

Determine which value(s) occur most often.

## 3 Multiple Modes

If multiple values tie for highest frequency, the data set is multimodal.

1 FOR NUMBER

1 \*  \* MOS

2 \*

3 +

4 \*

5 \*

# Calculating the Mode

---

- Exa: Find the modal values for the following data
- a) 22, 66, 69, 70, 73. (no modal value)
- b) 1.8, 3.0, 3.3, 2.8, 2.9, 3.6, 3.0, 1.9, 3.2, 3.5 (modal value = 3.0 kg)



# Practical Examples

Data Set	Mean	Median	Mode
2, 3, 4, 4, 5, 5, 5, 6, 7	4.56	5	5
10, 20, 30, 40, 1000	220	30	N/A
Red, Blue, Green, Blue, Red	N/A	N/A	Red, Blue



# Key Takeaways

## Choosing the Right Measure

Select the appropriate measure based on your data type and distribution. Consider the presence of outliers.

## Complementary Use

Using multiple measures together often provides a more comprehensive understanding of the data's central tendency.

## Data Interpretation

Understanding these measures is crucial for accurate data analysis and informed decision-making in various fields.

# Conclusion

1

## Versatile Tools

Measures of central tendency are versatile tools for data analysis across various disciplines.

2

## Informed Decisions

They enable better understanding of data distributions, leading to more informed decision-making.

3

## Continuous Learning

As data becomes increasingly important, understanding these measures is crucial for data literacy.

