# ADITYA ENGINEERING COLLEGE (A)

## PROBABILITY AND STATISTICS
## 191BS4T18

# ADITYA ENGINEERING COLLEGE (A)

# Descriptive Statistics and Methods of Data Science

By
D.V. L. Prasanna
Sr. Asst. Professor
Aditya Engineering College(A)

# Unit –I:

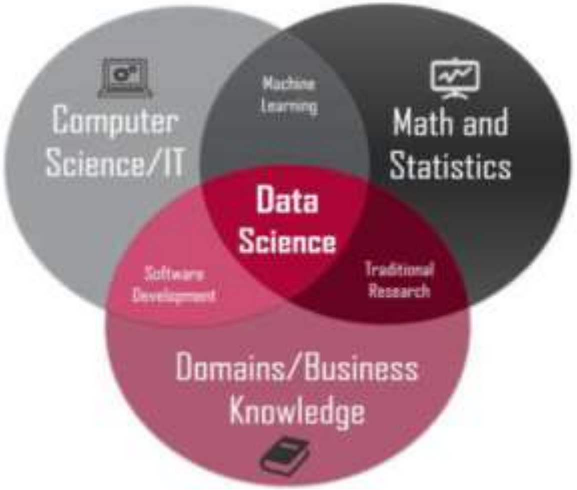
Data Science-Statistics Introduction-Population Vs Sample-Collection of data-Primary and secondary Data-Type of variable-:dependent and independent, Categorical and Continuous variables-Data visualization-Measures of Central tendency-Measures of variability(Spread or variance)

# Data Science

**Data science** is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

- a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.

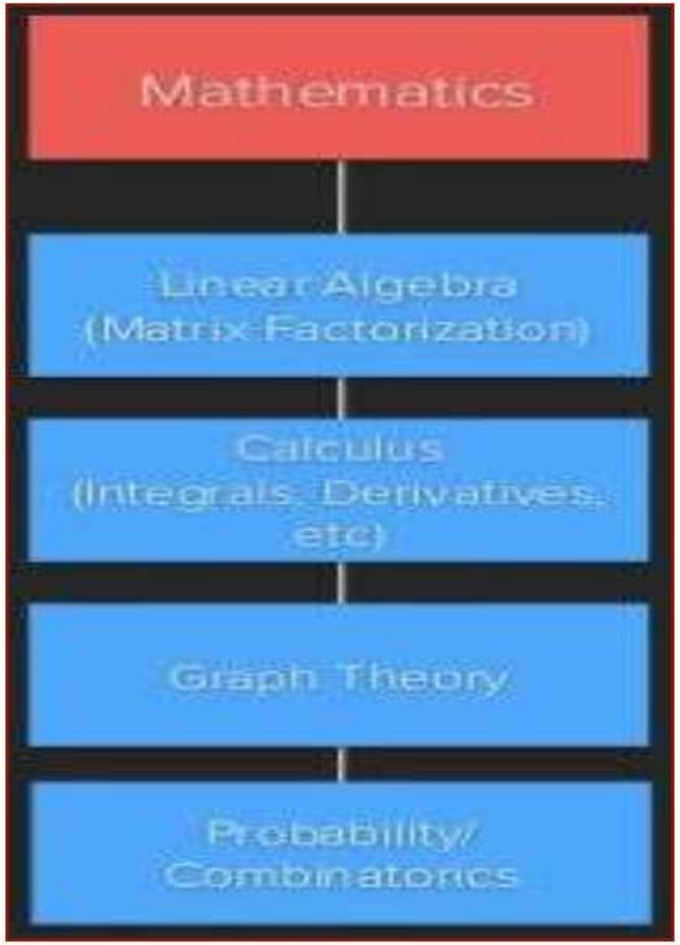
- employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science.



### Regular Data Science

- Data Analysis
- Modelling Statistics
- Engineering / Prototyping

# Data Scientist need to comfortable with:



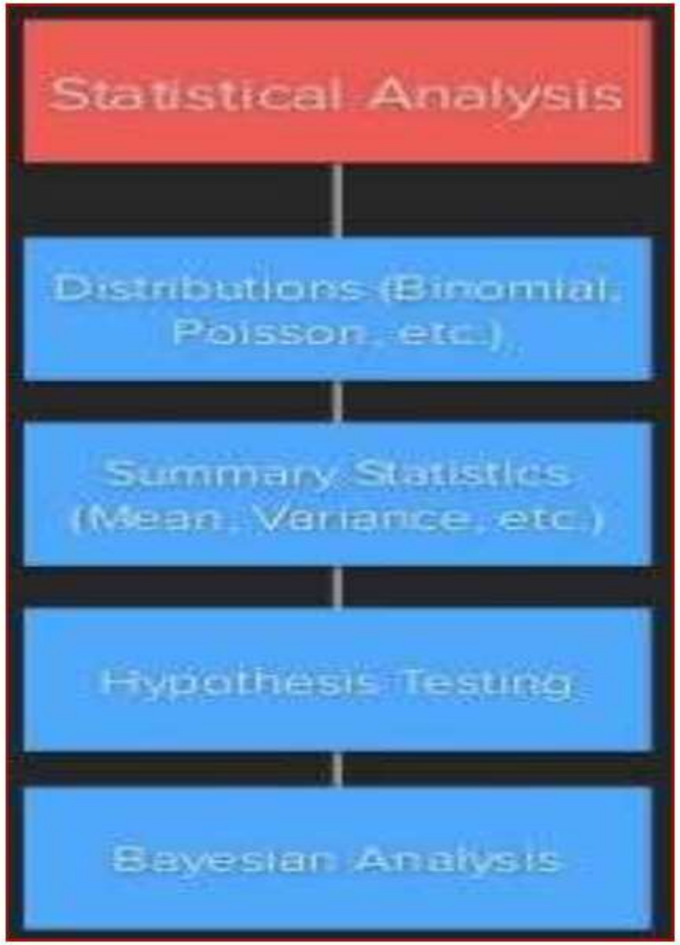
Mathematics

Linear Algebra (Matrix Factorization)

Calculus (Integrals, Derivatives, etc)

Graph Theory

Probability/ Combinatorics



Statistical Analysis

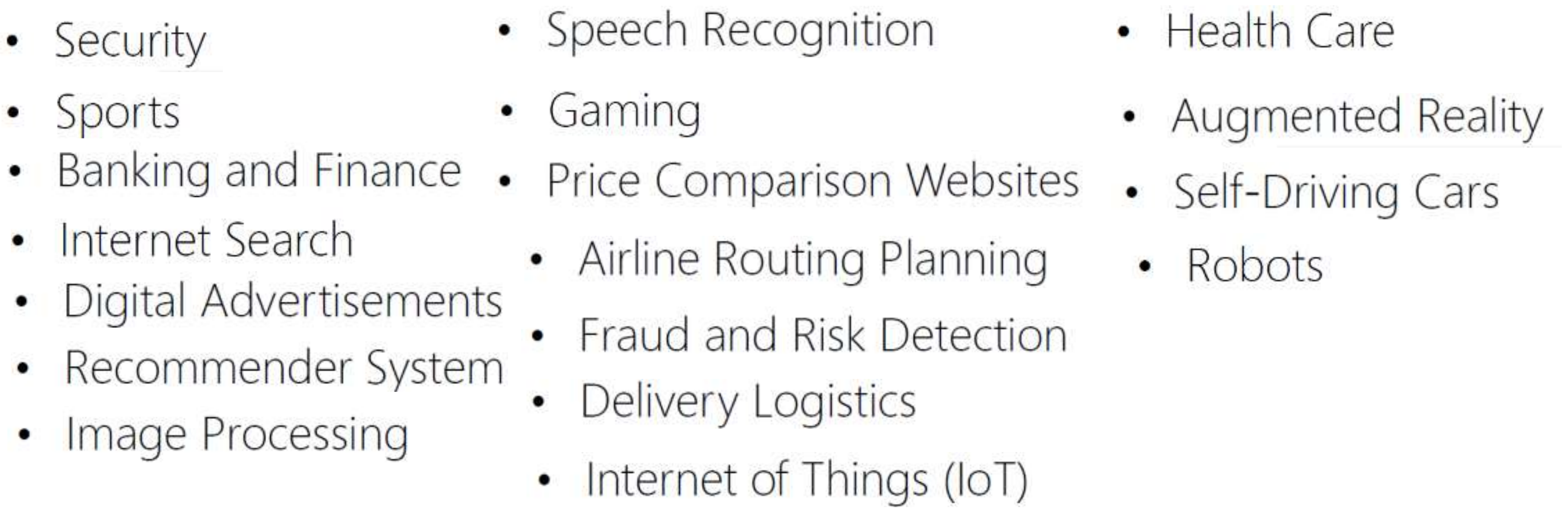
Distributions (Binomial, Poisson, etc.)

Summary Statistics (Mean, Variance, etc.)

Hypothesis Testing

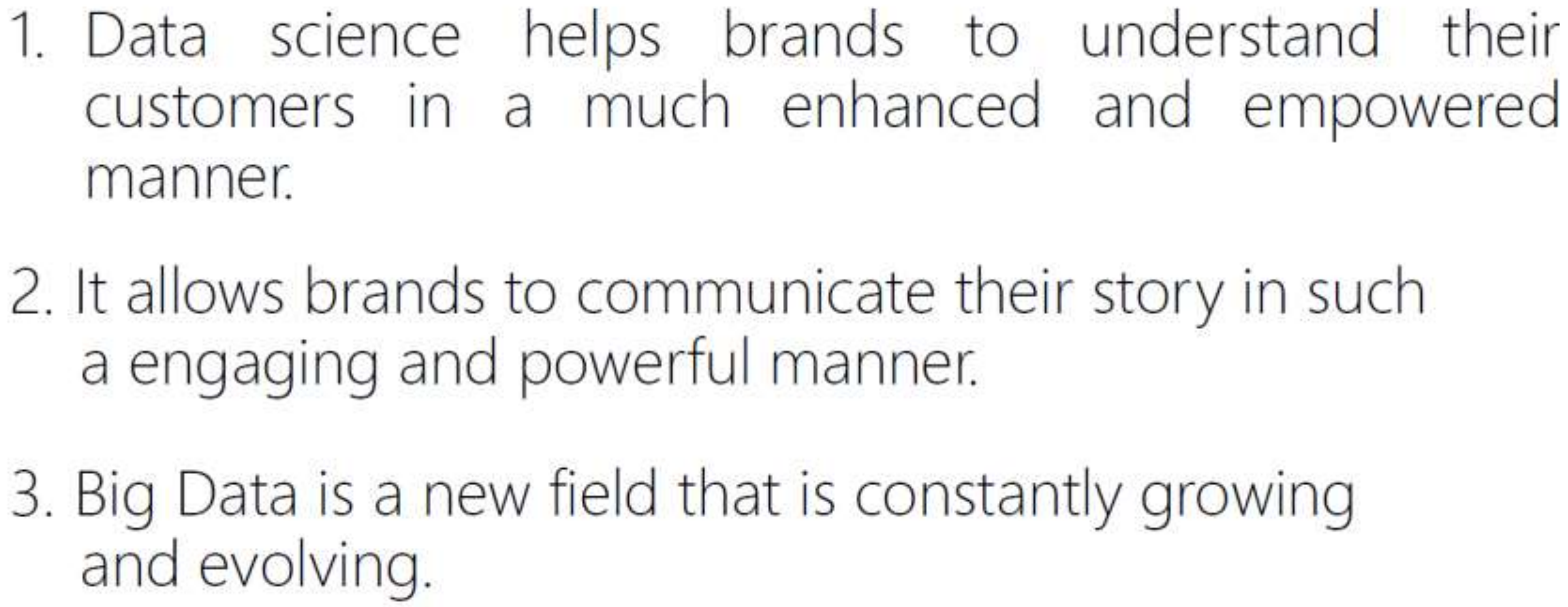
Bayesian Analysis

- **Applications of Data Science**



- Security
- Sports
- Banking and Finance
- Internet Search
- Digital Advertisements
- Recommender System
- Image Processing
- Speech Recognition
- Gaming
- Price Comparison Websites
- Airline Routing Planning
- Fraud and Risk Detection
- Delivery Logistics
- Internet of Things (IoT)
- Health Care
- Augmented Reality
- Self-Driving Cars
- Robots

# Importance Of Data Science



1. Data science helps brands to understand their customers in a much enhanced and empowered manner.

2. It allows brands to communicate their story in such a engaging and powerful manner.

3. Big Data is a new field that is constantly growing and evolving.

4. Its findings and results can be applied to almost any sector like travel, healthcare and education among others.

5. Data science is accessible to almost all sectors.

# STATISTICS

**Statistics** is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments

When census data cannot be collected, statisticians collect data by developing specific experiment designs and survey samples

- ➢ Descriptive statistics
- ➢ Inferential statistics.
- Organizing and summarizing data using informal methods like graphing and using numbers is called descriptive statistics.
- Organizing and summarizing data using formal methods is called Inferential statistics.

# ROLE OF STATISTICS

- Statistics is a language of data
- Statistics provides a scientific way to extract and retrieve the information hidden inside the data
- Statistics cannot do miracles
- Statistics cannot change the process or phenomenon
- Statistical tools provide forecasting buy not like astrologer’s parrot
- Statistics cannot do anything in no time .Tools can be developed but development needs time

# STATISTICAL TOOLS

Several tools and components are available

- Graphical tools

  provide visualization from first hand information

- Analytical Tools

  provide quantitative information

Both approaches work together and are inseparable

# STATISTICAL TOOLS

| Graphical Tools | Analytical Tools |
|---|---|
| 2D&3D plots | Central tendency of data: mean, median ,mode, geometric mean, harmonic mean, quartiles |
| Scatter Diagram | |
| Pie Diagram | Dispersion of data: Variance, standard deviation, standard error, mean deviation, absolute deviation, range etc., |
| Histogram | |
| Bar Chart | |
| Stem & Leaf plot | |
| Box Plot | |

# Population Vs Sample

**Population(or Universe):** A collection or aggregate or totality of persons, things or objects or statistical data under study is known as population.

**Size of the Population:** Number of units in the population is known as size of the population. It is denoted by N.

**Finite Population:** The number of units in the population is finite then it is known as finite population.

**Infinite Population: :** The number of units in the population is infinite then it is known as infinite population.

**Sample:** A portion(or subset) of the population is known as sample.

**Size of the sample:** Number of units in the sample is known as size of the sample. It is denoted by n.

**Large sample:** If the size of the sample is greater than 30 then it is known as large sample.

**Small sample:** If the size of the sample is less than 30 then it is known as small sample.

**Parameter:** The values that are obtained from the population data or a number that describes the property of the population is known as parameters

**Statistic:** The values that are obtained from the sample data or a number that describes the property of the sample is known as Statistics.

# Collection of Data

## Types of data:

Primary data

Secondary data

**Primary Data:**

Data originally collected by an investigator for the first time for any statistical investigation and it is the type of **data** that is collected by researchers directly from main **sources.**

**Sources of primary data:**

- Direct personal investigation
- Questionnaire received through postal mail, e-mail, e-forms(google forms), online surveys etc.,
- Questionnaire sent through surveyors

## Secondary data:

Data which has already been collected by some person or agency for any statistical investigation

## Sources of secondary data:

- Published sources ex: books, journals, articles, web pages etc.
- Data collected from survey agencies
- public reports the data Ex: municipalities
- blogs etc
- library, bots, internet sources etc.

**Examples:**

An example of primary data is the national census data collected by the government while an example of secondary data is the data collected from online sources.

The secondary data collected from an online source could be the primary data collected by another researcher.

For example, the government, after successfully the national census, they share the results in newspapers, online magazines, press releases, etc. Another government agency that is trying to allocate the state budget for healthcare, education, etc. may need to access the census results.

With access to this information, the number of children who needs education can be analyzed and hard to determine the amount that should be allocated to the education sector. Similarly, knowing the number of old people will help in allocating funds for them in the health sector.

## Advantages:

Some common advantages of primary data are its authenticity, specific nature, and up to date information while secondary data is very cheap and not time-consuming.

Primary data is very reliable because it is usually objective and collected directly from the original source. It also gives up to date information about a research topic compared to secondary data.

Secondary day, on the other hand, is not expensive making it easy for people to conduct secondary research. It doesn't take so much time and most of the secondary data sources can be accessed for free.

# Similarities Between Primary & Secondary Data

1. Contains Same Content:

Secondary data was once primary data when it was newly collected by the first researcher. The content of the data collected does not change and therefore has the same content with primary data.

It doesn't matter if it was further visualized in the secondary form, the content does not change. A common example of these are definitions, theorems, and postulates that were made years ago but still remain the same.

2.Primary data and secondary data both have applications in business and research. They may, however, differ from each other in the way in which they are collected, used, and analyzed.

## Differences between primary and secondary data:

1. Primary data is very expensive while secondary data is economical. When working on a low budget, it is better for researchers to work with secondary data, then analyze it to uncover new trends.
2. Primary data is more accurate and reliable while secondary data is relatively less reliable and accurate. This is mainly because the secondary data sources are not regulated and are subject to personal bias.
3. Primary data is available in crude form while secondary data is available in a refined form. That is, secondary data is usually made available to the public in a simple form for a layman to understand while primary data are usually raw and will have to be simplified by the researcher.

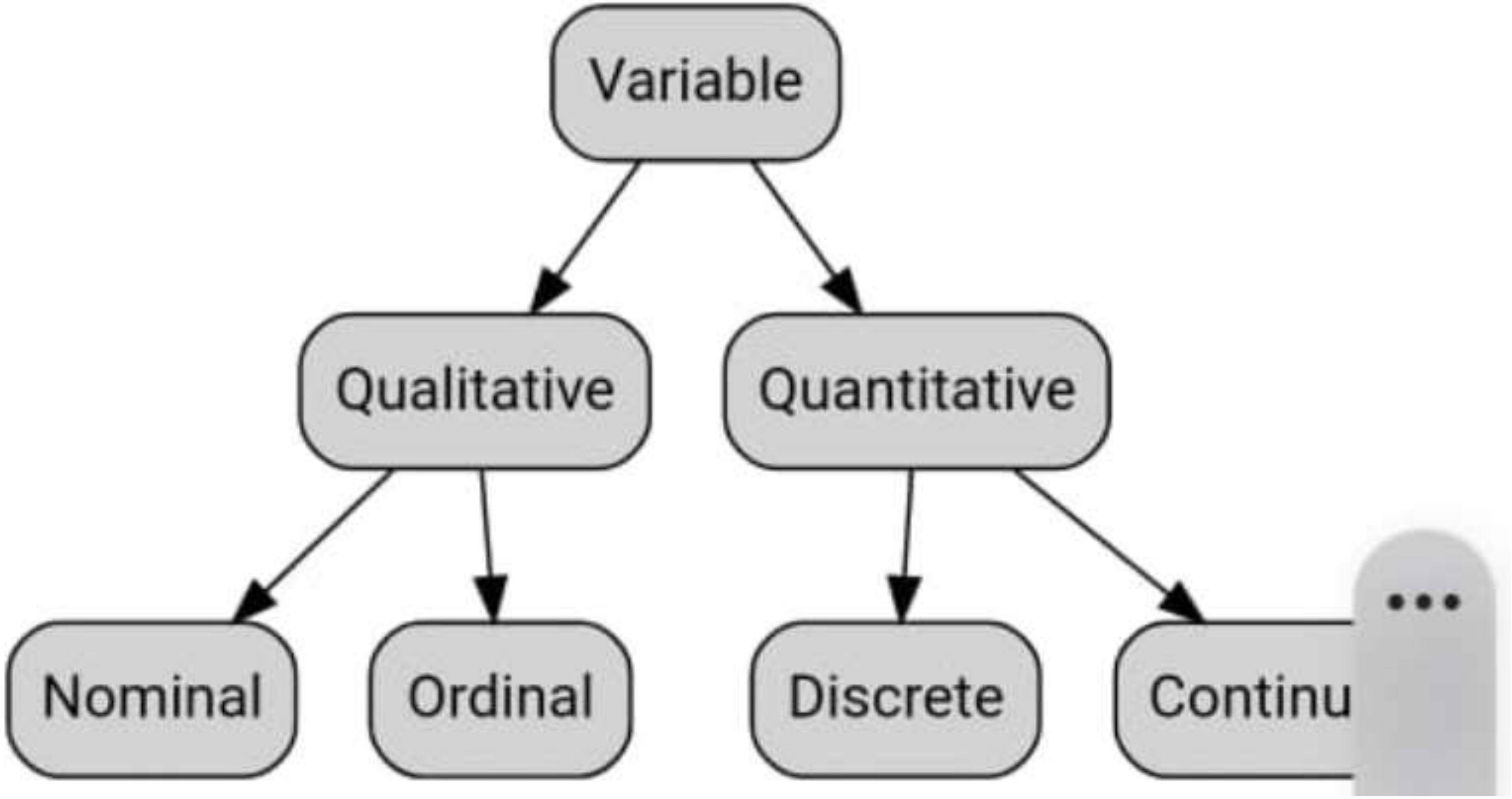
# Variables and types of variables

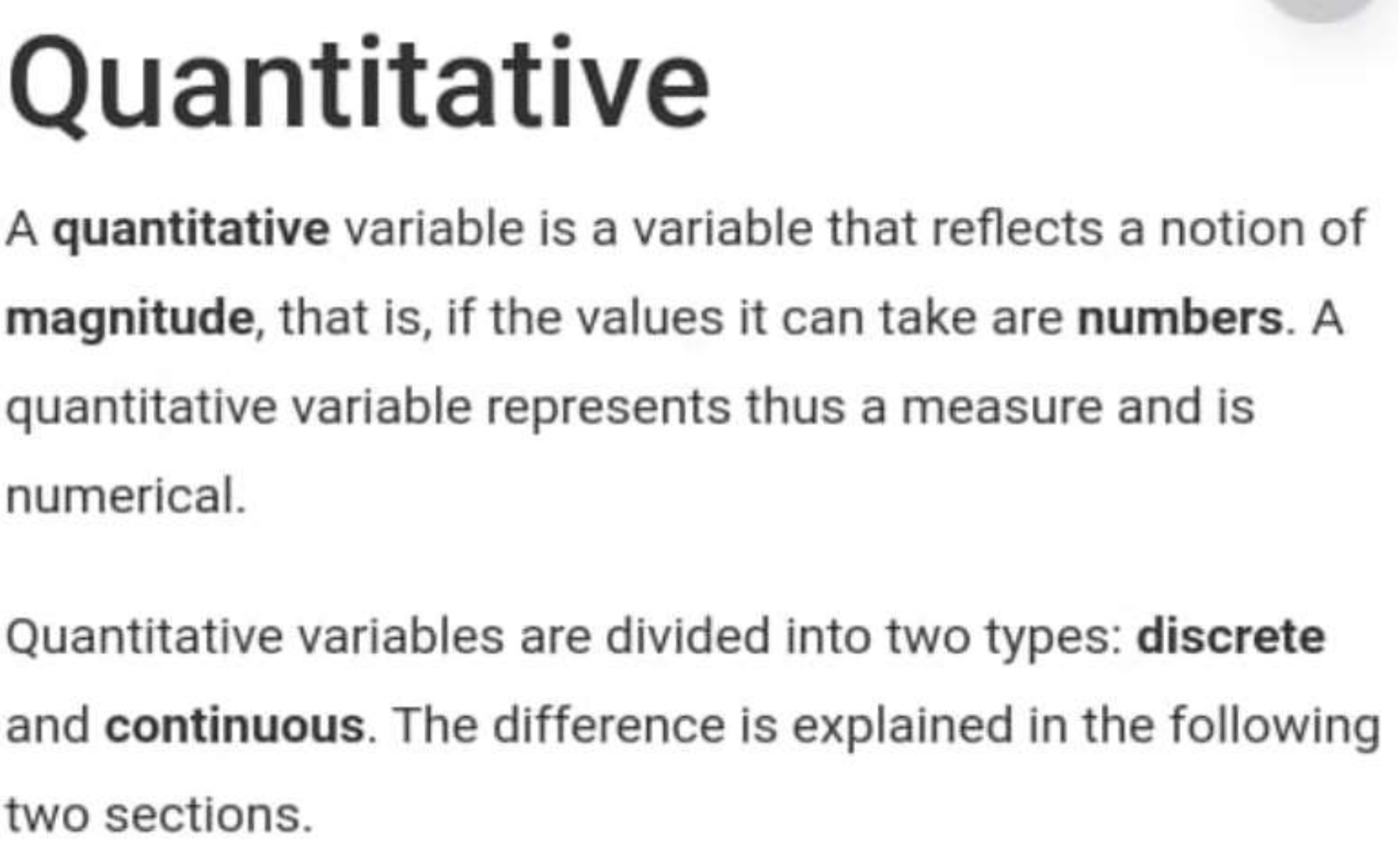
The values that are altering according to circumstances are referred to as variables. A variable can occurs in any form, such as trait, factor or a statement that will constantly be changing according to the changes in the applied environment.

Variables in statistics are broadly divided into four categories such as

1. Independent variables
2. Dependent variables
3. Categorical variables and
4. Continuous variables.

Apart from these, quantitative and qualitative variables hold data as nominal, ordinal, interval and ratio.

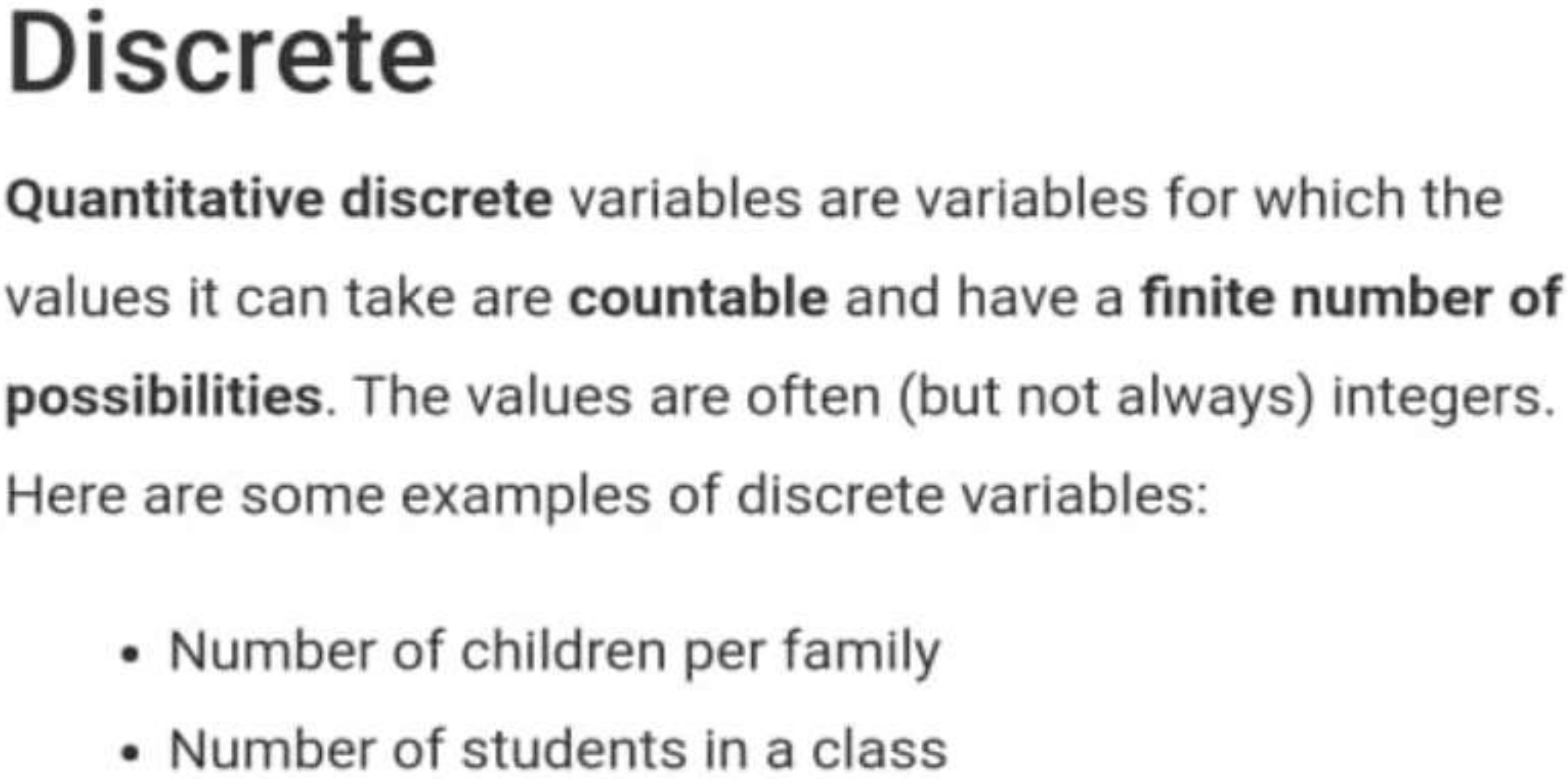

Variable
Qualitative
Quantitative
Nominal
Ordinal
Discrete
Continu



# Quantitative

A **quantitative** variable is a variable that reflects a notion of **magnitude**, that is, if the values it can take are **numbers**. A quantitative variable represents thus a measure and is numerical.

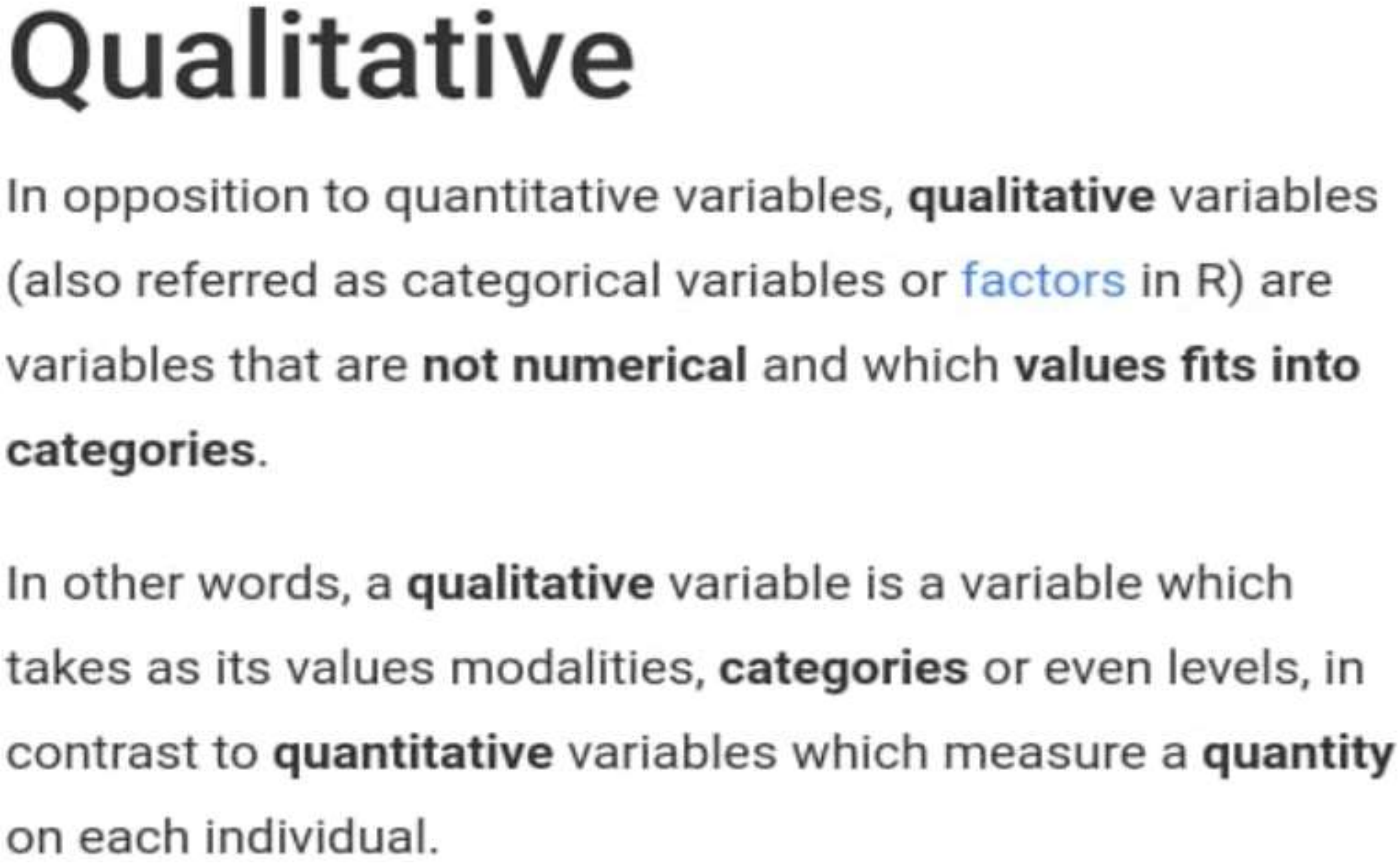
Quantitative variables are divided into two types: **discrete** and **continuous**. The difference is explained in the following two sections.



# Discrete

**Quantitative discrete** variables are variables for which the values it can take are **countable** and have a **finite number of possibilities**. The values are often (but not always) integers. Here are some examples of discrete variables:

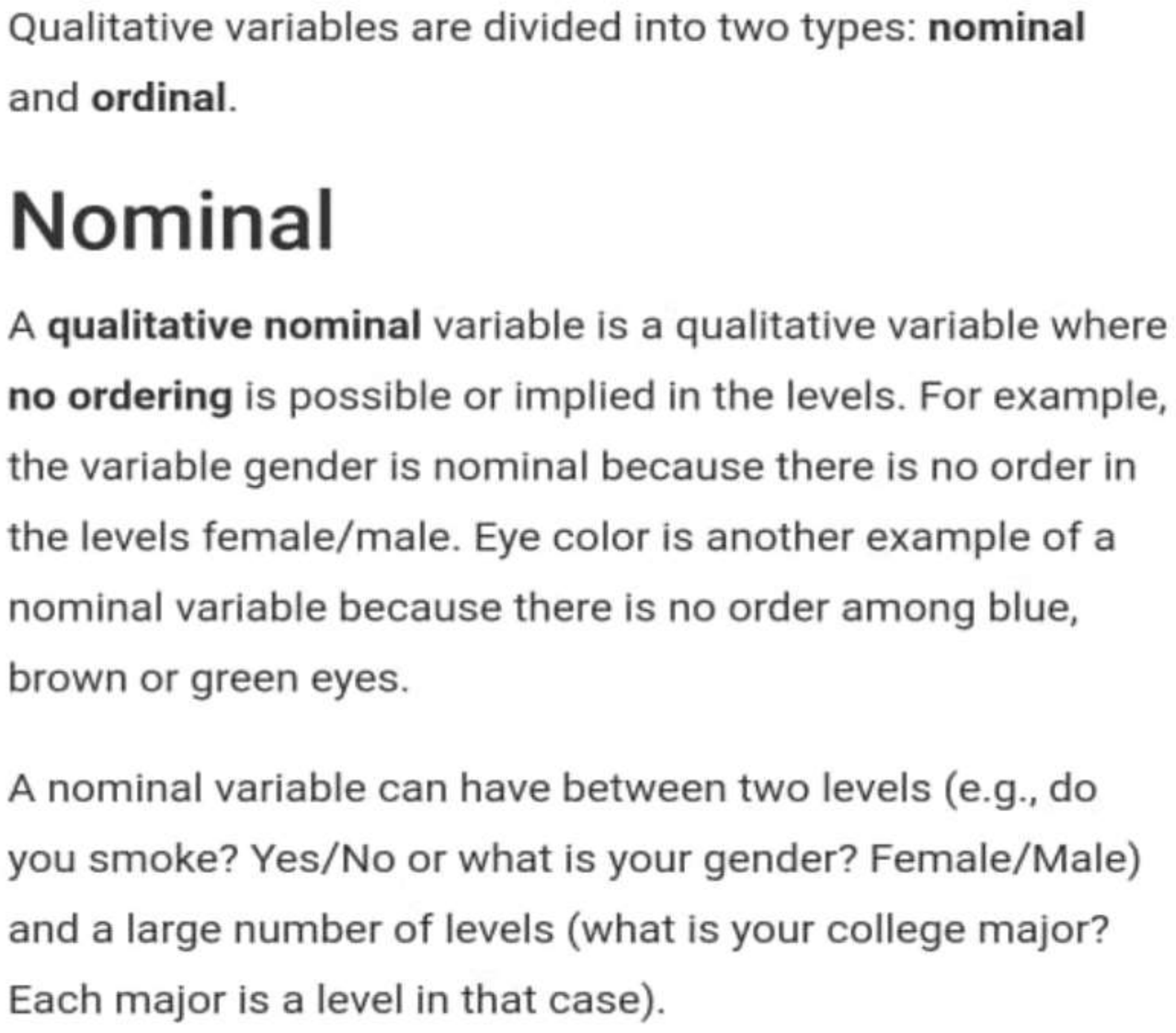
- Number of children per family
- Number of students in a class



# Qualitative

In opposition to quantitative variables, **qualitative** variables (also referred as categorical variables or factors in R) are variables that are **not numerical** and which **values fits into categories**.

In other words, a **qualitative** variable is a variable which takes as its values modalities, **categories** or even levels, in contrast to **quantitative** variables which measure a **quantity** on each individual.

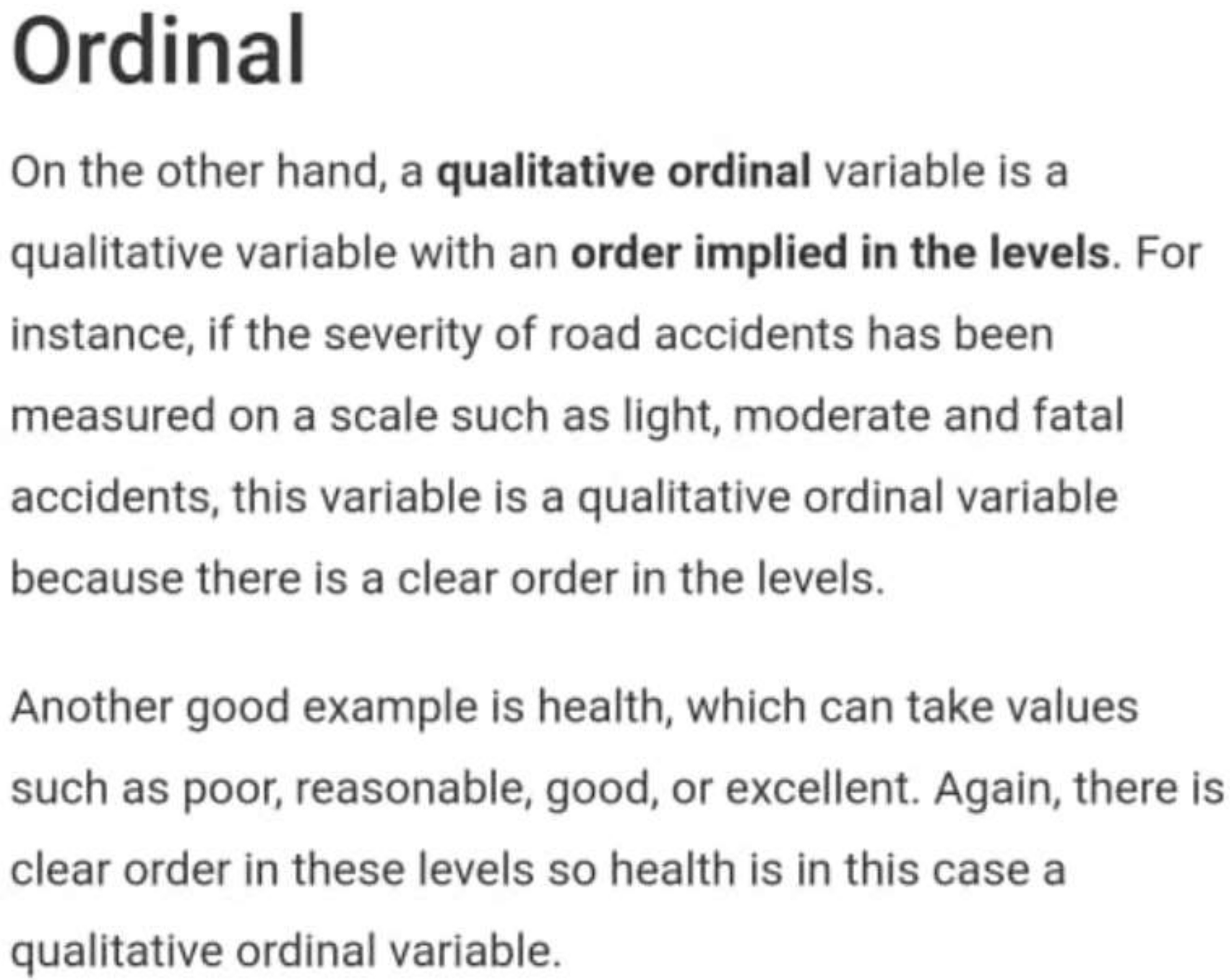


Qualitative variables are divided into two types: **nominal** and **ordinal**.

## Nominal

A **qualitative nominal** variable is a qualitative variable where **no ordering** is possible or implied in the levels. For example, the variable gender is nominal because there is no order in the levels female/male. Eye color is another example of a nominal variable because there is no order among blue, brown or green eyes.

A nominal variable can have between two levels (e.g., do you smoke? Yes/No or what is your gender? Female/Male) and a large number of levels (what is your college major? Each major is a level in that case).



# Ordinal

On the other hand, a **qualitative ordinal** variable is a qualitative variable with an **order implied in the levels**. For instance, if the severity of road accidents has been measured on a scale such as light, moderate and fatal accidents, this variable is a qualitative ordinal variable because there is a clear order in the levels.

Another good example is health, which can take values such as poor, reasonable, good, or excellent. Again, there is clear order in these levels so health is in this case a qualitative ordinal variable.

## *1. Independent Variables*

The independent variable is the one that is computed in research to view the impact of dependent variables. It is also called as resultant variables, predictor or experimental variables.
For example, A manager asks 100 employees to complete a project. He should know the capacity of the individual employee. He wants to know the reason behind smart guys and failure guys. The first reason is that some will be working hard for day and night to complete the project within the estimated time, and the other one is that some guys are born intelligent and smarter than others. The variable which is similar to an independent variable is called a covariate variable but is impacted by the dependent variable but not as common as a variable of interest.

## 2. *Dependent Variables*

The dependent variable is also called a criterion variable which is applied in non-experimental circumstances. The dependent variable has relied on the independent variable.
From the above-mentioned example, the project's productivity or completion is the main criteria that are dependent on estimated time and IQ. Here, the independent variables are IQ and estimated time, which may or may not reflect in an employee's productivity. So the extension of estimated time or enhancing the IQ of a person doesn't make any sense in employee's productivity as it is not predictable.
Hence, the managers' focus is to work on the independent variables such as allotted time and IQ that leads to certain changes in employee's productivity that are the dependent variables. So both the variables are connected in some measures. The variables which get affected by other variables in econometrics is termed as endogenous variables. A hidden variable impacts the relationship between the dependent and independent variable called lurking variables. When an independent variable is not impacted by any other variables and is restricted to a certain extent are called an explanatory variable.

## *3. Categorical Variables*

It is a wide category of variable which is infinite and has no numerical data. These variables are called as qualitative variables or attribute variable in terms of statistics software. Such variables are further divided into nominal variables, ordinal and dichotomous variables.

**Nominal variables** don't have any intrinsic order. For instance, a developer classifies his environment into different types of networks based on their structure, such as P2P, cloud computing, pervasive computing, IoT. So here, the type of network is a nominal variable comprised of four categories. The varied categories present in the nominal variable can be known as the nominal variable levels or groups.

**Dichotomous variables** are also called binary values, which have only two categories.

For example, if we question a person that he owns a car, he would reply only with yes or no. such types of two distinct variables that are nominal are called as dichotomous. It just accounts for only two values, such as 0 or 1. It could be yes or no, short or long, etc.

**Ordinal variables** are nominal variables that include two or multiple categories. If you see any hotel feedback form, it has five ratings such as excellent, good, better, poor and very poor. So we can rank the level with the help of ordinal variables that hold meaning to the research. It is unambiguous, and values can be considered for decision making.

## 4. *Continuous Variables*

The variables which measure some count or quantity and don't have any boundaries are termed as continuous variables. It can be segregated into ratio or interval, or discrete variables.

**Interval variables** have their centralized attribute, which is calibrated along with a range with some numerical values. The example can be temperature calibrated in Celsius or Fahrenheit doesn't give any two different meaning; they display the optimum temperature, and it's strictly not a ratio variable.

It can account for only a certain set of values, such as several bikes in a parking area are discrete as the floor holds only a limited portion to park bikes. Ratio variables occur with intervals; it has an extra condition that zero on any measurement denotes that there is no value of that variable. In simple, the distance of four meters is twice the distance of two meters. It operates on the ratio of measurements.

A factor that remains constant in an experiment is termed as a control variable. In an experiment, if the scientist wants to test the plant's light for its growth, he should control the value of water and soil quality. The additional variable which has a hidden impact on the obtained experimental values are called confounding variables.

# Data Visualization

**Data visualization** is the graphical representation of information and **data**.

The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.

By using visual elements like charts, graphs, and maps, **data visualization** tools provide an accessible way to see and understand trends, outliers, and patterns in **data**. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

# Benefits of data visualization :

- the ability to absorb information quickly, improve insights and make faster decisions
- an increased understanding of the next steps that must be taken to improve the organization
- an improved ability to maintain the audience's interest with information they can understand
- an easy distribution of information that increases the opportunity to share insights with everyone involved
- eliminate the need for data scientists since data is more accessible and understandable and
- an increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

## Common general types of data visualization:

- Charts
- Tables
- Graphs
- Maps
- Info graphics
- Dashboards

## More specific examples of methods to visualize data:

- Bar Chart
- Pie chart
- Histogram
- Stem leaf plots
- Scatter diagrams etc.

**Absolute frequency:**

Number of times that a value appears is known as absolute frequency. It is represented as $f_i$ where the subscript represents each of the values. The sum of the absolute frequencies is equal to the total number of data, represented bas N.

$$N = f_1 + f_2 + - - - + f_n = \sum_{i=1}^{n} f_i$$

**Relative frequency:** The result of dividing the absolute frequency of a certain value by the total number of data. It is represented as $n_i$ . The sum of the relative frequencies is equal to 1. We can prove this easily by factorizing N.

$$n_i = \frac{f_i}{N}$$

**Example**

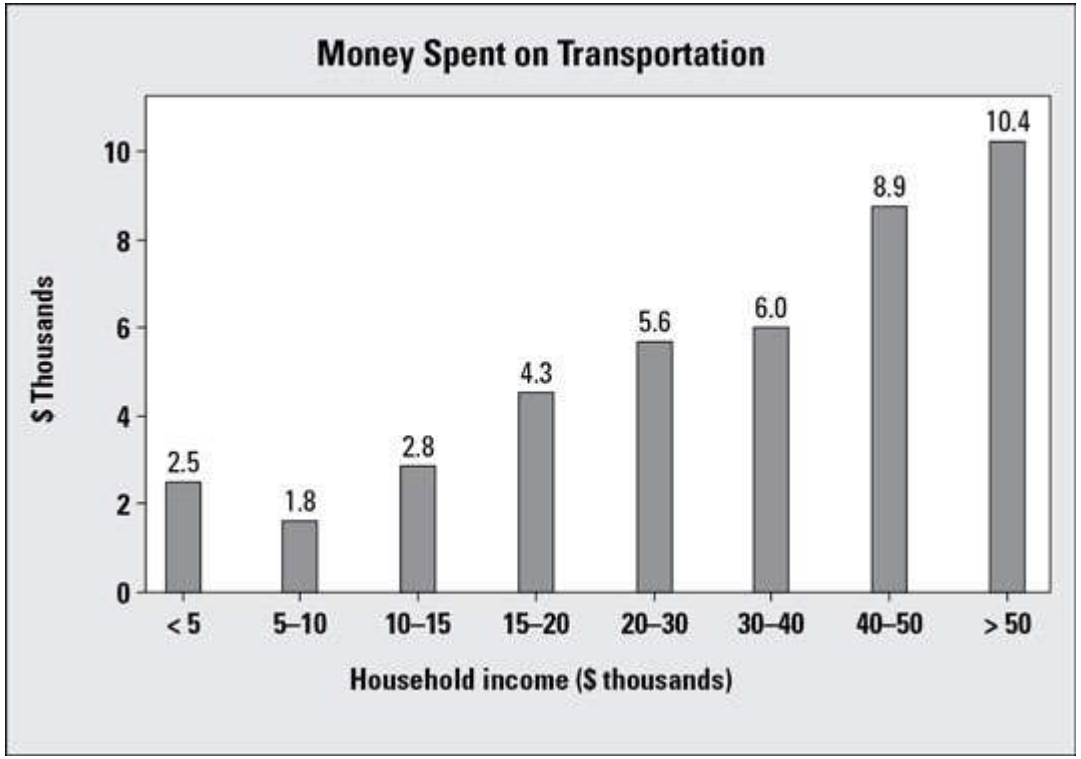
15 students answer the question of how many brothers or sisters they have. The answers are:
1,1,2,0,3,2,1,4,2,3,1,0,0,1,2
Then, we can construct a table of frequencies

Solution:
Here
N=3+5+4+2+1=15

| Brothers | Absolute frequency fi | Relative frequency ni | Cumulative frequency Fi | Relative cumulative frequency Ni |
|---|---|---|---|---|
| 0 | 3 | 3/15 | 3 | 3/15 |
| 1 | 5 | 5/15 | 3+5=8 | 3/15+5/15 =8/15 |
| 2 | 4 | 4/15 | 3+5+4=12 | 12/15 |
| 3 | 2 | 2/15 | 3+5+4+2=14 | 14/15 |
| 4 | 1 | 1/15 | 3+5+4+2+1=15 | 1 |

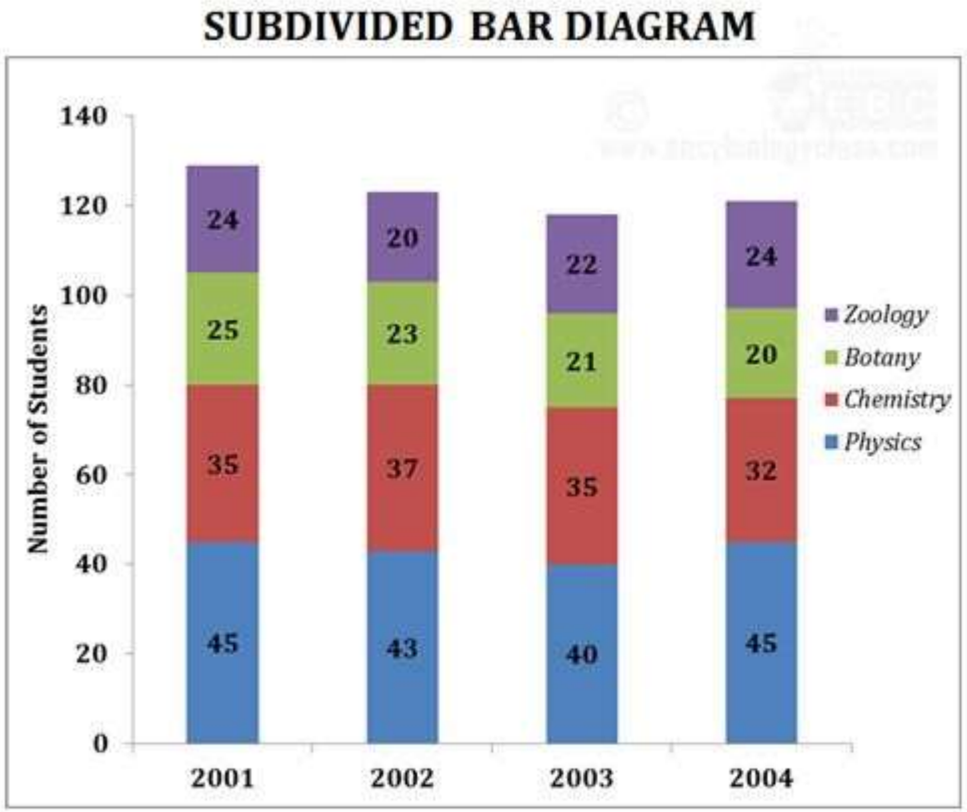
# Bar diagrams:

- Visualizes the relative or absolute frequencies of observed values of a variable
- Consists of one bar for each category either horizontal or vertical
- Height of each bar is decided by the frequency of the respective category shown in y-axis



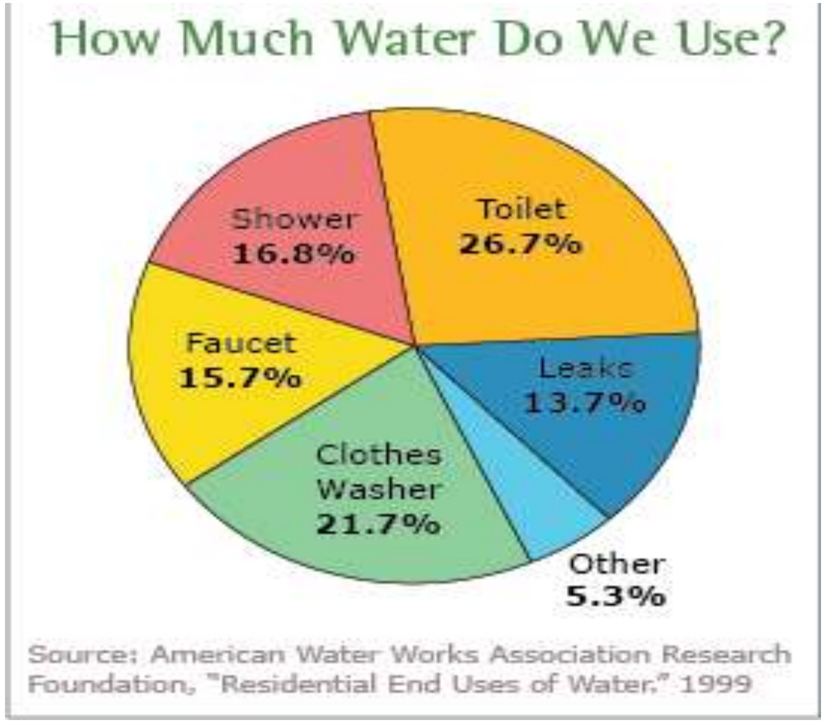
## Subdivided Bar diagrams:

- Divides the total magnitude of variables into various parts



## Pie diagrams:

- Visualizes the absolute and relative frequency
- A circle partitioned into segments where each of the segments represents a category
- Size of each segment depends upon relative frequency and is determined by the angle(relative frequency x 360°)

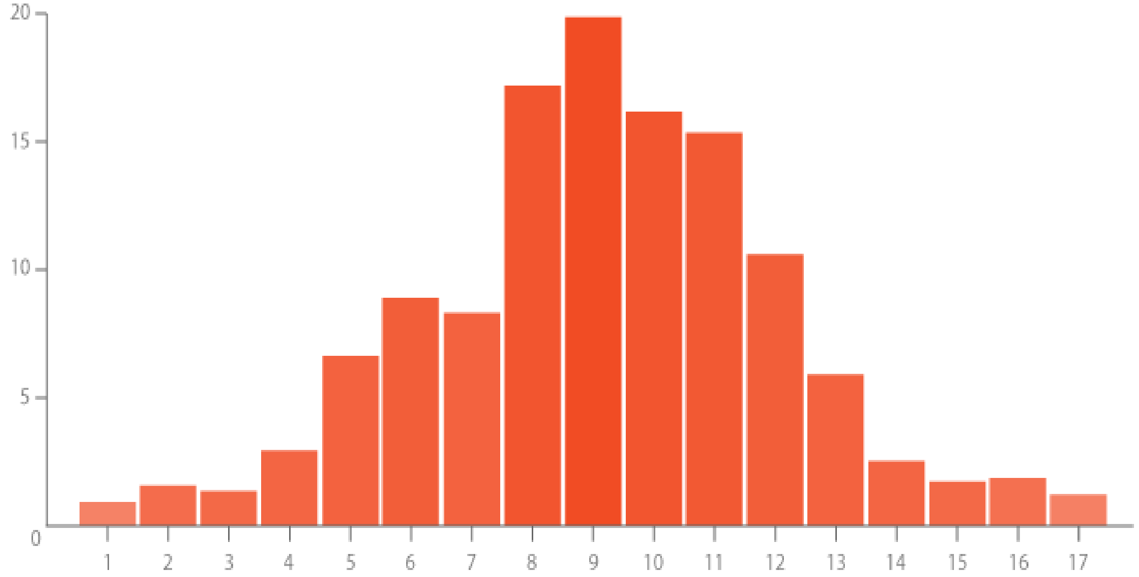


Source: American Water Works Association Research Foundation, "Residential End Uses of Water." 1999

# Histogram:

A Histogram visualizes the distribution of data over a continuous interval or certain time period. Each bar in a histogram represents the tabulated frequency at each interval/bin.

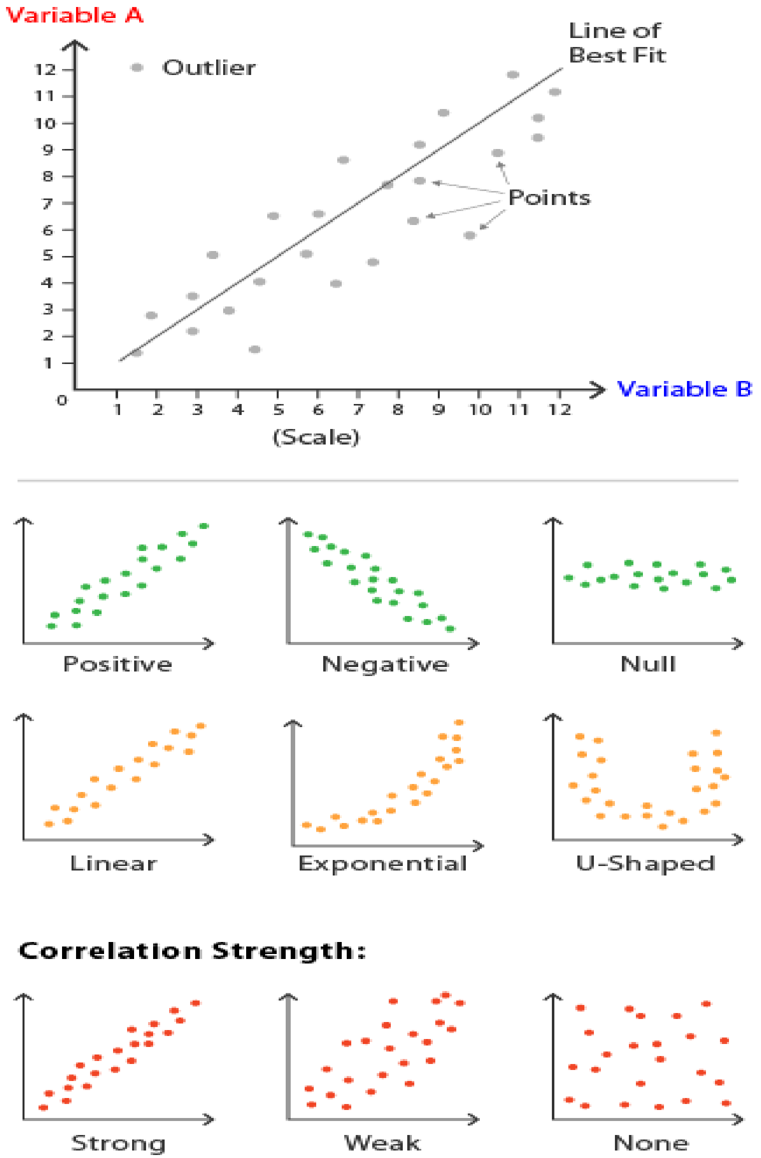
Histograms help give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps or unusual values. They are also useful for giving a rough view of the probability distribution.



# Scatter plot:

Also known as a *Scatter Graph, Point Graph, X-Y Plot, Scatter Chart or Scattergram*.

Scatterplots use a collection of points placed using Cartesian Coordinates to display values from two variables. By displaying a variable in each axis, you can detect if a relationship or correlation between the two variables exists.



Various types of correlation can be interpreted through the patterns displayed on Scatterplots. These are: **positive** (values increase together), **negative** (one value decreases as the other increases), **null** (no correlation), **linear**, **exponential** and **U-shaped**. The strength of the correlation can be determined by how closely packed the points are to each other on the graph. Points that end up far outside the general cluster of points are known as **outliers**.

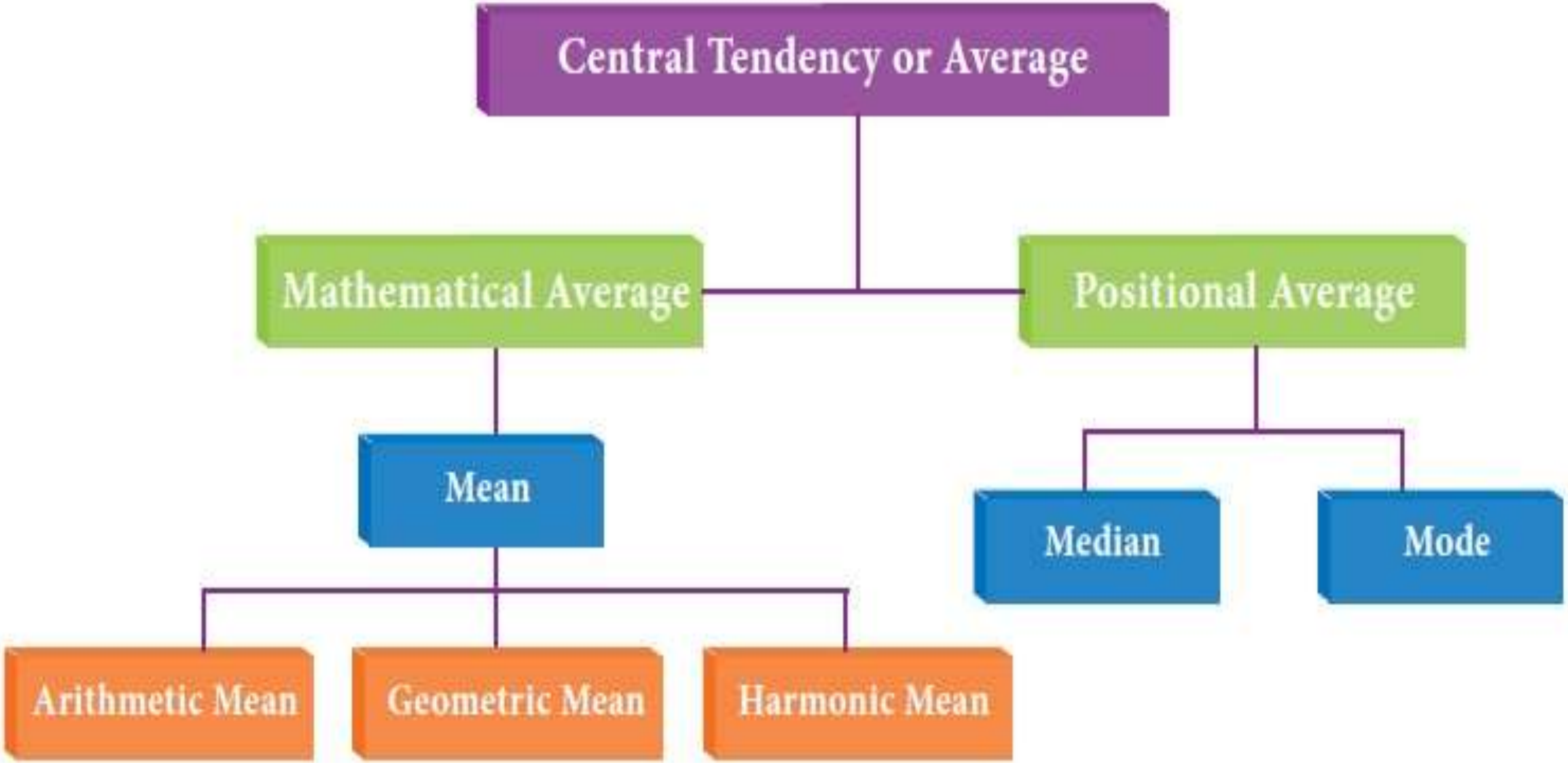
# Measures of Central tendency:

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data or the statistical measure that represents a single value of the entire dataset.

It is also called as measures of central location or summary statistics.

The five measures of central tendency are

1. Arithmetic mean or simple mean
2. Median
3. Mode
4. Geometric mean and
5. Harmonic mean



Central Tendency or Average
Mathematical Average
Positional Average
Mean
Median
Mode
Arithmetic Mean
Geometric Mean
Harmonic Mean

## 1. Arithmetic mean or simple mean:

It is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data.

The arithmetic mean of a set of observations is the sum of observations divided by the number of observations in the data. It is denoted by $\overline{x}$

For n values in a set of data namely as $x_1, x_{2,} x_{3, \ldots} x_n$, the mean of data is given as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots \ldots \ldots \ldots x_n}{n}$$

or

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For calculating the mean when the frequency of the observations is given, such that $x_1$, $x_{2,}$ $x_{3,\dots,}$ $x_n$ is the recorded observations, and $f_1$, $f_{2,}$ $f_3$ … $f_n$ is the respective frequencies of the observations then;

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 \dots \dots \dots f_n x_n}{f_1 + f_2 + f_3 \dots \dots \dots f_n}$$

or

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

**Example 1 :**

The marks obtained by 10 students in a test are 15, 75, 33, 67, 76, 54, 39, 12, 78, 11. Find the arithmetic mean.

**Solutior**

Mean $\bar{x} = \dfrac{x_1 + x_2 + x_3 \dots \dots \dots . . x_n}{n}$

=Total marks of 10 students / 10

= (15 + 75 + 33 + 67 + 76 + 54 + 39 + 12 + 78 + 11) / 10

= 460 / 10

Mean= 46

## Example 2 :

John studies for 4 hours, 5 hours and 3 hours respectively on three consecutive days. How many hours does he study daily on an average?

**Solution :**

The average study time of John

= Total number of study hours / Number of days for which he studied

= (4 + 5 + 3) / 3

= 12 / 3

= 4 hours

## Example 3

A batsman scored the following number of runs in six innings:

36, 35, 50, 46, 60, 55

Calculate the mean runs scored by him in an inning.

**Solution :**

To find the mean, we find the sum of all the observations and divide it by the number of observations.

Mean = Total runs / Number of innings

= (36 + 35 + 50 + 46 + 60 + 55) / 6

= 47

Thus, the mean runs scored in an inning are 47.

Arithmetic mean can be computed for both ungrouped data (raw data: data without any statistical treatment) and grouped data (data arranged in tabular form containing different groups).

| **Method's Name** | **Nature of Data** | |
|---|---|---|
| | **Ungrouped Data** | **Grouped Data** |
| **Direct Method** | $\bar{x} = \frac{\sum x}{n}$ | $\bar{x} = \frac{\sum fx}{N}$<br>$N = \sum f_i$ |
| **Indirect or Short-Cut Method** | $\bar{x} = A + \frac{\sum D}{n}$<br>$D = x - A$ | $\bar{x} = A + \frac{\sum fD}{N}$<br>$D = x - A$ |
| **Method of Step-Deviation** | $\bar{x} = A + \frac{\sum u}{n}.c$<br>$u = (x - A)/c$ or $h$ | $\bar{x} = A + \frac{\sum fu}{N}.c$ |

Provide the given distribution of the following frequency distribution of first year students of a particular college:

| **Age (Years)** | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|
| **Number of Students** | 2 | 5 | 13 | 7 | 3 |

**Solution:**

The given distribution is grouped data and the variable involved is ages of first year students, while the number of students represents frequencies.

| **Ages (Years)**<br>x | **Number of Students**<br>f | fx |
|---|---|---|
| 13 | 2 | 26 |
| 14 | 5 | 70 |
| 15 | 13 | 195 |
| 16 | 7 | 112 |
| 17 | 3 | 51 |
| **Total** | ∑f=30 | ∑fx=454 |

$$\bar{x} = \frac{\sum fx}{N}$$

$$N = \sum f_i$$

=454/30=15.13 years

**Example:**

The following data shows the distance covered by 100 people to perform their routine jobs.

| **Distance (Km)** | 0–10 | 10–20 | 20–30 | 30–40 |
|---|---|---|---|---|
| **Number of People** | 10 | 20 | 40 | 30 |

**Solution:**

The given distribution is grouped data and the variable involved is distance covered, while the number of people represents frequencies.

| **Distance (Km)** | **Number of People**<br>f | **Mid Points**<br>x | fx |
|---|---|---|---|
| 0–10 | 10 | 5 | 50 |
| 10–20 | 20 | 15 | 300 |
| 20–30 | 40 | 25 | 1000 |
| 30–40 | 30 | 35 | 1050 |
| **Total** | ∑f=100 | | ∑fx=2400 |

$$\bar{x} = \frac{\sum fx}{N}$$

$$N = \sum f_i$$

=2400/100=24 Km.

**3.** If the mean of the following distribution is 9, find the value of p.

| **x** | 4 | 6 | p + 7 | 10 | 15 |
|---|---|---|---|---|---|
| **f** | 5 | 10 | 10 | 7 | 8 |

**Solution:**

| $x_i$ | $f_i$ | $x_i f_i$ |
|---|---|---|
| 4 | 5 | 20 |
| 6 | 10 | 60 |
| p + 7 | 10 | 10(p + 7) |
| 10 | 7 | 70 |
| 15 | 8 | 120 |

$\sum f_i = 5 + 10 + 10 + 7 + 8 = 40$, Given that $\bar{x} = 9$

$\sum f_i x_i = 270 + 10(p + 7)$

Mean = $\bar{x} = \dfrac{\sum fx}{N}$

$$9 = \frac{270 + 10(p + 7)}{40}$$

⇒ 270 + 10p + 70 = 9 × 40

⇒ 340 +10p = 360

⇒ 10p = 360 - 340

⇒ 10p = 20

⇒ p = 20/10

⇒ p = 2

We have

**4.** The following table shows the number of plants in 20 houses in a group

| **Number of Plants** | 0 - 2 | 2 - 4 | 4 - 6 | 6 - 8 | 8 - 10 | 10 - 12 | 12 - 14 |
|---|---|---|---|---|---|---|---|
| **Number of Houses** | 1 | 2 | 2 | 4 | 6 | 2 | 3 |

Find the mean number of plants per house

**Solution:**

| **Number of Plant** | **Number of Houses ($f_i$)** | **Class Mark ($x_i$)** | **$f_ix_i$** |
|---|---|---|---|
| 0 - 2 | 1 | 1 | 1 × 1 = 1 |
| 2 - 4 | 2 | 3 | 2 × 3 = 6 |
| 4 - 6 | 2 | 5 | 2 × 5 = 10 |
| 6 - 8 | 4 | 7 | 4 × 7 = 28 |
| 8 - 10 | 6 | 9 | 6 × 9 = 54 |
| 10 - 12 | 2 | 11 | 2 × 11 = 22 |
| 12 -14 | 3 | 13 | 3 × 13 = 39 |

$\sum f_i$ = 1 + 2 + 2 + 4 + 6 + 2 + 3 = 20

$\sum f_i\, x_i$ =1 + 6 + 10 + 28 + 54 + 22 + 39 = 160

Therefore, mean = $\bar{x} = \dfrac{\sum fx}{N}$

= 160/20 = 8 plants

**Example**:

The following data shows distance covered by 100 people to perform their routine jobs. Calculate the arithmetic mean by step-deviation method.

| **Distance (Km)** | 0–10 | 10–20 | 20–30 | 30–40 |
|---|---|---|---|---|
| **Number of People** | 10 | 20 | 40 | 30 |

**Solution:**

The given distribution is grouped data and the variable involved is distance covered, while the number of people represents frequencies.

| **Distance Covered in (Km)** | **Number of People** f | **Mid Points** x | $u=(x-15)/10$ | fu |
|---|---|---|---|---|
| 0–10 | 10 | 5 | –1 | –10 |
| 10–20 | 20 | 15 | 0 | 0 |
| 20–30 | 40 | 25 | +1 | 40 |
| 30–40 | 30 | 35 | +2 | 60 |
| **Total** | $\sum f=100$ | | | $\sum fu=90$ |

A=15, ∑fu=90, ∑f=100 and h=10

$$\overline{x} = A + \frac{\sum fu}{N}.h$$

$$\overline{x} = 15 + \frac{90}{100}.10 = 24km$$

**Example**:

The following frequency distribution showing the marks obtained by 50 students in statistics at a certain college. Find the arithmetic mean using

(1) direct method

(2) short-cut method

(3) step-deviation.

| **Marks** | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 | 70–79 | 80–89 |
|---|---|---|---|---|---|---|---|
| **Frequency** | 1 | 5 | 12 | 15 | 9 | 6 | 2 |

**Solution**:

| | | | Direct Method | Short-Cut Method | | Step-Deviation Method | |
|---|---|---|---|---|---|---|---|
| **Marks** | f | x | fx | D=x–A | fD | u=(x–A)/h | fu |
| 20–29 | 1 | 24.5 | 24.5 | –30 | –30 | –3 | –3 |
| 30–39 | 5 | 34.5 | 172.5 | –20 | –100 | –2 | –10 |
| 40–49 | 12 | 44.5 | 534.5 | –10 | –120 | –1 | –12 |
| 50–59 | 15 | 54.5 | 817.5 | 0 | 0 | 0 | 0 |
| 60–69 | 9 | 64.5 | 580.5 | 10 | 90 | 1 | 9 |
| 70–79 | 6 | 74.5 | 447.5 | 20 | 120 | 2 | 12 |
| 80–89 | 2 | 84.5 | 169.5 | 30 | 60 | 3 | 6 |
| **Total** | 50 | | 2745 | | 20 | | 2 |

**1) Direct Method**:

$$\bar{x} = \frac{\sum fx}{N}$$

$$N = \sum f_i$$

=2745/50=54.9 or 55Marks

**(2) Short-Cut Method:**

$$\bar{x} = A + \frac{\sum fD}{N}$$

$$D = x - A$$

Where A=54.5

=54.5+20/50=54.9Marks

**(3) Step-Deviation Method:**

$$\bar{x} = A + \frac{\sum fu}{N}.h$$

Where A=54.5 h=10

=54.5+(2/50)×10

=54.5+0.4=54.9 Marks

# Merits and demerits of Arithmetic mean

## Merits

1. It is rigidly defined.

2. It is easy to understand and easy to calculate.

3. If the number of items is sufficiently large, it is more accurate and more reliable.

4. It is a calculated value and is not based on its position in the series.

5. It is possible to calculate even if some of the details of the data are lacking.

6. Of all averages, it is affected least by fluctuations of sampling.

7. It provides a good basis for comparison.

## Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

## 2. Median:

The median is a measure of central tendency, which denotes the value of the middle-most observation in the data.

**For ungrouped data:**

Median = $\left(\frac{n+1}{2}\right)^{th}$ observation, if n is odd.

Median = mean of $\left(\frac{n}{2}\right)^{th}$ observation and $\left(\frac{n}{2}+1\right)^{th}$ observation, if n is even.

**Procedure:**

**Step 1:** Arrange the given values in the ascending order.

**Step 2:** Find the number of observations in the given set of data. It is denoted by n.

**Step 3:** If n is odd, the median equals the $\left(\frac{n+1}{2}\right)^{th}$ observation.

**Step 4:** If n is even, then the median is given by the mean of $\left(\frac{n}{2}\right)^{th}$ observation and $\left(\frac{n}{2}+1\right)^{th}$ observation.

## Problems :

1. Alex timed 21 people in the sprint race, to the nearest second:
   59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 58, 62, 62, 68, 65, 56, 59, 68, 61, 67

**Solution:**

Arranging the given values in the ascending order

53, 55, 56, 56, 58, 58, 59, 59, 60, 61, **61**, 62, 62, 62, 64, 65, 65, 67, 68, 68, 70

Median = $\left(\frac{n+1}{2}\right)^{th}$ observation, if n is odd.

$$=\left(\frac{21+1}{2}\right)^{th}=11^{th}\text{ observation}$$

=61

Median=61

2. Find the median of the values **5, 7, 10, 20, 16, 12**

**Solution**:

Arrange the data in ascending order

**5, 7, 10, 12, 16, 20**

Since n=6 is even,

Median = mean of $\left(\frac{n}{2}\right)^{th}$ observation and $\left(\frac{n}{2}+1\right)^{th}$ observation, if n is even

= mean of 3rd and 4th observation

Median = $\frac{10+12}{2}=11$

Median = 11

3. Find the median of the values **4, 1, 8, 13, 11** **1,4,8,11,13**

4. Find the median of the following set of data

**13, 8, 19, 30, 15, 21, 9, 5** **5,8,9,13,15,19,21,30**

# For grouped data:

Median= $l+\frac{h}{f}\left(\frac{N}{2}-c\right)$

where

*l*= Lower class boundary of the median class

*f*= Frequency of the median class

N=∑f= Total frequency

c= Cumulative frequency of the class preceding the median class

h= Class interval size of the model class

**Step 1:** Make a table with 3 columns. First column for the class interval, second column for frequency f and the third column for cumulative frequency c.

**Step 2:** Write the class intervals and the corresponding frequency in the respective columns.

**Step 3:** Write the cumulative frequency in the column c. It is done by adding the frequency in each step

**Step 4:** Find the sum of frequencies N i.e., ∑f. It will be the same as the last number in the cumulative frequency column.

**Step 5:** Find N/2. Then find the class whose cumulative frequency is greater than and nearest to N/2. This is the median class.

**Step 6:** Now use the formula Median = $l+\frac{h}{f}\left(\frac{N}{2}-c\right)$

where

*l*= Lower class boundary of the model class

*f*= Frequency of the median class

N=∑f= Total frequency

c= Cumulative frequency of the class preceding the median class

h= Class interval size of the model class

## Problems:

**1.**Calculate the median from the following data:

| Group | 60 – 64 | 65 – 69 | 70 – 74 | 75 – 79 | 80 – 84 | 85 – 89 |
|---|---|---|---|---|---|---|
| Frequency | 1 | 5 | 9 | 12 | 7 | 2 |

## Solution:

| Group | *f* | Cumulative Frequency |
|---|---|---|
| 60 – 64 | 1 | 1 |
| 65 – 69 | 5 | 1+5=6 |
| 70 – 74 | 9 | 6+9=15 |
| 75 – 79 | 12 | 15+12=27 |
| 80 – 84 | 7 | 27+7=34 |
| 85 – 89 | 2 | 34+2=36 |
| | N=36 | |

← Median class

$$\because \left(\frac{N}{2}\right)^{th} = \frac{36}{2} = 18^{th}$$

*l*= Lower class boundary of the model class= 74.5

*f*= Frequency of the median class = 12

N= 36

c= Cumulative frequency of the class preceding the median class= 15

h= Class interval size of the model class= 5

Median = $l + \frac{h}{f}\left(\frac{N}{2} - c\right)$

$$= 74.5 + \frac{5}{12}\left(\frac{36}{2} - 15\right)$$

$$= 75.75$$

Median =75.75

## Median from Discrete Data

When the data follows a discrete set of values grouped by size, we use the formula $\left(\frac{N}{2}\right)^{th}$ item for finding the median. First form a cumulative frequency distribution, and the median is the value of x which corresponds to the cumulative frequency in which $\left(\frac{N}{2}\right)^{th}$ item lies.

**Problem:**

The following frequency distribution is classified according to the number of leaves on different branches. Calculate the median number of leaves per branch.

| No of Leaves | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| No of Branches | 2 | 11 | 15 | 20 | 25 | 18 | 10 |

## Solution:

| No of Leaves *X* | No of Branches *f* | Cumulative Frequency *C.* |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 11 | 13 |
| 3 | 15 | 28 |
| 4 | 20 | 48 |
| 5 | 25 | 73 |
| 6 | 18 | 91 |
| 7 | 10 | 101 |
| Total | N=101 | |

Median = Size of $\left(\frac{N}{2}\right)^{th}$ item

= $\left(\frac{101}{2}\right)^{th} = 50.5$

Median= Value of x corresponding to just greater than cumulative frequency 50.5= 5

3. If the median of the following frequency distribution is 28.5, find the missing frequencies.

| Class | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | Total |
|---|---|---|---|---|---|---|---|
| **Frequency** | f1 | 8 | 20 | 15 | 7 | f2 | 60 |

**Solution:**

| class | Frequency $f$ | Cumulative Frequency $C.$ |
|---|---|---|
| 0-10 | **f1** | **f1** |
| **10-20** | **8** | **f1+8** |
| **20-30** | **20** | **f1+28** |
| **30-40** | **15** | **f1+43** |
| **40-50** | **7** | **f1+50** |
| **50-60** | **f2** | **f1+f2+50** |
| **Total** | **N=60** | |

N=f1+8+20+15+7+f2=60

f1+f2=60-50

⇒f1+f2=10

It is given that the median is 28.5.

Clearly, 28.5 lies in the median class 20-30. So, 20-30 is the median class.

l=20, f=20, c=f1+8, N=60, h=10

Median = $l+\frac{h}{f}\left(\frac{N}{2}-c\right)$

$$28.5=20+\frac{10}{20}\left(\frac{60}{2}-(f1+8)\right)$$

$$28.5-20=\frac{1}{2}(30-f1-8)$$

$$8.5*2=(22-f1)$$

$$f1=22-17=5$$

$$f1+f2=10$$

$$f2=10-5=5$$

$$\therefore\ f1=5,\ f2=5$$

**Practice Problems:**

**1.**The number of magazines read by a group of women in a week is recorded. If the median of the distribution is 2, find the value of x.

| **Number of magazines** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **Number of women** | 5 | 2 | 1 | x |

**2.** The shoe size of 155 people was recorded and the raw data was presented in the form of the following frequency table:

| **Size of shoe** | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.5 | 7.0 | 7.5 | 8.0 |
|---|---|---|---|---|---|---|---|---|---|
| **Frequency** | 10 | 18 | 22 | 25 | 40 | 15 | 10 | 8 | 7 |

**3.** The marks obtained in English test by 17 students were recorded. What is the median marks of the students?

| **Marks out of 50** | **Frequency** |
|---|---|
| 0-10 | 2 |
| 10-20 | 4 |
| 20-30 | 5 |
| 30-40 | 4 |
| 40-50 | 2 |

## Merits of Median

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.

## Demerits of Median

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in calculating mean deviation.
4. It does not take into account all the observations.

# 3. Mode:

The mode is the value that appears most often in a set of data i.e. the value whose frequency is maximum. The mode of a discrete probability distribution is the value x at which its probability mass function takes the maximum value.

**Mode from Ungrouped Data**

Mode is calculated from ungrouped data by inspecting the given data. We pick out the value which occurs the greatest number of times in the data.

**Mode from Grouped Data**

With frequency distribution with equal class interval sizes, the class which has the maximum frequency is called the model class.

Mode= $l+\dfrac{f_m-f_1}{(f_m-f_1)+(f_m-f_2)}*h$

*l* = Lower class boundary of the model class

$f_m$ = Frequency of the model class (maximum frequency)

$f_1$ = Frequency preceding the model class frequency

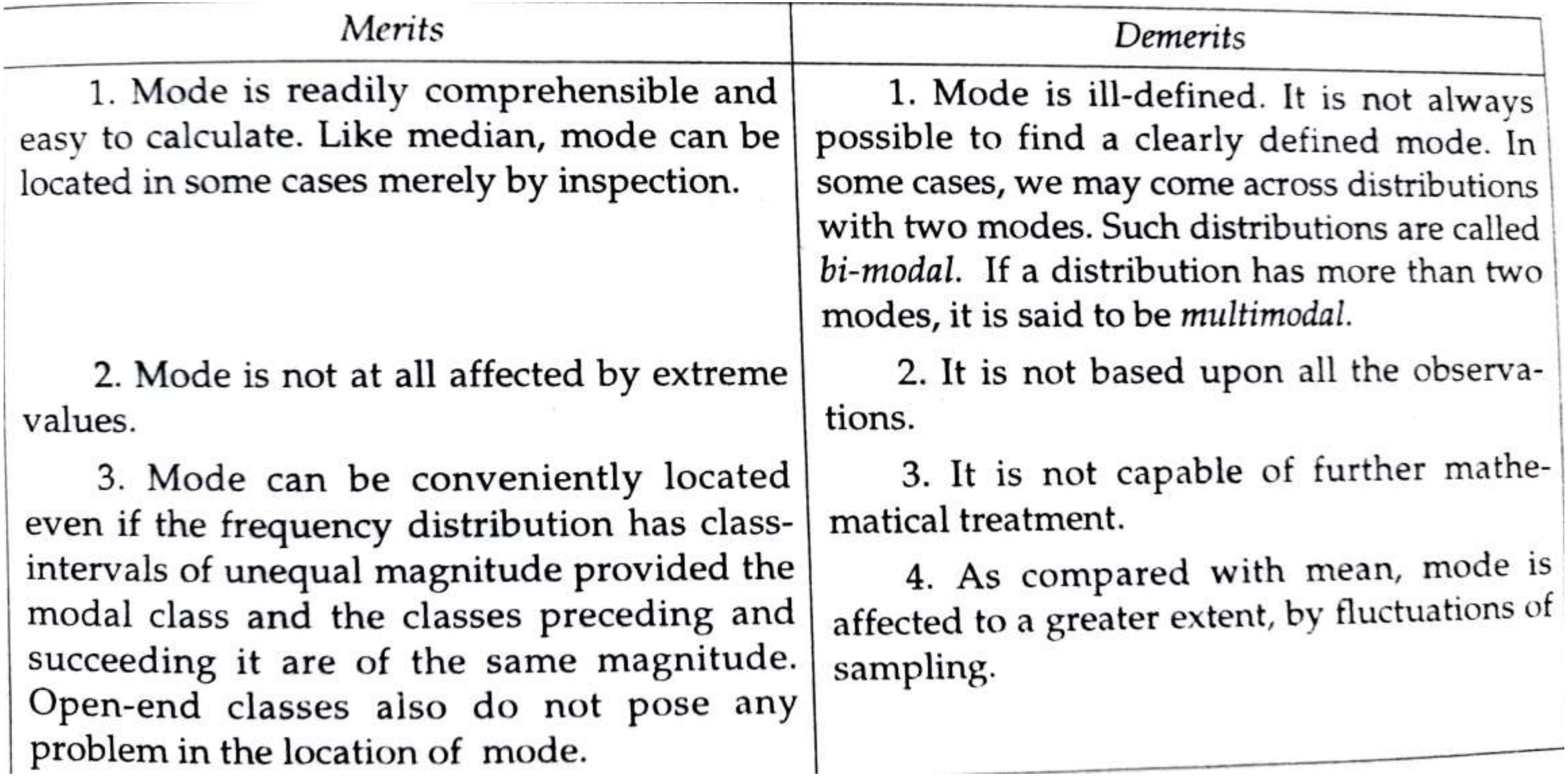
$f_2$ = Frequency following the model class frequency

h = Class interval size of the model class

## Mode from Discrete Data

When the data follows a discrete set of values, the mode may be found by inspection. The mode is the value of *X* corresponding to the maximum frequency.

# Merits and Demerits of Mode



| *Merits* | *Demerits* |
|---|---|
| 1. Mode is readily comprehensible and easy to calculate. Like median, mode can be located in some cases merely by inspection. | 1. Mode is ill-defined. It is not always possible to find a clearly defined mode. In some cases, we may come across distributions with two modes. Such distributions are called *bi-modal*. If a distribution has more than two modes, it is said to be *multimodal*. |
| 2. Mode is not at all affected by extreme values. | 2. It is not based upon all the observations. |
| 3. Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open-end classes also do not pose any problem in the location of mode. | 3. It is not capable of further mathematical treatment. |
| | 4. As compared with mean, mode is affected to a greater extent, by fluctuations of sampling. |

### 1. Example:

Find the mode of the values **5, 7, 2, 9, 7, 10, 8, 5, 7**

**Solution:**

The mode is **7** because it occurs the greatest number of times in the data.

### 2. Example:

The weights of 50 college students are given in the following table. Find the mode of the distribution.

| Weight (Kg) | 60 – 64 | 65 – 69 | 70 – 74 | 75 – 79 | 80 – 84 |
|---|---|---|---|---|---|
| No of Students | 5 | 9 | 16 | 12 | 8 |

**Solution:**

| Weight (Kg) | No of Students<br>*f* | Class Boundary |
|---|---|---|
| 60 – 64 | 5 | 59.5 – 64.5 |
| 65 – 69 | 9 | 64.5 – 69.5 |
| 70 – 74 | 16 | 69.5 – 74.5 |
| 75 – 79 | 12 | 74.5 – 79.5 |
| 80 – 84 | 8 | 79.5 – 84.5 |

← Highest frequency 16

Mode= $l+\frac{f_m - f_1}{(f_m - f_1)+(f_m - f_2)}*h$

$l$ = Lower class boundary of the model class = 69.5

$f_m$ = Frequency of the model class (maximum frequency)= 16

$f_1$= Frequency preceding the model class frequency=9

$f_2$ = Frequency following the model class frequency= 12

h = Class interval size of the model class= 5

$$=69.5+\frac{16-9}{(16-9)+(16-12)}*5$$

$$=72.68$$

Mode= 72.68

## Practice Problems:

1. Find the mode of the given data.

| Class | Frequency |
|---|---|
| 50-55 | 2 |
| 55-60 | 7 |
| 60-65 | 8 |
| 65-70 | 4 |

2. Alex timed 21 people in the sprint race, to the nearest second:

59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 58, 62, 62, 68, 65, 56, 59, 68, 61, 67

Find mode of the above distribution.

**3.** Find the mode of the data:

27,23,39,18,27,21,27,27,40,36,27

3. You grew fifty baby carrots using special soil. You dig them up and measure their lengths (to the nearest mm) and group the results:

| Length (mm) | Frequency |
|---|---|
| 150 - 154 | 5 |
| 155 - 159 | 2 |
| 160 - 164 | 6 |
| 165 - 169 | 8 |
| 170 - 174 | 9 |
| 175 - 179 | 11 |
| 180 - 184 | 6 |
| 185 - 189 | 3 |

Find mode of the above distribution.

# 4. Geometric mean:

The Geometric Mean (G.M.) of a set of n observations is the nth root of their product.

If $x_1, x_2, x_3, ----, x_n$ are $n$ observations then

$$G.M = \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n} = (x_1 \times x_2 \times x_3 \times \ldots \times x_n)^{\frac{1}{n}}$$

Taking the nth root of a number is difficult. Thus, the computation is done as

$$\log(G.M) = \log(x_1 \times x_2 \times x_3 \times \ldots \times x_n)^{\frac{1}{n}}$$

$$\log(G.M) = \frac{1}{n}\log(x_1 \times x_2 \times x_3 \times \ldots \times x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \ldots\ldots + \log x_n)$$

$$\log(G.M) = \frac{\sum_{i=1}^{n} \log x_i}{n}$$

$$G.M = anti\log\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right)$$

**<u>Example</u>:**

**1.** Find the geometric mean of the values 10, 5, 15, 8, 12.

**<u>Solution</u>:**

$$x_1 = 10, x_2 = 5, x_3 = 15, x_4 = 8\, and\; x_5 = 12$$

Here n=5

$$G.M = \sqrt[n]{x_1.x_2.x_3....x_n}$$

$$= \sqrt[5]{10\times 5\times 15\times 8\times 12}$$

$$= (72000)^{\frac{1}{5}} = 9.36$$

$$G.M = 9.36$$

(OR)

| x | log x |
|---|---|
| 10 | 1.0000 |
| 5 | 0.6990 |
| 15 | 1.1761 |
| 8 | 0.9031 |
| 12 | 1.0792 |
| **Total** | ∑log x=4.8573 |

$$G.M = anti\log\left(\frac{\sum_{i=1}^{n}\log x_i}{n}\right)$$

$$= anti\log\left(\frac{4.8573}{5}\right)$$

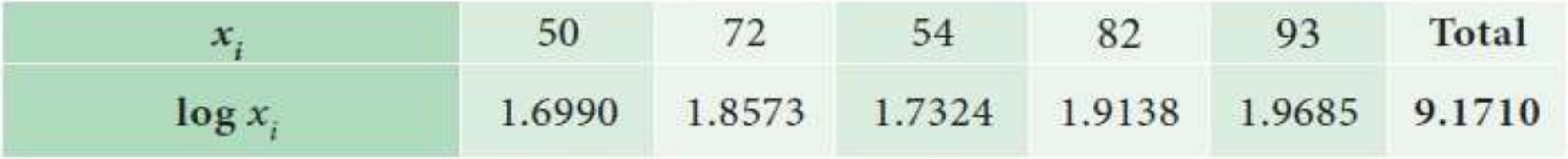
$$= anti\log(0.9715) = 9.36$$

$$G.M = 9.36$$

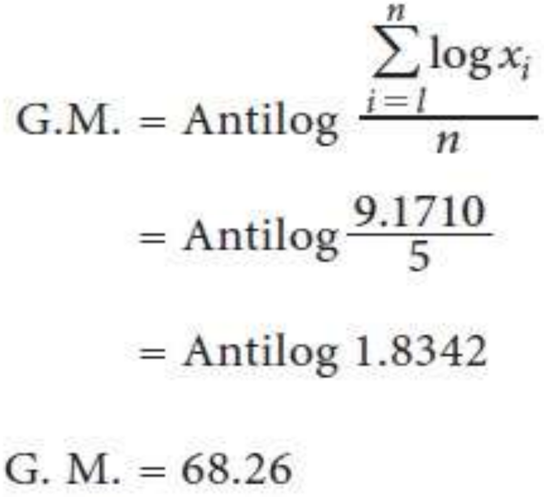
**2.** Calculate the geometric mean of the annual percentage growth rate of profits in business corporate from the year 2000 to 2005 is given below

50, 72, 54, 82, 93

***Solution:***



| $x_i$ | 50 | 72 | 54 | 82 | 93 | **Total** |
|---|---|---|---|---|---|---|
| **log $x_i$** | 1.6990 | 1.8573 | 1.7324 | 1.9138 | 1.9685 | **9.1710** |



$$\text{G.M.} = \text{Antilog}\,\frac{\sum_{i=1}^{n}\log x_i}{n}$$

$$= \text{Antilog}\,\frac{9.1710}{5}$$

$$= \text{Antilog } 1.8342$$

$$\text{G. M.} = 68.26$$

Geometrical mean of annual percentage growth rate of profits is 68.26

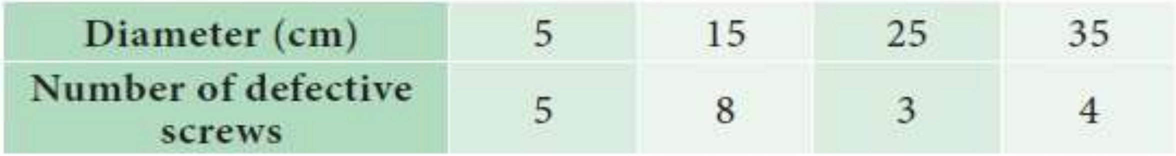
# Geometric mean for grouped data:

$$G.M = \sqrt[n]{{x_1}^{f_1} \times {x_2}^{f_2} \times {x_3}^{f_3} \times \ldots . \times {x_n}^{f_n}}$$

$or$

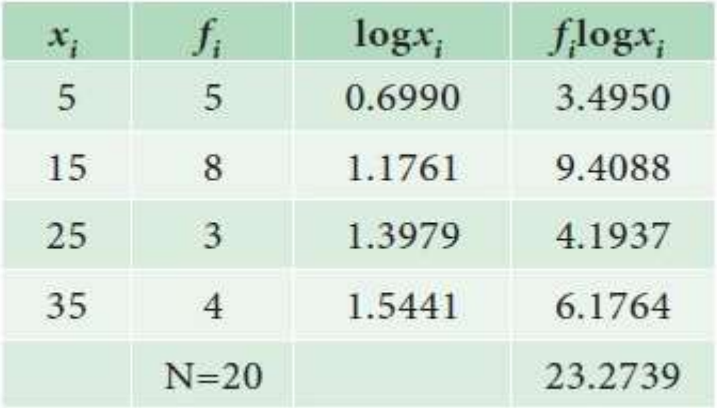
$$G.M = anti\log\left(\frac{\sum_{i=1}^{n} f_i \log x_i}{N}\right)$$

**1.** Find the G.M for the following data, which gives the defective screws obtained in a factory.



| Diameter (cm) | 5 | 15 | 25 | 35 |
|---|---|---|---|---|
| Number of defective screws | 5 | 8 | 3 | 4 |

***Solution:***



| $x_i$ | $f_i$ | $\log x_i$ | $f_i \log x_i$ |
|---|---|---|---|
| 5 | 5 | 0.6990 | 3.4950 |
| 15 | 8 | 1.1761 | 9.4088 |
| 25 | 3 | 1.3979 | 4.1937 |
| 35 | 4 | 1.5441 | 6.1764 |
| | N=20 | | 23.2739 |

$$\text{G.M} = \text{Antilog}$$

$$= \text{Antilog} \; \frac{\sum_{i=1}^{n} f_i \log x_i}{N}$$

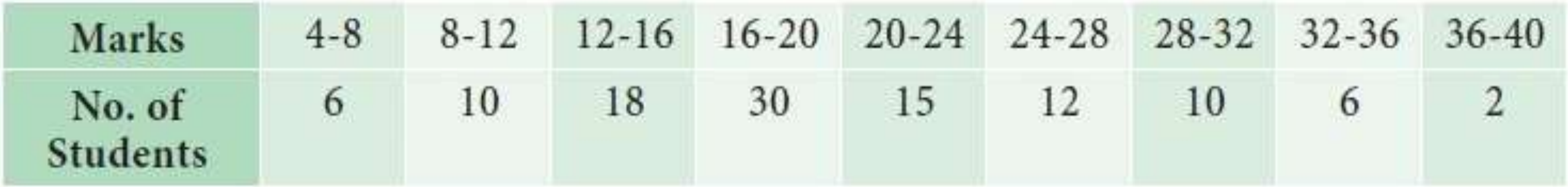
$$= \text{Antilog} \; \frac{23.2739}{20}$$

$$= \text{Antilog } 1.1637$$

$$\text{G.M} = 14.58$$

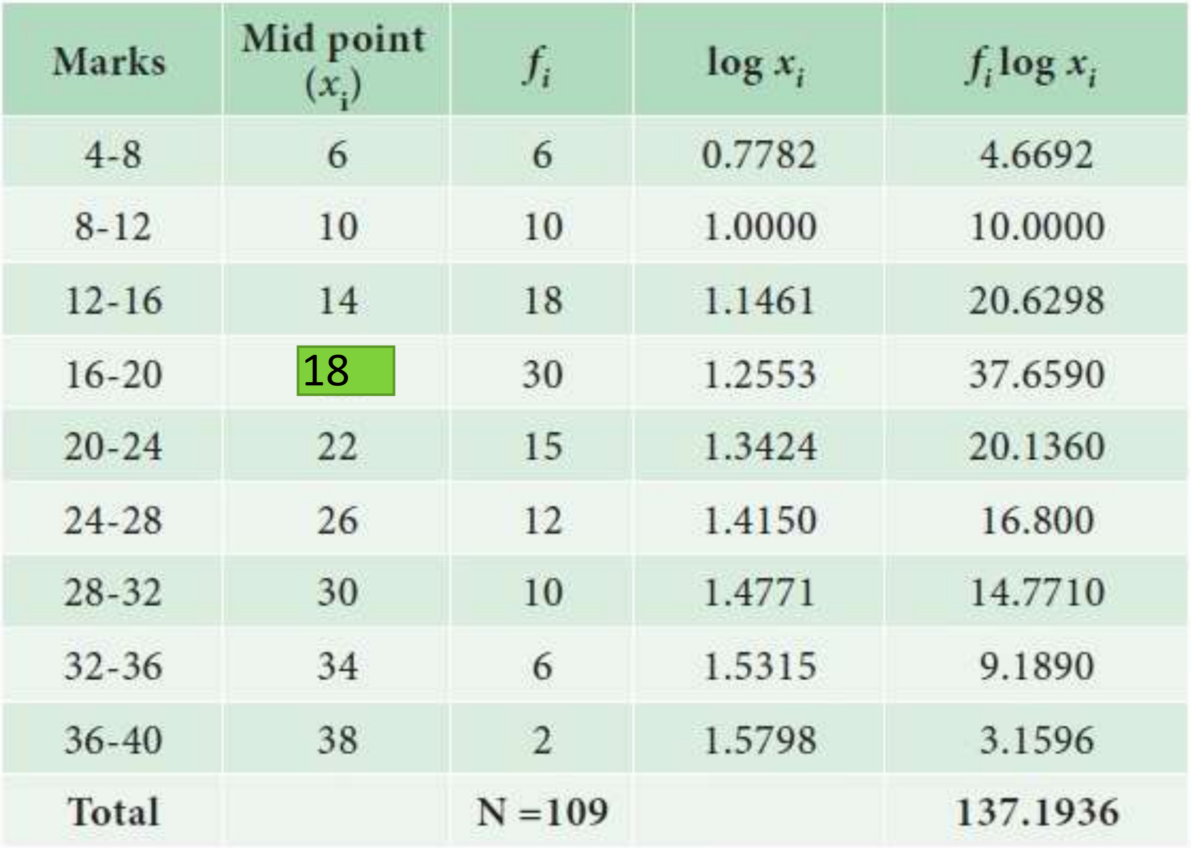
## (c) G.M. for Continuous grouped data

1. The following is the distribution of marks obtained by 109 students in a subject in an institution. Find the Geometric mean.



| Marks | 4-8 | 8-12 | 12-16 | 16-20 | 20-24 | 24-28 | 28-32 | 32-36 | 36-40 |
|---|---|---|---|---|---|---|---|---|---|
| No. of Students | 6 | 10 | 18 | 30 | 15 | 12 | 10 | 6 | 2 |

***Solution:***



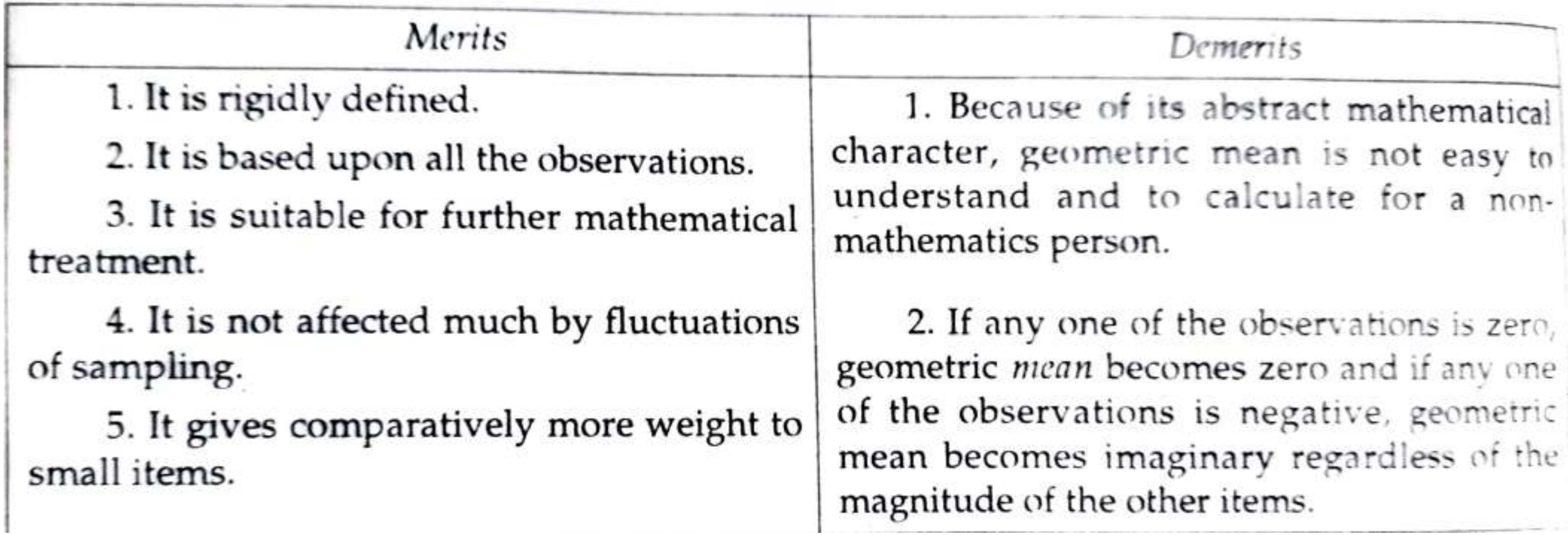
| Marks | Mid point $(x_i)$ | $f_i$ | $\log x_i$ | $f_i \log x_i$ |
|---|---|---|---|---|
| 4-8 | 6 | 6 | 0.7782 | 4.6692 |
| 8-12 | 10 | 10 | 1.0000 | 10.0000 |
| 12-16 | 14 | 18 | 1.1461 | 20.6298 |
| 16-20 | 18 | 30 | 1.2553 | 37.6590 |
| 20-24 | 22 | 15 | 1.3424 | 20.1360 |
| 24-28 | 26 | 12 | 1.4150 | 16.800 |
| 28-32 | 30 | 10 | 1.4771 | 14.7710 |
| 32-36 | 34 | 6 | 1.5315 | 9.1890 |
| 36-40 | 38 | 2 | 1.5798 | 3.1596 |
| **Total** | | **N =109** | | **137.1936** |

$$\text{G.M.} = \text{Antilog}\left[\frac{\sum_{i=l}^{n} f_i \log x_i}{N}\right]$$

$$= \text{Antilog}\left[\frac{137.1936}{109}\right] = \text{Antilog}\,[1.2587]$$

$$\text{G. M.} = 18.14$$

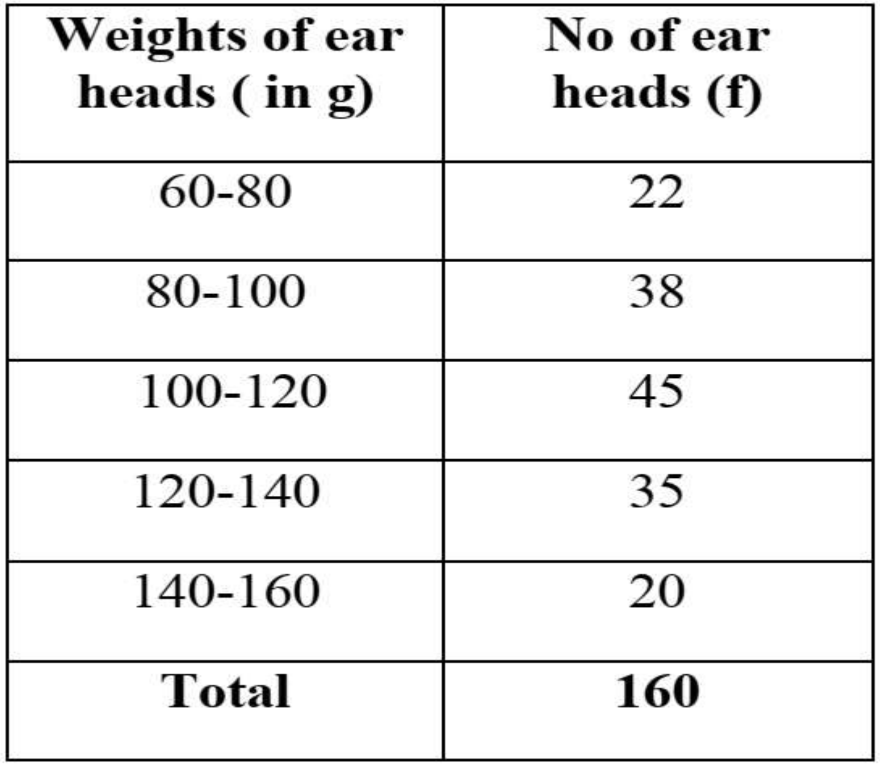
# Merits and Demerits of Geometric Mean:



| Merits | Demerits |
|---|---|
| 1. It is rigidly defined.<br>2. It is based upon all the observations.<br>3. It is suitable for further mathematical treatment.<br>4. It is not affected much by fluctuations of sampling.<br>5. It gives comparatively more weight to small items. | 1. Because of its abstract mathematical character, geometric mean is not easy to understand and to calculate for a non-mathematics person.<br>2. If any one of the observations is zero, geometric *mean* becomes zero and if any one of the observations is negative, geometric mean becomes imaginary regardless of the magnitude of the other items. |

# Practice problems:

1. For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the Geometric mean



| Weights of ear heads ( in g) | No of ear heads (f) |
|---|---|
| 60-80 | 22 |
| 80-100 | 38 |
| 100-120 | 45 |
| 120-140 | 35 |
| 140-160 | 20 |
| **Total** | **160** |

2. Find the geometric mean of the following data

| X | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|
| f | 2 | 5 | 13 | 7 | 3 |

**3)** Find the geometrical mean of the number 3, $3^2$, …, $3^n$.

# Harmonic mean (H.M)

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x1, x2…..xn are n observations,

$$\text{H.M} = \frac{n}{\sum_{i=n}^{n}\left(\frac{1}{x_i}\right)}$$

For a frequency distribution

$$\text{H.M} = \frac{N}{\sum_{i=n}^{n} f\left(\frac{1}{x_i}\right)}$$

## Problems:

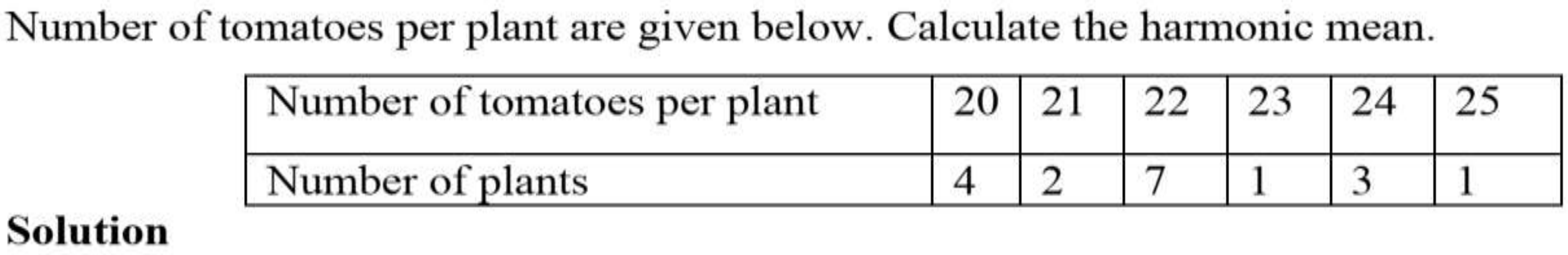
1. From the given data 5, 10,17,24,30 calculate H.M.

| X | $\frac{1}{x}$ |
|---|---|
| 5 | 0.2000 |
| 10 | 0.1000 |
| 17 | 0.0588 |
| 24 | 0.0417 |
| 30 | 0.4338 |

$$H.M = \frac{n}{\sum_{i=n}^{n}\left(\frac{1}{x_i}\right)}$$

$$H.M = \frac{5}{0.4338} = 11.526$$

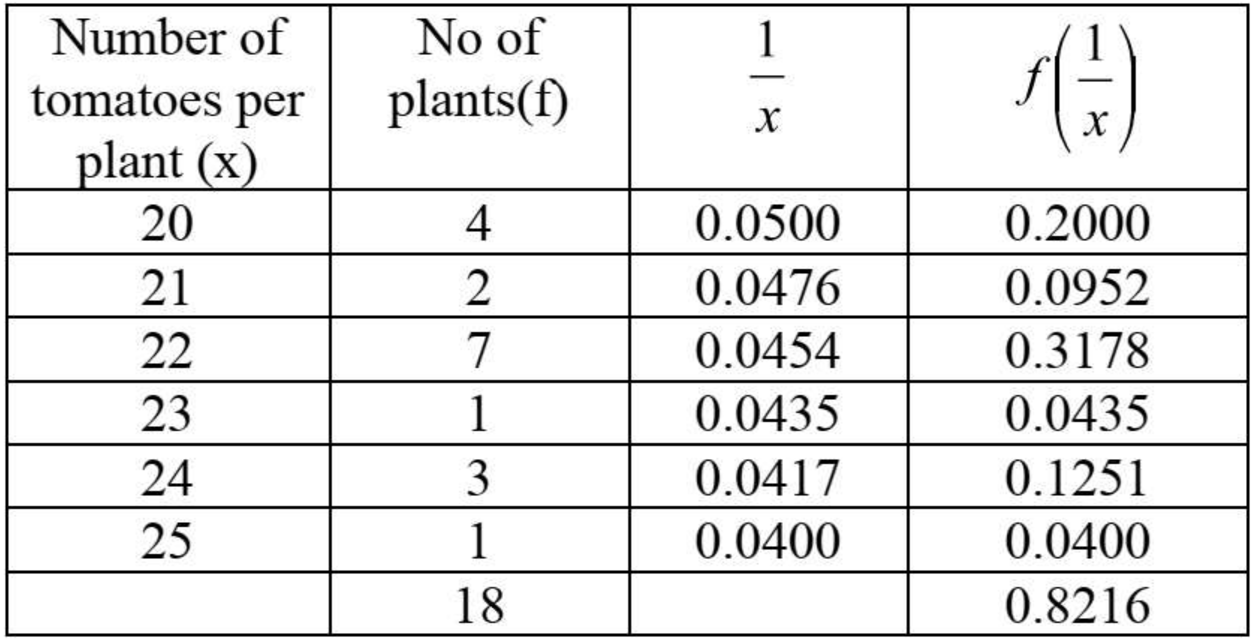
2.



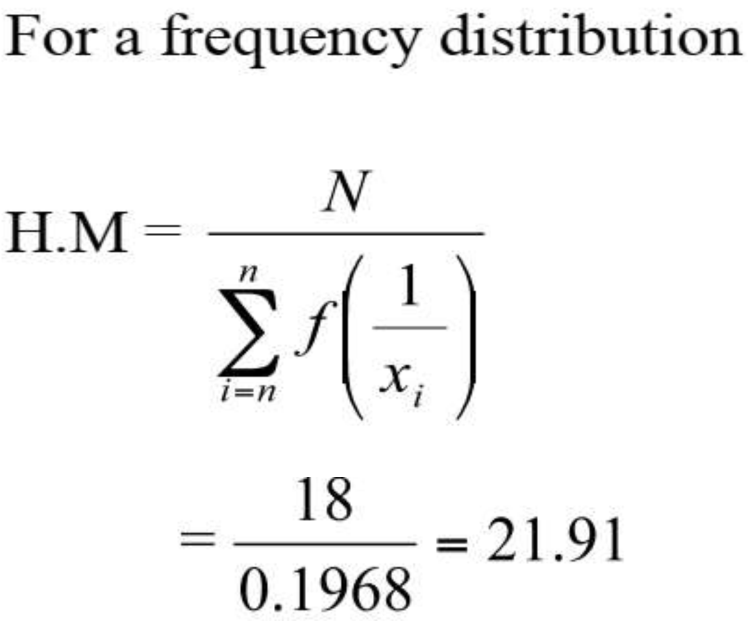
Number of tomatoes per plant are given below. Calculate the harmonic mean.

| Number of tomatoes per plant | 20 | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|---|
| Number of plants | 4 | 2 | 7 | 1 | 3 | 1 |

**Solution**



| Number of tomatoes per plant (x) | No of plants(f) | $\frac{1}{x}$ | $f\left(\frac{1}{x}\right)$ |
|---|---|---|---|
| 20 | 4 | 0.0500 | 0.2000 |
| 21 | 2 | 0.0476 | 0.0952 |
| 22 | 7 | 0.0454 | 0.3178 |
| 23 | 1 | 0.0435 | 0.0435 |
| 24 | 3 | 0.0417 | 0.1251 |
| 25 | 1 | 0.0400 | 0.0400 |
| | 18 | | 0.8216 |



For a frequency distribution

$$\text{H.M} = \frac{N}{\sum_{i=n}^{n} f\left(\frac{1}{x_i}\right)}$$

$$= \frac{18}{0.1968} = 21.91$$

**3.** Calculate the harmonic mean for the data given below:

| **Marks** | 30–39 | 40–49 | 50–59 | 60–69 | 70–79 | 80–89 | 90–99 |
|---|---|---|---|---|---|---|---|
| ff | 2 | 3 | 11 | 20 | 32 | 25 | 7 |

**Solution**:

| **Marks** | x | f | f/x |
|---|---|---|---|
| 30–39 | 34.5 | 2 | 0.0580 |
| 40–49 | 44.5 | 3 | 0.0674 |
| 50–59 | 54.5 | 11 | 0.2018 |
| 60–69 | 64.5 | 20 | 0.3101 |
| 70–79 | 74.5 | 32 | 0.4295 |
| 80–89 | 84.5 | 25 | 0.2959 |
| 90–99 | 94.5 | 7 | 0.0741 |
| **Total** | | ∑f=100 | ∑(f/x)=1.4368 |

For a frequency distribution

$$\text{H.M} = \frac{N}{\sum_{i=n}^{n} f\left(\frac{1}{x_i}\right)}$$

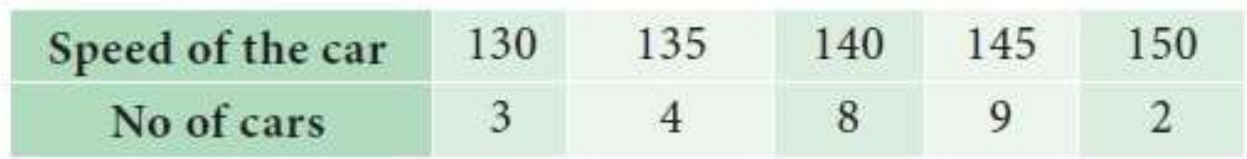
$$= \frac{100}{1.4368}$$

$$= 69.60$$

$$\therefore H.M = 69.6$$

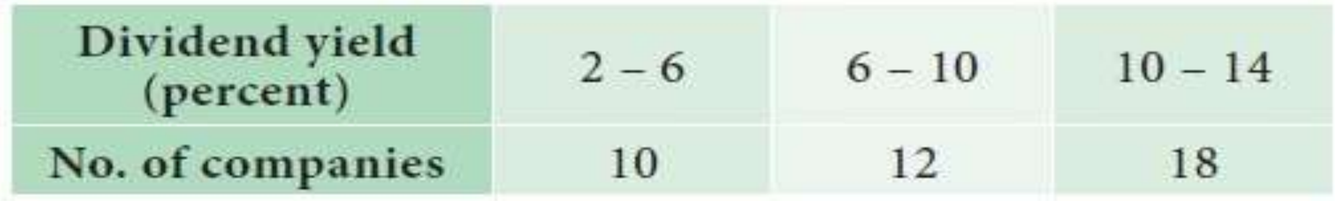
**Practice problems:**

1. The following data is obtained from the survey. Compute H.M



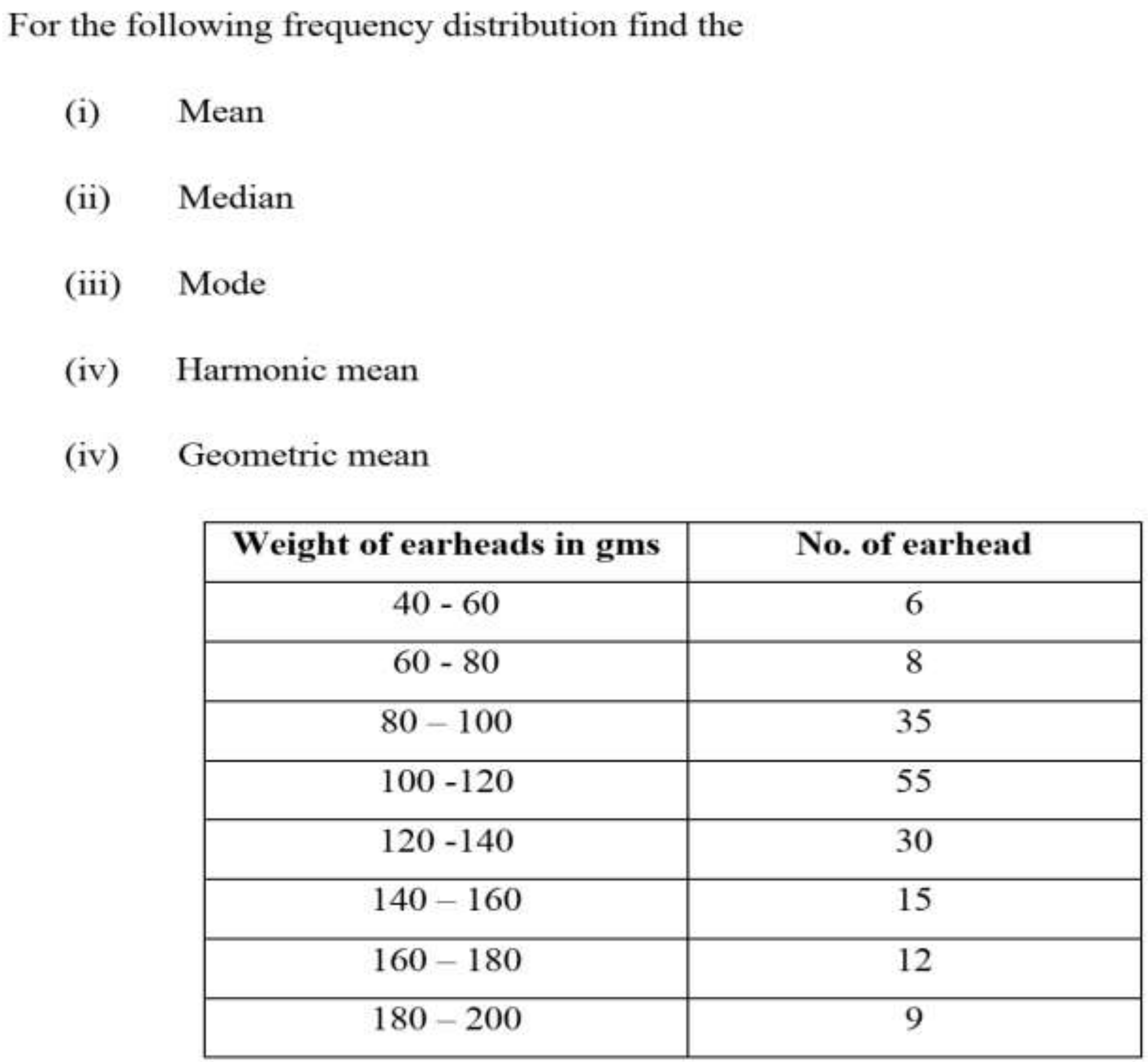
| Speed of the car | 130 | 135 | 140 | 145 | 150 |
|---|---|---|---|---|---|
| No of cars | 3 | 4 | 8 | 9 | 2 |

2. Find the harmonic mean of the following distribution of data



| Dividend yield (percent) | 2 – 6 | 6 – 10 | 10 – 14 |
|---|---|---|---|
| No. of companies | 10 | 12 | 18 |

3.



For the following frequency distribution find the

(i) Mean

(ii) Median

(iii) Mode

(iv) Harmonic mean

(iv) Geometric mean

| Weight of earheads in gms | No. of earhead |
|---|---|
| 40 - 60 | 6 |
| 60 - 80 | 8 |
| 80 – 100 | 35 |
| 100 -120 | 55 |
| 120 -140 | 30 |
| 140 – 160 | 15 |
| 160 – 180 | 12 |
| 180 – 200 | 9 |

## Merits of H.M

1. It is rigidly defined.

2. It is defined on all observations.

3. It is amenable to further algebraic treatment.

4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

## Demerits of H.M

1. It is not easily understood.

2. It is difficult to compute.

3. It is only a summary figure and may not be the actual item in the series

4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.

5. It is rarely used in grouped data.

## Measures of Variability:

A measure of variability is a summary statistic that represents the amount of dispersion in a dataset. While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the center.

There are many ways to describe variability or spread including:

1. Range
2. Interquartile range (IQR)
3. Variance and Standard Deviation

## Range:

The range is the difference in the maximum and minimum values of a data set. The maximum is the largest value in the dataset and the minimum is the smallest value. The range is easy to calculate but it is very much affected by extreme values.

Range=maximum−minimum

## Interquartile Range (IQR):

The interquartile range is the difference between upper and lower quartiles and denoted as **IQR**.

IQR=Q3−Q1=upper quartile−lower quartile=75th percentile−25th percentile

### Range and Interquartile range example:

1. You have 8 data points from Sample A.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Data (minutes) | 72 | 110 | 134 | 190 | 238 | 287 | 305 | 324 |

The highest value (*H*) is **324** and the lowest (*L*) is **72**.

$R = H – L$

$R$ = 324 – 72 = **252**

To find the interquartile range of 8 data points, first find the values at Q1 and Q3.

Multiply the number of values in the data set (8) by 0.25 for the 25th percentile (Q1) and by 0.75 for the 75th percentile (Q3).

Q1 position: 0.25 x 8 = 2

Q3 position: 0.75 x 8 = 6

Q1 is the value in the 2nd position, which is **110**. Q3 is the value in the 6th position, which is **287**.

IQR = Q3 – Q1

IQR = 287 – 110 = **177**

The interquartile range of given data is **177 minutes**.

# Variance:

Variance is the average squared difference of the values from the mean

Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

## Variance formula for populations

| Formula | Explanation |
|---|---|
| $\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$ | •$\sigma^2$ = population variance<br>•Σ = sum of…<br>•*X* = each value<br>•μ = population mean<br>•*N* = number of values in the population |

# Variance formula for samples

| Formula | Explanation |
|---|---|
| $s^2 = \dfrac{\Sigma\,(X - \bar{x})^2}{n-1}$ | •$s^2$ = sample variance<br>•Σ = sum of…<br>•*X* = each value<br>•x̄ = sample mean<br>•*n* = number of values in the sample |

## Standard deviation:

Square root of variance is called standard deviation.

## Standard deviation formula for Population:

| Formula | Explanation |
|---|---|
| $\sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$ | •σ = population standard deviation<br>•∑ = sum of…<br>•*X* = each value<br>•μ = population mean<br>•*N* = number of values in the population |

## Standard deviation formula for samples:

| Formula | Explanation |
|---|---|
| $s = \sqrt{\frac{\Sigma (X - \bar{x})^2}{n - 1}}$ | •*s* = sample standard deviation<br>•∑ = sum of…<br>•*X* = each value<br>•x̅ = sample mean<br>•*n* = number of values in the sample |

# Problems:

1. Calculate the variance and standard deviation for these final exam scores.

   61, 67, 73, 75, 76, 78, 81

**Solution:**

$$\bar{x} = \frac{sum\,of\ values}{no.of\ values}$$

$$= \frac{61+67+73+75+76+78+81}{7}$$

$$= 73$$

$$Variance(\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}$$

$$= \frac{(61-73)^2+(67-73)^2+(73-73)^2+(75-73)^2+(76-73)^2+(78-73)^2+(81-73)^2}{7}$$

$$= \frac{(-12)^2+(-6)^2+(0)^2+(2)^2+(3)^2+(5)^2+(8)^2}{7}$$

$$= \frac{144+36+4+9+25+64}{7}$$

$$= 40.28$$

$$Variance = 40.28$$

Standard deviation $= \sqrt{Variance}$

$$= \sqrt{40.28} = 6.3467$$

Standard deviation=6.3467

2.Find the variance, and the standard deviation for the following sample.

2 −3 6 0 3 1

Solution:

$$\bar{x} = \frac{2-3+6+0+3+1}{6}$$

$$= \frac{9}{6} = 1.5$$

$$\bar{x} = 1.5$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(2-1.5)^2 + (-3-1.5)^2 + (6-1.5)^2 + (0-1.5)^2 + (3-1.5)^2 + (1-1.5)^2}{6-1}$$

$$= \frac{(0.5)^2 + (-4.5)^2 + (4.5)^2 + (-1.5)^2 + (1.5)^2 + (-0.5)^2}{5}$$

$$= \frac{.25 + 20.25 + 20.25 + 2.25 + 2.25 + 0.25}{5}$$

$$= 9.1$$

$$Variance = 9.1$$

$$S\tan dard\ deviation = \sqrt{Variance}$$

$$= \sqrt{9.1} = 3.017$$

$$S\tan dard\ deviation = 3.017$$

3. Find mean , variance and standard deviation of the given dataset I

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Data Set I: | 46 | 37 | 40 | 33 | 42 | 36 | 40 | 47 | 34 | 45 |

## Variance and standard deviation for grouped data:

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i (x_i - A)^2}{N} \,(or)\, \frac{\sum_{i=1}^{n} f_i d_i{}^2}{N}$$

$$where\; N = \sum f_i$$

$$S\tan dard\; deviation = \sqrt{Variance}$$

1. Calculate the variance of the following table.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| x | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
| f | 4 | 4 | 5 | 15 | 8 | 5 | 4 | 5 |

**Solution :**

| x | f | d = x-8 | $d^2$ | $fd^2$ |
|---|---|---|---|---|
| 2 | 4 | 2-8 = -6 | 36 | 144 |
| 4 | 4 | 4-8 = -4 | 16 | 64 |
| 6 | 5 | 6-8 = -2 | 4 | 20 |
| 8 | 15 | 8-8 = 0 | 0 | 0 |
| 10 | 8 | 10-8 = 2 | 4 | 32 |
| 12 | 5 | 12-8 = 4 | 16 | 80 |
| 14 | 4 | 14-8 = 6 | 36 | 144 |
| 16 | 5 | 16-8 = 8 | 64 | 320 |

$\Sigma f = 50$ and $\Sigma fd^2 = 804$

$\sigma^2$ = $(\Sigma fd^2/\Sigma f)$

= (804/50)

= 16.08

Standard deviation= $\sqrt{Variance}$

$$= \sqrt{16.08} = 4.009$$

Standard deviation=4.009

**Question 2 :**

Find the variance and standard deviation of the following distribution

| Class interval | Frequency |
|---|---|
| 20-24 | 15 |
| 25-29 | 25 |
| 30-34 | 28 |
| 35-39 | 12 |
| 40-44 | 12 |
| 45-49 | 8 |

## Solution :

Here we consider the first data that is class interval as (x) and no of frequency as (f).

| Class interval | x | f | d = x-32 | $d^2$ | $fd^2$ |
|---|---|---|---|---|---|
| 20-24 | 22 | 15 | 22-32 = -10 | 100 | 1500 |
| 25-29 | 27 | 25 | 27-32 = -5 | 25 | 625 |
| 30-34 | 32 | 28 | 32-32 = 0 | 0 | 0 |
| 35-39 | 37 | 12 | 37-32 = 5 | 25 | 300 |
| 40-44 | 42 | 12 | 42-32 = 10 | 100 | 1200 |
| 45-49 | 47 | 8 | 47-32 = 15 | 225 | 1800 |

$\Sigma fd^2 = 5425$ and $\Sigma f = 100$

$\sigma^2 = (\Sigma fd^2/\Sigma f)$

$= (5425/100)$

Variance= 54.25

Standard deviation= $\sqrt{\text{var}\, iance} = \sqrt{54.25}$

=7.37

Standard deviation =7.37