BASIC STATISTICS

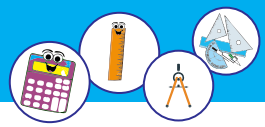
Unit 22

- Weightage = 9%

Students Learning Outcomes (SLOs)

After completing this unit, students will be able to:

- Know about frequency distribution, and learn how to:
  - Construct a grouped frequency table.
  - Construct histograms with equal class intervals.
  - Construct histograms with unequal class intervals.
  - Construct a frequency polygon.
- Know about cumulative frequency distribution, and learn how to:
  - Construct a cumulative frequency table.
  - Draw a cumulative frequency polygon.
- Know about measures of central tendency, and learn how to:
  - Calculate (for ungrouped and grouped data):
    - Arithmetic mean by definition and using deviations from assumed mean.
    - Median, mode, geometric mean and harmonic mean.
  - Recognize properties of arithmetic mean.
  - Calculate weighted mean and moving averages.
  - Estimate median, quartiles and mode, graphically.
- Know about measures of dispersion, and learn how to:
  - Define, identify and measure range, calculate variance, mean deviation and standard deviation.



Introduction:

Statistics is an important branch of Mathematics which deals with the collection, organization, representation, interpretation and analysis of data to derive meaningful insights and patterns which help take wiser decisions and frame efficient policies. We already know basic concepts of classification, graphical representation and some measures of central tendency of simpler data. Here, we explore these in more detail and extend the concepts towards moving averages and measures of dispersion.

Data are basis of any statistical investigation, and are gathered on a variable of interest by asking questions, counting and measuring, referring reports, news, articles. The very basic form of data are the **raw data**, which are **unclassified/ ungrouped**, and it is difficult to directly get any meaningful conclusion from it. Such data are obtained using surveys, interviews or questionnaires, also referred as **primary sources** of data. After classifying and organizing raw/ungrouped data, we can get useful information through **grouped/classified data**. Newspapers, reports and articles are referred as **secondary sources** of data because these lead to classified data.

Data are divided into two types: **qualitative (categorical)** or **quantitative (numeric)**. Data on gender, blood group, eyebrow color, grades, roll numbers, etc. are qualitative data. Data on height, weight, salary, pH values of solutions, annual profits of a company, marks of students in a subject etc. are quantitative.

Observations in qualitative data are purely attributes/categories without any numerical significance, but these may or may not possess ranking/ordering (ascending or descending). Qualitative data which does not possess ordering are **nominal**, otherwise **ordinal**. Gender, blood group, eye brow color, religion lead to nominal data. For example, we cannot rank male/female observations as to which is higher/lower. Data on grades of students (A, B, C, Fail), quality of food (excellent, good, fair, bad, worst) can be ordered/ranked, and lead to ordinal data.

Observations in quantitative data are numbers with numerical significance and ordering. The numbers may be integers only or also in decimal form. If observations in a quantitative data are only integers, then data are **discrete**, otherwise **continuous**. Data on salary (in Rs.), number of female students per class in a school, number of heads while tossing 4 coins simultaneously lead to discrete data. Data on measuring heights, weights, pH values of solutions are continuous data.

It is extremely important to distinguish between types of data to appropriately apply some statistical tests or to measure any indicators for further analysis of data.

22.1. Frequency distribution:

To extract meaningful information from ungrouped data with higher number of observations, we divide data into smaller **classes/groups**. The classes are constructed to include all similar observations at one place. The **discrete classes** comprise of a single number/attribute, whereas **continuous classes** contain a range of numbers. The number of observations falling in a particular class is called its **frequency**.

A **frequency distribution** is a tabular description of a grouped data, consisting of classes and corresponding frequencies. Data arranged into a frequency distribution provide clearer understanding for the basic analysis than the ungrouped/raw data.

The **relative** and **percentage frequencies** of a class with frequency "f" are: $\left(\frac{f}{n}\right)$ and $\left(\frac{f}{n}\times 100\right)$, respectively, where "$n$" is total number of observations. The sum of all relative frequencies is 1, and of percentage frequencies is 100.

22.1(i) Construct grouped frequency table:

The procedure to construct a grouped frequency table or a frequency distribution depends on the way we form groups/classes. We discuss the following two ways of construction.

(a) Frequency table with discrete classes:

Discrete classes refer to the classes with only a single number/attribute to include similar observations. We use discrete classes only when a fewer observations repeat more number of times in the data. Frequency table with discrete classes is also referred as discrete frequency table. Sometimes, the relative and percentage frequencies are also added in the table in last if required.

Given a dataset, we follow the following steps to construct its frequency table with discrete classes comprising of three columns.

1. Identify distinct observations as **classes/groups** and write these in the first column in any order if data are nominal. For ordinal and quantitative data, write the classes in ascending order.
2. Use **tally marks** to distribute data into the classes in the second column.
3. Count tally marks to write **frequency** of each class in third column.

Example 1:

Construct grouped frequency table of the blood groups of 30 students of a class.

A+ O– B+ A+ A– AB+ B+ O+ O+ A– A+ AB– A+ O+ B+

B+ O+ B+ O– AB+ O– A+ AB+ A+ O+ B– O+ A– AB– O+

Solution:

Here, n = 30 and data are nominal. The eight distinct blood groups: A+, B+, AB+, O+, A–, B–, AB– and O– are the eight classes which are written in first column. Using tally marks in the second column to distribute data and then counting these to write frequencies in the third column, we get the required grouped frequency table for the distribution of blood groups of students.

Grouped Frequency Table

Classes/Groups "Blood groups"	Tally marks	Frequency "Number of students"
A+	𝍸 |	6
B+	𝍸	5
AB+	|||	3
O+	𝍸 ||	7
A–	|||	3
B–	|	1
AB–	||	2
O–	|||	3
Total	---	n = 30

Example 2:

The responses of 40 randomly selected customers in a shopping mall when asked "How satisfied are you with the services?" are given below. Construct a grouped frequency table of their responses. Also compute relative and percentage frequencies.

Very	Not very	Not very	Very	Somewhat	Not at all	Very	Not very
Very	Not sure	Very	Very	Very	Not very	Very	Somewhat
Not sure	Very	Not at all	Very	Very	Very	Not very	Somewhat
Not at all	Not sure	Not very	Not very	Not sure	Very	Very	Very
Very	Somewhat	Very	Somewhat	Not very	Not sure	Very	Somewhat

Solution:

Here, $n = 40$, and data are ordinal. The five distinct satisfaction levels in ascending order (lowest to highest) are classes. With tally marks and frequencies we get the grouped frequency table. The relative and percentage frequencies are also computed in fourth and fifth columns.

Classes/Groups "Satisfaction levels"	Tally marks	Frequency (f)	Relative frequency $\left(\frac{f}{n}\right)$	Percentage frequency (%) $\left(\frac{f}{n}\times 100\right)$
Not at all	𝍸	3	$\frac{3}{40} = 0.075$	$\frac{3}{40}\times 100 = 7.5$
Not very	𝍸 |||	8	0.2	20
Not sure	𝍸	5	0.125	12.5
Somewhat	𝍸 |	6	0.15	15
Very	𝍸 𝍸 𝍸 |||	18	0.45	45
Total	---	$n = 40$	1	100

Note that:

a. From grouped frequency table of Example 2, we can easily conclude that highest number of customers were "Very satisfied", where as a fewer number of customers were "Not satisfied at all".

b. It is appropriate and understandable to say that 45% customers were very satisfied instead of 18 out of 40 were very satisfied.

Example 3:

The responses of 45 families in a city, when asked about number of mobiles they used, were noted as follows. Construct a discrete frequency table of the responses. Also obtain relative and percentage frequencies of the responses.

3 1 3 2 2 2 2 1 2 1 2 2 3 3 3 3 4 1 3 0 2 4 3

3 3 2 3 2 2 5 1 6 1 6 2 1 5 3 2 4 2 4 7 4 2

Solution:

Here, $n = 45$ and data are quantitative, and in particular discrete.

The distinct numbers in the data in ascending order are: 0, 1, 2, 3, 4, 5, 6, 7 which are fewer numbers repeating again and again. These are the 8 classes in the first column.

Writing and counting the tally marks in second column, we compute frequencies in the third column to get the required discrete frequency table of distribution of number of mobiles possessed by 45 families.

Discrete Frequency Table

Classes/Groups "Number of mobiles"	Tally marks	Frequency "Number of families"
0	|	1
1	𝍸 ||	7
2	𝍸 𝍸 𝍸	15
3	𝍸 𝍸 ||	12
4	𝍸	5
5	||	2
6	||	2
7	|	1
Total	---	$n = 45$

Example 4:

The amount of cold drink was measured (in liters, L) in 20 randomly selected 1.5L bottles of a company as given below. Construct a discrete frequency table of the measured amount of cold drink. Also compute relative and percentage frequencies.

1.48 1.51 1.50 1.49 1.49 1.49 1.51 1.48 1.49 1.52
1.51 1.49 1.51 1.50 1.50 1.50 1.51 1.49 1.49 1.50

Solution:

The data are quantitative, and in particular continuous. The distinct numbers: 1.48, 1.49, 1.50, 1.51 and 1.52 (in ascending order) are five classes. With tally marks and frequencies, we get the discrete frequency table. The relative and percentage frequencies are also computed in fourth and fifth columns.

Class limits "Amount of cold drink (L)"	Tally marks	Frequency "Number of bottles"	Relative frequency $\left(\frac{f}{n}\right)$	Percentage frequency (%) $\left(\frac{f}{n}\times 100\right)$
1.48	||	2	$\frac{2}{20} = 0.1$	$\frac{2}{20}\times 100 = 10$
1.49	𝍸 ||	7	0.35	35
1.50	𝍸	5	0.25	25
1.51	𝍸	5	0.25	25
1.52	|	1	0.05	5
Total	---	$n = 20$	1	100

(b) Frequency table with continuous classes:

When data are quantitative (discrete or continuous) and observations are split over a range of number with lesser repetitions and more variations, then we use **continuous classes**. These classes are not single numbers but a range of numbers (intervals) each consisting of all numbers within the limits of the class. A **continuous frequency table** comprises of continuous classes together with respective frequencies.

We have already studied frequency distribution for discrete data in previous class. Here, we will generalize and use the basic concepts for discrete and continuous data with any number

of decimal places. We first recall and define some important terms which are actively used in the procedure.

Range:

It is the difference between highest and lowest observations in the given data.

Range = [Highest observation] – [Lowest observation] (1)

Class limits:

These are numbers used to identify a class. For each class, there is a smallest number, called **lower class limit** (LCL), and a largest number called **upper class limit** (UCL). The observations starting from the LCL up to the UCL fall in a particular class.

Class interval/ width (h):

It is defined as the **size / length** of a class, and computed by finding difference between any two consecutive LCLs or UCLs. We usually use constant class width/interval, denoted by h for all classes and calculated as:

$$h = \frac{1}{10^m}\left\lceil \frac{\mathrm{R}}{\mathrm{K}} \times 10^m \right\rceil \quad (2)$$

where, K is number of classes, R is range, and m is number of maximum decimal places in the observations.

If data are discrete, then m = 0, if data are continuous with values up to one decimal places, then m = 1, and so on. $\lceil . \rceil$ denotes the **ceiling approximation** to get final integer value of h. The ceiling of an integer is the same integer, whereas of a decimal fraction is an integer immediately greater than it.

For example $\lceil 4 \rceil = 4, \lceil 1.6 \rceil = 2, \lceil 1.9 \rceil = 2, \lceil 7.5 \rceil = 8.$

Number of classes (K):

The **number of classes** is denoted by K and it ranges from 5 to 15 depending on number of observations and range. It should be carefully chosen/calculated, if not given. Much higher value of K results in poor grouping and loss of information. H. Sturges (1926) suggested a rule to calculate desirable number of classes K by using the numbers of observations n. The Sturges' rule is defined as:

$$\mathrm{K} = \lceil 1 + 3.322 \log(\mathrm{n}) \rceil \quad (3)$$

where, log (n) is logarithm of n with base 10, and $\lceil . \rceil$ is ceiling approximation. It should be noted that Sturges' rule is mostly appropriate for $15 \leq n \leq 200$. For example, if $n = 25$, then: $K = \lceil 1 + 3.322 \log(25) \rceil = \lceil 5.6439 \rceil = 6$ classes.

Spacing between the classes (d):

The constant difference between the LCL of a class and UCL of the next class is **spacing between the classes (*d*)**. If m is number of maximum decimal digits, then:

$$d = \frac{1}{10^m}, \; m = 0, 1, 2, \ldots \quad (4)$$

For discrete data, $m = 0$ and $d = 1$. If data are given up to one digit after the decimal, then $m = 1$ and $d = 0.1$. For data up to 2 decimal places, $m = 2$ and $d = 0.01$.

Procedure to construct continuous frequency table:

The steps to construct continuous frequency table for quantitative data are:

1. Note number of observations (n) and identify type of data: discrete or continuous. Also, note the maximum number of decimal places in data (m).
2. Calculate range (R) using equation (1), number of classes (K) using Sturges' rule in equation (3), if not given.
3. Calculate class interval (h) using equation (2) and spacing between the classes (d) using equation (4). In extreme conditions, we may have to take *"$h + d$"* as class width or add one more class.
4. Determine class limits to assure that all observations are included into the K classes. We can also use the lowest observation as starting LCL. The starting UCL is: $\text{UCL} = \text{LCL} + h - d$. Write LCLs and UCLs of other classes by continuously adding h in starting LCL and UCL.
5. Write class limits as classes in the first column of continuous frequency table.
6. Use **tally notation** / **list entries** to distribute data into classes in second column. Tally marks are preferred.
7. Count tally marks or entries to write frequencies of all classes in third column.

Example 1:

The electricity consumption (in kWh) in a shop was noted for 60 consecutive days as mentioned below. Construct frequency distribution using tally method.

106	107	76	82	109	107	115	93	187	95	139	119	115	128	115
123	125	111	92	86	70	126	68	130	129	194	82	90	158	118
123	146	80	136	137	110	141	152	104	111	140	184	204	178	75
113	162	131	99	185	181	84	486	100	98	148	90	110	107	78

Solution:

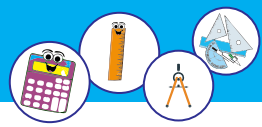
Here, $n = 60$, data are discrete with $m = 0$. $\text{R} = 204 - 68 = 136$. Using Sturges' rule for number of classes:

$\text{K} = \lceil 1 + 3.322 \log(60) \rceil = \lceil 6.9070 \rceil = 7.$

$h = \left\lceil \frac{136}{7} \right\rceil = \lceil 19.4285 \rceil = 20$, and $d = 1$. With the starting LCL=68, the starting UCL=68+20–1=87. Adding h=20 in these, we get all other limits. Writing class limits in first column, tally marks in second, and frequencies in third, we get the required frequency table.

Frequency Table

Class limits	Tally marks	Frequency
68 – 87	𝍸 𝍸	10
88 – 107	𝍸 𝍸 |||	13
108 – 127	𝍸 𝍸 𝍸	15
128 – 147	𝍸 𝍸	10
148 – 167	||||	4
168 – 187	𝍸 |	6
188 – 207	||	2
Total	---	$n = 60$



Example 2:

The diameters (in mm) of a reel of wire measured at 24 places correct to nearest 0.01mm are: 2.10, 2.29, 2.32, 2.21, 2.14, 2.22, 2.28, 2.18, 2.17, 2.20, 2.23, 2.13, 2.26, 2.10, 2.21, 2.17, 2.28, 2.15, 2.34, 2.27, 2.11, 2.23, 2.25, 2.16. Construct frequency distribution.

Solution:

$n = 24$, data are continuous up to two digits after decimal, so: $m = 2$.

$R = 2.34 - 2.10 = 0.24$, and $K = \lceil 1 + 3.322\log(24) \rceil = 6$.

$h = \frac{1}{100}\left\lceil \frac{R}{K} \times 100 \right\rceil = 0.04$, $d = \frac{1}{100} = 0.01$. Starting with LCL=2.10 and $h = 0.04$, six classes are: 2.10-2.13, 2.14-2.17, 2.18-2.21, 2.22-2.25, 2.26-2.29, 2.30-2.33. But, 2.34 cannot be included in any class. In this extreme case, we use class width of $h + d = 0.05$. The required six classes are: 2.10 - 2.14, 2.15 - 2.19, 2.20 - 2.24, 2.25 - 2.29, 2.30 - 2.34, 2.35 - 2.39, and the required frequency distribution of the diameters is:

Frequency Table

Class limits	2.10 – 2.14	2.15 – 2.19	2.20 – 2.24	2.25 – 2.29	2.30 – 2.34	2.35 – 2.39	Total
Tally mark	𝍸	𝍸	𝍸	𝍸	𝍸		---
Frequency	3	5	6	6	3	0	$n = 24$

Note that:

In Example 2, we could also add just one more class to include 2.34. The table becomes:

Class limits	2.10 – 2.13	2.14 – 2.17	2.18 – 2.21	2.22 – 2.25	2.26 – 2.29	2.30 -2.33	2.34 – 2.37
Tally mark	𝍸	𝍸	𝍸	𝍸	𝍸	𝍸	𝍸
Frequency	4	5	4	4	5	1	1

Some more important terms in grouped frequency tables with continuous classes are:

Class boundaries:

The numbers which separate a class form adjoining classes are called **class boundaries**. For a class, its **lower class boundary** (LCB) and **upper class boundary** (UCB) are obtained from the LCL and UCL of that class, respectively as:

For a class: $\text{LCB} = \text{LCL} - \frac{d}{2}$ and $\text{UCB} = \text{UCL} + \frac{d}{2}$

where, d is the spacing between the classes.

Class marks/midpoints:

Class marks or **mid-points**, denoted as x, in a grouped data with continuous classes are the arithmetic means of the class limits or class boundaries of the classes.

For a class: $x = \frac{\text{LCL}+\text{UCL}}{2}$ or $x = \frac{\text{LCB}+\text{UCB}}{2}$

Cumulative Frequency:

The word **cumulative** means something increasing/growing by successive additions. If we go on adding frequencies of classes, one after one, each time adding frequency of next class in the total, we get cumulative frequencies. For a particular class, the **cumulative frequency** is the sum of all frequencies up to that class.

Example 3:

Compute class boundaries, class marks and cumulative frequencies for the frequency table given below.

Class limits	7.1-7.3	7.4-7.6	7.7-7.9	8.0-8.2	8.3-8.5	8.6-8.8	8.9-9.1
Frequencies	3	5	9	14	11	6	2

Solution:

Here, $n = 50$ and $d = 0.1$. Class boundaries, class marks and cumulative frequencies are computed in the following table.

Class limits	Class boundaries	Class marks	Frequencies	Cumulative frequencies
7.1-7.3	7.05-7.35	7.2	3	3
7.4-7.6	7.35-7.65	7.5	5	3+5=8
7.7-7.9	7.65-7.95	7.8	9	17
8.0-8.2	7.95-8.25	8.1	14	31
8.3-8.5	8.25-8.55	8.4	11	42
8.6-8.8	8.55-8.85	8.7	6	48
8.9-9.1	8.85-9.15	9.0	2	n = 50
Total	**----**	**----**	**n = 50**	**----**

Note that:

1. We can observe that the UCB of a class and LCB of the next class are same.
2. Adding h successively in starting LCB, we get all remaining class boundaries.
3. We can get all remaining class marks by adding "h" in the starting class marks.
4. The cumulative frequency of the last class is the total number of observations.

EXERCISE 22.1

1. Identify type of data (discrete/ continuous/ nominal/ ordinal) and total number of observations in the following:

- **a.** Number of visitors in a shopping mall per day for 30 days.
- **b.** Time (in minutes per day) spent by a person in GYM for a month.
- **c.** Life (in years) of 45 car batteries.
- **d.** Electricity consumption (in kWh/day) for 2 months.
- **e.** Weight (in Kg) of 55 students in a class.
- **f.** Eyebrow colors of 20 individuals.
- **g.** Grades obtained by 70 students in a class.
- **h.** Result (Positive/ Negative) of COVID diagnosis test of 10 people.
- **i.** Marital status of 30 teachers of a school.
- **j.** District of domicile of students of your class.

2. Construct grouped frequency table of the educational levels (current class of study) of 20 students going to a country tour from a school. The educational levels are: VIII, X, X, VIII, VI, IX, X, IX, X, VIII, VII, VIII, VI, X, VI, V, IV, VII, IX, IX.

3. The mode of transport used by 20 students to attend a language course were noted as given below. Construct grouped frequency table with percentage frequencies.

Walk	Walk	Bike	Bus	Car	Walk	Bike	Walk	Walk	Bus
Car	Bus	Car	Walk	Walk	Bus	Car	Car	Bus	Bus

4. Classify responses of 40 families on how frequently they spend vacations out of city in a grouped frequency distribution. Also compute relative frequencies.

Never	Rarely	Often	Sometimes	Never	Rarely	Always	Often
Sometimes	Never	Sometimes	Always	Sometimes	Rarely	Never	Sometimes
Sometimes	Sometimes	Always	Always	Sometimes	Often	Often	Never
Often	Rarely	Sometimes	Always	Rarely	Always	Often	Often
Always	Never	Always	Sometimes	Never	Never	Often	Always

5. Organize the colors of Caps of prefects of various schools in frequency table.

Black	Blue	Black	Brown	Green	Blue	Black	Brown	Brown	Green
Blue	Green	Red	Black	Red	Blue	Red	Green	Black	Red

6. The marks secured by 25 students is a class test out of 5 were: 4.0, 0.5, 4.5, 1.0, 3.5, 3.5, 5.0, 2.5, 4.0, 2.5, 1.0, 2.0, 4.0, 3.5, 0.0, 3.0, 5.0, 1.0, 2.0, 2.0, 4.5, 2.5, 3.0, 3.5, 1.5. Obtain frequency table with discrete classes and relative frequencies.
7. The number of times the lectures of a teacher were downloaded per hour from a website was noted for a day before test, and results were: 1, 1, 2, 1, 3, 3, 4, 1, 4, 1, 2, 3, 3, 2, 4, 2, 3, 2, 3, 2, 1, 0, 0, 1. Classify data into a discrete frequency table.
8. Construct frequency distribution with continuous classes of the number of absentees in a class for 18 days: 4, 3, 0, 1, 2, 5, 6, 8, 10, 7, 11, 15, 13, 14, 3, 4, 12, 12.
9. Construct continuous frequency distribution of the weights recorded to nearest pound of 40 college students: 138, 148, 146, 146, 119, 164, 152, 173, 158, 154, 150, 144, 142, 140, 165, 132, 150, 147, 147, 153, 144, 156, 135, 126, 140, 125, 168, 157, 138, 135, 149, 136, 135, 142, 161, 145, 176, 163, 145, 128. Also compute class boundaries, class marks and cumulative frequencies.
10. The value of resistance in ohms of a batch of 48 resistors of similar value are: 21.0, 22.4, 22.8, 21.5, 22.6, 21.1, 21.6, 22.3, 22.9, 20.5, 21.8, 22.2, 21.0, 21.7, 22.5, 20.7, 23.2, 22.9, 21.7, 21.4, 22.1, 22.2, 22.3, 21.3, 22.1, 21.8, 22.0, 22.7, 21.7, 21.9, 21.1, 22.6, 21.4, 22.4, 22.3, 20.9, 22.8, 21.2, 22.7, 21.6, 22.2, 21.6, 21.3, 22.1, 21.5, 22.0, 23.4, 21.2. Form continuous frequency table. Also, compute cumulative and percentage frequencies.
11. The mass (in Kg) of 50 blocks of metal were measured collect to nearest 0.1kg and are listed below. Construct frequency distribution of the masses and compute relative and percentage frequencies.

8.0	8.3	7.7	8.1	7.4	8.6	7.1	8.4	7.4	8.2	8.4	8.8	7.9	8.1	8.2	7.5	8.3
8.8	8.0	7.7	8.3	8.2	7.9	8.5	7.9	8.0	8.4	7.2	8.7	8.0	9.1	8.5	7.6	8.2
7.8	7.8	8.7	8.5	8.4	8.5	8.1	7.8	8.2	7.7	7.5	8.5	8.1	7.3	9.0	8.6	

22.1(ii). Construct histograms with equal class intervals:

Histograms are used for graphical representation of quantitative data by using the frequency tables with discrete and continuous classes and related concepts. A **histogram** is drawn by using adjacent rectangles with bases of rectangles marked by distribution of numbers in the classes and areas/heights of rectangles being proportional to the class frequencies. In histograms with equal class intervals, both areas and heights of rectangles are proportional to frequency.

In histograms corresponding to discrete frequency tables, the bases of rectangles represent the distinct numbers in the data and the heights are usually taken as the frequencies of classes. The width of rectangles are taken with a fixed interval.

In histograms corresponding to continuous frequency tables with equal class intervals, the bases of rectangles represents class boundaries and the heights are usually taken as the frequencies of classes. The width of rectangles is equal to the class interval.

While drawing histograms, axes should be labelled clearly and scales should be defined. X-axis does not have to start from 0. Y-axis must start from 0. Frequencies should also be marked above the rectangles. The procedure to draw histograms is explained in following examples.

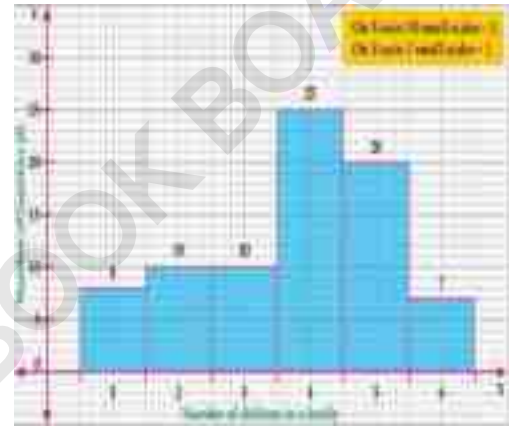
Example 1:

Draw histogram for the number of children per family in villages using the following data of 80 families.

Number of children in a family	1	2	3	4	5	6
Number of families	8	10	10	25	20	7

Solution:

Data represent a discrete frequency table. Writing distinct numbers: 1, 2, 3, 4, 5, 6 representing classes on X-axis with a fixed interval for all, then writing frequencies on Y-axis starting with "0" with suitable scale to include all frequencies up to 25. Drawing rectangles based at the distinct numbers with equal width and heights equal to frequencies, we get the following histogram. Axes are labeled and scales are defined.



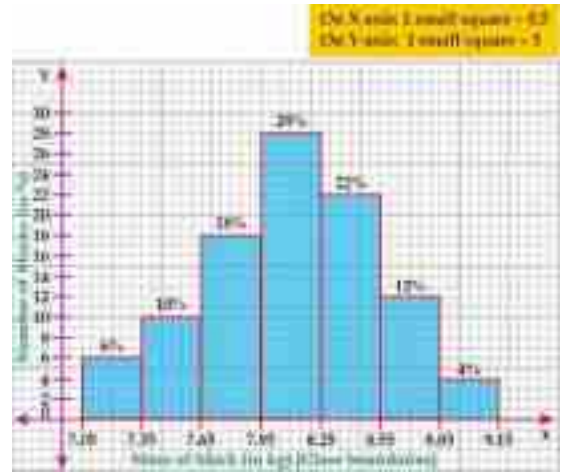
Example 2.

Draw percentage frequency histogram for the following distribution.

Mass of block (in kg)	7.1 - 7.3	7.4 - 76	7.7 - 7.9	8.0 - 8.2	8.3 - 8.5	8.6 - 8.8	8.9 - 9.1
Number of blocks	3	5	9	14	11	6	2

Solution:

Given is a continuous frequency table with continuous data up to one digit after the decimal, so $n = 50$, $m = 1$ and $d = 0.1$. The class boundaries are: 7.05 - 7.35, 7.35 - 7.65, 7.65 - 7.95, 7.95 - 8.25, 8.25 - 8.55, 8.55 - 8.85, 8.85 - 9.15. Writing these on X-axis and corresponding percentage frequencies: 6, 10, 18, 28, 22, 12, 4 on Y-axis with respective scales. Drawing adjacent rectangles based at class boundaries, of width $h = 0.3$ and heights equal to percentage frequencies, we get the required percentage frequency histogram.



22.1 (iii). Construct histograms with unequal class intervals:

For the case of **histograms with unequal class intervals**, the adjacent rectangles are constructed so that only their areas (not heights) are proportional to the frequencies. To do this, we define **adjusted heights** of rectangles to assure proportional areas and frequencies. If f^* and h^* are frequency and width of a class, respectively, then:

$$\text{Adjusted height of a rectangle}^* = \frac{f^*}{h^*} \qquad (1)$$

$$\text{So that, Area of a rectangle}^* = \text{width} \times \text{height} = h^* \times \frac{f^*}{h^*} = f^* \qquad (2)$$

From (1), we see that adjusted heights are not proportional to f^*, but areas in (2) are. This satisfies the requirement. The steps of drawing histograms with unequal class intervals are same as before, but we must write adjusted heights on Y-axis.

Example:

Construct histogram for the following data with unequal class intervals:

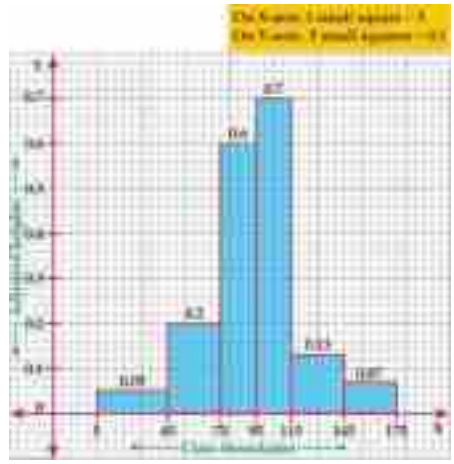
Class limits	10 - 40	50 - 70	80 - 90	100 - 110	120 - 140	150 - 170
Frequencies	2	6	12	14	4	2

Solution:

We can observe that the given data show a frequency table with unequal class intervals. We first compute class boundaries and adjusted heights. Here, $d = 10$.

Class limits	Class boundaries	Class intervals (h^*)	Frequencies (f^*)	Adjusted heights of rectangles (f^* / h^*)
10 – 40	5-45	40	2	$\frac{2}{40} = 0.05$
50 – 70	45-75	30	6	$\frac{6}{30} = 0.2$
80 – 90	75-95	20	12	0.6
100 – 110	95-115	20	14	0.7
120-140	115-145	30	4	≈ 0.13
150-170	145-175	30	2	≈ 0.07

Writing class boundaries on X-axis and adjusted heights on Y-axis, we draw rectangles for all classes with widths equal to unequal class intervals and heights equal to the adjusted as computed in above table. Thus, we get the required histogram with unequal intervals.



We can verify that areas of the rectangle are proportional, and particularly equal, to the frequencies here.

Area of rectangle-1 = 40× 0.05 = 2.

Area of rectangle-2 = 30× 0.2 = 6. Area of rectangle-3 = 20× 0.6 = 12.

and so on.

22.1 (iv). Construct a frequency polygon:

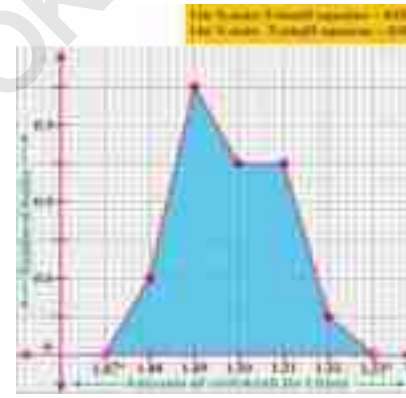
Frequency polygon is another way of graphical representation of a quantitative data grouped into frequency distribution. We often see such graphs in newspapers (especially in business section), in weather forecast news, and in cricket match stats (a commonly used abbreviation of statistics). The word **polygon** refers to a closed shape in a plane formed by connecting at least three line segments. For example, triangles, squares, pentagons, etc. Here, we will learn how this geometric concept is related with frequency distributions. The frequency polygons are used in advanced statistical analysis is to identify shape of the distribution.

To construct a frequency polygon, the points (x, y) are plotted, where the abscissas (the x-coordinates) are class marks and ordinates (the y-coordinates) are the class frequencies. Secondly, we add two dummy classes, one in the start and one in the end, with 0 frequencies, and mark corresponding points for these on X-axis. Finally, we join the scattered points on the graph by line segments to get the required closed shape, called **frequency polygon**, with base on X-axis and peaks showing frequencies. For clarity, filled marks (•) denote points of real classes and blank marks (○) denote points for dummy classes. It is obvious to label the axes and define suitable scales.

Example 1:

Draw a relative frequency polygon for the following data.

Amount of cold drink in liters	1.48	1.49	1.50	1.51	1.52
Numbers of bottles	2	7	5	5	1



Solution:

Given data show discrete frequency table. The distinct numbers: 1.48, 1.49, 1.50, 1.51, 1.52 are class marks (x). The frequencies (y) are: 2, 7, 5, 5, 1. First, we compute relative frequencies, which are: 0.1, 0.35, 0.25, 0.25 and 0.05. With two dummy classes: the plotted points are: (1.47*, 0), (1.48, 0.1), (1.49, 0.35), (1.50, 0.25), (1.51, 0.25), (1.52, 0.05) and (1.53*, 0). Joining points by line segments, we get the relative frequency polygon.

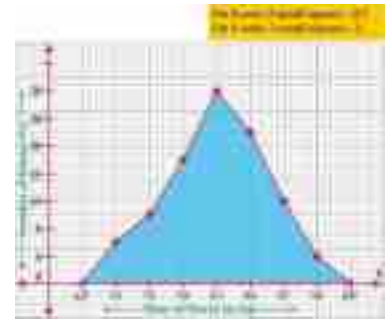
Example 2:

Draw percentage frequency polygon for the following distribution.

Mass of block (in kg)	7.1 - 7.3	7.4 - 7.6	7.7 - 7.9	8.0 - 8.2	8.3 – 8.5	8.6 – 8.8	8.9 – 9.1
Number of block	3	5	9	14	11	6	2

Solution:

Data show a continuous frequency distribution with equal class interval $h = 0.3$. The points to plot are (x, y), where x represent class marks: 7.2, 7.5, 7.8, 8.1, 8.4, 8.7, 9.0, and y are percentage frequencies: 6, 10, 18, 28, 22, 12, 4. Adding dummy class marks and the required percentage frequency polygon is plotted through (6.9*, 0), (7.2, 6), (7.5, 10), (7.8, 18), (8.1, 28), (8.4, 22), (8.7, 12), (9.0, 4) and (9.3*, 0) with suitable scales.



22.2. Cumulative frequency distribution:

A **cumulative frequency distribution** is a table listing upper class boundaries of classes together with the corresponding cumulative frequencies. We already know about the computation of cumulative frequencies. The cumulative frequencies computed by adding number of observations less than the UCBs of classes are called **"less than cumulative frequencies"**. These start with "0" and end at the total number of observations "n". We must add a dummy UCB in the start with cumulative frequency of 0.

"The greater than cumulative frequencies" are computed by summing all observations greater than the LCBs of classes. These start with "n" and end at "0". But we often use "less than cumulative frequencies" in practice. In the forthcoming discussion, we use the term **"cumulative frequency"** to refer the **"less than cumulative frequency"** for simplicity.

22.2 (i). Construct a cumulative frequency table:

Steps to construct cumulative frequency table of grouped quantitative data are:

1. Obtain upper class boundaries and cumulative frequencies.
2. Add a dummy upper class boundary in the start with 0 cumulative frequency.
3. Form a table to mention all UCBs with respective cumulative frequencies.

Example 1: Construct a cumulative frequency table using the following data.

Marks of Students	1	2	3	4	5
Number of Students	2	5	4	3	1

Solution: Given data shows a discrete frequency distribution.

Here, $d = 1$ and $n = 15$. Computing UCBs by adding $\frac{d}{2}$ in the class limits, and then finding cumulative frequencies we have:

Classes	UCBs	Frequencies	Cumulative Frequencies
1	1.5	2	2
2	2.5	5	7
3	3.5	4	11
4	4.5	3	14
5	5.5	1	15 = n

Adding a dummy UCB in start equal to 0.5* with cumulative frequency "0", the required cumulative frequency table is:

Marks of students (UCBs) Less than	0.5*	1.5	2.5	3.5	4.5	4.5
Number of students (Cumulative frequencies)	0	2	7	11	14	15

Example 2: Construct cumulative frequency table for electricity consumption data.

Electricity Consumption (kWh)	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution: Given data is a continuous frequency distribution, Here, $d = 1$ and $n = 60$. The UCBs are: 87.5. 107.5, 127.5, 147.5, 167.5, 187.5 and 207.5. Adding a dummy UCB of **67.5*** in start, we have the following cumulative frequency table.

Electricity Consumption Less than (kWh)	67.5*	87.5	107.5	127.5	147.5	167.5	187.5	207.5
Number of days	0	10	23	38	48	52	58	60

22.2 (ii). Draw a cumulative frequency polygon:

A cumulative frequency distribution is graphically represented by a **cumulative frequency polygon**, also called an **ogive**. We begin plotting the points (x, y) of the cumulative frequency table (x: UCBs and y: cumulative frequencies), then join these by line segments. Finally, we draw a perpendicular from the peak point on X-axis to get a closed shape, which is the required ogive or cumulative frequency polygon.

Example 1.

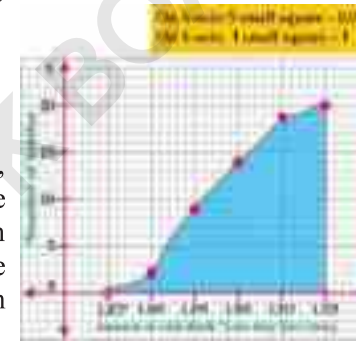
Draw an ogive (cumulative frequency polygon) for the following data.

Amount of cold drink (in liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution:

Here, $n = 20$ and $d = 0.01$. By computing UCBs and cumulative frequencies, we get the cumulative frequency table, where 1.475* is a dummy starting UCB.

Amount of cold drink (in liters)	1.475*	1.485	1.495	1.505	1.515	1.525
Number of bottles	0	2	9	14	19	20



Now, we plot the points (1.475*, 0), (1.485, 2), (1.495, 9), (1.505, 14), (1.515, 19) and (1.525, 20) on graph, join these through line segments, and, finally draw a perpendicular from the last point towards X-axis to get the required cumulative frequency polygon or ogive. Axes are labeled clearly with proper scaling.

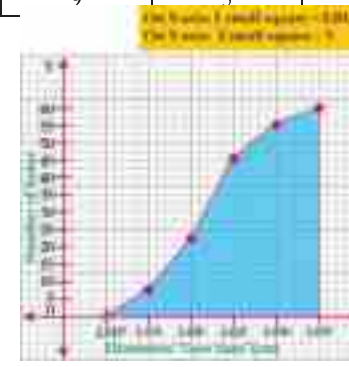
Example 2:

Draw ogive for the following distribution of diameters (in cm) of 60 holes bored in engine castings as measured in a certain study.

Diameters (cm)	2.011– 2.014	2.016 – 2.019	2.021 - 2.024	2.026 – 2.029	2.031 – 2.034
Number of holes	7	16	23	9	5

Solution:

Given is a continuous frequency table with $n = 60$ and $d = 0.002$. Using UCBs and cumulative frequencies, we get the cumulative frequency table with points: (2.010*, 0), (2.010, 7), (2.020, 23), (2.025, 46), (2.030, 55) and (2.035, 60). These points lead to the required ogive or the cumulative frequency polygon.



EXERCISE 22.2

1. Obtain percentage histogram, frequency polygon, cumulative frequency table and ogive for the following data.

Marks of Students	1	2	3	4	5
Number of Students	2	5	4	3	1

2. Construct histogram for the following data with unequal class intervals:

Class limits	0-39	40-49	50-79	80-99
Frequencies	6	8	12	4

3. Construct histogram, frequency polygon, cumulative frequency table and ogive.

Amount of money earned weekly	20-40	50-70	80-90	100-110	120-140	150-170
Number of people	2	6	12	14	4	2

4. Plot histogram, relative frequency polygon, and ogive for the following data.

Class limits	10.5-10.9	11.0-11.4	11.5-11.9	12.0-12.9	13.0-13.4
Frequencies	2	7	10	12	8

22.3. Measures of central tendency:

The ability of all observations in data to cluster around a **central point** is referred as the **central tendency**. A central point of the data is called a **measure of central tendency** or simply an **average**. A measure of central tendency represents the whole data by a single central point, and is a concise identification of whole data.

We usually talk about averages of numbers and categories frequently. For example, consider the following statements:

1. "The majority of students in a class secured Grade-B". Here, Grade B is an average.
2. "Ali spends 45 minutes in the GYM daily". Here, average time per day Ali spends in GYM is 45 minutes. Obviously, he may have spent more or less than 45 minutes.
3. "Average score of a batsman in T20", which is based on the scores of the batsman in all T20 matches he played. The average score is generalizes all individual scores.

22.3 (i). To calculate measures of central tendency:

We usually compute an average or measure central tendency of data by adding the observation and then dividing by the total numbers of observations, which is particularly the **arithmetic mean**. There are many more ways to compute averages. We discuss the following five types here:

(1) Arithmetic mean (2) Median (3) Mode (4) Geometric mean (5) Harmonic mean

These averages have some advantages and disadvantages when compared with each other. Also, the best use of an average depends on nature of data.

22.3 (i) (a) arithmetic mean by definition and using deviations from assumed mean

The **arithmetic mean** is based on the principle of equality among all observations. We mix all n observations by their sum, and then divide the sum in n equal parts.

If $x_1, x_2, x_3, ..., x_n$ are n numbers, then their arithmetic mean (denoted as $\bar{x}$) is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad \text{or} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1)$$

where "Σ" is an upper-cased Greek letter SIGMA used for summation.

Note that:

- Arithmetic mean is a unique number within the range of the data.
- If data are quantitative, then we directly find arithmetic mean. For nominal or ordinal, we may assign numeric codes or ranks, respectively.
- Arithmetic mean is not appropriate for nominal data.

Arithmetic mean by definition/ direct method:

The **direct method** or **by definition** refers to computing arithmetic mean directly from the given data without changing the location or scale of the observations.

In direct method/ by definition, we use following formulas:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n} \quad \text{(Ungrouped data)} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum f_i} = \frac{\sum xf}{\sum f} \quad \text{(Grouped data)}$$

where x_i are individual observations in ungrouped data and class marks in grouped data, and f_i are corresponding class frequencies. For brevity, we can omit subscripts in the formulas.

Example 1: Find arithmetic mean of the following datasets by definition.

(i). 2, 3, 7, 5, 5, 13, 1, 7, 4, 8, 3, 4, 3 (ii). 7, 5, 74, 10 (iii). 18.92, 27.9, 34.7, 39.68

Solution:

(i). By definition or direct method, we have:

$$\bar{x} = \frac{\sum x}{n} = \frac{2+3+7+5+5+13+1+7+4+8+3+4+3}{13} = \frac{65}{13} = 5.$$

(ii). By definition, we have: $\bar{x} = \frac{\sum x}{n} = \frac{7+5+74+10}{4} = \frac{96}{4} = 24.$

(iii). $\bar{x} = \frac{\sum x}{n} = \frac{18.92+27.9+34.7+39.68}{4} = \frac{121.2}{4} = 30.3.$

Note that:

- The arithmetic mean is affected by **extreme values** (or **outliers**). In Example 1 (ii), $\bar{x} = 24$ is far away from all observations: 7, 5, 74 and 10 due to the outlier 74.

Example 2: The grades of five students who studied in group were: A+, B+, C, B, A. Identify the average grade using arithmetic mean.

Solution: The data are ordinal and can be ranked by numbers. We represent the grades: C, C+, B, B+, A, A+ by labels: 1, 2, 3, 4, 5, 6 respectively, then:

$\bar{x} = \frac{\sum x}{n} = \frac{6+1+4+3+5}{5} = \frac{19}{5} = 3.8 \approx 4.$ So, the average grade is B+.

Example 3:

Find arithmetic mean amount of cold drink by definition in the following data:

Amount of cold drink(in liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution:

Data is grouped into discrete classes. The class marks (x) are: 1.48, 1.49, 1.50, 1.51, 1.52. The computations are shown in the table.

x	f	xf
1.48	2	$1.48\times 2 = 2.96$
1.49	7	10.43
1.50	5	7.5
1.51	5	7.55
1.52	1	1.52
	$\sum f = 20$	$\sum xf = 29.96$

Finally, $\bar{x} = \dfrac{\sum xf}{\sum f} = \dfrac{29.96}{20} = 1.498$ liters.

The average amount of cold drink using arithmetic mean is 1.498 liters.

Example 4:

Find arithmetic mean of satisfaction of customers using definition.

Satisfaction level	Not at all	Not very	Not sure	Somewhat	Very
Number of customers	3	8	5	6	18

Solution:

We represent attributes of the ordinal variable: "Not at all, Not very, Not sure, Somewhat and Very" by ranks, say, 1, 2, 3, 4. 5, respectively. By definition, we have:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{(1)(3)+(2)(8)+(3)(5)+(4)(6)+(5)(18)}{3+8+5+6+18} = \frac{148}{40} = 3.7 \approx 4.$$

So, average satisfaction level is 4^{th} rank, which is "Somewhat satisfied".

Example 5:

Find arithmetic mean of electricity consumption (in kWh) of a shop for 60 days:

Electricity consumption	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution:

Data show continuous frequency table. We find class marks and required sums in adjacent table.

By definition or direct method:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{7270}{60} = 121.166.$$

Arithmetic mean of electricity consumption of shop is 121.166 kWh.

Class limits	Class Marks (x)	Frequencies (f)	xf
68-87	77.5	10	775
88-107	97.5	13	1267.5
108-127	117.5	15	1762.5
128-147	137.5	10	1375
148-167	157.5	4	630
168-187	177.5	6	1065
188-207	197.5	2	395
Total	---	$\sum f = 60$	7270

Note that:

Arithmetic mean of an ungrouped data and its discrete frequency table are always same. But, the arithmetic mean of a data from continuous frequency distribution is not equal to (but closer to) the mean of corresponding ungrouped data.

Arithmetic mean by using deviations/ indirect methods:

If x_i are observations and A is **assumed/provisional mean** (any number within or outside the range of data), then deviations about A are: $D_i = x_i - A$.

There are two indirect methods based on deviations: **shortcut** and **coding** methods, to compute arithmetic mean when dealing with larger data with higher values.

Shortcut method:

The arithmetic mean using shortcut method is computed as:

$\bar{x} = A + \frac{\sum D}{n}$ (for ungrouped data) and $\bar{x} = A + \frac{\sum Df}{\sum f}$ (for grouped data)

For ungrouped data, x_i are observations. For grouped data, x_i are class marks. It is better to take x_i corresponding to the highest frequency as "A" for grouped data.

Coding method:

The formulae for computing arithmetic mean using coding method are:

$\bar{x} = A + h\frac{\sum u}{\sum n}$ (for ungrouped data) and $\bar{x} = A + h\frac{\sum uf}{\sum f}$ (for grouped data)

where, $u_i = \frac{x_i - A}{h}$ or, simply $u = \frac{x - A}{h}$ is referred as the coded variable. For ungrouped data, x_i are observations, and for grouped data x_i are class marks.

Example 1:

The total marks of six students who studied in group for the examination are: 610, 640, 685, 680, 710, 580. Find arithmetic mean marks using indirect methods by taking (i). $A = 650$ and $h = 10$, (ii). $A = 680$ and $h = 20$.

Solution:

The shortcut method states: $\bar{x} = A + \frac{\sum D}{n}$

The coding method states: $\bar{x} = A + h\frac{\sum u}{n}$, where $u = \frac{x - A}{h}$.

The required calculations are shown in tables separately for both parts.

(i). $A = 650$ and $h = 10$.

x	$D = x - 650$	$u = \frac{x - 650}{10}$
610	-40	-4
640	-10	-1
685	35	3.5
680	30	3
710	60	6
580	-70	-7
	$\sum D = 5$	$\sum u = 0.5$

By shortcut method, we have:

$\bar{x} = 650 + \frac{5}{6} = 650.833.$

By coding method, we have

$\bar{x} = 650 + 10 \times \frac{0.5}{6} = 650 + \frac{5}{6} = 650.833.$

(ii). $A = 680$ and $h = 20$.

x	$D = x - 680$	$u = \frac{x-680}{20}$
610	-70	-3.5
640	-40	2
685	5	0.25
680	0	0
710	30	1.5
580	-100	-5
	$\sum D = -175$	$\sum u = -8.75$

By shortcut method:

$$\bar{x} = 680 + \left(\frac{-175}{6}\right) = 680 - 29.1666$$

$\bar{x} = 650.833.$

By coding method:

$$\bar{x} = 680 + 20\left(\frac{-8.75}{6}\right)$$

$\bar{x} = 680 - 29.1666 = 650.833.$

The mean marks are 650.833 or 651 regardless of the choice of A and h.

Example 2: Find mean amount of cold drink (in liters) by shortcut and coding methods.

Amount of cold drink(in liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution: Here, we use $A = 1.49$ and $h = 0.01$. The calculations are detailed in the table.

x	f	$D = x - 1.49$	$u = \frac{x-1.49}{0.01}$	Df	uf
1.48	2	-0.01	-1	-0.02	-2
1.49	7	0	0	0	0
1.50	5	0.01	1	0.05	5
1.51	5	0.02	2	0.1	10
1.52	1	0.03	3	0.03	3
	$\sum f = 20$			$\sum Df = 0.16$	$\sum uf = 16$

By shortcut method: $\bar{x} = A + \frac{\sum Df}{\sum f} = 1.49 + \frac{0.16}{20} = 1.498.$

By coding method: $\bar{x} = A + h \times \frac{\sum uf}{\sum f} = 1.49 + 0.01\left(\frac{16}{20}\right) = 1.498.$

We note that the result by both indirect methods and by definition match.

Example 3: Use shortcut and coding methods to find mean mass of 50 metal blocks.

Mass of block(in Kg)	7.1-7.3	7.4-7.6	7.7-7.9	8.0-8.2	8.3-8.5	8.6-8.8	8.9-9.1
Number of blocks	3	5	9	14	11	6	2

Solution: Taking $A = 8.1$ and $h = 0.1$, we proceed with the calculations in the table.

Class limits	Class marks (x)	f	$D = x - 8.1$	$u = \frac{x-8.1}{0.1}$	Df	uf
7.1-7.3	7.2	3	-0.9	-9	-2.7	-27
7.4-7.6	7.5	5	-0.6	-6	-3	-30
7.7-7.9	7.8	9	-0.3	-3	-2.7	-27
8.0-8.2	8.1	14	0	0	0	0
8.3-8.5	8.4	11	0.3	3	3.3	33
8.6-8.8	8.7	6	0.6	6	3.6	36
8.9-9.1	9.0	2	0.9	9	1.8	18
		$\sum f = 50$			$\sum Df = 0.3$	$\sum uf = 3$

By shortcut method: $\bar{x} = A + \frac{\sum Df}{\sum f} = 8.1 + \frac{0.3}{50} = 8.106$.

By coding method: $\bar{x} = A + h \times \frac{\sum uf}{\sum f} = 8.1 + 0.1 \times \left(\frac{3}{50}\right) = 8.106$.

So, the average/mean mass is 8.106Kg.

22.3(i) (b). Median, mode, geometric mean and harmonic mean

Besides, the arithmetic mean (based on equality), there are some other ways to compute measures of central tendency of a data. For example, in our electoral system, everyone has right to vote, but the elected representative is finally chosen on the basis of **majority**, as widely said "majority is the authority". Such a case represents **mode**, the most frequent observation. Similarly, the **median** focuses on the **middle most part** of a ranked data set. The **geometric and harmonic means** are mathematical in nature like arithmetic mean in contrast to median and mode which are descriptive in nature.

Median:

The **median** in an ordered data is the value dividing it into two equal parts. By computing median, we assume that the central point of the whole data arranged into ascending or descending order lies at middle position.

Note:

- In a data, median is always unique.
- We can compute median of quantitative data.
- For qualitative data, it is meaningful to find median of only the ordinal data.
- The median does not rely equally on all observations
- The median is not affected by extreme values (outliers) in the data.

For an ordered ungrouped data with "n" observations, median "m" is the middle-most part of data, and can be defined through the following two cases:

$$m = \left(\frac{n+1}{2}\right)^{th} \text{ observation, when n is odd} \qquad (1)$$

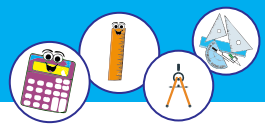
$$\text{Otherwise, } m = \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{ observations}\right], \text{when n is even} \qquad (2)$$

For grouped data with discrete classes, by the help of cumulative frequencies, we choose the median observation using (1) and (2).

For grouped data with continuous classes, the median lies in the class containing $\left(\frac{n}{2}\right)^{th}$ observation, which is also called the median class (m-class). The formula is:

$$m = L + \frac{h}{f}\left(\frac{n}{2} - c\right) \qquad (3)$$

Where, L is the LCB of m-class, h is width of m-class, f is frequency of m-class and c is cumulative frequency of the class just before the m-class.



Example 1: Find median in the following:

(i). 2, 3, 7, 5, 5, 13, 1, 7, 4, 8, 3, 4, 3, (ii). 7, 5, 74, 10 (iii). 18.92, 27.9, 37.4, 39.68

Solution:

(i). Writing data is ascending order: 1, 2, 3, 3, 3, 4, 4, 5, 5, 7, 7, 8, 13.

Here, $n = 13$ (odd), So: $m = \left(\frac{13+1}{2}\right)^{th}$ observation $= 7^{th}$ observation $= 4$.

(ii). In ascending order, we have: 5, 7, 10, 74. Here, $n = 4$ (even), so:

$$m = \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{observations}\right] = \frac{1}{2}\left[2^{nd} + 3^{rd} \text{ observations}\right] = \frac{1}{2}[7+10] = 8.5.$$

Here, n =4 (even). Data is already in ranked from.

$$m = \frac{1}{2}\left[2^{nd} + 3^{rd} \text{ observations}\right] = \frac{1}{2}[27.9 + 37.4] = 32.65.$$

Example 2: Identify the median grade of the following two groups of students:

(i). A+, B+, C, B, A (ii). C+, A, B, B, B+, C, B+, A, A+, C

Solution:

(i). The ranked data are: C, B, B+, A, A+. Also, $n = 5$ (odd), so:

$$m = \left(\frac{n+1}{2}\right)^{th} = \left(\frac{5+1}{2}\right)^{th} = 3^{rd} \text{ observation} = B+, \text{ which is the median grade.}$$

(ii). Arranging in ascending order, we have: C, C, C+, B, B, B+, B+, A, A, A+. Here, $n = 10$, the two middle observations are $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$, i.e. 5^{th} and 6^{th} observations. Which are: B and B+ grades (different). So, we conclude that the median grade is equally split between B and B+ grades.

Example 3:

What is the median satisfaction level in the following data of 40 customers?

Satisfaction level	Not at all	Not very	Not sure	Somewhat	Very
Number of customers	3	8	5	6	18

Solution:

Here, $n = 40$ (even). So, two middle observations are: 20^{th} and 21^{th}. These two lie in group "Somewhat". So, the median satisfaction level is "Somewhat satisfied".

Example 4: Find median amount of cold drink in the following data.

Amount of cold drink(in liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution:

Here, $n = \sum f = 20$ (even).

So, $m = \frac{1}{2}\left[10^{th} + 11^{th} \text{ observations}\right]$.

Finding cumulative frequencies, we see that 10^{th} and 11^{th} observations lie in group with cumulative frequency 14.

x	f	Cumulative frequencies
1.48	2	2
1.49	7	9
1.50	5	14
1.51	5	19
1.52	1	20
	$\sum f = 20$	

$m = \frac{1}{2}[1.50 + 1.50] = \frac{1}{2}(3) = 1.50$ liters.

Example 5:

Find the median electricity consumption of a shop for 60 days using the data:

Electricity consumption	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution:

Here, $n = 60$, $h = 20$. Data are a continuous frequency table, so we need to compute class boundaries and cumulative frequencies as done in the table below.

Class limits	Class boundaries	f	Cumulative frequencies
68-87	67.5-87.5	10	10
88-107	87.5-107.5	13	23
108-127	107.5-127.5	15	38 (m-class)
128-147	127.5-147.5	10	48
148-167	147.5-167.5	4	52
168-187	167.5-187.5	6	58
188-207	187.5-207.5	2	60
		$n = 60$	

Median class (m-class) is one containing $\left(\frac{n}{2}\right)^{th}$ observation, i.e. 30th observation. So, m-class is one with cumulative frequency of 38. The m-class is 108-127 (as highlighted in table). Using the formula:

$$m = \mathrm{L} + \frac{h}{f}\left(\frac{n}{2} - c\right) = 107.5 + \frac{20}{15}\left(\frac{60}{2} - 23\right)$$

$$m = 107.5 + \frac{20}{15}(30 - 23) = 107.5 + \frac{20}{15} \times 7 = 107.5 + 9.333 = 116.833 \text{ kWh.}$$

Here, L = LCB of m-class = 107.5, h = class interval m-class = 20, f = frequency of m-class = 15 and c = Cumulative frequency of class just before the median class = 23.

Quartiles:

An ordered data may be divided into four equal parts to define **quartiles**. There are three quartiles within the range of data which divide data into 4 equal parts.

The **first (or lower) quartile** Q_1, divides ordered data into 25% to 75% ratio.

For ungrouped data: $Q_1 = \left(\frac{n+3}{4}\right)^{th}$ observation, if $\frac{n}{4}$ is not integer (1)

Otherwise, $Q_1 = \frac{1}{2}\left[\left(\frac{n}{4}\right)^{th} + \left(\frac{n}{4}+1\right)^{th} \text{ observations}\right]$, if $\frac{n}{4}$ is an integer (2)

For grouped data with discrete classes, we also use (1) and (2). For grouped data with continuous classes, we use: $Q_1 = L + \frac{h}{f}\left(\frac{n}{4} - c\right)$ (3)

where, L is LCB, h is width, f is frequency of Q_1-class, and c is cumulative frequency of a class just before Q_1-class. The Q_1-class is one containing the $\left(\frac{n}{4}\right)^{th}$ observation.

The **second (or middle) quartile** Q_2, divides ordered data into 50% – 50% ratio. It is same as the median of data.

The **third (or upper) quartile** Q_3 divides ordered data into 75% – 25% ratio.

For ungrouped data: $Q_3 = \left(\frac{3n+1}{4}\right)^{th}$ observation, if $\frac{3n}{4}$ is not integer (4)

Otherwise, $Q_3 = \frac{1}{2}\left[\left(\frac{3n}{4}\right)^{th} + \left(\frac{3n}{4}+1\right)^{th} \text{ observations}\right]$, if $\frac{3n}{4}$ is an integer (5)

For grouped data with discrete classes (4)-(5) are used with cumulative frequencies.

For grouped data with continuous classes: $Q_3 = L + \frac{h}{f}\left(\frac{3n}{4} - c\right)$ (6)

where, L is LCB, h is width, f is frequency of Q_3–class, which contains $\left(\frac{3n}{4}\right)^{th}$ observation, and c is the cumulative frequency of the class just before Q_3–class.

Example 1:

Find lower and upper quartiles in the following datasets.

(i). 1000, 1200, 1600, 1500, 1200 (ii). 32, 36, 36, 37, 39, 41, 45, 46

Solution

(i). The ordered data set is: 1000, 1200, 1200, 1500, 1600. Here, n = 5.

For $Q_1 : \frac{n}{4} = \frac{5}{4} = 1.25$ is not integer,

so: $Q_1 = \left(\frac{n+3}{4}\right)^{th} = 2^{nd}$ observation $= 1200$.

For $Q_3 : \frac{3n}{4} = \frac{3\times 5}{4} = 3.75$ is not integer,

so: $Q_3 = \left(\frac{3n+1}{4}\right)^{th} = 4^{th}$ observation1500.

(ii): The data are already given in ordered form and $n = 8$.

For $Q_1 : \frac{n}{4} = \frac{8}{4} = 2$ is integer,

$$\text{so: } Q_1 = \frac{1}{2}\left[\left(\frac{n}{4}\right)^{th} + \left(\frac{n}{4}+1\right)^{th} \text{ observations}\right]$$

$$\Rightarrow Q_1 = \frac{1}{2}\left[2^{nd} + 3^{rd} \text{ observation}\right] = \frac{1}{2}[36+36] = 36.$$

For $Q_3 : \frac{3n}{4} = \frac{3\times 8}{4} = 6$ (integer),

$$\text{so: } Q_3 = \frac{1}{2}\left[6^{th} + 7^{th} \text{ observations}\right] = \frac{1}{2}[41+45] = 43.$$

Example 2:

Find lower and upper quartiles of the number of children in 80 families:

Number of children	1	2	3	4	5	6
Number of families	8	10	10	25	20	7

Solution:

Data are grouped with discrete classes. Here, $n = 80$.
Finding cumulative frequencies as in the following table.

x	f	Cumulative frequency
1	8	8
2	10	18
3	10	28 (Q_1-class)
4	25	53
5	20	73 (Q_3-class)
6	7	80

For $Q_1 : \frac{n}{4} = \frac{80}{4} = 20$ (an integer).

So, Q_1 is average of 20^{th} and 21^{st} observations.

These both lie in class with cumulative frequency 28, $Q_1 = \frac{1}{2}(3+3) = 3$.

For $Q_3 : \frac{3n}{4} = \frac{3\times 80}{4} = 60$.

60^{th} and 61^{st} observations lie in the class with cumulative frequency 73.

So, $Q_3 = \frac{1}{2}(5+5) = 5$.

Example 3:

Find quartile weights (in Kg) of 100 students reported below:

Weight (Kg)	70-74	75-79	80-84	85-89	90-94
Number of students	10	24	46	12	8

Solution:

Here, n =100, h = 5, and data are grouped with continuous classes. First we compute class boundaries and cumulative frequencies. The calculations are summarized in the following table.

Class limits	Class boundaries	f	Cumulative Frequencies
70-74	69.5-74.5	10	10
75-79	74.5-79.5	24	34 (Q_1-class)
80-84	79.5-84.5	46	80 (m *and* Q_3-class)
85-89	84.5-89.5	12	92
90-94	89.5-94.5	8	100

For $Q_1 : \frac{n}{4} = \frac{100}{4} = 25.$ 25^{th} observation lies in class with cumulative frequency 34,

So, Q_1 – class is 75 – 79.

In this class: $Q_1 = L + \frac{h}{f}\left(\frac{n}{4} - c\right) = 74.5 + \frac{5}{24}(25 - 10) = 77.625$ Kg.

For $Q_2 = m : \frac{n}{2} = \frac{100}{2} = 50.$

50^{th} observation lies in the class with cumulative frequency 80.

So, m-class is 80-84.

$m = L + \frac{h}{f}\left(\frac{n}{2} - c\right) = 79.5 + \frac{5}{46}(50 - 34) = 81.239$ Kg.

For $Q_3 : \frac{3n}{4} = \frac{3 \times 100}{4} = 75.$

75^{th} observation lies in the class with cumulative frequency of 80.

So, Q_3-class is same as m-class, which is: 80-84.

In this class:

$Q_3 = L + \frac{h}{f}\left(\frac{3n}{4} - c\right) = 79.5 + \frac{5}{46}(75 - 34) = 83.956$ Kg.

Note that:

The quartiles are positional in nature, their positions do not change as long as number of observations remain same. For example, for an ordered data with $n = 10$:

$Q_1 = 3^{rd}$ observation, $Q_2 = m = \frac{1}{2}\left[5^{th} + 6^{th} \text{ observations}\right], Q_3 = 8^{th}$ observation

Any ordered dataset with 10 observations will follow these positions, as in:

a. 1, 3, 5, 7, 9, 11, 13, 15, 17, 19 $Q_1 = 5,\ m = 10,\ Q_3 = 15$

b. 10, 12, 12, 13, 17, 20, 20, 20, 21, 25 $Q_1 = 12,\ m = 13.5,\ Q_3 = 20.$

c. 1, 1, 1, 3, 12, 12, 13, 17, 19, 30 $Q_1 = 1,\ m = 12,\ Q_3 = 17.$

Mode:

The most frequent observation in a data is referred as **mode**. Mode relies on the principle "majority is the authority". Mode can be computed for both quantitative and qualitative data. Data with no mode is non-modal, one mode is uni-modal, two or more modes is multi-modal.

For ungrouped data, observation(s) which occur most frequently than others, if any, are referred as mode(s).

For grouped data with discrete classes, mode is the observation (class) with highest frequency. If two or more than two classes have tie for highest frequency, then data is multi-modal, and all observations with same highest frequency are modes.

For grouped data with continuous classes, a class with highest frequency is called modal class, and a mode lies in the modal class. The formula to compute mode (M) in modal-class (M-class) is: Mode = $M = L + \dfrac{(f_m - f_{m-1}) \times h}{2f_m - f_{m-1} - f_{m+1}}$

Where, L is LCB, h is width and f_m is frequency of M-class. f_{m-1} and f_{m+1} are frequencies of the classes just before and after M-class, respectively.

Example 1: Find mode in the following datasets:

(i). 75, 76, 80, 80, 82, 82, 82, 85 (ii). 13, 14, 15, 11, 16, 10, 19, 20, 18, 17

(iii). 1.49, 1.50, 1.51, 1.50, 1.48, 1.51 (iv). 1, 2, 2, 2, 5, 5, 5, 8, 9, 9, 9, 6, 1, 10

(v). C+, A, B, B, B+, C, B, A, A+, C (grades) (vi). Good,Poor,Dull,Good,Fair,Fair (ratings)

Solution:

(i). Mode = 82 as 82 occurs 3 times, which is more than any other observation.

(ii). Data has no mode, it is non-modal data.

(iii). There are two modes: 1.50 and 1.51 as these equally occur more than others.

(iv). Data is tri-modal. There are three modes: 2, 5 and 9.

(v). Grade "B" occurs most frequently, so the modal grade is "B".

(vi). Data has two modes: Fair and Good as these are equally most frequent ratings.

Example 2. Find modal number of mobiles possessed by a family from data of 45 families.

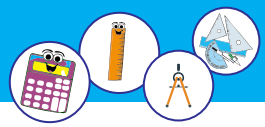
Numbers of mobiles	0	1	2	3	4	5	6	7
Numbers of families	1	7	15	12	5	2	2	1

Solution: Data are grouped into discrete classes. Here, "15" is the highest frequency, and it corresponds to the class "2". Hence, the modal number of mobiles per family is 2.

Example 3: From the blood groups of 30 students given below, find the modal blood group.

Blood group	A+	B+	AB+	O+	A-	B-	AB-	O-
Number of student	6	5	3	7	3	1	2	3

Solution: The modal blood group is O+ as it is associated with the highest frequency "7".



Example 4: Compute modal satisfaction level of 40 customers from the following data.

Satisfaction level	Not at all	Not very	Not sure	Some what	Very
Number of customers	3	8	5	6	18

Solution: The modal satisfaction level is "Very satisfied" as it occurs most frequently.

Example 5: Find modal kWh electricity consumption of a shop for 60 days using the data:

Electricity consumption	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution: Data are grouped in continuous classes. The highest frequency is 15. The modal class is 108-127. The class boundaries are computed in table.

Class limits	Class boundaries	Frequency
68 - 87	67.5 - 87.5	10
88 - 107	87.5 - 107.5	13
108 - 127	107.5 - 127.5	15
128 - 147	127.5 - 147.5	10
148 – 167	147.5 - 167.5	4
168 - 187	167.5 - 187.5	6
188 - 207	188.5 - 207.5	2

$h = 20, L = 107.5, f_m = 15, f_{m-1} = 13, f_{m+1} = 10$

$$\text{Mode} = M = L + \frac{(f_m - f_{m-1}) \times h}{2f_m - f_{m-1} - f_{m+1}}$$

$$M = 107.5 + \frac{(15-13)\times 20}{2(15)-13-10} = 107.5 + \frac{2\times 20}{30-23}$$

$$M = 107.5 + 5.71428 = 113.21428 \text{kWh}.$$

Note that:

The continuous frequency distribution in Example 5 was uni-modal as there was only one class with the highest frequency.

The following continuous frequency distribution representing diameters of a reel of wire using measured diameters at 24 places is bi-modal.

Diameters (mm)	2.10-2.13	2.14-2.17	2.18-2.21	2.22-2.25	2.26-2.29	2.30-2.33	2.34-2.37
Number of places	4	5	4	4	5	1	1

There are two modal classes: "2.14 – 2.17" and "2.26 – 2.29" corresponding to the highest frequency "5". Using formula in each class, we can get the respective modal diameters of 2.175mm and 2.263mm.

Geometric mean:

In **geometric mean**, all numbers are given **equal** importance. We mix the numbers by their product and then remove effect of mixing by finding n^{th} root of the product. It is meaningful to compute geometric mean (G.M.) of only quantitative data which are positive. If $x_1, x_2, x_3, ..., x_n$ are n numbers (all positive), then:

$$\text{G.M.} = (x_1 . x_2 ... x_n)^{\frac{1}{n}} = \left[\prod(x)\right]^{\frac{1}{n}} \qquad (1)$$

Here, Π is upper-cased Greek letter pi for product. G.M., if exists, is unique.

For ungrouped data, formula (1) is referred as the **basic formula** to find G.M. But, when observations are bigger in magnitude, we use **logarithmic formula** obtained by taking logarithm on both sides of (1) and using properties, which is:

$$\text{G.M} = \text{antilog}\left[\frac{\sum_{i=1}^{n}\log x_i}{n}\right] = \text{antilog}\left[\frac{\sum \log x}{n}\right] \qquad (2)$$

In (2), G.M. is the antilogarithm of the arithmetic mean of logarithms of data.
For grouped data, the **basic formula** is:

$$\text{G.M.} = \left[(x_1)^{f_1}\times(x_2)^{f_2}\times\ldots\times(x_n)^{f_n}\right]^{\frac{1}{\sum f}} = \left[\prod(x)^{f}\right]^{\frac{1}{\sum f}} \qquad (3)$$

and, **logarithmic formula** is: $$\text{G.M.} = \text{antilog}\left[\frac{\sum \text{f. log}\, x}{\sum \text{f}}\right] \qquad (4)$$

In (3)-(4), $f_1, f_2, \ldots, f_n$ are class frequencies and $x_1, x_2, \ldots, x_n$ are class marks, all positive. None of the frequencies should be zero in (4).

Example 1: Find G.M. using basic and logarithmic formula. Which one is better and why?
(i). 2, 4, 2, 16 (ii). 1.48, 1.52, 1.47, 1.50 (iii). 1000, 1200, 1600, 1500, 1200

Solution:

(i). Using basic formula: $\text{G.M.} = (2\times 4\times 2\times 16)^{\frac{1}{4}} = (256)^{\frac{1}{4}} = 4.$

Using logarithmic formula: $\text{G.M.} = \text{antilog}\left[\frac{\log(2)+\log(4)+\log(2)+\log(16)}{4}\right]$

$$\text{G.M.} = \text{antilog}\left[\frac{0.3010+0.6021+0.3010+1.2041}{4}\right] = \text{antilog}\,(0.60205) = 3.9999 \approx 4.$$

Observations were small in magnitude, so basic formula is better.

(ii). Using basic formula: $\text{G.M.} = (1.48\times 1.52\times 1.47\times 1.50)^{\frac{1}{4}} = (4-9604)^{\frac{1}{4}} = 1.4924$

By logarithmic formula: $\text{G.M.} = \text{antilog}\left[\frac{\log(1.48)+\log(1.52)+\log(1.47)+\log(1.5)}{4}\right]$

$$\text{G.M.} = \text{antilog}\left[\frac{0.1703+0.1818+0.1673+0.1761}{4}\right] = \text{antilog}\,(0.173875) = 1.4924.$$

Again, basic formula is suitable as observation were smaller in magnitude.

(iii). Using basic formula:

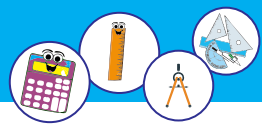
$$\text{G.M.} = (1000\times 1200\times 1600\times 1500\times 1200)^{\frac{1}{5}} = (3456000000000000)^{\frac{1}{5}} = 1281.48 \approx 1281.5$$

Using logarithmic formula:

$$\text{G.M.} = \text{antilog}\left[\frac{\log(1000)+\log(1200)+\log(1600)+\log(1500)+\log(1200)}{5}\right]$$

$$\text{G.M.} = \text{antilog}\left[\frac{3+3.0792+3.2041+3.1761+3.0792}{5}\right] = \text{antilog}(3.10772) = 1281.504 \approx 1281.5$$

The logarithmic formula is better as logarithms made higher magnitudes smaller to use.



Example 2:

Find average amount of cold drink using geometric mean in the following data.

Amount of cold drink (liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution:

Given data are grouped into discrete classes. Computations required in the basic and logarithmic formula are show in following table.

Classes (x)	f	$(x)^f$	log(x)	f × log(x)
1.48	2	2.1904	0.1703	0.3406
1.49	7	16.3044	0.1732	1.2124
1.50	5	7.5938	0.1761	0.8805
1.51	5	7.8503	0.1790	0.8950
1.52	1	1.52	0.1818	0.1818
---	$\sum f = 20$	$\prod x^f = 3236.0651$	---	$\sum f \log(x) = 3.5103$

Using basic formula: G.M. $= \left[\prod (x)^f\right]^{\frac{1}{\sum f}} = (3236.0651)^{\frac{1}{20}} = 1.49796 \approx 1.498$ liters

Using logarithmic formula: G.M. $=$ antilog $\left[\dfrac{\sum f \log x}{\sum f}\right] =$ antilog $\left[\dfrac{3.5103}{20}\right]$

G.M.$=$antilog $(0.175515) = 1.4980$ L. The G.M. amount of cold drink is 1.498 liters.

Example 3:

Find G.M. of the electricity consumption data using logarithmic method.

Electricity consumption (kWh)	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution:

The computation required in the formula are carried art in the following table.

Class limits	f	Class marks (x)	Log (x)	f × log(x)
68-87	10	77.5	1.8893	18.8930
88-107	13	97.5	1.9890	25.8570
108-127	15	117.5	2.0700	31.0500
128-147	10	137.5	2.1383	21.3830
148-167	4	157.5	2.1973	8.7892
168-187	6	177.5	2.2492	13.4952
188-207	2	197.5	2.2956	4.5912
	$\sum f = 60$			$\sum f \log x = 124.0586$

So, G.M. $=$ antilog $\left[\dfrac{\sum f \log x}{\sum f}\right] =$ antilog $\left[\dfrac{124.0586}{60}\right] =$ antilog $(2.0676) = 116.8422$ kWh.

Example 4:

Find average diameter of a rear of wire using G.M. Use both methods.

Diameter (mm)	2.10-2.13	2.14-2.17	2.18-2.21	2.22-2.25	2.26-2.29	2.30-2.33	2.34-2.37
Number of places	4	5	4	4	5	1	1

Solution:

Data are grouped into continuous classes. The computation follow in the table.

Class limits	Class marks (x)	f	x^f	log(x)	$f \times \log x$
2.10-2.13	2.115	4	20.0097	0.3253	1.3012
2.14-2.17	2.155	5	46.4768	0.3334	1.6670
2.18-2.21	2.195	4	23.2134	0.3414	1.3656
2.22-2.25	2.235	4	24.9523	0.3493	1.3972
2.26-2.29	2.275	5	60.9406	0.3570	1.7850
2.30-2.33	2.315	1	2.315	0.3646	0.3646
2.34-2.37	2.355	1	2.355	0.3720	0.3720
		$\sum f = 24$	$\prod x^f = 178967745.4$	---	$\sum f \log x = 8.2526$

By basic formula: G.M $= \left[\prod x^f\right]^{\frac{1}{\sum f}} = \left(178967745.4\right)^{\frac{1}{24}} = 2.2073$ mm

By logarithmic formula: G.M. = antilog $\left[\frac{\sum f \log x}{\sum f}\right]$ = antilog $\left[\frac{8.2526}{24}\right]$ = 2.2074 mm

Harmonic mean:

The **harmonic mean** (H.M.) is the reciprocal of the arithmetic mean of reciprocals of data. In H.M., we mix all n non-zero numbers by the sum of their reciprocals and then remove the effect of mixing by dividing the same into n **equal** parts. Finally, the H.M. is the reciprocal of the result. The H.M is usually computed for quantitative data, and it is always unique. If $x_1, x_2, ..., x_n$ are n non-zero numbers, then:

For ungrouped data: H.M. = Reciprocal of "arithmetic mean of reciprocals".

$$\text{H.M.} = \text{Reciprocal of } \left[\frac{\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}}{n}\right] = \text{Reciprocal of } \left[\frac{\sum\left(\frac{1}{x}\right)}{n}\right] = \frac{n}{\sum\left(\frac{1}{x}\right)} \qquad (1)$$

For grouped data, if x's are non-zero class marks, then $\text{H.M.} = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$ (2)

Example 1:

Find H.M. of the speeds of a vehicle measured for each 10 km distance. The speeds are: 15 km/h, 30 km/h, 22km/h, 30 km/h and 45 km/h.

Solution:

$$\text{H.M.} = \frac{n}{\sum\left(\frac{1}{x}\right)} = \frac{5}{\frac{1}{15}+\frac{1}{30}+\frac{1}{22}+\frac{1}{30}+\frac{1}{45}} = \frac{5}{0.0667+0.0333+0.0454+0.0333+0.0222}$$

$\text{H.M.} = \frac{5}{0.2009} = 24.8880$. So, average speed of vehicle is 12.4719 km/h using H.M.

Example 2:

The amount of cold drink (in liters) is given for 20 bottles. Find H.M. amount.

Amount of cold drink(in liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution:

Given data are grouped into discrete classes, so:

$$\text{H.M.} = \frac{\sum f}{\sum\left(\frac{f}{x}\right)} = \frac{2+7+5+5+1}{\frac{2}{1.48}+\frac{7}{1.49}+\frac{5}{1.50}+\frac{5}{1.51}+\frac{1}{1.52}}$$

$$\text{H.M.} = \frac{20}{1.3513+4.6980+3.3333+3.3112+0.6579} = \frac{20}{13.3517} = 1.4979 \text{ liters.}$$

Example 3:

Find harmonic mean mass (in Kg) of 50 blocks of metals given as::

Mass (Kg)	7.1-7.3	7.4-7.6	7.7-7.9	8.0-8.2	8.3-8.5	8.6-8.0	8.9-9.1
Numbers of blocks	3	5	9	14	11	6	2

Solution:

Data are grouped into continuous classes, so: $\text{H.M} = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$

Computing the required sums in the following table, we have:

Class limits	Class marks (x)	f	$\frac{f}{x}$
7.1-7.3	7.2	3	0.4167
7.4-7.6	7.5	5	0.6667
7.7-7.9	7.8	9	1.1538
8.0-8.2	8.1	14	1.7284
8.3-8.5	8.4	11	1.3095
8.6-8.8	8.7	6	0.6896
8.9-9.1	9.0	2	0.2222
Total	---	50	6.1869

$$\text{H.M} = \frac{50}{6.1869} = 8.0816 \text{ Kg.}$$

EXERCISE 22.3

1. Find A.M., G.M., H.M., median and mode in the following (wherever possible).
   - **a.** 3.2, 6, 10, 12, 12, -20, 25, 28, 30.8
   - **b.** 14, 12, 18, 19, 0, -19, -18, -12, -14
   - **c.** 6.5, 11, 12.3, 9, 8.1, 16, 18, 20.5, 25
   - **d.** 51, 55, 52, 54, 58, 60, 61, 62, 52, 57, 52, 64
   - **e.** A+, O-, AB+, O+, AB+, AB-, B+, AB+, O+, A- (blood groups)
   - **f.** North, South, East, West (directions)
   - **g.** B+, Fail, B+, A+, A-, C+, B+, A-, B-, Fail, A-, B+, C- (grades)

2. The daily earnings for ten workers in Rs. are: 188, 170, 172, 125, 115, 195, 181, 190, 195, 190. Find A.M. (by definition and deviations with A = 50, h = 10), G.M. (by definition and logarithmic method), H.M, median and mode.

3. Find average attitude for dogs using A.M. median and mode from 60 people data.

Attitude	Love dogs	Like dogs	No opinion	Dislike dogs	Hate dogs
Number of people	20	15	4	13	8

4. Find modal cause of death from the following data of mortality/month in a city.

Cause of death	T.B.	Diabetics	Malaria	Cholera	COVID	Cancer	B.P.	Heart Attack
Number of people	10	6	2	2	15	10	2	5

5. The sizes of shoe sold at a store on a 50% off price are listed. Calculate A.M., G.M., H.M., median, Q_1, Q_3 and modal shoe size sold that day.

Shoe size	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5
Number of pairs sold	2	5	15	30	60	40	23	11	4	1

6. Daily wages (in Rs. 100) for thousand employees in a factory are given. Find A.M., G.M., H.M., median, quartiles and modal wages.

Daily wages (in Rs. 100)	22	24	26	28	30	32	34	36	38	40	42	44
Number of employees	3	13	43	102	175	220	204	139	69	25	6	1

7. The profits earned by a company for a period of last 50 days are summarized below. Find the A.M. profit using shortcut and coding methods with

(a). A = 9000, h = 2000 (b). A = 11000, h = 2000

Profits (Rs.)	4000-6000	6000-8000	8000-10000	10000-12000	12000-14000
Number of days	5	7	11	21	6

8. The marks obtained by students in a subject (out of 50) are given in the following grouped table. Find A.M., G.M. (using direct and logarithmic methods), H.M., median and mode.

Marks	25-29	30-34	35-39	40-44	45-49
Number of students	9	18	35	17	5

9. The following data show number of devices resulting in observed values in appropriate ranges. Find A.M., G.M., H.M., median, quartiles and mode.

Class limits	10.5-10.9	11.0-11.4	11.5-11.9	12.0-12.4	12.5-12.9
Frequencies	2	7	10	12	8

22.3(ii) Recognize properties of arithmetic mean:

Here, we discuss some important properties of A.M. with examples.

1. The sum of all deviations of observations in an ungrouped data about A.M. is zero.
2. A.M. of a constant dataset is that constant itself.
3. If A.M. of a set X is $\bar{x}$, then of Y = aX + b is $\bar{y} = a\bar{x} + b$, where a, b are real numbers.
4. If A.M. = G.M. = H.M., then all observations in data are same or constant.
5. For a non-constant data, A.M. > G.M. > H.M.
6. For a non-constant data: $(\text{A.M})(\text{H.M}) \approx (\text{G.M.})^2$.
7. If A.M. = Median, then the data are **symmetric**, otherwise **asymmetric**.
8. In a uni-modal symmetric data: A.M. = Median = Mode.
9. In a constant data: A.M. = G.M. = H.M. = Median = Mode.

Example 1:

Find the arithmetic mean of the following datasets:

(i). X={19, 19, 19, 19, 19} (ii). Z = 3Y + 7, where Y={3, 4, 6, 1, 6}

Solution:

(i). X= $\{19,19,19,19,19\}$ is constant with all values equal to 19, so: $\bar{x} = 19$.

(ii). First we find: $\bar{y} = \frac{3+4+6+1+6}{5} = \frac{20}{5} = 4$

Now, as Z = 3Y + 7, so: $\bar{z} = 3\bar{y} + 7 = 3(4) + 7 = 19$.

We can verify that, when Z = 3Y + 7 = $\{16,19,25,10,25\}$, then $\bar{z} = 19$ is correct.

Example 2:

Show that sum of all deviations in $\{3,4,6,1,6\}$ about its A.M. is zero.

Proof:

Let X = $\{3,4,6,1,6\}$. Finding $\bar{x}$ first, which is: $\bar{x} = \frac{\sum x}{n} = \frac{20}{5} = 4$.

Now, the deviations about $\bar{x}$ are: (3-4), (4-4), (6-4), (1-4), and (6-4), or simply: -1, 0, 2, -3 and 2. Now, we observe the sum of all deviations about $\bar{x}$:

$\sum(X - \bar{x}) = (-1) + (0) + (2) + (-3) + (2) = 0. \Rightarrow \sum(X - \bar{x}) = 0.$ Hence shown.

Example 3:

Verify that A.M. = G.M. = H.M. = Median = Mode for $\{19,19,19,19,19\}$.

Solution:

Finding all averages for the given constant dataset. Here, n = 5 (odd).

$$\text{A.M.} = \frac{19+19+19+19+19}{5} = 19 \qquad \text{(i)}$$

$$\text{G.M.} = (19\times19\times19\times19\times19)^{1/5} = 19^{5\times\frac{1}{5}} = 19. \qquad \text{(ii)}$$

$$\text{H.M.} = \frac{5}{\frac{1}{19}+\frac{1}{19}+\frac{1}{19}+\frac{1}{19}+\frac{1}{19}} = \frac{5}{\frac{5}{19}} = 19. \qquad \text{(iii)}$$

$$\text{Median} = \left(\frac{5+1}{2}\right)^{th} \text{observation} = 3^{rd} \text{ observation} = 19. \qquad \text{(iv)}$$

Mode = 19 (Most frequent observation). (v)

From equations (i)-(v), we have: A.M. = G.M. = H.M. = Median = Mode. Hence proved

Example 4:

Use A.M. and median to see if the following data are symmetric/ asymmetric.

(i). 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10 (ii). 4, 8, 13, 14, 19, 20, 23

Solution:

(i). $$\text{A.M.} = \frac{4+5+6+6+6+7+7+7+7+7+7+8+8+8+9+10}{16} = \frac{112}{16} = 7$$

$$\text{Median} = \frac{1}{2}\left[8^{th} + 9^{th} \text{ observation}\right] = \frac{1}{2}[7+7] = 7$$

As A.M = Median, so data are symmetric.

(ii). $\text{A.M} = \frac{4+8+13+14+19+20+23}{7} = 14.42$

$\text{Median} = \left[\frac{7+1}{2}\right]^{th}$ observation $= 4^{th}$ observation $= 14$

As A.M. $\neq$ Median, so data are asymmetric.

Example 5:

For {4,5,6,6,6,7,7,8}, verify (i). A.M. > G.M. > H.M. and
(ii). $(\text{G.M})^2 \approx (\text{A.M})(\text{H.M})$.

Solution:

$\text{A.M.} = \frac{4+5+6+6+6+7+7+8}{8} = 6.125$, $\text{G.M.} = (4\times5\times6\times6\times6\times7\times7\times8)^{1/8} = 6.006$

$\text{H.M.} = \frac{8}{\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{7}+\frac{1}{7}+\frac{1}{8}} = 5.879$.

As 6.125 > 6.006 > 5.879, so: A.M. > G.M. > H.M. Hence verified (i)

As $(\text{G.M.})^2 = 36.072$ and (A.M.) (H.M.)=36.008:

So, $(\text{G.M.})^2 \approx (\text{A.M.})(\text{H.M.})$. Hence verified (ii).

22.3 (iii) Calculate weighted mean and moving averages

(a) Weighted mean:

When some observations are more important than others in a data, then we cannot give equal weight (relative importance) to all for computing the mean. The mean which is computing by using relative importance / weights of the observations is called **weighted mean**.

If $x_1, x_2, ..., x_n$ are numbers associated with observations in a data and

$w_1, w_2, ..., w_n$ are corresponding weights, then weighted A.M., G.M. and H.M. are:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i} = \frac{\sum wx}{\sum w}, \quad G_w = \text{antilog}\left[\frac{\sum w \log x}{\sum w}\right] \quad \text{and} \quad H_w = \frac{\sum w}{\sum\left(\frac{w}{x}\right)}$$

Example 1:

The marks of a student in a social sciences diploma course are given for each subject taught with relative weights. Find weighed mean marks:

Subject	English	French	History	Science	Mathematics
Marks	73	82	60	62	57
Weights	4	3	2	1	1

Solution:

The formulas are: $\bar{x}_w = \frac{\sum xw}{\sum w}$, $G_w = \text{antilog}\left[\frac{1}{\sum w}\left(\sum w \log x\right)\right]$ and $H_w = \frac{\sum w}{\sum\left(\frac{w}{x}\right)}$.

We compute the required sums in following table:

x	w	$x.w$	$\frac{w}{x}$	log x	w. log x
73	4	292	0.0548	1.8633	7.4532
82	3	246	0.0366	1.9138	5.7414
60	2	120	0.0333	1.7782	3.5564
62	1	62	0.0161	1.7924	1.7924
57	1	57	0.0175	1.7559	1.7559
Sum	11	777	0.1538	---	20.2993

$$\bar{x}_w = \frac{777}{11} = 70.6 \text{ marks.}$$

$$G_w = \text{anti log}\left[\frac{1}{11} \times 20.2993\right]$$

$$G_w = \text{antilog}(1.8454) = 70.04 \text{ marks.}$$

$$H_w = \frac{11}{0.1583} = 69.5 \text{ marks.}$$

Note that:

In Example 1, the unweighted means are: $\bar{x} = 66.8$, G.M.=66.17 and H.M.=65.58 which are lower than the weighted means. The weighted mean marks best describe the average performance of student by giving more share to important courses.

(b). Moving Averages:

When we want to examine average behavior of a variable of interest over fixed intervals of time (in days/months/years/etc) then it is important to track the change in average as time proceeds for specific periods. In this case, the average of specific periods/ intervals of time keep changing/ moving as time increases, and are referred as **moving averages**.

For the case of an odd-period moving averages, we place the moving averages at the middle cells of specific period only and rest of the cells are left blank. The placement of moving averages should be done at the already existing cells.

For the case of an even-period moving averages, we place the average of two middle cell. This way placement can be done at already existing cells.

The process is illustrated using following examples.

Example 1:

The data summarizes monthly price of an item in Rs./month for the last year.

Month	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Price (Rs../month)	180	87	91	102	108	139	150	200	220	307	289	330

Compute (i). 3-months, (ii). 5-months and (iii). 4-months moving averages.

Solution:

(i). First we compute 3-months moving averages by considering 3 consecutive months starting from: Jan.-Mar., Feb.-Apr., Mar.-May., and so on until the last 3-months period: Oct.-Dec. These are placed at middle month of the 3-months period. For example, the moving average of Jan.-Mar. is placed beside Feb.

The 3-months moving averages are summarized in the adjacent table.

Month	Price	3- months moving averages
Jan.	180	----------
Feb.	87	(180+87+91)/3=119.33
Mar.	91	93.33
Apr.	102	100.33
May.	108	166.33
Jun.	139	132.33
Jul.	150	163
Aug.	200	190
Sep.	220	242.33
Oct.	307	272
Nov.	289	308.66
Dec.	330	----------

(ii). For 5- months moving averages, we consider 5 consecutive months each time starting from Jan.-May, Feb.-June, and so on up to Aug.-Dec.

Moving averages are shown in the adjacent table.

We can see that there are total 8 such possibilities. The first and last two cells are left blank.

The 5-months moving averages are positioned at the middle months.

Month	Price	5- months moving averages
Jan.	180	----------
Feb.	87	----------
Mar.	91	(180+87+91+102+108)/5=113.6
Apr.	102	105.4
May.	108	118
Jun.	139	139.8
Jul.	150	163.4
Aug.	200	203.2
Sep.	220	233.2
Oct.	307	269.2
Nov.	289	----------
Dec.	330	----------

(iii). For 4- months moving averages, we consider groups of 4 consecutive months each time starting from Jan.-Apr., Feb.-May, … , Sep.-Dec.

Initially the averages are placed mid-way between the 4 months, then each pair of two 4-months averages are averaged further to get the required 4-months centered moving averages.

The results are summarized in the following table. We have to add two columns for even-period group averages always.

Months	Prices	4- months moving averages(Initial)	4- month moving averages(centered)
Jan.	180		----------
Feb.	87		----------
		115	
Mar.	91		106
		97	
Apr.	102		103.5
		110	
May.	108		117.375
		124.75	
Jun.	139		137
		149.25	
Jul.	150		163.25
		177.25	
Aug.	200		198.25
		219.25	
Sep.	220		236.625
		254	
Oct.	307		270.25
		286.5	
Nov.	289		----------
Dec.	330		----------

Note that:

The average price (arithmetic mean) of the 12 months for the year equals Rs. 183.58 per month. We can see that the prices were not constant, and the 3, 5, 4- months moving averages better summarize the variations in respective periods over the year.

22.3 (iv): Estimate medians, quartiles and mode, graphically.

Graphically, to locate and estimate median and quartiles, we use ogive (cumulative frequency polygon) of the data. The median, lower quartile (Q_1) and upper quartile (Q_3) in an ogive are simply the x-coordinates of the points on ogive whose y-coordinates (cumulative frequencies) are $\frac{n}{2}$, $\frac{n}{4}$ and $\frac{3n}{4}$, respectively.

To locate and estimate mode (if it exists) graphically, we use the histogram of data. The modal class is one for which the rectangle has highest height, and in this class, the mode is the x-coordinate of the point of intersection of two line segments. First segment is drawn by joining left peaks of the rectangles corresponding to the modal class and class after it. The second line segment is drawn by joining right peaks of the rectangles corresponding to the modal class and class before it. The perpendicular from point of intersection to the x-axis hits it at the location of mode. This procedure is used for a grouped data with continuous classes.

Example 1:

Locate and estimate median, quartiles and modal weights (in Kg) graphically.

Weights (in Kg)	70-74	75-79	80-84	85-89	90-94
Number of students	10	24	46	12	8

Solution:

We have n=100 and, and given data are grouped with continuous classes. First, computing the class boundaries and cumulative frequencies in the adjacent table to be able to draw ogive and histogram.

Class limits	Class boundaries	f	Cumulative frequencies
70-74	69.5-74.5	10	10
75-79	74.5-79.5	24	34
80-84	79.5-84.5	46	80
85-89	84.5-89.5	12	92
90-94	89.5-94.5	8	100

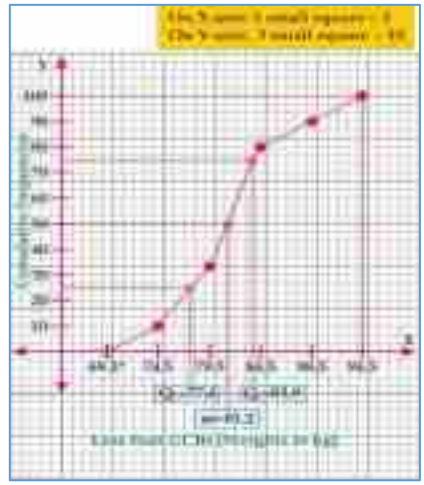
For median and quartiles, we identify $\frac{n}{2} = 50$, $\frac{n}{4} = 25$ and $\frac{3n}{4} = 75$ on y-axis.

Then, drawing perpendiculars through these to ogive, and then from ogive to x-axis lead to median, Q_1 and Q_3, respectively. We read values as:

Median ≈ 81.2,

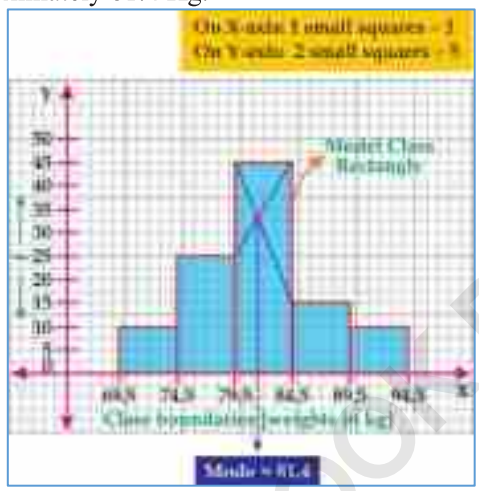
$Q_1 \approx 77.6$,

$Q_3 \approx 83.9$.



For mode, we first draw the histogram, as shown here. The modal class corresponds to the highest height rectangle as highlighted in the figure.

Joining left peaks of the rectangles of modal class and class after it, and then right peaks of the rectangles of the modal class and class before it to get two line segments. Now, the perpendicular from point of intersection to the x-axis hits at nearly 81.4 where mode is located. The modal weight is approximately 81.4 Kg.



EXERCISE 22.4

1. If W = {148, 145, 160, 157, 156, 160, 160, 165}, X = {-2, -2, -2, -2}, then:
   a. Show that sum of all deviations of W about its A.M. is zero.
   b. Show that A.M. = G.M. = H.M. = Median = Mode for X.
   c. Compute A.M. of Y = 3X.
   d. Compute A.M. of Z = 3W-11
   e. Show that H.M. < G.M. < A.M. < Median < Mode for W.
2. Check if X = {7,9,3,3,3,4,1,3,2,2} and Y = {2,1,4,4,4,6,6,5,7,1} are symmetric?
3. The required weights (Kg) and price (Rs./Kg) of monthly items required by a family are given. Find weighted mean prices using A.M., G.M. and H.M.

Item	A	B	C	D	E
Price (Rs./Kg)	300	200	600	250	650
Required weight (Kg)	25	5	4	8	3

4. Of 50 bricks bought, 21 bricks have mean mass of 24.2Kg, 29 bricks have mean mass of 23.6Kg, Find weighted mean mass of the 50 bricks.
5. Calculate 2, 3 and 4-days moving average from the following data of number of deaths monitored for a peak week in a province due to nCOVID'19.

Days	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Number of deaths	102	130	158	188	196	259	310

6. Calculate 4 and 5 years moving averages of sales (in million Rs.) over 11 years.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Sales	102	130	158	188	196	259	310	188	196	259	310

7. Locate and estimate median, quartiles and mode graphically in the data.

Overtime (hrs/week)	25-29	30-34	35-39	40-44	45-49	50-54	55-59
Number of days	5	4	7	11	12	8	1

8. Locate and estimate the indicated measures in the following.

a. Mode

b. Median and quartiles

22.4. Measures of Dispersion

The measures of central tendency only focus on the average or central behavior of data without focusing on the variations and consistency of data. When observations are much away from the average, then the measures of central tendency alone are not sufficient to have a thorough understanding of the behavior and properties of a data. **For example:**

The table below shows the marks obtained by two students in five subjects.

Subjects	I	II	III	IV	V
Marks of student-A	90	30	85	50	25
Marks of student-B	63	50	60	55	52

If we find average marks obtained by each student using A.M., then we observe that the A.M. marks of both students are 56, but is it sufficient to compare the performance of both students? The answer is No. Student-A marks have higher fluctuations about the mean, whereas Student-B marks are closer to the mean marks.

The performance of Student-B is more stable/ consistent/ reliable than of Student-A due to lesser fluctuations/ variations/ scatter/ dispersions about the mean.

We can also observe average marks of students using all types we learnt below:

Average	A.M.	Median	Mode	G.M.	H.M.
Marks of student A	56	50	None	49 (Approximately)	43 (Approximately)
Marks of student B	56	55	None	56 (Approximately)	56 (Approximately)

Due to variation in marks, the averages for Student-A are also much different than each other. The averages of the marks of Student-B are closer to each other.

Due to stable and consistent performance, we are in a better position to predict the marks of Student-B in a forthcoming subject as compared to Student-A.

The knowledge of variation in data matters a lot in decision making and forecasting. Therefore, it is important to compute the degree of variation (or dispersion) in a data besides computing the average to have a better understanding of data. The **measures of dispersion** are indicators of the extent to which the data are spread-out/scattered about the central point or average. A **measure of dispersion** indicates the average variation in a data. It is mostly used to compare two or more data sets from the view-point of stability/ consistency / reliability. Higher is the variation/ dispersion in a data, lower is the stability/ consistency/ reliability.

When we are interested in variation/dispersion in only one data set, we compute the measures of dispersion in the units of data, and are thus called **absolute measures of dispersion**. The variation of two or more data sets with same nature/units, number of observations and equal/closer averages can also be compared using absolute measures of dispersion.

For two or more data sets with different nature/units, number of observations or different averages, we use **relative measures of dispersion**, which are not in the units of data but show the relative behavior of the variation.

22.4(i) Define, identify and measure range, calculate variance, mean deviation and standard deviation:

We discuss some important measures of dispersion here.

(a) Range: Range of a data refers to the difference between the extreme observations.

For ungrouped data, range (R) is the difference between the largest observation (x_L) and smallest observation (x_S), i.e. $R = x_L - x_S$ (1)

For grouped data, R is the difference between UCB of the final group (UCB_F) and the LCB of the initial group (LCB_I), i.e. $R = UCB_F - LCB_I$ (2)

Formulas (1)-(2) calculate **absolute range**. For **relative range** we divide by $x_L + x_S$ and $UCB_F + LCB_I$ in equations (1) and (2), respectively for comparison.

Note that:

- Range does not properly measure variation of data with outliers.
- Range uses extremes in data only without using frequencies and other parameters.

Example 1: Find range of the data sets: W = {27.90, 34.70, 54.40, 18.92, 47.60, 39.68}, X = {3, 8, 10, 7, 5, 14, 2, 12, 8}, Y = {8, 8, 8, 8, 8} and Z = {1, 3, 2, 4, 3, 2, 43}. Also interpret the results.

Solution: We know that for an ungrouped data: $R = x_L - x_S$ .So,

$R_W = 54.40 - 18.92 = 35.48, R_X = 14 - 2 = 12, R_Y = 8 - 8 = 0$ and $R_Z = 43 - 1 = 42$.

We can interpret range for dataset, say W, as: "All observations in dataset W lie within a distance of 35.48 between the extreme values 18.92 and 54.40.

The observations in data W and X are well-spread around the extreme values, so range meaningfully represent the variation. The range of set Y is zero as all observations were same and there is no variation. The majority of observations in set Z lie within 1 and 4, but due to an outlier 43, the range equal to 42 is misleading.

Example 2:

Compare the variation of marks of two students in five subjects as given below using range, and interpret the results. Which student performs more consistently?

Marks of student A	90	30	85	50	25
Marks of student B	63	50	60	55	52

Solution:

Data for both students are ungrouped. The absolute range in marks are: $R_A = 90 - 25 = 65$ marks and $R_B = 63 - 50 = 13$ marks. Here, $x_L + x_S$ is same for both students, so relative range is not useful here. The performance of Student-A show higher variation as compared to Student-B on the basis of absolute range. So, Student-B performs more consistently than Student-A.

Example 3:

Using range compare variation of price and life of five similar rating batteries manufactured by two different companies. Interpret the results.

Price (in thousand Rs.)	8	13	18	23	30
Life (in years)	1.3	1.5	1.8	2.5	3.5

Solution:

The data of price (P) and life (L) of batteries are ungrouped, and different in nature and units. Also, $x_L + x_S$ are different for both datasets. So we use the relative range to compare variation.

$$\text{Relative } R_P = \frac{P_L - P_S}{P_L + P_S} = \frac{30 - 8}{30 + 8} = \frac{22}{38} = 0.5789. \quad \text{Relative } R_L = \frac{L_L - L_S}{L_L + L_S} = \frac{3.5 - 1.3}{3.5 + 1.3} = \frac{2.2}{4.8} = 0.4583.$$

The relative variation in price of batteries manufactured by five different companies is higher in comparison to the relative variation in lives of batteries.

Example 4:

Compare the range of number of mobiles possessed by a family for Karachi and Moro using the grouped data of some randomly selected families in the cities. Is it sufficient to compare the variation on basis of range?

Number of mobiles	0	1	2	3	4	5	6	7
Number of families (Karachi)	1	7	15	12	5	2	2	1
Number of families (Moro)	4	13	8	5	2	1	0	1

Solution:

The data are grouped into discrete classes with spacing between the classes d=1. So, the range is: $R = UCB_F - LCB_I = (7 + 0.5) - (0 - 0.5) = 7.5 + 0.5 = 8$ mobiles. The range of number of mobiles possessed by a family in Karachi and Moro is same, i.e. 8 mobiles. But, range does not consider number of families in the cities (frequencies). The variation between the number of mobiles possessed by an average family in Karachi and Moro are surely different, but this cannot be reflected by the range.

Example 5:

Find the range of electricity consumption (in kWh) of a shop using the data:

Electricity consumption	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution:

The data are grouped into continuous classes with spacing between the classes $d = 1$. So, $R = UCB_F - LCB_1 = \left[207 + \frac{1}{2}\right] - \left[68 - \frac{1}{2}\right] = 207.5 - 67.5 = 140$ kWh. Here, R does not use the frequencies of classes, so can be misleading.

(b) Variance, mean deviation and standard deviation

To better calculate the amount of dispersion in data, we use all observations instead of only the extreme ones along with their frequencies in the concepts of **variance, mean deviation and standard deviation**. The deviation of an observation x_i about the mean (usually A.M.) is defined as:

Deviation about mean = $D_i = x_i - \bar{x}$ (1)

Deviations can be negative, positive or zero. But, to calculate total of all deviations, simply summing them will not be useful as the result will always be zero.

To overcome this, we must avoid negative deviations, and this is done in two ways. We can use the squared deviations or the absolute deviations, defined as:

Squared deviation about mean = $D_i^2 = (x_i - \bar{x})^2$ (2)

Absolute deviation about mean = $|D_i| = |x_i - \bar{x}|$ (3)

The squared deviations are always are non-negative. The absolute value of a real number is always non-negative, for example: $|3| = 3, |-22| = 22, |0| = 0, |-1.5| = 1.5$. These two ways lead us to define more appropriate measures of dispersions.

The **variance** (Var or s^2) is the mean of squared deviations of observations about their A.M. The **mean (or absolute) deviation** ($M.D.$) is the mean of absolute deviations of observations about their A.M. Because the dispersion is usually meaningful in original units of data, and the variance finds it in squared units, so we find its square root to define the standard deviation. The **standard deviation** ($S.D.$ or s) is the positive square root of variance.

For ungrouped data with observations: $x_1, x_2, x_3, ..., x_n$ with their A.M. $= \bar{x}$:

$$Var = s^2 = \frac{\sum (x - \bar{x})^2}{n} \quad (4) \qquad M.D. = \frac{\sum |x - \bar{x}|}{n} \quad (5)$$

$$S.D. = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \quad (6) \qquad \text{where } \bar{x} = \frac{\sum x}{n}$$

For grouped data, if x_i are class marks and f_i are class frequencies, then:

$$Var = s^2 = \frac{\sum (x - \bar{x})^2 f}{\sum f} \quad (7) \qquad M.D. = \frac{\sum |x - \bar{x}| f}{\sum f} \quad (8)$$

$$S.D. = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}} \quad (9) \qquad \text{where } \bar{x} = \frac{\sum xf}{\sum f}$$

Formulae (4)-(9) calculate **absolute measures,** the **relative measures** are computed by dividing $\bar{x}$ in the units of data for comparison.

Example 1:

Find variance, mean deviation and standard deviation of the marks of students A and B. Which student performs more consistently?

Marks of student A	90	30	85	50	25
Marks of student B	63	50	60	55	52

Solution:

Marks of both students show ungrouped data. First, we compute mean marks.

For student-A (x) and student-B (y), we have:

$$\bar{x} = \frac{\sum x}{n} = \frac{280}{5} = 56 \quad \text{and} \quad \bar{y} = \frac{\sum y}{n} = \frac{280}{5} = 56$$

As means are same, so we use the absolute variation formulas variance, mean deviation and standard deviation. Computations are shown in the following tables.

x	$x-\bar{x}$	$(x-\bar{x})^2$	$\lvert x-\bar{x}\rvert$
90	34	1156	34
30	-26	676	26
85	29	841	29
50	-6	36	6
25	-31	961	31
Sum=280	0	3670	126

$$Var_A = \frac{\sum (x-\bar{x})^2}{n} = \frac{3670}{5} = 734.$$

$$(M.D.)_A = \frac{\sum |x-\bar{x}|}{n} = \frac{126}{5} = 25.2.$$

$$(S.D.)_A = \sqrt{734} = 27.0924.$$

y	$y-\bar{y}$	$(y-\bar{y})^2$	$\lvert y-\bar{y}\rvert$
63	7	49	7
50	-6	36	6
60	4	16	4
55	-1	1	1
52	-4	16	4
Sum=280	0	118	22

$$Var_B = \frac{\sum (y-\bar{y})^2}{n} = \frac{118}{5} = 23.6.$$

$$(M.D.)_B = \frac{\sum |y-\bar{y}|}{n} = \frac{22}{5} = 4.4.$$

$$(S.D.)_B = \sqrt{23.6} = 4.8578.$$

We observe that the variation in marks of student A is much higher than that in the marks of student B by all formulae, so performance of student B is more consistent.

Example 2: Compare the variation using standard deviation of price and life of five similar rating batteries manufactured by different companies. Also interpret the results.

Price (in thousand Rs.)	8	13	18	23	30
Life (in years)	1.3	1.5	1.8	2.5	3.5

Solution: Data of price (P) and life (L) are ungrouped and different in nature and units. First, we compute means.

$$\bar{x}_P = \frac{\sum x_P}{n} = \frac{92}{5} = 18.4 (1000 \text{ Rs}) \text{ and } \bar{x}_L = \frac{\sum x_L}{n} = \frac{10.6}{5} = 2.12 \text{ years.}$$

As means are different, we use the relative standard deviation formula for comparison. Computations are shown in the following tables for P and L.

Price (P)	$x_P - \bar{x}_P$	$(x_P - \bar{x}_P)^2$
8	-10.4	108.16
13	-5.4	29.16
18	-0.4	0.16
23	4.6`	21.16
30	11.6	134.56
Sum=92	0	293.2

$$s_P = \sqrt{\frac{\sum \left(x_P - \bar{x}_P\right)^2}{n}} = \sqrt{\frac{293.2}{5}} = 7.658 (1000 \text{ Rs})$$

$$\text{Relative } s_P = \frac{s_P}{\bar{x}_P} = \frac{7.658}{18.4} = 0.416.$$

Life (L)	$x_L - \bar{x}_L$	$(x_L - \bar{x}_L)^2$
1.3	-0.82	1.69
1.5	-0.62	2.25
1.8	-0.32	3.24
2.5	0.38	6.25
3.5	1.38	12.25
Sum=10.6	0	25.68

$$s_L = \sqrt{\frac{\sum \left(x_L - \bar{x}_L\right)^2}{n}} = \sqrt{\frac{25.68}{5}} = 2.266 \text{ years}$$

$$\text{Relative } s_L = \frac{s_L}{\bar{x}_L} = \frac{2.266}{2.12} = 1.069.$$

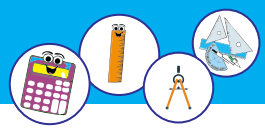
The relative standard deviation in life of batteries is higher than price. This means that data of price of batteries is more consistent than the lives of batteries. We note that if relative formula was not used then conclusion would be different and wrong.

Example 3: Find variance, mean deviation and standard deviation of amount of cold drink.

Amount of cold drink (in liters)	1.48	1.49	1.50	1.51	1.52
Number of bottles	2	7	5	5	1

Solution: Data are grouped with discrete classes. Calculations are given in table below.

Amounts (x)	Number of bottles (f)	xf	$(x - \bar{x})$	$(x - \bar{x})^2$	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2 f$	$\lvert x - \bar{x} \rvert f$
1.48	2	2.96	-0.018	0.0003240	0.018	0.000648	0.036
1.49	7	10.43	-0.008	0.0000640	0.008	0.000448	0.056
1.5	5	7.5	0.002	0.0000040	0.002	0.00002	0.01
1.51	5	7.55	0.012	0.0001440	0.012	0.00072	0.06
1.52	1	1.52	0.022	0.0004840	0.022	0.000484	0.022
Sum	20	29.96	--	--	--	0.00232	0.184



$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{29.96}{20} = 1.498 \text{ liters.} \qquad Var = s^2 = \frac{\sum \left(x - \bar{\bar{x}}\right) f}{\sum f} = \frac{0.00232}{20} = 0.000116 \left(\text{liters}\right)^2.$$

$$S.D. = s = \sqrt{0.000116} = 0.01078 \text{ liters.} \qquad \text{M.D.} = \frac{\sum \left|x - \bar{x}\right| f}{\sum f} = \frac{0.184}{20} = 0.0092 \text{ liters.}$$

The variation in amounts of cold drink is much smaller, or negligible.

Example 4: Find mean deviation and standard deviation of electricity consumption data.

Electricity consumption (kWh)	68-87	88-107	108-127	128-147	148-167	168-187	188-207
Number of days	10	13	15	10	4	6	2

Solution: The calculations are summarized in the table below, where x show class marks.

Class limits	Class marks (x)	f	xf	$(x-\bar{x})$	$(x-\bar{x})^2$	$\lvert x-\bar{x}\rvert$	$(x-\bar{x})^2 f$	$\lvert x-\bar{x}\rvert f$
68-87	77.5	10	775	-44.167	1950.724	44.167	19507.24	441.67
88-107	97.5	13	1267.5	-24.167	584.0439	24.167	7592.571	314.171
108-127	117.5	15	1762.5	-4.167	17.36389	4.167	260.4583	62.505
128-147	137.5	10	1375	15.833	250.6839	15.833	2506.839	158.33
148-167	157.5	4	630	35.833	1284.004	35.833	5136.016	143.332
168-187	177.5	6	1065	55.833	3117.324	55.833	18703.94	334.998
188-207	197.5	2	395	75.833	5750.644	75.833	11501.29	151.666
Sum	--	60	7270	--	--	--	65208.35	1606.672

First we need mean consumption, which is: $\bar{x} = \frac{\sum xf}{\sum f} = \frac{7270}{60} = 121.667 \text{kWh.}$

Now, using the required sums, the mean and standard deviations are:

$$\text{M.D.} = \frac{\sum \left|x - \bar{x}\right| f}{\sum f} = \frac{1606.672}{60} = 26.778 \text{kWh.} \quad S.D. = s = \sqrt{\frac{\sum \left(x - \bar{x}\right)^2 f}{\sum f}} \sqrt{\frac{65208.35}{60}} = 32.967 \text{kWh.}$$

Example 5:

Some randomly selected families in the Karachi and Moro, when asked about number of mobiles possessed by them, resulted in an average mobiles/family of 5.773 and 6.133 with standard deviations 1.412 and 3.408, respectively. Compare which city shows more consistent use of mobiles per family?

Solution:

The average mobiles/family for Karachi (K) and Moro (M) are different as: $\bar{x}_K = 5.773 \text{ mibiles/family}$ and $\bar{x}_M = 6.133 \text{mibiles/family}$.

The standard deviations are given as:

$s_K = 1.412$ mobiles/family and $s_M = 3.408$ mobiles/family.

As mean mobiles/family differ for both cities, we use relative measures:

$$\text{Relative } s_K = \frac{s_K}{\bar{x}_K} = \frac{1.412}{5.733} = 0.246 \text{ and Relative } s_M = \frac{s_M}{\bar{x}_M} = \frac{3.408}{6.133} = 0.555.$$

By observing relative standard variations of both cities, we conclude that the relative variation for number of mobiles possessed by a family in Karachi is smaller than that for Moro. So, average use of mobiles/ family is higher and more consistent for Karachi than Moro.

EXERCISE 22.5

1. Find range, variance, mean deviation and standard deviation of number of absentees in a class for last seven days: 3, 5, 3, 2, 4, 1, 8.
2. Find range, mean deviation, variance and standard deviations of scores of two batsmen in 6 innings. Who is more consistent player?

Batsman-A	12	15	0	185	7	19
Batsman-B	47	12	76	48	4	51

3. Compare variation using range and standard deviation of income and expenditure of six families. Also interpret the results.

Income (in thousand Rs.)	10	20	30	40	50	60
Expenditure (in thousand Rs.)	7	21	23	34	36	53

4. Goals scored by teams A and B in a football season are given. Use relative standard and mean deviations to find which team performed more consistently?

Goals scored	0	1	2	3	4
Number of games A played	27	9	8	5	4
Number of games B played	17	9	6	5	3

5. Find range, variance, standard deviation and mean deviation in mass of 50 blocks of metal as distributed in the following data:

Mass of block(in Kg)	7.1-7.3	7.4-7.6	7.7-7.9	8.0-8.2	8.3-8.5	8.6-8.8	8.9-9.1
Number of blocks	3	5	9	14	11	6	2

Review Exercise 22

1. Tick the correct answer.

(i). Arithmetic mean is based on the use of _____ observations of data.
(a). Middle (b). Extreme (c). All (d). None

(ii). The ungrouped data must be ordered first to find _____ .
(a). A.M. (b). Mode (c). Median (d). Range

(iii). The number of classes in a continuous frequency table lies between 5 and ___.
(a). 10 (b). 15 (c). 20 (d). 25

(iv). Mode of grouped data is obtained graphically using _____.
(a). Histogram (b). Polygon (c). Ogive (d). Bar chart

(v). If data has outliers, then _____ is misleading.
(a). A.M. (b). Range (c). Median (d). Mode

(vi). The A.M. of {0, 90, k, 10, 100} is 40, then k = ____ .
(a). 0 (b). 90 (c). 10 (d). 100

(vii). Median and quartiles are _____ in nature.
(a). Mathematical (b). Positional (c). Logical (d). None of these

(viii). Upper quartile divides data in _____ ratio.
(a). 50%–50% (b). 25%–25% (c). 75%–25% (d). 40%–60%

(ix). If data contain a number equal to 0, then ____ cannot be computed.
(a). A.M. (b). G.M. (c). H.M. (d). Median

(x). If all numbers in data are equal, then:
(a). A.M.=G.M.=H.M. (b). Range = 0 (c). S.D. = 0 (d). All of these

SUMMARY

- Quantitative data refers to numbers, and are further categorized into discrete (only integers) and continuous data (any real numbers).
- Qualitative data refers to attributes of a variable, and are further categorized into nominal (without ordering) and ordinal (with ordering) data.
- To get meaningful information, data are grouped into frequency distribution.
- Frequency distribution lists classes and respective frequencies.
- Classes defined by a single point are discrete, whereas by a range of numbers are continuous.
- Observations are distributed into classes using tally or list entries method.
- Number of classes can be obtained using Sturges' (1926) rule.
- The range is the difference between highest and lowest observations in a data.
- The number of observations that can fall in a class is class interval or width.
- Class limits are starting and ending points of a class.
- Class boundaries are obtained by averaging any two consecutive class limits.
- The spacing between classes is the distance between any two adjacent classes.
- Relative and percentage frequencies lie from 0 to 1 and 0 to 100, respectively.
- Cumulative frequencies are the sum of all frequencies upto a particular class.
- Histogram is a graph of adjacent rectangles with bases at class boundaries and heights proportional to class frequencies.
- Frequency polygon is formed by connecting points with line segments whose x-coordinates are class marks and y-coordinates are frequencies.
- Ogive or cumulative frequency polygon is obtained by joining the points with coordinates (class boundaries, cumulative frequencies).
- A measure of central tendency is capacity of data to cluster about a central point.
- A.M., G.M. and H.M. are based on the equality among all observations.
- Median is the middle most part of a ranked dataset.
- Mode is based on the principle of majority, and is the most frequent observation.
- A.M. $\geq$ G. M. $\geq$ H.M.
- If A.M. = median, then data are symmetric, otherwise asymmetric.
- Mode can be located and estimated graphically using histogram.
- Median and quartiles can be located and estimated graphically using ogive.
- A measure of dispersion refers to the amount of variation / scatter in a data.
- The higher the variation/ dispersion in a data, lesser is the stability/ consistency.
- Range measures variation in data by difference of two extreme observations.
- The sum of all deviations about mean in an ungrouped data is zero.
- Variance is the A.M. of squared deviations of data about A.M.
- Mean/absolute deviation is the A.M. of squared deviations of data about A.M.