

CHAPTER - 2

2. Summarization of Data

2.1 Measures of Central Tendency

The most important objective of a statistical analysis is to determine **a single value** for the entire mass of data, which describes the **overall** level of the group of observations and can be called a **representative** of the whole set of data. It tells us where the **center** of the distribution of data is located. The **most** commonly used measures of central tendencies are :

- ✓ The Mean (Arithmetic mean, Weighted mean, Geometric mean and Harmonic means)
- ✓ The Mode
- ✓ The Median

The Summation Notation:

- Let $X_1, X_2, \dots, X_N$ be the number of measurements where N is the total number of observation and X_i , is the i^{th} observation.
- Very often in statistics an algebraic expression of the form $X_1 + X_2 + \dots + X_N$ is used in a formula to compute a **statistic**. It is tedious to write an expression like this very often, so mathematicians have developed a shorthand notation to represent a sum of scores, called the **summation** notation.
- The symbol $\sum_{i=1}^N X_i$ is a mathematical shorthand for $X_1 + X_2 + \dots + X_N$.

Therefore, $\sum_{i=1}^N X_i = X_1 + X_2 + \dots + X_N \rightarrow \text{for the populations.}$

$\sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_n \rightarrow \text{for the samples.}$

Example: Suppose the following were scores (marks) made on the first assignment for five students in the class: 5, 7, 7, 6, and 8. Write their marks using summation notation.

Solution: $\sum_{i=1}^5 X_i = X_1 + X_2 + \dots + X_5 = 5 + 7 + 7 + 6 + 8 = 33$

Properties of summation

1. $\sum_{i=1}^n c = nc$ for any constant c .
2. $\sum_{i=1}^n cX_i = c \sum_{i=1}^n X_i$
3. $\sum_{i=1}^n (X_i \pm c) = \sum_{i=1}^n X_i \pm nc$
4. $\sum_{i=1}^n (Y_i + X_i) = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$
5. $\sum_{i=1}^n X_i Y_i = X_1 Y_1 + X_2 Y_2 + \dots + X_n Y_n$
6. $1 + 2 + 3 + 4 + \dots + n = \frac{n(n+1)}{2}$
7. $1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$

2.2 Types of Measure of Central Tendency

2.2.1 The Mean

2.2.1.1 Arithmetic mean

The arithmetic mean of a sample is the sum of all observations divided by the number of observations in the sample. i.e.

Sample mean or arithmetic mean = $\frac{\text{the sum of all values in the sample}}{\text{number of values in the sample}}$

Suppose that $x_1, x_2, \dots, x_n$ are n observed values in a sample of size n taken from a population of size N . Then the **arithmetic mean of the sample**, denoted by $\bar{x}$, is given by

$$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} \rightarrow \text{for samples.}$$

If we take an entire population, the **population mean** denoted by **μ** is given by

$$\mu = \frac{X_1+X_2+\cdots+X_N}{N} = \frac{\sum_{i=1}^N X_i}{N} \rightarrow \text{for populations.}$$

In general, the sample arithmetic mean is calculated by

$$\bar{X} = \begin{cases} \sum_{i=1}^n \frac{X_i}{n} & \rightarrow \text{for raw data} \\ \sum_{i=1}^k \frac{f_i X_i}{\sum f_i} & \rightarrow \text{for ungrouped data where } \sum f_i = n. \\ \sum_{i=1}^k \frac{M_i X_i}{\sum f_i} & \rightarrow \text{for grouped data where } \sum f_i = n. \end{cases}$$

Example 1: The net weights of five perfume bottles selected at random from the production line are 85.4, 85.3, 84.9, 85.4 and 85. What is the arithmetic mean weight of the sample observation?

Solution; Given $n = 5$ $x_1 = 85.4, x_2 = 85.3, x_3 = 84.9, x_4 = 85.4$ and $x_5 = 85$.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{85.4+85.3+84.9+85.4+85}{5} = \frac{426.6}{5} = 85.32.$$

Example 2: Calculate the mean of the marks of 46 students given below;

Marks (X_i)	9	10	11	12	13	14	15	16	17	18
Frequency (f_i)	1	2	3	6	10	11	7	3	2	1

Solution: $\sum f_i = n = 46$ is the sum of the frequencies or total number of observations.

To calculate $\sum_{i=1}^k f_i X_i$ consider the following table.

X_i	9	10	11	12	13	14	15	16	17	18	Total
f_i	1	2	3	6	10	11	7	3	2	1	46
$f_i X_i$	9	20	33	72	130	154	105	48	34	18	623

So $\bar{X} = \frac{\sum_{i=1}^k f_i X_i}{\sum f_i} = \frac{623}{46} = 13.54.$

Example 3: The net income of a sample of large importers of Urea was organized into the following table. What is the arithmetic mean of net income?

Net income	2-4	5-7	8-10	11-13	14-16
Number of importers	1	4	10	3	2

Solution: $\sum f_i = n = 20$ is the sum of the frequencies or total number of observations.

To calculate $\sum_{i=1}^k f_i m_i$ consider the following table.

Net income (CI)	2-4	5-7	8-10	11-13	14-16	Total
Number of importers (f_i)	1	4	10	3	2	20
Class marks (m_i)	3	6	9	12	15	
$f_i m_i$	3	24	90	36	30	183

So $\bar{X} = \frac{\sum_{i=1}^k f_i m_i}{\sum f_i} = \frac{183}{20} = 9.15.$

Example 4: From the following data, calculate the missing frequency? The mean number of tablets to cure ever was 29.18.

Number of tablets	19 – 21	22 – 24	25 – 27	28 – 30	31 – 33	34 – 36	37 – 39
Number of persons cured	6	13	19	f_4	18	12	9

Solution; $\sum f_i = n = 77 + f_4$ is the sum of the frequencies or total number of observations.

To calculate $\sum_{i=1}^k f_i m_i$ consider the following table.

CI	19 – 21	22 – 24	25 – 27	28 – 30	31 – 33	34 – 36	37 – 39	Total
f_i	6	13	19	f_4	18	12	9	$77 + f_4$
m_i	20	23	26	29	32	35	38	
$f_i m_i$	120	299	494	$29f_4$	576	420	342	$2251 + 29f_4$

Since mean is given to be 29.18 we have

$$\bar{X} = \sum_{i=1}^k \frac{f_i m_i}{\sum f_i} \Rightarrow 29.18 = \frac{2251 + 29f_4}{77 + f_4} \Rightarrow 29.18(77 + f_4) = 2251 + 29f_4$$

$$\Rightarrow 29.18f_4 - 29f_4 = 2251 - 2246.86 \Rightarrow 0.18f_4 = 4.14 \Rightarrow f_4 = \frac{4.14}{0.18} = 23.$$

Combined mean

If we have an arithmetic means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n$ of n groups having the same unit of measurement of a variable, with sizes $n_1, n_2, \dots, n_n$ observations respectively, we can compute the **combined mean** of the variant values of the groups **taken together** from the individual means by

$$\bar{X}_{com} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_n \bar{x}_n}{n_1 + n_2 + \dots + n_n} = \frac{\sum_{i=1}^n n_i \bar{x}_i}{\sum_{i=1}^n n_i}$$

Example 1: Compute the **combined** mean for the following two sets.

Set A: 1, 4, 12, 2, 8 and 6 ; **Set B:** 3, 6, 2, 7 and 4.

Solution: $n_1 = 6, \bar{x}_1 = \frac{\sum_{i=1}^6 X_i}{n_1} = \frac{33}{6} = 5.5$; $n_2 = 5, \bar{x}_2 = \frac{\sum_{i=1}^5 X_i}{n_2} = \frac{22}{5} = 4.4.$

$$\bar{X}_{com} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{6 \times 5.5 + 5 \times 4.4}{6 + 5} = \frac{55}{11} = 5.$$

Example 2: The mean weight of 150 students in a certain class is 60 kg. The mean weight of boys in the class is 70 kg and that of girl's is 55 kg . Find the number of boys and girls in the class?

Solution; Let n_1 be the number of boys and n_2 be the number of girls in the class.

Also let $\bar{x}_1, \bar{x}_2$ and $\bar{x}_{com}$ be the mean weights of boys, girls and the mean weights of all students respectively. Then $\bar{x}_1 = 70, \bar{x}_2 = 55$ and $\bar{x}_{com} = 60$. Clearly $n_1 + n_2 = 150$.

$$\bar{X}_{com} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \Rightarrow 60 = \frac{70 n_1 + 55 n_2}{n_1 + n_2} \Rightarrow 60 = \frac{70 n_1 + 55 n_2}{150}$$

$$\Rightarrow 9000 = 70 n_1 + 55 n_2 \quad \dots \dots (1) \text{ and}$$

$$150 = n_1 + n_2 \quad \dots \dots (2)$$

Solving equation (1) and (2) simultaneously, we get $n_1 = 50$ and $n_2 = 100$.

Disadvantages of the arithmetic mean

1. The mean is meaningless in the case of nominal or qualitative data.
2. In case of grouped data, if any class interval is open ended, arithmetic mean cannot be calculated, since the class mark of this interval cannot be found.

2.2.1.2 Weighted mean

In the computation of arithmetic mean, we had given an **equal importance** to each observation. Sometimes the individual values in the data **may not have an equally importance**. When this is the case, we assigned to each weight which is **proportional** to its relative importance.

- The weighted mean of a set of values $x_1, x_2, \dots, x_n$ with corresponding weights $w_1, w_2, \dots, w_n$ denoted by $\bar{x}_w$ is computed by:

$$\overline{X}_w = \frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

The calculation of cumulative grade point average (**CGPA**) in Colleges and Universities is a good example of *weighted mean*.

Example: If a student scores "A" in a **3 EtCTS** course, "B" in a **6 EtCTS** course, "B" in another **5 EtCTS** course and "D" in a **2 EtCTS** course. Compute his /her GPA for the semester.

Solution: Here the numerical values of the letter grades are the values (i.e. $A = 4, B = 3, C = 2$ and $D = 1$) and the corresponding EtCTS of the course are their respective weights. i.e.

Grade values (x_i)	4	3	3	1
Weight (w_i)	3	6	5	2

$$GPA = \overline{X}_w = \frac{w_1x_1+w_2x_2+\dots+w_nx_n}{w_1+w_2+\dots+w_n} = \frac{\sum_{i=1}^4 w_i x_i}{\sum_{i=1}^4 w_i} = \frac{4 \times 3 + 3 \times 6 + 3 \times 5 + 1 \times 2}{3 + 6 + 5 + 2} = \frac{12 + 18 + 15 + 2}{16} = \frac{47}{16} = 2.9375.$$

2.2.1.3 Geometric mean

In algebra geometric mean is calculated in the case of geometric progression, but in statistics we need not bother about the progression, here it is particular type of data for which the geometric mean is of great importance because it gives a good mean value. If the observed values are measured as **ratios, proportions or percentages**, then the geometric mean gives a better measure of central tendency than any other means.

- The Geometrical mean of a set of values $x_1, x_2, \dots, x_n$ of n positive values is defined as the n^{th} root of their product . That is,

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

Example: The G.M of 4, 8 and 6 is

$$G.M = \sqrt[3]{4 \times 8 \times 6} = \sqrt[3]{192} = 5.77.$$

In general, the sample geometric mean is calculated by

$$G.m = \begin{cases} \sqrt[n]{x_1 * x_2 * \dots * x_n} & \rightarrow \text{for raw data} \\ \sqrt[n]{x_1^{f_1} * x_2^{f_2} *, \dots, * x_k^{f_k}} & \rightarrow \text{for ungrouped data where } n = \sum f_i . \\ \sqrt[n]{m_1^{f_1} * m_2^{f_2} *, \dots, * m_k^{f_k}} & \rightarrow \text{for grouped data where } n = \sum f_i . \end{cases}$$

Example1: The man gets three annual raises in his salary. At the end of first year, he gets an increase of 4%, at the end of the second year, he gets an increase of 6% and at the end of the third year, he gets an increase of 9% of his salary. What is the average percentage increase in the three periods?

$$\textbf{Solution: } G.M = \sqrt[3]{1.04 * 1.06 * 1.09} = 1.0631 \quad \Rightarrow \quad 1.0631 - 1 = 0.0631.$$

Therefore, the average percentage increase is 6.31%.

Example 2: Compute the Geometric mean of the following data.

Values	2	4	6	8	10
Frequency	1	2	2	2	1

Solution: $g.m = \sqrt[8]{2^1 * 4^2 * 6^2 * 8^2 * 10^1} = \sqrt[8]{2 * 16 * 36 * 64 * 100} = \sqrt[8]{737280} = 5.41.$

Example 3: Suppose that the profits earned by a certain construction company in four projects were 3%, 2%, 4% & 6% respectively. What is the Geometric mean profit?

Solution: $g.m = \sqrt[4]{3 * 2 * 4 * 6} = \sqrt[4]{144} = 3.46.$

Therefore; the geometric mean profit is 3.46 percent.

2.2.1.4 Harmonic mean

Another important mean is the harmonic mean, which is suitable measure of central tendency when the data pertains to **speed, rates and price**.

➤ Let $x_1, x_2, \dots, x_n$ be n variant values in a set of observations, then simple harmonic mean is given by:

$$S.H.M = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

✓ **Note:** SHM is used for equal distances, equal costs and equal rates.

Example 1: A motorist travels for three days at a rate (speed) of 480 km/day. On the first day he travels 10 hours at a rate of 48 km/h, on the second day 12 hours at a rate of 40 km/h, on the third day 15 hours at a rate of 32 km/h. What is the average speed?

Solution: Since the distance covered by the motorist is equal (i.e. $s_1 = 480, s_2 = 480, s_3 = 480$), so we use SHM.

$$S.H.M = \frac{3}{\frac{1}{48} + \frac{1}{40} + \frac{1}{32}} = 38.92 \quad \text{so the required average speed} = 38.92 \text{ km/h.}$$

We can check this, by using the known formula for average speed in elementary physics.

$$\begin{aligned} \text{Check; Average speed } (V_{av}) &= \frac{\text{total distance covered}}{\text{total time taken}} = \frac{S_T}{t_T} \\ &= \frac{480\text{km} + 480\text{km} + 480\text{km}}{10\text{hr} + 12\text{hr} + 15\text{hr}} = \frac{1440\text{km}}{37\text{hr}} = 38.92 \text{ km/h.} \end{aligned}$$

Example 2: A business man spent 20 Birr for milk at 40 cents per liter in Mizan-Aman town and another 20 Birr at 60 cents per liter in Tepi town. What is the average price of milk at two towns.

Solution: Since the price on the two towns are equal (20 Birr), so we use SHM.

$$\text{Average price } (p_{av}) = S.H.M = \frac{2}{\frac{1}{40} + \frac{1}{60}} = 48 \text{ cents/liter.}$$

Weighted harmonic mean (WHM)

✓ WHM is used for different distance, different cost and different rate.

$$W.H.M = \frac{\sum w_i}{\sum \frac{w_i}{x_i}}$$

Example 1: A driver travel for 3 days. On the 1st day he drives for 10h at a speed of 48 km/h, on the 2nd day for 12h at 45 km/h and on the 3rd day for 15h at 40 km/h. What is the average speed?

Solution: since the distance covered by the driver is not equal, so we use WHM by taking the distance as weights (w_i).

$$v_{av} = w.h.m = \frac{\sum w_i}{\sum \frac{w_i}{x_i}} = \frac{(480 + 540 + 600)\text{km}}{\left(\frac{480}{40} + \frac{540}{45} + \frac{600}{40}\right)h_r} = 43.32 \text{ km/hr.}$$

Example 2: A business man spent 20 Birr for milk at rate of 40 cents per liter in Mizan-Aman town and 25 Birr at a price of 50 cents per liter in Tepi town. What is the average price ?

Solution: Since the price on the two towns are different , so we use *WHM* by taking the cost as weights (w_i).

$$p_{av} = w.h.m = \frac{\sum w_i}{\sum \frac{w_i}{x_i}} = \frac{(20 + 25) \text{ birr}}{\left(\frac{20}{40} + \frac{25}{50}\right) \text{ birr l/c}} = 45 \text{ c/l.}$$

✓ (Finally If all the observations are positive) $A.M \geq G.M \geq H.M$.

Corrected mean

$$\bar{x}_{corr} = \bar{x}_w + \frac{c - w}{n} \quad \text{where } \bar{x}_w \text{ is the wrong mean}$$

✓ c is the sum of correct values and w is the sum of wrong values.

Example: The mean age of a group of 100 persons was found to be 32.02 years. Later on, it was discovered that age of 57 was misread as 27. Find the corrected mean?

Solution: $n = 100$, $\bar{x}_w = 32.02$, $c = 57$ and $w = 27$.

$$\bar{x}_{corr} = \bar{x}_w + \frac{c - w}{n} = 32.02 + \frac{57 - 27}{100} = 32.02 + 0.3 = 32.32 \text{ years.}$$

Median and mode

2.2.2 The Median

Suppose we sort all the observations in numerical order, ranging from smallest to largest or vice versa. Then the median is the middle value in the sorted list. We denote it by $\tilde{x}$.

✚ Let $x_1, x_2, \dots, x_n$ be n **ordered** observations. Then the median is given by:

$$\tilde{x} = \begin{cases} X_{\frac{n+1}{2}} & \text{If } n \text{ is odd.} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} & \text{If } n \text{ is even.} \end{cases}$$

Example 1: Find the median for the following data.

23, 16, 31, 77, 21, 14, 32, 6, 155, 9, 36, 24, 5, 27, 19

Solution: First arrange the given data in increasing order. That is

5, 6, 9, 14, 16, 19, 21, 23, 24, 27, 31, 32, 36, 77, and 155.

$$n = 15 \Rightarrow \text{odd}, \quad \tilde{x} = X_{\frac{n+1}{2}} = X_{\frac{15+1}{2}} = X_8 = 23$$

Example 2: Find the median for the following data.

61, 62, 63, 64, 64, 60, 65, 61, 63, 64, 65, 66, 64, 63

Solution: First arrange the given data in increasing order. that is

60, 61, 61, 62, 63, 63, 63, 64, 64, 64, 65, 65 & 66.

$$n = 14 \Rightarrow \text{even}, \quad \tilde{x} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{\frac{14}{2}} + X_{\frac{14}{2}+1}}{2} = \frac{X_7 + X_8}{2} = \frac{63 + 64}{2} = \frac{127}{2} = 63.5$$

Median for ungrouped data

➤ Let $x_1, x_2, \dots, x_k$ have their corresponding frequencies $f_1, f_2, \dots, f_k$ then to find the median:

✓ First sort the data in ascending order.

✓ Construct the less than cumulative frequency (*lcf*) .

✓ If $n = \sum_{i=1}^k f_i$ is odd, find $\frac{n+1}{2}$ and search the **smallest lcf** which is $\geq \frac{n+1}{2}$. Then the **variant value** corresponding to this **lcf** is the median.

✓ If n is even, find $\frac{n}{2}$ & $\frac{n}{2} + 1$ and search the **smallest lcf** which is $\geq \left(\frac{n}{2}\right)$ & $\left(\frac{n}{2} + 1\right)$.

Then the **average of the variant values** corresponding to these **lcf** is the median.

Example 1: Find the median for the following data.

Values (x_i)	3	5	4	2	7	6
Frequency (f_i)	2	1	3	2	1	1

Solution: First arrange the data in increasing order and construct the lcf table for this data.

Values (x_i)	2	3	4	5	6	7
Frequency (f_i)	2	2	3	1	1	1
Lcf	2	4	7	8	9	10

$n = 10 \Rightarrow even$. So $\frac{n}{2} = \frac{10}{2} = 5$ and $\frac{n}{2} + 1 = \frac{10}{2} + 1 = 5 + 1 = 6$.
Then the smallest LCF which is ≥ 5 & 6 is 7 and the variant value corresponding to this LCF is 4. Thus the median is $\tilde{x} = \frac{4+4}{2} = 4$.

Example 2: Calculate the median of the marks of 46 students given below.

Values (x_i)	10	9	11	12	14	13	15	16	17	18
Frequency (f_i)	2	1	3	6	10	11	7	3	2	1

Solution: First arrange the data in ascending order and construct the LCF table for this data.

Values (x_i)	9	10	11	12	13	14	15	16	17	18
Frequency (f_i)	1	2	3	6	11	10	7	3	2	1
LCF	1	3	6	12	23	33	40	43	45	46

$n = 46 \Rightarrow even$. So $\frac{n}{2} = \frac{46}{2} = 23$ and $\frac{n}{2} + 1 = \frac{46}{2} + 1 = 23 + 1 = 24$.
The smallest LCF ≥ 23 & 24 are 23 & 33 respectively and the variant values corresponding to these LCF are 13 & 14 respectively. Thus the median is $\tilde{x} = \frac{13+14}{2} = 13.5$.

Median for grouped data

The formula for computing the median for grouped data is given by

$$median = \tilde{x} = lcb_{\tilde{x}} + \frac{\left(\frac{n}{2} - lcf_p\right) \times w}{f_m}$$

Where: $lcb_{\tilde{x}}$ – is the **lcb** of the median class.

- ✓ n – is the total number of observations.
- ✓ f_m is the frequency of the median class.
- ✓ w is the width of the median class. recall : $w = ucb - lcb$.
- ✓ lcf_p is the **lcf** corresponding to the class **immediately preceding** the median class.

- **Note:** The class corresponding to the **smallest LCF** which is $\geq \frac{n}{2}$ is called the **median class**. So that the median lies in this class.

Steps to calculate the median for grouped data

1. First construct the LCF table.
2. Determine the median class. To determine the median class, find $\frac{n}{2}$ and search the **smallest LCF** which is $\geq \frac{n}{2}$. Then the class corresponding to this **lcf** is the **median class**.

Example 1: Find the median for the following data.

Daily production	80 – 89	90 – 99	100 – 109	110 – 119	120 – 129	130 – 139
Frequency	5	9	20	8	6	2

Solution: First construct the LCF table.

Daily production(CI)	80 – 89	90 – 99	100 – 109	110 – 119	120 – 129	130 – 139
Frequency(fi)	5	9	20	8	6	2
Lcf	5	14	34	42	48	50

To obtain the median class , calculate $\frac{n}{2} = \frac{50}{2} = 25$. Thus the smallest lcf which is $\geq \frac{n}{2}$ is 34. So the class corresponding to this lcf is 100 – 109, *is the median class*.

Therefore, $lcb_{\tilde{x}} = 99.5$, $w = 10$, $f_m = 20$, $lcf_p = 14$.

$$median = \tilde{x} = lcb_{\tilde{x}} + \frac{\left(\frac{n}{2} - lcf_p\right) \times w}{f_m} = 99.5 + \frac{(25 - 14) \times 10}{20} = 105.$$

Properties of the median

1. The median is unique.
2. It can be computed for an open ended frequency distribution if the median does not lie in an open ended class.
3. It is not affected by extremely large or small values .
4. It is not so suitable for algebraic manipulations.
5. It can be computed for ratio level, interval level and ordinal level data.

2.2.3 The mode

In every day speech, something is “*in the mode*” if it is fashionable or popular. In statistics this “*popularity*” refers to frequency of observations.

Therefore, **mode** is the *most frequently* observed value in a set of observations.

Example: *Set A:* 10, 10, 9, 8, 5, 4, 5, 12, 10 *mode* = 10 → *unimodal*.

Set B: 10, 10, 9, 9, 8, 12, 15, 5 *mode* = 9 & 10 → *bimodal*.

Set C: 4, 6, 7, 15, 12, 9 *no mode*.

Remark: In a set of observed values, all values occur *once* or *equal* number of times, there is no mode. (See set C above).

Mode for a grouped data

If the data is grouped such that we are given frequency distribution of finite class intervals, we do not know the value of every item, but we easily determine the class with **highest** frequency. Therefore, the **modal class** is the class with the **highest** frequency. So that the mode of the distribution lies in this class.

➤ To compute the mode for a grouped data we use the formula:

$$mode = \hat{X} = lcb_{\hat{x}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times w$$

where; $\Delta_1 = f_m - f_p$, $\Delta_2 = f_m - f_s$

Where: $lcb_{\hat{x}}$ – is the **lcb** of the modal class.

f_m – is the frequency of the modal class.

f_p – is the frequency of the class **preceding** the modal class.
 f_s – is the frequency of the class **succeeding** the modal class.
 w – is the width of the modal class.

Example 1: The ages of newly hired, unskilled employees are grouped into the following distribution. Then compute the modal age?

Ages	18 – 20	21 – 23	24 – 26	27 – 29	30 – 32
Number	4	8	11	20	7

Solution: First we determine the modal class. The modal class is 27 – 29, since it has the highest frequency. Thus, $lcb_{\hat{x}} = 26.5$, $w = 3$, $\Delta_1 = 20 - 11 = 9$, $\Delta_2 = 20 - 7 = 13$.

$$\hat{X} = lcb_{\hat{x}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times w = 26.5 + \left(\frac{9}{9 + 13}\right) \times 3 = 26.5 + \left(\frac{27}{22}\right) = 26.5 + 1.2 = \mathbf{27.7}$$

Interpretation: The age of most of these newly hired employees is 27.7 (27 years and 7 months).

Example 2: The following table shows the distribution of a group of families according to their expenditure per week. The **median and the mode** of the following distribution are known to be 25.50 Birr and 24.50 Birr respectively. Two frequency values are however missing from the table. Find the missing frequencies.

Class interval	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
Frequency	14	f_2	27	f_4	15

Solution: The LCF table of the given distribution can be formed as follows.

Expenditure (CI)	1 – 10	11 – 20	21 – 30	31 – 40	41 – 50
Number of families (f_i)	14	f_2	27	f_4	15
LCF	14	$14 + f_2$	$41 + f_2$	$41 + f_2 + f_4$	$56 + f_2 + f_4$

Here: $n = 56 + f_2 + f_4$. Since the median and the mode are Birr 25.5 & 24.5 respectively then the class 21 – 30 is the **median class** as well as the **modal class**.

$$25.5 = 20.5 + \frac{\left(\frac{56+f_2+f_4}{2} - (14+f_2)\right) \times 10}{27} \tag{i}$$

$$24.5 = 20.5 + \frac{(27-f_2) \times 10}{(27-f_2)+(27-f_4)} \tag{ii}$$

Eqn. (i) and eqn. (ii) can be written as

$$5 = \frac{5 \times (56+f_2+f_4) - 10 \times (14+f_2)}{27} \text{ \& \& } 4 = \frac{(27-f_2) \times 10}{54-f_2-f_4}$$

Further simplifying the above we get

$$f_2 - f_4 = 1. \tag{iii} \qquad \&$$

$$3f_2 - 2f_4 = 27. \tag{iv}$$

Solving (iii) & (IV), we get the values of f_2 & f_4 as $f_2 = 25$ & $f_4 = 24$.

Properties of mode

1. it is not affected by extreme values.
2. It can be calculated for distribution with open ended classes.
3. It can be computed for all levels of data i.e. nominal, ordinal, interval and ratio.

4. The main drawback of mode is that often it does not exist.
5. Often its values are not unique.

2.3 Measure of non - central location (Quintiles')

There are three types of quintiles. These are:

1. Quartiles

The quartiles are the three points, which divide a given order data into *four equal* parts. These

$$Q_i = \frac{i \times (n+1)^{th}}{4}, \quad i = 1, 2, 3. \rightarrow \text{for raw data.}$$

Q_1 is the value corresponding to $(\frac{n+1}{4})^{th}$ ordered observation.

Q_2 is the value corresponding to $2 \times (\frac{n+1}{4})^{th}$ ordered observation.

Q_3 is the value corresponding to $3 \times (\frac{n+1}{4})^{th}$ ordered observation.

Example: Consider the age data given below and calculate Q_1 , Q_2 , and Q_3 .

19, 20, 22, 22, 17, 22, 20, 23, 17, 18

Solution: First arrange the data in ascending order, $n=10$.

17, 17, 18, 19, 20, 20, 22, 22, 22, 23

$$\begin{aligned} \text{✚ } Q_1 &= \left(\frac{n+1}{4}\right)^{th} = \left(\frac{10+1}{4}\right)^{th} = (2.75)^{th} \text{ observation} = 2^{nd} \text{ observation} + 0.75 (3^{rd} - 2^{nd}) \\ &\text{observation} = 17 + 0.75(18 - 17) = 17.75 \end{aligned}$$

Therefore 25% of the observations are below 17.75

$$\begin{aligned} \text{✚ } Q_2 &= 2 \times \left(\frac{n+1}{4}\right)^{th} = 2 \times \left(\frac{10+1}{4}\right)^{th} = (5.5)^{th} \text{ observation} = 5^{th} + 0.5(6^{th} - 5^{th}) = 20 + \\ &0.5(20 - 20) = 20. \end{aligned}$$

$$\begin{aligned} \text{✚ } Q_3 &= 3 \times \left(\frac{n+1}{4}\right)^{th} = 3 \times \left(\frac{10+1}{4}\right)^{th} = (8.25)^{th} \text{ observation} = 8^{th} + 0.25 \times (9^{th} - 8^{th}) = 22 + 0.25 \times (22 - 22) = 22 \end{aligned}$$

Calculation of quartiles for grouped data

- For the grouped data, the computations of the three quartiles can be done as follows:
- ✓ Calculate $\frac{ixn}{4}$ and search the minimum lcf which is $\geq \frac{ixn}{4}$, for $i = 1, 2, 3$.

The class corresponding to this **lcf** is called the **i^{th} quartile class**. This is the class where Q_i lies.

✚ The unique value of the i^{th} quartile (Q_i) is then calculated by the formula

$$Q_i = lcb_{q_i} + \frac{\left(\frac{i \times n}{4} - lcf_p\right) \times w}{f_{q_i}}, \quad \text{for } i = 1, 2, 3.$$

Where: lcb_{q_i} – is the **lcb** of the i^{th} quartile class.

✓ f_{q_i} – is the frequency of the i^{th} quartile class.

✓ lcf_p is the lcf corresponding to the class **immediately preceding** the i^{th} quartile class.

Note: $Q_2 = \text{median}$

2. Percentiles (P)

Percentiles are 99 points, which divide a given ordered data into 100 equal parts. These

$$P_m = \frac{m \times (n + 1)^{th}}{100}, \quad m = 1, 2, \dots, 99. \rightarrow \text{for raw data.}$$

Calculation of percentiles for grouped data

For the grouped data, the computations of the 99 percentiles can be done as follows:

- ✓ Calculate $\frac{mxn}{100}$ and search the **minimum lcf** which is $\geq \frac{mxn}{100}$, for $m = 1, 2, \dots, 99$.

The class corresponding to this **lcf** is called the **m^{th} percentile class**. This is the class where P_m lies.

- ✚ The unique value of the m^{th} percentile (P_m) is then calculated by the formula

$$P_m = lcb_{p_m} + \frac{\left(\frac{mxn}{100} - lcf_p\right) \times w}{f_{p_m}}, \quad \text{for } m = 1, 2, \dots, 99.$$

Where: lcb_{p_m} , f_{p_m} and lcf_p will have a similar interpretation as in quartiles.

3. Deciles (D)

Deciles are the nine points, which divide the given ordered data into 10 equal parts.

$$D_k = \frac{k \times (n + 1)^{\text{th}}}{10}, \quad k = 1, 2, \dots, 9.$$

For the grouped data, the computations of the 9 deciles can be done as follows:

- ✓ Calculate $\frac{kxn}{10}$ and search the **minimum lcf** which is $\geq \frac{kxn}{10}$, $k = 1, 2, \dots, 9$.

The class corresponding to this **lcf** is called the **k^{th} decile class**. This is the class where D_k lies.

- ✚ The unique value of the k^{th} decile (D_k) is calculated by the formula

$$D_k = lcb_{D_k} + \frac{\left(\frac{kxn}{10} - lcf_p\right) \times w}{f_{D_k}}, \quad \text{for } k = 1, 2, \dots, 9.$$

Where: lcb_{D_k} , f_{D_k} and lcf_p will have a similar interpretation as in quartiles and percentiles.

Note that: median = $Q_2 = D_5 = P_{50}$ and $D_1, D_2, \dots, D_9$ correspond to $P_{10}, P_{20}, \dots, P_{90}$

Q_1, Q_2 and Q_3 corresponding to P_{25}, P_{50} and P_{75} .

Example: For the following FD data, find

- a) Q_1, Q_2 and Q_3 b) P_{25}, P_{30}, P_{50} and P_{75} c) D_1, D_2, D_3 and D_5

interval	21 – 22	23 – 24	25 – 26	27 – 28	29 – 30
F	10	22	20	14	14

Solution: First find the lcf table

interval	21 – 22	23 – 24	25 – 26	27 – 28	29 – 30	total
F	10	22	20	14	14	80
Lcf	10	32	52	66	80	

- a) $Q_1 = ?$ $\frac{n}{4} = \frac{80}{4} = 20$. Thus, the minimum **lcf** just ≥ 20 is 32 so the class corresponding to this **lcf** is 23 – 24, is the **first quartile class**. $lcb_{q_1} = 22.5$, $w = 2$, $f_{q_1} = 22$, $lcf_p = 10$.

$$Q_1 = lcb_{q_1} + \frac{\left(\frac{n}{4} - lcf_p\right) \times w}{f_{q_1}} = 22.5 + \frac{(20 - 10) \times 2}{22} = 23.41.$$

- $Q_2 = ?$ $\frac{2n}{4} = \frac{160}{4} = 40$. Thus, the minimum **lcf** just ≥ 40 is 52 so the class corresponding to this **lcf** is 25 – 26, is the **second** quartile class. $lcb_{q_2} = 24.5$, $w = 2$, $f_{q_2} = 20$, $lcf_p = 32$.

$$Q_2 = lcb_{q_2} + \frac{\left(\frac{2n}{4} - lcf_p\right) \times w}{f_{q_2}} = 24.5 + \frac{(40 - 32) \times 2}{20} = 25.3.$$

$$Q_3 = 27.64.$$

b) $P_{25} = ? \frac{25 \times n}{100} = \frac{25 \times 80}{100} = 20$. Thus, the minimum **lcf** just ≥ 20 is 32 so the class corresponding to this **lcf** is 23 – 24, is the **25th** percentile class.

Thus, $\text{lcb}_{p_{25}} = 22.5, w = 2, f_{p_{25}} = 22, \text{lcf}_p = 10$.

$$p_{25} = \text{lcb}_{p_{25}} + \frac{\left(\frac{25 \times n}{100} - \text{lcf}_p\right) \times w}{f_{p_{25}}} = 22.5 + \frac{(20 - 10) \times 2}{22} = 23.41.$$

$$p_{20} = 23.045.$$

$$p_{30} = 23.77.$$

$$p_{50} = 25.3.$$

$$p_{75} = 27.64.$$

C) $D_1 = ? \frac{1 \times n}{10} = \frac{80}{10} = 8$. Thus, the **minimum lcf** just ≥ 8 is 10 so the class corresponding to this **lcf** is 21 – 22, is the **first** decile class. Thus, $\text{lcb}_{D_1} = 20.5, w = 2, f_{D_1} = 10, \text{lcf}_p = 0$.

$$D_1 = \text{lcb}_{D_1} + \frac{\left(\frac{1 \times n}{10} - \text{lcf}_p\right) \times w}{f_{D_1}} = 20.5 + \frac{(8 - 0) \times 2}{10} = 22.1 \Rightarrow \text{upto class boundary.}$$

$$D_2 = 23.045.$$

$$D_3 = 23.77.$$

$$D_5 = 25.3.$$

$\therefore Q_1 = P_{25}, Q_2 = P_{50}$ and $Q_3 = P_{75}$ $D_1 = P_{10}, D_2 = P_{20}, D_3 = P_{30}$ and $D_5 = P_{50}$
and median = $Q_2 = D_5 = P_{50}$

2.4. Measures of variation (dispersion)

Measures of central tendency locate the **center** of the distribution. But they **do not** tell how individual observations are **scattered** on either side of the center. **The spread** of observations around the **center** is known as **dispersion or variability**. In other words; the degree to which numerical data tend to spread about an **average value** is called dispersion or variation of the data.

➤ **Small** dispersion indicates **high** uniformity of the observation while **larger** dispersion indicates **less** uniformity.

Types of Measures of Dispersion

The **most** commonly used measures of dispersions are:

1. The Range (R)

The Range is the difference b/n the highest and the smallest observation. That is;

$$R = \begin{cases} X_{\max} - X_{\min} & \rightarrow \text{for raw and ungrouped data.} \\ UCL_{\text{last}} - LCL_{\text{first}} & \rightarrow \text{for grouped data.} \end{cases}$$

It is a **quick and dirty** measure of variability. Because of the range is greatly **affected** by extreme values, it may give a **distorted** picture of the scores.

Range is a measure of absolute dispersion and as such **cannot** be used for comparing variability of two distributions expressed in **different** units.

4. Variance and Standard Deviation

Variance: is the average of the squares of the deviations taken from the mean.

🌈 Suppose that $x_1, x_2, \dots, x_N$ be the set of observations on N populations. Then,

$$\text{Population variance} = \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N x_i^2 - N\mu^2}{N} \rightarrow \text{for population.}$$

$$\text{Sample variance} = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} \rightarrow \text{for sample.}$$

In general, the sample variance is computed by:

$$s^2 = \begin{cases} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} & \rightarrow \text{for raw data.} \\ \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\sum_{i=1}^k f_i - 1} = \frac{\sum_{i=1}^k f_i x_i^2 - n\bar{x}^2}{n-1} & \rightarrow \text{for ungrouped data.} \\ \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{\sum_{i=1}^k f_i - 1} = \frac{\sum_{i=1}^k f_i m_i^2 - n\bar{x}^2}{n-1} & \rightarrow \text{for grouped data.} \end{cases}$$

Standard Deviation: it is the square root of variance. Its advantage over the variance is that it is in the **same units** as the variable under the consideration. It is a measure of the average variation in a set of data. It is a measure of how far, on the average, an individual measurements is from the mean.

$$S.d = \sqrt{\text{variance}} = \sqrt{s^2} = S.$$

Example 1: Compute the variance for the sample: 5, 14, 2, 2 and 17.

Solution: $n = 5$, $\sum_{i=1}^n x_i = 40$, $\bar{x} = 8$, $\sum_{i=1}^n x_i^2 = 518$.

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{518 - 5 \times 8^2}{5-1} = 49.5, \quad S = \sqrt{49.5} = 7.04.$$

Example 2: Suppose the data given below indicates time in minute required for a laboratory experiment to compute a certain laboratory test. Calculate the mean, variance and standard deviation for the following data.

x_i	32	36	40	44	48	Total
f_i	2	5	8	4	1	20
$f_i x_i$	64	180	320	176	48	788
$f_i x_i^2$	2048	6480	12800	7744	2304	31376

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{n} = \frac{788}{20} = 39.4, \quad s^2 = \frac{\sum_{i=1}^n f_i x_i^2 - n\bar{x}^2}{n-1} = \frac{31376 - 20 \times (39.4)^2}{19} = 17.31, \quad S = \sqrt{17.31} = 4.16.$$

Properties of Variance

1. The variance is always non-negative ($s^2 \geq 0$).
2. If every element of the data is multiplied by a constant "c", then the new variance $s_{new}^2 = c^2 \times s_{old}^2$.
3. When a constant is added to all elements of the data, then the variance does not change.
4. The variance of a constant (c) measured in **n times** is zero. i.e. ($\text{var}(c) = 0$).

Exercise: Verify the above properties.

Uses of the Variance and Standard Deviation

1. They can be used to determine the spread of the data. If the variance or S.D is large, then the data are more dispersed.
2. They are used to measure the consistency of a variable.
3. They are used quit often in inferential statistics.

5. Coefficient of Variation (C.V)

Whenever the two groups have the **same units** of measurement, the variance and S.D for each can be compared directly. A statistics that allows one to compare two groups when the units of measurement are **different** is called **coefficient of variation**. It is computed by:

$$C.V = \frac{\sigma}{\mu} \times 100\% \quad \rightarrow \text{for population.}$$

$$C.V = \frac{S}{\bar{x}} \times 100\% \quad \rightarrow \text{for sample.}$$

Example: The following data refers to the hemoglobin level for 5 males and 5 female students. In which case, the hemoglobin level has high variability (less consistency).

For males (x_i)	13	13.8	14.6	15.6	17
For females (x_i)	12	12.5	13.8	14.6	15.6

Solution: $\bar{x}_{male} = \frac{13+13.8+14.6+15.6+17}{5} = \frac{74}{5} = 14.8$, $\bar{x}_{female} = \frac{12+12.5+13.8+14.6+15.6}{5} = \frac{68}{5} = 13.7$.

$$s^2_{males} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = 2.44, \quad S_{males} = \sqrt{2.44} = 1.56205,$$

$$s^2_{females} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = 2.19, \quad S_{females} = \sqrt{2.19} = 1.479865.$$

$$C.V_{males} = \frac{S_{males}}{\bar{x}_{male}} \times 100\% = \frac{1.56205}{14.8} \times 100\% = 10.56\%,$$

$$C.V_{females} = \frac{S_{females}}{\bar{x}_{female}} \times 100\% = \frac{1.479865}{13.7} \times 100\% = 10.8\%.$$

Therefore, the variability in hemoglobin level is higher for females than for males.

6. Standard Scores (Z-Scores)

- ✓ It is used for describing the **relative position** of a **single score** in the entire set of data in terms of the mean and standard deviation.
- ✓ It is used to compare **two observations** coming from different groups.
 - If X is a measurement (an observation) from a distribution with mean $\bar{x}$ and standard deviation S , then its value in standard units is

$$Z = \frac{X - \mu}{\sigma} \rightarrow \text{for population.}$$

$$Z = \frac{x - \bar{x}}{S} \rightarrow \text{for sample.}$$

- Z gives the number of standard deviation a particular observation lie above or below the mean.
- A **positive** Z -score indicates that the observation is **above** the mean.
- A **negative** Z -score indicates that the observation is **below** the mean.

Example: Two sections were given an examination on a certain course. For section 1, the average mark (score) was 72 with standard deviation of 6 and for section 2, the average mark (score) was 85 with standard deviation of 7. If student A from section 1 scored 84 and student B from section 2 scored 90, then who perform a better relative to the group?

Solution: Z score for A is $Z = \frac{x - \bar{x}}{S} = \frac{84 - 72}{6} = 2$.

$$Z \text{ score for B is } Z = \frac{x - \bar{x}}{S} = \frac{90 - 85}{7} = 0.71.$$

Since $Z_A > Z_B$ i.e. $2 > 0.71$, student A performed better relative to his group than student B. Therefore, student A has performed **better** relative to his group because the score's of student A is two standard deviation above the mean score of section 1 while the score of student B is only 0.71 standard deviation above the mean score of students in section 2.