III

DESCRIBING DATA WITH NUMERICAL MEASURES

Introduction

So far, we have learned the techniques of data collection, and condensing and summarizing data in the form of frequency distribution table and presenting data in the form of different diagrams and graphs. Diagrams and graphs, discussed in the preceding chapter, are the powerful and effective media for presenting statistical data, they can only represent a limited amount of information, and they are not much helpful when intensive analysis of the data is required. Now, we shall deal with some arithmetic procedures that can be used for analyzing, interpreting quantitative data both for ungrouped and grouped data. These measures and procedures relate to some properties and characteristics of data, which include measures of central tendency, or location of data, measures of dispersion of data in itself and around some central values, and the shape characteristics of the data. Broadly speaking there are four important characteristics of a set of data or its frequency distribution. These are:

(a) Location or central tendency,
(b) Dispersion,
(c) Skewness, and
(d) Kurtosis.

In this chapter we shall discuss how to measure the first type of characteristics of a distribution. Other measures will be discussed in subsequent two chapters.

Measures of Location or Central Tendency

When we have a set of quantitative data it is observed that most of the values of the data set cluster around some central value. This tendency of a set of quantitative data is called central tendency. It is more or less a central value and one of the principal characteristics of a frequency distribution. An average is a measure of central value of a set of data whether it is from a sample or a population. For example, we often talk of average income, average hourly product of a firm, average weight, average age of employees etc. Thus an average is a single value, which is considered as the most representative or typical value for the respective set of data.

According to Simpson and Kafca "A measure of central tendency is a typical value around which other figures congregate."

The purpose of an average is to get one single value that describes the characteristics of the entire data numerically and facilitates comparison with other distribution of similar nature.

The most commonly used measures of central tendency or location are (i)Mean, (ii)Median, and (iii) Mode.

...ese measures have the same unit as that of the variable. For example, if the variable height is measured in centimeter, the mean or median will also be in centimeter.

Characteristics of a good measure of location or central tendency. *According Yule and Kendall, a good measure of central tendency should have the following characteristics:*

i) It should be easy to understand,
ii) It should be easy to compute,
iii) It should be rigidly defined,
iv) It should be based on all the observations,
v) It should be capable of further algebraic treatment,
vi) It should have sampling stability,
vii) It should not be affected by the presence of extreme values.

Arithmetic Mean for Ungrouped Data

It is the most widely used average in statistics. It is commonly known as mean.

Definition. The arithmetic mean is the total or sum of the values of a set of observations divided by the total number of observations.

Actually, it is the center of gravity of a set of observations. Now, we shall state some formulae for finding arithmetic mean in different situations.

We may get data from a population or from a sample. When we have population data, we can compute population mean by the following formula.

Population Mean. If $X_1, X_2, \ldots\ldots\ldots\ldots, X_N$ are N values of a finite population, then the population mean denoted by μ (mu) is a parameter defined by

$$\mu = \frac{\Sigma X_i}{N}.$$

Here N is the population size or the total number of observations in the population. It is customary to represent the parameter by Greek letters. Population mean is denoted by the Greek letter μ.

Example: Suppose in a small city there are five drugstores. The numbers of employees at the five drugstores are 3, 5, 6, 4, and 7. Find the mean number of employees for the five stores.

Solution. We can consider it as a finite population with 5 observations. Here $N = 5$. The population mean is

$$\mu = \frac{3+5+6+4+7}{5} = \frac{25}{5} = 5 \text{ employees.}$$

Sample Mean or Sample Arithmetic Mean. If $x_1, x_2, \ldots, x_n$ are n observations of a sample, then the sample mean or sample arithmetic mean is a statistic defined by

$$\bar{x} = \frac{\sum x_i}{n}.$$

Here n is the sample size. A statistic is usually represented by ordinary letters of the English alphabet. A sample mean is denoted by $\bar{x}$.

Example: Suppose a factory has 150 workers. The monthly incomes (in taka) of 10 workers selected randomly from this factory are as follows:

3449, 3447, 3468, 3493, 3572, 3516, 3502, 3492, 3446, 3475.

Find the arithmetic mean using the appropriate symbol.

Solution. It is a sample data, since only 10 workers are selected from the 150. The sample arithmetic mean or simply sample mean is:

$$\bar{x} = \frac{3449+3447+3468+3493+3572+3516+3502+3492+3446+3475}{10} = \frac{34940}{10} = \text{Tk. } 3494.$$

Sample arithmetic mean for grouped data in case of discrete variable. Suppose $x_1, x_2, \ldots, x_k$ are the k values of a variable with corresponding frequencies $f_1, f_2, \ldots, f_k$ then the arithmetic mean is computed by the formula:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n}, \qquad \text{where } n = \sum f_i.$$

One can compute $\bar{x}$ by using the following table:

Value of the variable (x)	Frequency (f)	fx
x_1	f_1	f_1x_1
x_2	f_2	f_2x_2
x_3	f_3	f_3x_3
.	.	.
.	.	.
x_k	f_k	f_kx_k
Total	$\sum f_i = n$	$\sum f_ix_i$

Thus the arithmetic mean is calculated as:

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

Example: The following frequency distribution refers to the number of children per family of 75 workers of a factory:

Number of children (x)	0	1	2	3	4	5	6

Number of families (f)	3	5	10	16	25	12

Compute the average number of children per family of the workers of the factory.

Solution. This is a frequency distribution with discrete variable. Number of children is a variable.

Here, $x_1 = 0,\ x_2 = 1,\ x_3 = 2,\ x_4 = 3,\ x_5 = 4,\ x_6 = 5,\ x_7 = 6$ and

$f_1 = 3,\ f_2 = 5,\ f_3 = 10,\ f_4 = 16,\ f_5 = 25,\ f_6 = 12,\ f_7 = 4.$

Value of the variable (x)	Frequency (f)	fx
0	3	0
1	5	5
2	10	20
3	16	48
4	25	100
5	12	60
6	4	24
Total	$\Sigma f = 75$	$\Sigma fx = 257$

Here $k = 7,\quad n = \Sigma f_i = 75,\quad \Sigma f_i x_i = 257.$

Hence the average number of children per family is

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f_i x_i}{n} = \frac{257}{75} = 3.43\,.$$

Arithmetic mean from a frequency distribution with class interval or from a grouped data or from a continuous variable. Suppose $x_1, x_2, \ldots\ldots, x_k$ are the k mid-points of k classes with their corresponding frequencies $f_1, f_2, \ldots\ldots, f_k$, then the arithmetic mean is defined as:

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f_i x_i}{n}, \quad \text{where } n = \Sigma f_i\,.$$

Here mid-point of each class is taken as the representative value of the class. One can compute $\bar{x}$ by using the following table:

Mid-point of class interval (x)	Frequency (f)	fx
x_1	f_1	f_1x_1
x_2	f_2	f_2x_2
x_3	f_3	f_3x_3
.	.	.
.	.	.
x_k	f_k	f_kx_k
Total	$\Sigma f_i = n$	Σf_ix_i

Thus the arithmetic mean is calculated as: $\bar{x} = \frac{\Sigma x_i}{\Sigma i} = \frac{\Sigma x_i}{n}$

In order to compute the average from a frequency distribution with the given class intervals, the following three assumptions are made:

i) The class intervals must be closed.
ii) The values in each class must be uniformly distributed over the class interval.
iii) The mid-point of each class must represent the average of the class.

Example. The following data refer to the number of years worked by 37 employees of a firm:

Number of years worked	5–10	10–15	15–20	20–25	25–30	30–35
Number of employees	5	7	11	8	4	2

Find the average number of years worked by the employees of the firm.

Solution. The arithmetic mean of the number of years worked by the employees is the required answer. Here the mid-points of the class intervals are considered as values of the variable. To get the arithmetic mean, we required the following table:

Class interval	Mid-point (x)	Frequency (f)	fx
5 – 10	7.5	5	37.50
10 – 15	12.5	7	87.50
15 – 20	17.5	11	192.50
20 – 25	22.5	8	180.00
25 – 30	27.5	4	110.00
30 – 35	32.5	2	65.00
Total		37	672.50

Here, $\Sigma f = n = 37$, $\Sigma fx = 672.50$.

$$\text{Arithmetic mean} = \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f_i x_i}{n} = \frac{672.50}{37} = 18.18 \text{ years.}$$

That is the average number of years worked by the employees of the firm is 18.18 years. This method is called direct method for finding arithmetic mean from the grouped data.

Short-Cut Method for Calculating Arithmetic Mean. The method of computation for finding arithmetic mean discussed so far is time consuming and laborious, which usually needs calculator. Now, we shall discuss a method, which is known as short-cut method. In this method, the computation of mean can be done even manually. It saves both time and labour for computing arithmetic mean compared to the direct method discussed above. This method is useful when the data set is very large.

In this case, we introduce a new variable defined by $d_i = \frac{x_i - A}{c}$.

Here A is called assumed mean, c is the width of the class interval, n is the total number of observations, and f_i is the frequency corresponding to the value x_i. If we subtract or add any constant from a variable, its origin changes, and if the variable is divided or multiplied by some constant, the scale changes. Hence, the system is called shift of origin and change of scale of measurement.

Then, arithmetic mean becomes $\bar{x} = A + \frac{\Sigma f_i d_i}{n} \times c = A + c\bar{d}$.

This formula is known as short-cut method for computing arithmetic mean. The formula shows that arithmetic mean depends on the shift of origin and change of scale.

Remarks: A is usually taken as a middle value of the variable which has the highest frequency or near to the highest frequency just to get the maximum benefit of the calculation.

Example: The following frequency distribution refers to the number of hours worked per month by 50 workers of a factory:

Number of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Find the average number of hours worked per month by the workers of that factory by using the (i) Direct method and (ii) Short-cut method.

Solution. Table for Calculation of Arithmetic Mean

Number of hours worked	Mid -point (x)	(f)	fx	$d = \frac{x - 117.5}{25}$	fd
30 – 55	42.5	3	127.5	–3	-9
55 – 80	67.5	4	270.0	–2	-8
80 – 105	92.5	6	555.0	–1	-6
105 – 130	117.5	9	1057.5	0	0
130 – 155	142.5	12	1710.0	1	12
155 – 180	167.5	11	1843.5	2	22
180 – 205	192.5	5	962.5	3	15
Total		50	6525		26

(i) Direct Method. The formula for computing arithmetic mean by direct method is:

$$\bar{x} = \frac{\Sigma f_i x_i}{n}.$$

Here, x_i = mid point of the ith class, f_i = frequency of the ith class, and n = total frequency or total number of observations.

$$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{6525}{50} = 130.50 \text{ hours per month.}$$

That is, the worker of the factory worked on an average of 130.50 hours per month.

(ii) **Short-cut method.** The formula for finding arithmetic mean by short-cut method is

$$\bar{x} = A + \frac{\Sigma f_i d_i}{n} \times c.$$

Here, $d_i = \frac{x_i - A}{c}$, A = assumed mean and c = size of the class interval or width of the class interval. We take $A = 117.5$ as it is in the middle most value of x, and $c = 25$ as the width of the class interval is 25. We have $\Sigma fd = 26$, $n = 50$.

So, $$\bar{x} = A + \frac{\Sigma f_i d_i}{n} \times c = 117.5 + \frac{26}{50} \times 25 = 117.5 + 13 = 130.5 \text{ hours.}$$

It is seen that both the methods give the same results but the short-cut method is easier than the direct method from the computation point of view.

Example: The following frequency table gives the weekly number of hours worked including overtime by 70 workers of a factory:

Number of hours worked	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of workers	5	12	15	25	8	3	2

Compute arithmetic mean or average working hours by (a) direct method and (b) short-cut method.

Weighted Arithmetic Mean. In ordinary arithmetic mean, equal importance is given to all the observations. But in practice relative importance of all the observations are not the same. In such situation we compute weighted arithmetic mean. The term 'weight' stands for the relative importance of the different observations. Weighted mean is especially useful in problems relating to the construction of index numbers and standardized birth and death rates.

Definition. Suppose $x_1, x_2, \ldots\ldots\ldots, x_k$ are k values of a variable x whose relative importance are measured by the weights $w_1, w_2, \ldots\ldots\ldots, w_k$ respectively, then the weighted arithmetic mean is computed by the following formula:

$$\bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i}.$$

It is to be noted that there is similarity between the formula of weighted mean and the mean of a frequency distribution. Actually, the formula for mean of a frequency distribution can be considered as a special case of weighted mean. Frequency of a class is considered as weight of the mid-point of that class.

Example: A contractor employs three types of worker say male, female and children. To a male worker he pays Tk. 125 per day, to female worker Tk.100 per day and to a child worker

Tk. 75 per day. The numbers of male, female and child workers hired by the contractor are 15, 25 and 35 respectively. What is the average wage per day paid by the contractor?

Solution. The simple arithmetic mean of the wage is

$$\bar{x} = \frac{125+100+75}{3} = \text{Tk. } 100 \text{ per day.}$$

It is not the correct answer to the problem. If the numbers of male, female and child workers are the same, this answer would be correct.

For example, if the contractor hired 20 workers in each category, then the weighted mean is:

$$\bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{20\times125+20\times100+20\times75}{20+20+20} = \frac{6000}{60} = \text{Tk.}100 \text{ per day.}$$

It is the same as the simple arithmetic mean.

Here the numbers of male, female and child workers are different. The appropriate mean is the weighted mean. It is calculated as follows:

Worker	Wage per day (x)	No. of workers (w)	wx
Male	125	15	1875
Female	100	25	2500
Child	75	35	2625
		75	7000

$$\bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{7000}{75} = \text{Tk. } 93.33 \text{ per day.}$$

The weighted mean is less than the simple mean, since the weight of the child workers is more than the male and female.

Merits and demerits of Arithmetic Mean.

Merits. It is the most popular and widely used average in practice. It has the following merits:

i) It is easy to understand,
ii) It is easy to calculate,
iii) It is based on all the observations,
iv) It is rigidly defined,
v) It is capable of further algebraic treatment,
vi) It is less affected by sampling fluctuation.

It is the best measure of average among all the averages. However it has some limitations too.

Demerits or limitations of Arithmetic mean

i) It is affected by extreme values,
ii) It cannot be computed in case of open-ended class interval of a frequency distribution,
iii) It is not a good measure of central tendency in case of highly skewed distribution,

iv) It cannot be calculated for qualitative data,

v) It cannot be found graphically.

Some Mathematical Properties of Arithmetical Mean. Three most important mathematical properties of arithmetic mean are:

(1) The algebraic sum of the deviations of all the observations about the arithmetic mean is always zero. Symbolically, if $x_1, x_2, \ldots\ldots\ldots, x_n$ are n observations of a set of data and if $\bar{x}$ is the arithmetic mean, then $\Sigma(x_i - \bar{x}) = 0$.

(2) The sum of the squared deviations of all the observations from the arithmetic mean is minimum. Symbolically, $\Sigma(x-\bar{x})^2 \le \Sigma(x-a)^2$, where a is any arbitrary value other than $\bar{x}$.

(3) Arithmetic mean depends on the shift of origin and change of scale.

Proof. Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be n values of a variable x.

By definition, $\bar{x} = \frac{\Sigma x}{n}$.

Let, $u = \frac{x - A}{k}$, this means we have shift origin of x to A and change by scale k.

In this case, $x = ku + A$.

Then, $x_1 = ku_1 + A, x_2 = ku_2 + A, \ldots\ldots\ldots, x_n = ku_n + A$

$$x_1 + x_2 + \ldots\ldots + x_n = ku_1 + ku_2 + \ldots\ldots + ku_n + nA$$

$$\frac{\Sigma x}{n} = \frac{k\Sigma u}{n} + \frac{nA}{n}$$

Hence, $\bar{x} = k\bar{u} + A$.

Here it is seen that arithmetic mean $\bar{x}$ depends on both k and A. This means arithmetic mean depends on the shift of origin and change of scale.

Remarks: This formula is used for finding arithmetic mean by short cut method.

(4) If $\bar{x}_1$ and $\bar{x}_2$ are two arithmetic means of two related sets of observations, and n_1 and n_2 are the corresponding number of observations, then the combined arithmetic mean of the two sets of observations is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$\bar{x}$ = Combined mean of two sets,

$\bar{x}_1$ = Arithmetic mean of the first set, $\bar{x}_2$ = Arithmetic mean of the second set, n_1 = Number of observation in the first set, n_2 = Number of observation in the second set.

If there are k such groups with means $\bar{x}_1, \bar{x}_2, \ldots\ldots, \bar{x}_k$ with number of observations $n_1, n_2, \ldots\ldots, n_k$ respectively, then, the combined mean of the k groups is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots\ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots\ldots + n_k} = \frac{\sum n_k \bar{x}_k}{\sum n_k}.$$

Example In a garments factory, the mean wages of male and female workers per month are Tk. 4,000.00 and Tk. 3,500 respectively. The numbers of male and female workers are 65 and 125 respectively. Find the average monthly wage of the workers.

Solution. Here $\bar{x}_1 = 4000$, $\bar{x}_2 = 3500$, $n_1 = 65$ and $n_2 = 125$. The combined mean,

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{65 \times 4000 + 125 \times 3500}{65 + 125}$$

$$= \frac{260000 + 437500}{190} = \frac{697500}{190} = \text{Tk. } 3671.05 \text{ per month.}$$

Median

Definition. Median is the middle most value of a set of observations when the values are arranged in order of magnitude. That means, it divides the whole ordered observations into two equal parts, half of the observations is greater than or equal to it. It is also called a position or location measure of central tendency.

Median from ungrouped data. First arrange the observations in ascending or descending order of magnitude (both arrangement would give the same answer).

Rule 1. Now if the number of observations n is odd, then there will be a single middle value, which is the median, and its position will be $(\frac{n+1}{2})$th ordered observation of the series.

Rule 2. If the number of observations n of data set is even, then the position of the median will be the arithmetic mean of the $(\frac{n}{2})$th and $\left(\frac{n}{2}+1\right)$th ordered observations.

Another rule for computing median. When n/2 of the ordered array is integer, then median is the mean of the (n/2)th and {(n/2) +1}th ordered observations. But when n/2 is not an integer, round up it to the next higher integer and the observation corresponding to that integer is the median.

Example The following data give the monthly wages in taka of 7 workers of a factory:
Wage (in Taka): 2700, 2750, 2680, 2790, 2760, 2720, 2740.

Compute median wage of the workers.

Solution. First we arrange the data set in ascending order of magnitude. The ordered array is 2680, 2700, 2720, 2740, 2750, 2760, 2780.

Here n is odd. The median of the data set is $\left(\frac{n+1}{2}\right)$th ordered observation = $\left(\frac{7+1}{2}\right)$th ordered observation = 4th ordered observation = Tk. 2740 per month.

Example The following data refer to the profits of a store in thousand taka for the last 12 months are 3, 6, 8, 9, 6, 10, 5, 12, 9, 8, 11, 7. Compute median profit of the store.

Solution. First we arrange the observations in ascending order of magnitude. The ordered array is 3, 5, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12. Here $n = 12$ is even.

$$\text{Median} = \frac{\frac{n}{2}\text{th observation} + \frac{n+1}{2}\text{th observation}}{2}$$

$$= \frac{6\text{th observation} + 7\text{th observation}}{2} = \frac{8+8}{2} = \text{Tk. 8 thousand.}$$

Interpretation. 50% of the monthly profit is Tk. 8 thousand or less.

Computation of Median from grouped data of discrete variable. Median from a discrete frequency distribution can be obtained by using the rule1 and rule 2. But to locate the position of median we have to construct a cumulative frequency table.
We shall cite one example to clarify the matter.

Example A survey was conducted on 100 school teachers to know the median number of children of their family. The results of the survey are presented in the following frequency table:

No. of children	0	1	2	3	4	5
No. of family	5	15	25	35	16	4

Solution. First we have to construct a cumulative frequency table for locating the position of median:

Variable (No. of children) (x)	Frequency (No. of family) (f)	Cumulative frequency
0	5	5
1	15	20
2	25	45
3	35	80
4	16	96
5	4	100

Here $n = 100$ is even, and then the position of median will be the arithmetic mean of the 50th and 51th ordered observations. It is noted that the frequency distribution is always constructed orderly. It is seen from the third column (cumulative frequency column) that the

positions of both the observations 50th and 51th correspond to the no. of children 3. Hence median is $(3+3)/2 = 3$. That means, 50% of the teachers have children 3 or less.
The position of median can also be easily located from the stem and leaf plot.

Computation of median from a grouped data of continuous variable. First we have to construct a cumulative frequency table. Then we have to identify the median class. Median class is the most important class for computing median. The class which contains $\frac{n}{2}$ th observation is called the median class. Here we always use $\frac{n}{2}$ instead of $\frac{n+1}{2}$ to locate median because $\frac{n}{2}$ divides the whole area of the curve into two equal parts in case of continuous variable. The formula for computing median is

$$\boxed{M_e = L + \frac{n/2 - F}{f} \times c.}$$

Here M_e = median, L = lower limit of the median class, n = Total of observations, F = Cumulative frequency of pre-median class, f = Frequency of the median class, c = Width of the median class.

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

Number of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Compute median of the frequency distribution.

Solution. First we construct a cumulative frequency table with the frequency distribution. The cumulative frequency distribution table is

Class interval	Frequency (f)	Cumulative frequency (F)
30 – 55	3	3
55 – 80	4	7
80 – 105	6	13
105 – 130	9	22
130 – 155	12	34
155 – 180	11	45
180 – 205	5	50

Here, $n = 50$, then $\frac{50}{2} = 25$th observation lies in the class 130-155. Hence the median class is 130-155. That is 25th observation is in cumulative frequency 34 and the corresponding class is 130-155. Here, $L = 130$, $\frac{n}{2} = 25$, $F = 22$, $f = 12$ and $c = 25$. Hence

$$M_e = L + \frac{n/2 - F}{f} \times c = 130 + \frac{25-22}{12} \times 25$$

$= 130 + 6.25 = 136.26$ hours per month.

Interpretation. 50% of the workers worked for 136.26 hours or less per month.

Locating Median graphically.

Median from ogive Curve. Generally median is obtained graphically from ogive curve. It can be determined graphically by applying any of the following two methods:

1. Draw two ogives- one by 'less than' method and the other by 'more than' method. Draw a perpendicular on the x-axis from the point where the two curves intersect each other. The point where this perpendicular touches the x-axis gives the value of the median.
2. Draw one ogive usually by 'less than' method. Plot the upper limit of the variable on the x-axis and the cumulative frequency on the y-axis. Locate a point by $n/2$ on the y-axis and from this point draw a horizontal line parallel to the x-axis on the cumulative frequency curve. Draw a perpendicular on the x-axis from the point where it meets the ogive. The point at which the perpendicular cuts the x-axis is the median.

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

i) Draw ogives by 'less than' method and by 'more than' and locate median from them,
ii) Draw an ogive by 'less than' method and locate median.

Solution. First we construct a 'less than' and 'more than' cumulative frequency table

No. of hours	frequency	No. of hours	frequency
Less 30	0	More than 30	50
Less 55	3	More than 55	47
Less 80	7	More than 80	43
Less 105	13	More than 105	37
Less 130	22	More than 130	28
Less 155	34	More than 155	16
Less 180	45	More than 180	5
Less 205	50	More than 205	0

Now plot the class intervals on the x-axis and the cumulative frequency on the y-axis.

(i) Draw two ogives by 'less than' and by 'more than' methods on the same graph paper. Now draw a perpendicular from the intersecting point A on the x-axis. The point at which the perpendicular cuts the x-axis is the median.

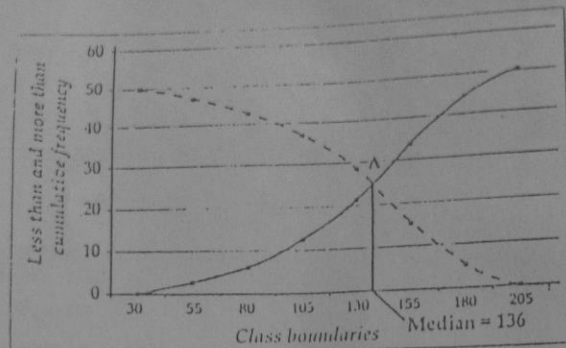

Fig. *Median from two ogives.*

It is observed that two ogives intersect at A, and the value at which the perpendicular from A to x-axis cuts the x-axis is about 136. So median is 136.

(ii) Now plot the class intervals on the x-axis and the' less than' cumulative frequency on the y-axis. Plot points above the class intervals according to their cumulative frequencies. Join the point's free hand to get the required ogive. Locate a point $n/2 = 50/2 = 25$ on the y-axis and from this point draw a line parallel to the x-axis on the ogive. Now draw perpendicular on the x-axis from the point at which the line cuts on the ogive. The point at which the perpendicular cuts the x-axis is the median. It is observed that the value at which the perpendicular from A cuts the x-axis is 136.So the median is again is 136.

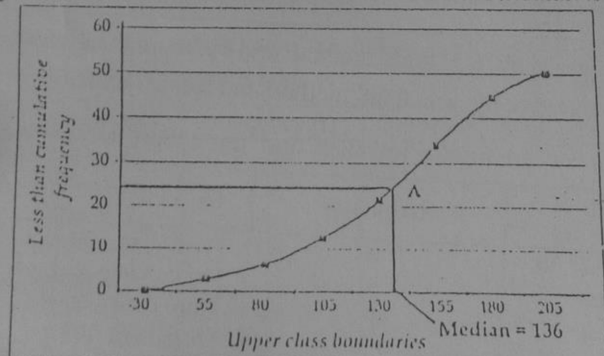

Fig. *Median from less than ogive.*

Merits and demerits of median

Merits: Median is a positional measure. It has the following merits:

i) It is easy to understand.
ii) It is easy to calculate.
iii) It is not affected by extreme values. That is, it is unaffected by outliers.
iv) It can be computed in open-end frequency distribution.

v) It can be obtained from ogive. That means it can be found graphically.
vi) It is a suitable measure of location in case of very skewed distribution.
vii) The position of median can be easily located when a qualitative variable is measured in ordinal scale.
viii) It is a unique value for a set of data like arithmetic mean.

Demerits or limitations:

i) It is not based on all the observations.
ii) It is not capable of algebraic treatment.
iii) It is more affected by sampling fluctuations.
iv) It cannot be calculated for nominal data.

Advantage of median over arithmetic mean. Arithmetic mean is the best measure of central tendency. But there some situations where median is superior to arithmetic mean.

1) In presence of outliers of a set of data, median is better than arithmetic mean.
2) For highly skewed distribution, median is superior to arithmetic mean.
3) For open-ended distribution, arithmetic mean is not possible to calculate unless some assumption is made but median can be easily computed.
4) The position of median can be easily located when a qualitative variable measured in ordinal scale but mean is not possible for an ordinal data.

Mode

Mode is another important measure of central tendency. It is a value of the variable which occurs the maximum number of times, i.e., having highest frequency.

Definition. Mode is that value of a variable, which has highest frequency. According to Zizek, "Mode is the value occurring most frequently in a series and around which the other items are distributed most densely". There may be a unique mode, several modes or essentially no mode. The distribution, for which there exists only one mode, is called unimodal distribution. Similarly, for two or more modes, the distributions are known as bimodal or multimodal distribution respectively. Like median mode is not influenced by extreme values.

Mode from ungrouped small set of data. First arrange the observations in ascending order of magnitude. Then count the number of times of repetition of each observation. The observation that has the highest frequency is called the mode.

Example Find Mode, median and mean of the data sets

(a) 4, 5, 5, 5, 6, 6, 7, 8, 12
(b) 1, 2, 3, 3, 3, 5, 6, 7, 7, 7, 23
(c) 1, 3, 5, 6, 7, 9, 11, 15, 16

Solution. The mode, median and mean of the three data sets are calculated and shown in the table given below:

Data Set	Mode	Median	Mean
(a) 4, 5, 5, 5, 6, 6, 7, 8, 12	5	6	6.44
(b) 1, 2, 3, 3, 3, 5, 6, 7, 7, 7, 23	3, 7	5	6.09
(c) 1, 3, 5, 6, 7, 9, 11, 15, 16	None	7	8.11

The data set (a) in example 5.5.1 is called unimodal, since there is one mode. Data set (b) is called bimodal, since there are two modes. On the other hand the data set (c) has no mode.

Remarks: For many sets of data median lies between mode and mean. But this is not always so.

Mode from grouped data in case of discrete variable. First we shall discuss how to find mode from grouped data of discrete variable. In this case mode is that value of the variable, which has the highest frequency. Now we shall cite one example.

Example The data given below are size of shoes sold by a shop in a week:

Size of shoes	5	6	7	8	9	10	11
Number of shoes sold	10	23	27	45	24	17	9

Find the modal size of the shoes sold by the shop per week.

Solution. It is easily seen that the modal size of the shoe is 8 since it's frequency is 45 which is the highest. That means 8 is the most popular size of the shoe.

Calculation of Mode from grouped data in case of continuous variable. When we have a grouped data of continuous variable, the formula for computing mode is

$$\boxed{M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i.}$$

Here M_o = Mode, L = Lower limit of the modal class, Δ_1 = Frequency difference between the modal class and pre-modal class, Δ_2 = Frequency difference between the modal class and post-modal class, i = size of the modal class.

It is seen from the formula that the modal group or class is the most important class for finding mode. Modal class is the class, which has the highest frequency.

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Compute mode.

Solution. It is obvious from the frequency table that the class 130-155 contains the highest frequency. Hence the modal class is 130-155. The formula for finding mode is

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i.$$

Here $L = 130$, $\Delta_1 = 12 - 9 = 3$, $\Delta_2 = 12 - 11 = 1$, $i = 25$. Hence,

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 130 + \frac{3}{3+1} \times 25$$

$$= 130 + 18.75 = 148.75 \text{ hours per month.}$$

Hence the modal working time is 148.75 hours per month. That means, most of the workers worked for 148.75 hours.

Sometimes, mode of a frequency distribution is difficult to compute. For example, when the highest frequency of a frequency distribution lies in the first class or in the last class, we cannot compute mode with the above formula. In that case mode can be obtained by the following approximate formula given by Karl Pearson:

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Usually this formula is applicable in case of moderately skewed distribution.

Example In a moderately skewed distribution arithmetic mean and mode are 24.6 and 26.1 respectively. Find the value of the median and explain the reason for the method employed.

Solution. The relationship among mean, median and mode in a moderately skewed distribution is:

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Median can be easily found from this relationship. Here

$$26.1 = 3 \text{ Median} - 2\,(24.6)$$

$$3 \text{ Median} = 26.1 + 49.2 = 75.3$$

$$\text{Median} = 75.3/3 = 25.1$$

Locating Mode graphically. Mode of a frequency distribution can be located graphically from a histogram. The steps in finding mode are:

1) First draw a histogram of the frequency distribution,
2) Then locate modal class and the rectangle over this class by inspecting highest frequency,
3) Then draw two lines diagonally on the inside of the modal class rectangle, starting from each upper corner of the rectangle to the upper corner of the adjacent rectangle,
4) Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis which gives us modal value.

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Draw histogram and locate mode from it.

Solution. We know from the Example 5.5.7 that the mode of the distribution is 148.75 hours per month.

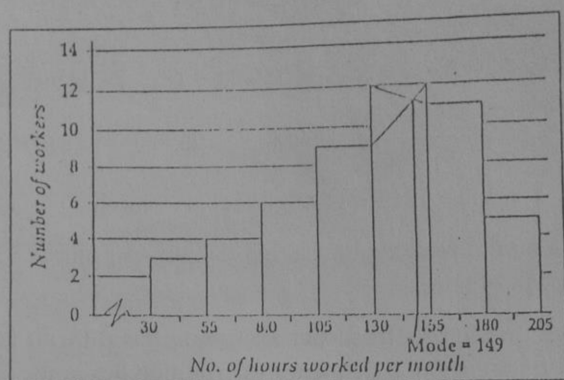

Fig. *Mode from histogram.*

From the histogram it is seen that the mode is 149.00 hours. Hence by both the methods we get approximately the same value of mode.

Merits and demerits of mode.

Merits:

i) It is easy to understand,
ii) It is easy to calculate,
iii) It is not affected by extreme values,
iv) It can be calculated for open-ended class interval,
v) It can be calculated graphically,
vi) It can be calculated both for qualitative and quantitative data.

Demerits:

i) It is not based on all the observations,
ii) Mode is not a rigidly defined measure as there are several formulae for finding mode, all of which usually give somewhat different answers.
iii) It is not clearly defined in case of bimodal or multimodal distribution,
iv) Mode cannot be defined if each value of the variable occurs only once in a set of data,
v) It is affected by sampling fluctuation,
vi) It is not suitable for further algebraic treatment,

vii) Mode cannot be calculated if the highest frequency lies in the first or last class in a frequency distribution.

There are many situations in which arithmetic mean and median fail to reveal the true characteristics of data. For example, when we talk of most common wage, most common income, and most common sales size of a shop or size of a ready-made garment, most common height etc. we have in mind mode and not the arithmetic mean or median discussed earlier.

Some Other Positional Measures

Median is the most important positional measure. It divides the whole distribution into two equal parts. That is 50% observations are equal to or smaller than median and 50% observations are equal to or larger than the median. Other important positional measures are (i) Quartiles, (ii) Deciles and (iii) Percentiles. In general, quartiles, deciles and percentiles are all known as quantiles.

Quartiles: Quartiles divide the ordered data into 4 equal parts. So there are three quartiles. They are Q_1, Q_2 and Q_3. Q_2 is the median. Q_1 and Q_3 are called the first and third quartiles.

First Quartile: The first quartile, Q_1, is a value for which 25% of the observations are equal to or smaller than Q_1 and 75% are equal or larger than Q_1.

Second quartile: Second quartile, Q_2, is a value for which 50% of the observations are equal to or smaller than Q_2 and 50% are equal to or larger than Q_2. That is it divides the ordered observations into two equal parts.

Third Quartile: The third quartile, Q_3, is a value for which 75% of the observations are smaller than or equal to it and 25% are equal to or larger than it.

Quartiles from ungrouped data:

Step 1. First we arrange the observations in ascending order of magnitude (smallest value to largest value).

Step 2. For finding ith quartile, Q_i $(i = 1, 2, 3)$, we compute an index:

$$j = \frac{in}{4}; \qquad i = 1, 2, 3$$

Step 3. If j is an integer, Q_i is the mean of the jth and $(j+1)$th ordered observations.

Step 4. If j is not an integer, Q_i is the value of the ordered observation corresponding to the next integer greater than j.

Different steps for finding Q_1:

Step 1. Arrange the observations in ascending order of magnitude.

Step 2. Compute the index $j = \frac{n}{4}$.

Step 3. If $j = \frac{n}{4}$ is an integer, Q_1 is the mean of the jth and $(j+1)$th observations of the ordered array.

Step 4. If $j = \frac{n}{4}$ is not an integer, Q_1 is the value corresponding to the next higher integer than j of the ordered array.

Example The following data refers to the monthly starting salaries in taka for a sample of 12 business school graduates: 7850, 7950, 8050, 7880, 7755, 7710, 7890, 8130, 7940, 8325, 7920, and 7880. Find Q_1, Q_2 and Q_3.

Solution. Stepwise procedure for finding quartiles is followed here:

Step 1.First we arrange the data in ascending order,

7710, 7755, 7850, 7880, 7880, 7890, 7920, 7940, 7950, 8050, 8130, 9325.

Step 2. For finding Q_1, we compute $j = \frac{n}{4}$.

Step 3. Here $j = \frac{n}{4} = \frac{12}{4} = 3$ is an integer.

Step 4. Q_1 is the mean of the 3rd and 4th observations of the ordered array. That is,

$$Q_1 = \frac{7850 + 7880}{2} = \text{Tk.}7865 \text{ per month.}$$

Similarly, for Q_2 or median, $\frac{2n}{4} = \frac{2 \times 12}{4} = 6$ is an integer. Hence median is the average of the 6th and 7th observations of the ordered array. That is,

$$\text{Median} = Q_2 = \frac{7890 + 7920}{2} = \text{Tk. } 7905 \text{ per month.}$$

For third quartile, $\frac{3n}{4} = \frac{3 \times 12}{4} = 9$ is an integer. Hence Q_3 is the mean of the 9th and 10th ordered observations. Here 9th and 10th observations are 7950 and 8050 respectively. Hence,

$$Q_3 = \frac{7950 + 8050}{2} = 8{,}000 \text{ Tk. per month.}$$

Deciles. Deciles divide the total ordered data into 10 equal parts. So there are nine deciles. They are denoted by $D_1, D_2, \ldots\ldots\ldots, D_9$. D_5 is the median.

Deciles from ungrouped data: Different steps for finding deciles are:

Step 1. As in case of median and quartiles, we arrange the data in ascending order of magnitude.

Step 2. For ith decile, D_i $(i = 1, 2, \ldots\ldots\ldots, 9)$, we compute an index:

$$j = \frac{in}{10}; \quad i = 1, 2, 3, \ldots\ldots, 9.$$

Step 3. If j is an integer, D_i is the mean of the jth and (j + 1)th ordered observations.

Step 4. If j is not an integer, D_i is the value of the ordered observation corresponding to the next integer greater than j.

Example The following data refer to the monthly starting salaries in taka for a sample of 12 business school graduates:

7850, 7950, 8050, 7880, 7755, 7710, 7890, 8130, 7940, 8325, 7920, 7880.

Compute D_1, D_5, D_8.

Solution. Step 1. First we arrange the data in ascending order of magnitude. The ordered array is 7710, 7755, 7850, 7880, 7880, 7890, 7920, 7940, 7950, 8050, 8130, 9325.

Step 2. For finding D_1, we compute, $j = \frac{n}{10} = \frac{12}{10} = 1.2$.

Step 3. Here $j = 1.2$ is not an integer, the next integer is 2. The second ordered observation is 7755. Hence D_1 is 7755. For finding D_5, we compute the index as

$$j = \frac{5n}{10} = 6.$$

Here j is an integer. The mean of the 6th and 7th ordered observations is D_5. Hence

$$D_5 = \frac{7890 + 7920}{2} = \text{Tk.}7905 \text{ per month.}$$

D_5 and Q_2 are the median which we got before.

For finding D_8, we compute the index as

$$j = \frac{8n}{10} = \frac{8 \times 12}{10} = 9.6.$$

Here $j = 9.6$ is not an integer. The next integer is 10. The 10th ordered observation is 8050. Hence the 8th decile is 8050 Tk. D_8 = Tk. 8050 suggests that salaries of 80% of the business graduates are less than or equal to Tk. 8050.

Percentiles. Percentiles divide the total ordered data into 100 equal parts. So there are 99 percentiles. They are denoted by P_1, P_2, ..., P_{99}. P_{50} is the median or 5th decile or 2nd quartile of the distribution, P_{25} is the first quartile , P_{75} is the third quartile , P_{20} is the second decile and so on.

Percentiles from ungrouped data: Different steps for finding percentiles are:

Step 1. As before we arrange the data in ascending order of magnitude.

Step 2. For ith percentile, P_i (i = 1, 2, ..., 99), we compute an index:

$$j = \frac{in}{100}; \qquad i = 1, 2, 3, \ldots, 99$$

Step 3. If j is an integer, P_i is the mean of the jth and (j + 1)th ordered observations.

Step 4. If j is not an integer, P_i is the value of the ordered observation corresponding to the next integer greater than j.

Example The following data refer to the monthly starting salaries in taka for a sample of 12 business school graduates:

7850, 7950, 8050, 7880, 7755, 7710, 7890, 8130, 7940, 8325, 7920, 7880.

Compute P_{11}, P_{50}, P_{80}.

Solution. Step 1. First we arrange the data in ascending order of magnitude,

7710, 7755, 7850, 7880, 7880, 7890, 7920, 7940, 7950, 8050, 8130, 9325.

Step 2. For finding P_{11}, we compute

$$j = \frac{11n}{100} = \frac{12 \times 11}{100} = 1.32 .$$

Step 3. Here $j = 1.32$ is not an integer. The next integer is 2. The second ordered observation is P_{11} which is 7755. Hence eleventh percentile, P_{11} is 7755.

For finding P_{50}, we compute the index:

$$j = \frac{50n}{100} = \frac{50 \times 12}{100} = 6 .$$

Here j is an integer. The mean of 6th and 7th ordered observations is P_{50}. Here it is

$$P_{50} = \frac{7890 + 7920}{2} = \text{Tk. } 7905 \text{ per month.}$$

P_{50}, D_5 and Q_2 are the median which are same.

For finding P_{80} we compute the index:

$$j = \frac{80n}{100} = \frac{80 \times 12}{100} = 9.6 .$$

Here j is not an integer. The next integer greater than 9.6 is 10. The 10th ordered observation is 8050. Hence 80th percentile is Tk.8050 which is the same as 8th decile.

Steps for computing quartiles, deciles and percentiles from ungrouped data:

Step 1. Arrange the observations in ascending order of magnitude.

Step 2. Compute the index j defined for quartiles, deciles and percentiles.

Step 3. If the index j is an integer for finding any quartile or decile or percentile, the mean of the jth and $(j+1)$th observations of the ordered array is the required quartile or decile or percentile.

Step 4. If the index j is not an integer, the ordered observation corresponding to the next greater integer than j is the required result.

Example A bank branch located in a commercial district of a city has developed an improved process for serving customers during the 12:00 P.M. to 1:00 P.M. peak lunch period. The waiting time in minutes of all customers during this hour is recorded over a period of one week. For this purpose a random sample of 20 customers is selected and the results are as follows: 4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.52, 3.20, 4.50, 6.10, 0.38, 5.12, 65.00, 6.19, 3.79, 4.51, 4.21, 2.31, 3.42, 3.45. Compute 1st quartile, 3rd quartile, 7th decile and 65th percentile and comment.

Solution. Step 1. First, we arrange the observations in ascending order of magnitude. The ordered array is

0.38, 2.31, 2.34, 3.02, 3.20, 3.42, 3.45, 3.52, 3.79, 4.21,
4.21, 4.50, 4.51, 4.77, 5.00, 5.12, 5.13, 5.55, 6.10, 6.19.

Step 2. For finding Q_1, compute $j = \frac{n}{4}$.

Step 3. Here, $j = \frac{n}{4} = \frac{20}{4} = 5$.

Step 4. Here, $j = 5$ is an integer. The mean of the 5th and 6th ordered observations is the first quartile. Hence, $Q_1 = \frac{3.20 + 3.42}{2} = 3.31$ minutes.

Comment: $Q_1 = 3.31$ means waiting time for 25% of the customers are less than or equal to 3.31 minutes and 75% of the customers are more than 3.31 minutes. It is to be noted that first quartile is also 25th percentile.

For finding Q_3, we compute j as $j = \frac{3n}{4}$. Here, $j = \frac{3n}{4} = \frac{3 \times 20}{4} = 15$.

Here $j = 15$ is an integer. The mean of the 15th and 16th ordered observations is the third quartile. Hence, $Q_3 = \frac{5.00 + 5.12}{2} = 5.06$ minutes.

Comment: $Q_3 = 5.06$ means waiting time of 75% of the customers are 5.06 minutes or less and 25% customers are more than 5.06 minutes. Third quartile is also 75th percentile.

Seventh Decile: For finding D_7, we compute the index j as $j = \frac{7n}{10}$. Hence, $j = \frac{7n}{10} = \frac{7 \times 20}{10} = 14$. Here $j = 14$ is an integer. The mean of the 14th and 15th ordered observations is the seventh decile. Hence, $D_7 = \frac{4.77 + 5.00}{2} = 4.87$ minutes.

Comment: $D_7 = 4.87$ means waiting time of 70% of the customers are 4.87 minutes or less and 30% are more than 4.87 minutes. It is the 70th percentile.

65th percentile: For finding P_{65}, we compute the index j as $j = \frac{65n}{100}$. Then $j = \frac{65n}{100} = \frac{65 \times 20}{100} = 13$.

Here $j = 13$ is an integer. Then the mean of the 13th and 14th ordered observations is the 65th percentile. Hence, $P_{65} = \frac{4.51 + 4.77}{2} = 4.64$ minutes.

Comment: $P_{65} = 4.64$ means waiting time of 65% of the customers is 4.64 minutes or less and 35% are more than 4.64 minutes.

Computation of quartiles, deciles and percentiles from grouped data of continuous type. The formulae for computing quartiles, deciles and percentiles are very much similar with the formula of median. For grouped data, the following formulae are used for finding quartiles, deciles and percentiles:

$$Q_i = L_i + \frac{in/4 - F_i}{f_i} \times c_i \qquad \text{For } i = 1, 2, 3.$$

Here Q_i is the ith quartile.

$$D_j = L_j + \frac{jn/10 - F_j}{f_j} \times c_j \qquad \text{For } j = 1, 2, \ldots\ldots, 9.$$

Here D_j is the jth decile.

$$P_k = L_k + \frac{kn/100 - F_k}{f_k} \times c_k \qquad \text{For } k = 1, 2, \ldots\ldots, 99.$$

Here P_k is the kth percentile.

The other symbols have their usual meanings and interpretation.

The above mentioned formulae are used in case of grouped data of continuous type.

Example The following frequency distribution refers to the number of hours worked per month of 50 workers of a factory:

No. of hours worked per month	30-55	55-80	80-105	105-130	130-155	155-180	180-205
Number of workers	3	4	6	9	12	11	5

Calculate Q_1, Q_3, D_3, P_{65} and interpret the values you obtained.

Solution. Table for computation

Class interval	Frequency (f)	Cumulative frequency (F)
30 - 55	3	3
55 - 80	4	7
80 - 105	6	13
105 - 130	9	22
130 - 155	12	34
155 - 180	11	45
180 - 205	5	50

Here $n = 50$. First quartile, Q_1 is the n/4th ordered observation = 12.5th ordered observation. It lies in the class interval 80-105. The formula for Q_1 is

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1 .$$

Here, $L_1 = 80$, $n/4 = 12.5$, $F_1 = 7$, $f_1 = 6$ and $c_1 = 25$.

$$Q_1 = L_1 + \frac{n/4 - F_1}{f_1} \times c_1 = 80 + \frac{12.5 - 7}{6} \times 25$$

$$= 80 + \frac{5.5 \times 25}{6} = 80 + 22.92 = 102.92 \text{ hours per month.}$$

Comment: 25% of the workers worked for 102.72 hrs or lees per month whereas 75% worked for more than 102.72 hrs.

Third quartile, Q_3 is the $3n/4$th ordered observation = 37.5th ordered observation. It lies in the class 155-180. Third quartile is

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3 .$$

Here, $L_3 = 155$, $3n/4 = 37.5$, $F_3 = 34$, $f_3 = 11$ and $c_3 = 25$.

$$Q_3 = L_3 + \frac{3n/4 - F_3}{f_3} \times c_3 = 155 + \frac{37.5 - 34}{11} \times 25$$

$$= 155 + \frac{3.5 \times 25}{11} = 155 + 7.95 = 162.95 \text{ hrs.}$$

Comment: 75% of the workers worked 162.95 hrs or less per month, whereas 25% worked more than 162.95 hrs per month.

Third decile, D_3 is the $(3n/10)$th ordered observation $= \frac{3 \times 50}{10} = 15$th ordered observation. It lies in the class 105-130. The third decile is

$$D_3 = L_3 + \frac{3n/10 - F_3}{f_3} \times c_3$$

Here, $L_3 = 105$, $3n/10 = 15$, $F_3 = 13$, $f_3 = 9$ and $c_3 = 25$.

Hence, $$D_3 = L_3 + \frac{3n/10 - F_3}{f_3} \times c_3 = 105 + \frac{3 \times 50/10 - 13}{9} \times 25$$

$$= 105 + \frac{2 \times 25}{9} = 110.56 \text{ hrs.}$$

Comment: 30% of the workers worked 105.56 hrs or less per month, whereas 70% worked more than 105.56 hours per month.

Sixty fifth percentile, P_{65} is the $(65n/100)$th ordered observation = $(65 \times 50)/100 = 32.5$th ≈ 33rd observation. 33rd observation lies in the class 130-155. Hence the 65th percentile is:

$$G.M. = Anti-\log\left(\frac{\Sigma \log x}{n}\right) = Anti-\log\left(\frac{5.4697}{6}\right) = Anti-\log(0.9116) = 8.159.$$

Comment: Again it is seen that A.M. is greater than G.M.

Calculation of Geometric mean for grouped data. In grouped data geometric mean is calculated with the help of the following formula:

$$G.M. = \text{Anti-log}\left(\frac{\Sigma f \log x}{n}\right).$$

Where x is the value of the variable in case of discrete data or x is the mid-point of the class interval in case of continuous data and f is the frequency. Now, we shall cite some examples.

Grouped data of discrete type

Example Find out geometric mean from the following data:

x	10	20	30	40	50	60
f	12	15	25	10	6	2

Solution. Calculation of Geometric Mean

x	f	log x	f log x
10	12	1.0000	12.0000
20	15	1.3010	19.5150
30	25	1.4771	36.9275
40	10	1.6021	16.0210
50	6	1.6921	10.1940
60	2	1.7782	3.5564
	n = 70		Σ f log x = 98.2139

$$G.M. = Anti-\log\left(\frac{\Sigma f \log x}{n}\right) = Anti-\log\left(\frac{98.2139}{70}\right) = Anti-\log(1.4031) = 25.30.$$

Grouped data of continuous type

Example Calculate geometric mean for the following distribution:

Weight(in lbs)	100-104	105-109	110-114	115-119	120-124
Number of employees	24	30	45	65	72

Solution. Calculation of Geometric Mean

Weight (in lbs)	f	Mid-point (x)	log x	f log x
100 – 104	24	102	2.0086	48.2064
105 – 109	30	107	2.0294	60.8820
110 – 114	45	112	2.0492	92.2140
115 – 119	65	117	2.0682	134.4330
120 – 124	72	122	2.0864	150.2208

		N = 236			$\Sigma f \log x = 485.9562$

$$G.M. = \text{Anti-log}\left(\frac{\Sigma f \log x}{n}\right) = \text{Anti-log}\left(\frac{485.9562}{236}\right) = \text{Anti-log}(2.059) = 114.6.$$

Application of Geometric Mean. Geometric mean is especially useful in the following cases:

1) It is used to find the average percent increase in sales, production, population or other economic or business data.
2) It is theoretically considered to be the best average in the construction of index number.
3) It is an average, which is most suitable when large weights have to be given to small values of observations and small weights to large values of observations, situation which we usually come across in social and economic fields.

Harmonic Mean

It is rarely used to analysis business and economic data. It is useful for computing the rate of increase of profits or average speed at which a journey has been performed or the average price in which an article has been sold.

Definition. Harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations. Suppose $x_1, x_2, \ldots\ldots, x_n$ are n non-zero observations of a data set, then it is computed by the formula:

$$H.M. = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots\ldots+\frac{1}{x_n}} = \frac{n}{\Sigma\left(\frac{1}{x}\right)}; \quad \text{for ungrouped data.}$$

$$\text{or,} \quad \frac{1}{H.M} = \frac{\frac{1}{x_1}+\frac{1}{x_2}+\ldots\ldots+\frac{1}{x_n}}{n} = \frac{\Sigma\left(\frac{1}{x}\right)}{n}$$

For grouped data, $H.M. = \frac{n}{\Sigma\left(f \times \frac{1}{x}\right)}$.

Here x is the value of the variable in case of discrete data or mid-pint in case of continuous data and f's are frequencies of x's. It is to be noted that the values of x must be non-zero in computing harmonic mean.

That means harmonic mean is the reciprocal of the arithmetic mean of the reciprocal of the individual observations.

In actual practice, the harmonic mean is most frequently used in averaging speeds for various distances covered where the distances remain constant, and also in finding the average cost of some commodity, such as mutual funds, when several different purchases are made by investing the same amount of money each time.

Some Simple Theorems and Problems on Measures of Location

Theorem For two positive non-zero quantities

$$A.M. \geq G.M. \geq H.M.$$

Here, A.M. = Arithmetic mean, G.M. = Geometric mean and H.M. = Harmonic mean.

Proof. Suppose x_1 and x_2 are two positive and non-zero quantities. Then,

$$A.M. = \frac{x_1 + x_2}{2}, \quad G.M. = \sqrt{x_1 \times x_2} \quad \text{and} \quad H.M. = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

Here, $\left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0$; since x_1 and x_2 are positive.

or, $x_1 + x_2 - 2\sqrt{x_1 x_2} \geq 0$

or, $x_1 + x_2 \geq 2\sqrt{x_1 \times x_2}$

or, $\frac{x_1 + x_2}{2} \geq \sqrt{x_1 \times x_2}$

Hence, A.M $\geq$ G.M ... (i)

Again, $\left(\frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{x_2}}\right)^2 \geq 0$

or, $\frac{1}{x_1} + \frac{1}{x_2} - 2\frac{1}{\sqrt{x_1 \times x_2}} \geq 0$

or, $\frac{1}{x_1} + \frac{1}{x_2} \geq 2\frac{1}{\sqrt{x_1 \times x_2}}$

or, $\sqrt{x_1 \times x_2} \geq \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$

Hence, G.M. $\geq$ H.M. ... (ii)

From (i) and (ii), we have A.M. $\geq$ G.M. $\geq$ H.M.

Theorem For two non-zero and positive quantities

$$G.M. = \sqrt{A.M. \times H.M.}$$

Proof. Suppose x_1 and x_2 are two positive and non-zero quantities. Then

$$A.M. = \frac{x_1 + x_2}{2}, \quad G.M. = \sqrt{x_1 \times x_2} \quad \text{and} \quad H.M. = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}.$$

Now, $\quad A.M. \times H.M. = \dfrac{x_1 + x_2}{2} \times \dfrac{2}{\dfrac{1}{x_1} + \dfrac{1}{x_2}} = \dfrac{x_1 + x_2}{2} \times \dfrac{2}{\dfrac{x_1 + x_2}{x_1 \times x_2}}$

$$= \frac{x_1 + x_2}{2} \times \frac{2(x_1 x_2)}{x_1 + x_2} = x_1 x_2 = \left(\sqrt{x_1 x_2}\right)^2 = (G.M.)^2.$$

Hence, $\quad G.M. = \sqrt{A.M. \times H.M.}$