Lecture 3

Graphing Quantitative Data

1. Histogram
2. Frequency polygon
3. Ogive

Histograms

- Grouped (quantitative: ratio and interval scale) data can be displayed in a *histogram* or a *polygon*.
- A **histogram** can be drawn for a frequency distribution, a relative frequency distribution, or a percentage distribution.
- To draw a histogram, we first mark class boundaries on the horizontal axis and frequencies (or relative frequencies or percentages) on the vertical axis.
- Next, we draw a bar for each boundary so that its height represents the frequency of that class.
- The bars in a histogram are drawn adjacent to each other with no gap between them.
- A histogram is called a **frequency histogram**, a **relative frequency histogram**, or a **percentage histogram** depending on whether frequencies, relative frequencies, or percentages are marked on the vertical axis.

Table: Class Boundaries, Class Widths, and Class Midpoints for Table 2.7

Class Limits	Class Boundaries	Frequency
30-39	29.5 to less than 39.5	**3**
40-49	39.5 to less than 49.5	1
50-59	49.5 to less than 59.5	8
60-69	59.5 to less than 69.5	10
70-79	69.5 to less than 79.5	7
80-89	79.5 to less than 89.5	7
90-99	89.5 to less than 99.5	4

Frequency

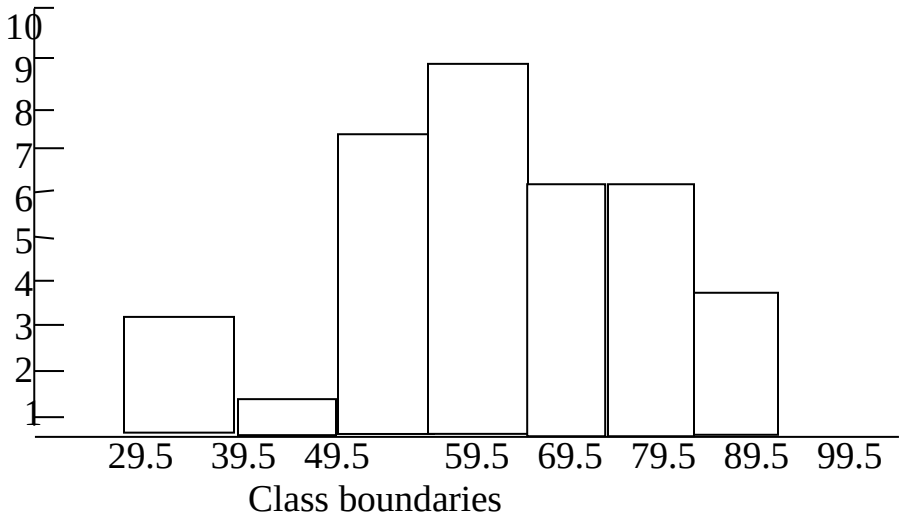
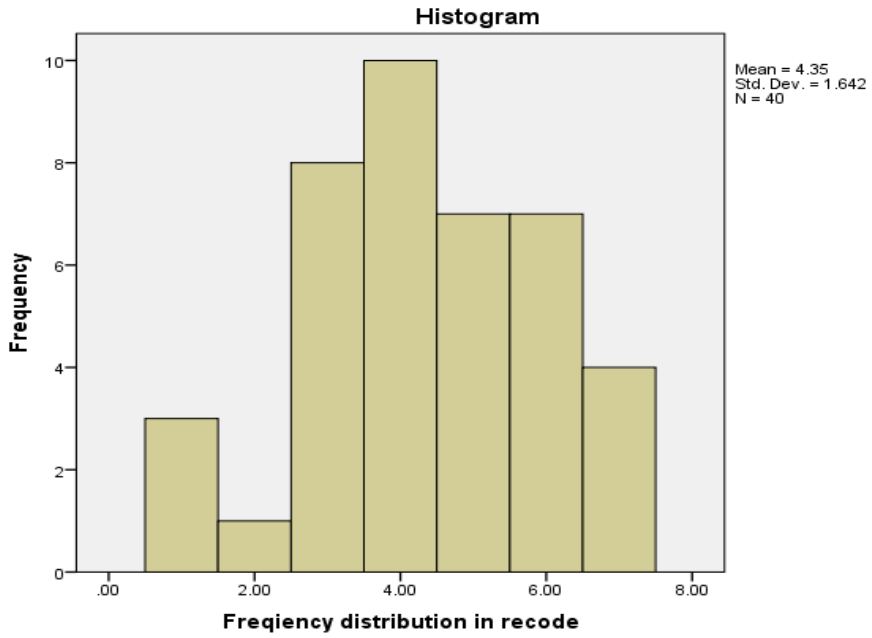


Figure: Frequency histogram for Table.

SPSS



Polygons

- A **polygon** is another device that can be used to present quantitative data in graphic form.

- A **polygon** can also be drawn for a frequency distribution, a relative frequency distribution, or a percentage distribution.
- To draw a **frequency polygon**, we first mark a dot above the **midpoint** of each class at a **height** equal to the **frequency** of that class.
- This is the **same** as marking the **midpoint at the top** of each bar in a histogram.
- Next, we mark remaining classes, one at each end, and mark their midpoints.
- In the last step, we join the adjacent dots with straight lines.
- The resulting line graph is called a frequency polygon or simply a polygon.
- A polygon **with relative frequencies** marked on the vertical axis is called a ***relative frequency polygon***.
- Similarly, a polygon **with percentages** marked on the vertical axis is called a ***percentage polygon***

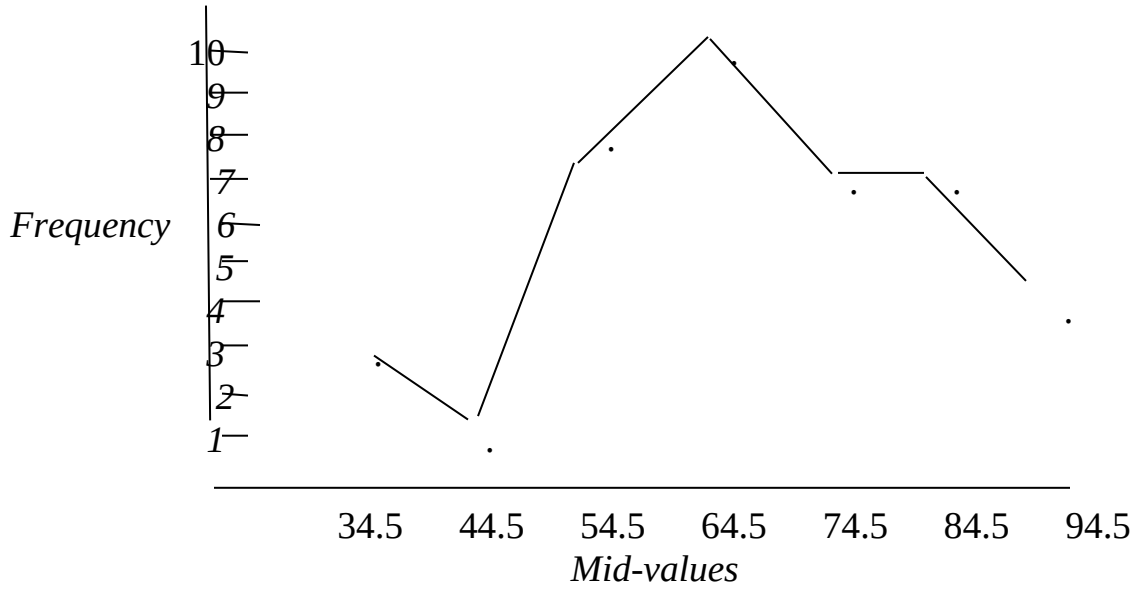


Figure: Frequency polygon for frequency distribution Table.

Shapes of Histogram

- Histogram usually shows the shape of the distribution.
- A histogram can assume any one of a large number of shapes. The most common of these shapes are
  1. Symmetric
  2. Skewed
  3. Uniform or rectangular

- A **symmetric histogram** is identical on both sides of its central point. The histograms shown in the following Figures are symmetric around the dashed lines that represent their central points.

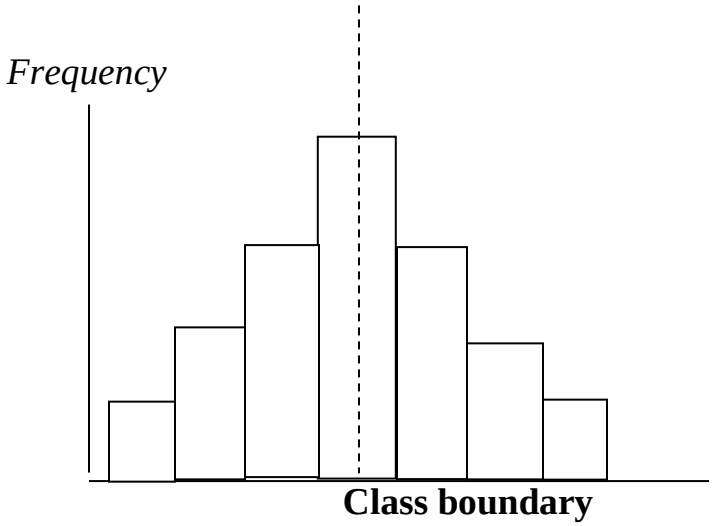
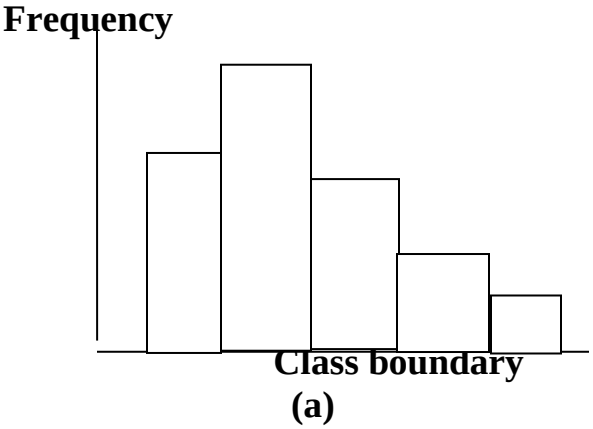


Figure: Symmetric histograms.

- A **skewed histogram** is nonsymmetric.
  - For a skewed histogram, the tail on one side is longer than the tail on the other side.
  - A **skewed-to-the-right histogram** has a longer tail on the right side (see Figure *a*).
  - A **skewed-to-the-left histogram** has a longer tail on the left side (see Figure *b*).



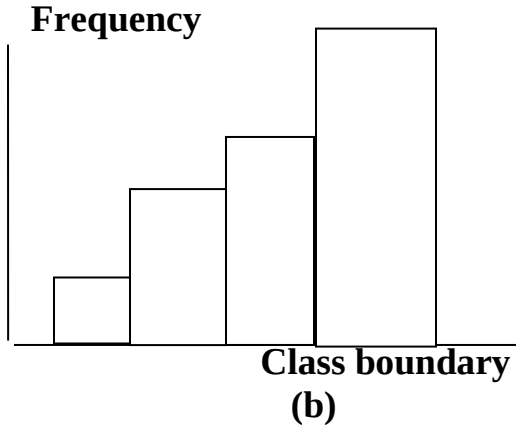


Figure (*a*) A histogram skewed to the right. (*b*) A histogram skewed to the left.

- A **uniform** or **rectangular histogram** has the same frequency for each class. The following Figure is an illustration of such a case.

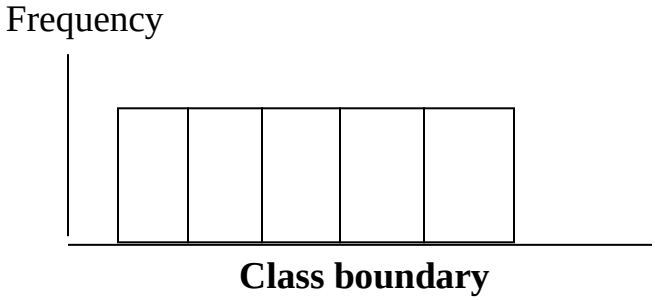


Figure: A histogram with uniform distribution.

Cumulative Frequency Distributions

A *cumulative frequency distribution* gives the total number of values that fall below the upper boundary of each class.

Table: Cumulative frequency

Class Limits	Frequency	Cumulative frequency
30-39	**3**	**3**
40-49	1	3+1=4
50-59	8	3+1+8=12
60-69	10	3+1+8+10=22
70-79	7	3+1+8+10+7=29
80-89	7	3+1+8+10+7+7=36
90-99	4	3+1+8+10+7+7+4=40

Ogive

An *ogive* is a curve drawn for the cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at heights equal to the cumulative frequencies of respective classes.

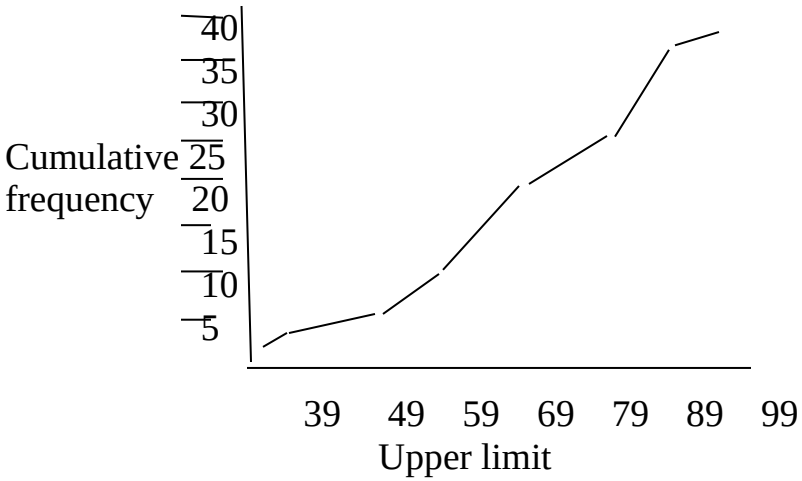


Figure: Ogive for the cumulative frequency distribution Table .

Stem-and-Leaf Display

In a *stem-and-leaf display* of quantitative data, each value is **divided into two portions**—a stem and a leaf.

- A stem-and-leaf display is an **exploratory data analysis** graph that is an **alternative to the histogram**.
- Data are **grouped** according to their **leading digits** (called the stem) while listing the **final/last** digits (called leaves) separately for each member of a class.
- The **leaves are displayed** individually in **ascending order** after each of the stems.

Example:

Table 1: Days to maturity 40 short-term investments.

71	64	99	55	64	89	87	65	62	38
67	70	60	69	78	39	75	56	71	51
99	68	95	86	57	53	47	50	55	81
80	98	51	36	63	66	85	79	83	70

Construct a stem-and-leaf display

Solution:

- To construct a stem-and-leaf display for these scores, we split each score into two parts.
- The first part contains the first digit, which is called the *stem*.
- The second part contains the second digit, which is called the *leaf*.
- Thus, for the score of the first student, which is 71, 7 is the stem and 1 is the leaf.
- To create a stem-and-leaf display, we draw a vertical line and list the stems on the left side of it, **arranged in increasing order**.
- After we have listed the stems, we **list the leaves** for all scores and record them next to the corresponding stems on the **right side of the vertical line**.

Stem (leading digit)	Leaf (last digit)
3	8 9 6
4	7
5	5 6 1 7 3 0 5 1
6	4 4 5 2 7 0 9 8 3 6
7	1 0 8 5 1 9 0
8	9 7 6 1 0 5 3
9	9 9 5 8

Figure: Stem-and-leaf display of the given data.

- The leaves for each stem of the stem-and-leaf display are *ranked* (in increasing order) and presented in the following Figure.

Stem	Leaf
3	6 8 9
4	7
5	0 1 1 3 5 5 6 7
6	0 2 3 4 4 5 6 7 8 9
7	0 0 1 1 5 8 9
8	0 1 3 5 6 7 9
9	5 8 9 9

Figure: Ranked stem-and-leaf display of the given data.

Dot Plot

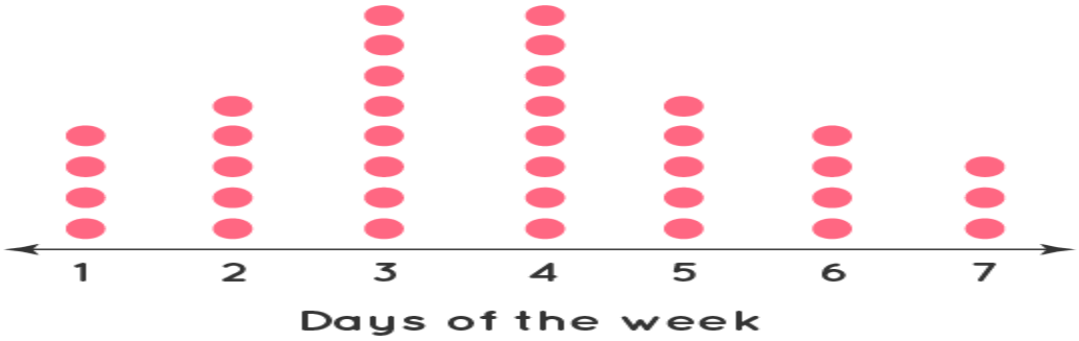
What is Dot Plot?

- A dot plot is used to display any data graphically in the form of dots or small circles.
- It is similar to a simplified histogram or a bar graph as the height of the bar formed with dots represents the numerical value of each variable.
- Dot plots are used to represent small amounts of data. What Is a Dot Plot?
- A Dot Plot is a graph for displaying the distribution of quantitative variable where each dot represents a value.
- For whole numbers, if a value occurs more than once, the dots are placed one above the other so that the height of the column of dots represents the frequency for that value.

Example: The number of hours that students did on homework in one week was recorded in the frequency table below. Draw a dot plot for the information given.

Day of the week	Number of hours of homework
1 (Monday)	4
2 (Tuesday)	5
3 (Wednesday)	8
4 (Thursday)	8
5 (Friday)	5
6 (Saturday)	4
7 (Sunday)	3

Solution: Dot plot is represented below.



Time Series Plot

What is a time series plot?

- A **time series graph** is a line graph that shows data such as measurements, sales or frequencies over a given time period.
- They can be used to show a pattern or **trend in the data** and are useful for making **predictions** about the future such as weather forecasting or financial growth.
- To draw a time series graph, we need a set of **axes**.
- The **horizontal axis** always shows the **time period**, and the **vertical axis** represents the **variable being recorded against time**.

For example,

This time series graph shows the temperature of a town recorded over two years at three-monthly periods known as **quarters**.

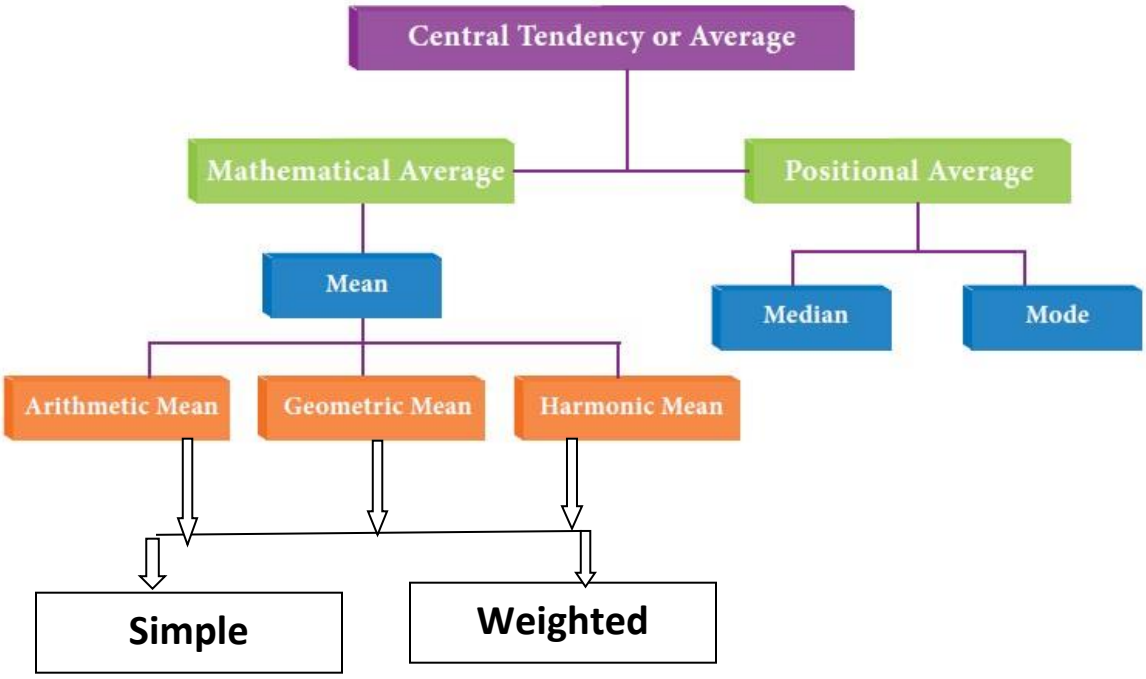
Summary Measures for Numerical Data

Measures of Central Tendency

- A measure of central tendency is ***a single value that attempts to describe a set of data by identifying the central position within that set of data***.
- It is a **summary statistic** that represents the **center point or typical value of a dataset.**
- You can think of it as the **tendency of data to cluster around a middle value.**

There are three main **measures of central tendency**: the mean, the median and the mode.

Various measures of central tendency



- Each of these **measures** describes a different indication of the typical or **central** value in the distribution.

<u>Simple Arithmetic Mean</u>
<u>Notations</u>

Population arithmetic mean is denoted by μ
Sample arithmetic mean is denoted by $\bar{x}$

Calculating Arithmetic Mean for Individual Observations

- The *mean for individual observations is* obtained by dividing the sum of all values by the number of values in the data set. Thus,

Mean for **population** data $=\mu = \frac{\sum x_i}{N}$

Mean for **sample** data $= \bar{x} = \frac{\sum x_i}{n}$

EXAMPLE

The following are the ages (in years) of all eight employees of a small company:

53, 32, 61, 27, 39, 44, 49, 57

Find the mean age of these employees.

Solution Because the given data set includes *all* eight employees of the company, it represents the population. Hence, $N = 8$. We have
The population mean is

$$\mu = \frac{\sum x_i}{N} = \frac{53+32+61+27+39+44+49+57}{8} = \frac{362}{8} = 45.25 \text{ years}$$

Thus, the mean age of all eight employees of this company is 45.25 years, or 45 years and 3 months.

EXAMPLE

Table 1 lists the total philanthropic givings (in million dollars) by six companies during 2021.

Table: Philanthropic Givings of Six Companies During 2021

Corporation	Money Given in 2020 (millions of dollars)
CVS	22.4
Best Buy	31.8
Staples	19.8
Walgreen	9.0
Lowe's	27.5
Wal-Mart	337.9

- ❖ Notice that the charitable contributions made by Wal-Mart are very large compared to those of other companies.
- ❖ Hence, it is an outlier.
- ❖ Show how the inclusion of this **outlier** affects the value of the mean.

Solution If we **do not include the charitable givings** of Wal-Mart (the outlier), the mean of the charitable contributions of the five companies is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{22.4+31.8+19.8+9.0+27.5}{5} = \frac{110.5}{5} = \$22.1\ million$$

Now, to see the impact of the outlier on the value of the mean, we **include the contributions of Wal-Mart** and find the mean contributions of the six companies. This mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{22.4+31.8+19.8+9.0+27.5+337.9}{6} = \frac{448.4}{6} = \$74.73\ million$$

Thus, including the contributions of Wal-Mart **causes more than a threefold increase** in the value of the mean, which changes from \$22.1 million to \$74.73 million.

- ➢ Among all the measures of central tendency, the arithmetic mean or mean is considered to be the best measure, because **it includes all the values of the data set**.
- ➢ If **any value changes** in the data set, this will **affect the mean value**, but it will not be in the case of median or mode.

But this example should encourage us to be cautious.

- We should remember that the mean is not always the best measure of central tendency because **it is heavily influenced by outliers**.
- Sometimes other measures of central tendency give a more accurate impression of a data set.
- For example, **when a data set has outliers**, instead of using the mean, we can use either the **trimmed mean** or the **median** as a measure of central tendency.

What is a 'Trimmed Mean'

- Trimmed Mean is an averaging method ***which eliminates*** *a partial percentage of the greatest and smallest values before evaluating the standard mean of the given data.*
- After removing the specified observations, the trimmed mean is found using a standard arithmetic averaging formula.
- The use of a trimmed mean helps eliminate the influence of data points on the tails that may unfairly affect the traditional mean.
- In above example, the mean obtained by discarding the givings of Wal-Mart is a trimmed mean.

Calculating Mean for Ungrouped Frequency Distribution

Suppose that a data set contains observation values $x_1, x_2, \ldots\ldots\ldots, x_k$ occurring with frequencies, $f_1, f_2, \ldots\ldots, f_k$ respectively.

(i) For a population of N observations, so that $N = \sum_{i=1}^{k} f_i$

the mean is $\mu = \dfrac{\sum_{i=1}^{k} f_i x_i}{N}$

(ii) For a sample of n observations, so that $n = \sum_{i=1}^{k} f_i$

the mean is $\bar{x} = \dfrac{\sum_{i=1}^{k} f_i x_i}{n}$

EXAMPLE: A random sample of 50 personal property insurance policies showed the following number of claims over the past two years.

No. of claims	No. of policies
0	21
1	13
2	5
3	4
4	2
5	3
6	2

Sum	50

Calculate the mean number of claims.

Solution:

No. of claims (x_i)	No. of policies (f_i)	$f_i \times x_i$
0	21	0
1	13	13
2	5	10
3	4	12
4	2	8
5	3	15
6	2	12
Sum	$\sum f_i = 50 = n$	$\sum f_i x_i = 70$

Hence, mean number of claims= $\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{70}{50} = 1.4$

Calculating Mean for Grouped Frequency Distribution

(i) For a population of N observations, so that $N = \sum_{i=1}^{k} f_i$

the mean is $\mu = \frac{\sum_{i=1}^{k} f_i m_i}{N}$

(ii) For a sample of n observations, so that $n = \sum_{i=1}^{k} f_i$

the mean is $\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n}$

where m_i is the midpoint and f_i is the frequency of the *ith* class.

- ❖ To calculate the mean for grouped data, first find the midpoint of each class and then multiply the midpoints by the frequencies of the corresponding classes.

EXAMPLE: Calculate arithmetic mean of data of days to maturity 40 short-term investments.

Table: Data of days to maturity 40 short-term investments.

Class interval	No. of investments
30—39	3
40—49	1
50—59	8
60—69	10
70—79	7
80—89	7
90—99	4
Total	40

Solution:

Table: Calculation of arithmetic mean of data of days to maturity 40 short-term investments.

Class interval	Midpoint (m_i)	Frequency (f_i)	$f_i \times m_i$
30—39	34.5	3	103.5
40—49	44.5	1	44.5
50—59	54.5	8	436
60—69	64.5	10	645
70—79	74.5	7	521.5
80—89	84.5	7	591.5
90—99	94.5	4	378
Total		40	2720

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n} = \frac{2720}{40} = 68.00$$

EXAMPLE

The following Table gives the frequency distribution of the daily commuting times (in minutes) from home to work for *all* 25 employees of a company.

Table

Daily Commuting Time (minutes)	**Time Number of Employees**
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean of the daily commuting times.

Solution Note that because the data set includes *all* 25 employees of the company, it represents the population.

Table

Daily Commuting Time (minutes)	$\boldsymbol{f_i}$	$\boldsymbol{m_i}$	$\boldsymbol{f_i m_i}$
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
Total	$\sum f_i = N = 25$		$\sum f_i m_i = 535$

Thus

Population mean = $\boldsymbol{\mu = \frac{\sum f_i m_i}{\sum f_i} = 21.40}$ **minutes**

Thus, the employees of this company spend an average of 21.40 minutes a day commuting from home to work.

EXAMPLE

The following Table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company.

Table

Number of Orders	**Number of Days**
10–12	4
13–15	12
16–18	20
19–21	14

Calculate the mean.

Solution

Because the data set includes only 50 days, it represents a sample.

Daily Commuting Time (minutes)	f_i	m_i	$f_i m_i$
10–12	4	11	44
13–15	12	14	168
16–18	20	17	340
19–21	14	20	280
Total	$\sum f_i = n = 50$		$\sum f_i m_i$ =832

The value of the sample mean is

$$\bar{x} = \frac{\sum f_i m_i}{n} = \frac{832}{50} = 16.64 \quad \text{orders}$$

Thus, this mail-order company received an average of 16.64 orders per day during these 50 days.

Weighted Mean

Some situations require a special type of mean called a weighted mean.

The weighted mean of a set of data is $\bar{x} = \dfrac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

where w_i = weight of the *ith* observation.

- One important situation that requires the use of a weighted mean is the calculation of Grade Point Average (GPA).

EXAMPLE: Grade Point Average (Weighted Mean)
Suppose that a student completed 15 credit hours during his first semester and received following grades:

Table: Semester Academic Record

Course	Grade	Grade Point x_i	Credit hours w_i
English	A	4	3
Math	B	3	3
Biology lab	C	2	4
Spanish	D	1	5
Total			15

Calculate the student's semester GPA.

Solution: GPA calculated by the simple mean is

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{4+3+2+1}{4} = 2.5$$

But this not the correct GPA. In computing simple mean we assume that each course is of equal importance or "weight," but this assumption ignores that fact that the numbers of credit hours are not the same for all courses. So, accurate measure is the weighted mean. This information is summarized in the following Table:

Course	Grade	Value x_i	Credit hours w_i	$w_i x_i$
English	A	4	3	*12*
Math	B	3	3	*9*
Biology lab	C	2	4	*8*
Spanish	D	1	5	*5*
Total			15	*34*

Then

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{3(4)+3(3)+4(2)+5(1)}{15} = \frac{34}{15} = 2.3$$

Thus the GPA is 2.3 and not 2.5.

Median

Median is the middle most observation in a set of ordered (increasing or decreasing) observations.

As is obvious from the definition of the median, it divides a ranked data set into two equal parts. The calculation of the median consists of the following two steps:

1. Rank the data set in increasing order.

2. Find the middle term. The value of this term is the median.

Median for Individual Observation

Find the Middle Term

The median will be located in the $\frac{(n+1)}{2}$ th ordered ***position.***

EXAMPLE

The following data give the prices (in thousands of dollars) of seven houses selected from all houses sold last month in a city.

312, 257, 421, 289, 526, 374, 497

Find the median.

Solution First, we rank the given data in increasing order as follows:

257, 289, 312, 374, 421, 497, 526

Since there are seven homes in this data set, the middle term is the $\frac{7+1}{2} = 4th$ term.

Thus, the median price of a house is 374, or **$374,000**.

EXAMPLE

The following Table gives the 2020 profits (rounded to billions of dollars) of 12 companies selected from all over the world.

Table: Profits of 12 Companies for 2020

Company	**2008 Profits (billions of dollars)**
Merck & Co	8
IBM	12
Unilever	7
Microsoft	17
Petrobras	14
Exxon Mobil	45
Lukoil	10
AT&T	13
Nestlé	17
Vodafone	13
Deutsche Bank	9
China Mobile	11

Find the median for these data.

Solution First we rank the given profits as follows:

7, 8, 9, 10, 11, 12, 13, 13, 14, 17, 17, 45

Since there are seven homes in this data set, the middle observation is the $\frac{12+1}{2} = 6.5th$ observation. Thus the middle observation is in between 6th and 7^{th} observations, i.e. in between 12 and 13. The median, which is given by the average of these two values, is calculated as

Median = (12 + 13)/2 = 12.5 = $12.5 billion

- Note that if the number of observations in a data set is *odd*, then the median is given by the value of the middle term in the ranked data.
- However, if the number of observations is *even*, then the median is given by the average of the values of the two middle terms.

Calculating Median from Ungrouped Frequency Distribution

EXAMPLE: The following table shows the number of family members of 30 families.

Number of family members	Number of families
2	2
3	8
4	10
5	6
6	3
7	1

Find the median for these data.

Solution: The frequency distribution is the presentation of data in organized (ordered/ranked) and summarized form. Since there are 30 families in this data set, the middle observation is the $\frac{n+1}{2}th = \frac{30+1}{2}th = 15.5th$ observation. The middle term of a frequency distribution can be detected by using the cumulative frequency distribution as follows:

Number of family members	Number of families (f)	Cumulative frequency
2	2	2
3	8	10
4	10	20
5	6	26
6	3	29
7	1	30

Since the observation in both the 15th and 16th position is 4, so the median number family member is 4.

Calculating the Median from Grouped Data

Statisticians use the following equation to determine the median of groped data:

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

l = Lower limit of the median class,
h = width of the median class,
f = frequency of the median class
c = cumulative frequency of the class preceding the median class
n = total frequency

EXAMPLE: Calculate the median days of maturity of 40-short term investments.

Table:

Class interval	Frequency (f_i)	Cumulative frequency
30—39	3	3
40—49	1	4
50—59	8	12
60—69	10	22
70—79	7	29
80—89	7	36
90—99	4	40
Total	40	

Since there are 40 observations in this data set, the middle observation is the $\frac{n+1}{2}th = \frac{40+1}{2}th = 20.5th$ observation. Thus median observation lies between 20th and 21st observation. These two observations belong to the class 60-69. Thus the median is calculated as follows:

$$\text{Median} = 60 + \frac{10}{10}(20 - 12) = 60 + 8 = 68$$

Thus the median number of days to mature the short term invest is 68 days.

Note:

- The median gives the center of a ordered data set, with half of the data values to the left of the median and half to the right of the median.
- The advantage of using the median as a measure of central tendency is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of central tendency for data sets that contain outliers.