Data Analytics

Lecture No: 03

Lecture Objectives

- Central tendency
- Measures of Dispersion

Measures of Central Tendency

- Measures of central tendency yield information about **“Particular places or locations in a group of number”.**

- A single number to describe the characteristics of a set of data.

Summary statistics

- **Central tendency** or measures of location:
  - Arithmetic mean
  - Weighted mean
  - Median
  - Percentile

- **Dispersion**
  - Skewness
  - Kurtosis
  - Range
  - Interquartile range
  - Variance
  - Standard score
  - Coefficient of variation;

Arithmetic Mean

- Commonly called “the mean”
- It is the average of a group of numbers.
- Applicable for interval and ratio data.
- Not applicable for nominal or ordinal data.
- Affected by each value in the data set, including extreme values.
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set.

$$Mean = \frac{Sum\ of\ all\ observartions}{Total\ number\ of\ observations}$$

Population Mean

- Arithmetic mean is expressed as
  - Population mean
  - Sample mean

$$\mu = \sum \frac{X}{N} = \frac{X_1 + X_2 + X_3 + \ldots\ldots + X_N}{N}$$

Example: What is the population mean for the given weights of persons- {50, 60, 70, 100, 80, 55, 60, 65}?

∑X = 10 + 20 + 30 + 40 + 50 + 55 + 45 + 35 + 25 + 15 = 325

∑X/N = 325/10 = 32.5.

Sample Mean

- The sample mean only considers a **selected number** of observations.
- Selected data are drawn from the population.
- A sample mean is determined to get an *approximate value* that represents the entire population.
- Computing the sample mean is easy, whereas computation of the population mean a tedious process.

$$X = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + x3 \ldots + xn}{n}$$

Cont'd

- **Advantage of the mean:**
  - The mean can be used for both continuous and discrete numeric data.
- **Limitations of the mean:**
  - The mean cannot be calculated for categorical data, as the value cannot be summed.

Mean of the Grouped Data

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + \ldots + f_nX_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

$$= \frac{\sum_{i=1}^{n} f_iX_i}{\sum_i f_i}$$

$$= \sum \frac{fx}{f}, \text{ where } \sum f = n \text{ (total number of observation)}$$

where, $f_1, f_2, \ldots, f_n$ are corresponding frequencies of $x_1, x_2, \ldots, x_n$

Cont'd

Example: Consider the following frequency distribution of the salaries of 50 employees of a certain University, compute arithmetic mean.

Salary (000)[x]	40	50	60	70	80	90	Total
Number of employees [f]	20	10	8	5	4	3	50
fx	800	500	480	350	320	270	2720

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum f_i} = \frac{2720}{50} = 54.4$$

It shows that each employee of the University has 54.4 (thousand) salary, on the average.

Example

- Using the following data showing the weights (in grams) of 60 apples, calculate the mean (average weight).

Weight	65-84	85-104	105-124	125-144	145-164	165-184	185-204	Total
Number of apples	9	10	17	10	5	4	5	60

To compute the mean, first we convert the class intervals into midpoint (X)

Weight	65-84	85-104	105-124	125-144	145-164	165-184	185-204	Total
Number of apples (f)	9	10	17	10	5	4	5	60
Mid point (X)	74.5	94.5	114.5	134.5	154.5	174.5	194.5	
fx	670.5	945	1946.5	1345	772.5	698	972.5	7350

$$X = \frac{\sum_{I=1}^{7} f_i x_i}{\sum f_i} = \frac{7350}{60} = 122.5$$

Weighted Average

- Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers.

- The average you need is the **“Weighted average”**

Cont'd

$$\text{weighted average} = \frac{\sum xw}{\sum w}$$

where x is a data value and w is the weight assigned to thta data value. The sum is taken overall data values.

Example

- Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average of your scores. If the minimum average for an A is 90, will you earn an A?

$$weighted\ average = \frac{(83)(0.40) + (95)(0.60)}{0.40 + 0.60}$$

$$= \frac{32 + 57}{1} = 33.2 + 57 = 90.2$$

Geometric Mean

- Geometric mean is the n-th positive root of the product of n-positive values. Mathematically, geometric mean is defined as:

$$GM = \sqrt{x_1 + x_2 + \ldots + x_n} = \left(x_1 + x_2 + \ldots + x_2\right)^{\frac{1}{n}}$$

$$OR$$

$$GM = Antilog\left[\frac{\sum_{i=1}^{n} \log x_i}{n}\right] \quad (For\ individual\ series)$$

$$GM = Antilog\left[\frac{\sum_{i=1}^{n} f_i \log x_i}{n}\right] \quad (For\ frequency\ distribution)$$

Cont'd

- Geometric mean is preferred as a measure of central tendency for rates and ratio data.
- Geometric mean vanishes is any values of a data set/series is zero
- Geometric mean can not be computed if any value of a series is negative.

Example

The following data indicate the consumption (x000) of 7 various households in a certain locality of Peshawar. Calculate Geometric mean.

2	4	7	3	9	11	12

$$GM = (2x4x7x3x9x11x12)^{\frac{1}{7}} = (199584)^{\frac{1}{7}} = 5.7169$$

$$GM = Antilog\left[\frac{\sum_{i}^{n} \log x_i}{n}\right]$$

$$= Antilog\left(\frac{5.3001}{7}\right)$$

$$= Antilog\,(0.7572) = 5.7169$$

X	Log X
2	0.3010
4	0.6021
7	0.8451
3	0.4771
9	0.9542
11	1.0414
12	1.0792
	5.3001

Example

- The following data indicate the consumption (000) of 50 different households in a certain locality of peshawar. Calculate GM.

$$GM = \text{Antilog}\left(\frac{\sum_{i=1}^{n} f_i \log x_i}{n}\right)$$

$$= \text{Antilog}\ (34.3310/50)$$

$$= \text{Antilog}\ (0.6866)$$

$$= 4.8595$$

X	f	LogX	fLogX
2	5	0.3010	1.5051
4	7	0.6021	4.2144
7	10	0.8451	8.4510
3	15	0.4771	7.1568
9	7	0.9542	6.6797
11	4	1.0414	4.1656
12	2	1.0792	2.1584
	50		**34.3310**

Practice Session

- Find the G.M of the values 10, 25,5 and 30?
- Calculate the geometric mean of the annual percentage growth rate of profits in business corporate from the year 2000 to 2005 is given below
  - **50 , 72, 54, 82, 93**
- Find the G.M for the following data, which gives the defective screws obtained in a factory.

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

Harmonic Mean

- Harmonic mean is the reciprocal of arithmetic mean and reciprocal of the values. Mathematically, Harmonic mean is defined as:

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

$$H.M = \frac{n}{\sum_{i=1}^{n}\left(\frac{f_i}{x_i}\right)} \quad (For\ frequency\ distribution)$$

Note:

- HM is used to average the speed and ratio data.
- Harmonic mean vanishes if any values of a data set/ series is zero

Harmonic Mean

Example: The following data indicate the consumption (x000) of 7 various households in a certain locality of Peshawar. Calculate Harmonic mean.

2	4	7	3	9	11	12

$$HM = \frac{n}{\sum_{i=1}^{n}(1/x_i)}$$

$$= \frac{7}{(1/2+1/4+1/7+1/3+1/9+1/11+1/12)}$$

$$= 7/1.5115 = 4.6310$$

X	1/X
2	0.5000
4	0.2500
7	0.1429
3	0.3333
9	0.1111
11	0.0909
12	0.0833
	1.5115

Cont'd

- The following data shows the weekly consumption (000) of 50 households, calculate the Harmonic Mean.

X	f	f/x
5	5	1.000
6	7	1.167
7	10	1.429
8	15	1.875
9	7	0.778
10	4	0.400
12	2	0.167
	50	6.815

$$HM = \frac{n}{\sum_{i=1}^{n}(f_i / x_i)}$$

$$= 50 / 6.815$$

$$= 7.337$$

Note: It indicates that on the average each household consume Rs. 6.815 (000) per week

Mean

The following data shows the frequency distribution of the salary of 50 employees of a firm. Calculate the following

1. Arithmetic mean
2. Geometric mean
3. Harmonic mean

Salary (000)	5-9	10-14	15-19	20-24	25-29	30-34
Number of employees	20	10	8	5	4	3

Median

- Middle value in ordered array of numbers.
- Applicable for ordinal, interval and ratio data.
- Not applicable for nominal data.
- Unaffected by extremely large and small values

Median: Computation Procedure

- First Procedure:
  - Arrange the observations in an ordered array.
  - If there is an odd number of terms, the median is the middle term of the ordered array.
  - If there is an even number of terms, the median is the average of the middle two terms
- Second Procedure:
  - The median’s position is an ordered array is given by (n+1)/2.

Median: Example with an odd numbers of terms

Ordered Array:

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

- There are 17 items in the ordered array.
- Position of median = (n+1)/2 = (17+1)/2 = 9
- The median is the 9th term i.e., **15**
- If the 22 is replaced with 100, the median will still be 15
- If the 3 is replaced with -103, the median will remain the same.

Median: Example with an Even number of terms

Ordered Array:

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

- There are 16 items in the ordered array.
- Position of the median = (n+1)/2 = (16+1)/2 = 8.5
- The median will be between the 8th and 9th terms, **14.5**
- If 21 is replaced with 100, the median will still be 14.5
- Similarly if 3 is replaced with -90, the median will remain the same.

Median of Grouped data

$$Median = L + \frac{\frac{N}{2} - cfp}{f_{med}}(W)$$

Where:

L = *The lower limit of the median class*

Cfp = *Cumulative frequency of class preceeding the median class*

$fmed$ = *frequency of the median class*

W = *width of the median class*

N = *total number of frequencies*

Example

Class interval	Frequency	Cumulative Frequency
20-30	6	6
30-40	18	24
40-50	11	35
50-60	11	46
60-70	3	49
70-80	1	50
	N=50	

$$Median = L + \frac{\frac{N}{2} - cf\,p}{f\,med}(W)$$

$$= 40 + \frac{\frac{50}{2} 24}{11}(10)$$

$$= 40.909$$

Example: Solve by yourself – For continuous distribution

Calculate the median of the following data.

Marks	0 – 20	20 – 40	40 – 60	60 – 80	80 –100
No of Students	6	20	37	10	7

Practice:

- The information on the observed lifetimes (in hours) of 225 electrical components are given in the following frequency table. Find the mode and median of the lifetimes of the electrical components.

The lifetime of electrical components (in hours)	0-20	20-40	40-60	60-80	80-100	100-200
Frequency	10	35	52	61	38	29

Cont'd

- The following distribution table shows the number of runs scored by some top batsmen of the world in one-day international cricket matches. Find the mode and median of the given data

Runs scored by Top Batsmen	Number of Batsmen
3000 – 4000	4
4000 – 5000	18
5000 – 6000	9
6000 – 7000	7
7000 – 8000	6
8000 – 9000	3
9000 – 10000	1
10000- 11000	1

MODE

- The most frequently occurring value in a data set.
- Applicable to all the levels of data measurements
- Bimodal – Data sets that have two modes.
- Multimodal – Data set that contain more than two nodes.

Mode:

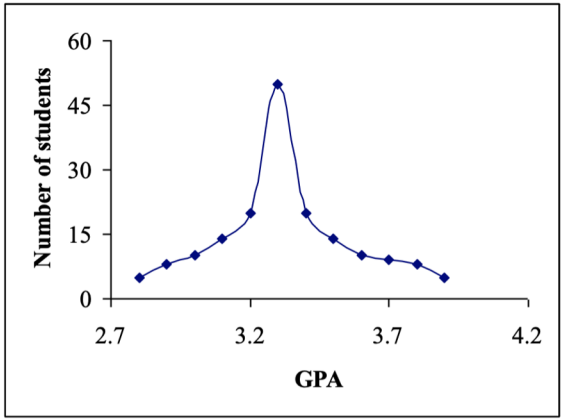
For example:

- 2, 4, 6, 4, 8, 10 (mode = 4)
- 2, 4, 6, 4, 8, 10, 8 (mode = 4 and 8)
- 2, 4, 6, 4, 8, 10, 8, 10 (mode = 4, 8 and 10)

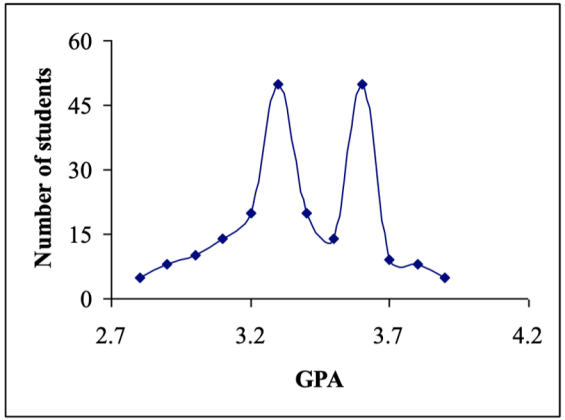
- If all the observations of a data set have the same frequencies (repeated the same number of times), the data set will have no mode.
- For example: 2, 4, 6, 4, 8, 10, 8, 10, 6: this data set has no mode.

Cont'd

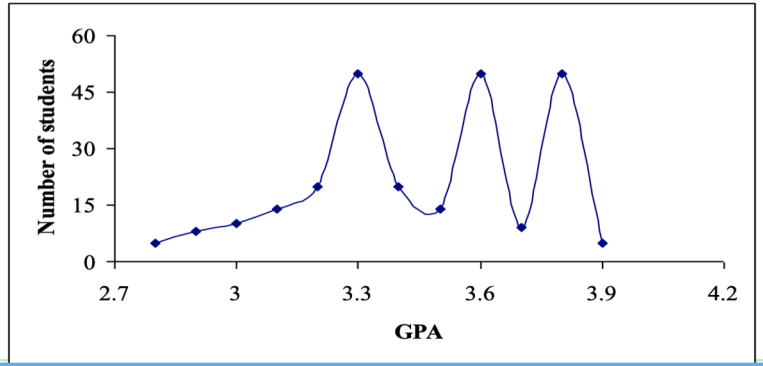
Uni-modal distribution



Bi-modal distribution



Tri-modal distribution



Mode of Grouped data

$$Mode = L + \left(\frac{f\,m - f\,1}{2f\,m - f\,1 - f\,2} \right) \times h$$

Where

$L = Lower\ limit\ of\ the\ modal\ class$

$f\,m = Frequency\ of\ the\ modal\ class$

$f\,1 = f\,requency\ of\ the\ class\ preceedinf\ modal\ class$

$f\,2 = f\,requency\ of\ the\ class\ succeeding\ modal\ class$

$f = f\,requency\ of\ median\ class$

$h = width\ of\ the\ modal\ class$

Class interval	Frequency
20-30	6
30-40	18
40-50	11
50-60	11
60-70	3
70-80	1

$$Mode = 30 + \left(\frac{18 - 6}{2 \times 18 - 6 - 11} \right) \times 10$$

$$= 36.31$$

Practice

- Find the mode of the give continuous distribution.

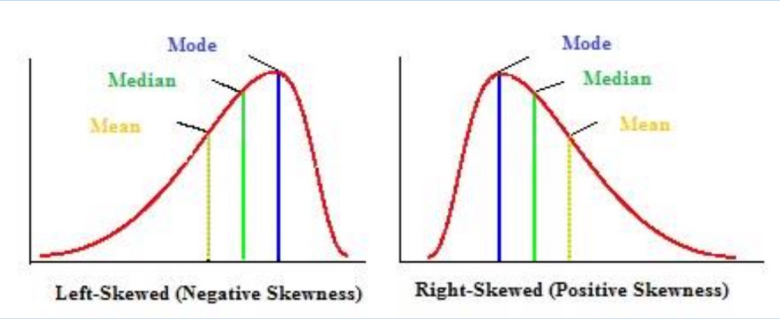
Consumption	f
4.5-9.5	10
9.5-14.5	8
14.5-19.5	20
19.5-24.5	5
24.4-29.5	4
29.5-34.5	3
	50

Skewed Distribution

- If one tail is longer than another, the distribution is skewed.
- **Skewness** defines the asymmetry of a distribution. Unlike the familiar normal distribution with its bell-shaped curve, these distributions are asymmetric.
- The two halves of the distribution are not mirror images because the data are not distributed equally on both sides of the distribution's peak.
- These distributions are sometimes called ***asymmetric*** or **asymmetrical distributions** as they don't show any kind of symmetry.
- For example, the normal distribution is a symmetric distribution with no skew. The tails are exactly the same.

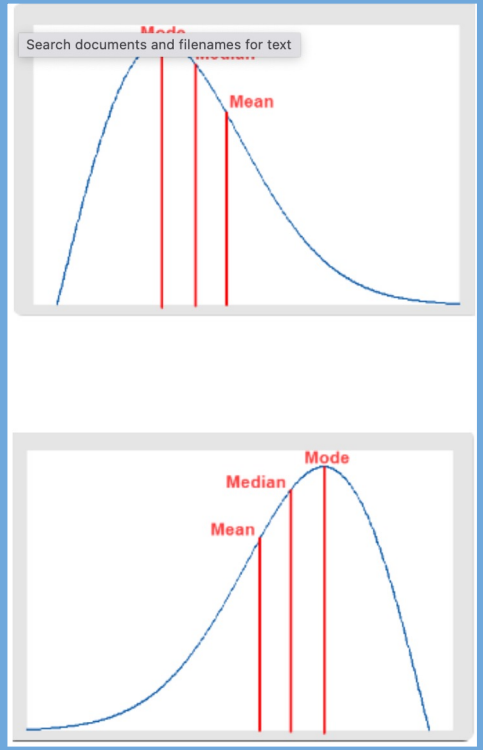
Cont'd

- A **left-skewed distribution** has a long left tail. Left-skewed distributions are also called negatively-skewed distributions. That's because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak.
- A **right-skewed distribution** has a long right tail. Right-skewed distributions are also called positive-skew distributions. That's because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak.



Cont'd

- **Right skewed:** The mean is greater than the median. The mean overestimates the most common values in a positively skewed distribution.
- **Left skewed:** The mean is less than the median. The mean underestimates the most common values in a negatively skewed distribution.



Calculations

Salary (000)	Class boundary	f	cf	X	fX	f/x	log X	f log X
5-9	4.5-9.5	20	20	7	140	2.857	0.845	16.902
10-14	9.5-14.5	10	30	12	120	0.833	1.079	10.792
15-19	14.5-19.5	8	38	17	136	0.471	1.230	9.844
20-24	19.5-24.5	5	43	22	110	0.227	1.342	6.712
25-29	24.4-29.5	4	47	27	108	0.148	1.431	5.725
30-34	29.5-34.5	3	50	32	96	0.094	1.505	4.515
		50			**710**	**4.630**		**54.4904**

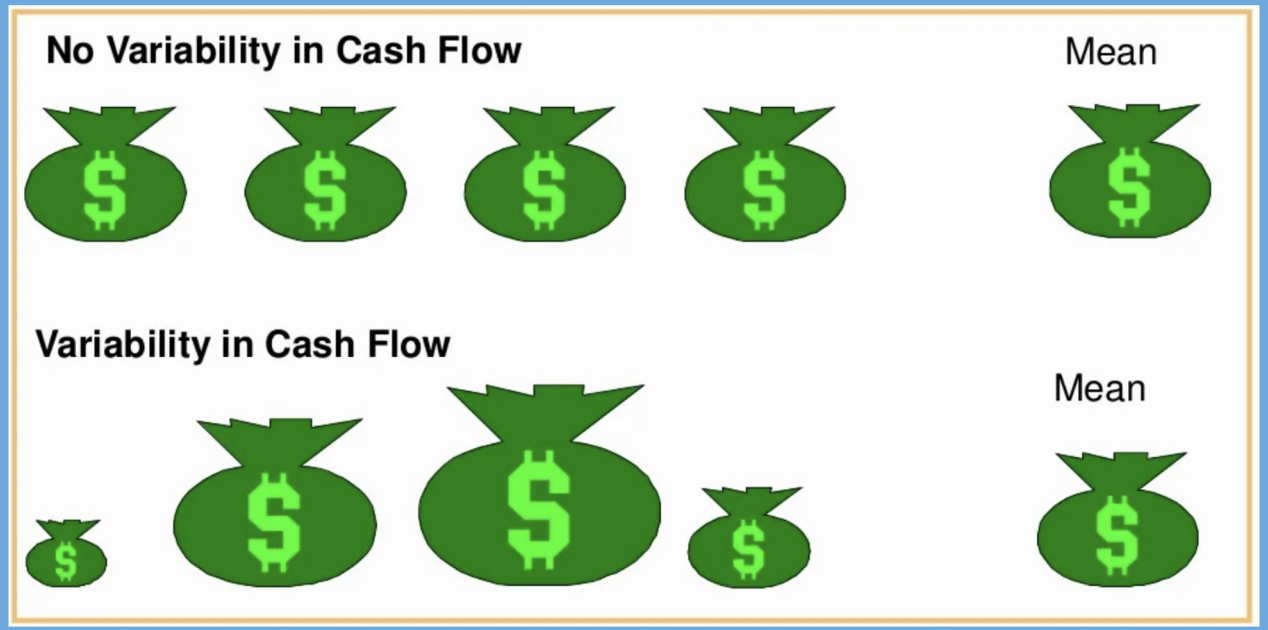
1. $$AM = \frac{\sum fX}{\sum f} = \frac{710}{50} = 14.2$$

2. $$HM = \frac{\sum f}{\sum (f/X)} = \frac{50}{4.63} = 10.8$$

3. $$GM = \text{Antilog}\left(\frac{\sum f \log X}{\sum f}\right) = \frac{54.4904}{50} = 1.090 = 12.30$$

DISPERSION

Variability

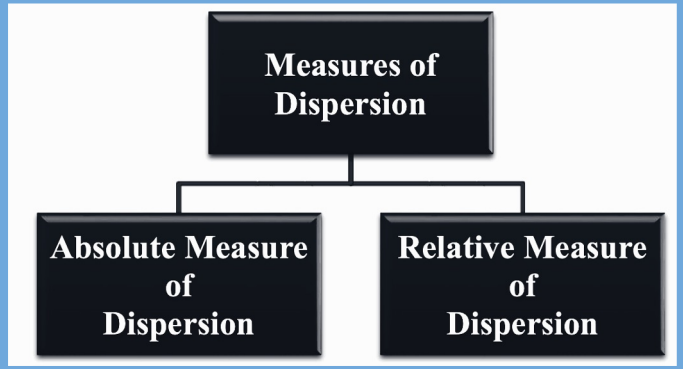


Dispersion

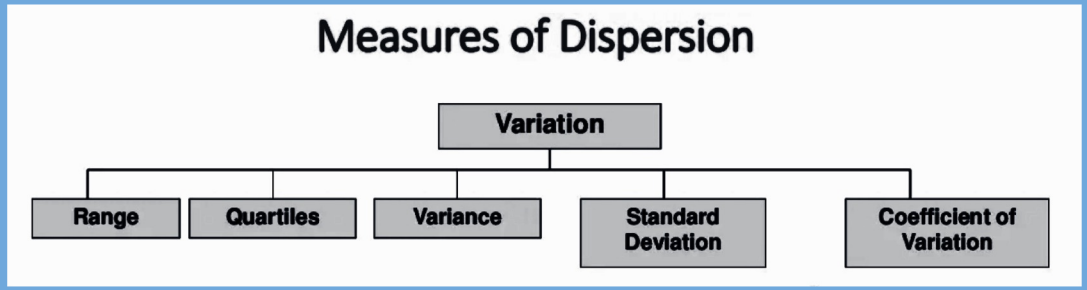
- Measure of variability describe the spread or the dispersion of a set of data.
- Reliability of measure of central tendency.
    - Measures of **central tendency (mean, median, mode, GM and HM)** do not provide all information about the observations contained in a data set
- To compare dispersion of various samples.

Cont'd

- A **single value** which measure that how the individual observations of a data set are **scattered/dispersed** around the central value, is called measure of dispersion.
- Measures of dispersion are classified as **"Absolute measures"** and **"Relative measures"** of dispersion.



Absolute measure of dispersion



- A type of dispersion which can be expressed in the same unit of measurement in which the **original series/data set/** distribution is given, is called **"Absolute measure of dispersion".**
  - Range
  - Interquartile range
  - Semi interquartile range or Quartile Deviation (QD)
  - Mean deviation
  - Variance and
  - Standard deviation

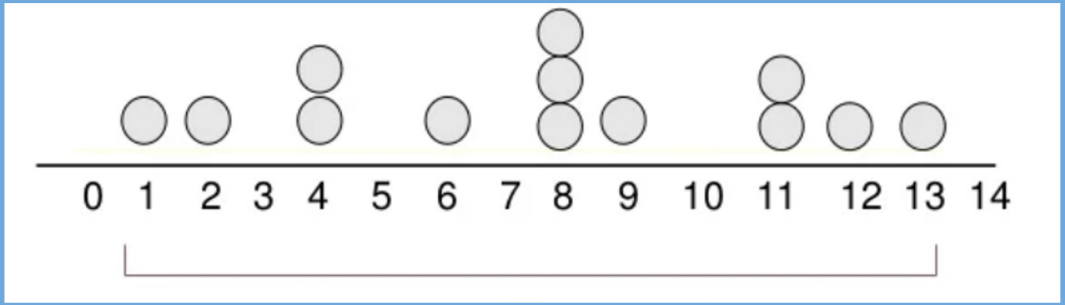
Relative measure of dispersion

- A type of dispersion which is **independent** of unit of measurement is called **"Relative measure"** of dispersion.
  - Coefficient of Range
  - Coefficient of Inter quartile range
  - Coefficient Semi interquartile range
  - Coefficient Mean deviation
  - Coefficient of Variation

Range

- Simplest measure of dispersion.
- Difference between the largest and the smallest values:

$$Range = X_{largest} - X_{smallest}$$



Range

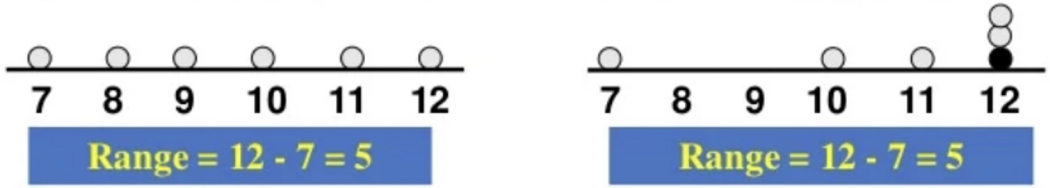
- Its relative measure is called coefficient of range and can be defined as:

$$Coef\,f\,icient\ of\ \ range = \frac{X_{largest} - X_{smallest}}{X_{largest} + X_{smallest}}$$

For Example: The following data indicate the amount of fill (in ml) of 5 different bottle by a soft drink company. The data is 12.5, 12.3, 12, 13, 12.8
So, Range = 13-12 = 1 ml.
Coefficient of Range = (13-12)/(13+12) = 1/25.

Why the range can be misleading

- Ignores the way in which data are distributed



- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Range = 120 - 1 = 119

Range

Advantages

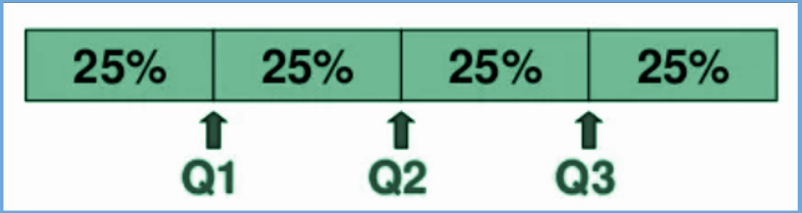
- Best for symmetric data with no outliers.
- Easy to compute and understand.
- Good option for ordinal data.

Disadvantages

- Doesn't use all of the data, only the extremes.
- Very much affected if the extremes are outliers.
- Only shows maximum spread, does not show shape.

Quartile

- Quartile split the ranked data into 4 segments with an equal number of values per segment:



- The first quartile, **Q1**, is the value of which 25% of the observation are smaller and 75% are larger.
- **Q2** is the same as the median (50% of the observations are smaller and 50% are larger.
- Only 25% of the observations are greater than the third quartile **Q3.**

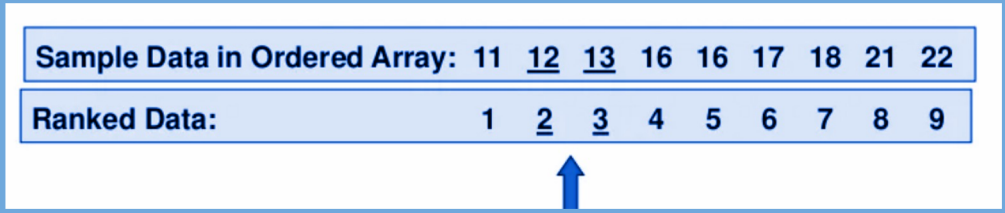
Quartile measures: Locating Quartiles

- Find the quartile by determining the value in the appropriate position in the ranked data, where:
  - First quartile position: Q1 = (n+1)/4 ranked value.
  - Second quartile position: Q2 = (n+1)/2 ranked value.
  - Third quartile position: Q3 = 3(n+1)/4 ranked value.

  Where n is the number of observed values.

Quartiles

n=9



Q1 is in the (9+1)/4 = 2.5 position of the ranked data, so use the value half way between 2nd and 3rd values. So, Q1 = 2.5.

Q2 = 16, Q3 = 19.5

Q1 and Q2 are the measure of non-central location.

Q2 = median, a measure of central tendency

Inter Quartile Range and Quartile Deviation

- **Inter Quartile Range (IQR):** is an absolute measure of dispersion. “It is the difference between upper quartile (Q3) and lower quartile (Q1) of a data set”. Mathematically, IQR is defined as:
  - **IQR = Q3 – Q1**
- **Quartile Deviation (Semi inter quartile range):** It is half of the inter quartile range. It is also called quartile deviation (QD) and is expressed as:
  - **SIQR = QD = (Q3 – Q1)/2**

Inter Quartile Range and Quartile Deviation

- A **relative measure of IQR and SIQR** is called coefficient of IQR and coefficient of SIQR (coefficient of QD), respectively and can be expressed as:
  - **Coefficient IQR = (Q3 − Q1)/ (Q3 + Q1)**
  - **Coefficient of SIQR =Coefficient of QD = (Q3 − Q1)/ (Q3 + Q1)**