Descriptive Statistics

1. measures of central tendancy
   - ↳ mean, mode, median
2. measures of dispersion / variation
   - ↳ range, quartile, IQR, variation, std deviation
3. measures of asymmetry
   - ↳ skewness, kurtosis
4. measures of relationship
   - ↳ covariance & correlation, Regression

Measures Of Central Tendency

central value of series

Measures of Central Tendency

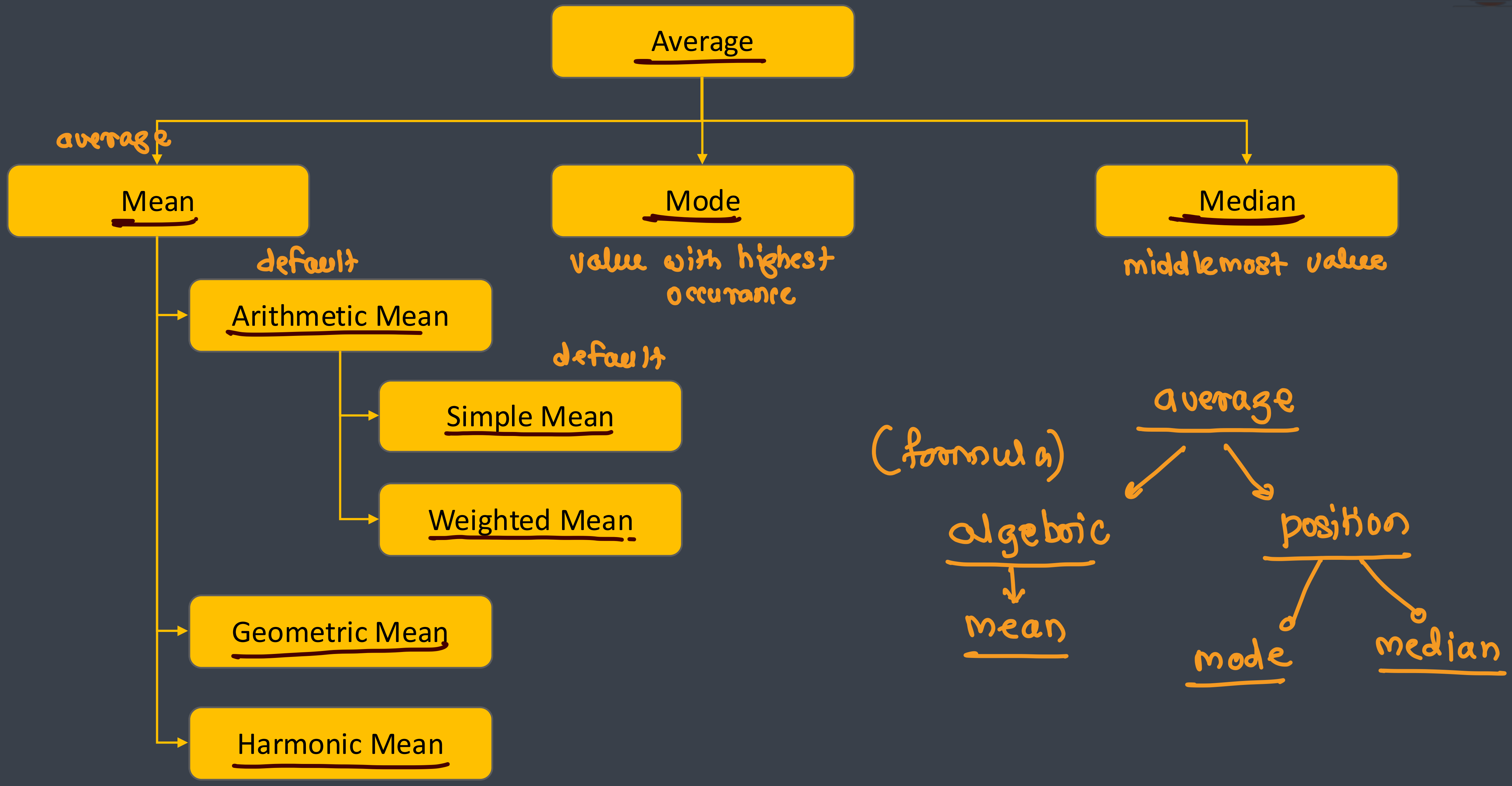
central

- One of the important objectives of statistical analysis is to get one single value that describes the characteristic of entire mass of selected data
- Such value is called as "Central Value" or "Average" or expected value of the variable
- Average
  - Average is an attempt to find one single figure to describe the whole of figures
  - Average is a single value selected from a group of values to represent them in some way
  - Average is sometimes described as a number which is typical of the whole group
- Objectives of averaging
  - To get single value that describes the characteristics of the entire group
  - To facilitate comparison

Requisites of good average

- Easy to understand
- Simple to compute
- Based on all the items
- Not be unduly affected by extreme observations
- Rigidly defined → formula
- Capable of further algebraic treatment → computed
- Sampling stability

Types of Averages



Mean

Simple Arithmetic Mean – Individual Series

- Direct method
- Steps
  - Add all the observations together and obtain the total $\sum X$
  - Divide the total by number of observations

10, 20, 30, 40, 35, 38, 29, 41,
50, 60, 65, 55, 53

$$\text{mean} = \frac{\text{sum}(\ldots)}{13}$$

mean = 40.46

$$\bar{X} = \frac{X1 + X2 + X3 \ \ldots + Xn}{N}$$

OR

$$\bar{X} = \frac{\sum X}{N}$$

Simple Arithmetic Mean – Individual Series

- Shortcut method (Using Assumed Mean)
- Steps
  - Take an assumed mean and denote it as A
  - Take the deviations of items from assumed mean and denote them by d
  - Obtain the sum of these deviations i.e. $\sum d$
  - Apply the formula

$d = x_i - A$

$$\bar{X} = A + \frac{\sum d}{N}$$

Simple Arithmetic Mean – Individual Series

- Following are the monthly income of 10 employees in an office
  - 14780, 15760, 26690, 27750, 24840, 24920, 16100, 17810, 27050, 16950
- Calculate arithmetic mean of income

using direct method

$$\text{mean} = \frac{\Sigma x}{N}$$

$$\text{mean} = \frac{14780 + 15760 + \ldots + 16950}{10}$$

$$\text{mean} = \frac{212650}{10} = \boxed{21265}$$

Using assumed mean

$A = 10000$

X	$X - A$
14780	4780
15760	5760
26690	16690
27750	17750
24840	14840
24920	14920
16100	6100
17810	7810
27050	17050
16950	6950

$\Sigma d = 112650$

$$\text{mean} = A + \frac{\Sigma d}{N}$$

$$= 10000 + \frac{112650}{10}$$

$$\boxed{\text{mean} = 21265}$$

Simple Arithmetic Mean – Discrete Series

- Direct method
- Steps
  - Multiply the frequency of each row with the variable and obtain the total $\sum fX$
  - Divide the total by number of observation that is the total frequency

$$\bar{X} = \frac{\sum fX}{N}$$

$N = \Sigma f$

- Where
  - f = frequency
  - X = observations
  - N = total frequency

Simple Arithmetic Mean – Discrete Series

- Shortcut method - Using Assumed mean
- Steps
  - Take an assumed mean and denote it by A
  - Take the deviations of the variable X from the assumed mean and denote the deviations by d
  - Multiply this deviation by respective frequency and take the total $\sum fd$
  - Apply the formula

$d = x_i - A$, $N = \Sigma f$

$$\bar{X} = A + \frac{\sum fd}{N}$$

$N = \Sigma f$

- Where
  - f = frequency
  - d = deviation from Assumed mean
  - A = assumed mean
  - N = total frequency

Simple Arithmetic Mean – Discrete Series

- From the following data of marks obtained by students, calculate arithmetic mean

Marks	20	30	40	50	60	70
# students	8	12	20	10	6	4

marks (x)	# students (f)	f·x
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
N =	60	2460

$$\text{mean} = \frac{\Sigma f \cdot x}{N}$$

$$N = 60 \ (\Sigma f)$$

$$\text{mean} = \frac{2460}{60} = 41$$

$$\boxed{\text{mean} = 41}$$

Simple Arithmetic Mean – Continuous Series

- Direct method
- Steps

$\frac{\text{lower} + \text{upper}}{2}$

  - Obtain the mid point of each class and denote it by m
  - Multiply these mid points by the respective frequency of each class and obtain $\sum fm$
  - Divide the total obtained by the sum of frequency (N)

$$\bar{X} = \frac{\sum fm}{N}$$

$N = \Sigma f$

- Where
  - f = frequency
  - m = mid point of each class
  - N = total frequency

Simple Arithmetic Mean – Continuous Series

- Shortcut method - Using Assumed mean
- Steps
  - Take an assumed mean and denote it by A
  - From the mid point of each class deduct the assumed mean
  - Multiply the respective frequencies of each class by the deviations and obtain $\sum fd$
  - Apply formula

$d = x'_i - A$, $N = \Sigma f$

$$\bar{X} = A + \frac{\sum fd}{N}$$

- Where
  - f = frequency
  - d = deviation of class mid point from assumed mean
  - A = assumed mean
  - N = total frequency

Simple Arithmetic Mean – Continuous Series

- From the following data of marks obtained by students, calculate arithmetic mean

Marks	0-10	10-20	20-30	30-40	40-50	50-60
# students	5	10	25	30	20	10

marks (x)	# students	m	fm
0-10	5	5	25
10-20	10	15	150
20-30	25	25	625
30-40	30	35	1050
40-50	20	45	900
50-60	10	55	550
	100		3300

$$\text{mean} = \frac{\Sigma fm}{N}$$

$N = 100$, $\Sigma fm = 3300$

$$\text{mean} = \frac{3300}{100} = 33$$

$$\boxed{\text{mean} = 33}$$

Mathematical Properties of Arithmetic Mean

- Sum of the deviations of the items from the arithmetic mean (taking sign into account) is always zero
- Sum of the squared deviations of the items from arithmetic mean is minimum, that is, less than the sum of squared deviations of the items from any other value
- Including the mean value in the series multiple times wont change the mean
- If we have arithmetic mean and number of items of two or more than two related groups, we can compute combined mean of these groups using formula

→ Replace all NA (missing) values by mean

2, 3, 4, 5, 6 mean = 4

x	x-mean	$(x\text{-mean})^2$
2	-2	4
3	-1	1
4	0	0
5	1	1
6	2	4
	0	10

$$\overline{X_{12}} = \frac{N_1\overline{X_1} + N_2\overline{X_2}}{N_1 + N_2}$$

2, 3, 4, 5, 6, 4, 4

m = 4 m = 4

x	x - 2	$(x-2)^2$
2	0	0
3	1	1
4	2	4
5	3	9
6	4	16
		8

$S1 = 2, 3, 4, 5, 6$, $m1 = 4$

$S2 = \underline{1, 7, 8}$ $m2 = 5.3$

$S3 = 1, 2, 3, 4, 5, 6, 7, 8 = m3 = \underline{4.5}$

$N_1 = 5, \quad N2 = 3$

$$m3 = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{5\times 4 + 3\times 5.3}{5+3}$$

$$= \frac{20 + 15.3}{8} = \underline{4.487}$$

Merits

- It is simplest average to understand and easiest to compute
- It is affected by value of every item in the series
- It is defined by rigid mathematical formula with the result that everyone who computes the average gets the same answer
- It lends itself to subsequent algebraic treatment better than median or mode
- The mean is typical in the sense that it is the center of gravity, balancing the values on the either sides of it
- It is calculated values and not based on the positions

Geometric Mean

- Steps
  - Multiply all the values and get the result
  - Get the square root to the Nth power to find the geometric mean

$$\bar{X} = \sqrt[N]{x_1 * x_2 * \ldots .* x_n}$$

Harmonic Mean

- Steps
  - Get reciprocal of each number and add together
  - Divide the number of values by the total calculated eariler

$$\bar{X} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Weighted Mean

- Steps
  - Multiply every value with corresponding weight
  - Add the values together
  - Divide the total by sum of all the weights

$$\bar{X} = \frac{\sum WiXi}{W1 + W2 \ldots. + Wn}$$

Median

Median

- By definition, it refers to the middle value in a distribution
- The median is just 50^{th} percentile value below which 50% of the values in the sample fall
- It splits the observations into two halves
- Unlike the mean, median is calculated by position (which refers to the place of the value in the series)

Median – Individual Series

- Steps
  - Arrange the data in the ascending or descending order of magnitude
  - In a group composed of an odd number of values such as 7, add 1 to the total number of values and divide it by 2. Thus 7 + 1 would be 8 which divided by 2 gives 4 – the position used to calculate the mean
  - In a group composed of even number of values such as 10, use the average of middle two values. Thus 10 / 2 gives 5 – which will produce a median by taking average of 5^{th} and 6^{th} position values

$$median = \frac{N+1}{2}$$

Median – Individual Series

- E.g. 1:
  - find median of 14100, 14150, 16080, 17120, 15200, 16160, 17400
  - Arrange them in ascending order
    - 14100, 14150, 15200, 16080, 16160, 17120, 17400
  - Median = (N + 1) / 2th item
  - Median = 7 + 1 / 2 = 4th item => 16080

- E.g. 2:
  - Find median of 19, 28, 40, 10, 29, 50, 37, 89, 90, 60
  - Arrange them in ascending order
    - 10, 19, 28, 29, 37, 40, 50, 60, 89, 90
  - Median = (N + 1)/ 2 the item
  - Median = average of 5th and 6th items => Average(37, 40) =>38.50

Median – Discrete Series

- Steps
  - Arrange the data in ascending or descending order of magnitude
  - Find out cumulative frequencies
  - Apply the formula (N + 1) / 2 the item
  - Now look at the cumulative frequency and find the total which is either equal to (N + 1) /2 or next higher to that and determine the value of variable corresponding to it
  - This gives the value of median

Median – Discrete Series

Marks	20	30	40	50	60	70
# students	8	12	20	10	6	4

Marks	#students	Cumulative frequency
20	8	8
30	12	20
40	20	40
50	10	50
60	6	56
70	4	60

- Median is (N + 1) / 2 th item => (60 + 1) / 2 = 30.5 th item
- Since the value at 30.5^{th} (or just higher than it) is 40
- Median = 40

Median – Continuous Series

- Steps
  - Determine the particular class in which the value of median lies, consider this as median class
  - Calculate the cumulative frequencies
  - Use N/2 as the rank of the median
  - Use the formula

$$median = L + \frac{\frac{N}{2} - cf}{f} * i$$

- Where
  - L = Lower limit of the median class (the class in which middle item of the distribution lies)
  - cf = cumulative frequency of the class preceding the median class
  - f = frequency of the median class
  - i = class interval of the median class

Median – Continuous Series

Marks	0-10	10-20	20-30	30-40	40-50	50-60
# students	5	10	25	30	20	10

- The median class is => 100 /2 => 50 lies in (30-40)
- Median = 30 + ((100/2 – 40) / 30) * 10
- Median = 30 + (10/30) * 10 = 33.33

Marks	#students	cf
0-10	5	5
10-20	10	15
20-30	25	40
30-40	30	70
40-50	20	90
50-60	10	100

Merits

- It is useful in case of open-end classes since only the position and not the values of the items must be known
- Median is recommended if the distribution has unequal classes
- Extreme values do not affect the median as strongly as they do the mean
- It is most appropriate average in dealing with qualitative data
- Value of median can be calculated graphically
- It represents clear-cut the middle value in the distribution

Limitations

- For calculating median, it is necessary to arrange the data in a specific order
- Since it is a middle value, its value is not determined by each and every observation
- It is not capable of algebraic treatment
- The value of median is affected more by fluctuations than the value of the arithmetic mean
- It is erratic if the number of observations is very small

Mode

Mode

- The mode or modal value is that value in a series which occurs most frequently
- That is the mode always will have the highest frequency in the data
- There are many situations where mean and median fails to reveal the true middle value, in such scenarios mode is used to find the central value

Mode – Individual Series

- Steps
  - Count the number of times the various values repeate themselves and the value occurring maximum number of times is the modal value

- E.g. 10, 28, 39, 40, 10, 20, 40, 50, 10 => mode = [10]
- E.g. 10, 20, 40, 50, 10, 20, 30, 40, 50 => mode = [10, 20, 50]
- E.g. 10, 20, 30, 40, 50, 60, 70, 80, 90 => mode = []

Mode – Discrete Series

- Steps
  - Mode can be determined just be inspection
  - i.e. by looking to that value of the variable around which the items are most heavily concentrated

- E.g.

Marks	20	30	40	50	60	70
# students	8	12	20	10	6	4

- The mode here is 40

Mode – Continuous Series

- Steps
  - Find the modal class by finding the largest value
  - Determine the value of mode by applying the following formula

$$mode = L + \frac{\Delta_1}{\Delta_1 - \Delta_2} * 1$$

- Where
  - L = Lower limit of modal class
  - Δ_1 = difference between the frequency of modal class and frequency of pre-modal class
  - Δ_2 = difference between the frequency of modal class and frequency of post-modal class

Mode – Continuous Series

Marks	0-10	10-20	20-30	30-40	40-50	50-60
# students	5	10	25	30	20	10

- Modal class here is: 30-40
- Using the formula
  - Mode = 30 + ((30-25) / ((30-25) + (30-20))) * 10
  - Mode = 30 + (5 / (5 + 10)) * 10
  - Mode = 30 + 3.33 = 33.33

Merits

- Mode is the most typical or representative value of the distribution
- Like median, mode is not unduly affected by extreme values
- It can be used to describe the qualitative phenomenon
- The value of mode can be calculated graphically

Limitations

- The value of mode can not always be determined
- It is not capable of algebraic manipulation
- The value of mode is not based on each and every value of distribution