STATISTICAL DESCRIPTION OF DATA

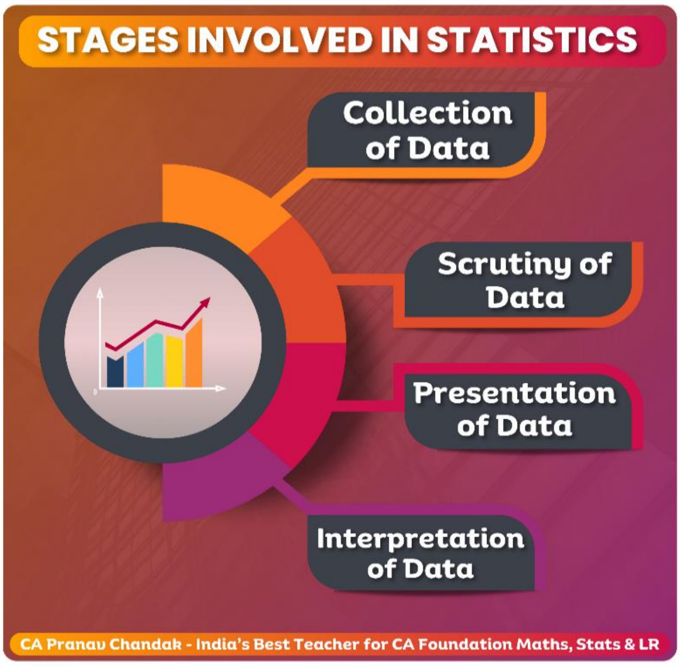
ORIGIN OF STATISTICS

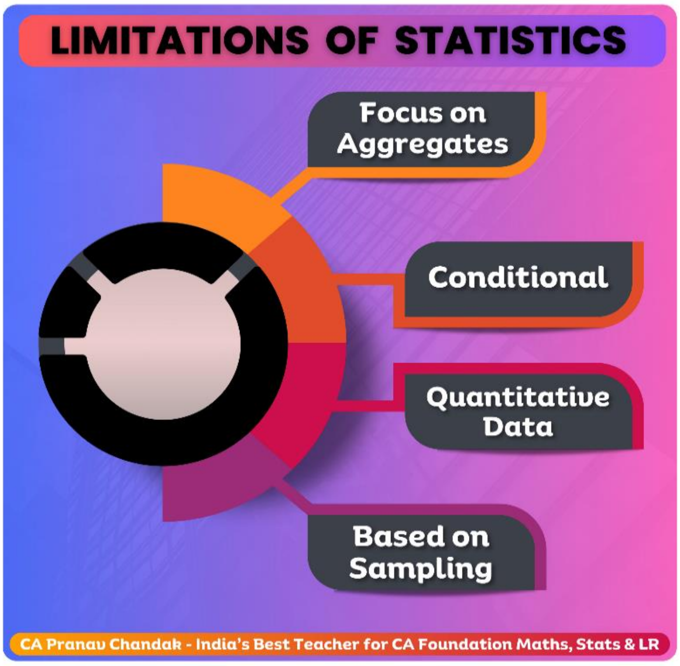
- Parent Word
  - Latin word → Status
  - German word → Statistik
  - Italian word → Statista
  - French word → Statistique

Meaning ⇒ Political State

DEFINITION OF STATISTICS

- Singular Sense → Scientific Methods
- Plural Sense → Data Collected





MEANING & TYPES OF DATA

1 Quantitative Data (Cardinal Data) ⇒ Expressed Numerically.

Variable: **Quantitative characteristic** which vary for different data.

Discrete Variable	▪ Discrete variable can have only whole number as its value. ☞ Number of students, Number of Misprints in a book, No. of Accidents. ☞ Marks of a student, Annual Income of a person. ☞ No. of shares distributed/owned. *PC Note: Discrete Variable is always a Whole Number. It cannot have fractional values.*
Continuous Variable	▪ Continuous variable can have any value (whole number or even a fraction) ☞ Height, Weight & Age of a person. ☞ Sales (revenue), profit (income) of a company. ☞ Time, Speed, Temperature. *PC Note: Continuous Variable can have fractional values.*

2	Qualitative Data (Ordinal Data) ⇒ Cannot be measured Numerically.
	❖ Attributes: Qualitative characteristic which varies for different data. ☞ Beauty of a girl, gender of a person, nationality of a person, Drinking habits of a person; ☞ Sweetness of a dish, Blood group, Hair Color, Caste/Religion, Locality of a city.

COLLECTION OF DATA

Primary Data	▪ Data collected for the first time by an investigator or agency.
Secondary Data	▪ It is a data which is already collected & then is used by a different person or agency.

Sources of Secondary Data

- International sources like WHO, ILO, IMF, World Bank etc.
- Government sources, Books, magazines, official records, census etc.
- Private & quasi-government sources like ISI, ICAR, NCERT etc.
- Unpublished sources of various research institutes, researchers etc.

METHODS OF COLLECTION OF PRIMARY DATA

1	Interview Method	
	Personal Interview	▪ Investigator meets respondents directly & collects data. ▪ Most Accurate method of data collection. ▪ Used in natural calamity (earthquake) or epidemic like plague. ▪ cannot cover a large area, time-consuming & expensive.
	Indirect Interview	▪ Investigator collects data from 3rd party who have knowledge about the situation. ▪ Less accurate than Personal Interview. ▪ Used if there are some problems in reaching respondents directly [Ex: Rail accident] ▪ It cannot cover a large area, time-consuming & expensive [Same as personal interview].
	Telephone Interview	▪ Researcher contacts interviewee over the phone to gather relevant information ▪ Pros: It is a quick & non-expensive method and has a wide coverage. ▪ Cons: Number of non-responses is maximum.
2	Observation Method	
	▪ Data are collected by direct observation or using instrument. Ex: Obtaining data on height of the students. ▪ It is the best method for data collection but it is time consuming, laborious & covers only a small area.	
3	Mailed Questionnaire Method	
	▪ Framing a well-drafted & sequenced questionnaire covering all important aspects & sending to respondents with pre-paid stamp with all necessary guidelines for filling. ▪ Pros: It has a most wide coverage (even more than telephonic interview) ▪ Cons: Non-responses are Maximum (even more than telephonic interview).	
4	Questionnaires filled & sent by enumerators (person employed to take a census of the population).	

CQ. Which method covers the widest area?

(a) Telephone interview (b) Mailed questions (c) Direct interview (d) All of these

CQ. The number of non-responses is likely to be maximum in ____ method of collecting data.

(a) Telephone interview (b) Personal interview (c) Mailed questionnaire (d) Observation

SCRUTINY OF DATA

- Scrutiny means checking the accuracy & consistency of data collected.
- To check internal consistency of data, we need to have related series so that we can cross-check them.

CLASSIFICATION OF DATA

1	Temporal or Chronological or Time Series Data	Frequency Data
	▪ Data is classified according to time.	
2	Geographical or Spatial Series Data	
	▪ Data is classified according to region.	
3	Qualitative Data [Ordinal Data] → Already Studied Earlier	Non- Frequency Data
	▪ Data is classified according to attributes/non-measurable characteristics.	
4	Quantitative Data [Cardinal Data] → Already Studied Earlier	
	▪ Data is classified according to a variables/measurable numerical size.	

Objectives of Classification of Data

- It puts data in **neat, precise & condensed form**.
- It makes **comparison possible** between various characteristics.
- Statistical analysis is possible only for the **classified data**.

PRESENTATION OF DATA

1 Textual presentation ⇒ Paragraphs [Ex: All Official reports]

- Merits of Textual Presentation
  - Very simple. Even layman can understand.
  - Observations with exact magnitude can be presented with the help of textual presentation.
  - It is the first step towards other methods.
- Demerits of Textual Presentation
  - It is dull & monotonous. **Comparison b/w different observations is not possible.**
  - For manifold classification, it is not recommended. (Ex: Population classified on gender & religion)

2 Tabular Presentation or Tabulation ⇒ Statistical table [Ex: All Official reports]

- It facilitates comparison between rows & columns.
- Statistical analysis of data is not possible without tabulation.

Structure of a Table [Very IMP for Theory MCQs]

Table should be divided into caption, Box-head, Stub & Body. (4 Parts)

1. Caption: Upper part of the table, describing columns & sub-columns.
2. Box-head: Entire upper part which includes columns & sub-column, units of measurement & caption.
3. Body: Main part of the table that contains the numerical figures.
4. Stub: Left part of the table describing the rows.

⇒ Headnote: Information about unit of measurement of data like, 'amount in rupees or $'.
⇒ Source Note: indicating the source of data presented in the table.
⇒ Foot Note: Specific feature of the table which is not self-explanatory.

3 Diagrammatic Representation of Data

- This method is **most attractive** in which data is provided by **charts, diagrams & pictures.**
- Any **hidden trend** present in the data can be **noticed**.

Line Diagram or Line Chart or Line Graph or Historiagram

- When the data **vary over time**, line diagram is **preferred.**
- **A line chart is a comparison of two variables** shown on x-axis (generally time) & y-axis (value).

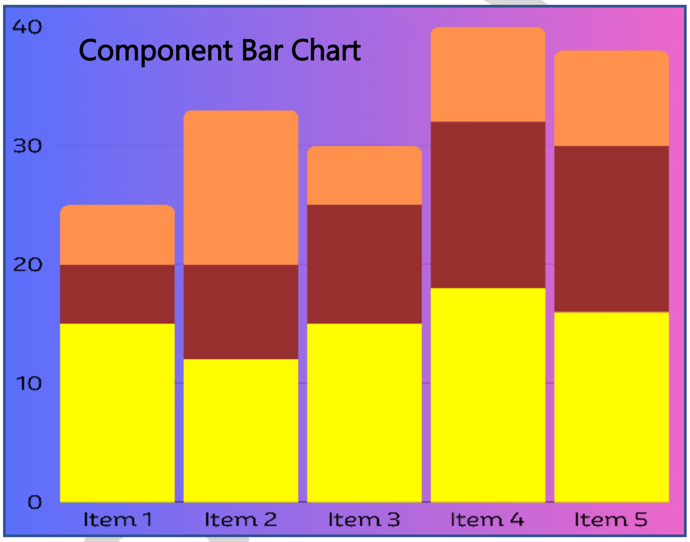
⇒ **Logarithmic or Ratio Chart:** Used when time series exhibits **wide range of fluctuations**.
⇒ **Multiple Line Chart:** Used when **two or more** related time series data is to be expressed in **same unit**.
⇒ **Multiple Axis Chart:** Used when **two or more** related time series data is to be expressed in **different unit.**

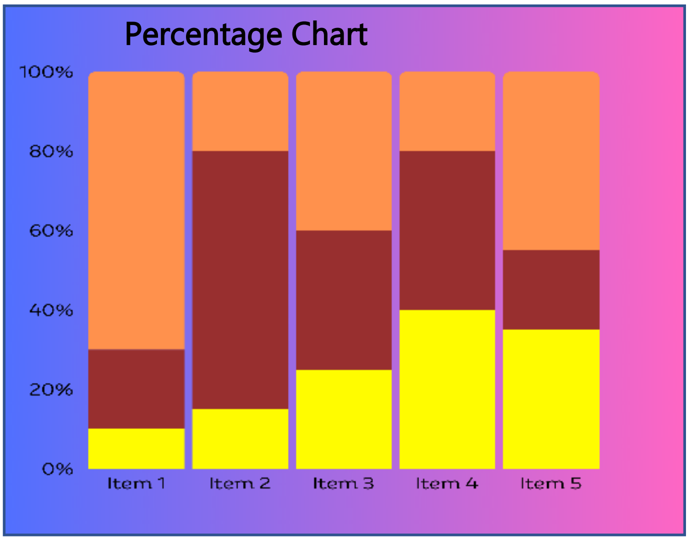
Pie Chart [Two Dimensional]

- A pie chart is a type of graph that **represents the data in the circle.**
- Whole data is **divided into 360^0** of a circle.
- Angle allotted to a Component = $\frac{\text{Value of the Component}}{\text{Total of All Components}} \times 360$
- **Pie chart is used when we want to show various components (usually 5 or more components) in a diagram.**

Bar Diagram [One Dimensional]

- Bars are rectangles of equal width & **lengths vary as per the value of a variable.**
- **Vertical Bar diagram** ⇒ Used for **Quantitative data or time series data.**
- **Horizontal Bar diagram** ⇒ Used for **Qualitative data or data varying over space.**
- **Multiple Bar Chart or Grouped Bar Chart** are used to compare related series.
- **Component or sub-divided Bar diagrams** are used for representing data which is divided into a number of components.
- **Divided Bar charts or Percentage Bar** diagrams are used for **comparing different components of a variable** & also **relating of the components to the whole.**



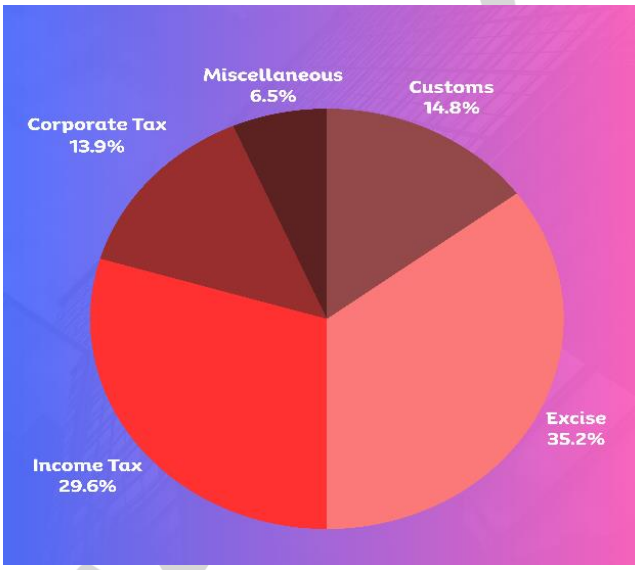


CQ3. Draw an appropriate diagram with a view to represent the following data:

Source	Customs	Excise	Income Tax	Corporate Tax	Miscellaneous
Revenue (in Rs. Millions)	80	190	160	75	35

Solution: Pie chart or divided bar chart would be ideal diagram in this case. We consider Pie chart.

Source	Revenue	Central Angle
Customs	80	$\frac{80}{540} \times 360^\circ = 53^\circ$
Excise	190	$\frac{190}{540} \times 360^\circ = 127^\circ$
Income Tax	160	$\frac{160}{540} \times 360^\circ = 107^\circ$
Corporate Tax	75	$\frac{75}{540} \times 360^\circ = 50^\circ$
Miscellaneous	35	$\frac{35}{540} \times 360^\circ = 23^\circ$
Total	**540**	360°



FREQUENCY DISTRIBUTION

- Frequency (f) of a particular value is the number of times the value occurs in the data.
- It arranges observations in increasing order, in terms of no. of groups & relates to measurable characteristic.
- Frequency distributions are portrayed as frequency tables.

Discrete (Single) Frequency Distribution		Continuous (Grouped) Frequency Distribution	
Marks	Frequency	Class Intervals	Frequency
25	12	0-5	15
30	8	5-10	12
35	4	10-15	17
40	3	15-20	19
45	2	20-25	11
50	1	25-30	26
PC Note: Used in Mutually Inclusive Classification.		PC Note: Used in Mutually Exclusive Classification.	

Continuous Frequency Distribution is of two types			
Mutually Inclusive Class Intervals (Non-Overlapping)		Mutually Exclusive Class Intervals (Overlapping)	
Weight (in Kgs)	Frequency	Weight (in Kgs)	Frequency
44 – 48	3	43.5 – 48.5	3
49 – 53	4	48.5 – 53.5	4
54 – 58	5	53.5 – 58.5	5
59 – 63	7	58.5 – 63.5	7
64 – 68	9	63.5 – 68.5	9
69 – 73	8	68.5 – 73.5	8
Total	36	Total	36

SOME IMPORTANT TERMS ASSOCIATED WITH A FREQUENCY DISTRIBUTION	
Class Limit	▪ Lower Class limit (LCL) = Minimum value of a class interval. ▪ Upper Class limit (UCL) = Maximum value of a class interval. PC Note: For Overlapping Class Intervals, Class limits & Class boundaries are same.
Class Boundary	▪ Class boundaries may be defined as the **actual class limit of a class interval.** ▪ Overlapping classification [Mutually exclusive] → Class boundaries & Class Limits are same. Ex: 1-10, 10-20, 20-30 & so on. ▪ Non-overlapping classification [Mutually inclusive] Ex: 0-9, 10-19, 20-29 & so on. Class boundaries & Class Limits are **different.** LCB = LCL – 0.5 & UCB = UCL + 0.5
Mid-Point	▪ Also known as Mid-Value or Class Mark. Mid-point = $\frac{LCL+UCL}{2} = \frac{LCB+UCB}{2}$
Width of CI	▪ Width of a Class interval = UCB - LCB of that class interval. Also known as "**Class Length**".
Frequency Density	▪ Ratio of the frequency of that CI to corresponding class length. $\frac{\text{Frequency of CI}}{\text{Class Length}}$
Relative frequency	▪ Ratio of class frequency to total frequency. $\frac{\text{Class Frequency}}{\text{Total Frequency}}$
Percentage Frequency	▪ Ratio of class frequency to total frequency, expressed as percentage. $\frac{\text{Class Frequency}}{\text{Total Frequency}} \times 100$

CUMULATIVE FREQUENCY

- Cumulative frequency is calculated by adding each frequency from a frequency distribution table to the sum of its predecessors. **Cumulative frequency may be (i) Less than CF or (ii) More than CF.**
- **Let us take an example:**

Weight (in Kgs)	No. of Students
44 – 48	3
49 – 53	4
54 – 58	5
59 – 63	7
64 – 68	9
69 – 73	8
Total	36

[First, we will have to convert CIs into Mutually Exclusive (overlapping) CI.

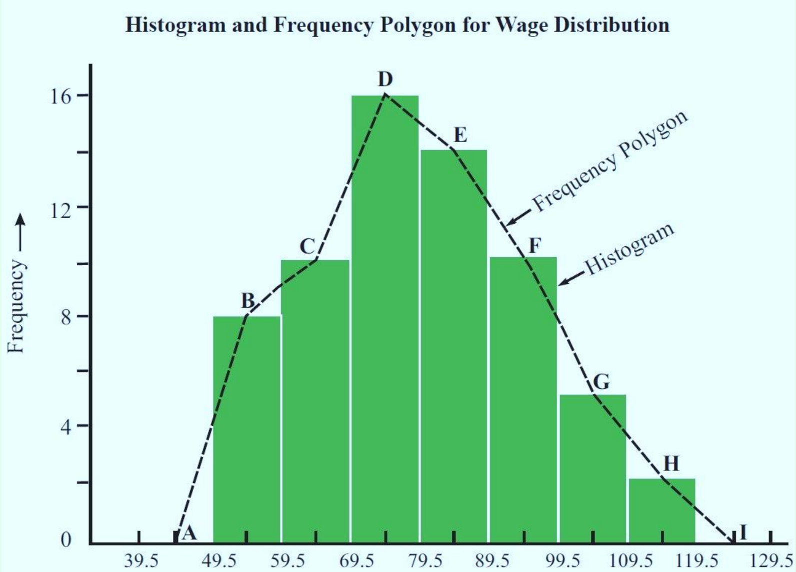
Mutually Exclusive CI	Weight in Kg	Less than CF	More Than CF
43.5 – 48.5	43.5	0	33+3 or 36
48.5 – 53.5	48.5	0+3 or 3	29+4 or 33
53.5 – 58.5	53.5	3+4 or 7	24+5 or 29
58.5 – 63.5	58.5	7+5 or 12	17+7 or 24
63.5 – 68.5	63.5	12+7 or 19	8+9 or 17
68.5 – 73.5	68.5	19+9 or 28	0+8 or 8
	73.5	28+8 or 36	0

GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTION

Histogram or Area Diagram

- Histogram is graphical representation of a **continuous (grouped) frequency distribution.**
- It is an **area diagram. Mode** can be obtained using a **histogram.**

  PC Note: Length = Frequency of CI & Width = Class Interval.
- **Comparison among class frequencies for different CIs** is possible only in Histogram.



PC Note:

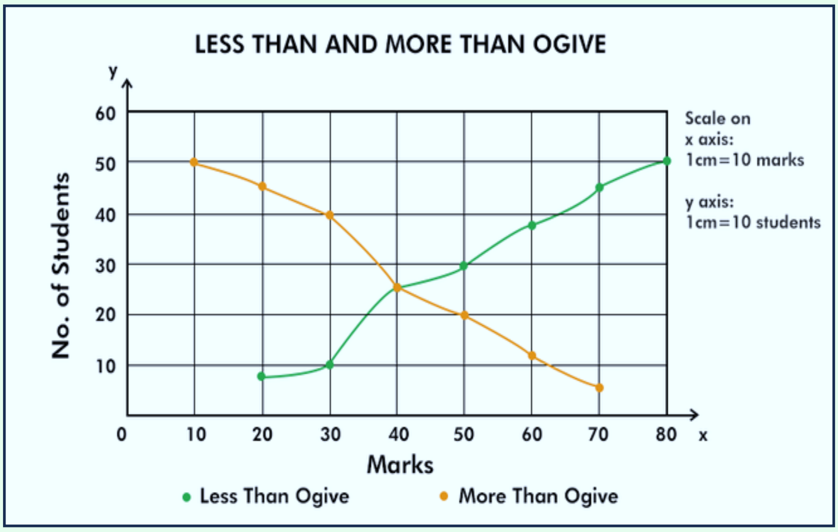
- For constructing a histogram, **class-intervals** may be **equal or unequal.**
- To draw Histogram, frequency distribution should be exclusive type. (No Gap)
- We use **frequency density** to plot histograms.

Frequency Polygon

- It is meant for **single (discrete) frequency distribution.**
- However, it can also be used for **grouped frequency distribution if width of class interval is same.**
- **A frequency curve can be regarded as a limiting form of frequency polygon.**
- It can be drawn by **plotting mid-points** of each class on the graph at the **height of frequency on y-axis** & then connecting dots with a line & joining 2 extreme ends to lowest & highest value.

Ogives or CF Graph

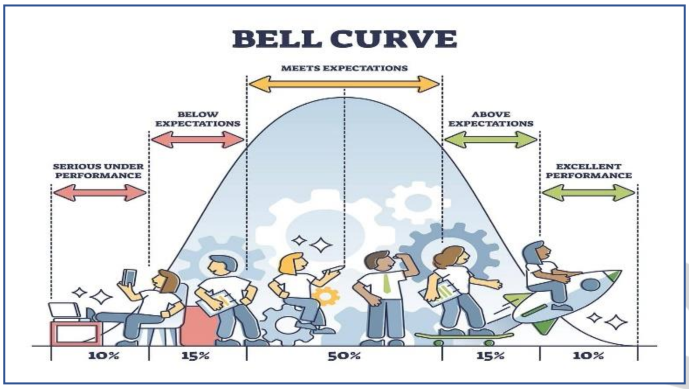
- **Ogive is a graph of a cumulative frequency distribution**, which explains data values on horizontal axis (i.e x-axis) & either the cumulative relative frequencies, cumulative frequencies or cumulative per cent frequencies on the vertical axis (i.e y-axis).

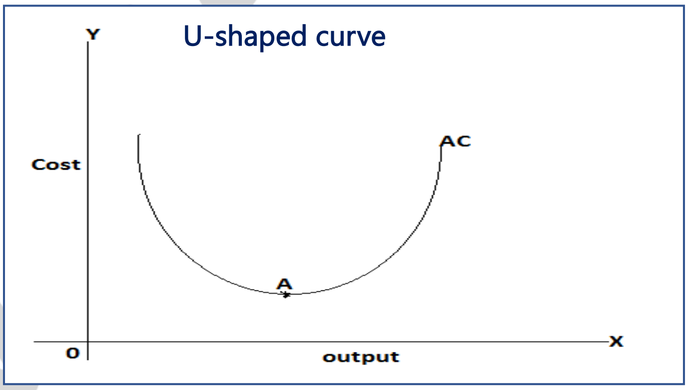


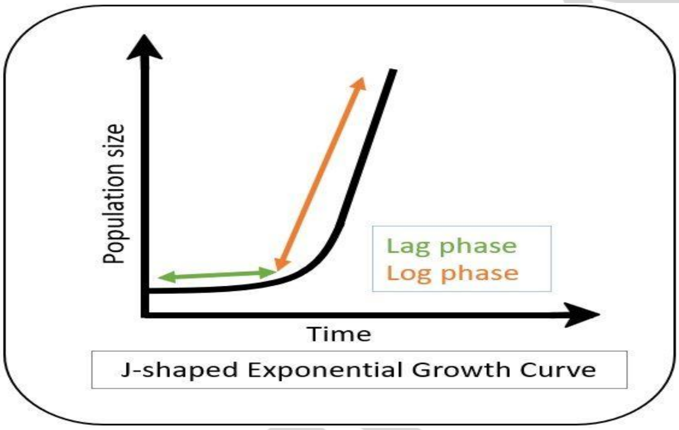
Point of Intersection of Ogives = **Median**

Frequency Curve

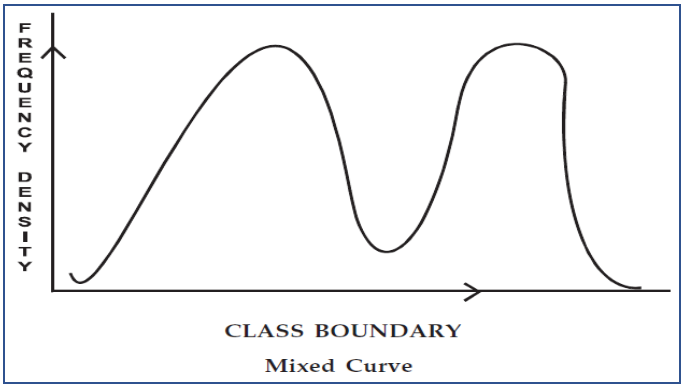
- A frequency curve is a **smooth curve** for which the **total area is taken to be unity (i.e 1)**.
- **It is a limiting form of histogram or frequency polygon.** [MCQ Point]
- It is formed by drawing a **smooth & free hand curve through the mid-points of histogram.**
- There exist **four types of frequency curves** namely
  1. **Bell-Shaped Curve:** Frequency starts from a low value, gradually reaches maximum value, somewhere near central part & then gradually decreases to reach its lowest value at the other extremity. **Ex: Distribution of height, weight, mark, profit of a company.**
  2. **U-shaped curve:** Frequency is minimum near central part & frequency slowly but steadily reaches its maximum at two extremities. **Ex: Distribution of Kolkata bound commuters as there are maximum commuters during peak hours in morning & in evening.**
  3. **J-shaped curve:** It starts with a minimum frequency & then gradually reaches its maximum frequency at the other extremity. **Ex: Distribution of commuters coming to Kolkata from early morning hours to peak morning hour.**
  4. **Mixed curve:** A combination of these frequency curves is known as mixed curve.







J-shaped Exponential Growth Curve



Mixed Curve

Q1. 20% of total employees were females & 40% of them were married. 30 female workers were not members of Trade Union. Compared to this, out of 600 male workers 500 were members of Trade Union & 50% of male workers were married. Unmarried non-member male employees were 60 which formed 10% of total male employees. Unmarried non-members of employees were 80. Find ratio of married male non-members to married female non-members.

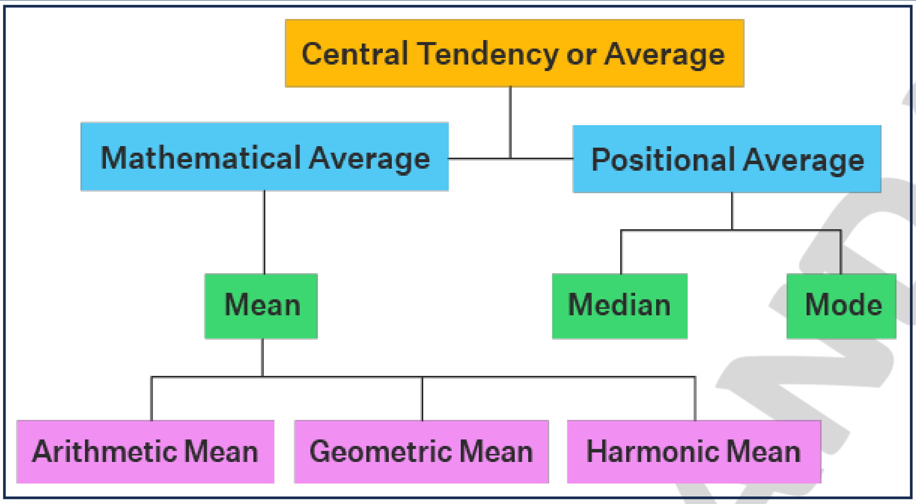
(a) 1:3 (b) 3:1 **(c) 4:1** (d) 5:1

Q2. Find the number of observations between 250 and 300:

Value (Greater than)	200	250	300	350
Frequency	56	38	15	0

(a) 56 **(b) 23** (c) 15 (d) 8

MEASURES OF CENTRAL TENDENCY



ARITHMETIC MEAN

1	**Individual Series**	$\frac{\text{Sum of All Observations}}{\text{Number of Observations}} = \frac{\Sigma X}{N}$
2	**Discrete Series (with frequency)**	$\frac{\text{Sum of All [f.x]}}{\text{Sum of Frequency}} = \frac{\Sigma(f.x)}{\Sigma f}$ **Use Calculator M+ & MRC to calculate**
3	**Continuous Series (with Class Intervals)**	**Formula is same as 2.** **PC Note: Take midpoint of CI as 'x' & use same formula as in 2.**
PC Note: We don't use Step deviation method as we are allowed to use the calculator.		

COMBINED AM $\Rightarrow \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ [For two groups containing n_1 & n_2 observations, $\overline{x_1}$ & $\overline{x_2}$ as their AMs,]

PROPERTIES OF ARITHMETIC MEAN

1	**AM of a constant series:** If all observations are same (let's say **a**), then AM is also 'a'.
2	**Sum of Deviations from AM = Zero** ▪ Individual series ⇒ $\sum(x-\bar{x}) = 0$ ▪ Continuous Series ⇒ $\sum f.(x-\bar{x}) = 0$
3	**Weighted AM** = $\frac{\Sigma(W_i.X_i)}{\Sigma(W_i)}$ [Where **'w'** → **weights** assigned to observations]

Q1. Compute the mean of Marks obtained by students: **[Ans: 45/10 = 4.5]**

Marks (x)	3	4	5	7
No. of Students (f)	4	1	3	2

Q2. Compute mean weight of a group of BBA students of St. Xavier's College from following data: **[Ans: 61.42]**

Weight (in kgs)	44 - 48	49 - 53	54 - 58	59 - 63	64 - 68	69 - 73
No. of Students	3	4	5	7	9	8

Q3. Compute mean weight of a group of BBA students of St. Xavier's College from following data: **[Ans: 416.71]**

Class Interval	350 - 369	370 - 389	390 - 409	410 - 429	430 - 449	450 - 469	470 - 489
Frequency	23	38	58	82	65	31	11

Q4. Mean salary for 40 female workers is Rs. 5200 & for 60 male workers is Rs. 6800. Find their combined salary?

MODE ⇒ Maximum Repetition or Highest frequency

1	**Individual Series**	**Mode** = Observation which is **repeated highest number of times.**
2	**Discrete Series (with frequency)**	**Mode = Observation corresponding** to the **highest Frequency.**
3	**Continuous Series (with Class Intervals)**	$L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C$

L = LCB of Modal class; **C** = Class length of Modal class
f_1 = Frequency of Modal class; f_0 = Frequency of Pre-modal class
f_2 = Frequency of Post-modal class; **Modal Class** = Max. Frequency

PC Note: Mode can have **no value, one** (uni-modal) or **more than one** value (multi-modal). **[Not Uniquely Defined]**

Q5. Find Mode for the given set of observations:

Marks (x)	3	4	5	7	6	8
No. of Students (f)	4	1	3	5	4	2

Q6. Find Mode for the given set of observations.

Marks (x)	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students (f)	4	1	3	5	4	2

Q7. Find Mode for the given set of observations: **[Ans: Modal Class is 409.50 - 429.50 & Mode = 421.21]**

CI	349.5-369.5	369.5-389.5	389.5-409.5	409.5-429.5	429.5-449.5	449.5-469.5	469.5-489.5
F	23	38	58	82	65	31	11

MEDIAN

❖ **'Middlemost value'** [Observations are arranged in ascending order].

Best for Open-ended Class

❖ 50% observations are smaller than median & 50% are larger than median.

❖ **How to find the Median:** **[Only for Theory MCQs]**
- If No. of given observations are **odd** → **Median = Middle observation.**
- If No. of given observations are **even** → **Median = Average of 2 middle observations.**

1	**Individual Series**	**Value of** $\left(\frac{n+1}{2}\right)^{th}$ **observation**
2	**Discrete Series (with frequency)**	**Value of** $\left(\frac{n+1}{2}\right)^{th}$ **frequency**
3	**Continuous Series (with Class Intervals)**	$L + \frac{\frac{N}{2} - m}{f} \times c$

L = Lower Class Limit of Median Class;
f = frequency of Median Class
m = CF of Class preceding Median Class
c = Class width of Median Class

How to Calculate Fractional Term
- $T_{3.5} = T_3 + 0.5 \times (T_4 - T_3)$
- $T_{7.2} = T_7 + 0.2 \times (T_7 - T_6)$

Q8. Compute the median of Marks obtained by students:

Marks (x)	3	4	5	7	6	2
No. of Students (f)	4	1	3	2	4	5

Q9. Compute the median for the given distribution: **[Ans: 418.0366]**

Class Intervals	350-369	370-389	390-409	410-429	430-449	450-469	470-489
Frequency	23	38	58	82	65	31	11

PC Note: Mean Deviation is Minimum when taken from Median.

PARTITION VALUES – QUARTILES, DECILES & PERCENTILES

Particulars	Quartiles	Deciles	Percentiles
Divide series into	**4** equal parts	**10** equal parts	**100** equal parts
No. of Partition values	**3**	**9**	**99**
Denoted as	$Q_1, Q_2, \& Q_3$	$D_1, D_2 \ldots\ldots, D_9$	$P_1, P_2 \ldots\ldots, P_{99}$
Each partition contains	25% of data	10% of data	1% of data
For Individual Series	**$k(N+1)/4^{th}$ term**	**$k(N+1)/10^{th}$ term**	**$k(N+1)/100^{th}$ term**
For Frequency Data	Value of **$k(N+1)/4^{th}$** frequency	Value of **$k(N+1)/10^{th}$** frequency	Value of **$k(N+1)/100^{th}$** frequency
For Continuous Data	$Q_k = L + \frac{\frac{kN}{4} - m}{f} \times c$	$D_k = L + \frac{\frac{kN}{10} - m}{f} \times c$	$D_p = L + \frac{\frac{kN}{100} - m}{f} \times c$
Value of K	**K= 1, 2, 3**	**K= 1, 2……,9**	**K= 1, 2……,99**

Q10. Find Q_1, D_6, P_{82} from: 82, 56, 90, 50, 120, 75, 75, 80, 130, 65 **[Ans: Q_1 = 62.75; D_6 = 81.20; P_{82} = 120.20]**

Q11. Following distribution is of distribution of monthly wages of 100 workers. Compute **Q_3, D_7 & P_{23}**.

Wages in Rs.	< 500	500 - 699	700 - 899	900 - 1099	1100 - 1499	> 1500
No. of Workers	5	23	29	27	10	6

CHANGE OF ORIGIN & CHANGE OF SCALE

- **Mean, Median & Mode ⇒ Affected by 'Change of Origin & Change of Scale'**
- If **Y = a + b.X,** then **AM(Y) is given by $\bar{Y} = a + b\bar{X}$**
- **PC Note:** We have to write 'y' in terms on 'x' & then substitute $\bar{X}$ as x & $\bar{Y}$ as y.

Q12. x & y are related by **2x + 3y + 7 = 0** & $\bar{x}$ = 4, then $\bar{y} = \frac{-7 - 2(4)}{3} = \frac{-15}{3} = -5$

EMPIRICAL RELATIONSHIP BETWEEN MEAN, MEDIAN & MODE

- Symmetrical Distribution: **Mean = Median = Mode.**
- Asymmetrical Distribution/Moderately Skewed Distribution: **Mode = 3 Median - 2 Mean.**

Q13. For a moderately skewed distribution of marks in statistics for a group of 200 students, the mean mark & median mark were found to be 55.60 & 52.40. What is the modal mark? **[Ans =** 46]

GEOMETRIC MEAN ⇒ (Product of All Observations)$^{1/n}$

1. Individual Series ⇒ GM = $(x_1 . x_2 . x_3 \ldots\ldots\ldots . x_n)^{1/n}$

Q14. Find the GM of 3, 6 & 12.

2. Discrete Series ⇒ GM = $[(x_1)^{f_1} . (x_2)^{f_2} . (x_3)^{f_3} \ldots . (x_n)^{f_n}]^{\frac{1}{N}}$

Q15. Find GM of the given set of observations:

X	2	4	8	16
f	2	3	3	2

PROPERTIES OF GM

- Geometric mean is **best for reporting average inflation, percentage change & growth rates.**
- Geometric Mean = 0 if one of the observations is 0.
- **If z = xy then GM (z) = GM (x) × GM (y)]**
- **If** $z = \frac{X}{Y}$ **then GM of z =** $\frac{\text{GM of x}}{\text{GM of y}}$]
- **Weighted GM** = $\text{Antilog} \frac{\sum W_i \times \log x_i}{\sum w_i}$

HARMONIC MEAN ⇒ Reciprocal of AM of the reciprocals of the observations.

1. Individual Series

HM = $\frac{N}{\frac{1}{x_1}+\frac{1}{x_2}....+\frac{1}{x_n}} = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$

Q16. Find HM for 4, 6 & 10.

2. Discrete/Frequency Data Series

HM = $\frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_n}{x_n}}$

Combined HM = $\frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}$

Q17. Find the Harmonic Mean:

X	2	4	8	16
f	2	3	3	2

PROPERTIES OF HM

- For all the observations at constant (say a), HM = a
- If any of the values of a given series is 0, then HM cannot be determined as the reciprocal of 0 doesn't exist.
- HM can also be evaluated for the series having any negative values.
- **Weighted HM** = $\frac{\sum w_i}{\sum\left[\frac{w_i}{x_i}\right]}$ [Where w, represents the weights assigned to observations]

RELATION BETWEEN AM, GM & HM

- For set of non-zero positive values (& all values are not equal) ⇒ **AM > GM > HM**
- If all values are equal, **AM = GM = HM**
- If a & b are two positive numbers, then
- **$(GM)^2$ = AM x HM**

AM = $\frac{a+b}{2}$	GM = $\sqrt{ab}$	HM = $\frac{2}{\frac{1}{a}+\frac{1}{b}}$ =

CQ18. If AM & HM of the data sets are 25 & 4 respectively, then find GM.

MEASURES OF DISPERSION

- **Dispersion** ⇒ **Amount of deviation of the observations** from an appropriate measure of CT.
- Dispersion = Zero **if all the observations are same.**
- **Two Distributions may be identical i.r.o central tendency but they may differ on measures of dispersion.**

DIFFERENT MEASURES OF DISPERSION

Absolute Measures	Relative Measures
▪ Range ▪ Mean Deviation ▪ Standard Deviation ▪ Quartile Deviation	▪ Coefficient of Range. ▪ Coefficient of Mean Deviation. ▪ Coefficient of Variation. ▪ Coefficient of Quartile Deviation.
1. Dependent on the unit of Variable. 2. Not used to compare two or more distributions.	1. Unit Free [They do not have unit]. 2. Used for comparing two or more distributions

RANGE & CO-EFFICIENT OF RANGE

Range	Coefficient of Range
Range = Largest – Smallest (L – S)	**Co-efficient of Range =** $\frac{L-S}{L+S}$ **x 100**

PC Note: For Continuous Frequency Distribution, **Range = UCB of Highest Class – LCB of Smallest Class.**

MEAN DEVIATION & CO-EFFICIENT OF M.D

Type of Series	Mean Deviation (M.D)	Coefficient of M.D
Individual Series	$M.D = \frac{\sum [\,\lvert x-A \rvert\,]}{N}$ **[A is any central tendency]**	$\frac{MD}{CT}$ x 100
Series with Frequency	$M.D = \frac{\sum [f.\ \lvert x-A \rvert\,]}{N}$ **[A is any central tendency]**	$\frac{MD}{CT}$ x 100

PC Note: For Continuous Distribution, **(series with class intervals), take mid-point as 'x' & use same formula.**

QUARTILE DEVIATION & ITS COEFFICIENT

Quartile Deviation	Coefficient of QD
$\frac{Q_3 - Q_1}{2}$	$\frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

STANDARD DEVIATION & COEFFICIENT OF VARIATION

Type of Series	Standard Deviation (S.D)	Coefficient of Variation
Individual Series	$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$ **OR** $\sqrt{\frac{\sum x^2}{n} - (\bar{x})^2}$	$\frac{SD}{AM}$ x 100
Series with Frequency	$SD = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}}$ **OR** $\sqrt{\frac{\sum f(x)^2}{N} - (\bar{x})^2}$	$\frac{SD}{AM}$ x 100

PC Note: For Continuous Distribution, **(series with class intervals), take mid-point as 'x' & use same formula.**

CHANGE OF ORIGIN & CHANGE OF SCALE

- **Range, MD, QD, SD ⇒ Not affected by change of origin but affected by change of scale.**
- **If two variables x & y are related as y = a + bx & Range of X is given, then Range of $y = |b|.R_x$.**

PC Note: Write the Equation in terms of Y. Find Change of Scale & Multiply by CT of X. (Ignore Origin)

Q1. If the relationship between x & y is given by 2x + 3y = 10 & range of x is Rs. 15. Find range of y.

Q2. If the quartile deviation of x is 6 & 3x + 6y = 20, what is the quartile deviation of y?

PC NOTE FOR THEORY MCQS

- Range is **always positive**.
- If all observations are equal, Range = Zero.
- MD is AM of absolute deviations of all items of distribution from a measure of central tendency.
- QD is the **best measure of dispersion for open-end classification.**
- QD is less affected due to sampling fluctuations.
- SD is known as **ROOT MEAN SQUARE DEVIATION. It is the "Best" measure** of dispersion.
- If all observations are **equal, then SD= Zero.**

SD of only 2 observations $= \frac{L-S}{2} = \frac{Range}{2}$

⇒ **SD of first 'n' consecutive number** $= \sqrt{\frac{n^2-1}{12}}$

Combined SD = $\sqrt{\frac{n_1s_1^2+n_2s_2^2+n_1d_1^2+n_2d_2^2}{n_1+n_2}}$

⇒ **Variance = (SD)^2** or **SD =** $\sqrt{Variance}$

Q3. What is the range & its coefficient for following distribution?

Weights in Kgs.	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
No. of Students	12	18	23	10	3

Q4. Find mean deviation about AM & also coefficient of MD:

X	1	3	5	7	9
F	5	8	9	2	1

Q5. Compute the coefficient of mean deviation about median:

Weight in kgs.	40-50	50-60	60-70	70-80
No. of persons	8	12	20	10

Q6. Find an appropriate measures of dispersion from the following data:

Daily wages	upto 20	20 - 40	40-60	60-80	80-100
No. of workers	5	11	14	7	3

Q7. Find the SD of the following distribution:

Weight (kgs.)	50 – 52	52 – 54	54 – 56	56 – 58	58 – 60
No. of Students	17	35	28	15	5

Q8. For a group of 60 boy students, mean & SD of marks are 45 & 2 respectively. For a group of 40 girls, mean & SD of marks are 55 & 3 respectively. What is mean & SD of marks if two groups pooled together?

Q9. Calculate mode if Median 23, variance 100, Coefficient of variation 50%.

Q10. If AM & coefficient of variation of x are 10 & 40 respectively, what is the variance of (15 – 2x)?

TRY AT HOME MASTER QUESTIONS

Q11. Find co-efficient of variation from the following data:

Age	Under 10	Under 20	Under 30	Under 40	Under 50	Under 60
No. of persons Dying	10	18	30	45	60	80

Q12. The mean & SD of the salaries of the two factories are provided below:

Factory	No. of Employees	Mean Salary	SD of Salary
A	30	Rs. 4,800	Rs.10
B	20	Rs. 5,000	Rs. 12

(i) Find the combined mean salary and standard deviation of salary?
(ii) Examine which factory has more consistent structure so far as satisfying its employees are concerned?

Q13. A student computes the AM & SD for a set of 100 observations as 50 & 5 respectively. Later on, she discovers that she has made mistake in taking one observation as 60 instead of 50. Find correct mean & SD if
(a) The wrong observation is left out? (b) The wrong observation is replaced by the correct observation?

CORRELATION & REGRESSION

MAGIC

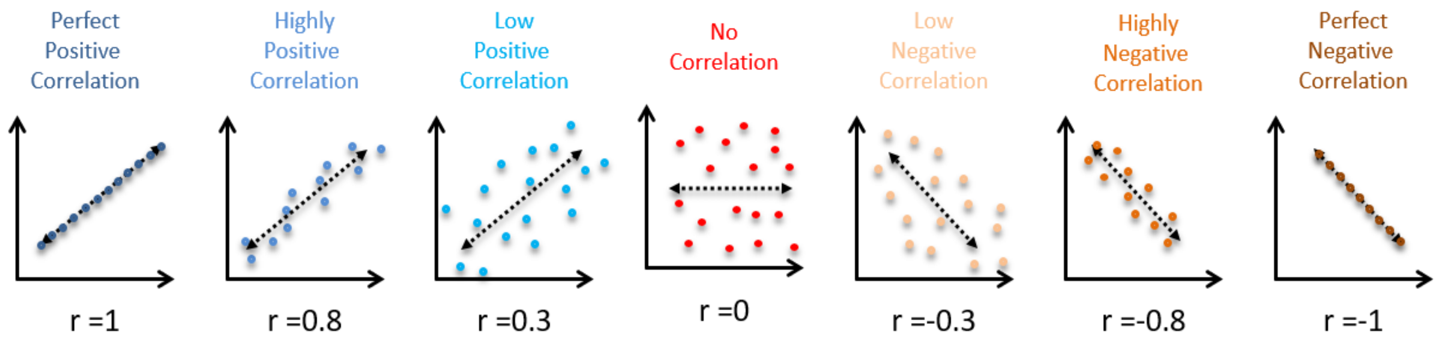
CORRELATION CO-EFFICIENT (r)

- It measures the **direction (positive or negative)** & **extent (-1 to 1)** of relationship among two variables.
- Value of 'r' lies between -1 to +1.

Types of Correlation

Types	Relation among 2 variables	Examples
Positive	Move in **same** direction	Height & Weight, yield & rainfall, profit & investment etc.
Negative	Move in **opposite** direction	Price & demand, Profits of insurance company & no. of claims
Uncorrelated	**No relation** between two variables	shoe size & intelligence.

SCATTER DIAGRAM



PC Note:

- Scatter Diagram ⇒ Simplest way to represent bivariate data.
- Used for Linear, non-Linear (curvilinear) both.
- **Curvilinear Correlation: r = 0** [If plotted points is not in the constant ratio, it is said to be curvilinear correlation]

KARL PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

- It is defined as ratio of **covariance between two variables** to the **product of standard deviations of two variables.**
- **Best method** for finding correlation between two variables **(linear relationship).**

$$r_{xy} = \frac{\mathrm{Cov}(x,y)}{SD_x \times SD_y} \quad \text{OR} \quad r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$\mathrm{Cov}(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = \frac{\sum xy}{n} - \bar{x}\bar{y}; \qquad SD_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} \qquad SD_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}} = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}$$

Q1. Compute correlation coefficient b/w x & y: n = 10, $\sum x = 40$, $\sum y = 50$, $\sum xy = 220$, $\sum x^2 = 200$, $\sum y^2 = 262$.

Q2. Find product moment correlation coefficient from following data:

X	2	3	5	5	6	8
Y	9	8	8	6	5	3

Q3. Coefficient of correlation b/w x & y for 20 items is 0.4. AM's & SD's of x & y are known to be 12 & 15 & 3 & 4 respectively. Later on, it was found that pair (20, 15) was wrongly taken as (15, 20). Find correct value of correlation coefficient.

PC Note: No impact due to change of origin. No impact due to change of scale on value, but it may affect sign of 'r'.

1. Sign of co-efficient of 'x' & 'y' should be same. [If it is not same, pahle usko same karo by multiplying by (-1)].
2. Then check the sign of u & v.
   - Sign of x & y must be same.
   - $r_{xy} = r_{uv}$ [If u & v have same sign] & $r_{xy} = -r_{uv}$ [If u & v have opposite sign]

Q4. Correlation coefficient between x & y is 0.8, write down correlation coefficient between u & v. [IMP]

(a) 2u + 3x + 4 = 0 & 4v + 16y + 11 = 0
(b) 2u – 3x + 4 = 0 & 4v + 16y + 11 = 0
(c) 2u – 3x + 4 = 0 & 4v – 16y + 11 = 0
(d) 2u + 3x + 4 = 0 & 4v – 16y + 11 = 0

SPEARMAN'S RANK CORRELATION COEFFICIENT

- Useful in finding correlation between two qualitative characteristics. Ex: Beauty & intelligence.
- It is easier to compute. $r_R = 1 - \frac{6\sum d^2}{n(n^2-1)}$ [d = difference in ranks, n = total no. of variables]

Q5. Compute coefficient of rank correlation between sales & advertisement (Rs. in Rs. 000):

Sales	90	85	68	75	82	80	95	70
Advertisement	7	6	2	3	4	5	8	1

Tied Rank Formula: $r_R = 1 - \frac{6\left[\sum d^2 + \sum\frac{t^3-t}{12}\right]}{n(n^2-1)}$ [t = number of variables are involved in tie]

Q6. Compute coefficient of rank correlation between Eco & Stats marks:

Eco Marks	80	56	50	48	50	62	60
Stats Marks	90	75	75	65	65	50	65

Space for Pranav Sir's Class Note:

Q7. For a number of towns, the coefficient of rank correlation between the people living below the poverty line & increase of population is 0.50. If sum of squares of the differences in ranks awarded to these factors is 82.50, find number of towns.

COEFFICIENT OF CONCURRENT DEVIATIONS

- Method of finding 'r' when we are not serious about the magnitude of these two variables.
- Very simple & casual method of finding correlation. It lies between -1 to +1.

$$r_C = \pm\sqrt{\pm\frac{2c-m}{m}}$$

m = n -1 & C = no. of concurrent deviations.

Space for Pranav Sir's Class Note:

Q8. Find coefficient of concurrent deviation from the following data:

Year	1990	1991	1992	1993	1994	1995	1996	1997
Price	25	28	30	23	35	38	39	42
Demand	35	34	35	30	29	28	26	23

REGRESSION

- It is concerned with **establishing a mathematical relationship between two variables.**
- Regression Analysis → Used for forecasting.

Regression Equation

If y depends on x [denoted by b_{yx}]	If x depends on y [denoted by b_{xy}]
Formula: $(y-\bar{y}) = b_{yx}(x-\bar{x})$ (y → dependent variable, x → independent variable) ⇒ y = a + b.x	Formula: $(x-\bar{x}) = b_{xy}(y-\bar{y})$ (y → independent variable, x → dependent variable) ⇒ x = a + b.y

Formula for b_{xy}	Formula for b_{yx}
$b_{xy} = r.\frac{SD_x}{SD_y}$	$b_{yx} = r.\frac{SD_y}{SD_x}$
$b_{xy} = \frac{Cov(x,y)}{(SD_y)^2}$	$b_{yx} = \frac{Cov(x,y)}{(SD_x)^2}$
$b_{xy} = \frac{n\sum xy - \sum x.\sum y}{n\sum y^2 - (\sum y)^2}$	$b_{yx} = \frac{n\sum xy - \sum x.\sum y}{n\sum x^2 - (\sum x)^2}$

Q9. Find two regression equations & Estimate y when x is 13 & estimate also x when y is 15.

X	2	4	5	5	8	10
Y	6	7	9	10	12	12

Q10. Following data relate to mean & SD of prices of two shares in a stock Exchange:

Share	Mean	SD in Rs.
HDFC	44	5.60
ICICI	58	6.30

Coefficient of correlation between the share price = 0.48. Find the most likely price of share of HDFC corresponding to a price of Rs. 60 of ICICI and also the most likely price of share of ICICI corresponding to a price of Rs. 50 of HDFC.

Properties of Regression Lines:

(a) Regression coefficients remain unchanged due to shift of origin but change due to a shift of scale.

(b) Two regression lines always **intersect** at the **mean $(\bar{x}, \bar{y})$.**

(c) Coefficient of correlation between two variables x & y is GM of two regression coefficients $r = \pm\sqrt{b_{yx} \times b_{xy}}$

If Regression Coefficients are negative, r would be negative & if both are positive, r would be positive.

PC Note: $b_{uv} = b_{xy} \times \frac{\text{Change of scale of } x}{\text{Change of scale of } y}$ & $b_{vu} = b_{yx} \times \frac{\text{change of scale of } y}{\text{change of scale of } x}$

Q11. If relationship between two variables x and u is u + 3x = 10 and between two other variables y and v is 2y + 5v = 25, and the regression coefficient of y on x is known as 0.80, what would be the regression coefficient of v on u?

Q12. For variables x & y, the regression eq^n are given as 7x - 3y - 18 = 0 & 4x - y - 11 = 0

(a) Find arithmetic means of x and y

(b) Identify the regression equation of y on x.

(c) Compute correlation coefficient b/w x & y

(d) Given the variance of x is 9, find the SD of y.

PROBABLE ERROR (PE)

- It is a method of obtaining correlation coefficients of population.
- It is defined as $\mathbf{P.E.} = \mathbf{0.674} \times \frac{1-r^2}{\sqrt{N}}$ Where r = correlation coefficients, N = No. of pairs observe

  $\mathbf{PE} = \frac{2}{3}\mathbf{SE}$ [Where SE = Standard Error of correlation coefficients] & $\mathbf{SE} = \frac{1-r^2}{\sqrt{N}}$
- **Limits of correlation** is given by $p = r \pm PE$ [Where p = Correlation coefficient of the population]

Assumption while probable errors are significant

(a) If **r < PE**, there is no evidence of correlation

(b) If value of 'r' is more than 6 times of the probable error, then the presence of correlation coefficient is certain.

(c) Since 'r' lies between -1 & +1 ($-1 \le r \le 1$), **probable error is never negative.**

Q13. Compute Probable Error assuming correlation coefficient of 0.8 from sample of 25 pairs of items.

Q14. If r = 0.7 & n = 64, find out PE of 'r' & determine the limits for the population correlation coefficient.

For Theory MCQs

- If two variables x & y are **independent or uncorrelated, correlation coefficient between x & y is zero (0).**
- **If Karl Pearson's correlation coefficient = zero, then we cannot conclude that the two variables are independent.**
- There are some cases when we may a correlation between two variables are not casually related. This is due to existence of a third variable which is related to **spurious correlation or non-sense correlation.**
- Coefficient of determination = r^2 = Explained Variance/Total Variance
- Coefficient of non-determination = $1 - r^2$

INDEX NUMBER

MEANING & DEFINITION

- Index Number ⇒ **Ratio of two or more time periods** involved (one is base period & other is current period)
- **Base period (IN = 100) ⇒ Standard point of comparison.**
- Year in which comparison is made = **Current year** & Year w.r.t. which the comparison is made = **Base year**.
- Index numbers are of **two types: (1) Simple (2) Composite.** Most Index numbers are composite in nature.
- Index number are always unit free.

Basically, Index Number = $\frac{\text{Current Price}}{\text{Base Price}} \times 100 = \frac{P_1}{P_0} \times 100$

PRICE RELATIVE & QUANTITY RELATIVE

- **Price Relative** ⇒ Ratio of Current Year's price to the base year's price (expressed in terms of percentage).

$$PR = \frac{P_1}{P_0} \times 100\%$$

P_1 = Price of Current Year & P_0 = Price in base year

- **Quantity Relative** ⇒ Ratio of Current year's quantity to base year's quantity expresses in terms of percentage.

$$QR = \frac{Q_1}{Q_0} \times 100\%$$

Q_1 = Quantity in Current Year & Q_0 = Quantity in base year

Q1. P_{10} is the index for time on:

(a) 0 on 1 (b) 1 on 0 (c) 1 on 1 (d) 0 on 0

Q2. P_{01} is the index for time on:

(a) 0 on 1 **(b) 1 on 0** (c) 1 on 1 (d) 0 on 0

---

1. SIMPLE AGGREGATIVE METHOD ⇒ $\frac{\Sigma P_1}{\Sigma P_0} \times 100\% = \frac{\text{Sum of Prices of All Commodities in Current Year}}{\text{Sum of Prices of All Commodities in Base Year}} \times 100\%$

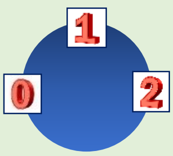
2. WEIGHTED AGGREGATIVE METHOD

SN	Name of the Method	Price Index Number	Quantity Index Number
1	**Paasche's Price Index No.**	$\frac{\sum P_1.Q_1}{\sum P_0.Q_1}$	$\frac{\sum Q_1.P_1}{\sum Q_0.P_1}$
2	**Lasperyes Price Index No.** **[Same formula for Cost of Living Index]**	$\frac{\sum P_1.Q_0}{\sum P_0.Q_0}$	$\frac{\sum Q_1.P_0}{\sum Q_0.P_0}$
3	**Marshall-Edgeworth Price Index No.**	$\frac{\sum P_1.(Q_1+Q_0)}{\sum P_0.(Q_1+Q_0)}$	$\frac{\sum Q_1.(P_1+P_0)}{\sum Q_0.(P_1+P_0)}$
4	**Fisher's Price Index No. [GM of P & L]**	$\sqrt{\frac{\sum P_1.Q_1}{\sum P_0.Q_1} \times \frac{\sum P_1.Q_0}{\sum P_0.Q_0}}$	$\sqrt{\frac{\sum Q_1.P_1}{\sum Q_0.P_1} \times \frac{\sum Q_1.P_0}{\sum Q_0.P_0}}$
5	**Bowley's Price Index No. [AM of P & L]**	$\frac{\text{Lasperyes's} + \text{Paasche's}}{2}$	$\frac{\text{Lasperyes's} + \text{Paasche's}}{2}$

Q3. Compute Price Index Number for the current year using the following data:

Commodity	Base Year		Current Year	
	Price in Rs.	Quantity in Kg.	Price in Rs.	Quantity in Kg.
A	1	6	5	8
B	2	7	4	7
C	3	8	3	6
D	4	9	2	5

TEST OF ADEQUACY OF INDEX NUMBER FORMULA

1	**Unit Test**	☛ **Formula must be independent of unit used.** ☛ Satisfied by all Index Numbers Except Simple Aggregate Index Method. ☛ **Thus, Simple Aggregate Index Method does not satisfy unit test.**
2	**Time Reversal Test**	☛ $P_{01} \times P_{10} = 1$ ☛ **Not satisfied** by **Paasche's & Laspeyre's** Index No. ☛ Other Index Numbers satisfy this test.
3	**Factor Reversal Test**	☛ $P_{01} \times Q_{01} = V_{01}$. ☛ **Only Fisher Index No. satisfy this test.**
4	**Circular Test** 	☛ It is an Extension of Time Reversal Test. We change Base Year. ☛ $P_{01} \times P_{12} \times P_{20} = 1$ ☛ **Only Simple G.M. of Price Relative & weighted aggregative with fixed weights method satisfy this test**

DEFLATING TIME SERIES USING INDEX NUMBER

⇒ **Deflated value** = $\dfrac{Current\ Value}{Price\ Index\ of\ Current\ year}$ or $\dfrac{Current\ Value \times Base\ Price(P_0)}{Current\ Price\ (P_1)}$

Year	Wholesale Price Index	GNP at Current Prices	Real GNP = [GNP at Current Prices/Index No]
1970	113.1	7499	**6630** [7499/113.2]
1971	116.3	7935	**6823** [7935/116.3]
1972	121.2	8657	**7143** [8657/121.2]
1973	127.7	9323	**7301** [9323/127.7]

PC Note: Real economic growth is determined by deflating GNP values using price index in terms of base year.

CHAIN INDEX NUMBER

❖ Fixed base is not advisable when conditions change quite fast. In such case, it is advisable to change base for example, 1998 for 1999, and 1999 for 2000, and so on.

❖ **Under this method, price relatives of each year are first related to the preceding year called the link relatives and then they are chained together by successive multiplication to form a chain index.**

❖ **Chain Index** = $\frac{\text{Link relative of current year} \times \text{Chain Index of Previous Year}}{100}$

Q4.

Year	Price	Link Relatives	Chain Indices
1991	50	100	100
1992	60	$\frac{60}{50}$ x 100 = 120.0	$\frac{120 \times 100}{100}$ = 120.0
1993	62	$\frac{62}{60}$ x 100 = 103.3	$\frac{103.3 \times 120}{100}$ = 124.0
1994	65	$\frac{65}{62}$ x 100 = 104.8	$\frac{104.8 \times 124}{100}$ = 129.9
1995	70	$\frac{70}{65}$ x 100 = 107.7	$\frac{107.7 \times 129.9}{100}$ = 139.9
1996	78	$\frac{78}{70}$ x 100 = 111.4	$\frac{111.4 \times 139.9}{100}$ = 155.8
1997	82	$\frac{82}{78}$ x 100 = 105.1	$\frac{105.1 \times 155.8}{100}$ = 163.7
1998	84	$\frac{84}{82}$ x 100 = 102.4	$\frac{102.4 \times 163.7}{100}$ = 167.7
1999	88	$\frac{88}{84}$ x 100 = 104.8	$\frac{104.8 \times 167.7}{100}$ = 175.7
2000	90	$\frac{90}{88}$ x 100 = 102.3	$\frac{102.3 \times 175.7}{100}$ = 179.7

Q5. From the following data:

Group	A	B	C	D	E	F
Group Index	120	132	98	115	108	95
Weight:	6	3	4	2	1	4

The general price index is given by: (a) 113.54 (b) 115.30 (c) 117.92 **(d) 111.30**

Q6. If a person was earning Rs. 2,050 in the base period. What should be his salary in current period if his standard of living is to remain the same? Given ∑W = 25 & ∑IW = 3544: (a) Rs. 2096 **(b) Rs.2906** (c) Rs. 2106 (d) Rs. 2306

Q7. If the prices of all commodities in a place has increased 20% in comparison to the base period prices, then the index number of prices for the place is now _________. (a) 100 **(b) 120** (c) 20 (d) 150

Q8. If $\Sigma P_0Q_0 = 116$, $\Sigma P_0Q_1 = 140$, $\Sigma P_1Q_0 = 97$, $\Sigma P_1Q_1 = 117$, then Fisher's ideal index number is __
(a) 184 **(b) 83.59** (c) 119.66 (d) 120

Q9. Net monthly salary of an employee was ₹ 3,000. The consumer price index number in 1985 is 250 with 1980 as base year. If he has to be rightly compensated then additional dearness allowance to be paid to employee is:
(a) ₹ 4,000 (b) ₹ 4,800 (c) ₹ 5,500 **(d) ₹ 4,500**

Q10. Bowley's index = 150, Laspeyer's index = 180, then Paasche's index = ____
(a) 120 (b) 30 (c) 165 (d) None

SHIFTING & SPLICING OF INDEX NUMBER

- These refer to two technical points:
  (i) how the base period of the index may be shifted
  (ii) how two index covering different bases may be combined into single series by splicing.

- **Shifted Price Index** = $\frac{\text{Original Price Index}}{\text{Price index of the year on which it has to be shifted}} \times 100$

- **Example: Splicing Two Index Number Series**

Year	Old Price Index [1990 = 100]	Revised Price Index [1995 = 100]	Spliced Price Index [1995 = 100]
1990	100.0		87.6
1991	102.3		89.6
1992	105.3		92.2
1993	107.6		94.2
1994	111.9		98.0
1995	114.2	100.0	100.0
1996		102.5	102.5
1997		106.4	106.4
1998		108.3	108.3
1999		111.7	111.7
2000		117.8	117.8

THEORETICAL DISTRIBUTION

BINOMIAL DISTRIBUTION (BD) → Bi-parametric (n & p)

- No. of trials is **too large but finite.**
- Trials are independent.
- Outcomes → Mutually Exclusive & Exhaustive
- **Success → 'p' & Failure → q = (1 – p)**
- Probability = $^{n}C_{r} \cdot P^{r} \cdot q^{n-r}$ [n = No. of trials, p = probability of success, r = success required]

Application of Binomial Distribution:
- Number of defectives in a lot size 'n'.
- Number of students in class.
- Number of boys married.
- Number of accidents on road on a day.

Q1. A coin is tossed 10 times. Assuming the coin to be unbiased, what is the probability of getting
(a) 4 heads? (b) At least 4 heads? (c) At most 3 heads?

Properties of BD

- Mean (μ) = np
- Variance (σ^2) = npq
- SD = $\sqrt{\text{Variance}}$
- Mode → Depends on Value of p(n+1)

Q2. Find BD if mean = 6 & SD = 2

- Variance is always less than mean.
- Variance will be highest When $p = q = 0.5$
- Highest Value of Variance = n/4

If Value of p(n+1) is integer ⇒ Bimodal	If Value of p(n+1) is fraction ⇒ Unimodal
Mode 1 = **p(n+1)** & Mode 2 = **p(n+1) – 1**	Mode = Integral part of **p(n+1)**
Ex: n = 9 & p = 0.5 ⇒ p(n+1) = 0.5 x (9+1) = 5 ∴ 1st Mode = 5 & 2nd Mode = 5 -1 = 4	**Ex.** n=10 & p = 1/3. ∴ (n+1) x 1/3 = 11/3 = 3.67 ∴ **Mode = 3**

Q2. What is the mode of the distribution for which mean and SD are 10 and $\sqrt{5}$ respectively.

Q3. The incidence of occupational disease in an industry is such that the workmen have a 10% chance of suffering from it. What is the probability that out of 5 workmen, 3 or more will contract the disease?

Q4. Experiment succeeds thrice as after it fails. If experiment is repeated 5 times, probability of having no success is __.

Q5. If x & y are 2 independent binomial variables with parameters 6 & 1/2 and 4 & 1/2 respectively, what is **P(x+y ≥ 1)**?

Additive property of BD: If X & Y are 2 independent variables s.t **X~B(n_1, P) & Y~B(n_2, P)**, then **(X+Y) ~B (n_1 + n_2, P)**

POISSON DISTRIBUTION (PD) → Uni-parametric (m)

- **Trials are too large. [Tends to infinity] &** independent.
- **Probability of success → Too low & Probability of failure → Too high**

Probability Mass Function = $\frac{e^{-m} \times m^{r}}{r!}$ [m (average) = np; r = No. of success required; e = 2.71828]

Properties of PD

- Mean (μ) = $np = m$
- Variance = npq
- Mode ⇒ Value of m

Integer	Fraction
Bimodal ⇒ m & m-1	Unimodal ⇒ Largest integral

Application of Poisson Distribution:
- Number of printing mistakes per page of large book
- Number of road accidents on a busy road per minute
- Number of radio-active elements per minute in a fusion process.
- Number of demands per minute for health-centre and so on.

Q6. If 2 % of electric bulbs manufactured by a company are defectives, what is probability that a sample of 150 electric bulbs taken from the production process of that company would contain:
(a) exactly one defective bulb? (b) more than 2 defective bulbs? [Given e^{-3} = 0.04978]

Q7. The manufacturer of a certain electronic component is certain that two per cent of his product is defective. He sells the components in boxes of 120 and guarantees that not more than 2% in any box will be defective. Find the probability that a box, selected at random, would fail to meet the guarantee? [$e^{-2.40}$ = 0.0907]

Q8. Find mean & standard deviation of x where x is a Poisson variate satisfying the condition P(x=2) = P(x=3).

Q9. Probability that a random variable 'x' following PD would assume a positive value is $(1 - e^{-2.7})$. What is the mode?

Q10. Standard deviation of a PD is 1.732. What is the probability that the variate lies between –2.3 to 3.68?

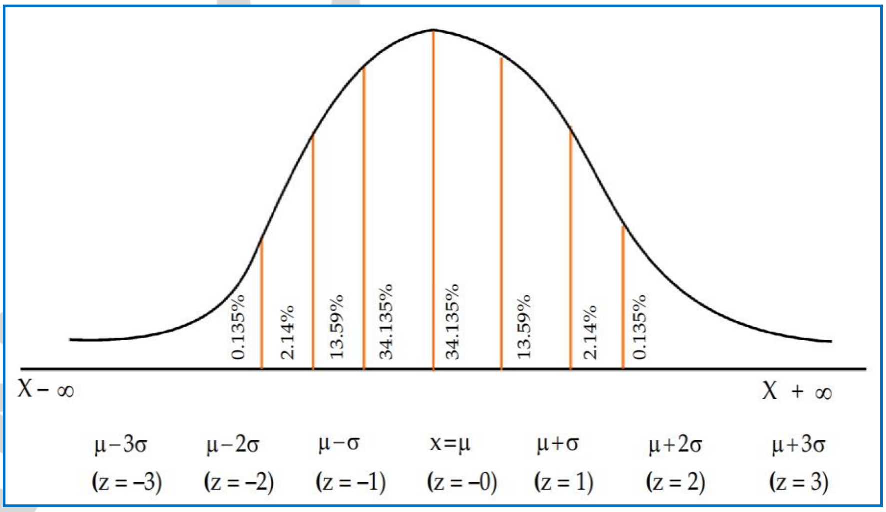
NORMAL DISTRIBUTION or GAUSSIAN DISTRIBUTION

$P(x) = \frac{1}{\sigma\sqrt{2\pi}} . e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ [$\sigma = SD$; $\pi = \frac{22}{7} = 3.14$; $e = 2.71818$; x = No. of success required]

Q11. For random variable x, probability density function is given by $f(x) = \frac{e^{-(x-4)^2}}{\sqrt{\pi}}$ $for - \infty < x < \infty$. Find mean & variance

Properties of ND

- Mean = Median = Mode
- Perfectly symmetrical about mean(μ).
- MD = 0.80 x SD
- 4SD = 5MD = 6QD
- $Q_1 = \mu - 0.675\sigma$ & $Q_3 = \mu + 0.675\sigma$
- QD = $\frac{Q3 - Q1}{2}$ = 0.675σ
- It is unimodal. Total Area = 1.
- Point of Inflexion
  - $\mu - \sigma \rightarrow$ Concave to Convex
  - $\mu + \sigma \rightarrow$ Convex to Concave



- Two tails of normal curve extend infinitely on both sides of curve & both tails (left & right) never touch 'x' axis.
- z = $\frac{x - \mu}{\sigma}$
- Additive Property: If z = x + y, then **Mean (z) = Mean (x) + Mean (y)**; SD(z) = $\sqrt{SD_x^2 + SD_y^2}$

Q12. x & y are independent normal variables with mean 100 & 80 respectively and standard deviation as 4 and 3 respectively. What is the distribution of (x + y)?

Q13. If two quartiles of a normal distribution are 47.30 and 52.70 respectively, what is the mode of the distribution? Also find the mean deviation about median of this distribution.

Q14. Find the points of inflexion of the normal curve $f(x) = \frac{1}{4\sqrt{2\pi}} . e^{\frac{-(x-10)^2}{32}}$ $for - \infty < x < \infty$

Q15. X follows normal distribution with μ = 50 & $(SD)^2$ = 100. What is P(x ≥ 60)? [Given z(1) = 0.8413]

Q16. Mean of a normal distribution is 500 & 16% of the values are greater than 600. What is standard deviation of the distribution? (Given that the area between z = 0 to z = 1 is 0.34)

STANDARD NORMAL DISTRIBUTION [denoted by z]

- Mean = Median = Mode = Zero
- Standard deviation = 1; MD = 0.8; QD = 0.675
- Points of inflexion are –1 & 1.
- Standard normal distribution is symmetrical about z = 0.

$$P(x) = \frac{1}{\sqrt{2\pi}}.e^{\frac{-1}{2}\left(\frac{x}{1}\right)^2} = \frac{1}{\sqrt{2\pi}}.e^{\frac{-(x)^2}{2}}$$

$$Z = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0,1)$$

Q17. The mean height of 2000 students at a certain college is 165 cms & SD 9 cms. What is the probability that in a group of 5 students of that college, 3 or more students would have height more than 174 cm? [Best Question]

Ans: X = Height of the Students, μ = 165 & σ = 9

p = P (x>174) = 1 - P (x ≤ 174) = 1 - $P(\frac{x-165}{9} < \frac{174-165}{9})$ = 1 - P (z ≤ 1) = 1 - z (1) = 1 – 0.8413 = 0.1587

n = 5, p = 0.1587

Probability that 3 or more students would be more than 174 cm = P(≥3) = P(3) + P(4) + P(5)

= 5_{C_3} $(0.1587)^3$. $(0.8413)^2$ + $5_{C_4}(0.1587)^4$ x (0.8413) + $5_{C_5}(0.1587)^5$

= 0.02829 + 0.002668 + 0.000100 = **0.03106.**

Q18. In a sample of 500 workers, mean wage = Rs. 500 & SD of wages = Rs. 48. Find the number of workers having wages: (a) more than Rs. 600; (b) less than Rs. 450; (c) between Rs. 548 & Rs. 600. [Given z(2.08) = 0.9812 & z(1.04)= 0.8508]

Ans: X = Wage of the workers, μ = 500 & σ = 48

(a) Probability that a worker selected at random would have wage more than Rs. 600

= P (x>600) = 1 – P(x≤600) = 1 - $P(\frac{x-500}{48} < \frac{600-500}{48})$ = 1 – P(z ≤ 2.08) = 1 – z (2.08) = 1 – 0.9812 = 0.0188

∴ No. of workers having wages more than Rs. 600 = 500 x 0.0188 = 9.4 ≅ **9**

(b) Probability of a worker having wage less than Rs. 450 = P (x<450) = $P(\frac{x-500}{48} < \frac{450-500}{48})$

= P (z < -1.04) = z (-1.04) = 1 - z (1.04) = 1 – 0.8508 = 0.1492

∴ No. of workers having wages less than Rs. 450 = 500 x 0.1492 ≅ **75**

(c) Probability of a worker having wage between Rs. 548 & Rs. 600 = $P(120 < x \le 150)$

= $P(\frac{548-500}{48} < \frac{X-500}{48} < \frac{600-500}{48})$ = $P(0 < z < 2.08)$ = z(2.08) – z(1) = 0.9812 – 0.8413 = 0.1399

∴ No. of workers having wages between Rs. 548 & Rs. 600 = 500 x 0.1399 ≅ **70**

Q19. If a random variable 'x' follows normal distribution with mean as 120 & standard deviation as 40, what is the probability that P (x ≤ 150 / x > 120)? [Given that area of the normal curve between z = 0 to z = 0.75 is 0.2734]

Ans: μ = 25 & σ = 10

⇒ P(25 < x < b) = 0.4772 ⇒ $P(\frac{25-25}{10} < \frac{X-25}{10} < \frac{b-25}{10})$ = 0.4772 ⇒ $P(0 < z < \frac{b-25}{10})$ = 0.4772

⇒ $z(\frac{b-25}{10})$ – z(0) = 0.4772 ⇒ $z(\frac{b-25}{10})$ – 0.5 = 0.4772 ⇒ $z(\frac{b-25}{10})$ = 0.9772

⇒ $z(\frac{b-25}{10})$ = z(2) ⇒ $\frac{b-25}{10}$ = 2 ⇒ b = 2 x 10 + 25 = **45**

Q20. The distribution of wages of a group of workers is known to be normal with mean Rs. 500 & SD Rs. 100. If the wages of 100 workers in the group are less than Rs. 430, what is the total number of workers in the group? [Z(0.70) = 0.758]

Ans: X = Wage of the workers, μ = 500 & σ = 100

⇒ $P(\frac{x-500}{100} < \frac{430-500}{100}) = \frac{100}{N}$ ⇒ P (z < -0.7) = $\frac{100}{N}$ ⇒ z (-0.70) = $\frac{100}{N}$

⇒ 1 - z (0.70) = $\frac{100}{N}$ ⇒ 1 – 0.758 = $\frac{100}{N}$ ⇒ 0.242 = $\frac{100}{N}$ ⇒ **N ≅ 413**