Probability

Know about Probability

→ First use of Probability was made 300 years back in Europe by a group of mathematicians to enhance their chances of winning in gambling
→ It is a full-fledged subject and become an integral part of statistics
→ Theories of Testing Hypothesis and Estimation are based on probability

Types

Subjective Probability	Dependent on personal judgment, useful in decision making. It is out scope of our syllabus
Objective Probability	This is based on Mathematical Rules and not judgment based. We will study this section in our chapter.

Random Experiment

Experiment	A performance that produces certain results
Random Experiment	An experiment is defined to be random if the results of the experiment depend on chance only.
Examples	Tossing a coin, throwing a dice, drawing cards from a pack

Events

The **results or outcomes** of a **random experiment** are known as events

Types of Events

Based on Combination of Events

Simple or Elementary	If the event cannot be decomposed into further events
Composite or Compound	An event that can be decomposed into two or more simple events

Based on nature of occurrence (applicable for set of events)

Mutually Exclusive or Incompatible Events	A set of events A1, A2 … is said to be mutually exclusive if they cannot occur simultaneously. Occurrence of one implies non occurrence of other.
Exhaustive Events	A set of events A1, A2 … is said to be exhaustive if one of these must necessarily occur on a random experiment
Equally Likely or Equi-Probable Events or Mutually Symmetric	If it is evident that from the set of events, none of the events is expected to occur more frequently than others.

Classical Definition of Probability

Also called Prior Definition of Probability, this formula is Event (Result) Based. It is given by Bernoulli and Laplace.

$$P(A) = \frac{\text{no. of events favorable to A}}{\text{total number of events}}$$

More about Classical Probability	Demerits or Limitations	→ Applicable only when events are finite and are equally likely → Limited application of this definition like in tossing coin, throwing dice, cards etc.
	Other Notes	→ $0 \leq P(A) \leq 1$, P(A) = 1 means Sure Event, P(A) = 0 means impossible event → Probability of non-occurrence of an event A is denoted by P(A') or P($\bar{A}$) is called as complimentary event of A. P(A') = 1−P(A)
	Odds in Favor of an Event	$\frac{\text{no. of favorable events}}{\text{no. of unfavorable events}}$
	Odds Against an Event	$\frac{\text{no. of unfavorable events}}{\text{no. of favorable events}}$
Special Formula	If an experiment results in p outcomes and if it is repeated q times then $\text{Total no. of outcomes} = p^q$	
Terms used in 52 Cards Deck	Suits (four)	Spades - ♠ Hearts - ♡ Diamond - ♢ Clubs - ♣
	Ranks (13)	A (Ace), K (King), Q (Queen), J (Jack), 10, 9, 8, 7, 6, 5, 4, 3, 2
Relative Frequency Definition of Probability	$\text{Relative Frequency} = \frac{\text{no.of times the event occured during experimental trials}}{\text{total no.of trials}} = \frac{f_A}{n}$ Probability by this method is defined as $P(A) = \lim_{n \to \infty} \frac{f_A}{n}$ (Relative Frequency on infinite no. of trials is equal to probability)	
Set Based Probability	Sample Space (denoted by **S** or **Ω**-omega)	a non-empty set containing all the elementary events of a random experiment as sample points
	Event A	Event which is under consideration for probability calculations is defined as a non empty subset of Set S (Sample Space)
	Probability Formula	$P(A) = \frac{\text{no. of sample points in A}}{\text{no. of sample points in S}} = \frac{n(A)}{n(S)}$
Axiomatic Or Modern Definition of Probability	This definition is also based on Sets Concepts. Here Probability is not a simple ratio like above, but can be said as function P defined on S known as Probability Measure. P(A) is defined as the probability of A as per this function only if below conditions are satisfied:	
	Condition 1	$P(A) \geq 0$, for every $A \subseteq S$
	Condition 2	$P(S) = 1$
	Condition 3	For any sequence of mutually exclusive events A1, A2, A3... $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$

Addition Theorems	Theorem 1	$P(A \cup B) = P(A + B) = P(A\ or\ B) = P(A) + P(B)$ If A and B are mutually exclusive events
	Theorem 2	For set of mutually exclusive events $A_1, A_2, A_3, \ldots$ $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$
	Theorem 3	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ For any two events A and B
	Theorem 4	$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$
Expected Frequency	No. of sample points n(S) × P(A)	
Conditional Probability or Compound Theorem	Dependent Events	If occurrence of one event is influenced by occurrence of another event, then two events are dependent.
	Independent Events	Two events are said to be independent if occurrence of one event do not influence the occurrence of other.
	Probability in case of Dependent Events A and B	Conditional Probability of B/A: means probability of event B given that event A has already been occurred $P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)}$, provided $P(A) > 0$ Similarly, Conditional Probability of A/B: $P\left(\frac{A}{B}\right) = \frac{P(B \cap A)}{P(B)}$, provided $P(B) > 0$ **Compound Theorem: P(A∩B) = P(B) × P(A/B) or P(A∩B) = P(A) × P(B/A)**
	Probability in case of Independent Events	Since there is no dependency, Conditional Probability = Normal Probability i.e. P(B/A) = P(B) and P(A/B) = P(A) Here, P(A∩B)= P(A) × P(B) And for three events, A, B, C P(A∩B∩C) = P(A) × P(B) × P(C) Also, if A and B are independent, then below are also independent : A and B', A' and B, A' and B'
Random Variable: Probability Distribution	Random Variable	It is a function defined on Sample Space of a random experiment that can take any value (Real Number)
	Discrete Random Variable	RV that can take only discrete values. RV on a discrete sample space
	Continuous Random Variable	RV that can take any values within an interval. [infinite no. of sample points in a sample space]
	Probability Distribution	If is defined as the statement/ table that shows no. of different value taken by Random Variable and their corresponding probabilities
	Conditions of Probability Dist.	If X (a random variable) takes n finite values like $X_1, X_2, X_3, \ldots, X_n$ and probabilities are $P_1, P_2, P_3, \ldots, P_n$ then, $P_i \geq 0$ for every i and $\sum P_i = 1$

Expected Value	Expected Value	It is defined as the sum of products of different values taken by Random Variable and corresponding probabilities. $E(x) = \sum p_i \, x_i$ (this formula is similar to AM of frequency distribution)
	Mean of Probability Distribution	Since this is mean, we can say that Expected value is equal to arithmetic mean of probability distribution. Here mean is denoted by μ, hence $\mu = E(x) = \sum p_i \, x_i$
	Variance of Probability Distribution	$V(x) = \sigma^2 = E(x-\mu)^2 = E(x)^2 - \mu^2$
	Properties of E.V.	→ E.V. of a constant is constant → $E(x+y) = E(x) + E(y)$ → $E(k.x) = E(x).k$ → $E(x.y) = E(x).E(y)$

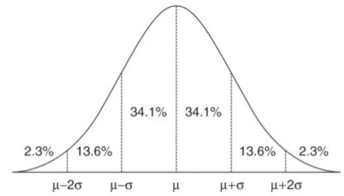
Theoretical Distribution

Binomial Distribution (bi-parametric discrete probability distribution)	Bernoulli's Trial	→ Each trial is associated with two mutually exclusive and exhaustive outcomes [one is success and other one is failure] → Trials are independent → Probability of success (*p*) and failure (*q=1-p*) will remain unchanged throughout the process → No. of trials is a positive integer
	Binomial Variable	It is a discrete random variable X that follows binomial distribution and is denoted by $\boldsymbol{X \sim B(n, p)}$
	Probability Mass Function	$\boldsymbol{f(x) = P(X = x) = {}^{n}C_{x}\, p^{x} q^{n-x}}$ for $x = 0,1,2,3, \dots, n$ and $f(x) = 0$ if x is otherwise
	Mean	$\boldsymbol{\mu = np}$
	Variance	$\boldsymbol{\sigma^2 = npq}$, also Variance is always less than mean, maximum value of variance is $^{n}/_{4}$
	Mode	Calculate $(n+1)p$, if the resulting value is integer then Bi-modal: $\boldsymbol{\mu_0} = (n+1)p$ *and* $[(n+1)p - 1]$ If the resulting value is non-integer then Uni-modal: $\boldsymbol{\mu_0}$ = largest integer contained in $(n+1)p$
	Additive Property	If X and Y are two independent variables such that $\boldsymbol{X \sim B(n_1, p)}$ and $\boldsymbol{Y \sim B(n_2, p)}$, then $\boldsymbol{(X+Y) \sim B(n_1 + n_2, p)}$
Poisson Distribution (uni-parametric discrete probability distribution)	History	Simon Denis Poisson of France introduced this distribution way back in the year 1837
	Conditions	It is a limiting form of Binomial Distribution, where n→∞, p→0. It is also a discrete distribution
	Poisson Variable	It is a discrete random variable that follows Poisson Distribution denoted as $\boldsymbol{X \sim P(m)}$
	Probability Mass Function	$\boldsymbol{f(x) = P(X = x) = \frac{(e^{-m}.m^{x})}{x!}}$ for $x = 0, 1, 2, \dots \infty$
	Mean	$\boldsymbol{\mu = m}$
	Variance	$\boldsymbol{\sigma^2 = m}$
	Mode	Calculate m, if the resulting value is integer then Bi-modal: $\boldsymbol{\mu_0 = m}$ *and* $[\boldsymbol{m-1}]$ If the resulting value is non-integer then Uni-modal: $\boldsymbol{\mu_0}$ = largest integer contained in $\boldsymbol{m}$
	Additive Property	If X and Y are two independent variables such that $\boldsymbol{X \sim P(m_1)}$ and $\boldsymbol{Y \sim P(m_2)}$, then $\boldsymbol{(X+Y) \sim P(m_1 + m_2)}$

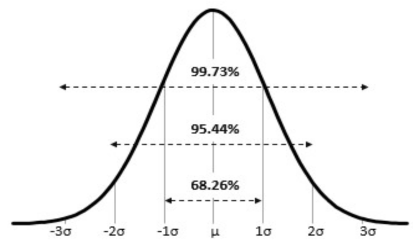
Normal Distribution (bi-parametric continuous probability distribution)	Basics	Various Mathematical experiments have proved that most of the continuous random variables will follow normal distribution. It is universally accepted distribution.
	Probability Density Function	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-(\frac{x-\mu}{\sigma})^2 \times \frac{1}{2}}$ It is defined for $-\infty < x < \infty$
Normal Distribution Properties	Mean = Median = Mode	μ
	Standard Deviation	σ
	Mean Deviation	$\sigma \times \sqrt{2/\pi} = 0.8\,\sigma$
	Quartile Deviation	$Q_1 = \mu - 0.675\sigma$ and $Q_3 = \mu + 0.675\sigma$
	Shape of Normal Curve	Bell Shaped
	Normal Variable	$X \sim N(\mu, \sigma^2)$
	Additive Property	Only applicable when two different random variables are independent. Assume we have two variables X and Y such that $\boldsymbol{X \sim N(\mu_1, \sigma_1{}^2)}$, $\boldsymbol{Y \sim N(\mu_2, \sigma_2{}^2)}$ then $\boldsymbol{X + Y \sim N(\mu_1 + \mu_2, \sigma_1{}^2 + \sigma_2{}^2)}$
	Normal Curve is symmetrical at	$x = \mu$
	Points of Inflexion	$\mu - \sigma$ & $\mu + \sigma$
	Ratio between QD:MD:SD	10:12:15
Standard Normal Distribution	Conditions	Parameter | Value Mean μ | 0 Standard Deviation σ | 1
	Standard Normal Variate	The variable used in this distribution is called as Standard Normal Variate and is denoted by ~~Z~~ *[Striked Z]*
	Area from X=-3σ to X=3σ	99.73%
	Z Table	This table gives us the probability of values from *X=μ=0* to *X=any value up to 3*
	Z Score	$Z = \frac{x-\mu}{\sigma}$
	Cumulative Distribution Function	$\phi(x) = P(X \le x$
	Probability Function	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-(z)^2 \times \frac{1}{2}}$ for $-\infty < z < \infty$
	Mean, Median, Mode	μ=0
	SD, Variance	$\sigma=1, \sigma^2=1$
	Points of Inflexion	-1, 1
	Mean Deviation	0.8
	Quartile Deviation	0.675
	Probability Function	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-(z)^2 \times \frac{1}{2}}$ for $-\infty < z < \infty$

Area under Normal Curve

From	To	Area/Probability
μ	+σ	34.135%
+σ	+2σ	13.59%
+2σ	+3σ	2.14%
3σ	∞	0.135%



From	To	Area/Probability
-σ	+σ	68.3%
-2σ	+2σ	95.5%
-3σ	+3σ	99.7%



A. INTRODUCTION TO STATS

Definition

Singular Sense:

- Scientific method that is used for collecting, analyzing and presenting data
- Used to draw statistical inferences
- Inferences means conclusion reached on the basis of evidence and reasoning

Example:

After applying statistical methods we have arrived at a conclusion that in last 5 years crime rate is reduced.

Plural Sense:

- Data qualitative or quantitative collected to do statistical analysis

Example: Based on Cricket Match statistic of this stadium, chasing team wins mostly

History of Stats

- Word Origin
  - ✓ Latin word – Status
  - ✓ Italian word – Statista
  - ✓ German word – statistic
  - ✓ French word – statistique
- Publication:
  - ✓ Koutilya's book Arthashastra
  - ✓ Stat records on Agriculture found in Ain-i-Akbari (author Abu Fezal)
- Census: First ever census done in Egypt (300 years BC to 2000 BC)

Application of Stats

There are various but we will confine to below:

1. Economics: Time Series analysis, index, demand analysis, econometrics, regression analysis
2. Business Management: business decisions rely upon QT
3. Commerce/ Industry: Sales, Purchase, RM, Salary Wages etc. data are analyze for business decisions and policy making

Limitation of Stats:

1. Relevant for aggregate data and not individual data
2. Quantitative data can only be used, however for qualitative – it needs to be converted into quantitative
3. Projections are based on conditions/ assumptions and any change in that will change the projection
4. Sampling based conclusions are used, improper sampling leads to improper results

B. COLLECTION OF DATA

Data and Variable

- Variable = measurable quantity
  - Discrete variable: when a variable assumes a finite or count ably infinite isolated values. Example: no. of petals in a flower, no. of road accident in locality
  - Continuous variable: when a variable assumes any value from the given interval (can also be in decimals, fractions). Example: height, weight, sale, profit
  - Attribute: qualitative characteristics. Example: Gender of a baby, nationality of a person
- Data = quantitative information shown as number. These are of two types:
  - Primary : first time collected by agency/ investigator
  - Secondary: collected data used by different person/ agency

How to collect Primary Data?

1. **Interview Method**:
   a. Personal Interview: directly from respondents. Example: Natural Calamity, Door to Door Survey
   b. Indirect Interview: when reaching to person difficult, contact associated persons. Example: Rail accident
   c. Telephone Interview: over phone, quick and non-responsive

Type of Interview/ Parameters	Personal	Indirect	Telephone
Accuracy	High	Low	Low
Coverage	Low	Low	High
Non Response	Low	Low	High

2. **Mailed Questionnaire Method:**
   a. Mailed means by Post or Email
   b. Well drafted + properly sequenced + with guidelines
   c. Non Response is Maximum

3. **Observation Method:**
   a. Data collected by direct observation or using instrument
   b. Example: Height check, Weight check,
   c. Although more accurate but it is time consuming, low coverage and laborious

4. **Questionnaire filled and sent by Enumerators**
   a. Enumerator: Person who directly interact with respondent and fill the questionnaire
   b. Generally used in Surveys

Sources of Secondary Data

1. International sources like World Health Organization (WHO), International Monetary Fund (IMF), International Labor Organization (ILO), World Bank
2. Government Sources – In India – Central Statistics Office (CSO), National Sample Survey Office- NSSO, Regulators – RBI, SEBI, RERA, IRDA
3. Private or Quasi-government sources like Indian Statistical Institute (ISI), Indian Council of Agriculture, NCERT
4. Research Papers and other unpublished sources

Scrutiny of Data

1. Scrutiny – checking accuracy and consistency of data
2. Finding of errors by enumerators while filling or receiving questionnaire
3. Internal consistency check: when two or more series of related data are given check each other
4. Consider enumerators' bias while using data

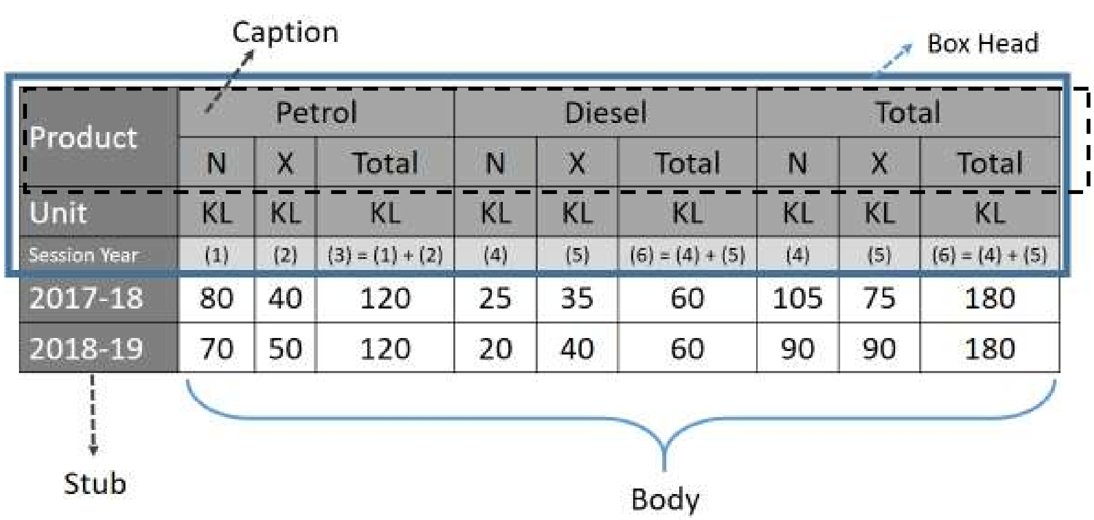
C. PRESENTATION OF DATA

Classification and organization of Data:

- means process of arranging data based on some logic
- there are four types of classification of data
    a. Chronological/ Temporal/ Time Series Data (ex. Profit YoYi.e year on year)
    b. Geographical or Spatial Series Data (ex. Weather in North India and South India)
    c. Qualitative or Ordinal Data (ex. Rating Top 20 songs by Radio Mirchi)
    d. Quantitative or Cardinal Data (no. of left handed batsmen in cricket teams playing CWC19)

Mode of Presentation

1. **Textual:** where text is used in the form of para or sentence. Example: Height of A,B and C is 160cm, 165cm, 175cm respectively
2. **Tabular/ Tabulation:**
    - Data shown in the form of table
    - Some important terms about Table (we will understand by example - next page figure)
    - It is preferred over textual form because
        - Useful in easy comparison
        - Complicated data can be presented
        - Table is must to create a diagram
        - No analysis possible without diagram



Product	Petrol			Diesel			Total		
	N	X	Total	N	X	Total	N	X	Total
Unit	KL	KL	KL	KL	KL	KL	KL	KL	KL
Session Year	(1)	(2)	(3) = (1) + (2)	(4)	(5)	(6) = (4) + (5)	(4)	(5)	(6) = (4) + (5)
2017-18	80	40	120	25	35	60	105	75	180
2018-19	70	50	120	20	40	60	90	90	180

3. **Diagrammatic representation of data**
   - Can be helpful for layman (without having much knowledge of numbers)
   - Hidden trend can be traced
   - Table is more accurate than diagrams
   - Types of Diagram below:

Line Diagram/ Histogram:

- plotting points in graph and join them to make a line
- used generally for time series (variable y is plotted against time t)
- for wide fluctuation, log chart or ratio chart is used (log y is plotted against t)
- for two or more series of same unit – multiple line chart is used
- for two or more series of distinct unit – multiple axes chart is used
- Refer Material for Diagram

Bar Diagram

- Bar means rectangle of same width and of varying length drawn horizontally or vertically
- For comparable series – multiple or grouped bar diagrams can be used
- For data divided into multiple components – subdivided or component bar diagrams
- For relative comparison to whole, percentage bar diagrams or divided bar diagrams

Pie Chart

- Used for circular presentation of relative data (% of whole)
- Summation of values of all components/segments are equated to 360 Degree (total angle of circle)
- Segment angle = $\frac{\text{segment value x } 360°}{\text{total value}}$

D. FREQUENCY DISTRIBUTION

What is Frequency Distribution?

Frequency means number of times a particular observation is repeated. This applies to both variable and attribute. It is shown in tabular form with class interval or the observation in one column and its frequency in the other.

These are of two types

- Ungrouped/ Simple Frequency Distribution
- Grouped Frequency Distribution

Important Terms

1. **Mutually exclusive classification or Overlapping Classification**: This is usually applicable for continuous variable. An observation as UCL is excluded from the class interval and taken in the class where it is LCL.
   Example: in the below class interval where will the observation 20 fall?

Class	Class where 20 will fall
10-20	No – excluded
20-30	Yes
30-40	No

2. **Mutually inclusive classification or Non Overlapping Classification**: This is usually applicable to discrete variable. All observation including UCL and LCL will be taken in the same class interval as there is no confusion.
   Example:

Class	Class where 20 will fall
10-19	No
20-29	Yes
30-39	No

3. **Class Limit:** for a class interval CL is the minimum and maximum value the class interval may contain. Minimum = Lower Class Interval (LCL) and Maximum = Upper Class Interval (UCL)
   Example:

Class	Type	LCL	UCL	Class	Type	LCL	UCL
10-19	Mutually Inclusive	10	19	10-20	Mutually Exclusive	10	20
20-29	Mutually Inclusive	20	29	20-30	Mutually Exclusive	20	30
30-39	Mutually Inclusive	30	39	30-40	Mutually Exclusive	30	40

4. **Class Boundary:** These are actual class limits of a class interval
   a. **For Mutually Exclusive / Overlapping :** Class Boundary = Class Limit
      LCL = LCB, UCL = UCB
   b. **For Mutually Inclusive / Non Overlapping:** Mid of the two class limits
      LCB = LCL – D/2, UCB = UCL + D/2

Example:

Class	Type	LCL	UCL	LCB	UCB	Class	Type	LCL	UCL	LCB	UCB
10-19	Mutually Inclusive	10	19	9.5	19.5	10-20	Mutually Exclusive	10	20	10	20
20-29	Mutually Inclusive	20	29	19.5	29.5	20-30	Mutually Exclusive	20	30	20	30
30-39	Mutually Inclusive	30	39	29.5	39.5	30-40	Mutually Exclusive	30	40	30	40

5. **Mid Point/ Mid Value of Class / Class Mark**

$$\frac{\text{LCL+UCL}}{2} \text{ or } \frac{\text{LCB+UCB}}{2}$$

6. **Width / Size of Class Interval**
   UCB – LCB

7. **Cumulative Frequency**

Class	Frequency	Less than type CF	More than type CF
10-20	5	5	18
20-30	2	7	13
30-40	8	15	11
40-50	3	18	3
Total	18		

8. **Frequency Density**

$$\frac{\text{Frequency of class}}{\text{Class length of that class}}$$

9. **Relative Frequency or % Frequency**

$$\frac{\text{Frequency of class}}{\text{Total Frequency of table}}$$

Class	Frequency	Class Length	Frequency Density	Relative Frequency	Percent Frequency
10-20	5	10	0.5	5/18	27.7%
20-30	2	10	0.2	2/18	11.11%
30-40	8	10	0.8	8/18	44.44%
40-50	3	10	0.3	3/18	16.67%
Total	18				

Graphical Presentation of Frequency Distribution

1. **Histogram/ Area Diagram** [refer study material page 14.20 for diagram]
   a. It is a convenient way to represent FD
   b. Comparison between frequency of two different classes possible
   c. It is useful to calculate mode also
   d. Steps to create
      - Covert CL into CB and plot in x axis

- Form rectangles taking class interval as base (x axis)
- And frequency as length (y axis) | Use frequency density in case of uneven length

2. **Frequency Polygon**
   a. Usually preferable for ungrouped frequency distribution
   b. Can be used for grouped also but only if class lengths are even
   c. Steps to create
      - Plot (x_i, f_i) where x_i = class value (in case of ungrouped), mid value (in case of grouped) and f_i = frequency
      - Join all plotted points to make line segments which eventually will become a polygon (a shape with multiple number of line segments)

3. **Ogives/ Cumulative Frequency Graph**
   a. Create a table where cumulative frequency is mapped against each CB (Class Boundary) and make a curve by plotting and joining points by line segments. (curve is called Ogive)
   b. This graph can be made by both type of Cumulative Frequency and called as Less than Ogive or More than Ogive
   c. It can be used for calculating quartiles also
   d. If we plot both ogives in same graph, perpendicular line drawn from their intersection towards x axis is cutting axis at Median

4. **Frequency Curve**
   a. It is a limiting form of Area Diagram (Histogram) or frequency polygon
   b. It is obtained by drawing smooth and free hand curve though the mid points
   c. These are of below four types:
      - Bell Shaped
      - U-Shaped
      - J-Shaped
      - Combination of Curves as Mixed Curve

Central Tendency

Meaning	Central Tendency is the tendency of a given set of observations to cluster around a single central or middle value and the single value that represents the given set of observations is described as a measure of central tendency or, location, or average.	
Arithmetic Mean	Definition	the sum of all the observations divided by the number of observations
	Formula for discrete distribution	$\bar{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n}$ or $\frac{\Sigma x}{n}$
	Formula for frequency distribution	$\bar{x} = \frac{\Sigma fx}{N}$ $N = \Sigma f$, x = mid-point in case of grouped frequency distribution
	Deviation Method	$\bar{x} = A + \frac{\Sigma fd}{N} \times C$, where $d = \frac{(x-A)}{C}$ A = assumed mean, C = class length
	Properties	→ If all the observations are constant, AM is also constant → the algebraic sum of deviations of a set of observations from their AM is zero → AM is affected both due to change of origin and scale → Combined Mean: $\bar{x}_c = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$
Median (one of the partition values)	Definition	the middle-most value when the observations are arranged either in an ascending order or a descending order of magnitude
	For Discrete Distribution	Step 1: Arrange data in ascending (or descending) order Step 2: Use the formula $\left[\frac{n+1}{2}\right]^{th}$ term
	For Frequency Distribution *(refer example 15.1.6 – Page 15.8 Study Mat)*	**Step 1**: Prepare a less than type cumulative frequency distribution with Class boundaries as base. **Step 2:** Calculate N/2 and check between which class boundaries it falls. Mark LCB as l_1 and l_2 and corresponding cumulative FD as N_l and N_u **Step 3**: Apply the below formula $Me = l_1 + \left[\frac{\frac{N}{2} - N_l}{N_u - N_l}\right] \times Class\ length$
	Properties	→ Median is affected by both change of origin and scale → For a set of observations, the sum of absolute deviations is minimum, when the deviations are taken from the median.

Partition Values

Meaning	values dividing a given set of observations into a number of equal parts
Median	Median is also a quartile that divides the set of observations into two equal parts.

Quartiles

Number of equal parts	Four (4)
Number of Quartiles	Three (3)
Denoted by	Q_1, Q_2, Q_3

Deciles

Number of equal parts	Ten (10)
Number of Deciles	Nine (9)
Denoted by	$D_1, D_2, D_3, \dots, D_9$

Percentiles

Number of equal parts	Hundred (100)
Number of Percentiles	Ninety Nine (99)
Denoted by	$P_1, P_2, P_3, \dots, P_{99}$

How to calculate Partition Values

P^{th} Quartile	$(n+1)P^{th} term,$ $here\ p = \frac{1}{4}, \frac{2}{4}, \frac{3}{4},$
P^{th} Decile	$(n+1)P^{th} term,$ $here\ p = \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \dots, \frac{9}{10}$
P^{th} Percentile	$(n+1)P^{th} term,$ $here\ p = \frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \dots, \frac{99}{100}$

Mode

Definition	Mode is the value that occurs the maximum number of times.
Type of Mode	A distribution can be uni-modal, bi-modal or multi-modal
For Frequency Distribution	$Mode = l_1 + \left[\frac{f_0 - f_{-l}}{2f_0 - f_{-l} - f_1}\right] \times Class\ length$ Here, $f_0 = frequency\ of\ the\ modal\ class,$ $f_{-1} = frequency\ of\ pre-modal\ class,$ $f_1 = frequency\ of\ the\ post\ modal\ class$

Empirical Relationship

For a moderately skewed distribution,

$$Mean - Mode = 3 \times (Mean - Median)$$

Geometric Mean

Definition	For a given set of n positive observations, the geometric mean is defined as the n^{th} root of the product of the observations
Formula	$G = (x_1 \times x_2 \times x_3 \ldots \times x_n)^{1/n}$
Properties	→ $\log G = \frac{1}{n}\Sigma \log x$ → If all observations are constant GM is also constant → $GM\ of\ xy = GM\ of\ x \times GM\ of\ y$ → $GM\ of \frac{x}{y} = \frac{GM\ of x}{GM\ of\ y}$

Harmonic Mean

Definition	For a given set of non-zero observations, harmonic mean is defined as the reciprocal of the AM of the reciprocals of the observation
Formula	$H = \frac{n}{\Sigma(1/x)}$
Properties	→ If all observations are constant HM is also constant → Combined HM: $\bar{x}_c = \frac{n_1+n_2}{\frac{n_1}{H_1}+\frac{n_2}{H_2}}$

When to use GM and HM

In case of rates like speed, hours per day, etc.	HM is used
In case of % and ratios	GM is used

Relationship between AM, GM and HM

General	AM ≥ GM ≥ HM
When all the observations are same	AM = GM = HM
When all the observations are distinct	AM > GM > HM

Ideal Measure of Central Tendency

Best Measure – Overall	AM
Best Measure for Open End Class	Median
Based on all observations	AM, GM, HM
Based on 50% values	Median
Not affected by Sampling fluctuations	Median
Rigidly defined, easy to comprehend	AM, Median, GM, HM
No Mathematical Property	Mode

Dispersion

Definition		Dispersion for a given set of observations may be defined as the amount of deviation of the observations, usually, from an appropriate measure of central tendency
Types of Dispersion	Absolute Measures of Dispersion	These are with units and not useful for comparison of two variables with different units. Example: Range, Mean Deviation, Standard Deviation, Quartile Deviation
	Relative Measures of Dispersion	These are unit free measures and useful for comparison of two variables with different units. Example: Coefficient of Range, Coefficient of Mean Deviation, Coefficient of variation, Coefficient of Quartile Deviation
Range	Definition	Difference between the largest and smallest of observations.
	Formula	Range = L – S
	Relative Measure	Coefficient of Range = $\frac{L-S}{L+S} \times 100$
	Properties	→ No effect of change of origin but affected by change of scale in the magnitude (ignore sign).
Mean Deviation	Definition	Mean deviation is defined as the arithmetic mean of the absolute deviations of the observations from an appropriate measure of central tendency
	Formula	$MD_A = \frac{1}{n}\Sigma\lvert x - A\rvert$ Here A is mean or median as given in question
	Relative Measure	Coefficient of Mean Deviation $= \frac{\text{Mean Deviation about A}}{\text{A}} \times 100$
	Properties	→ No effect of change of origin but affected by change of scale in the magnitude (ignore sign).
Standard Deviation	Definition	It is defined as the root mean square deviation when the deviations are taken from the AM of the observations
	Formula	$SD_x \text{ or } \sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{\text{n}}} \text{ or } \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}$
	Relative Measure	Coefficient of Variation $= \frac{\text{SD}}{\text{AM}} \times 100$
	Standard Result	For any two numbers, a and b: SD $= \frac{\lvert a-b\rvert}{2}$ SD of first n natural numbers: $\sqrt{\frac{(n^2-1)}{12}}$
	Properties of SD	→ If all the observations are constant, SD is Zero → No effect of change of origin but affected by change of scale in the magnitude (ignore sign) → Combined SD $= \sqrt{\frac{n_1{s_1}^2+n_2{s_2}^2+n_1{d_1}^2+n_2{d_2}^2}{n_1+n_2}}$

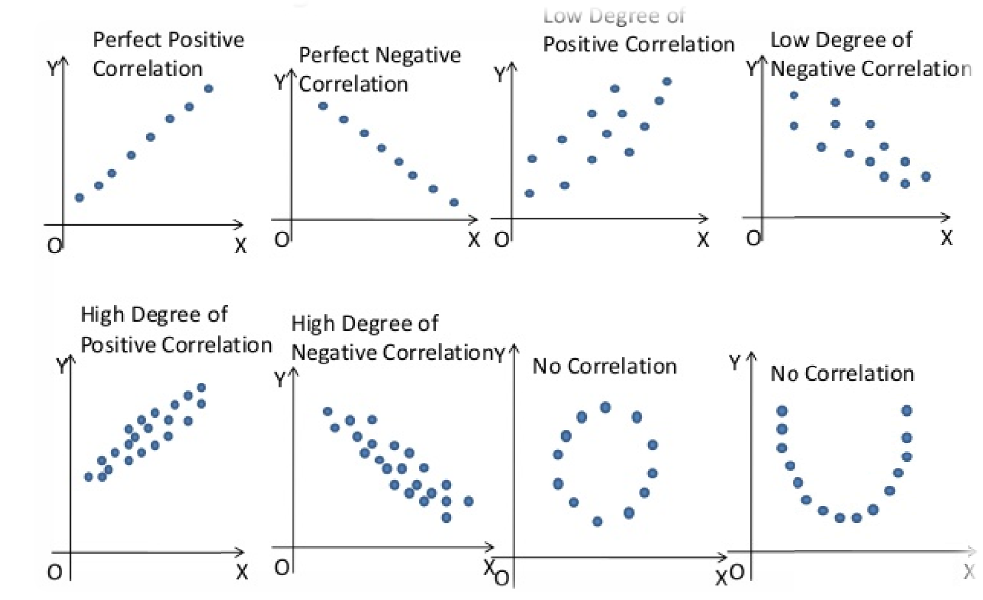
Quartile Deviation	Definition	It is defined as the semi-inter quartile range
	Formula	$Q_d = \frac{Q_3 - Q_1}{2}$
	Relative Measure	Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

Ideal Measure of Dispersion	Best Measure – Overall	SD
	Best Measure for Open End Class	QD
	Quickest to compute	Range
	Not based on all observations	Range
	Difficult to comprehend and less Mathematical	Mean Deviation
	Rigidly defined, easy to comprehend	Mean Deviation, SD, QD
	Not affected by Sampling fluctuations	QD

CORRELATION

<table>
<tr><td>Bi-Variate Data</td><td>When data are collected on two discrete variables simultaneously, they are known as Bi-Variate data</td></tr>
<tr><td>Bi-Variate Distribution</td><td>Distribution of Bi-Variate data is called as Bivariate Distribution</td></tr>
<tr><td>Bi-Variate Frequency Distribution</td><td>
<table>
<tr><td>Meaning</td><td>Frequency distribution involving two discrete variables.</td></tr>
<tr><td>Marginal Distribution</td><td>If we make a separate distribution from bi-variate frequency distribution where we take aggregate of only one variable at a time. Total no. of marginal distributions = 2</td></tr>
<tr><td>Conditional Distribution</td><td>If we make a separate distribution from bi-variate frequency distribution where we take one variable related one class interval of another variable. Total no. of conditional distributions = m + n (m = no. of rows, n = no. of columns)</td></tr>
</table>
</td></tr>
<tr><td>Correlation</td><td>While studying two variables at the same time, if it is found that the change in one variable leads to change in the other variable either directly or inversely, then the two variables are known to be associated or correlated.
<table>
<tr><td>Positive Correlation</td><td>If two variables move in the same direction</td></tr>
<tr><td>Negative Correlation</td><td>If two variables move in the opposite direction</td></tr>
<tr><td>No Correlation</td><td>If no change due to each other</td></tr>
</table>
</td></tr>
<tr><td>Measure of Correlation</td><td>A measurement or formula that represents the nature/ direction and/or magnitude of correlation.
<table>
<tr><th>Method</th><th>Helps in obtaining</th></tr>
<tr><td>Scatter Diagram</td><td>Only direction of correlation</td></tr>
<tr><td>Karl Pearson's Product moment correlation coefficient</td><td>Direction as well as strength of correlation. Best Method – Most accurate</td></tr>
<tr><td>Spearman's rank correlation co-efficient</td><td>Direction as well as strength of correlation. Useful for attributes.</td></tr>
<tr><td>Co-efficient of concurrent deviations</td><td>Direction as well as strength of correlation. Only preferred for direction and not magnitude. Quickest method.</td></tr>
</table>
</td></tr>
</table>

Scatter Diagram



Karl Pearson's Product moment correlation coefficient

Defined as	the ratio of covariance between the two variables to the product of the standard deviations of the two variables
Main Formula	$r_{xy} = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$
Formula for Covariance	$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$ or $\frac{\Sigma xy}{n} - \bar{x}.\bar{y}$
Formula for Standard Deviation σ_x or σ_y	$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{\text{n}}}$ or $\sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}$
Properties	→ It is a unit-free measurement → Value of r lies from -1 to +1 both inclusive → Change of origin or Scale Change of Origin: No impact Change of Scale: No impact of value, but if change of scale of both variables are of different sign then sign r will also change

Change of Origin	No impact
Change of Scale	No impact of value, but if change of scale of both variables are of different sign then sign r will also change

Spearman's Rank Correlation coefficient

Applied to	find the level of agreement (or disagreement) between two judges so far as assessing a qualitative characteristic is concerned
Main Formula	$r_R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$, here d means difference in ranks
Adjustment Value in case of Tie Rank	$\Sigma \frac{(t^3-t)}{12}$ here t is a tie length and we need to do summation of all ties
Formula in case of Tie length	$r_R = 1 - \frac{6(\Sigma d^2 + \text{value of adjustment})}{n(n^2-1)}$

Co-efficient of concurrent deviations	Use	A very simple and casual method of finding correlation when we are not serious about the magnitude of the two variables
	Steps in this method	This method involves in attaching a positive sign for a x-value (except the first) if this value is more than the previous value and assigning a negative value if this value is less than the previous value. Applies to both variable and then these signs are compared. If signs match – pair is counted as concurrent deviation.
	Formula	$$r_c = \pm\sqrt{\pm\frac{2c-m}{m}}$$ Here, *m = total no. of deviations (it is one less than total no. of pairs under observation i.e m=n-1), c = no. of concurrent deviations,* r_c *also lies between -1 and 1 incl.*

REGRESSION

Regression Analysis	Estimation of one variable for a given value of another variable on the basis of an average mathematical relationship between the two variables

Estimation of Y (when it is dependent on X)	Line	Regression line of Y on X
	Regression Coefficient	Regression Coefficient of Y on X denoted by $\boldsymbol{b_{yx}}$
	Form	$\boldsymbol{Y - \bar{Y} = b_{yx}\,(X - \bar{X})}$, $\bar{X}$ *and* $\bar{Y}$ *are means of X series and Y series*
Estimation of X (when it is dependent on Y)	Line	Regression line of X on Y
	Regression Coefficient	Regression Coefficient of X on Y denoted by $\boldsymbol{b_{xy}}$
	Form	$\boldsymbol{X - \bar{X} = b_{xy}\,(Y - \bar{Y})}$, $\bar{X}$ *and* $\bar{Y}$ *are means of X series and Y series*

Important Theory Points	When linear relationship exists between two variables (i.e. correlation is perfect, $r_{xy} = -1\ or\ +1$)	The linear equation so arrived can be used both ways for Y on X and X on Y. It means regression lines are identical.
	When no linear relationship exist between two variables	In that case, we need to estimate the regression lines with the help of Method of Least Squares
	To derive regression line of y on x	The minimisation of vertical distances in the scatter diagram is to be done
	To derive regression line of x on y	The minimisation of horizontal distances in the scatter diagram is to be done

Regression Coefficient	Defined as the ratio of	$\frac{\text{Covariance between two variables}}{\text{Variance of Independent variable}}$
	Regression Coefficient of Y on X	$b_{yx} = r.\frac{\sigma_y}{\sigma_x}$ or $b_{yx} = \frac{Cov(x,y)}{\sigma_x{}^2}$
	Regression Coefficient of X on Y	$b_{xy} = r.\frac{\sigma_x}{\sigma_y}$ or $b_{xy} = \frac{Cov(x,y)}{\sigma_y{}^2}$
	r used here is Karl Pearson's Correlation Coefficient	
Properties of Regression lines and coefficient	Change of origin	The regression coefficients remain unchanged
	Change of scale	: If original pair is X, Y and modified pair is U, V where U = $\frac{X-m}{p}$ and V = $\frac{Y-n}{q}$, then $b_{vu} = b_{yx}\frac{q}{p}$, $b_{uv} = b_{xy}\frac{p}{q}$
	Intersection of two regression lines	Two regression (if not identical) will intersect at the point $(\bar{x}, \bar{y})$ [means]
	Relation between correlation and regression coefficients	$r = \pm\sqrt{\pm b_{xy} \times b_{yx}}$ b_{xy}, b_{yx} and r all will have same sign
Coefficient of Determination	Coefficient of Determination	r^2 (square of correlation coefficient)
	Interpretation of value of r^2	It explains the percentage of variation in dependent variable due to variation in independent variable
	Example: if $r_{xy} = 0.8, then\ r^2 = 0.64$	It means 64% of variation in ***x*** is due to variation in **y** and remaining 36% due to other factors. It shows the reliability of correlation coefficient.
Probable Error	Formula	Probable Error [P.E] = $^2/_3$ × Standard Error [S.E.]
	Standard Error	$\frac{1-r^2}{\sqrt{n}}$
	Use	Probable Error is used to test the reliability of ***r***
	Test	If r is less than PE: The value of r is not significant. Not reliable If r is greater than six times of PE: The value of r is significant and there is evidence of correlation