Question – 1:

1. Types of Data

A. Qualitative Data (Categorical Data)

Definition: Non-numerical data that describes characteristics or qualities.

Used for: Categorization and labeling, not for mathematical operations.

Examples: Eye color (Blue, Green), Gender (Male, Female), Blood type (A, B, AB, O)

B. Quantitative Data (Numerical Data)

Definition: Numerical data that can be measured and quantified.

Used for: Mathematical calculations and statistical analysis.

Types of Quantitative Data:

Discrete Data: Countable values (e.g., Number of students = 25)

Continuous Data: Measurable values (e.g., Height = 5.4 ft)

2. Scales of Measurement

Scale Type	Type of Data	Ordered?	Equal Intervals?	True Zero?	Example
Nominal	Qualitative	No	No	No	Gender, Eye Color
Ordinal	Qualitative	Yes	No	No	Movie Ratings, Education Level
Interval	Quantitative	Yes	Yes	No	Temperature (°C), IQ Score
Ratio	Quantitative	Yes	Yes	Yes	Age, Weight, Salary

Question – 2:

Measures of Central Tendency

Measures of central tendency are statistical values that represent the center or typical value of a dataset. They summarize data by identifying a single value that best describes the entire distribution.

1. **Mean** (Arithmetic Average)

Definition: The sum of all data values divided by the number of values.

Formula: Mean = (Sum of all data points) / (Number of data points)

When to use: Use the mean when data is quantitative, continuous, and symmetrically distributed without outliers.

Example: Test scores: 80, 85, 90, 95, 100. Mean = (80+85+90+95+100) / 5 = 90

Limitations: The mean is sensitive to extreme values (outliers), which can skew it.

2. **Median**

Definition: The middle value when data are ordered from smallest to largest. If there's an even number of observations, it's the average of the two middle numbers.

When to use: Use the median when data is skewed or contains outliers, or for ordinal data.

Example: House prices (in $1000s): 200, 220, 240, 5000, 5200. Median = 240 (middle value). Mean would be skewed higher because of the expensive houses.

Strength: It is not affected by outliers and provides a better central location for skewed data.

3. **Mode**

Definition: The value that appears most frequently in the data set.

When to use: Use the mode for categorical data or to identify the most common item. Also useful in quantitative data where the most frequent value matters.

Example: Shoe sizes sold: 7, 8, 8, 9, 10. Mode = 8 (most common size).

Strength: Can be used with nominal data where mean and median cannot.

Note: There can be more than one mode (bimodal or multimodal data), or no mode if all values are unique.

Summary Table

Measure	Definition	When to Use	Example	Strengths	Limitations
Mean	Average of all values	Symmetrical quantitative data without outliers	Average test score	Uses all data points	Sensitive to outliers

Median	Middle value in ordered data	Skewed data or data with outliers	Median house price	Not affected by outliers	Doesn't use all data information
Mode	Most frequent value	Categorical data or to find common value	Most common shoe size	Works with nominal data	May be no mode or multiple modes

---

Question – 3:

Concept of Dispersion, Variance, and Standard Deviation

Concept of Dispersion

Dispersion (also called variability or spread) describes how much the data points in a dataset differ from the central value (like the mean) and from each other

- If data points are close to each other and the mean, dispersion is low.
- If data points are spread out over a wider range, dispersion is high.

Variance

Definition: Variance measures the average of the squared differences between each data point and the mean.

It gives a numeric value representing how far each point is from the mean on average, squared to avoid negative differences canceling out positive ones.

Formula for Population Variance (σ^2):

$\sigma^2 = \Sigma (x_i - \mu)^2 / N$

where x_i = each data point, μ = population mean, and N = number of data points.

Formula for Sample Variance (s^2):

$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$

where $\bar{x}$ = sample mean, and n = sample size.

Interpretation: Higher variance means greater spread in the data.

Standard Deviation

Definition: The standard deviation is the square root of the variance.

Formula for Population Standard Deviation (σ):

$\sigma = \sqrt{\sigma^2}$

Formula for Sample Standard Deviation (s):

$s = \sqrt{s^2}$

Interpretation: Standard deviation expresses spread in the same units as the original data, making it easier to understand.

Measure	Description	Formula	Units	Usefulness
Variance	Average squared deviation from mean	$\Sigma (x_i - \mu)^2 / N$ or $\Sigma (x_i - \bar{x})^2 / (n - 1)$	Squared units (e.g., m^2)	Quantifies spread but less interpretable due to squared units
Standard Deviation	Square root of variance	√variance	Same units as data	Easier to interpret; shows average distance from mean

---

Question – 4:

A **box plot** (also called a **box-and-whisker plot**) is a graphical way to display the distribution of a dataset based on five summary statistics:

Minimum (smallest value, excluding outliers)

First quartile (Q1) (25th percentile)

Median (Q2) (50th percentile)

Third quartile (Q3) (75th percentile)

Maximum (largest value, excluding outliers)

The “box” shows the middle 50% of the data between Q1 and Q3, called the **interquartile range (IQR)**. The “whiskers” extend to the smallest and largest values within 1.5 times the IQR from the quartiles. Points outside this range are considered **outliers** and are plotted individually.

Question – 5:

Random sampling is the process of selecting a subset of individuals from a population such that every member has an equal chance of being included.

Aspect	Role of Random Sampling
Bias	Minimizes selection bias
Representativeness	Ensures sample reflects population diversity
Statistical Assumptions	Validates assumptions underlying inference
Estimation	Provides unbiased estimates of population parameters
Error Measurement	Enables calculation of sampling error
Generalization	Allows valid conclusions about the population

Question – 6:

Skewness and Its Types

Skewness measures the asymmetry of a data distribution.

Positive Skew (Right Skew):
Tail extends to the right; mean > median. Example: Income data.

Negative Skew (Left Skew):
Tail extends to the left; mean < median. Example: Early retirement ages.

Zero Skew (Symmetrical):
Balanced tails; mean ≈ median ≈ mode. Example: Heights in a population.

Impact of Skewness on Interpretation

Mean is pulled toward the tail; median is a better central measure for skewed data.

Skewness indicates potential outliers or asymmetry affecting statistical analysis.

May require special statistical techniques if data is heavily skewed.

Question – 7:

What is the Interquartile Range (IQR)?

The **Interquartile Range (IQR)** is a measure of statistical dispersion that represents the range within which the middle 50% of the data lie.

It is calculated as:
IQR = Q3 − Q1
where **Q1** is the 25th percentile (first quartile) and **Q3** is the 75th percentile (third quartile).

How is IQR Used to Detect Outliers?

Outliers are data points that lie unusually far from the rest of the data.

A common rule to detect outliers:

Any value **below** Q1 − 1.5 × IQR

Any value **above** Q3 + 1.5 × IQR
is considered an outlier.

Question – 8:

Conditions for Using the Binomial Distribution

- **Fixed Number of Trials (n):**
The experiment consists of a set number of trials.

- **Two Possible Outcomes:**
Each trial results in either a "success" or a "failure."

- **Constant Probability of Success (p):**
The probability of success remains the same for each trial.

- **Independent Trials:**
The outcome of any trial does not affect others.

Question – 9:

Properties of the Normal Distribution

Symmetrical and bell-shaped: The curve is perfectly symmetric around the mean.

Mean = Median = Mode: All three measures of central tendency are equal.

Defined by mean (μ) and standard deviation (σ): The shape depends on these two parameters.

Asymptotic tails: The curve approaches, but never touches, the horizontal axis.

Total area under the curve = 1: Represents total probability.

Empirical Rule (68-95-99.7 Rule)

For data following a normal distribution:

About **68%** of values lie within **±1σ** of the mean.

About **95%** lie within **±2σ** of the mean.

About **99.7%** lie within **±3σ** of the mean.

This helps quickly estimate the spread and probability of values in a normal distribution.

Question – 10:

Example of a Poisson Process

Scenario:
A call center receives an average of 3 calls per hour. The number of calls arriving in each hour follows a Poisson distribution.

Calculate the Probability

Question:
What is the probability that exactly 5 calls will be received in the next hour?

Solution

Average rate (λ) = 3 calls per hour

Number of calls (k) = 5

Poisson probability formula:

$$P(X = k) = \{\lambda^k e^{\{-\lambda\}}\}/\{k!\}$$

$$P(X = 5) = \{3^5 \times e^{\{-3\}}\}/\{5!\} = \{243 \times e^{\{-3\}}\}/\{120\}$$

The probability of receiving exactly 5 calls in an hour is approximately **0.1008** (or 10.08%).

Question 11:

A **random variable** is a numerical outcome of a random phenomenon or experiment. It assigns a number to each possible outcome in a sample space.

Feature	Discrete	Continuous
Values	Countable (finite or infinite)	Infinite within a range
Examples	0, 1, 2, 3…	1.5, 2.01, 3.141…
Probability	Calculated for **exact values**	Calculated over an **interval**

Question 12:

Covariance

X (Study Hours)	Y (Test Score)
2	65
3	70
5	75
6	80
8	95

Step 1: Calculate the Mean

Mean of X: (2+3+5+6+8)/5 = 4.8

Mean of Y: (65+70+75+80+95)/5 = 77

Step 2: Compute Covariance

Cov(X,Y) = (1/n) * Σ(Xi - $\bar{X}$)(Yi - $\bar{Y}$)

X	Y	X - $\bar{X}$	Y - $\bar{Y}$	(X - $\bar{X}$)(Y - $\bar{Y}$)
2	65	-2.8	-12	33.6
3	70	-1.8	-7	12.6
5	75	0.2	-2	-0.4
6	80	1.2	3	3.6
8	95	3.2	18	57.6

Cov(X,Y) = (33.6 + 12.6 - 0.4 + 3.6 + 57.6)/5 = 21.4

Step 3: Compute Correlation

$\text{Corr}(X,Y) = \text{Cov}(X,Y) / (\sigma X * \sigma Y)$

$\sigma X = \sqrt{[(1/5) * \Sigma(Xi - \bar{X})^2]} \approx \sqrt{5.36} \approx 2.31$

$\sigma Y = \sqrt{[(1/5) * \Sigma(Yi - \bar{Y})^2]} \approx \sqrt{116} \approx 10.77$

$\text{Corr}(X,Y) \approx 21.4 / (2.31 * 10.77) \approx 0.86$
