ANSWERS :-

ANS-1 DATA- DATA is collection of facts or piece of information that can be stored , measured and re-accessed. Types of data-Data can be broadly categorized into qualitative and quantitative .

#1 Qualitative Data (Categorical Data) Definition: Qualitative data describes qualities or characteristics. It is non-numerical and used to categorize or label attributes. Examples: Education Level (High School, Bachelor's, Master's, PhD) Customer Satisfaction (Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied) Pain Level (None, Mild, Moderate, Severe)

#2 Quantitative Data (Numerical Data) Definition: Quantitative data represents quantities and can be measured numerically. Examples: Height (e.g., 180 cm) Weight (e.g., 70 kg) Age (e.g., 25 years) Income (e.g., $50,000)

ANS-2 Measures of central tendency are statistical metrics that summarize a set of data by identifying the central point within that dataset. The three most common measures of central tendency are the mean, median, and mode. The mean is the average of a set of numbers, calculated by adding all the values together and dividing by the number of values.The median is the middle value of a dataset when the numbers are arranged in ascending or descending order.The mode is the value that appears most frequently in a dataset. A dataset may have one mode, more than one mode (bimodal or multimodal), or no mode at all.

ANS-3 Dispersion refers to how spread out or scattered the data values are around the central tendency (like the mean). It shows how much the values differ from each other.

Variance and standard deviation both measure how spread out the data is.

Standard deviation is more commonly used because it's in the original data units.

ANS-4 A box plot (or box-and-whisker plot) is a graphical summary that shows the distribution of a dataset through its five-number summary:

Minimum, First quartile (Q1),Median (Q2), Third quartile (Q3), Maximum.

A box plot helps you quickly see if the data is symmetrical, skewed, or has outliers.

ANS-5 Random sampling plays a crucial role in making inferences about populations in statistical research. It is a method used to select a subset of individuals from a larger population in such a way that every individual has an equal chance of being chosen.By ensuring representativeness, reducing bias, and allowing for the application of statistical methods, random sampling is essential for drawing accurate conclusions and making informed decisions based on sample data.

ANS-6 Skewness is a statistical measure that describes the asymmetry of a probability distribution. It indicates the direction and degree of distortion from the symmetrical bell curve (normal distribution). In simpler terms, skewness helps us understand whether the data points in a dataset are concentrated on one side of the mean or are evenly distributed around it. Types of Skewness Positive Skewness (Right Skewness) Negative Skewness (Left Skewness) Zero Skewness (Symmetrical Distribution)

Effects of Skewness on Data Interpretation Mean vs. Median Data Analysis and Decision Making Outlier Detection risk assesment.

ANS-7 The interquartile range (IQR) is a measure of statistical dispersion that represents the range within which the central 50% of a data set lies. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1):

[\text{IQR} = Q3 - Q1]

Q1 (First Quartile): The median of the lower half of the data set (25th percentile). Q3 (Third Quartile): The median of the upper half of the data set (75th percentile). How IQR is Used to Detect Outliers Outliers are data points that are significantly different from the rest of the data. The IQR is commonly used to identify these outliers through the following steps:

Calculate Q1 and Q3: Determine the first and third quartiles of the data set.

Calculate the IQR: Subtract Q1 from Q3.

Determine the Outlier Boundaries:

Calculate the lower boundary: (Q1 - 1.5 \times \text{IQR}) Calculate the upper boundary: (Q3 + 1.5 \times \text{IQR}) Identify Outliers: Any data point that falls below the lower boundary or above the upper boundary is considered an outlier.

ANS-8 The conditions under which the binomial distribution is used are as follows:

Fixed Number of Trials (n): The experiment consists of a predetermined number of trials, denoted as (n). This number must be constant throughout the experiment.

Two Possible Outcomes: Each trial results in one of two outcomes, commonly referred to as "success" and "failure." For example, in a coin toss, the outcomes could be heads (success) and tails (failure).

Constant Probability of Success (p): The probability of success, denoted as (p), remains constant for each trial. This means that the likelihood of achieving a success does not change from one trial to the next.

Independence of Trials: The trials must be independent, meaning the outcome of one trial does not affect the outcome of another. For instance, the result of one coin toss does not influence the result of the next toss.

ANS-9 Normal Distribution The normal distribution is a symmetric, bell-shaped curve that describes how many natural phenomena are distributed. Empirical Rule (68-95-99.7 Rule) In a normal distribution: 68% of the data falls within 1 standard deviation of the mean (μ ± 1σ). 95% falls within 2 standard deviations (μ ± 2σ). 99.7% falls within 3 standard deviations (μ ± 3σ). The normal distribution describes many real-world variables. The empirical rule helps quickly understand the spread and likelihood of values within a normal dataset.

ANS-10 Let's calculate the probability that the call center receives exactly 5 calls in one hour.

Set (\lambda = 10) (the average number of calls). Set (k = 5) (the number of calls we want to find the probability for). Using the formula:

[P(X = 5) = \frac{e^{-10} \cdot 10^5}{5!}]

Calculating each component:

(e^{-10} \approx 0.0000453999) (10^5 = 100000) (5! = 120) Now plug these values into the formula:

[P(X = 5) = \frac{0.0000453999 \cdot 100000}{120}]

Calculating the numerator:

[0.0000453999 \cdot 100000 \approx 4.53999]

Now divide by (120):

[P(X = 5) \approx \frac{4.53999}{120} \approx 0.03783]

ANS-11 A random variable is a variable that represents the possible outcomes of a random experiment. It assigns numerical values to those outcomes. Random variables can be classified into two main types: discrete random variables and continuous random variables.

Discrete Random Variable

Takes on countable values (finite or countably infinite). Often involves whole numbers. Examples: Number of students in a class Number of goals in a match Number of heads in 10 coin flips

Continuous Random Variable

Takes on infinite values within a given range. Can take any value (fractions, decimals) in an interval. Examples: Height of a person (e.g., 172.5 cm) Time taken to run a race Temperature in a city

ANS-12 **covariance** and **correlation**.

Example Dataset

We have two variables:

- X: Hours Studied
- Y: Exam Scores

Student	Hours Studied (X)	Exam Score (Y)
A	2	65
B	4	70
C	6	75
D	8	85
E	10	95

Step 1: Calculate Means

$$\acute{X} = \frac{2+4+6+8+10}{5} = 6$$

$$\acute{Y} = \frac{65+70+75+85+95}{5} = 78$$

Step 2: Calculate Covariance

$$\text{Cov}(X,Y) = \frac{\sum (X_i - \acute{X})(Y_i - \acute{Y})}{n-1}$$

$$¿ \frac{(2-6)(65-78)+(4-6)(70-78)+(6-6)(75-78)+(8-6)(85-78)+(10-6)(95-78)}{4}$$

$$¿ \frac{(-4)(-13)+(-2)(-8)+(0)(-3)+(2)(7)+(4)(17)}{4}$$

$$¿ \frac{52+16+0+14+68}{4} = \frac{150}{4} = 37.5$$

Step 3: Calculate Correlation

$$\text{Correlation (r)} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

First, calculate the **standard deviations** of X and Y:

σ_X:

$$¿ \sqrt{\frac{(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2}{4}} = \sqrt{\frac{64}{4}} = \sqrt{16} = 4$$

$\$ \sigma_Y$:

$$¿ \sqrt{\frac{(65-78)^2+(70-78)^2+(75-78)^2+(85-78)^2+(95-78)^2}{4}} = \sqrt{\frac{770}{4}} = \sqrt{192.5} \approx 13.87$$

Now compute the correlation:

$$r = \frac{37.5}{4 \times 13.87} \approx \frac{37.5}{55.48} \approx 0.676$$

Interpretation

- Covariance = 37.5: Positive, meaning as - hours studied increases, exam scores also tend to increase. But it’s hard to interpret scale-wise.
- Correlation ≈ 0.676: Shows a moderate to strong positive linear relationship between hours studied and exam scores. Correlation is scale-free and ranges from -1 to 1.