STATISTICS

Q.20. What do you mean by Measure of Central Tendency and Measures of Dispersion .How it can be calculated.

ANS>> Definition: Measures of central tendency are statistical metrics that describe the center point or typical value of a dataset. They provide a summary statistic that represents the middle of the data distribution.

Definition: Measures of dispersion describe the spread or variability of a dataset. They indicate how much the data points differ from the central tendency (mean, median, or mode).

Calculation Examples:

For a dataset: [2, 4, 4, 4, 5, 5, 7, 9]

1. **Mean:**

   MEAN = (2+4+4+4+5+7+9) / 8 = 5

2. **Median:**

   Arranged data: [2, 4, 4, 4, 5, 5, 7, 9]

   Median: (4 + 5)/2 = 4.5 (since there are 8 data points)

3. **Mode:**

   The mode is 4 (it appears most frequently).

4. **Range:**

   Range=9−2=7

5. **Variance:**

   Variance = [(2−5)^2+(4−5)^2+(4−5)^2+(4−5)^2+(5−5)^2+(5−5)^2+(7−5)^2+(9−5)2^2] / 8

   Variance = 4

6. **Standard Deviation:**

Standard Deviation = (4)^1/2 = 2

7. **IQR:**

   Q1 (25th percentile): 4

   Q3 (75th percentile):

   IQR = 6-4 = 2

Q 21. What do you mean by skewness.Explain its types.Use graph to show.

ANS >> ## Skewness

Definition: Skewness refers to the degree of asymmetry in a distribution of data. It indicates whether the data points are concentrated more on one side of the mean than the other, giving the distribution a "skewed" appearance.

Positive Skew (Right-Skewed):

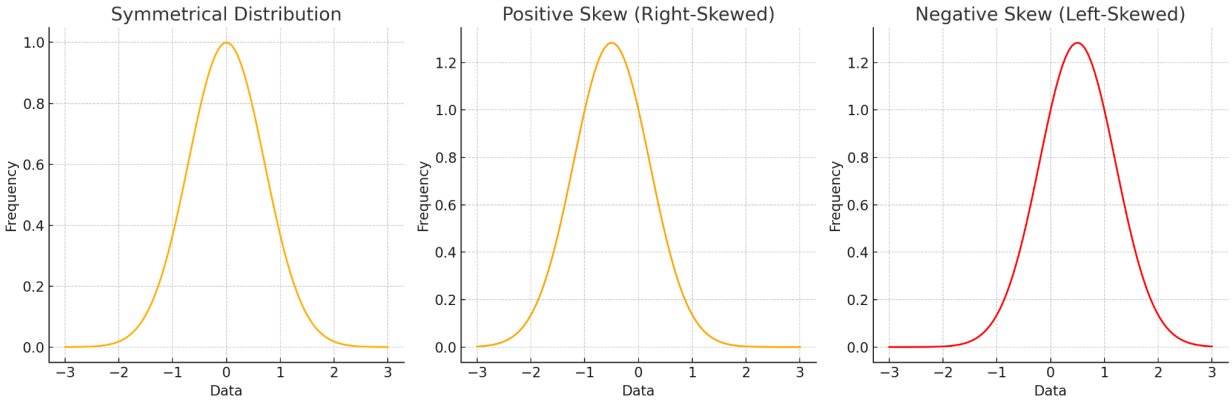
- **In a positively skewed distribution, the tail on the right side is longer or fatter than the left side. The bulk of the data points are concentrated on the left side, and the mean is greater than the median.**
- **Example: Income distribution, where a small number of people have very high incomes, skewing the distribution to the right.**

Negative Skew (Left-Skewed):

- **In a negatively skewed distribution, the tail on the left side is longer or fatter than the right side. The bulk of the data points are concentrated on the right side, and the mean is less than the median.**
- **Example: Scores on a difficult exam, where most students score low but a few score very high, skewing the distribution to the left.**

No Skew (Symmetrical or Zero Skewness):

- **In a symmetrical distribution, the left and right sides are mirror images of each other, with the mean, median, and mode all being equal or very close.**
- **Example: Normal distribution (bell curve), where data points are evenly distributed around the mean.**



Q 22. Explain PROBABILITY MASS FUNCTION (PMF) and PROBABILITY DENSITY FUNCTION (PDF). and what is the difference between them?

ANS>>

Probability Mass Function (PMF) :

Definition: A Probability Mass Function (PMF) is a function that provides the probability of discrete outcomes for a random variable. It assigns a probability to each possible outcome of a discrete random variable, such that the sum of all probabilities equals 1.

Properties:

1. **Discrete Random Variable: PMF applies to discrete random variables, which have a countable number of possible outcomes.**
2. **Probability Values: The PMF gives the probability P(X=x)P(X = x)P(X=x) for each outcome xxx.**
3. **Sum of Probabilities: The sum of the probabilities for all possible outcomes is 1:**

Probability Density Function (PDF) :

Definition: A Probability Density Function (PDF) describes the likelihood of a continuous random variable taking on a range of values. Unlike the PMF, the PDF does not give probabilities directly but instead provides a density value. The probability of the random variable falling within a particular interval is found by integrating the PDF over that interval.

Properties:

1. **Continuous Random Variable: PDF applies to continuous random variables, which have an infinite number of possible values.**
2. **Density Values: The value of the PDF at any given point does not represent a probability but a density. Probabilities are obtained by integrating the PDF over an interval.**
3. **Total Area Under Curve: The total area under the PDF curve is 1, representing the**

Key Differences Between PMF and PDF

1. **Applicability:**
    - **PMF: Used for discrete random variables.**
    - **PDF: Used for continuous random variables.**
2. **Probability Calculation:**
    - **PMF: Directly gives the probability for each outcome.**
    - **PDF: Probabilities are calculated by integrating the PDF over an interval, not directly from the density value.**
3. **Values and Units:**
    - **PMF: The values of a PMF are probabilities and thus lie between 0 and 1.**
    - **PDF: The values of a PDF can be greater than 1 since they represent density, not probability. The area under the curve, however, represents probability.**
4. **Summation vs. Integration:**
    - **PMF: The total probability is obtained by summing the probabilities of all outcomes.**
    - **PDF: The total probability is obtained by integrating the PDF over the entire range of possible values.**

Q 23. What is correlation. Explain its type in details.what are the methods of determining correlation

ANS >>

Correlation :

Definition: Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It quantifies the degree to which changes in one variable predict or are associated with changes in another. The correlation coefficient, denoted as rrr, ranges from -1 to +1.

r = +1: Perfect positive linear correlation, meaning as one variable increases, the other increases proportionally.

r = −1r = -1r = −1: Perfect negative linear correlation, meaning as one variable increases, the other decreases proportionally.

r = 0r = 0r = 0: No linear correlation, meaning there is no linear relationship between the variables..

Types of Correlation

1. **Positive Correlation:**
    - **Definition: Both variables move in the same direction. As one variable increases, the other also increases, and vice versa.**
    - **Example: Height and weight; typically, as height increases, weight also increases.**
    - **Range: 0 < r <= 1**
2. **Negative Correlation:**
    - **Definition: Variables move in opposite directions. As one variable increases, the other decreases.**
    - **Example: Speed of a vehicle and travel time; as speed increases, travel time decreases.**
    - **Range: -1 = < r <0**
3. **No Correlation:**
    - **Definition: No discernible linear relationship between the variables. Changes in one variable do not predict changes in the other.**
    - **Example: The number of books read and the color of cars owned; there is no linear relationship.**
    - **Value: r = 0**

Methods of Determining Correlation

1. **Pearson Correlation Coefficient:**
    - **Definition: Measures the strength and direction of the linear relationship between two continuous variables.**

Assumptions: Data should be normally distributed, and the relationship between variables should be linear.

Range: -1 to +1

Spearman's Rank Correlation Coefficient:

- **Definition: A non-parametric measure of rank correlation that assesses the relationship between two variables using their ranked values.**
-

Use Case: Useful when the data does not meet the assumptions of the Pearson correlation, such as when the data is ordinal or not normally distributed.

Range: -1 to +1

Kendall's Tau:

- **Definition: A non-parametric measure of the strength and direction of association between two variables based on the ranks of data.**

Point-Biserial Correlation:

- **Definition: Measures the relationship between a binary variable and a continuous variable.**
- **Use Case: Used when one variable is dichotomous (e.g., male/female, yes/no) and the other is continuous.**
- **Range: -1 to +1**

24. Calculate coefficient of correlation between the marks obtained by 10 students in Accountancy and

Statistics:

STUDENT	:	1	2	3	4	5	6	7	8	9	10
ACCOUNTANCY	:	45	70	65	30	90	40	50	75	85	60
STATISTICS	:	35	90	70	40	95	40	60	80	80	50

Use Karl Pearson's Coefficient of Correlation Method to find it.

ANS >> 0.903 , P.E.= 0.039

Q 25. Discuss the 4 differences between correlation and regression.

ANS >> Correlation and regression are both statistical methods used to examine the relationship between variables, but they serve different purposes and have distinct characteristics. Here are four key differences between correlation and regression:

1. Purpose and Interpretation

- **Correlation:**
  - **Purpose: Measures the strength and direction of the linear relationship between two variables.**
  - **Interpretation: The correlation coefficient (r) indicates how closely the variables are related. It ranges from -1 to +1, where values close to +1 or -1 indicate a strong linear relationship, and values close to 0 indicate a weak or no linear relationship.**

Regression:

- **Purpose: Describes the nature of the relationship between two variables by establishing a predictive model. It estimates how one variable (dependent) changes as another variable (independent) changes.**
- **Interpretation: Regression provides an equation that can be used to predict the dependent variable based on the independent variable(s). The slope of the regression line indicates the direction and magnitude of the relationship.**

2. Causality

- **Correlation:**
  - **Causality: Does not imply causation. A correlation between two variables indicates a relationship but does not confirm that changes in one variable cause changes in the other.**
- **Regression:**
  - **Causality: While regression analysis can suggest a cause-and-effect relationship by predicting the dependent variable from the independent variable(s), it does not establish causation definitively. Other methods and experiments are needed to confirm causal relationships.**

3. Symmetry

- **Correlation:**
  - **Symmetry: Correlation is symmetric, meaning the correlation coefficient between X and Y is the same as that between Y and X.**
- **Regression:**
  - **Symmetry: Regression is not symmetric. The regression line for predicting Y from X is different from the regression line for predicting X from Y. This asymmetry is due to the specific roles of the dependent and independent variables in the analysis.**

4. Scope of Analysis

- **Correlation:**
  - **Scope: Typically focuses on the relationship between two variables, indicating how they vary together. It does not provide a model for prediction.**
- **Regression:**
  - **Scope: Goes beyond correlation by providing a mathematical model that describes the relationship between variables. Regression can involve multiple independent variables (multiple regression) and is used for making predictions and forecasting.**

Q 26. Find the most likely price at Delhi corresponding to the price of Rs. 70 at Agra from the following data: Coefficient of correlation between the prices of the two places +0.8.

ANS >> WHERE IS DATA QUESTION INCOMPLETE

Q.27. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of x = 9, Regression equations are: (i) 8x−10y = −66; (ii) 40x − 18y = 214. What are (a) the mean values of x and y, (b) the coefficient of correlation between x and y, (c) the σ of y.

ANS>>

(a) The mean values of X and Y are :x bar = 13 , y bar = 17

(b) The coefficient of correlation r is approximately : 0.8

(c) The standard deviation of Y (σy) is approximately : 3.75

Q 28. What is Normal Distribution? What are the four Assumptions of Normal Distribution? Explain in detail.

ANS>> The normal distribution is a fundamental concept in statistics, commonly known for its bell-shaped curve. It is a continuous probability distribution that is symmetric about its mean, showing that data near the mean are more frequent in occurrence than data far from the mean. Here's a detailed explanation of the normal distribution and its assumptions:

Normal Distribution

1. **Definition: The normal distribution is defined by its probability density function (PDF)**
2. **μ is the mean of the distribution, which also represents the peak of the bell curve.**
3. **σ2\sigma^2σ2 is the variance, which measures the spread of the distribution. σ\sigmaσ is the standard deviation, the square root of the variance.**

Assumptions of Normal Distribution

1. **Linearity of Relationships:**
    - **In a normal distribution, the relationship between variables should be linear. For example, if you're analyzing two variables, their relationship should be linear, which implies that any scatterplot of the two variables should show a straight-line relationship if they are normally distributed.**
2. **Independence:**
    - **The observations or data points should be independent of each other. This means that the occurrence of one observation should not affect the occurrence of another. In practical terms, each data point should come from a different, non-overlapping sample or unit.**
3. **Homoscedasticity:**
    - **The variance of the data should be constant across the range of values of the independent variable. This implies that the spread or dispersion of the data should be uniform across all levels of the independent variable. In simpler terms, the variability of the data should be the same regardless of the value of the predictor.**
4. **Normality of Residuals:**

- **When modeling data, the residuals (the differences between observed and predicted values) should be normally distributed. This means that any errors or deviations in predictions should follow a normal distribution. This assumption is crucial for valid hypothesis testing and confidence interval estimation in linear regression models.**

Q 29.Write all the characteristics or Properties of the Normal Distribution Curve.

ANS >> The normal distribution curve, also known as the Gaussian distribution or bell curve, has several key characteristics and properties that define its shape and behavior. Here's a detailed overview:

Symmetry:

- **The curve is perfectly symmetrical about its mean, μ\muμ. This means that the left half of the curve is a mirror image of the right half.**

Mean, Median, and Mode:

- **The mean (μ\muμ), median, and mode of the distribution are all equal and located at the center of the curve. This central point is where the peak of the bell curve occurs.**

Bell-Shaped Curve:

- **The shape of the normal distribution curve is bell-shaped. It rises gradually to a peak at the mean and then falls off symmetrically on both sides.**

Asymptotic:

- **The tails of the normal distribution curve approach the horizontal axis asymptotically. This means the curve never actually touches the horizontal axis but gets infinitely close to it.**

Q 30.Which of the following options are correct about Normal Distribution Curve.

(a) Within a range 0.6745 of σ on both sides the middle 50% of the observations occur i,e. mean ±0.6745σ

covers 50% area 25% on each side.

(b) Mean ±1S.D. (i,e.μ ± 1σ) covers 68.268% area, 34.134 % area lies on either side of the mean.

(c) Mean ±2S.D. (i,e. μ ± 2σ) covers 95.45% area, 47.725% area lies on either side of the mean.

(d) Mean ±3 S.D. (i,e. μ ±3σ) covers 99.73% area, 49.856% area lies on the either side of the mean.

(e) Only 0.27% area is outside the range μ ±3σ.

ANS >> (a), (b), and (e) are correct

Q 31. The mean of a distribution is 60 with a standard deviation of 10. Assuming that the distribution is normal, what percentage of items be (i) between 60 and 72, (ii) between 50 and 60, (iii) beyond 72 and (iv) between 70 and 80?

ANS>> Given: Mean (μ\muμ) = 60 , Standard deviation (σ\sigmaσ) = 10

(i) Between 60 and 72

1. **Calculate the Z-scores:**
   - **For X = 60**
     Z = (60-60) /10 = 0
   - **For X = 72**
   - **Z = (72-60) / 10 = 1.2**

Look up the Z-scores in the Z-table:

- **The Z-score of 0 corresponds to 0.5000 (50%).**
- **The Z-score of 1.2 corresponds to approximately 0.8849 (88.49%).**

Find the percentage between the two Z-scores:

Percentage = (0.8849-0.5000) * 100 = 38.49%

(i) Between 60 and 72: Approximately 38.49%

(ii) Between 50 and 60: Approximately 34.13%

(iii) Beyond 72: Approximately 11.51%

(iv) Between 70 and 80: Approximately 13.59%

Q 32. 15000 students sat for an examination. The mean marks was 49 and the distribution of marks had a standard deviation of 6. Assuming that the marks were normally distributed what proportion of students scored (a) more than 55 marks, (b) more than 70 marks

ANS>> (a) More than 55 marks: Approximately 15.87%

(b) More than 70 marks: Approximately 0.02%

Convert the raw scores to Z-scores using the formula: $Z = [X - \mu] / \sigma$

Q 33. If the height of 500 students are normally distributed with mean 65 inch and standard deviation 5 inch. How many students have height : a) greater than 70 inch. b) between 60 and 70 inch.

ANS>> (a) Greater than 70 inches: Approximately 79 students

(b) Between 60 and 70 inches: Approximately 341 students

Z-scores using the formula: $Z = [X - \mu] / \sigma$

34. What is the statistical hypothesis? Explain the errors in hypothesis testing.b)Explain the Sample. What are Large Samples & Small Samples?

ANS>> Statistical Hypothesis

Definition: A statistical hypothesis is a statement or an assumption about a population parameter or a relationship between variables that can be tested using statistical methods. The hypothesis serves as the basis for statistical inference and allows researchers to draw conclusions about the population based on sample data

Types of Hypotheses:

1. **Null Hypothesis (H0):**
    - **The null hypothesis is a statement of no effect, no difference, or no relationship. It represents the default or baseline assumption that there is no significant effect or relationship in the population.**
    - **Example: "There is no difference in mean test scores between two teaching methods."**
2. **Alternative Hypothesis (H1 OR Ha):**

- **The alternative hypothesis is a statement that contradicts the null hypothesis. It represents the presence of an effect, difference, or relationship.**
- **Example: "There is a difference in mean test scores between two teaching methods."**

Errors in Hypothesis Testing

In hypothesis testing, two types of errors can occur:

1. **Type I Error (α Error):**
   - **Definition: Occurs when the null hypothesis is incorrectly rejected when it is actually true.**
   - **Example: Concluding that a new drug is effective when, in reality, it is not.**
   - **Probability: The significance level (α) of the test is the probability of making a Type I error. Common values are 0.05, 0.01, etc.**
2. **Type II Error (β Error):**
   - **Definition: Occurs when the null hypothesis is not rejected when the alternative hypothesis is actually true.**
   - **Example: Concluding that a new drug is not effective when, in reality, it is effective.**
   - **Probability: The probability of making a Type II error is denoted by β\betaβ. The complement of β (i.e., 1−β1) is called the power of the test.**

Sample

Definition: A sample is a subset of individuals or observations selected from a larger population, used to estimate characteristics of the entire population. The sample should be representative of the population to provide accurate and valid estimates.

Types of Samples:

1. **Random Sample:**
   - **Every individual in the population has an equal chance of being selected. This minimizes bias and ensures that the sample is representative.**
2. **Stratified Sample:**

- The population is divided into strata (subgroups) based on certain characteristics, and random samples are taken from each stratum.

3. Systematic Sample:
   - A sample is selected using a systematic approach, such as every nth individual from a list.
4. Convenience Sample:
   - Samples are taken from individuals who are easiest to reach. This method may introduce bias and is less representative.

Large Samples & Small Samples

Large Samples:

Definition: A sample is considered large when it contains a sufficiently large number of observations, generally 30 or more. The exact definition can vary based on the context and statistical method

Small Samples:

Definition: A sample is considered small when it contains fewer than 30 observations. The exact cutoff can vary, but this is a common guideline.

Q 35.A random sample of size 25 from a population gives the sample standard derivation to be 9.0. Test the hypothesis that the population standard derivation is 10.5.Hint(Use chi-square distribution).

Ans >> The chi-square test statistic is approximately 17.64. Given the critical values for df=24 and α=0.05 (12.40 and 39.36), we fail to reject the null hypothesis. There is insufficient evidence to conclude that the population standard deviation is different from 10.5 at the 5% significance level

37.100 students of a PW IOI obtained the following grades in Data Science paper :

Grade :[A, B, C, D, E]

Total Frequency :[15, 17, 30, 22, 16, 100]

Using the χ^ 2 test , examine the hypothesis that the distribution of grades is uniform.

ANS >> Expected frequency for each grade = (Total number of students) /(Number of grades) = 20

Calculate the Chi-Square Test Statistic:

The chi-square test statistic is calculated using the formula:

X^2 = SUMMATION[{O-E} / E]

For each grade:

Grade A: [(15-20)^2/20] = 1.25

SIMILAR CALCULATE : B,C,D,E

Summing these values:

$\chi 2$=1.25+0.45+5.00+0.20+0.80=7.70

Df = 5−1 = 4

for df = 4 and α=0.05 The critical value is approximately 9.488.

Here, the test statistic $\chi 2$=7.70 is less than the critical value of 9.488

Q 38 .Anova Test :

To study the performance of three detergent and three different

Water temperatures the following whiteness reading were obtained with specially designed equipment.

Water temp :	detergent A	detergent B	detergentC
Cold water :	57	55	67
Worm water :	49	52	68
Hot water :	54	46	58

ANS >>

Calculate the Mean of Each Group:

- **Detergent A:**
  - **Cold Water: 57**
  - **Warm Water: 49**
  - **Hot Water: 54**
  - **Mean for Detergent A: 57+49+543 / 3 = 53.33**
- **Detergent B:**
  - **Cold Water: 55**
  - **Warm Water: 52**
  - **Hot Water: 46**
  - **Mean for Detergent B: 55+52+46 / 3 = 51**
- **Detergent C:**
  - **Cold Water: 67**
  - **Warm Water: 68**
  - **Hot Water: 58**
  - **Mean for Detergent C: 67+68+58 / 3 = 64.33**
- **Water Temperature (Overall Means):**
  - **Cold Water: 57+55+67/3=59**
  - **Warm Water: 49+52+68/3=56.33**
  - **Hot Water: 54+46+58/3=52.6**
  - **Grand Mean:**
    Grand Mean=57+55+67+49+52+68+54+46+58/9=56

Calculate Sum of Squares (SS):

SSTotal=(57−56)^2+(55−56^)2+(67−56)^2+⋯+(58−56)^2 = 424

Sum of Squares for Detergent (SS Detergent):

SSDetergent = 3 · [7.11+25.00+70.70] = 3 · 102.81 = 308.43

Sum of Squares for Water Temperature (SS Temperature):

SSTemperature = 3 · [9.00+0.11+11.78] = 3 · 20.89 = 62.67

Sum of Squares for Interaction (SS Interaction):

SSInteraction = 424−308.43−62.67 = 53.90

Calculate Degrees of Freedom (df):

df(Error) = (9−1)−2−2−4 = 0

Calculate Mean Squares (MS):

MSdetergent : 308.43/2 = 154.22

MStemperature = 62.67/2 = 31.33

MSinteraction = 53.90/4 = 13.48

Calculate the F-ratios:

Since,MSerror cannot be calculated with Dferror = 0