Nikita kakde…

Statistics Basics| **Assignment**

**Question 1:** What is the difference between descriptive statistics and inferential statistics? Explain with examples.

**Answer:**

|  |
| --- |
| Descriptive Statistics   1. It is the branch of statistics that describes and summarizes data in a simple, understandable form. 2. It organizes, simplifies, and presents data without making any predictions. 3. It is limited to the collected data only and does not go beyond it. 4. It uses measures like mean, median, mode, range, standard deviation, and visual tools like graphs, pie charts, bar diagrams, and tables. 5. It only tells what is happening in the data, not why or what will happen. 6. A teacher calculates that the average score of her class of 50 students is 72 marks. This is just describing the data she has.   Inferential Statistics   1. It is the branch of statistics that draws conclusions, predictions, or generalizations about a larger population using a sample of data. 2. It helps to make decisions, test hypotheses, and forecast results for the whole population. 3. It goes beyond the available data and involves probability. 4. It uses methods like hypothesis testing, confidence intervals, regression, correlation, and probability models. 5. It tells what may be true for the population or what might happen in the future. 6. From a survey of 100 students, we predict that the average marks of all students in the school lie between 70–74 marks. |

**Question 2:** What is sampling in statistics? Explain the differences between random and stratified sampling.

**Answer:**

|  |
| --- |
| Sampling in Statistics  Sampling is the process of selecting a part of the population (sample) to study and draw conclusions about the whole population. Since studying every individual in a population is often difficult, time-consuming, and costly, sampling is used to save effort and resources while still giving reliable results  Random Sampling   1. Every individual in the population has an equal chance of being selected. 2. Selection is completely by chance, without any grouping. 3. Simple to use and avoids personal bias. 4. Example: Picking 100 students randomly from a school list of 1000 students.   Stratified Sampling   1. The population is divided into groups (called strata) based on some characteristic like gender, age, income, etc. 2. Then, samples are taken proportionally from each group. 3. Ensures that all groups are represented properly. 4. Example: If a school has 60% boys and 40% girls, and we take a sample of 100 students, then 60 boys and 40 girls are selected |

**Question 3:** Define mean, median, and mode. Explain why these measures of central tendency are important.

**Answer:**

|  |
| --- |
| Mean  Mean is the average of all observations.  It is found by adding all values and dividing by the number of values.  Example: For 10, 20, 30 → Mean = (10+20+30)/3 = 20.  Median  Median is the middle value when all data is arranged in order (ascending or descending).  If there are two middle values, the median is the average of those two.  Example: For 10, 20, 30, 40, 50 → Median = 30.  Mode  Mode is the value that occurs most frequently in the data set.  A data set can have one mode, more than one mode, or no mode.  Example: For 2, 4, 4, 6, 7 → Mode = 4.  Importance of Mean, Median, and Mode   1. They are called measures of central tendency because they show the “central value” or “typical value” of data. 2. They help to summarize large data into a single representative figure. 3. They are useful in comparison between different groups or datasets. 4. They are widely used in business, economics, education, research, and daily life for decision making. |

**Question 4**: Explain skewness and kurtosis. What does a positive skew imply about the data?

**Answer:**

|  |
| --- |
| Skewness  Skewness tells us about the asymmetry of a distribution (whether data is spread more to the left or right).  If data is perfectly symmetrical, skewness = 0.  Types:  Positive skew (Right skewed) → Tail is longer on the right side, most values are on the left.  Negative skew (Left skewed) → Tail is longer on the left side, most values are on the right.  Kurtosis  Kurtosis tells us about the peakedness or flatness of a distribution compared to a normal distribution.  High kurtosis = sharper peak and fatter tails (more extreme values).  Low kurtosis = flatter peak and thinner tails (fewer extreme values).  Positive Skew Meaning   1. In a positively skewed distribution, the tail is stretched to the right. 2. The mean is usually greater than the median. 3. It shows that a majority of values are small, but there are a few very large values pulling the tail to the right. 4. Example: Income distribution in a country → most people earn low to moderate income, but a few earn very high income. |

**Question 5:** Implement a Python program to compute the mean, median, and mode of a given list of numbers.

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

(*Include your Python code and output in the code box below.*)

**Answer:**

|  |
| --- |
| # Python program to compute mean, median, and mode  import statistics as stats  numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]  mean\_value = stats.mean(numbers)  median\_value = stats.median(numbers)  mode\_value = stats.mode(numbers)  print("Mean:", mean\_value)  print("Median:", median\_value)  print("Mode:", mode\_value)  **Output:**  Mean: 19.6  Median: 19  Mode: 12 |

**Question 6:** Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

list\_x = [10, 20, 30, 40, 50]

list\_y = [15, 25, 35, 45, 60]

(*Include your Python code and output in the code box below.*)

**Answer:**

|  |
| --- |
| import numpy as np  # Given lists  list\_x = [10, 20, 30, 40, 50]  list\_y = [15, 25, 35, 45, 60]  cov\_matrix = np.cov(list\_x, list\_y, ddof=0)  cov\_xy = cov\_matrix[0, 1]  corr\_xy = np.corrcoef(list\_x, list\_y)[0, 1]  print("Covariance:", cov\_xy)  print("Correlation Coefficient:", corr\_xy)  **Output:**  Covariance: 220.0  Correlation Coefficient: 0.995893206467704 |

**Question 7**: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

(*Include your Python code and output in the code box below.*)

**Answer:**

|  |
| --- |
| import numpy as np  data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]  Q1 = np.percentile(data, 25)  Q3 = np.percentile(data, 75)  IQR = Q3 - Q1  lower\_bound = Q1 - 1.5 \* IQR  upper\_bound = Q3 + 1.5 \* IQR  outliers = [x for x in data if x < lower\_bound or x > upper\_bound]  print("Q1:", Q1)  print("Q3:", Q3)  print("IQR:", IQR)  print("Lower Bound:", lower\_bound)  print("Upper Bound:", upper\_bound)  print("Outliers:", outliers)  Explanation   1. The boxplot summarizes the distribution of the data using median, quartiles, and spread. 2. The Interquartile Range (IQR) is 6 (24 − 18). 3. Values above 33 or below 9 are considered outliers. 4. Hence, 35 is the only outlier in this dataset. 5. Most values lie between 14 and 29, with a balanced spread around the median. |

**Question 8**: You are working as a data analyst in an e-commerce company. The marketing team wants to know if there is a relationship between advertising spend and daily sales.

* Explain how you would use covariance and correlation to explore this relationship.
* Write Python code to compute the correlation between the two lists:

**advertising\_spend = [200, 250, 300, 400, 500]**

**daily\_sales = [2200, 2450, 2750, 3200, 4000]**

(*Include your Python code and output in the code box below.*)

**Answer:**

|  |
| --- |
| **Covariance** tells us whether the two variables move in the same direction or not.   1. Positive covariance → when advertising spend increases, sales also increase. 2. Negative covariance → when advertising spend increases, sales decrease. 3. The limitation: it does not show the **strength** of the relationship.   **Correlation** standardizes covariance to a scale between **-1 and +1**.   1. **+1** → perfect positive relationship. 2. **-1** → perfect negative relationship. 3. **0** → no linear relationship. 4. This helps us measure both the **direction and strength** of the relationship between advertising spend and sales.   import numpy as np  # Data  advertising\_spend = [200, 250, 300, 400, 500]  daily\_sales = [2200, 2450, 2750, 3200, 4000]  # Compute correlation matrix  correlation = np.corrcoef(advertising\_spend, daily\_sales)  print("Correlation Coefficient:", correlation[0,1])  Interpretation:   1. The correlation coefficient is 0.991, which is very close to +1. 2. This shows a strong positive relationship between advertising spend and daily sales. 3. Meaning → as the company spends more on advertising, sales increase significantly. |

**Question 9**: Your team has collected customer satisfaction survey data on a scale of 1-10 and wants to understand its distribution before launching a new product.

* Explain which summary statistics and visualizations (e.g. mean, standard deviation, histogram) you’d use.
* Write Python code to create a histogram using Matplotlib for the survey data:

survey\_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

(*Include your Python code and output in the code box below.*)

**Answer:**

|  |
| --- |
| To understand the distribution of customer satisfaction survey data (scale 1–10):   * **Summary Statistics** to use:   + **Mean** → average satisfaction score.   + **Median** → middle score, useful if the data is skewed.   + **Mode** → most common satisfaction score.   + **Standard Deviation (SD)** → how spread out the scores are.   + **Minimum & Maximum** → range of scores. * **Visualizations** to use:   + **Histogram** → shows frequency distribution of survey scores.   + **Boxplot** → highlights median, spread, and outliers.   + These help identify whether scores are **clustered** or **widely spread**.   import matplotlib.pyplot as plt  import numpy as np  from statistics import mean, median, mode, pstdev  # Data  survey\_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]  # Summary Statistics  print("Mean:", mean(survey\_scores))  print("Median:", median(survey\_scores))  print("Mode:", mode(survey\_scores))  print("Standard Deviation:", pstdev(survey\_scores))  print("Min:", min(survey\_scores))  print("Max:", max(survey\_scores))  # Histogram  plt.hist(survey\_scores, bins=6, edgecolor='black', alpha=0.7)  plt.title("Histogram of Customer Satisfaction Scores")  plt.xlabel("Survey Scores (1-10)")  plt.ylabel("Frequency")  plt.show() |