**Assignment Code: DS-AG-005**

# Statistics Basics| **Assignment**

**Instructions:** Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

**Total Marks**: 200

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

**Answer:**

|  |
| --- |
| 1. **Descriptive Statistics**  Meaning: Descriptive statistics refers to methods of summarizing, organizing, and presenting data in a meaningful way.  Purpose: To describe the main features of a dataset without making conclusions beyond the given data.  Techniques Used:   * Measures of central tendency → Mean, Median, Mode * Measures of dispersion → Range, Variance, Standard Deviation * Measures of symmetricity → skewness, kurtosis   Example:  Suppose you collected the marks of 50 students in a class.  Average marks = 72  Highest marks = 95  Standard deviation = 10  These numbers only describe the data of those 50 students.  2. **Inferential Statistics**  Meaning: Inferential statistics uses sample data to make predictions, generalizations, or decisions about a larger population.  Purpose: To draw conclusions and test hypotheses about a population based on sample data.  Techniques Used:   * Hypothesis testing * Confidence intervals * Regression analysis * ANOVA (Analysis of Variance)   Example:  From the same class of 50 students, you take a sample of 10 students and find their average marks = 74.  Using inferential statistics, you estimate that the average marks of all 50 students (population) is likely between 70 and 78 (confidence interval).  Here, you are going beyond the sample to make conclusions about the population.  ***Key Difference Summarized***  Descriptive statistics: describe a dataset.  Inferential statistics: generalize from a sample to a population. |

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

**Answer:**

|  |
| --- |
| **Sampling** is the process of selecting a subset of individuals (sample) from a larger group (population) to study and make conclusions about the population.  Purpose: Since studying the entire population is often time-consuming, costly, or impractical, sampling provides a manageable way to gather information.  The Differences between random and stratified sampling:   * ***Random Sampling***   Definition: In random sampling, every individual in the population has an equal chance of being selected.  Method: Selection is purely by chance (like a lottery draw).  Advantage: Simple and unbiased.  Limitation: May not represent different subgroups proportionally.  Example:  From a college of 1,000 students, if you randomly pick 100 students using a lottery or random number generator, that’s random sampling.   * ***Stratified Sampling***   Definition: In stratified sampling, the population is divided into subgroups (strata) based on some characteristic (like age, gender, income), and samples are taken proportionally from each stratum.  Advantage: Ensures representation of all groups; gives more accurate results when the population is diverse.  Limitation: Requires prior knowledge of population characteristics.  Example:  In the same college of 1,000 students:  600 are girls, 400 are boys.  If you want a sample of 100 students, you take 60 girls and 40 boys (proportionally).  This ensures both groups are represented → that’s stratified sampling.  Key Difference Summarized :  Simple random sampling selects individuals from the population with an equal chance of inclusion, whereas stratified sampling first divides the population into distinct, non-overlapping subgroups (strata) based on shared characteristics and then randomly selects a sample from each stratum. |

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

**Answer:**

|  |
| --- |
| **Definition**   * Mean (Arithmetic Average) : The mean is the sum of all values in a dataset divided by the number of values. * Median: The median is the middle value when all observations are arranged in ascending (or descending) order. * Mode : The mode is the value that occurs most frequently in the dataset.   **Importance of Mean, Median, and Mode**  These are called measures of central tendency because they indicate the center or typical value of data.  Why important?  1. Summarization: They condense a large dataset into a single representative value.  2. Comparison: Help compare two or more groups (e.g., average income of two cities).  3. Decision-making: Useful in business, economics, and research for policy and planning.  4. Handling different data:  Mean → useful when data is evenly distributed.  Median → useful when data has outliers (e.g., incomes where few rich people skew the mean).  Mode → useful for categorical data (e.g., most preferred product brand). |

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

**Answer:**

|  |
| --- |
| **Skewness**  *Definition*: Skewness measures the degree of asymmetry of a distribution around its mean.  *Types*:  1. Symmetrical (Zero Skewness): Mean = Median = Mode  2. Positive Skew (Right Skew): Tail on the right-hand side is longer → Mean > Median > Mode  3. Negative Skew (Left Skew): Tail on the left-hand side is longer → Mean < Median < Mode  Example:  Income distribution in a country is usually positively skewed (a few very high incomes pull the mean upwards).  **Kurtosis**  *Definition*: Kurtosis measures the peakedness or flatness of a distribution compared to a normal distribution.  Types:  1. Mesokurtic (k = 3): Normal bell-shaped curve.  2. Leptokurtic (k > 3): More peaked, heavy tails → data has more extreme values.  3. Platykurtic (k < 3): Flatter curve, lighter tails → fewer extreme values.  **Positive Skew (Right Skew) – What it Implies**  *Meaning*: A positive skew implies that the distribution has a longer right tail.  *Implication*:   * Most values are concentrated on the lower side (left side), * A few very high values pull the mean to the right. * Hence, Mean > Median > Mode.   Example: Distribution of salaries → most people earn low-to-moderate incomes, while a few earn extremely high salaries, making the data positively skewed. |

**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:**

***Paste your code and output inside the box below:***

|  |
| --- |
|  |

**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:**

***Paste your code and output inside the box below:***

|  |
| --- |
|  |

**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:**

|  |
| --- |
| **Boxplot & Outlier Detection Method**  A boxplot shows median, quartiles, and possible outliers.  IQR (Interquartile Range) = Q3 – Q1  Lower Fence = Q1 – 1.5 × IQR  Upper Fence = Q3 + 1.5 × IQR  Any value outside these fences is considered an outlier.  **Python Code**  *Explanation of Result*   * The boxplot will show most data clustered between 17 and 24. * The value 35 lies above the upper fence (32), so it is identified as an outlier. * No values fall below the lower fence (10), so no low outliers exist. |

**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:**

|  |
| --- |
| 1. **Using Covariance and Correlation**   **Covariance**  Definition: Covariance measures the direction of the relationship between two variables.  Interpretation:  Positive covariance → when advertising spend increases, sales also increase.  Negative covariance → when advertising spend increases, sales decrease.  Value depends on units, so not easy to interpret magnitude.  **Correlation**  Definition: Correlation measures the strength and direction of the relationship between two variables.  Range: -1 ≤ r ≤ +1  r close to +1 → strong positive relationship  r close to -1 → strong negative relationship  r ≈ 0 → no linear relationship  Better than covariance because it is unit-free and standardized.  ***Application to the problem***:  Compute covariance → tells if ad spend and sales move in the same direction.  Compute correlation → tells how strongly ad spend is related to sales.  If correlation is very high (close to +1), we can say that higher advertising spend strongly drives higher sales.   1. **Python Code to Compute Correlation**     *Expected Output (approximately):*   * Covariance: Large positive value (indicating both variables move together). * Correlation: Close to +0.99 or +1 → very strong positive relationship. |

**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:**

|  |
| --- |
| 1. **Summary Statistics to Use**  To understand customer satisfaction survey data (scale 1–10), we should calculate:  Mean → Average satisfaction score.  Median → Middle score (less affected by outliers).  Mode → Most frequent score (common opinion).  Standard Deviation (SD) → Spread of scores; higher SD means more varied opinions.  Range (Min–Max) → Lowest and highest ratings.  **Visualizations to Use**  Histogram → Shows frequency distribution of scores (how many customers gave 7, 8, etc.).  Boxplot → Shows median, quartiles, and potential outliers.  Bar Chart (for categorical frequencies) → Useful if you want to see counts of each rating.   1. **Python Code (with Histogram)**   *The Output Will Show*:   * Summary Statistics printed: Mean ≈ 7.4, Median = 7, Mode = 7, SD ≈ 1.5 (approx). * Histogram: A bar graph showing most responses are clustered around 7–9, meaning customers are generally satisfied. |