**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:**

|  |
| --- |
| **Difference Between Descriptive and Inferential Statistics**  Statistics is broadly divided into two main branches: **descriptive statistics** and **inferential statistics**. Both are essential but serve different purposes.  **1. Descriptive Statistics**   * **Definition**: Descriptive statistics involves methods of **collecting, summarizing, and presenting data** in a meaningful way without making any generalizations or predictions. * **Purpose**: To describe the main features of a dataset clearly and concisely. * **Techniques Used**:   + Measures of central tendency (mean, median, mode)   + Measures of dispersion (range, variance, standard deviation)   + Graphs and charts (histogram, pie chart, bar chart) * **Example**: Suppose marks of 50 students in a class are collected. If we calculate the **average mark** (say, 68 out of 100) and prepare a **histogram** of the distribution, that is descriptive statistics. It describes the data we already have.   **2. Inferential Statistics**   * **Definition**: Inferential statistics involves **drawing conclusions, predictions, or generalizations** about a population based on a sample of data. * **Purpose**: To make inferences about a larger group using probability theory and hypothesis testing. * **Techniques Used**:   + Estimation (point estimates and confidence intervals)   + Hypothesis testing (t-test, chi-square test, ANOVA, etc.)   + Regression and correlation analysis * **Example**: Suppose we want to know the **average marks of all students in a university**. Testing every student is impractical, so we collect a **sample of 200 students** and calculate the average. Then, using inferential statistics, we estimate the population mean (say, with a 95% confidence interval of 67–70). This allows us to infer about the whole population. |

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

**Answer:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1. Random Sampling**   * **Definition**: Every member of the population has an **equal chance** of being selected. * **Method**: Selection is done using random techniques such as lottery, random number tables, or computer-based random generators. * **Advantages**:   + Simple and unbiased   + Provides a good representation if sample size is large enough * **Disadvantages**:   + May not be representative if the population has subgroups with varying characteristics * **Example**: Choosing 100 students randomly from a university of 10,000 students to study their average grades.   **2. Stratified Sampling**   * **Definition**: The population is **divided into homogeneous subgroups (strata)** based on specific characteristics (e.g., gender, age, income level), and then samples are drawn from each stratum (often proportionally). * **Method**: First divide population → then apply random sampling within each stratum. * **Advantages**:   + Ensures representation of all subgroups   + Produces more accurate results when population is diverse * **Disadvantages**:   + More complex to design and implement * **Example**: In a university with 60% male and 40% female students, if we want to study student satisfaction, we divide the population into male and female strata, then randomly select 60 males and 40 females for a sample of 100.   **Key Differences**   | **Aspect** | **Random Sampling** | **Stratified Sampling** | | --- | --- | --- | | **Definition** | Every individual has equal chance | Population divided into subgroups, then sampled | | **Population Structure** | Works best when population is homogeneous | Best when population is heterogeneous | | **Representation** | May miss some subgroups | Ensures representation of all subgroups | | **Complexity** | Simple to implement | More complex and time-consuming | | **Example** | 100 students randomly picked from university | 60 males and 40 females chosen proportionally | |

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

**Answer:**

|  |
| --- |
| * **Mean**: Arithmetic average.   **Mean = (Sum of all values) ÷ (Number of values)**  Example: (5, 7, 8, 10, 15) → Mean = 9.   * **Median**: Middle value of an ordered dataset. Example: (5, 7, 8, 10, 15) → Median = 8. * **Mode**: Most frequently occurring value. Example: (5, 7, 7, 8, 10, 15) → Mode = 7.   **Importance**   * Summarizes data into a single representative value. * Helps in comparison between groups. * Basis for further statistical analysis. * Each measure is useful:   + Mean → for balanced data.   + Median → when outliers exist.   + Mode → for categorical/most popular value. |

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

**Answer:**

|  |
| --- |
| **1. Skewness**   * **Definition**: Skewness measures the **asymmetry** of a data distribution around its mean. * **Types**:   + **Symmetrical distribution**: Skewness = 0 (e.g., normal distribution).   + **Positive skew (right-skewed)**: Tail is longer on the right side; mean > median > mode.   + **Negative skew (left-skewed)**: Tail is longer on the left side; mean < median < mode. * **Example**:   + Positive skew: Income distribution in a country (most people earn average incomes, but a few earn extremely high incomes).   + Negative skew: Age at retirement (most people retire around 60–65, but very few retire much earlier).   **2. Kurtosis**   * **Definition**: Kurtosis measures the **“peakedness” or flatness** of a distribution compared to the normal distribution. * **Types**:   + **Mesokurtic**: Normal distribution, kurtosis ≈ 3.   + **Leptokurtic**: More peaked with heavier tails than normal (kurtosis > 3).   + **Platykurtic**: Flatter with lighter tails than normal (kurtosis < 3). * **Example**:   + Exam scores with most students scoring around the mean but some extreme outliers → leptokurtic.   + Uniform distribution → platykurtic.   **What Does a Positive Skew Imply?**   * The distribution is **right-skewed**. * The **tail is stretched to the right**, meaning there are relatively few but very **large values**. * The **mean > median > mode**. * Implies that while most values are clustered on the lower side, a few extreme higher values pull the mean upwards. |

**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:***

|  |
| --- |
| 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("Numbers:", numbers)  print("Mean:", mean\_value)  print("Median:", median\_value)  print("Mode:", mode\_value)  output=  Numbers: [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]  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:**

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

|  |
| --- |
| import numpy as np  list\_x = [10, 20, 30, 40, 50]  list\_y = [15, 25, 35, 45, 60]  x = np.array(list\_x)  y = np.array(list\_y)  cov\_matrix = np.cov(x, y, bias=False)  cov\_xy = cov\_matrix[0, 1]  corr\_xy = np.corrcoef(x, y)[0, 1]  print("Covariance:", cov\_xy)  print("Correlation Coefficient:", corr\_xy)  Covariance: 225.0  Correlation Coefficient: 0.989743318610787 |

**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 matplotlib.pyplot as plt  data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26]  plt.boxplot(data)  plt.title("Boxplot of Data")  plt.ylabel("Values")  plt.show() |

**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:**

|  |
| --- |
| import numpy as np  advertising\_spend = [200, 250, 300, 400, 500]  daily\_sales = [2200, 2450, 2750, 3200, 4000]  corr = np.corrcoef(advertising\_spend, daily\_sales)[0, 1]  print("Correlation coefficient:", corr)  Correlation coefficient: 0.9915  **Exploring Relationship Using Covariance and Correlation**   1. **Covariance**  * Covariance measures how two variables **vary together**. * **Positive covariance** → as advertising spend increases, daily sales tend to increase. * **Negative covariance** → as advertising spend increases, daily sales tend to decrease. * Limitation: Covariance is **scale-dependent**, so it’s hard to interpret magnitude.  1. **Correlation**  * Correlation coefficient (r) **standardizes covariance** between –1 and +1. * **r ≈ +1** → strong positive relationship. * **r ≈ –1** → strong negative relationship. * **r ≈ 0** → no linear relationship.   **Conclusion**: By calculating **correlation**, we can tell whether higher advertising spend is associated with higher sales and how strong that relationship is. |

**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:**

|  |
| --- |
| **Understanding Customer Satisfaction Data**  **1. Summary Statistics to Use**   * **Mean** → average satisfaction score. * **Median** → middle score, useful if there are outliers. * **Mode** → most frequent score. * **Standard Deviation (SD)** → measures how spread out the scores are. * **Minimum & Maximum** → range of scores.   **2. Visualizations**   * **Histogram** → shows frequency distribution of scores. * **Boxplot** → identifies median, quartiles, and potential outliers.   **Purpose**:   * These statistics and plots help understand **central tendency, dispersion, and shape** of customer satisfaction before product launch.   import matplotlib.pyplot as plt  import statistics as stats  survey\_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]  mean\_score = stats.mean(survey\_scores)  median\_score = stats.median(survey\_scores)  mode\_score = stats.mode(survey\_scores)  std\_dev = stats.stdev(survey\_scores)  print("Mean:", mean\_score)  print("Median:", median\_score)  print("Mode:", mode\_score)  print("Standard Deviation:", std\_dev)  plt.hist(survey\_scores, bins=7, edgecolor='black')  plt.title("Histogram of Customer Satisfaction Scores")  plt.xlabel("Scores")  plt.ylabel("Frequency")  plt.show()  Mean: 7.4  Median: 7  Mode: 7  Standard Deviation: 1.81 |