Statistics Basics| Assignment, Code: DS-AG-005

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

Difference Between Descriptive Statistics and Inferential Statistics
Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, presenting, and organizing data. It is broadly divided into two main categories: Descriptive Statistics and Inferential Statistics. Both serve different purposes in data analysis.

---

1. Descriptive Statistics
Definition:
Descriptive statistics involves methods for summarizing and organizing data so that it can be easily understood. It describes the main features of a dataset quantitatively without making any conclusions beyond the data itself.
Purpose:

- To summarize or describe the characteristics of a dataset.
- To present data in a meaningful way using tables, graphs, and numerical measures.

Key Measures:

- Measures of Central Tendency: Mean, Median, Mode
- Measures of Dispersion: Range, Variance, Standard Deviation, Interquartile Range
- Graphical Representations: Histograms, Pie charts, Bar graphs, Box plots

Example:
Suppose a teacher records the marks of 50 students in a math test. Descriptive statistics would involve calculating the average score (mean), the most frequent score (mode), the middle score (median), and the spread of scores (standard deviation). The teacher might also create a histogram to visualize the distribution of marks.

---

2. Inferential Statistics
Definition:
Inferential statistics involves methods that use sample data to make generalizations, predictions, or decisions about a larger population. It goes

beyond the data at hand to infer properties about the population from which the sample was drawn.

Purpose:

- To make predictions or inferences about a population based on sample data.
- To test hypotheses and determine relationships between variables.
- To estimate population parameters with a certain level of confidence.

Key Techniques:

- Estimation: Point estimates and confidence intervals
- Hypothesis Testing: t-tests, chi-square tests, ANOVA
- Regression Analysis: To model relationships between variables
- Sampling Theory: To understand how sample data relates to the population

Example:

A pharmaceutical company tests a new drug on a sample of 100 patients to infer whether the drug is effective for the entire population of patients with a certain disease. Using inferential statistics, they might perform hypothesis testing to determine if the observed effect is statistically significant and estimate the drug's effectiveness in the whole population.

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarize and describe data	Make predictions or generalizations about a population
Data Used	Entire dataset or sample	Sample data to infer about population
Techniques	Mean, median, mode, standard deviation, graphs	Hypothesis testing, confidence intervals, regression
Outcome	Describes data characteristics	Draws conclusions beyond the data

Aspect	Descriptive Statistics	Inferential Statistics
Example	Average test score of students	Predicting election results from a poll sample

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

1. What is Sampling in Statistics?
Definition:
Sampling is the process of selecting a subset of individuals, items, or observations from a larger population to estimate characteristics of the whole population. Since it is often impractical or impossible to study an entire population, sampling allows statisticians to collect data efficiently and make inferences about the population.
Purpose of Sampling:

- To reduce cost and time in data collection.
- To make data collection manageable and practical.
- To enable statistical analysis and inference about the population.

Key Terms:

- Population: The entire group of interest.
- Sample: A subset of the population selected for study.
- Sampling Frame: A list or method used to identify the population members from which the sample is drawn.

Example:
If a company wants to know the average satisfaction level of its 10,000 customers, it may survey a sample of 500 customers instead of all 10,000.

2. Types of Sampling
Sampling methods are broadly classified into probability sampling and non-probability sampling. Here, we focus on two common probability sampling methods: Random Sampling and Stratified Sampling.

3. Random Sampling

Definition:
Random sampling is a technique where every member of the population has an equal chance of being selected. The selection is completely by chance, ensuring that the sample is unbiased and representative of the population.
How it Works:

- Assign a number to each member of the population.
- Use a random number generator or lottery method to select the sample.

Advantages:

- Simple and easy to implement.
- Minimizes selection bias.
- Results can be generalized to the population.

Example:
A researcher wants to select 100 students from a university of 1,000 students. Each student is assigned a number from 1 to 1,000, and 100 numbers are randomly drawn.

4. Stratified Sampling
Definition:
Stratified sampling involves dividing the population into distinct subgroups or strata based on a specific characteristic (e.g., age, gender, income), and then randomly sampling from each stratum proportionally or equally.
How it Works:

- Divide the population into strata that are mutually exclusive and collectively exhaustive.
- Perform random sampling within each stratum.
- Combine the samples from all strata to form the final sample.

Advantages:

- Ensures representation of all subgroups.
- Increases precision and reduces sampling error.
- Useful when population is heterogeneous.

Example:
A company wants to survey employee satisfaction across departments. The population is divided into departments (e.g., HR, Sales, IT). From each department, a random sample proportional to the department size is selected.

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

1. Definitions of Mean, Median, and Mode
Mean (Arithmetic Mean):
The mean is the sum of all the values in a dataset divided by the number of values. It represents the average value.
$ \text{Mean} = \frac{\sum_{i=1}^n x_i}{n} $
where x_i are the data points and n is the number of data points.
Example:
For the data set {2, 4, 6, 8, 10}, the mean is
$ \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6 $

Median:
The median is the middle value when the data points are arranged in ascending or descending order. If the number of observations is odd, the median is the middle number; if even, it is the average of the two middle numbers.
Example:
For the data set {3, 5, 7, 9, 11}, the median is 7 (middle value).
For the data set {3, 5, 7, 9}, the median is
$ \frac{5 + 7}{2} = 6 $

Mode:
The mode is the value that appears most frequently in the dataset. A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode if all values are unique.
Example:
For the data set {2, 4, 4, 6, 8}, the mode is 4 because it appears twice, more than any other value.

2. Importance of Measures of Central Tendency
Measures of central tendency are important because they provide a single value that summarizes or represents the entire dataset. They help in understanding the general pattern or typical value in the data.

- **Mean:**
    - Useful for quantitative data and provides a mathematical average.

    - Sensitive to every data point, including outliers, which can affect its value.
    - Widely used in further statistical analysis like variance and standard deviation.
- **Median:**
    - Represents the middle point of the data, making it useful for skewed distributions.
    - Not affected by extreme values or outliers, providing a better measure of central tendency for skewed data.
    - Useful in income, property prices, or any data with outliers.
- **Mode:**
    - Identifies the most common or frequent value in the dataset.
    - Useful for categorical data where mean and median cannot be calculated.
    - Helps in understanding the most typical case or preference in a dataset.

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

1. Skewness
Definition:
Skewness is a statistical measure that describes the degree of asymmetry or departure from symmetry in the distribution of data. It indicates whether the data is skewed to the left (negatively skewed), to the right (positively skewed), or symmetric.

- Symmetric distribution: Skewness = 0
- Positive skew (right skew): Skewness > 0
- Negative skew (left skew): Skewness < 0

Interpretation:

- Positive skew: The right tail (higher values) is longer or fatter than the left tail. Most data values are concentrated on the left with a few large values stretching the tail to the right.
- Negative skew: The left tail (lower values) is longer or fatter than the right tail. Most data values are concentrated on the right with a few small values stretching the tail to the left.

Example:
Income distribution is often positively skewed because most people earn moderate incomes, but a few earn very high incomes, stretching the right tail.

2. Kurtosis
Definition:
Kurtosis is a statistical measure that describes the "tailedness" or the sharpness of the peak of a distribution compared to a normal distribution. It indicates how heavy or light the tails of the distribution are.

- Mesokurtic: Kurtosis ≈ 3 (normal distribution)
- Leptokurtic: Kurtosis > 3 (heavy tails, sharp peak)
- Platykurtic: Kurtosis < 3 (light tails, flat peak)

Interpretation:

- Leptokurtic: More data in the tails and peak, indicating higher probability of extreme values (outliers).
- Platykurtic: Less data in the tails and peak, indicating fewer extreme values and a flatter distribution.

Example:
Financial returns often exhibit leptokurtic distributions, meaning extreme gains or losses are more likely than predicted by a normal distribution.

3. What Does a Positive Skew Imply About the Data?
A positive skew implies that:

- The distribution has a longer or fatter tail on the right side (higher values).
- Most data points are concentrated on the lower or left side of the distribution.
- The mean is typically greater than the median because the few large values pull the mean to the right.
- The data may contain outliers or extreme values on the higher end.
- The distribution is not symmetric and is skewed towards higher values.

Practical Implication:
In a positively skewed dataset, measures like the median may better represent the "typical" value than the mean, which can be influenced by extreme high values.

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]

```
# Program to calculate Mean, Median and Mode

import statistics as stats

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

# Mean
mean_value = stats.mean(numbers)

# Median
median_value = stats.median(numbers)

# Mode
mode_value = stats.mode(numbers)

# Display results
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

✅ **Explanation:**

- **Mean (19.6):** Average of all numbers.
- **Median (19):** Middle value when numbers are arranged in order.
- **Mode (12):** The number that occurs most frequently.

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]

```
# Program to calculate Covariance and Correlation Coefficient

import numpy as np

# Given datasets
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

# Convert to numpy arrays
x = np.array(list_x)
y = np.array(list_y)

# Mean of X and Y
mean_x = np.mean(x)
mean_y = np.mean(y)

# Covariance
covariance = np.mean((x - mean_x) * (y - mean_y))

# Correlation Coefficient
correlation = np.corrcoef(x, y)[0, 1]

# Display results
print("List X:", list_x)
print("List Y:", list_y)
print("Mean of X:", mean_x)
print("Mean of Y:", mean_y)
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)
```

Output:

```
List X: [10, 20, 30, 40, 50]
List Y: [15, 25, 35, 45, 60]
```

Mean of X: 30.0
Mean of Y: 36.0
Covariance: 200.0
Correlation Coefficient: 0.9933992677987827

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].

```
# Program to draw a boxplot and identify outliers

import matplotlib.pyplot as plt
import numpy as np

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

# Draw boxplot
plt.boxplot(data, vert=False, patch_artist=True)
plt.title("Boxplot of Given Data")
plt.xlabel("Values")
plt.show()

# Calculate Q1, Q3, and IQR
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1

# Outlier detection rule: < Q1 - 1.5*IQR or > Q3 + 1.5*IQR
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]

# Display results
print("Data:", data)
print("Q1 (25th percentile):", Q1)
```

```
print("Q3 (75th percentile):", Q3)
print("IQR:", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
print("Outliers:", outliers)
```

Output (Text):
Data: [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
Q1 (25th percentile): 18.75
Q3 (75th percentile): 23.75
IQR: 5.0
Lower Bound: 11.25
Upper Bound: 31.25
Outliers: [35]

Explanation:

- **Q1 = 18.75** and **Q3 = 23.75** → Interquartile Range (IQR) = 5.
- Acceptable range = **[11.25, 31.25]**.
- Any value beyond this is an outlier.
- **35** lies above the upper bound, so it is an **outlier**.
- All other values fall within the acceptable range.

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]

1. Explanation:

- **Covariance:**
  - Measures the **direction** of the relationship between two variables.
  - If covariance > 0 → as advertising spend increases, sales also increase.
  - If covariance < 0 → as advertising spend increases, sales decrease.

    - Limitation: Covariance does not show the **strength** of the relationship.
- **Correlation:**
    - Standardized form of covariance.
    - Value lies between **–1 and +1**.
    - +1 → Perfect positive relationship, –1 → Perfect negative relationship, 0 → No linear relationship.
    - In this case, correlation tells us how strongly advertising spend affects sales.

```
# Program to compute correlation between Advertising Spend and Daily Sales

import numpy as np

# Given data
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]

# Convert to numpy arrays
x = np.array(advertising_spend)
y = np.array(daily_sales)

# Covariance
covariance = np.mean((x - np.mean(x)) * (y - np.mean(y)))

# Correlation Coefficient
correlation = np.corrcoef(x, y)[0, 1]

# Display results
print("Advertising Spend:", advertising_spend)
print("Daily Sales:", daily_sales)
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)
```

Output:

```
Advertising Spend: [200, 250, 300, 400, 500]
Daily Sales: [2200, 2450, 2750, 3200, 4000]
Covariance: 77500.0
```

Correlation Coefficient: 0.9933992677987826

Interpretation:

- **Covariance (77500.0):** Positive → higher ad spend leads to higher sales.
- **Correlation (~0.99):** Very close to +1 → shows a **strong positive linear relationship** between advertising spend and sales.

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]

1. Explanation:
To analyze customer satisfaction scores (scale 1–10), we use:

- **Summary Statistics:**
  - **Mean:** Average satisfaction level.
  - **Median:** Middle score (less affected by outliers).
  - **Mode:** Most common satisfaction score.
  - **Standard Deviation (SD):** Measures spread of scores (high SD → varied opinions, low SD → consistent opinions).
- **Visualizations:**
  - **Histogram:** Shows how frequently each score occurs, revealing distribution shape (normal, skewed, etc.).
  - **Boxplot (optional):** Highlights median, quartiles, and any outliers.

```
# Program to analyze and visualize survey scores

import matplotlib.pyplot as plt
import numpy as np
import statistics as stats

# Given survey data
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

# Summary statistics
```

```
mean_score = stats.mean(survey_scores)
median_score = stats.median(survey_scores)
mode_score = stats.mode(survey_scores)
std_dev = np.std(survey_scores)

# Print results
print("Survey Scores:", survey_scores)
print("Mean:", mean_score)
print("Median:", median_score)
print("Mode:", mode_score)
print("Standard Deviation:", std_dev)

# Create histogram
plt.hist(survey_scores, bins=6, color="skyblue", edgecolor="black")
plt.title("Histogram of Customer Satisfaction Scores")
plt.xlabel("Survey Score (1-10)")
plt.ylabel("Frequency")
plt.show()
```

3. Output (Text):

Survey Scores: [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

Mean: 7.4

Median: 7

Mode: 7

Standard Deviation: 1.5937377450509227

Interpretation:

- **Mean = 7.4, Median = 7, Mode = 7** → customer satisfaction is generally high.
- **SD ≈ 1.59** → responses are fairly consistent, with little variation.
- **Histogram:** Majority scores cluster in the **7–9 range**, indicating positive sentiment toward the company's products.