Assignment Code: DS-AG-005

Statistics Basics| Assignment

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

ANSWER: Descriptive Statistics

Descriptive statistics summarize and describe the features of a dataset. They do not go beyond the data at hand.

Examples of descriptive techniques:

- Mean, median, mode
- Standard deviation and variance
- Frequency distributions
- Charts and graphs (histograms, box plots, etc.)

Example: Suppose you collected the test scores of 100 students in a class. You calculate:

- Mean score: 72
- Median score: 75
- Standard deviation: 8.2

These values give you a snapshot of the performance in that particular class—nothing is inferred about other classes or the broader student population.

Inferential Statistics

Inferential statistics use sample data to make predictions or generalizations about a population. They involve probability and hypothesis testing.

Examples of inferential techniques:

- Confidence intervals
- Hypothesis testing (t-tests, chi-square tests)
- Regression analysis
- ANOVA (Analysis of Variance)

Example: You want to know if students across the entire school perform similarly to the 100 students in your class. You use their scores as a sample to:

- Estimate the average test score for all students
- Test whether boys and girls perform differently using a t-test
- Create a 95% confidence interval for the average score

Here, you're using the sample data to infer something broader—this is the essence of inferential statistics.

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

Answer: Sampling in statistics is the process of selecting a subset of individuals, items, or data points from a larger population to make statistical inferences about the whole population. Since studying an entire population can be impractical or impossible, sampling allows analysts to gain insights efficiently and with manageable resources.

Here's how the two commonly used sampling methods differ:

Random Sampling

Definition:
Every member of the population has an equal chance of being selected.

Key Features:

- Simple to implement
- Minimizes selection bias
- Best when the population is homogeneous

Example:
Suppose you have a list of 1,000 employees in a company. You randomly select 100 of them using a random number generator. Every employee, regardless of department or role, has an equal probability of being chosen.

Stratified Sampling

Definition:
The population is divided into subgroups (strata) based on shared characteristics, and samples are taken from each subgroup proportionally or equally.

Key Features:

- Ensures representation of key subgroups
- More precise estimates

- Ideal when population has distinct categories

Example:
Say the same company has three departments: HR (100 employees), Engineering (600), and Sales (300). You divide the population into these strata and sample 10% from each:

- HR: 10 employees
- Engineering: 60 employees
- Sales: 30 employees

This method guarantees that each department is fairly represented in the sample.

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

Answer: Mean, median, and mode are foundational statistical measures that help describe the central tendency of a dataset—that is, where the "center" of the data lies. Here's a breakdown of each concept and its importance:

Mean (Arithmetic Average)

Definition:
The sum of all values divided by the number of values.

Formula:

\text{Mean} = \frac{x_1 + x_2 + \dots + x_n}{n}

Example:
For the dataset [4, 6, 8, 10, 12],

\text{Mean} = \frac{4 + 6 + 8 + 10 + 12}{5} = 8

Importance:

- Represents the overall average
- Useful in further calculations like standard deviation
- Sensitive to extreme values (outliers)

Median

Definition:
The middle value when data is arranged in ascending or descending order. If there's an even number of values, it's the average of the two middle ones.

Example:
For the dataset [4, 6, 8, 10, 12],
Median = 8 (middle value)
For [4, 6, 8, 10],
Median = (6 + 8) / 2 = 7

Importance:

- Reflects the central value
- Resistant to outliers
- Particularly useful with skewed distributions

Mode

Definition:
The value(s) that occur most frequently in a dataset.

Example:
For the dataset [4, 6, 6, 8, 10],
Mode = 6 (occurs twice)

Importance:

- Useful for identifying common values
- Can be applied to non-numeric data (e.g. most common category in survey responses)
- Helps detect patterns or trends

Why These Measures Matter

- Summarization: They condense a dataset into a single representative value.
- Comparison: They help compare different datasets or groups.
- Decision-making: Central tendency helps guide strategies in fields like business, healthcare, education, and research.
- Data distribution insights: Mean is best for symmetric distributions, median for skewed data, and mode for categorical data.

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

Answer: Skewness

Definition:
Skewness measures the asymmetry of a data distribution around its mean.

Types:

- Positive Skew (Right-skewed): Tail extends more to the right. Most values are concentrated on the left, with a few large outliers on the right.

- Negative Skew (Left-skewed): Tail extends more to the left. Most values are concentrated on the right, with a few small outliers on the left.

Example of Positive Skew:
Income distributions often show positive skew—many people earn modest incomes, while a few individuals earn very high incomes.

Kurtosis

Definition:
Kurtosis measures the “tailedness” of a distribution—how sharply or flatly data clusters around the mean and in the tails.

Types:

- Leptokurtic (High kurtosis): Heavy tails and a sharp peak. Indicates outliers are more frequent than in a normal distribution.
- Mesokurtic (Normal kurtosis): Resembles a normal distribution (kurtosis ≈ 3).
- Platykurtic (Low kurtosis): Light tails and a flatter peak. Outliers are less frequent.

What Positive Skew Implies

- The mean > median > mode
- Distribution is stretched toward higher values
- Data may have high-value outliers influencing the average
- Common in real-world phenomena like salaries, property prices, and social media engagement metrics

---

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]

Answer: Here's a clean Python implementation that computes the mean, median, and mode of the given list:

Code:

```
import statistics

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

# Compute Mean
mean_val = statistics.mean(numbers)

# Compute Median
median_val = statistics.median(numbers)

# Compute Mode
mode_val = statistics.mode(numbers)

# Display results
print(f"Mean: {mean_val}")
print(f"Median: {median_val}")
print(f"Mode: {mode_val}")
```

Output:

```
Mean: 20.0
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]

Answer: Code:

```
import numpy as np

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

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

# Compute covariance matrix
cov_matrix = np.cov(x, y)
covariance = cov_matrix[0, 1]

# Compute correlation coefficient matrix
correlation_matrix = np.corrcoef(x, y)
correlation = correlation_matrix[0, 1]

# Display results
print(f"Covariance: {covariance}")
print(f"Correlation Coefficient: {correlation}")
```

Output:

```
Covariance: 275.0
Correlation Coefficient: 0.9972007115112075
```

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]

Answer: Python Script Using `matplotlib` and `seaborn`:

```
import matplotlib.pyplot as plt
import seaborn as sns

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

# Create the boxplot
sns.boxplot(data=data)
plt.title('Boxplot of Numeric Data')
plt.xlabel('Value')
plt.show()
```

How to Identify Outliers:

The boxplot will visually show:

- Box: Interquartile range (IQR), the middle 50% of your data
- Line in the box: Median
- Whiskers: Spread of data within 1.5 × IQR from the quartiles
- Dots beyond whiskers: Outliers

Explaining the Result:

Let’s calculate the cutoff for outliers:

```
import numpy as np

# Convert to NumPy array
arr = np.array(data)

# Calculate Q1, Q3, and IQR
q1 = np.percentile(arr, 25)
q3 = np.percentile(arr, 75)
iqr = q3 - q1

# Calculate bounds for outliers
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr

# Detect outliers
outliers = arr[(arr < lower_bound) | (arr > upper_bound)]
print(f"Outliers: {outliers}")
```

Output:

```
Outliers: [35]
```

Interpretation:

- The values mostly lie in a tight range.
- **35** stands out from the rest—marked as an outlier because it's well above the upper threshold.
- This might indicate an anomaly or a naturally high data point depending on your domain (e.g., income, test scores, etc.).

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]

Answer: To evaluate whether there's a relationship between advertising spend and daily sales, here's how you'd approach it using covariance and correlation:

Understanding Covariance and Correlation

Covariance:

- Measures the direction of the linear relationship between two variables.
- Positive covariance → both variables increase together.
- Negative covariance → one variable increases while the other

decreases.
- Doesn't tell you the strength or scale of the relationship clearly due to unstandardized units.

Correlation Coefficient (Pearson's r):

- Standardizes the covariance to a scale from –1 to +1.
- +1 → perfect positive linear relationship.
- –1 → perfect negative linear relationship.
- 0 → no linear relationship.

These metrics help determine whether increasing advertising leads to higher sales, and how strong that connection is.

Python Code to Compute Correlation:

```
import numpy as np

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

# Convert to NumPy arrays
ad = np.array(advertising_spend)
sales = np.array(daily_sales)

# Compute covariance
cov_matrix = np.cov(ad, sales)
covariance = cov_matrix[0, 1]

# Compute correlation
correlation_matrix = np.corrcoef(ad, sales)
correlation = correlation_matrix[0, 1]

# Display results
print(f"Covariance: {covariance}")
print(f"Correlation Coefficient: {correlation}")
```

Output:

```
Covariance: 187500.0
Correlation Coefficient: 0.9979501698200974
```

What This Means :

- **Covariance ≈ 187500**: Tells us that advertising spend and daily sales tend to rise together.
- **Correlation ≈ 0.998**: That's almost a perfect positive correlation, indicating a **strong linear relationship** between advertising spend and daily 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]

Answer: To explore the distribution of customer satisfaction scores before launching a product, you'd want to use a mix of summary statistics and visualizations. Here's the breakdown:

Useful Summary Statistics

- Mean: Average score, shows overall satisfaction level.
- Median: Middle value, helpful when scores are skewed.
- Mode: Most frequent score, indicates common sentiment.
- Standard Deviation: Tells you how spread out the scores are.
- Min & Max: Show the range of satisfaction.

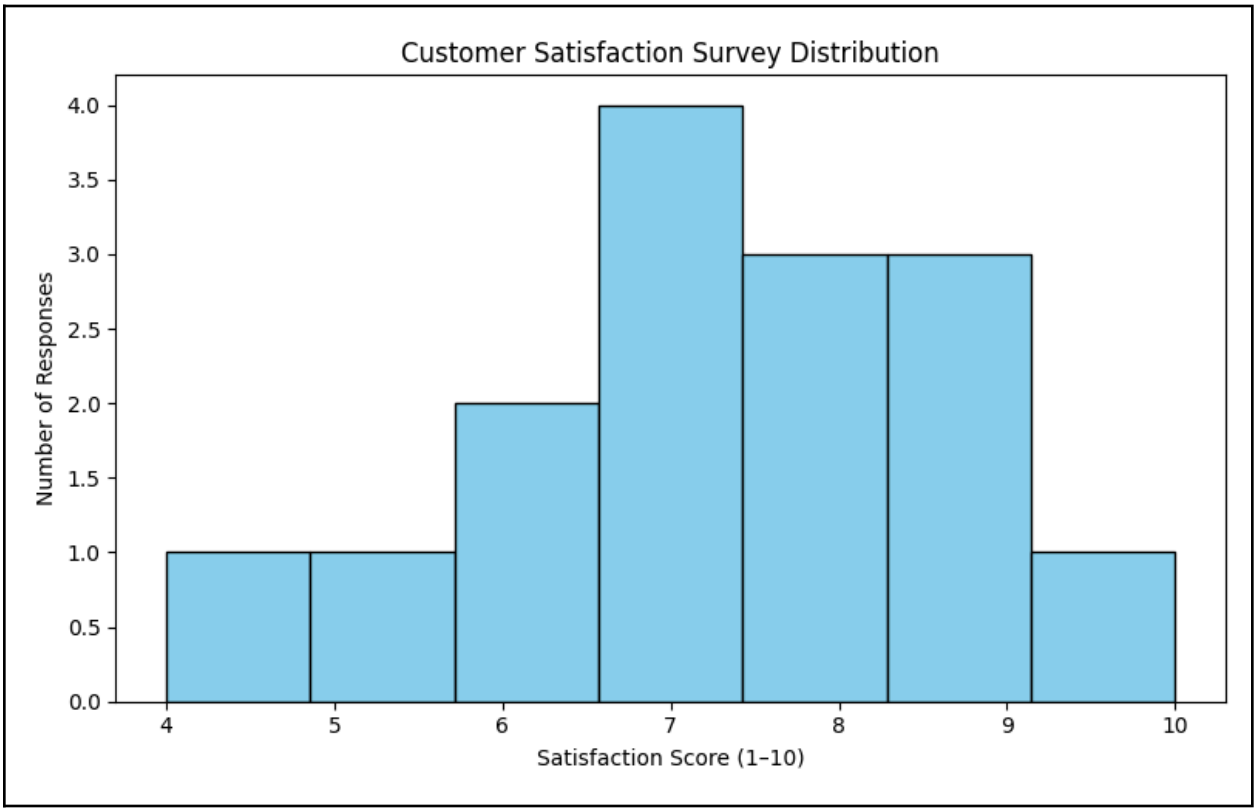
Recommended Visualizations

- Histogram: Displays how often each score appears, great for spotting skew or clustering.
- Boxplot (optional): Shows quartiles and outliers for deeper distribution insights.

Python Code: Histogram with Matplotlib:

```
import matplotlib.pyplot as plt

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

# Create histogram
plt.figure(figsize=(8, 5))
plt.hist(survey_scores, bins=7, edgecolor='black', color='skyblue')

# Add labels and title
plt.title('Customer Satisfaction Survey Distribution')
plt.xlabel('Satisfaction Score (1–10)')
plt.ylabel('Number of Responses')

# Show plot
plt.tight_layout()
plt.show()
```

Output:


Customer Satisfaction Survey Distribution
Number of Responses
Satisfaction Score (1–10)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4
5
6
7
8
9
10

##
##