Statistics Basics| Assignment

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

- Descriptive and inferential statistics are two fundamental branches of statistics, each serving distinct purposes in data analysis. Below are examples to illustrate their differences and applications.
- **Descriptive Statistics Example**
- Descriptive statistics summarize and describe the characteristics of a dataset. For instance, consider the test scores of 1,000 students in a school. Using descriptive statistics, you could calculate:
- Mean (Average): The average test score is 82.13, providing a central measure of performance.
- Median: The middle score is 84, indicating that half of the students scored above and half below this value.
- Range: The scores range from 45 to 100, showing a spread of 55 points.
- Visualization: A histogram could reveal that most students scored between 70 and 90, forming a bell-shaped distribution.
- **Inferential Statistics Example**
- Inferential statistics use a sample to make predictions or draw conclusions about a larger population. For example, suppose you want to estimate the average height of adult women in a country. Instead of measuring every woman, you could:
- Sample: Measure the heights of 1,000 randomly selected women.
- Confidence Interval: Calculate a 95% confidence interval, such as [155 cm, 163 cm], indicating you are 95% confident the true average height lies within this range.
- Hypothesis Testing: Test whether the average height is significantly different from a global average, using statistical tests like a t-test.

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

- **Sampling in statistics is the process of selecting a subset of individuals from a larger population to estimate characteristics of the whole population. Random sampling and stratified sampling are two common methods used to achieve this.**
- **Random Sampling**
- **Definition**: Random sampling is a method where each member of the population has an equal chance of being selected. This technique is often used when the population is homogeneous, meaning that individuals have similar characteristics.

- Process: Members are chosen entirely by chance, often using random number generators or lottery methods. This approach helps minimize bias and ensures that the sample is representative of the population as a whole.
- **Advantages**: It is simple to implement and provides a straightforward way to gather data without needing to consider specific subgroups within the population.
- **Stratified Sampling**
- **Definition:** Stratified sampling involves dividing the population into distinct subgroups, or strata, based on shared characteristics (e.g., age, gender, income). A random sample is then taken from each stratum in proportion to its representation in the population.
- **Process:** This method ensures that all subgroups are adequately represented in the final sample. For example, if a population consists of 60% females and 40% males, the sample will reflect this ratio.
- **Advantages:** Stratified sampling provides more accurate and reliable results, especially when the population is heterogeneous. It allows for detailed analysis of specific subgroups, which can be crucial for studies focusing on particular characteristics or behaviors.
- **Key Differences**
- **Selection Method:** Random sampling selects individuals entirely by chance, while stratified sampling divides the population into subgroups and samples from each subgroup proportionally.
- **Representation:** Random sampling may not guarantee representation of all groups within the population, whereas stratified sampling ensures that each subgroup is represented according to its size in the population.
- **Complexity:** Random sampling is generally simpler to execute, while stratified sampling can be more complex due to the need to identify and categorize the strata before sampling.

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

- The measures of central tendency are important because they provide a summary of the data and help in understanding the data's distribution. The mean, median, and mode each have their own advantages and are used in different contexts. The mean is sensitive to extreme values, while the median is not. The mode is useful for categorical data, but it

may not exist for all types of data. Understanding these measures allows for better decision-making and analysis in various fields, including statistics, research, and business.

- **Mean:** The mean, or arithmetic average, is calculated by summing all the values in a dataset and dividing by the total number of values.
- **Median:** The median is the middle value in a dataset when the numbers are arranged in ascending order. If there is an even number of values, the median is the average of the two middle numbers.
- **Mode:** The mode is the value that occurs most frequently in a dataset. A dataset may have one mode, more than one mode, or no mode at all.

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

- Skewness measures the asymmetry of a data distribution, while kurtosis measures the "tailedness" or peak of the distribution.
- A positive skew indicates that the data has a longer right tail, suggesting that most values are concentrated on the left with a few high values pulling the mean to the right.
- **Skewness**
- Definition: Skewness quantifies the degree of asymmetry of a distribution around its mean. It indicates whether the data points are more concentrated on one side of the mean than the other.
- Types of Skewness:

1. **Positive Skew (Right Skew)**: The right tail of the distribution is longer or fatter than the left.
- This means that most data points are clustered on the left, with a few larger values pulling the mean to the right. In this case, the relationship is typically Mean > Median > Mode. Examples include income distribution and exam scores where a few high scores can significantly affect the average.
2. **Negative Skew (Left Skew):** The left tail is longer, indicating that most data points are on the right, with a few smaller values pulling the mean to the left. Here, the relationship is Mean < Median < Mode. An example could be test scores on an easy exam where most students score high, but a few score very low.

- **Kurtosis**

- Definition: Kurtosis measures the "tailedness" of a distribution, indicating how much of the data is in the tails versus the center.
- It provides insights into the likelihood of extreme values (outliers) in the dataset.
- Types of Kurtosis:

1. Leptokurtic: High kurtosis (>3), indicating a sharp peak and heavy tails, suggesting a higher likelihood of outliers.
2. Mesokurtic: Normal kurtosis (≈3), similar to a normal distribution, indicating moderate tails and a moderate peak.

- Platykurtic: Low kurtosis (<3), indicating a flat peak and light tails, suggesting fewer outliers.

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]

Code:

```
import numpy as np

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

np.mean(numbers)
```

- Mean: Average of all numbers.
- Mean value is : `np.float64(19.6)`

  `np.median(numbers)`

- Median: Middle value when sorted (or average of two middle values if even count).
- Median value is : `np.float64(19.0)`

  `Import statistics`

```
statistics.mode(numbers)
```

- Mode: Most frequently occurring value.
    - If there's no unique mode, it returns `"No unique mode"`.
    - Mode value is: 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]

Code:

```
import numpy as np

# Given data

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)

# Covariance matrix

cov_matrix = np.cov(x, y, ddof=1)

covariance = cov_matrix[0, 1]


# Correlation coefficient

correlation = np.corrcoef(x, y)[0, 1]
```

```
print("Covariance:", covariance)

print("Correlation Coefficient:", correlation)
```

Output

- Covariance: 250.0
- Correlation Coefficient: ~0.997

🔍 Interpretation

- Covariance of 250.0 indicates a strong positive relationship: as x increases, y tends to increase.
- Correlation coefficient close to 1 (≈ 0.997) confirms a very strong linear relationship.

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]

Code:

```
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]

# Create boxplot

plt.boxplot(data, vert=False, patch_artist=True, boxprops=dict(facecolor='lightblue'))

plt.title('Boxplot of Given Data')

plt.xlabel('Value')

plt.grid(True)
```

```
plt.show()

# Identify outliers using IQR method

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("Outliers:", outliers)
```

Explanation of the Output

- Q1 (25th percentile): 18.25
- Q3 (75th percentile): 24.25
- IQR (Q3 - Q1): 6.0
- Lower Bound: 18.25 - 1.5×6 = 9.25
- Upper Bound: 24.25 + 1.5×6 = 33.25

Outliers:

- The only value outside the upper bound is 35, so it's identified as an outlier.

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]

How Covariance and Correlation Help

- **Covariance** measures how two variables change together:
    - A **positive covariance** means that as advertising spend increases, daily sales tend to increase.
    - A **negative covariance** means that as advertising spend increases, daily sales tend to decrease.
    - However, covariance is **scale-dependent**, so it's hard to interpret the strength of the relationship.
- **Correlation coefficient** (Pearson's *r*) standardizes this relationship:
- Ranges from **-1 to +1**
- **+1** = perfect positive linear relationship
- **0** = no linear relationship
- **-1** = perfect negative linear relationship
- It's **scale-independent**, making it easier to interpret.

Code:

```
import numpy as np

# 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)

# Covariance

cov_matrix = np.cov(ad, sales, ddof=1)

covariance = cov_matrix[0, 1]
```

```
# Correlation coefficient

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

print("Covariance:", covariance)

print("Correlation Coefficient:", correlation)
```

Interpretation of Results

- **Covariance:** Large positive value (e.g., 225000) → indicates that both variables increase together.
- **Correlation Coefficient:** Close to **+1** (e.g., ~0.997) → strong positive linear relationship.

This suggests that **higher advertising spend is strongly associated with higher daily sales**, which is valuable insight for the marketing team.

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]

Summary Statistics to Use

- **Mean (Average):** Indicates the central tendency of satisfaction.
- **Median:** Shows the middle score, useful if the data is skewed.
- **Mode:** Most frequent score, helpful for identifying common sentiment.
- **Standard Deviation:** Measures how spread out the scores are.
- **Minimum and Maximum:** Show the range of satisfaction.

Visualizations to Use

- **Histogram:** Displays the frequency of scores across bins (e.g., scores 1–10).

- **Boxplot (optional):** Highlights median, quartiles, and outliers.
- **Density Plot (optional):** Smooth curve to show distribution shape.

Code:

```
import matplotlib.pyplot as plt

import numpy as np

# Survey data

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

# Summary statistics

mean_score = np.mean(survey_scores)

std_dev = np.std(survey_scores)

min_score = np.min(survey_scores)

max_score = np.max(survey_scores)

print(f"Mean: {mean_score:.2f}")

print(f"Standard Deviation: {std_dev:.2f}")

print(f"Min: {min_score}, Max: {max_score}")

# Histogram

plt.hist(survey_scores, bins=7, color='skyblue', edgecolor='black')

plt.title('Customer Satisfaction Survey Distribution')

plt.xlabel('Satisfaction Score')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()
```

Output:

- If the histogram is centered around 7–9, it suggests generally high satisfaction. A low standard deviation means most customers gave similar scores.
- If scores are skewed (e.g., more 4s or 10s), it may indicate polarized opinions.