Name: Sharandeep Meharwal

Assignment : Statistics Basics

Email : sharanmeharwal@gmail.com

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

Answer:

Descriptive statistics involve methods for summarizing and organizing data, such as calculating mean, median, mode, standard deviation, and visualizing data with graphs like histograms and boxplots.

For example, finding the average height of students in a classroom is a descriptive statistic.

Inferential statistics, on the other hand, use sample data to make generalizations or predictions about a population, often involving hypothesis testing, confidence intervals, and regression analysis.

For example, estimating the average height of all students in a school by analyzing data from one classroom is 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 (sample) from a larger population to estimate characteristics of the whole population.

Random sampling means each member of the population has an equal chance of being selected, ensuring representation and minimizing bias.

Stratified sampling divides the population into groups (strata) based on specific characteristics, and samples are then drawn proportionally or equally from each strata. This improves representation of subgroups and can increase accuracy versus simple random sampling.

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

Answer:
Mean is the arithmetic average of a data set, calculated by dividing the sum of all values by the number of values. Median is the middle value when data are ordered from least to greatest; if there's an even number of values, it's the average of the two middle values. Mode is the value that appears most frequently in a dataset.

These measures are important because they summarize a large dataset using single values that describe the central position, helping to understand general trends and compare datasets effectively.

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

Answer:
Skewness measures asymmetry in the distribution of data. Positive skew (right-skewed) indicates that the data has a longer right tail; most values are concentrated on the left but extreme high values stretch out the distribution to the right.

Kurtosis measures the 'tailedness' of the distribution. High kurtosis suggests more outliers. Positive skew implies most data points are below the mean, with a few higher values pulling the mean to the right.

Question 5

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:

```python
import statistics

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
mean_val = statistics.mean(numbers)
median_val = statistics.median(numbers)
mode_val = statistics.mode(numbers)
print("Mean:", mean_val)
print("Median:", median_val)
print("Mode:", mode_val)
```

```
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.

listx: 10, 20, 30, 40, 50

listy: 15, 25, 35, 45, 60

Answer:

```python
import numpy as np

list_x = np.array([10, 20, 30, 40, 50])
list_y = np.array([15, 25, 35, 45, 60])

covariance = np.cov(list_x, list_y)[0][1]
correlation = np.corrcoef(list_x, list_y)[0][1]

print("Covariance:", covariance)
print("Correlation coefficient:", correlation)
```

```
Covariance: 275.0
Correlation coefficient: 0.995893206467704
```

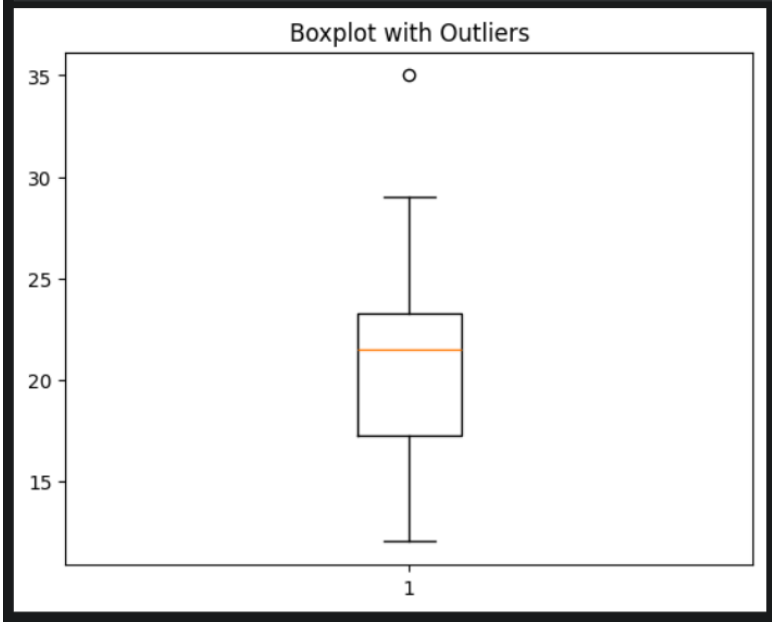
Question 7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result.

data: 12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35

Answer:

```python
import matplotlib.pyplot as plt

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
plt.boxplot(data)
plt.title("Boxplot with Outliers")
plt.show()
```



- The box spans from 17.25 to 23.25, covering the middle 50% of values.
- The orange line inside the box marks the median = 22, showing the central tendency.
- The whiskers extend to the minimum (12) and maximum within the allowed range (29).
- The value 35 lies beyond the upper bound (32.25) and is plotted as a circle (outlier).

Question 8

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:

advertisingspend = 200, 250, 300, 400, 500

dailysales = 2200, 2450, 2750, 3200, 4000

Answer:

Covariance measures how two variables change together—if they increase or decrease in the

same direction, the covariance is positive. Correlation, on the other hand, standardizes this measurement and quantifies the strength and direction of the linear relationship between two variables (between -1 and 1). To explore the relationship between advertising spend and daily sales, first calculate covariance, then calculate correlation. A strong positive correlation close to 1 suggests that an increase in advertising spend is associated with an increase in daily sales.

Python code to compute correlation:

```
]: import numpy as np

advertisingspend = [200, 250, 300, 400, 500]
dailysales = [2200, 2450, 2750, 3200, 4000]

correlation = np.corrcoef(advertisingspend, dailysales)[0][1]
print("Correlation coefficient:", correlation)

Correlation coefficient: 0.9935824101653329
```

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:
surveyscores = 7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7

Answer:

To understand the distribution, use summary statistics such as mean (average score), median (middle score), mode (most frequent score), and standard deviation (spread of the scores). These help identify typical satisfaction, variability, and the most common opinions.

A histogram visualizes how frequently each score occurs and whether the distribution is skewed or symmetric, pinpointing trends or potential outliers

```
0]: import matplotlib.pyplot as plt

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

    plt.hist(surveyscores, bins=range(4,12), edgecolor='black')
    plt.title('Survey Score Distribution')
    plt.xlabel('Score')
    plt.ylabel('Frequency')
    plt.show()
```