Statistics Assignment Question and Answer

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

.

Answer Descriptive statistics - Descriptive statistics are used to **summarize and describe** the main features of a dataset

Inferential statistics - Inferential statistics are used to **make predictions, inferences, or generalizations** about a larger population based on a sample of data.

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

Answer Sampling : Sampling is the process of **selecting a small group (sample)** from a **larger population** in order to **study or analyze** it and make conclusions about the whole population.

Random sampling :In random sampling, **every member of the population has an equal chance** of being selected.

Stratified : In stratified sampling, the population is **divided into groups (called strata)** based on a certain characteristic (like gender, age, income, etc.), and then **random samples** are taken **from each group**

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

Answer Mean The mean is the sum of all observations divided by the number of observations.

Mean :The median is the middle value when the data is arranged in ascending or descending order.

If there’s an even number of observations, the median is the average of the two middle values.

Mode : The mode is the value that appears most frequently in the dataset.

Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?
Answer Skewness is a statistical measure that tells us how symmetric or asymmetric a data distribution is around its mean.

Kurtosis : Skewness is a statistical measure that tells us how symmetric or asymmetric a data distribution is around its mean.

Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers.

Answer
```
import statistics as stats

numbers = [10, 20, 30, 40, 50, 50, 60]

mean_value = stats.mean(numbers)
median_value = stats.median(numbers)
mode_value = stats.mode(numbers)

print("Given Numbers:", numbers)
print("Mean:", mean_value)
print("Median:", median_value)
print("Mode:", mode_value)
```

```
Given Numbers: [10, 20, 30, 40, 50, 50, 60]
Mean: 37.142857142857146
Median: 40
Mode: 50
```

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
```
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)

# Calculate covariance matrix
cov_matrix = np.cov(x, y, bias=False)

# Covariance between x and y
cov_xy = cov_matrix[0, 1]

# Calculate correlation coefficient
corr_xy = np.corrcoef(x, y)[0, 1]

# Display results
print("List X:", list_x)
print("List Y:", list_y)
print("Covariance:", cov_xy)
print("Correlation Coefficient:", corr_xy)
```

```
List X: [10, 20, 30, 40, 50]
List Y: [15, 25, 35, 45, 60]
Covariance: 237.5
Correlation Coefficient: 0.9933992677987826
```

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

Answer import matplotlib.pyplot as plt
import seaborn as sns

```
# Given numeric list
data = [10, 12, 15, 18, 19, 21, 22, 23, 24, 30, 45, 55, 60, 100]

# Create a boxplot
plt.figure(figsize=(8, 4))
sns.boxplot(data=data, color='skyblue')
```

```
# Add title and labels
plt.title("Boxplot of Given Data", fontsize=14)
plt.xlabel("Data Values")

# Display the plot
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]
```

Answer covariance :Covariance tells us whether two variables move together
Corelation : Correlation is a standardized form of covariance, always between –1 and +1.

It measures the strength and direction of the linear relationship.

```
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)

# Compute Covariance Matrix
cov_matrix = np.cov(x, y, bias=False)
cov_xy = cov_matrix[0, 1]

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

# Display results
print("Advertising Spend:", advertising_spend)
```

```
print("Daily Sales:", daily_sales)
print("Covariance:", cov_xy)
print("Correlation Coefficient:", corr_xy)
```

```
Advertising Spend: [200, 250, 300, 400, 500]
Daily Sales: [2200, 2450, 2750, 3200, 4000]
Covariance: 85500.0
Correlation Coefficient: 0.9957880632554369
```

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 If mean ≈ median, distribution is fairly symmetric.
If standard deviation is low, customers mostly agree; if high, opinions are very mixed.

1. Histogram

- Shows the frequency distribution of scores.
- Helps visualize whether ratings are skewed, uniform, or bell-shaped.

2. Boxplot (optional)

- Useful to detect outliers and see spread at a glance.

```
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_val = stats.mean(survey_scores)
median_val = stats.median(survey_scores)
mode_val = stats.mode(survey_scores)
std_dev = np.std(survey_scores)
data_range = max(survey_scores) - min(survey_scores)

print("Mean:", mean_val)
print("Median:", median_val)
print("Mode:", mode_val)
print("Standard Deviation:", std_dev)
print("Range:", data_range)

# --- Visualization: Histogram ---
plt.figure(figsize=(8, 4))
plt.hist(survey_scores, bins=6, color='skyblue', edgecolor='black')
plt.title("Customer Satisfaction Score Distribution", fontsize=14)
plt.xlabel("Satisfaction Score (1-10)")
plt.ylabel("Frequency")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

```
Mean: 7.4
Median: 7
Mode: 7
Standard Deviation: 1.62
Range: 6
```