Statistics Basic Assignment

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

Answer - Statistics plays a vital role in analyzing and interpreting data in various fields such as business, economics, science, healthcare, and social research. It is mainly divided into two broad categories: **Descriptive Statistics** and **Inferential Statistics**. Both these branches serve different purposes but are equally important in the overall process of statistical analysis.

- **Descriptive Statistics** refers to the methods used for summarizing and organizing data in a meaningful way. It deals with presenting raw data in a simplified and interpretable form. This includes calculating measures like mean, median, mode (measures of central tendency), and range, standard deviation, or variance (measures of dispersion). Descriptive statistics also involve the use of graphs, charts, and tables to visually represent the data. The key point about descriptive statistics is that it is limited to the dataset at hand and does not go beyond that data. For example, if a teacher records the scores of 50 students in a test and finds the average (mean) to be 72, the highest score to be 95, and the standard deviation to be 8, then these values describe the performance of **only those 50 students**. No generalizations are made beyond this group.
- **Inferential Statistics** involves making predictions, estimations, or generalizations about a larger population based on a sample of data. Since it is often not feasible to study an entire population due to time or cost constraints, a smaller group (sample) is studied, and conclusions are drawn about the population as a whole. Inferential statistics uses more complex methods such as hypothesis testing, confidence intervals, t-tests, ANOVA, correlation, and regression analysis. For instance, a researcher may survey 500 people in a city to estimate the average income of the entire population of that city. Based on the sample data, they can make an estimate of the average income for the whole city and also provide a confidence interval that shows the reliability of the estimate. In this way, inferential statistics allows decision-making beyond the data available.

The main difference between the two is that descriptive statistics are confined to summarizing the existing data, while inferential statistics allow one to make predictions or decisions about a broader population based on a sample. Descriptive statistics answers the question “what is happening in the data?”, whereas inferential statistics addresses the question “what could happen in the population based on this data?”

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

Answer- **Sampling** is a fundamental concept in statistics. It refers to the process of selecting a subset (called a sample) from a larger group (called a population) in order to gather information and make inferences about the whole population. Since it is often impractical or impossible to collect data from an entire population due to time, cost, or accessibility, sampling provides an efficient and reliable way to study and analyze data.

A **sample** should ideally be representative of the entire population so that the conclusions drawn from it are valid. There are various methods of sampling, among which **random sampling** and **stratified sampling** are commonly used.

- **Random sampling** is a sampling method in which every individual or item in the population has an equal chance of being selected. It is considered one of the most basic and unbiased sampling methods. The selection is purely based on chance, and no preferences or groupings are considered during selection.

  Example:
  Suppose a school has 1,000 students, and the principal wants to survey 100 students about the quality of cafeteria food. Using random sampling, the principal can assign numbers to all students and use a random number generator to select 100 students. Here, each student had an equal chance of being chosen.

  Advantages:

- Minimizes bias
- Simple to understand and implement
- Suitable when the population is homogeneous

  Disadvantages:

- May not be representative if the sample size is small
- Could accidentally overrepresent or underrepresent certain groups

- **Stratified sampling** is a method where the population is divided into distinct subgroups or strata based on certain characteristics (e.g., age, gender, income level), and then a random sample is taken from each subgroup. This ensures that all key subgroups are fairly represented in the final sample.

  Example:
  Using the same school example, suppose the school has 600 boys and 400 girls. If

the principal wants to ensure both genders are proportionally represented in the sample of 100 students, he can divide students into two strata — boys and girls — and randomly select 60 boys and 40 girls. This way, the sample maintains the gender ratio of the population.

Advantages:

- Ensures representation of all key subgroups
- More accurate and reliable results, especially when population is diverse
- Reduces sampling error

Disadvantages:

- Requires detailed knowledge of population characteristics
- Slightly more complex to design and implement

Q3. Define mean, median, and mode. Explain why these measures of central tendency are important.

Answer: In statistics, measures of central tendency are numerical values that describe the center point or average of a dataset. The three most commonly used measures of central tendency are mean, median, and mode. These values help us understand and summarize a large set of data with a single representative value.

1. Mean The mean (also called the average) is the sum of all values in a dataset divided by the total number of values.

Formula:
Mean = (Sum of all values) / (Number of values)

Example:
If the marks obtained by five students are 60, 70, 75, 80, and 85, then:
Mean = (60 + 70 + 75 + 80 + 85) / 5 = 370 / 5 = 74

2. Median The median is the middle value in a dataset when the numbers are arranged in ascending or descending order. If the number of values is even, the median is the average of the two middle values.

Example:
For the values 10, 20, 30, 40, and 50, the median is 30 (the middle value).
For an even number of values like 10, 20, 30, 40, the median = (20 + 30)/2 = 25

3. 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.

Example:
For the values 5, 7, 7, 9, 10, the mode is 7 (as it appears twice).

Importance of Mean, Median, and Mode

These three measures are essential because they give us insights into the distribution and central value of data:

- Mean is useful when the data is evenly distributed and there are no extreme values (outliers). It is commonly used in economics, education, and general analysis.

- Median is more reliable when the data has outliers or is skewed, as it is not affected by extreme values. For example, in income data, where a few people may earn significantly more, the median gives a more accurate picture of the typical income.

- Mode helps identify the most common or popular item in a dataset, which is useful in market research, voting analysis, and inventory management..

Q4. Explain skewness and kurtosis. What does a positive skew imply about the data?

Answer: In statistics, skewness and kurtosis are two important concepts used to describe the shape and distribution of a dataset. While other measures like mean and standard deviation tell us about the center and spread of the data, skewness and kurtosis help us understand its symmetry and peakedness.

Skewness refers to the asymmetry of the data distribution around its mean. If a distribution is perfectly symmetrical, the skewness is zero. A positive skew, also known as right skew, means that the tail on the right side of the distribution is longer or fatter than the left side. In such cases, most of the data values are concentrated on the lower end, and a few very high values stretch the tail to the right. When data is positively skewed, the mean is greater than the median, and the median is greater than the mode.

For example, income distribution is often positively skewed because most people earn within a certain range, but a small number of individuals earn extremely high incomes, which raises the average.

Kurtosis, on the other hand, measures the peakedness or flatness of a distribution compared to a normal distribution. A normal distribution has a kurtosis value of approximately 3 and is referred to as mesokurtic. If a distribution has a higher peak and fatter tails than normal, it is called leptokurtic (kurtosis > 3), indicating more extreme values or outliers. If it has a flatter peak and thinner tails, it is called platykurtic (kurtosis < 3), indicating fewer outliers and less variation in extreme values.

In summary, skewness helps identify the direction of the distribution's tail, and a positive skew means the data is spread out more on the right side due to high outlier values. Kurtosis indicates how sharp or flat the peak is and whether the data has heavy or light tails. Both are essential in understanding the overall shape and behavior of a dataset.

Q5. 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 import statistics

```
# Given list of numbers

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

# Calculate Mean

mean_value = statistics.mean(numbers)

# Calculate Median

median_value = statistics.median(numbers)

# Calculate Mode

mode_value = statistics.mode(numbers)

# Display the results

print("Mean:", mean_value)

print("Median:", median_value)

print("Mode:", mode_value)
```

Output:

```
Mean: 19.6

Median: 19

Mode: 12
```

Q6. 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

```
# Given lists

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)

# Calculate covariance matrix

cov_matrix = np.cov(x, y, bias=False)  # bias=False uses sample formula (N-1)

# Extract covariance value

covariance = cov_matrix[0][1]

# Calculate correlation coefficient

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

# Display results

print("Covariance:", covariance)

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

Output

```
Covariance: 275.0

Correlation Coefficient: 0.9946917938265513
```

Q7. 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

```
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)

plt.title("Boxplot of Data")

plt.xlabel("Values")
plt.grid(True)

plt.show()

# Identify outliers manually using IQR

Q1 = np.percentile(data, 25)

Q3 = np.percentile(data, 75)

IQR = Q3 - Q1

# Outlier thresholds

lower_bound = Q1 - 1.5 * IQR

upper_bound = Q3 + 1.5 * IQR

# Identify outliers

outliers = [x for x in data if x < lower_bound or x > upper_bound]

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)
```

Q8. 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 **1. Explanation: Using Covariance and Correlation**

To explore the relationship between advertising spend and daily sales, we can use **covariance** and **correlation** as statistical tools.

- **Covariance** tells us the **direction** of the relationship between two variables:
    - A **positive covariance** means that as advertising spend increases, daily sales also tend to increase.

    - A **negative covariance** means that as advertising spend increases, daily sales tend to decrease.
    - However, covariance **does not show the strength or scale** of the relationship.
- **Correlation coefficient** (specifically **Pearson correlation**) measures both the **direction and strength** of the relationship:
    - It ranges from **-1 to 1**.
    - A value close to **1** indicates a strong positive relationship.
    - A value close to **0** indicates no relationship.
    - A value close to **-1** indicates a strong negative relationship.

In our case, we would calculate both to see if more advertising spend is truly linked to higher daily sales, and **how strong** that relationship is.

```
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

cov_matrix = np.cov(x, y, bias=False)

covariance = cov_matrix[0][1]


# Compute correlation coefficient
```

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


# Output

print("Covariance:", covariance)

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

Output

```
Covariance: 87500.0

Correlation Coefficient: 0.9983003825919913
```

Interpretation:

- The **positive covariance** indicates that advertising spend and sales increase together.
- The **correlation coefficient (~0.998)** is very close to **1**, showing a **very strong positive linear relationship**.
  This suggests that **increasing advertising spend is strongly associated with higher daily sales**, and the marketing team can use this insight to plan budget allocations.

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:

Answer **1. Explanation: Summary Statistics and Visualizations to Use**

To understand the distribution of customer satisfaction scores, we would use a combination of **summary statistics** and **visualizations**.

Key Summary Statistics:

- **Mean**: Shows the average satisfaction level.
- **Median**: Tells us the middle score, useful if the data is skewed.
- **Mode**: Shows the most frequently selected score.
- **Standard Deviation**: Tells us how spread out the scores are (variability).

- **Minimum and Maximum**: Show the range of satisfaction levels.

Recommended Visualizations:

- **Histogram**: To visualize the frequency distribution of scores across the 1–10 scale. It helps identify the shape (e.g., skewed, uniform, bell curve).
- **Boxplot**: To view median, quartiles, and detect any outliers.
- **Bar chart of frequencies**: Can be used if the data is discrete and you want to see exact counts per score.

These tools together provide both numerical and visual understanding of how customers feel, which is critical before launching a new product.

```
import matplotlib.pyplot as plt


# Example customer satisfaction data (scale 1 to 10)

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


# Create histogram

plt.hist(survey_data, bins=10, edgecolor='black', range=(1, 10))

plt.title("Customer Satisfaction Survey Distribution")

plt.xlabel("Satisfaction Score (1–10)")

plt.ylabel("Number of Customers")

plt.grid(True)

plt.show()
```