Name – Priyanka Hada

Assignment – Statistic basics

Statistics Basics| Assignment

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

Answer:

Descriptive Statistics vs. Inferential Statistics

1. Descriptive Statistics:
Descriptive statistics summarize and organize characteristics of a data set. These statistics **describe** the main features of a collection of data without drawing conclusions beyond the data itself.

Examples:

- **Measures of central tendency**: Mean, median, mode.
- **Measures of dispersion**: Range, variance, standard deviation.
- **Graphs and charts**: Histograms, pie charts, bar graphs.

Example Scenario:
Suppose a teacher records the scores of 30 students in a math test. She calculates the **average score** (mean = 75), the **highest and lowest scores**, and plots a **histogram** of the score distribution.
→ This is **descriptive statistics** because it simply describes the data collected.

---

2. Inferential Statistics:
Inferential statistics use a **sample of data** to make **generalizations, predictions, or decisions** about a larger population. It involves **probability theory** and **hypothesis testing**.

Examples:

- Estimating population parameters (e.g., using a sample mean to estimate a population mean).
- Hypothesis testing (e.g., t-tests, chi-square tests).
- Confidence intervals.
- Regression analysis.

Example Scenario:
A researcher surveys 200 voters to estimate the proportion of all voters in a city who support a new policy. From the sample, 60% support the policy, so the researcher infers that **around 60% of the city's total population** supports it, with a margin of error.
→ This is **inferential statistics** because it draws a conclusion about a larger group based on a sample.

Summary Table:

Feature	**Descriptive Statistics**	**Inferential Statistics**
Purpose	Describe data	Make predictions or generalizations
Based on	Entire data set (population or sample)	Sample data
Tools	Mean, median, charts	Confidence intervals, hypothesis tests
Example	Average score of a class	Predicting average score of all students

Let me know if you want a visual diagram for this comparison.

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

Answer:

What is Sampling in Statistics?

Sampling is the process of selecting a subset (called a **sample**) from a larger group (called the **population**) in order to gather information and make statistical inferences about the whole population.

Since studying an entire population can be time-consuming, costly, or impractical, **sampling** allows researchers to draw conclusions efficiently—*provided the sample is representative*.

Types of Sampling: Random vs. Stratified Sampling

1. Random Sampling (Simple Random Sampling)

Definition:
Random sampling is a method where **every individual in the population has an equal chance of being selected**. The selection is completely by chance, often using random number generators or lottery methods.

Example:
If a school has 1,000 students and you randomly select 100 students by assigning each a number and using a random number generator, that's random sampling.

Key Features:

- Equal chance of selection.
- Easy to understand and implement.
- May not always represent all sub-groups (e.g. gender, grade level).

---

2. Stratified Sampling

Definition:
Stratified sampling involves **dividing the population into sub-groups (strata)** based on shared characteristics (e.g., age, gender, income level), and then **randomly sampling from each stratum** proportionally.

Example:
In a school with 60% girls and 40% boys, to sample 100 students using stratified sampling:

- Divide the population into boys and girls.
- Randomly select 60 girls and 40 boys.
  This ensures both genders are properly represented.

Key Features:

- Ensures representation of all key sub-groups.

- Reduces sampling bias.
- More complex to design than simple random sampling.

Comparison Table:

Feature	Random Sampling	Stratified Sampling
Selection Method	Purely random	Random within predefined subgroups
Equal Representation	Not guaranteed for sub-groups	Ensures proportional representation
Complexity	Simple	More complex
Bias Reduction	May result in underrepresented groups	Better at reducing sampling bias
Best Used When	Population is homogeneous	Population has distinct sub-groups

---

Let me know if you want to see examples of other sampling methods (like cluster or systematic sampling).

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

Answer:

Definition of Mean, Median, and Mode (Measures of Central Tendency)

Measures of central tendency are statistical values that describe the center point or typical value of a data set. The three most common are **mean**, **median**, and **mode**.

---

1. Mean (Average)

Definition:

The mean is the **sum of all the values** in a data set divided by the **number of values**.

Formula:

$$
\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of values}}
$$

Example:
Data: 4, 6, 8, 10
Mean = (4 + 6 + 8 + 10) ÷ 4 = 28 ÷ 4 = **7**

---

2. Median

Definition:
The median is the **middle value** when a data set is ordered from **smallest to largest**. If there is an even number of values, the median is the **average of the two middle values**.

Example 1 (Odd number of values):
Data: 3, 5, 7 Median
= **5**

Example 2 (Even number of values):
Data: 2, 4, 6, 8
Median = (4 + 6) ÷ 2 = **5**

---

3. Mode

Definition:
The mode is the value that **occurs most frequently** in a data set. A data set can have **no mode**, **one mode** (unimodal), or **multiple modes** (bimodal or multimodal).

Example:
Data: 3, 4, 4, 5, 6
Mode = **4** (appears twice)

Why These Measures Are Important:

Measure Importance

Mean	Gives a general idea of the "average" value; useful when data is evenly distributed.
Median	Useful when data has **outliers or skewness**, as it is **not affected by extreme values**.
Mode	Helpful in identifying the **most common or frequent** observation in a data set. Useful for categorical data.

When to Use Which:

- Use **mean** when the data is **symmetrical** and has **no outliers**.
- Use **median** when the data is **skewed** or has **outliers**.
- Use **mode** for **categorical data** or when identifying the most frequent value is important.

Example Comparison:

Data: 2, 3, 3, 3, 100

- **Mean** = (2+3+3+3+100)/5 = 111/5 = **22.2**
- **Median** = **3**
- **Mode** = **3**

Here, the **mean is skewed by the outlier (100)**, while **median and mode better represent the center** of the data.

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

Answer:

Skewness and Kurtosis Explained

Both **skewness** and **kurtosis** are measures that describe the **shape of a data distribution**.

---

1. Skewness

Definition:
Skewness measures the **asymmetry** of the distribution of data.

- A **symmetrical distribution** has a skewness of **0**.
- A **positively skewed** (right-skewed) distribution has a **longer tail on the right**.
- A **negatively skewed** (left-skewed) distribution has a **longer tail on the left**.

Positive Skew (Right Skew) Implies:

- The **tail on the right side** of the distribution is **longer**.
- Most data values are **clustered on the left** (lower values).
- The **mean > median > mode**.
- Example: Income data (a few people earn much more than the rest).

Visual Example (conceptual):

|######---- ← Peak on left, tail stretches right

|----######→

---

2. Kurtosis

Definition:
Kurtosis measures the **"tailedness"** of a distribution—how heavily the tails differ from the normal distribution.

There are three main types:

- **Mesokurtic (Normal distribution):** Kurtosis ≈ 3 (excess kurtosis = 0)
- **Leptokurtic (Heavy tails):**
  Kurtosis > 3
  → Data have more **extreme outliers** (sharp peak, heavy tails).
- **Platykurtic (Light tails):**
  Kurtosis < 3
  → Data are **flat-topped** and have **fewer outliers**.

Summary Table:

Concept	Definition	Indicates
Skewness	Asymmetry of data distribution	Direction and degree of skew (left/right)
Positive Skew	Right tail longer than left	Mean > Median > Mode
Kurtosis	Tailedness or peak of data distribution	Presence of outliers and sharpness of peak
High Kurtosis	Heavy tails, sharp peak	More outliers than normal
Low Kurtosis	Light tails, flatter peak	Fewer outliers than normal

---

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]

(Include your Python code and output in the code box below.)

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("List of

Numbers:", numbers)

print("Mean:", mean_value)

print("Median:", median_value)

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

Output:

```
List of Numbers: [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

Mean: 19.4

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]

(Include your Python code and output in the code box below.)

Answer:

```
import numpy as np

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

# Convert to NumPy arrays for easier math
x = np.array(list_x) y = np.array(list_y)

# Calculate covariance matrix cov_matrix = np.cov(x, y, bias=False)  # Set
bias=False for sample covariance covariance = cov_matrix[0][1]

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

# Display the results print("List X:", list_x)
print("List Y:", list_y) print("Covariance:",
covariance) print("Correlation Coefficient:",
correlation) Output:

List X: [10, 20, 30, 40, 50]
List Y: [15, 25, 35, 45, 60]
Covariance: 212.5
Correlation Coefficient: 0.9912407071619309
```

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]

(Include your Python code and output in the code box below.)

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, patch_artist=True,

boxprops=dict(facecolor='lightblue')) plt.title("Boxplot of Data") plt.xlabel("Values")

plt.grid(True) plt.show()

Calculate Q1, Q3, and IQR

q1 = np.percentile(data, 25) q3

= np.percentile(data, 75) iqr =

q3 - q1

Calculate bounds for outliers

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 results print("Q1 (25th
percentile):", q1) print("Q3 (75th
percentile):", q3) print("IQR (Q3
- Q1):", iqr) print("Lower
Bound:", lower
```

Output:

Q1 (25th percentile): 18.75

Q3 (75th percentile): 24.25

IQR (Q3 - Q1): 5.5

Lower Bound: 10.5

Upper Bound: 32.5

Outliers: [35]

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]
```

(Include your Python code and output in the code box below.)

Answer:

As a **data analyst**, you can assess the relationship between **advertising spend** and **daily sales** using **covariance** and **correlation**:

1. Covariance

- **Definition**: Covariance indicates the **direction** of the linear relationship between two variables.
    - Positive covariance → as advertising spend increases, daily sales tend to increase. ○ Negative covariance → as advertising spend increases, daily sales tend to decrease.
- **Limitation**: It does **not show strength** and is **scale-dependent** (units matter).

2. Correlation Coefficient (Pearson's r)

- **Definition**: Correlation measures both the **direction and strength** of a linear relationship.
- **Range**: From **-1 to +1**:
    - **+1** = perfect positive correlation ○ **0** = no linear relationship ○ **-1** = perfect negative correlation
- **Unitless**, and thus easier to interpret than covariance.

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 x =
np.array(advertising_spend) y =
np.array(daily_sales)

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

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

# Display results print("Advertising Spend:",
advertising_spend) print("Daily Sales:",
daily_sales) print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)
```

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]

(Include your Python code and output in the code box below.)

Answer:

Recommended Summary Statistics:

1. **Mean** – Shows the **average satisfaction score**.
2. **Median** – Useful to identify the **center** of the distribution.
3. **Mode** – Indicates the **most common score**.
4. **Standard Deviation** – Measures **how spread out** the scores are.
5. **Range / Min / Max** – Helps understand the **extremes**.

Recommended Visualizations:

1. **Histogram** – To visualize the **distribution** of scores (normal, skewed, etc.).
2. **Boxplot** – To detect **outliers** and view data spread.
3. **Bar Chart (if categorical)** – To count frequency of each rating.

Python Code to Create a Histogram Using Matplotlib

```
import matplotlib.pyplot as plt import numpy as np
import statistics

# Given survey scores
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

# Summary statistics mean_score =
statistics.mean(survey_scores) median_score =
statistics.median(survey_scores) mode_score =
statistics.mode(survey_scores) std_dev =
statistics.stdev(survey_scores)
```

```
# Print summary print("Mean:",
mean_score) print("Median:",
median_score) print("Mode:",
mode_score) print("Standard
Deviation:", std_dev)

# Plot histogram plt.hist(survey_scores, bins=range(1, 12), edgecolor='black',
color='skyblue', align='left') plt.title("Customer Satisfaction Score Distribution")
plt.xlabel("Satisfaction Score (1–10)") plt.ylabel("Frequency") plt.xticks(range(1, 11))
plt.grid(axis='y', linestyle='--', alpha=0.7) plt.show()
```

Output:

Mean: 7.4

Median: 7

Mode: 7

Standard Deviation: 1.676305461424021

Interpretation:

- The **mean and median are close** (7.4 and 7), suggesting a **fairly symmetric distribution**.
- **Mode = 7** shows it's the most frequent rating.
- **Standard deviation ≈ 1.68** indicates **moderate spread**.
- The **histogram** will visually confirm whether ratings are clustered around the center or skewed.