Name :- Ansh Sharma
Email :- ansh16042001@gmail.com
Assignment Name:- Statistics Basics

Q1. What is the difference between descriptive statistics and inferential statistics? Explain with examples.
Ans:- Difference Between descriptiv statistics nd inferential statistics are:-

1. Descriptive statistics:- Descriptive statistics are used to summarize, organize, and describe a dataset.

What it does:

- Describes what the data shows
- Does not make predictions or generalizations
- Works only with the given data

Common tools:

- Mean, Median, Mode
- Range, Variance, Standard Deviation
- Tables, Graphs, Charts (bar chart, histogram, pie chart)

Example:
Suppose the marks of 5 students are:

```
60, 70, 75, 80, 85
```

- Mean = (60 + 70 + 75 + 80 + 85) / 5 = 74
- Highest mark = 85
- Lowest mark = 60

2. Inferential statistics:- Inferential statistics are used to make predictions or conclusions about a population based on a sample.

What it does:

- Uses sample data
- Makes estimates, predictions, or decisions
- Includes uncertainty

Common tools:

- Hypothesis testing
- Confidence intervals
- Regression analysis
- t-test, z-test, ANOVA

Example:
From the same 5 students' marks, we estimate:

"The average marks of all students in the class are around 74."

Q2. What is sampling in statistics? Explain the differences between random and stratified sampling.
Ans:- Sampling is the process of selecting a small group (sample) from a large group (population) to study and analyze, so that we can draw conclusions about the entire population.
Example:
If a college has 5,000 students, studying all students is difficult.
Difference Between Random and stratified sampling are:-

Random Sampling	Stratified Sampling
1.Purely by chance	1. From each subgroup
2.Not divided	2.Divided into strata
3.May miss some groups	3.All groups represented
4.Less accurate	4.More accurate
5.Homogeneous population	5.Heterogeneous population

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

Ans:- Measures of Central Tendency:- Measures of central tendency are statistical values that represent the center or typical value of a dataset. The three most common measures are mean, median, and mode.

1. Mean

Definition:
The mean is the average of all observations in a dataset.

Formula:

Mean=Sum of all valuesNumber of values\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}Mean=Number of valuesSum of all values

Example:
Data: 2, 4, 6, 8, 10

Mean=52+4+6+8+10/5=6

2. Median

Definition:
The median is the middle value of a dataset when the data is arranged in ascending or descending order.

Rules:

- If the number of observations is odd, the median is the middle value.
- If the number of observations is even, the median is the average of the two middle values.

Example:
Data: 3, 5, 7, 9, 11
Median = 7

Data: 4, 6, 8, 10
Median = (6 + 8) / 2 = 7

3. Mode

Definition:
The mode is the value that occurs most frequently in a dataset.

Example:
Data: 2, 3, 3, 5, 7
Mode = 3

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

Ans:- 1. Skewness

Definition:
Skewness measures the degree of asymmetry of a distribution around its mean.

Types of Skewness:

1. Zero Skewness (Symmetrical Distribution)
- Left and right sides are mirror images
- Mean = Median = Mode
- Example: Normal distribution
2. Positive Skewness (Right-skewed distribution)
- Long tail on the right side
- Most data values are concentrated on the left
- Mean > Median > Mode

Example:
Income distribution:
Most people earn low to moderate income, but a few people earn very high income.

3. Negative Skewness (Left-skewed distribution)
- Long tail on the left side
- Most data values are concentrated on the right
- Mean < Median < Mode

Example:
Marks in an easy exam where most students score high marks.

---

2. Kurtosis

Definition:
Kurtosis measures the peakedness or flatness of a distribution compared to a normal distribution.

Types of Kurtosis:

1. Mesokurtic
- Normal distribution
- Moderate peak
2. Leptokurtic
- High, sharp peak
- Heavy tails (more extreme values)
3. Platykurtic
- Flat peak
- Light tails (fewer extreme values)

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]

Ans:-

```
from collections import Counter

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

# Mean

mean = sum(numbers) / len(numbers)

# Median

numbers_sorted = sorted(numbers)

n = len(numbers_sorted)

mid = n // 2

median = numbers_sorted[mid]   # odd number of elements

# Mode

count = Counter(numbers)

max_freq = max(count.values())

mode = [num for num, freq in count.items() if freq == max_freq]

print("Mean:", mean)

print("Median:", median)

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

OUTPUT:-

makefile

Copy code

Mean: 19.6

Median: 19

Mode: [12, 19, 24]

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]

Ans:- import math

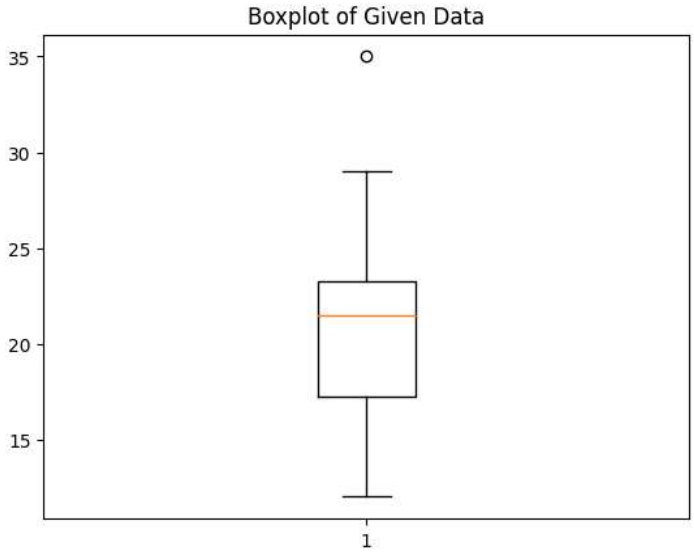
```
list_x = [10, 20, 30, 40, 50]

list_y = [15, 25, 35, 45, 60]

n = len(list_x)

# Mean

mean_x = sum(list_x) / n

mean_y = sum(list_y) / n

# Covariance (sample covariance)

covariance = sum((list_x[i] - mean_x) * (list_y[i] - mean_y) for i in range(n)) / (n - 1)

# Standard deviations

std_x = math.sqrt(sum((x - mean_x) ** 2 for x in list_x) / (n - 1))

std_y = math.sqrt(sum((y - mean_y) ** 2 for y in list_y) / (n - 1))

# Correlation coefficient

correlation = covariance / (std_x * std_y)

print("Covariance:", covariance)

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

OUTPUT:-

Covariance: 275.0

Correlation Coefficient: 0.996

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]



Ans:-

Python Boxplot and Outlier Analysis

You asked to draw a boxplot, identify outliers, and explain the result.
The boxplot for your data has been generated above.

Given Data

```
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
```

Numerical Results (from computation)

- Q1 (First Quartile) = 17.25
- Q3 (Third Quartile) = 23.25
- Interquartile Range (IQR) = 6.0

Outlier limits (1.5 × IQR rule):

- Lower bound = 8.25
- Upper bound = 32.25

Identified Outlier

```
Outlier = [35]
```

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]
```

Ans:-

```
import math

advertising_spend = [200, 250, 300, 400, 500]

daily_sales = [2200, 2450, 2750, 3200, 4000]

n = len(advertising_spend)

# Mean

mean_x = sum(advertising_spend) / n

mean_y = sum(daily_sales) / n

# Covariance

covariance = sum(

    (advertising_spend[i] - mean_x) * (daily_sales[i] - mean_y)
```

```
    for i in range(n)

) / (n - 1)

# Standard deviations

std_x = math.sqrt(sum((x - mean_x) ** 2 for x in advertising_spend) /
(n - 1))std_y = math.sqrt(sum((y - mean_y) ** 2 for y in daily_sales)
/ (n - 1))

# Correlation coefficient

correlation = covariance / (std_x * std_y)

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

Q9. 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]
```

Ans:-

```
import matplotlib.pyplot as plt



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

plt.figure()

plt.hist(survey_scores, bins=7)
```

```
plt.xlabel("Survey Score (1-10)")

plt.ylabel("Frequency")

plt.title("Histogram of Customer Satisfaction Survey Scores")

plt.show()
```