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

Ans- **Descriptive statistics** and **inferential statistics** are two fundamental branches of statistics, each serving a different purpose in analyzing data. Here's a breakdown of the differences along with examples:

Descriptive Statistics

Definition:
Descriptive statistics involve methods for organizing, summarizing, and presenting data in a meaningful way. It does **not** make predictions or generalizations beyond the data at hand.

Purpose:
To describe the **main features** of a data set.

Tools used:

- Measures of central tendency (mean, median, mode)
- Measures of dispersion (range, variance, standard deviation)
- Charts and graphs (bar charts, histograms, pie charts)

Example:
Suppose a teacher records the exam scores of 30 students in a class.

- Mean score = 72%
- Median score = 75%
- Standard deviation = 10%
- Histogram showing the distribution of scores

These statistics **describe** the performance of that particular group of 30 students.

Inferential Statistics

Definition:
Inferential statistics involve methods that use data from a **sample** to make **predictions or generalizations** about a larger **population**.

Purpose:
To draw **conclusions** and make **inferences** about a population based on sample data.

Tools used:

- Hypothesis testing
- Confidence intervals
- Regression analysis
- ANOVA (Analysis of Variance)

Example:
Suppose a researcher takes a random sample of 200 voters from a city to estimate the proportion of people who support a new policy.

- From the sample, 60% support the policy.
- The researcher uses inferential statistics to estimate that **between 56% and 64%** of **all voters** in the city support the policy (with 95% confidence).

Here, conclusions are made about the entire population **based on the sample**.

Feature	**Descriptive Statistics**	**Inferential Statistics**
Purpose	Describe data	Make predictions or generalizations
Based on	Entire population or sample	Sample only
Type of analysis	Summary and visualization	Probability-based conclusions
Examples	Mean, median, graphs	Hypothesis tests, confidence intervals

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

Ans- **Sampling** in statistics is the process of selecting a **subset of individuals or observations** from a larger **population** to estimate characteristics of the whole population.

Since it is often impractical or too expensive to study an entire population, **sampling** allows researchers to draw conclusions using data from just a portion of that population.

Random Sampling (Simple Random Sampling)

Definition:
Every individual or item in the population has an **equal chance** of being selected.

How it works:

- Assign a number to every member of the population.
- Use a random method (like a lottery or computer generator) to select the sample.

Example:
If you have a population of 1000 students, you randomly pick 100 students using a random number generator, giving each student an equal chance of being chosen.

Advantages:

- Minimizes bias
- Easy to understand and implement (for small populations)

Disadvantages:

- May not represent subgroups well, especially in diverse populations

Stratified Sampling

Definition:
The population is divided into **strata** (subgroups) based on shared characteristics (e.g., age, gender, income), and **samples are randomly selected from each stratum**.

How it works:

1. Divide the population into strata (e.g., males and females).
2. Randomly sample a proportionate number from each stratum.

Example:
In a school with 60% girls and 40% boys, to get a sample of 100 students:

- Divide the population into two strata: boys and girls.
- Randomly select 60 girls and 40 boys to maintain proportionality.

Advantages:

- Ensures representation of **all subgroups**
- More accurate and reliable if subgroups differ significantly

Disadvantages:

- More complex to design
- Requires knowledge of population structure

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

Ans- **Mean, median, and mode** are three key measures of **central tendency** in statistics. They help summarize and describe the center point of a dataset.

1. Mean (Average)

Definition:
The mean is the **sum of all values** in a dataset divided by the **number of values**.

Formula:

Mean= sum of all values/ no of values

Example:
Scores: 70, 75, 80

mean= 70+75+80/3 = 225/3=75

Use:
Best used when data is **symmetrical** and has **no extreme outliers**.

2. Median

Definition:
The median is the **middle value** in an **ordered** dataset.
If the number of observations is even, the median is the average of the two middle numbers.

Example:
Scores: 60, 70, 75, 80, 90 → Median = **75**
Scores: 60, 70, 75, 80 → Median =70+75/2=72.5

Use:
Useful when data is **skewed** or contains **outliers** (e.g., income, housing prices).

3. Mode

Definition:
The mode is the value that **occurs most frequently** in a dataset.
A dataset can have:

- No mode (if all values occur equally)
- One mode (unimodal)
- More than one mode (bimodal or multimodal)

Example:
Scores: 70, 75, 75, 80 → Mode = **75**

Use:
Best for **categorical data** or understanding the most **common** value.

Mean	Gives an overall average. Useful for further calculations like variance and standard deviation.
Median	Gives the midpoint. It's **not affected by outliers**, so it gives a better "typical" value in skewed data.
Mode	Identifies the most frequent value. Useful for **qualitative data** or finding common trends.

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

Ans- **Definition:**

Skewness measures the **asymmetry** of a data distribution.

- If data is **symmetrical**, skewness = 0 (normal distribution).
- If data is **skewed**, it means one tail is longer or fatter than the other.

Types of Skewness:

Type	Description	Tail Direction	Mean vs Median
Positive Skew (Right-skewed)	Tail is stretched to the right (higher values)	Right	Mean > Median
Negative Skew (Left-skewed)	Tail is stretched to the left (lower values)	Left	Mean < Median

What does a positive skew imply?

- **Most values are concentrated on the lower end**, but **a few high outliers** pull the mean to the right.
- The **mean is greater than the median**.

Example: Income data — most people earn average salaries, but a few very high earners increase the mean.

2. Kurtosis

➢ Definition:

Kurtosis measures the **"tailedness"** or **peakedness** of a data distribution — how heavily the tails differ from the normal distribution.

Types of Kurtosis:

Type	Description	Tail Behavior	Peak Shape
Mesokurtic	Normal kurtosis (≈ 3)	Normal tails	Normal peak
Leptokurtic	High kurtosis (> 3)	Heavy/fat tails	Sharp peak
Platykurtic	Low kurtosis (< 3)	Light/thin tails	Flat/broad peak

Why is kurtosis important?

- **High kurtosis** indicates more **outliers** than normal (risky or volatile data).
- **Low kurtosis** means fewer extreme values (more consistent data).

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]

Untitled0.ipynb

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]

statistics basics.ipynb

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]

statistics basics.ipynb

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]

Ans- **Covariance:**

- **Tells the direction** of the 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**: Covariance doesn't show the **strength** or **scale** of the relationship.

Correlation Coefficient (Pearson r):

- Measures both **direction and strength** of the linear relationship.
- Ranges from **-1 to +1**:
    - **+1**: Perfect positive correlation
    - **0**: No linear correlation
    - **–1**: Perfect negative correlation
- **Better for interpretation** than covariance because it's **scale-independent**.

statistics basics.ipynb

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:

Ans- **Summary Statistics to Use:**

Statistic	Purpose
Mean	Shows the **average satisfaction score**

Median	Reveals the **middle score** (less affected by outliers)
Mode	Indicates the **most frequent score**
Standard Deviation	Measures how **spread out** the scores are
Minimum & Maximum	Show the **range** of responses
Skewness & Kurtosis *(optional)*	Help understand **shape** of the distribution

Visualizations to Use:

Chart Type	Purpose
Histogram	Shows the **frequency distribution** (are scores clustered or spread?)
Boxplot	Visualizes **median, quartiles, and outliers**
Bar Chart	Useful if survey responses are discrete categories (e.g. Likert scale)

statistics basics.ipynb