Statistics Basic - Assignment

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

Soln.

1. Descriptive Statistics: Descriptive statistics aim to summarize and organize data so it's easy to understand. You aren't trying to look beyond the data you have in front of you; you are simply painting a clear picture of it.

Key Measures: Mean (average), Median, Mode, Range, and Standard Deviation.
Visuals: Pie charts, histograms, and bar graphs.

Example: You survey a class of 30 students and find that their average test score is **85%**.
By reporting this average and noting that scores ranged from **70% to 98%**, you are using descriptive statistics. You are only talking about those specific 30 students.

2. Inferential Statistics: Inferential statistics take data from a small group (a **sample**) and use it to make predictions or generalizations about a much larger group (the **population**). Since you can't measure everyone in the world, you use math to calculate the probability that your sample reflects reality.

Key Measures: P-values, Hypothesis tests (t-tests, ANOVA), and Confidence Intervals.
Goal: To test theories and reach conclusions that extend beyond the immediate data.

Example: You want to know the average test score of every student in the country. It's impossible to test millions of kids, so you test a random sample of **1,000 students**. If their average is **85%**, you use inferential statistics to estimate that the national average is also likely around **85%** (within a certain margin of error).

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

Soln.

Sampling in statistics is the process of selecting a representative subset (a sample) from a larger population to analyze and draw inferences about the entire group. It is a practical, cost-effective, and time-efficient alternative to studying an entire population.

Simple Random Sampling

In a simple random sample, every member of the population has an **equal chance** of being selected. It is the purest form of probability sampling, often compared to pulling names out of a hat or using a random number generator. It's easy to implement and eliminates researcher bias. You might accidentally miss out on specific subgroups. For example, if you randomly pick 10 people from a company of 100, you might accidentally pick 10 managers and zero entry-level employees just by "luck" of the draw.

Stratified sampling

Stratified sampling involves dividing the population into smaller subgroups, known as strata, based on shared characteristics (like age, gender, income, or job role). You then take a random sample from each of those subgroups. If a company is 60% women and 40% men, you split the list by gender and randomly pick 6 women and 4 men to ensure your sample perfectly mirrors the company's gender split. It ensures that minority groups are represented and usually provides more precise data than simple random sampling. It requires more prior knowledge about the population and is more time-consuming to organize.

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

Soln.

Mean

The mean is the most common measure of central tendency. You calculate it by adding up all the values in a dataset and dividing by the total number of values.Its best for data that is spread out evenly without extreme outliers (like height or weight in a general population).

Median

The median is the literal middle point of a dataset when the numbers are arranged from smallest to largest. If there is an even number of values, it is the average of the two middle numbers. It’s best for data with "outliers" (extreme values) that might skew an average like housing prices or annual salaries.

Mode

The mode is the value that appears most often in a dataset. A dataset can have one mode, multiple modes (bimodal or multimodal), or no mode at all if every number is unique. It’s best for Categorical data where you want to know the most popular choice (e.g., "What is the most common car color in the parking lot?").

Without these measures, a large list of numbers is just "noise." They are vital for three main reasons:

1. They allow you to summarize thousands of data points into a single, digestible value. It’s easier to say "The average score was 80" than to list 500 individual grades.
2. They provide a standardized way to compare different groups. For example, you can compare the median household income of two different cities to see which is more affordable.
3. Identifying Skewness: By comparing the mean and median, you can tell if your data is "lopsided." If the mean is much higher than the median, you know a few very large numbers are pulling the average up

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

Soln.

Skewness measures the asymmetry of the probability distribution. It tells you which way the "tail" of the data is pointing.

- **Zero Skew:** The data is perfectly symmetrical (like a standard normal distribution). The mean, median, and mode are all equal.
- **Positive Skew (Right-skewed):** The long tail extends toward the **higher (right)** side of the scale.
- **Negative Skew (Left-skewed):** The long tail extends toward the **lower (left)** side of the scale.

While skewness is about side-to-side balance, **kurtosis** is about the "heaviness" of the tails and the sharpness of the peak. It describes how much of the data is concentrated in the extremes versus the center

When a dataset has a positive skew, it indicates that:

- **Outliers are on the high end:** Most of the data points are clustered at lower values, but there are a few exceptionally high values pulling the "tail" to the right.
- **The Mean is greater than the Median:** Because the mean is sensitive to extreme values, those high outliers pull the average upward.
- **Real-world Example: Wealth distribution.** Most people earn a modest income (the hump on the left), while a small number of billionaires create a long tail stretching far to the right.

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]

Soln.

```
import statistics

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

mean = statistics.mean(numbers)

median = statistics.median(numbers)

mode = statistics.mode(numbers)

print(f'Mean: {mean}')
print(f'Median: {median}')

print(f'Mode: {mode}')
```

```
Mean: 19.6
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]

Soln.

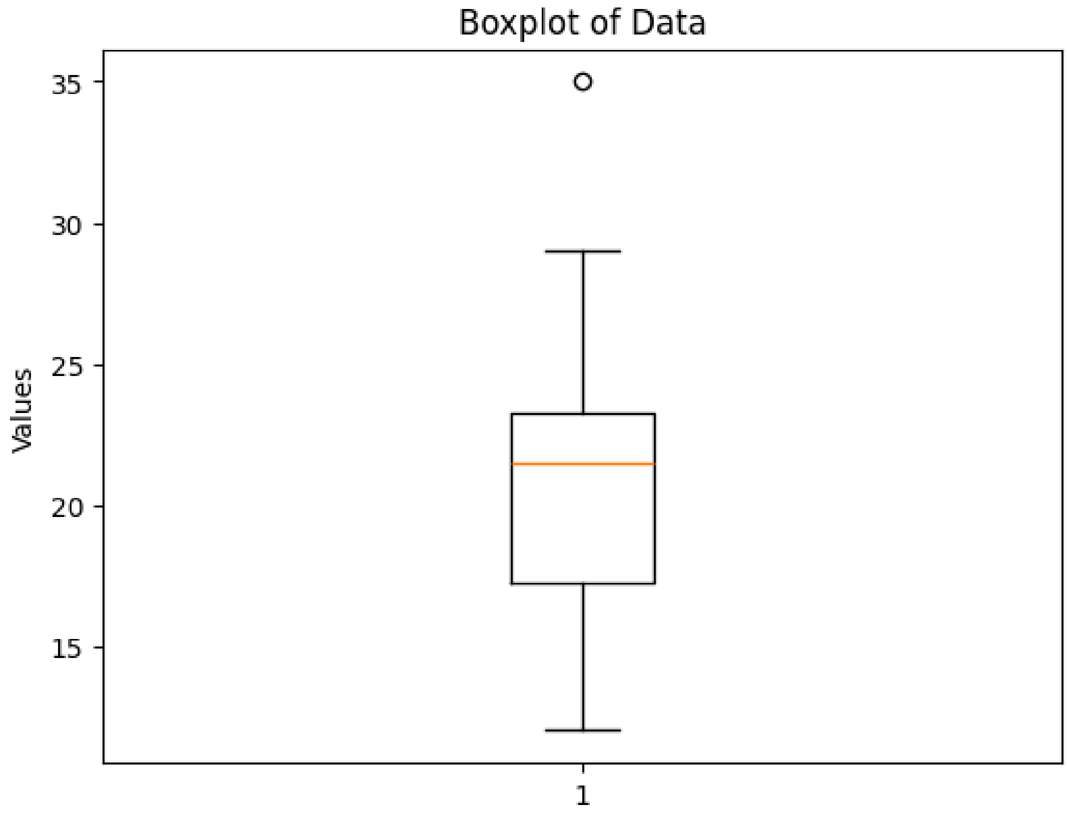
```
import statistics
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]
covariance = statistics.covariance(list_x, list_y)
correlation_coefficient = statistics.correlation(list_x, list_y)
print(f'Covariance: {covariance}')
print(f'Correlation Coefficient: {correlation_coefficient}')
```

```
Covariance: 275.0
Correlation Coefficient: 0.9958932064677039
```

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]

Soln.

```
import matplotlib.pyplot as plt
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
plt.boxplot(data)
plt.title('Boxplot of Data')
plt.ylabel('Values')
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]

Soln.

Positive Covariance: If analysis shows a positive value, it implies that as advertising spend increases, daily sales also tend to increase.
Negative Covariance: A negative value would suggest an inverse relationship, where higher ad spend corresponds to lower daily sales.
Zero Covariance: This indicates there is no clear linear relationship; the variables move independently of each other.

```
import statistics
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]
correlation = statistics.correlation(advertising_spend, daily_sales)
```

```
print(f'Correlation: {correlation}')
```

```
Correlation: 0.9935824101653327
```

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]

Soln.

Mean (Average): This gives the overall "mood" of the customer base. If the mean is 8.5, sentiment is generally high. However, the mean is sensitive to outliers (extremely unhappy customers).

Median (The Middle Value): In a 1–10 scale, the median is crucial. If the median is significantly different from the mean, it tells us the data is skewed (e.g., a few 1s are dragging down an otherwise happy group).

Standard Deviation:

Low Std Dev: Most customers feel the same way (everyone rates it a 7 or 8).

High Std Dev: Opinions are polarized (some people love it with 10s, others hate it with 2s). This is a red flag for a new product launch.

Mode: The most frequent score. On a 1–10 scale, knowing the most common rating helps identify the "typical" customer experience.

Histogram

This is the most important tool for this task. It shows the frequency of each score from 1 to 10.

Is it a "Bell Curve" Or is it Bimodal , suggesting your product works for one group but fails another?

Box Plot

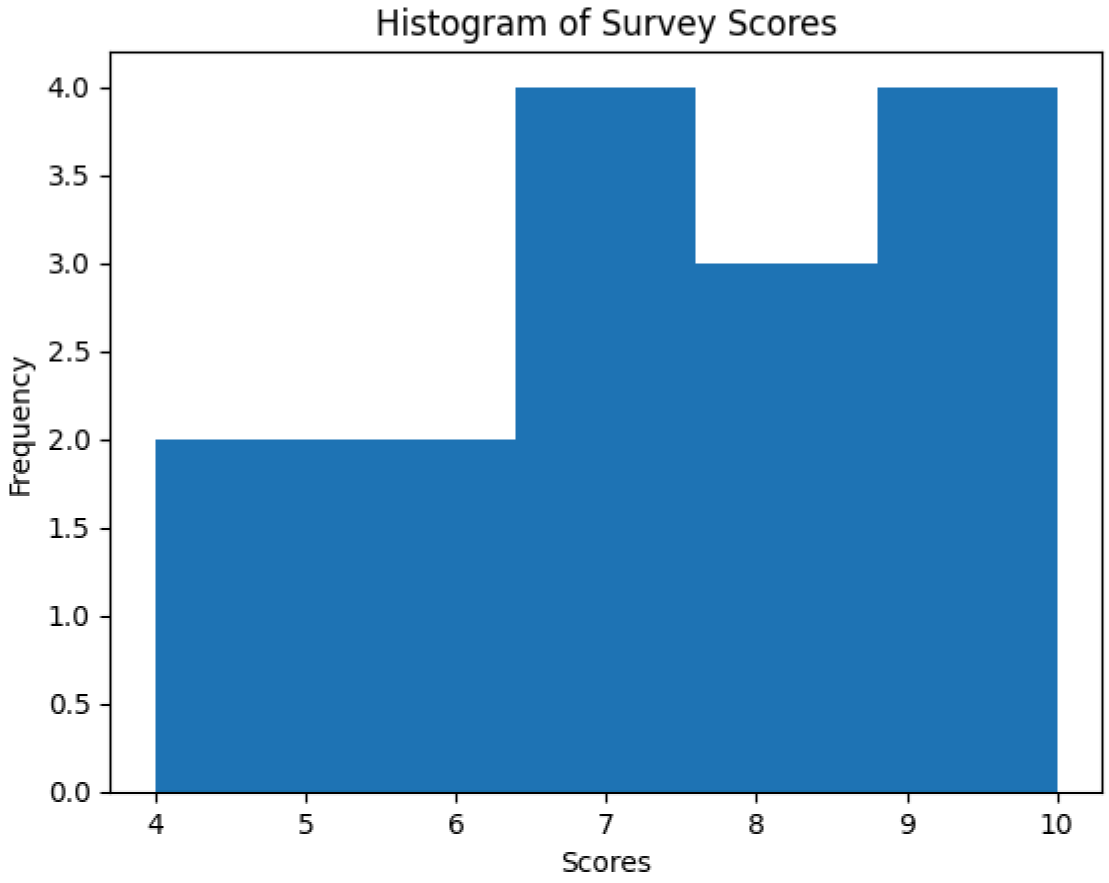
This is excellent for identifying outliers and the "interquartile range" (where the middle 50% of customers sit).

It quickly shows the "whiskers" (the range of typical responses) and any "dots" that represent extreme outliers customers who had an exceptionally bad or good experience that doesn't represent the norm.

Bar Chart

Since a 1–10 scale is "discrete" (customers usually pick whole numbers), a simple bar chart of the count for each score is often more readable for stakeholders than a smoothed histogram.

```python
import matplotlib.pyplot as plt
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
plt.hist(survey_scores, bins=5)
plt.title('Histogram of Survey Scores')
plt.xlabel('Scores')
plt.ylabel('Frequency')
plt.show()
```


Histogram of Survey Scores
Frequency
Scores