Question 1

1. Qualitative (Categorical) Data describes **qualities or characteristics** that classify items into categories.

- **Nominal data** involves categories with **no inherent order**;

  examples include gender (male, female) and colors (red, blue, green).

- **Ordinal data** has categories with a **meaningful order** but without consistent intervals

  examples include satisfaction levels (satisfied, neutral, unsatisfied) and education levels (high school, college, graduate).

2. Quantitative (Numerical) Data represents **numerical values** that allow for mathematical calculations.

- **Interval data** has numeric scales with **equal intervals** but **no true zero**

  examples temperature in Celsius or Fahrenheit (e.g., 10°C, 20°C, but no absolute zero).

- **Ratio data** includes numeric scales with **equal intervals** and a **true zero**, allowing for meaningful comparison of magnitudes;

  examples include height, weight, age, and income.

Question 2

Measures of central tendency are statistics that represent the **center** or **typical value** of a given dataset.

1. **Mean**: The **average** of all data points, calculated by summing the values and dividing by the total number of items. Use the mean when data is **symmetrical** and **continuous** (e.g., heights, weights). However, it is **sensitive to outliers**.
   - **Example**: In a class with test scores of 70, 80, 90, the mean is 80.
   - **Use Case**: Suitable for data without extreme values(outliers), like average temperatures.
2. **Median**: The **middle value** when data is sorted in ascending order. If the dataset has an **outlier**, the median is more reliable than the mean. Use the median with **skewed** distributions or **ordinal data**.
   - **Example**: For incomes of $30,000, $35,000, $100,000, the median is $35,000.
   - **Use Case**: Suitable for data with outliers, like income or home prices.

3. **Mode**: The **most frequent value** in the dataset. It’s useful for **categorical data** and for identifying the **most common value** in a dataset, even if there’s no numerical order.
    - **Example**: For survey responses {good, good, fair, excellent}, the mode is “good.”
    - **Use Case**: Best for nominal data(categorical), like survey responses or product preferences.

Question 3

Dispersion refers to the **spread** or **variability** of data points in a dataset. It shows how much the data points differ from each other and from the central value.

Variance

Variance measures the **average squared difference** between each data point and the mean.

A high variance indicates data points are far from the mean, meaning a greater spread, while a low variance indicates they are close to the mean, meaning less spread.

Standard Deviation

Standard Deviation is the **square root of the variance**.

Like variance, a high standard deviation indicates greater spread, while a low standard deviation means data points are close to the mean.

Question 4

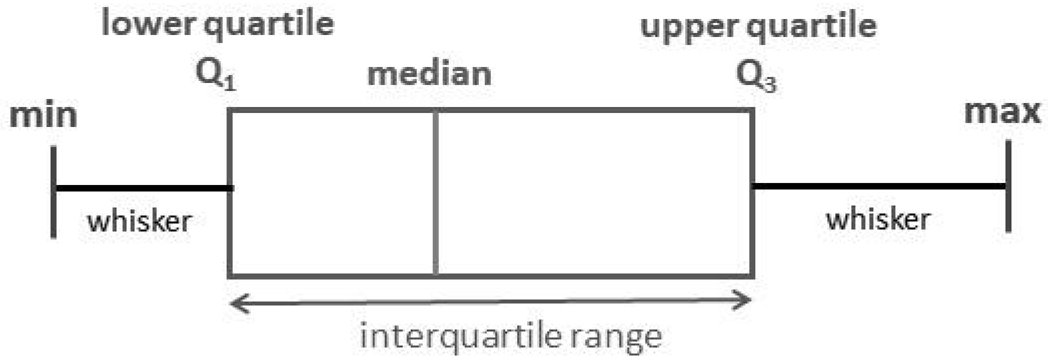
A **box plot** (or **box-and-whisker plot**) is a visual representation of the distribution of a dataset, showing its spread and skewness. It summarizes data using five key points: **minimum (Q0)**, **first quartile (Q1)**, **median (Q2)**, **third quartile (Q3)**, and **maximum(Q4)**.

What it Shows:

1. **Center**: The line inside the box shows the **median** (middle value).
2. **Spread**: The **box width** shows the **interquartile range (IQR)**, which covers the middle 50% of the data. IQR=Q3-Q1
3. **Whiskers**: Lines extending from the box show the **range** (spread) of the data, typically up to **1.5 × IQR** above Q3 and below Q1.
4. **Outliers**: Points outside the whiskers are **outliers**, indicating values unusually far from the majority of data.

What it Tells About Distribution:

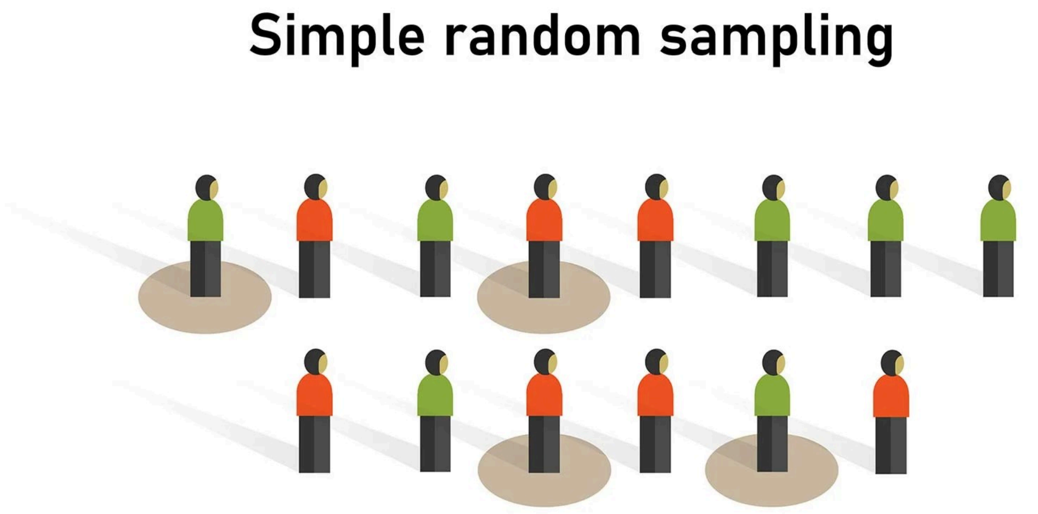
- **Symmetry**: If the box and whiskers are even around the median, the data is likely symmetric.
- **Skewness**: If the box is stretched to one side or if one whisker is longer, the data is skewed.
- **Outliers**: Points outside the whiskers show potential outliers, suggesting unusual data points.



Question 5

Random sampling is essential for understanding a population because it allows us to draw conclusions about a large group without studying everyone (or the whole population). Here's why it's so valuable:

1. **Better Representation**: With random sampling, each person in the population has an equal chance of being picked. This gives us a sample that reflects the whole population's diversity, making our findings more accurate.
2. **Less Bias**: Since we're picking randomly, we avoid favoring any specific group. This reduces bias, meaning our sample is more likely to capture the true picture of the population.
3. **Generalization**: A good random sample lets us confidently apply what we learn from the sample to the larger population. This way, we can estimate things like average age or income across the whole group.



Question 6

Skewness describes the **asymmetry** in a dataset's distribution. It tells us if data values are spread more to one side of the average (mean) than the other, helping to identify any imbalance in the data.

Types of Skewness:

1. **Positive Skew (Right Skew)**: The **tail is longer on the right** side. In this case, most data values are on the lower end, and a few high values pull the mean to the right of the median.
   - **Example**: Income data, where most people earn lower to moderate amounts, with a few earning very high incomes.
2. **Negative Skew (Left Skew)**: The **tail is longer on the left** side. Here, most data values are on the higher end, and a few low values pull the mean to the left of the median.
   - **Example**: Test scores where most students score well, but a few score very low.
3. **Symmetric (No Skew)**: Both sides of the distribution are **balanced and mirror each other**. The mean, median, and mode are roughly equal.
   - **Example**: Heights in a well-sampled population often form a symmetric distribution.

How Skewness Affects Interpretation:

- **Measures of Central Tendency**: In a skewed distribution, the mean is pulled in the direction of the skew. For example, in a positively skewed distribution, the mean is higher than the median.
- **Choosing the Right Measure**: In skewed data, the median is often a better measure of central tendency than the mean, as it isn't affected by extreme values.
- **Data Interpretation**: Knowing the skewness helps in understanding the spread of values. For instance, a positive skew might indicate that a few high outliers are raising the average, which is common in income or housing prices.

Question 7

The **Interquartile Range (IQR)** measures the **spread of the middle 50%** of a dataset. It's calculated as the difference between the third quartile (Q3) and the first quartile (Q1):

IQR=Q3−Q1

Using IQR to Detect Outliers:

Outliers are extreme values that are much higher or lower than the rest of the data. To identify them, we calculate the **lower fence** and **upper fence**:

- **Lower fence**: Q1−1.5×IQR
- **Upper fence**: Q3+1.5×IQR

Values below the lower fence and values above the upper fence are outliers.

Question 8

The **binomial distribution** is used to model the number of successes in a fixed number of independent trials, each with the same probability of success. It applies in the following conditions:

1. **Fixed Number of Trials**: There is a set number, n, of trials (e.g., flipping a coin 10 times).
2. **Two Possible Outcomes**: Each trial has only two possible outcomes, typically labeled as **success** and **failure** (e.g., heads or tails in a coin flip).
3. **Constant Probability of Success**: The probability of success, ppp, remains the same for each trial (e.g., the probability of getting heads remains 0.5 for each flip).
4. **Independence**: Each trial's outcome is independent of the others, meaning the result of one trial does not affect the others (e.g., each coin flip is independent of the last).

Question 9

The **normal distribution** is a symmetric, bell-shaped curve that describes how data points are distributed around the mean.

Properties of the Normal Distribution:

1. **Symmetry**: It's perfectly symmetric around the mean, so the left and right sides are mirror images.
2. **Mean = Median = Mode**: All three measures of central tendency are equal and located at the center of the distribution.
3. **Asymptotic**: The tails approach, but never touch, the horizontal axis, meaning values theoretically extend infinitely in both directions.
4. **Defined by Mean and Standard Deviation**: The shape and spread are determined by the mean (center) and standard deviation (spread).

The Empirical Rule (68-95-99.7 Rule):

This rule describes how data is distributed in a normal distribution:

- **68%** of the data falls within **1 standard deviation** from the mean.
- **95%** of the data falls within **2 standard deviations** from the mean.
- **99.7%** of the data falls within **3 standard deviations** from the mean.

Question 10

A **Poisson process** models the occurrence of **independent events** happening at a **constant average rate** over time or space. For example, the number of cars passing through a toll booth per hour can be modelled using a Poisson process.

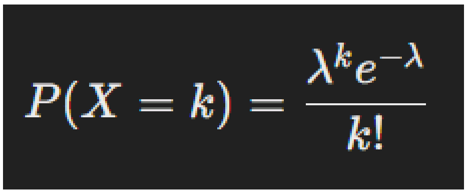
Example:

If **3 cars** pass through a toll booth every hour ($\lambda = 3$), and we want to calculate the probability of **4 cars** passing, we get:

$P(X=4)\approx 0.168$ or 16.8%

This means there's a **16.8%** chance of exactly **4 cars** passing the toll booth in one hour.

Formula:



$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where:

- **λ** = average number of events (rate),
- **k** = number of events we're calculating for,
- **e** = Euler's number (≈ 2.718).

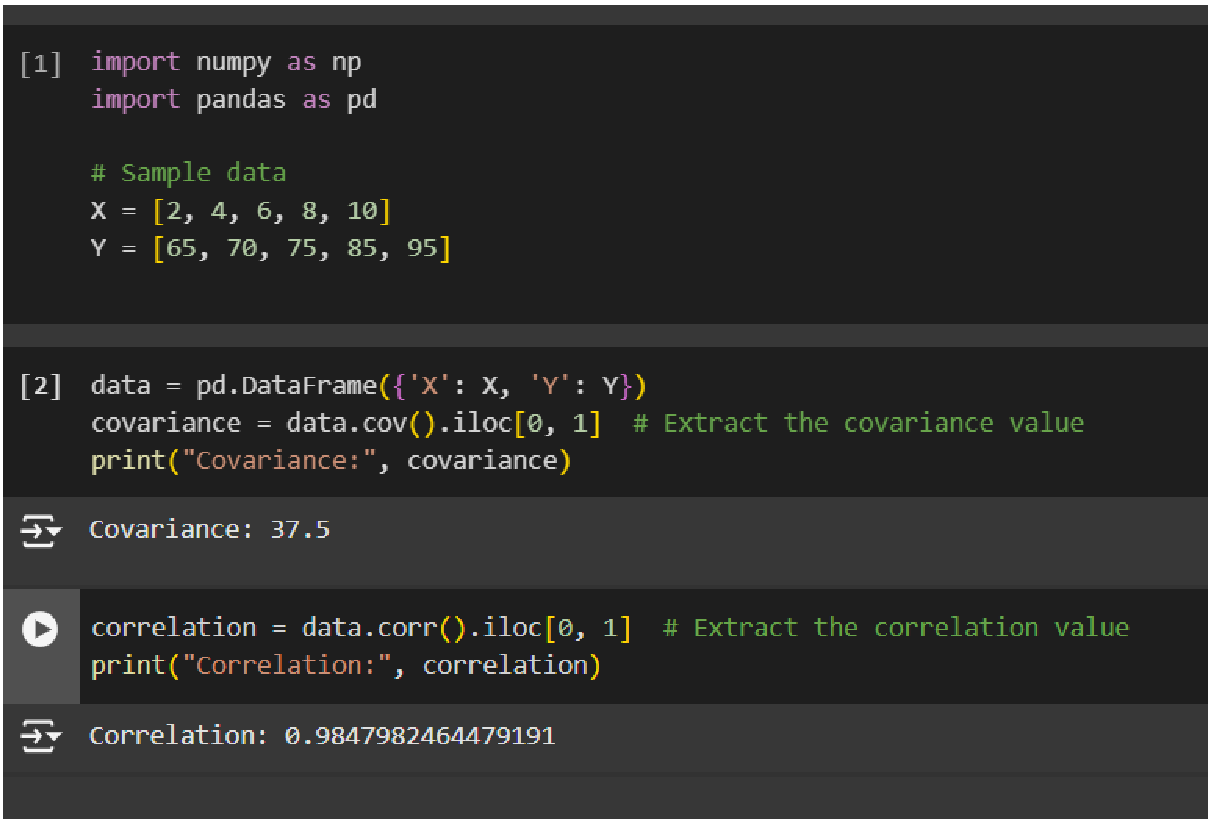
Question 11

A **random variable** is a variable that represents possible outcomes of a random phenomenon. It assigns a numerical value to each outcome of an experiment, allowing us to analyze probabilities mathematically.

Types of Random Variables:

1. **Discrete Random Variable**: Takes on **specific, countable values** (often whole numbers). It has a **finite** or **countably infinite** number of possible outcomes.
   - **Example**: The number of heads when flipping a coin three times (0, 1, 2, or 3 heads) or the number of customers in a line.
2. **Continuous Random Variable**: Takes on **any value within a range**, meaning it has an **infinite** number of possible values. These values are often measured rather than counted.
   - **Example**: The exact height of students in a class or the time it takes to run a race.

Question 12



```
[1] import numpy as np
    import pandas as pd

    # Sample data
    X = [2, 4, 6, 8, 10]
    Y = [65, 70, 75, 85, 95]
```

```
[2] data = pd.DataFrame({'X': X, 'Y': Y})
    covariance = data.cov().iloc[0, 1]  # Extract the covariance value
    print("Covariance:", covariance)
```

```
Covariance: 37.5
```

```
correlation = data.corr().iloc[0, 1]  # Extract the correlation value
print("Correlation:", correlation)
```

```
Correlation: 0.9847982464479191
```

Interpretation:

Covariance tells you whether the variables move together (positive covariance) or in opposite directions (negative covariance).

Correlation tells you how strongly and in what direction the variables are related, on a scale from -1 to +1. A positive correlation close to +1 means a strong direct relationship.