**Statistics Basics**

1. **Explain the different types of data (qualitative and quantitative) and provide examples of each. Discuss nominal, ordinal, interval, and ratio scales.**

**Qualitative Data**

Represents non-numeric information that categorizes or labels elements.

**Examples:**

Colors of cars: Red, Blue, Green.

Types of cuisines: Italian, Chinese, Indian.

Gender: Male, Female, Non-binary.

**Quantitative Data**

Represents numeric values that measure quantities.

**Examples:**

Height of individuals: 160 cm, 175 cm, 180 cm.

Age: 25 years, 30 years, 45 years.

Temperature: 22°C, 25°C, 30°C.

1. **What are the measures of central tendency, and when should you use each? Discuss the mean, median, and mode with examples and situations where each is appropriate?**

The most common measures are **mean**, **median**, and **mode**. Each has its unique use depending on the nature of the data and the analytical goals.

**1.The Mean (Arithmetic Average)**

**Definition:** The sum of all values divided by the number of values.

**Formula:** Mean=Sum of all values/Number of values

**Example:**

Scores in a test: 85,90,75,80,9585, 90, 75, 80, 9585,90,75,80,95.

Mean = (85+90+75+80+95)/5=85(85 + 90 + 75 + 80 + 95) / 5 = 85(85+90+75+80+95)/5=85.

**When to Use:**

When data is continuous and not heavily skewed.

Suitable for datasets without extreme outliers.

**Example Situation:**

Calculating the average height of students in a class.

**2.The Median (Middle Value)**

**Definition:** The middle value in a dataset when arranged in ascending or descending order. If there’s an even number of values, the median is the average of the two middle values.

**Steps to Calculate:**

* 1. Arrange data in order.
  2. Identify the middle value(s).

**Example:**

Dataset: 75,80,85,90,9575, 80, 85, 90, 9575,80,85,90,95. Median = 858585.

If even number of values: 70,75,80,8570, 75, 80, 8570,75,80,85, Median = (75+80)/2=77.5(75 + 80) / 2 = 77.5(75+80)/2=77.5.

**When to Use:**

When data is skewed or has outliers.

Median is robust to extreme values.

**Example Situation:**

Determining the typical income in a region with a few very high earners.

**3. The Mode (Most Frequent Value)**

**Definition:** The value that appears most frequently in a dataset.

**Characteristics:**

A dataset can have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode if all values occur with equal frequency.

**Example:**

Dataset: 1,2,2,3,4,4,4,51, 2, 2, 3, 4, 4, 4, 51,2,2,3,4,4,4,5. Mode = 444.

**When to Use:**

When data is categorical or when identifying the most common value is important.

**Example Situation:**

Determining the most popular ice cream flavor sold at a store.

* 1. Explain the concept of dispersion. How do variance and standard deviation measure the spread of data?

**Low Dispersion:** Data points are clustered closely around the central value.

Example: Test scores of students: 78, 79, 80, 81, 82.

**High Dispersion:** Data points are spread out widely.

Example: Test scores of students: 50, 60, 80, 90, 100.

**1. Variance**

**Definition:** Variance measures the average squared deviation of each data point from the mean. It indicates how much the data points differ from the mean, emphasizing larger deviations.

**Formula:**

Variance(σ2)=∑(xi−μ)2n\n​

Where:

xix\_ixi​: Each data point.

μ\muμ: Mean of the data.

nnn: Number of data points.

**Example:**

Data: 2,4,6,8,102, 4, 6, 8, 102,4,6,8,10.

Mean (μ\muμ): (2+4+6+8+10)/5=6(2+4+6+8+10)/5 = 6(2+4+6+8+10)/5=6.

Variance: (2−6)2+(4−6)2+(6−6)2+(8−6)2+(10−6)25=8\frac{(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2}{5} = 85(2−6)2+(4−6)2+(6−6)2+(8−6)2+(10−6)2​=8.

**Interpretation:** Larger variance indicates greater spread.

**2. Standard Deviation**

**Definition:** Standard deviation is the square root of variance. It measures the average distance of each data point from the mean, using the same units as the data.

**Formula:**

Standard Deviation(σ)=square root of Variance

**Example:**

Using the variance example above (σ2=8\sigma^2 = 8σ2=8).

Standard Deviation (σ\sigmaσ): 8≈2.83\sqrt{8} =2.83

**Interpretation:**

A low standard deviation means most data points are close to the mean.

A high standard deviation means data points are spread out.

4.What is a box plot, and what can it tell you about the distribution of data?

A **box plot** is a graphical representation of data that summarizes the distribution by showing key descriptive statistics. It provides a visual snapshot of the data’s central tendency, spread, and variability, while also highlighting potential outliers.

5.Discuss the role of random sampling in making inferences about populations.

**Random sampling** is a method of selecting a subset of individuals from a population where every member has an equal chance of being chosen. It plays a crucial role in making inferences about populations by:

**1.Ensuring Representativeness:** Reduces selection bias, making the sample reflect the population.

**2.Facilitating Generalization:** Allows findings from the sample to be applied to the entire population.

**3.Enabling Statistical Analysis:** Supports valid conclusions using probability- based techniques.

6.Explain the concept of skewness and its types. How does skewness affect the interpretation of data?

**Skewness** measures the asymmetry of a dataset's distribution. It indicates whether the data is symmetrically distributed around the mean or if it leans more towards one side.

**Types**

Positive Skew

Negative Skew

Symmetric Distribution

**Effect of Skewness on Data Interpretation**

**Central Tendency:** Skewed data affects the mean more than the median.

**Outliers:** Skewness can indicate the presence of extreme values.

**Decision-Making:** Positive skew implies more high values, while negative skew suggests more low values.

7.What is the interquartile range (IQR), and how is it used to detect outliers?

The Interquartile Range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1), representing the middle 50% of the data:

IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1

**Detecting Outliers:**

Lower Bound=Q1−1.5×IQR

Upper Bound=Q3+1.5×IQR

8. Discuss the conditions under which the binomial distribution is used.

1.There are a fixed number of trials (nnn).

2.Each trial has only two outcomes (success or failure).

3.The probability of success (ppp) remains constant for all trials.

4.The trials are independent of each other.

**9. Explain the properties of the normal distribution and the empirical rule (68-95-99.7 rule).**

**1.Symmetry:** Bell-shaped and symmetric around the mean (μ\muμ).

**2.Mean = Median = Mode:** All are equal and located at the center.

**3.Asymptotic Tails:** Tails approach but never touch the horizontal axis.

**4.Standard Deviation (σ\sigmaσ):** Determines the spread; larger σ\sigmaσ means wider distribution.

**5.Total Area Under the Curve:** Equals 1, representing 100% probability.

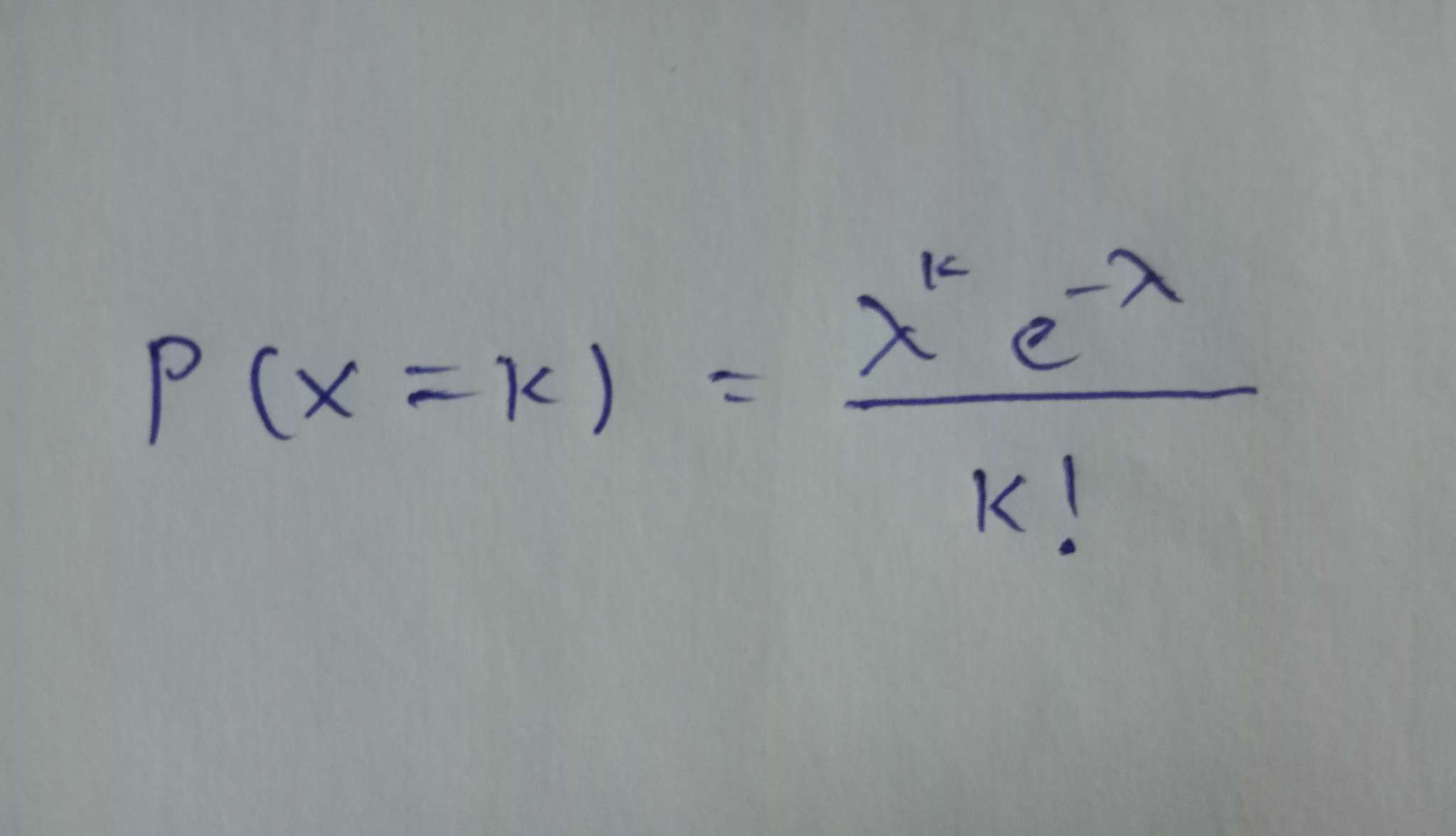
**10.Provide a real-life example of a Poisson process and calculate the probability for a specific event**.

A customer service center receives an average of 3 calls per minute. This follows a Poisson process since calls are independent, random, and occur at a constant average rate.

**Probability Calculation**

**Find:** Probability of receiving exactly 5 calls in a minute.

**Formula:**

****

λ=3 (average rate of calls),

k=5 (specific number of calls).

**Solution:**

P(X=5)= 3^5 e^-3 / 5! = 0.1008

**Result:**

The probability of receiving exactly 5 calls in a minute is approximately **10.08%**.

**11. Explain what a random variable is and differentiate between discrete and continuous random variables.**

A **random variable** is a numerical value determined by the outcome of a random process or experiment.

**Types**

1. **Discrete Random Variable:**

Takes specific, countable values.

**Examples:** Number of heads in 10 coin flips, number of students in a class.

1. **Continuous Random Variable:**

Takes any value within a range or interval.

**Examples:** Height of individuals, time to complete a task.

**12. Provide an example dataset, calculate both covariance and correlation, and interpret the results.**

**Example Dataset**

**X y**

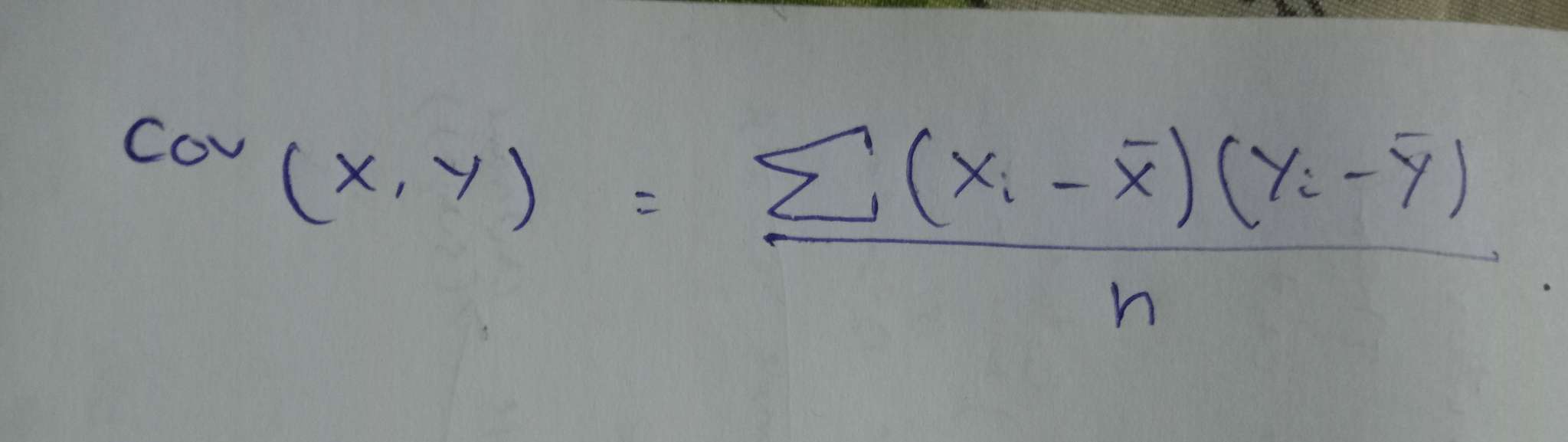
1 2

2 4

3 6

4 8

**Calculate Covariance**

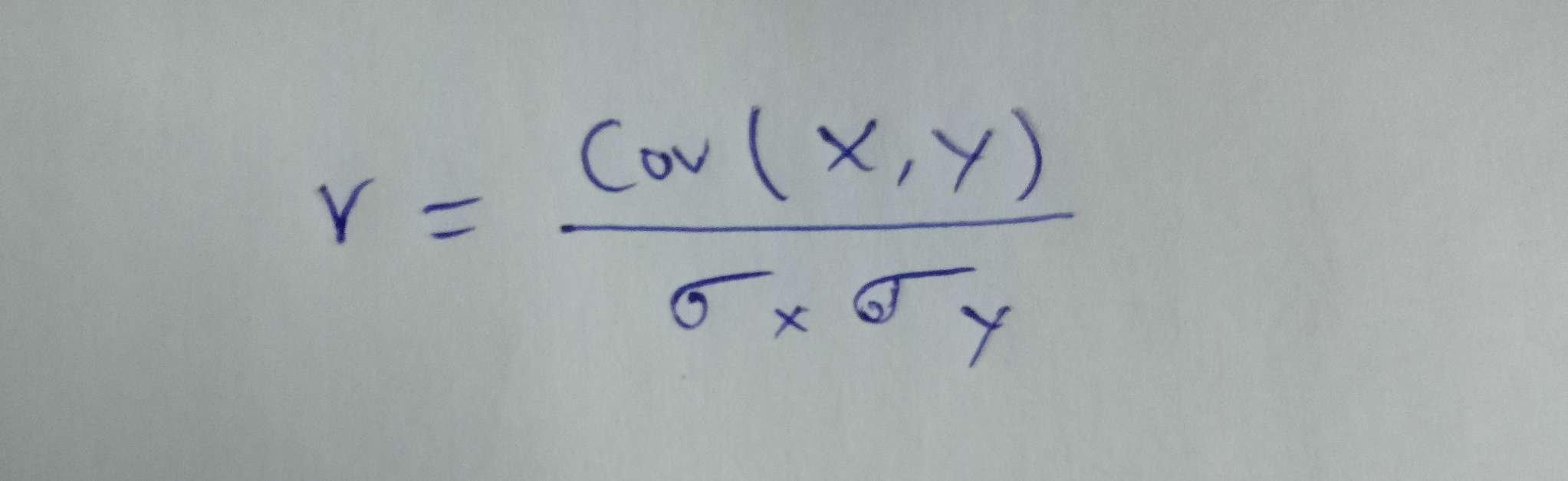
****

**Means:** Xˉ=2.5\bar{X} = 2.5Xˉ=2.5, Yˉ=5\bar{Y} = 5Yˉ=5.

**Cov(X,Y**)= (1−2.5)(2−5)+(2−2.5)(4−5)+(3−2.5)(6−5)+(4−2.5)(8−5)​/4

**=3.5**

**Calculate Correlation**

****

**Standard Deviations:  
σX**=1.12 **, σY**​=2.24.

**r**=3.5/1.12×2.243 = 1