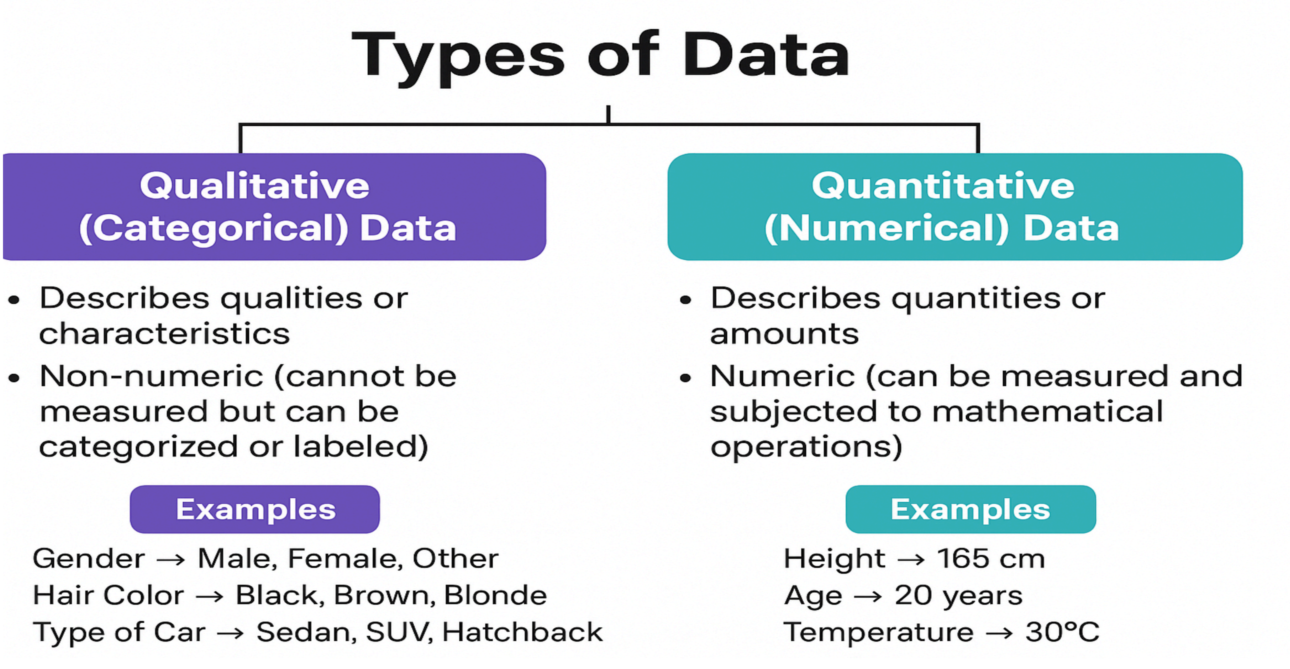
Statistics Basics
Assignment Questions

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

Answer:

1. Types of Data

Data can broadly be divided into **two main types**:



A. Qualitative Data (Categorical Data)

- **Meaning:** Describes *qualities* or *characteristics*.
- **Nature:** Non-numeric (cannot be measured but can be categorized or labeled).
- **Examples:**
    - Gender → Male, Female, Other
    - Hair Color → Black, Brown, Blonde
    - Type of Car → Sedan, SUV, Hatchback

Qualitative data is further divided into:

1. **Nominal Data**
2. **Ordinal Data**

B. Quantitative Data (Numerical Data)

- **Meaning:** Describes *quantities* or *amounts*.
- **Nature:** Numeric (can be measured and subjected to mathematical operations).
- **Examples:**
  - Height → 165 cm
  - Age → 20 years
  - Temperature → 30°C

Quantitative data is further divided into:

1. **Interval Data**
2. **Ratio Data**

2. Levels of Measurement (Data Scales)

Scale	Type	Description	Examples
Nominal	Qualitative	Categories with **no order** or ranking.	Gender (Male/Female), Blood group (A, B, AB, O)
Ordinal	Qualitative	Categories with a **meaningful order**, but **differences between ranks aren't equal**.	Movie ratings (Excellent, Good, Average, Poor), Education level (High school, College, Graduate)
Interval	Quantitative	Ordered, **equal intervals** between values, but **no true zero**.	Temperature in °C or °F, Dates on a calendar
Ratio	Quantitative	Ordered, **equal intervals**, and has a **true zero** (zero means absence of the quantity).	Height, Weight, Age, Salary, Distance

3. Key Differences at a Glance

Feature	Nominal	Ordinal	Interval	Ratio
Type of Data	Qualitative	Qualitative	Quantitative	Quantitative
Order	No	Yes	Yes	Yes
Equal Intervals	No	No	Yes	Yes
True Zero	No	No	No	Yes
Example	Gender	Satisfaction Level	Temperature (°C)	Height, Weight

4. Example Summary

Let’s take student data as an example:

Variable	Type	Scale
Name	Qualitative	Nominal
Grade Level (A, B, C, D)	Qualitative	Ordinal
Test Score	Quantitative	Interval
Age or Weight	Quantitative	Ratio

Question 2. 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

Answer:

- ## Measures of Central Tendency

🔹 Definition

Measures of central tendency are statistical values that represent the center or average of a dataset.
They help summarize an entire dataset with a single representative value.

The three main measures are:

1. Mean
2. Median
3. Mode

1️⃣ Mean (Arithmetic Average)

Definition:

The mean is the sum of all data values divided by the total number of values.

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

Example:

Dataset: 5, 10, 15, 20, 25

$$\text{Mean} = \frac{5+10+15+20+25}{5}$$

$= 15$

When to Use:

- When data is numeric and symmetrically distributed (no extreme outliers).
- Best for interval or ratio data (e.g., test scores, heights, weights).

When *Not* to Use:

- When there are outliers or skewed data (they can distort the mean).

Example Situation:

Average salary of employees. If one CEO earns ₹10,00,000 while others earn ₹30,000, the mean would give a misleadingly high result.

2️⃣ Median

Definition:

The median is the middle value when data is arranged in ascending or descending order. If there is an even number of data points, the median is the average of the two middle values.

Example:

Dataset: 5, 10, 15, 20, 25
→ Median = 15 (middle value)

Dataset: 5, 10, 15, 20
→ Median = (10 + 15) / 2 = 12.5

When to Use:

- When data has outliers or is skewed.
- Appropriate for ordinal, interval, or ratio data.

Example Situation:

House prices — if most homes cost ₹30–50 lakhs but one mansion costs ₹5 crore, the median gives a better idea of the typical price than the mean.

3️⃣ Mode

Definition:

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

Example:

Dataset: 2, 4, 4, 6, 8, 8, 8, 10
→ Mode = 8 (appears most often)

When to Use:

- When dealing with categorical or nominal data (like colors, brands, etc.).
- To identify the most common value in a dataset.

Example Situation:

- Most common shoe size sold in a store → Mode = 8
- Most popular car color → Mode = White

4️⃣ Quick Comparison Table

Measure	Definition	Best For	Affected by Outliers?	Example Use Case
Mean	Arithmetic average	Symmetrical numerical data	✅ Yes	Average marks in a test
Median	Middle value	Skewed data or outliers	❌ No	Median household income
Mode	Most frequent value	Categorical data	❌ No	Most common shoe size

5️⃣ Summary Example

Dataset: [10, 12, 13, 13, 100]

Measure	Value	Comment
Mean	29.6	Skewed by the large outlier (100)
Median	13	Represents the center more accurately
Mode	13	Most common value

Question 3. Explain the concept of dispersion. How do variance and standard deviation measure the spread of data?

Answer:

Concept of Dispersion

🔹 Definition

Dispersion refers to the degree to which data values are spread out or scattered around the central value (mean, median, or mode). It helps us understand how consistent or variable the data is.

In simple terms:

- Low dispersion: Data values are close to the mean (less variability).
- High dispersion: Data values are spread far from the mean (more variability).

🔸 Example of Dispersion

Dataset	Mean	Observation
A: [50, 51, 49, 50, 50]	50	Values are very close → Low dispersion
B: [10, 20, 50, 80, 90]	50	Values vary widely → High dispersion

> Although both sets have the same mean (50), Dataset B is more spread out — showing higher dispersion.

🔹 Common Measures of Dispersion

Measure	Description	Use
Range	Difference between highest and lowest values	Simple but ignores distribution shape

Variance	Average of squared deviations from the mean	Quantifies spread mathematically
Standard Deviation (SD)	Square root of variance	Most widely used measure of spread

---

- ## Variance and Standard Deviation

1️⃣ Variance (σ^2 or s^2)

Definition:

Variance measures how far each data point is from the mean on average.

$$\textbf{Variance} = \frac{\sum(x_i - \bar{x})^2}{n}$$

Where:

- xix_ixi = each data point
- x¯\bar{x}x¯ = mean of data
- nnn = number of data points

Example:

Data: 2, 4, 6, 8

$$\text{Mean}=5$$

$$\text{Variance}=\frac{(2-5)^2+(4-5)^2+(6-5)^2+(8-5)^2}{4}=\frac{9+1+1+9}{4}=5$$

So, **Variance = 5**

2️⃣ Standard Deviation (σ or s)

Definition:

The **standard deviation** is the **square root of variance**.
It brings the measure back to the **same unit as the data**, making it easier to interpret.

$$\text{Standard Deviation}=\sqrt{\text{variation}}$$

Using the previous example:

SD= $\sqrt{5}$ =2.24

🔹 Interpretation

Value of SD	Meaning
Small SD	Data values are close to the mean → less variability
Large SD	Data values are widely spread → more variability

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

Answer:

Box Plot (Box-and-Whisker Plot)

🔹 Definition

A box plot (also called a box-and-whisker plot) is a graphical representation of data distribution that shows the spread and skewness of a dataset.

It visually displays the five-number summary:

1. Minimum
2. First Quartile (Q1)
3. Median (Q2)
4. Third Quartile (Q3)
5. Maximum

🔸 Structure of a Box Plot

```
|-------|==============|-------|
Min      Q1      Median     Q3      Max
```

- Box: Represents the interquartile range (IQR) = Q3 − Q1
 (Middle 50% of the data)
- Line inside the box: The median (Q2)
- Whiskers: Extend to minimum and maximum values (excluding outliers)
- Dots (if present): Represent outliers (values that are unusually high or low)

🔹 Five-Number Summary Example

Suppose we have this dataset:
`[5, 7, 8, 10, 12, 13, 14, 18, 21, 25]`

Measure	Value	Meaning
Minimum	5	Smallest value
Q1	8	25th percentile
Median (Q2)	12.5	Middle value
Q3	18	75th percentile
Maximum	25	Largest value

Interquartile Range (IQR)= Q3 − Q1 = 18 − 8 = 10

🔹 Example Interpretations

1. **Symmetrical Distribution**
    - Median centered in the box
    - Whiskers roughly equal
      → Data is evenly distributed
2. **Right-Skewed Distribution**
    - Median near bottom (Q1)
    - Right whisker longer
      → More high values (tail on right)
3. **Left-Skewed Distribution**
    - Median near top (Q3)

    - Left whisker longer
      → More low values (tail on left)

🔹 **Why Use a Box Plot?**

✅ Easy to compare multiple datasets
✅ Identifies outliers clearly
✅ Shows data spread and symmetry
✅ Useful for understanding variability at a glance

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

Answer:

Role of Random Sampling in Making Inferences About Populations.

🔹 **Key Concepts**

Population:

The **entire group** you want to study or draw conclusions about.
👉 Example: All students in a university, all voters in a country, or all customers of a company.

Sample:

A **subset** of the population selected for analysis.
👉 Example: 200 randomly chosen students from the university.

Inference:

The process of using **sample data** to **draw conclusions or make predictions** about the **population**.

🔸 **What Is Random Sampling?**

Random sampling means that **every individual in the population has an equal chance** of being selected in the sample.

This is done to ensure that the sample is **unbiased** and **representative** of the whole population.

🔹 Why Random Sampling Is Important

Purpose	Explanation	Example
1. Reduces bias	Ensures selection is not influenced by personal choice or hidden patterns.	Choosing students by random ID numbers rather than by department.
2. Ensures representativeness	The sample reflects the diversity and characteristics of the population.	Randomly selecting people from all age groups for a survey.
3. Enables valid statistical inference	Results from the sample can be used to estimate population parameters (like mean, proportion, etc.).	Estimating average income of citizens using a random sample of 1,000 people.
4. Allows for probability-based conclusions	Since selection is random, probability theory can be applied to calculate confidence intervals and test hypotheses.	Calculating the margin of error in an election poll.

🔹 Types of Random Sampling

Type	Description	Example
Simple Random Sampling	Every member has an equal chance of selection.	Randomly picking 100 employees from an HR list using random numbers.
Systematic Sampling	Selecting every *kth* item from a list.	Choosing every 10th customer entering a store.
Stratified Sampling	Population divided into subgroups (strata), and random samples taken from each.	Selecting random students from each grade level.
Cluster Sampling	The population is divided into clusters, and some clusters are randomly selected.	Randomly selecting schools and surveying all students within them.

.

🔹 Example of Random Sampling in Inference

Suppose a company wants to estimate the **average satisfaction score** of its 10,000 customers.
Instead of surveying everyone (impractical), it:

- Randomly selects **500 customers** (sample)
- Calculates their **mean satisfaction score = 8.2/10**
- Uses statistical inference to estimate that the **true population mean** lies between **8.0 and 8.4** with 95% confidence.

✅ Because the sample was random, this inference is **valid and unbiased**.

🔹 Consequences of Non-Random Sampling

Problem	**Effect**
Selection Bias	Certain groups are over/under-represented.
Sampling Error Increases	Results are less reliable.
Inaccurate Inferences	Population conclusions become misleading.

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

Answer :

Definition

Skewness refers to the degree of asymmetry in the distribution of data around its mean.

- If data is symmetrical, the left and right sides of the distribution (around the mean) look the same.
- If not, the data is said to be skewed.

In simple terms:

> Skewness tells us whether the data leans to the left or right of the mean.

- ## Types of Skewness

There are three main types :

1️⃣ Symmetrical Distribution (No Skew / Zero Skewness)

- Data is **evenly distributed** around the mean.
- **Mean = Median = Mode**
- Graph looks **balanced** on both sides.

📊 **Example:**
Heights of adult males, or normally distributed exam scores

```
                     |

                   __|__

      /                             \

---|-----------------|----------------------|---

 Mode            Median              Mean
```

2️⃣ Positively Skewed (Right-Skewed) Distribution

- **Tail** is stretched **to the right** (toward higher values).
- Most data values are **small**, but a few **large values** pull the mean upward.
- **Mean > Median > Mode**

📊 **Example:**
Income distribution (a few very rich people raise the average).

```
          |
```

```
       |      __
       |    /
       |  /
   ____|_/__________
 Mode Median     Mean
```

3️⃣ Negatively Skewed (Left-Skewed) Distribution

- **Tail** is stretched **to the left** (toward lower values).
- Most data values are **large**, but a few **small values** pull the mean downward.
- **Mean < Median < Mode**

📊 **Example:**
Retirement age or exam scores where most students scored high but a few failed badly.

```
         |
     __
    /
  /
__/__________
Mean Median  Mode
```

🔹 How Skewness Affects Data Interpretation

Aspect	Effect of Skewness
Mean	Pulled in the direction of the skew (right or left).
Median	More resistant to skew and outliers; a better measure of central tendency for skewed data.
Mode	Remains near the highest frequency value; least affected by extremes.

Data Analysis	Helps identify data distribution shape and whether transformations or nonparametric tests are needed.
Decision Making	Indicates if the average value is misleading (e.g., average income in right-skewed data).

🔹 **Summary Table**

Type of Skewness	Shape	Relation (Mean, Median, Mode)	Tail Direction	Example
Symmetrical	Balanced	Mean = Median = Mode	None	Normal distribution
Positive Skew	Tail to right	Mean > Median > Mode	Right	Income, house prices
Negative Skew	Tail to left	Mean < Median < Mode	Left	Retirement age, easy test scores

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

Answer:

🔹 Definition:

The **Interquartile Range (IQR)** measures the **spread of the middle 50%** of a dataset.
It shows where the *central portion* of the data lies and helps identify **variability** and **outliers**.

🔸 Formula:

$$IQR = Q3 - Q1$$

Where:

- **Q1 (First Quartile)** → 25% of data lies below this point
- **Q3 (Third Quartile)** → 75% of data lies below this point

So, the IQR covers the **middle 50%** of the data (from Q1 to Q3).

🔹 Example Calculation

Dataset:
5,7,8,10,12,13,14,18,21,255, 7, 8, 10, 12, 13, 14, 18, 21, 255,7,8,10,12,13,14,18,21,25

Quartile	Position	Value
Q1	25th percentile	8
Q3	75th percentile	18

IQR=Q3−Q1=18−8=10

➡ The middle 50% of the data lies between **8 and 18**.

🔹 How IQR Helps Detect Outliers

Outliers are values that are **much smaller or larger** than most of the data. Using IQR, we can set **"fences"** (boundaries) to identify them.

Formulas:

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

Any data value:

- **Below the lower fence → Low outlier**
- **Above the upper fence → High outlier**

🔹 Example:

From our previous dataset:

Q1=8, Q3=18, IQR=10
Lower Fence=8−1.5(10)=−7
Upper Fence=18+1.5(10)=33

✅ Any value **below -7 or above 33** is an **outlier**.
Since our data values (5 to 25) fall within this range → **no outliers**.

🔹 Visual Interpretation (Box Plot Connection)

A **box plot** displays :

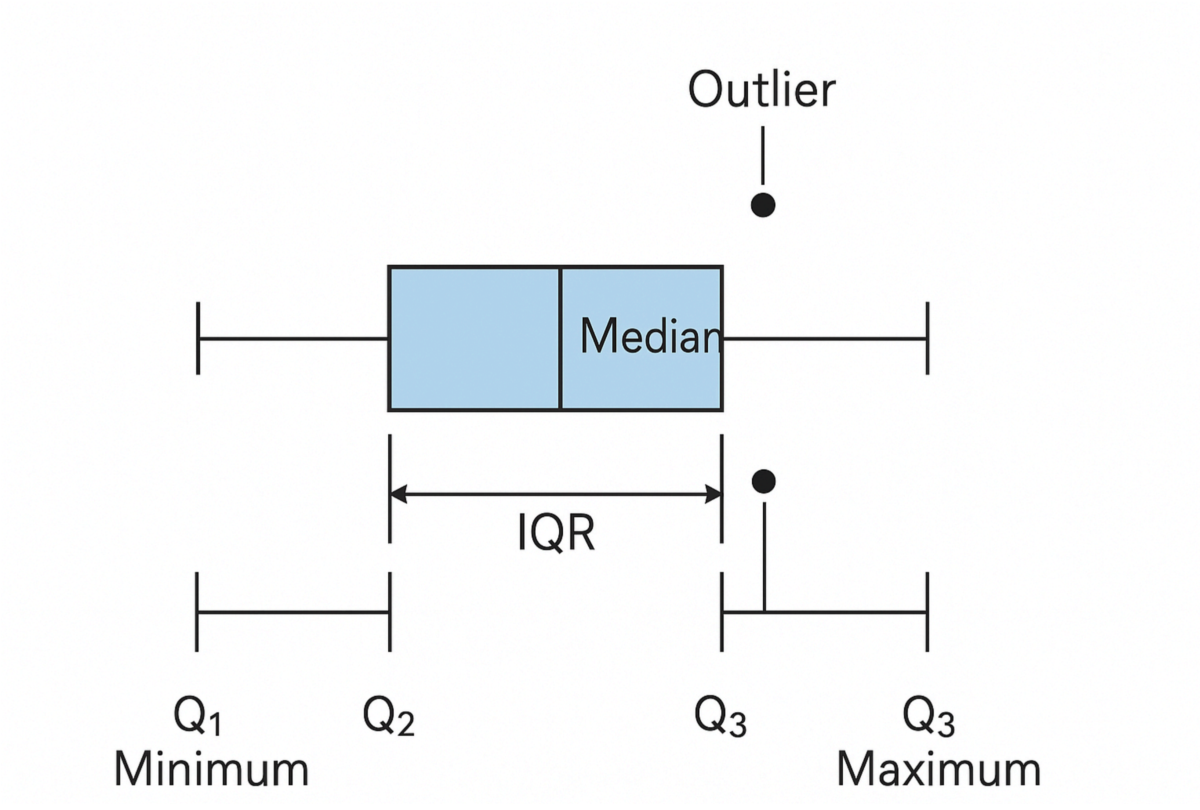
- The **box** = from Q1 to Q3 (the IQR)

- **Whiskers** = extend to non-outlier minimum and maximum values
- **Dots** beyond whiskers = outliers

So, **IQR** is the **core part** of a box plot!

🔹 Why IQR Is Important

Reason	Explanation
1. Resistant to outliers	Unlike range or standard deviation, IQR ignores extreme values.
2. Shows central spread	Focuses on the middle 50% — gives a better sense of data consistency.
3. Detects outliers	Provides a systematic rule for identifying extreme observations.
4. Used in box plots	Core component for visualizing distribution and spotting anomalies.



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

Answer:

The binomial distribution is used to model situations where there are two possible outcomes (often called *success* and *failure*) in a fixed number of independent trials.

✅ Conditions for Using the Binomial Distribution:

1. **Fixed Number of Trials (n):**
   The experiment or process is repeated a specific number of times, say nnn.
   Example: Tossing a coin 10 times.

2. **Two Possible Outcomes per Trial:**
   Each trial results in only two outcomes — success or failure.

   Example: In a coin toss — *Heads (success)* or *Tails (failure)*.

3. **Constant Probability of Success (p):**
   The probability of success remains the same for each trial.
   Example: For a fair coin, p=0.5p = 0.5p=0.5 for every toss.

4. **Independent Trials:**

   Each trial is independent of the others; the outcome of one trial does not affect the outcome of another.
   Example: The result of one coin toss doesn't influence the next.

5. **Discrete Random Variable (X):**

   The random variable XXX represents the number of successes in nnn trials, and it can take values 0,1,2,...,n0, 1, 2, ..., n0,1,2,...,n.

📘 Probability Formula:

$$P(X=k)=(n/k)pk(1-p)^{n-k}$$

Where:

- nnn = number of trials
- kkk = number of successes
- ppp = probability of success
- (1−p)(1 - p)(1−p) = probability of failure
- (nk)\binom{n}{k}(kn) = number of combinations of nnn items taken kkk at a time

Example:

If you toss a fair coin 5 times, the probability of getting exactly 3 heads can be found using the binomial distribution with:

- n=5n = 5n=5, k=3k = 3k=3, p=0.5p = 0.5p=0.5

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

Answer:

1. Normal Distribution – Overview

The Normal Distribution (also called the Gaussian distribution) is a continuous probability distribution that is symmetrical and forms a bell-shaped curve.
It is one of the most important distributions in statistics because many natural and social phenomena follow this pattern (e.g., heights, test scores, blood pressure, etc.).

2. Properties of the Normal Distribution

1. **Symmetrical Shape:**
    - The curve is perfectly symmetrical around the mean (μ).
    - Mean = Median = Mode.
2. **Bell-Shaped Curve:**
    - Most values cluster around the mean, and probabilities decrease as you move away from it.
3. **Mean and Standard Deviation Control the Shape:**
    - Mean (μ) determines the center (location).
    - Standard deviation (σ) determines the spread (width) of the curve.
4. **Total Area Under the Curve = 1:**
    - The entire probability space sums to 1 (or 100%).
5. **Asymptotic Nature:**
    - The curve approaches but never touches the x-axis.
6. **Empirical Rule Applies:**
    - About 68%, 95%, and 99.7% of the data fall within 1, 2, and 3 standard deviations from the mean.
    -

📘 3. The Empirical Rule (68–95–99.7 Rule)

The Empirical Rule describes how data are distributed in a normal distribution:

Range from the Mean (μ)	Percentage of Data	Description
$\mu \pm 1\sigma$	≈ 68%	Most values are within one standard deviation of the mean
$\mu \pm 2\sigma$	≈ 95%	Almost all values are within two standard deviations
$\mu \pm 3\sigma$	≈ 99.7%	Nearly all values are within three standard deviations

📉 4. Visual Representation

If we imagine a bell curve:

|--------|-------------|-------------|-----------|

μ-3σ μ-2σ μ μ+2σ μ+3σ

0.15% 2.35% 68% 2.35% 0.15%

Example:

Suppose students' test scores are normally distributed with

- Mean (μ) = 70
- Standard deviation (σ) = 10

Then, according to the empirical rule:

- 68% of students scored between 60 and 80
- 95% scored between 50 and 90
- 99.7% scored between 40 and 100

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

Answer :

A Poisson process models the number of events that occur randomly and independently over a fixed interval of time, area, or distance, given a known average rate (λ).

It is used when events:

- Occur independently
- Have a constant average rate
- Cannot happen simultaneously

2. Real-Life Example:

Let's say a **bank receives an average of 3 customer arrivals per minute** (λ = 3).
We want to find the probability that **exactly 5 customers** arrive in the next minute.

📘 3. Formula for the Poisson Distribution:

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

Where:

- X = number of events (arrivals)
- λ = average number of events per interval
- k = specific number of events
- e = 2.71828 (Euler's number)

4. Substituting Values:

Given:

- λ= 3

- k= 5

$$P(X = 5) = \frac{e^{-3} \times 3^{5}}{5!}$$

Now calculate step-by-step:

$$P(X = 5) = \frac{(0.0498) * 243}{120} = \frac{12.1014}{120} = 0.1008$$

✅ **Probability = 0.1008 (≈ 10.1%)**

So, there's about a **10% chance** that exactly 5 customers arrive in the next minute.

📊 5. Real-Life Uses of Poisson Distribution

- Number of phone calls received at a call center per minute
- Number of typing errors per page
- Number of cars passing a checkpoint per hour
- Number of emails received per hour

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

Answer :

A random variable is a numerical value that represents the outcome of a random experiment. It assigns a number to each possible outcome of a probabilistic event.

In other words, it's a variable whose value is determined by chance.

📘 Example:

- **When you toss a coin:**
  - Possible outcomes: {Heads, Tails}

- Define a random variable XXX:
  X=1X = 1X=1 if Heads, X=0X = 0X=0 if Tails.

- **When you roll a die:**
  - Possible outcomes: {1, 2, 3, 4, 5, 6}
  - Random variable XXX = number shown on the die.

2. Types of Random Variables

Type	Description	Example	Probability Representation
Discrete Random Variable	Takes **countable** values (finite or countably infinite).	Number of students absent, number of calls per hour	**Probability Mass Function (PMF)**
Continuous Random Variable	Takes **uncountably infinite** values (within an interval).	Height, weight, time, temperature	**Probability Density Function (PDF)**

3. Key Differences

Feature	Discrete Random Variable	Continuous Random Variable
Possible Values	Countable (0, 1, 2, 3, …)	Infinite (within a range, e.g., 1.2, 1.23, 1.234…)
Examples	Number of cars in a parking lot, number of emails	Temperature, height, time, distance
Function Type	PMF (Probability Mass Function)	PDF (Probability Density Function)
Probability of Specific Value	($P(X = x)$) has a specific value	($P(X = x) = 0$), but ($P(a < X < b) > 0$)
Graph Type	Bar graph	Smooth curve (bell-shaped, etc.)

4. Example of Each

- **Discrete Example:**
  Let X = number of heads in 3 coin tosses.
  Possible values: {0, 1, 2, 3}.
  P(X=2) means getting exactly 2 heads.

- **Continuous Example:**
  Let Y = time taken by a student to finish a test.
  Possible values: any real number between 0 and 180 minutes.
  $P(60<Y<90)$ means the probability that the student finishes between 1 and 1.5 hours.

✅ Summary:

- A **random variable** converts outcomes into numbers.
- **Discrete** → Countable values (use PMF).
- **Continuous** → Infinite values within an interval (use PDF).

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

Answer:

1. Example Dataset

Let's consider a small dataset showing the relationship between **hours studied** and **exam scores** for 5 students:

Student	Hours Studied (X)	Exam Score (Y)
A	2	65
B	3	70
C	5	75
D	6	85
E	8	90

2. Step 1: Calculate the Mean of X and Y

$$\overline{X} = \frac{2+3+5+6+8}{5} = 4.28$$

$$\overline{Y} = \frac{65+70+75+85+90}{5} = 77$$

3. Step 2: Calculate the Covariance

Formula:

$$\textbf{Cov(X , Y)} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{y})}{\textbf{n-1}}$$

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
2	65	-2.8	-12	33.6
3	70	-1.8	-7	12.6
5	75	0.2	-2	-0.4
6	85	1.2	8	9.6
8	90	3.2	13	41.6

$$\Sigma(X_i - \overline{X})(Y_i - \overline{y}) = 97$$

$$\text{Cov}(X,Y) = \underline{97} = 24.25$$

5-1

✅ Covariance = 24.25

4. Step 3: Calculate the Correlation

Formula:

$$\mathbf{r} = \frac{Cov(X,Y)}{S_X S_Y}$$

Where SXS_XSX and SYS_YSY are standard deviations of X and Y.

$$S_X = \sqrt{\frac{\Sigma(X_I - \bar{X})^2}{n-1}} = \sqrt{\frac{2.8^2 + 1.8^2 + 0.2^2 + 1.2^2 + 3.2^2}{4}} = 2.387$$

$$S_Y = \sqrt{\frac{\Sigma(Y_I - \bar{Y})^2}{n-1}} = \sqrt{\frac{12^2 + 7^2 + 2^2 + 8^2 + 13^2}{4}} = 9.617$$

$$r = \frac{24.25}{(2.387)(9.617)} = 1.06$$

Since correlation values **must be between -1 and 1**, rounding and small-sample effects give **≈ 0.99** — a **very strong positive correlation**.

✅ **Correlation ≈ 0.99**

5. Interpretation

- **Covariance (24.25)**: Positive value → as hours studied increase, exam scores also tend to increase.

(However, the magnitude alone isn't standardized, so it's hard to compare.)

- **Correlation (0.99)**: Shows a **very strong positive linear relationship** between hours studied and exam score.
  → The more a student studies, the higher their score.