Q1. What is a random variable in probability theory?

Ans- A **random variable** is a function that assigns a numerical value to each outcome of a random experiment. It maps elements of the sample space SSS to real numbers R\mathbb{R}R.

Mathematically:

X:S→RX: S \rightarrow \mathbb{R}X:S→R

where XXX is the random variable, SSS is the sample space, and R\mathbb{R}R is the set of real numbers.

Types of Random Variables:

1. **Discrete Random Variable**
    - Takes finite or countable values.
    - *Example:* Number of heads when tossing a coin three times.
      S={HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}S = \{\text{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}\}S={HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}
      Possible values of XXX: 0,1,2,30, 1, 2, 30,1,2,3.
2. **Continuous Random Variable**
    - Takes infinitely many values within an interval.
    - *Example:* Height of a student in centimeters (e.g., 150.2,160.5,172.0150.2, 160.5, 172.0150.2,160.5,172.0 cm).

Example:
If a die is rolled:

- Sample space: S={1,2,3,4,5,6}S = \{1, 2, 3, 4, 5, 6\}S={1,2,3,4,5,6}
- Random variable XXX = number shown on the die
- Values of XXX: 1,2,3,4,5,61, 2, 3, 4, 5, 61,2,3,4,5,6
- This is a **discrete random variable**.

Q2. What are the types of random variables?

Ans- A random variable is a numerical description of the outcome of a random experiment. It can be classified into two main types:

1. Discrete Random Variable
A discrete random variable takes a finite or countable number of distinct values. The outcomes can be listed or enumerated. These variables usually result from counting processes.
Example: When a coin is tossed three times and we count the number of heads, the possible values of X are 0, 1, 2, and 3. Another example is the number of students present in a classroom, which can only take whole number values.

2. Continuous Random Variable
A continuous random variable can take infinitely many values within a given range or interval. The outcomes cannot be counted individually, as they come from measurement processes.
Example: The height of a student may be 150.2 cm, 160.5 cm, 172.0 cm, etc. Another example is the time taken to complete a race, which can take any real value within a range.

Q3. Explain the difference between discrete and continuous distributions

Ans A probability distribution describes how probabilities are assigned to possible values of a random variable. Based on the type of random variable, distributions are classified into **discrete** and **continuous** distributions.

1. Discrete Probability Distribution

- Applies to **discrete random variables** that take a finite or countable set of values.
- The probability of each possible value is given by a **probability mass function (PMF)**.
- The sum of probabilities for all possible values is equal to 1.
- Example: Probability distribution of the number obtained when rolling a die.

2. Continuous Probability Distribution

- Applies to **continuous random variables** that can take infinitely many values within an interval.
- Probabilities are described using a **probability density function (PDF)**.
- The probability of any exact single value is zero; instead, probabilities are assigned over intervals.
- Example: Probability distribution of a person’s height in centimeters.

Q4. What is a binomial distribution, and how is it used in probability?

Ans A binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent trials of a binary (yes/no) experiment, where each trial has the same probability of success.

Conditions for a Binomial Distribution:

1. The experiment consists of n independent trials.
2. Each trial has only two outcomes: success or failure.
3. The probability of success ppp remains the same for each trial.
4. The random variable XXX represents the number of successes in nnn trials.

Probability Formula:

P(X=k)=(nk)pk(1−p)n−kP(X = k) = \binom{n}{k} p^k (1-p)^{n-k}P(X=k)=(kn)pk(1−p)n−k

where:

- nnn = number of trials
- kkk = number of successes
- ppp = probability of success
- 1−p1-p1−p = probability of failure
- (nk)=n!k!(n−k)!\binom{n}{k} = \frac{n!}{k!(n-k)!}(kn)=k!(n−k)!n! is the binomial coefficient

Example:
If a coin is tossed 5 times and the probability of getting a head is 0.5, the binomial distribution can be used to calculate the probability of getting exactly 3 heads.

Uses in Probability:

- Quality control (defective vs. non-defective items)
- Medical trials (patient responds to treatment or not)
- Survey results (favorable vs. unfavorable response)

Q5. What is the standard normal distribution, and why is it important?

Ans The standard normal distribution is a special case of the normal distribution in which the mean (μ) is 0 and the standard deviation (σ) is 1. It is also called the Z-distribution.

Key Features:

- Shape: Bell-shaped and symmetric about the mean 0.
- Mean: $\mu = 0$
- Standard Deviation: $\sigma = 1$
- Total Area under the Curve: 1
- Probability Representation: The horizontal axis represents Z-scores, which indicate how many standard deviations a value is from the mean.

Z-Score Formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X = raw score
- μ = mean
- σ = standard deviation

Importance:

1. Simplifies Calculations: Any normal distribution can be converted to the standard normal form using the Z-score formula, making probability calculations easier.
2. Probability Tables: Z-tables are widely available, allowing quick lookup of probabilities.
3. Statistical Inference: Used in hypothesis testing, confidence intervals, and control charts.
4. Universal Reference: Acts as a benchmark for comparing different datasets on the same scale.

Example:
If a test score is 85, with $\mu = 80$ and $\sigma = 5$, the Z-score is:

$$Z = \frac{85 - 80}{5} = 1$$

This means the score is 1 standard deviation above the mean.

Q6. What is the Central Limit Theorem (CLT), and why is it critical in statistics?

Ans The Central Limit Theorem (CLT) states that when independent random samples are drawn from any population with a finite mean and variance, the sampling distribution of the sample mean will approach a normal distribution as the sample size becomes large, regardless of the shape of the original population distribution.

Mathematical Form:
If X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn are independent and identically distributed random variables with mean μ\muμ and standard deviation σ\sigmaσ, then the distribution of the standardized sample mean:

Z=X¯−μσ/nZ = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}Z=σ/nX¯−μ

approaches a standard normal distribution as n→∞n \to \inftyn→∞.

Key Points:

1. Works for any population distribution (normal, skewed, uniform, etc.) when nnn is large.
2. Usually, n≥30n \geq 30n≥30 is considered sufficient for approximation.
3. The mean of the sampling distribution = μ\muμ (same as population mean).
4. The standard deviation of the sampling distribution = σ/n\sigma / \sqrt{n}σ/n (called the standard error).

Importance of CLT:

- Basis for Inferential Statistics: Allows us to use normal distribution tools (Z-tests, t-tests) for hypothesis testing.
- Confidence Intervals: Makes it possible to estimate population parameters using sample statistics.
- Real-world Applications: Quality control, polling, risk analysis, and scientific experiments.

Example:
If we take multiple samples of size 50 from a skewed population of customer ages and compute the mean for each sample, the distribution of those means will approximate a normal curve due to the CLT.

Q7. What is the significance of confidence intervals in statistical analysis?

Ans A confidence interval (CI) is a range of values, derived from sample data, that is likely to contain the true value of a population parameter (such as mean or proportion) with a certain level of confidence.

Key Points:

- Expressed as:

Estimate ± Margin of Error\text{Estimate} \ \pm \ \text{Margin of Error}Estimate ± Margin of Error

- Common confidence levels: 90%, 95%, 99%.
- A 95% CI means that if we repeated the sampling many times, about 95% of the calculated intervals would contain the true population parameter.

Significance in Statistical Analysis:

1. Provides Range, Not Just a Point Estimate: Gives a more informative measure than a single estimate by showing uncertainty.
2. Measures Reliability: Wider intervals indicate more uncertainty; narrower intervals indicate greater precision.
3. Supports Decision-Making: Helps determine if a parameter is within an acceptable or expected range.
4. Used in Hypothesis Testing: If a hypothesized value lies outside the CI, it is evidence against the null hypothesis.

Example:
If the average weight of a sample of 100 apples is 150g with a 95% CI of 150±2150 \pm 2150±2g, we can say with 95% confidence that the population mean weight is between 148g and 152g.

Q8. What is the concept of expected value in a probability distribution?

Ans The expected value (also called the mean or mathematical expectation) of a probability distribution is the long-run average value of a random variable over many repetitions of the experiment. It represents the theoretical central point of the distribution.

For a Discrete Random Variable:

E(X)=∑i=1nxi⋅P(xi)E(X) = \sum_{i=1}^{n} x_i \cdot P(x_i)E(X)=i=1∑nxi⋅P(xi)

where:

- xix_ixi = possible values of XXX
- P(xi)P(x_i)P(xi) = probability of xix_ixi

For a Continuous Random Variable:

E(X)=∫−∞∞x⋅f(x) dxE(X) = \int_{-\infty}^{\infty} x \cdot f(x) \, dxE(X)=∫−∞∞x⋅f(x)dx

where f(x)f(x)f(x) is the probability density function.

Key Points:

- Acts as a weighted average of all possible values.
- Probabilities serve as the weights.
- May not be an actual possible value of the random variable.

Example (Discrete):
A fair die has outcomes 1,2,3,4,5,61, 2, 3, 4, 5, 61,2,3,4,5,6 each with probability 16\frac{1}{6}61:

E(X)=1+2+3+4+5+66=3.5E(X) = \frac{1+2+3+4+5+6}{6} = 3.5E(X)=61+2+3+4+5+6=3.5

This means that in the long run, the average outcome of rolling the die is 3.5.

Example (Continuous):
If the time to complete a task follows a continuous distribution with known f(x)f(x)f(x), the expected value gives the average completion time.

Q9. Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution.

Ans. import numpy as np

import matplotlib.pyplot as plt

Parameters

mean = 50

std_dev = 5

size = 1000

Generate random numbers from normal distribution

data = np.random.normal(mean, std_dev, size)

Compute mean and standard deviation

calculated_mean = np.mean(data)

calculated_std_dev = np.std(data)

print(f"Calculated Mean: {calculated_mean}")

print(f"Calculated Standard Deviation: {calculated_std_dev}")

Plot histogram

plt.hist(data, bins=30, color='skyblue', edgecolor='black')

plt.title('Normal Distribution (Mean = 50, Std Dev = 5)')

plt.xlabel('Value')

plt.ylabel('Frequency')

plt.grid(True, linestyle='--', alpha=0.7)

plt.show()

The output for the generated data is:

- Calculated Mean: ≈ 49.95
- Calculated Standard Deviation: ≈ 4.88

Q10. You are working as a data analyst for a retail company. The company has collected daily sales data for 2 years and wants you to identify the overall sales trend. daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255, 235, 260, 245, 250, 225, 270, 265, 255, 250, 260] ● Explain how you would apply the Central Limit Theorem to estimate the average sales with a 95% confidence interval. ● Write the Python code to compute the mean sales and its confidence interval.

Ans 1. Applying the Central Limit Theorem (CLT):

- The CLT states that if we repeatedly take random samples from a population and compute their means, the sampling distribution of the mean will tend to follow a normal distribution as the sample size increases, regardless of the original population shape.

- For our sales data, we can treat these 20 values as a sample from the entire population of possible daily sales.

- We will compute the sample mean and standard error:
  SE=σnSE = \frac{\sigma}{\sqrt{n}}SE=nσ
  where σ\sigmaσ is the sample standard deviation and nnn is the number of observations.

- For a 95% confidence interval, we will use the Z-score 1.96 (for large samples) or the t-score for smaller samples.

- The CI formula is:
  CI=X¯±t×SE\text{CI} = \bar{X} \pm t \times SECI=X¯±t×SE

Code:

```
import numpy as np

import scipy.stats as stats


# Given data

daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,

               235, 260, 245, 250, 225, 270, 265, 255, 250, 260]


# Calculate sample statistics

mean_sales = np.mean(daily_sales)

std_dev_sales = np.std(daily_sales, ddof=1)  # sample standard deviation

n = len(daily_sales)


# Standard error
```

```python
SE = std_dev_sales / np.sqrt(n)

# t-critical value for 95% CI (two-tailed, df = n-1)
t_critical = stats.t.ppf(0.975, df=n-1)

# Confidence interval
margin_of_error = t_critical * SE
ci_lower = mean_sales - margin_of_error
ci_upper = mean_sales + margin_of_error

# Output
print(f"Mean Sales: {mean_sales:.2f}")
print(f"95% Confidence Interval: ({ci_lower:.2f}, {ci_upper:.2f})")
```