Question 1: What is a random variable in probability theory?
Answer: In probability theory, a random variable is a variable that assigns a numerical value to each possible outcome of a random experiment.

Question 2: What are the types of random variables?
Answer:Types of Random Variables are :

1. Discrete Random Variable

- Takes on a finite or countable set of possible values.

- Example:
  - Number of heads when tossing 3 coins → {0,1,2,3}\{0, 1, 2, 3\}{0,1,2,3}
  - Roll of a die → {1,2,3,4,5,6}\{1, 2, 3, 4, 5, 6\}{1,2,3,4,5,6}

2. Continuous Random Variable

- Can take any value within an interval or collection of intervals on the real number line.

- Example:
  - Height of a person (e.g., 165.2165.2165.2 cm, 170.45170.45170.45 cm)
  - Time taken for a bus to arrive (e.g., 4.324.324.32 minutes)

3. Mixed Random Variable

- Has both discrete and continuous components.

- Example:
  - Time until the next train: could be exactly at a scheduled time (discrete) or somewhere in between due to delays (continuous).

Question 3: Explain the difference between discrete and continuous distributions.
Answer:

Feature	Discrete Distribution	Continuous Distribution
Values	Countable (finite or countably infinite)	Uncountably infinite (any real number in a range)
Function used	PMF (Probability Mass Function)	PDF (Probability Density Function)
Probability of a point	Can be non-zero	Always zero
Example	Rolling a die, number of emails received	Height, weight, time, temperature

Question 4: What is a binomial distribution, and how is it used in probability?

Answer:A binomial distribution is a probability distribution that models the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes — success or failure — and the probability of success is the same for each trial.

Binomial Distribution used in probability in certain steps :

Step 1: Calculating Exact Probabilities

If you know n (number of trials), p (probability of success), and k (desired number of successes), you can find:

$P(X=k)=(n/k)p^k(1-p)^n-k$Example:

Probability of getting exactly 7 heads in 10 fair coin tosses.

Step 2: Finding Cumulative Probabilities

Sometimes you need:

a)Probability of at most k successes: $P(X \le k)$

b)Probability of at least k successes: $P(X \ge k)$

You add the probabilities for the relevant k values.

Example: Probability of getting at least 3 customers who buy a product out of 5 who walk in.

Step 3: Predicting Likely Outcomes

The binomial distribution tells us which outcomes are most probable.

Example: In a survey where 60% say “Yes,” the distribution shows that the most likely number of “Yes” answers in 20 people is around $20\times0.6=12$

Step 4: Approximations & Decision-Making

a)For large n, the binomial can be approximated by the normal distribution (Central Limit Theorem) or Poisson distribution (if p is small).

b)Helps in hypothesis testing, risk assessment, and quality control.

Question 5: What is the standard normal distribution, and why is it important?

Answer:The standard normal distribution is a special case of the normal (Gaussian) distribution where:

- Mean (μ) = 0
- Standard deviation (σ) = 1
- It’s symmetric, bell-shaped, and centered at 0.

The variable that follows it is usually called Z and is said to have a Z-distribution.

It’s important for various things like :

1)Basis for Z-Scores

Any normal distribution can be converted to the standard normal distribution using:

$$x=(x-\mu)/\sigma$$

This tells you how many standard deviations a value is from the mean.

2)Universal Reference Table
Once values are converted to Z-scores, you can use the standard normal table to find probabilities without creating a new table for every possible mean and standard deviation.

3)Simplifies Probability Calculations
Instead of integrating different normal curves, we convert them to the standard one and look up values in Z-tables or use software.

4)Foundation for Statistical Tests
Many inferential statistics methods (like hypothesis tests and confidence intervals) rely on the standard normal distribution.

5)Real-Life Applications

- Quality control
- Standardized testing (SAT, IQ scores)
- Finance (risk modeling, stock returns)
- Natural and social sciences (measurement errors, biological data)

Question 6: What is the Central Limit Theorem (CLT), and why is it critical in statistics?

Answer:The Central Limit Theorem (CLT) is one of the most important results in statistics because it explains *why* the normal distribution appears so often in real-world data analysis.

It's critical in statistics because :
1)Foundation for Inferential Statistics
CLT allows us to make probability-based conclusions about population parameters (mean, proportion) even if the population isn't normally distributed.

2)Enables Hypothesis Testing & Confidence Intervals
Because sample means are (approximately) normal, we can use Z-tests, t-tests, and build confidence intervals.

Question 7: What is the significance of confidence intervals in statistical analysis?

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

Significance of Confidence Intervals in statistical analysis are:

1. Gives a Range, Not Just a Point

- A single estimate (like a sample mean) doesn't tell you how precise it is.
- A CI provides a range of plausible values for the parameter, showing both the estimate and its uncertainty.

2. Quantifies Uncertainty

- The width of a CI reflects the uncertainty in your estimate.

- Narrow interval → more precise estimate.
- Wide interval → less precise, more uncertainty.

- Influenced by:
  - Sample size (larger samples → narrower CI)
  - Variability in data
  - Confidence level chosen (e.g., 95%, 99%)

3. Interpretable Probabilistically

- A 95% CI means: If we took many random samples and built a CI for each, about 95% of those intervals would contain the true parameter.

4. Supports Decision-Making

- In research, business, and science, CIs show the reliability of estimates.
- Example: If a drug increases recovery rate by [4%, 10%], the CI shows the possible real effect range — helping decide if the drug is effective.

5. Used in Hypothesis Testing

- If a 95% CI for a mean difference does not contain 0, it suggests the difference is statistically significant at the 5% level.

Question 8: What is the concept of expected value in a probability distribution?
Answer:The expected value (also called the mean or expectation) of a probability distribution is the long-run average of a random variable's outcomes, weighted by their probabilities.
It's essentially what you'd expect to get on average if you repeated a random experiment many, many times.
Example :You roll a fair six-sided die:

- Possible values: 1,2,3,4,5,6
- Probability of each: 1/6

E[X]=(1)(1/6)+(2)(1/6)+⋯+(6)(1/6)=(1+2+3+4+5+6)/6 = 3.5
This means if you roll the die many times, the average roll will approach 3.5.

Question 9: 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.
Answer: Python Code

```
import numpy as np
import matplotlib.pyplot as plt

# Set seed for reproducibility (optional)
np.random.seed(42)

# Generate 1000 random numbers from a normal distribution
mean = 50
std_dev = 5
```

```
size = 1000
data = np.random.normal(mean, std_dev, size)

# Compute mean and standard deviation
computed_mean = np.mean(data)
computed_std = np.std(data)

# Display results
print(f"Computed Mean: {computed_mean:.2f}")
print(f"Computed Standard Deviation: {computed_std:.2f}")

# Plot histogram
plt.hist(data, bins=30, edgecolor='black', alpha=0.7)
plt.title("Normal Distribution Histogram")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.axvline(computed_mean, color='red', linestyle='dashed', linewidth=2, label=f"Mean =
{computed_mean:.2f}")
plt.legend()
plt.show()
```

Question 10: 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.

Answer: Steps to apply CLT Theorem are ;

1. Restating the Problem

We have 20 daily sales figures (sample) and want to estimate the true average daily sales for the full 2 years of data. Since we don’t have all 730 days, we use this sample to make an inference.

2. Applying the Central Limit Theorem

- The sample size n=20 is less than 30, so instead of the Z-distribution, we’ll use the t-distribution (which accounts for extra uncertainty in small samples).
- The sample mean will be our best estimate of the population mean.
- The spread of sample means is measured by the Standard Error (SE):
  $SE=s/\sqrt{n}$
- where s is the sample standard deviation.

3. Constructing a 95% Confidence Interval

For a 95% confidence level:

- We find the t-critical value from the t-distribution with n−1n-1n−1 degrees of freedom.

- The formula for the CI is: $CI = \bar{x} \pm t_{\alpha/2,\, df=n-1} \times SE$,Where:
- $\bar{x}$ = sample mean
- $t_{\alpha/2,df}$ = t-value for given confidence level
- SE = standard error

4. Interpretation
The resulting CI will give us a range of values that we are 95% confident contains the true average daily sales for the entire 2-year period.
For example:
If we get a CI of (240,255), it means that based on this sample, there is a 95% probability that the real average sales per day lies between 240 and 255 units.

Python Code for computing the mean sales and confidence interval :

```
import numpy as np
from scipy import stats

# Daily sales data
daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
               235, 260, 245, 250, 225, 270, 265, 255, 250, 260]

# Convert to NumPy array
sales_array = np.array(daily_sales)

# Sample size, mean, and standard deviation
n = len(sales_array)
mean_sales = np.mean(sales_array)
std_sales = np.std(sales_array, ddof=1)  # ddof=1 for sample std deviation

# Standard Error
SE = std_sales / np.sqrt(n)

# t-critical value for 95% confidence interval (since n < 30)
t_crit = stats.t.ppf(1 - 0.025, df=n-1)

# Confidence Interval
lower_bound = mean_sales - t_crit * SE
upper_bound = mean_sales + t_crit * SE

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