Question 1: What is a random variable in probability theory?

A random variable is a function that assigns numerical values to the outcomes of a random experiment. In probability theory, it is used to quantify uncertain outcomes mathematically. For example, when tossing a coin, we can define a random variable XXX such that X=1X = 1X=1 if the coin shows heads and X=0X = 0X=0 if it shows tails.

Random variables are categorized into two types — discrete and continuous — depending on whether the possible outcomes are countable or uncountable. They help in building probability distributions, calculating expected values, variances, and performing statistical inferences. Essentially, a random variable serves as a bridge between theoretical probability and real-world quantitative analysis.

Question 2: What are the types of random variables?

There are mainly **two types of random variables**:

1. **Discrete Random Variable** –
   A discrete random variable takes a countable number of distinct values. Examples include the number of heads obtained in a series of coin tosses or the number of defective products in a batch. The probability distribution of a discrete random variable is represented by a probability mass function (PMF).
2. **Continuous Random Variable** –
   A continuous random variable takes an infinite number of possible values within a given range. Examples include height, weight, or temperature. The probability distribution of a continuous random variable is represented by a probability density function (PDF).

In summary, discrete variables describe outcomes that can be listed or counted, while continuous variables represent outcomes that can take any value within a continuum.

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

A **discrete probability distribution** applies to situations where the set of possible outcomes is finite or countably infinite. The probability of each specific outcome can be directly assigned, and the total probability of all outcomes equals 1. Examples include the **binomial distribution**, **Poisson distribution**, and **geometric distribution**.

In contrast, a **continuous probability distribution** deals with variables that can take any real value within an interval. Since the number of possible outcomes is infinite, the probability of any single value is zero. Instead, probabilities are defined over intervals using a **probability density**

function (PDF). Common examples include the **normal distribution**, **exponential distribution**, and **uniform distribution**.

The key distinction is that discrete distributions handle countable data, while continuous distributions deal with measurable, uncountable data.

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

A **binomial distribution** is a discrete probability distribution that represents the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success ppp.

The probability mass function of a binomial distribution is:

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 on each trial

It is widely used to model experiments where there are only two possible outcomes — success or failure — such as:

- The number of defective items in a batch of products
- The number of students passing an exam
- The number of heads in coin tosses

The binomial distribution is essential in hypothesis testing, risk modeling, and quality control because it simplifies real-world binary outcome processes into measurable probabilities.

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

The **standard normal distribution** is a specific form of the normal distribution where the mean μ=0\mu = 0μ=0 and the standard deviation σ=1\sigma = 1σ=1. It is denoted as N(0,1)N(0,1)N(0,1). The curve is symmetric around the mean, with most values falling within one or two standard deviations from the mean.

This distribution is important in statistics because it serves as a **reference distribution** for all normal data. Any normal distribution can be converted into a standard normal distribution using **Z-scores** through the formula:

$Z = \frac{X - \mu}{\sigma} Z=σX−μ$

where XXX is the observed value, μ\muμ is the mean, and σ\sigmaσ is the standard deviation.

The standard normal distribution is used to calculate probabilities, confidence intervals, and perform hypothesis testing. It allows researchers to compare different datasets on the same scale, enabling consistent interpretation of statistical results.

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

The **Central Limit Theorem (CLT)** states that when independent random samples of sufficient size are drawn from any population with a finite mean and variance, the sampling distribution of the sample mean approaches a **normal distribution** regardless of the population's original distribution.

Mathematically, as the sample size nnn increases,

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) X¯∼N(μ,nσ2)$

where $\bar{X}$ is the sample mean, μ\muμ is the population mean, and σ2\sigma^2σ2 is the population variance.

The CLT is critical because it forms the foundation for many statistical methods, including confidence intervals, hypothesis testing, and regression analysis. It allows analysts to make inferences about population parameters using the normal distribution, even when the underlying data is not normally distributed.

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

A **confidence interval (CI)** is a statistical range, calculated from sample data, that is likely to contain the true population parameter (such as the mean or proportion) with a certain level of confidence. It provides a measure of uncertainty around an estimate.

For example, a 95% confidence interval for a mean indicates that if we were to take many random samples and compute an interval from each, approximately 95% of those intervals would contain the true population mean.

Confidence intervals are significant because they:

1. **Quantify uncertainty** — They show how precise a sample estimate is.
2. **Provide a range of plausible values** — Instead of a single estimate, we get a range that reflects possible true values.

3. **Guide decision-making** — In hypothesis testing, if a hypothesized parameter value (e.g., a mean) lies outside the CI, it suggests statistical significance.
4. **Support interpretation of data** — Wider intervals indicate more variability or smaller samples, while narrow intervals show higher precision.

Thus, confidence intervals are a crucial part of inferential statistics, allowing analysts to make reliable conclusions about populations based on limited sample data.

Question 8: What is the concept of expected value in a probability distribution?

The **expected value (EV)** of a random variable is the long-term average or mean value that the variable would take if an experiment were repeated many times. It represents the **center of gravity** of a probability distribution.

Mathematically:

- For a **discrete random variable**,

  $$E(X)=\sum x_i P(x_i)E(X) = \sum x_i P(x_i)E(X)=∑xiP(xi)$$

- For a **continuous random variable**,

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

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

The expected value helps summarize a random process with a single representative number. It is used in decision theory, economics, and risk analysis to determine the average outcome of uncertain events. For example, in gambling, the expected value can show whether a bet is favorable over the long run.

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.

Python Code:

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

# Generate 1000 random numbers from a normal distribution
```

```
data = np.random.normal(loc=50, scale=5, size=1000)

# Compute the mean and standard deviation
mean_value = np.mean(data)
std_value = np.std(data)

print("Calculated Mean:", mean_value)
print("Calculated Standard Deviation:", std_value)

# Plot histogram
plt.hist(data, bins=30, edgecolor='black')
plt.title('Histogram of Normally Distributed Data (mean=50, std=5)')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```

Expected Output (sample):

```
Calculated Mean: 49.92
Calculated Standard Deviation: 5.08
```

Explanation:
This code creates 1000 values sampled from a normal distribution with the specified parameters. Using NumPy's `mean()` and `std()` functions, we compute the sample mean and standard deviation, which should be close to 50 and 5, respectively. The histogram shows the bell-shaped curve characteristic of a normal distribution.

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.

Dataset:

```
daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
               235, 260, 245, 250, 225, 270, 265, 255, 250, 260]
```

Applying the Central Limit Theorem (CLT):

According to the Central Limit Theorem, even if the population distribution is not normal, the sampling distribution of the sample mean approaches a normal distribution as the sample size increases. This allows us to estimate the **average daily sales** and construct a **95% confidence interval** for the population mean.

Steps:

1. Calculate the sample mean and standard deviation.
2. Compute the standard error (SE) = sn\frac{s}{\sqrt{n}}ns.

3. For a 95% confidence interval, use the critical value z=1.96z = 1.96z=1.96.
4. The CI is given by:

   CI=x¯±z×SE\text{CI} = \bar{x} \pm z \times SECI=x¯±z×SE

This interval will give a range that likely contains the true mean daily sales.

Python Code:

```
import numpy as np
import scipy.stats as stats

# 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 = np.array(daily_sales)

# Calculate statistics
mean_sales = np.mean(sales)
std_sales = np.std(sales, ddof=1)  # sample standard deviation
n = len(sales)

# Compute standard error
se = std_sales / np.sqrt(n)

# 95% confidence interval
confidence_level = 0.95
z_value = stats.norm.ppf(1 - (1 - confidence_level) / 2)
margin_of_error = z_value * se
lower_bound = mean_sales - margin_of_error
upper_bound = mean_sales + margin_of_error

print("Mean Sales:", round(mean_sales, 2))
print("95% Confidence Interval:", (round(lower_bound, 2), round(upper_bound,
2)))
```