Statistic Advance

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

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 an experiment .Two main types:

1. Discrete random variable:
   Takes on a countable set of possible values (e.g., the number of heads in 3 coin flips).

2. Continuous random variable:
   Takes on values in an interval or continuous range (e.g., the time it takes for a bus to arrive).

Example:

If you roll a fair six-sided die. Possible outcomes {1,2,3,4,5,6} Here, the random variable x =" the number that appears on the die". So, x can take any value from 1 to 6.

Question 2. What are the types of random variables?

Answer → There are main two types of random variables.

1. Discrete Random Variable :

A discrete random variable can take a countable number of distinct values. It usually arises from counting outcomes.

Ex : - Number of heads in 3 coin tosses → {0, 1, 2, 3} , Number of students present in a class → {0, 1,2,...., n}

Probability distribution : Each value of the variable has a certain probability.

2. Continuous Random Variable :

A continuous random variable can take infinitely many values within a range or interval. It usually arises from measurement rather than counting.
Ex : - Height of students in class (e.g., 150.2cm , 151.7cm etc.) , Time taken to complete a task etc.
Probability Density Function : For continuous random variables , we use a curve instead of a table. The area under the curve between two points givens the probability.

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

Answer →

	Discrete	**Continuous**
Possible Values	Countable (finite or countably infinite)	Uncountable(intervals of real numbers)
Type of random variable	Discrete (countable)	Continuous (uncountable)

Example	Number of coin tosses showing heads	Time taken to finish a race
Probability Representation	Probability Mass Function (PMF)	Probability Density Function (PDF)
Probability at a point	P(X=x)P(X = x)P(X=x) can be > 0	P(X=x)=0P(X = x) = 0P(X=x)=0
Probability function	PMF (Probability Mass Function)	PDF (Probability Density Function)
Cumulative probability	P(X≤x)=∑P(X=xi)P(X \le x) = \sum P(X = x_i)P(X≤x)=∑P(X=xi)	P(X≤x)=∫−∞xf(t) dtP(X \le x) = \int_{-\infty}^x f(t)\,dtP(X≤x)=∫−∞xf(t)dt

Examples	Binomial, Poisson, Geometric	Normal, Exponential, Uniform

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

Answer → The binomial distribution is one of the most important discrete probability distributions in statistics and probability theory. It models the number of successes in a fixed number of independent Bernoulli trials, where each trial has only two possible outcomes: success or failure.

Formula : P(X=x)=(xn)px(1−p)n−x,for x=0,1,2,…,n

Mean and Variance

- Mean (Expected Value) : np
- Variance: np(1−p)

```
from scipy.stats import binom

n = 5
p = 0.5
k = 3

prob = binom.pmf( k, n, p)
print( f"P( X = 3 ) = {prob: .4f}")

mean, var = binom.stats( n, p)
print( f"Mean = {mean} , Variance = {var}")
```

Output :-
```
P( X = 3 ) =  0.3125
Mean = 2.5 , Variance = 1.25
```

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 distribution that has:

Mean (μ) = 0
Standard deviation (σ) = 1
It is denoted as $Z \sim N(0, 1)$, and its probability density function (PDF) is .

Importance of the Standard Normal Distribution:-

1. Simplifies Calculations:
   Many statistical methods and tests rely on the normal distribution.
   By converting any normal variable X to a standard normal variable Z using
   $Z=X-\mu\sigma$, $Z=\sigma X-\mu$,
   we can use standard normal tables (Z-tables) to find probabilities easily.

2. Foundation for Statistical Inference:
   Many inferential statistics methods (like confidence intervals and hypothesis tests) use the Z-distribution as a basis, especially when population parameters are known.

3. Central Role in Probability Theory:
   The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases—often approximated using the standard normal.

4. Universal Benchmark:
   The Z-score (standard score) tells how many standard deviations a value is from the mean, making it useful for comparing data across different scales.

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

Answer → When we take many random samples of a sufficiently large size (n) from any population with a finite mean (μ) and variance (σ^2), the sampling distribution of the sample mean will approach a normal distribution, regardless of the shape of the original population.

Why the CLT Is Critical in Statistics :-

1. Foundation for Inference:
   It allows us to make inferences about population parameters using sample data.
   Many tests (e.g., Z-test, t-test, confidence intervals) rely on the assumption that the sampling distribution of the mean is normal.

2. Simplifies Probability Calculations:
   Even if the population distribution is unknown or non-normal, we can use normal probability models for sample means.

3. Supports Real-World Decision Making:
   In practice, populations are rarely perfectly normal, but thanks to the CLT, we can still use normal-based methods for large samples.

4. Enables Standardization:
   Through the CLT, we can convert sample means to Z-scores and use standard normal tables.

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

Answer →

A confidence interval 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.

where:

- X^{-} = sample mean
- $Z\alpha$ = critical value from the standard normal distribution
- σ = population standard deviation
- n = sample size

Significance in Statistical Analysis :-

1. Quantifies Uncertainty:
   Confidence intervals provide a range instead of a single estimate, showing the precision of the estimate.

2. Informs Decision-Making:
   Wider intervals indicate more uncertainty; narrower intervals mean more precise estimates.
   They help in determining whether estimates are statistically significant.

3. Alternative to Hypothesis Testing:
   Confidence intervals give more information than a simple “reject” or “fail to reject” outcome in hypothesis tests.
   They show both the direction and magnitude of an effect.

4. Used Across Many Fields:
   Common in research, economics, medicine, and social sciences to express reliability of estimated parameters.

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

Answer → The expected value (EV) of a probability distribution represents the long-run average or mean outcome of a random variable if an experiment is repeated many times. It tells us what value we can *expect* on average.

>>For a discrete random variable (X) with possible values x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn and corresponding probabilities P(x1),P(x2),...,P(xn)P(x_1), P(x_2), ..., P(x_n)P(x1),P(x2),...,P(xn):

E(X)=∑i=1nxi P(xi)

>>For a continuous random variable, the expected value is defined as:

E(X)=∫−∞∞x f(x) dx

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

>> The expected value is the weighted average of all possible outcomes, where each outcome is weighted by its probability.

It doesn’t necessarily have to be a value the variable can actually take — it’s a theoretical mean.

Example :

Suppose a fair six-sided die is rolled.

E(X)=1(1/6)+2(1/6)+3(1/6)+4(1/6)+5(1/6)+6(1/6)

E(X)=3.5

So, the expected value of a fair die roll is 3.5 — not a possible outcome, but the average result over many rolls.

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.

(Include your Python code and output in the code box below.)

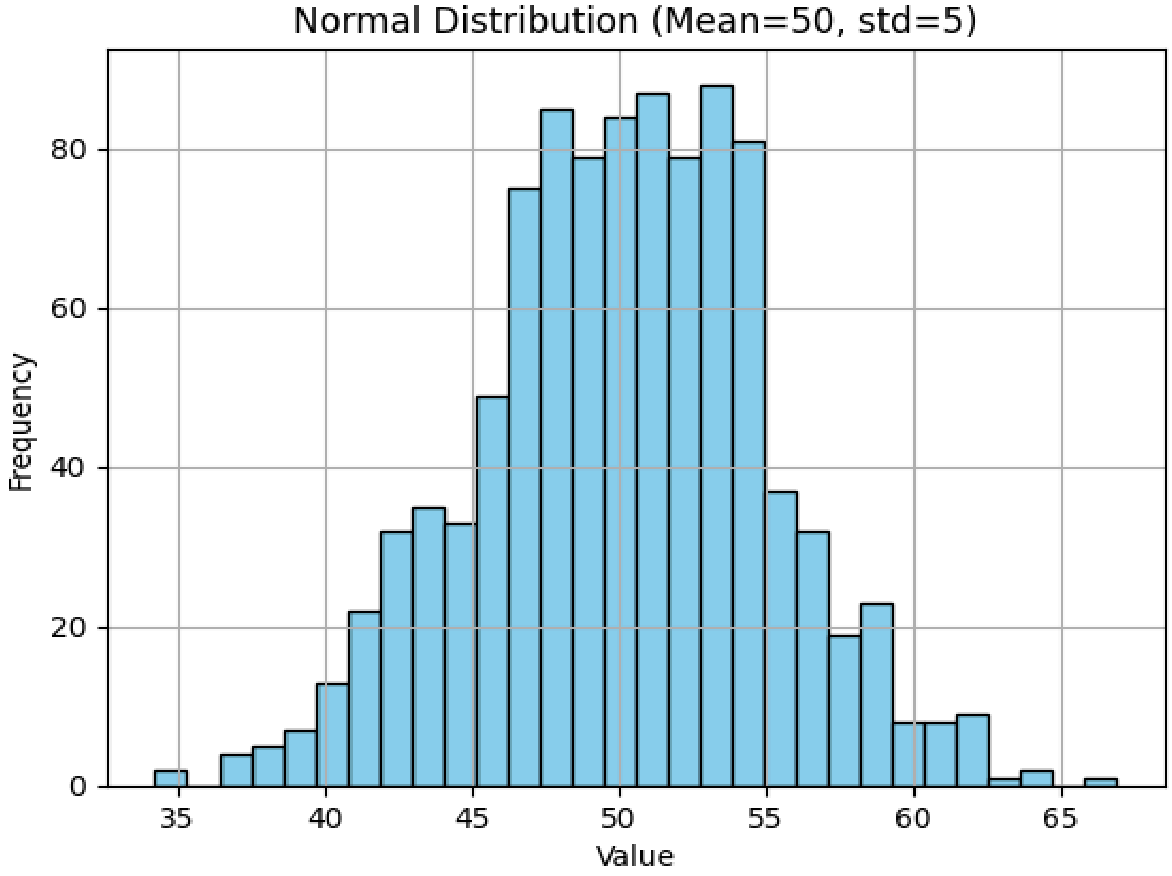
Answer → import numpy as np

```
import matplotlib.pyplot as plt


mean = 50

std_dev = 5

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


calculated_mean = np.mean(data)

calculated_std = np.std(data)


print("Calculated Mean :", round(calculated_mean, 2))

print("Calculated Standard Deviation :", round (calculated_std, 2))
```

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

plt.title("Normal Distribution (Mean=50, std=5)")

plt.xlabel("Value")

plt.ylabel("Frequency")

plt.grid(True)

plt.show()
```

Output :-

```
Calculated Mean : 50.11

Calculated Standard Deviation : 4.94
```



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. (Include your Python code and output in the code box below.)

Answer →

```
import numpy as np
from scipy import stats

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

mean_sales = np.mean(daily_sales)
std_sales = np.std(daily_sales, ddof=1)  # Sample standard deviation
n = len(daily_sales)

confidence_level = 0.95
alpha = 1 - confidence_level
z_score = stats.norm.ppf(1 - alpha/2)

margin_of_error = z_score * (std_sales / np.sqrt(n))
confidence_interval = (mean_sales - margin_of_error, mean_sales + 
margin_of_error)

print("Sample Mean:", round(mean_sales, 2))
print("Sample Standard Deviation:", round(std_sales, 2))
print("95% Confidence Interval:", (round(confidence_interval[0], 2), 
round(confidence_interval[1], 2)))
```

Output :-

```
Sample Mean: 248.25

Sample Standard Deviation: 17.27

95% Confidence Interval: (np.float64(240.68), np.float64(255.82))
```