Statistics Advanced - 1| **Assignment**

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

Answer:

A **random variable** is a variable that represents the **numerical outcome** of a random experiment.

It is like a function that assign a real number to each possible outcome of a random experiment.

Example :
Toss a fair coin : outcome : head : assign >> 1
Outcome : tail : assign >> 0
Notation :

Example : P(x = head) it means probability that x equal to head

Question 2: What are the types of random variables?

Answer:

Type of Random variables :
1 : Discreate Random Variable :- Take countable values (1,2,3,4,5 etc.)
2 : Continuous Random Variable :- Number in the range (height of student in the class)

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

Answer:

Discrete Distributions :

A discrete distribution describes the probabilities of a discrete random variable — one that takes countable distinct values.

Possible values : countable
Each possible value has a specific probability
Sum of probabilities= 1

Example : Tossing a coin

Possbile outcome (head , tail)
P(Head) = ½
P(Tail) = ½

Sum of probabilities = P(Head) + P(Tail)
= ½ + ½
= 1

Continuous Distributions :

Continuous Distributions that show the probabilities of continuous random variable.

Possible value : infinite
Probability is found using area under the curve

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

Answer:

A binomial distribution is the one of the most common discreate probability distribution
A random variable follows Binomial distribution if
There is n independent trail
Each trail has only possible outcomes (success or failure)
Probability of successs in each trail is always same
Sum of probability of both trail is 1
P success and q failure
Q = 1-p

Formula :

$P(x = k) = {}^{n}C_{k}(p^{k}\ (1-p)^{n-k}\ )$

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’s a bell-shaped, symmetric probability distribution centered at 0.

Importance of Normal distribution :

➔ **Standardization**

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

➔ **Simplifies Probability Calculation**

- Instead of calculating probabilities for every different normal curve, we **convert all data** to the standard normal form (Z-scores) and use a single **Z-table**.
- This makes it fast and consistent.

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

Answer:

If you take many random samples from any population with a finite mean and variance, the distribution of the sample means will become approximately normal (bell-shaped) as the sample size grows — even if the original population is not normal.

when $n \geq 30$ (sample size).
Mean of sample means = population mean.
Standard deviation of sample means (standard error) = $\sigma/n^{1/2}$.

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

Answer:

The **significance of confidence intervals (CIs)** is that they give us a **range of values** that is likely to contain the true population parameter (like mean or proportion) instead of just a single estimate.

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

Answer:

The expected value in a probability distribution is the long-term average outcome you would expect if you repeated a random process many times.
It's like the center of gravity of the distribution — a weighted average where each possible value is weighted by its probability.

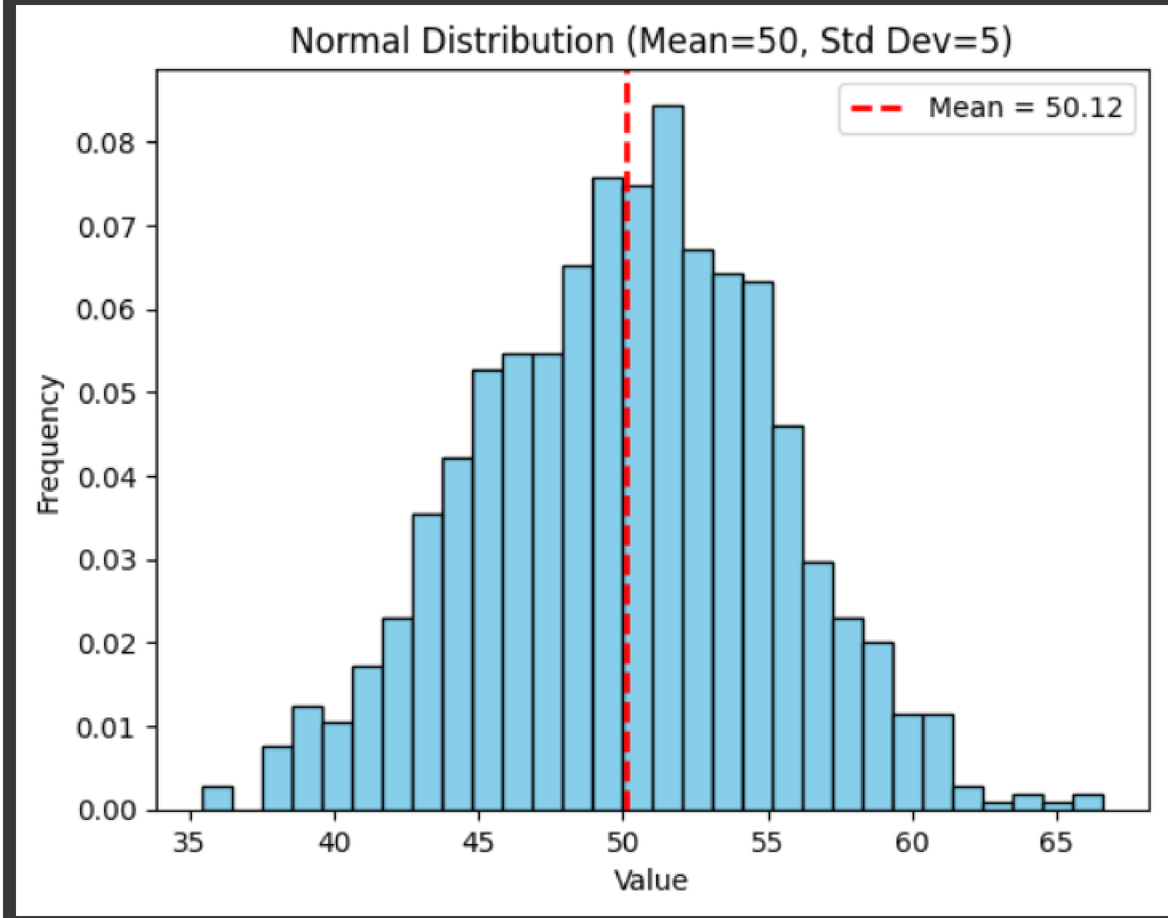
Question 9: Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution.

Answer:

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

# Parameters
mean = 50
std_dev = 5
n_samples = 1000

# Generate 1000 random numbers from normal distribution
data = np.random.normal(mean, std_dev, n_samples)

# Compute mean and standard deviation
calculated_mean = np.mean(data)
calculated_std_dev = np.std(data)

# Print results
print(f"Calculated Mean: {calculated_mean:.2f}")
print(f"Calculated Standard Deviation: {calculated_std_dev:.2f}")

# Plot histogram
plt.hist(data, bins=30, color='skyblue', edgecolor='black', density=True)
plt.title("Normal Distribution (Mean=50, Std Dev=5)")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.axvline(calculated_mean, color='red', linestyle='dashed', linewidth=2,
            label=f'Mean = {calculated_mean:.2f}')
plt.legend()
plt.show()
```

Output :

```
Calculated Mean: 50.12
Calculated Standard Deviation: 5.15
```



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.*)

CLT use: Even with just 20 days of sales data, CLT lets us estimate the **population mean** using the sample mean. Since n<30n < 30n<30, we use the **t-distribution** to create a 95% confidence interval.

Formula:

$$CI = \bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

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

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

n = len(sales)
mean = np.mean(sales)
std = np.std(sales, ddof=1)
SE = std / np.sqrt(n)
t_val = stats.t.ppf(0.975, df=n-1)

CI = (mean - t_val*SE, mean + t_val*SE)

print(f"Mean: {mean:.2f}")
print(f"95% CI: ({CI[0]:.2f}, {CI[1]:.2f})")
```

```
Mean: 248.25
95% CI: (240.17, 256.33)
```