Statistics Advanced - 1| Assignment

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

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

It is not random itself—its value depends on the result of the experiment.

Formally: It's a function that maps each possible outcome of a sample space to a real number.

Example:

Toss a coin → Assign 1 for heads, 0 for tails → The variable X X is the random variable.

Question 2: What are the types of random variables?

Two main types:

Discrete random variable – Takes countable values (e.g., number of heads in 5 coin tosses).

Continuous random variable – Takes uncountably infinite values within an interval (e.g., height, weight, time).

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

Feature	**Discrete Distribution**	**Continuous Distribution**
Possible values	Countable	Infinite, uncountable
Probability representation	Probability mass function (PMF)	Probability density function (PDF)
Example	Binomial, Poisson	Normal, Exponential
Probability of a specific value	P(X=x)>0P(X = x) > 0 possible	P(X=x)=0P(X = x) = 0, probability is in intervals

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

The binomial distribution models the probability of getting exactly k k successes in n n independent Bernoulli trials (success/failure), where each trial has a success probability p p.

Used in:

Quality control

Success/failure experiments

Survey results

Example: Probability of getting exactly 3 heads in 5 coin tosses.

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

Standard normal distribution: A normal distribution with mean $\mu=0$ and standard deviation $\sigma=1$.

Denoted as $Z \sim N(0,1)$

Importance:

- Used in z-scores to compare data from different normal distributions.
- Many statistical tests rely on standard normal values.

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

CLT: For a large sample size n , the sampling distribution of the sample mean approaches a normal distribution, regardless of the original population's distribution, as long as the population has a finite variance.

Importance:

Allows using normal distribution–based methods even for non-normal data.

Forms the basis of confidence intervals and hypothesis tests.

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

A confidence interval (CI) is a range of values, derived from sample data, that is likely to contain the true population parameter with a given confidence level (e.g., 95%).

Significance:

Quantifies uncertainty in estimates.

Gives a range instead of a single number, making conclusions more reliable.

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

Expected value (EV): The long-run average value of a random variable over many repetitions of the experiment.

Formula for discrete variable:

$$E[X]=\sum xi \cdot P(xi)$$

Formula for continuous variable:

$$E[X] = \int_{\{-\infty\}}^{\{\infty\}} x \cdot f(x), dx$$

Measures the "center" or "average" of the distribution.

Used in decision-making, economics, and risk assessment.

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.)

CODE –

```
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 = np.std(data)


print("Calculated Mean:", calculated_mean)

print("Calculated Standard Deviation:", calculated_std)


# Plot histogram

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

plt.title('Histogram of Normally Distributed Data')

plt.xlabel('Value')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()
```

Output –

Calculated Mean: 50.12

Calculated Standard Deviation: 4.97

The histogram will look like a bell-shaped curve centered around 50 with a spread of about 5.

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.)

CODE –

```
import numpy as np

import scipy.stats as 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)

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

# Z-score for 95% confidence
z_score = 1.96

# Margin of error
margin_error = z_score * (std_sales / np.sqrt(n))

# Confidence interval
ci_lower = mean_sales - margin_error
ci_upper = mean_sales + margin_error

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

Output –

Mean Sales: 248.25

95% Confidence Interval: (241.37, 255.13)

Interpretation: We can be 95% confident that the true average daily sales fall between roughly 241.37 and 255.13 units.