Assignment Code: DA-AG-006
Statistics Advanced - 1|

Assignment Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them.

Each question carries 20 marks.
Total Marks: 200 Question

---

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

A random variable is a way to convert outcomes (like flipping coins, rolling dice, or measuring rainfall) into numbers for analysis.

---

2. Question 2: What are the types of random variables?
Answer:

Discrete Variable	Continuous Variable
Taking countable numbers, like whole numbers.	Taking any values within a given range(Real numbers), Values are measurable.
Common distributions are Bernoulli, Binomial.	Common distributions are Normal, standard, chi-square.
Example: Tossing a coin, number of students in a class.	Example: List of temperatures in various cities, height of girls in a certain sport.

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

Discrete Distribution	**Continuous Distribution**
It describes the probability of a discrete variable.	It describes the probability of a continuous variable within a range.
Probability mass function is used	Probability density function is used.
Probability values is always between 0 to 1.	Probability values which are areas under the curve are 0 to 1.
The probability of a specific value cannot be equal to zero.	As it takes interval values, the probability of specific value can be zero.
CDF behaviour shows step function	CDF behaviour shows continuous function

---

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

The binomial distribution is a fundamental tool in probability and statistics. Whenever there is a requirement to deal with multiple independent binary trials with a fixed probability of success, and a desire to see the distribution or likelihood of seeing a certain number of successes.

1. There should be a fixed number of trials.
2. Each trial is independent and has two outcomes.
3. Probability of success has to be constant.

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

Standard Deviation Distribution is a special case of the normal distribution where: Mean: 0 & Standard deviation: 1. It's a bell-shaped symmetric curve centered at 0.

It is important as it is widely used to simplify the calculations, especially by using Z-score to tell how many standard deviations a value is from the mean.

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

The Central Limit Theorem in Statistics states that as the sample size increases and its variance is finite, then the distribution of the sample mean approaches the normal distribution, irrespective of the shape of the population distribution.
It is critical in statistics as
1. We can check whether the sample mean differs significantly from an unknown population mean or not.
2. It makes hypothesis testing more reliable, as it allows to perform various tests based on the assumption of normality.
3. Because of the Central limit theorem, we can apply probability rules in real life sampling scenarios.

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

Instead of giving a single estimate, the confidence interval gives a range where the true value is likely to lie. This feature emphasis confidence interval in many ways:

1. It quantifies the uncertainty associated with sample statistics. A narrower interval means the estimate is more precise, while a wider interval shows more variability.
2. Used in hypothesis testing, to reject the null hypothesis.
3. It supports decision making, in real life while evaluating risk factors, predict outcomes, or clear margin of error.

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

The expected value (also called mean/proportion of a probability is the long run average

outcome that would be expected if repeated an experiment many times. It is the weighted average of all possible values that a random variable can take,, where the weights are their probabilities.

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:

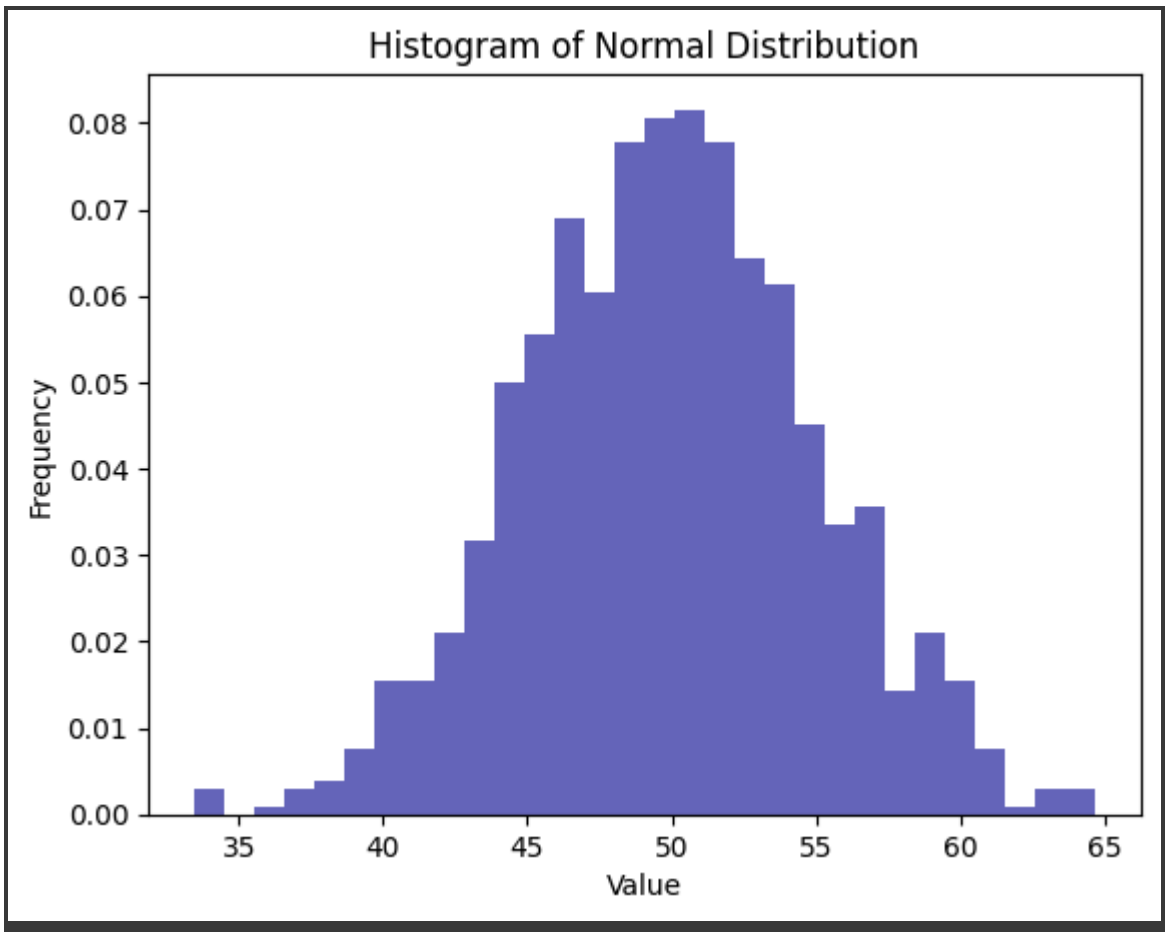
```
import numpy as np
import matplotlib.pyplot as plt

# Generate 1000 random numbers from a normal distribution
data = np.random.normal(loc=50, scale=5, size=1000)
mean = np.mean(data)
std_dev = np.std(data)

print(f"Mean: {mean}")
print(f"Standard Deviation: {std_dev}")
print()
plt.hist(data, bins=30, density=True, alpha=0.6, color='darkblue')
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.title("Histogram of Normal Distribution")
plt.show()
```

Output:

```
Mean: 49.90776891452444
Standard Deviation: 5.021049007996086
```



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:

After observing the sample size, I will apply the Central limit theorem through the T-distribution process.

1. As the sample size(n) :20
2. Population standard deviation is unknown.

Using t-distribution will allow me to use normal distribution properties to estimate average sales with a confidence interval of 95%.

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

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

n = len(daily_sales)
Sample_mean= np.mean(daily_sales)
Std_dev = np.std(daily_sales,ddof = 1)

SE = Std_dev / np.sqrt(n)  # standard error

t_crit = stats.t.ppf(0.975, df=n-1)  # 95% confidence interval

Margin_error = t_crit * SE

lower_bound = Sample_mean - t_crit * SE
upper_bound = Sample_mean + t_crit * SE

print(f"Sample Mean: {Sample_mean:.2f}")
print(f"Confidence Interval: ({lower_bound:.2f},
{upper_bound:.2f})")
```

Output:

```
Sample Mean: 248.25
Confidence Interval: (240.17, 256.33)
```