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: What is a random variable in probability theory?

Answer:

A **random variable** is a function that maps the possible outcomes of a random experiment (the sample space) to numerical values

Question 2: What are the types of random variables?

Answer:

Discrete Random Variables: Can only take a countable number of distinct values (e.g., counts).
Continuous Random Variables: Can take any value within a given interval (e.g., measurements).

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

Answer:

Distribution	Random Variable	Probability Function	Calculation
Discrete	Countable values (e.g., integers)	**Probability Mass Function (PMF)**, $P(X=x)$	Summation of individual probabilities
Continuous	Any value within a range (uncountable)	**Probability Density Function (PDF)**, $f(x)$	Integration (area under the curve)

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

Answer:

A **binomial distribution** is a **discrete** distribution modeling the number of **successes** in a **fixed number** (n) of independent trials, each with the same probability of success (p). It is used in probability to assess outcomes in binary (success/failure) experiments, like quality control or coin flips

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

Answer:

The **standard normal distribution** (or **z-distribution**) is a specific normal distribution with a **mean (μ) of 0** and a **standard deviation (σ) of 1**.

It is important because **any normal distribution can be standardized** to this form using **z-scores** ($z = \frac{x - \mu}{\sigma}$), allowing for universal comparison and forming the basis for many inferential statistical tests.

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

Answer:

The **Central Limit Theorem (CLT)** states that for a sufficiently large sample size ($n \ge 30$), the distribution of the **sample means** (the sampling distribution) will be **approximately normally distributed**, regardless of the original population's distribution. It is critical because it allows us to use the well-understood properties of the normal distribution to perform **statistical inference** (like confidence intervals and hypothesis testing) on a population mean, even if we don't know the population's true distribution.

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

Answer:

A **confidence interval (CI)** is a range of values that is likely to contain the true population parameter. Its significance is in:

- **Quantifying Uncertainty:** It expresses the statistical accuracy of a sample estimate by providing a plausible range, not just a single point estimate.
- **Inference:** It allows for conclusions about the population. If a CI does not contain a specific null value (e.g., zero), the result is statistically significant.

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

Answer:

The **expected value** $E(X)$ is the **long-term average** or **mean** of a probability distribution. It is a **weighted average** of all possible outcomes, where each outcome is weighted by its probability.

- **Discrete Formula:** $E(X) = \sum x P(x)$

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

# Parameters: mean=50, std_dev=5, n=1000
random_data = np.random.normal(loc=50, scale=5, size=1000)

sample_mean = np.mean(random_data)
sample_std = np.std(random_data)

print(f"Computed Sample Mean (x̄): {sample_mean:.4f}")
print(f"Computed Sample Standard Deviation (s): {sample_std:.4f}")

# Visualization (will vary)
# plt.hist(random_data, bins=30); plt.title('Normal Distribution Histogram'); plt.show()
```

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:

We assume the CLT allows us to use the **t-distribution** (for small samples $n=20$ and unknown population σ) to create a 95% Confidence Interval for the true average sales (μ). The interval is calculated using the sample mean ($\bar{x}$) and sample standard deviation (s).

```
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]

sample_mean = np.mean(daily_sales)
sample_std = np.std(daily_sales, ddof=1)
n = len(daily_sales)

# Calculate 95% CI using t-distribution
confidence_interval = stats.t.interval(
    0.95, df=n-1, loc=sample_mean, scale=sample_std / np.sqrt(n)
)

print(f"Sample Mean Sales (x̄): {sample_mean:.2f}")
print(f"95% Confidence Interval: [{confidence_interval[0]:.2f}, {confidence_interval[1]:.2f}]")
```