Statistics Advanced - 1| Assignment

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

Answer : A random variable in probability theory is a function that assigns a numerical value to each possible outcome in a sample space of a random phenomenon. It is not "random" itself, but rather a variable whose value is determined by the outcome of a random event. It allows us to apply mathematical analysis to the results of chance experiments. Random variables can be discrete (taking countable values) or continuous (taking any value in a range). Essentially, it's a way to translate non-numerical outcomes (like "heads") into numbers for calculation.

Question 2: What are the types of random variables?

Answer : The two main types of random variables are Discrete and Continuous. A discrete random variable can only take on a finite or countably infinite number of values (e.g., the number of heads in two coin flips: 0, 1, or 2). A continuous random variable can take on any value within a specified range or interval (e.g., the height or weight of a person). The type determines how its probability distribution is calculated and represented.

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

Answer: A discrete distribution is used for variables that can only take on a countable number of values (like integers), and probability is assigned to each individual value using a Probability Mass Function (PMF). In contrast, a continuous distribution is for variables that can take on any value within a given range (uncountable), and the probability of a specific, single value is zero. Continuous distributions use a Probability Density Function (PDF), and probabilities are calculated as the area under the curve over an interval. The sum of all probabilities in a discrete distribution equals 1, similar to the total area under the curve in a continuous distribution.

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

Answer: A binomial distribution is a discrete probability distribution that models the number of "successes" in a fixed number (n) of independent trials. Each trial must have only two possible outcomes (success/failure), and the probability of success (p) must remain constant across all trials. It is used in probability to calculate the likelihood of getting an exact number of successes (e.g., getting exactly 6 heads in 10 coin flips) in scenarios that fit these strict criteria, such as quality control, survey analysis, and medical trials.

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

Answer: The standard normal distribution (or Z-distribution) is a special continuous probability distribution where the mean (mue) is 0 and the standard deviation (sigma) is 1.

It is a form of the bell-shaped **Normal Distribution**. It is important because any normal distribution can be transformed into the standard normal using the **Z-score** formula: Z = (X - mue) / sigma . This standardization allows statisticians to use a single, universal table (the Z-table) to easily **calculate probabilities** and compare data from completely **different datasets** and scales.

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

Answer: The **Central Limit Theorem (CLT)** states that when you take sufficiently **large random samples** (typically n >= 30) from **any population** (regardless of its original distribution—be it uniform, skewed, etc.), the distribution of the **sample means** will be approximately a **normal distribution** (bell-shaped). The mean of this sampling distribution will equal the population mean (mue).

It is **critical** because it allows statisticians to use the powerful, well-understood properties of the normal distribution to perform **statistical inference** (like constructing confidence intervals and hypothesis testing) on a population, **even if the population's true distribution is unknown or non-normal**.

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

Answer: Confidence intervals (CIs) are significant because they provide a range of plausible values for an unknown population parameter (like the mean), not just a single point estimate. They quantify the precision of a sample estimate; a narrower interval implies a more precise estimate. They are used for statistical inference, indicating, with a specified confidence level (e.g., 95%), how often this interval-constructing method will capture the true parameter in the long run. CIs also help assess statistical significance in hypothesis testing: if the interval excludes the null hypothesis value (e.g., zero difference), the result is significant.

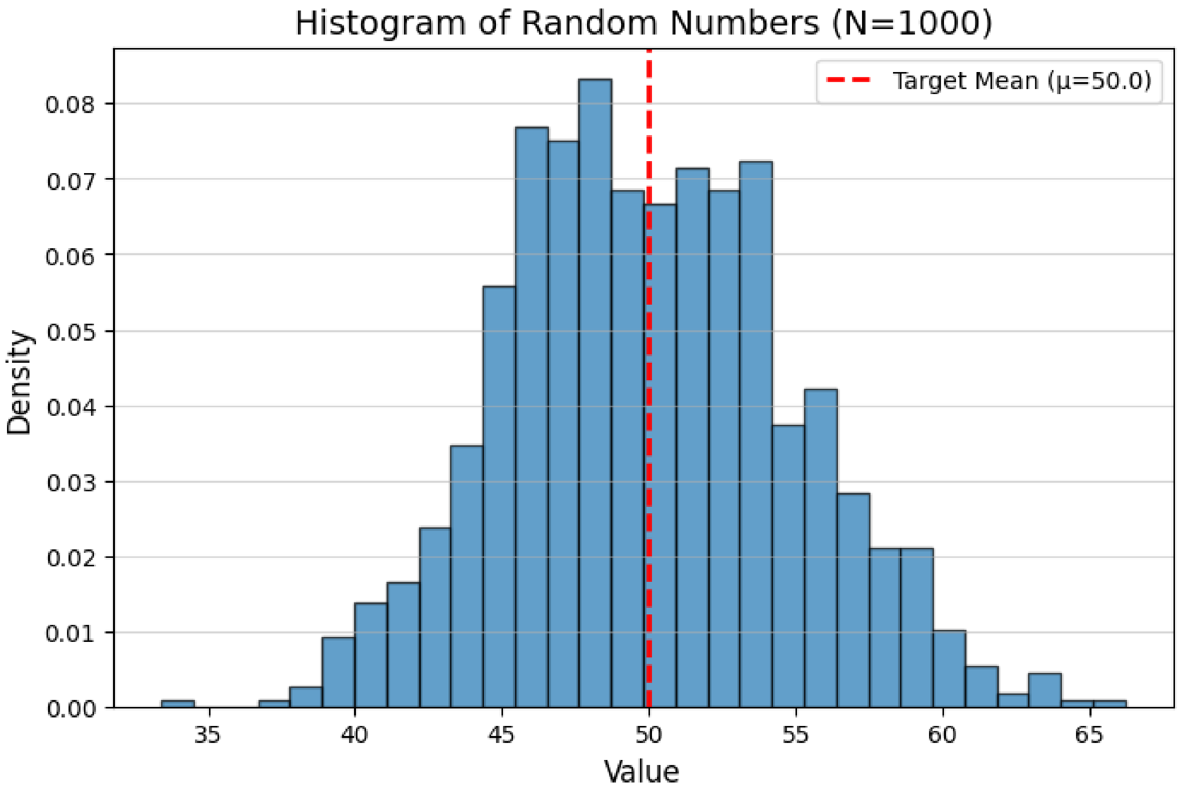
Question 8: What is the concept of expected value in a probability distribution?

Answer: The **expected value** (E[x]) of a random variable X represents the **long-term average** value of the variable if the random experiment were repeated many times.[3] It is essentially the **weighted average** of all possible outcomes, where each outcome is weighted by its probability of occurrence.[4] For a discrete variable, it's calculated as the sum of each outcome multiplied by its probability: E[X] = submersion of x .p(x) The expected value is not necessarily an outcome that will occur, but a crucial measure of the **central tendency** of the distribution.

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

MU = 50.0
SIGMA = 5.0
N_SAMPLES = 1000
data = np.random.normal(loc=MU, scale=SIGMA, size=N_SAMPLES)

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

print(f"--- Statistical Analysis of Generated Data ---")
print(f"Number of Samples (N): {N_SAMPLES}")
print(f"Target Distribution: N(μ={MU}, σ={SIGMA})")
print(f"")
print(f"Calculated Sample Mean: {sample_mean:.4f}")
print(f"Calculated Sample Standard Deviation: {sample_std:.4f}")

plt.figure(figsize=(8, 5))
plt.hist(data, bins=30, density=True, alpha=0.7, color='#1f77b4', edgecolor='black')
plt.title(f'Histogram of Random Numbers (N={N_SAMPLES})', fontsize=14)
plt.xlabel('Value', fontsize=12)
plt.ylabel('Density', fontsize=12)
plt.axvline(MU, color='r', linestyle='dashed', linewidth=2, label=f'Target Mean (μ={MU})')
plt.legend()

plt.grid(axis='y', alpha=0.5)
```



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.)

Answer:

```
import numpy as np

from scipy.stats import t


# Provided daily sales data (sample)

daily_sales = np.array([220, 245, 210, 265, 230, 250, 260, 275, 240, 255,

                        235, 260, 245, 250, 225, 270, 265, 255, 250, 260])


# 1. Calculate the necessary statistics
```

```
sample_mean = np.mean(daily_sales)
# Use ddof=1 for sample standard deviation (unbiased estimator)
sample_std = np.std(daily_sales, ddof=1)
n = len(daily_sales)
degrees_of_freedom = n - 1
confidence_level = 0.95

# 2. Compute the 95% Confidence Interval using the t-distribution
# FIX: The argument for the confidence level has been corrected from 'alpha' to 'confidence'.
confidence_interval = t.interval(
    confidence=confidence_level, # Corrected parameter name
    df=degrees_of_freedom,
    loc=sample_mean,
    scale=sample_std / np.sqrt(n) # Scale is the Standard Error of the Mean (SEM)
)

# --- Output the results ---
print(f"--- Sales Analysis Results ---")
print(f"Sample Size (n): {n}")
print(f"Calculated Sample Mean (X̄): ${sample_mean:.2f}")
print(f"Calculated Sample Standard Deviation (s): ${sample_std:.2f}")
print(f"Degrees of Freedom: {degrees_of_freedom}")
print(f"")
print(f"95% Confidence Interval for Average Sales:")
print(f"Lower Bound: ${confidence_interval[0]:.2f}")
print(f"Upper Bound: ${confidence_interval[1]:.2f}")
print(f"CI Range: (${confidence_interval[0]:.2f}, ${confidence_interval[1]:.2f})")
```