Statistics Advanced - 1| Assignment

Q1. What is a random variable in probability theory?
A1. A random variable is a function that assigns a numerical value to each outcome of a random experiment. The value of a random variable depends on chance and varies according to the outcome of the experiment.Random variables are used to quantify outcomes so that probability distributions can be defined and analyzed mathematically.

Q2. What are the types of random variables?
A2. There are two main types of random variables:
Discrete Random Variable : A discrete random variable takes a finite or countable number of values. Ex number shown on a dice.
Continuous Random Variable : A continuous random variable takes any value within a given range. Ex height of a person.

Q3. Explain the difference between discrete and continuous distributions?
A3. Discrete distributions deal with countable outcomes and assign probabilities to exact values, while continuous distributions deal with ranges of values and use probability density functions.
A discrete distribution represents a random variable that can take a finite or countable number of distinct values.For example, the outcomes of a dice roll or the number of students in a class follow a discrete distribution.
A continuous distribution represents a random variable that can take any value within a given range.Examples include height, weight, and time.

Q4. What is a binomial distribution, and how is it used in probability?
A4. A binomial distribution is a discrete probability distribution that represents the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes: success or failure, and the probability of success remains constant.
Binomial distribution is used to calculate the probability of a certain number of successes in repeated experiments such as number of heads in coin tosses ,number of students passing an exam.

Q5. What is the standard normal distribution, and why is it important?
A5. The standard normal distribution is a special type of normal distribution in which the mean is 0 and the standard deviation is 1. Any normal distribution can be converted into a standard normal distribution using standardization, which helps in simplifying probability calculations.
The standard normal distribution is important because it provides a common scale for analyzing and comparing data and makes statistical inference easier.

Q6. What is the Central Limit Theorem (CLT), and why is it critical in statistics?
A6. The Central Limit Theorem (CLT) states that when large samples are taken from any population (regardless of its original distribution), the distribution of the sample means will approach a normal distribution, provided the sample size is sufficiently large (generally $n \geq 30$ n≥30).
The Central Limit Theorem is critical because it justifies the use of normal distribution in statistical analysis and inference using sample data.

Q7. What is the significance of confidence intervals in statistical analysis?
A7. A confidence interval is a range of values used to estimate a population parameter (such as mean or proportion) based on sample data. It provides both an estimate and a measure of the uncertainty associated with that estimate.
Confidence intervals are significant because they quantify uncertainty, improve interpretation of results, and support reliable statistical inference.

Q8. What is the concept of expected value in a probability distribution?
A8. The expected value is a statistical measure that represents the long-term average value of a random variable over many repeated trials of a random experiment. It indicates what value we can expect on average from a probability distribution.
The expected value summarizes a probability distribution into a single representative value that reflects the average outcome over repeated experiments.

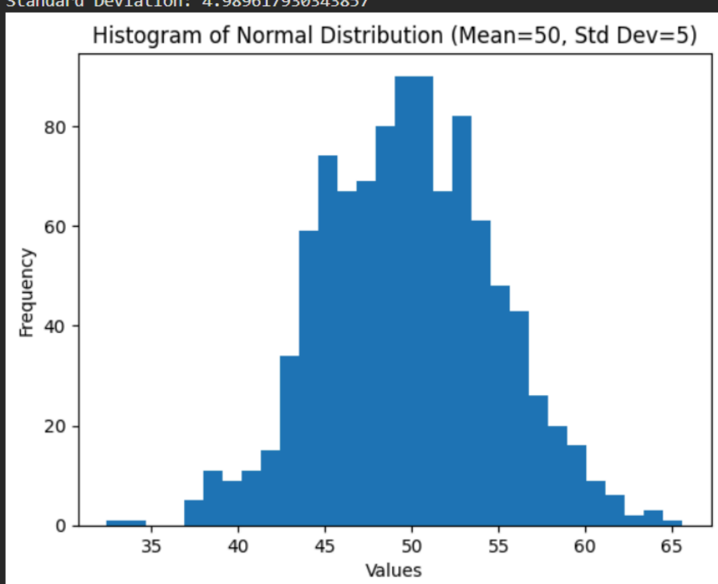
Q9. Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution.

A9. Code :

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

data = np.random.normal(loc=50, scale=5, size=1000)

mean_value = np.mean(data)
std_dev_value = np.std(data)

print("Mean:", mean_value)
print("Standard Deviation:", std_dev_value)

plt.hist(data, bins=30)
plt.title("Histogram of Normal Distribution (Mean=50, Std Dev=5)")
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.show()
```

Output :

```
Mean: 49.85610370533946
Standard Deviation: 4.989617930343857
```



Q10. 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.

A10. According to the CLT, when the sample size is reasonably large, the sampling distribution of the sample mean is approximately normal, regardless of the original distribution of the data. This allows us to estimate the population mean and construct a confidence interval using the normal distribution.
By applying the Central Limit Theorem, the sampling distribution of the mean becomes approximately normal, allowing us to estimate the average daily sales with a 95% confidence interval.

```
[6]
0s
import numpy as np
from math import sqrt

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

data = np.array(daily_sales)

mean_sales = np.mean(data)
std_dev = np.std(data, ddof=1)
n = len(data)

z = 1.96

standard_error = std_dev / sqrt(n)

lower_bound = mean_sales - z * standard_error
upper_bound = mean_sales + z * standard_error

print("Mean Daily Sales:", float(mean_sales))
print("95% Confidence Interval:",
      (float(lower_bound), float(upper_bound)))
```

```
Mean Daily Sales: 248.25
95% Confidence Interval: (240.68312934041109, 255.81687065958891)
```

Code in written format:
A9:

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

data = np.random.normal(loc=50, scale=5, size=1000)
```

```
mean_value = np.mean(data)
std_dev_value = np.std(data)

print("Mean:", mean_value)
print("Standard Deviation:", std_dev_value)

plt.hist(data, bins=30)
plt.title("Histogram of Normal Distribution (Mean=50, Std Dev=5)")
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.show()
```

A10.

```
import numpy as np
from math import sqrt

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

data = np.array(daily_sales)

mean_sales = np.mean(data)
std_dev = np.std(data, ddof=1)
n = len(data)

z = 1.96

standard_error = std_dev / sqrt(n)

lower_bound = mean_sales - z * standard_error
upper_bound = mean_sales + z * standard_error

print("Mean Daily Sales:", float(mean_sales))
print("95% Confidence Interval:",
      (float(lower_bound), float(upper_bound)))
```