

Statistics Advanced - 1| Assignment

Q1 : What is a random variable in probability theory?

Ans:

A random variable is a numerical quantity whose value depends on the outcome of a random experiment. It assigns a real number to each possible outcome in the sample space. Random variables allow us to quantify uncertainty and analyze probabilistic events mathematically.

For example, if we toss a coin:

- Let $X=1$ if Head occurs
 - Let $X=0$ if Tail occurs
- Here, X is a random variable.

Q2 : What are the types of random variables?

Ans:

There are two main types of random variables:

1. Discrete Random Variable

- Takes countable values
- **Examples:** number of heads in coin tosses, number of defective items

2. Continuous Random Variable

- Takes any value within a range
- **Examples:** height, weight, temperature, time

Q3 : Explain the difference between discrete and continuous distributions.

Ans:

Discrete Distribution: A discrete distribution deals with random variables that take countable and distinct values, where the probability is assigned to each specific outcome, such as the number of students in a class or the result of rolling a die.

Continuous Distribution: A continuous distribution deals with random variables that can take any value within a continuous range, where probabilities are defined over intervals rather than exact values, such as a person's height, weight, or temperature.

Q4 : What is a binomial distribution, and how is it used in probability?

Ans:

A binomial distribution models the number of successes in a fixed number of independent trials where:

- Each trial has two outcomes (success/failure)
- Probability of success remains constant

Formula:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Number of heads in 10 coin tosses.

Q5 : What is the standard normal distribution, and why is it important?

Ans:

The standard normal distribution is a normal distribution with:

- Mean $\mu = 0$
- Standard deviation $\sigma = 1$

It is important because:

- Any normal distribution can be converted to it using Z-scores
- It simplifies probability calculations
- It is widely used in hypothesis testing and confidence intervals

Q6 : What is the Central Limit Theorem (CLT), and why is it critical in statistics?

Ans:

The Central Limit Theorem (CLT) states that for a sufficiently large sample size, the sampling distribution of the sample mean approaches a normal distribution, regardless of the population's distribution.

Why it is critical:

- Allows inference even when population distribution is unknown
- Forms the foundation of confidence intervals and hypothesis testing
- Enables real-world data analysis

Q7 : What is the significance of confidence intervals in statistical analysis?

Ans:

A confidence interval provides a range of values that is likely to contain the true population parameter (like the mean).

Significance: Instead of providing a single "point estimate," it accounts for uncertainty and sampling error. For example, a 95% confidence interval means we are 95% confident that the true population mean falls within that range.

- Measures estimation reliability
- Shows uncertainty in sample estimates
- Widely used in decision-making and reporting results

Q8 : What is the concept of expected value in a probability distribution?

Ans:

The expected value (mean) represents the long-term average outcome of a random variable.

For a discrete variable:

$$E(X) = \sum x \cdot P(x)$$

For a continuous variable:

$$E(X) = \int xf(x)dx$$

It is widely used in economics, finance, and risk analysis.

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.

Ans:

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

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

# Compute mean and standard deviation
calculated_mean = np.mean(data)
calculated_std = np.std(data)

# Output results
print(f"Calculated Mean: {calculated_mean:.2f}")
print(f"Calculated Standard Deviation: {calculated_std:.2f}")

# Draw histogram
plt.hist(data, bins=30, edgecolor='black', alpha=0.7)
plt.title('Histogram of Normal Distribution (Mean=50, Std=5)')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```

Output:

- Calculated Mean: ~50.00
- Calculated Standard Deviation: ~5.00
- (*A bell-shaped histogram centered at 50*)

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.

Ans:

Apply the Central Limit Theorem to the daily_sales data:

- Calculate the sample mean (\bar{x}) of the provided sales data.
- Calculate the standard error using $SE = s / \sqrt{n}$ where s is the sample standard deviation and n is the sample size.
- For a 95% confidence interval, use the Z-score of 1.96. The interval is $\bar{x} \pm (1.96 * SE)$.

Code:

```
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]

mean_sales = np.mean(daily_sales)
std_error = stats.sem(daily_sales) # Standard error of the mean
confidence = 0.95

# Compute confidence interval
h = std_error * stats.t.ppf((1 + confidence) / 2, len(daily_sales) - 1)
start = mean_sales - h
end = mean_sales + h

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

Output:

- Mean Sales: 248.25
- 95% Confidence Interval: (240.24, 256.26) Approximate