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 numerical function that assigns a real number to each outcome of a random experiment. It provides a mathematical way to represent uncertain events quantitatively. Instead of describing outcomes in words, random variables allow outcomes to be analyzed using numbers and probability distributions.

Random variables are used to model real-life random phenomena such as the number of defective items in a batch, the marks obtained by a student, or the time taken to complete a task. They form the foundation of probability theory and statistics because all statistical analysis is based on random variables.

For example, if a coin is tossed three times, the random variable Xmay represent the number of heads obtained. The possible values of $X$are 0, 1, 2, and 3. Each value has a specific probability associated with it.

Thus, a random variable connects real-world random experiments with mathematical analysis, making it a fundamental concept in statistics.

Question 2: What are the types of random variables?

Answer:
Random variables are broadly classified into two main types:
1. Discrete Random Variable
A discrete random variable takes countable and finite or countably infinite values. The probability of each value can be listed individually. Discrete random variables are usually associated with counting processes.
Examples include:

- Number of students present in a class
- Number of defective products in a lot
- Number of heads in coin tosses

Discrete random variables are described using a probability mass function (PMF), where the sum of all probabilities equals 1.
2. Continuous Random Variable
A continuous random variable can take any value within a given range or interval. The number of possible values is infinite, and probabilities are assigned over intervals rather than exact values.
Examples include:

- Height of students
- Weight of fruits
- Time taken to complete a task

Continuous random variables are described using a probability density function (PDF), and the total area under the curve equals 1.
Thus, discrete variables are countable, while continuous variables are measurable.

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

Answer:

The difference between discrete and continuous distributions lies in the nature of the random variable they describe.
A discrete distribution represents random variables that take specific, countable values. Probabilities are assigned directly to each possible value. Examples include the binomial and Poisson distributions. In discrete distributions, the probability of a single outcome can be non-zero.
A continuous distribution, on the other hand, represents random variables that take values over a continuous range. Probabilities are calculated over intervals using probability density functions. Examples include the normal and exponential distributions. In continuous distributions, the probability of a single exact value is always zero.
In summary, discrete distributions involve counting, while continuous distributions involve measurement.

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

Answer:
The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent trials of a random experiment. Each trial has only two possible outcomes: success or failure, and the probability of success remains constant.
The conditions for a binomial distribution are:

1. Fixed number of trials
2. Each trial is independent
3. Two possible outcomes
4. Constant probability of success

The probability mass function of a binomial distribution is given by:

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

where $n$is the number of trials, $x$is the number of successes, and $p$is the probability of success.
Binomial distribution is widely used in quality control, medical testing, survey analysis, and decision-making processes.

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

Answer:
The standard normal distribution is a special case of the normal distribution with a mean of 0 and a standard deviation of 1. It is symmetric about the mean and follows a bell-shaped curve.
Any normal distribution can be converted into a standard normal distribution using the z-score formula:

$$z = \frac{x - \mu}{\sigma}$$

The importance of the standard normal distribution lies in its universal applicability. It allows probabilities from any normal distribution to be determined using a single standard normal table. It is widely used in hypothesis testing, confidence interval estimation, and statistical inference.

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 the sample size is sufficiently large, the sampling distribution of the sample mean approaches a normal distribution, regardless of the shape of the population distribution, provided the population has a finite mean and variance.
The CLT is critical because it justifies the use of normal probability models in real-world situations. It enables statisticians to make inferences about population parameters using sample data. Without the CLT, many statistical methods such as hypothesis testing, confidence intervals, and regression analysis would not be possible.
Thus, the Central Limit Theorem forms the backbone of inferential statistics

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

Answer:
A confidence interval (CI) is a range of values constructed from sample data that is likely to contain the true population parameter, such as the mean or proportion, with a specified level of confidence. Common confidence levels are 90%, 95%, and 99%.
The significance of confidence intervals lies in the fact that they provide more information than a single point estimate. While a point estimate gives only one possible value of a population parameter, a confidence interval shows the range within which the true value is expected to lie, along with the degree of certainty associated with that estimate.
Confidence intervals are important because:

1. They measure the reliability and precision of an estimate. A narrower interval indicates higher precision.
2. They help in decision-making, especially in scientific research, business, and quality control.
3. They are widely used in hypothesis testing, as they can indicate whether a parameter value is statistically significant.
4. They account for sampling variability, making results more realistic and interpretable.

For example, a 95% confidence interval for the mean indicates that if many samples were taken and intervals constructed, approximately 95% of them would contain the true population mean. Thus, confidence intervals play a crucial role in statistical inference and real-world data analysis.

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

Answer:
The **expected value** of a probability distribution is the theoretical average or mean value of a random variable. It represents the long-run average outcome of a random experiment if it is repeated a large number of times.
For a **discrete random variable**, the expected value is calculated as:

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

For a **continuous random variable**, it is given by:

$$E(X) = \int x f(x)\, dx$$

where $f(x)$is the probability density function.
The expected value is significant because:

1. It provides a **central measure** of a probability distribution.
2. It is used extensively in **decision theory**, economics, finance, and risk analysis.
3. It forms the basis for many advanced concepts such as variance, standard deviation, and regression analysis.

For example, in a game of chance, the expected value helps determine whether the game is fair or biased. Even though individual outcomes may vary, the expected value reflects the overall 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:

Python program to generate random numbers, compute statistics, and draw a histogram

Python Code:

```
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)

# Compute mean and standard deviation
mean_value = np.mean(data)
std_value = np.std(data)

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

# Plot histogram
plt.hist(data, bins=30)
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.title("Histogram of Normal Distribution (Mean=50, SD=5)")
plt.show()
```

Sample Output:

```
Mean: 49.92
Standard Deviation: 5.01
```

Explanation:

In this program, NumPy is used to generate 1000 random numbers from a normal distribution with a mean of 50 and a standard deviation of 5. The sample mean and standard deviation are calculated using NumPy functions. A histogram is drawn using Matplotlib to visualize the distribution, which shows a bell-shaped curve characteristic of a normal distribution.

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
import scipy.stats as st

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

# Calculate mean and standard deviation
mean_sales = np.mean(daily_sales)
std_sales = np.std(daily_sales, ddof=1)
n = len(daily_sales)

# 95% confidence interval
confidence_interval = st.t.interval(
   0.95,
   df=n-1,
   loc=mean_sales,
   scale=std_sales / np.sqrt(n)
)

print("Mean Sales:", mean_sales)
print("95% Confidence Interval:", confidence_interval)
```

OUTPUT:

```
Mean Sales: 248.25
```

95% Confidence Interval: (240.17, 256.33)

Using the Central Limit Theorem, the distribution of the sample mean of daily sales can be approximated as a normal distribution. Since the population standard deviation is unknown and the sample size is limited, the t-distribution is used to estimate the mean.

First, the sample mean and standard deviation of daily sales are calculated. Then, the standard error is obtained by dividing the standard deviation by the square root of the sample size. A 95% confidence interval is constructed using the formula:

$$\text{Mean} \pm t_{0.025} \times \frac{s}{\sqrt{n}}$$

Result:

- Mean daily sales = 248.25
- 95% confidence interval = (240.17, 256.33)

This means we are 95% confident that the true average daily sales of the company fall within this range.