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 in probability theory is a numerical description of the outcome of a random phenomenon or statistical experiment. It's essentially a function that assigns a real numerical value to each possible outcome in a sample space

Question 2: What are the types of random variables?

Answer: The two main types of random variables in probability theory are Discrete and Continuous random variables.

1.Discrete Random Variables
A discrete random variable is a variable that can take on only a finite or a countably infinite number of distinct values. These variables are typically the result of counting.

2. Continuous Random Variables
A **continuous random variable** is a variable that can take on any value within a given interval. These variables are typically the result of **measuring**

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

Answer: The core distinction is whether the possible outcomes can be counted or must be measured.

Feature	Discrete Distribution	Continuous Distribution
Random Variable	Discrete (Countable, distinct values)	Continuous (Measurable, infinite values in a range)
Probability Function	Probability Mass Function (PMF) $P(X=x)$	Probability Density Function (PDF) $f(x)$
Probability at a Point	Possible (The PMF gives a non-zero probability for an exact value, e.g., $P(X=2)=0.1$)	Zero ($P(X=x)=0$ for any specific value, because there are infinitely many possibilities)
Total Probability	The sum of all probabilities must equal 1 ($\sum P(X=x)=1$)	The area under the curve must equal 1 ($\int_{-\infty}^{\infty} f(x)dx=1$)
Probability of an Interval	Found by summing the PMF over the values in the interval.	Found by integrating the PDF over the interval (finding the area).
Real-World Examples	Number of defective items, count of customers, result of a die roll.	Height, weight, temperature, time.

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 probability of obtaining a certain number of "successes" in a fixed number of identical, independent trials.
It is one of the most fundamental distributions in probability and is used anytime an experiment has only two possible outcomes per trial (like a coin flip or a pass/fail test).

The binomial distribution is used to calculate the probability of observing exactly x successes in n trials.
1. Calculation of Probability
The probability mass function (PMF) for a binomial distribution $X \sim B(n,p)$ is:
$P(X=x)=(xn)px(1-p)n-x$
Where:

- $P(X=x)$ is the probability of getting exactly x successes.
- n is the fixed number of trials.
- x is the desired number of successes (where $x=0,1,2,\ldots,n$).
- p is the probability of success on a single trial.
- $(1-p)$ is the probability of failure on a single trial (often denoted as q).
- $(xn)=x!(n-x)!n!$ (the binomial coefficient) is the number of ways to choose x successes from n trials.

Measures
The distribution also allows for the easy calculation of the expected value (mean) and variance:

- Mean (Expected Number of Successes): $\mu=n \cdot p$
- Variance: $\sigma 2=n \cdot p \cdot (1-p)$

This allows analysts to quickly determine the expected number of successes without complex summation.

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

Answer: The standard normal distribution, also known as the z-distribution, is a specific type of normal distribution with the following fixed parameters:

- Mean (μ) is 0.
- Standard Deviation (σ) is 1.

Like all normal distributions, it is a continuous probability distribution characterized by its symmetrical, bell-shaped curve, where the mean, median, and mode are all equal to 0. The total area under the curve is equal to 1, representing 100% of the probability.

Importance of the Standard Normal Distribution
The standard normal distribution is crucial in statistics because it acts as a standardized scale for comparing and calculating probabilities for *any* normal distribution.
1. Standardization and Comparison
The standard normal distribution allows for standardization of data from different normal distributions. Any value (x) from a normal distribution with mean μ and standard deviation σ can be converted into a z-score using the formula:
$z=\sigma x-\mu$

- A z-score indicates precisely how many standard deviations a raw score (x) is above or below the mean.
- By converting scores from different distributions (e.g., SAT scores and ACT scores) into z-scores, you can compare observations on a common scale, regardless of their original units or variability.

2. Probability Calculations
Because the standard normal distribution is fixed, a single table (the z-table or standard normal table) can be used to find the probability (the area under the curve) associated with any z-score. This simplifies complex probability calculations:

- Finding the Probability: You can determine the probability of a value falling above, below, or between any two points in the distribution simply by converting the x-values to z-scores and looking up the corresponding area in the z-table.
- Statistical Inference: This forms the basis for many common statistical tests, like z-tests and the creation of confidence intervals, which rely on comparing observed data to the expected distribution defined by the standard normal curve.

3. The Empirical Rule (68-95-99.7)
The properties of the standard normal distribution define the Empirical Rule for all normal distributions:

- Approximately 68% of the data falls within ±1 standard deviation (z-scores of −1 to 1) of the mean.
- Approximately 95% of the data falls within ±2 standard deviations (z-scores of −2 to 2) of the mean.
- Approximately 99.7% of the data falls within ±3 standard deviations (z-scores of −3 to 3) of the mean**.**

.

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

Answer: The Central Limit Theorem (CLT) is a foundational concept in statistics that addresses the characteristics of sampling distributions. It states that if you take sufficiently large random samples from any population (regardless of its original distribution—be it normal, skewed, uniform, etc.), the distribution of the sample means will be approximately a normal distribution (a bell-shaped curve).
Key aspects of the CLT's conclusion are:

- The mean of the sampling distribution of the sample means ($\mu_{\bar{x}}$) will be equal to the population mean (μ).

- The standard deviation of the sampling distribution (known as the standard error of the mean, $\sigma_{\bar{x}}$) will be equal to the population standard deviation (σ) divided by the square root of the sample size (n).

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

Answer: The significance of confidence intervals (CIs) in statistical analysis lies in their ability to quantify the uncertainty and precision of an estimate, moving beyond a single guess to provide a plausible range of values for a population parameter.
A confidence interval is a range of values (e.g., 9.5 to 10.5) that is likely to contain the true value of an unknown population parameter (like the mean or a proportion), accompanied by a specified confidence level (e.g., 95%).
Key Significance of Confidence Intervals
1. Quantifying Precision and Uncertainty
CIs provide a measure of how reliable a sample-based estimate is.

- Precision: A narrower confidence interval (e.g., [9.5,10.5]) indicates a more precise estimate, suggesting that the true population value is likely very close to the sample estimate. This usually results from a larger sample size or lower data variability.
- Uncertainty: A wider confidence interval (e.g., [5,15]) indicates less precision and more uncertainty, suggesting that the sample estimate could be far from the true population value.

2. Informative Estimation (The Plausible Range)
CIs are preferred over single-value point estimates (like a sample mean) because they communicate the entire range of values that are plausible for the population parameter.

- Instead of just stating "the average is 10," a 95% CI of [9.5,10.5] tells the reader: "We are 95% confident that the true average lies somewhere between 9.5 and 10.5." This is much more informative for decision-making.

3. Alternative to p-values (Hypothesis Testing)
Confidence intervals provide a way to conduct hypothesis testing without solely relying on p-values.

- Statistical Significance: If a confidence interval for a difference between two groups does not contain the null value (which is usually 0 for a difference in means or 1 for a ratio like an odds ratio), the result is considered statistically significant at the chosen confidence level (α).
- Clinical/Practical Significance: Beyond just saying a result is "significant," the CI shows the entire range of likely effects. Researchers can use this to judge whether the entire range of plausible values is practically or clinically important. For example, a treatment that is statistically significant but has a CI of [0.1,0.2] may be deemed too small an effect to be practically useful.

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

Answer: The expected value (E[X]) of a probability distribution is the theoretical long-run average of a random variable (X). It represents the value you would expect to approach if you repeated the random experiment a very large number of times.

It is also known as the mean (μ) of the probability distribution, and it serves as a measure of central tendency.

Calculation and Meaning
The expected value is calculated as a weighted average of all possible outcomes, where each outcome's weight is its probability.
For a Discrete Random Variable
For a discrete random variable X with possible outcomes xi and corresponding probabilities P(xi):
$E[X]=\mu=\sum x_i \cdot P(x_i)$
This formula instructs you to:

1. Multiply each possible outcome by its probability.
2. Sum all those products.

Example: For a fair six-sided die, the possible outcomes are {1,2,3,4,5,6}, each with a probability of 1/6.
$E[X]=(1\cdot 61)+(2\cdot 61)+(3\cdot 61)+(4\cdot 61)+(5\cdot 61)+(6\cdot 61)=621=3.5$
Interpretation: An expected value of 3.5 does not mean you will ever roll a 3.5. It means that if you roll the die thousands of times, the average of all your rolls will get closer and closer to 3.5.
For a Continuous Random Variable
For a continuous random variable X with a probability density function f(x):
$E[X]=\int_{-\infty}^{\infty} x \cdot f(x)dx$

Practical Significance
The expected value is a crucial concept for decision-making under uncertainty:

- Risk Analysis and Finance: It's used to determine the anticipated value of an investment, a stock's return, or the average payoff of a lottery ticket or an insurance policy. A positive expected value suggests a long-term gain, while a negative value suggests a long-term loss.

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

Generate random numbers

mu = 50 # mean

```
sigma = 5   # standard deviation
random_numbers = np.random.normal(mu, sigma, 1000)

# Compute mean and standard deviation using NumPy
mean_rn = np.mean(random_numbers)
std_dev_rn = np.std(random_numbers)

print("Mean:", mean_rn)
print("Standard Deviation:", std_dev_rn)

# Draw a histogram

plt.hist(random_numbers, bins=30, density=True, alpha=0.6, color='g')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.title('Histogram of Random Numbers from Normal Distribution')
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:
The goal is to estimate the true average daily sales (μ) for the entire two-year period, based only on the provided sample data.

1. The Role of the Central Limit Theorem

The CLT is the foundational principle here. It states that if we take many random samples from *any* population distribution (even one that isn't normal), the distribution of the sample means will tend to follow a normal distribution.

- Application: Our small list of 20 daily_sales acts as one sample. The CLT tells us that the mean of this sample ($\bar{X}$) is a good point estimate for the true population mean (μ). Furthermore, the shape of the distribution of potential sample means allows us to determine a margin of error.

2. Estimating the Confidence Interval

Since we have a small sample size (n=20) and the true population standard deviation (σ) is unknown (we only have the sample standard deviation, s), we must use the t-distribution instead of the standard normal (Z) distribution for greater accuracy.
The confidence interval formula is:
Confidence Interval=$\bar{X}\pm t_{\alpha/2,n-1}\cdot n$
Where:

- $\bar{X}$: The sample mean (our best estimate of the true average sales).
- $t_{\alpha/2,n-1}$: The t-critical value for our desired confidence level (95%) and degrees of freedom (n−1=19).
- s: The sample standard deviation.
- n s: The standard error of the mean (SEM), which measures the variability of the sample means.

3. Interpretation of the 95% Confidence Interval

Once calculated, the 95% confidence interval (CI) means that if we were to take 100 different

random samples of daily sales and compute a CI for each one, approximately 95 of those intervals would contain the true, but unknown, average daily sales (μ) for the retail company.

---

2. Python Code Implementation

```
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]

# Calculate the sample mean

sample_mean = np.mean(daily_sales)

# Calculate the sample standard deviation

sample_std = np.std(daily_sales, ddof=1) # Use ddof=1 for sample standard deviation

# Calculate the standard error of the mean

n = len(daily_sales)
sem = sample_std / np.sqrt(n)

# Set the confidence level and find the corresponding Z-score

confidence_level = 0.95
alpha = 1 - confidence_level
z_score = stats.norm.ppf(1 - alpha/2) # Percent point function (inverse of cdf)

# Calculate the margin of error

margin_of_error = z_score * sem

# Calculate the confidence interval

confidence_interval_lower = sample_mean - margin_of_error
confidence_interval_upper = sample_mean + margin_of_error

print(f"Sample Mean: {sample_mean:.2f}")
print(f"Standard Error of the Mean: {sem:.2f}")
print(f"Z-score for {confidence_level*100}% confidence: {z_score:.2f}")
print(f"Margin of Error: {margin_of_error:.2f}")
print(f"95% Confidence Interval for Average Sales: ({confidence_interval_lower:.2f},
```

{confidence_interval_upper:.2f})")