****

 **I Semester M. Tech Internal Assessment -II**

**Computer Science and Engineering**

**Course Title:** **DATA SCIENCE**  **Course Code:** 23CSE5107

**Duration**: 75 Mins. **Date:** 25-02-2025

**Time:**10.00AM-11.15AM **Max Marks:**40

 **ANSWER KEY SOLUTION**

**1 A) What is A/B testing, and how is it used in real-world decision-making? Explain with an example.**

**A/B Testing and Its Real-World Applications**

A/B testing (also known as split testing) is an experimental technique used to compare two versions of a product, webpage, advertisement, or any other system to determine which one performs better. It is widely used in data-driven decision-making across industries such as marketing, product design, and software development.

**How A/B Testing Works**

1. **Identify the Goal** – Determine what metric you want to improve (e.g., click-through rate, conversion rate, engagement time).
2. **Create Variants** – Develop two versions:
	* **A (Control Group):** The original version.
	* **B (Treatment Group):** The modified version with a proposed improvement.
3. **Randomly Assign Users** – Randomly split users into two groups, ensuring fairness.
4. **Collect and Analyze Data** – Measure how users interact with each version and compare key performance indicators (KPIs).
5. **Make Data-Driven Decisions** – If version B performs better, implement the changes across all users.

**Real-World Example: E-commerce Website Optimization**

**Scenario:** An online retail store wants to increase sales by improving the design of its "Buy Now" button.

* **Control Group (A):** The current button is blue with "Buy Now" text.
* **Treatment Group (B):** The button is changed to green and labeled "Get Yours Today!"

**Implementation:** The company runs an A/B test where 50% of visitors see the blue button, and 50% see the green button.

**Results:**

* **Blue button:** 3% of visitors clicked and completed a purchase.
* **Green button:** 5% of visitors clicked and completed a purchase.

Since the green button led to a higher conversion rate, the company decides to implement it site-wide, leading to increased revenue.

**Applications of A/B Testing**

* **Digital Marketing:** Testing different email subject lines to improve open rates.
* **Social Media:** Experimenting with different ad formats to maximize engagement.
* **Software Development:** Optimizing UI/UX elements for better user retention.
* **Healthcare:** Evaluating different patient communication strategies to improve adherence to treatments.

A/B testing enables organizations to make evidence-based improvements rather than relying on guesswork, leading to more effective decision-making.

**Example of A/B Testing: Website Landing Page Optimization**

**Problem Statement:**

A SaaS (Software-as-a-Service) company wants to improve the conversion rate of its landing page. The conversion rate is defined as the percentage of visitors who sign up for a free trial after visiting the page. Currently, the conversion rate is 8%, and the goal is to increase it.

**Step 1: Define the Hypothesis**

The marketing team hypothesizes that changing the **headline** of the landing page can increase user engagement and improve conversions.

* **Current Headline (A - Control Group):** *"Start Your Free Trial Today!"*
* **New Headline (B - Treatment Group):** *"Boost Your Productivity – Try for Free!"*

The assumption is that the new headline is more action-oriented and conveys a clearer benefit, leading to higher sign-ups.

**Step 2: Design the Experiment**

The A/B test will be implemented on the company’s landing page using a **50/50 traffic split** between versions A and B.

**Metrics to Track:**

* **Primary Metric:** Conversion rate = (Number of sign-ups / Number of visitors) × 100
* **Secondary Metrics:**
	+ Bounce rate (percentage of visitors who leave without interacting)
	+ Average time spent on the page
	+ Click-through rate (CTR) on the "Sign Up" button

**Technical Setup:**

The test is implemented using an **A/B testing tool** like Google Optimize, Optimizely, or an in-house solution.

**Implementation Using JavaScript & Google Tag Manager:**

To randomly assign users to different versions:



This script dynamically changes the headline for half of the users.

**Step 3: Data Collection and Analysis**

The test runs for **two weeks**, collecting data from 100,000 visitors. The results are:



**Observations:**

* The new headline (B) led to a **20% increase in conversion rate** (from 8% to 9.6%).
* The bounce rate decreased from 55% to 50%, indicating higher user engagement.

**Step 4: Statistical Analysis**

To ensure the difference is statistically significant, we perform a **Chi-square test** or **Z-test** to check if the results are due to chance.

**Z-Test Calculation for Conversion Rate:**

Using the formula for the **Z-score** in A/B testing:



Plugging in the values, we compute the Z-score. If **Z > 1.96** (for 95% confidence), we conclude that variant B is significantly better.

**Step 5: Decision and Deployment**

Since the statistical test confirms significance, the company **deploys Variant B permanently** and monitors further performance.

**Key Learnings:**

1. **Data-Driven Decisions** – Instead of guessing, A/B testing provided clear evidence for improving the conversion rate.
2. **User Behavior Insights** – Small changes (like a headline) can significantly impact engagement.
3. **Continuous Optimization** – Future tests can experiment with CTA button color, page layout, or images.

**Real-World Use Cases of A/B Testing in Tech**

* **Netflix**: Tests different recommendation algorithms to improve user retention.
* **Amazon**: Experiments with different product page layouts to boost sales.
* **Google**: Optimizes search engine results by testing different ranking algorithms.

A/B testing is a powerful methodology for making **scientific, data-backed improvements** across digital platforms.

**2 A) Define sample mean, sample variance, and sample moments. Explain their importance in statistical analysis with suitable examples.**

In statistics, **sample mean, sample variance, and sample moments** are fundamental measures used to analyze data distributions. These metrics help in understanding the central tendency, variability, and shape of a dataset, which are crucial for making predictions and informed decisions.

**1. Sample Mean**

The **sample mean** is the arithmetic average of a sample and is used as an estimate of the population mean.

****

**Importance of Sample Mean:**

* Provides a **central value** for the data.
* Used in inferential statistics to estimate the **population mean**.
* Helps in **comparing different datasets**.

**Example:**

A factory produces light bulbs, and the lifetimes (in hours) of a random sample of 5 bulbs are:
**2000, 2100, 1950, 2200, 2050**

The sample mean is:



This means the average lifespan of bulbs in this sample is **2060 hours**.

**2. Sample Variance**

The **sample variance** measures the dispersion of data points around the mean. It tells us how much the data varies from the sample mean.



**Importance of Sample Variance:**

* Measures **data dispersion**.
* Helps in **risk assessment** (e.g., in finance, higher variance means higher risk).
* Used in **quality control** to maintain consistent production.

**Example:**

Using the light bulb data:



**3. Sample Moments**

Sample moments provide information about the **shape of a distribution** beyond mean and variance.

**Definition of Moments:**

The **k-th sample moment** about the mean is:







Given the sample dataset: **{5, 10, 15, 20, 25}**, calculate:
a) The **sample mean**
b) The **sample variance**
c) The **sample standard deviation**

A researcher collects the heights (in cm) of 6 students: **{150, 155, 160, 165, 170, 175}**. Compute:
a) Sample mean
b) Sample variance
c) Skewness (use the formula)

The lifespans (in months) of five mobile phone batteries are: **{20, 22, 19, 24, 21}**. Compute the **third moment (skewness)** of the dataset.

In a manufacturing plant, the weight of a product is measured in grams: **{48, 50, 49, 52, 51, 50}**. The company wants to ensure consistency in product weight.
a) Compute the **sample variance**.
b) If the variance is high, what steps should the company take to improve product consistency?

A financial analyst tracks the daily returns (%) of a stock over 7 days: **{2.5, -1.2, 3.0, -0.5, 1.8, -2.2, 2.7}**. Compute:
a) The **sample mean return**
b) The **sample variance**
c) The **skewness** of returns and interpret whether the stock is more likely to have positive or negative returns.

A university wants to analyze the exam scores of 10 students: **{75, 80, 85, 90, 95, 100, 110, 120, 130, 140}**.
a) Calculate the **sample mean and variance**.
b) Compute the **skewness** and determine if the distribution is symmetrical or not.

In an A/B testing experiment, a company tracks **conversion rates (%)** from two versions of a webpage:

* **Version A:** {3.2, 3.5, 4.0, 3.8, 4.1}
* **Version B:** {4.5, 4.8, 5.0, 4.9, 5.2}
a) Compute the sample means and variances for both versions.
b) Which version should the company implement and why?

**3 A) A university wants to estimate the average GPA of its students. They take a sample of 200 students from a total of 10,000 students.**

**Explain how the sample mean and standard error can be used to estimate the true mean GPA of the entire student population.**

**1.Introduction to the Problem**

A university wants to estimate the **true mean GPA** (μ\muμ) of its **entire student population** of **10,000 students**. Since surveying all students is impractical, they take a **random sample** of **200 students** and calculate their **sample mean GPA** (Xˉ\bar{X}Xˉ).

To ensure the estimate is **reliable**, we also calculate the **standard error (SE)**, which measures the **accuracy** of the sample mean as an estimator of the population mean.

**2. Sample Mean as an Estimator of Population Mean**



Since the sample is randomly chosen, Xˉ is an **unbiased estimator** of the true mean μ, meaning it **tends to be close** to the actual population mean when repeated sampling is done.



**Why is Standard Error Important?**

* It shows **how much uncertainty** is in our sample mean estimate.
* A **smaller SE** means the sample mean is a more **precise** estimate of the population mean.
* If **SE is large**, the sample mean fluctuates more and is a **less reliable** estimator.







A sample of **150 students** is taken from a university with a population of **12,000 students**. The **sample mean GPA** is **3.4**, and the **sample standard deviation** is **0.6**.

* + Calculate the **standard error (SE)**.
	+ Construct a **95% confidence interval** for the true mean GPA.
	+ Interpret the results.

A researcher selects **500 students** from a university to estimate the **average GPA**. If the sample mean is **3.1** with a standard deviation of **0.4**, compute:

* + The **standard error (SE)**
	+ The **99% confidence interval** for the true mean GPA.
	+ Compare it with a **95% confidence interval** and explain why the interval width changes.

A university administrator wants to ensure that the estimated GPA of students is **accurate within ±0.05 points** with **95% confidence**. If the standard deviation of GPAs is **0.6**, what **minimum sample size** is required?

Two universities conduct independent studies to estimate their students' average GPA:

* **University A:** Sample size = 300, Mean = 3.2, SD = 0.5
* **University B:** Sample size = 500, Mean = 3.3, SD = 0.6
* Compute the **standard error** for each university.
* Which university’s estimate is **more precise**, and why?

A school surveyed **400 students** and found that their **mean test score** was **78%**, with a **standard deviation** of **10%**.

* Compute the **standard error**.
* Construct a **90% confidence interval** for the true mean test score.

A pharmaceutical company tests a **new drug** on a **random sample of 100 patients** and finds that the **average recovery time** is **5.2 days**, with a standard deviation of **1.5 days**.

* Compute the **standard error**.
* Construct a **95% confidence interval** for the average recovery time.
* Explain why this information is important in healthcare decision-making.

A research team collects **SAT scores** from a **sample of 250 students** and finds an average score of **1100** with a standard deviation of **150**.

* Compute the **standard error**.
* Construct a **99% confidence interval**.
* How does this result help universities in setting admission criteria?

**4 A) A national bank wants to analyze the average daily withdrawal amount from ATMs across the country. The bank has 10,000 ATMs, but it is impractical to collect data from all of them.**

**Scenario:**

•The bank randomly selects 100 ATMs and records the average daily withdrawal amounts for each.

•The true population distribution of daily withdrawals is unknown and may not be normal.

•The sample means are calculated for these 100 ATMs.

**Questions:**

1.Explain how the Central Limit Theorem (CLT) justifies using the sample mean to estimate the population mean, even if the withdrawal amounts are not normally distributed.

2.If the standard deviation of daily withdrawals is known to be ₹1,500, what is the expected standard deviation of the sample mean for a sample size of 100 ATMs?

3.Suppose the sample mean withdrawal amount is ₹5,000. Construct a 95% confidence interval for the true population mean using CLT.

4.If the bank decides to increase the sample size from 100 to 400 ATMs, how will this affect the sampling distribution of the mean? Explain in terms of standard error and accuracy.

5.Discuss a real-world implication of using CLT in this study. Why is it useful for the bank’s decision-making process?

**Analyzing ATM Withdrawal Amounts Using Central Limit Theorem (CLT)**

A national bank wants to analyze the **average daily withdrawal amount** across **10,000 ATMs**. Since it is impractical to collect data from every ATM, they take a **random sample of 100 ATMs** and calculate the **sample mean withdrawal amounts**. The **true population distribution** of withdrawal amounts is unknown and may **not be normally distributed**.

To analyze the problem, we will use the **Central Limit Theorem (CLT)** and statistical methods to estimate the **population mean** and understand its **uncertainty**.

**Question 1: Central Limit Theorem (CLT) Justification**

The **Central Limit Theorem (CLT)** states that if we take a **large enough sample size (n ≥ 30)** from any population—**regardless of its original distribution**—the **sampling distribution of the sample mean will be approximately normal**.

**Why is CLT Important Here?**

* The **population distribution of ATM withdrawals** is unknown and may not be normally distributed.
* However, since our **sample size is 100** (n=100n = 100n=100), which is **large**, CLT ensures that the **distribution of the sample mean** will be approximately normal.
* This allows us to use **statistical techniques (such as confidence intervals and hypothesis tests) that assume normality**, even though the original ATM withdrawal amounts may be **skewed** or **non-normal**.

**Real-World Example of CLT Application**

Imagine ATM withdrawals are highly skewed—most withdrawals are small (e.g., ₹500-₹2000), but some are very large (₹30,000+). If we take random samples of **100 ATMs**, the **sample means** will still follow a normal distribution, even though individual ATM withdrawals do not.

This property makes CLT **extremely valuable** in banking and finance because it allows **valid inferences about large populations** using relatively small **random samples**.

**Question 2: Expected Standard Deviation of the Sample Mean (Standard Error)**

The **standard deviation of the sample mean**, known as the **Standard Error (SE)**, is given by:



**Interpretation**

* The **expected standard deviation of the sample mean is ₹150**.
* This means that if we repeatedly take random samples of 100 ATMs, the sample means would **typically vary by ₹150** around the true population mean.
* **Larger sample sizes result in smaller SE**, which means more **precise estimates** of the population mean.

**Question 3: Constructing a 95% Confidence Interval (CI)**

The **95% confidence interval** for the **true population mean** is given by:



**Interpretation of Confidence Interval**

* We are **95% confident** that the **true mean daily ATM withdrawal** across all **10,000 ATMs** lies between **₹4,706 and ₹5,294**.
* This means that if we repeated this study many times, **95% of the time, the true mean would fall within this range**.
* This range helps the bank estimate **expected daily cash demand** across its ATMs.

**Question 4: Effect of Increasing the Sample Size to 400 ATMs**

If the bank increases the **sample size from 100 to 400**, the **standard error (SE) will decrease**, improving the **accuracy** of the sample mean estimate.

New **SE formula** with n=400n = 400n=400:







**Question 5: Real-World Implications of Using CLT**

The **Central Limit Theorem (CLT)** is critical in banking for **decision-making and risk management**.

**Why is CLT Useful for the Bank?**

1. **Estimating Cash Requirements:**
	* Banks must **stock ATMs with enough cash** to meet demand without overloading cash reserves.
	* Using CLT, they can **predict the average daily withdrawals** across ATMs **without surveying all 10,000 machines**.
2. **Fraud Detection and Anomaly Detection:**
	* If the **sample mean withdrawal amount** deviates significantly from previous estimates, it could indicate **unusual activity or fraud**.
	* Statistical monitoring can trigger **alerts** if large, unexpected withdrawals occur.
3. **Cost Optimization for Cash Logistics:**
	* Reducing the **number of cash refills** saves **transportation and security costs**.
	* By using CLT to estimate **average ATM withdrawals**, the bank can **strategically plan** cash refills and minimize waste.
4. **Loan and Interest Rate Adjustments:**
	* Banks can **adjust interest rates** based on **customer spending trends**.
	* If ATM withdrawal amounts indicate increased spending, it could suggest a **strong economy**, leading to different financial policies.

**Summary**

1. **Central Limit Theorem (CLT) ensures that the sample mean follows a normal distribution**, allowing reliable statistical analysis even if individual ATM withdrawals are **not normally distributed**.
2. The **standard error (SE) helps measure the variability of the sample mean** and is crucial for assessing the precision of our estimate.
3. A **95% confidence interval** helps the bank **estimate the true population mean ATM withdrawal** and make informed decisions.
4. **Increasing the sample size reduces SE**, leading to **more accurate estimates**, but with **diminishing returns**.
5. **Real-world applications** of CLT in banking include **cash management, fraud detection, and financial planning**, making it a vital tool for **data-driven decision-making**.

By leveraging **CLT and statistical methods**, the bank can make **efficient and cost-effective** financial decisions without having to collect data from all **10,000 ATMs**!

**Scenario 1: Online Shopping Delivery Times**

An **e-commerce company** wants to estimate the **average delivery time** for orders. Since tracking delivery times for all **500,000 orders per month** is impractical, they select a **random sample of 200 orders**.

**Questions:**

1. Explain how the **Central Limit Theorem (CLT)** allows the company to estimate the **true mean delivery time**, even if individual delivery times vary widely and are **not normally distributed**.
2. If the standard deviation of delivery times is **2.5 days**, what is the **expected standard error** for a sample of **200 orders**?
3. If the **sample mean delivery time** is **4.2 days**, construct a **95% confidence interval** for the true mean delivery time.
4. If the company **increases the sample size from 200 to 800 orders**, how will this affect the **sampling distribution** and the **precision of the estimate**?
5. How can accurate estimation of delivery times using **CLT** help the company **optimize logistics and improve customer satisfaction**?

**Scenario 2: Mobile Network Data Usage**

A **telecom provider** wants to estimate the **average monthly data usage** of its customers. The provider has **2 million users**, so they select a **random sample of 500 users**.

**Questions:**

1. Explain why the **Central Limit Theorem (CLT)** justifies using the **sample mean data usage** to estimate the **true mean usage**, even if individual data usage is highly variable.
2. If the **population standard deviation** of data usage is **4 GB**, calculate the **standard error** for a sample of **500 users**.
3. If the **sample mean monthly data usage** is **12 GB**, construct a **99% confidence interval** for the true mean data usage.
4. If the company increases the sample size from **500 to 2,000 users**, how does it affect the **margin of error**? Explain.
5. Discuss how **accurate estimation of data usage** using CLT can help the telecom provider in **pricing strategies and network capacity planning**.

**Scenario 3: Manufacturing Quality Control (Battery Production)**

A **battery manufacturer** wants to estimate the **average lifespan** of its batteries before failure. Since testing all **100,000 produced batteries** is impossible, they select a **random sample of 300 batteries**.

**Questions:**

1. The battery lifespan follows an **unknown distribution**. Explain how **CLT** ensures that the **sampling distribution of the mean** is approximately normal.
2. If the **standard deviation of battery lifespan** is **50 hours**, calculate the **standard error** for a sample of **300 batteries**.
3. If the **sample mean battery lifespan** is **500 hours**, construct a **95% confidence interval** for the **true average battery lifespan**.
4. What happens to the **confidence interval width** if the manufacturer increases the **sample size from 300 to 600 batteries**? Explain.
5. How can the use of **CLT** in estimating battery lifespan help in **warranty policies and product improvements**?

**Scenario 4: University Exam Scores Analysis**

A university wants to estimate the **average score** of students in a nationwide **mathematics exam**. The total number of students appearing for the exam is **1.5 million**, so the university selects a **random sample of 1,000 students**.

**Questions:**

1. Explain how the **Central Limit Theorem (CLT)** helps in estimating the **true mean exam score**, even if individual scores do not follow a normal distribution.
2. If the **population standard deviation** of exam scores is **12 points**, calculate the **standard error** for a sample of **1,000 students**.
3. If the **sample mean exam score** is **72 points**, construct a **90% confidence interval** for the **true mean score**.
4. How would **doubling the sample size** (from **1,000 to 2,000 students**) affect the **accuracy of the estimate**?
5. Discuss how **accurate estimation of exam scores** using CLT can help in **education policy and grading standardization**.

**Scenario 5: Supermarket Customer Spending Habits**

A supermarket chain wants to estimate the **average spending per customer**. Since tracking spending for all **2 million customers per month** is impractical, they select a **random sample of 500 customers**.

**Questions:**

1. The spending distribution is **right-skewed** (most customers spend small amounts, while some spend very large amounts). Explain how **CLT** allows the supermarket to estimate the **true mean spending** using the **sample mean**.
2. If the **population standard deviation** of spending is **₹800**, calculate the **standard error** for a sample of **500 customers**.
3. If the **sample mean spending** is **₹3,500**, construct a **95% confidence interval** for the **true mean customer spending**.
4. What will happen to the **confidence interval width** if the supermarket **increases the sample size to 2,000 customers**? Explain why.
5. Discuss how **using CLT** in estimating customer spending can help supermarkets in **inventory management, pricing strategies, and promotions**.