USN | | | | | | | | | | |

InternalAssessmentTest1
Sep 2025

Sub:	Statistical Machine Learning for Data Science (SML)					SubCode:	BAD702	Branch:	AINDS/CS(DS)
Date:	29/09/25	Duration:	**90minutes**	MaxMarks:	**50**	Sem	**VII**		**OBE**

		AnsweranyFIVEQuestions	**MARKS**	**CO**	**RBT**
1		**A startup has collected data on the daily time (in minutes) that 50 users spend on their new mobile learning app. The company wants to estimate the average daily usage in the population.** **Since the sample size is not very large and the underlying distribution looks skewed, the data science team decides to use bootstrap resampling.** **Questions:** **1. How would you apply the bootstrap method to estimate a 95% confidence interval for the mean daily usage?** **2. Why might bootstrap be preferred here over using a standard formula for the confidence interval?** **3. If after 1,000 bootstrap resamples, the 2.5th percentile of the means is 32 minutes and the 97.5th percentile is 48 minutes, what would you conclude about the population mean?**	10	CO1	L3

1) How to apply the bootstrap to estimate a 95% CI for the mean

Procedure (step-by-step):

1. Start with your observed sample of daily usage times (size n=50). Call it x.
2. Repeat the following **B** times (common choices: 1,000 — 10,000):
   - Draw a bootstrap sample of size n **with replacement** from x.
   - Compute the sample mean of the bootstrap sample; store it.
3. After B resamples you have a bootstrap distribution of the mean: $\{\bar{x}1*,\ldots,\bar{x}B*\}$.
4. The **percentile bootstrap 95% CI** is simply the empirical 2.5th and 97.5th percentiles of that bootstrap-means distribution:

$$\mathrm{CI}_{95\%} = \left(\text{percentile}_{2.5}(\bar{x}^*),\ \text{percentile}_{97.5}(\bar{x}^*)\right)$$

5. Optionally compute the bootstrap standard error:

$SE_{boot} = \mathrm{sd}(\bar{x}^*)$. You can also build a normal-approx CI using , $\bar{x} \pm z_{0.975} \cdot SE_{boot}$,but for skewed data the percentile or BCa intervals are preferred.

Quick checklist / assumptions: sample observations should be roughly independent and the sample should be representative of the population you want to infer about.

2) Why bootstrap is preferred here

- **Skewed underlying distribution:** Classical formula for the mean uses the Central Limit Theorem (CLT) and normal-based CIs. For small-to-moderate n and **skewed** data the CLT approximation can be poor; bootstrap captures the actual shape of the sampling distribution empirically.
- **Small-ish sample size (n = 50):** 50 is borderline; with heavy skew, normal approximations may be biased. Bootstrap does not require normality.
- **No need to assume parametric form:** If you're unsure about the population distribution (which you are), bootstrap is nonparametric and more robust.
- **Flexible to statistics beyond the mean:** Bootstrap easily extends to medians, percentiles, differences, etc.
- **Practical caveats:** bootstrap relies on the sample being representative and having independent observations. If those are violated (e.g., strong time dependence), bootstrap results will be unreliable.

If you need higher accuracy in skewed cases, consider the **BCa (bias-corrected and accelerated)** bootstrap interval rather than plain percentile.

3) Interpretation of the given bootstrap percentiles

We performed 1,000 bootstrap resamples. The 2.5th and 97.5th percentiles of the bootstrap means are **32** and **48** minutes, respectively.

Conclusion:
A 95% bootstrap percentile confidence interval for the population mean is [32, 48] minutes. Practically, this means that based on the observed sample and the bootstrap procedure, the best empirical estimate is that the true **average daily usage** in the population lies between **32 and 48 minutes** with *approximate* 95% confidence.

Short interpretation we can report:
"We estimate the population mean daily usage to be between **32 and 48 minutes (95% CI, bootstrap percentile)**."

Extra notes / caveats you should state:

- This interval is conditional on the sample being representative and

observations independent.

- With only B = 1,000 resamples the interval is usually fine, but for more precise endpoints you can increase to B = 5,000–10,000 (computationally cheap).
- If the bootstrap distribution is strongly skewed, consider the BCa interval which corrects bias and skewness.

Bonus — Jupyter / Python code

```
import numpy as np
import pandas as pd
def bootstrap_mean_ci(data, B=10000, alpha=0.05, seed=None):
    rng = np.random.default_rng(seed)
    n = len(data)
    means = np.empty(B)
    for b in range(B):
        sample = rng.choice(data, size=n, replace=True)
        means[b] = sample.mean()
    lower = np.percentile(means, 100*(alpha/2))
    upper = np.percentile(means, 100*(1-alpha/2))
    se_boot = means.std(ddof=1)
    return {
        "mean_observed": np.mean(data),
        "se_boot": se_boot,
        "ci_percentile": (lower, upper),
        "bootstrap_means": means
    }
```

2(a)	What do you mean by Bootstrap samples? How does it help in building a Confidence Interval?	5+5	CO1	L3

Suppose you have a dataset of size n:

$$x = \{x_1, x_2, \ldots, x_n\}$$

- A **bootstrap sample** is a new sample of size n drawn **with replacement** from the original dataset.
  - "With replacement" means the same observation can appear multiple times in the bootstrap sample, or not at all.
  - For example, if your dataset is [10,20,30,40], a bootstrap sample could be [20,20,40,10] or [30,30,10,40].
- We can generate many such bootstrap samples (say B=1000, 5000, or more).
- For each bootstrap sample, we compute the statistic of interest (mean, median, variance, regression coefficient, etc.).

This gives a **bootstrap distribution** of that statistic.

How does it help in building a Confidence Interval?

The key idea:

- In classical statistics, confidence intervals often rely on theoretical formulas (like assuming normality of errors, using t- or z-distributions, etc.).
- But what if the data are skewed, sample size is small, or you don't want to assume a specific distribution?

Bootstrap helps because:

1. It **mimics repeated sampling** from the population by resampling from the observed data.
2. The variability in the bootstrap distribution reflects the sampling variability of the statistic.
3. To build a **95% confidence interval**, we take the middle 95% of the bootstrap distribution (usually the 2.5th and 97.5th percentiles).

2(b) A pharmaceutical trial compares a new drug with a placebo. Suppose you reject the null hypothesis and claim the drug works, but later it turns out the drug has no effect.

- What type of statistical error have you made?
- How could you reduce the chances of this error in future experiments?

Question 1: What type of statistical error is this?

- The null hypothesis H0: "The drug has no effect."
- You rejected H0 and concluded the drug works.
- Later, it turned out the drug really has **no effect** → your rejection of H0 was wrong.

 This is a **Type I Error** (false positive).

Definition: Type I Error occurs when you reject a true null hypothesis.

Question 2: How to reduce the chances of this error?

The probability of making a Type I Error is the **significance level (α)** you choose for your test.

- Common choices: $\alpha=0.05$ (5%) or $\alpha=0.01$ (1%).

To reduce Type I error:

1. **Lower the significance level (α)**:
    - For example, test at 1% instead of 5%.
    - Makes it harder to reject H0.
2. **Use corrections for multiple testing** (e.g., Bonferroni correction) if many comparisons are being made.
3. **Improve experimental design**:

		○ Increase sample size for more reliable estimates. ○ Reduce bias and noise (randomization, blinding, proper controls). 4. **Replication**: Repeating the experiment independently reduces the chance that one false positive drives the conclusion.			

3		Explain the practical relevance of Binomial and Poisson Distribution. Give the formula for their Pdf. What is output of the following Python code : from scipy import stats print(stats.binom.pmf(2, n=5, p=0.5)) **1. Practical Relevance** **Binomial Distribution** • Used when there are a **fixed number of independent trials** with two possible outcomes (Success/Failure). • Examples: ○ Tossing a coin 10 times and counting the number of heads. ○ Quality control: number of defective items in a batch of 50. ○ Medical trials: number of patients who respond positively to a new drug out of a fixed sample. **Poisson Distribution** • Used when we count the number of events happening in a **fixed interval of time/space**, assuming events occur independently at a constant average rate. • Examples: ○ Number of customer arrivals at a bank per hour. ○ Number of phone calls received at a call center per minute. ○ Number of emergency cases arriving at a hospital per day. **2. Probability Density Function (PMF)** • **Binomial PMF:** $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$ where • n = number of trials, • p = probability of success, • k = number of successes.	10	CO2	L2

Poisson PMF:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where

- λ = average rate of occurrence,
- k = number of events.

3. Python Code Output

from scipy import stats

print(stats.binom.pmf(2, n=5, p=0.5))

This computes the probability of getting exactly **2 successes** in **5 trials** with probability of success p=0.5.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \qquad \cdot (0.125) = 0.3125$$

4	a	Convergence Always the same shape Approaches the normal distribution as df → ∞	5	CO3	L1

Difference between Student's t-distribution and Normal distribution

Feature	Normal Distribution	Student's t-Dis
Shape	Symmetrical, bell-shaped	Symmetrical, bell-shaped
Parameters	Mean (μ), Standard deviation (σ)	Degrees of freedom (df = n-
Tails	Thinner tails	Heavier tails (more probabi
Usage	When population σ is known, or sample size is very large	When σ is unknown and we sample (small/medium sam
Convergence	Always the same shape	Approaches the normal dist

What is Standard Error (SE)?

- **Definition:** The **standard error** is the standard deviation of a **sampling distribution** of a statistic (like the sample mean).
- For the sample mean:

$$SE = \frac{s}{\sqrt{n}}$$

where:

- s = sample standard deviation,
- n = sample size.
- **Interpretation:** SE measures how much the sample mean is expected to vary from one sample to another.
- **Key role:** Smaller SE → the sample mean is a more precise estimate of the population mean.

b State the Central Limit Theorem? How is it relevant to Statistical Inference. | 5 | CO3 | L1

Central Limit Theorem (CLT)

Statement:
If we take many random samples of size n from any population with mean μ and finite variance σ2, then as nnn becomes large, the **sampling distribution of the sample mean** $\bar{X}$ approaches a **Normal distribution**, regardless of the population's original distribution.

Formally:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty$$

Relevance to Statistical Inference

The CLT is the backbone of modern statistics because it justifies why we can use **normal-based methods** (like z-tests, t-tests, confidence intervals) even when the underlying population is **not normal**:

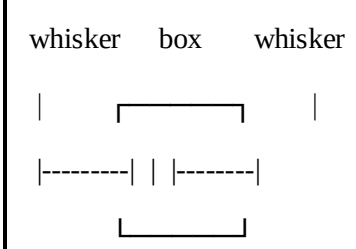
1. **Approximate Normality of Sample Mean:**
   Even if data is skewed or irregular, the mean of a reasonably large sample (usually n≥30) will be approximately normal.
2. **Confidence Intervals:**
   CLT allows us to construct confidence intervals for population means using the normal (or t) distribution.
3. **Hypothesis Testing:**
   Test statistics (like the standardized sample mean) rely on the CLT to follow approximately normal distributions under H0.
4. **Practical Applications:**
   - Polling: estimating population proportions.
   - Quality control: average defect rates.
   - Finance: average returns over time.

5	a	State the difference between Violin plot and Box plot with diagram. When do we use Hexagonal Binning.	5	CO3	L1

Feature	Box Plot	Violin P
Shows	Summary statistics (median, quartiles, whiskers, outliers)	Summary statistics **and** ful
Shape	Rectangular box with whiskers	Symmetric, violin-shaped
Data distribution	Hides the shape of the distribution (only gives 5-number summary)	Displays the **kernel densit** showing whether the data i multimodal, etc.
Use case	Good for comparing medians and spread across groups	Good for seeing detailed di especially multimodal data

Diagram (conceptual):

Box Plot:



Violin Plot:

So violin plots = box plot + distribution shape.

When do we use Hexagonal Binning?

- **Hexbin plots** are used when you have **two continuous variables** and a **large dataset**, where scatter plots would suffer from **overplotting** (points overlapping, making it hard to see density).
- Instead of plotting each point, the data space is divided into **hexagonal bins**, and the color of each hexagon represents the **count (frequency)** of points in that bin.

Use cases:

- Visualizing correlations in large datasets.
- Example: plotting 100,000 values of `height` vs `weight`.
- Density-based patterns become clear compared to a messy scatter plot.

b

What do you mean by Test Statistic? Differentiate between Sample Parameter and Population parameter.

5 CO3 L1

What do you mean by a Test Statistic?

- A **test statistic** is a numerical value calculated from sample data that is used in hypothesis testing.
- It measures how far the sample statistic is from the hypothesized population parameter, relative to the variation in the data.
- The test statistic is then compared to a theoretical distribution (Normal, t, F, Chi-square, etc.) to decide whether to reject the null hypothesis.

Examples:

- **z-test statistic:**

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- **t-test statistic:**

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- **Chi-square test statistic (goodness-of-fit):**

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Difference between **Sample Parameter** and **Population Parameter**

Aspect	Population Parameter	Sample
Definition	A numerical summary that describes the **entire population**	A numerical summa **sample** (subset of p
Examples	Population mean (μ), population variance (σ^2), population proportion (P)	Sample mean ($\bar{x}$\ba variance (s^2), sampl
Known / Unknown	Usually unknown (we can't measure the whole population)	Known (we comput collected)
Role	Fixed value (does not change)	Varies from sample **estimate** the popula
Use in Inference	True value we want to infer about	Provides the basis f testing about the po

6	a	Explain the intuition behind Hypothesis Testing, Level of Significance and p-value
		
		Intuition behind Hypothesis Testing
		
		• Imagine you're a detective. You start with a presumption of innocence (**null hypothesis, H_0**).
		• You collect evidence (sample data).
		• You ask: *"Is this evidence strong enough to reject H_0 and believe the alternative hypothesis, H_1?"*
		
		So, **hypothesis testing** is a structured decision-making process:
		
		1. Assume **H_0 is true** (status quo).
		2. Calculate a **test statistic** from your sample.
		3. Compare it to a reference distribution (normal, t, chi-square, etc.).
		4. If the evidence is too unlikely under H_0, you reject H_0 in favor of H_1.
		
		Level of Significance (α)
		
		• Denoted by **α**, it is the **threshold for risk** you are willing to take of making a **Type I Error** (rejecting a true H_0).
		• Common values:
		  ○ $\alpha = 0.05 \rightarrow$ 5% chance of wrongly rejecting H_0.
		  ○ $\alpha = 0.01 \rightarrow$ 1% chance.
		
		Example: If $\alpha = 0.05$, you're saying *"I'm okay with being wrong 5 times out of 100 in rejecting H_0."*
		
		p-value
		
		• The **p-value** is the probability of observing a test statistic as extreme as (or more extreme than) the one you got, **assuming H_0 is true**.
		• It answers: *"If the null hypothesis were true, how surprising is my sample?"*
		
		Interpretation:
		
		• Small p-value ($\leq \alpha$) $\rightarrow$ evidence is strong against H_0 $\rightarrow$ reject H_0.
		• Large p-value ($> \alpha$) $\rightarrow$ not enough evidence $\rightarrow$ fail to reject H_0.
		
		Example:
		
		• $p = 0.03$, $\alpha = 0.05 \rightarrow$ reject H_0 (evidence suggests effect exists).
		• $p = 0.40$, $\alpha = 0.05 \rightarrow$ fail to reject H_0 (data consistent with no effect).
			5	CO3	L3

	b	A sports scientist is tracking the performance of professional runners. In the first race of the season, one athlete unexpectedly runs **much faster than their usual average time** (an unusually good performance).	5	CO3	L3

The scientist predicts that in the next race, the athlete's performance will be **slower** and closer to their long-term average.

Questions:

1. What statistical concept explains why the athlete's performance is likely to decline toward their usual average in the next race?
2. Does this mean the athlete is "getting worse"? Explain why or why not.
3. Give another real-world example (outside sports) where regression to the mean commonly occurs.

1.What statistical concept explains this?

The concept is **Regression to the Mean**.

 It means that if a random variable shows an **extreme value** (very high or very low) in one measurement, the next measurement is likely to be closer to its average, simply because of natural variability.

In this case, the athlete's unusually fast race is partly due to chance factors (good weather, perfect mindset, competition, etc.). These lucky factors won't all align again, so the next performance is expected to move closer to the average.

2. Does this mean the athlete is "getting worse"?

✗No, it doesn't.

The athlete is not suddenly performing worse. The unusually good performance was an **outlier** influenced by random variation. Their "true" performance ability is reflected by their long-term average.

So when performance drops back toward the average, it's not a decline in skill — it's just the natural balancing out of random fluctuations.

3. Another real-world example of Regression to the Mean

 Education:
Students who score extremely high or extremely low on a test often score closer to the class average on the next test. This doesn't mean smart students got "dumber" or struggling students got "smarter" — it's just that random factors (luck, question fit, mood, etc.) don't repeat in the same way.