## DS203-2025-S2: E02

- **Submissions due by:** Aug 28, 2025; 23:55 Hrs.
- **Follow the submission guidelines** given at the end of this document.
- **Late / non-submissions:** (-1) mark will be added to your account.
- **Copied / fraudulent submissions:** (-10) marks will be added to your account.
- Blank and woefully inadequate / irrelevant submissions will be considered fraudulent.

---

1. Review and execute the Notebook `sampling-distribution-mean-Aug-2025.ipynb` and record your observations and conclusion related to the following questions: (be sure to include plots as part of your explanations!)
    a. What is the relationship between the population standard deviation and that of the sampling distribution?
    b. What happens when you use different distributions to model the population?
    c. What happens when you change the sample size?
    d. Choose at least one other statistic (eg. Median / Variance). Modify the code as required and record your observations and conclusions in the context of the above questions.
2. For the dataset `E2.csv`:
    a. Visualize it, and record your observations and conclusions.
    b. Calculate and analyze the Pearsons Correlation Coefficient. Link it with your above observations.
    c. Modify the 'y' values as follows to create 5 distinct datasets as described below:

    - Two datasets where the standard deviation of y is reduced to 5 and 10, respectively.
    - The original dataset
    - Two datasets where the standard deviation of y is increased to 20 and 25, respectively.

    Use the following formulas to modify y:

    - To reduce the spread:

$$y_{\text{new}} = y \times \left( \frac{\text{desired std}}{\text{current std}} \right)$$

    - To increase the spread:

$$y_{\text{new}} = y + \text{noise}, \quad \text{where noise} \sim N(0, \sigma_{\text{noise}})$$

$$\sigma_{\text{noise}} = \sqrt{\text{desired std}^2 - \text{current std}^2}$$

    For each variant, fit a linear regression model and report the following in a Table for both, **train and test** data:

    - $R^2$, `p-values`, F-statistic, and Root Mean Squared Error (RMSE).

    d. Analyze the Table to understand the impact of variance / standard deviation on the model and its metrics. Create plots (metrics v/s standard deviation) to aid your analysis.
    e. In each noisy variant, explore the implications of the standard deviation on the confidence intervals associated with the regression coefficients. Specifically, interpret the width of the confidence intervals and what it suggests about parameter uncertainty.

**Submission guidelines:**

- Submit a concise, well-structured report with answers to all questions. Include tables, plots, and analyses.
- Attach the spreadsheet and/or Python notebook used for calculations.
- File naming: E2-YourRollNo.