DSC 241: Statistical Models, Assignment 1

Winter 2026

Instructions

1. **Deadline.** The submission deadline is *January 16 at 22:00:00.* Late submissions will incur a penalty of 1 point per hour (0.8 from Problems and 0.2 from the Project), applied for up to 50 hours. Submissions received after *January 18 at 23:59:59* will not be accepted under any circumstances.
2. **Use of Large Language Models (LLMs).** The use of LLMs is *permitted and encouraged.* You may use LLMs to help explain the questions, obtain hints, and/or generate draft solutions for reference. *LLMs may produce incorrect or misleading content, and you remain fully responsible for the correctness of your submitted answers.* If you use an LLM, you must explicitly disclose this and clearly describe how it was used. If you do not use an LLM for one or more questions, you must explicitly state this as well. You are not required to provide verbatim transcripts; instead, briefly summarise your LLM usage and the type of assistance it provided. *Failure to provide an accurate and genuine disclosure constitutes a violation of academic integrity and may result in point deductions or a nullified grade.*
3. **Clarity and Conciseness.** Answers should be clear, concise, and directly address the questions. *Redundant, irrelevant, unclear, or unnecessarily verbose responses may result in point deductions.*
4. **Quality of graphics.** Where applicable, present high-quality graphics, charts, and tables with appropriate axis labels, legends, colors, line styles, and font sizes. Avoid unnecessary duplication of figures when results can be clearly and effectively combined for comparison. *Redundant, irrelevant, unclear, or unnecessary graphics may result in point deductions.*

Problems (80 points)

1. **(20 points) [Theory]** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. Consider the following family of estimators for λ:

$$\hat{\lambda}_{\alpha,\beta} = \frac{\sum_{i=1}^n X_i + \alpha}{n+\beta} \qquad \alpha, \beta \geq 0.$$

Note that the sample mean estimator $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$ is recovered as a special case when $\alpha = \beta = 0$.

(i) Compute the expectation $\text{E}[\hat{\lambda}_{\alpha,\beta}]$ and the variance $\text{Var}[\hat{\lambda}_{\alpha,\beta}]$.

(ii) Determine for which values of (α, β) the estimator $\hat{\lambda}_{\alpha,\beta}$ is biased.

(iii) Compute the mean squared error (MSE) of $\hat{\lambda}_{\alpha,\beta}$, defined as $\text{MSE}[\hat{\lambda}_{\alpha,\beta}] = \text{E}\left[(\hat{\lambda}_{\alpha,\beta} - \lambda)^2\right]$.

(iv) Use the previous results to explain why sometimes one might prefer using $\hat{\lambda}_{\alpha,\beta}$ over $\hat{\lambda}$ or viceversa.

2. **(30 points) [Coding]** Using the setup from Problem 1, consider $\lambda \in \{1, 10, 100, 1000\}$ and set $n = 100$.

(i) For each value of λ, plot the mean squared error $\text{MSE}[\hat{\lambda}_{a,b}]$ as a function of $\beta \in [1, 1000]$ on a log–log scale, for each $\alpha \in \{1, 10, 100, 1000\}$. Each value of λ should correspond to a single plot containing multiple curves (one for each α). Present your code and plots.

(ii) Briefly describe the main observations.

(iii) Do the empirical trends agree with the theoretical conclusions obtained in Problem 1?

3. **(30 points) [Coding]** Using the setup from Problem 1, fix $\lambda = 1$. Perform a Monte Carlo simulation with $N = 1000$ independent repetitions for every pair of $\alpha, \beta \in \{1, 10\}$. In each repetition, simulate $\hat{\lambda}_{\alpha,\beta}$ with $n = 10$ generated samples.

    (i) Plot the sampling distribution of $\hat{\lambda}_{\alpha,\beta}$. Indicate its central 95% sampling confidence interval (CI) (i.e., 2.5% and 97.5% quantiles of the samplings). Present your code and plots.

    (ii) Compute its empirical expectation $\hat{\mathrm{E}}[\hat{\lambda}_{\alpha,\beta}]$ and its empirical variance $\widehat{\mathrm{Var}}[\hat{\lambda}_{\alpha,\beta}]$. Indicate its central 95% Wald CI, and compare it with the sampling CI in part (i). Present your code.

    (iii) Briefly describe the main observations.

    (iv) Do the Monte Carlo simulations agree with the theoretical conclusions obtained in Problem 1?

    (v) Are there any peculiar features showing up in the plots? If yes, describe them and explain why they happen.

Project (20 points)

1. **(10 points)** Choose a dataset from those offered in the Canvas project page. Write a few sentences explaining why you found this dataset particularly interesting.

2. **(10 points)** Find the research article associated with the data. Based on the article, identify the variables in the dataset. Write a table summary of the variables indicating their name, a short description of what they measure, the variable type (e.g., nonnegative integer, real number, binary, categorical, ordinal, etc.) and the units of measurement where relevant.