Math 126 Problem 6: Set Size Estimate Accuracy

Spring 2027

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Monday, February 15, 2027 (Pacific time).**

- You are encouraged to discuss ideas and work with your classmates. However, you *must acknowledge* your collaborators at the top of each solution on which you collaborated with others and you *must write* your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown.
- I recommend that you write your solutions in a Python notebook using Google Colab.
- You should submit your solutions as a *single PDF* via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by *not* grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.
Original notebook: Google Colab.

Set Size Estimate Accuracy

Recall the set size estimation problem, where we draw m uniform samples from a set of size n. Let D be the number of duplicates in these samples that we observed. On the notes, we claimed that:

If we make $m \geq c\frac{\sqrt{n}}{\epsilon}$ samples for a particular constant c, then the estimate $\widehat{n} = \frac{m(m-1)}{2D}$ satisfies $(1-\epsilon)n \leq \widehat{n} \leq (1+\epsilon)n$ with probability 9/10.

Show this claim using a similar approach to our distinct elements analysis.