CSCI 145 Problem 1: The Correlation Floor

July 19, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 1: The Correlation Floor

A poll asks $n = 10{,}000$ randomly selected people a yes/no question. The textbook margin of error for a poll this size is about one point, and it shrinks like $1/\sqrt{n}$: pay for more respondents, get a better answer. Yet real polls of this size routinely miss the true answer by three points or more. In this problem, we will find out why, and discover that no number of respondents can fix it.

Model person i's answer as a random variable X_i with $\mathrm{Var}(X_i) = \sigma^2$ (for example, $X_i = 1$ if person i says yes and $X_i = 0$ otherwise). The poll reports the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Part A: Variance of a Sum

In class, we proved that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$. Generalize this identity to n random variables:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j),$$

where the second sum ranges over all ordered pairs $i \neq j$. (Both tools are in the reading: $\mathrm{Var}(Z) = \mathrm{Cov}(Z,Z)$, and covariance is *bilinear*, so it distributes over sums in each of its two arguments just like foiling a product.)

Part B: Correlated Voices

Independence between respondents is a fiction: they share news sources, the poll reaches some groups more easily than others, and people are more willing to admit some opinions to a stranger than others. Model this shared error by making the answers *equicorrelated*: $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$ for all $i \neq j$. Since all the X_i share the standard deviation σ, this ρ is exactly the correlation from lecture; take $0 \le \rho < 1$. Show that

$$\mathrm{Var}(\bar{X}) = \frac{(1-\rho)\sigma^2}{n} + \rho\sigma^2.$$

(Lecture's $\mathrm{Var}(aZ) = a^2\,\mathrm{Var}(Z)$ handles the $\frac{1}{n}$ out front.) Check your formula on the independent case: $\rho = 0$ should give $\mathrm{Var}(\bar{X}) = \sigma^2/n$, which for a yes/no question ($\sigma \le \frac{1}{2}$) and $n = 10{,}000$ is a standard deviation of at most half a point, the textbook margin. What happens to each of the two terms as $n \to \infty$?

Part C: The Floor

Take $\sigma = \frac{1}{2}$ and $n = 10{,}000$. How large must ρ be for the standard deviation of $\bar{X}$ to reach three points, i.e., 0.03? (Part B tells you which term you may ignore.) Your ρ should be far too small to notice between two individual respondents. Conclude in one sentence why the pollster cannot fix the miss by polling more people, and say what *would* have to change instead.

Where this goes next. This calculation seeds the course's variance-reduction thread. Next lecture correlation switches sides: a quantity *correlated* with the one we care about becomes a tool for driving variance down, priced by exactly the ρ you met here. The floor returns when we study stochastic gradient descent, where near-duplicate examples in a batch are correlated voices.