Discretionary Note

Anish Krishna Lakkapragada

IF YOU USE THIS FILE TO CHEAT, YOU ARE NOT ONLY STUPID BUT YOU ARE CHEATING YOURSELF OUT OF THE ABILITY TO FALL IN LOVE WITH MATH. Furthermore, I am not smarter than you and my solutions did not always get a perfect score.

CONTENT STARTS ON NEXT PAGE.

To access the general instructions for this repository head `here`.

S&DS 242/542: Homework 1

Due Wednesday, January 22, at 1PM

Gradescope and Piazza. Enroll in our course on `gradescope.com/courses/945275` (access code VDN8J5) and `piazza.com/yale/spring2025/sds242542`.

Late days. If you are using late days for this assignment, please indicate at the top of your assignment the number of late days used.

1. **Stratified sampling.** Let N be the number of registered voters in Connecticut. Suppose, for some unknown $\theta \in (0, 1)$ and $\delta > 0$, that $N/2$ of the voters are female, a fraction $\theta + \delta$ of whom support Kamala Harris. The remaining $N/2$ voters are male, a fraction $\theta - \delta$ of whom support Harris. Thus the total fraction of Harris supporters is θ, but there is more support for Harris among female voters.

We survey n voters by choosing a simple random sample of $n/2$ female voters and (independently) a second simple random sample of $n/2$ male voters. Let $\hat{\theta}$ be the total fraction of Harris supporters among all n surveyed voters.

(a) What is the bias of $\hat{\theta}$ as an estimate of θ?

(b) What is the variance of $\hat{\theta}$?

(c) Recall from lecture that, if we had chosen a single simple random sample of size n from the total population of all N voters, then we would have

$$\mathrm{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Supposing that N is much larger than n, is your variance in part (b) larger or smaller than this quantity? By approximately how much?

2. **Survey bias.** Let N be the number of registered voters in Connecticut, a fraction θ of whom support Kamala Harris. We survey n voters, for simplicity here *with replacement*. (Thus each surveyed person is chosen independently at random, and the same person might be chosen more than once.)

Unknown to us, the survey may have a bias, where each Harris supporter is chosen with probability p/N and each non-Harris supporter with probability q/N, with $\theta p + (1-\theta)q = 1$. Let $\hat{\theta}$ be the fraction of Harris supporters among our n surveyed voters.

(a) What is the bias of $\hat{\theta}$ as an estimate of θ? For what value of (p, q) is $\hat{\theta}$ unbiased?

(b) Suppose $\theta = 0.5$, $p = 1.02$, $q = 0.98$. Apply the Central Limit Theorem to approximate the sampling distribution of $\hat{\theta}$ by a normal distribution, when the sample size n is large. Under this approximation, what is the value of $\mathbb{P}[\hat{\theta} > 0.5]$ when $n = 100$? When $n = 1000$? When $n = 10000$? Comment briefly on how this changes with n.

[For these calculations, you may use any software or website that computes tail probabilities of the normal distribution.]

3. **Uncorrelated but not independent (Rice 4.59.)** Let (X, Y) be a random point uniformly distributed on the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$. Show that $\mathrm{Cov}[X, Y] = 0$. Explain briefly why X and Y are not independent.

4. **Bivariate normal.** Let $X, Y \sim \mathcal{N}(0, 1)$ be two independent standard normal random variables. Compute $\mathbb{P}[X + Y > 0 \mid X > 0]$. (Hint: Visualize the joint PDF of (X, Y) on the plane, and use that it is rotationally symmetric about the origin.)