Homework 3

Due Feb 18 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages).

1. (Markov's and Chebyshev's inequalities are tight) In this problem we show that Markov's and Chebyshev's inequalities cannot be improved without further assumptions, because there exist random variables for which they are tight.

   (a) For any $c > 0$ and any $0 < \theta < 1$, build a nonnegative random variable $\tilde{a}$ such that
   $$\mathrm{P}\left(\tilde{a} \geq c\right) = \theta = \frac{\mathrm{E}\left[\tilde{a}\right]}{c}. \tag{1}$$

   (b) For any $c > 0$, any $0 < \theta < 1$ and any $\mu \in \mathbb{R}$, build a random variable $\tilde{b}$ with mean μ and finite variance, such that
   $$\mathrm{P}\left(|\tilde{b} - \mu| \geq c\right) = \theta = \frac{\mathrm{Var}[\tilde{b}]}{c^2}. \tag{2}$$

2. (Online poll) In online polls, young people are often overrepresented. In this problem we study how to correct for this. When answering the questions use the following notation: α is the proportion of young people (between 18 and 35 years old) in the population, θ_1 the proportion of young people in the population who will vote for the Democratic candidate, θ_2 the proportion of old people in the population who will vote for the Democratic candidate, n_1 the number of young people in the poll, and n_2 the number of old people in the poll. Assume that α is known.

   (a) Derive an estimator of the proportion of voters that will vote for the Democratic candidate, as a function of the number of young people y and the number of old people o in the poll that intend to vote Democrat.

   (b) Evaluate your estimator for a poll with 100 participants where 60 intend to vote for the Democratic candidate. Out of the 100 participants, 70 are young, and 50 of them intend to vote for the Democratic candidate. The fraction of young people among voters in general is 25%.

   (c) Under what assumptions is your estimator unbiased? Justify your answer mathematically.

   (d) Show that your estimator is consistent as $n_1 \to \infty$ and $n_2 \to \infty$.

3. (Blood Pressure) The table in `cardio.csv` records the systolic blood pressure (*"ap_hi"*) of patients. Randomly sample subsets consisting of $0.1\%, 0.2\%, \cdots, 99.9\%$ of the full dataset.

(a) Compute and plot the bias and variance of these subsets as a function of the number of samples. Interpret your findings.

(b) Approximate the probability that *ap_hi* deviates from the corresponding population mean via Monte Carlo simulations, and compare it to the Chebyshev bound that we use to prove the law of large numbers.