# Homework 8

Due November 21 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.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. 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). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Short questions)

   (a) Give an example of a nonnegative random variable $\tilde{a}$ and a constant $c > 0$ for which $\mathrm{P}(\tilde{a} \geq c) = \mathrm{E}(\tilde{a})/c$. What does this say about Markov's inequality?

   (b) In the notes, we have defined the sample variance of a dataset $X := \{x_1, x_2, \ldots, x_n\}$ as

   $$\sigma_X^2 := \frac{\sum_{i=1}^n (x_i - \mu_X)^2}{n}, \tag{1}$$

   where $\mu_X$ is the sample mean. Show that this is not an unbiased estimator of the true variance if the data are i.i.d. samples from a distribution with zero mean and variance $\sigma^2$. Explain how to fix the estimator so that it is unbiased.

2. (Poll) In an online poll before an election, 60 participants intend to vote for the Democratic candidate, and the remaining 40 intend to vote for the Republican candidate.

   (a) The number of young people (between 18 and 35 years old) in the poll is 70. 50 intend to vote for the Democratic candidate. The fraction of young people among voters in general is 25%. Provide an estimate of the proportion of voters that will vote for the Democratic candidate.

   (b) Under what assumptions is your estimate unbiased? Justify your answer mathematically.

   (c) Let $\alpha$ be the proportion of young people in the population, $\theta_1$ the proportion of young people who vote for the Democratic candidate, and $\theta_2$ the proportion of old people who vote for the Democratic candidate. If the number of young people in a poll is $n_1$ and the number of old people is $n_2$, what is the variance of your estimator if the assumptions from the previous question hold?

3. (Length of confidence interval) We are interested in estimating the mean height in a population from a finite set of random samples. We would like to have a 95% confidence interval for our estimate of width equal to 5 cm.

(a) Use Chebyshev's inequality to determine how many samples we need to take. Explain any assumptions you make.

(b) Use the central limit theorem to determine how many samples we need to take, assuming that the sample standard deviation of the data equals 10 cm.

4. (Radioactive sample) Consider the following experiment. We have a radioactive sample situated at unit distance from a line of sensors. Each time a sensor detects a particle emitted from the sample we obtain a reading of the position of the sensor in the $x$ axis (we assume that we have so many sensors that you can model this position as a continuous random variable). We model the measurements as an i.i.d. sequence distributed as a random variable $\tilde{m} = c + \tilde{x}$ where the pdf of $\tilde{x}$ is symmetric around the origin, that is $f_{\tilde{x}}(x) = f_{\tilde{x}}(-x)$ for all real numbers $x$. Your task is to estimate the position of the sample $c$ from these data.

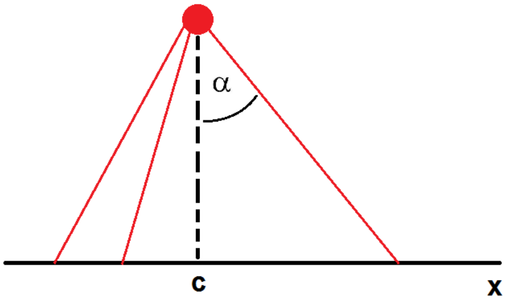

Figure 1: Diagram of the experiment.

(a) The file *radioactive_sample_1.txt* contains a vector of measurements $m_1, m_2, \ldots$. Plot a moving average of the measurements $\frac{1}{n}\sum_{i=1}^{n} m_i$ for $n = 1, 2, 3, \ldots$ (and submit the plot). Use the plot to give an estimate for the value of $c$. (Hint: What is the expected value of $\tilde{m}$?)

Under what assumptions on $\tilde{x}$ can you prove that the estimation method you propose work?

(b) The file *radioactive_sample_2.txt* contains a vector of measurements corresponding to a different radioactive sample. Does the estimation method described above work in this case? Submit the plot of the new moving average.

(c) A colleague suggests that the angle $\alpha$ between the trajectory of the particles emitted by the new sample and the vertical axis (illustrated in Figure 1) might be well modeled by a random variable $\tilde{a}$ that is uniformly distributed between $-\pi/2$ and $\pi/2$. Compute the pdf and mean of $\tilde{x}$ under this assumption.
(Hint: remember the trigonometric function tan and its inverse arctan.)
Would such model explain your observations in (b)?

(d) The sample mean can be affected by extreme values and outliers, whereas the sample median is more robust. The sample median converges to the median of an iid sequence of random variables even when the mean is not well defined. Use the sample median from *radioactive_sample_2.txt* to estimate $c$.