Homework 5

Due Tuesday, October 20

Please either give the assignment to Loraine at the CDS or send it via email to the graders **before noon**.

1. *Roulette* Bob is playing roulette in Las Vegas. He has to choose between betting on a number and on a color. His criterion is that he wants to maximize the probability of making money if he plays 100 times in a row, betting a dollar each time. In a roulette there are 38 numbers: 18 black, 18 red and 2 green (0 and 00).
   a. Approximate the probability Bob is interested in if he bets on a number by using the Central Limit Theorem. This bet pays 35:1 (you win $35 for each dollar you bet). Explain every assumption you make clearly.
   b. Approximate the probability Bob is interested in if he bets on red or black by using the Central Limit Theorem. This bet pays 1:1 (you win $1 dollar for each dollar you bet).
   c. Which bet is better asymptotically, for a very large number of bets? To evaluate which bet is *better*, compare the amount of money gained per bet as the number of bets tends to infinity.
   d. What bet will Bob choose? He chooses that bet and actually makes money. Approximate his expected gain conditioned on this information using the Central Limit Theorem.
2. *Weight* The file *weights.txt* contains the weights of 25 000 individuals. We are interested in determining an interval of width 5 lb that includes the mean weight in the population with 95% confidence (i.e. we want a 95% confidence interval of width 5 lb), sampling as few individuals as possible.
   a. How many individuals do we need to sample if we want to make sure that we have a confidence interval that holds rigorously? Jon Brower Minnoch, the heaviest man ever recorded, weighed 1400 lbs.
   b. You compute the sample variance using 1000 samples and it turns out to be 144. How many individuals do we need to sample to compute an approximate confidence interval using this sample variance?
   c. Complete the script *confidence_intervals.py* and run it. The script first plots 100 approximate confidence intervals using 20 and 1000 samples and compares them to the true mean. Then it generates 10 000 more confidence intervals for both values and counts the number of confidence intervals that do not contain the true mean. Finally, it plots the empirical distribution of the sample variance. Report the results (you don't need to include the plots or the code), compare them to the theoretical prediction and explain any discrepancies.
3. *Convergence in probability implies convergence in distribution.* Prove that convergence in probability implies convergence in distribution for continuous cdfs following these steps:
   a. Show that for any random variables A_n and A, any real number a and any $\epsilon > 0$,

$$\mathrm{P}\left(A \leq a - \epsilon\right) - \mathrm{P}\left(|A_n - A| > \epsilon\right) \leq \mathrm{P}\left(A_n \leq a\right) \leq \mathrm{P}\left(A \leq a + \epsilon\right) + \mathrm{P}\left(|A_n - A| > \epsilon\right)$$

b. Assume that $A_n \to A$ in probability. Use the expression derived in (a) to prove that $A_n \to A$ in distribution. Hint: For points a at which F_A is continuous,

$$\lim_{\epsilon \to 0} F_A(a \pm \epsilon) = F_A(a).$$

c. Does convergence in distribution imply convergence in probability? Prove it, or provide a counterexample.

4. *Radioactive sample.* Consider the following experiment. We have a radioactive sample situated at unit distance from a line of sensors. We assume that there are so many sensors that they cover the entire x axis. 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 model the measurements as i.i.d. samples of a random variable $M = c + X$ where the pdf of X is symmetric around the origin. Your task is to estimate the position of the sample c from these data.

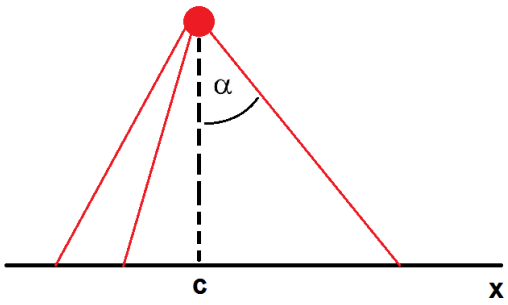

Figure 1: Diagram of the experiment.

a. The file *radioactive_sample_1.txt* contains a vector of measurements $m_1, m_2, \ldots$. Plot the running sample mean $\bar{m}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$ for $n = 1, 2, \ldots$. Give an estimate for c. Under what assumptions does this estimation method work?

b. The file *radioactive_sample_2.txt* contains a vector of measurements corresponding to a different radioactive sample. Does your method work? Submit the plot of the new running sample mean.

c. A colleague suggests that the angle α 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 A that is uniformly distributed between $-\pi/2$ and $\pi/2$. Compute the pdf and mean of X under this assumption. Would this model explain your observations in (b)?

d. Sample four different i.i.d. vectors of length 10^4 from the distribution of X that you computed in (c). Plot their sample running averages in the same graph. What do you observe?

5. *Election poll* You are hired to analyze a poll right before an election. You model the ith person in the poll using the random variable

$$R_i := \begin{cases} 1 & \text{if Republican,} \\ 0 & \text{if Democrat.} \end{cases} \tag{1}$$

a. How would you use the sample mean of the sequence $R_1, R_2, \ldots$ to estimate the outcome of the election? Under what assumptions would this work?

b. 49000 people said they would vote Democrat, 45000 said they would vote Republican. Who do you think will win the election? In order to quantify the accuracy of your estimate, upper bound the probability of the sample mean deviating from the true mean so much that you could be wrong (the upper bound should hold rigorously, do not use an approximation)? Does it make sense to say that this is an upper bound on *the probability that you are right*? (Hint: You can bound the variance of a Bernouilli random variable by 1.)

c. You realize that part of the poll was carried out online and you suspect that young people (under 35 years old) might be overrepresented. You decide to model the ith young person and the ith old person using the random variables

$$Y_i := \begin{cases} 1 & \text{if Republican,} \\ 0 & \text{if Democrat,} \end{cases} \qquad O_i := \begin{cases} 1 & \text{if Republican,} \\ 0 & \text{if Democrat.} \end{cases} \tag{2}$$

Then you consider the estimator

$$X_{n_1,n_2} = a \sum_{i=1}^{n_1} Y_i + b \sum_{j=1}^{n_2} O_i, \tag{3}$$

where n_1 is the number of young people and n_2 the number of old people. What should a and b be equal to so that X_{n_1,n_2} is an unbiased estimator of the fraction of people that vote Republican? Express your answer in terms of the number of young/old people that vote Republican or Democrat.

d. 21000 out of a total of 35000 old people and 24000 out of 59000 young people taking the poll said they would vote Republican. Also according to the census, about 20% of the American population is between 18 and 35 years old, whereas 55% is above 35. Now, who would you predict will win the election? State any assumptions you make and quantify the precision of your estimate as in part b (the upper bound should hold rigorously, do not use an approximation).