EE 325: Probability and Random Processes

Assignment 1

Submission deadline on Moodle: Midnight, Thursday, August 11.

Three simple computation experiments are outlined below. You can use any programming language that you are comfortable with. The key objective of this experiment is to make you think about many of the questions that are given at the end. If you have done some of this formally through other means, this will be good revision. If you are seeing these questions for the first time, then do spend time thinking about them as a means of knowing the motivations for many of the concepts that we develop in this course. We will formally address them as we proceed through the course. Of course, increasing degrees of complexity will be developed. Finally, real life data is not as cleanly available as the examples suggest. Some examples will be introduced as we progress through the course.

1. There are 10,000 students in a IITB. A cellphone company is considering designing some deals for the students and wants to determine the average data consumption (in GB) per month per student on the campus. This is called the *population average.* Collecting the data from every student is costly, time consuming, and can be prone to errors. It may not even be possible. The obvious 'short cut' is to obtain the data from a small number of students, take the mean of the collected samples, and declare that that is the true value of the data that is being sought. However, this is clearly a guess. For the guess to be good, we need to answer at least two questions. Assume that the cost of collecting the data is an increasing function of K. The cellphone company has a budget and would like to keep K low and yet be accurate.

   - How to select the K students to collect their data?
   - How does the guess 'improve' as a function of K? And hence, what is a good K?

   Here are some three options for choosing the K students from whom to collect the data.

   (a) Ask the first K students that you can find as soon as you enter the campus.

   (b) Choose an *arbitrary* point in the campus and ask K students from there. Here arbitrary means that you can pick any point that you like.

   (c) *Randomly* select K from the 10,000 people in the colony. You can visualise a random selection to be the result of the following experiment. Put all the 10,000 names in a pot, mix the pot thoroughly and pick a name. Repeat K times. You can use a random number generator to pick the name.

   File `hw1a.txt` contains the 'actual' weekly data consumption from the 10,0000 users. Assume that this data is not available to you. For a given K, write a program to simulate the three scenarios described above.

(a) For each of the three scenarios, write a program to obtain the K samples from this list, and calculate the average of the K samples. The program should repeat this experiment fifty times and make a scatter-plot for each of the three scenarios, i.e., mark a dot for every one of the fifty points on a suitable scale.

(b) Repeat the experiment for $K = 10, 20, 50, 100, 200$. There is one scatter plot for every combination of K and the method of selecting the K. Now answer the following questions.

i. What is your guess for the actual average and the actual standard deviation? And what is the actual value?

ii. Each of the fifty repetitions can be seen to be a separate survey. If you could do the survey only once for a given K,

A. Which of the above three schemes would you use in practice to determine the best guess?

B. If you were allowed to choose the value of K, what value would you choose? And how sure would you be of the actual values of the average? What kind of quantitative measure would you use to describe your "sureness of the estimate from the single survey of K samples?"

Submit the following: the program, the scatter plots, and the answers to the two questions, and sub questions, above.

2. Thre is a believe that when Kohli was the captain and had to call at the toss, he always called Heads. What do you think he will assume the probability of the coin toss coming up heads to be? Assume that he believes that his opponent does not have the powers or Shakuni or is not using the coin that Amitabh Bachchan used in Sholay (if you do not know what that is, see around 3:50 *and* 13:05 `https://www.youtube.com/watch?v=QmLiyVT_lLc`. Now imagine that he sees the sequence in `hw1b1.txt` that contains the result of 100 tosses of a coin.

(a) At what point in the sequence does he begin to *doubt* his initial assumption?

(b) And when can he be sure that his initial assumption is wrong, if it is indeed wrong?

(c) Repeat for three more coins. The results of 100 tosses of these coins are in files `hw1b2.txt`–`hw1b4.txt`.

(d) At every point in the sequence, when you are determining whether you are sure of your initial assumption or not, think about how you would describe the "degree of sureness," i.e., provide a quantitative description to your 'hunch.'

Submit your program, and the written answers to the four questions above.

3. Given N pairs of data (x_i, y_i) for $i = 1, N$ you have done straight line fitting on this data of the form $y = ax + b$. Recall that you determine a and b to minimise root mean square error $\sum_{i=1}^{N}(y_i - ax_i - b)^2$. Consider the data given in `hw1c1.txt`. Here x_i represents the height of person i in centimeters of IITB student i and y_i is the weight of the person in kgs. Fifty IIT students were sampled and their height and weight recorded in the file. Write a program to fit the straight line to this data, i.e., determine a and b. The data in `hw1c1.txt` is used construct a model. Now let us see how to use this data.

   (a) Assume now that you have determined the model described above. You see a person of height x. You can use the model to predict this person's height. Let your prediction be denoted by $\hat{y}$. When you measure the actual weight, now denoted by y, what can you say about the difference between the true height y and your intelligent guess $\hat{y}$ using the linear model.

   (b) Now assume that you meet many people of height x and for each of these you guess their weight $\hat{y}$, measure the true value y and obtain the error $y - \hat{y}$. What kind of properties would you like for these errors. Specifically, comment on the average and the standard deviation of the errors. You can use the data in file `hw1c2.txt` that contains 25 values of y for three different values of x to help arrive at your answer.

   Submit the program and the answers to the the questions above.

Create one PDF file for this assignment and submit on Moodle. Every one should submit, and each member of a group can submit the same file.