EE 325: Probability and Random Processes

Programming Assignment 1

Submission deadline on Moodle: Midnight, Monday, 01 Sept 2024

Some 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.

The questions are necessarily imprecise. There is no one right answer. The idea of this homework is for you to think about the questions in a systematic and quantitative manner. Discuss among your group members. **Do not Google or use any of the AI Tools for any aspect of this homework.** Doing so defeats the purpose of the exercise.

Part 1

In the style of the American TV game "Family Feud," IITB Demographic Project asked one hundred people the following question "Where do most IITB students come from?" and we found the following data: 41 said Andhra & Telengana, 28 said Maharastra, 10 said Rajasthan, 10 said UP & Bihar, 9 said West Bengal, and 2 said that they did not know. Note that we are not saying anything about how the 100 people were selected.

1. Is the opinion from the 100 people representative of the opinion of India? Explain.
2. What is your belief about the home state most IITB students? Write the belief from each member of your batch; it is reiterated that there is no shame in being wrong.
3. To determine the 'true' answer to the question above would require collecting demographic data from students over the years and then determining the population average. Since that can be very costly, may be even impossible, an obvious 'short cut' is attempted—data from a small number from one graduating batch is collected by calling them up. The mean of the data from ths small sample of K students of one particular batch is declared as the true value of the data that is being sought. However, this is clearly a guess. The Project has a budget and would like to keep K low and yet make a guess as accurate as possible. For the guess to be good, we need to answer at least two questions.
    - 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 options for choosing the K students from whom to collect the data.

- Ask the first K students in the list for the batch.
- Choose an *arbitrary* point in the list and ask K students from that point on in the list. Here arbitrary means that you can pick any point that you like.
- *Randomly* select K from the list. You can visualise a random selection to be the result of the following experiment. Put all the 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.

The actual demographic data for the 1500 students from one such batch is available in the file `Demographic.csv`. Our interest is in knowing (1) the top three home states and (2) the combined total fraction of students from these three top states.

(a) For each of the three scenarios described above, 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.

(c) By looking at data of the entire list of 1500 students, what are the top three states and what is the fraction of students from these students?

(d) 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,

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

   ii. 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 questions and sub questions, above.

Part 2

Now you can use the data from the full batch and answer the following question. For each question below, write a short paragraph describing your calculation method and the conclusion. To help answer questions in this part, you can use the data about the demographics of the different states that is provided to you. The following is reiterated: *'do not search the web or ask AI tool and there is no such thing as the right answer.*

1. Is the IITB UG population a good sample of India at the granularity of states? If not, how badly is it skewed in favour of some states? Define **your own measure for skew** (remember, no cheating or being 'inspired' by anything that Google/ChatGPT/... can throw up) and apply to the data that you have. Justify the measure.

2. Considering the population and the per capita income of the states, is the distribution of the student body among the states/regions fair? Once again, define **your own measure for skewfairness** and apply to the data that you have. Justify the measure.

3. How strongly does the state affect the JEE rank, the graduating CPI, and the first salary?

4. How strongly does the family income affect the JEE rank, the graduating CPI, and the first salary?

The report should contain answers to the three questions in Part 1 and the four questions in Part 2. **Every one should submit their reports. Members of a batch can submit identical reports.**