**EE 325: Probability and Random Processes Programming Assignment 1**

**NOTE:**  GitHub repo link: [github.EE325\_Assignment](https://github.com/rookie-apoorv/Assignment_1.git) . Refer to the README.txt for easy navigation

**Part-1**

1. **Is the opinion from the 100 people representative of the opinion of India? Explain.**

* The following is the given opinion

|  |  |
| --- | --- |
| State | NO. of students |
| Andhra and Telangana | 41 |
| Maharashtra | 28 |
| Rajasthan | 10 |
| UP and Bihar | 10 |
| West Bengal | 9 |
| Don't Know | 2 |
| Total | 100 |

* + The above opinion is not representative of the opinion of India.
  + This is so because this set of people does not include people from many parts of the country which are known as prominent education centers, like Delhi NCR, and it also does not have any representatives from far eastern states or the south eastern part.

1. **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.**

**View of Member2 (Apoorv)**

* My belief is similar to what opinion we have been given in the first question, only that I disagree with the numbers.
* This is my belief:

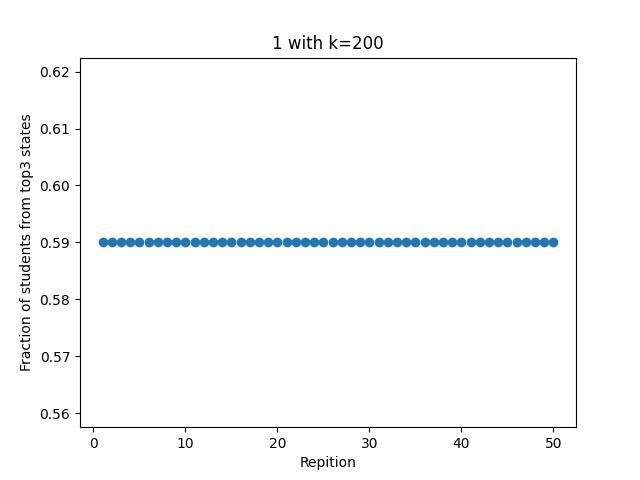
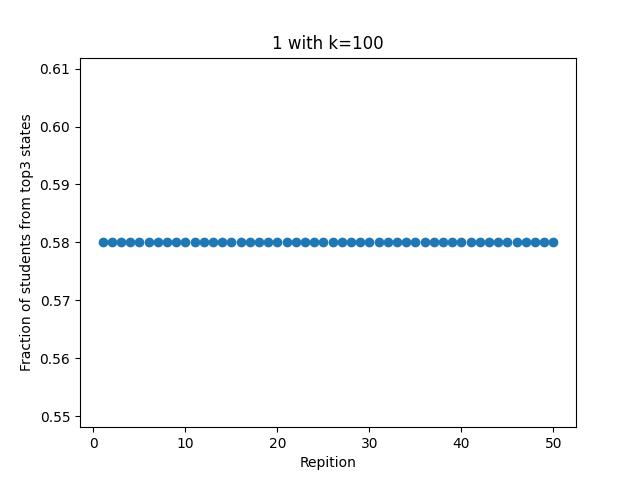
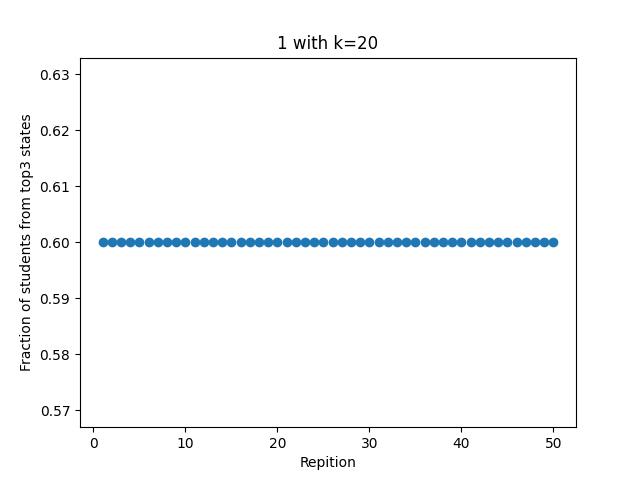
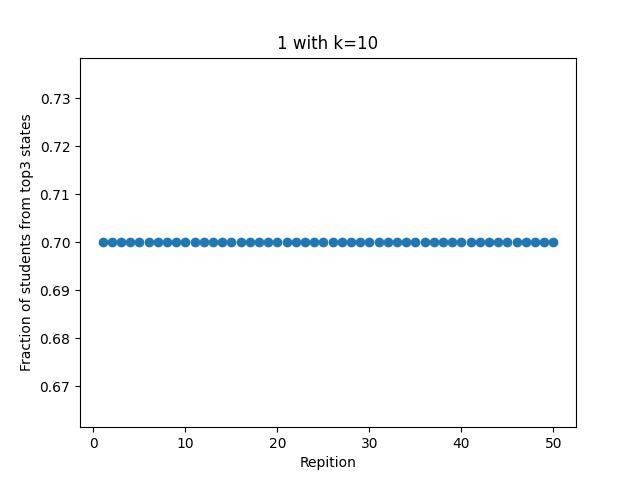
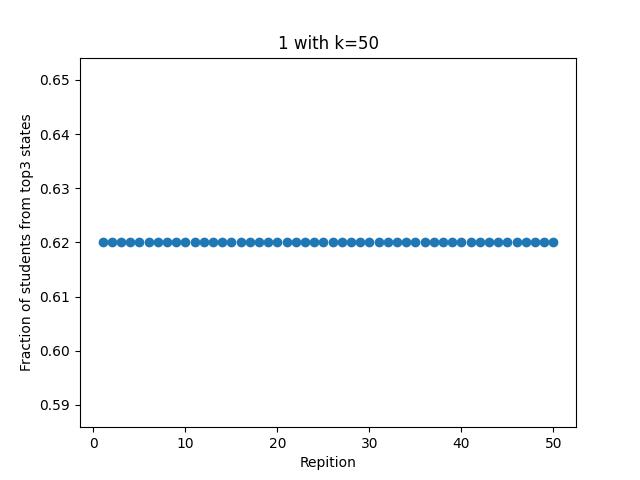
|  |  |
| --- | --- |
| **State** | **Percentage of students** |
| The Southern States  (Andhra, Telangana, Tamil Nadu, Kerala) | 30% |
| Rajasthan and Delhi NCR | 25% |
| Central and Northern States  (MP, Punjab, Haryana, UP, Bihar) | 20% |
| Maharashtra and Gujarat | 15 % |
| Others | 10% |

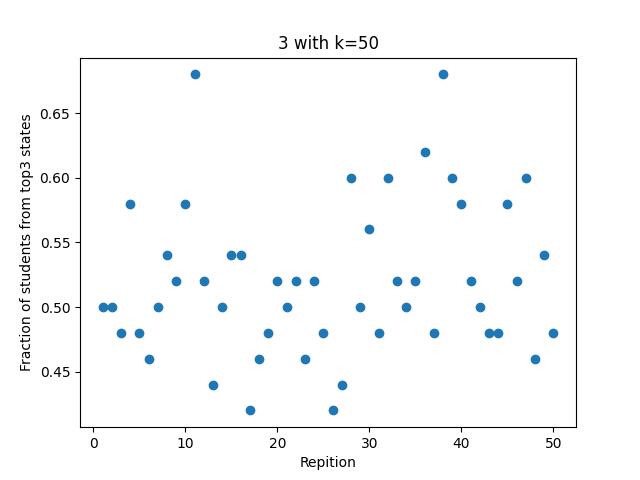
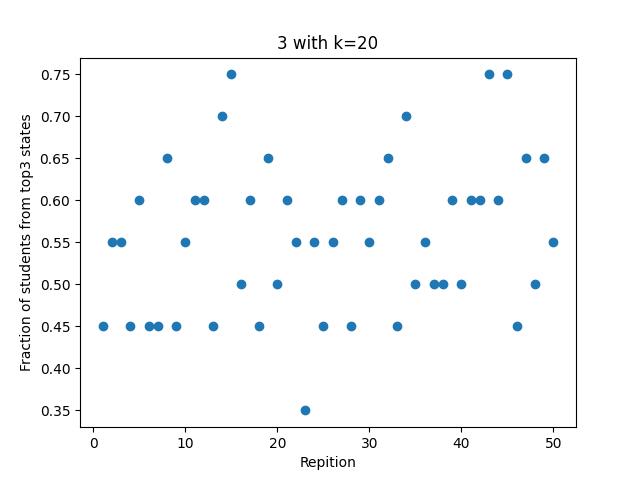
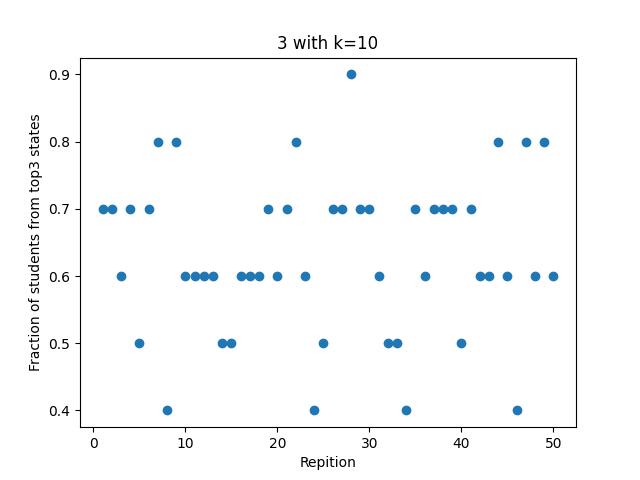
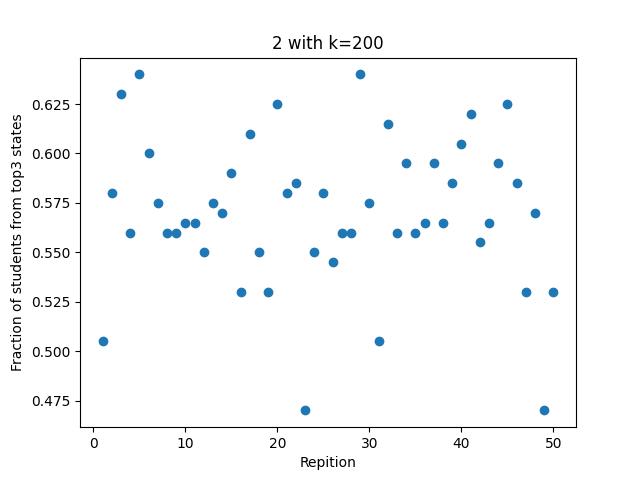
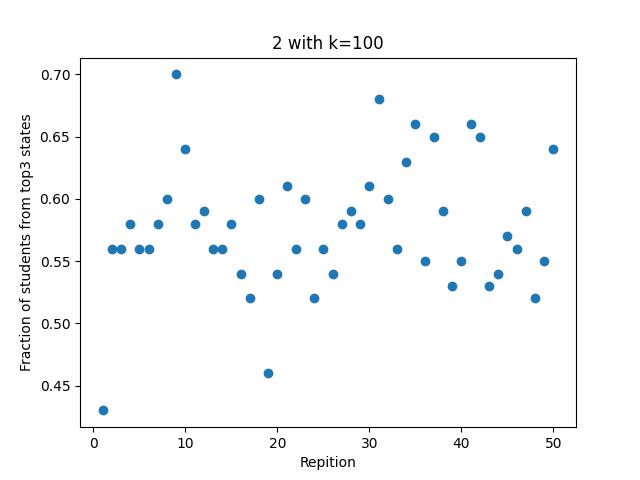
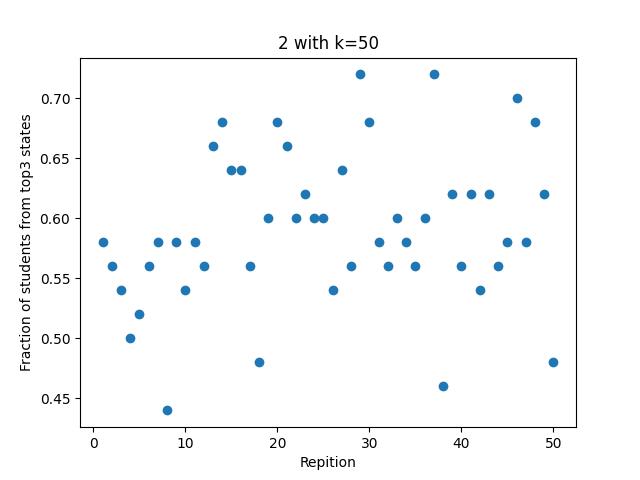
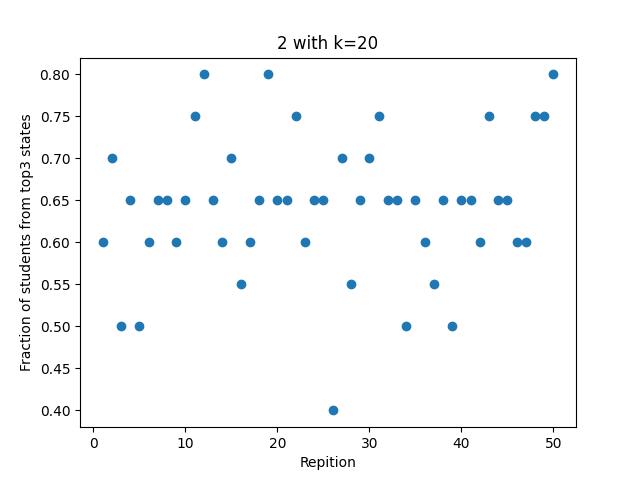
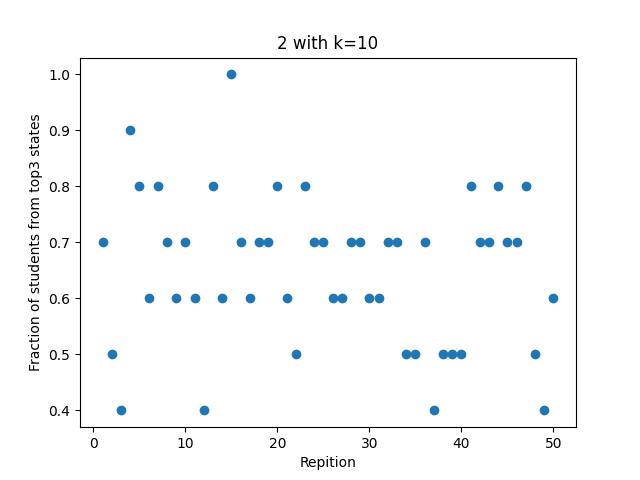
* The above belief is based on my personal experience from the people I have met in the institute.

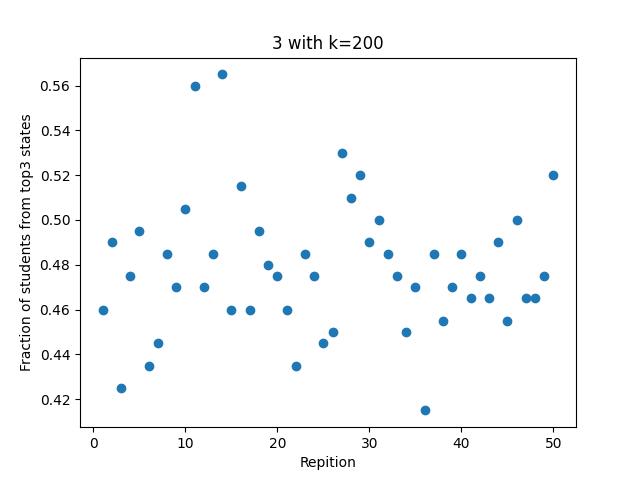
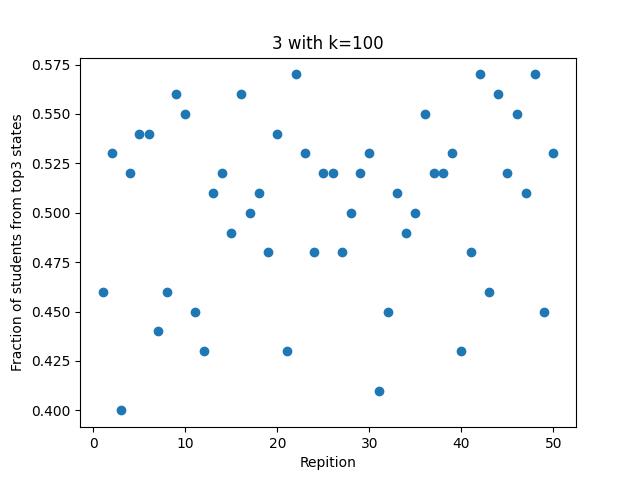
**View of Member1(Hari)**

|  |  |
| --- | --- |
| State | Percentage Students |
| The Southern states | 30% |
| Maharashtra | 30% |
| Rajasthan & Delhi | 20% |
| Others | 20% |

**3.Do the experiment stated in the assignment 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.**

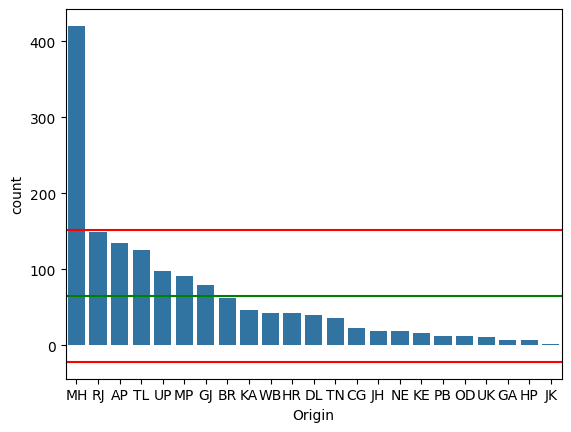
The following are the scatter plots for each way of selecting k people for each k





**NOTE:**  For better clarity these plots along with the code are also uploaded on the GitHub repository whose link is shared at the top of this document.

**3(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?**



The above shows a histogram of no. of students from each state.

* The top 3 states are:

|  |  |  |
| --- | --- | --- |
| State | No. of students | Fraction |
| Maharashtra | 420 | 0.28 |
| Rajasthan | 149 | 0.1 |
| Andhra Pradesh | 135 | 0.09 |
|  |  |  |
| Total | 704 | 0.47 |

**3(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**

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

An important point to note is that our result should be independent of in what order the data is collected.

* First k people: This is not an accurate way of measurement as this might a bias data collection. As we increase k for this model, the fraction obtained approaches the actual value of 0.47, but for this we would have make the value of k very high (maybe even 700 or more). The value greatly depends on the order of data. Hence this is not a good way.
* First K people from a random point: Although this model introduces some degree of randomness to data collection, still taking k subsequent values assigns some weight to the order in which the data was collected. Hence this is also not a good way.
* Random k data points: This has the most randomness, and the values obtained from this can be considered to be very close to the actual value. As you can see from the scatter plots, the range of values for the 5 repetitions contains the actual value and this range keeps centering towards the actual value as we increase k.

Hence, I would choose the 3rd way of selecting random k people, as this may give the best results.

1. **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?**

* We are choosing the third way of sample selection so we will focus on the results from that only.
* What is the situation?
  + We don’t know the actual value (i.e. the value 0.47).
  + Now we do the survey once (i.e. we get one of the values from the 50 repetitions)
  + We need to determine for what k would we be very sure of the correctness of the obtained value.
  + The following is the table consisting of the average of the 50 repetitions for each value of k in model 3.

|  |  |
| --- | --- |
| Value of K | Average of 50 repetitions |
| 10 | 0.65 |
| 20 | 0.58 |
| 50 | 0.52 |
| 100 | 0.49 |
| 200 | 0.48 |
| 1200 | 0.47 |

* What we are trying to do here is called sampling distribution in statistics.
* In this we don’t know the actual average of the population, so we take various samples, take the average of each sample and then take the expected value of these averages.
* We now hypothesize that the actual mean should be very close to this expected value.
* So, what we have on the right in this table is that expected value, as you can see increasing k has a very nice effect on it and it approaches the actual value.
* But is increasing k after a certain point beneficial?? Not really as you can see from the table.
* Hence 200 is a good enough k, where we are very close to the actual value with not much cost for data collection.

**How sure are we?**

* Our sureness depends on the fact of how many values of those 50 repetitions for k=200 is actually close to the expected value of 0.48 of k=200.
* For this our quantitative measurement would be standard deviation of this data.
* We calculate the standard deviation and it comes out to be
  + Sigma = 0.03
* In statistics we say that with 95% confidence the value lies in the interval expected value +- 2\*sigma.
* So, we say with 95% confidence that our value is in 2\*sigma proximity of the actual value.