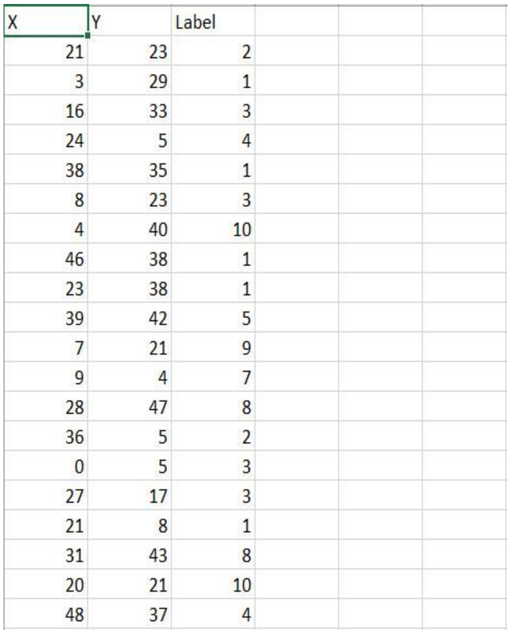
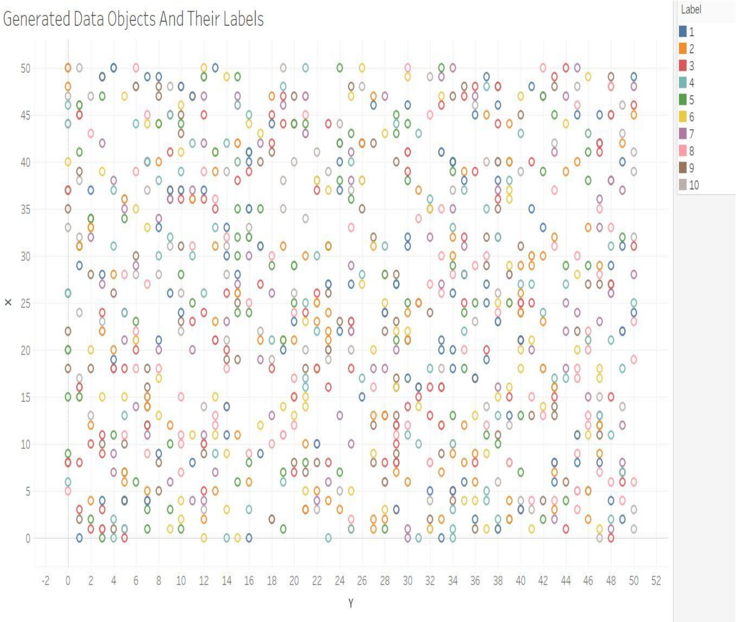
**Vineet Joshi**
**MT19020**
**Assignment 1 - Question 2**

## The Dataset:

The dataset used for these experiments, consists of 1000 data objects with two features (X and Y),each divided among 10 class labels (1-10).
A small snippet of the dataset is given below.



| X | Y | Label |
|---|---|---|
| 21 | 23 | 2 |
| 3 | 29 | 1 |
| 16 | 33 | 3 |
| 24 | 5 | 4 |
| 38 | 35 | 1 |
| 8 | 23 | 3 |
| 4 | 40 | 10 |
| 46 | 38 | 1 |
| 23 | 38 | 1 |
| 39 | 42 | 5 |
| 7 | 21 | 9 |
| 9 | 4 | 7 |
| 28 | 47 | 8 |
| 36 | 5 | 2 |
| 0 | 5 | 3 |
| 27 | 17 | 3 |
| 21 | 8 | 1 |
| 31 | 43 | 8 |
| 20 | 21 | 10 |
| 48 | 37 | 4 |



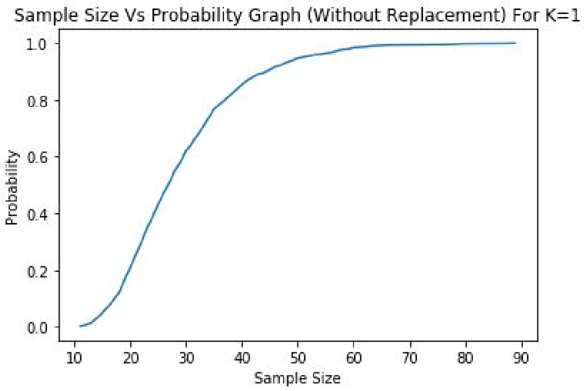
Creating the dataset was a pretty straightforward task. We filled the features for the 1000 objects with random values, between 0 to 50, using *random.randint()* function in python. The class labels are then simply distributed 100 objects at a time and the entire dataset is shuffled.

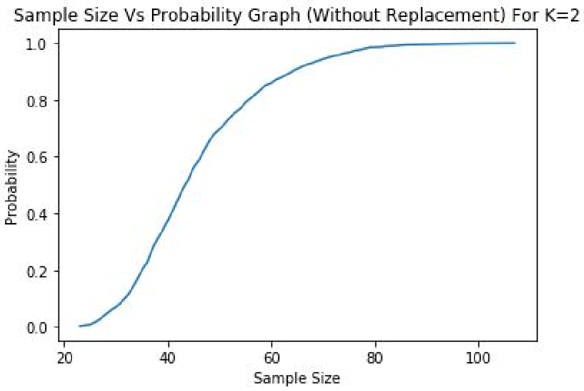
## Sampling Without Replacement

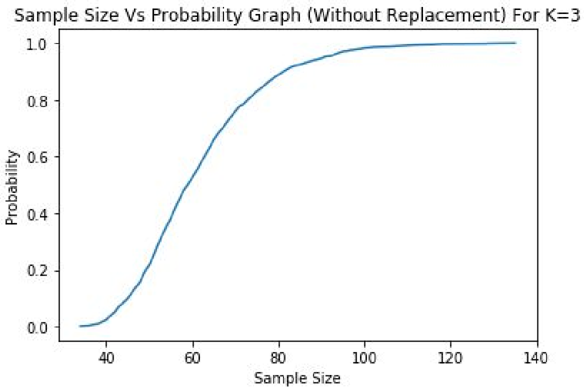
In order to plot a graph between probability and sample size, we conducted a series of 1000 experiments, in which the result of each experiment denotes the sample size for which the number of instances of each class satisfies the given K.

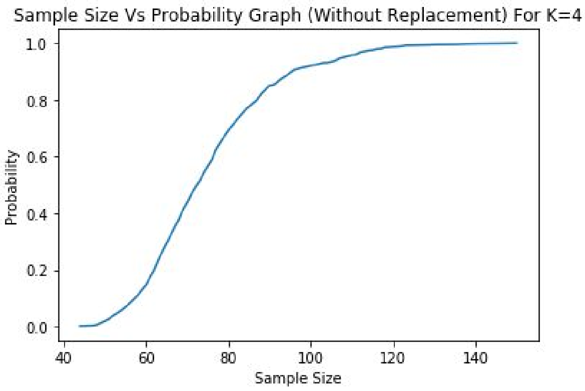
We created the sample space after 1000 such trials and calculated the cumulative probability associated with each sample size.

Using *matplotlib* library in Python, we were able to plot the graph, sample size vs. probability for K = 1,2,3 and 4.
These graphs are shown below.









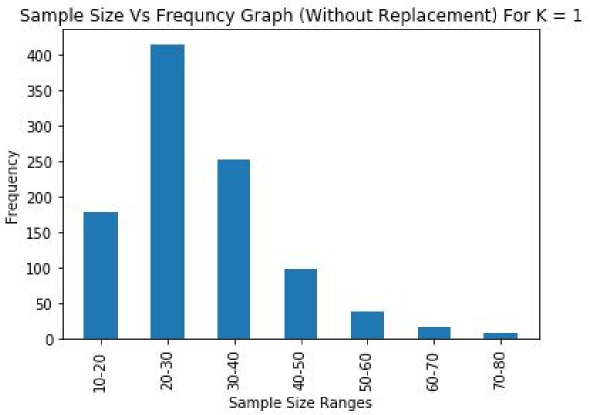
**Inferences :**

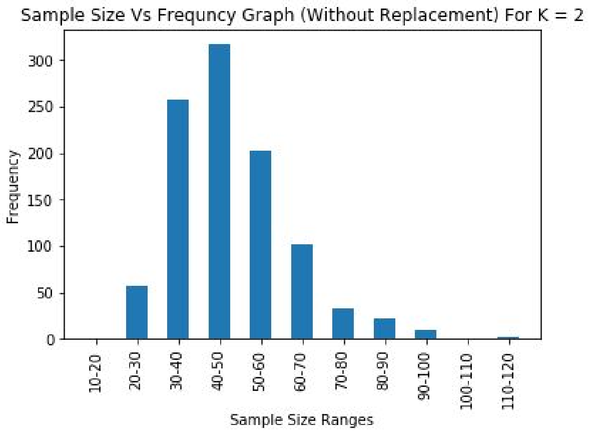
The following are a few inferences we can draw from the above graphs:

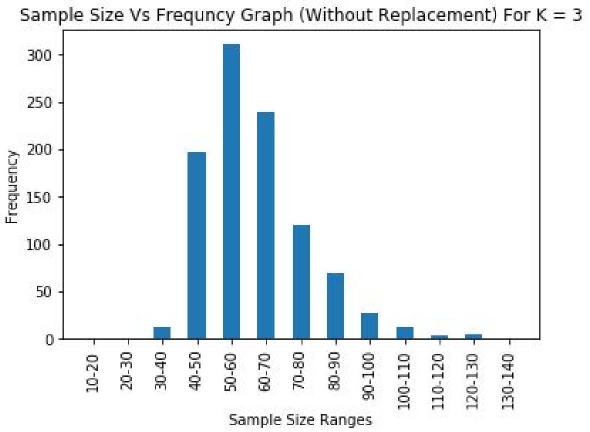
- For each value of K, there is a threshold sample size. After this, the probability of occurrence of each class at least K times in the sample is K. For example, for K=1, the threshold sample size is close to 85.
- As the values of K increases, the threshold sample size increases. This is quite understandable as we know we will now need more samples to ensure at least K instances of each class occur.

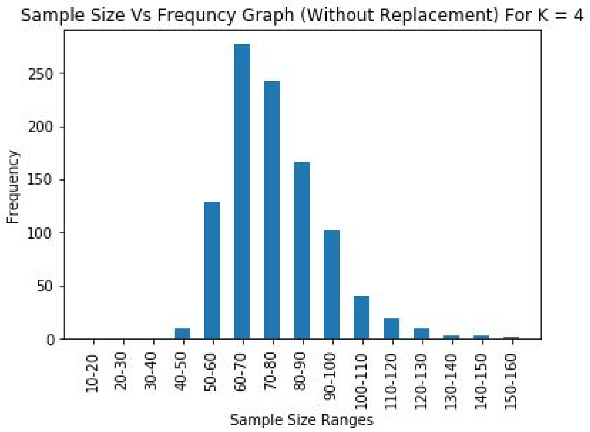
**NOTE:** This experiment was performed without replacement i.e. a data object once selected will not be selected again.
In order to achieve this, we are using *random.sample()* library in Python.

Our next task is plotting the frequency of the sample sizes in our sample space of 1000.
Below are the graphs which shows the result for various values of K.









**Inferences :**

The following are a few inferences we can draw from the above graphs:

- As the number of K increases, the variance in sample space increases.
- With the increase in values of K, the graph is shifting more and more towards the right.
- The graph follows a bell curve.

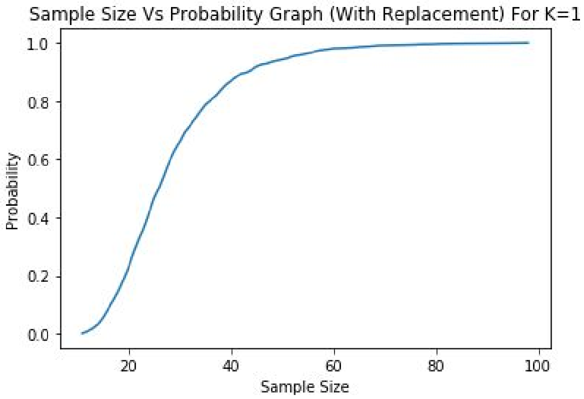
Inferences from both the graph are quite interrelated. Both of them are based on the core idea that as K increases we have to increase the sample size in order to satisfy the condition of at least K instances of each class and as sample size increases, the probability moves to 1.

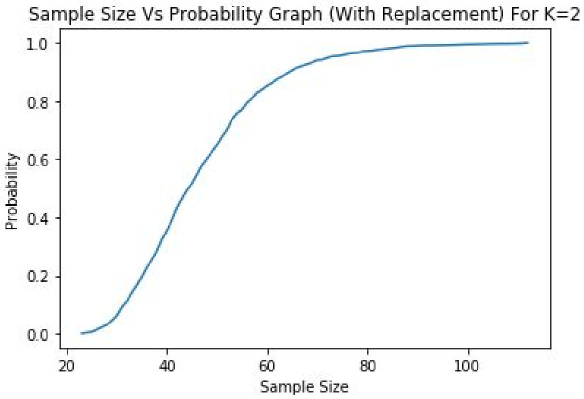
## Sampling  With Replacement

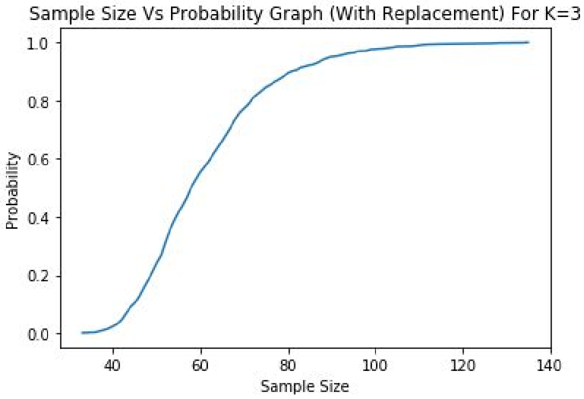
The exact same procedure is followed in this case also. The only difference being the fact that we are performing sampling with replacement, that is a data object can be selected more than once.

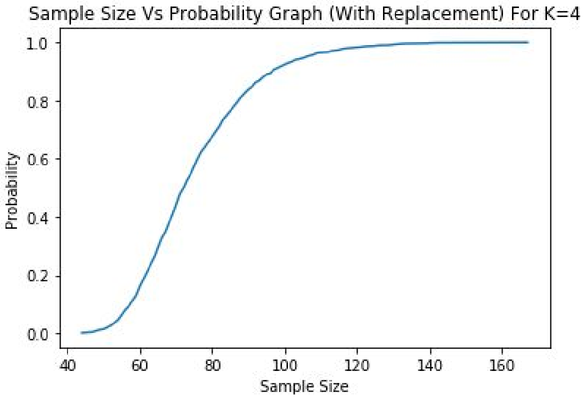
In order to sample data objects with replacement, we used *random.choices()* function in Python.

The probability vs. sample size graph for different values of K are below.



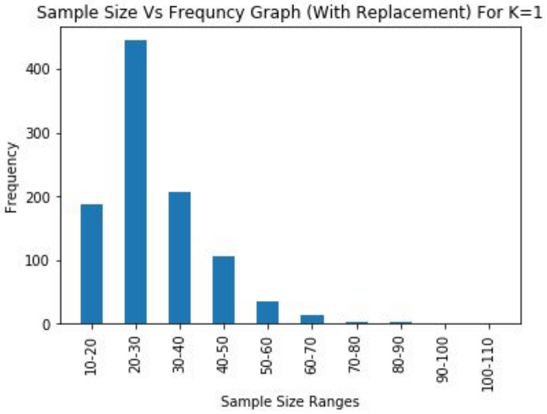


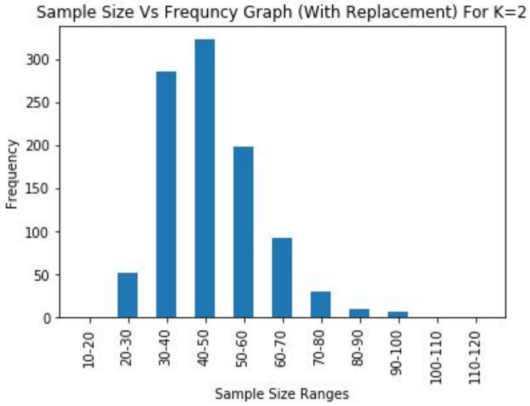


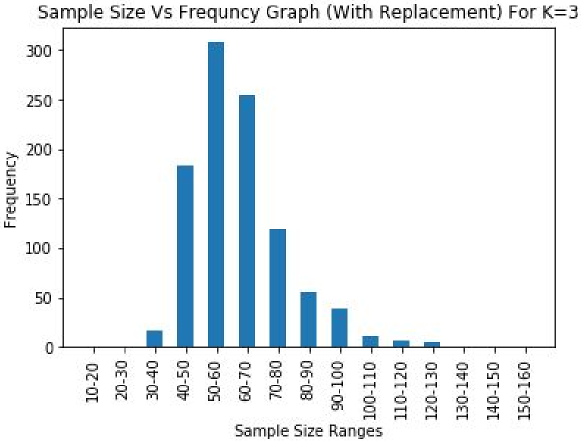


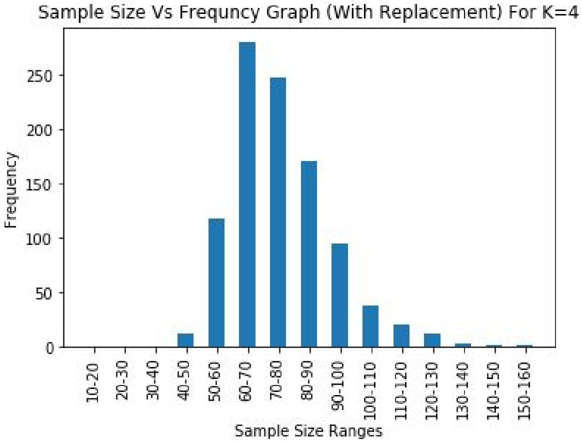
These graphs are quite similar to the once we obtained from without replacement sampling. Only difference being the fact that the threshold is increased slightly. This is mostly due to the fact that now a data object can get repeated and hence we need more samples in order to be 100% sure that all classes are included at least K times.

Our next task is plotting the frequency of the sample sizes in our sample space of 1000.
In the next page are the graphs which shows the result for various values of K.


Sample Size Vs Frequncy Graph (With Replacement) For K=1


Sample Size Vs Frequncy Graph (With Replacement) For K=2


Sample Size Vs Frequncy Graph (With Replacement) For K=3


Sample Size Vs Frequncy Graph (With Replacement) For K=4

Here also, the graphs are quite similar to the once without replacement, except for the fact that the values have shifted to the right, due to the chance of data objects getting repeated.