Sampling

How I created data set?

Firstly, I wanted data to be unique so, I decided to use two attributes. I chose one attribute to be string, so that value can be unique. Another value I choose to be a random number. In such way I created 1000 sample point. After it I equally divided my sample into 10 classes such that each class has equal number of rows. At last I shuffled my data stored it in csv file.

```
"CDTIAXQOFK",29478,2
"AHRGDWUPVM",42854,0
"KUPCRIOMNZ",32978,0
"IVKNOHZTUR",60396,7
"CKLXITPYJZ",57108,1
"JIXMRKQPNO",31378,8
"XRUQKNYPWD",43050,1
"IGXCQTKPAM",30136,7
"BZQEXIKDLO",13716,7
"FUGEDRWONX",8154,4
"CTAZEQSVXJ",58582,1
"PRCTOIGZDH",70576,8
"PUKGHORLZM",9840,7
"LEWPYMZRSF",83842,4
"UFRYEAONIT",35636,5
"MEUTDLNBRA",26034,1
"AMJBXYHUDR",85132,4
"HZWRVGYQIO",28556,5
"NCVIHQBKJO",40500,6
```

Figure 1 Sample of data

Sampling without replacement.

We have to find the minimum number of samples that must be taken so that point from each of the classes are there. For this purpose, we used Cumulative distribution function.

For K =1, I performed the experiment 1000 times. For each experiment I took the number of samples for which we got all the class. For this I used sample method if random package in python.

Then I plotted the probability vs sample size graph.

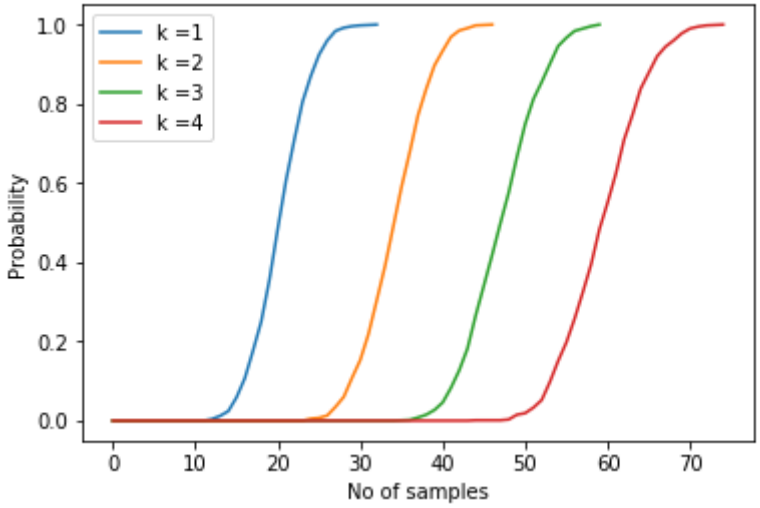
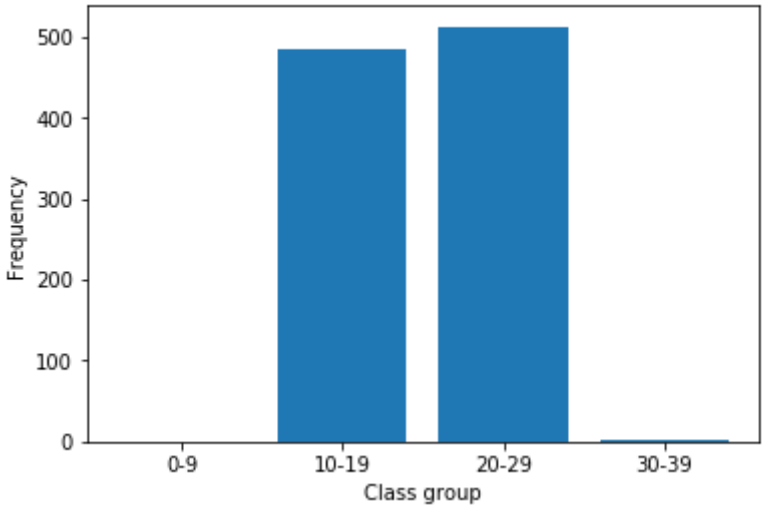
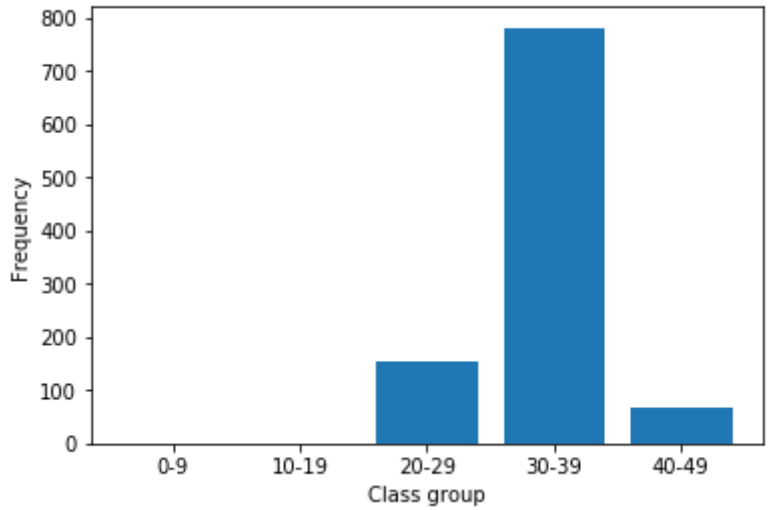
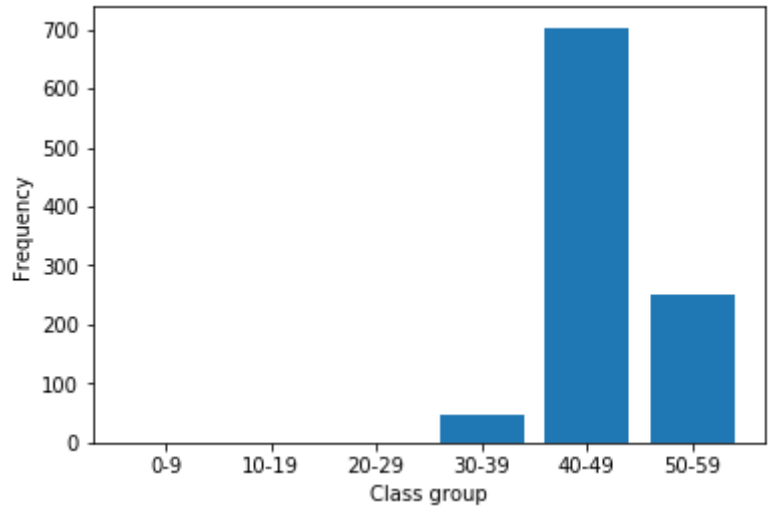
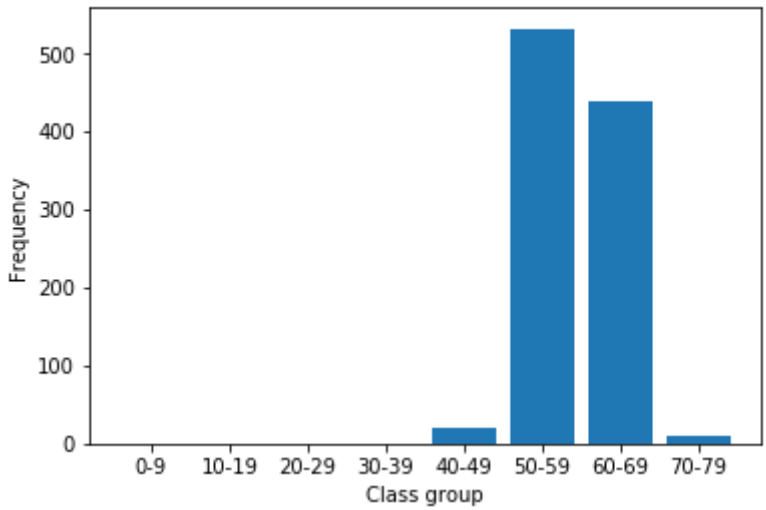


Figure 2 shows the graph for cumulative sampling for various k

After this I plotted bar graph for frequency vs sample size group

For k =1	
For k =2	
For k =3	



Sampling with replacement

I this I followed the same process I followed in without replacement. The only difference was instead of Sample I used choices().

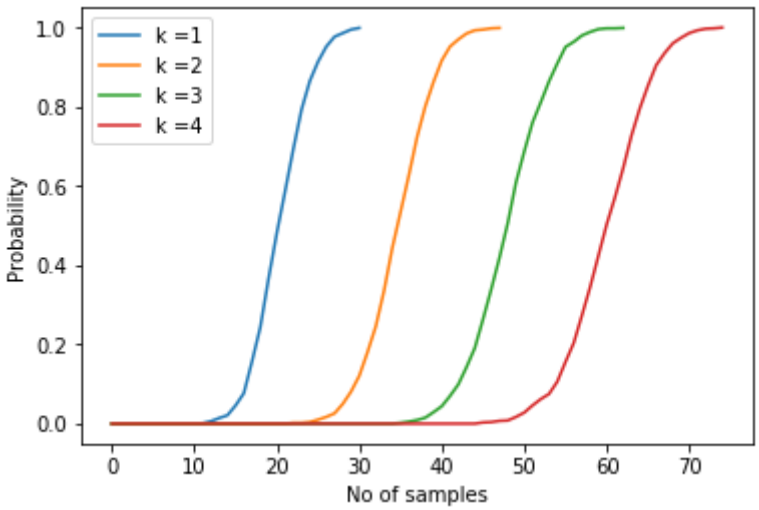
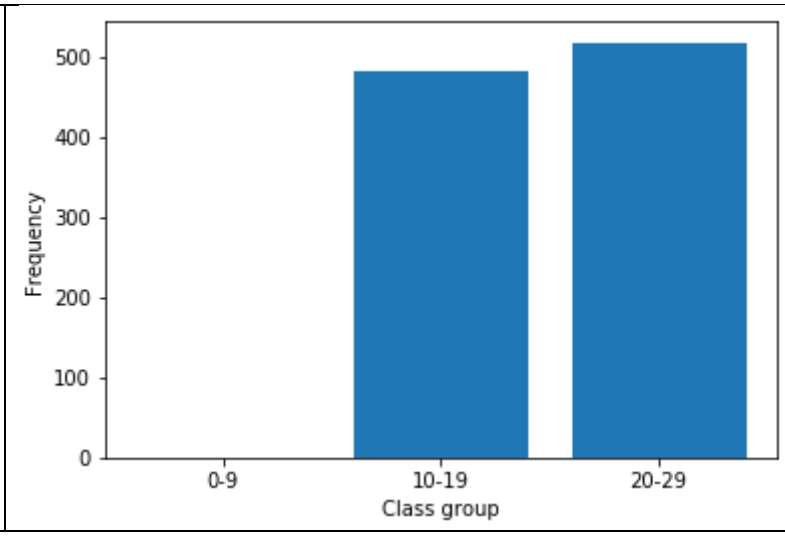
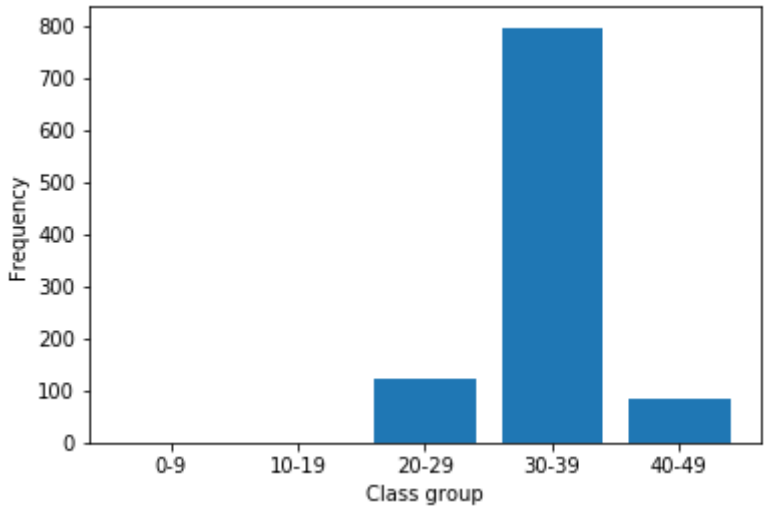
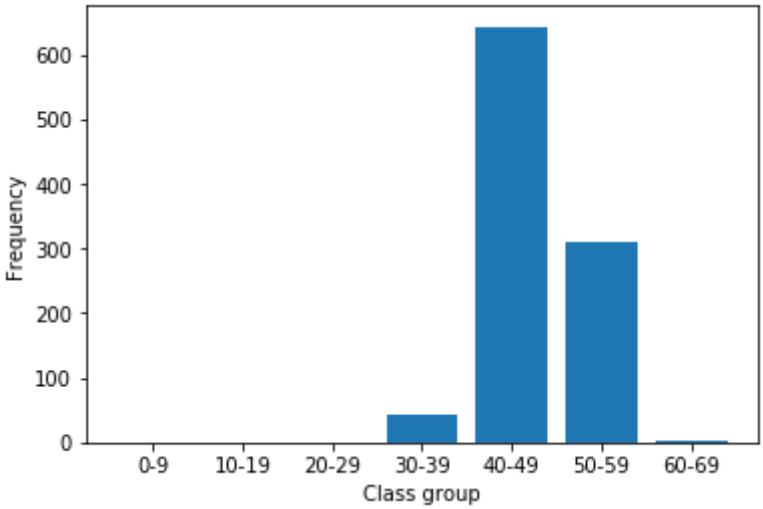
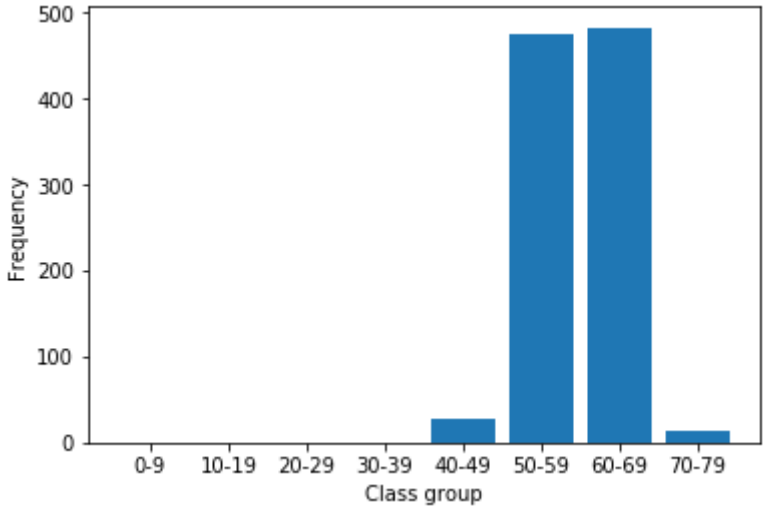


Figure 3 shows the graph for cumulative sampling for various k

Bar plot graph for frequency vs sample size group

For k =1



For k=2	
For k =3	
For k = 3	

Inferences

For the above experiment we can conclude that the sample size depend upon the value of K. the more the value of k the more point we take. And the bar graph will peak will shift toward right.

In case of replacement and non-replacement graph structure will be same and with replacement sample size can be greater than that of without replacement.