Assignment 1

-Suraj Prathik Kumar(2016101)

Question 1

To obtain uniform sample of size s from N such that s ≪ N. Let N be the number of such tweets. For it to occur with equal probability,we can use 2 types of sampling techniques:

1. Sampling without replacement - In this type of sampling technique the tweets are selected randomly with equal probability and added to the sample but not add it back to the tweets. So, all sample units are unique.

2. Sampling with replacement - In this type of sampling technique the tweets are selected randomly with equal probability and added to sample and then added it back to the tweets. So, repetition is allowed.

Question 2

Code submitted for this Question

Assumptions made for the sampling process are :

- Number of tweets considered 100,500,1000,10000(N)
- The tweets are Integers from 1-N
- Stream size 20 selected at random without repetition
- For sampling,100 iterations are done for the items of the steam which is selected at random which may be repeated.
- Length of sample is 5. Values from sample wont be removed until there are 5 items in the sample.

Observations -

```
N = 100
A1 = 2  A2 = 3  A3 = 8  A4 = 6  A5 = 9  A6 = 3  A7 = 8  A8 = 3  A9 = 2  A10 = 2 A11 = 4 A12 = 4 A13 = 1 A14 = 0 A15 = 2 A16 = 1 A17 = 3 A18 = 1 A19 = 2 A20 = 1

N = 500
A1 = 8  A2 = 4  A3 = 4  A4 = 3  A5 = 7  A6 = 3  A7 = 2  A8 = 5  A9 = 5  A10 = 4 A11 = 2 A12 = 3 A13 = 2 A14 = 1 A15 = 2 A16 = 4 A17 = 0 A18 = 2 A19 = 1 A20 = 2

N = 1000
A1 = 3  A2 = 6  A3 = 9  A4 = 6  A5 = 5  A6 = 1  A7 = 4  A8 = 4  A9 = 3  A10 = 2 A11 = 3 A12 = 0 A13 = 1 A14 = 1 A15 = 2 A16 = 1 A17 = 1 A18 = 1 A19 = 0 A20 = 1

N = 10000
A1 = 3  A2 = 2  A3 = 4  A4 = 2  A5 = 4  A6 = 7  A7 = 3  A8 = 10 A9 = 3  A10 = 2 A11 = 2 A12 = 4 A13 = 2 A14 = 1 A15 = 2 A16 = 3 A17 = 3 A18 = 1 A19 = 2 A20 = 2
```

e) It is Unbiased because when the stream is created randomly from tweets,all the tweets have same probability to occur in the stream.While creating sample each item of the stream has a similar count after a sufficient number of iterations.

f) According to the question, the size of the stream is n and the size of the sample is s. In the following proof, we establish that the sampling process used is unbiased and the probability of each stream point to be included in the sample is (s/n)

Proof :

- Let s be the size of the sample such that n>s.
- We assume that the current elements have been included in the sample of size s from a stream of size n with a probability s/n.
- Thus, we prove n+1 stream elements will have s/n+1 probability. (By the principle of mathematical induction). The streams are generated randomly from the given number of tweets and this sampling process is carried out for all the streams hence,generated.
- The elements in a sample have 1/s probability of getting replaced. This is because they are removed randomly with equal probability when a new stream item is added during the 100 iterations.
- Therefore, Probability that an element already in the n-stream sample will not be replaced by the one from the that of the n+1-stream sample will be 1 - (1/s)*s/(n+1) = 1 - 1/(n+1) = n/(n+1)
- Eventually we prove that the probability that the n+1 stream sample contains the element already present in the n stream sample without being removed by a new one = (s/n)*n/(n+1)=s/(n+1)

Thus, we proved that the probability of each stream point to be included in the sample is (s/n).