Assignment – II
(k-mer Analysis, Dotplots)

Deadline: 2nd Apr

1. Simulate observations having the binomial distribution with p = 0.25 and *n* = 1000. What is the probability of observing at least 240 A's in such a sequence? [Hint: Obtain 10,000 simulations and compute the number of A's in each run]. Compare your result with the normal approximation to binomial distribution.

2. Suppose X has a binomial distribution with p = 0.3 and *n* = 10. Compute P(X=0), P(X=2), E(X) and Var(X).

3. Briefly discuss the applications of k-mer analysis. (done)

4. Show a dotplot of the following two sequences and give the conserved region:
   (Make a n x m table and put '.' or 'x' for match)

GGCTGCAACTAGCTC
GGGTAAGCTTGC

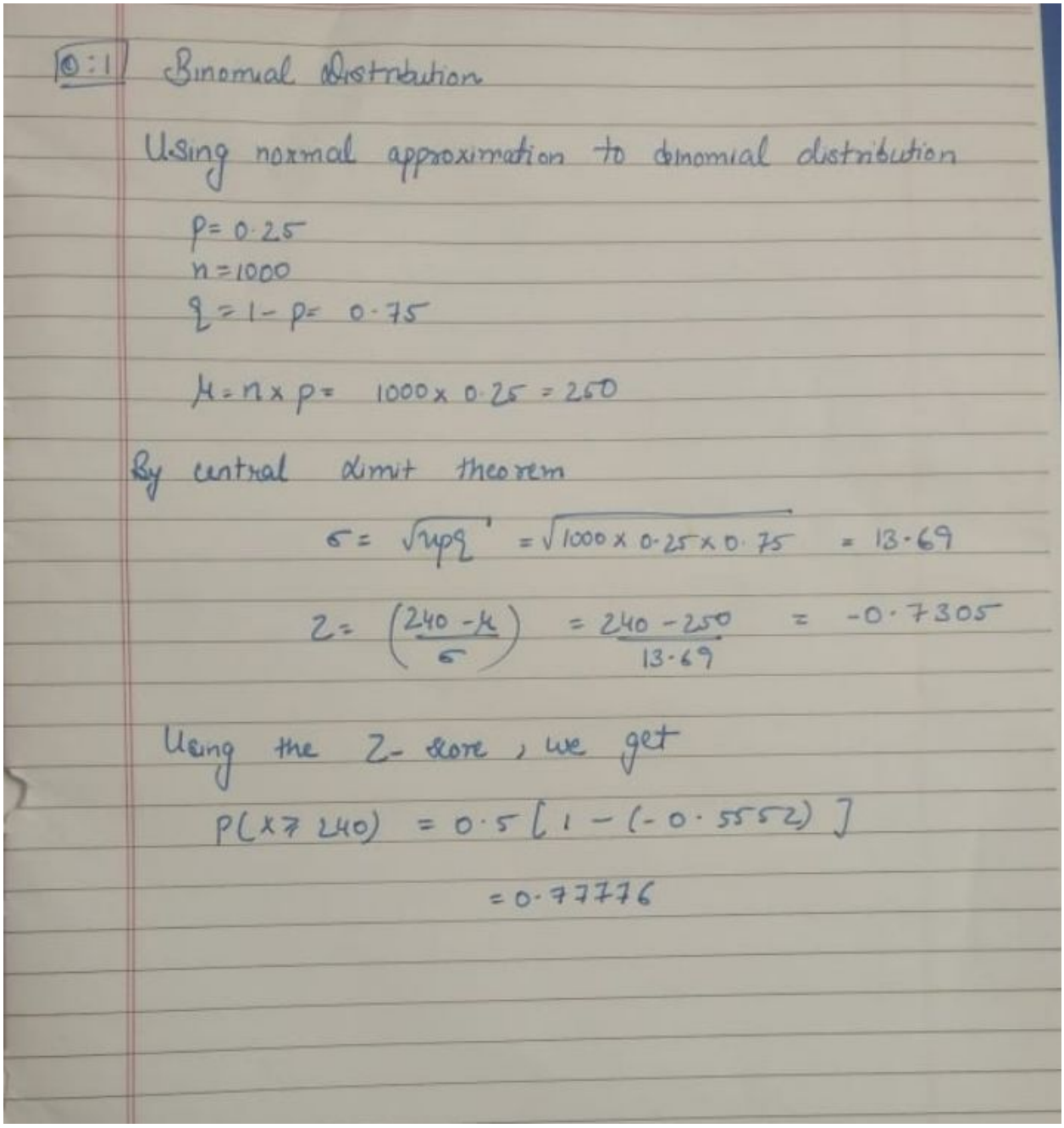
5. Obtain the self-dotplot of the following sequence to identify repeat region:

TGGCACACTCACACCACACAGACAGTTA

6. Find self-complementary regions in the following RNA sequence:
AUGUGGCAUGCCAGG

Answer 1 :

Taking the normal approximation to the Binomial Distribution, we get results as under:



Q:1 Binomial Distribution

Using normal approximation to binomial distribution

$p = 0.25$

$n = 1000$

$q = 1 - p = 0.75$

$\mu = n \times p = 1000 \times 0.25 = 250$

By central limit theorem

$$\sigma = \sqrt{npq} = \sqrt{1000 \times 0.25 \times 0.75} = 13.69$$

$$Z = \left(\frac{240 - \mu}{\sigma}\right) = \frac{240 - 250}{13.69} = -0.7305$$

Using the Z-score, we get

$$P(X \geq 240) = 0.5 [1 - (-0.5552)]$$

$$= 0.77776$$

Using Code (for running code, run Python3 ques1.py)

We get the average value of P (X >= 240) as 0.7747

Comparison :

We see that both the values are very very close. The difference is very minute indeed. If we increase the number of simulations (going from 10,000 to even more) ,then , we shall observe that the obtained average values will be even more closer to the obtained normal approximation values.

Answer 2 :

The solution will is written on paper and uploaded as under:
(Following are two images that contain the solutions to the 4 parts of the question)

Q:2

X has a binomial distribution

$p = 0.3$

$n = 10$

$P(X=a) = {}^{n}C_{a} (p)^{a} (1-p)^{n-a}$

① $P(X=0)$

$= ({}^{10}C_{0}) (p)^{0} (1-p)^{10}$

$\boxed{P(X=0) = (0.7)^{10}} \quad (= 0.0282)$

② $P(X=2)$

$\Rightarrow P(X=2) = ({}^{10}C_{2}) (0.3)^{2} (0.7)^{8}$

$\Rightarrow P(X=2) = \left(\frac{10 \times 9}{2}\right) (0.3)^{2} (0.7)^{8}$

$\Rightarrow P(X=2) = 45 \times (0.09) (0.7)^{8}$

$\Rightarrow P(X=2) = 0.2335$

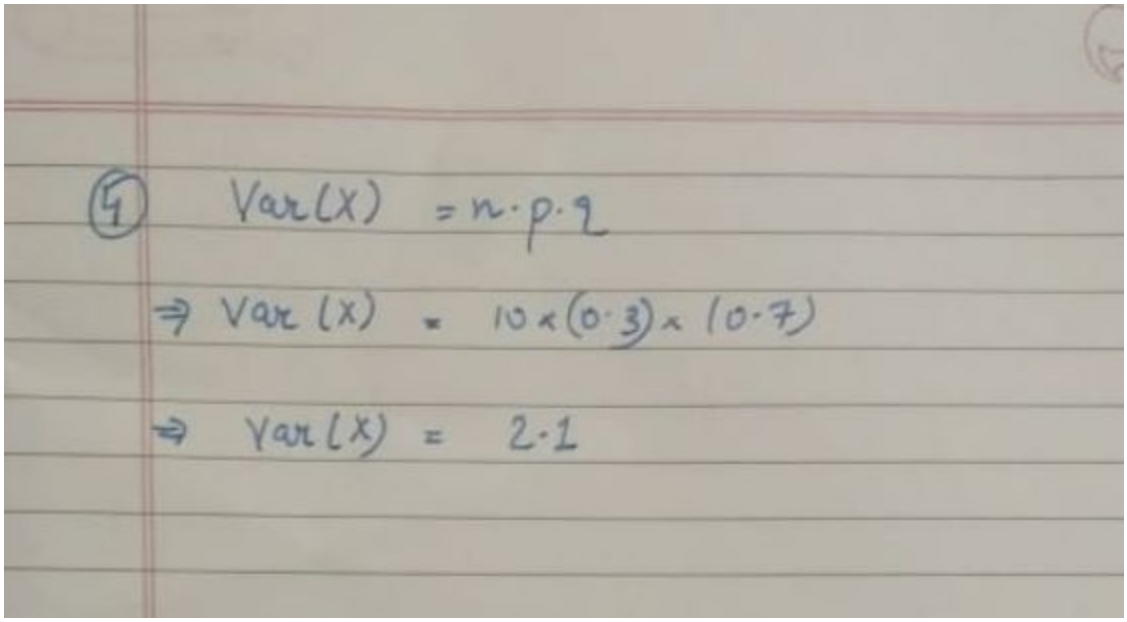
③ $E(X)$

$E(X) = n \cdot p$

$\Rightarrow E[X] = 10 \cdot (0.3)$

$\Rightarrow E[X] = 3$

PTO →



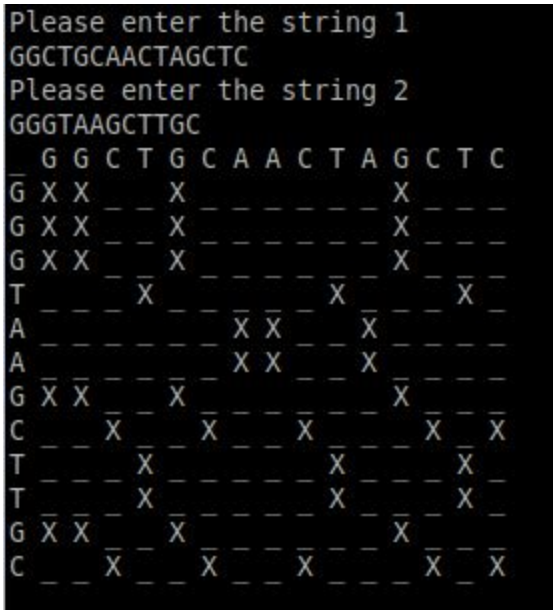
Answer 3 :

The applications of K-mer Analysis are as follows :

1. K-mers are used to identify regions having aberrant base compositions that may indicate genome segments acquired by lateral transfer.
2. G C skew is even used to predict locations of replication origins and termini in prokaryotes.
3. Parametric methods at the gene level like Codon Usage Bias, Amino Acid Usage Bias, GC content at Codon positions etc are all based on K-Mer analysis
4. K-Mer distributions are well preserved among related strains/species. Bacterial genomes can be clustered into natural groups according to K - mer distribution similarities.
5. The frequencies of K-Tuples have a number of applications. Like, for eukaryotes, gene regions, in general, have a different base composition than non-genic regions.

6. Different gene classes have different codon usage frequencies (for instance, the case of the highly expressed genes)
7. Codon usage differs from organism to organism - useful in identifying horizontally transferred genes => Another application of K-mer analysis.
8. Sometimes, observed frequencies of K - words can be used to make inferences about DNA sequences .
    a. For eg : Simply by noting the frequency of stop codons in all the three reading frames, and knowing that a typical bacterial gene contains, on average, more than 300 codons, or that the typical human exon, contains around 50 codons, one can make a reasonable inference as to whether a given sequence is or is not code for a protein.
    b. Which is another direct application of the studies done under K-mer analysis

Answer 4 :



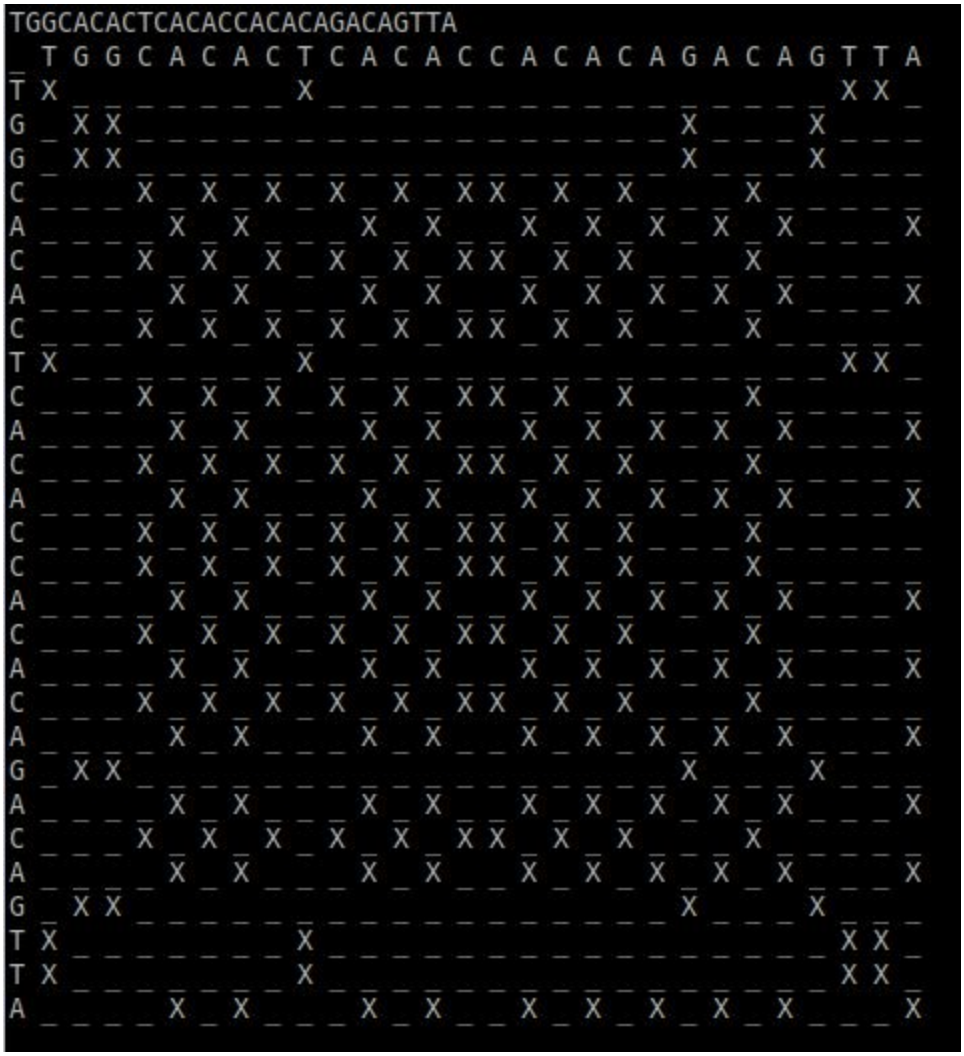
From this plot, we gather that the conserved region is

1. AGTC
2. GCT
3. TGC

These are the only conserved regions (i.e. diagonal (top left to bottom right) that are of length 3 or more) . We are ignoring lengths less than or equal to 2 for the cause of this question.

Answer 5 :

Obtaining the self dotplot for the given sequence and finding the self repeat regions :



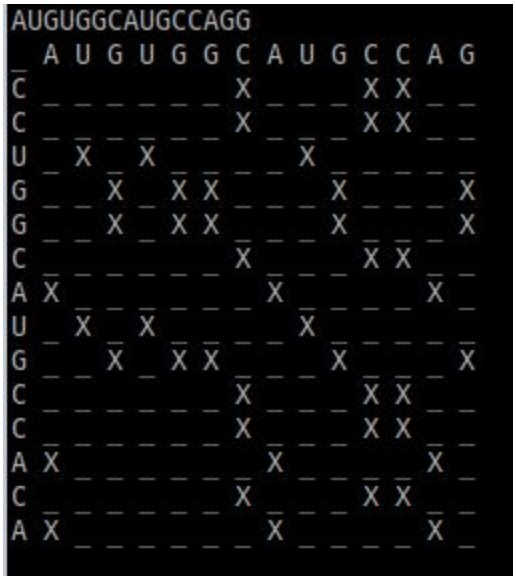
From this plot, we can see that the repeat regions are :

1. CAC

2. ACA
3. CAG
4. CACA
5. ACAC
6. CACAC

These are the only repeated regions (i.e. diagonal (top left to bottom right) that are of length 3 or more) . We are ignoring lengths less than 2 for the cause of this question.

ANSWER 6 :



The self complementary regions in the Rna sequence are :

1. UGGCAUGCCA

It’s substrings are also coming under the self complementary regions in the RNA but for sake of convenience, they are not listed here.