Home Work

1. Please give an estimate of the upper bound for the total number of words spoken by all *Homo sapiens* ever lived and now living on Earth. (Hint: whenever a quantity is bounded from above, there must exist a precise upper bound. The precise upper bound may never be known in our case, but by putting together a few reasonable assumptions one may get some estimate which is surely greater than the precise bound. Better assumptions may yield a better estimate.)
2. Please calculate the number of DNA molecules in 10 μg (microgram) of *E. coli* genome. Note: an *E. coli* genome consists of 4 639 221 base pairs; the molecular weight of a base pair is approximately 620 D; the Avogadro's number is 6.023×10^{23}.
3. In a Bernulli test the probability of success is p and that of failure is $1 - p$. Let random variable Z takes value $z = 1$ at success and $z = -1$ at failure. Please calculate the expectation and variance of Z.
4. Please calculate the expectation and variance of binomial distribution

$$B(N, p) = \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k}.$$

5. Suppose in a casino the probability of using a fair dice is $P(D_{\text{fair}}) = 0.99$ and the probability of using a loaded dice is $P(D_{\text{loaded}}) = 0.01$. For a fair dice

$$P(i|D_{\text{fair}}) = 1/6$$

for all $i = 1, 2, \cdots, 6$. For a loaded dice

$$P(i|D_{\text{loaded}}) = \left(\begin{array}{ll} \frac{1}{10}, & \forall i = 1, 2, 3, 4, 5 \\ \frac{1}{2}, & \text{for } i = 6 \end{array} \right.$$

Please calculate the probability that the dice is a loaded one when three consecutive '6' has been observed, i.e., the conditional probability $P(D_{\text{loaded}}|666)$. (Hint: Use Bayes' formula. Taken from the book by R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison)

6. Period-3 fifth-order Markov models are often used to describe coding sequences in DNAs. How many independent parameters are there in such a model?
7. Given two sequences $ALGORITHMIC$ and $ARITHMETIC$. Please get the highest-score alignment(s) of them according to the following scoring scheme:

Match	2
Mismatch	-1
Gap	-2

8. Suppose that a specific string of length K does not appear in a DNA sequence. Show that at a longer length $K + i$ the number of strings that will not appear due to the absence of the K-string is given by $4^i \times (i + 1)$ to a good approximation. (Hint: use mathematical induction.) Give an example when this formula does not lead to an exact result.