Homework #2

Please present 4 significant figures in your final answers for probabilities. Also, make sure to explain your thought process as if the reader is one of your classmates.

1. (7 points) Mutations to DNA occurs when the cell makes mistakes as it makes copies of its DNA. Suppose the rate of mutation per base, per generation is about 2.5×10^{-8} and there are about 3.2×10^{9} sites in the human genome. Assume that mutations at different sites are independent of each other.

(a) Calculate the rate λ at which mutations occur in a cell with one copy of the human genome.

Answer: $\lambda = (3.2 \times 10^{9})(2.5 \times 10^{-8}) = 80$

(b) If one observes the cell for 1 generation what is the probability that at least 80 and at most 85 mutations will occur?

Answer: Let X~Pois(80)

$P(85 <= X <= 80) = P(X=80) + P(X=81) + P(X=82) + P(X=83) + P(X=84) + P(X=85)$

$= 0.04456 + 0.04401 + 0.04293 + 0.04138 + 0.03941 + 0.03709$

$= 0.24938$

2. (7 points) For some reason a lot of people hate raisins in cookies, but you love them. So, when you were baking 100 cookies for a party, you only put in 100 raisins in the batter. Your best friend who absolutely hates raisins picks 3 cookies at random. What is the probability that at least one of the cookies is raisin-free? (Hint: Use both Poisson and Binomial)

Answer: Expected number of raisins per cookie λ=1. Probability that a randomly picked cookie has no raisins is $P(X=0) = e^{-\lambda} = e^{-1} = 0.3679$

Probability that at least one cookie out of 3 has no raisins is

$1 - P(Y=0) = 1 - {}^{3}C_{0}\,(0.3679)^{0}\,(0.6321)^{3} = 0.2526$

3. (11 points) Sequencing technologies can only "read" short fragments from a genome. Given that the process through which the sequences are generated is random, it is possible that certain parts of the genome will remain uncovered unless an impractical amount of sequences are generated.

We know that the size of the human genome is 3×10^{9} bp. Now a new human genome has been sequenced and it's randomly covered by 30 million reads (read length is 300 bp). We assume that the number of times a base in the human genome is covered follows a Poisson distribution.

(a) What is the probability that a particular base is covered by at least one read?

Answer: the average time a base is covered is $30 \times 10^6 \times \frac{300}{3\times10^9} = 3$. The probability that a particular base is not covered by any read is $P(X = 0) = e^{-3} = 0.04979$ So,
P(X>=1) = 1 – 0.04979 = 0.95021

(b) Calculate the number of contigs
Answer: $N_{contigs} = N * e^{-\lambda}$ = $30 \times 10^6 * e^{-3}$= 1.494 x 10^6

(c) What is the average length of a contig?
Answer: $G_{covered}$ / $N_{contigs}$ = G * P(X>0) /1.494 x 10^6
= 3x10^9 * 0.95021 / 1.494 x 10^6
= 1908