Minor 1: SBL701 Biometry

Max Marks: 30

Submission deadline: 2359 hrs on 25th September, 2022 (NO EXTENSION PLEASE!)

Submission details (deviations are not acceptable):

I. *Only soft copy submissions on the email ids amittal@bioschool.iitd.ac.in will be accepted (i.e. no hard copy). No late submission will be accepted (regardless of server problems etc.). To be safe, submit well in advance.*
II. *Submissions should be compiled in form of one Adobe pdf file. Multiple files or files in any other software are not acceptable. One way to do this is "print" all your files as "adobe pdf" rather than sending them to a printer.*
III. *It is preferable, but not mandatory, to use Microsoft Equation 3.0 or Equation Editor or Math Type in MS word if you need to type any equations.*
IV. *Solutions should contain each and every step followed for calculations, properly visualized data – e.g. tables or graphs and programming codes (e.g. MATLAB or listing of steps followed in MS Excel/any other software used) along with code output (in form of numbers in command window or figures).*
V. *Copying is not acceptable. While you are free to consult anyone, the solutions, interpretations, answers should all be in your own language. Any material that is found to be copied from another source will automatically lead to a "zero" out of 30. Any material found to be identical in different submissions will also lead to "zero" out of 30 for all the submissions appearing identical.*

Questions

1. Owing to a severe coronavirus outbreak in China, several companies are working towards development of a diagnostic kit for rapid detection of infection in patients. Using genome sequencing, which takes at least three days, it is found that of the randomly sampled 1000 people in China, 20 were infected with the virus on January 15, 2020. Epidemiologists collaborating with the local medical personnel also report that the virus is spreading at a very rapid constant rate – an increase of exactly 1 percent of the population is confirmed to be infected patients per day. To contain the virus, Chinese government quarantines the complete population, regardless of whether one is infected or not. You are a founder of a diagnostics company at IIT Delhi through FITT. Considering the urgency, you develop a diagnostic test-kit for detection of the virus on January 20, 2020. You immediately courier the test-kit to your contacts in China who receive your kit on January 21, 2020. Regulatory approval mechanisms in China consider any test kit to be reliable only if there is a 90% or more chance of being infected with the virus if the test is positive. Your contacts in China find that of the 100 randomly selected patients infected with the virus, your test is positive in 82. They also find that of the 50 randomly selected non-infected people, your test is positive only in 2. If it takes only one day to clear all the regulatory approval mechanisms, by what date do you expect your diagnostic kit to be launched in the market?

(**3**)

2. The evolutionary process of amino acid substitutions in proteins is sometimes described by the Poisson probability distribution function. The probability $p_s(t)$ that exactly s substitutions at a given amino acid position occur over an evolutionary time t is

$$p_s(t) = \frac{e^{-\lambda t}(\lambda t)^s}{s!}$$

where λ is the rate of amino acid substitutions per site per unit time. Fibrinopeptides evolve rapidly: λ_F = 9.0 substitutions per site per 10^9 years. Lysozyme is intermediate: λ_L = 1.0. Histones evolve slowly: λ_H = 0.010 substitutions per site per 10^9 years.

(a) What is the probability that a fibrinopeptide has no mutations at a given site in t = 1 billion years?

(b) What is the probability that lysozyme has three mutations per site in 100 million years?

(c) We want to determine the expected number of mutations ‹s› that will occur in time t. We will do this in two steps. First, using the fact that probabilities must sum to one, write $\alpha = \sum_{s=0}^{\infty} \frac{(\lambda t)^s}{s!}$ in a simpler form.

(d) Now write an expression for ‹s›. Note that

$$\sum_{s=0}^{\infty} \frac{s(\lambda t)^s}{s!} = (\lambda t) \sum_{s=1}^{\infty} \frac{(\lambda t)^{s-1}}{(s-1)!} = \lambda t \alpha$$

(e) Using your answer to part (d), determine the ratio of the expected number of mutations in a fibrinopeptide to expected number of mutations in histone protein, $\langle s \rangle_{fib} / \langle s \rangle_{his}$.

(1 + 1 + 1 + 1 + 1 = **5**)

3. In forensic science, DNA fragments found at the scene of a crime can be compared with DNA fragments from a suspected criminal to determine that the probability that a match occurs by chance. Suppose that DNA fragment A is found in 1% of the population, fragment B is found in 4% of the population, and fragment C is found in 2.5% of the population.

(a) If the three fragments contain independent information, what is the probability that a suspect's DNA will match all three of these fragment characteristics by chance?

(b) Some people believe that such a fragment analysis is flawed because different DNA fragments do not represent independent properties. As before, suppose fragment A occurs in 1% of the population. But now suppose that p(B/A) = 0.40 (instead of p(B) = 0.04), and p(C/A) = 0.25 (instead of p(C) = 0.025). There is no additional information about any relationship between B and C. What is the probability match now?

(1 + 1= **2**)

4. Many bacteria are motile and are able to move because of small propeller like membrane proteins called flagella. Let us assume that the distance covered by a bacterial cell in a particular area is given by an exponential distribution: $f_X(x) = \lambda e^{-\lambda x}$.

(a) Applying the concepts for continuous distributions, derive the expression for E(X), Var(X).

(b) Assume E(X) = 1/λ, Var(X) = $1/\lambda^2$. Thus, given a population mean or variance, one can estimate λ and vice versa. Now assume that λ = 0.2. In MS Excel or MATLAB (or any other software), generate the pdfs for the exponential distribution and normal distribution for sample sizes of 10, 50 and 100, with estimators based on the above assumptions, wherever applicable. Compare the pdfs (in form of histogram plots, i.e. bin/range of values on the X-axis, probability density = frequency/sample size on the Y-axis). Comment on the observations.

(c) Using the assumptions from (b) above, in MATLAB or MS Excel, generate 10,000 random samples of 10 values each and calculate the sample mean and sample variance for each sample. Plot the histogram of the frequencies (probability densities) for the obtained sample means and sample variances. Now repeat the process for 10,000 random samples of 100 values each. (HINT: One can use "hist" function in MATLAB and generate a similar plot in MS Excel as discussed in class).

(d) For the simulations in (c), if $\bar{x}$ represents the mean of means for the 10,000 samples, using the formulae for mean and variance of discrete random variables to approximate the expected value and variances of the estimators, complete the following table:

	$E(\bar{x})$	$Var(\bar{x})$	$E(S_x^2)$	$Var(S_x^2)$
n = 10				
n = 100				

(e) Comment on the results/observations in (b), (c) and (d) in context of the central limit theorem.

(1 + 3 + 5 + 4 + 1 = **14**)

5. Monty-Hall Problem: You are a contestant on a game show. There are three closed doors: one hides a car, and two hide goats. You point to one door, call it C. The gameshow host, knowing what's behind each door, now opens either door A or B, to show you a goat; say it's door A. To win a car, you now get to make your final choice: should you stick with your original choice C, or should you now switch and choose door B? Why or why not?

(0.5 + 1.5 = **2**)

6. Consider the probability distribution $p(x) = ax^n$, $0 \leq x \leq 1$, for a positive integer n.
   (a) Determine "a" in terms of n.
   (b) Compute the central tendency, E(x), of the distribution as a function of n.
   (c) Compute the dispersion, Var(x), of the distribution as a function of n.
   (d) Is $\sigma^2 = E(x^2) - [E(x)]^2$ the same as Var(x)?

(1.5 + 1 + 1 + 0.5 = **4**)