Individual: Probability and statistics

Problem 1. Suppose you are buying an item that is needed for your factory. There are three stores nearby. Each store will tell you the price X_i ($i = 1, 2, 3$). After you ask around, you will buy from the shop with the lowest price. Suppose that you believe $X_i \sim$ Uniform$(100, 150)$ for all $i = 1, 2, 3$, independently of one another.

a Suppose you will always ask the first two stores for quotes (for free), but it costs \$3 to ask the third quote. Show that the expected saving due to asking for the third quote is strictly positive.

b Suppose when you reported your results in part (a) to your factory director, he was confused as to whether you were recommending that one should always go for the third quote. In order to provide him with better guideline for a stepwise decision making, present your decision rule as (let $Y = \min\{X_1, X_2\}$): If $Y \geq c$, then we should get the third quote; find c.

c What is the probability that you end up not saving money at all when asking the third quote?

Problem 2. Assume there are N short fragments, each of length L, sampled randomly from a long sequence of length G ($G >> L$). Specifically, ignoring boundary effects, we assume the left-hand ends of the fragments are independently distributed according to a uniform distribution over $(0, G)$.

These N fragments may overlap. Overlapping fragments can be merged to form longer contiguous stretches of sequence. A **contig** is one such assembled stretch (that cannot be further extended) in which all the fragments connect unambiguously (i.e., with no unresolved gaps or uncertainties within the sequence). Given N random fragments of length L, the sequencing **coverage** is defined as $a = \frac{NL}{G}$.

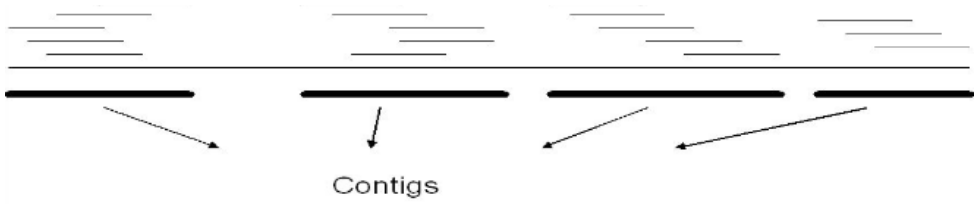

Figure 1: There are four Contigs in the above long sequence

a To ensure that the mean proportion of the long sequence covered by at least one fragment is 0.99, what is the approximate minimum coverage a required?

b What is the mean number of contigs that can be formed for the long sequence?

c Prove that the mean contig size is $\frac{L(e^a-1)}{a}$ with $a = \frac{NL}{G}$

Problem 3. Let $(X_n)_{n\geq 0}$, with $X_0 = 0$, be a discrete time simple random walk on $\mathbb{Z}$ in a dynamic random environment defined as follows. Fix $a > 0$. At each time $n \geq 0$, every undirected edge $e := \{i, i+1\}$ is assigned a conductance $C_n(e)$ with $C_n(e) = 1$ if e has not been crossed by time n, and $C_n(e) = a$ if e has been crossed before. Given $X_n = x \in \mathbb{Z}$ and the conductance configuration $C_n(\cdot)$ at time n, the random walk jumps to either $x + 1$ or $x - 1$ with probability

$$P(X_{n+1} = x \pm 1 | X_n = x, C_n) = \frac{C_n(\{x, x \pm 1\})}{C_n(\{x, x+1\}) + C_n(\{x, x-1\})}.$$

Show that, almost surely, X will return to 0 infinitely many times.

Problem 4. Let $\epsilon_i, x_{ij}, i = 1, \ldots, n, j = 1, \ldots, n$ be i.i.d. $N(0, 1)$ random variables. Define

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \epsilon_i, \quad i = 1, \ldots, n.$$

Suppose we only observe (y_1, x_{11}), (y_2, x_{21}, x_{22}), $\ldots$, $(y_n, x_{n1}, \ldots, x_{nn})$. Obtain estimators of $\beta_1, \ldots, \beta_n$. What desirable properties do these estimators possess? Are they optimal in some sense? If yes, why; if no, do you have any suggestions on how to improve, especially when n is large? Hint: You may consider estimation individual β_j separately.