AML5103 | Applied Probability and Statistics | Coding Problem Set-3

1. It is possible to identify a person behind the screen by analyzing how they type on the keyboard. Coursera uses such Biometric Keystroke signatures for plagiarism detection. If a person cannot write a sentence with the same statistical distribution of key press timings as in their previous work, a red flag is raised. In this problem, you will implement such a plagiarism detection model. For that purpose, the following three files are provided:

    1. **personKeyTimingA.txt** has keystroke timing information for a user A writing a template assage. The first column is the time in milliseconds (since the start of writing) when the user hit each key. The second column is the key that the user hit.
    2. **personKeyTimingB.txt** has keystroke timing information for a second user (user B) writing the same template passage as the user A. Even though the content of the passage is the same, the timing of how the second user wrote the passage is different.
    3. **email.txt** has keystroke timing information for an unknown user

    The goal of this problem is to find out who among A and B wrote the email. To that end, do the following:

    (a) Plot the density histograms for the keystroke times for users A and B. How are the histograms shaped?

    (b) Let T_A and T_B be random variables for the duration of time, in milliseconds, for users A and B (respectively) to type a key. Using the histograms, use an appropriate distribution (with parameters) to model the random variables T_A and T_B.

    (c) Calculate the ratio of the likelihood that user A wrote the email over the likelihood that user B wrote the email.