TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Mid-Term Assignment 2019-20 STU33009: Statistical Methods for Computer Science

Submitting Your Report

- **Reports must be typed (no handwritten answers please) and submitted on Blackboard.**
- **Reports should be no more than 2 pages in length (extra pages beyond this will be ignored when marking).**
- **You will need to use matlab to calculate values from the assignment dataset, or alternatively write a short program in python to do this. In either case give the code used as an appendix to the report (it doesn't count towards the 2 page limit), but please keep the code short.**
- **It is important that you explain/justify how you obtained your results in order to obtain full credit.**

Downloading Dataset

- Download the assignment dataset from `https://www.scss.tcd.ie/doug.leith/ST3009/midterm2020.php`. Important: You must fetch your own copy of the dataset, do not use the dataset downloaded by someone else. Keep the dataset that you download as I might request it to validate your results.
- The data file consists of columns of data. Each column corresponds to one user of a REST API. The values in the column are measurements of the time (rounded to the nearest millisecond) taken to serve a request by that user. The first line of the file gives the probability that a system request comes from each user (needed in part 3 of the assignment, see below).

Assignment

You can assume that the service times for the same user are independent and identically distributed. Let X_i be an indicator random variable that takes value 1 when the time to serve a request for user i exceeds 10ms, and 0 otherwise.

1. For the first column of data only (i.e. the data for user 0 only):

   (a) Plot a histogram showing the PMF of the time taken to serve requests for user 0.

   (b) Estimate $Prob(X_0 = 1)$. Hint: Recall that for an indicator RV $Prob(X_0 = 1) = E[X_0]$, so map each service time to a 0 or 1 indicator value and calculate the empirical mean of these 0/1 values.

   (c) Derive confidence intervals for your estimate $Prob(X_0 = 1)$ using the CLT, Chebyshev Inequality and Bootstrapping. Discuss the pros and cons of each of these methods.

2. Estimate $Prob(X_i = 1)$ for each of the remaining users. There's no need to plot the PMF or give confidence intervals, just report the estimates of $Prob(X_i = 1)$ for all users.

The server receives a sequence of requests from users. Let Z_n be a random variable whose value is equal to the time taken to serve the n'th request. Let U_n be the index (i.e. the column number in the data file) of the user who submitted the n'th request. The first line of the data file you downloaded gives $P(U_n = i)$.

3. Using your calculated values of $Prob(X_i = 1)$ and the values given for $P(U_n = i)$, calculate the probability that Z_n exceeds 10ms. Hint: Use marginalisation.

4. Calculate $P(U_n = 0|Z_n > 10)$. Hint: Use Bayes Rule.

5. Write a stochastic simulation of this setup. Namely, there is a sequence of requests Z_n, $n = 1, 2, \ldots$. The probability that request n comes from user i is $P(U_n = i)$. Given that a request came from user i the probability that it takes more than 10ms to server is $P(Z_n > 10|U_n = i) = Prob(X_i = 1)$. Using this simulation estimate $P(Z_n > 10)$ and compare against the value you calculated above. Discuss.