Indian Institute of Technology Dharwad, Karnataka, India

EE 227 / EE 201: Data Analysis
Project (Autumn 2021)
Due date: 16th November 2021, 20:00:00 IST

Instructions

1. Write a **short note on the approach followed and results obtained** for each problem. Hand-written note is fine.
2. As usual, discussion is allowed. But copying is not allowed. **Any kind of copying**, like copying of solutions, copying the project code by simply changing variable names, etc. will result in **zero marks** for this entire project component.
3. **Late submission (even one second after the deadline)** will get **zero marks** for the entire project because you have got 5.5 hours for an exam requiring around 3-4 hours. No last minute requests will be entertained.

Questions

1. (20 points) **Simulating RVs:** You are in a situation where you can only generate uniform random variables between 0 and 1 using in-built functions (for example: using *rand* in Python, MATLAB and C/C++).
   (a) (14 points) Generate samples from a standard normal distribution using one or more uniform RVs. Compare your results with those obtained an in-built function generating standard normal RVs (for example: *randn* in Python, MATLAB).
   (b) (6 points) Generate samples from a Chi-Squared distribution with 4 degrees of freedom using the one or more Gaussian RVs obtained from the method developed in part (a).
2. (20 points) **Estimators:** You have been given the data (1000 samples) drawn from a uniform distribution between a and b, where a and b are positive integers. However, the data is corrupted with noise following a standard normal distribution.
   (a) (10 points) How will you jointly estimate a and b? Justify whether your estimation strategy is unbiased and consistent.
   (b) (10 points) Let us say that you were not aware that the samples were noisy. Find separate estimators for a and b. Write code (or use MS Excel or any data analysis tool) to check whether these estimators are unbiased and consistent.

   The data is at `DA_P2/DA_P2_data_<your_roll_number>.txt` on the Google Drive Link sent through e-mail.
3. (30 points) **Test after test after test:** You have been given the data corresponding to two actual populations (say A and B). Each row of data contains two entries, where the first entry is from Population A and the second entry is from Population B. There are 10000 such rows. You have also been given 50 measured samples from one of these two populations.

Both the populations follow different Gaussian distributions. Compute their mean and variance and use them as "actual mean" and "actual variance" to answer the following.

(a) (10 points) Find 99% confidence intervals for your estimated mean and variance. Determine to which population do your measured samples belong to.

(b) (10 points) Let us say that you did not know the actual mean and variance, but you know that the measured samples also follow a Gaussian distribution. Write code (or use excel or any data analysis tool) to perform a 5% significance level test for mean and variance.

(c) (10 points) Pick the first m samples from the 50 measured samples and call it "Set 1". Denote the remaining $(50 - m)$ samples as "Set 2". Perform F-ratio tests for different values of m (starting from 4 up to 46) and find out the range of values of m for which the F-ratio test indicates that the variances of both the sets are equal.

The population data is at `DA_P3/P3_Pop/DA_P3_data_Pop_<your_roll_number>.txt` on the Google Drive Link sent through e-mail.

The measurement data is at `DA_P3/P3_Samp/DA_P3_data_Samp_<your_roll_number>.txt` on the Google Drive Link sent through e-mail.

4. (10 points) **On your explorations:** Generate your own data and write code (or use MS excel or any data analysis tool) to

(a) (5 points) illustrate one-way ANOVA using data spanning 3 columns and 10 samples per column.

(b) (5 points) illustrate Spearman's rank correlation using data spanning 2 columns and 20 samples per column.

Clearly indicate how you generated the data.