Indian Institute of Technology Madras

Biotechnology Department

BT5450: Data-driven Modeling and Optimization of Bioprocesses

Assignment II

Due date: Sept 19, 2021, Online

Instructions

1. Assignment shall be submitted before the due date. Late submissions will not be entertained. If you cannot submit the assignment due to some reasons, please contact the instructor by the email.
2. Assignment has to be submitted in a single .zip file with the codes (in .ipynb and .pdf format) and the report (.pdf format) in it. Please indicate the name of all the group members and anyone of the group member can upload the assignment in moodle site.
3. All the assignment must be the students own work. The students are encouraged to discuss or consult class mates or friends within or outside their groups. However, they have to submit their own work.
4. If you find the solution in the book or article or on the website, please indicate the reference in the solution.

Problem

Q1 A scientist performs an experiment for collecting data on a biological markers for different patients. The data are available in 'Q1.csv'. Each measurement can be modelled as a random variable X. The scientist proposes the following cumulative distribution function (CDF) for the X based on her experience:

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x^3}{a} & \text{for } x \in [0, 3) \\ \frac{x-2}{b} & \text{for } x \in [3, 5] \\ 1 & \text{for } x \geq 5 \end{cases}$$

Find the probability density function (PDF) for the X.

Q2 Randomly collect N(say, 100) samples from a Bernoulli's distribution ($p = 0.2$, 0=Failure, 1=Success) and plot the histogram of the collected samples. Now, repeat these experiments R (say,1000) times. Then, we can construct 1000×100-dimensional matrix. Use it to plot the following histograms:

(a) Count the number of successes in each of the row and plot the histogram of the number of successes in 1000 experiments.

(b) Collect 1000 samples from the Poisson distribution with $\lambda = N\,p$ (Use $N = 100$ and $p = 0.2$). Plot the histogram of these samples.

(c) Compare the histograms obtained in Q2(a) and Q2(b). Report your inference.

(d) Count the number of failures occurred before two successes in each of the realization and plot the histogram of the same. For example if a realisation has samples as $[0, 0, 1, 0, 1, 0, 0, 0, 1]$ the number of failures before 2 successes is 3.

(e) Collect 1000 samples from the negative binomial distribution ($n = 2$, $p = 0.2$) and plot the histogram of the samples. Compare the plot with the plot in Q2(d). Report your inference.

Hint: Use numpy.random.binomial(), numpy.random.poisson() and numpy.random.negative_binomial() to collect the samples.

Q3 Benford's law or law of first digits states that the leading digits of most of the naturally occurring numbers are most probably small. It has been shown that most of the datasets from electricity bills, height of buildings, length of the rivers, death rates, stock prices follows this law (refer https://en.wikipedia.org/wiki/Benford%27s_law for more details). A set of numbers is said to follow Benford's law if the probabilities of the numbers in the first and second digit are as shown in Table 1 Use the daily data (cases and death) reported for the second wave of COVID-19 in India

Table 1: Probability values of numbers occurring in 1st and 2nd digit according to benford's law

Digits	0	1	2	3	4	5	6	7	8	9
1st	NA	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
2nd	0.12	0.114	0.109	0.104	0.1	0.097	0.093	0.09	0.088	0.085

(say, 15/03/2021 to 16/07/2021) from https://api.covid19india.org/csv/latest/states.csv and create the following plots

(a) Histograms of 1st and 2nd digits of daily confirmed COVID-19 cases for all the states during the second wave.

(b) Histograms of 1st and 2nd digits of daily deceased cases for all the states during the second wave.

Using the histograms in Q3(a) and (b), report your observations on states that follows the Benford's law approximately and the states that don't. State your reason for the same.

Note that this data gives cumulative confirmed cases and deaths, and hence, it has to be preprocessed to get the daily cases and death.

Q4 Two independent random variables X and Y follow the Gaussian distribution. Estimate the parameters of the distributions from the given dataset 'Q4.csv' using (i) the maximum likelihood estimation, (ii) the method of moments, and (iii) Bootstrap estimation method. Using the estimated parameters of the probability distribution function calculate the following quantities:

(a) $E[W] + E[V]$.

(b) $E[|X - Y|]$ and $E[|X| - |Y|]$

where $W = \min(X, Y)$, $V = \max(X, Y)$ and $E[\cdot]$ is the expected value of a random variable.

Q5 Generate samples from the random variables X_1, X_2, X_3, X_4, and X_5. Such that $X_1 \sim \mathcal{N}(4, 9)$ and $X_2 \sim \mathcal{N}(5, 7)$. Also, $X_3 = 3X_1 - 2X_2$; $X_4 = X_1 + 2X_3$; $X_5 = 5X_1 - X_2$.

(a) Estimate the covariance matrix from the obtained samples.

(b) Calculate the rank of the covariance matrix. What does it tell us about the number of independent variables in the given dataset?

Generate the random variables X_6 and X_7 such that $X_6 = X_1^2 + X_2^3$; $X_7 = \frac{X_1^2}{X_2}$.

(a) Estimate the covariance matrix from the obtained samples of seven random variables.

(b) Calculate the rank of the covariance matrix. Does the rank change or remain the same compared to the above one. Does this rank have any information on number of independent variables in the new dataset?

Q6 Let X_1, X_2, ..., X_7 denote a random sample from a population having mean μ and variance σ^2. Consider the following estimations of μ

$$\hat{\boldsymbol{\theta}}_{\mathbf{1}} = \frac{X_1 + X_2 + \ldots + X_7}{7}$$

$$\hat{\boldsymbol{\theta}}_{\mathbf{2}} = \frac{2X_1 - X_6 + X_4}{2}$$

(a) Is either estimator unbiased?

(b) Which is the best estimator and justify?