Indian Institute of Technology Madras

Biotechnology Department

BT5450: Data-driven Modeling and Optimization of Bioprocesses

Assignment III

Due date: Oct 31, 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 data scientist developed a clustering algorithm to cluster the biochemical reactions based on the distribution of the corresponding gene expression levels across tissues. (S)he applied that algorithm for reactions occurring in a human cell (say 4551 reactions) and grouped the reactions in to 26 clusters using the developed algorithm (look into the dataset 'Cluster_rxn_set.csv'). Now, the scientist is interested in finding out which clusters are enriched with more number of House keeping reactions (look into the dataset 'HK_rxns.csv' for housekeeping reactions in humans).
Perform a hyper-geometric test on each of the clusters and report the clusters which are enriched with more number of House keeping reactions (use $\alpha = 0.05$)

Q2 On different media, it has been reported that COVID-19 death rates are lower in India due to poor and water quality in India. We would like to understand how factually correct this statement with data provided by the publication `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7444648/`. This study is based on the hygiene hypothesis. According to wikipedia, The hygiene hypothesis can be described by the following statement:
Early childhood exposure to particular microorganisms (such as the gut flora and Helminth parasites) protects against diseases by contributing to the development of the immune system. In particular, a lack of exposure is thought to lead to defects in the establishment of immune tolerance (source: wikipedia)

A data scientist wants to understand the above hypothesis by comparing the average COVID deaths per million of various countries. Based on the news and the hypothesis, the scientist makes the following assumptions:

(a) People from low income countries have a larger exposure to microbes than any other country.

(b) COVID deaths per million of a particular country is independent of COVID deaths per million of other countries.

(c) The datasets have no bias to any particular population.

Help the scientist by comparing the average COVID deaths per million of the following populations using appropriate hypothesis testing. You should state appropriate null and alternative hypotheses, and choose appropriate test statistics. Note that you have to compare two populations.

Table 1: Populations to be compared for average COVID deaths per million

Si.No	Population 1	Population 2
1	High income	Rest of the world
2	High income and upper middle	Lower middle and low income
3	Upper middle	lower middle and low income

Explain whether or not the hygiene hypothesis is true in the case of COVID-19.

Also, the hygiene hypothesis cannot be true for all the disease cases. So, check whether the hygiene hypothesis is true in the case of diarrhoea. Use 'Fraction of diarrhoea due to lack of sanitation' in the provided dataset and make a hypothesis test by comparing average diarrhoea cases for the populations described in Table 1 (All the assumptions made for COVID deaths per million are true in this case too). (use Hygiene.csv dataset)

Q3 Mr. B has made a claim that average percentage of rural population for low income countries is 66. Choose appropriate test statistic assuming that the distribution is normal and answer the following questions.(use Hygiene.csv dataset)

(a) Is there evidence to support a claim that the average percentage of rural population from low income countries exceeds 66 with a significance level of 0.05?

(b) Compute power if true mean is 67.5.

(c) Explain how the question in (a) can be answered by constructing a two-sided confidence interval.

(d) How many samples are to be taken in order to have a true mean score of 70 if power of the test is to be 0.6?

Q4 For a biological system of nine metabolites (A, B, C, D, E, F, G, H and I), a researcher proposed two different possible topology for the biochemical reaction network as shown in the Figure 1. Note that the stoichiometric coefficient of the metabolites is 1 in all the reactions. Also, v_1, v_3, v_{14}, *and* v_{15} are exchange reactions that takes the respective metabolites in and out of the system. To identify the correct network, (S)he measured the rate of the reactions in triplicates (Data: 'Q4.csv') and made the following assumptions,

(a) The measured rate of the reaction can be written as $V_{m,i} = V_{a,i} + \epsilon$.
Where, i = 1,...,15; V_m is the measured flux values; V_a is the actual flux values; ϵ is the measurement error given by $\epsilon = \mathcal{N}(\mu = 0, \sigma^2 = 1)$

(b) There is no accumulation of the metabolites in the system i.e. $\frac{dX}{dt} = 0$. Where X is the vector of metabolite concentrations. At each node (denoted as metabolite), the fluxes in= the fluxes out.

(c) The measurement error, ϵ of a given rate of the reaction is independent of measurement errors of other flux values.

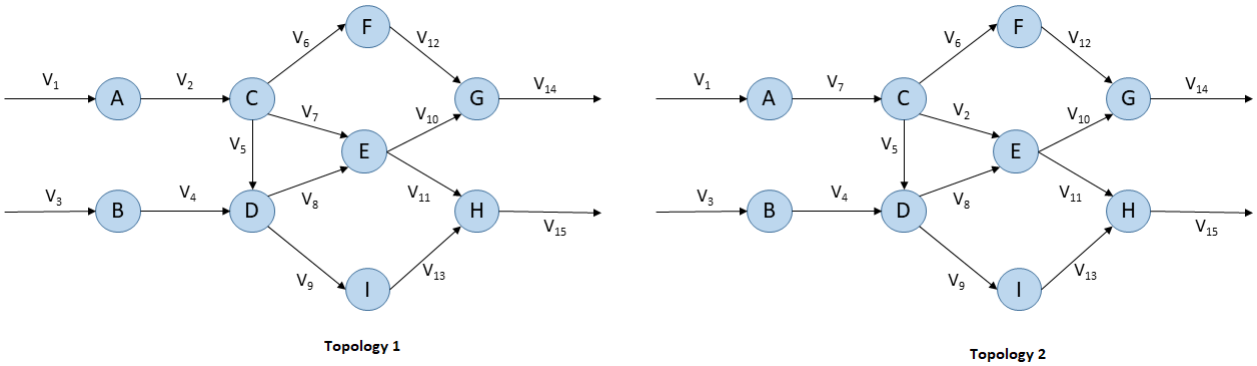

Figure 1: Two different reaction topology as proposed by the researcher

Help him/her to identify the correct topology by doing the following,

(a) Define the mass-balance equations for each of the metabolite in a given topology.

(b) Identify the parameters of the distribution of each of the derived mass-balance equations.

(c) state the null and alternate hypothesis for each of the mass-balance equations.

(d) Choose appropriate test statistic and accept/reject the null hypothesis with $\alpha = 0.05$ significance level.

(e) Based on the obtained p-values for each of the topology, report the correct bio-chemical network.

Q5 For a set of observations of a random variable (data: 'Q5.csv') , a function $\hat{f}$ is defined as,

$$\hat{f} = \frac{\sum_{i=0}^{i=N} \frac{exp(z_i)}{z_i^2}}{N}$$

Use the given dataset to accept/reject the null hypothesis with $\alpha = 0.05$ significance level as stated below,

$$\text{H0: } \hat{f} = 17$$
$$\text{H1: } \hat{f} \neq 17$$

Hint: Use bootstrap method to obtain confidence bounds.