MANIPAL SCHOOL OF INFORMATION SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

MANIPAL SCHOOL OF INFORMATION SCIENCES
25 years
MSIS
MAHE, MANIPAL
1998-2023

AML5103 | Applied Probability and Statistics | Coding Problem Set-1

1. After the Ebola outbreak of 2015 there was an urgent need to learn more about the virus. You have been asked to uncover how a particular group of bat-genes impact an important trait: whether the bat can carry the Ebola virus. Nobody knows the underlying mechanism; however, data could give us an insight into it. You have sampled data from 100,000 bats comprising the following information:

   - whether or not five genes are expressed in that bat;
   - and whether or not the bat can carry Ebola.

   If a gene is expressed, it can affect both the chance of other genes being expressed and the chance of Ebola virus being present or absent. You can find the data in a file called “bats.csv” in the `Codes/Data/` folder on the *APS* channel on Teams. Each row in the file corresponds to one bat and has 7 columns:

   - column-1 contains numerical identifier for the sampled bats;
   - column-2 through column-5 represent whether a particular gene is expressed in the bat (`True`) or not (`False`);
   - column-6 represents whether the bat carries Ebola (`True`) or not (`False`).

   Write a program to analyze the data you have collected and report the following:

   (a) What is the chance of a random bat carrying the Ebola virus?

   (b) For each gene, calculate the likelihood that it is expressed in a random bat.

   (c) Is the presence or absence of any of the genes indicative of a random bat potentially carrying the Ebola virus?