BIN504 Fall '23 Homework I

Due: December 3 by 23:59

Note: You can submit your answer to the programming problems (2 and 4) as R code only (at least with comments please). But it would be preferable if you created a R Markdown report containing both your code and your discussion of the solution (assumptions, approach etc.). R Markdown was discussed in the lecture; you can get further information about R Markdown at `https://rmarkdown.rstudio.com`. You can also do the non-programming problems (1 and 3) in R Markdown as well, if you wish.

Problem 1 (25 %)

Tay-Sachs disease is an *autosomal recessive disorder*[1] in children and infants which is fatal. Person A knows that both their parents are carriers of the disease.

a (5%) What is the probability that person A is a carrier of Tay-Sachs?

b (10%) Assume that the person is planning to have a child with another person (B), who is known to be a carrier of Tay-Sachs. What is the probability that the child is homozygous for the disease allele?

c (10%) Let's assume their first child did not have the disease. Update the probability in the first part of the question (that A is a carrier) given this information.

Problem 2 (25 %)

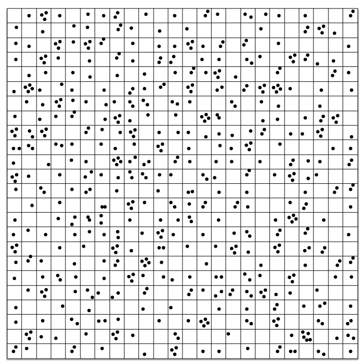
We will calculate the results for Probem 1, but this time you will be using R and creating a numerical simulation. Please estimate the following probabilities by simulating e.g. 10,000 random outcomes for Person A and their child and use the joint probability table thus created to calculate answers to all parts of the previous question.

[1] An autosomal recessive disorder develops only when two copies of a faulty gene are present (i.e. the person is homozygous for the disease allele). Being a carrier for the disease means that a person has only one copy of the disease allele, and can transmit it to offspring, but is not affected by the disease.

Problem 3 (25 %)

DeMine Inc. is developing a strain of bacteria that glow in the presence of TNT. The idea is to spray spores of this strain on a minefield and detect the location of landmines, so that they can be removed.

In order to test spore viability and effectiveness, they have rented a large square-shaped warehouse $24m \times 24m$ and have divided the floor into squares that are $1m$ by $1m$ each. In each square they spread a nutrient gel, spray the hangar with an aerosol containing the spores, and close up the hangar. After a certain amount of time, they open up the hangar, count the number of bacterial colonies in each square and find the following situation:



After they count, they find that:

- 1 square has 5 colonies,
- 7 squares have 4 colonies,
- 35 squares have 3 colonies,
- 93 squares have 2 colonies,
- 211 squares have 1 colony, and,
- 229 squares have no colonies.

The assumption is that each viable spore in the aerosol has formed a colony. Based on this, answer the following:

a (10%) If $\mathbf{X}$ is a random variable representing the number of spore colonies in each square, what do you think is the distribution that governs $\mathbf{X}$?

b (5%) Using your answer from part (a), find the expected value of $\mathbf{X}$ (the expected number of spores per m^2)

c (10%) The concentration of the aerosol used was C. Assuming the expected value in part (b) increases linearly with increasing aerosol concentration, what should the minimum concentration be if we want at least a 99.5% chance of there being at least one viable spore in each grid square?

Problem 4 (25 %)

Suppose the lifetime of a certain kind of fly follows an exponential distribution with unknown λ in days. Some flies have a mutation that shortens this to λ'. A researcher observes N flies continuously and records the lifetime of each fly as $y_1, y_2, \ldots, y_N$. You will use the EM algorithm to find an estimate of λ and λ'. The lifetimes recorded are given in the attached file `fly_lifetimes.txt`.

Prepare a program in R to estimate λ and λ' with EM. Your program should also estimate the ratio of mutated flies. Run your program with the data $y_1, y_2, \ldots y_{120} (N = 120)$ given in `fly_lifetimes.txt` and report the results.