Statistical Foundations of Data Science
Final Exam
June 2023
Time: 2 hours

---

1. One of the two functions given below is a probability density function (i.e. a distribution).

   A

   $$f(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & elsewhere \end{cases}$$

   B

   $$f(x) = \begin{cases} 3e^{-x} & x > 0 \\ 0 & elsewhere \end{cases}$$

   (a) Determine which one is a distribution and state why the other one isn't. [3]

   (b) Compute the mean of the distribution[3]

   (c) Compute the variance of the distribution[4]

2. A certain primary school institutes a swimming test for all grade six and seven learners. At this particular school there are twice as many grade sixes as grade sevens due to an expansion six years ago. Sixty percent of grade sevens pass the test the first time while only forty percent of grade sixes do. Find:

   a The probability that a randomly selected learner who passes the swim test on the first go is a grade seven. [5]

   b The probability that a randomly selected learner who doesn't pass the swim test on the first go is a grade six. [5]

3. Five chess players are travelling to a tournament. There are seven trains that can take them from the city centre to the suburb the tournament is at. Each player chooses a train to take at random (each train equally likely) and without consulting the others. Compute:

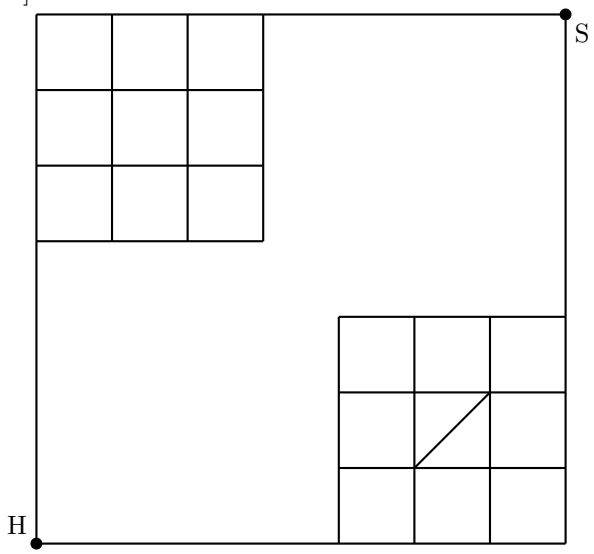
   a The probability that they all take different trains [5]

   b The probability that they all take the same train [3]

   c The probability that they are not all of the same train but that at least two share a train [2]

4. A chocolate company sells chocolate bars that are advertised as weighing 100g. More precisely they claim that these bars have weights that are normally distributed with mean 100g and standard deviation 5g in accordance with industry standards. For routine maintenance they want to check that this is the case. They weigh 25 bars and find that this sample has a mean weight is 98.5g. Perform a two sided hypothesis test at the five percent level to see if the machines are still performing up to standard. [10]

5. A learner walks from home to school. The roads between home (H) and school (S) are shown below. In how many ways can the learner move from home to school without moving left or down at any point. [10]



6. Five models are trained to make predictions on the results of a chess match. Each model is known to have an accuracy of sixty percent. You decide to try to make a more accurate predictor by choosing the player in each game who was chosen by the majority of the five original models. Compute the accuracy of your ensemble model if:

   a The predictions are mutually independent [6]

   b The predictions are correlated in a way which maximizes the likelihood of your predictor succeeding. [2]

   c The predictions are correlated in a way that minimizes the probability of your predictor succeeding.[2]

7. A sweet shop at Wits sells three types of ice-cream. Chocolate, Strawberry and Vanilla. In total 392 students who visit the shop and each one enjoys at least one type of ice-cream. 307 students enjoy vanilla and 196 students enjoy strawberry of these 125 students enjoy both strawberry and vanilla. 129 students enjoy only vanilla and 29 enjoy only strawberry and 34 enjoy all three flavours

   (a) How many students enjoy either strawberry or vanilla? [2]

   (b) How many enjoy only chocolate? [2]

   (c) How many students enjoy strawberry or vanilla but not chocolate? [4]

   (d) How many students enjoy chocolate [2]

8. A certain test tries to predict weather or not someone has the deadly X virus.

   Below is a table of the predictions made on a thousand patients

	Actually Infected	Actually Uninfected
Predicted Infected	800	60
Predicted Uninfected	80	60

   Compute :

   a The model's accuracy [3]

   b The model's balanced accuracy [3]

   c The model's F_1 score, using the infected group as the positive class [4]

9. A bag contains 5 green marbles, 7 yellow marbles, 6 blue marbles and 4 red marbles. 8 marbles are drawn from it. Compute:

   a The probability that at least two colours of marble are drawn [2]

   b The probability that no green marble is drawn [3]

   c The probability that no yellow marble is drawn [3]

   d The probability that no green or yellow marble is drawn [2]

10. A certain weighted coin is said to come up heads with a probability of only one percent. It's owner decides to flip it until the first head and this ends up takings 451 flips (coming up heads on flip 451 and tails the 450 times before that). Hearing about this you decide to perform a one sided hypothesis test at the three percent level. Perform this test. Hint: Recall that $a + ar + ar^2 + ar^3 + .. = \frac{1}{1-r}$

    (a) Remeber to state your null hypothesis, alternate hypothesis and significance level. [3]

    (b) Calculate the p-value and decide weather or not to reject the null hypothesis [7]