Data Analysis and Exploration
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]
   A. $\int_0^\infty 2e^{-2x}dx = 1$ while $\int_0^\infty 3e^{-x}dx = 3$

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

$$\begin{aligned} \int_0^\infty 2xe^{-2x}dx &= -xe^{-2x}|_0^\infty + \int_0^\infty e^{-2x}dx \\ &= 0 + \frac{1}{2} \\ &= \frac{1}{2} \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}[x^2] &= \int_0^\infty 2x^2e^{-2x}dx \\ &= -x^2e^{-2x}|_0^\infty + \int_0^\infty 2xe^{-2x}dx \\ &= 0 + 2\int_0^\infty xe^{-2x}dx \\ &= \frac{1}{2} \end{aligned}$$

   This last step uses part b

$$\begin{aligned} V(X) &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\ &= \frac{1}{2} - (\frac{1}{2})^2 \\ &= \frac{1}{4} \end{aligned}$$

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]

$$\begin{aligned}\mathbb{P}(\text{grade 7}|\text{passes}) &= \frac{\mathbb{P}(\text{grade 7 and passes})}{\mathbb{P}(\text{grade 7 and passes}) + \mathbb{P}(\text{grade 6 and passes})} \\ &= \frac{\frac{1}{3} \times 0.6}{\frac{1}{3} \times 0.6 + \frac{2}{3} \times 0.4} \\ &= \frac{3}{7} \\ &= 0.4285714..\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(\text{grade 6}|\text{fails}) &= \frac{\mathbb{P}(\text{grade 6 and fails})}{\mathbb{P}(\text{grade 7 and fails}) + \mathbb{P}(\text{grade 6 and fails})} \\ &= \frac{\frac{2}{3} \times 0.6}{\frac{1}{3} \times 0.4 + \frac{2}{3} \times 0.6} \\ &= \frac{3}{4} \\ &= 0.75\end{aligned}$$

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:

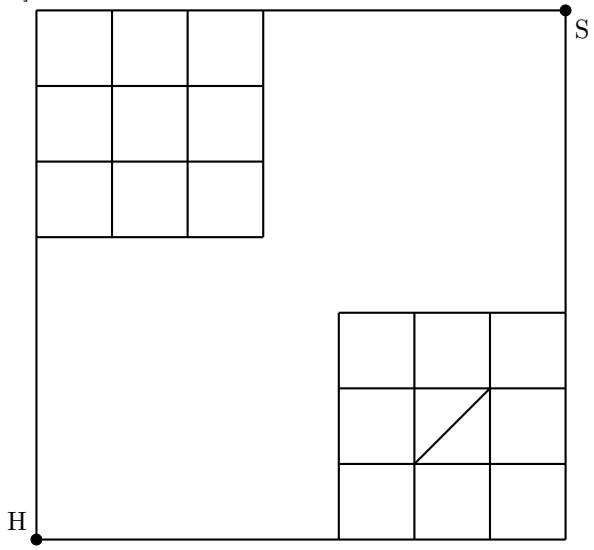
   a The probability that they all take different trains [5]
   We imagine them choosing trains one at a time. The first is guareenteed to be in his own train, the second has a $\frac{6}{7}$ chance of avoiding the first the third a $\frac{5}{7}$ and so on.
   $\frac{7}{7} \cdot \frac{6}{7} \cdot \frac{5}{7} \cdot \frac{4}{7} \cdot \frac{3}{7} = 0.14993752603$

   b The probability that they all take the same train [3]
   After the first player takes a seat the other four have a one in seven chance of joining the same train so $\frac{1}{7^4} = 0.00041649312$

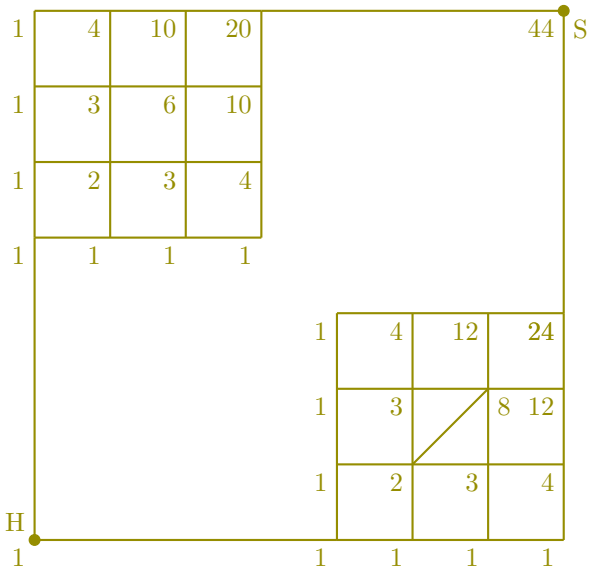
   c The probability that they are not all of the same train but that at least two share a train [2]
   $1 - 0.14993752603 - 0.00041649312 = 0.84964598085$

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]
   $H_0 : \mu = 100$
   $H_1 : \mu \neq 100$
   $\alpha = 0.05$
   Under the null hypothesis our sample mean $\overline{X}$ is distributed normally with mean 100 and standard deviation 1g. This is 1.5 standard deviations below the expected mean ans corresponds to a p-value of 0.1336. We fail to reject the null hypothesis and may continue without immediate repairs

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]



   44. This can be seen by recursive computation.



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]
   In this case the number of correct models is distributed binomially. Which leads to a probability of $\binom{5}{3}0.6^3 0.4^2 + \binom{5}{4}0.6^4 0.4^1 + \binom{5}{5}0.6^5 0.4^0 = 0.66304$

   b The predictions are correlated in a way which maximizes the likelihood of your predictor succeeding. [2]
   The expected number of correct predictions is $5 \times 0.6 = 3$. This can be evenly spread to give a winning probability of 1

   c The predictions are correlated in a way that minimizes the probability of your predictor succeeding.[2]
   The expected number of correct predictions is 3. The worst way for this to be spread is to have it always be 2 or 5 correct votes. To keep the expected value at 3 we should get 5 with probability $\frac{1}{3}$

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]
   We use the principle of inclusion exclusion. $307 + 196 - 125 = 378$

   (b) How many enjoy only chocolate? [2]
   Every student either likes only chocolate or likes at least one of the other two flavours. $392 - 378 = 14$.

   (c) How many students enjoy strawberry or vanilla but not chocolate? [4]
   Of our 392 we have 14 who enjoy only chocolate, 129 who enjoy only vanilla and 29 who enjoy only strawberry. These may be subtracted off to get 220 students who enjoy at least two flavours. Subtracting off the 34 who enjoy all three flavours gives 186 students who enjoy exactly two

flavours. We notice that $125 - 34 = 91$ students enjoy strawberry and vanilla only. Finally we have that $91 + 129 + 29 = 249$ students enjoy strawberry or vanilla but not chocolate.

(d) How many students enjoy chocolate [2]
$392 - 249 = 143$

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	60	60
Predicted Uninfected	80	800

   Compute :

   a The model's accuracy [3]

$$\begin{aligned}\text{Accuracy} &= \frac{\text{True Positives} + \text{True Negatives}}{\text{All Observations}} \\ &= \frac{800 + 60}{800 + 80 + 60 + 60} \\ &= \frac{860}{1000} \\ &= 0.86\end{aligned}$$

   b The model's balanced accuracy [3]

$$\begin{aligned}\text{Balanced Accuracy} &= \frac{\text{Sensitivity} + \text{Specificity}}{2} \\ &= \frac{\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} + \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}}{2} \\ &= \frac{\frac{60}{140} + \frac{800}{860}}{2} \\ &= 0.679402\end{aligned}$$

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

$$\begin{aligned}\text{F1 -score} &= \frac{\text{True Positives}}{\text{True Positives} + 0.5 \times (\text{False Positives} + \text{False Negatives})} \\ &= \frac{60}{60 + \frac{140}{2}} \\ &= \frac{\frac{60}{140} + \frac{800}{860}}{2} \\ &= 0.46153846153\end{aligned}$$

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]

1. No individual colour has 8 marbles.

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

$$\begin{aligned}\frac{\binom{17}{8}}{\binom{22}{8}} &= \frac{\frac{17\times16\times15\times14\times13\times12\times11\times10}{8!}}{\frac{22\times21\times20\times19\times18\times17\times16\times15}{8!}} \\ &= \frac{14\times13\times12\times11\times10}{22\times21\times20\times19\times18} \\ &= 0.07602339181\end{aligned}$$

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

$$\begin{aligned}\frac{\binom{15}{8}}{\binom{22}{8}} &= \frac{\frac{15\times14\times13\times12\times11\times10\times9\times8}{8!}}{\frac{22\times21\times20\times19\times18\times17\times16\times15}{8!}} \\ &= \frac{14\times13\times12\times11\times10\times9\times8}{22\times21\times20\times19\times18\times17\times16} \\ &= 0.020123839\end{aligned}$$

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

$$\begin{aligned}\frac{\binom{10}{8}}{\binom{22}{8}} &= \frac{\frac{10\times9\times8\times7\times6\times5\times4\times3}{8!}}{\frac{22\times21\times20\times19\times18\times17\times16\times15}{8!}} \\ &= \frac{10\times9\times8\times7\times6\times5\times4\times3}{22\times21\times20\times19\times18\times17\times16\times15} \\ &= 0.00014072614\end{aligned}$$

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 $a + ar + ar^2 + ar^3 + .. = \frac{a}{1-r}$

(a) Remember 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]

$$H_0 : p = 0.01$$

$$H_1 : p = 0.01$$

$$\alpha = 0.03$$

We compute our p-value as follows the probability of getting at least 451 flips before a head. This works out to:

$$
\begin{aligned}
(1-p)^{450}p + (1-p)^{451}p + (1-p)^{452}p + (1-p)^{453}p + ... &= \\
(1-p)^{450}p(1 + (1-p) + (1-p)^2 + ...) &= \\
(1-p)^{450}p\frac{1}{1-(1-p)} &= \\
(1-p)^{450} &= \\
0.99^{450} &= \\
0.01086019363
\end{aligned}
$$

This is lower than our cut-off so we reject the null hypothesis. This could also be seen by observing that it's the same as the first 450 flips being tails.