MTHM601 Fundamentals of Data Science
Assessed Exercises

Please make sure that the submitted work is your own. This is NOT a group assignment.

1. *[4 marks]* Consider the following equations, where η is a real constant.

$$x + y + z = 1$$
$$x + 2y + 4z = \eta$$
$$x + 4y + 10z = \eta^2$$

   (a) By inspection or considering a determinant, show that these equations will not have a unique solution for any value of η.

   (b) By Gaussian elimination, show that these equations will only have solutions for two values of η. Give a complete characterisation of the solutions in both of these cases.

2. *[6 marks]* Two dodgeball players, Alice and Ben, take alternate throws at each other, and a player wins the dodgeball game when their throw hits the other player. Alice's throws have a probability α of hitting Ben, while Ben's throws have a probability β of hitting Alice.

   (a) Consider Alice going first. Let P_1 denote the probability that Alice wins when she is first to throw. Give, in terms of P_1, α, and β, the probability of Alice missing Ben with her first throw and going on to win.

   (b) In terms of α and β, what is the probability P_1 of Alice winning the dodgeball game if Alice is first to throw.

   (c) In terms of α and β, what is the probability P_2 of Alice winning the dodgeball game if Ben is first to throw.

3. *[12 marks]* The company Bayeseinz makes tins of baked beans. It advertises that each tin weighs 426g, which it claims is "more than other brands!". The weight of a single tin of baked beans follows a normal distribution with a mean of 426g and a standard deviation of 21g. The tins are packed in boxes of 100. To protect the consumer, the Advertising Standards Agency are considering the following three approaches to fining Bayeseinz for misrepresenting the weight of their tins:

   A: The entire box of tins is weighed and the vendor is fined £3000 if the average weight of these tins is less than 420.75g

   B: Twenty five tins of baked beans are selected at random from the box. The company is fined £120 if the average weight of these tins is less than 421.8g.

   C: Tins of baked beans are removed one at a time, at random, until one has been found that weighs more than 426g. The company is fined £$5n(n-1)$, where n is the number of tins removed. If no such tin is found by the end of a box, the next box is opened.

   Complete the following:

   (a) By considering that the sum of an infinite geometric series for $|r| < 1$ is given by

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \tag{1}$$

give an expression for the sum

$$\sum_{n=0}^{\infty} n(n-1)r^{n-2}.$$

(b) Calculate the expected fine per box for schemes A, B, and C.

4. *[13 marks]* The Penryn Probability Society must elect a new chairperson to lead the Society for the coming year of meetings. At the time of the election, there are 363 members. Two of these members, M and B, join the election contest, where the aim is to get the majority support of the remaining 361 members of the Society.
We further assume that on the day the election for chairperson is announced 186 of these members support M, and the remaining 175 members support B in becoming the next chairperson. The announcement is followed by an election campaign, during which members can decide to change their allegiance. In particular, we know that on any given day, there is a probability of 0.005 that a member who has been supporting M will become a B supporter by the end of the day, while the probability that a member who has been supporting B will become an M supporter by the end of the day is 0.004. Each member makes their decision independently of each other, and independently of the decision they made the day before.

(a) Introduce the following random variables:

$$X_i^{(1)} = \begin{cases} 1, & \text{B supporter number } i \text{ still supports B at the end of day 1,} \\ 0, & \text{B supporter number } i \text{ changes to an M supporter at the end of day 1,} \end{cases}$$

for $i = 1, \ldots, 175$; and

$$X_i^{(2)} = \begin{cases} 1, & \text{M supporter number } i \text{ changes to a B supporter at the end of day 1,} \\ 0, & \text{M supporter number } i \text{ still supports M at the end of day 1,} \end{cases}$$

for $i = 1, \ldots, 186$.
Using the random variables $X_i^{(1)}$ $(i = 1, \ldots, 175)$ and $X_i^{(2)}$ $(i = 1, \ldots, 186)$, express the number of B supporters at the end of the first day, then use your formula to find the expected number of B supporters at the end of the first day. Justify every step of your argument.

(b) Using the random variables $X_i^{(1)}$ $(i = 1, \ldots, 175)$ and $X_i^{(2)}$ $(i = 1, \ldots, 186)$, express the number of M supporters at the end of the first day. What is the expected number of M supporters at the end of the first day?

(c) `R`: The election campaign is set to last for 2 weeks. This means that each member would vote according to the allegiance they have at the end of day 14, that is, the candidate they would vote for is the one they are supporting after the first 14 days of the campaign. Using simulation, find the probability that in this election B would hold the majority of the votes among the 361 members, and thereby win the election.

(d) `R`: Now suppose that the election had to be postponed, and with the new date, candidates now have a 60 day long campaign period (as opposed to 14 days). Adjust your code from part 4c to find the probability that B will win the delayed election. How does this probability compare to the one computed in part 4c?

5. *[24 marks]* The following data are the observed frequencies of the number of strikes y in a ship building company in the UK during 1948-1961:

No. of Strikes	0	1	2	3	4
Frequency	137	33	10	1	1

Thus $n = 182$ and $\sum_{i=1}^{182} y_i = 60$. Assume a Poisson model for these data with parameter λ.

(a) Suggest a point estimator for the Poisson parameter λ, and obtain a standard error for this estimator. Next, use the provided data to calculate an approximate 95% confidence interval for λ. Does the confidence interval support the hypothesis at the 5%-level that $\lambda = 1$?

(b) Suppose only those periods during which there was at least one strike are of interest (note that in this case the zero entries in the dataset are not considered!). An appropriate model here has the following probability mass function

$$f(y;\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^y}{(1-\mathrm{e}^{-\lambda})\,y!}, \quad y = 1, 2, 3, \ldots \text{ and } \lambda > 0.$$

Obtain the log-likelihood function for this case and show that the equation for the maximum likelihood estimator $\hat{\lambda}$ reduces to solving:

$$\frac{\hat{\lambda}}{1-\mathrm{e}^{-\hat{\lambda}}} = \frac{4}{3}.$$

That is, the maximum likelihood estimate is the root of the function

$$f(\hat{\lambda}) = \frac{\hat{\lambda}}{1-\mathrm{e}^{-\hat{\lambda}}} - \frac{4}{3}. \tag{2}$$

Note that you're not expected to solve the above equation manually.

(c) `R`: Plot the derivative of the loglikelihood function from Q5(b) and identify an interval of length one that contains the root of this function.

(d) `R`: Use the `optimise` function to determine the maximum likelihood estimate $\hat{\lambda}$ to three decimal places. (Looking at the help page of `optimise` reveals that `optimise` needs a function to minimise, furthermore it has a `lower` and `upper` argument that specifies the two endpoints of the interval to be searched). The function that needs to be minimised in order to obtain the maximum likelihood estimate is the **absolute value of the function** (2) **from Q5(b)** (you need to define this function), while a suitable interval can be the one identified in Q5(c).

6. *[17 marks]* The random variable Y has a Binom(n,p) distribution where n is known and $p \in [0,1]$ is to be estimated.

(a) Show that the method of moments estimate of p on observing $Y = y$ is $\hat{p} = \frac{y}{n}$.

From now on we shall write the method of moments estimator of p as $T = T(Y) = \frac{Y}{n}$.

(b) Find the mean and variance of T. What are the bias and mean square error of T as an estimator of p?

(c) An investigator is interested in the parameter $\theta = p^2$ and proposes to use $S = S(Y) = \frac{Y^2}{n^2}$ to estimate θ. Show that S is a biased estimator of θ.

(d) Find an expression of the form $aS + bT$ which gives an unbiased estimator of θ.

7. *[24 marks]* `R`: The file 'ozone.csv', available on the course ELE page, contains information on ozone levels recorded over 111 days in an urban location. The variables measured were:

`ozone`	Ozone levels in parts per billion (ppb)
`radiation`	in langleys
`temperature`	in farenheit
`wind`	in miles per hour (mph)

Read these data into `R` and answer the following questions.

(a) `R`: Carry out exploratory data analysis, and produce a matrix scatterplot of the dataset. Comment on your findings and what these plots suggest about the likely relationships between the response variable (`ozone`) and the other variables.

(b) `R`: Fit a multiple regression of `ozone` as the response variable, against `radiation`, `temperature` and `wind` as the explanatory variables (use all three, when fitting the model). Comment on the summary of the model. What do these coefficients suggest about the relationship between `ozone` and the other variables? Are these findings consistent with your earlier descriptive plots? Also include suitable residual plots, commenting as appropriate.

(c) `R`: A colleague suggests you implement the following model,

$$\log(\text{ozone}_i) = \beta_0 + \beta_1 \log(\text{radiation}_i) + \beta_2 \log(\text{temperature}_i) + \beta_3 \log(\text{wind}_i) + \epsilon_i \quad \text{where } \epsilon_i \sim N(0, \sigma^2).$$

Fit this new model to the data to obtain estimates for the regression coefficients. Produce a plot of the residuals against the fitted values, and a Q-Q plot of the residuals. Comment on the outputs from the modelling (comparing it to the previously fitted model), paying particular attention to the interpretation of the coefficients. Express the impact of the explanatory variables on the ozone levels, with the latter expressed on the original (untransformed) scale.

Total for paper = 100 marks.

The questions apart from those marked with " `R:`" are theoretical exercises, and should be solved using results we covered in lectures. Show all of your working and make sure you justify each step of the theoretical reasoning by clearly stating the theorem/property you are using (marks will be awarded for these). For the R questions, make sure that you add comments to each section of your R code, explaining what you're doing. And all of the relevant output from your R code (computed probabilities, plots, and so forth) should also be included in your submission. A pdf document with your R code and the solutions to this exercise sheet should be submitted through ELE by 12 noon on 27th November. The solutions to the theoretical parts may be a scanned in version of handwritten work (be sure it is clear and legible) or typeset with Rmarkdown or LaTeX. Note that late submissions will be penalised unless there is an approved mitigation request according to the standard University procedures.

We encourage you to discuss your work in groups, however your submitted assessment must be your own work. The first page of your solutions must include the statement "I have familiarised myself with the academic misconduct and plagiarism guidelines in the *Academic Honesty and Plagiarism* module and the MTHM601 (Fundamentals of Data Science) ELE site's Assessment Information tab. This submission constitutes my own work, I have explicitly referenced and acknowledged those parts that draw on the literature and online sources, and I have not discussed my answers directly with others." You must also state your use of Generative AI with a Gen-AI statement like the one on ELE.