THE UNIVERSITY OF MANCHESTER

One Hour

COURSE TITLE: STATISTICS AND MACHINE LEARNING I

Date: ???
Time: ???

Answer all questions
In total you may achieve **FORTY** (40) marks.

Please enter your answers in the Answer Book provided.

Electronic calculators may be used, provided that they cannot store text

P.T.O

Answer **ALL** questions.
For maximal points answers have to be correct and well-motivated

1. Assume an intercept-free simple linear regression model

$$Y_i = \beta x_i + \epsilon_i \quad \text{for } i = 1, \dots, n$$

where

- Y_i is the response,
- $x_1, \dots, x_n$ are fixed observables representing predictors,
- $\epsilon_1, \dots, \epsilon_n$ are normally distributed and independent random variables with expected value zero and variance σ^2, and
- β and σ^2 are unknown parameters (constants).

For example, x_i could be the height of an individual and Y_i the weight.

(a) What is the expected values of Y_i? (2 marks)

(b) What is the variance of Y_i? (2 marks)

(c) When is it justified to use a linear regression model without an intercept term? What are the consequences of excluding this term? (2 marks)

(d) A suggested estimator for β is

$$\hat{\beta}' = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

Show that $\hat{\beta}'$ is an unbiased estimator of β. (3 marks)

(e) Another suggested estimator for β is

$$\hat{\beta}'' = \frac{\bar{Y}}{\bar{x}}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Show that $\hat{\beta}''$ is an unbiased estimator of β. (3 marks)

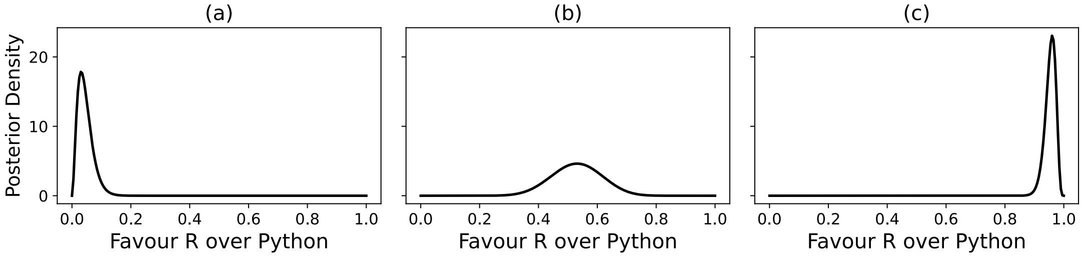
(f) Now you have to choose one of the two estimators $\hat{\beta}'$ and $\hat{\beta}''$. What would you base this choice on and why? (3 marks)

2. The expressions below are probability mass or density functions for distributions we studied during the term: one is for the Beta distribution, while the other is for the Binomial distribution.

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, p^{\alpha - 1}(1 - p)^{\beta - 1} \qquad g(k) = \frac{N!}{k!\,(N - k)!}\, p^{k}(1 - p)^{N - k}$$

(a) Say which function corresponds to which distribution and explain the roles of the symbols p, α, β, k and N. (5 marks)

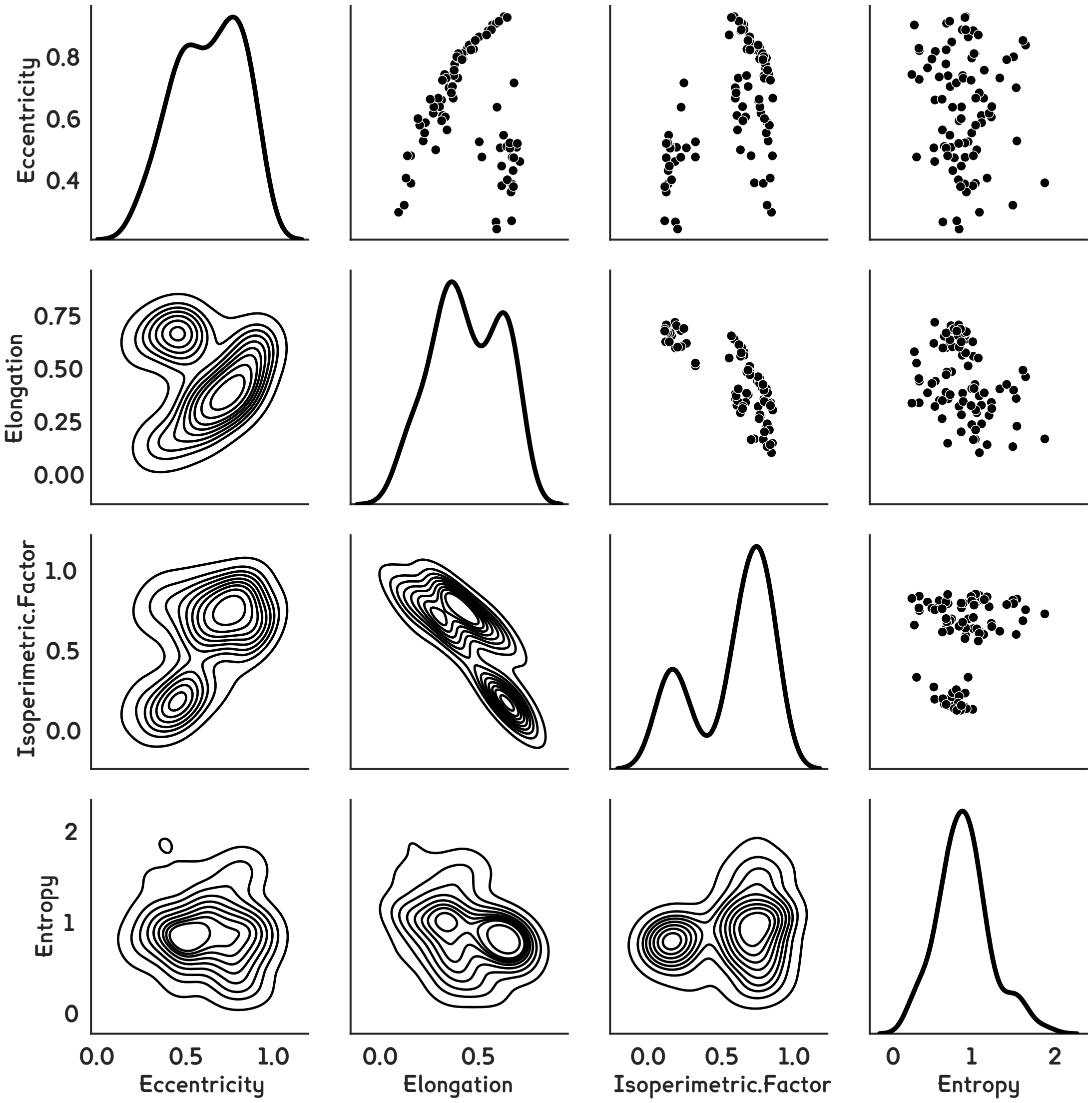
The panels below show posterior distributions for the proportion of Data Science students who prefer R to Python in three different MSc programmes. In all three cases, an uninformative prior was updated in light of the results of a poll.



(b) Answer the following questions: (5 marks)

- Which poll do you think had the largest sample? Explain your reasoning.
- Which of these densities would be best estimated by a suitable Laplace approximation? Again, explain your reasoning.

The figure below is part of an exploratory data analysis on a data set consisting of measurements on photographs of leaves from several populations.



(c) What insights can you draw from this figure? How many populations do you think there are, and why? (5 marks)

3. In statistical learning, there are two unwanted outcomes: *under-fitting* and *over-fitting*.

(a) Explain what the under-fitting and the over-fitting mean, respectively, in the context of statistical learning. (4 marks)

(b) Describe one method that can detect under-fitting and over-fitting in statistical learning. It is essential to give the main steps in your chosen method.

(3 marks)

(c) Describe one generic method that tends to prevent under-fitting and over-fitting in statistical learning. It is essential to give the main steps in your chosen method. (3 marks)

END OF EXAMINATION