DATA9001 Fundamentals of Data Science

Term 2, 2025

Assignment 1

Note:

- This assignment is due **Friday 20th June, 5pm (Week 3)** and must be uploaded to Moodle.
- Make sure you submit your work in **one** PDF format using the following name `A1-z1234567-FirstName-Surname.pdf`.
- Assignments without a signed plagiarism declaration (below) will not be accepted, and late assignments will not be accepted unless accompanied by medical certificates.
- This assignment counts for 15% of the final course mark.
- There are a total of **4 exercises** and **20 marks**.
- Worked solutions are required for full marks.

I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere. I acknowledge that the assessor of this item may, for the purpose of assessing this item reproduce this assessment item and provide a copy to another member of the University; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

Name: Student No: Signature: Date:

Tutorial Time: Tutor's name:

Exercise 1 [7 marks]

Part A [3 marks]

In the following exercise, a standard deck of cards is split into 2:

- Deck D_1 contains all honour cards (all 4 Jacks, Queens, Kings, and Aces, a total of 16 cards),
- Deck D_2 contains all non-honour cards (all 4 Twos, Threes, ..., and Tens, a total of 36 cards).

We roll two fair dice. If the number from the two dice added up is > 7, we then draw two cards from D_1 (without replacement), otherwise we draw one card from D_1 and one card from D_2.

1. What is the probability of drawing 2 Jacks? **[1 mark]**
2. What is the probability of drawing 2 non-honour cards? **[1 mark]**
3. What is the probability of drawing at least 1 honour card? **[1 mark]**

Part B [4 marks]

In the following exercise, we are interested in the content of School Dinners for 3 consecutive days: Day 1, Day 2 and Day 3. Suppose the probability of the dinner containing radishes on Day 1 is 0.43. If the dinner contained radishes on a particular day, the probability of it containing radishes again the next day is given as 0.71. Furthermore, if the dinner did not contain radishes on a particular day, the probability of it containing radishes on the next day is given as 0.36.

You may assume that the content of the School Dinner for any particular day only depends on the content of the School Dinner for the previous day and nothing else.

1. What is the probability that the dinners contain radishes on all 3 days? **[2 marks]**
2. What is the probability that the dinner contains radishes on Day 1 given that it will contain radishes on Day 2? **[2 marks]**

Exercise 2 [5 marks]

A treasure chest contains 3 diamonds and 5 lumps of coal. A treasure hunter successively draws 6 objects from the chest without replacement. Let X be the random variable that takes the value k if the first diamond appears on the k-th draw.

1. What is the expected value of X? **[2 marks]**
2. What is the standard deviation value of X? **[3 marks]**

Exercise 3 [3 marks]

Let $X_1, \ldots, X_n$ be a random sample from a N(0, 1) distribution, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let X_{n+1} be another independent observation from the same population.

1. What is the distribution of $\sum_{i=1}^{n}(X_i - \bar{X})^2 + X_{n+1}^2$? Why? **[1 mark]**
2. What is the distribution of $\sqrt{n}X_{n+1}/\sqrt{\sum_{i=1}^{n} X_i^2}$? Why? **[1 mark]**
3. What is the distribution of $(n-1)(\sum_{i=1}^{n}(X_i - \bar{X})^2 + X_{n+1}^2)/(n\sum_{i=1}^{n}(X_i - \bar{X})^2)$? Why? **[1 mark]**

Exercise 4 [5 marks]

Consider the `College` dataset available in the `R` package `ISLR`. This dataset contains statistics for a large number (777) of US Colleges from the 1995 issue of US News and World Report.
If needed, install the `ISLR` package by typing the following command in the console:

```
install.packages("ISLR")
```

Then load the package as well as the dataset using the commands:

```
library(ISLR)
data(College)
```

In the console, we can display the dataset by typing

```
College
```

and for example, access the variable `Accept` describing the number of applications accepted by each College by typing

```
College$Accept
```

Answer the following questions in `R`, providing your code.

1. What is the minimum and maximum number of applications accepted? **[1 marks]**
2. What is the mean and standard deviation of the number of applications accepted? **[2 marks]**
3. Use the `hist()` function and its argument `breaks` to draw a 17-bin histogram of the number of applications accepted. Consider typing `?hist` to access the help page. Describe what you observe. **[1 marks]**
4. Assume that the number of applications accepted by a College can be modelled as an exponential random variable with parameter β given by the observed mean from part 1. What is the probability that a College accepts more than $2,000$ applications?**[1 marks]**