MAST7866
Foundations of Data Science
Assessment 1

Attempt all questions and Please submit a PDF file electronically on Moodle via submission box **by** 5th November 2025. Note that

- Please attempt all questions.
- This assessment will be marked out of 100.
- You need to show all of your working.
- Following the instruction, if R is used, please provide the R-codes and the corresponding R-outputs.

1. Let A, B and C be three events associated with a random experiment. Express the following verbal statements in set notation:
   (a) at least one of the events occurs; [**2 marks**]
   (b) exactly one of the events occurs; [**2 marks**]
   (c) exactly two of the events occur; [**2 marks**]
   (d) not more than two of the events occur simultaneously. [**2 marks**]

2. (a) A lot consists of 5 good articles, 4 articles with minor defects and 2 with major defects. One article is chosen at random from the lot. Find the probability that:
      (i) it has no defects, [**3 marks**]
      (ii) it has no major defects, [**3 marks**]
      (iii) it is either good or has major defects. [**4 marks**]
   (b) Rings are mass-produced in a factory. The target internal diameter is 45 mm but records show that the diameters are normally distributed with mean 45 mm and standard deviation 0.04 mm. An acceptable diameter is one within the range 44.94 mm to 45.04 mm. Use R commands to calculate the probability that the output is unacceptable.
   [**5 marks**]

3. The following information is known about the three events A, B and C:

$$\Pr(A) = \tfrac{1}{3}\ , \qquad \Pr(B) = \tfrac{1}{4}\ , \qquad \Pr(C) = \tfrac{1}{2}\ , \quad \Pr(A \cup B \cup C) = 1\ ,$$
$$\Pr(A \cap C) = 0\ , \quad \Pr(B \cap C) = 0\ .$$

Show that

(a) $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$, **[5 marks]**

(b) $\Pr(A \cup C) = \Pr(A) + \Pr(C)$. **[5 marks]**

4. Choose the correct R function to put elements of vector x in increasing numerical order in the following list:

(a) order(x),

(b) mean(x),

(c) max(x),

(d) sort(x).

[5 marks]

5. The cost of car insurance (in £'s) for people living in suburban and urban areas are as follows:

Suburban: 58.5, 60.8, 60.6, 64.3, 64.1, 40.7, 43.7, 48.6, 49.3, 49.5

Urban: 97.8, 68.3, 109.2, 78.1, 113.7, 78.6, 122.0, 84.4, 125.5, 85.3

Read the data in R and use R commands:

(a) to calculate the median of the suburban car insurance costs; **[2 marks]**

(b) to calculate the inter-quartile range of the urban car insurance costs (rounded to the nearest integer); **[2 marks]**

(c) to calculate the sample standard deviations of both the urban and suburban datasets and to comment your results; **[2 marks]**

(d) to plot box-plots of both variables and to comment the distribution shapes of the costs in these areas; **[4 marks]**

(e) to state true or false for the following statements:

- The car insurance prices are higher in urban areas than in suburban areas.
- There are outliers in the urban dataset.

[4 marks]

6. Forbes magazine conducts an annual survey of the salaries of chief executive officers. In addition to salary information, Forbes collects and reports personal data on the CEOs, including their level of education and age. Data are saved in the text file **forbes.txt** which is available from moodle.

   (a) Import the data into R using the commands read.table() and attach(). [**2 marks**]

   (b) What are the results of the following operations? Illustrate the numerical results.

      (i) length(age), length(salary) and length(degree); [**3 marks**]

      (ii) sum(age$<$ 55), sum(salary[age $<$ 55])/sum(age $<$ 55) and sum(salary|age $<$ 55)/sum(age $<$ 55); [**3 marks**]

      (iii) salary/age. [**2 marks**]

   (c) Use the R-command to calculate the minimum, maximum, and average ages of these CEOs. Similarly, calculate the minimum, maximum and average salaries of these CEOs.[**5 marks**]

   (d) Table the degrees in the last column of the file using the command table(). [**4 marks**]

   (e) Use both a bar chart and a pie chart to portray the degrees of the CEOs. Do most CEOs have advanced degrees, such as masters or PhD? [**6 marks**]

   (f) Make a scatter plot of the ages of these CEOs against their salaries. How the CEO's salaries are related to their ages? Linear or non-linear? [**5 marks**]

   (g) Calculate the log-transformations of the CEO's ages and salaries, and make a scatter plot of these transformed data. Comment on your findings. [**5 marks**]

7. Your cell phone bill varies from month to month. Suppose your bill has the following monthly amounts 49, 36, 42, 43, 49, 30, 48, 32, 49, 35, 30, 48. Enter this data into a variable called bill. Use R commands to find the amount you spent this year on the cell phone, and the smallest and largest amounts you spent in a month. How many months was the amount greater than 41? What percentage was this? [**13 marks**]