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: Vansh Jain Student No: z5608613 Signature: Vansh Date: 20/06/2025

Tutorial Time: Thursday 7pm-8pm Tutor's name: Prof S Sisson

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**]

DATA9001 Statistics Assignment 1

Exercise 1

Part A

1. Probability of drawing 2 Jacks

Drawing two Jacks requires the sum of the dice to be greater than 7 so that two cards are drawn from D1, and both of the drawn cards must be Jacks.

The probability the total sum of the dice is > 7:
Total sample space: 36
Possible sums greater than 7 are 8, 9, 10, 11, 12. Total outcomes$= 15$

$$P(\text{sum} > 7) = \frac{15}{36} = \frac{5}{12}$$

Given that we're drawing from D1 which 16 cards including the 4 Jacks, the probability of drawing two Jacks:

$$P(\text{two Jacks} \mid \text{sum} > 7) = \frac{\binom{4}{2}}{\binom{16}{2}} = \frac{6}{120} = \frac{1}{20}$$

the total probability:

$$P(\text{two Jacks}) = \frac{5}{12} \times \frac{1}{20} = \frac{1}{48}$$

2. Probability of drawing 2 non-honor cards

This is impossible because:

- If sum > 7: Both cards come from D1 which are all honor cards
- If sum ≤ 7: One card from D1 (honor) and one from D2 (non-honor)

So, the probability is $\boxed{0}$

3. Probability of drawing at least 1 honor card

This always occurs because:

- If sum > 7: Both cards come from D1 which are all honor cards
- If sum ≤ 7: One card from D1 (honor) and one from D2 (non-honor)

Every possibility guarantees at least 1 honor card. So, the probability is $\boxed{1}$

Part B

1. Probability that the dinners contain radishes on all 3 days

Let X be the event that radishes are served on Day 1.
Let Y be the event that radishes are served on Day 2.
Let Z be the event that radishes are served on Day 3.

Given:

- $P(X) = 0.43$
- $P(Y \mid X) = 0.71$ (probability of radishes on Day 2 given radishes on Day 1)
- $P(Z \mid Y) = 0.71$ (probability of radishes on Day 3 given radishes on Day 2. It is given the dinner for any particular day only depends on the dinner for the previous day and nothing else.)

The probability of radishes on all three days is:

$$P(X \cap Y \cap Z) = P(X) \cdot P(Y \mid X) \cdot P(Z \mid X, Y)$$

This changes to:

$$P(X \cap Y \cap Z) = P(X) \cdot P(Y \mid X) \cdot P(Z \mid Y)$$

Substituting the values:

$$P(X \cap Y \cap Z) = 0.43 \times 0.71 \times 0.71 = 0.216763$$

Thus, probability the dinners contain radishes on all three days is $\boxed{0.217}$.

2. Probability that the dinner contains radishes on Day 1 given that it contains radishes on Day 2

$P(\text{Radishes are served on Day 1}|\text{Radishes are served on Day 2})$

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

$$P(X \cap Y) = P(X) \cdot P(Y \mid X) = 0.43 \times 0.71 = 0.3053$$

To find $P(Y)$, it depends on whether radishes were served on Day 1 or not.

- $P(\overline{X}) = 1 - P(X) = 1 - 0.43 = 0.57$
- $P(Y \mid \overline{X}) = 0.36$ (Given probability of radishes on Day 2 if there were no radishes on Day 1)

Using the law of total probability:

$$P(Y) = P(Y \mid X) \cdot P(X) + P(Y \mid \overline{X}) \cdot P(\overline{X}) = (0.71 \times 0.43) + (0.36 \times 0.57)$$

$$P(Y) = 0.3053 + 0.2052 = 0.5105$$

$$P(X \mid Y) = \frac{0.3053}{0.5105} \approx 0.598039 \approx 0.598$$

Thus, probability the dinner contains radishes on Day 1 given that it contains radishes on Day 2 is $\boxed{0.598}$.

Exercise 2

1. Expected Value of X

Total items = 3 diamonds + 5 coal = 8

Possible values for X are $k = 1, 2, 3, 4, 5, 6$.

$$\begin{aligned}
P(X = 1) &= \frac{3}{8} = \frac{21}{56} \\
P(X = 2) &= \frac{5}{8} \times \frac{3}{7} = \frac{15}{56} \\
P(X = 3) &= \frac{5}{8} \times \frac{4}{7} \times \frac{3}{6} = \frac{10}{56} \\
P(X = 4) &= \frac{5}{8} \times \frac{4}{7} \times \frac{3}{6} \times \frac{3}{5} = \frac{6}{56} \\
P(X = 5) &= \frac{5}{8} \times \frac{4}{7} \times \frac{3}{6} \times \frac{2}{5} \times \frac{3}{4} = \frac{3}{56} \\
P(X = 6) &= \frac{5}{8} \times \frac{4}{7} \times \frac{3}{6} \times \frac{2}{5} \times \frac{1}{4} \times 1 = \frac{1}{56}
\end{aligned}$$

The expected value is:

$$\begin{aligned}
E[X] &= \sum_{k=1}^{6} k \cdot P(X = k) \\
&= 1 \cdot \frac{21}{56} + 2 \cdot \frac{15}{56} + 3 \cdot \frac{10}{56} + 4 \cdot \frac{6}{56} + 5 \cdot \frac{3}{56} + 6 \cdot \frac{1}{56} \\
&= \frac{126}{56} = \boxed{\frac{9}{4}} = \boxed{2.25}
\end{aligned}$$

2. Standard Deviation of X

Using

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

$$
\begin{aligned}
E[X^2] &= \sum_{k=1}^{6} k^2 \cdot P(X = k) \\
&= 1 \cdot \frac{21}{56} + 4 \cdot \frac{15}{56} + 9 \cdot \frac{10}{56} + 16 \cdot \frac{6}{56} + 25 \cdot \frac{3}{56} + 36 \cdot \frac{1}{56} \\
&= \frac{1}{56}(21 + 60 + 90 + 96 + 75 + 36) \\
&= \frac{378}{56} = \frac{27}{4}
\end{aligned}
$$

The variance:

$$
\begin{aligned}
\mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\
&= \frac{27}{4} - \left(\frac{9}{4}\right)^2 \\
&= \frac{27}{4} - \frac{81}{16} \\
&= \frac{108}{16} - \frac{81}{16} \\
&= \frac{27}{16}
\end{aligned}
$$

The standard deviation is:

$$
\begin{aligned}
\mathrm{SD}(X) &= \sqrt{\mathrm{Var}(X)} \\
&= \sqrt{\frac{27}{16}} \\
&= \frac{\sqrt{27}}{4} \\
&= \boxed{\frac{3\sqrt{3}}{4}}
\end{aligned}
$$

Exercise 3

1. Distribution of $\sum_{i=1}^{n}(X_i - \bar{X})^2 + X_{n+1}^2$

$\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$ => For a normal distribution, $\sum(X_i - \bar{X})^2$ is the sample variance with $n-1$ degrees of freedom.

$X_{n+1}^2 \sim \chi^2_1$ => If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$
These are independent because X_{n+1} is independent of $X_1, \ldots, X_n$
So $\chi^2_{n-1} + \chi^2_1 = \chi^2_n$

$$
\boxed{\sum_{i=1}^{n}(X_i - \bar{X})^2 + X_{n+1}^2 \sim \chi^2_n}
$$

2. Distribution of $\frac{\sqrt{n}X_{n+1}}{\sqrt{\sum_{i=1}^n X_i^2}}$

$$\frac{\sqrt{n}X_{n+1}}{\sqrt{\sum_{i=1}^n X_i^2}} = \frac{X_{n+1}}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}}$$

We know:

- $X_{n+1} \sim N(0,1)$
- $\sum_{i=1}^n X_i^2 \sim \chi_n^2$
- Numerator and denominator are independent

This is the Student-t distribution

$$\boxed{\frac{N(0,1)}{\sqrt{\chi_n^2/n}} \sim t_n}$$

3. Distribution of $\frac{(n-1)(\sum_{i=1}^n (X_i-\bar{X})^2+X_{n+1}^2)}{n\sum_{i=1}^n (X_i-\bar{X})^2}$

- $\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2$
- $X_{n+1}^2 \sim \chi_1^2$

The expression becomes:

$$\frac{(n-1)\left(\sum_{i=1}^n (X_i-\bar{X})^2 + X_{n+1}^2\right)}{n\sum_{i=1}^n (X_i-\bar{X})^2} = \frac{n-1}{n}\left(\frac{\chi_n^2}{\chi_{n-1}^2}\right)$$

From the third property of the F-distribution

$$\frac{\chi_n^2/n}{\chi_{n-1}^2/(n-1)} \sim F_{n,n-1}$$

The distribution of the given expression is:

$$\boxed{\frac{(n-1)(\sum_{i=1}^n (X_i-\bar{X})^2 + X_{n+1}^2)}{n\sum_{i=1}^n (X_i-\bar{X})^2} \sim F_{n,n-1}}$$

Exercise 4

1. Min and Max applications accepted

```
# R code
min(College$Accept)  # Output: 72       Minimum
max(College$Accept)  # Output: 26330    Maximum
```

2. Mean and Standard Deviation

```
# R code
mean(College$Accept)  # Output: 2018.804    Mean
sd(College$Accept)    # Output: 2451.114    Standard Deviation
```

3. Histogram of Applications Accepted

```
# R code
hist(College$Accept, breaks=17,
     main="Applications Accepted",
     xlab="Num Applications Accepted")
```

Most of the colleges accept less than 5000 applications making the the graph right-skewed.

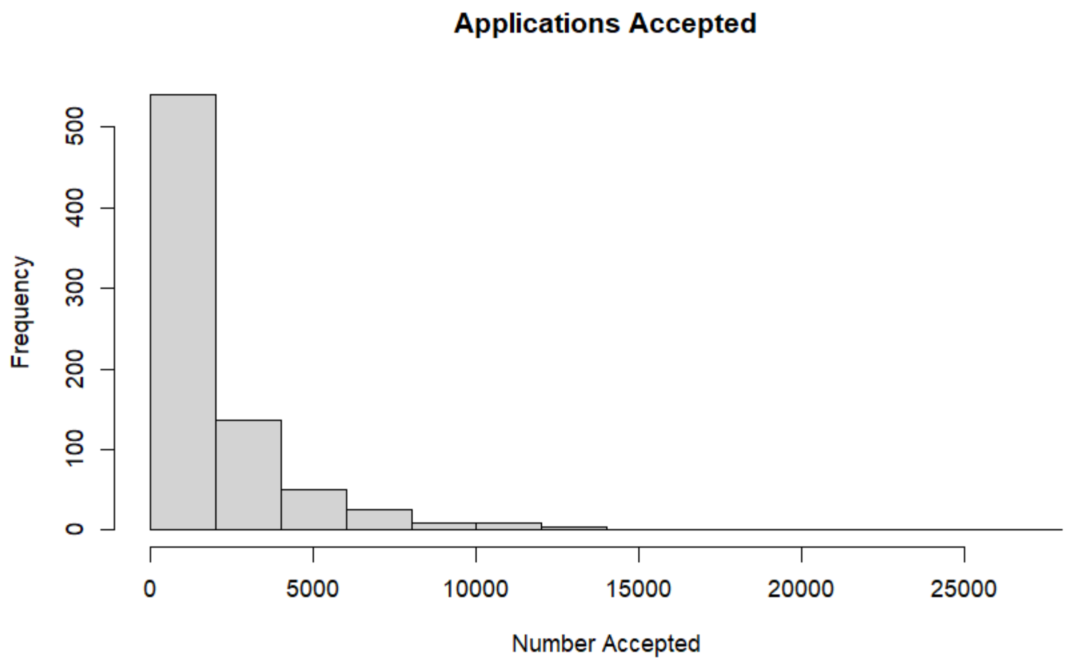

Figure 1: Histogram of the Number of Applications Accepted

4. Exponential Model Probability

$$P(X > 2000) = exp(-2000/\beta) = exp(-2000/2018.804) \approx 0.371$$

```
# R code
exp(-2000/2018.804)  # Output: 0.371322
```

The probability that a College accepts more than 2000 applications is $\boxed{37.1\%}$.