FIT5197 Assignment 2 Semester 1, 2020

Do & Don't

1. Some of these policies are similar to your assignment 1; however, please read carefully as there are some differences between this assignment and the first assignment.
2. These questions are meant for you to solve independently, we encourage students to figure out the questions themselves as it would be good for their understandings of the topics; however, please feel free to consult your tutors if needed. Plagiarism (either from using online sources or copying the answers from your classmates) will be penalised accordingly.
3. Requests for special consideration or extension must be submitted at least 2 days BEFORE THE DEADLINE. The due date is on Sunday, so the latest day you can ask for extension is on Friday (the last official working day of the week for the teaching team). Please follow Monash guidelines to request for extensions (medical certificates, doctor or GP letter, etc). Emergencies are to be adjusted individually.
4. Please show all working when answering questions, you will not get full marks for a question if you don't comply.
5. Assignments need to be submitted in PDF and ipynb file format. Failure to comply will result in 20% penalty on each missing file.
6. Filename format for submitting the assignment should be “Assignment2_StudentId.pdf” and "Assignment2_StudentId.ipynb". Files with the wrong format incurs 20% penalty each.
7. This assignment has 10 marks for presentation, this includes presenting your explanation in **Markdown**, writing and commenting on code efficiently, creating good plots with clear labels on the axis, etc.
8. Only answers with correct methodology will be considered for consequential marks. Meaning if you attempt the question and your answer is incorrect, but your methodology is correct, you will still receive partial marks for subsequent questions. However, answers with incorrect methodology (misunderstanding the questions) will generate no marks for subsequent questions.
9. Challenge questions are for students aiming to get a HD for the assignment. We don't advise for students to spend time on these questions before finishing all the other parts in the assignment. This assignment is designed in a way that students can get up to 80 (HD) without attempting the challenge questions.
10. Please don’t send emails to tutors asking for suggestions, we have Moodle and consultations for that. When writing your inquiries on Moodle please try to be clear in your problem and not revealing your working to others as this might be counted as plagiarism on your part. A good format for inquiry topic would be e.g. “Assignment 2 – Tutorial 10 (your tutorial slot) – Question about median"
11. Handwritten answers incur a penalty of 10% on your assignment, you have Markdown, please learn how to use it as it will be an useful skill for you going through the degree as well as in real life situation.
12. This assignment will contribute towards 20% of your total score.
13. Late submission is 5% per day, after 10 days you will be given no marks. Late submission is calculated as follows: If you get 70% on this assignment and you are late for 2 days (you submit on Tuesday), your score is now 70% -10% (2x5% per day) = 60%. This is done to ensure that the teaching team can release your result as soon as possible so that you can review on your mistakes and have a better study experience.
14. Assignments shall be marked completely in two weeks’ time according to Monash Policies. If there are any changes to the marking time, we will duly inform you. Solutions will not be released for this assignment; you can come to the tutorial and ask for explanation about how to solve the questions after scores are released.

Question 1 - Probabilities (10 Marks)

A box contains n pairs of shoes ($2n$ shoes in total). If $2r$ (with the assumption that $2r \leq n$) shoes are selected at random, find the probability for the following scenarios:

Question 1a. (2 Marks)

A_0 = 'No matching pair'

ANSWER:

Number of ways of selecting 2r shoes from 2n shoes:

$$\text{Total probability} = \binom{2n}{2r}$$

Number of ways of selecting r pairs i.e 2r shoes from n pairs: =

$$\binom{n}{2r}$$

Number of ways of selecting separate shoes from these r pairs i.e 2r shoes =

$$2^{2r}$$

$$\text{P(No matching pair)} = \frac{\binom{n}{2r}2^{2r}}{\binom{2n}{2r}}$$

Question 1b. (3 Marks)

A_1 = 'only one matching pair'

ANSWER:

Number of ways of selecting 2r shoes from 2n shoes:

$$\text{Total probability} = \binom{2n}{2r}$$

Number of ways of selecting 1 pair from n pairs: =

$$\binom{n}{1}$$

Now,
We need to select **(r - 1)** pairs from **(n - 1)** pairs. Thus,
Number of ways of selecting (r - 1) pairs from **(n - 1) pairs =

$$\binom{n-1}{2r-2}$$

Now,
Number of ways of selecting separate shoes from these (r -1) pairs i.e (2r - 2) shoes =

$$2^{2r-2}$$

$$\text{P(Only one matching pair)} = \frac{n\binom{n-1}{2r-2}2^{2r-2}}{\binom{2n}{2r}}$$

Question 1c. (2 Marks)

A_2 = ‘exactly two matching pairs’

ANSWER:

Number of ways of selecting 2r shoes from 2n shoes:

$$\text{Total probability} = \binom{2n}{2r}$$

Number of ways of selecting 2 pairs from n pairs: =

$$\binom{n}{2}$$

Now,
We need to select **(r - 2)** pairs from **(n - 2)** pairs. Thus,
Number of ways of selecting (2r - 4) pairs from (n - 2) pairs =

$$\binom{n-2}{2r-4}$$

Now,
Number of ways of selecting separate shoes from these (r -2) pairs i.e 2r - 4 =

$$2^{2r-4}$$

$$\text{P(Exactly two pairs)} = \frac{\binom{n}{2}\binom{n-2}{2r-4}2^{2r-4}}{\binom{2n}{2r}}$$

Question 1d. (3 Marks)

A_r = ‘exactly r matching pairs’

ANSWER:

Number of ways of selecting 2r shoes from 2n shoes:

$$\text{Total probability} = \binom{2n}{2r}$$

Number of ways of selecting 0 pairs from the remaining (n - r) pairs: =

$$\binom{n-r}{0}$$

Number of ways of selecting r pairs from n pairs: =

$$\binom{n}{r}$$

$$\text{P(Exactly r pairs)} = \frac{\binom{n}{r}(n-r)}{\binom{2n}{2r}}$$

Question 2 - Conditional Probabilities & Entropy (30 Marks)

Warning, no built in functions to calculate probability or entropy from R should be used for this part. The only help you can get from R should be dataframe manipulation. Answers using functions will not be marked even if the answer is correct.

Sports analytics (i.e., the application of data science techniques to competitive sports) is a rapidly growing area of data science. In this question we will look at some very basic analytics applied to the outcomes of consecutive games of English Premier League (EPL). The file chelsea.csv contains a record of the outcomes of games of EPL played by **Chelsea football club (CFC)** in the seasons from 1993 to 2018. The data is sequential, in the sense that each row recorded the result whether the home team wins (H), the away team wins (A), or there is a draw (D).

Please show all working including code and presentation for this question

Part 1: Analyzing Home/Away performance (17 Marks)

Question 2.a (3 Marks)

Find out the probabilities **P(Chelsea Wins), P(Chelsea Loses), and P(Chelsea Draws)**. This includes all the results both home and away.

In [598]:

```
# Reading the dataset into data
data <- read.csv("chelsea.csv")

# Calculating the total number of games
total_number_of_games <- nrow(data)

# Number of games won
chelsea_wins <- subset( data, (home == "Chelsea" & result == "H") |
                              (away == "Chelsea" & result == 'A') )
chelsea_wins_games <- nrow(chelsea_wins)
chelsea_wins_probability <- chelsea_wins_games/total_number_of_games
round( chelsea_wins_probability, digits = 4)

# Number of games lost
chelsea_lost <- subset( data, (home == "Chelsea" & result == "A") |
                              (away == "Chelsea" & result == 'H') )
chelsea_lost_games <- nrow(chelsea_lost)
chelsea_lost_probability <- chelsea_lost_games/total_number_of_games
round( chelsea_lost_probability, digits = 4)

# Number of games in draw
chelsea_draw <- subset( data, (home == "Chelsea" & result == "D") |
                              (away == "Chelsea" & result == 'D') )
chelsea_draw_games <- nrow(chelsea_draw)
chelsea_draw_probability <- chelsea_draw_games/total_number_of_games
round( chelsea_draw_probability, digits = 4)
```

```
0.5459

0.2098

0.2443
```

1. **P(Chelsea Wins)** = chelsea_wins_games/total_number_of_games = 0.5459

2. **P(Chelsea Loses)** = chelsea_lost_games/total_number_of_games = 0.2098
3. **P(Chelsea Draws)** = chelsea_draw_games/total_number_of_games = 0.2443

Question 2.b (6 Marks)

Find out the conditional probabilities:

1. **P(Chelsea Wins| Playing at Home)**
2. **P(Chelsea Wins| Playing away)**
3. **P(Chelsea Draws| Playing at Home)**
4. **P(Chelsea Draws| Playing away)**
5. **P(Chelsea Loses| Playing at Home)**
6. **P(Chelsea Loses| Playing away)**

Please make comparison and a general conclusion.

In [599]:

```
# Reading the dataset into data
data <- read.csv("chelsea.csv")

# Calculating the total number of games
total_number_of_games <- nrow(data)

# Games at home
home <- subset( data, (home == "Chelsea"))
total_games_home <- nrow(home)

# Games away
away <- subset( data, (away == "Chelsea"))
total_games_away <- nrow(away)

# P(Chelsea Wins| Playing at Home)
wins_playing_at_home <- subset( data, (home == "Chelsea" & result == "H") )
wins_playing_at_home <- nrow(wins_playing_at_home)
wins_playing_at_home_probability <- wins_playing_at_home/total_games_home
round( wins_playing_at_home_probability, digits = 4)

# P(Chelsea Wins| Playing Away)
wins_playing_away <- subset( data, (away == "Chelsea" & result == "A") )
wins_playing_away <- nrow(wins_playing_away)
wins_playing_away_probability <- wins_playing_away/total_games_away
round( wins_playing_away_probability, digits = 4)

# P(Chelsea Draws| Playing at Home)
draws_playing_at_home <- subset( data, (home == "Chelsea" & result == "D") )
draws_playing_at_home <- nrow(draws_playing_at_home)
draws_playing_at_home_probability <- draws_playing_at_home/total_games_home
round( draws_playing_at_home_probability, digits = 4)

# P(Chelsea Draws| Playing away)
draws_playing_away <- subset( data, (away == "Chelsea" & result == "D") )
draws_playing_away <- nrow(draws_playing_away)
draws_playing_away_probability <- draws_playing_away/total_games_away
round( draws_playing_away_probability, digits = 4)

# P(Chelsea Loses| Playing at Home)
loses_playing_at_home <- subset( data, (home == "Chelsea" & result == "A") )
loses_playing_at_home <- nrow(loses_playing_at_home)
loses_playing_at_home_probability <- loses_playing_at_home/total_games_home
round( loses_playing_at_home_probability, digits = 4)

# P(Chelsea Loses| Playing away)
loses_playing_away <- subset( data, (away == "Chelsea" & result == "H") )
loses_playing_away <- nrow(loses_playing_away)
loses_playing_away_probability <- loses_playing_away/total_games_away
round( loses_playing_away_probability, digits = 4)
```

```
0.6347

0.4572

0.238

0.2505

0.1273
```

```
0.2923
```

1. **P(Chelsea Wins| Playing at Home)** = 0.6347
2. **P(Chelsea Wins| Playing away)** = 0.4572
3. **P(Chelsea Draws| Playing at Home)** = 0.238
4. **P(Chelsea Draws| Playing away)** = 0.2505
5. **P(Chelsea Loses| Playing at Home)** = 0.1273
6. **P(Chelsea Loses| Playing away)** = 0.2923

Comparison : We can observe the following from the data :

```
* Chelsea wins more number of matches while playing at home than playing away.
* There is only a slightly higher probability of the match ending in a draw, when
 Chelsea is playing away, then at home.
* Similarly, the probability of Chelsea losing is higher when playing away, than p
laying at home.
```

General Conclusion :

```
* Thus, we can conclude that for Chelsea phas it's best performance when the team
 is playing at home as it has it's familiar with the ground and turf and has it's
 crowd for support.
```

Question 2.c (3 Marks)

Find $H(\text{Chelsea Results})$ this includes results in both home and away games

In [600]:

```python
#log(value, base)
# We have the probability values from the above calculations
pw = chelsea_wins_probability
pl = chelsea_lost_probability
pd = chelsea_draw_probability

p1 = pw*(log2(1/pw))
p2 = pl*(log2(1/pl))
p3 = pd*(log2(1/pd))

Chelsea_Results_Entropy = round( p1 + p2 + p3, digits = 4)
Chelsea_Results_Entropy
```

```
1.4461
```

$H(\textbf{Chelsea Results})$ = 1.4461

Question 2.d (5 Marks)

Is knowing whether Chelsea plays at home or away a good indicator in knowing the result of CFC games? Show your justification (answering just yes or no will not be given any marks) by calculating all the information necessary using the knowledge you have learnt so far in the unit.

Note that this is a challenge question, only attempt if you are comfortable with your progress. You should not use consultations to ask for this question as the tutors will proritize answering queries about other questions.

In [601]:

```
#Calculations which will used in the answer:

chelsea_home <- subset( data, (home == "Chelsea") )
chelsea_home_probability <- ( nrow( chelsea_home ))/total_number_of_games
cat("\n Probability of Chelsea playing at home:",chelsea_home_probability )

chelsea_away <- subset( data, (away == "Chelsea") )
chelsea_away_probability <- ( nrow( chelsea_away ))/total_number_of_games
cat("\n Probability of Chelsea playing at home:",chelsea_away_probability )
```

```
 Probability of Chelsea playing at home: 0.5
 Probability of Chelsea playing at home: 0.5
```

ANSWER:

Yes, knowing whether Chelsea plays at home or away a good indicator in knowing the result of CFC games.

Step 1:

- Clearly, we know from **2a** that the Chelsea has **higher chances of winning**, than losing or the match resulting in a draw.
- Similarly, from 2b, we know that when Chelsea, plays at home, the probability of winning is higher than the game resulting in a draw or a loss.

Step 2:
Now, we calculate the following:

- Probability of playing at home: P(H)

$$P(Home) = 0.5$$

Step 3:
Now, we can calculate the following:

- The probability that Chelsea played at home, given it won the game.

$$P(Home/Wins) = \frac{P(Home) \cdot (Wins/Home)}{P(Wins)}$$
$$= 0.5812$$

- The probability that Chelsea played at home, given it lost the game.

$$P(Home/Loses) = \frac{P(Home) \cdot (Loses/Home)}{P(Lost)}$$
$$= 0.3033$$

- The probability that Chelsea played at home, given the match resulted in a draw.

$$P(Home/Draws) = \frac{P(Home) \cdot (Draws/Home)}{P(Lost)}$$
$$= 0.5129$$

Conclusions:

- Thus, from we can conclude that:

- Given, Chelsea has won a game, the probability that Chelsea played the match at **home** is quite high.

Part 2: Analyzing effects of previous results on future results (13 Marks)

This is a new part of the question, this part will focus on a different aspects compared to part I. Objectively speaking, this part is harder compared to the previous one; thus, it has lower mark allocation, students are advised to spend time on this part if they want to achieve desirable outcome.

Based on the data given to you, please create another column named "binary". This column will record a win $(\text{corresponding to } 1)$ or a loss/draw $(\text{corresponding to } 0)$ in the order in which the games were played by CFC.

A simple question regarding this type of data might be regarding the existence of (de)motivativing effects on a team if they have won / not won their previous game. Let W_t denote the binary outcome of a game in round t and W_{t-1} denote the outcome of the game played in the previous round. Answer the following questions; **you must provide working/justification.**

In [602]:

```
# Adding a new column named binary to the dataframe

data <- read.csv("chelsea.csv")

data$binary <- ifelse( ( data$home == "Chelsea" & data$result == 'H') |
                       ( data$away == "Chelsea" & data$result == 'A'), 1,0 )
```

Question 2.e (4 Marks)

Using the data in **chelsea.csv** and the new column you just created for this task, write R code to **find the frequency** with which CFC won / did not win a game after it won / did not win its previous game. Using these frequencies, calculate the joint distributions $P(W_t = 0, W_{t-1} = 0)$, $P(W_t = 1, W_{t-1} = 0)$, $P(W_t = 1, W_{t-1} = 1)$, and $P(W_t = 0, W_{t-1} = 1)$. We suggest students create another column from the original dataframe to solve this question. Please read this question carefully before attempting.

In [603]:

```
# Creating counters
count_LL <- 0
count_LW <- 0
count_WW <- 0
count_WL <- 0

for (i in 2:total_number_of_games)
{
    if ( data$binary[i] == 0 & data$binary[i-1] == 0)
    {
     data$current_previous[i] <- "LL"
     count_LL <- count_LL + 1
    }
    else if ( data$binary[i] == 0 & data$binary[i-1] == 1 )
    {
      data$current_previous[i] <- "LW"
      count_LW <- count_LW + 1
    }
    else if ( data$binary[i] == 1 & data$binary[i-1] == 1 )
    {
      data$current_previous[i] <- "WW"
      count_WW <- count_WW + 1
    }
    else if (data$binary[i] == 1 & data$binary[i-1] == 0 )
    {
      data$current_previous[i] <- "WL"
      count_WL <- count_WL + 1
    }
}
```

In [604]:

```
# Frequencies
cat("\n (Wt = 0, Wt-1 = 0) :",count_LL)
cat("\n (Wt = 0, Wt-1 = 1) :",count_LW)
cat("\n (Wt = 1, Wt-1 = 1) :",count_WW)
cat("\n (Wt = 1, Wt-1 = 0) :",count_WL)
```

```
 (Wt = 0, Wt-1 = 0) : 198
 (Wt = 0, Wt-1 = 1) : 236
 (Wt = 1, Wt-1 = 1) : 287
 (Wt = 1, Wt-1 = 0) : 236
```

The different **frequencies** of the Joint Distributions are as follows:

$P(W_t = 0, W_{t-1} = 0)$ = 198
$P(W_t = 1, W_{t-1} = 0)$ = 236
$P(W_t = 1, W_{t-1} = 1)$ = 287
$P(W_t = 0, W_{t-1} = 1)$ = 236

In [605]:

```
# Calculating the probabilities
LL <- subset( data, (current_previous == "LL") )
WL <- subset( data, (current_previous == "WL") )
WW <- subset( data, (current_previous == "WW") )
LW <- subset( data, (current_previous == "LW") )

w00 <-  round ( ( nrow( LL ) ) / total_number_of_games, digits = 4 )
w10 <-  round ( ( nrow( WL ) ) / total_number_of_games, digits = 4 )
w11 <-  round ( ( nrow( WW ) ) / total_number_of_games, digits = 4 )
w01 <-  round ( ( nrow( LW ) ) / total_number_of_games, digits = 4 )

# Displaying the answers
w00
w10
w11
w01
```

```
0.2067

0.2463

0.2996

0.2463
```

The different **probabilities** of the Joint Distributions are as follows:

$P(W_t = 0, W_{t-1} = 0)$ = 0.2067
$P(W_t = 1, W_{t-1} = 0)$ = 0.2463
$P(W_t = 1, W_{t-1} = 1)$ = 0.2996
$P(W_t = 0, W_{t-1} = 1)$ = 0.2463

Question 2.f (2 Marks)

What is the probability that CFC will win a game given that they won their previous game?

In [606]:

```
n_wins <- nrow( subset( data, (current_previous == "WL" | current_previous == "WW")) )

n_WW <- nrow( subset( data, (current_previous == "WW")) )

n_win_won <- round( n_WW/n_wins, digits = 4 )

n_win_won
```

```
0.5488
```

Thus, the probability that CFC will win a game given that they won their previous game is:

$$\begin{aligned} Required Probability &= \frac{P(W/W)}{P(W/W) + P(W/L)} \\ &= 0.5488 \end{aligned}$$

Question 2.g (2 Marks)

What is the probability that CFC will win a game given that they didn't win their previous game?

In [607]:

```
n_losses <- nrow( subset( data, (current_previous == "LL" | current_previous == "LW")) )

n_WL <- nrow( subset( data, (current_previous == "WL")) )

n_win_loss <- round( n_WL/n_losses, digits = 4 )

n_win_loss
```

```
0.5438
```

Thus, probability that CFC will win a game given that they didn't win their previous game:

$$\begin{aligned} Required Probability &= \frac{P(W/L)}{P(L/W) + P(L/L)} \\ &= 0.5438 \end{aligned}$$

Question 2.h (2 Marks)

Do you think winning/not winning the previous game had an effect on the CFC players in their next game? Justify your answer? **(Note that this is different compared to 2.d)**

Yes, **winning/not winning** the previous game had an effect on the CFC players in their next game.

- Clearly, we can see from the above numbers that -
  - the probability of CFC winning, given they have won their previous game is **higher**.
- The contributing factors could be -
  - High Energy, morale after the previous win!
  - Greater confidence levels to keep th momemtum of winning high.

Question 2.i (3 Marks)

Calculate the probability of CFC not winning their next two games given that they won their previous game.

In [608]:

```
count <- 0

for( i in 2:total_number_of_games)
{
        if( data$binary[i - 1] == 1 & data$binary[i] == 0 & data$binary[i + 1] == 0 )
        {
                count <- count + 1
        }
}

answer <- round( count/total_number_of_games, digits = 4)
answer
```

```
0.1086
```

Thus, the probability of CFC not winning their next two games given that they won their previous game

$$\begin{aligned} RequiredProbability &= \frac{NumberOfBinaryOccurrencesof(1,0,0)}{TotalNumberOfGamesPlayed} \\ &= 0.1086 \end{aligned}$$

Question 3 - Expectation - Challenge Question (10 Marks)

Randomly place a point P inside triangle ABC. Let X be the continuous random variable representing the distance from point P to AB. Find out $E(X)$ & $Var(X)$ (**5 Marks each**)

Note that this is a challenge question, only attempt if you are comfortable with your progress. You should not use consultations to ask about this question as the tutors will prioritize answering queries about other questions.

ANSWER:

Calculating the pdf of x:

- Let us draw a line from the vertex C to base AB.
- Since Point P can lie anywhere inside the triangle, let us assume that P lies on this line.
- Let x be the the distance of the point P from the base AB.
- Let h be the height of the triangle.
- Hence, from the question,
    - we know that X is the continuous random variable representing the distance of this point P to the base AB.
- Now,

The pdf can be calculated as:

$$\begin{aligned} f(x) &= \frac{d}{dx} \cdot P[X <= x] \\ &= \frac{d}{dx} \cdot (\frac{h-x}{h})^2 \\ &= \frac{2(h-x)}{h^2} \end{aligned}$$

Hence, The pdf of X can be given as:

$$f(x) = \begin{cases} \frac{2(h-x)}{h^2} \text{ if } 0 \leq x \leq h \\ 0 \text{ otherwise} \end{cases}$$

where h = Distance from vertex C to AB

Now, Mean of X can be given as:

$$\begin{aligned} E[X] &= \int_0^h x \cdot f(x) \cdot dx \\ &= \int_0^h \frac{2x(h-x)}{h^2} \cdot dx \\ &= \frac{h}{3} \end{aligned}$$

Now,

$$\begin{aligned} E[X^2] &= \int_0^h x^2 \cdot f(x) \cdot dx \\ &= \int_0^h \frac{2x^2(h-x)}{h^2} \cdot dx \\ &= \frac{h^2}{6} \end{aligned}$$

Hence, Variance of x can be given as:

$$\begin{aligned} V[X] &= E[X^2] - E[X] \\ &= \frac{h^2}{18} \end{aligned}$$

Question 4 - Distribution (10 Marks)

The teaching team of FIT5197 is required to prepare 4 questions each week for the next week's tutorial. The number of questions created in a week is said to have a Poisson distribution with mean 6.

Question 4.a (2 Marks)

Find the probability that the teaching team manages to write enough questions for the following week?

ANSWER:

Here, it is given that the teaching team manages to write enough questions.
Hence, $x = 4$ and λ = 6

$$\begin{aligned} P(x = 4|\lambda) &= \frac{\lambda^4 \exp(-\lambda)}{4!} \\ &= \frac{6^4 \exp(-6)}{4!} \\ &= 0.1339 \end{aligned}$$

In [609]:

```
# Checking with the r function
round( dpois(4,6), digits = 4)
```

```
0.1339
```

Question 4.b (4 Marks)

Since some of the tutors in the teaching team are also responsible for other units from FIT, for each week, there is a probability of 40% that only half of the team will work on the questions. If that is the case, the teaching team can only create 3 questions on average. If the teaching team fails to finish 4 questions one week, what is the probability that only half of the team works that week?

YOUR ANSWER HERE

In [610]:

```
# Probability of failing, when only half the staff works.
print ( dpois(0,3) + dpois(1,3) + dpois(2,3) + dpois(3,3))

# Probability of failing, when the entire the staff works.
print ( dpois(0,6) + dpois(1,6) + dpois(2,6) + dpois(3,6))
```

```
[1] 0.6472319
[1] 0.1512039
```

ANSWER:

P(Full) :- Probability of full staff working = 0.6
P(Half) :- Probability of half staff working = 0.4
Now,
Here : **Failing** means making less than 4 questions in one week.

- When half staff works, lambda = 3
  - Probability of half staff failing = 0.6472
  - Which Implies, Probability of half staff not failing i.e passing = 1 - 0.6472 = 0.3528
- When full staff works, lambda = 6
  - Probability of full staff failing = 0.1512
  - Which Implies, Probability of full staff not failing i.e passing = 1 - 0.1512 = 0.8488

Here, it is given that the teaching team fails.
Hence, the required probability can be calculated as:

$$\begin{aligned} P(Half/Fail) &= \frac{P(Half)\cdot P(Fail/Half)}{P(Half)\cdot P(Fail/Half) + P(Full)\cdot P(Fail/Full)} \\ &= \frac{(0.4)(0.6472)}{(0.4)(0.6472) + (0.6)(0.1512)} \\ &= 0.7405 \end{aligned}$$

Question 4.c (4 Marks)

On week 12, the teaching team decides to no longer limit to 4 questions, and instead use every question they create. If a student has a 40% chance of correctly answering questions, and this student is expected to answer 2 questions correctly in the coming tutorial, what is the probability that the whole teaching team worked on creating the questions that week?

ANSWER:

P(Full) :- Probability of full staff working = 0.6
P(Half) :- Probability of half staff working = 0.4
Now,
Here, let the number of questions generated by the teaching team be N.
And, student has a 40% chance of correctly answering questions.
Also, student is expected to answer 2 questions correctly.
Hence,

$$\begin{aligned} &\Rightarrow 0.4N = 2 \\ &\Rightarrow \ N \ = 5 \end{aligned}$$

Hence, we now know that the number of questions generated are 5.
Which, implies the teaching team has passed.

Probability of full staff failing = 0.1512
Hence, Probability of full staff **passing** =1 - 0.1512 = 0.8488
Probability of half staff failing = 0.6472
Hence, Probability of half staff **passing** =1 - 0.6472 = 0.3528

Here, we have found out that the teaching team has passed.
Hence, the required probability can be calculated as:

$$\begin{aligned} P(Full/Pass) &= \frac{P(Full) \cdot P(Pass/Full)}{P(Full) \cdot P(Pass/Full) + P(Half) \cdot P(Pass/Half)} \\ &= \frac{(0.6)(0.8488)}{(0.4)(0.3528) + (0.6)(0.8488)} \\ &= 0.7830 \end{aligned}$$

Question 5 - Maximum Likelihood Estimation (15 Marks)

The exponential distribution is a probability distribution for non-negative real numbers. It is often used to model waiting or survival times. The version that we will look at has a probability density function of the form

$p(y|v) = exp(-e^{-v} y - v)$

where $y \in R_+$, i.e., y can take on the values of non-negative real numbers. In this form it has one parameter: a log-scale parameter v. If a random variable follows an `exponential distribution` with log-scale v we say that $Y \sim Exp(v)$. If $Y \sim Exp(v)$, then $E[Y] = e^v$ and $V[Y] = e^{2v}$.

Question 5.a (4 Marks)

Imagine we are given a sample of n observations $y = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from an exponential distribution with log-scale parameter v (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working.

ANSWER:

The likelihood is equal to the product of the individual probabilities.

Step 1: Multiplying the individual probabilities.
Step 2: Taking out the common values outside.
Step 3: Simplifying the expression.

Thus, we have

$$\begin{aligned} p(\mathbf{y}|v) &= exp(-e^{-v}y1 - v) \cdot exp(-e^{-v}y2 - v) \cdots \exp(-e^{-v}yn - v) \\ &= xe^{-xy1} \cdot xe^{-xy2} \cdots xe^{-xyn} \\ &= x^n e^{-xy1} \cdot e^{-xy2} \cdots e^{-xyn} \\ &= x^n e^{-xy1-xy2\cdots-xyn} \\ &= x^n e^{-x(y1+y2\cdots+yn)} \\ &= x^n e^{-xm} \end{aligned}$$

where $m = \sum_{i=1}^{n} y_i$
where $x = e^{-v}$

Question 5.b (2 Marks)

Take the negative logarithm of your likelihood expression and write down the negative loglikelihood of the data y under the exponential model with log-scale v. Simplify this expression

ANSWER:

The negative loglikelihood of the data y under the exponential model with log-scale v is given as follows:

Step 1: Take the negative logarithm of the likelihood
Step 2: Taking out the common values outside.
Step 3: Substitute the value of $x = e^{-v}$

Thus, we have

$$\begin{aligned} -ln(p(\mathbf{y}|v)) &= -ln(x^n e^{-xm}) \\ &= -nln(x) - nln(e^{-xm}) \\ &= -nln(x) + xm \\ &= -nln(e^{-v}) + e^{-v}m \\ &= nv + e^{-v}m \end{aligned}$$

where $m = \sum_{i=1}^{n} y_i$
where $x = e^{-v}$

Question 5.c (4 Marks)

Derive the maximum likelihood estimator $\hat{v}$ for v. That is, find the value of v that minimises the negative log-likelihood. You must provide working.

ANSWER:

The maximum likelihood estimator $\hat{v}$ for v for a given data sample $\mathbf{y}$ is given as follows:

Step 1: Differentiate the negative log-likelihood function with respect to υ.
Step 2: Set the derivative to zero.
Step 3: Solve for υ.

Begin by differentiate the negative log-likelihood with respect to our parameter of interest υ:

$$\begin{aligned}\frac{d\,\ell(\mathbf{v};\upsilon)}{d\upsilon} &= \frac{d\,(n\upsilon)}{d\upsilon} + \frac{d\,(me^{-\upsilon})}{d\upsilon}\\ &= n + m(-e^{-\upsilon})\\ &= n - me^{-\upsilon}\end{aligned}$$

Then we need to set this to zero and solve for υ:

$$\begin{aligned} n - me^{-\upsilon} &= 0\\ \Rightarrow \quad n \quad &= me^{-\upsilon}\\ \Rightarrow \quad e^{-\upsilon} \quad &= \frac{n}{m}\\ \Rightarrow \quad \upsilon \quad &= ln(\frac{m}{n})\end{aligned}$$

$$\hat{\upsilon}_{\text{ML}}(\mathbf{y}) = ln(\frac{m}{n})$$

where $m = \sum_{i=1}^{n} y_i$

Question 5.d (5 Marks)

Determine the approximate bias and variance of the maximum likelihood estimator $\hat{\upsilon}$ of υ for the exponential distribution.

Note that this is a challenge question, only attempt if you are comfortable with your progress. You should not use consultations to ask about this question as the tutors will proritize answering queries about other questions.

ANSWER:

Here, $Y \sim Exp(\upsilon)$,then
Mean: $E[Y] = e^{\upsilon}$
Variance: $V[Y] = e^{2\upsilon}$.

We can **approximate** the maximum likelihood estimator to be the **probability density function**
The maximum likelihood estimator helps us in determining the parameters that best help us describe the data
This data is also given by the probability density function which provides the likelihood of the variable in the data

Step 1: Calculating the Mean and Bias
Here, it is given that y follows and exponential distribution.
And, y takes values from y_1 to y_n.
Hence, the Mean is given as follows:

$$\begin{aligned}E[\hat{\upsilon}_{\text{ML}}(\mathbf{y})] &= \int_0^{\upsilon} x \cdot f(x) \cdot dx\\ &= \int_0^{\upsilon} x \cdot ln(\frac{m}{n}) \cdot dx\\ &= \frac{\upsilon^2}{2} \cdot ln(\frac{m}{n})\end{aligned}$$

Hence, Bias can be given as:

$$\begin{aligned} b[\hat{v}_{\text{ML}}(\mathbf{y})] &= E[\hat{v}_{\text{ML}}(\mathbf{y})] - v \\ &= \frac{v^2}{2} \cdot ln(\frac{m}{n}) - v \end{aligned}$$

Step 2: Calculating the Varince

- The variance of the maximum likelihood estimator can be given as:

$$\begin{aligned} V[\hat{v}_{\text{ML}}(\mathbf{y})] &= E[(\hat{v}_{\text{ML}}(\mathbf{y}) - E[\hat{v}_{\text{ML}}(\mathbf{y})])^2] \\ &= E[ln(\frac{m}{n}) - \frac{v^2}{2} \cdot ln(\frac{m}{n}))^2] \\ &= E[t - \frac{v^2}{2} \cdot t))^2] \\ &= E[\frac{t^2(4 + v^4 - 4v^2)}{4}] \\ &= \int_0^v v \cdot \frac{t^2(4 + v^4 - 4v^2)}{4} \cdot dx \\ &= \frac{v^2 t^2(12 + v^4 - 6v^2)}{24} \end{aligned}$$

where $m = \sum_{i=1}^{n} y_i$
where $t = ln(\frac{m}{n})$

Question 6 - Central Limit Theorem (15 Marks)

Sampling Process: Assume that we randomly select samples of the same size n an infinite number of times from a population that follows a Poisson distribution with mean of λ, and then, we calculate the mean of scores in each sample.

Question 6.a (2 Marks)

What does Central Limit Theorem tell us about the sampling distribution of the sample mean?

ANSWER:

The **Central Limit Theorem** states that if you have a population with **mean μ** and **standard deviation σ** and take sufficiently large random samples from the population with replacement , then the **distribution** of the **sample means** will be approximately **normally distributed as follows** for large values of n.

$$= N\left[\mu, \frac{1}{n}\sigma^2\right]$$

In Poisson Distribution, we know:

- **Mean =** λ
- **Variance =** λ

Pois(λ**)** approaches **N(**λ**,**λ**)** for large λ

Question 6.b (3 Marks)

For three different Poisson populations with mean of λ_1 = 1, λ_2 = 5 and λ_3 = 20, we will do the sampling four separate times -- for small samples (n=10), for samples of 100 subjects (n=100) and 1000 subjects (n=1000), and once for big samples (n=10000).

Based on your answer from 6.a, compute the parameter values for each sampling distribution in R.

In [623]:

```
# Function to calculate mean and standard deviation
calculate_mean_sd <- function(n, lambda)
{
        # Displaying the n and lambda values
        cat("\n n =",n)
        cat("\n lambda =",lambda)

        cat("\n")
        # Storing the mean and sd values in a list names result
        result <- list()
        result$mean_value <- mean(rpois(n,lambda))
        result$sd_value <- sd(rpois(n,lambda))

        #return (result)
        cat("\n Mean =",result$mean_value)
        cat("\n Standard Deviation =",result$sd_value)
        cat("\n --------------------------------\n ")

}
```

In [624]:

```
# Calling the function to calculate the mean and sd values

for (lambda in c(1,5,20))
 for (n in c(10,100,1000))
        calculate_mean_sd(n,lambda)
```

```
n = 10
lambda = 1

Mean = 1
Standard Deviation = 0.5163978
--------------------------------

n = 100
lambda = 1

Mean = 1.09
Standard Deviation = 0.8348471
--------------------------------

n = 1000
lambda = 1

Mean = 1.036
Standard Deviation = 0.9763982
--------------------------------

n = 10
lambda = 5

Mean = 5
Standard Deviation = 3.198958
--------------------------------

n = 100
lambda = 5

Mean = 5.14
Standard Deviation = 2.38336
--------------------------------

n = 1000
lambda = 5

Mean = 4.994
Standard Deviation = 2.258932
--------------------------------

n = 10
lambda = 20

Mean = 19.4
Standard Deviation = 3.683296
--------------------------------

n = 100
lambda = 20
```

```
Mean = 19.91
Standard Deviation = 4.398932
--------------------------------

n = 1000
lambda = 20

Mean = 19.925
Standard Deviation = 4.404919
--------------------------------
```

Question 6.c (5 Marks)

In this question, you are asked to experimentally justify the result in the CLT Theorem.

For different sample sizes of n = 10, 100 and 1000, use 50000 simulations (i.e. to approximate the infinite times we drew samples as mentioned before) to implement the sampling process.

From those 50000 sample means, compute the mean and standard deviation parameters (3 sample sizes and 3 λ rates, 9 pairs of parameters in total).

Discuss how the results reflect the CLT. Plot the results (mean and standard deviation separately) to demonstrate any effects you want to discuss.

How the results reflect the CLT:

- As the sample size **increases**, the mean and standard deviation **decreases.**
  - This is because as the sample size increases, the sample means cluster more and more around the true mean.
  - Hence, the standard error i.e standard deviation **decreases.**

In [625]:

```
# Function to calculate the parameters
simulate <- function(n, lambda)
{
    # To store the different mean values
    sample_mean = rep(0,50000)
    for (i in 1:50000)
    {
        sample_mean[i] = mean(rpois(n,lambda))

        # Storing the mean and sd values in result
        result <- list()
        result$mean_value <- mean(sample_mean)
        result$sd_value <- sd(sample_mean)

        # Returning this result value
        return (result)
    }

}
```

In [626]:

```
# Calling the function to calculate the parameters
for (lambda in c(1,5,20))
{
    for (n in c(10,100,1000))
    {
        print( (simulate(n,lambda)) )
        cat("\n ------------------- \n ")
    }

}
```

```
$mean_value
[1] 1.4e-05

$sd_value
[1] 0.003130495


 -------------------
 $mean_value
[1] 1.96e-05

$sd_value
[1] 0.004382693


 -------------------
 $mean_value
[1] 1.946e-05

$sd_value
[1] 0.004351388


 -------------------
 $mean_value
[1] 7.2e-05

$sd_value
[1] 0.01609969


 -------------------
 $mean_value
[1] 0.0001014

$sd_value
[1] 0.02267373


 -------------------
 $mean_value
[1] 0.00010184

$sd_value
[1] 0.02277212
```

```
 -------------------
 $mean_value
[1] 0.000456

$sd_value
[1] 0.1019647


 -------------------
 $mean_value
[1] 0.0004018

$sd_value
[1] 0.08984521


 -------------------
 $mean_value
[1] 0.000397

$sd_value
[1] 0.0887719


 -------------------
```

In [627]:

```
# Function to plot the mean
plot_mean <- function(n,lambda)
{
    sample_mean = rep(0,n)

    for (i in 1:50000)
        sample_mean[i] = mean(rpois(n,lambda))

        # Plotting an histogram of the means
        h = hist(sample_mean, xlab ="Mean",
        col=ifelse(n<11, 'blue', ifelse( n<101,'red', ifelse(n<1001,'green','yellow') )),
        breaks = 30, freq = F,
        main=sprintf("n=%s, lambda=%s", n, lambda),
        cex.lab=2, cex.axis=1.5, cex.main=2.5, cex.sub=1.5)
}
```

In [628]:

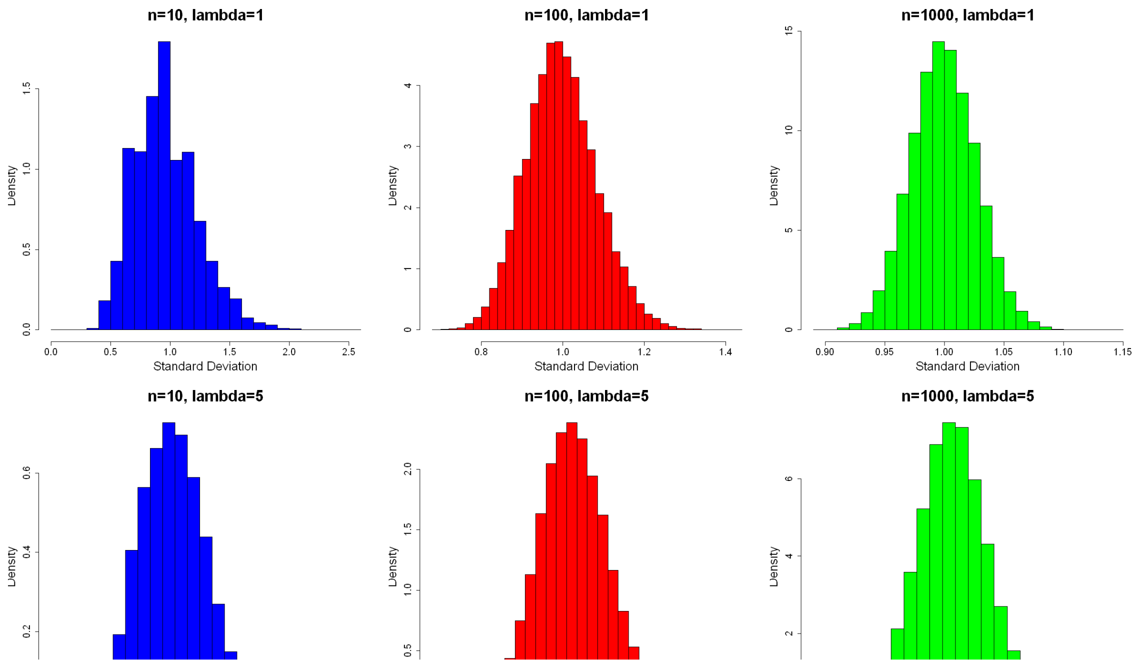
```
# Function to plot the standard deviation
plot_sd <- function(n,lambda)
{
    sample_deviation = rep(0,n)

    for (i in 1:50000)
        sample_deviation[i] = sd(rpois(n,lambda))

        # Plotting an histogram of the standard deviations
        h = hist(sample_deviation, xlab ="Standard Deviation",
        col=ifelse(n<11, 'blue', ifelse( n<101,'red', ifelse(n<1001,'green','yellow') )),
        breaks = 30, freq = F,
        main=sprintf("n=%s, lambda=%s", n, lambda),
        cex.lab=2, cex.axis=1.5, cex.main=2.5, cex.sub=1.5)
}
```

In [629]:

```
# Calling the function to plot means

par(mfrow=c(3, 3))
options(repr.plot.width=20, repr.plot.height=20)

for (lambda in c(1,5,20)) {
    for (n in c(10,100,1000)) {
        plot_mean(n, lambda)
    }
}
```

In [630]:

```
# Calling the function to plot standard deviations

par(mfrow=c(3, 3))
options(repr.plot.width=20, repr.plot.height=20)

for (lambda in c(1,5,20)) {
    for (n in c(10,100,1000)) {
        plot_sd(n, lambda)
    }
}
```



Question 6.d (5 Marks)

When rate λ_1 = 1 and λ_2 = 5 and sample size n is 10 or 100, obtain the z scores of the sample means (from 50000 simulations). Plot their distributions in a histogram with the theoretical Gaussian curve overlaid.

Note that for sample size 100, the plots overlay very nicely. But what happens with sample size 10? Explain the differences between the four plots.

For each simulation: the z score of the mean can be calculated as:

$$\frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

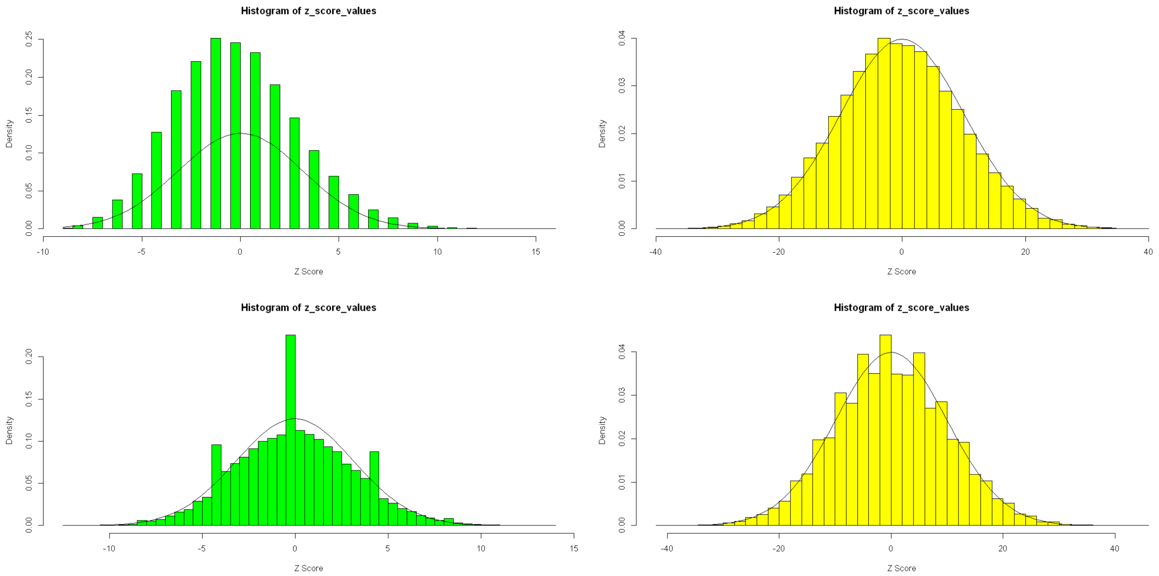
where $\bar{X}$ is the mean of the sample, μ is the population mean and σ is the population standard deviation.

In [631]:

```
# Function for calculating and plotting the zscore
plot_z_score <- function(lambda,n)
{
    # Calculating the population mean and standard deviation values
    mean_value = lambda
    sd_value = sqrt(lambda/n)

    # To store the values of the z_Score and sample means
    z_score_values = rep(0,50000)
    sample_means = rep(0,50000)

    for (i in 1:50000)
    {
        sample_means[i] = mean(rpois(n,lambda))
        z_score_values[i] = (sample_means[i] - mean_value)/(sd_value/sqrt(n))
    }
    #f = function(x, mean = mean_value, sd = sd_value) dnorm(x, mean = mean, sd = sd)

    hist(z_score_values, xlab ="Z Score", col=ifelse(n<99,'green','yellow'), breaks = 40,fr
    #curve(f, from = min(z_score), to = max(z_score), add = T)
    curve( dnorm(x, mean = mean(z_score_values), sd = sd(z_score_values)), add = T )

}
```

In [620]:

```
# Calling the zscore function to plot the graphs
par(mfrow=c(2, 2))
options(repr.plot.width=20, repr.plot.height=10)

for (lambda in c(1,5)) {
    for (n in c(10,100)) {
        plot_z_score(lambda,n)
    }
}
```



Difference between the four plots:

- When N = 10, The calculated values of the zscore are **slightly different** from the theoretical values of zscore.
- Also, When N = 10, the values of zscore lie in between 0 and 0.20
- When N = 100, We can notice than an increase in the value of N, causes a **significant** decrease in the value of zscore.
- Hence, When N = 100, the values of zscore lie in between 0 and 0.04
- As the value of lambda increases, there are **higher variations** in the values of zscore.
- For lambda = 1, the values of zscore increase evenly and then they decrease after a point.
- However, for lambda = 5, the values of zscore do not follow any pattern.
    - They gradually increase till a point, then there is a sudden spike and then they decrease suddenly.
    - These variations of sudden increase followed by a sudden decrease are greater as the value of N increases.

Also, we notice that the graph is not smooth when N=10. However, as the value of N increases, the graph becomes smoother.