Due Date: October 1st, 2021 *Problems Due: 1,3*

1. Security guard Hasini has a log book of the CMI-Shuttle. In the log book she keeps track of the kilometer reading before each time driver Sakshi fills petrol. The last 10 readings are:

   65311, 65624, 65908, 66219, 66499, 66821, 67145, 67447, 67786, 68103

   (a) Enter these numbers into R as a variable `kreading`. Use the function `diff` on the data. What does it give?

   ```
   > kreading = c(65311, 65624, 65908, 66219, 66499, 66821, 67145, 67447)
   > differences = diff(kreading)
   ```

   Write down,x, the number of kilometers between each time Sakshi fills up pertrol.

   (b) Use the `max` to find the maximum number of kilometers, the `mean` function to find the average number of kilometers and the `min` to get the minimum number of kilometers Sakshi has driven between two fill-ups.

2. Super Mani's quiz scores in Data science are given below

   7, 6, 10, 8, 7, 9, 9, 6, 4, 10, 8, 6, 9, 10

   (a) Enter this into R as a variable `scoreMani`. Use the function `max` to find the highest score, the function `mean` to find the average and the function `min` to find the minimum.

   (b) When confronted by Looser Siva, Mani realises that entry 4 was a mistake. It should have been 5. How can you fix this? Do so, and then find the new average.

   (c) What does the below command provide in R ?

   ```
   > sum( scoreMani >= 9)
   ```

   (d) What do you get? What percent of your scores are less than 17 ? How can you answer this with R?

3. Naina's cell phone bill varies from month to month. Suppose in her first year of Super DATA (hons.) program, under the Drop-atmost 10-calls monthly plan, the following monthly amounts were incurred:

   460, 330, 390, 370, 460, 300, 480, 320, 490, 350, 300, 480

   (a) Enter this data into a variable called `Nainabill`. Use the `sum` command to find the amount spent by Naina that year on the cell phone.

   (b) Using R find out what is the smallest amount she spent in a month and the largest amount she spent in a month ?

   (c) How many months was the amount greater than Rs 400? What percentage was this?

   (d) If her monthly loan from NOmoney Bank was Rs 3000. Using R store her balance(after paying her phone bill) in a variable called `freemoney`. Find the average amount available each month for her other expenses.

Name: Rishika Tibrewal **M.Sc. in Data Science, CMI**

Roll No: MDS202135 **Semester-1**

1. a) kreading=c(65311,65624,65908,66219,66499,66821,67145,67447,67786,68103);

   > differences=diff(kreading); ##Diff function gives the value by which two consecutive values of a vector differ, subtracting the first number from the next.

   > x=differences;

   > x

   [1] 313 284 311 280 322 324 302 339 317

   b) > max(x)

   [1] 339

   > mean(x)

   [1] 310.2222

   > min(x)

   [1] 280

2. a) > scoreMani=c(7, 6, 10, 8, 7, 9, 9, 6, 4, 10, 8, 6, 9, 10)

   > max(scoreMani)

   [1] 10
   > mean(scoreMani)
   [1] 7.785714
   > min(scoreMani)
   [1] 4
   b) > replace(scoreMani,9,5)

   [1] 7 6 10 8 7 9 9 6 5 10 8 6 9 10

   > mean(scoreMani)
   [1] 7.785714
   c) > sum(scoreMani>=9) ##this command counts the number of entries in scoreMani which are greater than 9.
   [1] 6
   d) > length(scoreMani[scoreMani<17])*100/length(scoreMani)

[1] 100

3. a) Nainabill=c(460, 330, 390, 370, 460, 300, 480, 320, 490, 350, 300, 480);

   > sum(Nainabill);

   [1] 4730

   b) > min(Nainabill);

   [1] 300

   > max(Nainabill);

   [1] 490

   c) > length(Nainabill[Nainabill>400]);

   [1] 5

   > length(Nainabill[Nainabill>400])*100/length(Nainabill);

   [1] 41.66667

   d) > loan=3000;

   > freemoney=loan-Nainabill;

   > freemoney;

   [1] 2540 2670 2610 2630 2540 2700 2520 2680 2510 2650

   [11] 2700 2520

   > mean(freemoney);

   [1] 2605.833

1. This worksheet will be graded.

2. Worksheet is due at 945am.

3. **Experiment 1:**

   (a) **Perform the following Experiment two times**: Using the (Play button) online Coin at: http://www.randomservices.org/random/apps/BinomialCoin.html
   Toss a fair coin 5 times and note down the outcome of each toss.

   (b) **Fill in the following Table accordingly**

Trial	Outcome of toss 1	Outcome of toss 2	Outcome of toss 3	Outcome of toss 4	Outcome of toss 5	**Y:= Number of Heads**
1						
2						

   Enter the data in the sheet on Experiment 1 in Google Drive link

4. **Experiment 2:**

   (a) **Perform the following Experiment two times**: Using the (Play button) online Dice at: http://www.randomservices.org/random/apps/Dice.html
   Roll a fair die 5 times and note down the outcome of each roll.

   (b) **Fill in the following Table accordingly**

Trial	Outcome of Roll 1	Outcome of Roll 2	Outcome of Roll 3	Outcome of Roll 4	Outcome of Roll 5	**Y:= Sum of the Rolls**
1						
2						

   Enter the data in the sheet on Experiment 2 in Google Drive Link

P.T.O.

5. **Experiment 3:**

   (a) **Perform the following Experiment two times**: Using the (Play button) online Dice at: http://www.randomservices.org/random/apps/DieCoin.html

   Roll a fair die once and Toss a fair coin as many times as the outcome on the roll. Note down the outcome of both the die and the roll.

   (b) **Fill in the following Table accordingly**

Trial	Outcome of Roll	Outcome of Toss	Outcome of Toss	Outcome of Toss	Outcome of Toss	**Y:= Number of Heads**
1						
2						

   Enter the data in the sheet on Experiment 3 in Google drive LINK

6. **Experiment 4:**

   (a) **Perform the following Experiment two times**: Using the (Play button) online Dice at: http://www.randomservices.org/random/apps/CoinDie.html
   Toss a fair coin: if head roll a 1-6 flat die (i.e 1,6 have probability $\frac{1}{4}$ and 2,3,4,5 have probability $\frac{1}{8}$); and if tail roll a 3-4 flat die (i.e 3,4 have probability $\frac{1}{4}$ and 1,2,5,6 have probability $\frac{1}{8}$).

   (b) **Fill in the following Table accordingly**

Trial	Outcome of Toss	Outcome of Roll 1	**Y:= Outcome of Roll 1**
1			
2			

   Enter the data in the sheet on Experiment 4 in Google Drive Link

7. On a sheet of paper write down the 4 tables and Scan the paper as a pdf file and upload into Moodle.

29/9/21

Name : Rishika Tibrewal Roll - MDS202135 .

Worksheet - 1

3).

Trial	Outcome of toss 1	Outcome of toss 2	Outcome of toss 3	Outcome of toss 4	Outcome of toss 5	Y = no. of heads
1	H	H	T	H	H	4
2	T	T	T	H	T	1

4)

Trial	Outcome of roll 1	Outcome of roll 2	Outcome of roll 3	Outcome of roll 4	Outcome of roll 5	Y = sum of rolls
1	3	1	2	5	1	12
2	3	1	1	5	4	14

5)

Trial	Outcome of roll	Outcome of toss	Outcome of toss	Outcome of toss	Outcome of toss	Y = no. of heads
1	3	T	T	H		1
2	1	H				1

6)

Trial	outcome of toss	outcome of roll 1	Y = outcome of roll 1 .
1	H	3	3
2	H	5	5

Due Date: October 8th, 2021
Problems Due: 1,3,5

1. Suppose `x` is a vector. Describe what each of the below commands do.

```
> length(x)
> x[2]
> x[-2]
> x[1:5]
> x(length(x) -5 : length(x))
> x[c(1,3,5)]
> x[x>3]
> x[x<-2 | x>2]
> which(x == max(x))
```

2. Consider the dataset `diamonds` in `ggplot2` inR.
   (a) In two to three lines describing the dataset.
   (b) Write down the list of categories considered.
   (c) Construct a Bar Plot using the below command:
      i.
```
> library(ggplot2)
>  ggplot(data=diamonds) +
+    geom_bar(mapping=aes(x=cut,  fill=clarity))+
+      scale_fill_viridis_d()
```
      ii.
```
> ggplot(data=diamonds) +
+     geom_bar(mapping=aes(x=cut,  fill=clarity), position="dodge") +
+     scale_fill_viridis_d()
```
   and describe the differences in the outputs.

3. Load the package `UsingR` consider the dataset `cavendish`.
   (a) In two to three lines describing the dataset.
   (b) Provide the five number summary of the three variables considered using the `summary` function.

4. Suppose we roll a dice five times. Let Y be the sum of the outcomes in each roll.Find the distribution of Y.

5. Toss a fair coin: if head roll a 1-6 flat die (i.e 1,6 have probability $\frac{1}{4}$ and 2,3,4,5 have probability $\frac{1}{8}$); and if tail roll a 3-4 flat die (i.e 3,4 have probability $\frac{1}{4}$ and 1,2,5,6 have probability $\frac{1}{8}$). Let X be the outcome of the toss of a coin. Let Y be the outcome of the roll of the die.

(a) Find the conditional distribution of $Y|X = Head$

(b) Find the conditional distribution of $Y|X = Tail$

(c) Find the $P(X = Head|Y = 3)$

6. Complete Worksheet 2.

Book-Keeping Exercises

From Probability and Statistics with Examples Using R

1. Ex 1.1.3
2. Ex 1.2.12
3. Example 1.3.10,1.3.12
4. Ex 1.3.9
5. Ex 1.3.10
6. Ex 1.3.13

Homework-2

1) Let x be a vector, say x = c(1,4,6,7,3,6,8,2,4,1,4,6,7)

(a) length(x) gives the size of the vector 'x'. Here, it will give 13

(b) x[2] gives the second element from the start. Here, 4.

(c) x[-2] gives all the elements of x by deleting the element in 2nd position. Here, output will be 1 6 7 3 6 8 2 4 1 4 6 7

(d) x[1:5] gives all the elements of x from position 1 to position 5. Here, 1 4 6 7 3.

(e) x[length(x)-5 : length(x)] gives all elements from length(x)-5 position to start in reverse order. i.e, from end. Here,

2 8 6 3 7 6 4 1.

f). x[c(1,3,5)] gives elements in the 1st, 3rd & 5th position of x
Here, 1 6 3

g) x[x>3] gives all elements of x which are greater than 3
Here, 4 6 7 6 8 4 4 6 7.

h) x[x<-2 | x>2] gives all elements of x as it assigns x with value 2 which is a true value.

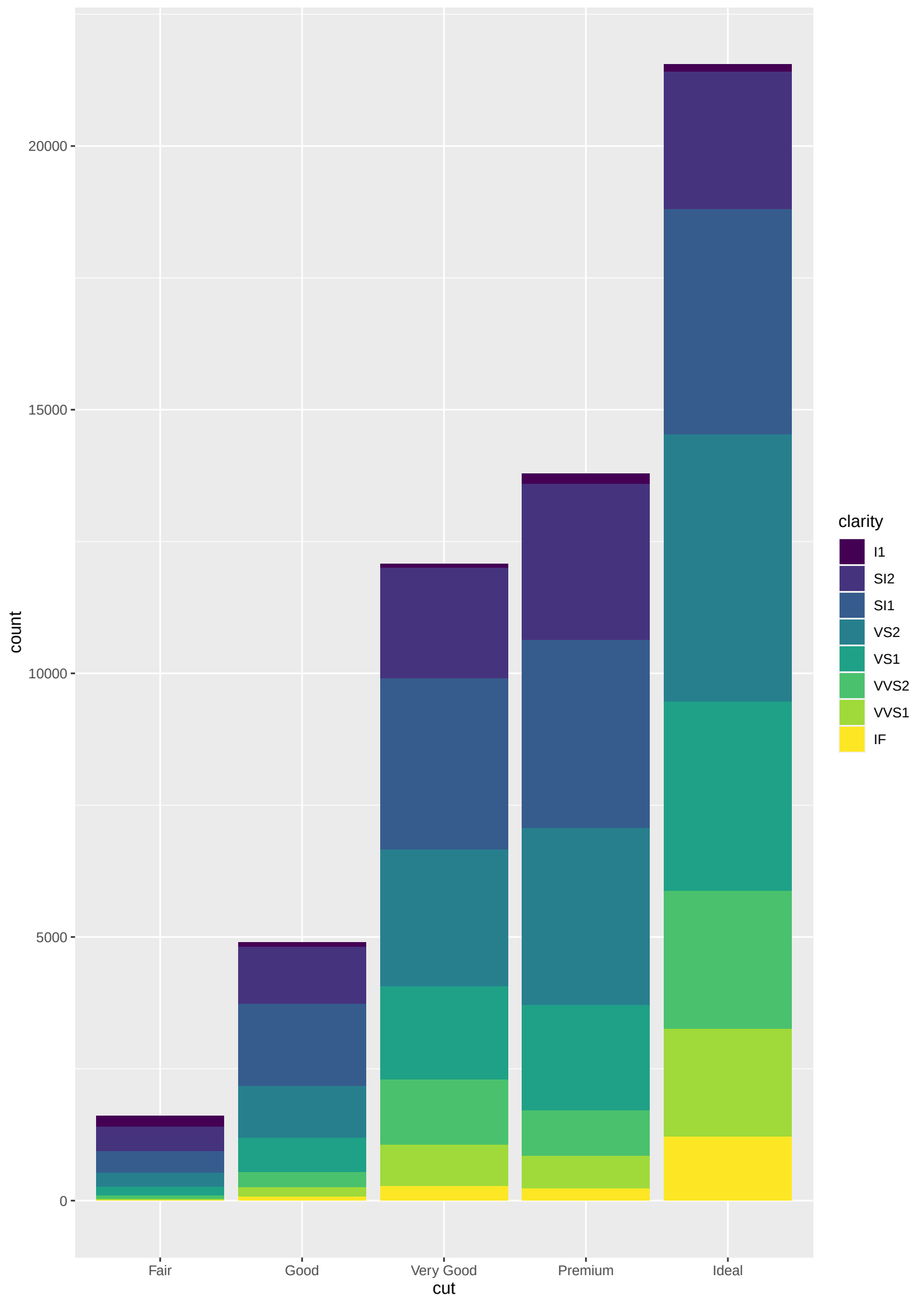
i) which (x == max(x)) gives all the positions of the maximum element of the vector. Here it will give 7.

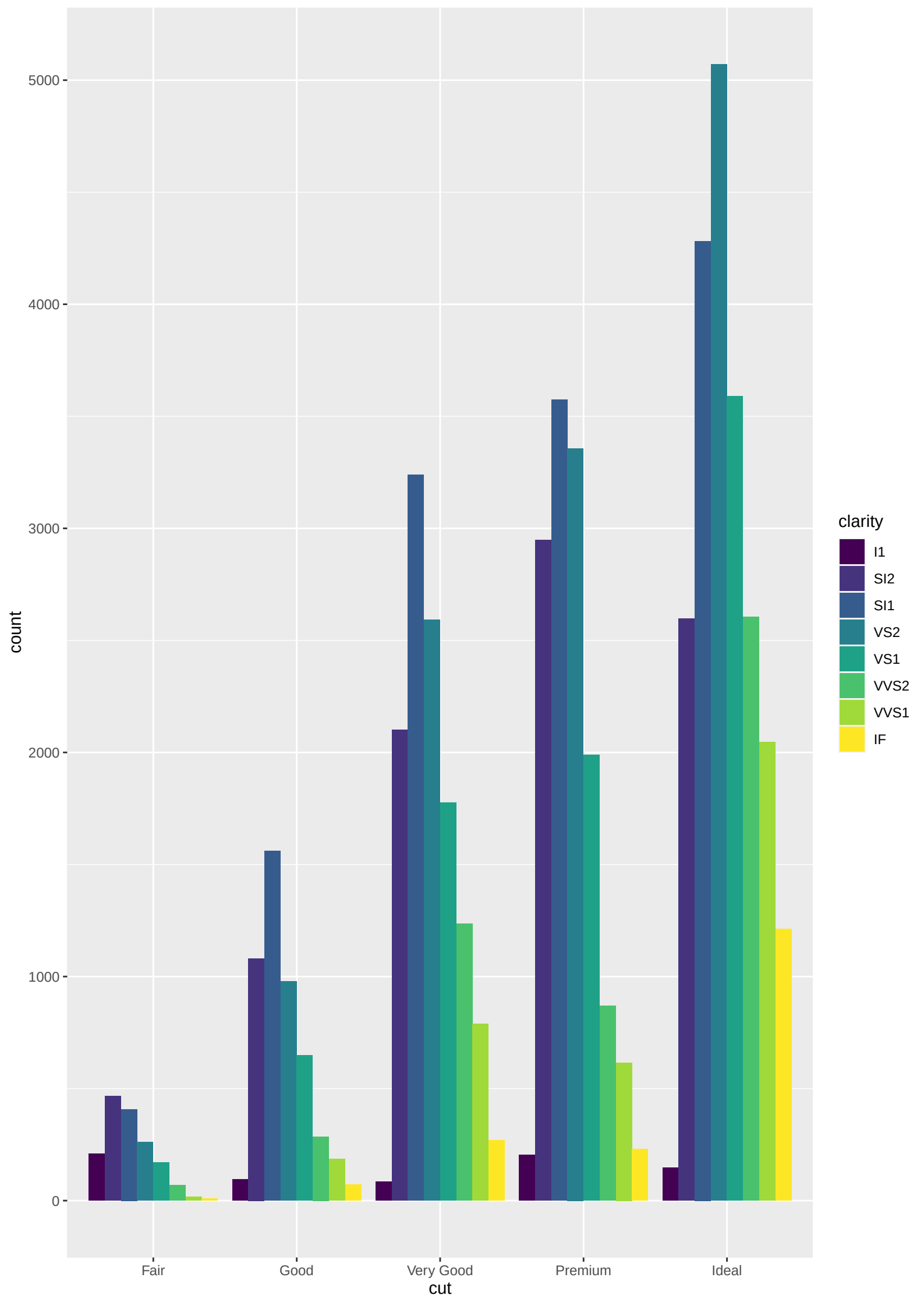
2) (a) The dataset 'diamonds' in 'ggplot2' contains data about different types of diamonds. The characteristics taken into consideration are price, carat (weight), cut, color, clarity, depth percentage, length, width and depth. It has 53940 rows and 10 variables.

(b) 'Diamonds' contains a data frame with 53940 rows and 10 variables. Below is given the ~~the~~ list of categories considered, with what stand for.

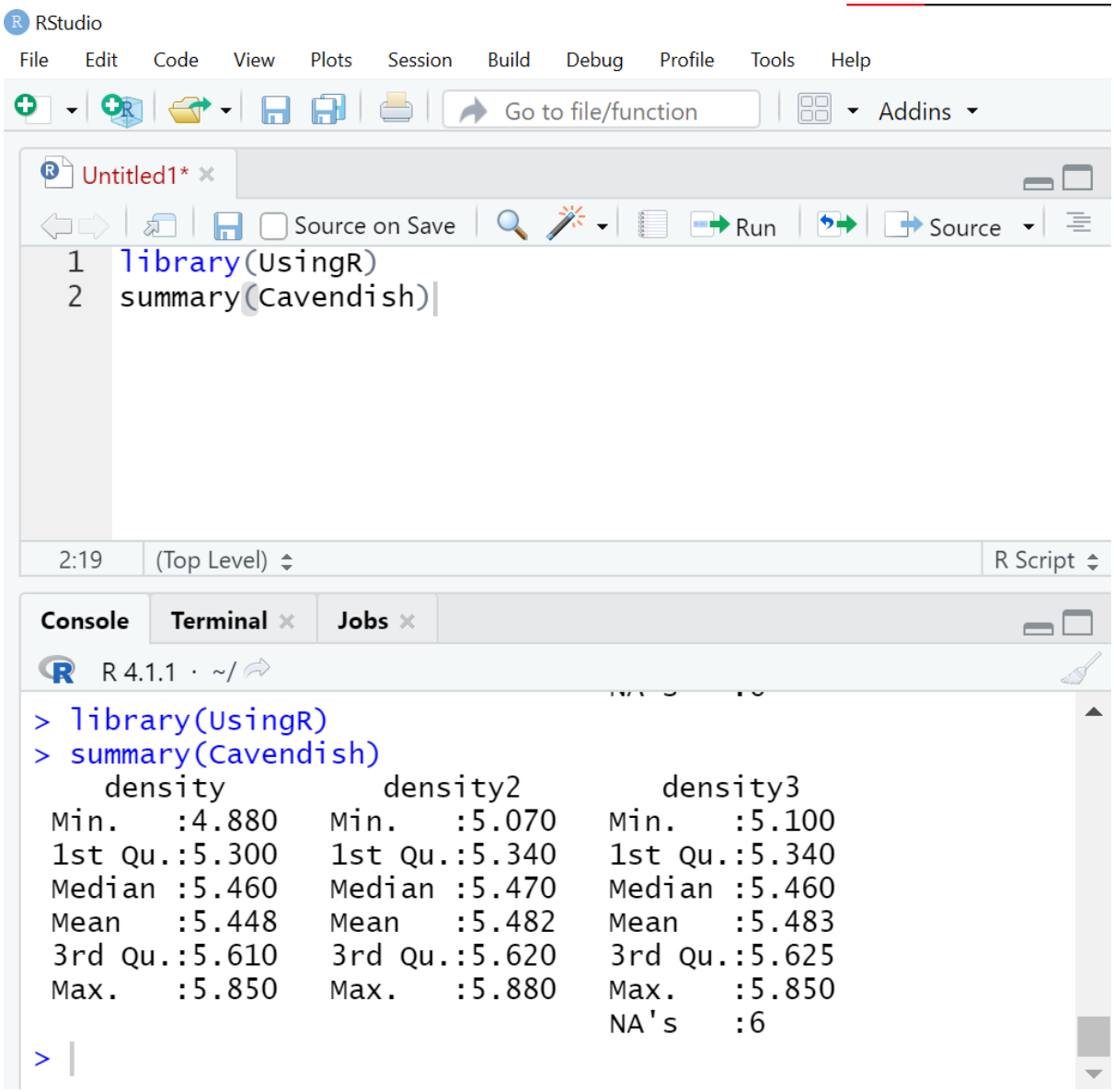
- Price (in US dollars)
- carat - weight of diamond (0.2 to 5.01)
- cut - quality of diamond (fair, good, very good, premium, ideal)
- color - (D being the best & J being the worst)
- clarity - (I1 being the best, IF being the worst, SI2, SI1, VS2, VVS2, VVS1)
- x - length (0 to 10.74 mm)
- y - width (0 to 58.9 mm)
- z - depth (0 to 31.8 mm)
- depth - depth % $= \frac{z}{\text{mean}(x,y)} = \frac{2z}{(x+y)}$ (43-79%)
- table - width of top of diamond w.r.t to widest point (43-95)

(c) The only difference between the two graphs is the dodge = "position" which basically preserves the vertical position of an geom (layout of a ggplot2) while adjusting the horizontal position. In the first graph ~~on top of~~ bars are arranged in horizontal manner while in 2nd one, bars are arranged vertically next to each other for a particular kind of cut.


20000
15000
10000
5000
0
count
Fair
Good
Very Good
Premium
Ideal
cut
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF


5000
4000
3000
2000
1000
0
count
Fair
Good
Very Good
Premium
Ideal
cut
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

3) cas Cavendish is a dataset in R which contains the observations of the experiments carried out by Henry Cavendish in 1798 to find the mean density of the Earth for calculation of ~~the~~ gravitational constant G. Stigler in 1977 used these data to determine the properties of the estimators. He believed that trimmed ~~mean~~ means equally or better than robust estimators. It has 29 observations with 3 variables: density, density2 and density3.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
Source on Save
Run
Source
1 library(UsingR)
2 summary(Cavendish)
2:19
(Top Level)
R Script
Console
Terminal
Jobs
R 4.1.1 · ~/
> library(UsingR)
> summary(Cavendish)
density
Min. :4.880
1st Qu.:5.300
Median :5.460
Mean :5.448
3rd Qu.:5.610
Max. :5.850
density2
Min. :5.070
1st Qu.:5.340
Median :5.470
Mean :5.482
3rd Qu.:5.620
Max. :5.880
density3
Min. :5.100
1st Qu.:5.340
Median :5.460
Mean :5.483
3rd Qu.:5.625
Max. :5.850
NA's :6
>

```
> k=NULL;x1=1:6;x2=1:6;x3=1:6;x4=1:6;x5=1:6;c=1;
> for(i1 in x1){
+   for(i2 in x2){
+     for(i3 in x3){
+       for(i4 in x4){
+         for(i5 in x5){
+           k[c]=i1+i2+i3+i4+i5;
+           c=c+1;
+         }
+       }
+     }
+   }
+ }
> range_y=unique(k)
> range_y
 [1]  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
[18] 22 23 24 25 26 27 28 29 30
> prob_y=function(g)
+ {
+   sum(g==k)/length(k);
+ }
> prob_y(19)
[1] 0.0945216
> prob_y(15)
[1] 0.08371914
> prob_y(5)
[1] 0.0001286008
>
```

5) 5) $X \to$ outcome of a fair coin
$Y \to$ outcome of die.

(a) $P(Y=y \mid X=H) = \begin{cases} 1/4 & ; y=1,6 \\ 1/8 & ; y=2,3,4,5 \\ 0 & ; \text{o.w.} \end{cases}$ (1-6 flat die)

(b) $P(Y=y \mid X=T) = \begin{cases} 1/4 & ; y=3,4 \\ 1/8 & ; y=1,2,5,6 \\ 0 & ; \text{o.w.} \end{cases}$ (3-4 flat die)

(c) $$P(X=H \mid Y=3) = \frac{P((X=H) \cap (Y=3))}{P(Y=3)}$$

$$= \frac{P(X=H)P(Y=3 \mid X=H)}{P(X=H)P(Y=3 \mid X=H) + P(X=T)P(Y=3 \mid X=T)}$$

$$= \frac{1/2 \cdot 1/8}{1/2 \cdot 1/8 + 1/2 \cdot 1/4} = \frac{1/2}{1/2+1} = \frac{1/2}{3/2}$$

$$= 1/3.$$

1. You can create your own functions in R. These are created using the `function` command. For example, we can design our own function to calculate mean.

```
> MYMEAN = function(x) { sum(x)/length(x)}
```

Then you can say

```
> x = c(1,2,3)
> MYMEAN = function(x) { sum(x)/length(x)}
> MYMEAN(x)


[1] 2
```

A function in R is another object, with the class `function`. It typically will return the last value computed in the body. Compute the output of `MYMEAN` for

```
> x = 1:100
> y = x[x<50 | x >2]
```

2. Suppose roll a fair die two times and let X_1 and X_2 be denote outcomes on each of the rolls. Let $Y = X_1 + X_2$.
   (a) Find the Range of Y
   (b) For each $y \in Y$, find the $f_Y(y) = P(Y = y)$.
   (c) Write an `R`-function that returns $f_Y(\cdot)$ for any given value y in Range of Y.

3. Suppose roll a fair die once and let X denote the outcome. Then we toss a biased coin, with $0 < p < 1$ being probability of obtaining heads, X times. Let Y denote the number of heads in X tosses.
   (a) Find the Range of Y
   (b) For each $y \in Y$, find the $f_Y(y) = P(Y = y)$.
   (c) Write an `R`-function that returns $f_Y(\cdot)$ for any given value y in Range of Y.

Worksheet2

Rishika Tibrewal

07/10/2021

```
1)

MYMEAN = function(x) { sum(x)/length(x)}
x = 1:100
x

##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
18
##  [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
36
##  [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
54
##  [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
72
##  [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89
90
##  [91]  91  92  93  94  95  96  97  98  99 100

MYMEAN(x)

## [1] 50.5

y = x[x<50 & x >2]
y

##  [1]  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
26 27
## [26] 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49

MYMEAN(y)

## [1] 26
```

2)(a) $X_1, X_2 \rightarrow$ outcome on roll of a die

$Y = X_1 + X_2$

Clearly, range of X_1 & $X_2 = \{1,2,3,4,5,6\}$

$\therefore$ Range $(Y) = \{2,3,4,\cdots,11,12\}$

(b) $P(Y=y) = P(X_1 + X_2 = y) = P((X_1 = x), (X_2 = y-x))$

($X_1 = x$, say)

Now, $P(X_1 = x) \neq 0 \;\forall\; x \in \{1,2,3,4,5,6\}$

$P(X_2 = y-x) \neq 0 \;\forall\; y-x \in \{1,2,3,4,5,6\}$

$\therefore P(Y=y) \neq 0$.

Now,

$P(Y=2) = \frac{1}{36}$ $\quad P(Y=3) = \frac{2}{36}$ $\quad P(Y=4) = \frac{3}{36}$

$\cdots$ $P(Y=7) = \frac{6}{36}$ $\quad P(Y=8) = \frac{5}{36}$

$P(Y=9) = \frac{4}{36}$ $\quad \cdots \quad P(Y=12) = \frac{1}{36}$.

In general, $P(Y=y) = \begin{cases} \frac{y-1}{36} & ; y \leq 7 \\ \frac{13-y}{36} & ; y \geq 8 \end{cases}$.

2)

```
x1=1:6; x2=1:6; c=1; k=NULL;
for(i1 in x1){
  for(i2 in x2){
    k[c]=i1+i2
    c=c+1
  }
}
range_y=unique(k)
range_y
```

```
##  [1]  2  3  4  5  6  7  8  9 10 11 12
```

```
probab_y=function(y){
  sum(y==k)/length(k)
}
probab_y(5)
```

```
## [1] 0.1111111
```

```
probab_y(10)
```

```
## [1] 0.08333333
```

```
probab_y(3)
```

```
## [1] 0.05555556
```

3) (a) $Y \to$ no. of heads in X tosses where X is the outcome on a die

$\therefore$ Range $(Y) = \{0, 1, 2, \cdots, X\}$. where Range $(X) = \{1, 2, \cdots 6\}$

(b) By law of total probability, $P(Y=y) = \sum_{x=1}^{6} P(Y=y \mid X=x) P(X=x) = \sum_{x=1}^{6} \binom{x}{y} p^y (1-p)^{x-y} \cdot \frac{1}{6}$.

But $\binom{x}{y}$ makes sense only when $y \leq x$, so,

$$P(Y=y) = \sum_{x=y}^{6} \binom{x}{y} p^y (1-p)^{x-y} \cdot \frac{1}{6}$$

$$= \frac{p^y}{6(1-p)^y} \sum_{x=y}^{6} \binom{x}{y} (1-p)^x .$$

```
3)

prob_y=function(y,p){
   s = 0
  for (j in y:6) {
    s = s + choose(j,y)*((1-p)^j)
  }
    return(s*((p^y)/(6*((1-p)^y))))
}
prob_y(4,0.2)

## [1] 0.003893333

prob_y(6,0.5)

## [1] 0.002604167
```

Due Date: October 15th, 2021
Problems Due: 2,4,6

From Probability and Statistics with Examples Using R

1. Ex. 4.1.2
2. Ex 4.1.3
3. Ex 4.1.6
4. Ex 4.2.1
5. Ex 4.2.2
6. Ex 4.2.3
7. Complete Worksheet 3.

Book-Keeping Exercises

From Probability and Statistics with Examples Using R

1. Example 4.1.4
2. Example 4.1.5
3. Example 4.1.16
4. Example 4.1.17

Homework 3

Let $X \sim Geo(p)$ where p is the probability of success in a single trial, $0 < p < 1$.

$$E(X) = \sum x f(x) = \sum_{k=1}^{\infty} k \cdot p(1-p)^{k-1}$$

$$= \lim_{n\to\infty} \sum_{k=1}^{n} k\,p(1-p)^{k-1} = \lim_{n\to\infty} \sum_{k=1}^{n} k\,[1-(1-p)](1-p)^{k-1}.$$

$$= \lim_{n\to\infty} \left\{ \sum_{k=1}^{n} (1-p)^{k-1} \cdot k - \sum_{k=1}^{n} k(1-p)^{k} \right\}.$$

$$= \lim_{n\to\infty} \left\{ \sum_{k=1}^{n} (1-p)^{k-1} - n(1-p)^{n} \right\}.$$

$$= \lim_{n\to\infty} \left\{ \frac{1-(1-p)^{n}}{p} - n(1-p)^{n} \right\}$$

$\because 0 < p < 1 \Rightarrow (1-p)^{n} \to 0$ and $(1-p)^{n} \cdot n \to 0$.

$\therefore E(X) = \lim_{n\to\infty} \frac{1}{p}$ as $n \to \infty$.

Homework-3.

2) (a) Let Y denote the number of questions asked until the contestant does not know the correct answer. So, $Y \sim Geo(p)$ where $p = 0.12$ which is the probability that the contestant does not know the answer correctly.

(b) $E(Y) = \frac{1}{p} = \frac{1}{0.12} = 8.333$.

So, expected number of questions asked until the contestant does not know the answer for the next question

(c) $P(Y=y) = (1-p)^{y-1} p$.

$\Rightarrow \frac{dP}{dy} = p(1-p)^{y-1} \cdot \log(1-p)$

$\because 0 < p < 1 \Rightarrow (1-p) > 0$ but $(1-p) < 1 \Rightarrow \log(1-p) < 0$.

$\Rightarrow \frac{dP}{dy} < 0$ So, P is a decreasing function

Now, Range of $Y = \{1, 2, 3, \ldots\}$ so the maximum value of P will be for the first value of Y which is 1.

So, mode = 1.

(d) Probability that the contestant is able to answer all 12 questions correctly and wins $= (0.88)^{12}$.

$= 0.2156711$.

4) $P(X=0) = 0.2$; $P(X=1) = 0.5$; $P(X=2) = 0.2$; $P(X=3) = 0.1$

$E(X) = \sum x f(x) = (0 \times 0.2) + (1 \times 0.5) + (2 \times 0.2) + (3 \times 0.1) = 1.2$

$Var(X) = E(X^2) - [E(X)]^2$

$E(X^2) = \sum x^2 f(x) = (0 \times 0.2) + (1 \times 0.5) + (4 \times 0.2) + (9 \times 0.1) = 2.2$

$Var(X) = 2.2 - (1.2)^2 = 2.2 - 1.44 = 0.76$.

$SD(X) = \sqrt{Var(X)} = \sqrt{0.76} = 0.871779$.

$P(|X - E(X)| > SD(X)) = P(|X - 1.2| > 0.871779)$

$$= P(X-1.2 < -0.871779) + P(X-1.2 > 0.871779)$$

$$= P(X < 0.328221) + P(X > 2.071779)$$

$$= P(X=0) + P(X=3)$$

$$= 0.2 + 0.1 = 0.3 .$$

b) X denotes the number of rolls needed before we see the first 3. Here, getting a 3 is considered as a success. So, $X \sim Geo(1/6)$ where $p = 1/6$.

(a) $E(X) = \frac{1}{p} = \frac{1}{1/6} = 6$. (proved above).

(b) ~~$E(X^2) = \sum x_i f(x_i) = 0 \times 1/6$.~~

$$SD(X) = \sqrt{Var(X)} = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-1/6}{(1/6)^2}}$$

$$= \frac{\sqrt{5/6}}{1/6} = 6\sqrt{\frac{5}{6}} = \sqrt{30} = 5.4772.$$

(c) Given that $X \in (E(X) - SD(X), E(X) + SD(X))$, we get $X \in (0.5228, 11.4772)$ which means, Range of $X = \{1, 2, \ldots, 11\}$. This implies that we may have to roll the die at most 11 times to get the first 3 on the die. So, according to the given condition, it does not seem unusual to roll the die more than 9 times to get the first 3.

(d) $P(X > 9) = 1 - P(X \leq 9) = 1 - \sum_{i=1}^{9} (5/6)^{i-1} (1/6)$

$$\left[\because X \sim Geo(p) \Rightarrow f(x) = q^{x-1} \cdot p \text{ where } q = 1-p \text{ \& } x \in \text{Range}(X)\right]$$

$$= 1 - \left(\frac{1}{6}\right) \sum_{i=1}^{9} (5/6)^{i-1} = 1 - \frac{1}{6}\left[1 + \frac{5}{6} + \left(\frac{5}{6}\right)^2 + \ldots + \left(\frac{5}{6}\right)^8\right]$$

$$= 1 - \left(\frac{1}{6}\right)\left[\frac{1-(5/6)^9}{1-(5/6)}\right] \quad \left[\because \text{GP series with } r = 5/6, \; n = 9\right]$$

$$= 1 - \frac{(1/6)}{(1/6)}\left[1-(5/6)^9\right] = 1 - 1 + (5/6)^9 = (5/6)^9 = 0.1938$$

(e) $P[E(X) - SD(X) \leq X \leq E(X) + SD(X)] = P(0.5228 \leq X \leq 11.4772)$

$= P(X \leq 11.4772) - P(X \leq 0.5228) = P(X \leq 11.4772) - 0$

$= (1/6) \sum_{i=1}^{11} (5/6)^{i-1} = \frac{1}{6} \times \frac{\left[1-(5/6)^{11}\right]}{1-(5/6)} = 1 - (5/6)^{11} = 0.8654$

1. Rolling a die.

```
> x = c(1,2,3,4,5,6)
> probx= c(1/4,1/8,1/8,1/8,1/8,1/8,1/4)
> F16=sample(x, size=1500,  replace=T, prob=probx)
```

   (a) Describe what each `R` command is peforming in the above.

   (b) Using the `mean` and `var` command find the mean and variance of `F16`. From this information alone what would you conclude is the range of the random variable `F16`.

   (c) Does the mean and variance from the sample generated compare closely with the true mean and variance of `F16`.

2. Tossing a coin 10 times.

```
> b1 = rbinom(100,10,0.5)
> b2 = rbinom(100,10,0.25)
> b3 = rbinom(100,10,0.75)
```

   (a) Using the `?rbinom` explain what each of the above commands is performing in `R`

   (b) Using the `mean` and `var` command find the mean and variance of `b1,b2,b3`. Compare them with the true mean and variance of the Binomial distribution.

3. `geom_hist` command.

```
> library(ggplot2)
> df1=data.frame(b1)
> p11= ggplot(df1) + geom_histogram(mapping=aes(x=b1), color="black", fill="NA", binwidth=1)
> p21= ggplot(df1) +
+         geom_histogram(mapping=aes(x=b1, y=..density..), color="black", fill="NA", binwidth=1)
```

   (a) Explain what are the plots p11,p21 providing.

   (b) Rewrite the code to provide the plots for `b2` and `b3`.

   (c) What can you say about the three plots ?

4. Density Approximation. The below code plots the function `density` fin the interval $(0, 10)$ with $a = 5, s = \sqrt{2.5}$ along with the plot `p21`.

```
> library(ggplot2)
> density = function(x,a,s){ (1/((2*pi)^(0.5)*s ))* exp(-(x-a)^2/(2*s^2))}
> df1=data.frame(b1)
> p21= ggplot(df1) +
+          geom_histogram(mapping=aes(x=b1, y=..density..), color="black", fill="NA", binwidth=1) +
+          xlim(0,10) +
+          geom_function(fun=density, args=list(a=5,s=(2.5)^(0.5)))
```

   (a) From the picture what does $\int_3^6 \texttt{density}(x, 5, \sqrt{2.5})dx$ approximate ?

   (b) If

$$\text{Area under the histogram between 3 and 7} \approx \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx,$$

   then what would be your guess for a and b

(c) How would you try the same idea for `b2` and `b3` ? Would you get the same result ?

5. (Sums of Rolls) Suppose we wish to simulate in R the experiment that we did in class last week of Rolling a die and noting down its sum. We can use the `sample`, `matrix` and `apply`.

```
> x = c(1,2,3,4,5,6)
> probx= c(1/6,1/6,1/6,1/6,1/6,1/6)
> Rolls=sample(x, size=1500,  replace=T, prob=probx)
> Rollm=matrix(Rolls, nrow = 5)
> Rollsums = apply(Rollm, 2, sum)
```

(a) Describe the commands `matrix` and `apply`

```
> library(ggplot2)
> density = function(x,a,s){ (1/((2*pi)^(0.5)*s ))* exp(-(x-a)^2/(2*s^2))}
> dfrolls = data.frame(Rollsums)
> mu = mean(dfrolls$Rollsums)
> sigma= sd(dfrolls$Rollsums)
> ggplot(data=dfrolls) + geom_histogram(mapping=aes(x=Rollsums,y=..density..), color="#00846b", fill=NA, binwidth=1) + xlim(5,30)+geom_function(fun=density, args=list(a=mu, s= sigma), color="blac
```

(a) From the picture what does $\int_{12}^{21}$`density`$(x, mu, sigma)dx$ approximate ?

(b) If

$$\text{Area under the histogram between 12 and 21} \approx \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx,$$

then what would be your guess for a and b

Q1 a)

First 2 commands create a vector and store the mentioned values. THird command generates a sample of size 1500 with replacemet considering the vector x and probabilities of its elements.

```
 x = c(1,2,3,4,5,6)
 probx= c(1/4,1/8,1/8,1/8,1/8,1/4)
 F16=sample(x, size=1500, replace=T, prob=probx)

mean=mean(F16)
var=var(F16)
mean

## [1] 3.49

var

## [1] 3.839126
```

Looking at the mean and variance, range of F16 is approximately [2,5]. [Mean - S.D , Mean + S.D]

```
true_mean=sum(x*probx)
x2=sum((x**2)*probx)
true_var=x2-(true_mean)**2
true_mean

## [1] 3.5

true_var

## [1] 3.75
```

Hence, sample and true mean/var are close to each other.

Q2 (a)

```
?rbinom

## starting httpd help server ... done
```

rbinom generates random values with arguments n,size,prob in order where n = total observations, size = total trials and prob= probability of success

```
b1 = rbinom(100,10,0.5)
b2 = rbinom(100,10,0.25)
b3 = rbinom(100,10,0.75)

mean_b1=mean(b1)
mean_b2=mean(b2)
mean_b3=mean(b3)
```

```
mean_b1
```

```
## [1] 4.88
```

```
mean_b2
```

```
## [1] 2.38
```

```
mean_b3
```

```
## [1] 7.54
```

```
var_1=var(b1)
var_2=var(b2)
var_3=var(b3)
```

```
var_1
```

```
## [1] 2.248081
```

```
var_2
```

```
## [1] 1.955152
```

```
var_3
```

```
## [1] 1.806465
```

```
true_mean_b1 = 10*0.5
true_mean_b2 = 10*0.25
true_mean_b3 = 10*0.75

true_var_b1 = true_mean_b1*0.5
true_var_b2 = true_mean_b2*0.75
true_var_b3 = true_mean_b3*0.25
```

```
true_mean_b1
```

```
## [1] 5
```

```
true_mean_b2
```

```
## [1] 2.5
```

```
true_mean_b3
```

```
## [1] 7.5
```

```
true_var_b1
```

```
## [1] 2.5
```

```
true_var_b2
```

```
## [1] 1.875
```

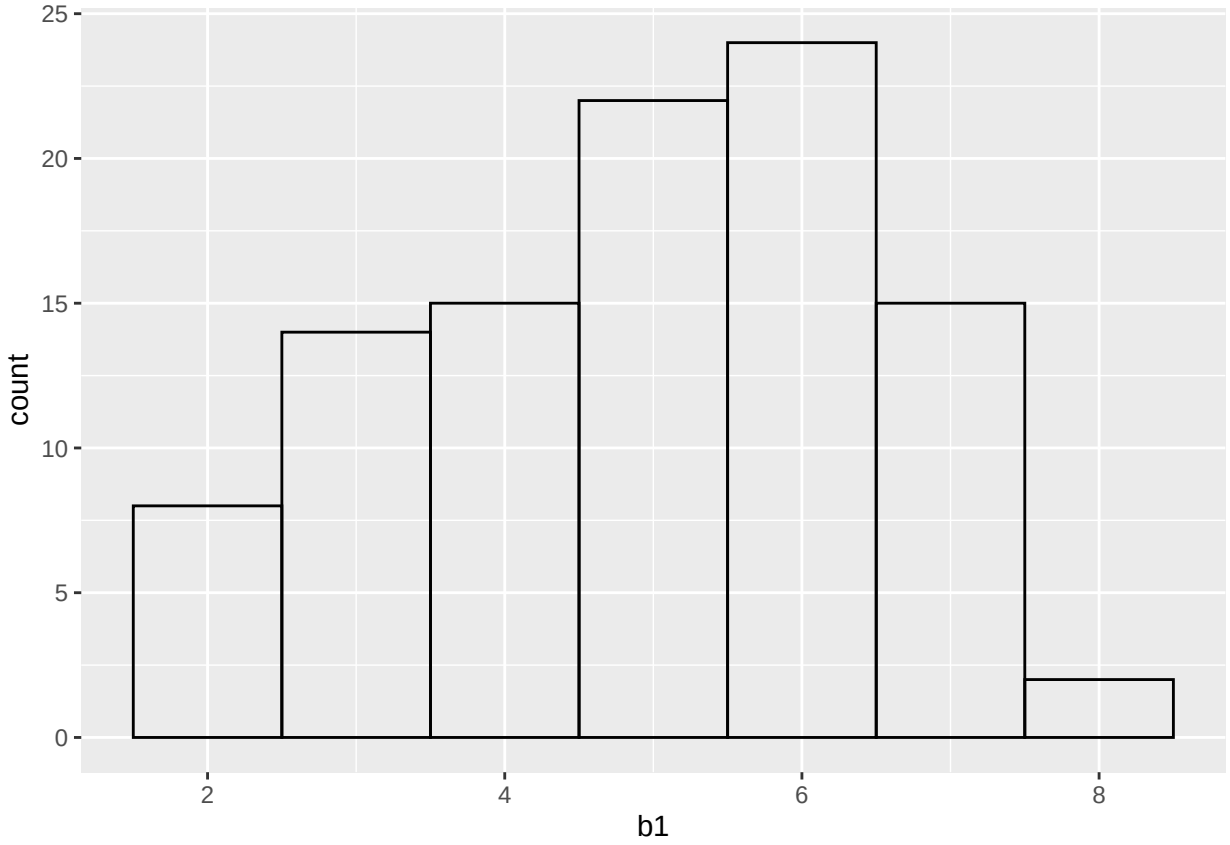
```
true_var_b3
```

```
## [1] 1.875
```

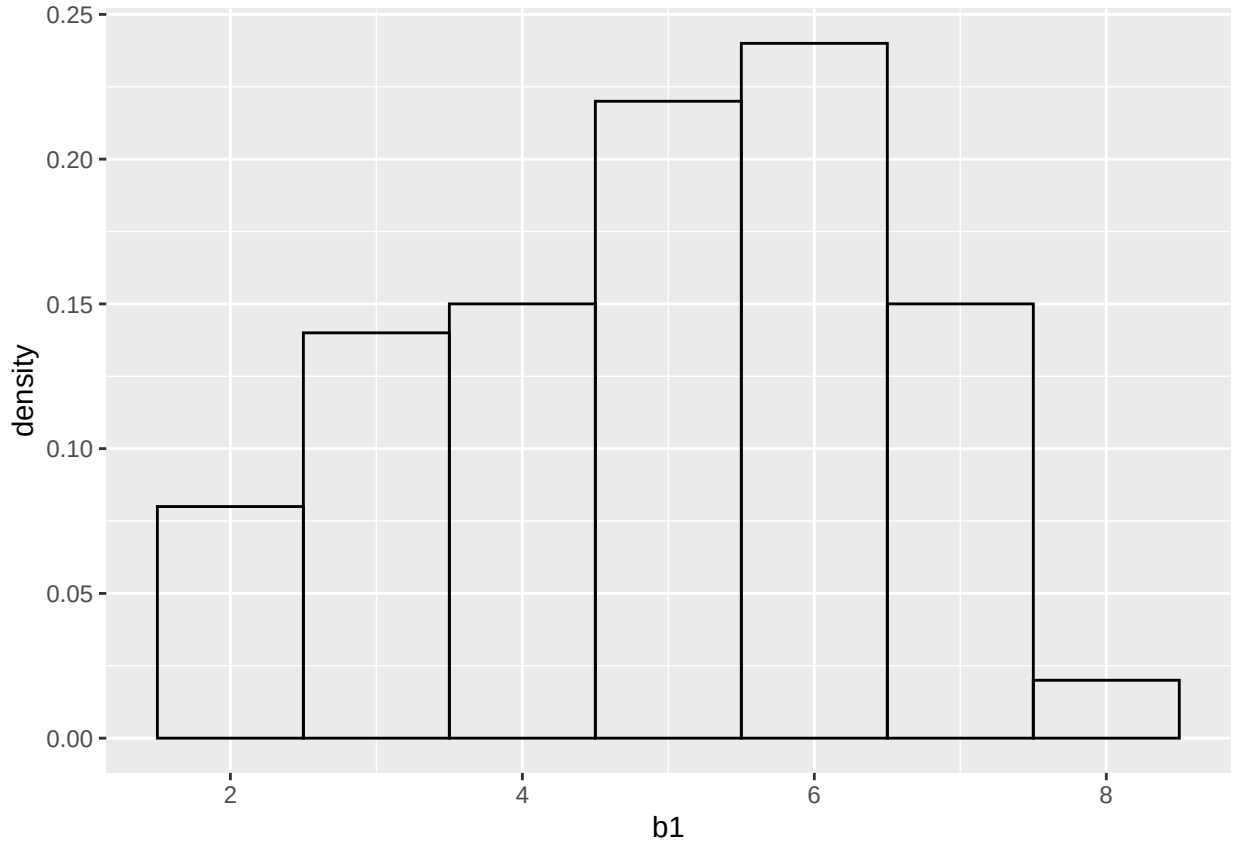
Clearly,true mean and variance are close to each other.

Q3

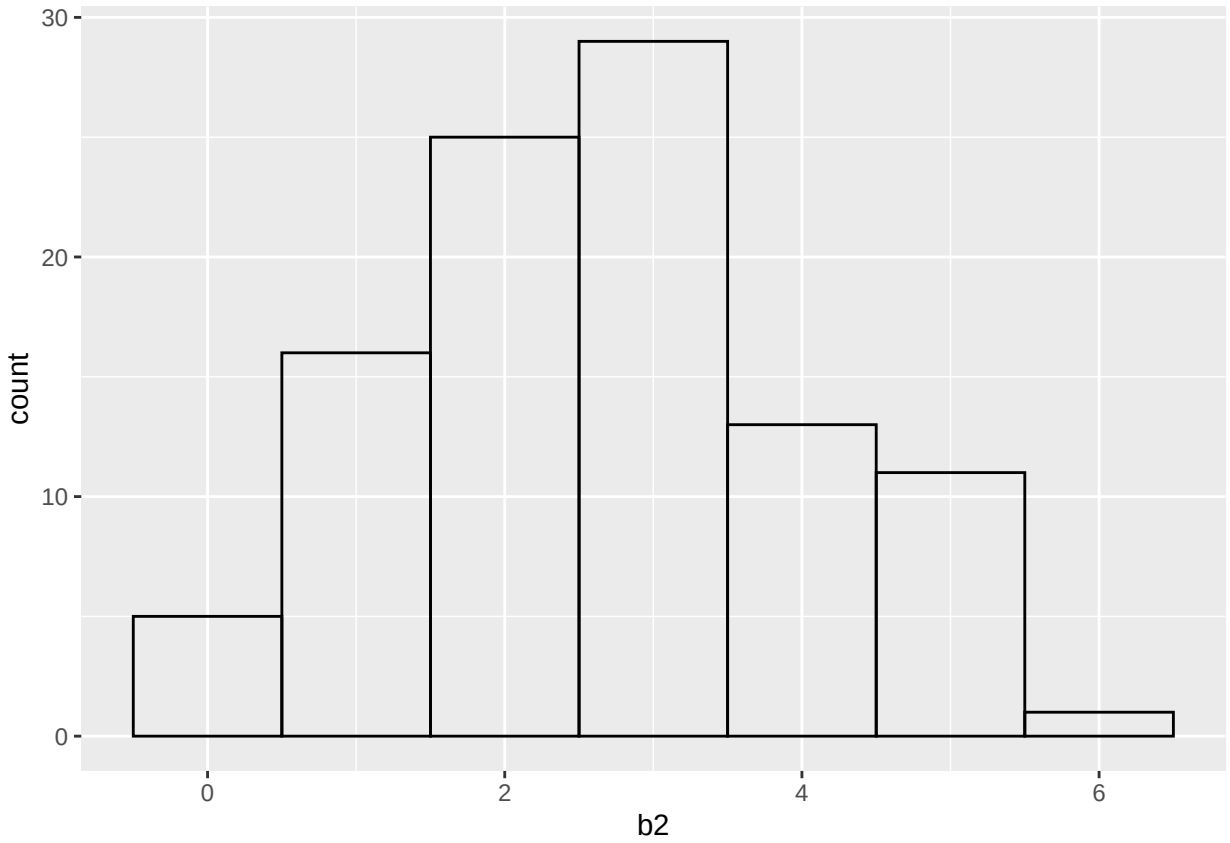
```
library(ggplot2)
b1 = rbinom(100,10,0.5)
df1=data.frame(b1)
p11= ggplot(df1) + geom_histogram(mapping=aes(x=b1), color="black", fill="NA", binwidth=1)
p21= ggplot(df1) + geom_histogram(mapping=aes(x=b1, y=..density..), color="black", fill="NA", binwidth=
p11
```



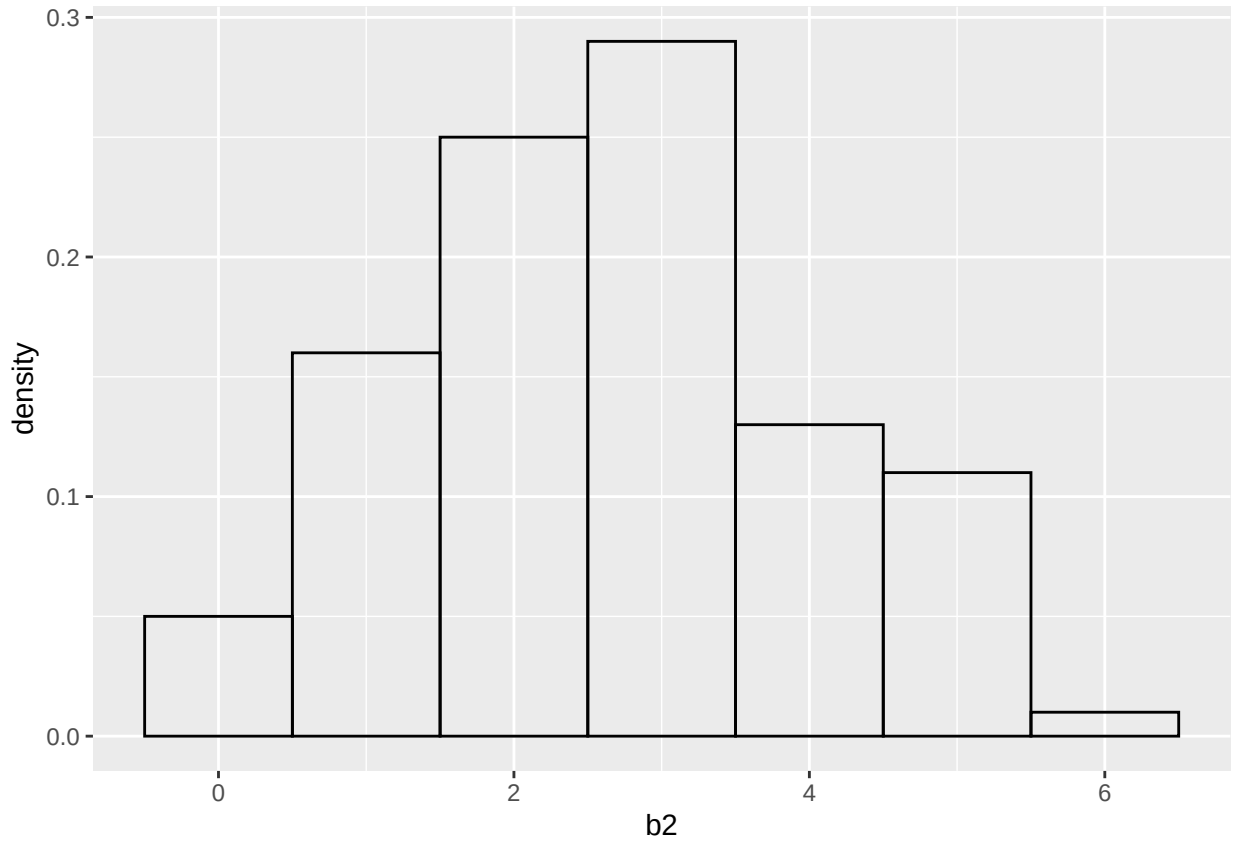

```
p21
```



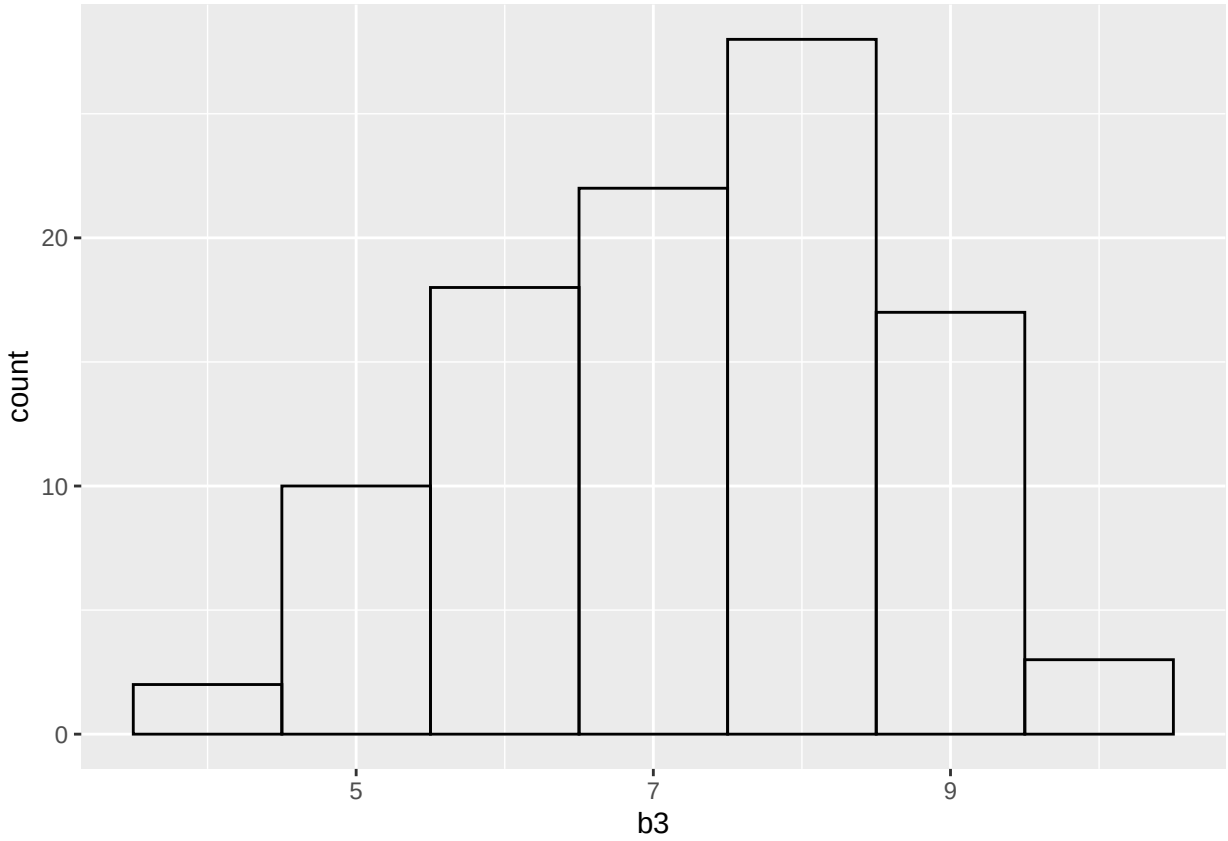

```
b2 = rbinom(100,10,0.25)
df1=data.frame(b2)
p11= ggplot(df1) + geom_histogram(mapping=aes(x=b2), color="black", fill="NA", binwidth=1)
p21= ggplot(df1) + geom_histogram(mapping=aes(x=b2, y=..density..), color="black", fill="NA", binwidth=
p11
```



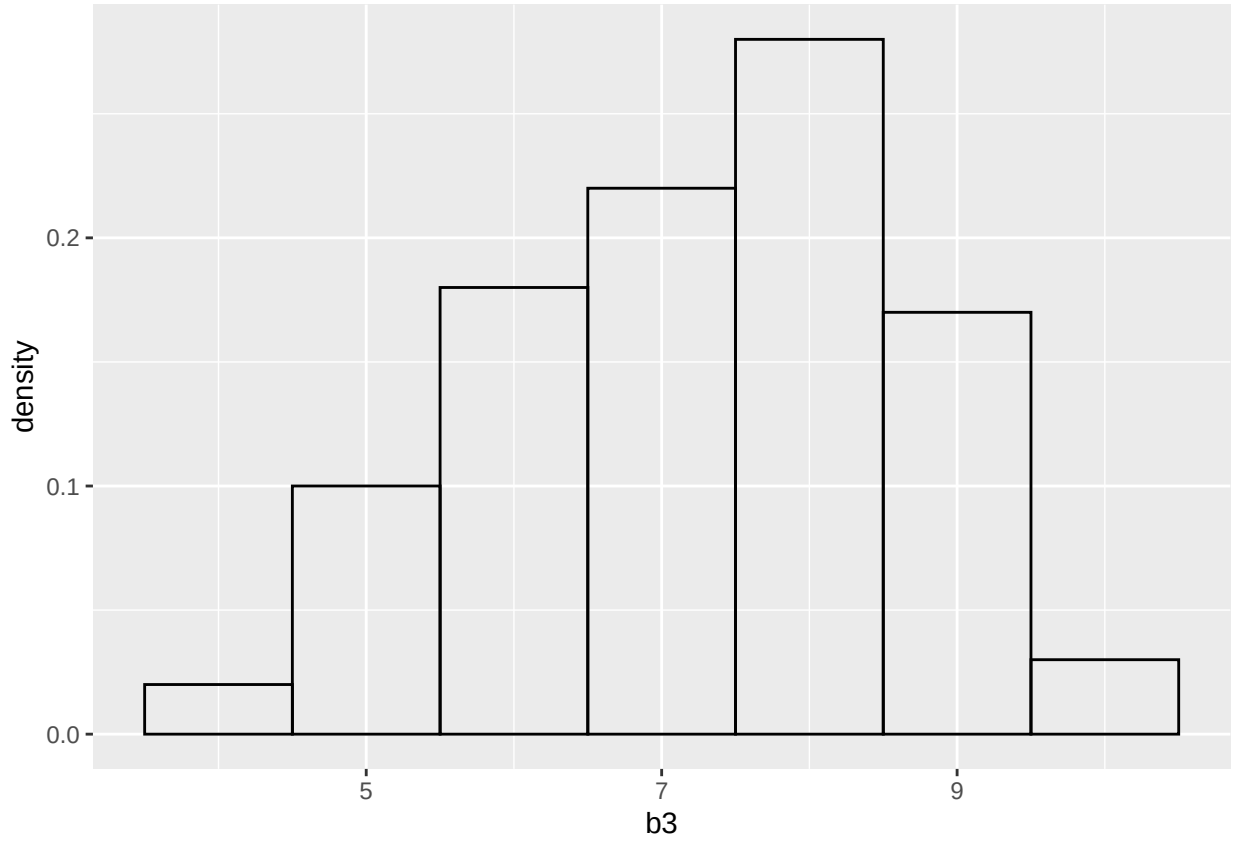

```
p21
```




```
b3 = rbinom(100,10,0.75)
df1=data.frame(b3)
p11= ggplot(df1) + geom_histogram(mapping=aes(x=b3), color="black", fill="NA", binwidth=1)
p21= ggplot(df1) + geom_histogram(mapping=aes(x=b3, y=..density..), color="black", fill="NA", binwidth=
p11
```




```
p21
```

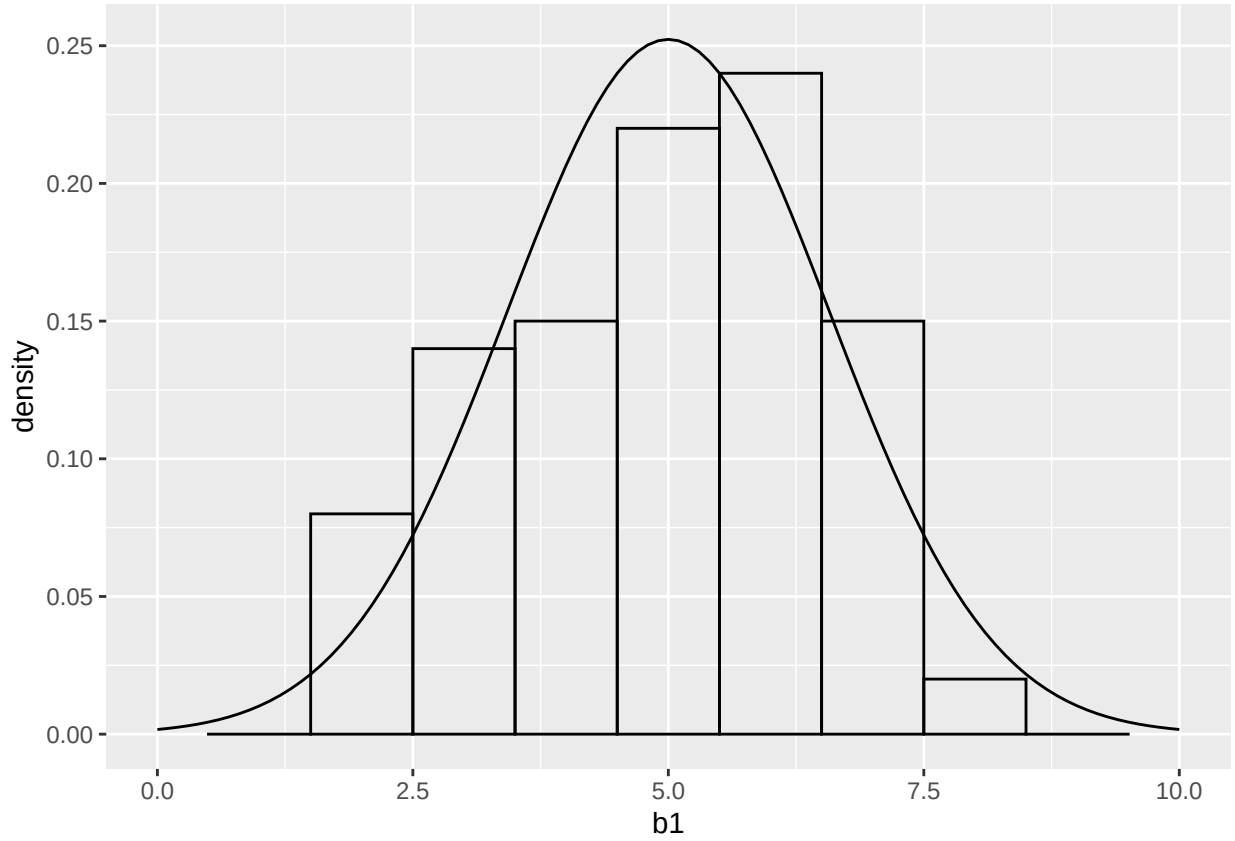


p11 creates a histogram of the binomial distribution with total counts on the y axis & p21 creates a histogram of binomial distribution with pdf in the y-axis. The three different plots show how the distribution changes when the probability of success is different. b1 is close to normal distribution,b2 is positively skewed and b3 is negatively skewed.

Q4

```
density = function(x,a,s){ (1/((2*pi)^(0.5)*s ))* exp(-(x-a)^2/(2*s^2))}
df1=data.frame(b1)
p21= ggplot(df1) +geom_histogram(mapping=aes(x=b1, y=..density..), color="black", fill="NA", binwidth=1
p21
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



a) Since, area of histograms wud be area under curve, the value will be close to 0.6 looking at the graph.

b) a & b will be approx -1.25 and 1.25 respectively.

c) Since, b2 and b3 are skewed and are not norma;, we wont get same results.

Q5

```
x = c(1,2,3,4,5,6)
probx= c(1/6,1/6,1/6,1/6,1/6,1/6)
Rolls=sample(x, size=1500, replace=T, prob=probx)
Rollm=matrix(Rolls, nrow = 5)
Rollsums = apply(Rollm, 2, sum)
Rollm
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    5    6    4    6    1    2    2    4    3     2     6     1     2     1
## [2,]    6    3    1    2    5    6    2    1    1     6     6     2     5     6
## [3,]    2    5    5    5    5    4    6    2    3     1     5     6     2     6
## [4,]    6    6    3    1    2    3    5    4    6     6     5     3     5     4
## [5,]    2    3    5    5    6    1    5    5    2     2     3     3     5     6
##      [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
## [1,]     5     6     5     5     4     1     3     6     3     4     2     6
## [2,]     4     1     1     2     5     6     2     5     3     1     1     3
## [3,]     1     3     1     6     4     4     4     1     1     6     5     3
## [4,]     2     4     1     4     6     2     2     4     1     5     1     1
```

```
## [5,]    2    1    4    1    4    2    6    4    1    2    2    3
##      [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38]
## [1,]    4    3    2    5    6    5    2    2    5    2    4    5
## [2,]    2    3    6    4    3    4    2    1    3    3    3    1
## [3,]    2    1    6    6    5    3    4    6    1    1    5    6
## [4,]    4    5    4    5    2    1    6    4    1    4    3    1
## [5,]    3    6    5    1    2    3    2    6    6    3    4    5
##      [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49] [,50]
## [1,]    5    3    3    1    2    2    6    3    3    5    6    1
## [2,]    6    6    5    2    4    3    2    3    6    5    5    3
## [3,]    2    3    1    2    3    4    2    2    5    3    2    2
## [4,]    3    6    3    2    5    1    3    6    5    6    2    4
## [5,]    5    5    1    2    2    4    4    5    3    1    1    4
##      [,51] [,52] [,53] [,54] [,55] [,56] [,57] [,58] [,59] [,60] [,61] [,62]
## [1,]    2    2    3    2    1    1    5    2    1    3    5    4
## [2,]    1    2    1    2    4    6    1    1    4    2    4    5
## [3,]    5    4    4    6    6    4    1    1    2    2    2    4
## [4,]    3    5    2    4    5    4    4    2    6    3    5    1
## [5,]    4    4    5    6    1    3    5    6    6    1    5    1
##      [,63] [,64] [,65] [,66] [,67] [,68] [,69] [,70] [,71] [,72] [,73] [,74]
## [1,]    5    4    1    3    5    3    1    1    5    1    3    5
## [2,]    4    1    5    1    3    5    3    2    6    6    4    2
## [3,]    4    6    5    5    3    1    4    1    2    6    5    2
## [4,]    5    1    1    2    3    2    4    5    5    1    6    1
## [5,]    4    3    2    2    3    2    3    2    2    2    3    3
##      [,75] [,76] [,77] [,78] [,79] [,80] [,81] [,82] [,83] [,84] [,85] [,86]
## [1,]    1    4    4    3    3    2    6    2    3    2    2    1
## [2,]    5    5    3    3    5    3    2    5    2    6    4    2
## [3,]    1    6    3    5    3    2    6    4    3    5    1    2
## [4,]    2    2    3    6    5    6    2    5    1    4    5    2
## [5,]    2    6    6    4    4    4    5    1    5    1    3    6
##      [,87] [,88] [,89] [,90] [,91] [,92] [,93] [,94] [,95] [,96] [,97] [,98]
## [1,]    6    6    1    2    5    2    2    2    1    1    2    6
## [2,]    1    5    1    2    1    5    1    2    3    2    4    6
## [3,]    4    3    2    6    2    1    3    1    2    6    6    6
## [4,]    6    2    2    5    5    3    5    6    1    6    3    4
## [5,]    5    4    1    1    2    5    5    2    1    2    1    4
##      [,99] [,100] [,101] [,102] [,103] [,104] [,105] [,106] [,107] [,108]
## [1,]    1      3      2      2      4      1      2      5      3      1
## [2,]    1      3      3      5      3      1      6      3      5      6
## [3,]    1      2      2      2      2      3      4      2      5      1
## [4,]    2      5      3      3      4      6      5      2      6      5
## [5,]    2      2      3      6      1      3      3      3      2      6
##      [,109] [,110] [,111] [,112] [,113] [,114] [,115] [,116] [,117] [,118]
## [1,]      6      3      4      4      4      6      2      3      2      1
## [2,]      4      2      3      1      1      4      4      6      1      6
## [3,]      3      5      4      4      1      6      5      2      4      6
## [4,]      5      1      1      4      4      1      2      3      4      3
## [5,]      6      1      4      4      2      1      1      2      5      4
##      [,119] [,120] [,121] [,122] [,123] [,124] [,125] [,126] [,127] [,128]
## [1,]      4      4      3      5      6      2      6      3      5      2
## [2,]      3      4      2      5      2      5      5      6      1      4
## [3,]      6      5      3      3      1      4      1      5      2      5
## [4,]      3      6      2      6      5      4      4      2      1      5
```

```
## [5,]      1      4      3      6      5      2      6      3      3      4
##      [,129] [,130] [,131] [,132] [,133] [,134] [,135] [,136] [,137] [,138]
## [1,]      6      5      5      2      1      5      5      6      6      5
## [2,]      2      5      6      6      2      1      5      6      3      1
## [3,]      5      6      6      5      5      1      2      1      1      5
## [4,]      1      5      1      5      5      6      5      3      6      1
## [5,]      2      2      4      1      2      2      5      3      6      4
##      [,139] [,140] [,141] [,142] [,143] [,144] [,145] [,146] [,147] [,148]
## [1,]      2      6      4      5      6      1      2      2      2      2
## [2,]      3      6      2      2      2      4      1      4      6      5
## [3,]      6      5      1      3      4      6      5      5      6      2
## [4,]      6      4      5      3      3      5      3      6      3      5
## [5,]      2      2      4      2      6      3      3      3      4      6
##      [,149] [,150] [,151] [,152] [,153] [,154] [,155] [,156] [,157] [,158]
## [1,]      5      1      6      5      3      1      1      6      1      5
## [2,]      4      1      3      2      1      1      4      4      5      5
## [3,]      2      3      4      1      6      1      5      5      6      6
## [4,]      3      1      4      6      1      5      6      4      4      4
## [5,]      3      4      5      5      2      1      4      3      6      2
##      [,159] [,160] [,161] [,162] [,163] [,164] [,165] [,166] [,167] [,168]
## [1,]      3      6      3      3      2      3      4      4      3      3
## [2,]      6      2      1      4      6      6      3      6      2      1
## [3,]      5      5      5      3      2      4      5      4      4      4
## [4,]      5      3      6      5      6      2      3      2      5      4
## [5,]      4      6      2      5      2      4      1      3      5      3
##      [,169] [,170] [,171] [,172] [,173] [,174] [,175] [,176] [,177] [,178]
## [1,]      3      5      5      6      2      2      5      2      5      3
## [2,]      5      6      6      6      6      4      6      4      6      5
## [3,]      6      5      1      3      3      3      6      5      1      2
## [4,]      1      5      6      4      3      1      2      4      3      6
## [5,]      2      4      6      1      3      6      2      3      6      3
##      [,179] [,180] [,181] [,182] [,183] [,184] [,185] [,186] [,187] [,188]
## [1,]      2      2      5      1      4      4      2      1      3      2
## [2,]      4      4      1      3      2      3      3      5      4      2
## [3,]      1      4      6      1      5      6      1      2      5      4
## [4,]      2      5      2      3      1      5      2      6      5      3
## [5,]      4      6      5      5      5      1      4      4      4      4
##      [,189] [,190] [,191] [,192] [,193] [,194] [,195] [,196] [,197] [,198]
## [1,]      6      3      4      1      1      4      4      5      1      4
## [2,]      5      4      4      6      1      5      1      4      4      3
## [3,]      3      6      2      6      2      2      1      4      4      1
## [4,]      2      6      2      3      6      4      6      2      5      2
## [5,]      1      4      1      3      3      4      5      1      1      3
##      [,199] [,200] [,201] [,202] [,203] [,204] [,205] [,206] [,207] [,208]
## [1,]      2      4      5      2      2      5      6      5      3      3
## [2,]      4      4      5      1      4      6      2      6      1      6
## [3,]      5      3      5      1      4      3      6      2      3      4
## [4,]      6      2      2      3      1      5      6      1      5      5
## [5,]      6      1      1      2      5      5      3      1      4      6
##      [,209] [,210] [,211] [,212] [,213] [,214] [,215] [,216] [,217] [,218]
## [1,]      3      2      2      6      4      4      6      1      1      4
## [2,]      5      6      6      5      6      3      4      3      3      3
## [3,]      5      4      6      1      3      6      4      2      4      3
## [4,]      2      2      1      6      5      5      5      4      2      6
```

```
## [5,]      5      5      1      2      4      6      4      5      1      2
##      [,219] [,220] [,221] [,222] [,223] [,224] [,225] [,226] [,227] [,228]
## [1,]      5      2      2      3      3      1      6      3      2      2
## [2,]      5      6      3      5      4      3      5      5      4      5
## [3,]      3      1      1      3      4      4      6      2      4      1
## [4,]      3      1      1      5      6      6      3      4      4      6
## [5,]      1      6      5      1      6      1      2      3      3      5
##      [,229] [,230] [,231] [,232] [,233] [,234] [,235] [,236] [,237] [,238]
## [1,]      3      5      3      2      1      6      2      5      6      3
## [2,]      3      3      5      6      4      3      4      5      5      2
## [3,]      2      2      1      5      3      6      5      3      6      2
## [4,]      2      4      1      1      3      6      5      1      1      6
## [5,]      1      2      3      2      2      2      1      1      6      1
##      [,239] [,240] [,241] [,242] [,243] [,244] [,245] [,246] [,247] [,248]
## [1,]      5      5      6      2      3      6      4      3      2      1
## [2,]      6      1      5      1      6      5      2      2      3      3
## [3,]      4      5      3      4      1      3      4      4      5      4
## [4,]      1      5      1      4      6      6      1      1      1      4
## [5,]      5      6      4      1      2      6      5      4      6      6
##      [,249] [,250] [,251] [,252] [,253] [,254] [,255] [,256] [,257] [,258]
## [1,]      5      3      4      5      2      1      2      4      1      4
## [2,]      4      6      6      2      5      3      2      3      2      5
## [3,]      5      6      2      2      6      1      5      5      2      4
## [4,]      5      4      5      4      6      4      2      5      5      5
## [5,]      2      5      2      3      6      3      5      3      3      1
##      [,259] [,260] [,261] [,262] [,263] [,264] [,265] [,266] [,267] [,268]
## [1,]      6      4      1      6      6      4      2      3      1      3
## [2,]      4      3      1      3      4      4      3      5      2      4
## [3,]      5      3      2      3      5      4      5      1      2      4
## [4,]      6      4      6      1      2      1      5      3      4      4
## [5,]      4      5      3      1      2      4      1      4      3      3
##      [,269] [,270] [,271] [,272] [,273] [,274] [,275] [,276] [,277] [,278]
## [1,]      6      2      2      5      2      4      1      6      4      6
## [2,]      2      2      6      4      2      3      2      3      1      5
## [3,]      4      4      2      2      3      6      6      1      3      5
## [4,]      6      1      3      4      3      1      6      2      1      1
## [5,]      5      6      3      5      1      5      5      4      1      2
##      [,279] [,280] [,281] [,282] [,283] [,284] [,285] [,286] [,287] [,288]
## [1,]      6      3      4      6      5      4      5      6      6      4
## [2,]      6      5      2      3      2      4      2      3      2      2
## [3,]      5      2      2      3      4      3      4      6      6      2
## [4,]      6      4      2      5      5      1      3      6      3      1
## [5,]      4      3      4      4      2      4      1      6      1      4
##      [,289] [,290] [,291] [,292] [,293] [,294] [,295] [,296] [,297] [,298]
## [1,]      4      3      4      6      5      2      1      1      6      2
## [2,]      1      6      3      1      4      5      6      1      5      5
## [3,]      2      2      6      2      5      2      3      5      6      2
## [4,]      3      1      3      4      6      5      5      3      1      2
## [5,]      6      3      1      3      3      1      1      6      1      2
##      [,299] [,300]
## [1,]      4      1
## [2,]      3      5
## [3,]      4      2
## [4,]      2      3
```

```
## [5,]      2      1
```

```
Rollsums
```

```
##   [1] 21 23 18 19 19 16 20 16 15 17 25 15 19 23 14 15 12 18 23 15 17 20  9 18 11
##  [26] 16 15 18 23 21 18 16 16 19 16 13 19 18 21 23 13  9 16 14 17 19 22 20 16 14
##  [51] 15 17 15 20 17 18 16 12 19 11 21 15 22 15 14 13 17 13 15 11 20 16 21 13 11
##  [76] 23 19 21 20 17 21 17 14 18 15 13 22 20  7 16 15 16 16 13  8 17 16 26  7 15
## [101] 13 18 14 14 20 15 21 19 24 12 16 17 12 18 14 16 16 20 17 23 13 25 19 17 22
## [126] 19 12 20 16 23 22 19 15 15 22 19 22 16 19 23 16 15 21 19 14 20 21 20 17 10
## [151] 22 19 13  9 20 22 22 22 23 22 17 20 18 19 16 19 19 15 17 25 24 20 17 16 21
## [176] 18 21 19 13 21 19 13 17 19 12 18 21 15 17 23 13 19 13 19 17 16 15 13 23 14
## [201] 18  9 16 24 23 15 16 24 20 19 16 20 22 24 23 15 11 18 17 16 12 17 23 15 22
## [226] 17 17 19 11 16 13 16 13 23 17 15 24 14 21 22 19 12 18 26 16 14 17 18 21 24
## [251] 19 16 25 12 16 20 13 19 25 19 13 14 19 17 16 16 12 18 23 15 16 20 11 19 20
## [276] 16 10 19 27 17 14 21 18 16 15 27 18 13 16 15 17 16 23 15 16 16 19 13 15 12
```

```
mean(Rollsums)
```

```
## [1] 17.43
```

```
var(Rollsums)
```

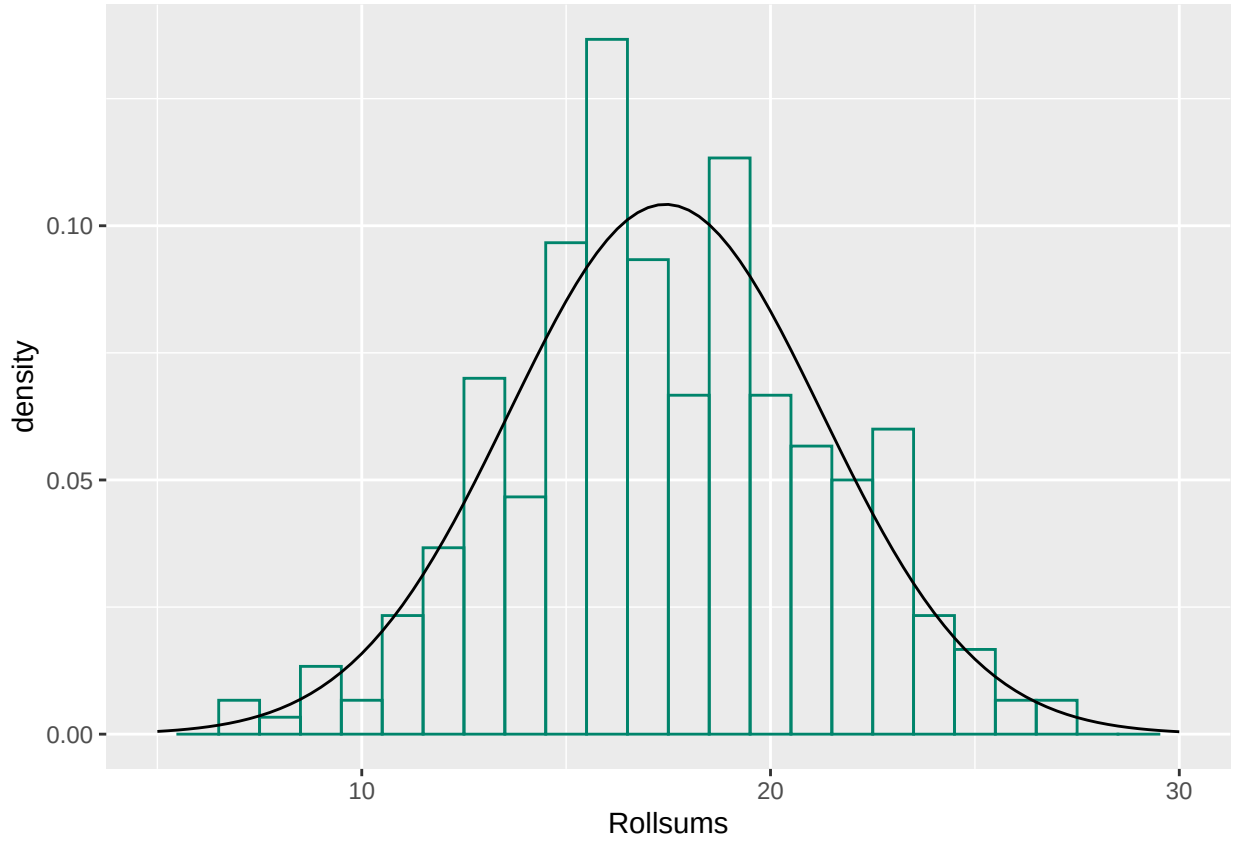
```
## [1] 14.65395
```

matrix function creates matrix with outcomes of die as matrix elements and apply function creates a vector which stores sum of each column.

Q6

```
library(ggplot2)
density = function(x,a,s){ (1/((2*pi)^(0.5)*s ))* exp(-(x-a)^2/(2*s^2))}
dfrolls = data.frame(Rollsums)
mu = mean(dfrolls$Rollsums)
sigma= sd(dfrolls$Rollsums)
ggplot(data=dfrolls) + geom_histogram(mapping=aes(x=Rollsums,y=..density..), color="#00846b", fill=NA,
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



From the graph,value will be approx 0.75.

Approx values would be 1 and -1.

Due date: October 22nd, 2021
Problems Due: 1,5

From Probabiliy and Statistics with Examples using R.

1. Do Worksheet Problem 3
2. Ex 5.2.10 (c)
3. Ex 5.2.11
4. Example 5.2.11
5. Example 5.2.12

HOMEWORK 4

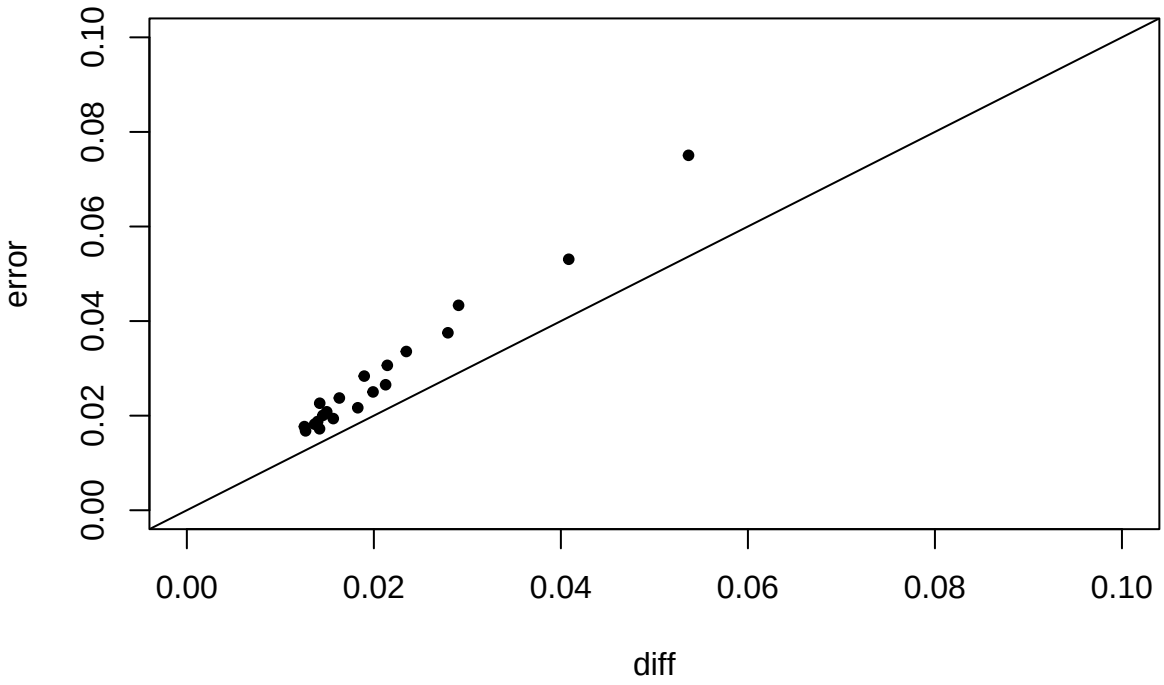
Probability and Statistics with R

Name - Aditya Anandkumar

Roll No. - MDS202102

Q1

```
m=seq(0,1000,50)
m[1]=1
x=seq(-2,2,.1)
z=pnorm(x)
y=NULL
diff=NULL
error=NULL
p=0.6
for(i in m)
{
  for(k in x)
  {
    a=NULL
    for (j in 1:100)
    {
      B=rbinom(1000,i,p)
      SB=(B-i*p)/((i*p*(1-p))^.5)
      a=c(a,length(SB[SB<=k])/length(SB))
    }
    y=c(y,mean(a))
  }
  diff=c(diff,max(abs(y-z)))
  error=c(error,(p^2+(1-p)^2)/(2*(i*p*(1-p))^0.5))
  y=NULL
}

plot(diff,error,pch=20,xlim=c(0,0.1),ylim=c(0,0.1))
lines(x,x)
```



Clearly, the points lie above x=y line which indicates that error values are greater than the diff values. Hence, Berry-Eseen Type bound is proved.

To improve this, we can store all values of diff which are greater than its corresponding error value in a vector. If length of the vector is zero, the bound is proved.

```
a=NULL

for (i in 1:length(m))
{

  if (diff[i]>error[i])
  {
    a=c(a,diff[i])
  }

}

length_a=length(a)

length_a
```

```
## [1] 0
```

CLearly, therte are no elements in diff which are greater than error.

Thus, the bound is proved.

Q5

```
mean=200
sd=4


#For Standardizing

X = (195-mean)/sd
X
```

```
## [1] -1.25
```

We need to find P(Z < -1.25).

```
probabilty = pnorm(-1.25)
probabilty
```

```
## [1] 0.1056498
```

Hence, there's a 10.56% chance of producing a bag which weighs less than 195g.

1. De Moivre's Central Limit Theorem.

   (a) Using the `rbinom` generate 100 samples of Binomial$(20, 0.5)$ and plot the histogram of the data-set.

   (b) Using the `rnorm` generate 100 samples of Normal(10, 5) and plot the histogram of the data set.

   Think of ways you can enhance the above exercise to come up with a computer proof of the Central Limit Theorem.

2. Poisson Approximation

   (a) Using the `rbinom` generate 100 samples of Binomial(2000,0.001), save it in a dataframe `dfbinomial` and plot the histogram of the data-set.

   (b) Using the `rpois` generate 100 samples of Poisson(2), save it in a dataframe `dfnormal` and plot the histogram of the data set

   Think of ways you can enhance the above exercise to come up with a computer proof of the Poisson Approximation, even though we have seen a proof in class.

3. The following result is a Berry-Eseen Type bound.

   Theorem: *Let $X_n \sim$ Binomial(n,p), then there exists $C > 0$ such that*

$$\sup_{x \in R} \left| P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq x \right) - \int_{\infty}^{x} \frac{\exp(-\frac{y^2}{2})}{\sqrt{2\pi}} dt \right| \leq \frac{(p^2 + (1-p)^2)}{2\sqrt{np(1-p)}}$$

   We shall prove it by simulation by the below algorithm.

```
For x = -2,-1.9,-1,8,...0,...,1.9,2
    using inbuilt pnorm find z[x]:- pnorm(x)
Set p
For m = 1,50,100,150,...,1000
    For x = -2,-1.9,-1,8,...0,...,1.9,2
      1)Generate B: 1000 Samples of Binomial (m,p) using inbuilt rbinom function
        Compute SB:  (B-m*p)/((m*p(1-p))^(0.5))
      2)Compute y[x] : the proportion of samples  in SB less than equal to x
      3)Repeat steps 1) and 2) 100 times and compute average --  my[x] over each trial.
      4)Calculate   diff[m]=  max(abs(my[x]- z[x]))

For m = 1,50,100,150,...,1000
    Calculate  error(m)= [p^2+(1-p)^2]/[2*(m*p*(1-p))^0.5]

Plot diff and error.
```

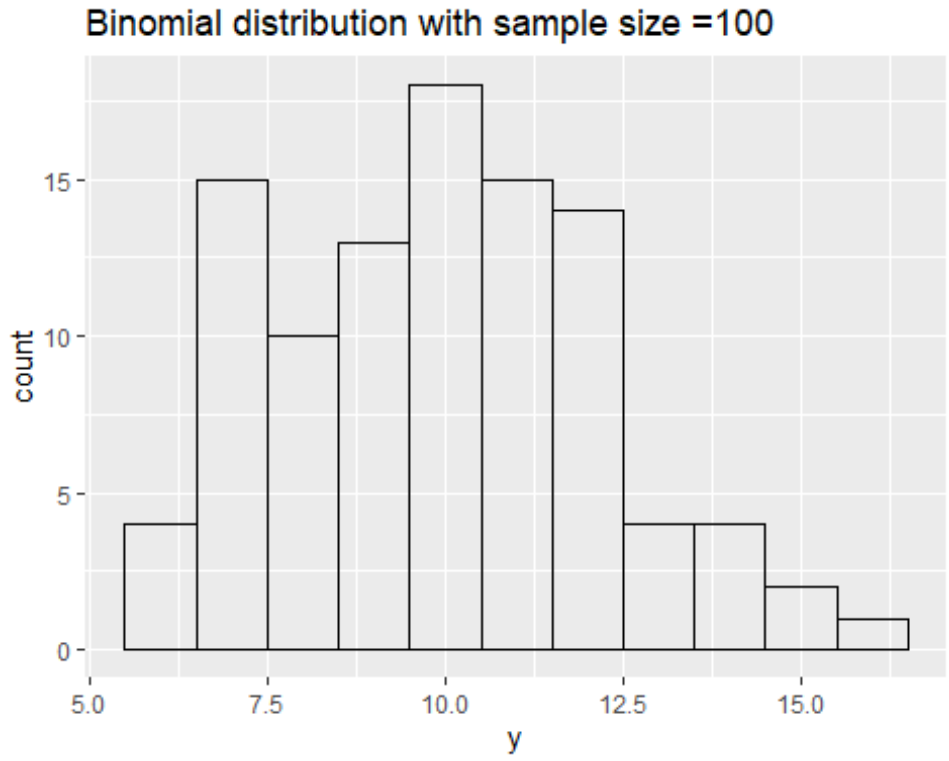
   See if result is verified by picture. Can you do anything additional to verify the Theorem ?

Worksheet 4

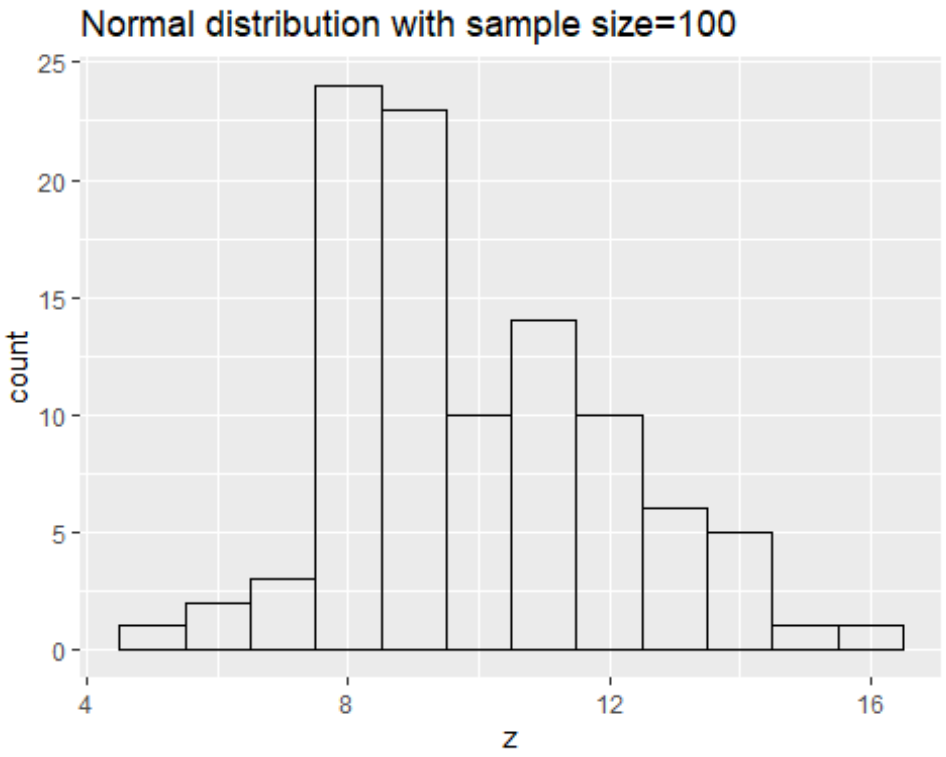
Rishika Tibrewal

15/10/2021

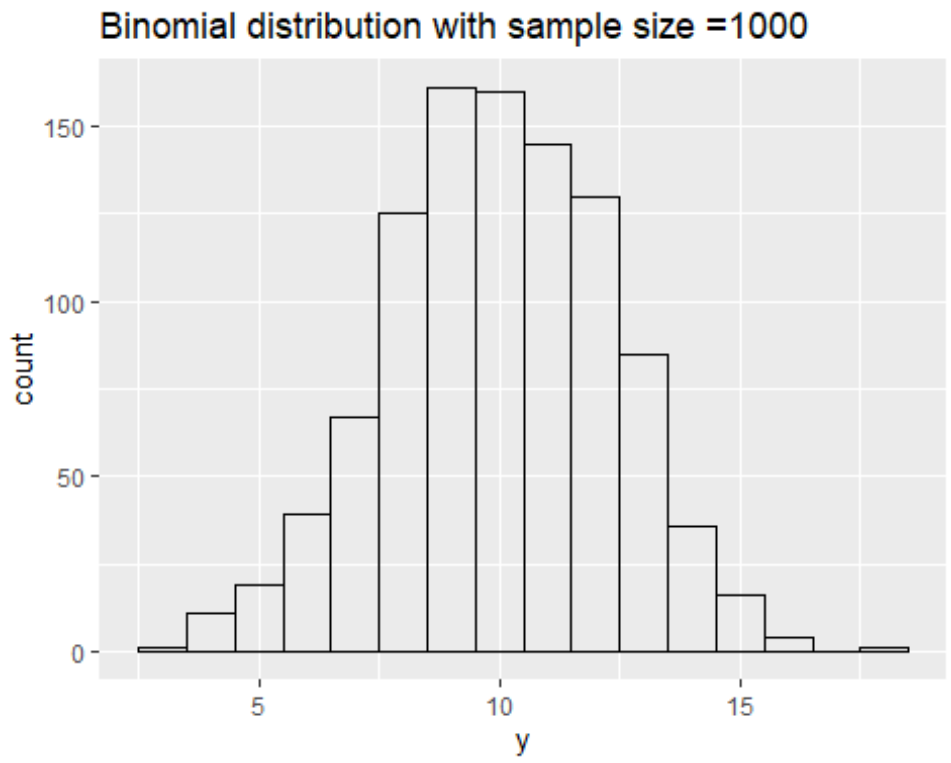
```
library(ggplot2)

#1)a)
y=rbinom(100,20,0.5)
df1= data.frame(y)
ggplot(df1)+geom_histogram(mapping = aes(x=y), color='black', fill='NA',
binwidth = 1)+ ggtitle("Binomial distribution with sample size =100")
```



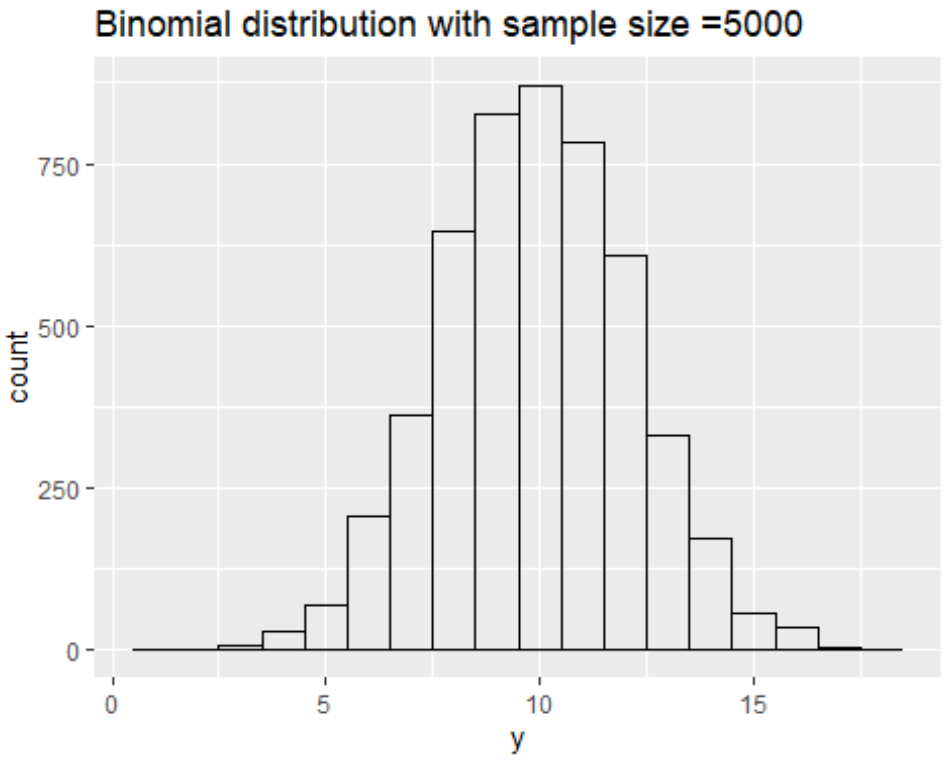

```
z=rnorm(100, mean=10, sd=5**(0.5))
df2= data.frame(z)
ggplot(df2)+geom_histogram(mapping = aes(x=z), color='black', fill='NA',
binwidth = 1)+ ggtitle("Normal distribution with sample size=100")
```



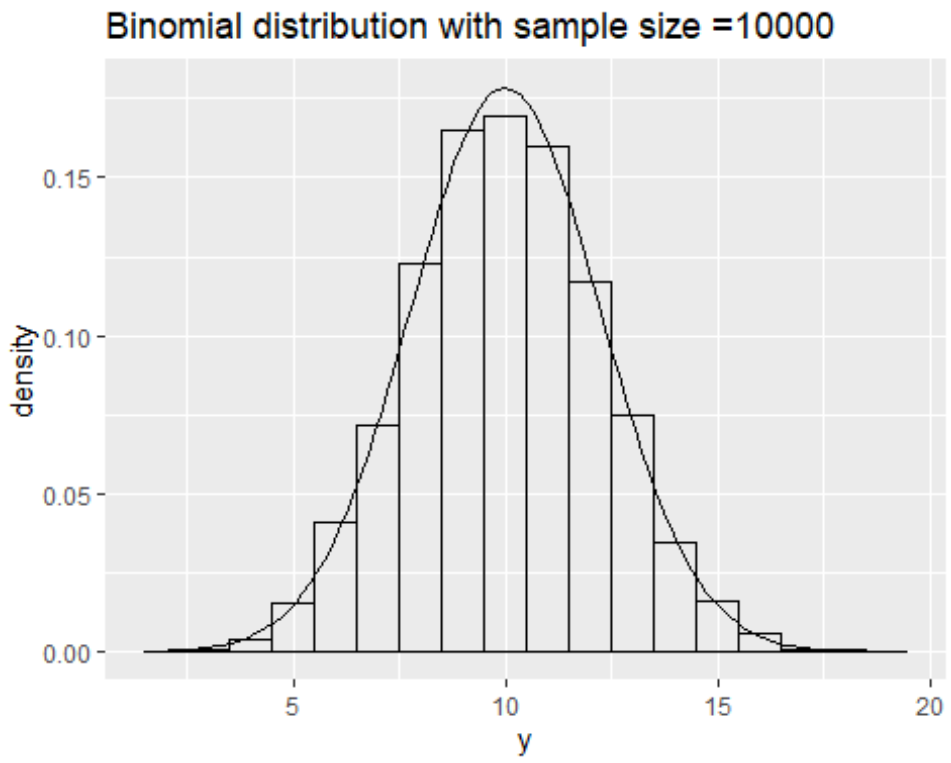

```r
#1)b)
y=rbinom(1000,20,0.5)
df3= data.frame(y)
ggplot(df3)+geom_histogram(mapping = aes(x=y), color='black', fill='NA',
binwidth = 1)+ ggtitle("Binomial distribution with sample size =1000")
```



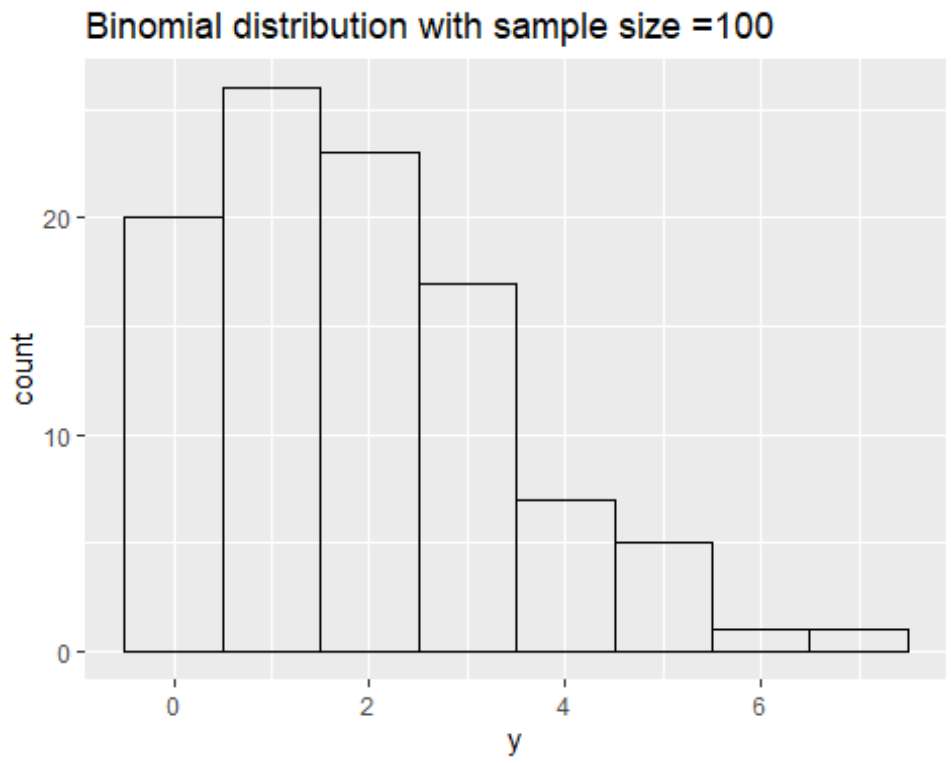

```r
y=rbinom(5000,20,0.5)
df4= data.frame(y)
ggplot(df4)+geom_histogram(mapping = aes(x=y), color='black', fill='NA',
binwidth = 1)+ ggtitle("Binomial distribution with sample size =5000")
```



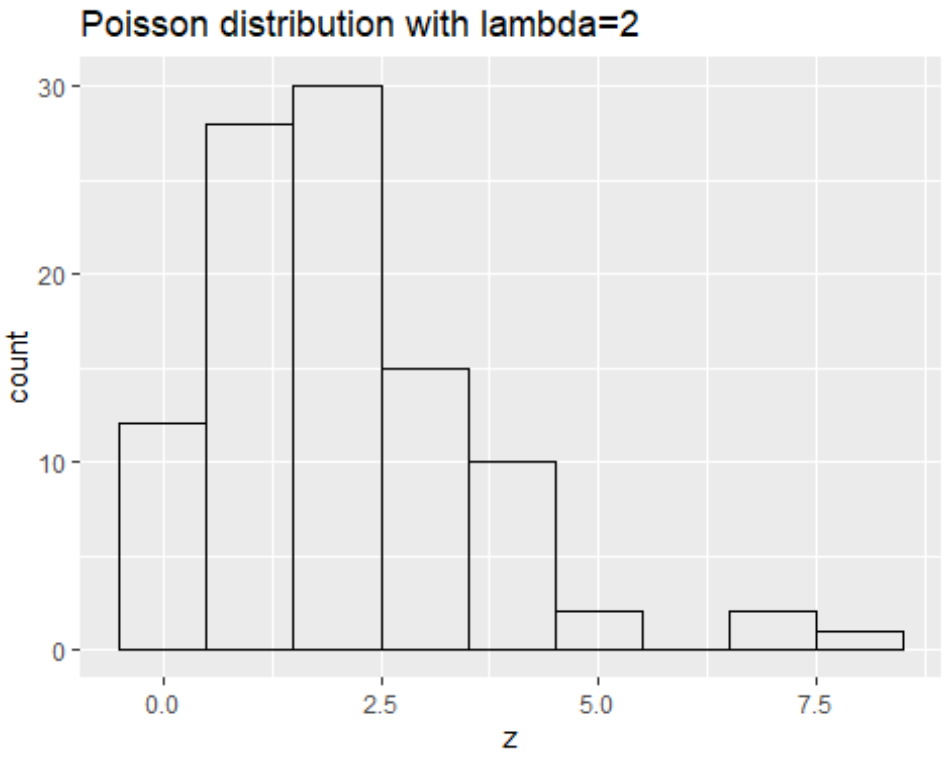

```
y=rbinom(10000,20,0.5)
df5= data.frame(y)
pdfnormal = function(x,a,s){ (1/((2*pi)^(0.5)*s ))* exp(-(x-a)^2/(2*s^2))}
ggplot(df5)+geom_histogram(mapping = aes(x=y, y=..density..), color='black',
fill='NA', binwidth = 1)+ ggtitle("Binomial distribution with sample size
=10000")+geom_function(fun=pdfnormal, args=list(a=10,s=(5)^(0.5)))
```



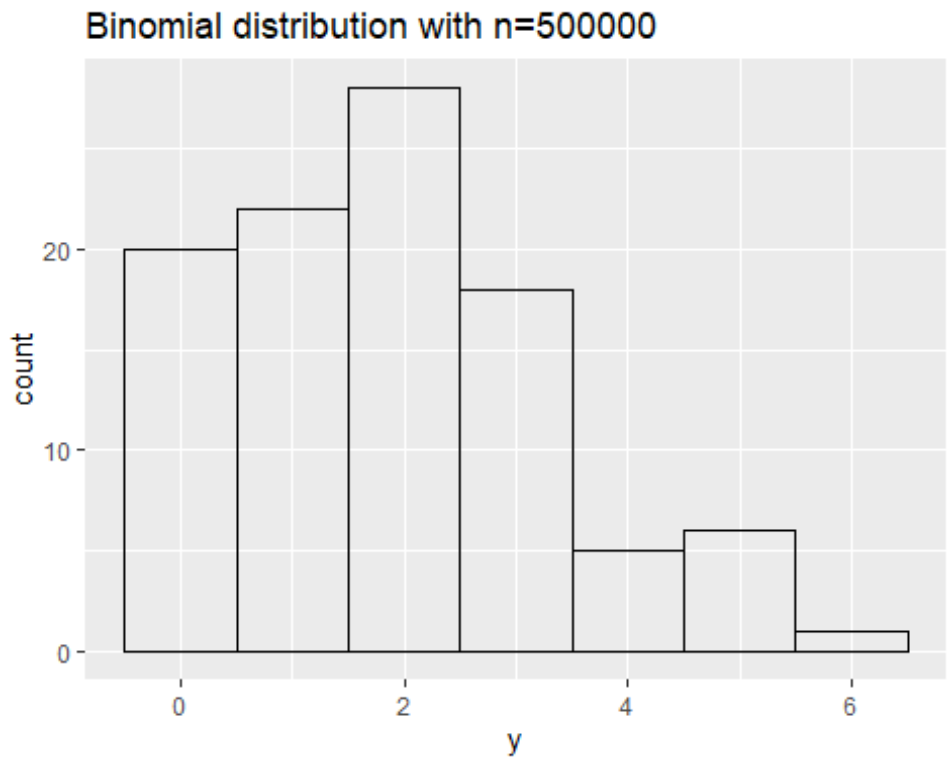

```
#From the above graphs we see that as sample size of the binomial
distribution increases, the graphs start looking similar to the Normal
distribution. Hence it is true by CLT that, as n tends to infinity, Binomial
distribution can be well approximated by a Normal Distribution.

#2)a)
y=rbinom(100,2000,0.001)
dfbinomial=data.frame(y)
ggplot(dfbinomial) + geom_histogram(mapping=aes(x=y), color="black",
fill="NA", binwidth=1)+ ggtitle("Binomial distribution with sample size
=100")
```



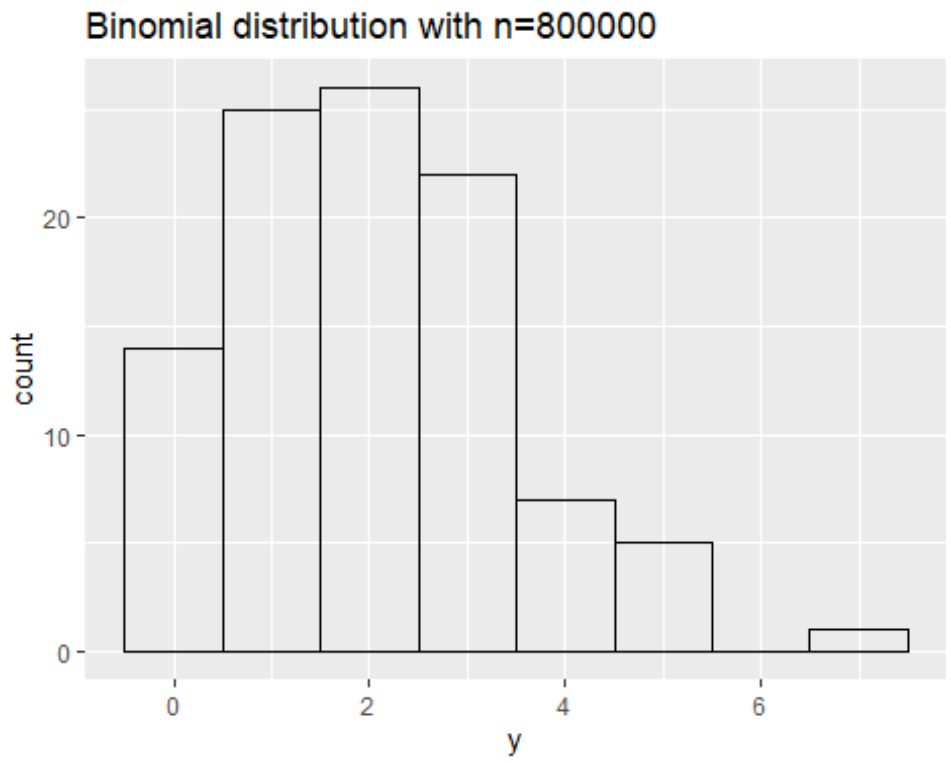

```
z=rpois(100,2)
dfnormal=data.frame(z)
ggplot(dfnormal) + geom_histogram(mapping=aes(x=z), color="black", fill="NA",
binwidth=1)+ ggtitle("Poisson distribution with lambda=2")
```



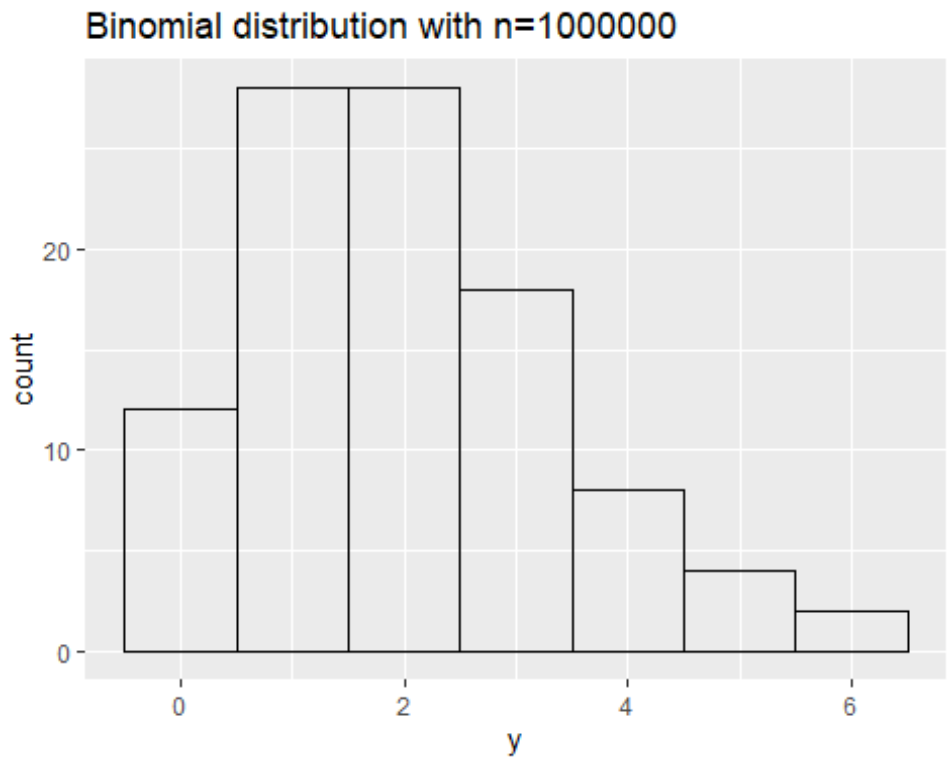

```r
#2)b)
y=rbinom(100,500000,0.000004)
dfbinomial=data.frame(y)
ggplot(dfbinomial) + geom_histogram(mapping=aes(x=y), color="black",
fill="NA", binwidth=1)+ ggtitle("Binomial distribution with n=500000")
```



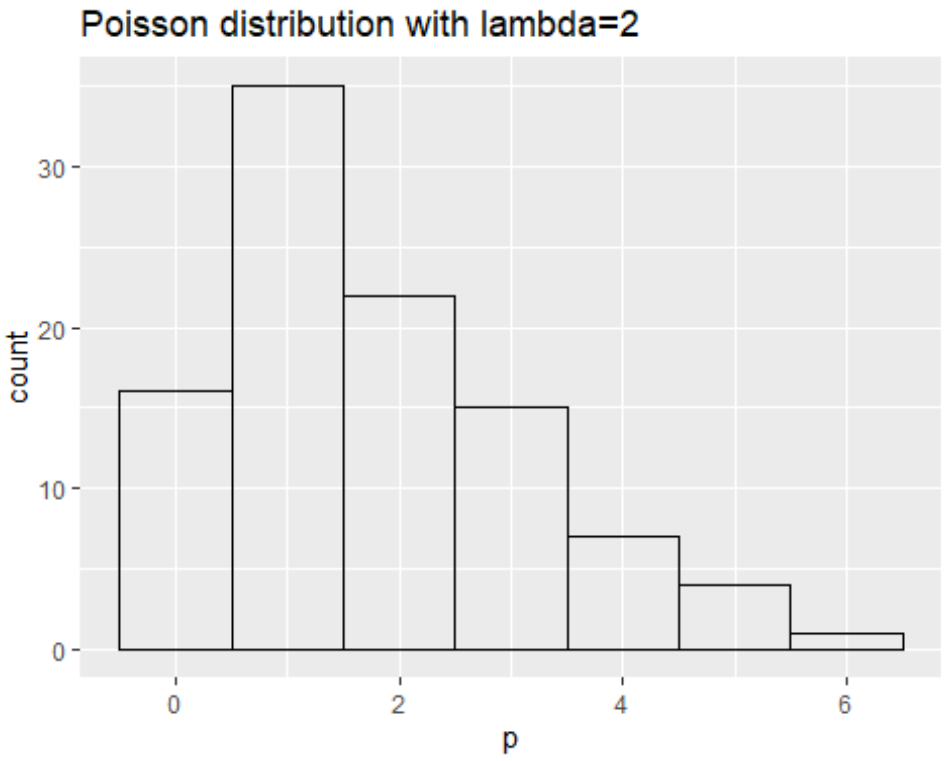

```
y=rbinom(100,800000,0.0000025)
dfbinomial=data.frame(y)
ggplot(dfbinomial) + geom_histogram(mapping=aes(x=y), color="black",
fill="NA", binwidth=1)+ ggtitle("Binomial distribution with n=800000")
```




```
y=rbinom(100,1000000,0.000002)
dfbinomial=data.frame(y)
ggplot(dfbinomial) + geom_histogram(mapping=aes(x=y), color="black",
fill="NA", binwidth=1)+ggtitle("Binomial distribution with n=1000000")
```




```
p= rpois(100,2)
df= data.frame(p)
ggplot(df)+geom_histogram(mapping = aes(x=p), color='black', fill='NA',
binwidth = 1)+ ggtitle("Poisson distribution with lambda=2")
```




```
#From the above graphs we see that as n increases and p decreases keeping the
mean constant, the graphs start looking similar to the Poisson distribution.
Hence it is true by CLT that, for  a large n and small p, keeping np
constant, Binomial distribution can be well approximated by a Poisson
Distribution.
```

Due date: October 29th, 2021
Problems Due: 2,4,6

From Probabiliy and Statistics with Examples using R.

1. Exercise 5.1.5
2. Exercise 5.1.7
3. Exercise 5.1.8
4. Exercise 5.2.3
5. Exercise 5.2.6
6. Exercise 5.2.7
7. Exercise 5.2.10

Homework - 5

2) $f: \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1/x^2 & ; \ 1 < x \\ 0 & ; \text{o.w} \end{cases}$$

(a) We know, for f to be a probability density function

i) $f(x) \geq 0$

ii) f is piecewise continuous

iii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

i) According to the definition of the function,

$f(x) = \frac{1}{x^2}$ for $x > 1$ So, $f(x) > 0 \ \forall \ x > 1$

Also, for other values, i.e., for $x \leq 1$, $f(x) = 0 \geq 0$.

$\therefore f(x) \geq 0$.

ii) ~~Let (x_n) be a sequence in $\mathbb{R}$ which converges to c~~ .

$\forall \ x > 1$, $f(x) = \frac{1}{x^2}$ is continuous

Also, $\forall \ x \leq 1$ $f(x) = 0$ is continuous (being a constant function)

$\therefore f(x)$ is continuous over $(-\infty, 1]$ and $(1, \infty)$ and hence piecewise continuous.

iii) $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{1} 0 \cdot dx + \int_{1}^{\infty} \frac{1}{x^2}\,dx = 0 + \left[\frac{-1}{x}\right]_1^{\infty}$

$= 0 + 1 = 1$.

$\therefore f$ is a probability density function .

(b) Let $a > 1$.

$$P((a,\infty)) = \int_a^{\infty} f(x)\,dx = \int_a^{\infty} \frac{1}{x^2}\,dx = \left[\frac{-1}{x}\right]_a^{\infty} = 0 + \frac{1}{a}.$$

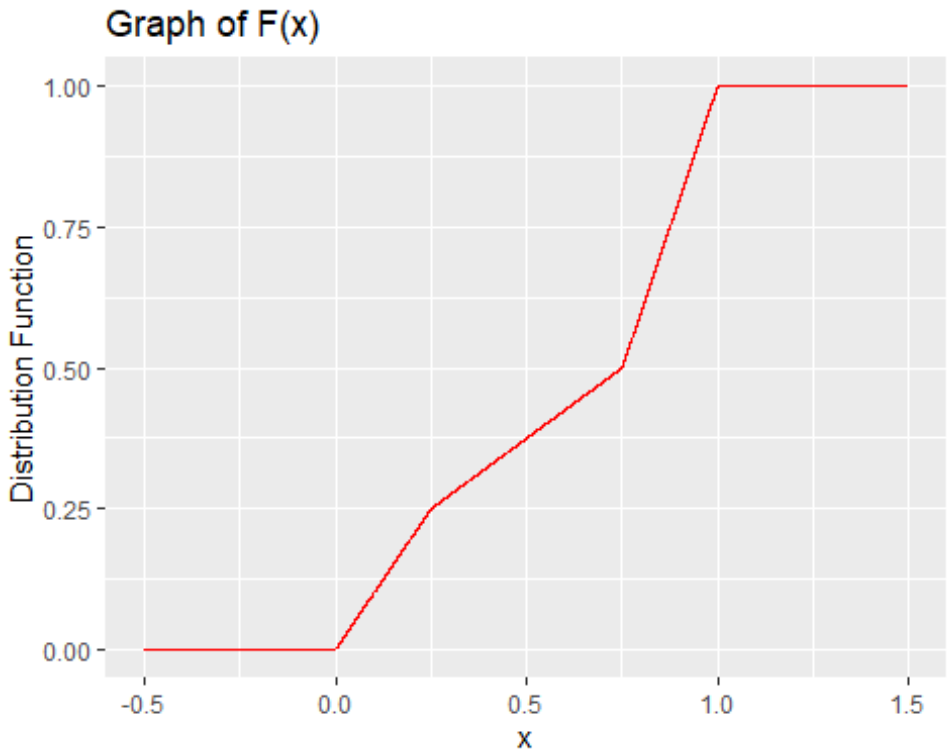
$\Rightarrow P((a,\infty)) = 1/a$

Assignment 5

Rishika Tibrewal

24/10/2021

```
# 4 a)

F=function(x)
{
  if (x<0)
    return(0)
  else if (0<=x & x<(1/4))
    return(x)
  else if ((1/4)<=x & x<(3/4))
    return((x/2)+(1/8))
  else if ((3/4)<=x & x<1)
    return((2*x)-1)
  else
    return(1)
}

y=seq(-0.5,1.5,0.0001)
F_y=rep(NA,length(y))
for(i in 1:length(y))
{
  F_y[i]=F(y[i])
}

df=data.frame(F_y,y)

library(ggplot2)
ggplot(df)+geom_line(aes(x=y,y=F_y),color="red")+labs(x='x',y='Distribution
Function')+ggtitle("Graph of F(x)")
```


Graph of F(x)
Distribution Function
1.00
0.75
0.50
0.25
0.00
-0.5
0.0
0.5
1.0
1.5
x

4)
$$F(x) = \begin{cases} 0 & ; x < 0 \\ x & ; 0 < x < 1/4 \\ \frac{x}{2} + \frac{1}{8} & ; 1/4 \le x < 3/4 \\ 2x - 1 & ; 3/4 \le x < 1 \\ 1 & ; x \ge 1 \end{cases}$$

(b) $P([0, 1/4]) = P((-\infty, 1/4]) - P((-\infty, 0])$

$= F(1/4) - F(0)$

$= \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{8} - 0 = \frac{2}{8} = \frac{1}{4}$.

$\Rightarrow P([0, 1/4]) = 1/4$

$P([1/8, 3/2]) = P((-\infty, 3/2]) - P((-\infty, 1/8])$.

$= F(3/2) - P(1/8)$

$= 1 - 1/8$

$\Rightarrow P([1/8, 3/2]) = 7/8$

$P((3/4, 7/8]) = P((-\infty, 7/8]) - P((-\infty, 3/4])$

$= F(7/8) - F(3/4)$

$= 2 \cdot \frac{7}{8} - 1 - 2 \cdot \frac{3}{4} + 1$

$\Rightarrow P((\frac{3}{4}, \frac{7}{8}]) = \frac{1}{4}$

(c) We can find the pdf of X by, $f(x) = \frac{d}{dx}[F(x)]$.

$$f(x) = \begin{cases} 1 & ; 0 < x < 1/4 \\ 1/2 & ; 1/4 < x < 3/4 \\ 2 & ; 3/4 < x < 1 \\ 0 & ; \text{o.w.} \end{cases}$$

6)
$$F(x) = \begin{cases} 0 & ; x \le 0 \\ \frac{2}{\pi} \sin^{-1}(\sqrt{x}) & ; 0 < x < 1 \\ 1 & ; x \ge 1 \end{cases}$$

($F : \mathbb{R} \to [0,1]$.)

For finding the pdf of the given $F(x)$, we need to differentiate it.

So, $f(x) = \begin{cases} \frac{1}{\pi} \cdot \frac{1}{\sqrt{x(1-x)}} & ; \ 0 < x < 1 \\ 0 & ; \ 0 \text{ i.w} \end{cases}$

$\because \frac{d}{dx}\left[\frac{2}{\pi} \sin^{-1}(\sqrt{x})\right] = \frac{2}{\pi} \cdot \frac{1}{\sqrt{1-(\sqrt{x})^2}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{\pi\sqrt{x}} \cdot \frac{1}{\sqrt{1-x}}$

This distribution of X is known as the standard arcsine law.

1. Let X be a Normal Random variable with mean $\mu = 3$ and standard deviation $\sigma = 1$.
   (a) Using the Normal tables provided in page 2 compute the $P(2.1 < X < 3.4)$
   (b) Use the `pnorm` functioin in `R` to compute the $P(2.1 < X < 3.4)$
2. Let X be a Normal Random variable with mean $\mu = 3$ and standard deviation $\sigma = 4$. Using the `pnorm` command compute
   (a) Compute $P(\mid X - \mu \mid < \sigma)$
   (b) Compute $P(\mid X - \mu \mid < 2\sigma)$
   (c) Compute $P(\mid X - \mu \mid < 3\sigma)$
3. Let X be a Normal Random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$, $x = -2, -1.9, -1, 8, ...0, ..., 1.9, 2$, $m = 100, p = 0.4$ and $Y \sim$ Binomial(m, p).
   (a) Using inbuilt `pnorm` plot the distribution function $P(X \leq x)$.
   (b) Let $Z = \frac{Y - mp}{\sqrt{mp(1-p)}}$, using inbuilt `pbinom` plot the distribution function $P(Z \leq x)$
4. Consider the Exponential (1) distribution.
   (a) Generate 100 trials of $5, 50, 5000$ samples respectively.
   (b) In each case $5, 50, 5000$, compute the sample mean for each of the 100 trials.
   (c) Compute the average value and variance of the sample mean (over the 100 trails) for each case $5, 50, 5000$.
   (d) What do you observe about the average value and variance comparing each case ?

Normal Tables

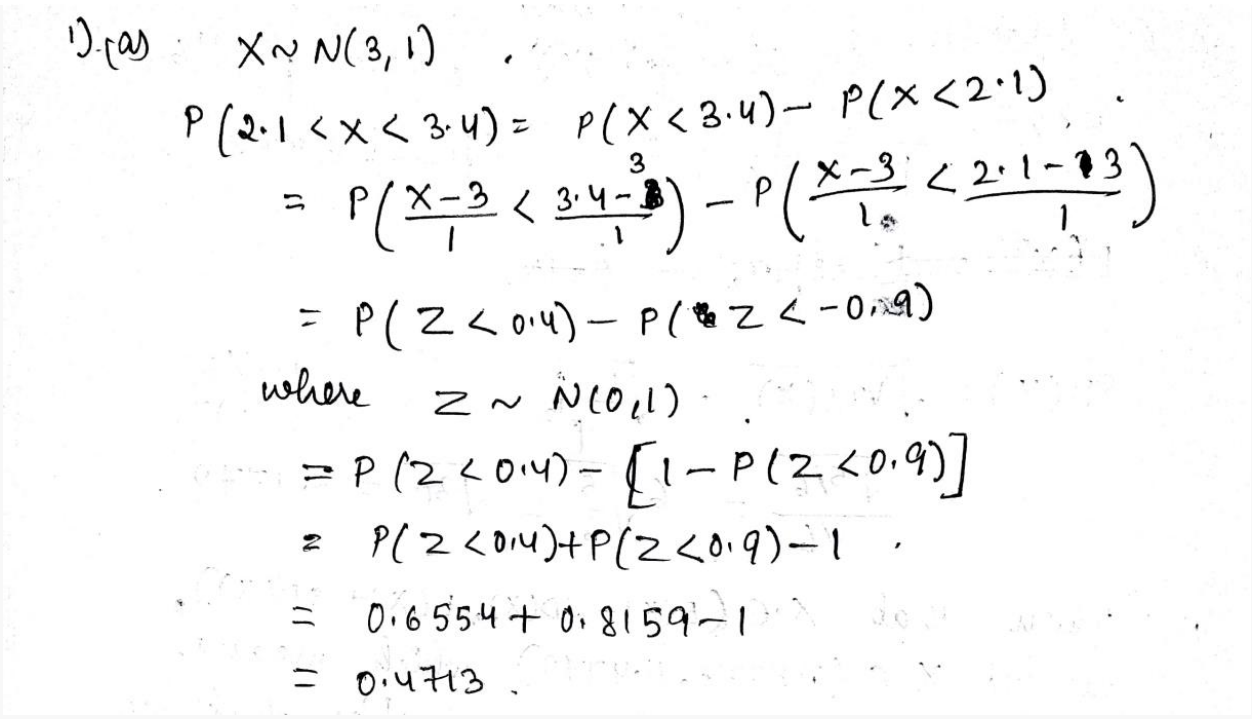
$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.5398
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.5793
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.6179
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.6554
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.6915
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.7257
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.7580
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7881
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8159
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.8413
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	0.8643
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	0.8849
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	0.9032
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	0.9192
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	0.9332
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	0.9452
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	0.9554
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	0.9641
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	0.9713
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	0.9772
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	0.9821
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	0.9861
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	0.9893
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	0.9918
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	0.9938
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	0.9953
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	0.9965
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	0.9987
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.9998

Worksheet 5

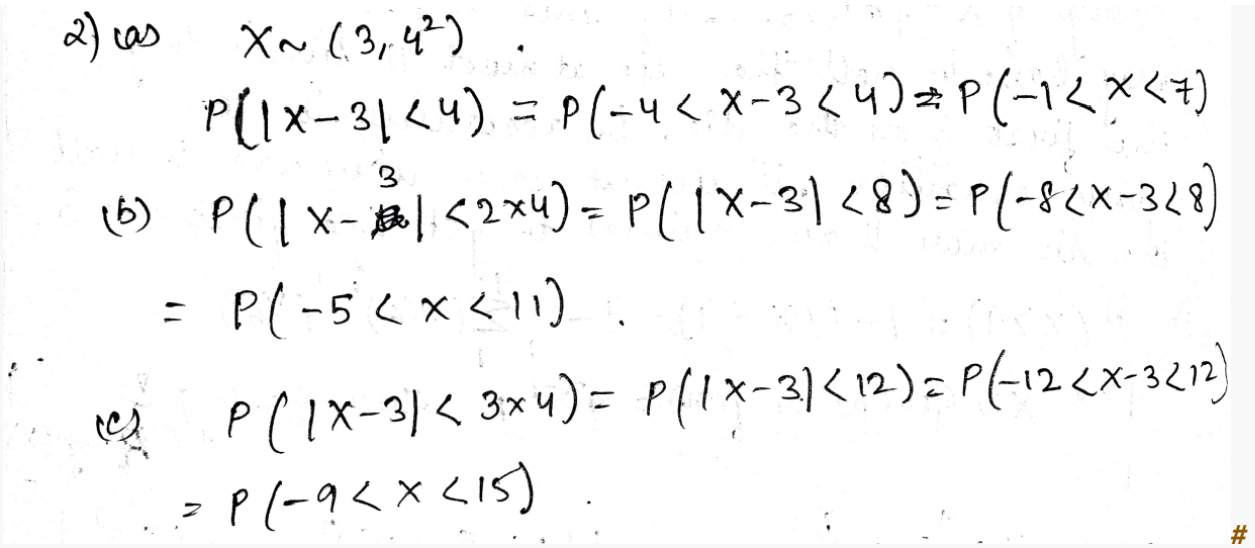
Rishika Tibrewal

20/10/2021




```
## 1 b)
pnorm(3.4,mean=3,sd=1)-pnorm(2.1,mean=3,sd=1)
```

```
## [1] 0.4713616
```




```
# 2 a)
pnorm(7,mean=3,sd=4)-pnorm(-1,mean=3,sd=4)
```

```
## [1] 0.6826895
```

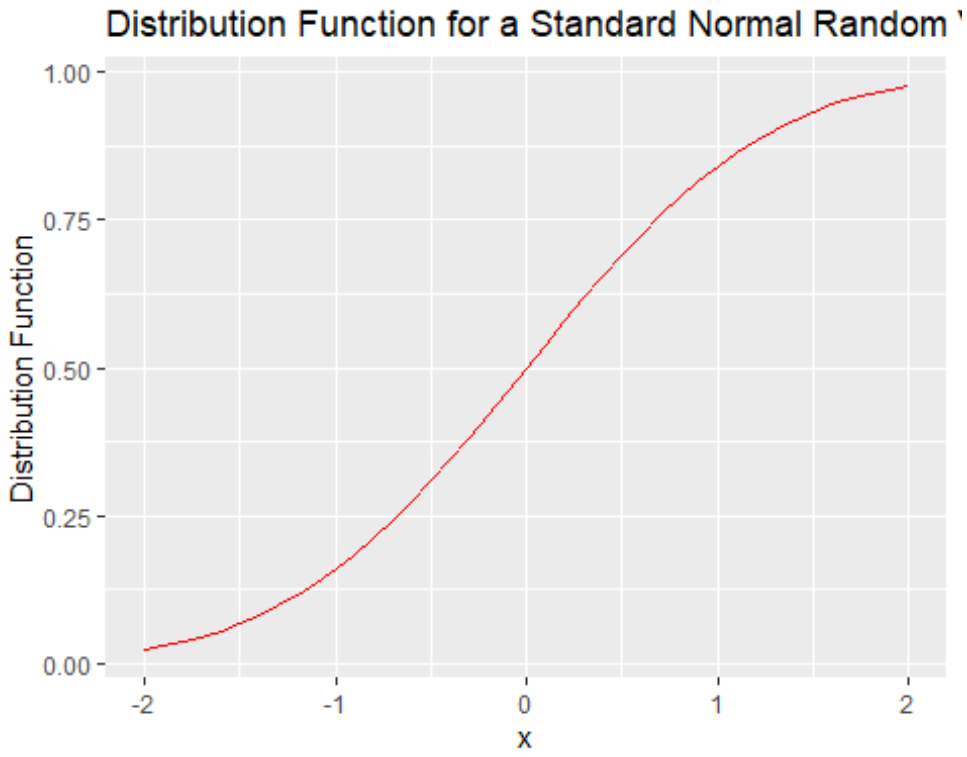
```
## 2 b)
pnorm(11,mean=3,sd=4)-pnorm(-5,mean=3,sd=4)
```

```
## [1] 0.9544997
```

```
## 2 c)
pnorm(15,mean=3,sd=4)-pnorm(-9,mean=3,sd=4)
```

```
## [1] 0.9973002
```

```
## 3 a)
x=seq(-2,2,0.1)
x1=pnorm(x)
df=data.frame(x1,x)
library(ggplot2)
ggplot(df)+geom_line(aes(x=x,y=x1),color="red")+labs(x='x',y='Distribution Fu
nction',title='Distribution Function for a Standard Normal Random Variable')
```



3) $X \sim N(0,1)$. $Y \sim Bin(m,p)$

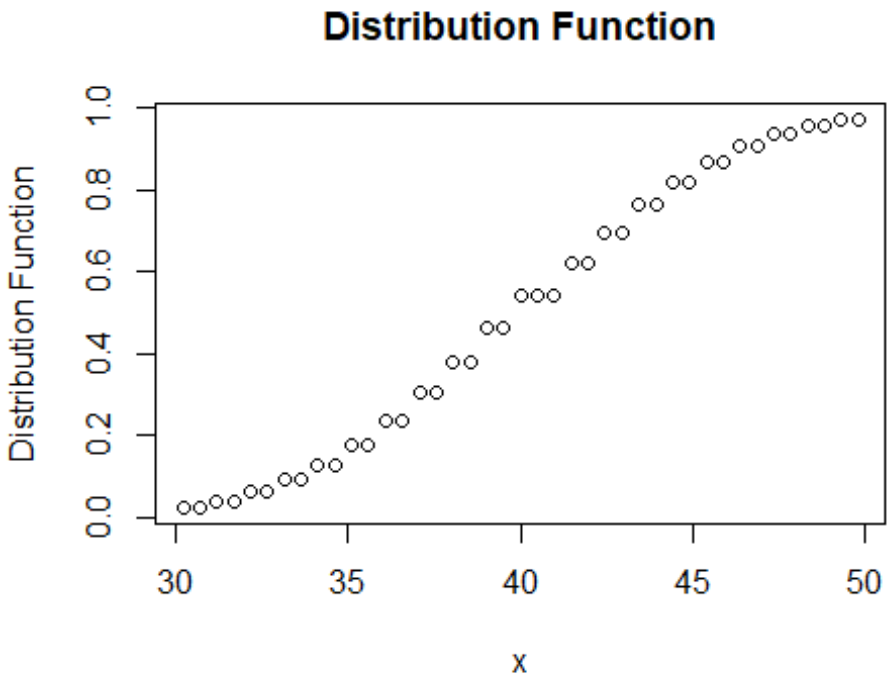
For Y, $\therefore \mu = mp$ & $\sigma = \sqrt{mp(1-p)}$

Given, $Z = \frac{Y-\mu}{\sigma} \Rightarrow Z\sigma = Y-\mu$

$\Rightarrow Y = Z\sigma + \mu$.

$\therefore P(Z \leq x) = P(Z\sigma+\mu \leq x\sigma+\mu) = P(Y \leq x2)$

```r
## 3 b)
x2=(x*(24^0.5))+40
z=pbinom(x2,100,0.4)
plot(x2,z,xlab='x',ylab='Distribution Function',main='Distribution Function')
```


Distribution Function
Distribution Function
0.0
0.2
0.4
0.6
0.8
1.0
30
35
40
45
50
x

Due date: November 5th, 2021
Problems Due: 2,4,6

From Probabiliy and Statistics with Examples using R.

1. Ex 1.4.6
2. Ex 1.4.7
3. Ex 3.3.3
4. Ex 3.3.6
5. Ex 3.3.7
6. Ex 4.4.1
7. Ex 4.5.2

Homework-6

2) Let U be the set of all students in the class.

$n(U) = 150$.

$A_1 = \{$the student is a female$\}$

$A_2 = \{$the student uses a pencil$\}$.

$A_3 = \{$the student is wearing eye glasses$\}$.

$n(A_1) = 90$ $\quad n(A_2) = 60$ $\quad n(A_3) = 30$.

$\Rightarrow P(A_1) = \frac{90}{150}$ $\quad \Rightarrow P(A_2) = \frac{60}{150}$ $\quad \Rightarrow P(A_3) = \frac{30}{150}$

$\Rightarrow P(A_1) = 0.6$ $\quad \Rightarrow P(A_2) = 0.4$ $\quad \Rightarrow P(A_3) = 0.2$

(a) For A_1, A_2, A_3 to be mutually independent, $P(A_1 \cap A_2 \cap A_3)$ must be equal to $P(A_1)P(A_2)P(A_3)$

Now, $P(A_1)P(A_2)P(A_3) = 0.6 \times 0.4 \times 0.2 = 0.048$

Let us assume it to be true,

then $P(A_1 \cap A_2 \cap A_3) = 0.048$

$\Rightarrow \frac{n(A_1 \cap A_2 \cap A_3)}{n(U)} = 0.048$

$\Rightarrow n(A_1 \cap A_2 \cap A_3) = 0.048 \times n(U)$

$= 0.048 \times 150$

$= 7.2$

which is not possible as $n(A_1 \cap A_2 \cap A_3)$ must be a non-negative integer lying between 0 and $\min\{n(A_1), n(A_2), n(A_3)\}$ i.e., $0 \leq n(A_1 \cap A_2 \cap A_3) \leq 30$. and must be an integer. So, we get a contradiction Hence, A_1, A_2, A_3 can't be mutually independent events.

(b) Let A_1, A_2, A_3 be pairwise independent then,

$P(A_1 \cap A_2) = P(A_1)P(A_2) = 0.6 \times 0.4 = 0.24 \Rightarrow \frac{n(A_1 \cap A_2)}{n(U)} = 0.24$

$\Rightarrow n(A_1 \cap A_2) = 0.24 \times 150 = 36$.

$P(A_2 \cap A_3) = P(A_2)P(A_3) = 0.4 \times 0.2 = 0.08 \Rightarrow \frac{n(A_2 \cap A_3)}{n(U)} = 0.08$

$\Rightarrow n(A_2 \cap A_3) = 0.08 \times 150 = 12$

$P(A_1 \cap A_3) = P(A_1)P(A_3) = 0.6 \times 0.2 = 0.18 \Rightarrow \frac{n(A_1 \cap A_3)}{n(U)} = 0.12$

$\Rightarrow n(A_1 \cap A_3) = 150 \times 0.12 = 18.$

which are all in the correct ranges as,

$0 \leq n(A_1 \cap A_2) \leq 60$, $0 \leq n(A_2 \cap A_3) \leq 30$, $0 \leq n(A_1 \cap A_3) \leq 30$

$\therefore$ It may be possible that A_1, A_2, A_3 are pairwise independent events.

4) Let X denote the number of heads on one flip of a single fair coin. Let Y denote the number of tails in the flip.

(a) $\because$ probability of obtaining a head on one flip of a single fair coin $= \frac{1}{2}$ = probability of obtaining a tail on one flip of a single fair coin.

Also, The number of trial is 1 as

So, for the success of getting a head, $p_1 = 1/2$.

$\Rightarrow X \sim$ Bernoulli $(1/2)$; $X \in \{0, 1\}$.

Similarly, for the success as getting a tail, $p_2 = 1/2$

$\Rightarrow Y \sim$ Bernoulli $(1/2)$; $Y \in \{0, 1\}$.

(b) $Z = X + Y$.

Outcome	X	Y	Z
head	1	0	1
tail	0	1	1

Suppose we toss and we get a head so, $X = 1$ and $Y = 0$. $\Rightarrow Z = 1$ Similarly, if we get a tail so, $X = 0$ and $Y = 1$. $\Rightarrow Z = 1$.

$\therefore Z = 1$ becomes a sure event and hence $P(Z = 1) = 1$.

(c) Z cannot be a binomial random variable as X, Y are not independent random variables. The probability of obtaining a head depends on the outcome of obtaining a tail. Also, numerically,

$P(X=1, Y=1) = 0$ [∵ we can't get both head and tail in the same trial]

But $P(X=1) \cdot P(Y=1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

∴ $P(X=1, Y=1) \neq P(X=1) \cdot P(Y=1)$

So, X, Y are not independent random variables and hence this example does not contradict the fact that if X, Y are independent random variables following Bernoulli(p), $X+Y \sim Bin(2, p)$.

c) Let $X \sim Geo(p)$ Let A be the event that $X \leq 3$

$$E[X|A] = \sum_{t \in Range(X)} t P(X=t|A)$$

$$= 1 \cdot P(X=1|X\leq 3) + 2 \cdot P(X=2|X\leq 3) + 3 \cdot P(X=3|X\leq 3)$$

[$P(X=t|X\leq 3)$ for $t > 3$ will be 0]

Now, $P(X \leq 3) = P(X=1) + P(X=2) + P(X=3)$

$$= p + (1-p)p + (1-p)^2 p$$

$$= p(p^2 - 3p + 3)$$

So, $E[X|A] = 1 \cdot \frac{P(X=1, X\leq 3)}{P(X\leq 3)} + 2 \cdot \frac{P(X=2, X\leq 3)}{P(X\leq 3)} + 3 \cdot \frac{P(X=3, X\leq 3)}{P(X\leq 3)}$

$$= \frac{1 \cdot p}{p(p^2-3p+3)} + \frac{2 \cdot (1-p)p}{p(p^2-3p+3)} + \frac{3 \cdot (1-p)^2 p}{p(p^2-3p+3)}$$

$$= \frac{1}{p^2-3p+3} + \frac{2-2p}{p^2-3p+3} + \frac{3+3p^2-6p}{p^2-3p+3}$$

$$\Rightarrow E[X|A] = \frac{6-8p+3p^2}{p^2-3p+3} = \gamma \text{ (say)}.$$

$$\text{Var}[X|A] = E[(X - E[X|A])^2 | A] = E[X^2 - 2XE[X|A] + (E(X|A))^2 | A]$$

$$= E[X^2|A] + E(E(X|A))^2 - E[2 \cdot E[X|A]|A].$$

$$= E[X^2|A] - 2E[X|A]^2 + (E[X|A])^2$$

$$= E[X^2|A] - (E[X|A])^2$$

$$E[X^2|A] = \sum_{t \in \text{Range}(X)} t^2 P(X=t|A) = \frac{p + 4p(1-p) + 9p(1-p)^2}{p + p(1-p) + p(1-p)^2}$$

$$= \frac{14p - 22p^2 + 9p^3}{3p - 3p^2 + p^3} = \frac{14 - 22p + 9p^2}{3 - 3p + p^2}$$

$$\text{Var}[X|A] = \frac{14 - 22p + 9p^2}{3 - 3p + p^2} - \left(\frac{6 - 8p + 3p^2}{3 - 3p + p^2}\right)^2$$

$$= \frac{(14 - 22p + 9p^2)(3 - 3p + p^2) - (6 - 8p + 3p^2)^2}{(3 - 3p + p^2)^2}$$

$$= \frac{1}{(3 - 3p + p^2)^2} \left[(42 - 42p + 14p^2 - 66p + 66p^2 - 22p^3 + 27p^2 - 27p^3 + 9p^4) - (36 + 64p^2 + 9p^4 - 96p - 48p^3 + 36p^2) \right]$$

$$= \frac{-p^3 + 7p^2 - 12p + 6}{(3 - 3p + p^2)^2}$$

$$\Rightarrow \text{Var}[X|A] = \frac{-p^3 + 7p^2 - 12p + 6}{(3 - 3p + p^2)^2}$$

1. Recall the Experiment 2 performed by all of you in Worksheet 1. Let X be the out come of the roll of a die and Y be the number of Heads in X tosses.

   (a) In R, using the the data set `dicetoss.csv` from the experiment, write an R-code that computes the conditional distribution of Y given X, for $X = 1, 2, 3, 4, 5, 6$. The output should be saved as a csv with entry of each row providing the conditional distribution Y for each value of X.

   (b) In R, using the the data set `dicetoss.csv` from the experiment, write an R-code that computes the conditional distribution of X given Y , for $Y = 1, 2, 3, 4, 5, 6$. The output should be saved as a csv with entry of each row providing the conditional distribution X for each value of Y.

Q1

```
dicetoss=read.csv("dicetoss.csv")
a=dicetoss[-c(28,57,58,60,104),]
```

a)

```
Conditional = data.frame(matrix(nrow=7,ncol=6))
rownames(Conditional)=c("Y=0","Y=1","Y=2","Y=3","Y=4","Y=5","Y=6")
colnames(Conditional)=c("X=1","X=2","X=3","X=4","X=5","X=6")

for (i in 0:6)
{
  for(j in 1:6)
  {
    Conditional[i+1,j]=(sum(a$Outcome.of.Roll==j &  a$Y..Number.of.Heads==i)/sum(a$Outcome.of.Roll==j))
  }
}

write.csv(Conditional,"1(a).csv")
```

b)

```
Conditional2 = data.frame(matrix(nrow=6,ncol=7))
colnames(Conditional2)=c("Y=0","Y=1","Y=2","Y=3","Y=4","Y=5","Y=6")
rownames(Conditional2)=c("X=1","X=2","X=3","X=4","X=5","X=6")

for (i in 1:6)
{
  for(j in 0:6)
  {
    Conditional2[i,j+1]=(sum(a$Outcome.of.Roll==i &  a$Y..Number.of.Heads==j)/sum(a$Y..Number.of.Heads==
  }
}

write.csv(Conditional2,"1(b).csv")
```

1. (Tschebychev Inequality)

   (a) Find a random variable X with Range$(X) = \{-1, 0, 1\}$ such that

   $$P(\mid X - \mu \mid \geq 2\sigma) = \frac{1}{4},$$

   with $\mu = E[X]$ and $\sigma^2 = \text{Var}[X]$.

   (b) Construct another random variable Y (different from X) with Range $(Y) = \{y_1, y_2, y_3\}$, mean μ and with

   $$P(\mid Y - \mu \mid > 2\sigma) > P(\mid X - \mu \mid > 2\sigma),$$

   so as to get

   $$P(\mid Y - \mu \mid > 2\sigma) > \frac{1}{4}$$

   Decide whether Tschebychev Inequality is violated ?

   (c) Write an R-code that takes an input k, and constructs a random variable X with Range$(X) = \{-1, 0, 1\}$ such that

   $$P(\mid X - \mu \mid \geq k\sigma) = \frac{1}{k^2},$$

   with $\mu = E[X]$ and $\sigma^2 = \text{Var}[X]$. Further the R-code should construct a random variable Y (different from X) with Range $(Y) = \{y_1, y_2, y_3\}$, mean μ so that

   $$P(\mid Y - \mu \mid > k\sigma) > \frac{1}{k^2}$$

   and (using replications) verifies your conclusion about Tschebychev's inequality in (b). It should save the entire output as a (suitably designed) csv file.

Due date: November 19th, 2021
Problems Due:1,3,5,7

From Probabiliy and Statistics with Examples using R.

1. Ex 3.2.4
2. Ex 3.2.5
3. Ex 3.2.9
4. Ex 3.3.7
5. Ex 3.3.11
6. Ex 3.3.15
7. Ex 4.4.3
8. Ex 4.4.4

1) (a) Range$(X) = \{-1, 0, 1\}$.

$P(|X - \mu_X| \geq 2\sigma_X) = 1/4$.

By Tschebychev's inequality, we know, $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

Comparing, we get that this is the equality case of Tschebychev's inequality when $k = 2$.

Let the probabilities of X taking values $-1, 0, 1$ be p_{-1}, p_0 and p_1 respectively.

Now, these events are exhaustive so, $p_{-1} + p_0 + p_1 = 1$

So, if we know value of p_{-1} & p_1, $p_0 = 1 - p_{-1} - p_1$.

For simplicity, lets take $p_1 = p_{-1} = \alpha \Rightarrow p_0 = 1 - 2\alpha$

Now, $E(X) = (-1)p_{-1} + 0\,p_0 + 1\,p_1 = -\alpha + \alpha = 0 \Rightarrow \mu_X = 0$

$\text{Var}(X) = (-1)^2\alpha + 0^2(1-2\alpha) + 1^2(\alpha) - 0^2$

$\Rightarrow \text{Var}(X) = 2\alpha \Rightarrow \sigma_X^2 = 2\alpha \Rightarrow \sigma_X = \sqrt{2\alpha}$.

So, we need α such that, $P(|X| \geq 2\sqrt{2\alpha}) = 1/4$.

If $\sqrt{\alpha} > \frac{1}{2\sqrt{2}} \Rightarrow 2\sqrt{2\alpha} > 1 \Rightarrow P(|X| \geq 2\sqrt{2\alpha}) = 0$.

(as X takes values $-1, 0, 1$)

So, $\sqrt{\alpha} \leq \frac{1}{2\sqrt{2}} \Rightarrow 2\sqrt{2\alpha} \leq 1 \Rightarrow |X| \geq 2\sqrt{2\alpha}$

$\Rightarrow |X| = 1 \Rightarrow X = -1, 1$.

So, we know $\sqrt{\alpha} \leq \frac{1}{2\sqrt{2}}$ or $\alpha \leq 1/8$.

Now, $P(|X| \geq 2\sqrt{2\alpha}) = P(X = -1) + P(X = 1)$

$= \alpha + \alpha = 2\alpha$

$\Rightarrow 1/4 = 2\alpha \Rightarrow \alpha = 1/8$.

$\Rightarrow p_{-1} = p_1 = 1/8 \Rightarrow p_0 = 1 - \frac{2}{8} = \frac{3}{4}$.

$\therefore X = \begin{cases} -1, & \text{with probability } \frac{1}{8} \\ 0, & \text{with probability } 3/4 \\ 1, & \text{with probability } 1/8 \end{cases}$

(b) Let Y be a random variable such that $Y \in \{-2, 0, 2\}$ with probability mass function

$$f(y) = \begin{cases} \beta & ; y = -2 \\ 1-2\beta & ; y = 0 \\ \beta & ; y = 2 \end{cases}$$

For simplicity we took $P(Y=-2) = P(Y=2)$ and $P(Y=0) = 1 - P(Y=-2) - P(Y=2)$

$$\mu_Y = \beta(-2) + 0(1-2\beta) + \beta(2) = 0$$

$$\sigma_Y^2 = (-2)^2\beta + 0^2(1-2\beta) + 2^2\beta = 4\beta + 4\beta = 8\beta$$

$$\Rightarrow \sigma_Y = \sqrt{8\beta} = 2\sqrt{2\beta}$$

Now, we also know $\sigma_X = \frac{1}{2}$. (from previous sum)

We know, $P(|X - \mu_X| \geq 2\sigma_X) = 1/4$

$$P(|Y - \mu_X| > 2\sigma_X) = P(|Y| > 1) \quad [\because \mu_X = 0, \sigma_X = 1/2]$$

$$= P(|Y| = 2) = P(Y=-2) + P(Y=2)$$

$$= 2\beta.$$

Now, if $P(|Y - \mu_X| > 2\sigma_X) > P(|X - \mu_X| > 2\sigma_X)$

$$\Rightarrow P(|Y - \mu_X| > 2\sigma_X) > 1/4.$$

So, clearly, $2\beta > 1/4 \Rightarrow \beta > 1/8$.

So, we can take any such value of β such that $\beta > 1/8$

$[\because 1 - 2\beta \geq 0 \Rightarrow 1 \geq 2\beta \Rightarrow 1/2 \geq \beta]$

* However, this does not violate Tschebyshev's inequality as here we are taking the mean and standard deviation of X for values of Y, instead to check if it violates we need to consider the mean and standard deviation of Y. $(1-2\beta \geq 0 \Rightarrow \beta \leq \frac{1}{2})$

$P(|Y - \mu_X| \geq 2\sigma_Y) = P(|Y| \geq 4\sqrt{2\beta})$. Hence we take $\beta \in (1/8, \frac{1}{2}] \Rightarrow 4\sqrt{2\beta} \in (2, 4]$

a) $P(|Y-\mu_Y| \geq 2\sigma_Y) = P(|Y| \in (2,4]) = 0 \leq \frac{1}{4}$.

Thus, Tschebycher's inequality holds.

So, we say,

$P(|Y-\mu_Y| \geq 2\sigma_Y) = 0$ when $\beta \in \left(\frac{1}{8}, \frac{1}{2}\right]$.

For $\beta \in [0, 1/8]$, $4\sqrt{2\beta} \in [0,2]$.

$\Rightarrow P(|Y-\mu_Y| \geq 2\sigma_Y) = P(|Y| \in [0,2])$

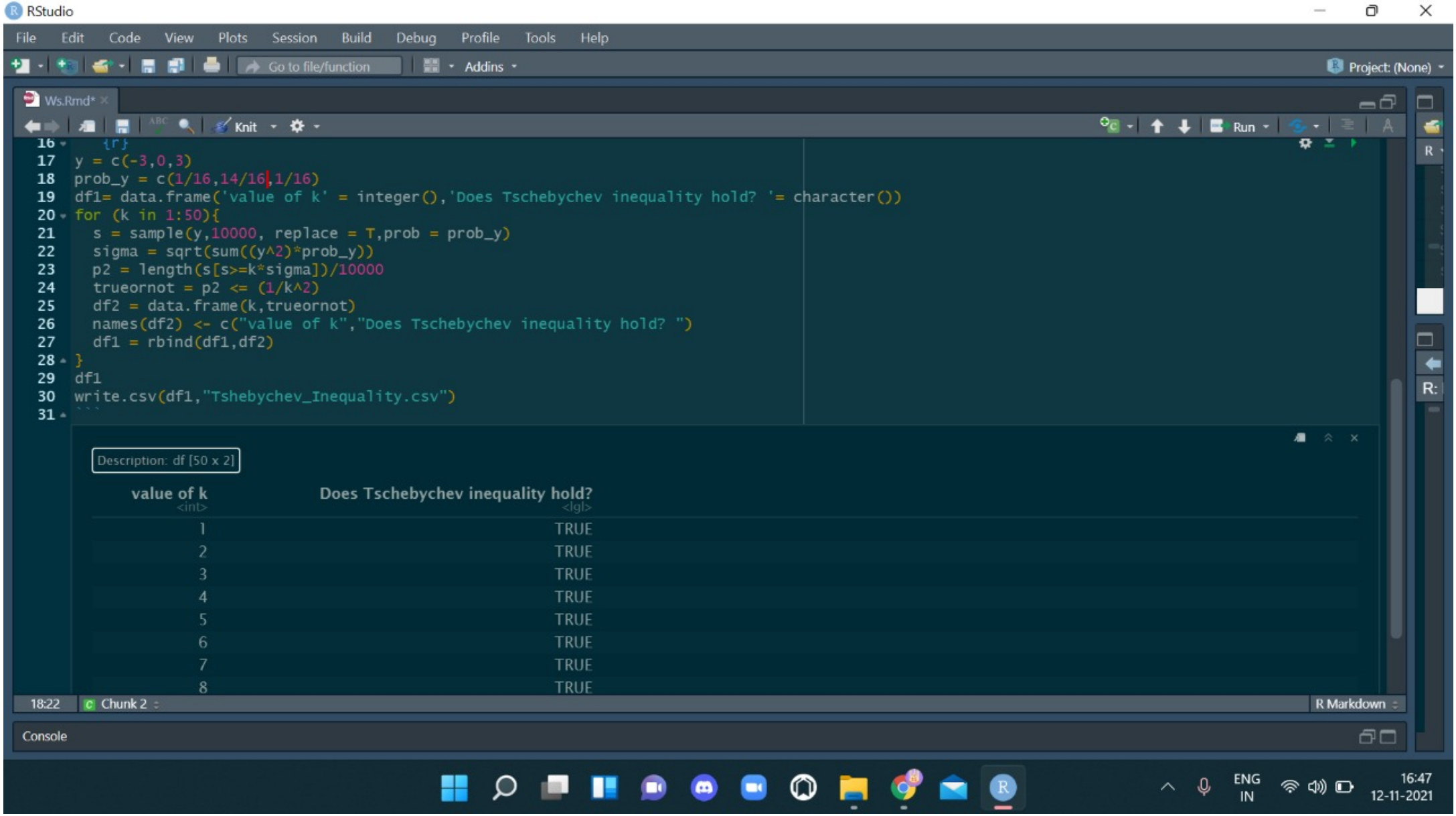
$= P(Y=-2) + P(Y=2) = 2\beta \leq 1/4$

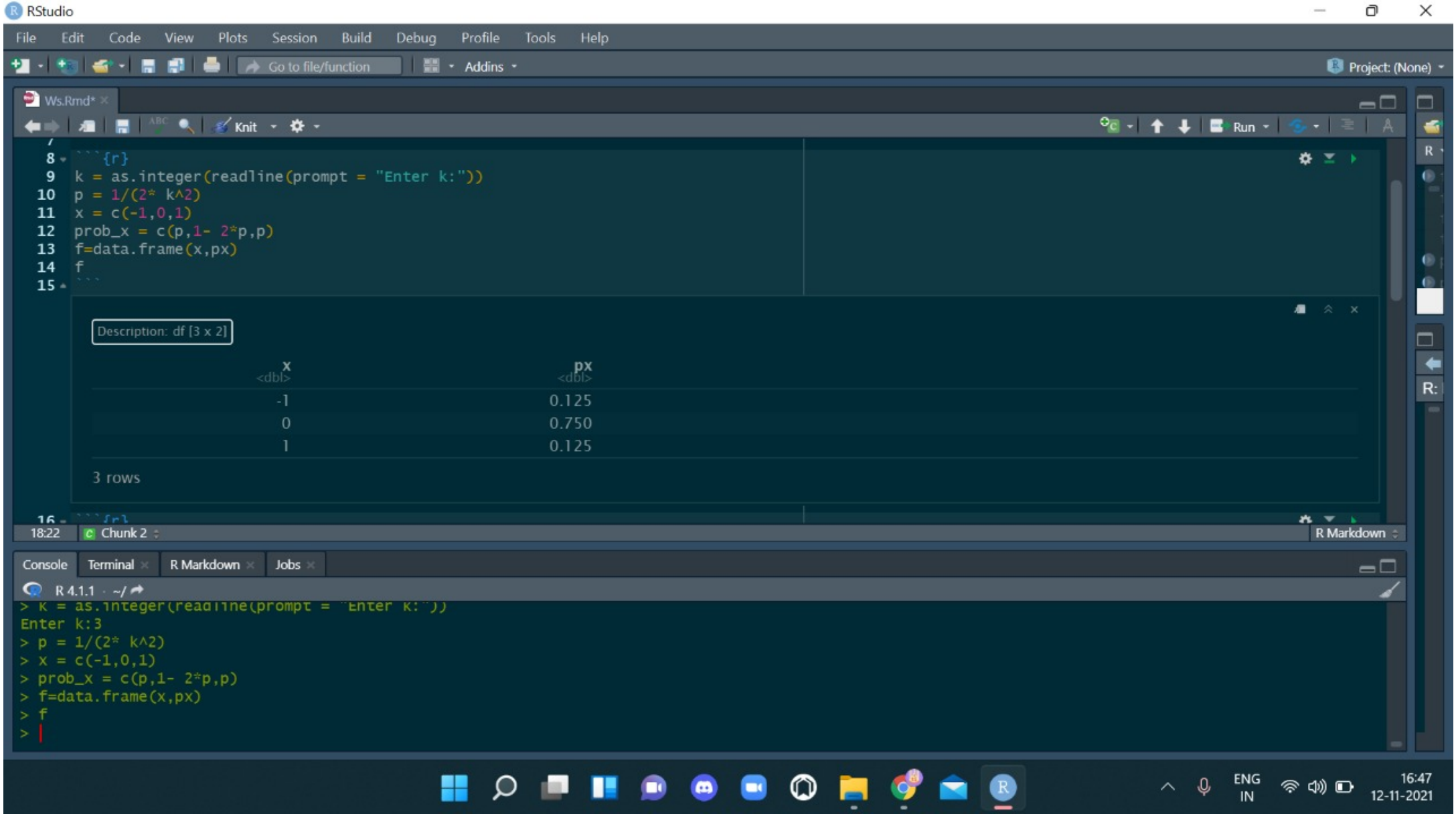
as $\beta \leq 1/8$.

We can take $\beta = \frac{2}{8} = 1/4$.

So, we get,

y	-2	0	2
$P(Y=y)$	$1/4$	$1/2$	$1/4$


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Ws.Rmd*
Knit
Run
```{r}
y = c(-3,0,3)
prob_y = c(1/16,14/16,1/16)
df1= data.frame('value of k' = integer(),'Does Tschebychev inequality hold? '= character())
for (k in 1:50){
s = sample(y,10000, replace = T,prob = prob_y)
sigma = sqrt(sum((y^2)*prob_y))
p2 = length(s[s>=k*sigma])/10000
trueornot = p2 <= (1/k^2)
df2 = data.frame(k,trueornot)
names(df2) <- c("value of k","Does Tschebychev inequality hold? ")
df1 = rbind(df1,df2)
}
df1
write.csv(df1,"Tshebychev_Inequality.csv")
```
Description: df [50 x 2]
value of k <int> | Does Tschebychev inequality hold? <lgl>
1 TRUE
2 TRUE
3 TRUE
4 TRUE
5 TRUE
6 TRUE
7 TRUE
8 TRUE
18:22
Chunk 2
R Markdown
Console
ENG IN
16:47
12-11-2021


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Ws.Rmd*
Knit
Run
```{r}
k = as.integer(readline(prompt = "Enter k:"))
p = 1/(2* k^2)
x = c(-1,0,1)
prob_x = c(p,1- 2*p,p)
f=data.frame(x,px)
f
```
Description: df [3 x 2]
x <dbl> px <dbl>
-1 0.125
0 0.750
1 0.125
3 rows
18:22
Chunk 2
R Markdown
Console
Terminal
R Markdown
Jobs
R 4.1.1 · ~/
> k = as.integer(readline(prompt = "Enter k:"))
Enter k:3
> p = 1/(2* k^2)
> x = c(-1,0,1)
> prob_x = c(p,1- 2*p,p)
> f=data.frame(x,px)
> f
ENG
IN
16:47
12-11-2021

Homework - 7

1) Let X, Y be random variables with joint distribution:

	$X=0$	$X=1$	$X=2$	Sum $(P(Y=y))$
$Y=0$	$1/12$	0	$3/12$	$4/12$
$Y=1$	$2/12$	$1/12$	0	$3/12$
$Y=2$	$3/12$	$1/12$	$1/12$	$5/12$
$(P(X=x))$ Sum	$6/12$	$2/12$	$4/12$	

(a) The marginal distribution of X is given by

$$P_X(x) = \begin{cases} 6/12 & ; x=0 \\ 2/12 & ; x=1 \\ 4/12 & ; x=2 \\ 0 & ; \text{o.w} \end{cases}$$

The marginal distribution of Y is given by

$$P_Y(y) = \begin{cases} 4/12 & ; y=0 \\ 3/12 & ; y=1 \\ 5/12 & ; y=2 \\ 0 & ; \text{o.w} \end{cases}$$

(b) $P(X=x|Y=2) = \dfrac{P(X=x, Y=2)}{P(Y=2)}$

So, $P(X=0|Y=2) = \dfrac{3/12}{5/12} = \dfrac{3}{5}$

$P(X=1|Y=2) = \dfrac{1/12}{5/12} = \dfrac{1}{5}$

$P(X=2|Y=2) = \dfrac{1/12}{5/12} = \dfrac{1}{5}$.

$$\therefore P(X=x|Y=2) = \begin{cases} 3/15 & ; x=0 \\ 1/5 & ; x=1,2 \\ 0 & ; \text{o.w} \end{cases}$$

(c) $P(Y=y|X=2) = \dfrac{P(X=2, Y=y)}{P(X=2)}$.

$$P(Y=0 \mid X=2) = \frac{3/12}{4/12} = \frac{3}{4}.$$

$$P(Y=1 \mid X=2) = \frac{0}{4/12} = 0.$$

$$P(Y=2 \mid X=2) = \frac{1/12}{4/12} = \frac{1}{4}.$$

$$\therefore P(Y=y \mid X=2) = \begin{cases} 3/4 & ; y=0 \\ 1/4 & ; y=2 \\ 0 & ; \text{O.W} \end{cases}.$$

(d) It is enough to show that $P(X=x, Y=y)$ is not equal to $P(X=x).P(Y=y)$ for any value of $x \in \text{Range}(X)$ & $y \in \text{Range}(Y)$.

Consider, $P(X=0, Y=0) = 1/12$.

and $P(X=0)P(Y=0) = \frac{6}{12} \cdot \frac{4}{12} = \frac{1}{6}$.

clearly, $P(X=0, Y=0) \neq P(X=0)P(Y=0)$ proving X, Y are dependent variables.

3) Let N be a random variable denoting the number of earthquakes in a year. Given, $N \sim \text{Poisson}(\lambda)$. Also that the probability that a given earthquake has magnitude ≥ 5 is p. (independent of all other quakes).

Given that M is a random variable denoting the number of earthquakes in a year with magnitude atleast 5 such that $(M \mid N=n) \sim \text{Bin}(n, p)$.

(a) So, $$P(M=m \mid N=n) = \begin{cases} \binom{n}{m} p^m (1-p)^{n-m} & ; 0 \leq m \leq n \\ 0 & ; \text{O.W}. \end{cases}$$

Also, $P(N=n) = \frac{\lambda^n e^{-\lambda}}{n!}$, $n \geq 0$.

Now, $P(M=m \mid N=n) = \frac{P(M=m, N=n)}{P(N=n)}$

$\Rightarrow P(M=m, N=n) = P(M=m \mid N=n)\, P(N=n)$

$$\Rightarrow P(M=m, N=n) = \begin{cases} \binom{n}{m} p^m (1-p)^{n-m} \cdot \frac{\lambda^n e^{-\lambda}}{n!} & ; \ 0 \leq m \leq n, \ n \geq 0. \\ 0 & ; \text{O.W.} \end{cases}$$

(b) $P(M=m) = \sum_n P(M=m \mid N=n) P(N=n)$.

$(m>0)$

$$= \sum_{n=0}^{\infty} \binom{n}{m} p^m (1-p)^{n-m} \cdot \frac{\lambda^n e^{-\lambda}}{n!}.$$

$$= \sum_{n=m}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \binom{n}{m} p^m (1-p)^{n-m} \quad \left[\because n \geq m \text{ for a particular } m\right]$$

$$= e^{-\lambda} p^m \sum_{n=m}^{\infty} \frac{\lambda^{n-m} \lambda^m n!}{(n-m)!\, m!\, n!} (1-p)^{n-m}$$

$$= \frac{e^{-\lambda} (p\lambda)^m}{m!} \sum_{n=m}^{\infty} \frac{\lambda^{n-m}}{(n-m)!} (1-p)^{n-m}$$

$$\therefore P(M=m) = \frac{e^{-\lambda} (p\lambda)^m}{m!} \sum_{n=m}^{\infty} \frac{\lambda^{n-m}}{(n-m)!} (1-p)^{n-m}.$$

(c) $$P(M=m) = \frac{e^{-\lambda} (\lambda p)^m}{m!} \sum_{n=m}^{\infty} \frac{\lambda^{n-m} (1-p)^{n-m}}{(n-m)!}.$$

$$= \frac{e^{-\lambda} (\lambda p)^m}{m!} \sum_{k=0}^{\infty} \frac{\lambda^k \cdot (1-p)^k}{k!}$$

where $k = n-m$, $n \in [m, \infty]$

$\Rightarrow n-m = k \in [0, \infty]$

(d) We know, $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

So, $\sum_{k=0}^{\infty} \frac{\lambda^k (1-p)^k}{k!} = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ where $x = \lambda(1-p)$

$= e^{\lambda(1-p)}$.

Hence, $P(M=m) = \frac{1}{m!} e^{-\lambda} \cdot (\lambda p)^m \cdot e^{\lambda - \lambda p}$.

$= \frac{(\lambda p)^m e^{-\lambda p}}{m!}$.

$\Rightarrow P(M=m) = \frac{\lambda_1^m e^{-\lambda_1}}{m!}$ where $\lambda_1 = \lambda p$.

(pmf of a Poisson distribution) $m!$ $m > 0$.

So, $M \sim$ Poisson(λ_1) i.e., Poisson(λp).

5) Let X_1, X_2, X_3, X_4 be an iid sequence of Bernoulli(p) random variables. Let $Y = X_1 + X_2$ & $Z = X_3 + X_4$.

$\because X_1, X_2$ are independent random variables with $X_1, X_2 \sim$ Bernoulli$(p) \Rightarrow Y = X_1 + X_2$ will follow Bin$(2, p)$. Similarly, $Z = X_3 + X_4$ will follow Bin$(2, p)$.

(a) $P(Z=0) = (1-p)^2$

$P(Z=1) = 2p(1-p)$

$P(Z=2) = p^2$.

$P(X=x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & 0 \le x \le n \\ 0 & ; 0 \text{ w} \end{cases}$

for $X \sim$ Bin(n, p).

Similarly, $P(Y=0) = (1-p)^2$

$P(Y=1) = 2p(1-p)$

$P(Y=2) = p^2$.

$$P(Y=y \mid Z=z) = P(X_1+X_2=y \mid X_3+X_4=z)$$

$$= P(X_1=x_1, X_2=y-x_1 \mid X_3=x_2, X_4=z-x_2)$$

$$= \frac{P(X_1=x_1, X_2=y-x_1, X_3=x_2, X_4=z-x_2)}{P(X_3=x_2, X_4=z-x_2)}$$

We know, X_i's are rid variable, $i=1(1)4$

$$= \frac{P(X_1=x_1)P(X_2=y-x_1)P(X_3=x_2)P(X_4=z-x_2)}{P(X_3=x_2)P(X_4=z-x_2)}$$

$$= P(X_1=x_1)P(X_2=y-x_1) = P(X_1+X_2=y) = P(Y=y).$$

We use, $P(Y=y|Z=z) = P(Y=y) \quad \forall\, z \in \text{Range}(Z)$

to get: $P(Y=0|Z=z) = (1-p)^2$

$P(Y=1|Z=z) = 2p(1-p)$

$P(Y=2|Z=z) = p^2$.

Similarly, $P(Z=z|Y=y) = P(Z=z) \quad \forall\, y \in \text{Range}(Y)$.

$P(Z=0|Y=y) = (1-p)^2$

$P(Z=1|Y=y) = 2p(1-p)$

$P(Z=2|Y=y) = p^2$.

We use, $P(Z=z|Y=y) = \dfrac{P(Z=z, Y=y)}{P(Y=y)}$.

$\Leftrightarrow P(Z=z, Y=y) = P(Z=z|Y=y)\,P(Y=y)$.

So, we get the following joint distribution table

	$Z=0$	$Z=1$	$Z=2$	$Y=y$
$Y=0$	$(1-p)^4$	$2p(1-p)^3$	$p^2(1-p)^2$	$(1-p)^2$
$Y=1$	$2p(1-p)^3$	$4p^2(1-p)^2$	$2p^3(1-p)$	$2p(1-p)$
$Y=2$	$p^2(1-p)^2$	$2p^3(1-p)$	p^4	p^2.
$Z=z$	$(1-p)^2$	$2p(1-p)$	p^2	

(b) Now, from table we get that,

$P(Z=0, Y=0) = (1-p)^4 = (1-p)^2(1-p)^2$

$= P(Z=0)\,P(Y=0)$.

$P(Z=2, Y=1) = 2p^3(1-p) = p^2 \cdot 2p(1-p)$

$= P(Z=2)\,P(Y=1)$

So, similarly, $P(Z=z, Y=y) = P(Z=z)P(Y=y)$

$\Rightarrow Y, Z$ are independent random variables

Moreover, from above computation,

$P(Z=z|Y=y) = P(Z=z) \Rightarrow$ Y, Z are independent random variables.

(c) $Y = X_1 + X_2 = f(X_1, X_2) \equiv$ function of X_1 & X_2

$Z = X_3 + X_4 = g(X_3, X_4) \equiv$ function of X_3, X_4.

$\therefore$ Z, Y are functions of mutually independent random variables, we can conclude directly from the given theorem that Y, Z will be independent random variables.

d). Estimated mean = \$ 2 million

Estimated standard deviation = \$3 million (if economy weakens or strengthens)

= \$ 2 million (if economy is stagnant)

Let X be the random variable denoting the return on the investment. Let A denote the economy strengthens event, B denote the stagnant economy event and C denote the weakening economy event in next quarter.

$SD[X|A] = 3$; $SD[X|B] = 2$; $SD[X|C] = 3$.

$E[X|A] = 3$; $E[X|B] = 1$; $E[X|C] = 1$

We know, $Var[X] = (Var[X|A] + (E[X|A])^2) P(A) +$

$(Var[X|B] + (E[X|B])^2) P(B) +$

$(Var[X|C] + (E[X|C])^2) P(C) - (E[X])^2$.

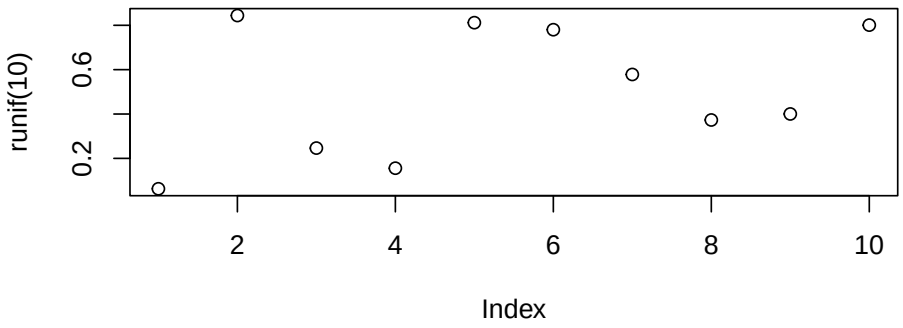
$= (3^2 + 3^2)(0.1) + (2^2 + 1^2)(0.4) + (3^2 + 1^2)(0.5) - (0.2)^2$

$= 1.8 + 2 + 5 - 0.04 = 8.76$

[$\because E[X] = 0.2$, $P(A) = 0.1$, $P(B) = 0.4$; $P(C) = 0.5$]

1. Consider the below outputs generated in R.

(a) (5 Points) Please write the down the R command that will provide the below plot. Describe in detail what the points in the plot represent.



(b) The following R code simulates a random variable X

```
> L = 10
> i = 0
> U = runif(1, min=0, max =1)
> Y = -log(U)/L
> Sum = Y
> while (Sum<1) {
+     U = runif(1, min=0, max =1)
+     Y = -log(U)/L
+     Sum = Sum +Y
+     i = i + 1
+ }
> X = i
```

(i) (10 points) Find $P(X = 0)$ and $P(X \geq 1)$.

(ii) (10 points) Suppose for $\lambda > 0$ and $T_1, T_2, \ldots T_n$ being i.i.d. Exp(λ) random variables it is known that for all $a > 0$,

$$P(\sum_{i=1}^{n} T_i \leq a) = \int_0^a \frac{\lambda^n}{n-1!} e^{-\lambda z} z^{n-1} dz$$

then find $P(X = n)$ for $n \geq 1$.

Page No.
Date: / /

8/11/21

Athul Prakash
athul@cmi.ac.in

Prob & Stat with R

Mid Term (Part - II)

1. a) plot(runif(10, min=0, max=1)) [continued on page 4] ✓ (5)

b) ~~[illegible]~~

The R-code simulates choosing 'k' values from Uniform(0,1) ~~[illegible]~~ $\{x_1, x_2 \cdots x_k\}$ such that

$$\sum_{\forall i} -\frac{\ln(x_i)}{10} = 1 \quad \text{(equal to or slightly greater than)}$$

$$\Rightarrow \prod_{\forall i} (x_i) = e^{-10} \quad (e^z \text{ on both sides} \cdot e^{\sum z_i} = \prod e^{z_i}) \; ?$$

With the R.V. $X = k-1$

(i) $x=0 \Rightarrow k=1$ values are sampled.

$$\Rightarrow -\frac{\ln(x_1)}{10} \geq 1$$

$$\Rightarrow \ln(x_1) \leq -10$$

$$\Rightarrow x_1 \leq e^{-10}$$

$$P(X=0) = P(K=1) = \frac{e^{-10}}{1-0} \quad \text{(uniform dist.)}$$

$$= e^{-10} \simeq 4.56 \times 10^{-5} \; ✓$$

(ii) Range(X) = 0, 1, 2, ...

~~[illegible]~~

(since in the R-code it is initialized to 0 and then counts up)

$$P(x=0 \cup x \geq 1) = 1$$

$$\Rightarrow P(x=0) + P(x \geq 1) = 1 \quad \text{(mutually exclusive)}$$

Page No.
Date: / /

Prob + Stat

(cont i.t.o (i))

$\Rightarrow \quad e^{-10} + P(X \geq 1) = 1$

$\Rightarrow P(X \geq 1) = 1 - e^{-10} = (0.9999546\ldots)$ ✓ 10

(b)(ii) $P\left(\sum T_i \leq a\right) = \int_0^a \frac{\lambda^n}{(n-1)!} e^{-\lambda x} x^{n-1}\, dx$ — (1)

- If $A \sim \text{Uniform}(0,1)$

  $P(A < a) = a$

- Let $B = \frac{-\ln(A)}{10}$

$P(B > b) = P\left(\frac{\ln(A)}{10} < -b\right) = P(\ln(A) < -10b)$

$= P(A < e^{-10b})$

$= e^{-10b}$

- $\Rightarrow B \sim \text{Exponential}(1/10)$

In the R-code, U is sampled from Uniform $(0,1)$

Hence Y is sampled from Exponential $\left(\frac{1}{10}\right)$

$X = n \Rightarrow \{T_1, T_2 \ldots T_{n+1}\}$ [$T_i \sim$ Exponential $\left(\frac{1}{10}\right)$] such that

$\sum_1^n T_i < 1$ — (2) and $\sum_1^{n+1} T_i \geq 1$ — (3)

(1) & (2) $\Rightarrow P(\Sigma T_i < 1) = P(\Sigma T_i \leq 1)$

$= \int_0^1 \frac{\left(\frac{1}{10}\right)^n}{(n-1)!} e^{-z/10} z^{n-1}\, dz$ — (4)

$= I$, say

(2) $\Rightarrow P(\Sigma T_i \geq 1) = 1 - P(\Sigma T_i \leq 1) = 1 - \int_0^1 \frac{\left(\frac{1}{10}\right)^{n+1}}{n!} e^{-z/10} z^n\, dz$ — (5)

$= 1 - J$ ('J')

(2)

$$I = \frac{(\frac{1}{10})^n}{(n-1)!} \int_0^1 \underbrace{e^{-z/10}}_{I} \underbrace{z^{n-1}}_{II} dz \qquad \text{(integrating by parts)}$$

$$= \left[\frac{(\frac{1}{10})^n}{(n-1)!} e^{-z/10} \frac{z^n}{n} \right]_0^1 - \int_0^1 \frac{(\frac{1}{10})^n}{(n-1)!} \left(\frac{-1}{10}\right) e^{-z/10} \frac{z^n}{n} dz$$

$$= \left[\frac{(\frac{1}{10})^n}{n!} e^{-1/10} - 0 \right] + \frac{(\frac{1}{10})^{n+1}}{n!} \int_0^1 e^{-z/10} \cdot z^n \cdot dz$$

$$I = \frac{(\frac{1}{10})^n}{n!} e^{-\frac{1}{10}} + J \qquad \text{(6)}$$

Then, if $I_k = P\left(\sum_1^k T_i \leq \frac{1}{10}\right)$

Then $I_k = \frac{(\frac{1}{10})^k}{k!} e^{-\frac{1}{10}} + I_{k+1} \qquad - (6)$

Let us reformulate, choose $(n+1)$ samples T_i from the above exponential.

$E_n \rightarrow$ Event that $\sum_1^n T_i \leq 1$

~~$P(E_n) = I_n$ (integral written above)~~

$E_{n+1} \rightarrow$ Event that $\sum_1^{n+1} T_i \leq 1$

$P(E_{n+1}) = I_{n+1}$ (integral above)

$P(E_n) = I_n$ (since T_i are i.i.d, $\sum_1^n T_i < 1$ is not affected by T_n or $\sum_1^{n+1} T_i$)

$E_{x=n} \rightarrow$ Event that $x=n$, ie event that $\sum_1^n T_i < 1$ and $\sum_1^{n+1} T_i \geq 1$

$\therefore E_n = E_{n+1} \cup E_{x=n}$ (if n variables sum <1, then either $n+1$ sum <1 or $n+1$ sum >1)

$P(E_n) = P(E_{n+1}) + P(x=n)$ (they are mutually exclusive since they map to different values of T_n)

$\Rightarrow P(x=n) = P(\epsilon_{n+1}) - P(\epsilon_{n+1})$

$= I_{n+1} - I_{n+1}$

$= \frac{\left(\frac{1}{10}\right)^n}{n!} e^{-\frac{1}{10}}$

1a) (continued)

~~The~~ The points represent each of 10 samples of a Random Variable ~~drawn from~~ which has a distribution of Uniform (0, 1).

This means that each point is sampled as a real number ~~in~~ from the interval [0, 1] such that the probability of sampling any number in that interval is identical

Due date: November 26th, 2021
Problems Due: 2,3

1. For each of the distributions: Beta(10,2) and Beta(10,10)
   (a) Generate 100 trials of $5, 50, 500$ samples respectively.
   (b) Using the data decide if the conclusion of the Central Limit Theorem applies in each of the three cases, $5, 50, 500$.

2. Consider the Poisson(1) distribution.
   (a) Generate 100 trials of 500 samples respectively.
   (b) Find the 95%-confidence interval for the mean in each trial.
   (c) Compute the number of trials in which the true mean lies in the interval.

3. The dataset `BangaloreRain.csv` in the course website at https://www.isibang.ac.in/ athreya/Teaching/PaSwR/Bngalor
   (a) Decide if any month's 100 year rainfail is Normally distributed.
   (b) Calculate the yearly total rain fall for each of the 100 years.
   (c) Plot the histogram and Decide if the distribution is Normal.
   (d) Find a 95% confidence interval for the expected annual rainfall in Bangalore.

4. Two types of coin are produced at a factory: a fair coin and a biased one that comes up heads 55% of the time. We have one of these coins but do not know whether it is a fair or biased coin. In order to ascertain which type of coin we have, we shall perform the following statistical test. We shall toss the coin 1000 times. If the coin comes up heads 525 or more times we shall conclude that it is a biased coin. Otherwise, we shall conclude that it is fair. If the coin is actually fair, what is the probability that we shall reach a false conclusion? What would it be if the coin were biased?

5. The length of time (in appropriate units) that a certain type of component functions before failing is a random variable with probability density function

$$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   Once the component fails it is immediately replaced with another one of the same type. Using the central limit theorem approximation, can you find, how many components would one need to have on hand to be approximately 90% certain that the stock would last at least 35 units of time ?

Homework 8

Rishika Tibrewal

19/11/2021

2a)

```
library("plotrix")
pois_data=rep(NA,50000)
pois_data = replicate(100, rpois(500,1),simplify=FALSE) #100 iterations of
generating 500 samples from a Poisson(1) distribution
```

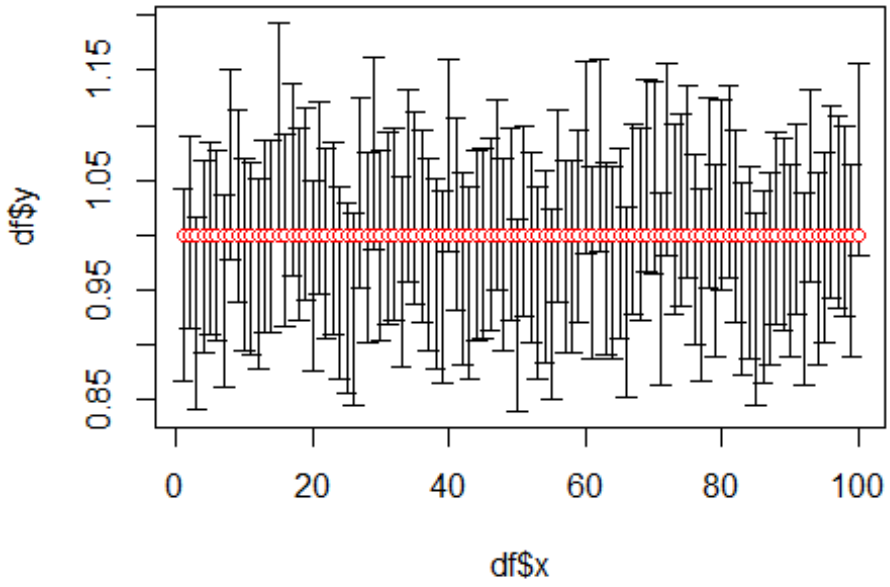
2b)

```
cifn = function(x, alpha=0.95){
z = qnorm( (1-alpha)/2, lower.tail=FALSE)
sdx = sqrt(1/length(x))
return(c(mean(x) - z*sdx, mean(x) + z*sdx))
} #function to calculate the 95% confidence intervals

cidata = sapply(pois_data, cifn) #applies the function cifn to pois_data and
stores it in cidata

df=data.frame(x=1:100,y=1,lower=cidata[1,],upper=cidata[2,]) #dataframe with
cols x,y,lower,upper where x is the number of iteration i, y is the expection
of Poisson(1) distribution-1, lower contains the lower limit of the
confidence interval of ith iteration and upper contains the upper limit of
the confidence interval of ith iteration
```

2c)

```
TRUEORNOT=(cidata[1,]<1 & cidata[2,]>1) #condition if the given interval
contains the expectation of Poisson(1) distribution-1 i.e, lower limit should
be less than 1 and upper limit should be greater than 1 so as to contain the
value 1 in the interval

res=sum(TRUEORNOT) #storing the number of iterations where the confidence
intervals contain 1 in res

table(TRUEORNOT) #printing the number of iterations where the confidence
intervals contain 1 and where not in a table format

## TRUEORNOT
## FALSE  TRUE
##     1    99
```

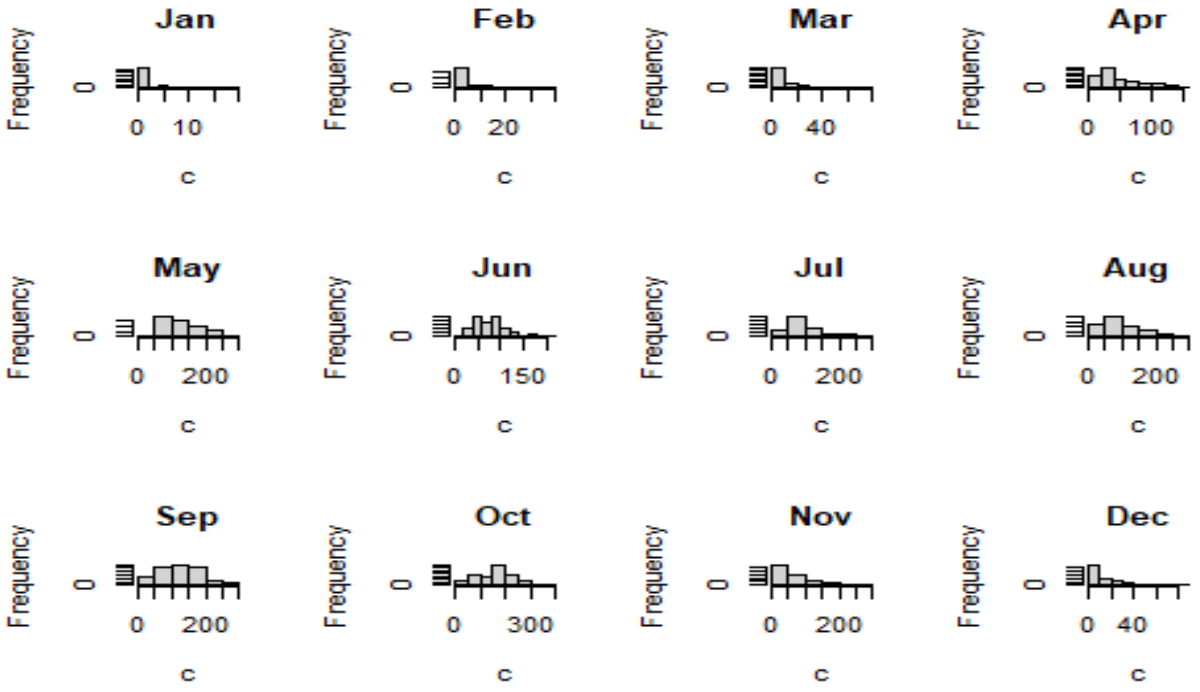
```
plotCI(df$x,df$y,li=df$lower,ui=df$upper,col="red",pch=21,scol="black")
#plotting the graph to visualise which confidence intervals contain 1 and
which not
```



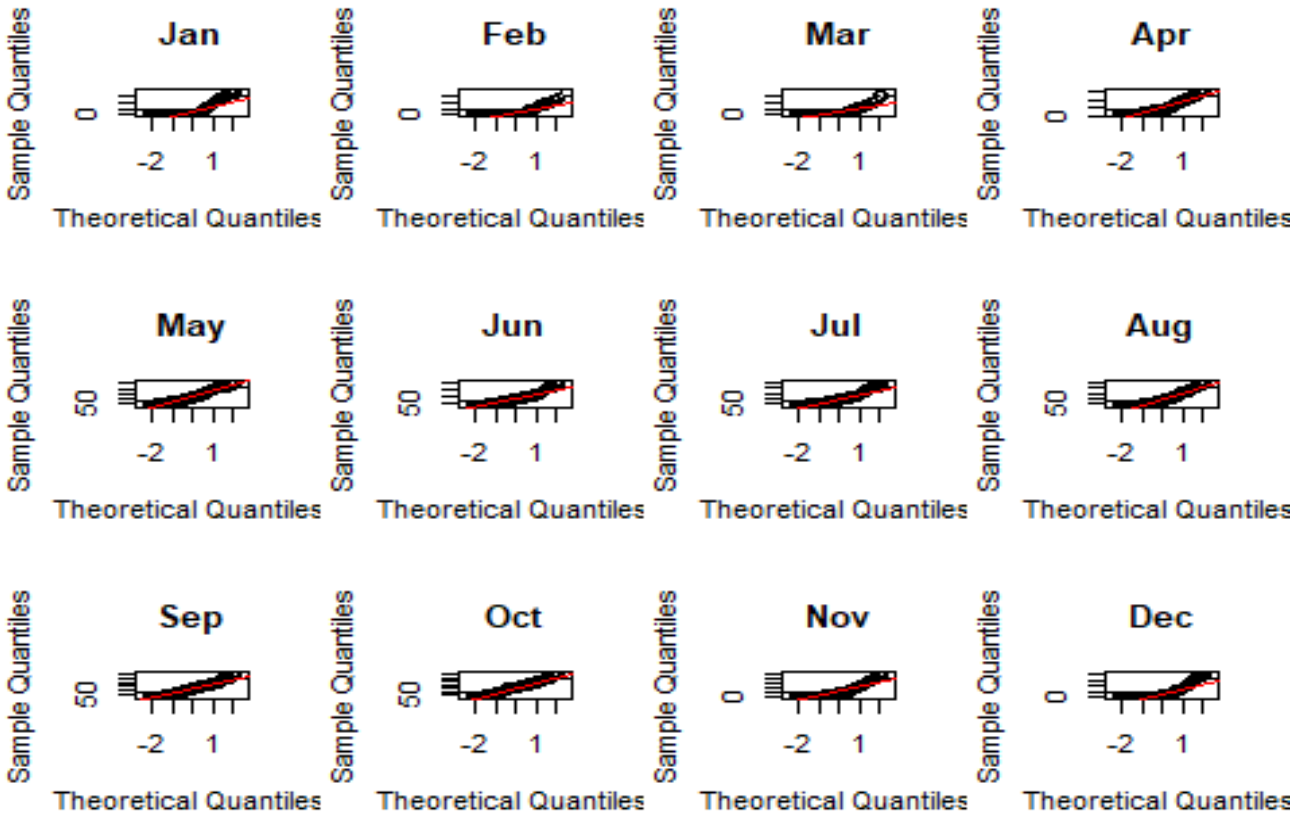

```
cat("\nThe number of intervals in which the true mean lies in the interval
is",res) #printing the number of iterations where the confidence interval
contained the expectation in it

##
## The number of intervals in which the true mean lies in the interval is 99
```

3a)

```
library("moments")
rain_data =
read.csv("https://www.isibang.ac.in/~athreya/Teaching/PaSwR/BangaloreRain.csv
", sep = "\t") #reading the data from the URL and separated by tabs

i = 2
par(mfrow = c(3,4))
for ( c in rain_data[,2:13] ) {
month = colnames(rain_data[i]) #storing the month name in month
hist(c,main=month) #plotting histogram for each month
i = i+ 1
}
```




```
skew=rep(NA,12)
kurt=rep(NA,12)

i = 2
par(mfrow = c(3,4))
for ( c in rain_data[,2:13] ) {
month = colnames(rain_data[i]) #storing the month name in month
qqnorm(c, main = month) #plotting qqplot for each month
qqline(c,col="red") #plotting qqline for each month
skew[i]=skewness(c)
kurt[i]=kurtosis(c)
i = i+ 1
}
```




```
skew[-1]

##  [1] 1.4676670 2.1011032 2.1158152 1.0178677 0.5080775 1.0284120 1.1804632
##  [8] 0.9439829 0.5098643 0.2289602 1.3967519 1.6914228

kurt[-1]

##  [1] 3.978936 7.635518 7.764698 3.209079 2.349746 4.254960 4.125412
3.200551
##  [9] 3.160818 2.875377 4.533403 5.278147
```

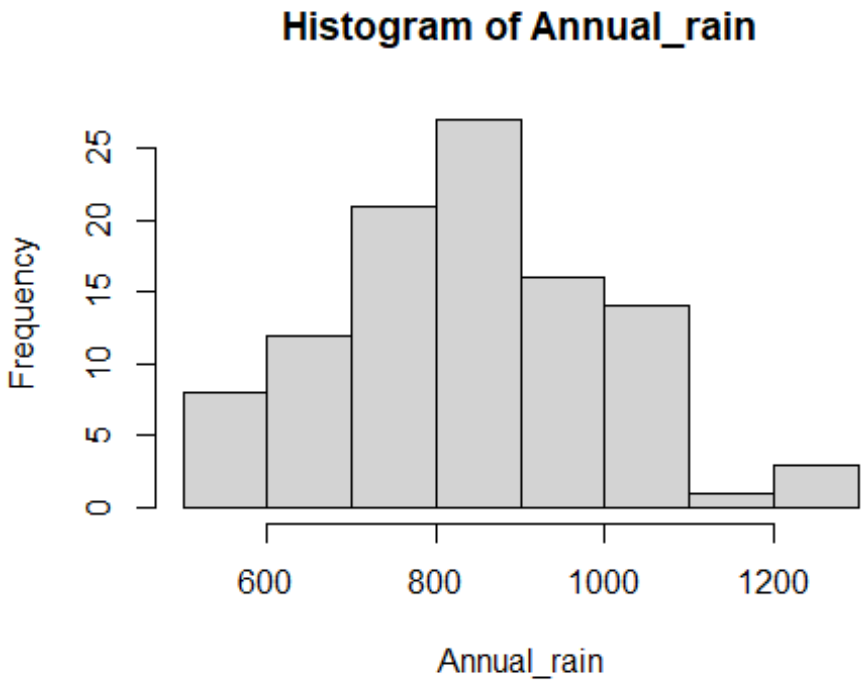
From the histograms and qqplots, it is evident that the above data is more or less normal for October. According to the value of skew and and kurt, we observe that the skewness and kurtosis of October lie in a range of (-0.5,0.5) and (2.5-3.5) respectively.

3b)

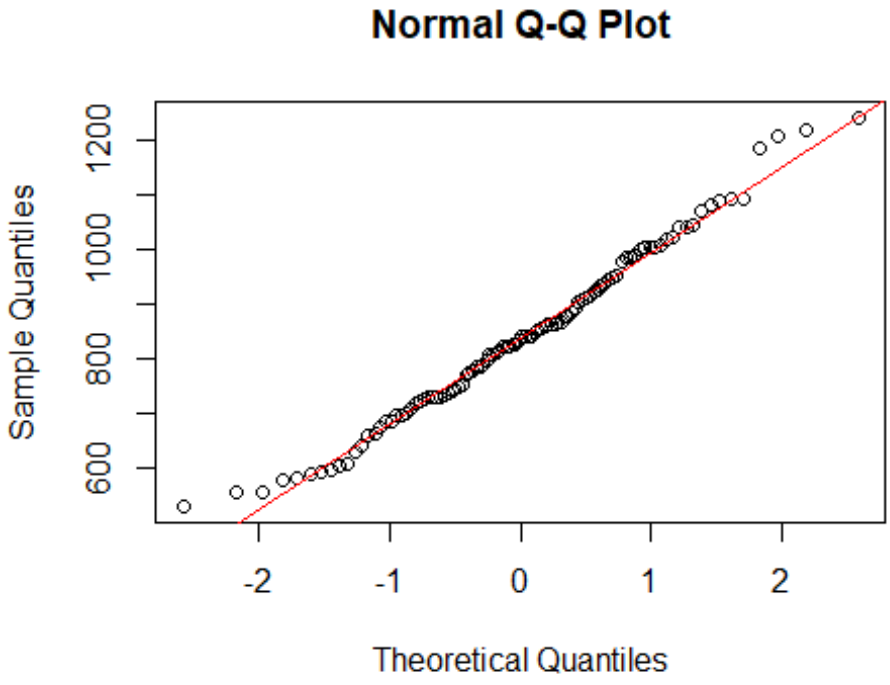
```
Annual_rain = rowSums(rain_data[,2:13]) #adding the amount of rainfall of
each month in a year to get for that particular year
```

3c)

```
hist(Annual_rain) #plotting the histogram of the Annual rainfall over the
years
```




```
qqnorm(Annual_rain)
qqline(Annual_rain,col="red")
```



We see from the above two graphs that the annual rainfall over the years is more or less normal, but towards the ends, the points on the qqplot move away from the normal line

3d)

```
z=qnorm((1-0.95)/2,lower.tail=FALSE)
lower_interval=mean(Annual_rain)-
z*sqrt(1/length(Annual_rain))*sd(Annual_rain)

upper_interval=mean(Annual_rain)+
z*sqrt(1/length(Annual_rain))*sd(Annual_rain)

cat("The 95% confidence interval for annual rainfall is
[",lower_interval,',',upper_interval,"]")
```

```
## The 95% confidence interval for annual rainfall is [ 807.9882 , 869.3301 ]
```

1. Use the data set `Scores` from the course website and decide if it is Normal using the following steps.
   (a) Check `summary` of scores
   (b) Compute the proportion of data that is 1− Standard Deviation, 2-Standard Deviation and 3-Standard Deviation far from the mean.
   (c) Plot: Histogram, Boxplot and Q-Q plot
   (d) Using the `moments` package, compute Skewness and Kurtosis.
2. Use the inbuilt-data sets in `R`, namely `ToothGrowth` and `faithful` USA.
   (a) Describe the `eruptions` variable in the data set codefaithful and `len` variable in the data set codeToothGrowth
   (b) Using the descriptive methods discussed so far, try to infer as much as you can about the distribution.
3. Consider the `Beta` (a, b) distribution. Discuss descriptive properties of the distribution when
   (a) $a = 10, b = 10$
   (b) $a = 10, b = 2$
   (c) $a = 2, b = 10$

Worksheet 8

Rishika Tibrewal

16/11/2021

```
#1)

Scores = scan("C:/Users/Rishika Tibrewal/OneDrive/Desktop/PBSR
WS/Scores.txt")

# (a) Checking summary of Scores
summary(Scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   22.00   30.00   29.28   40.00   50.00
```

```
# Finding standard deviation of Scores
sd(Scores)
```

```
## [1] 12.06816
```

```
# standardising the values in the scores
cs = (Scores-mean(Scores))/(sd(Scores))

# Finding the mean and sd of cs to check if they are 0 and 1 like a standard
Normal distribution

mean(cs)   # the mean approximates to 0
```

```
## [1] 7.922276e-17
```

```
sd(cs)     # the sd is 1
```

```
## [1] 1
```

```
# Hence the standardised vector cs can be essentially compared to a N(0,1)
distribution

# (b)

onesdcs = cs[cs >-1& cs <1] # data within one sd from mean
twosdcs = cs[cs >-2& cs <2] # data within two sd from mean
threesdcs = cs[cs >-3& cs <3] # data within three sd from mean

length(onesdcs)/length(cs) # Proportion of data within one sd from mean
```

```
## [1] 0.628866
```

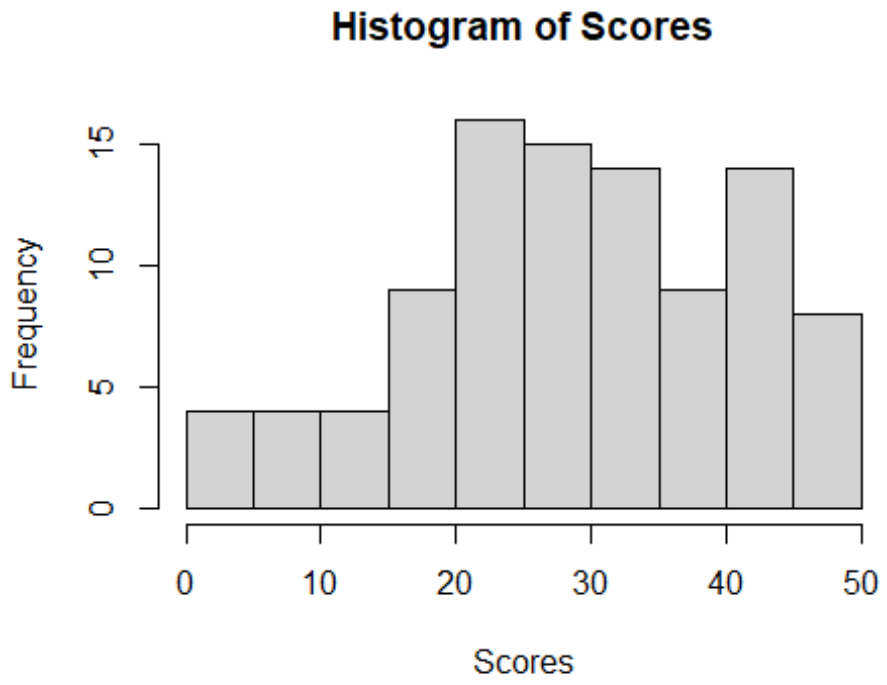
```
length(twosdcs)/length(cs) # Proportion of data within two sd from mean
```

```
## [1] 0.9587629
```

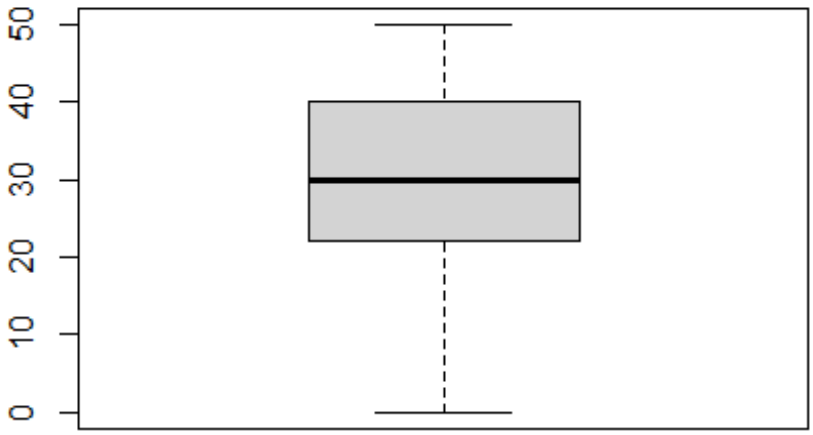
```
length(threesdcs)/length(cs) # Proportion of data within three sd from mean
```

```
## [1] 1
```

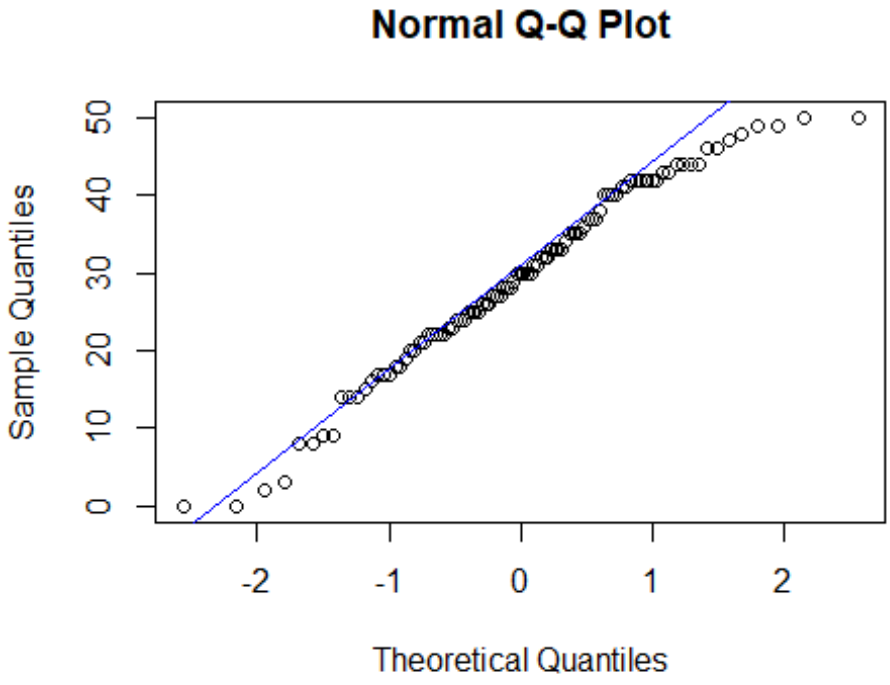
```
# (c)

# Histogram:
hist(Scores)
```




```
# Boxplot:
boxplot(Scores)
```




```r
# Q-Q plot:
qqnorm(Scores)
qqline(Scores, col = 'blue')
```



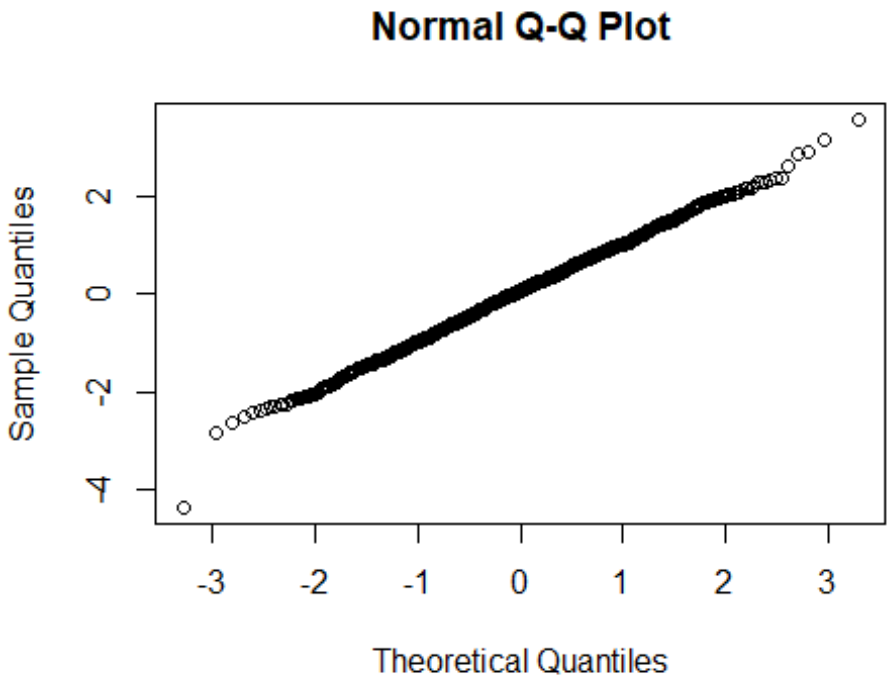
```r
# (d)

library(moments)

Kurtosis = kurtosis(Scores)
Kurtosis
```

```
## [1] 2.591014
```

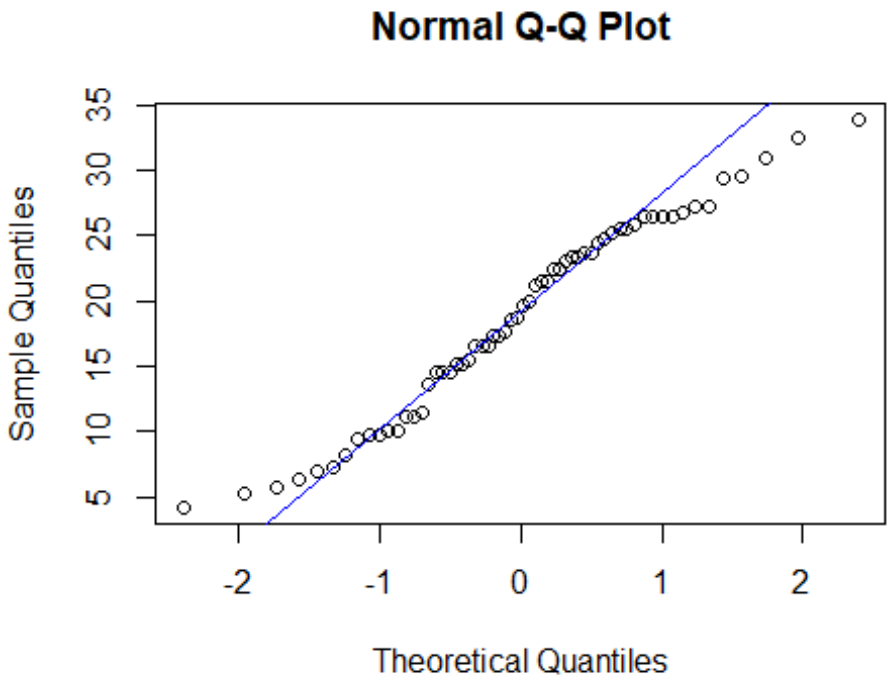
```r
Skewness = skewness(Scores)
Skewness
```

```
## [1] -0.3548957
```

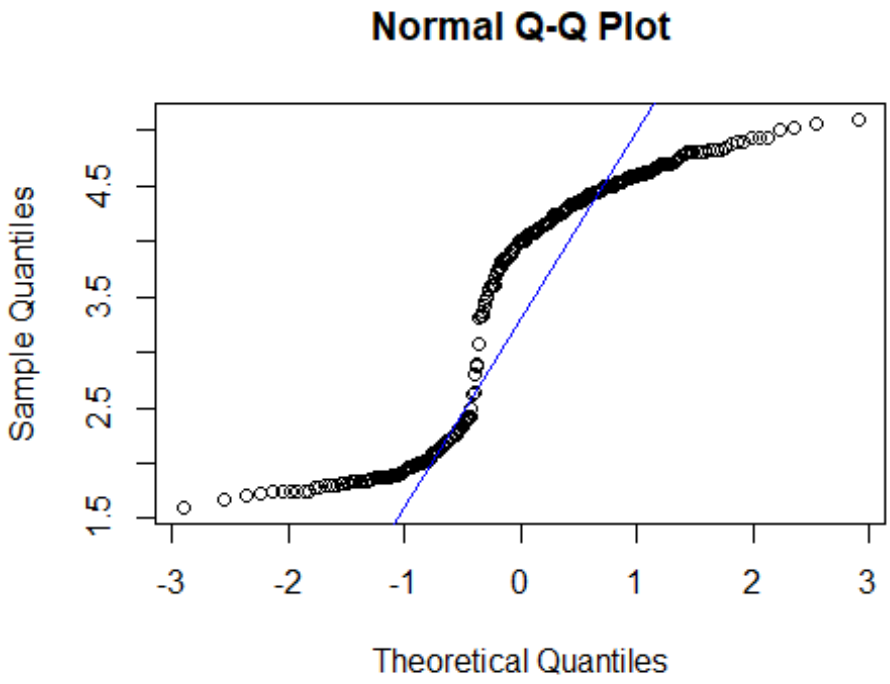
```r
# It can be seen that Scores data deviate from the line in the Q-Q plots. And also using the 68-95-99.7 rule, we get that this data is not likely to come from a normal distribution



# 2)a)

# The eruption variable in faithful dataset gives the numeric   Eruption time in mins (for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

# The len variable in ToothGrowth dataset numeric   Tooth length odontoblasts in 60 guinea pigs as a result of experiments on them where they received varying dose levels of Vitamin C using different delivery methods.

# (b) We can determine whether the eruption and len data are taken from a distribution that is close to normal distribution. We can make qq plot and compute proportion of data away 1,2,3 std dev from mean to use the 68-95-99.7 rule

# Plotting Q-Q plot for random values coming from normal distribution
x=rnorm(1000)
qqnorm(x)
```



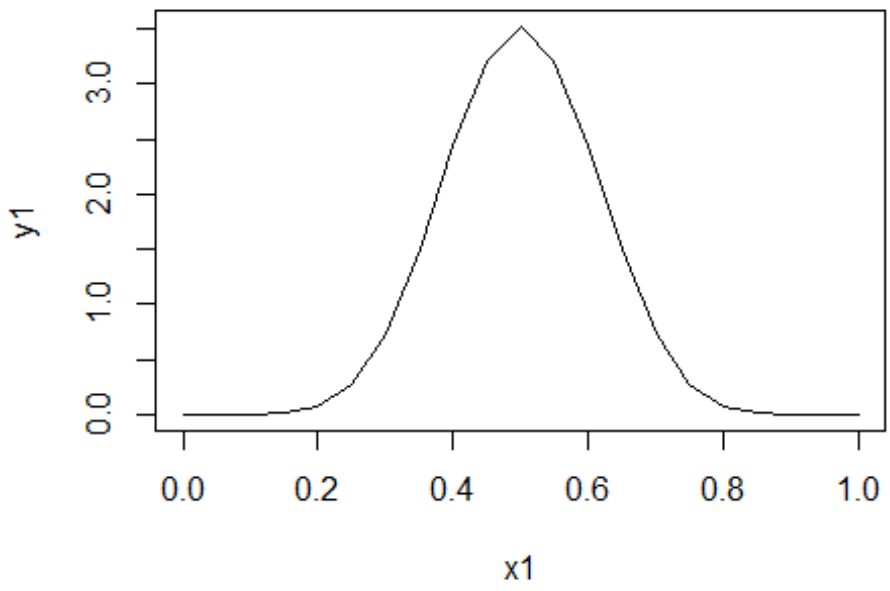

```
# Plotting Q-Q plot for len variable values in ToothGrowth
y = ToothGrowth
qqnorm(y$len)
qqline(y$len, col="blue") # adds a reference line
```



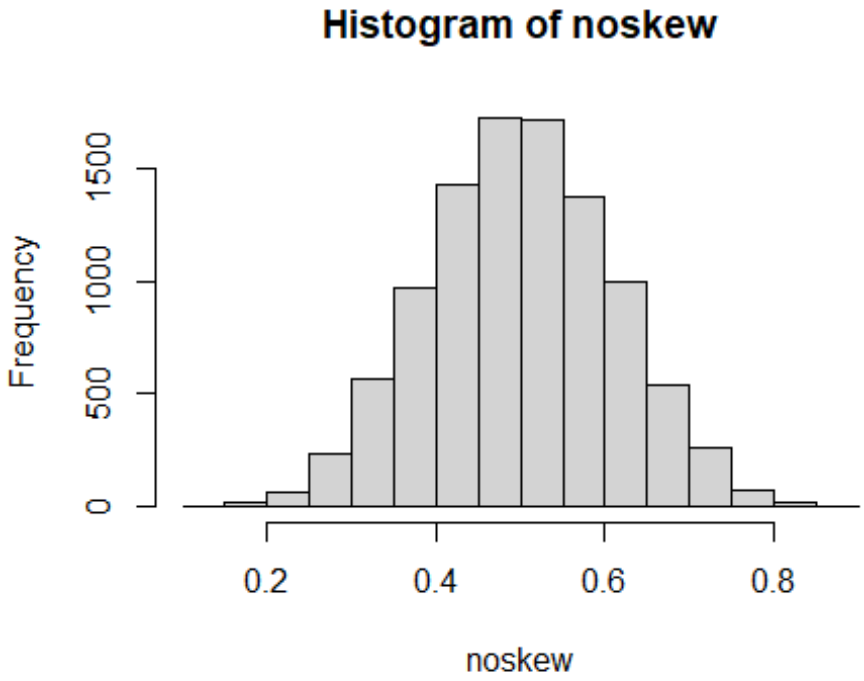

```
# Proportion of data within one sd from mean
Z=(y$len - mean(y$len))/sd(y$len)
sum(-1 < Z & Z < 1)/length(y$len)

## [1] 0.6666667

# Proportion of data within two sd from mean
sum(-2 < Z & Z < 2)/length(y$len)

## [1] 1

# Proportion of data within three sd from mean
sum(-3 < Z & Z < 3)/length(y$len)

## [1] 1

# Plotting Q-Q plot for eruption variable values in faithful
z = faithful
qqnorm(z$eruptions)
qqline(z$eruptions, col="blue") # adds a reference line
```



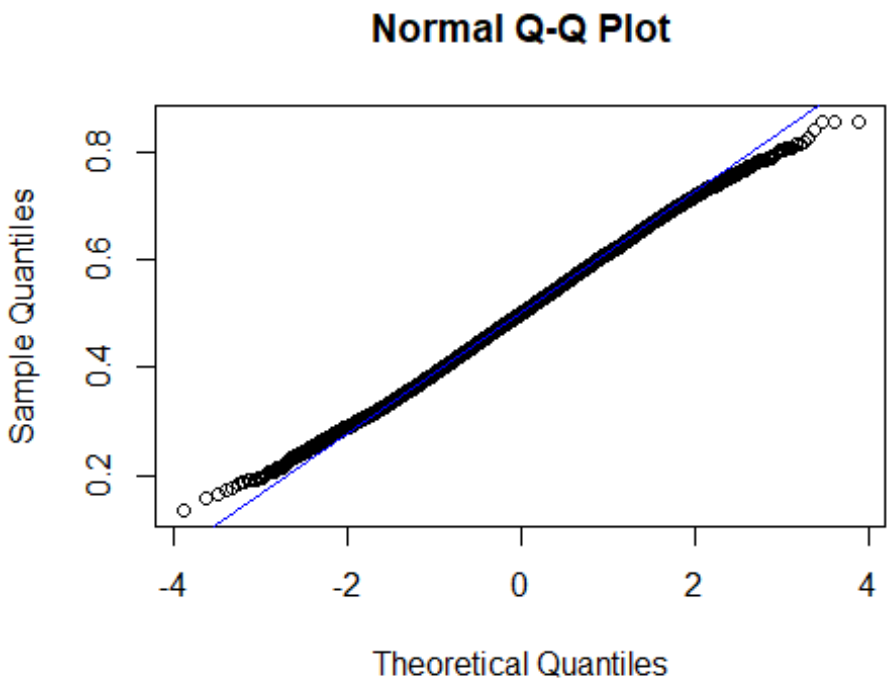

```
# Proportion of data within one sd from mean
Z1=(z$eruptions - mean(z$eruptions))/sd(z$eruptions)
sum(-1 < Z1 & Z1 <1)/length(z$eruptions)

## [1] 0.5514706

# Proportion of data within two sd from mean
sum(-2 < Z1 & Z1 <2)/length(z$eruptions)

## [1] 1

# Proportion of data within three sd from mean
sum(-3 < Z1 & Z1 <3)/length(z$eruptions)

## [1] 1

# It can be seen that they deviate from the line in the Q-Q plots. In case of
eruption variable there is significant deviations the proportions are not
close to the 68-95-99.7 rule, so these data points are not likely to come
from a normal distribution

#3)

# Beta(10,10)
x1 = seq(0,1, by = 0.05) #Creates a vector containing values from 0 to 1 with
0.05 spacing
y1 = dbeta(x1, 10,10) #Evaluates the pdf of Beta(10,10) for x1
plot(x1,y1, type="l") # Plots the pdf of Beta(10,10) for the values in x1
```



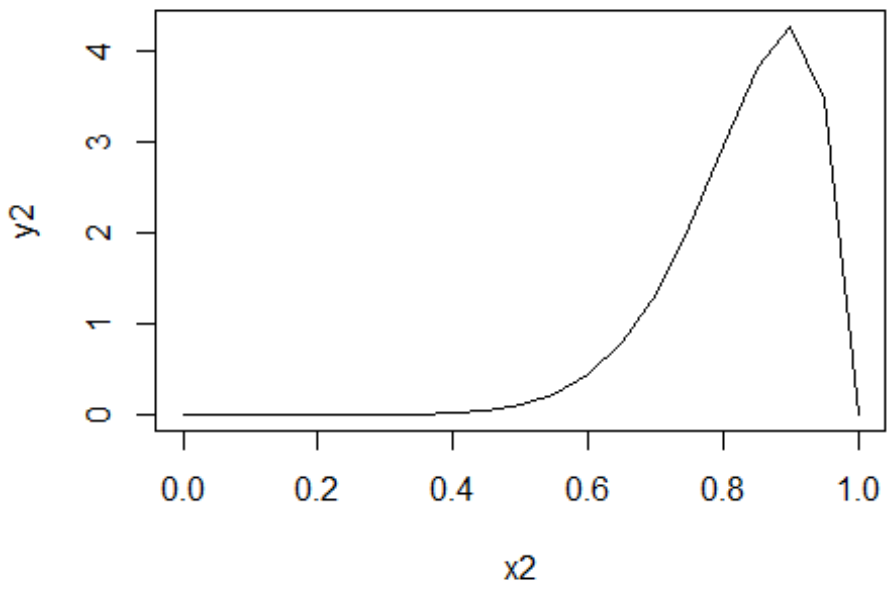

```
noskew = rbeta(10000,10,10) # Defines a vector of random values from
Beta(10,10) distribution
hist(noskew) # Creating histogram of the random deviates
```



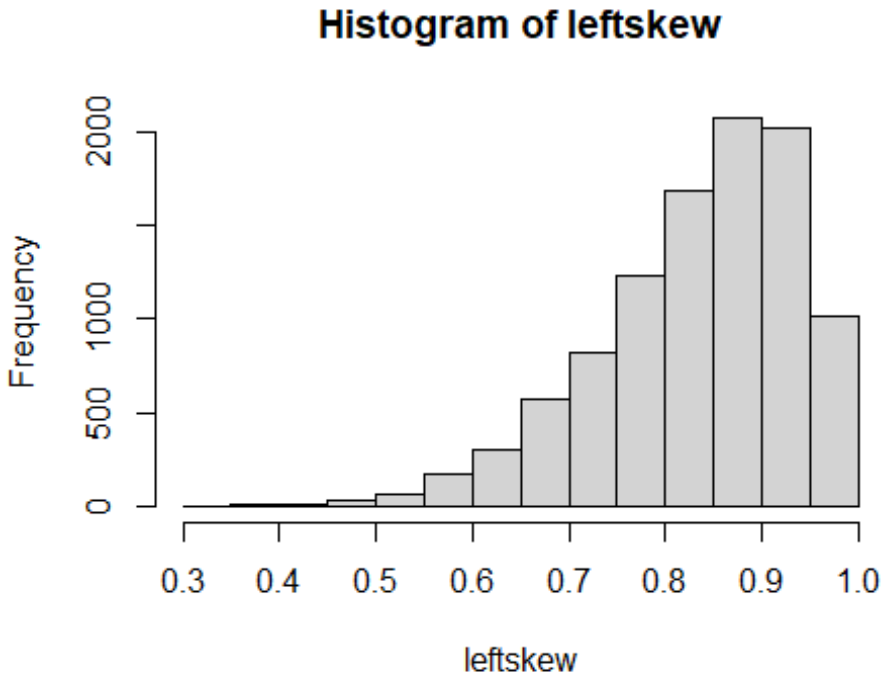
```r
qqnorm(noskew) # Creating Q-Q plot of the random deviates
qqline(noskew, col="blue")
```



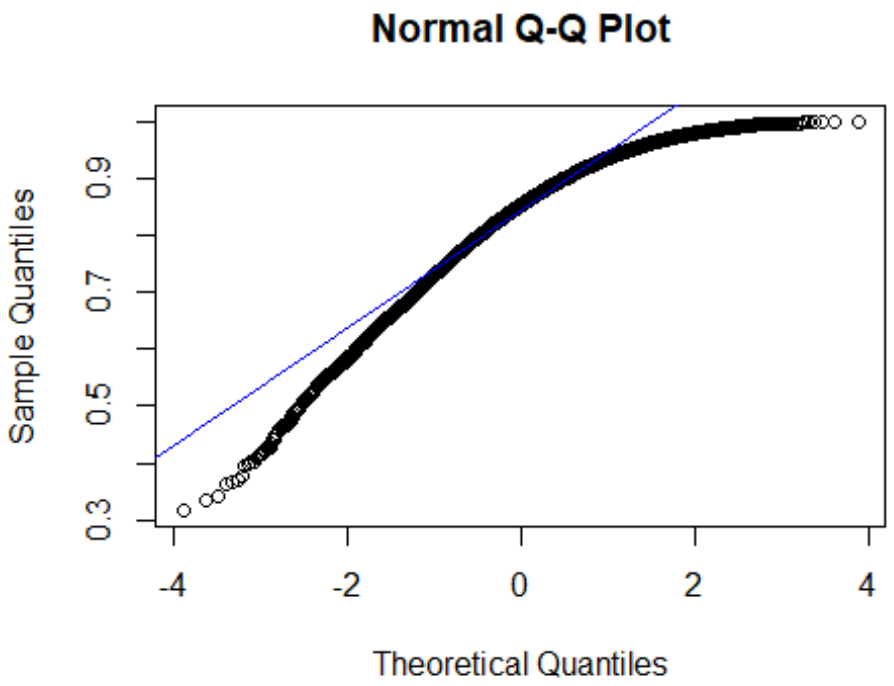

```r
# We observe that the plot and histogram are unskewed and is symmetric about mean just like Normal distribution.

# Beta(10,2)
x2 = seq(0,1, by = 0.05) # Creates the vector containing values from 0 to 1 with 0.05 spacing
y2 = dbeta(x2, 10,2) # Evaluates the pdf of Beta(10,2) for x2
plot(x2,y2, type="l") # Plots the pdf of Beta(10,2) for the values of vector x2
```



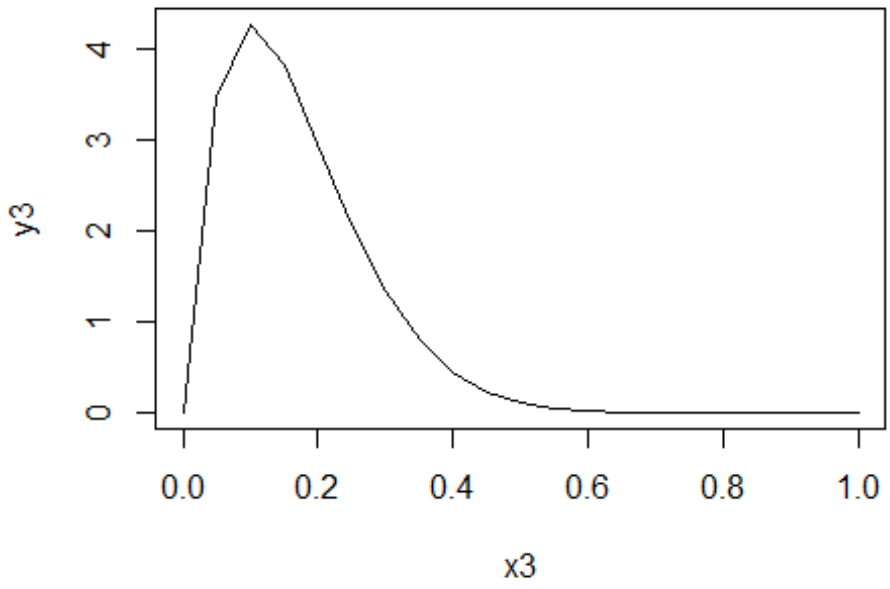

```
leftskew = rbeta(10000,10,2) # Defines a vector of random values from
Beta(10,2) distribution
hist(leftskew) # Creating histogram of the random deviates
```



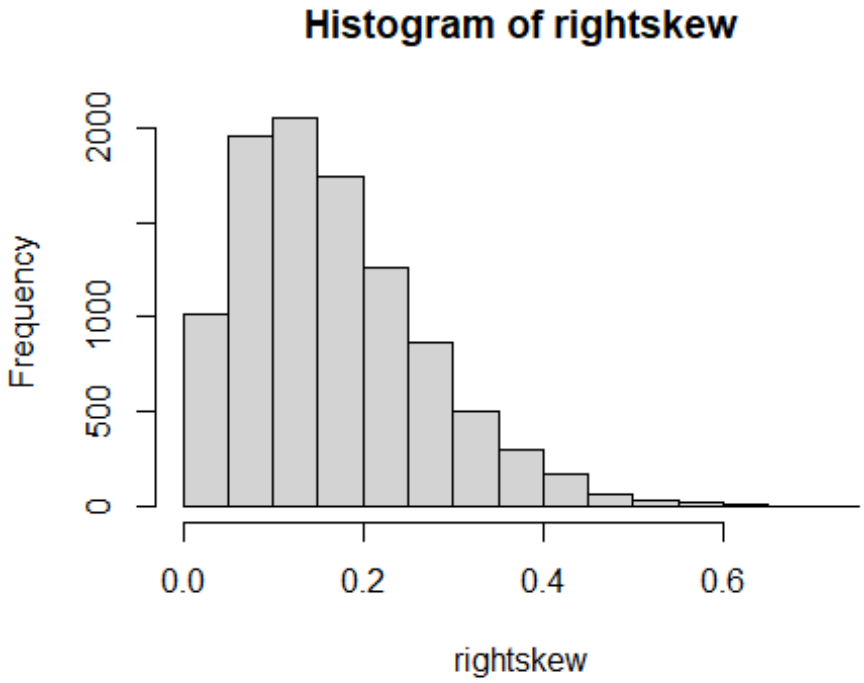
```r
qqnorm(leftskew) # Creating Q-Q plot of the random deviates
qqline(leftskew, col="blue")
```



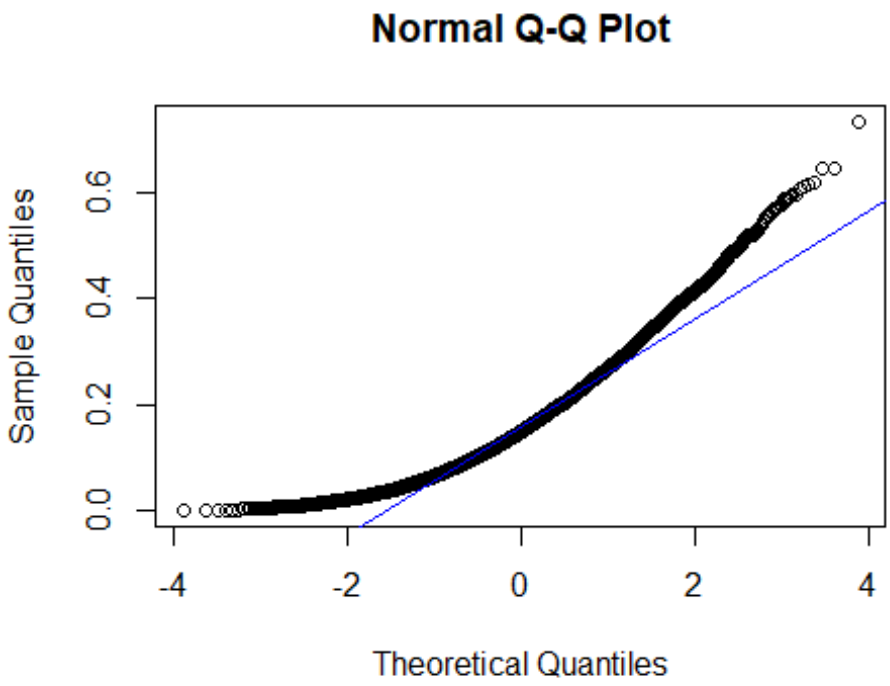

```r
# In this case the plot and histogram is left skewed and hence not normal.

# Beta(2,10)
x3 = seq(0,1, by = 0.05) # Creates a vector containing values from 0 to 1
with 0.05 spacing
y3 = dbeta(x3,2,10) # Evaluates the pdf of Beta(2,10) for x3
plot(x3,y3, type="l") # Plots the pdf of Beta(2,10) for the values of vector
x3
```




```
rightskew = rbeta(10000,2,10) # Defines a vector of random values from
Beta(2,10) distribution
hist(rightskew) # Creating histogram of the random deviates
```



```r
qqnorm(rightskew) # Creating Q-Q plot of the random deviates
qqline(rightskew, col="blue")
```




```r
# In this case the plot and histogram is right skewed and hence not normal
```

Problems due: 1,2

1. Suppose p is the unknown probability of an event A, and we estimate p by the sample proportion $\hat{p}$ based on an i.i.d. sample of size n.

   (a) Write $Var[\hat{p}]$ and $SD[\hat{p}]$ as functions of n and p.

   (b) Using the relations derived above, determine the sample size n, as a function of p, that is required to acheive $SD(\hat{p}) = 0.01$. How does this required value of n vary with p?

   (c) Design and implement the following simulation study to verify this behaviour. For $p = 0.01$, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99,

      (i) Simulate 1000 values of $\hat{p}$ with $n = 500$.

      (ii) Simulate 1000 values of $\hat{p}$ with n chosen according to the formula derived above.

      In each case, you can think of the 1000 values as i.i.d. samples from the distribution of $\hat{p}$, and use the sample standard deviation as an estimate of $SD[\hat{p}]$. Plot the estimated values of $SD(\hat{p})$ against p for both choices of n.

2. Consider Poisson λ distribution.

   (a) Show that both the sample mean and the sample variance of a sample obtained from the Poisson(λ) distribution will be unbiased estimators of λ.

   (b) For $\lambda = 10, 20, 50$ simulate $100, 500, 1000$ random observations from the Poisson(λ) distribution for various values of λ using the inbuilt function `rpois`.

   (c) Explore the behaviour of the two estimates for each λ as well as three sample sizes.

3. Biologists use a technique called "capture-recapture" to estimate the size of the population of a species that cannot be directly counted.

   Suppose the unknown population size is N, and fifty members of the species are selected and given an identifying mark. Sometime later a sample of size twenty is taken from the population, and it is found to contain X of the twenty previously marked. Equating the proportion of marked members in the second sample and the population, we have $\frac{X}{20} = \frac{50}{N}$, giving an estimate of $\hat{N} = \frac{1000}{X}$.

   (a) Show that the distribution of X has a hypergeometric distribution that involves N as a parameter.

   (b) Using the function `rhyper`. For each $N = 50$, 100, 200, 300, 400, and 500, simulate 1000 values of $\hat{N}$ and use them to estimate $E[\hat{N}]$ and $Var[\hat{N}]$. Plot these estimates as a function of N.

M.Sc. Data Science
Probability and Statistics with R - Homework 9

Name: Soham Biswas
Roll No.: MDS202147
Mail ID: sohamb@cmi.ac.in

1. Given that p is the unknown probability of an event A, and we estimate p by the sample proportion $\hat{p}$ based on an i.i.d. sample of size n.

   (a) We need to write $Var[\hat{p}]$ and $SD[\hat{p}]$ as functions of n and p.

   Let X be a random variable denoting the success of the event A.

   Therefore, $P(X \in A) = p$

   Let us consider $X_1, X_2, ..., X_n$ to be i.i.d. samples with the same distribution as a random variable X.

   Therefore, $\hat{p} = \dfrac{\#\{X_i \in A\}}{n} \qquad i = 1(1)n$

   Now, let $W = \#\{X_i \in A\}$. That is, W is a random variable denoting the success of the event A in n i.i.d. trials.

   Therefore, $W \sim Bin(n, p)$

   $$E[\hat{p}] = E\left[\frac{W}{n}\right] = \frac{1}{n}E[W] = \frac{1}{n} \cdot np = p = P(X \in A)$$

   $$Var[\hat{p}] = Var\left[\frac{W}{n}\right] = \frac{1}{n^2}Var[W] = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$$

   $$SD[\hat{p}] = \sqrt{Var[\hat{p}]} = \sqrt{\frac{p(1-p)}{n}}$$

   (b) Using the relations derived above, we are required to determine the sample size n, as a function of p, that is required to achieve $SD(\hat{p}) = 0.01$. We also need to state how this required value of n varies with p.

   That is, $\sqrt{\dfrac{p(1-p)}{n}} = 0.01$

   Squaring on both sides, we get:

   $$\sqrt{\frac{p(1-p)}{n}}^2 = 0.01^2$$

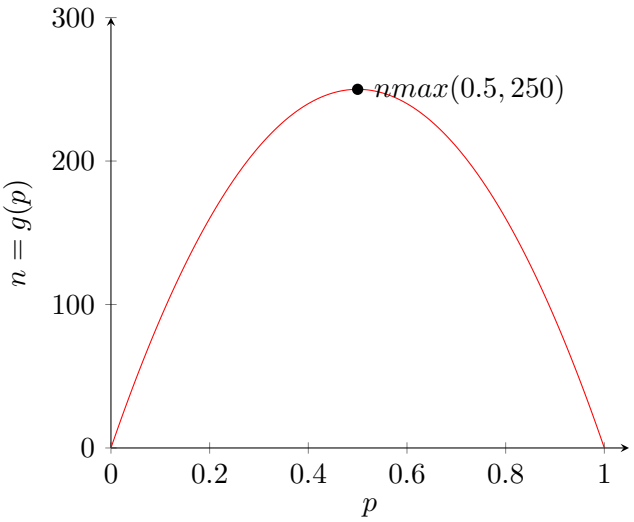
   $$\implies \frac{p(1-p)}{n} = 0.0001$$

   $$\implies n = 10000p(1-p)$$

Therefore, we can write n as a function (say g) of p as $n = g(p) = 10000p(1-p)$

Since p denotes the probability of success of the event A, $0 \leq p \leq 1$

The following graph is sketched to see how n varies with p based on the equation derived for the given conditions above.



So, we observe that n increases from 0 to 250 as p goes from 0 to 0.5 and then symmetrically decreases down to 0 as the value of p moves from 0.5 to 1. A maximum value of n(= 250) is seen when $p = 0.5$.

(c) We are required to design and implement the following simulation study to verify this behaviour, for $p = 0.01, 0.1, 0.25, 0.5, 0.75, 0.9$, and 0.99.

i. Simulate 1000 values of $\hat{p}$ with $n = 500$.

ii. Simulate 1000 values of $\hat{p}$ with n chosen according to the formula derived above.

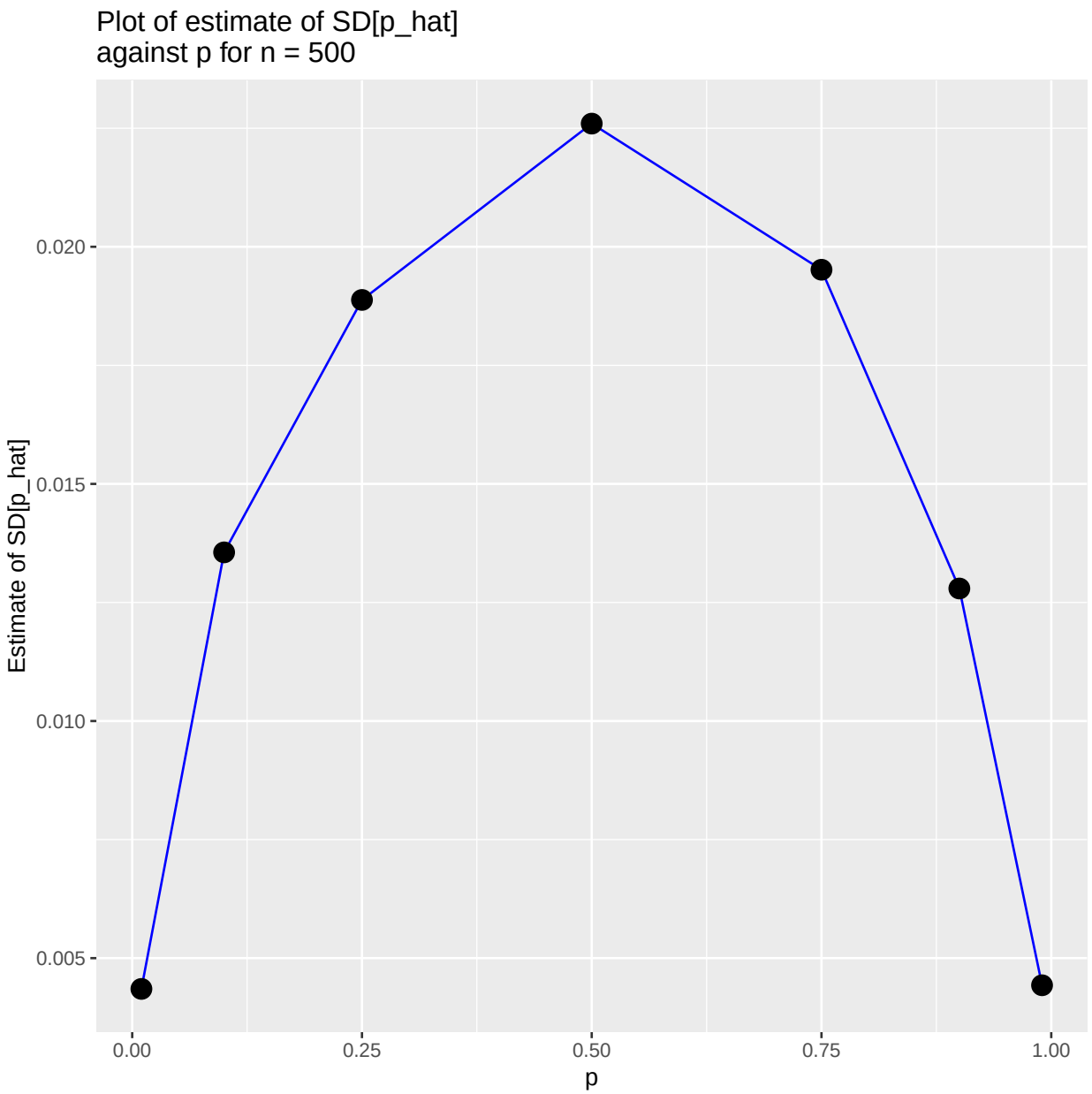
In each case,we are to consider the 1000 values as i.i.d. samples from the distribution of $\hat{p}$, and use the sample standard deviation as an estimate of $SD[\hat{p}]$. We then need to plot the estimated values of $SD[\hat{p}]$ against p for both choices of n.

i. First we simulate 1000 values of $\hat{p}$ with $n = 500$.

We generate $n = 500$ samples from the Bernoulli(p) distribution and find the sample proportion $\hat{p}$. We then find the corresponding sample standard deviation as an estimate of $SD[\hat{p}]$. This step is repeated 1000 times for the given values of p and then a plot of the estimate of $SD[\hat{p}]$ against p is displayed.

```
library(ggplot2)
p = c(0.01,0.1,0.25,0.5,0.75,0.9,0.99)
sd_500 = array(dim = 1)
p_hat = matrix(,nrow = 1000,ncol = 7)
for(i in 1:7)
```

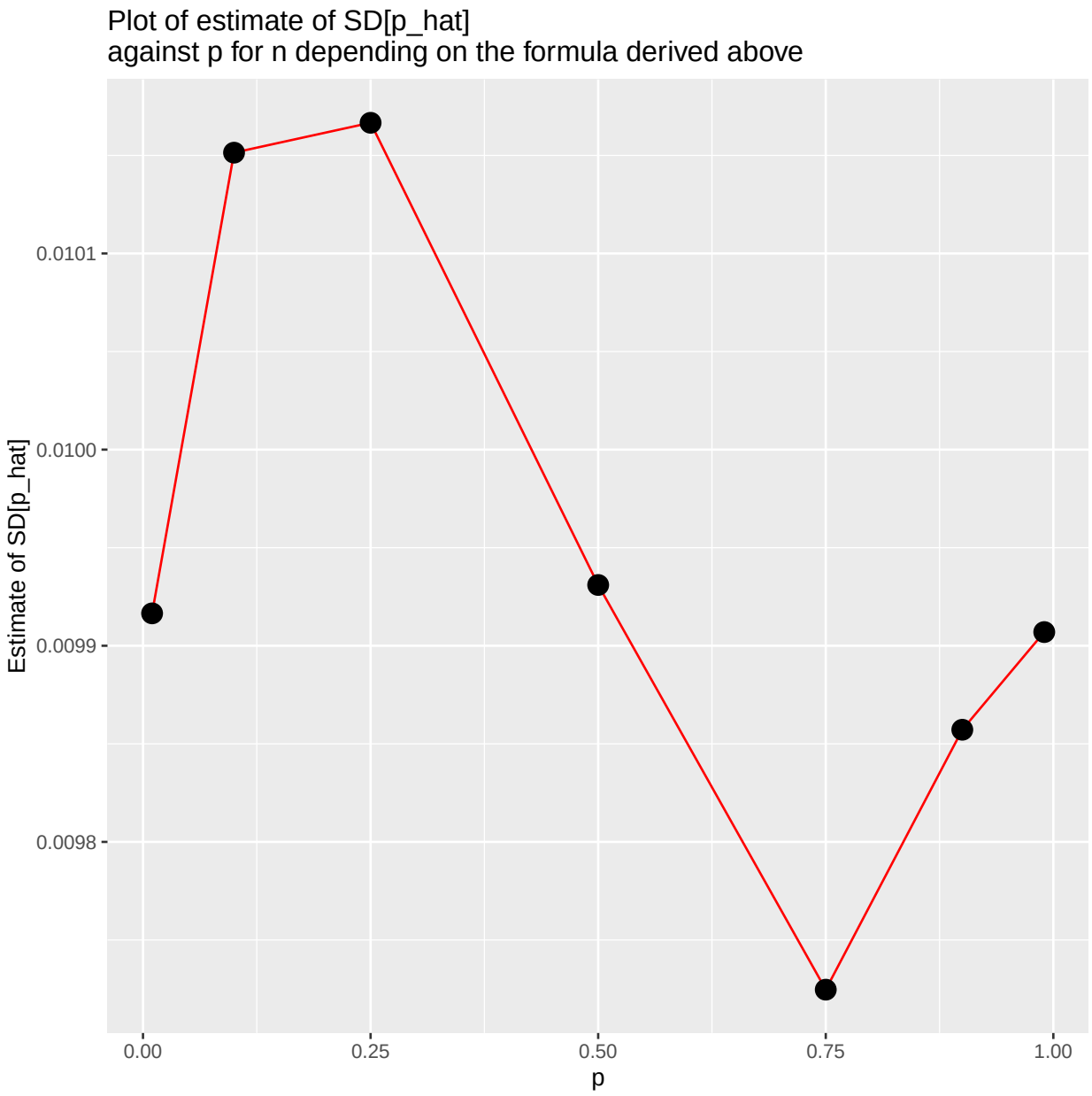
```
   {
     for(j in 1:1000)
       {
          p_hat[j,i] = sum(rbinom(500,1,p[i]))/500
       }
     sd_500[i] = sd(p_hat[,i])
   }
df_500 = data.frame(p,sd_500)

ggplot(data = df_500, aes(x = p, y = sd_500)) +
geom_line(color = "blue") +
labs(title = "Plot of estimate of SD[p_hat]
against p for n = 500", x = "p", y = "Estimate of SD[p_hat]") +
geom_point(size=4, color = "black")
```



ii. Next we simulate 1000 values of $\hat{p}$ with n chosen according to the formula derived above.

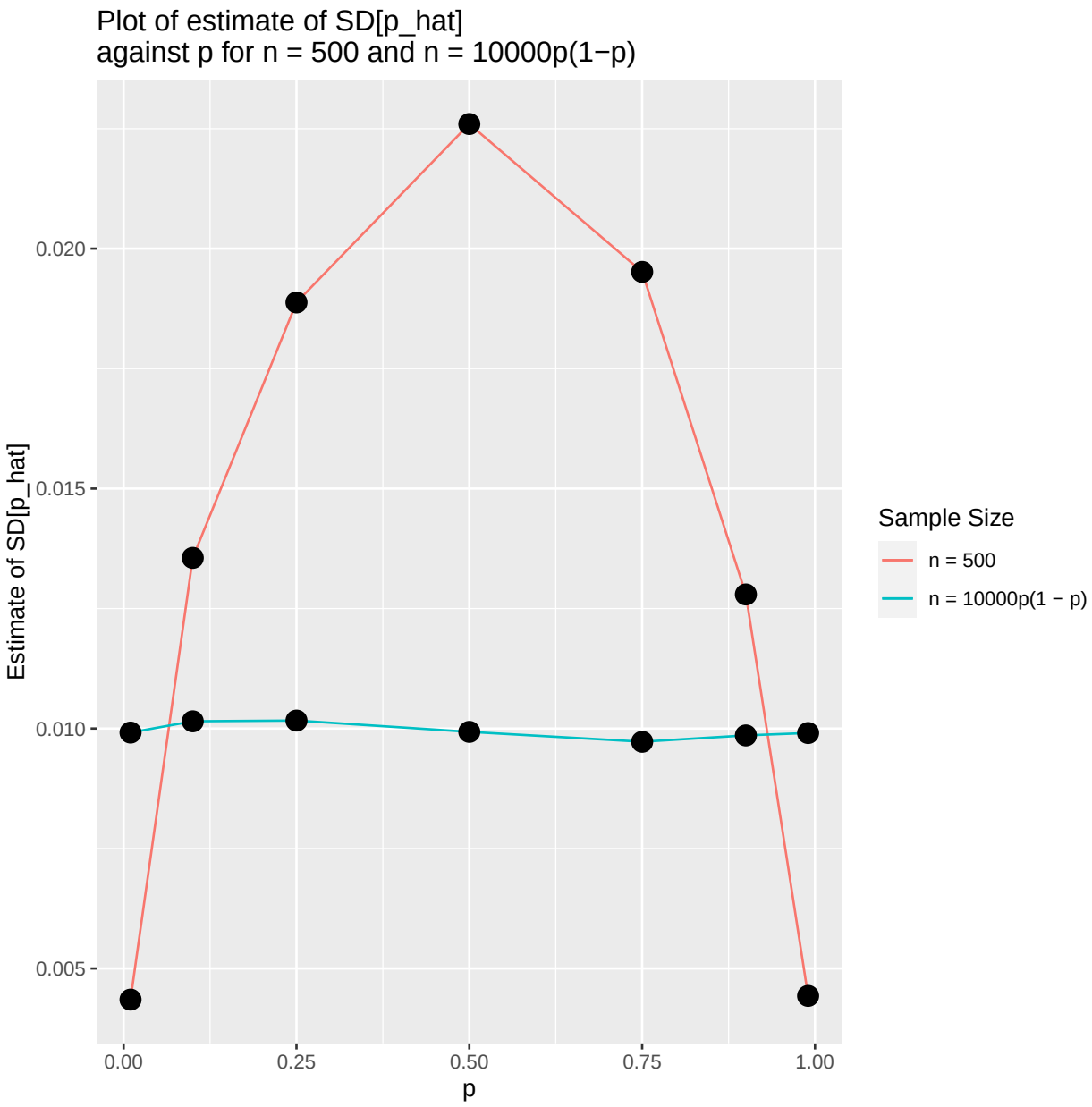
We generate samples of size based on the formula derived above from the Bernoulli(p) distribution and find the sample proportion $\hat{p}$. We then find the corresponding sample standard deviation as an estimate of $SD[\hat{p}]$. This step is repeated 1000 times for the given values of p and then a plot of the estimate of $SD[\hat{p}]$ against p is displayed.

```
library(ggplot2)
p = c(0.01,0.1,0.25,0.5,0.75,0.9,0.99)
sd_n = array(dim = 1)
p_hat_n = matrix(,nrow = 1000,ncol = 7)
for(i in 1:7)
   {
     for(j in 1:1000)
       {
          n = 10000*p[i]*(1 - p[i])
          p_hat_n[j,i] = sum(rbinom(n,1,p[i]))/n
       }
     sd_n[i] = sd(p_hat_n[,i])
   }
df_n = data.frame(p,sd_n)

ggplot(data = df_n, aes(x = p, y = sd_n)) +
geom_line(color = "red") +
labs(title = "Plot of estimate of SD[p_hat]
against p for n depending on the formula derived above", x = "p",
y = "Estimate of SD[p_hat]") +
geom_point(size=4, color = "black")
```



The 2 graphs are also plotted together for better comparision.

```
df_combined = data.frame(p,sd_500,sd_n)
p1 = ggplot(data = df_combined, aes(x = p)) +
geom_line(aes(y = sd_500,color = "blue")) +
geom_line(aes(y = sd_n,color = "red")) +
labs(title = "Plot of estimate of SD[p_hat]
against p for n = 500 and n = 10000p(1-p)", x = "p",
y = "Estimate of SD[p_hat]") +
geom_point(aes(y = sd_n),size=4) +
geom_point(aes(y = sd_500),size=4) +
scale_color_discrete(name = "Sample Size",labels=c("n = 500","n = 10000p(1 - p)"))

p1
```



Since in the second case the sample size n was calculated such that an $SD = 0.01$ was achieved corresponding to the given p, we see that the estimate of SD$[\hat{p}]$ is more or less around 0.01, while when the sample size is fixed at 500, we observe that the estimate of SD$[\hat{p}]$ is increasing as the value of p increases from 0 to 0.5 and again symmetrically decreases as the value of p increases from 0.5 to 1. We observe that a symmetric, bell-shaped curve is formed here and it seems to follow a normal distribution.

2. We are required to consider the Poisson λ distribution.

   (a) We are required to show that both the sample mean and the sample variance of a sample obtained from the $Poisson(\lambda)$ distribution will be unbiased estimators of λ.

   Let us consider $X_1, X_2, ..., X_n$ to be i.i.d. random samples from a Poisson(λ) distribution, $\lambda > 0$.

Sample Mean$(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample Variance$(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

Since, $X_1, X_2, ..., X_n \sim$ Poisson(λ) distribution and are i.i.d., $E[X_i] = \lambda$ and $V[X_i] = \lambda \; \forall \; i = 1(1)n$

Now, $E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\lambda = \frac{n\lambda}{n} = \lambda$

Therefore, we can conclude that sample mean is an unbiased estimator of λ for the Poisson(λ) distribution.

Also, $Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\lambda = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$

Now, $E\left[S^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \lambda + \lambda - \bar{X}\right)^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left\{\left(X_i - \lambda\right) - \left(\bar{X} - \lambda\right)\right\}^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left\{\left(X_i - \lambda\right)^2 + \left(\bar{X} - \lambda\right)^2 - 2\left(X_i - \lambda\right)\left(\bar{X} - \lambda\right)\right\}\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left\{\left(X_i - \lambda\right)^2\right\} + \sum_{i=1}^{n}\left\{\left(\bar{X} - \lambda\right)^2\right\} - \sum_{i=1}^{n}\left\{2\left(X_i - \lambda\right)\left(\bar{X} - \lambda\right)\right\}\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \lambda\right)^2 + n\left(\bar{X} - \lambda\right)^2 - 2\left(\bar{X} - \lambda\right)\sum_{i=1}^{n}\left(X_i - \lambda\right)\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \lambda\right)^2 + n\left(\bar{X} - \lambda\right)^2 - 2\left(\bar{X} - \lambda\right)\left\{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\lambda\right\}\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \lambda\right)^2 + n\left(\bar{X} - \lambda\right)^2 - 2\left(\bar{X} - \lambda\right)\left(n\lambda - n\lambda\right)\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \lambda\right)^2 + n\left(\bar{X} - \lambda\right)^2 - 2n\left(\bar{X} - \lambda\right)^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \lambda\right)^2 - n\left(\bar{X} - \lambda\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left\{\left(X_i - \lambda\right)^2\right\} - E\left\{n\left(\bar{X} - \lambda\right)^2\right\}\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} Var\left(X_i\right) - nVar\left(\bar{X}\right)\right]$$

$$= \frac{1}{n-1}\left(n\lambda - n\frac{\lambda}{n}\right)$$

$$= \frac{1}{n-1}\lambda\left(n-1\right)$$

$$= \lambda$$

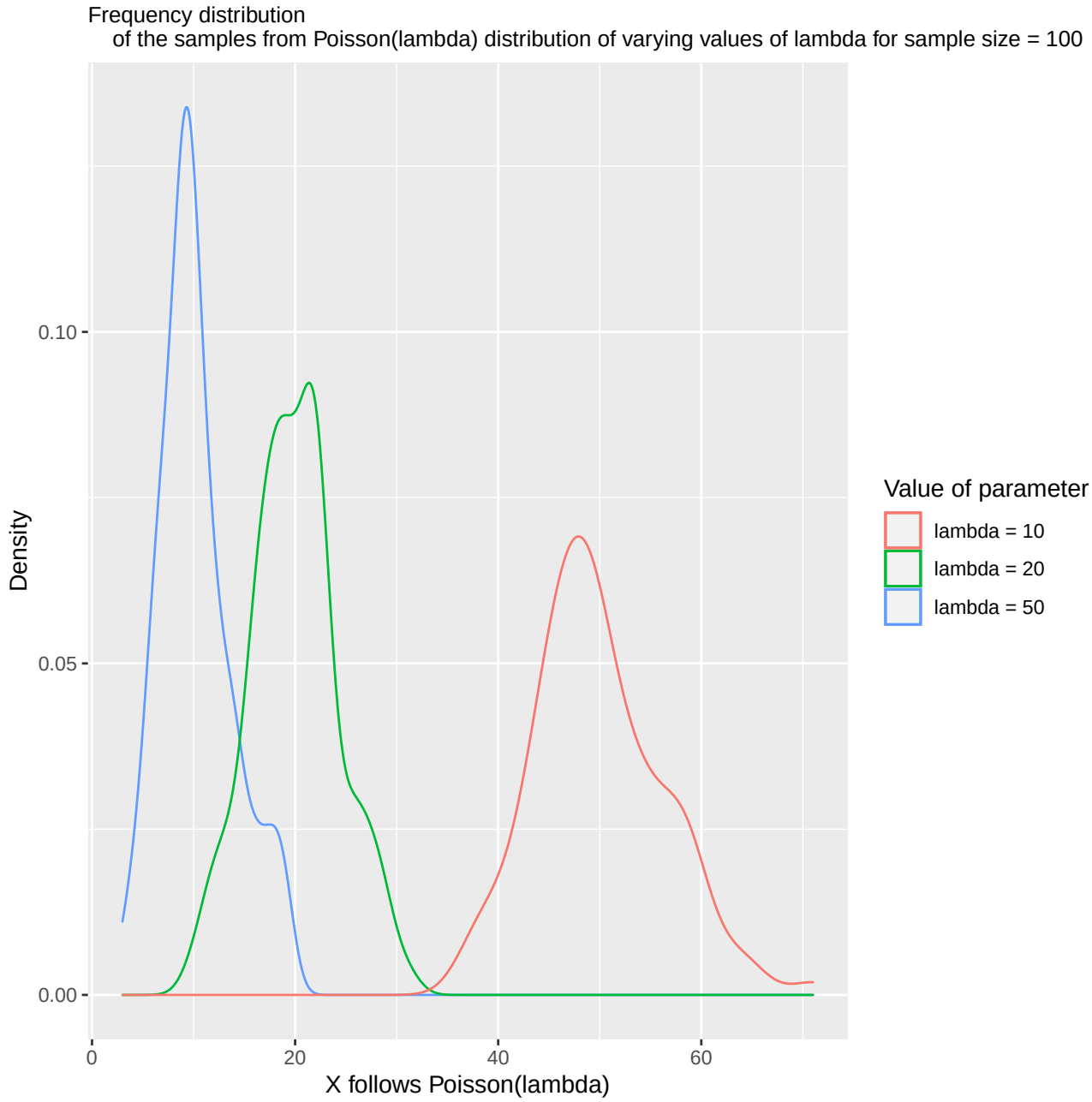
Therefore, we can conclude that sample variance is an unbiased estimator of λ for the Poisson(λ) distribution.

(b) For $\lambda = 10, 20, 50$ we are required to simulate $100, 500, 1000$ random observations from the Poisson() distribution for various values of λ using the inbuilt function rpois.

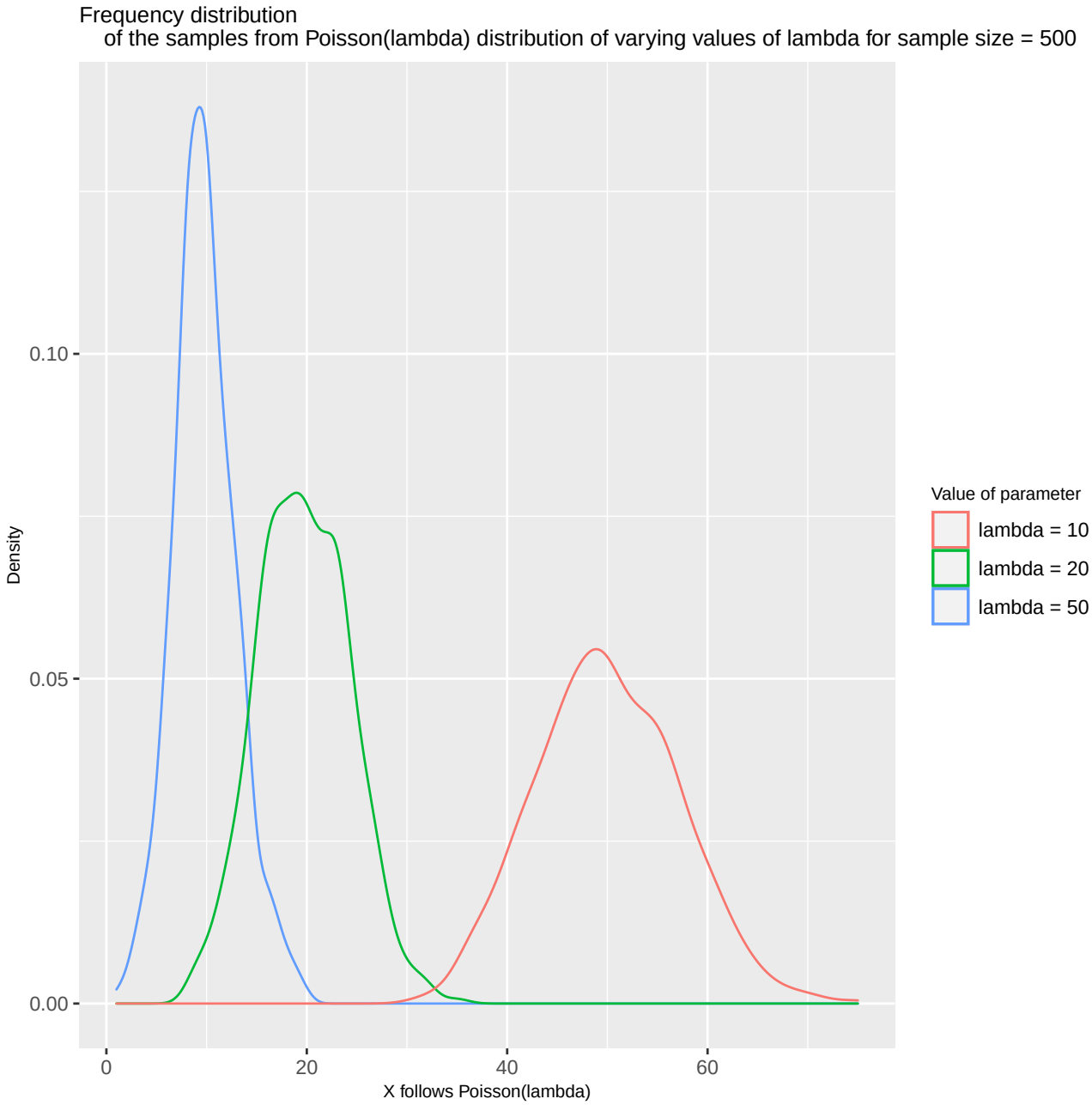
```
lambda = c(10, 20, 50)
n = c(100, 500, 1000)
samp_100 = matrix(,nrow = 100,ncol = 3)
samp_500 = matrix(,nrow = 500,ncol = 3)
samp_1000 = matrix(,nrow = 1000,ncol = 3)

for(i in 1:3)
  {
   samp_100[,i] = rpois(100,lambda[i])
   samp_500[,i] = rpois(500,lambda[i])
   samp_1000[,i] = rpois(1000,lambda[i])
  }
```

We now plot these samples:

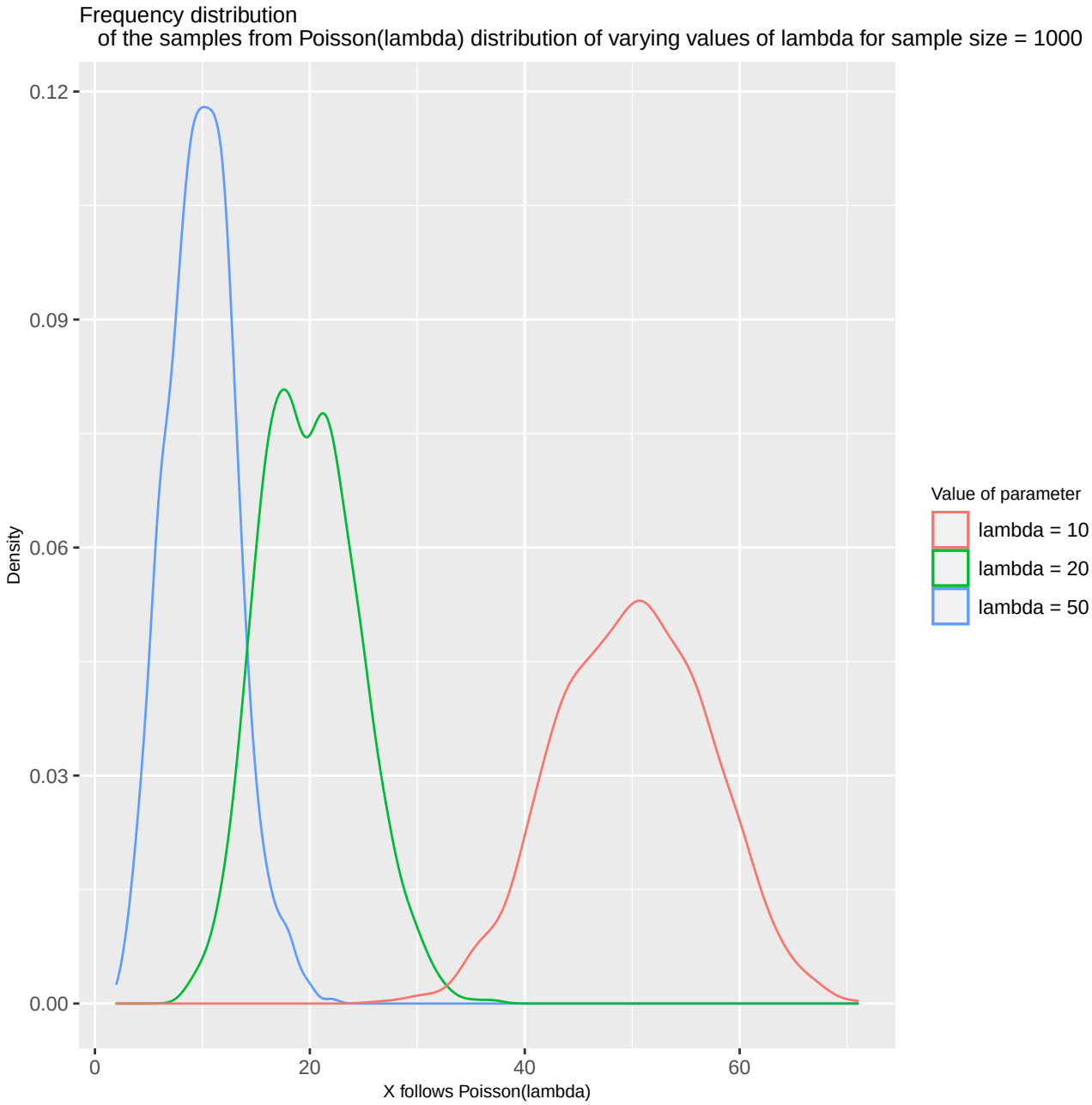
```
df_10 = data.frame(samp_100[,1],samp_100[,2],samp_100[,3])
ggplot(data = df_10) +
    geom_density(aes(x = samp_100[,1],color = "red"),stat = "density") +
    geom_density(aes(x = samp_100[,2],color = "blue"),stat = "density") +
    geom_density(aes(x = samp_100[,3],color = "black"),stat = "density") +
    labs(title = "Frequency distribution
    of the samples from Poisson(lambda) distribution of varying values of lambda for sample
    x = "X follows Poisson(lambda)",y = "Density") +
    scale_color_discrete(name = "Value of parameter",labels=c("lambda = 10","lambda = 20","l
    theme(plot.title = element_text(size = 10))
```




```
df_20 = data.frame(samp_500[,1],samp_500[,2],samp_500[,3])
ggplot(data = df_20) +
    geom_density(aes(x = samp_500[,1],color = "red"),stat = "density") +
    geom_density(aes(x = samp_500[,2],color = "blue"),stat = "density") +
    geom_density(aes(x = samp_500[,3],color = "black"),stat = "density") +
    labs(title = "Frequency distribution
    of the samples from Poisson(lambda) distribution of varying values of lambda for sample
    x = "X follows Poisson(lambda)",y = "Density") +
    scale_color_discrete(name = "Value of parameter",
    labels=c("lambda = 10",
    "lambda = 20","lambda = 50")) +
    theme(title = element_text(size = 8))
```




```
      df_30 = data.frame(samp_1000[,1],samp_1000[,2],samp_1000[,3])
      ggplot(data = df_30) +
geom_density(aes(x = samp_1000[,1],color = "red"),stat = "density") +
geom_density(aes(x = samp_1000[,2],color = "blue"),stat = "density") +
geom_density(aes(x = samp_1000[,3],color = "black"),stat = "density") +
labs(title = "Frequency distribution
of the samples from Poisson(lambda) distribution of varying values of lambda for sample s
x = "X follows Poisson(lambda)",y = "Density") +
scale_color_discrete(name = "Value of parameter",
labels=c("lambda = 10",
"lambda = 20","lambda = 50")) +
theme(title = element_text(size = 8))
```



We observe that as the sample size increases, the distribution seems to tend to normality.

(c) We need to explore the behaviour of the two estimates for each λ as well as three sample sizes.

We first find the sample mean and sample variance of all the cases and compare them with the true mean and true variance.

Now, the true mean for a Poisson(λ) distribution is λ

And the true variance for a Poisson(λ) distribution is also λ

We now calculate and tabulate the estimates of the sample means and sample variance for each n and λ by repeating the above process 500 times.

```r
lambda = c(10, 20, 50)
n = c(100, 500, 1000)
means_100 = matrix(,nrow = 100,ncol = 3)
means_500 = matrix(,nrow = 500,ncol = 3)
means_1000 = matrix(,nrow = 1000,ncol = 3)
for(j in 1:1000)
  {
    for(i in 1:3)
     {
       means_100[,i] = mean(rpois(100,lambda[i]))
       means_500[,i] = mean(rpois(500,lambda[i]))
       means_1000[,i] = mean(rpois(1000,lambda[i]))
     }
  }
A = c(mean(means_100[,1]),mean(means_100[,2]),
mean(means_100[,3]))
B = c(mean(means_500[,1]),mean(means_500[,2]),
mean(means_500[,3]))
C = c(mean(means_1000[,1]),mean(means_1000[,2]),
mean(means_1000[,3]))
df_means = data.frame(c("lambda = 10",
"lambda = 20","lambda = 50"),A,B,C)
names(df_means) = c("lambda / Sample Size",
"n = 100", "n = 500", "n = 1000")
df_means
```

```
##   lambda / Sample Size n = 100 n = 500 n = 1000
## 1          lambda = 10    9.76  10.110   10.036
## 2          lambda = 20   19.53  19.812   19.677
## 3          lambda = 50   50.00  50.168   50.236
```

```r
lambda = c(10, 20, 50)
n = c(100, 500, 1000)
var_100 = matrix(,nrow = 100,ncol = 3)
var_500 = matrix(,nrow = 500,ncol = 3)
var_1000 = matrix(,nrow = 1000,ncol = 3)
for(j in 1:1000)
  {
    for(i in 1:3)
     {
       var_100[,i] = var(rpois(100,lambda[i]))
       var_500[,i] = var(rpois(500,lambda[i]))
       var_1000[,i] = var(rpois(1000,lambda[i]))
     }
  }
A = c(mean(var_100[,1]),mean(var_100[,2]),
mean(var_100[,3]))
B = c(mean(var_500[,1]),mean(var_500[,2]),
mean(var_500[,3]))
C = c(mean(var_1000[,1]),mean(var_1000[,2]),
```

```
mean(var_1000[,3]))
df_means = data.frame(c("lambda = 10",
"lambda = 20","lambda = 50"),A,B,C)
names(df_means) = c("lambda / Sample Size",
"n = 100", "n = 500", "n = 1000")
df_means
```

```
##   lambda / Sample Size   n = 100   n = 500 n = 1000
## 1          lambda = 10  8.781717  9.728942 10.21021
## 2          lambda = 20 19.836263 20.083090 20.52653
## 3          lambda = 50 42.410000 51.619154 50.66766
```

The sample sizes $n = 100, 500, 1000$ are itself sufficiently big. But we can still see a general trend that as the sample size increases the estimate of the sample means seem to converge to the true mean that is λ slightly more accurately.

The general trend of a slight increase in the accuracy of the convergence as the sample size increases for the sample variance to the true variance, that is λ is once again seen for the variance case.

These results validate the proofs in part (a)

———X———

1. Suppose X follows Bernoulli(p) distribution. Let $p = 1/3$
   (a) Simulate for $n = 100$ $X_1, X_2, X_3, \ldots X_n$ i.i.d X
   (b) Demonstrate the Law of Large numbers by plotting the sample mean $\bar{X}_n$ as a function of n.
   (c) Using `replicate` command plot 15 independent trials of the above.
   (d) Do the same when $p = 0.001$, $n = 100$, $p = 0.5, n = 100$, $p = 0.99$ on different plots

2. We wish to compute
$$\int_a^b f(x)dx$$
using the Law of Large numbers.
   (a) Generate samples of $X_1, X_2, \ldots X_n$ i.i.d. Uniform (a, b). Justify
$$(b-a)\sum_{i=1}^{n} \frac{f(X_i)}{n} \approx \int_a^b f(x)dx$$
   (b) Write an `R`-code to estimate the $\int_0^7 \frac{16+\sin(x)}{x^2+4}dx$ using the procedure described in the previous part with $n = 400$.
   (c) Repeat the estimate 100 times and find the mean of these 100 simulations.
   (d) Use the `integrate` command in `R` to evaluate the integral. Compare the two answers.

3. Simulate 500 samples from each of the below distributions using their respective distribution function F_X and/or the inbuilt `runif`.
   (a) $X \sim$ Poisson(10)
   (b) $X \sim$ p.d.f f given by
$$f(x) = \begin{cases} x & 0 \le x \le \sqrt{2} \\ 0 & \text{otherwise.} \end{cases}$$
   (c) X, Y i.i.d $\sim$ Normal $(3, 4)$

Problems due: 1,2

1. Suppose p is the unknown probability of an event A, and we estimate p by the sample proportion $\hat{p}$ based on an i.i.d. sample of size n.

   (a) Design and implement the following simulation study to verify this behaviour. For $p = 0.01$, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99,

   (i) Simulate 1000 values of $\hat{p}$ with $n = 500$.

   (ii) Simulate 1000 values of $\hat{p}$ with n chosen according to the formula derived above.

   In each case, you can think of the 1000 values as i.i.d. samples from the distribution of $\hat{p}$, and use the sample standard deviation as an estimate of $SD[\hat{p}]$. Plot the estimated values of $SD(\hat{p})$ against p for both choices of n.

2. Consider Poisson λ distribution.

   (a) Show that both the sample mean and the sample variance of a sample obtained from the Poisson(λ) distribution will be unbiased estimators of λ.

   (b) For $\lambda = 10, 20, 50$ simulate $100, 500, 1000$ random observations from the Poisson(λ) distribution for various values of λ using the inbuilt function `rpois`.

   (c) Explore the behaviour of the two estimates for each λ as well as three sample sizes.

3. Biologists use a technique called "capture-recapture" to estimate the size of the population of a species that cannot be directly counted.

   Suppose the unknown population size is N, and fifty members of the species are selected and given an identifying mark. Sometime later a sample of size twenty is taken from the population, and it is found to contain X of the twenty previously marked. Equating the proportion of marked members in the second sample and the population, we have $\frac{X}{20} = \frac{50}{N}$, giving an estimate of $\hat{N} = \frac{1000}{X}$.

   (a) Show that the distribution of X has a hypergeometric distribution that involves N as a parameter.

   (b) Using the function `rhyper`. For each $N = 50$, 100, 200, 300, 400, and 500, simulate 1000 values of $\hat{N}$ and use them to estimate $E[\hat{N}]$ and $Var[\hat{N}]$. Plot these estimates as a function of N.

4. Suppose p is the unknown probability of an event A, and we estimate p by the sample proportion $\hat{p}$ based on an i.i.d. sample of size n.

   (a) Write $Var[\hat{p}]$ and $SD[\hat{p}]$ as functions of n and p.

   (b) Using the relations derived above, determine the sample size n, as a function of p, that is required to acheive $SD(\hat{p}) = 0.01$. How does this required value of n vary with p?

Worksheet 9

Rishika Tibrewal

26/11/2021

Q1 a)

```
#generate a random sample of size 100 from Bin(1,1/3) or Ber(1/3)
rbinom(100,size=1,prob=1/3)
```

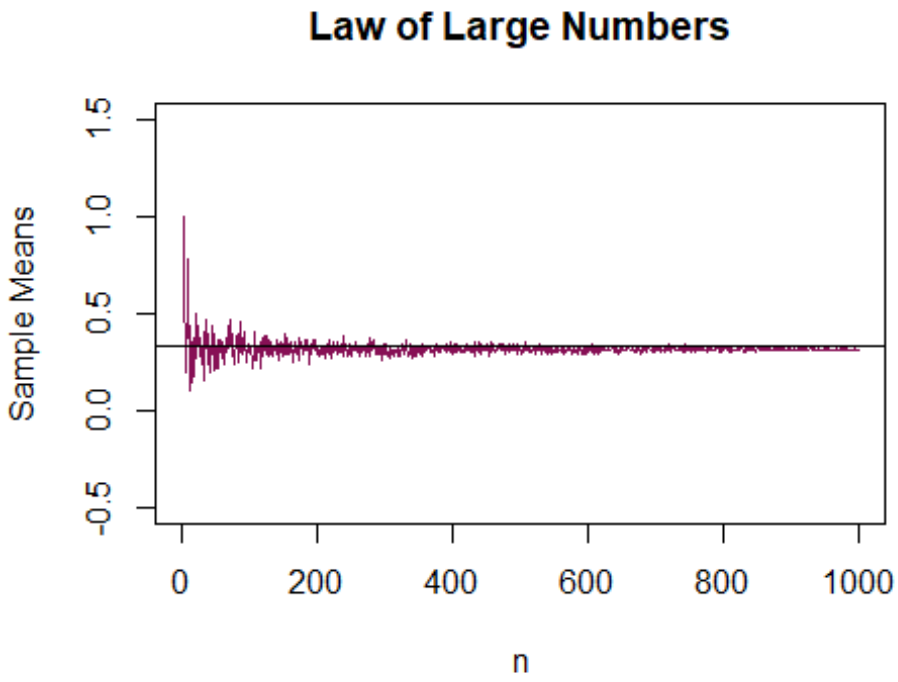
```
##   [1] 0 0 0 1 1 0 1 0 0 0 1 0 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 0 1 0 0 0
0 1 1
##  [38] 0 1 1 0 0 1 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0
1 1 0
##  [75] 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0
```

Q1 b)

```
#function to get the running mean and plot layout
runningmean=function(x,n,p){
  sam=sample(x,n,replace=TRUE,p)
  #taking sample of size n from x with probability p, with replacement
  cum=rep(NA,n)
  for (i in 1:n){
    cum[i]=sum(sample(sam,i,replace=FALSE))/i
  }
  par(new = 'TRUE')
  plot(cum, type='l',col=rgb(runif(3),runif(3),runif(3)), xlab='n', ylab='Sam
ple Means', ylim=c(-0.5,1.5),main="Law of Large Numbers")
}

runningmean(c(0,1),1000,c(2/3,1/3)) #plotting the graph of the running mean f
or n=1000 with p=1/3
```

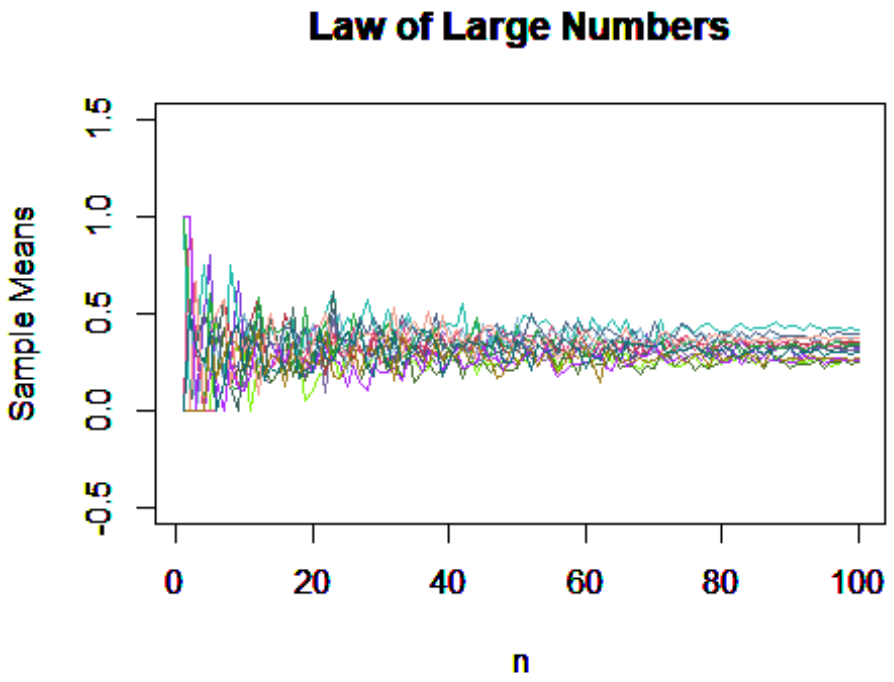
```
## Warning in par(new = "TRUE"): calling par(new=TRUE) with no plot
```

```
abline(h=1/3,col="black") #to get a horizontal line at y=1/3
```



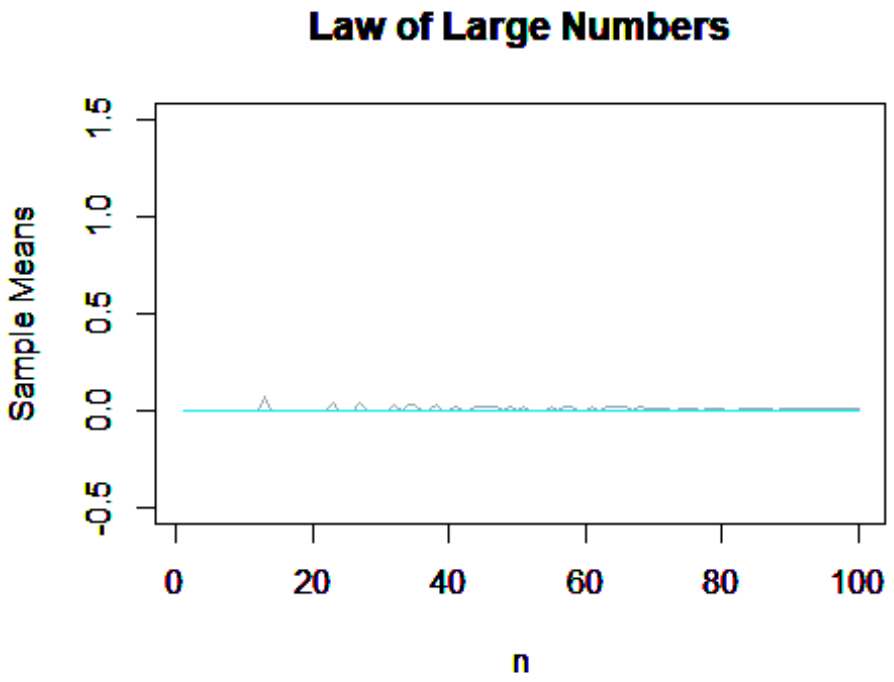
Q1 c)

```
replicate(15, (runningmean(c(0,1), 100, c(2/3,1/3)))) #15 iterations of plott
ing the graph of the running mean for n=100 with p=1/3

## Warning in par(new = "TRUE"): calling par(new=TRUE) with no plot
```

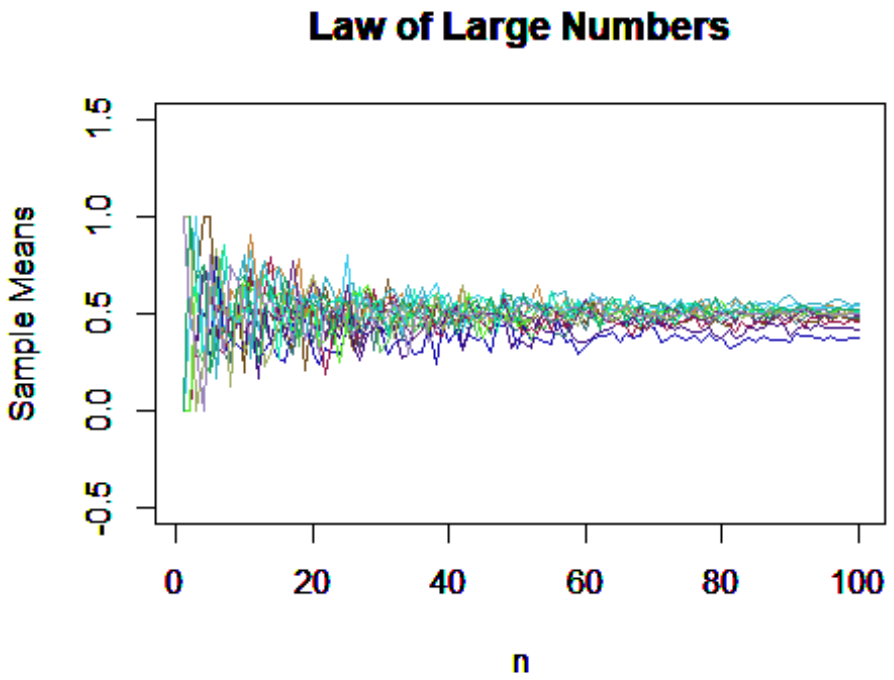


Q1 d)

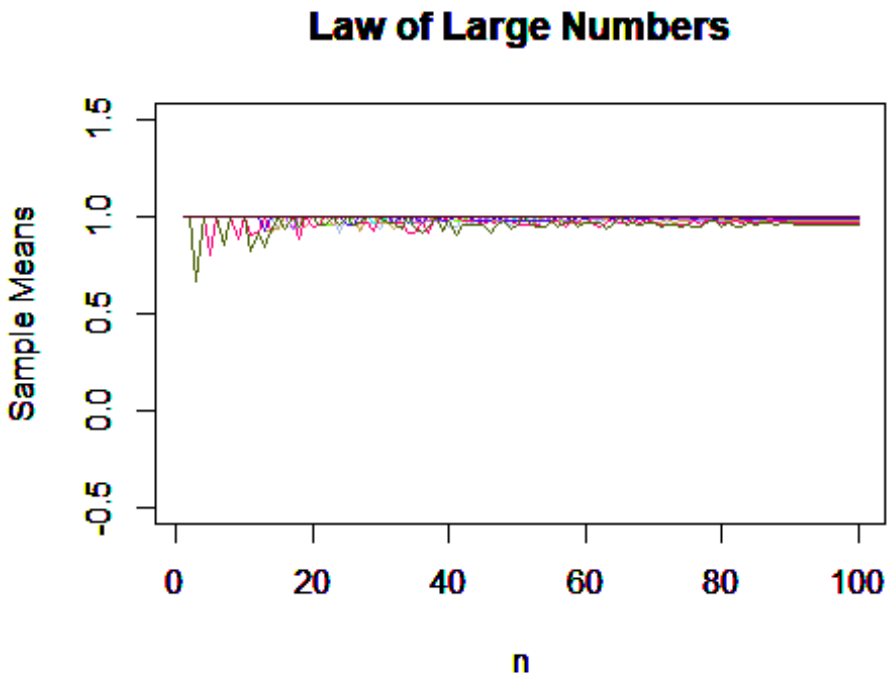
```
replicate(15, (runningmean(c(0,1), 100, c(0.999, 0.001)))) #15 iterations of
plotting the graph of the running mean for n=100 with p=0.001

## Warning in par(new = "TRUE"): calling par(new=TRUE) with no plot
```




```
r=replicate(15, (runningmean(c(0,1), 100, c(1/2,1/2)))) #15 iterations of plo
tting the graph of the running mean for n=100 with p=0.5

## Warning in par(new = "TRUE"): calling par(new=TRUE) with no plot
```




```
replicate(15, (runningmean(c(0,1), 100, c(0.01,0.99)))) #15 iterations of plo
tting the graph of the running mean for n=100 with p=0.99

## Warning in par(new = "TRUE"): calling par(new=TRUE) with no plot
```



Q2 a)

```
x=runif(100,0,8) #generating 100 samples from U(0,8) distribution
fx=(3*x)+8
xsum=mean(fx)*(8-0)
xsum # LHS value in the expression

## [1] 161.3876

func= function(x) {
  (3*x) + 8
}

integrate(func, lower = 0, upper = 8) #finding the integral from 0 to 8 i.e, RHS value

## 160 with absolute error < 1.8e-12
```

So, we see that the LHS value approximates the RHS value, and hence the result is verified.

Q2 b)

```
large_law = function(a,b){
  x=runif(400, min = a, max = b) #generating a random sample of size 400 from U(a,b) Distribution
  cum_sum = rep(NA,400)
  for (i in 1:400){
    cum_sum[i] = ((16+sin(x[i]))/(x[i]^2 + 4))
```

```
  }
  return ((b-a)*sum(cum_sum)/400)
}
large_law(0,7) #calling the function large_law with parameters 0 and 7

## [1] 11.00717
```

Q2 c)

```
means = rep(NA,100)
for (i in 1:100){
  means[i] =large_law(0,7)
} #generating 100 trials of large_law

mean(means) #finding the mean of the simulated means

## [1] 10.5903
```

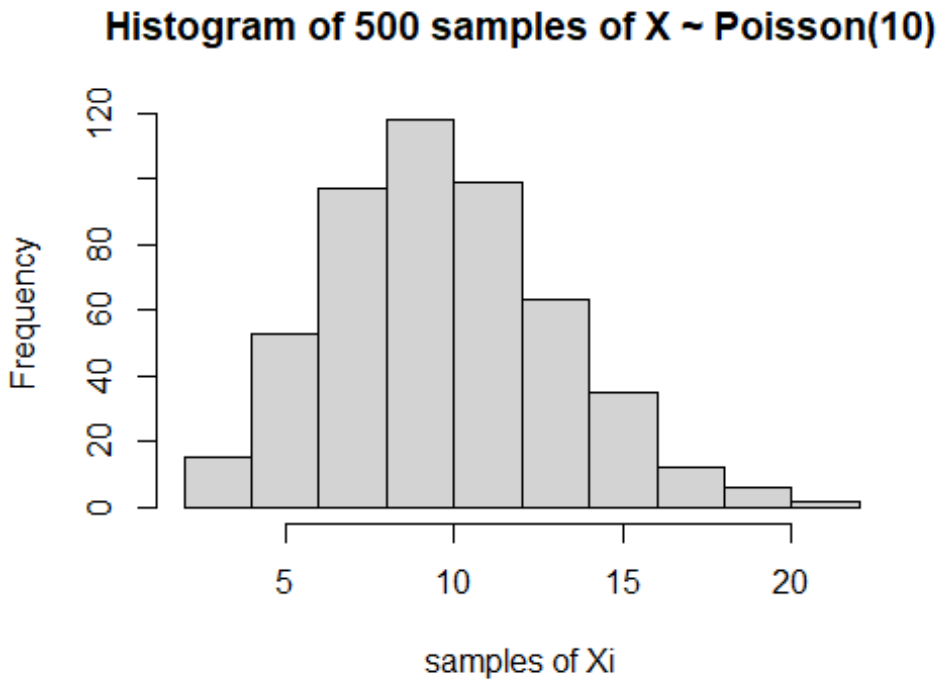
Q2 d)

```
funct = function(x) {
  (16+sin(x))/(x^2 + 4)
}
integrate(funct, lower = 0, upper = 7) #finding the integral from 0 to 7

## 10.58204 with absolute error < 1.9e-06
```

Since the value after integration is approximately equal to that obtained value using law of large numbers, we have verified 2a for the given function.

Q3 a)

```
Xp = rep(NA,500)
for (j in 1:500) {
  lambda = 10;
  i = 0;
  U = runif(1);
  cdf = exp(-lambda)
  while(U >= cdf){
    i = i + 1
    cdf = cdf + exp(-lambda)*lambda^i/gamma(i + 1) }
  X = i
  Xp[j] = X }
summary(Xp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   10.00   10.19   12.00   22.00

hist(Xp, main = "Histogram of 500 samples of X ~ Poisson(10)", xlab = "sample
s of Xi")
```



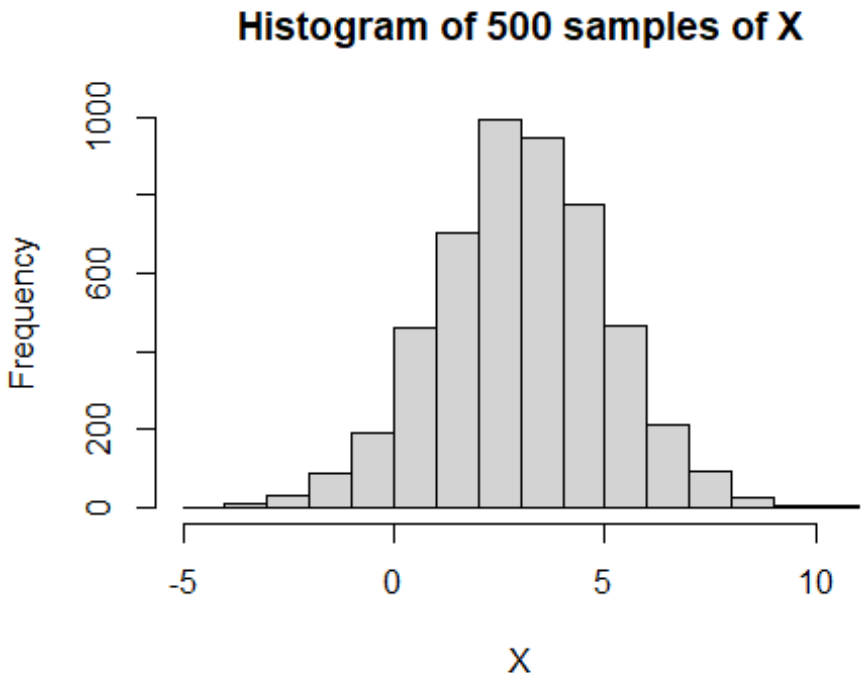
From the summary, we get that the mean of X approximates to actual mean value, 10.

3 b)

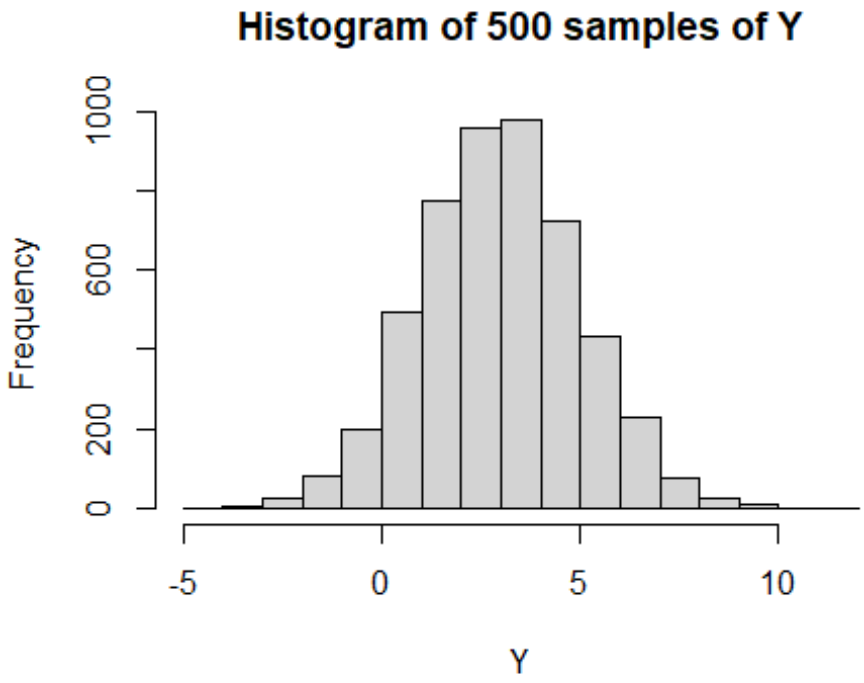
```
U = runif(500,0,1)
X = sqrt(2*U) #storing square root of 500 samples from U(0,1) distribution
summary(X)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04489 0.73763 0.99283 0.94917 1.21403 1.41215
```

Q3 c)

```
n = 5000;
Un1 = runif(n, min = 0, max = 1)
Un2 = runif(n, min = 0, max = 1)
Z1 = sqrt(-2 * log(Un1)) * cos(2 * pi * Un2)
Z2 = sqrt(-2 * log(Un1)) * sin(2 * pi * Un2)

X = 2*Z1 +3
Y = 2*Z2 +3
hist(X, main = "Histogram of 500 samples of X")
```




```
hist(Y, main = "Histogram of 500 samples of Y")
```



X and Y are iid random variable following N(3,4) distribution.

1. From the graph $G(10, \frac{x}{6})$ or from A that you constructed in the worksheet:
   (a) fill in the following table from the data in worksheet:

x	# Edges

   (b) Let E denote the number of edges in a realisation of $G(10, \frac{x}{6})$. Find the likelihood $L(x; E)$ that E edges occur in the random Graph $G(10, \frac{x}{6})$.

   (c) Find x^* that maximizes $L(x; E)$ with respect to x. You may assume $x \in [1, 5]$.
   (d) Substitute your value of E from Question 1, into the expression for x^*. Is the resulting x^* close to your chosen x ?
2. Example 9.2.1.
3. Example 9.2.2.
4. Example 9.3.2.
5. Example 9.3.3.
6. Example 9.4.2.

2). Let $X_1, X_2, \ldots, X_{10}$ be an i.i.d sample with distribution $Bin(N,p)$ where both N & p are unknown. Given, the empirical values of $X_1, \ldots, X_{10}$ as $8, 7, 6, 11, 8, 5, 3, 7, 6, 9$.

We find the first 2 moments of empirical distribution

Now, $m_1 = \frac{X_1 + \cdots + X_{10}}{10} = \frac{8+7+\cdots+9}{10} = \frac{70}{10} = 7$.

and $m_2' = \frac{X_1^2 + X_2^2 + \cdots + X_{10}^2}{10} = \frac{(8)^2 + (7)^2 + \cdots + (9)^2}{10} = \frac{534}{10}$

$= 53.4$.

So, by method of moments estimator for (N,p), we get, $m_1 = \hat{N}\hat{p}$ (i) & $m_2 = \hat{N}\hat{p}(1-\hat{p}) + \hat{N}^2\hat{p}^2$ (ii)

[$\because E(X) = Np$ & $Var(X) = E(X^2) - E^2(X) \Rightarrow E(X^2) = Var(X) + E^2(X)$
$\Rightarrow E(X^2) = Np(1-p) + N^2p^2$]

Substituting value from (i) in (ii), we get.

$m_2 = m_1(1-\hat{p}) + m_1^2 \Rightarrow m_2 - m_1^2 - m_1 = -m_1\hat{p}$

$\Rightarrow \hat{p} = \frac{m_1 + m_1^2 - m_2}{m_1} = \frac{7 + 49 - 53.4}{7} = \frac{2.6}{7} = 0.371$.

$\Rightarrow \hat{p} = 0.371$.

Putting in (i), $m_1 = \hat{N}\cdot\left(\frac{m_1 + m_1^2 - m_2}{m_1}\right) \Rightarrow \hat{N} = \frac{m_1^2}{m_1 + m_1^2 - m_2}$

$\Rightarrow \hat{N} = \frac{7^2}{7 + 7^2 - 53.4} = \frac{49}{2.6} = 18.84$

$\Rightarrow \hat{N} \simeq 19$. (rounded off to nearest integer)

$\therefore$ By method of moments, we get that the distribution from which the sample was drawn as $Bin(19, 0.371)$.

5) Let $p \in (0,1)$ and $X_1, X_2, \ldots, X_n$ be i.id. samples drawn from a Bernoulli(p) distribution. The pmf is given by:

$$f(x|p) = \begin{cases} p & ; \text{if } x = 1 \\ 1-p & ; x = 0 \\ 0 & ; \text{o.w.} \end{cases}$$

$$\Rightarrow f(x|p) = \begin{cases} p^x (1-p)^{1-x} & ; x = 0 \text{ or } 1 \\ 0 & ; \text{o.w.} \end{cases}$$

The likelihood function is given by:

$$L(p|X_1, \ldots, X_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \prod_{i=1}^{n} f(x_i|p)$$

$$= p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n - \sum_{i=1}^{n} x_i}$$

$$\Rightarrow \log L(p|X_1, \ldots, X_n) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

To maximise likelihood function, it is enough to maximise the log likelihood function as log is a monotonically increasing function.

$$\frac{d \log L(p|X_1, \ldots, X_n)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-p}.$$

Now $\frac{d \log L}{dp} = 0$ gives $\sum_{i=1}^{n} x_i - p \sum_{i=1}^{n} x_i = np - p \sum_{i=1}^{n} x_i$

$$\Rightarrow \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

$$\frac{d^2 \log L}{dp^2}(p|X_1, \ldots, X_n) = -\frac{\sum_{i=1}^{n} x_i}{p^2} + \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{(1-p)^2}.$$

At $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $= \left(\frac{-n^2}{\sum_{i=1}^{n} x_i} - \frac{n^2}{n - \sum_{i=1}^{n} x_i} \right) < 0$.

$\therefore \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$ provides a global maximum for the distribution Bernoulli(p) given the sample $X_1, X_2, \ldots, X_n$.

Erdös Renyi Graph $G(n,p)$ is constructed in the following manner:

1. Consider n vertices labeled $\{1,2,\ldots,n\}$.

2. Corresponding to each distinct pair $\{i,j\}$ we perform an independent Bernoulli (p) experiment and insert an edge between i and j with probability p. Note that all edges are *undirected* and hence there are total of $\binom{n}{2}$ possible edges, each occuring with probability p.

3. In this group worksheet you will simulate an Erdös Renyi Graph and find the M.L.E. for the relevant p. Your groups are available at:

https://docs.google.com/spreadsheets/d/1dqH5BvvYID43fK0Syx29CMFvo4hlyQvDcg0iReaO-Ns/edit?usp=sharing

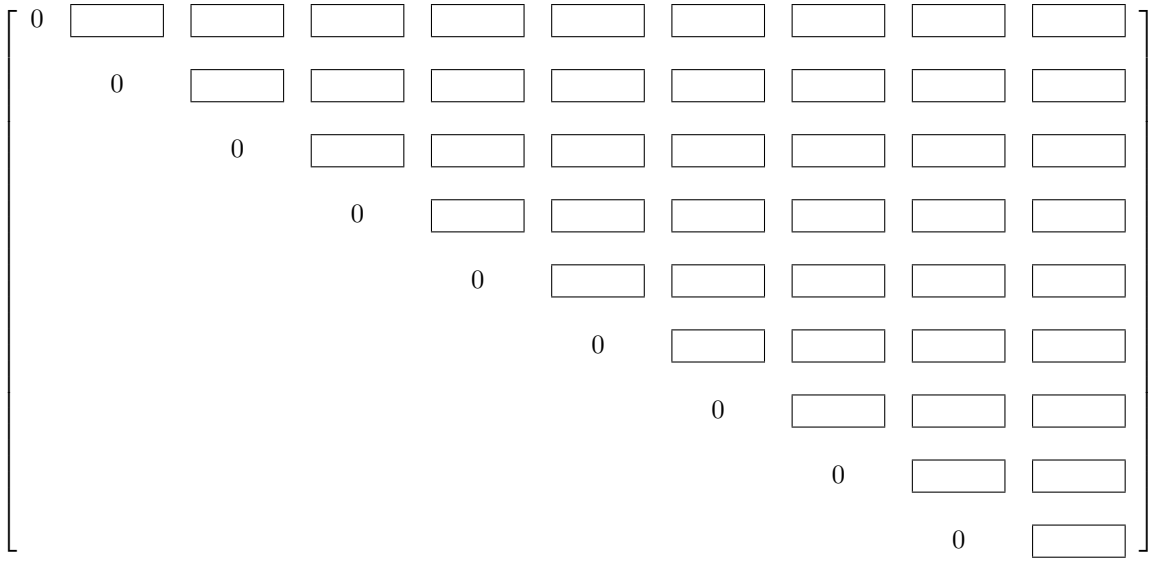
1. Choosing x: Write a simple R-code to generate a number uniformly from $\{1,2,3,4,5\}$. Let x denote the chosen number. Record x in the box: .
2. Consider the experiment of rolling a die and (choose) specify an event from that experiment which occurs with probability $x/6$. *All three persons together* decide on that event, and let it be called B. Write out the description of the event B in the box below:

   .
3. The set of vertices for the graph you are about to construct are $\{1,2,\ldots,10\}$. The graph has no self-edges (i.e Self-loops). What is the total number of possible edges ?

   Record answer in the box:
4. Construct the *random* adjacency matrix A for the graph as follows. For each pair $1 \le i < j \le 10$:
   (a) Roll your die(using one at home or at http://www.randomservices.org/random/apps/Dice.html) and observe if the event B has occured.
   (*Take turns with each person Rolling the die 15 times.*)
   (b) Designate

$$a_{ij} = \begin{cases} 1 & \text{if } B \text{ occured.} \\ 0 & \text{if } B \text{ did not occur} \end{cases}$$

All three persons in respective sheets fill in the matrix entries accordingly:



5. Using the `igraph` package draw the random graph , denote by $G(10, \frac{x}{6})$, corresponding to the above adjacency matrix (i.e draw an edge between i and j if $a_{ij} = 1$).

M.Sc. Data Science
Probability and Statistics with R - Worksheet 10

Name: Soham Biswas
Roll No.: MDS202147
Mail ID: sohamb@cmi.ac.in
Group: 17

1. We need to consider n vertices labeled $\{1, 2, ..., n\}$.

2. Then corresponding to each distinct pair $\{i, j\}$ we perform an independent Bernoulli(p) experiment and insert an edge between i and j with probability p. We note that all edges are undirected and hence there are total of $\binom{n}{2}$ possible edges, each occurring with probability p

3. We will be simulating an Erdos Renyi Graph and find the M.L.E. for the relevant p.

1. Writing a simple R-code to we generate a number uniformly from $\{1, 2, 3, 4, 5\}$ we choose x.

```
x = sample(1:5,1)
x
```

   The value of x is taken as 4.

2. We are required to consider the experiment of rolling a die and specify an event from that experiment which occurs with probability $\frac{x}{6} = \frac{4}{6} = \frac{2}{3}$.

   Therefore the event B is taken as follows:

   Let B denote the event that we get a number less than or equal to 4 on rolling the dice once.

   $P(B) = \frac{2}{3}$

3. Given that the set of vertices for the graph we are to construct are $\{1, 2, ..., 10\}$. The graph has no self-edges.

   Therefore, the total number of edges $= \binom{n}{2} = \binom{10}{2} = 45$

4. Next we need to construct the random adjacency matrix A for the graph as follows. For each pair $1 \leq i < j \leq 10$:

   (a) The die is rolled 15 times by each person in the group and the results obtained are as follows:

   Outcomes of 15 dice rolls for Srijit $= \{4, 2, 2, 1, 3, 3, 1, 4, 4, 2, 6, 4, 2, 6, 6\}$

Outcomes of 15 dice rolls for Soham Pyne $= \{3, 5, 5, 6, 2, 4, 6, 3, 2, 5, 5, 1, 5, 3, 2\}$

Outcomes of 15 dice rolls for Soham Biswas $= \{1, 6, 1, 3, 1, 4, 5, 6, 3, 5, 6, 4, 6, 2, 1\}$

(b) We now designate:

$$a_{ij} = \begin{cases} 1, & if\ B\ occurred \\ 0, & if\ B\ did\ not\ occur \end{cases}$$

We now obtain the adjacency matrix as follows:

```
srijit = c(4,2,2,1,3,3,1,4,4,2,6,4,2,6,6)
sohamp = c(3,5,5,6,2,4,6,3,2,5,5,1,5,3,2)
sohamb = c(1,6,1,3,1,4,5,6,3,5,6,4,6,2,1)
arr = c(srijit,sohamp,sohamb)

A = matrix(,nrow = 10, ncol = 10)
count = 1

for(i in 1:10)
  {
    for(j in 1:10)
      {
        if(i < j)
         {
           if(arr[count] < 5)
             A[i,j] = 1
           else
             A[i,j] = 0

           count = count + 1
         }


        else if(i == j)
        A[i,j] = 0
      }
  }

for(i in 1:10)
  {
    for(j in 1:10)
      {
        if(i > j)
          A[i,j] = A[j,i]
      }
  }
A
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    0    1    1    1    1    1    1    1    1     1
```

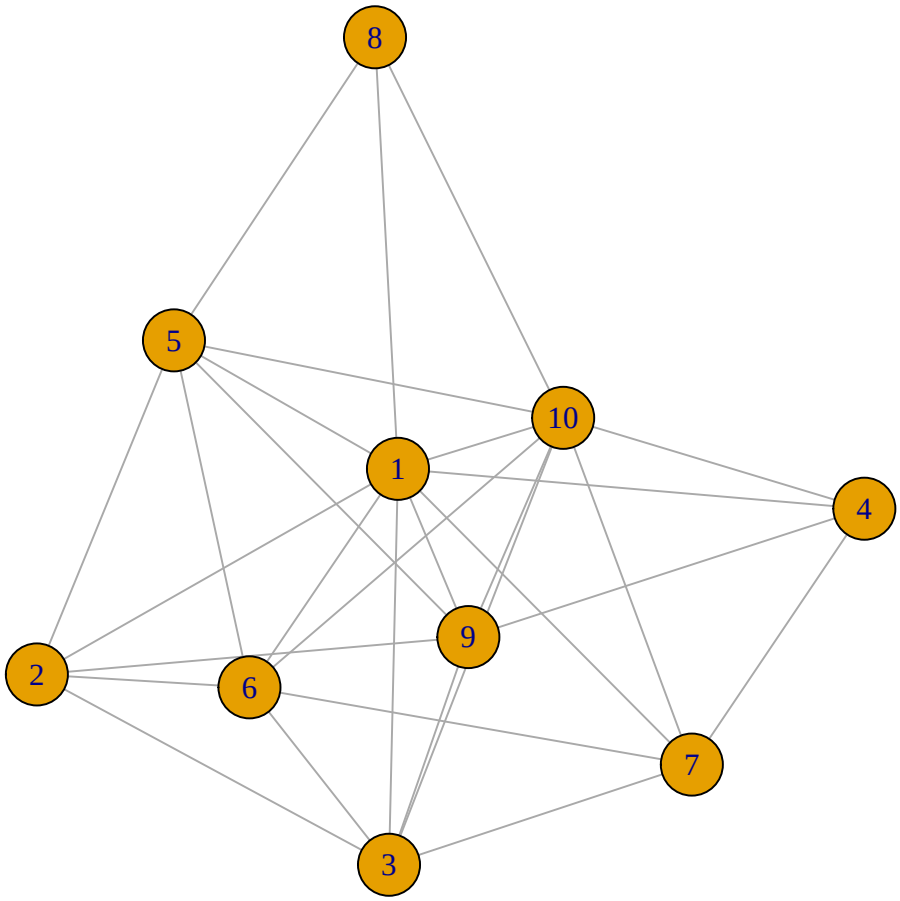
```
##  [2,]    1    0    1    0    1    1    0    0    1     0
##  [3,]    1    1    0    0    0    1    1    0    1     1
##  [4,]    1    0    0    0    0    0    1    0    1     1
##  [5,]    1    1    0    0    0    1    0    1    1     1
##  [6,]    1    1    1    0    1    0    1    0    0     1
##  [7,]    1    0    1    1    0    1    0    0    0     1
##  [8,]    1    0    0    0    1    0    0    0    0     1
##  [9,]    1    1    1    1    1    0    0    0    0     1
## [10,]    1    0    1    1    1    1    1    1    1     0
```

(c) Using the igraph package we need to draw the random graph, denote by $G\left(10, \frac{x}{6}\right)$, corresponding to the above adjacency matrix.

```
library(igraph)
```

```
##
## Attaching package:  'igraph'
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
## The following object is masked from 'package:base':
##
##     union
```

```
plot(graph_from_adjacency_matrix(A, mode = c("undirected"), diag = TRUE))
```




```
length(arr[arr <= 4])
```

```
## [1] 29
```

In the sample generated, we observe that the event B has occurred 29 times in a total of 45 trials.

Let Y_i, $i = 1(1)45$ be a random variable denoting the outcome of a single dice roll.

$Y_i \sim$ i.i.d. Bernoulli(p), $i = 1(1)45$.

Therefore, $\sum_{i=1}^{45} Y_i \sim$ Binomial$(45, p)$.

We consider the following likelihood function:

$L(\sum Y_i|p) = \binom{45}{\sum Y_i}\ p^{\sum Y_i}(1-p)^{45-\sum Y_i}$

Taking logarithm on both sides we get:

$\log L = \log\ \binom{45}{\sum Y_i}\ p^{\sum Y_i}(1-p)^{45-\sum Y_i}$

$= \log\ \binom{45}{\sum Y_i}\ + \sum Y_i \log p + (45 - \sum Y_i)\log(1-p)$

Differentiating both sides w.r.t. p we get:

$$\frac{d\log L}{dp} = \frac{d}{dp}\left(\log\binom{45}{\sum Y_i} + \sum Y_i \log p + (45 - \sum Y_i)\log(1-p)\right)$$

$$= \frac{d}{dp}\left(\log\binom{45}{\sum Y_i}\right) + \frac{d}{dp}\left(\sum Y_i \log p\right) + \frac{d}{dp}\left((45 - \sum Y_i)\log(1-p)\right)$$

$$= 0 + \frac{\sum Y_i}{p} + \frac{45 - \sum Y_i}{1-p}\cdot(-1)$$

$$= \frac{\sum Y_i(1-p) - (45 - \sum Y_i)p}{p(1-p)}$$

$$= \frac{\sum Y_i - 45p}{p(1-p)}$$

Now, we obtain the M.L.E. when $\frac{d\log L}{dp} = 0$

That is, $\frac{\sum Y_i - 45p}{p(1-p)} = 0$

$\implies \sum Y_i - 45p = 0$

$\implies p_{M.L.E.} = \frac{\sum Y_i}{45}$

The observed value of $\sum_{i=1}^{45} Y_i = 29$

Therefore, the estimate of the M.L.E¿ of p is $\hat{p}_{M.L.E.} = \frac{29}{45}$

———X———

Problems due: 1,2,3,6

1. Each of you who has an elder sibling enter into google spreadsheet (link to be provided in class on Tuesday) the following:
   (a) your gender and height (in cm)
   (b) gender and height of your oldest sibling
2. Each of you survey your parent(s) who have an elder sibling and enter into google spreadsheet (link to be provided in class on Tuesday) the following:
   (a) parent gender and height (in cm)
   (b) gender and height of parent's oldest sibling
3. Simulate 1000 samples from Normal$(0, 1)$ in `R`. Implement `R`-codes to do the following:
   (a) Assume that variance known to be 1. Find a 95% confidence interval for the mean μ.
   (b) Assume that variance unknown to be 1. Find a 95% confidence interval for the mean μ.
4. Implement via an `R`-code to perform 100 trials of question 1. Compute the following
   (a) In both cases (known and unknown variance) find the number of trials in which the intervals contain the true mean.
   (b) Plot the difference of the length of intervals observed in both cases (known and unknown variance) across trials.
5. Cracker-Free-rang-dal wants to understand the noise level of firecracker 10000 strip. Measuring the noise level of a random sample of 12 crackers, it gets the following data (in decibels).

   94.0, 98.6, 96.8, 95.5, 93.8, 95.6, 99.3, 95.8, 93.9, 90.2, 91.0, 93.9

   Find a 95% confidence interval for the average noise level of such crackers. Do not round the final answer. Enter the data with 1 decimal place.
6. Use the inbuilt `iris` data set in `R`. For each of the species `setosa, versicolor, virginica` find a 95% confidence interval for: `Sepal.Length`, `Sepal.Length`, `Sepal.Width`, `Petal.Length`, and `Petal.Width`

Homework 11

Rishika Tibrewal

14/12/2021

3 a)

Generating a normal sample of size 1000 from N(0,1)

```
set.seed(1)
norm_samp=rnorm(1000)
```

We write a function `zcinf` to calculate the 95% confidence intervals when variance is known.

```
zcinf = function(x, alpha=0.95, sigma_sq){
z = qnorm( (1-alpha)/2, lower.tail=FALSE)
lower = mean(x) - z*sqrt(sigma_sq/length(x))
upper= mean(x) + z*sqrt(sigma_sq/length(x))
c(lower,upper)
}
```

Using the function to calculate the 95% confidence interval when variance is known to be 1.

```
c_int=zcinf(norm_samp,sigma_sq=1)
cat("The 95% confidence interval for the mean is (",c_int[1],",",c_int[2],"),
provided variance is known to be 1.")
```

```
## The 95% confidence interval for the mean is ( -0.07362765 , 0.05033136 ),
provided variance is known to be 1.
```

3 b)

We write a function `tcinf` to calculate the 95% confidence intervals when variance is unknown.

```
tcinf = function(x, alpha=0.95){
t = qt(p = (1-alpha)/2,df = length(x)-1 , lower.tail=FALSE)
lower= mean(x) - t*sqrt(1/length(x))*sd(x)
upper= mean(x) + t*sqrt(1/length(x))*sd(x)
c(lower,upper)
}
```

Using the function to calculate the 95% confidence interval when variance is unknown.

```
c_int_t=tcinf(norm_samp)
cat("The 95% confidence interval for the mean is
(",c_int_t[1],",",c_int_t[2],"), provided variance is unknown.")
```

```
## The 95% confidence interval for the mean is ( -0.07586952 , 0.05257324 ),
provided variance is unknown.
```

6)

```
data(iris)
```

i)Setosa

```
x1 = iris$Sepal.Length[which(iris$Species == 'setosa')]
cat("The 95% confidence interval for the sepal length of setosa is
(",tcinf(x1)[1],",",tcinf(x1)[2],").")

## The 95% confidence interval for the sepal length of setosa is ( 4.905824 ,
5.106176 ).

x2 = iris$Sepal.Width[which(iris$Species == 'setosa')]
cat("\nThe 95% confidence interval for the sepal width of setosa is
(",tcinf(x2)[1],",",tcinf(x2)[2],").")

##
## The 95% confidence interval for the sepal width of setosa is ( 3.320271 ,
3.535729 ).

x3 = iris$Petal.Length[which(iris$Species == 'setosa')]
cat("\nThe 95% confidence interval for the petal length of setosa is
(",tcinf(x3)[1],",",tcinf(x3)[2],").")

##
## The 95% confidence interval for the petal length of setosa is ( 1.412645 ,
1.511355 ).

x4 = iris$Petal.Width[which(iris$Species == 'setosa')]
cat("\nThe 95% confidence interval for the petal width of setosa is
(",tcinf(x4)[1],",",tcinf(x4)[2],").")

##
## The 95% confidence interval for the petal width of setosa is ( 0.2160497 ,
0.2759503 ).
```

ii)versicolor

```
x1 = iris$Sepal.Length[which(iris$Species == 'versicolor')]
cat("The 95% confidence interval for the sepal length of versicolor is
(",tcinf(x1)[1],",",tcinf(x1)[2],").")

## The 95% confidence interval for the sepal length of versicolor is (
5.789306 , 6.082694 ).

x2 = iris$Sepal.Width[which(iris$Species == 'versicolor')]
cat("\nThe 95% confidence interval for the sepal width of versicolor is
(",tcinf(x2)[1],",",tcinf(x2)[2],").")
```

```
##
## The 95% confidence interval for the sepal width of versicolor is ( 2.68082
, 2.85918 ).
```

```
x3 = iris$Petal.Length[which(iris$Species == 'versicolor')]
cat("\nThe 95% confidence interval for the petal length of versicolor is
(",tcinf(x3)[1],",",tcinf(x3)[2],").")
```

```
##
## The 95% confidence interval for the petal length of versicolor is (
4.126453 , 4.393547 ).
```

```
x4 = iris$Petal.Width[which(iris$Species == 'versicolor')]
cat("\nThe 95% confidence interval for the petal width of versicolor is
(",tcinf(x4)[1],",",tcinf(x4)[2],").")
```

```
##
## The 95% confidence interval for the petal width of versicolor is (
1.269799 , 1.382201 ).
```

iii)virginica

```
x1 = iris$Sepal.Length[which(iris$Species == 'virginica')]
cat("The 95% confidence interval for the sepal length of virginica is
(",tcinf(x1)[1],",",tcinf(x1)[2],").")
```

```
## The 95% confidence interval for the sepal length of virginica is (
6.407285 , 6.768715 ).
```

```
x2 = iris$Sepal.Width[which(iris$Species == 'virginica')]
cat("\nThe 95% confidence interval for the sepal width of virginica is
(",tcinf(x2)[1],",",tcinf(x2)[2],").")
```

```
##
## The 95% confidence interval for the sepal width of virginica is ( 2.882347
, 3.065653 ).
```

```
x3 = iris$Petal.Length[which(iris$Species == 'virginica')]
cat("\nThe 95% confidence interval for the petal length of virginica is
(",tcinf(x3)[1],",",tcinf(x3)[2],").")
```

```
##
## The 95% confidence interval for the petal length of virginica is (
5.395153 , 5.708847 ).
```

```
x4 = iris$Petal.Width[which(iris$Species == 'virginica')]
cat("\nThe 95% confidence interval for the petal width of virginica is
(",tcinf(x4)[1],",",tcinf(x4)[2],").")
```

```
##
## The 95% confidence interval for the petal width of virginica is ( 1.947945
, 2.104055 ).
```

1. Suppose Somadev finds that his weight in kgs during each month of year to be

   75 76 73 75 74 73 73 76 73 79 77 75

   (a) Write a function called `zcinf` that takes in the weights above as a vector `x`, assumes a known standard deviation of 1.5 and produces default 95% confidence interval.

   (b) Write a function called `tcinf` that takes in the weights above as a vector `x`, assumes that variance is unknown and produces default -95% confidence interval.

   (c) Use the inbuilt `t.test` command on the vector `x` (as above), describe each output of the command `t.test(x)` Please explain all the inferences you can make from the output.

2. Distinguishing between Students-$t$distribution and Normal distribution:

   (a) Using `rnorm` and `rt` generate 100 samples of Normal$(0, 1)$ and $t_2 5$ distribution. Compare them using the inbuilt `boxplot`, `qqnorm` and `qqline` functions.

   (b) Using range of $[-4.4]$ (in same frame) plot the densities of Normal$(0, 1)$ and tk distributions for $k = 3, 33, 66$ and 99 using the `dnorm`, `dt` and `plot` function.

3. Suppose we wish to test if the coin given to us is fair. We toss it a 100 times and find that there are 45 heads. Using the inbuilt `prop.test` in `R`, describe each output of the command `prop.test`$(45, 100)$. Please explain all the inferences you can make from the output.

4. In the previous example if we toss the coin a 10000 times and find that there are 4500 heads. Then will you conclude that the coin is fair ?

5. Suppose Doddapple manufactures claims that their batteries last 25 years. Students from CMI's Data Science programs sample 10 users and find the sample mean time for battery life was 21 with a sample standard deviation of 1.7. Is Doddapple claim believable ?

Worksheet 13

Rishika Tibrewal

15/12/2021

1 a)

```
alpha=0.95
zcinf=function(x){
  sd=1.5
  n=length(x)
  lower_z=mean(x)-(sd/sqrt(n))*abs(qnorm((1-alpha)/2))
  upper_z=mean(x)+(sd/sqrt(n))*abs(qnorm((1-alpha)/2))
  c(lower_z,upper_z)
}
```

1 b)

```
tcinf=function(x){
  s=sd(x)
  n=length(x)
  tscore=qt((1-alpha)/2, df=n-1)
  lower_t=mean(x)-(s/sqrt(n))*abs(tscore)
  upper_t=mean(x)+(s/sqrt(n))*abs(tscore)
  c(lower_t,upper_t)
}
```

1 c)

```
x=c(75, 76, 73, 75, 74, 73, 73, 76, 73, 79, 77, 75)

#95% confidence interval for known sd = 1.5
zcinf(x)

## [1] 74.06798 75.76536

#95% confidence interval for unknown sd
tcinf(x)

## [1] 73.72158 76.11175

#t-test
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 137.97, df = 11, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  73.72158 76.11175
## sample estimates:
## mean of x
##  74.91667
```

From Q1 we infer that the default confidence interval in t test on R is the same as that for unknown sd using function tcinf.

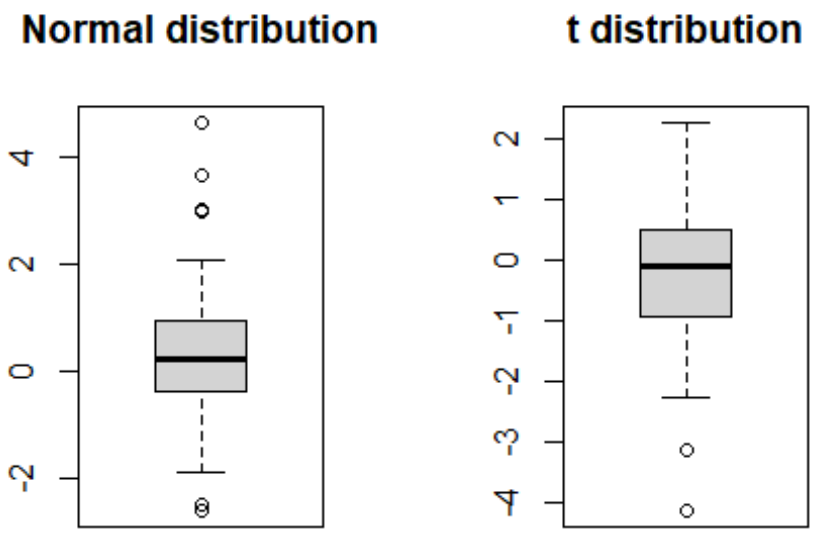
From 1 c) we infer the following:

- t = 137.97 is the calculated test statistic with degree of freedom = 11
- the p-value is < 2.2e-16 which is very small when compared to 0.05 (alpha value)
- the alternate hypothesis when true mean isn't equal to 0
- 95% confidence intervals (which is the same as calculated in 1 b))
- Estimate of sample means being 74.92

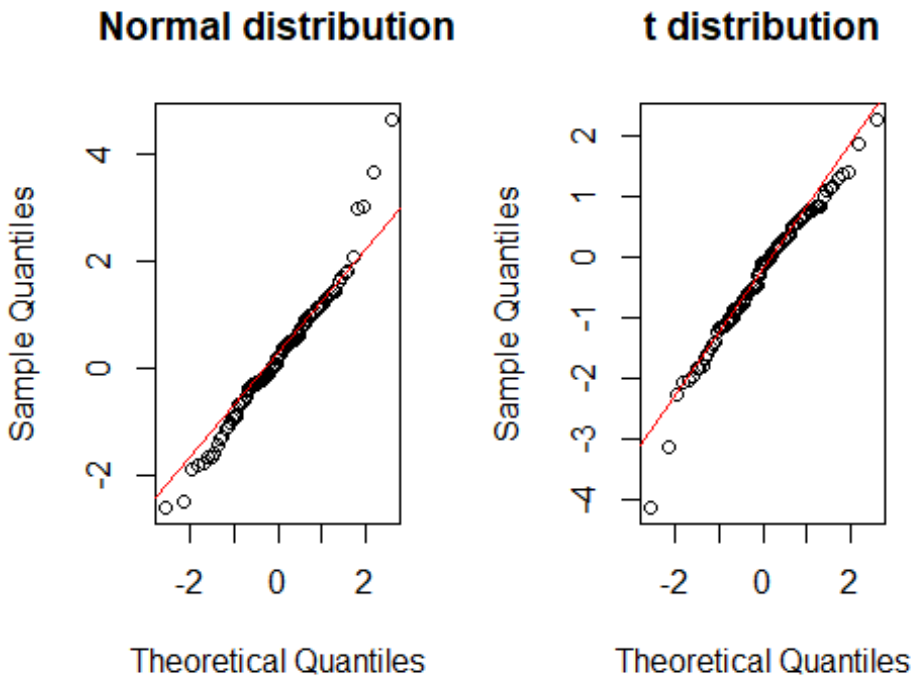
The p-value of the test is < 2.2e-16, which is very less than the significance level alpha = 0.05. We can conclude that the true mean weights are significantly different from sample mean and so, we reject the null hypothesis.

2 a)

```
n_samp=rnorm(100)
t_samp=rt(100,25)
par(mfrow=c(1,2))
boxplot(n_samp,main="Normal distribution")
boxplot(t_samp,main="t distribution")
```



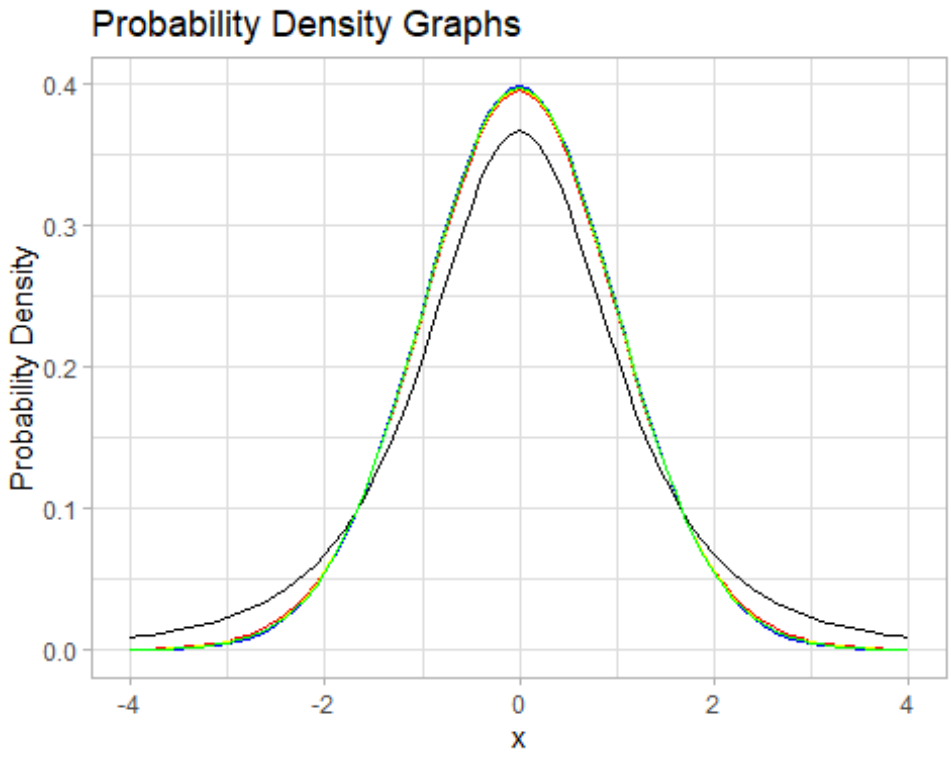

```
par(mfrow=c(1,2))
qqnorm(n_samp,main="Normal distribution")
qqline(n_samp,col="red")
qqnorm(t_samp,main="t distribution")
qqline(t_samp,col="red")
```



- Box plot of t- distribution has longer tails than that of Normal distribution samples.

2 b)

```
library(ggplot2)
x = seq(-4, 4,0.1)
samp = dnorm(x)
df_3 = dt(x,df = 3)
df_33 = dt(x,df = 33)
df_66 = dt(x,df = 66)
df_99 = dt(x,df = 99)
data = data.frame(x,df_3,df_33,df_66,df_99)
ggplot(data,aes(x))+geom_line(aes(y = samp), color = 'blue')+geom_line(aes(y
= df_3),color = 'black')+geom_line(aes(y = df_33),color =
'red')+geom_line(aes(y = df_66),color = 'yellow')+geom_line(aes(y = df_99),
color = 'green')+ylab("Probability Density")+ggtitle("Probability Density
Graphs")+theme_light()
```



3)

```
prop.test(45,100)

##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 100, null probability 0.5
## X-squared = 0.81, df = 1, p-value = 0.3681
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3514281 0.5524574
## sample estimates:
##    p
## 0.45
```

From the above proportion test, we infer the following:

- data as 45 heads appear out of 100
- the null probability (probability of getting a head) is 0.5 which is the probability in case of a fair coin
- the value of Pearson's Chi-squared test statistic with 1 degree of freedom as 0.81
- the p-value as 0.3681
- alternative hypothesis H1: the value of p isn't same as 0.5
- the 95% confidence interval which is (0.3514281,0.5524574)
- an estimated probability of success (i.e., the proportion of heads).

The p-value of the test which is 0.3681, is greater than the significance level, alpha = 0.05, so we can conclude that the proportion of heads on tossing a coin is not significantly different from 0.5. Hence, we accept the null hypothesis stating that the coin is fair.

4)

```
prop.test(4500,10000)

##
##  1-sample proportions test with continuity correction
##
## data:  4500 out of 10000, null probability 0.5
## X-squared = 99.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4402205 0.4598181
## sample estimates:
##    p
## 0.45
```

From the above proportion test, we infer the following:

- data as 4500 heads appear out of 10000
- the null probability (probability of getting a head) is 0.5 which is the probability in case of a fair coin
- the value of Pearson's Chi-squared test statistic with 1 degree of freedom as 99.8
- the p-value as < 2.2e-16
- alternative hypothesis H1: the value of p isn't same as 0.5
- the 95% confidence interval which is (0.4402205,0.4598181)
- an estimated probability of success (i.e., the proportion of heads)

The p-value of the test is < 2.2e-16, which is very small when compared to the significance level, alpha = 0.05, so we can conclude that the proportion of heads on tossing a coin is significantly different from 0.5 . Hence, we reject the null hypothesis which is why it can be concluded that the coin is not fair.

5) Given that Doddapple manufacturers claim that their batteries last 25 years. We sample 10 users, so $n = 10$.

Given, mean time of battery life (from sample) = 21.

and standard deviation (from sample) = 1.7

So, $\bar{X} = 21$ & $S = 1.7$.

To check if the claim is believable or not, we perform a t-test where $\mu_0 = 25$.

$H_0 : \mu = 25$ ag. $H_1 : \mu < 25$.

Given, $n = 10 (< 30)$, so we use the test statistic, $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1;\alpha}$ at 0.05 level of significance

Pulling the values of $\bar{X}$, μ_0, S and n, we get:

$$T_{obs} = \frac{21 - 25}{1.7/\sqrt{10}} = -7.4407.$$

$$\text{p-value} = P(T < -T_{obs}) = P(T < 7.4407)$$

$$= 0.00002$$

So, we reject H_0 as p-value $< 0.05 (\alpha)$. Hence, the claim that Doddapple makes about their batteries lifes being more than 25 years is not believable.