IST772– Problem Set 2

Sathish Kumar Rajendiran

1. I did this homework by myself, with help from the book and the professor.

```
set.seed(772)
```

Chapter 2, Exercise 1

Flip an actual physical fair coin by hand seven times and write down the number of heads obtained (1 pt). Now repeat this process 50,000 times. Obviously you don't want to have to do that by hand, so create the necessary lines of R code to do it for you. Hint: You will need both the rbinom() function and the table() function (1 pt). Write down the results and explain in comments in your own words what they mean (2 pts).

```
cat("\n*************  Chapter 2, Exercise 1   **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 1   **************************************
```

```
#  Chapter 2, Exercise 1
cat("\n Table 2.1 (1) Physical fair coin Heads-obtained Trial, 7 times per Trial:\n\n")
```

```
##
##  Table 2.1 (1) Physical fair coin Heads-obtained Trial, 7 times per Trial:
```

```
coin_flip <- matrix(c(1,1,0,1,0,1,1),ncol = 7,byrow = TRUE)
colnames(coin_flip) <- c(0,1,2,3,4,5,6)
rownames(coin_flip) <- c(" Head obtained count")
coin_flip <- as.table(coin_flip)
cbind(coin_flip, total = rowSums(coin_flip))
```

```
##                        0 1 2 3 4 5 6 total
##  Head obtained count 1 1 0 1 0 1 1     5
```

```
cat("\n Total number of Heads Obtained from , 7 times is:",rowSums(coin_flip))
```

```
##
##  Total number of Heads Obtained from , 7 times is: 5
```

```
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

```
cat("\n*************  Chapter 2, Exercise 1 - Continued  ***************************************\n")
```

```
##
## *************  Chapter 2, Exercise 1 - Continued  **************************************
```

```
#  Chapter 2, Exercise 1  (2)
set.seed(772)
cat("\n Table 2.1 (2) 50,000 Physical fair coin Heads-obtained Trial, 7 times per Trial:\n\n")
```

```
##
##  Table 2.1 (2) 50,000 Physical fair coin Heads-obtained Trial, 7 times per Trial:
```

```
# rbinom(n=50000,size=7,prob = 0.5)
random_trial50k <- rbinom(n=50000,size=7,prob = 0.5)  #rbinom() in action  | prob = 0.5 ;
#assuming its a fair coin with only 2 possible outcomes (probability 50%))
table_trial50k <- table(random_trial50k) #table() in action
# table_trial50k
trials <- c(0,1,2,3,4,5,6,7)
table_matix <- rbind(trials,table_trial50k)
rownames(table_matix) <- c("Head-obtained count","Number of Trials with Head-obtained count")
table_matix
```

```
##                                             0    1    2     3     4    5    6
## Head-obtained count                         0    1    2     3     4    5    6
## Number of Trials with Head-obtained count 404 2702 8205 13591 13689 8246 2773
##                                             7
## Head-obtained count                         7
## Number of Trials with Head-obtained count 390
```

```
cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

```
cat("\n*************  Chapter 2, Exercise 1 - Continued  ***************************************\n")
```

```
##
## *************  Chapter 2, Exercise 1 - Continued  **************************************
```

```
#  Chapter 2, Exercise 1  (3)
cat("\n Table 2.1 (3) 50,000 Physical fair coin Heads-obtained Trial, 7 times per Trial Summary:\n
 From above table, shows the number of times that head is obtained from each trial.

* set.seed(772)  -- function would retain the same result each time random function is
trying generate of sequence of random numbers. Its pseudo random number maintained at R Studio level.

* rbinom() - function generates random sequence of binomial distribution numbers with number
of trials specified. binomial meaning 'two names'.
In this case fair coin toss would have two answers as 'head' or 'tail'
")
```

```
##
##  Table 2.1 (3) 50,000 Physical fair coin Heads-obtained Trial, 7 times per Trial Summary:
##
##  From above table, shows the number of times that head is obtained from each trial.
##
## * set.seed(772)  -- function would retain the same result each time random function is
## trying generate of sequence of random numbers. Its pseudo random number maintained at R Studio level
##
## * rbinom() - function generates random sequence of binomial distribution numbers with number
## of trials specified. binomial meaning 'two names'.
## In this case fair coin toss would have two answers as 'head' or 'tail'
```

```
cat("
\n * size =7 As, each trial is consists of 7 times - its possible to have either 0 heads
or all heads (7). So, it ranges from 0,1,2,3,4,5,6 & 7. Here, trial refers to set of
tests/experiments/group event.

* n=50000 -- mentions total number of trials the fair coin experiments repeated.
i.e 50,000 times of 7 trial events.

* prob =0.5 assuming its a fair coin with only 2 possible outcomes (probability 50%))

* rbinom(n=50000,size=7,prob = 0.5) function return results as below,
# like below
")
```

```
##
##
##  * size =7 As, each trial is consists of 7 times - its possible to have either 0 heads
## or all heads (7). So, it ranges from 0,1,2,3,4,5,6 & 7. Here, trial refers to set of
## tests/experiments/group event.
##
## * n=50000 -- mentions total number of trials the fair coin experiments repeated.
## i.e 50,000 times of 7 trial events.
##
## * prob =0.5 assuming its a fair coin with only 2 possible outcomes (probability 50%))
##
## * rbinom(n=50000,size=7,prob = 0.5) function return results as below,
## # like below
```

```
cat("\n #  [1] 3 2 2 5 3 3 2 2 3 6 5 6 2 3 4 5 3 4 1 3 5 2 2 4 6 2 5 2 4 3 4 5 3 2 3 4 5 3 3 5 2 1 1 3
#   [79] 1 4 1 3 4 4 3 3 3 4 2 2 2 2 4 1 4 4 5 4 2 4 3 4 4 4 3 4 4 2 3 5 6 2 5 2 3 6 6 2 4 6 3 5 4 3 4 1
#  [157] 3 2 4 2 5 5 4 2 4 3 4 4 3 3 4 4 6 4 3 2 3 3 6 3 4 4 4 2 1 3 5 4 4 2 4 3 4 6 5 4 5 2 6 2 5 4 4
#  [235] 3 4 3 5 4 4 4 3 4 4 1 3 2 6 5 1 4 5 4 2 2 5 6 2 5 3 3 3 4 5 3 4 4 1 4 1 4 4 3 4 6 3 5 3 4 4 2
#  [313] 3 4 2 4 2 5 7 4 6 4 3 6 3 3 4 6 3 4 4 3 4 2 5 4 6 1 3 3 4 5 3 2 4 6 4 3 6 4 2 5 2 4 6 3 3 5 5
#  [391] 4 1 4 3 7 1 5 2 6 4 4 5 3 6 5 1 5 5 4 3 5 3 2 3 3 2 5 1 4 3 1 3 2 4 4 3 5 6 2 4 3 3 4 4 2 3 1
#  [469] 3 4 5 3 3 5 3 3 5 3 4 3 3 4 4 3 6 6 3 4 3 1 3 3 2 3 4 4 3 3 3 4 2 3 4 3 2 2 5 6 4 1 3 4 2 3 4
#  [547] 4 2 2 2 2 3 4 3 3 3 4 4 3 3 4 1 5 4 1 4 3 3 2 5 5 2 5 6 3 3 4 4 2 4 2 1 4 5 3 2 4 2 2 4 5 5 3
#  [625] 4 5 3 5 6 5 2 3 5 4 5 3 4 4 3 1 4 3 5 3 2 3 4 3 3 5 0 6 5 3 3 5 4 4 4 3 6 4 4 6 3 3 3 5 2 4 4
#  [703] 3 5 4 4 4 2 2 4 3 2 3 5 3 6 2 2 2 4 2 4 6 5 4 1 4 2 5 5 5 3 5 3 3 4 2 3 4 2 5 5 5 1 1 3 4 3 3
#  [781] 4 2 5 2 3 4 4 3 3 4 6 5 1 5 3 4 5 4 5 4 4 4 5 2 3 3 4 1 0 2 5 4 4 1 3 3 4 1 5 3 2 5 2 6 4 6 5
#  [859] 6 1 5 5 6 3 4 4 4 4 5 5 4 3 3 4 2 3 5 3 4 4 4 3 3 1 4 4 5 3 5 2 3 5 3 4 4 6 5 1 3 2 5 3 5 3 5
#  [937] 2 4 5 3 4 2 4 4 2 3 3 3 4 3 4 6 2 5 1 4 3 2 3 2 5 2 4 4 5 4 4 2 4 4 2 5 3 5 5 4 4 3 4 3 4 3 4
```

```
##
##  #   [1] 3 2 2 5 3 3 2 2 3 6 5 6 2 3 4 5 3 4 1 3 5 2 2 4 6 2 5 2 4 3 4 5 3 2 3 4 5 3 3 5 2 1 1 3 4 3
## #   [79] 1 4 1 3 4 4 3 3 3 4 2 2 2 2 4 1 4 4 5 4 2 4 3 4 4 3 4 4 2 3 5 6 2 5 2 3 6 6 2 4 6 3 5 4 3 4
## #  [157] 3 2 4 2 5 5 4 2 4 3 4 4 3 3 4 4 6 4 3 2 3 3 6 3 4 4 4 2 1 3 5 4 4 2 4 3 4 6 5 4 5 2 6 2 5 4
## #  [235] 3 4 3 5 4 4 4 3 4 4 1 3 2 6 5 1 4 5 4 2 2 5 6 2 5 3 3 3 4 5 3 4 4 1 4 1 4 4 3 4 6 3 5 3 4 4
## #  [313] 3 4 2 4 2 5 7 4 6 4 3 6 3 3 4 6 3 4 4 3 4 2 5 4 6 1 3 3 4 5 3 2 4 6 4 3 6 4 2 5 2 4 6 3 3 5
## #  [391] 4 1 4 3 7 1 5 2 6 4 4 5 3 6 5 1 5 5 4 3 5 3 2 3 3 2 5 1 4 3 1 3 2 4 4 3 5 6 2 4 3 3 4 4 2 3
## #  [469] 3 4 5 3 3 5 3 3 5 3 4 3 3 4 4 3 6 6 3 4 3 1 3 3 2 3 4 4 3 3 3 4 2 3 4 3 2 2 5 6 4 1 3 4 2 3
## #  [547] 4 2 2 2 2 3 4 3 3 3 4 4 3 3 4 1 5 4 1 4 3 3 2 5 5 2 5 6 3 3 4 4 2 4 2 1 4 5 3 2 4 2 2 4 5 5
## #  [625] 4 5 3 5 6 5 2 3 5 4 5 3 4 4 3 1 4 3 5 3 2 3 4 3 3 5 0 6 5 3 3 5 4 4 4 3 6 4 4 6 3 3 3 5 2 4
## #  [703] 3 5 4 4 4 2 2 4 3 2 3 5 3 6 2 2 2 4 2 4 6 5 4 1 4 2 5 5 5 3 5 3 3 4 2 3 4 2 5 5 5 1 1 3 4 3
## #  [781] 4 2 5 2 3 4 4 3 3 4 6 5 1 5 3 4 5 4 5 4 4 4 5 2 3 3 4 1 0 2 5 4 4 1 3 3 4 1 5 3 2 5 2 6 4 6
## #  [859] 6 1 5 5 6 3 4 4 4 4 5 5 4 3 3 4 2 3 5 3 4 4 4 3 3 1 4 4 5 3 5 2 3 5 3 4 4 6 5 1 3 2 5 3 5 3
## #  [937] 2 4 5 3 4 2 4 4 2 3 3 3 4 3 4 6 2 5 1 4 3 2 3 2 5 2 4 4 5 4 4 2 4 4 2 5 3 5 5 4 4 3 4 3 4 3
```

```
cat("\n\n* table() -- As you can see from result above -
it would be difficult to count number
of heads obtained in each trial ( 0,1,2 to 7) on all 50,000 times. Hence,
we need  table() function to summarize the results by trials.
 below result is in much better readable format. Having 13814 times max
 of 4 heads obtained; 396 times no heads were returned.

# random_trial50k
#     0     1     2     3     4     5     6     7
#   396  2749  8244 13481 13814  8181  2729   406

# In addition, last couple of R lines code is to beautify the results as below
Head-obtained count                          0    1    2     3     4    5    6   7
Number of Trials with Head-obtained count 396 2749 8244 13481 13814 8181 2729 406")
```

```
##
##
## * table() -- As you can see from result above -
## it would be difficult to count number
## of heads obtained in each trial ( 0,1,2 to 7) on all 50,000 times. Hence,
## we need  table() function to summarize the results by trials.
##  below result is in much better readable format. Having 13814 times max
##  of 4 heads obtained; 396 times no heads were returned.
##
## # random_trial50k
## #     0     1     2     3     4     5     6     7
## #   396  2749  8244 13481 13814  8181  2729   406
##
## # In addition, last couple of R lines code is to beautify the results as below
## Head-obtained count                          0    1    2     3     4    5    6   7
## Number of Trials with Head-obtained count 396 2749 8244 13481 13814 8181 2729 406
```

```
cat("\n*************  Chapter 2, Exercise 1  - End **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 1  - End **************************************
```

Chapter 2, Exercise 2

Using the output from Exercise 1, summarize the results of your 50,000 trials of 7 flips each in a bar plot using the appropriate commands in R. Convert the results to probabilities and represent that in a bar plot as well (1 pt for the two bar plots). Write a brief interpretive analysis that describes what each of these bar plots signifies and how the two bar plots are related (1 pt). Make sure to comment on the shape of each bar plot and why you believe that the bar plot has taken that shape. Also make sure to say something about the center of the bar plot and why it is where it is (1 pt for shape and centre; 1 pt for explanation of shape).

```
cat("\n*************  Chapter 2, Exercise 2   **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 2   **************************************
```

```
#  Chapter 2, Exercise 2  (1)
cat("\nTable 2.2 (1) 50,000 Physical fair coin Heads-obtained Trial, 7 times per Trial:\n\n")
```

```
##
## Table 2.2 (1) 50,000 Physical fair coin Heads-obtained Trial, 7 times per Trial:
```

```
cat("\nTrial Summary:\n\n")
```

```
##
## Trial Summary:
```

```
table_trial50k
```

```
## random_trial50k
##     0     1     2     3     4     5     6     7
##   404  2702  8205 13591 13689  8246  2773   390
```

```
# trials <- c(0,1,2,3,4,5,6,7)
# matrix_trial50k <- rbind(trials,table_trial50k)
# rownames(matrix_trial50k) <- c("Head-obtained count","Number of Trials with Head-obtained count")
# matrix_trial50k

cat("\nProbability Summary:\n\n")
```

```
##
## Probability Summary:
```

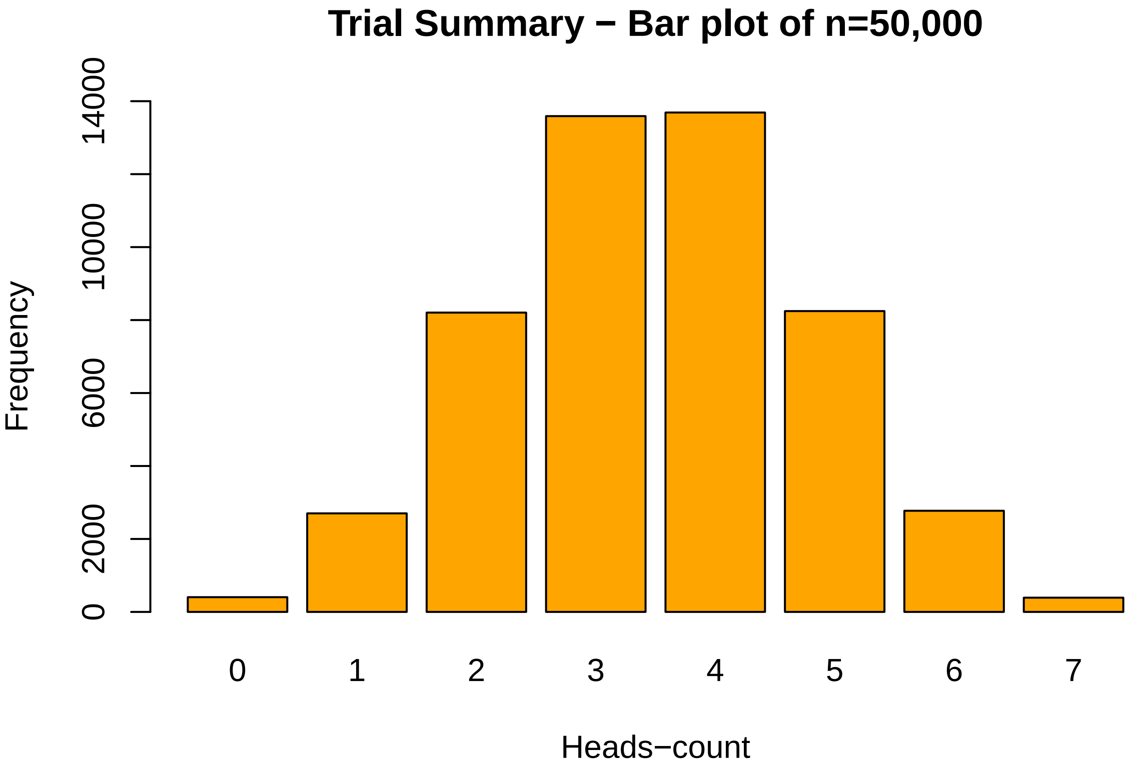
```
# table_trial50k
prob_trial50k <- table_trial50k/50000
prob_trial50k
```

```
## random_trial50k
##       0       1       2       3       4       5       6       7
## 0.00808 0.05404 0.16410 0.27182 0.27378 0.16492 0.05546 0.00780
```

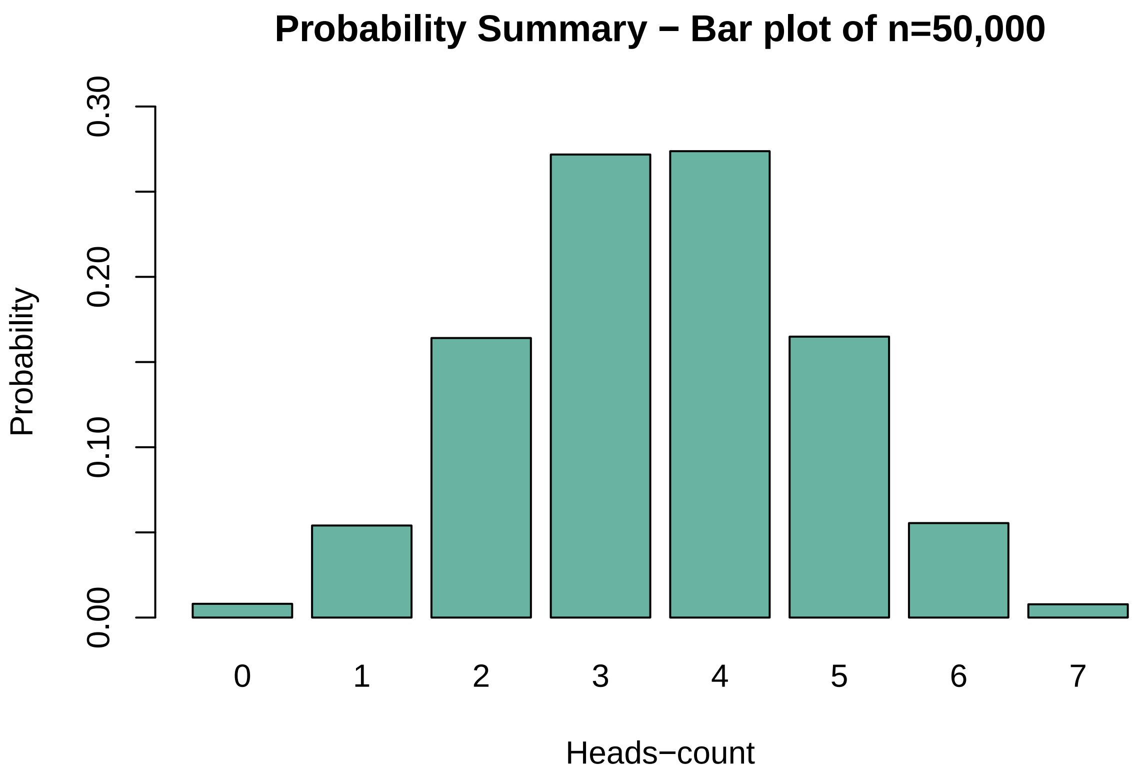
```r
# matrix_prob_trial50k <- rbind(rownames(prob_trial50k),prob_trial50k)
# rownames(matrix_prob_trial50k) <- c("Head-obtained count","Number of Trials with Head-obtained count",
# matrix_prob_trial50k


cat("\nTable 2.2 (1) barplots:\n\n")
```

```
##
## Table 2.2 (1) barplots:
```

```r
barplot(table_trial50k,main="Trial Summary - Bar plot of n=50,000",col="orange",xlab = "Heads-count",yla
```



```r
barplot(prob_trial50k,main="Probability Summary - Bar plot of n=50,000",col="#69b3a2",xlab = "Heads-cou
```



```
# cat("\nCumulative Probability Summary:\n\n")
# cumsum_trial50k <- cumsum(probtable)
# cumsum_trial50k

# cat("\nTable 2.2 (1) barplots:\n\n")
# barplot(cumsum_trial50k)

cat("\n ****************************************************************************")
```

```
##
##  ****************************************************************************
```

```
cat("\n*************  Chapter 2, Exercise 2  - Continued **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 2  - Continued **************************************
```

```
cat("\nMeaning and Relation of the Bar Charts:")
```

```
##
## Meaning and Relation of the Bar Charts:
```

```
cat("
  Both bar charts represents 50,000 Physical fair coin Trials with 7 times per Trial -
  shows number of head obtained counts spread in binomial distribution.Each bar of varying height
  represents number of occurances on each trial. Example, Trial 3 and 4
  having most number of heads obtained with 13591 & 13689 heads out of 50,000 tosses respectively.
x - same values in x-axis on both bar plots has the trial events ranging 0,1,2,3,4,5,6,7 -
confirming there are both 0 head obtained events as well 7 head obtained events.
y - slight variation in heights and totally differnt range on y-axis; Trial summary bar
plot has values in actual counts between 0 to 14,000 with an interval of 2000 events as bin;
Howver, the probability bar plot shows values in the range of 0 to .30 .
i.e probabilities instead of raw counts with ylim as 0.05. This is by simply dividing the
random events counts by 50,000 ( total number of events) I reduced the bar width
as max of probability is 0.27378 and 0.27182 respectively.
When we add the cumulative sum of these counts it would result in 1 ( max probability)
Probabilty and Trial summary bar plots and one and the same, except that the trial summary
plot values are 50,000 times that of probability trials.
It also, suggests that probability of getting 3 or 4 heads is far more than all heads or 0 heads.

")
```

```
##
##   Both bar charts represents 50,000 Physical fair coin Trials with 7 times per Trial -
##   shows number of head obtained counts spread in binomial distribution.Each bar of varying height
##   represents number of occurances on each trial. Example, Trial 3 and 4
##   having most number of heads obtained with 13591 & 13689 heads out of 50,000 tosses respectively.
## x - same values in x-axis on both bar plots has the trial events ranging 0,1,2,3,4,5,6,7 -
## confirming there are both 0 head obtained events as well 7 head obtained events.
## y - slight variation in heights and totally differnt range on y-axis; Trial summary bar
## plot has values in actual counts between 0 to 14,000 with an interval of 2000 events as bin;
## Howver, the probability bar plot shows values in the range of 0 to .30 .
## i.e probabilities instead of raw counts with ylim as 0.05. This is by simply dividing the
## random events counts by 50,000 ( total number of events) I reduced the bar width
## as max of probability is 0.27378 and 0.27182 respectively.
## When we add the cumulative sum of these counts it would result in 1 ( max probability)
## Probabilty and Trial summary bar plots and one and the same, except that the trial summary
## plot values are 50,000 times that of probability trials.
## It also, suggests that probability of getting 3 or 4 heads is far more than all heads or 0 heads.
##
```

```
cat("\nShape and Centre of the Bar Charts:")
```

```
##
## Shape and Centre of the Bar Charts:
```

```
cat("
   Shape of both bar plots represents a bell-shaped curve with symmetric tails on both ends
   and gradual curves toward a peak in the middle forming a normal distribution.
 Centre of the bar charts has two almost equal size of bars representing peak values and also
 reason for bell-shaped curve as both tails almost touching 0 ( 0.007 and 0.008 on the
 probability plot and  404, 390 against 50,000 counts on trial summary charts.)
Yes, the bell-shaped curve proves that the  that probability of getting
3 or 4 heads is far more than all heads or 0 heads.

")
```

```
##
##    Shape of both bar plots represents a bell-shaped curve with symmetric tails on both ends
##    and gradual curves toward a peak in the middle forming a normal distribution.
##  Centre of the bar charts has two almost equal size of bars representing peak values and also
##  reason for bell-shaped curve as both tails almost touching 0 ( 0.007 and 0.008 on the
##  probability plot and  404, 390 against 50,000 counts on trial summary charts.)
## Yes, the bell-shaped curve proves that the  that probability of getting
## 3 or 4 heads is far more than all heads or 0 heads.
##
```

```
cat("\nWhy Shape?:")
```

```
##
## Why Shape?:
```

```
cat("
  As these two plots are basically a result of binomial distribution events with fair
  coin trial the probability value of the trial p=0.5. Hence it forms a symmetric
  curve with n being large at 50,000. In addition, the mean of a binomial distribution
  is p (0.5) and standard deviation is sqr(p(1-p)/n)
So, probability of success p is same for each outcome
forming a bell shaped curve through tops of the bars.
")
```

```
##
##   As these two plots are basically a result of binomial distribution events with fair
##   coin trial the probability value of the trial p=0.5. Hence it forms a symmetric
##   curve with n being large at 50,000. In addition, the mean of a binomial distribution
##   is p (0.5) and standard deviation is sqr(p(1-p)/n)
## So, probability of success p is same for each outcome
## forming a bell shaped curve through tops of the bars.
```

```
cat("\n*************  Chapter 2, Exercise 2  - End **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 2  - End **************************************
```

Chapter 2, Exercise 6

One hundred students took a statistics test. Fifty of them are high school students and 50 are college students. Eighty-two students passed and 18 students failed. You now have enough information to create a two-by-two contingency table with all of the marginal totals specified (although the four main cells of the table are still blank). You may want to draw that table and write in the marginal totals to see what's happening with the data. I'm now going to give you one additional piece of information that will fill in one of the four blank cells: only 3 college students failed the test. With that additional information in place, you should now be able to fill in the remaining cells of the two-by-two table (2 pts for the table). Comment on why that one additional piece of information was all you needed in order to figure out all four of the table's main cells (1 pt). Next, create a second copy of the complete table, replacing the counts of students with probabilities. Finally, what is the pass rate for high school students? In other words, if one focuses only on high school students, what is the probability that a student will pass the test? (1 pt)

```
#  Chapter 2, Exercise 6  (1)

cat("\n*************  Chapter 2, Exercise 6   **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 6   **************************************
```

```
stat_exam <- matrix(c(35,15,47,3),ncol = 2,byrow = TRUE)
colnames(stat_exam) <- c("Pass","Fail") # Label the columns
rownames(stat_exam) <- c("HighSchool","College") # Label the rows

# margin.table(stat_exam) # This is the grand total of stat exam attempts
# margin.table(stat_exam,1) # These are the marginal totals for rows
# margin.table(stat_exam,2) # These are the marginal totals for columns

stat_examMargin <- addmargins(stat_exam) # Add marginal totals on rows and columns

cat("\nContingency Tables Creation:\n")
```

```
##
## Contingency Tables Creation:
```

```
# stat_exam # Report Contingency results to console
cat("\nContingency Table of Stat exam with marginal results:\n")
```

```
##
## Contingency Table of Stat exam with marginal results:
```

```
stat_examMargin # Report Marginal results to console
```

```
##            Pass Fail Sum
## HighSchool   35   15  50
## College      47    3  50
## Sum          82   18 100
```

```
cat("\nWhy enough data?\n")
```

```
##
## Why enough data?
```

```
cat("
  Below are the findings from the quetion
  * Total Number of Students is 100 (sum of all edges)
  * Marginal Row totals for High School students is  50
  * Marginal Row totals for College students is  50
  * Marginal Column totals for Pass students is  82
  * Marginal Column totals for Fail students is  18

    Lets build a 2 *2 contingency table based on the observations;
```

```
                      Pass  | Fail      |    Marginal
  ----------    |-----------|-----------|----------
  High School   |       P1    |      F1   |     50
  College         |     P2    |      F2   |     50
  Marginal      |     82    |      18   |     100

  In addition, It was mentioned that Only 3 College Students have failed.
  given,  F2 = 3; then P2 = 50 - 3 = 47; and F1 = 18-3 = 15;

  if P2 = 47; then P1 = 82-47 = 35; We can also verify P1 + F1 = 35 + 15 = 50; P2 + F2 = 47 + 3 = 50

                    Pass  | Fail  |  Marginal
  ----------    |---------|-------|----------
  High School   |     35    |    15   |   50
  College         |   47    |  03   |     50
  Marginal      |   82    |  18   |   100

  Next step is to calculate probability of each events. So, simply divide the values
  P1,P2, F1, F2 by 100. i.e total probability is 1 ( sum of edges)

  Probablity Values of Students taken the Stat Exam,

                    Pass  | Fail  |  Marginal
  ----------    |---------|-------|----------
  High School   |     .35   |   .15   |   .50
  College         |   .47   | .03   |     .50
  Marginal      |   .82   | .18   |   1

")
```

```
##
##   Below are the findings from the quetion
##   * Total Number of Students is 100 (sum of all edges)
##   * Marginal Row totals for High School students is  50
##   * Marginal Row totals for College students is  50
##   * Marginal Column totals for Pass students is  82
##   * Marginal Column totals for Fail students is  18
##
##     Lets build a 2 *2 contingency table based on the observations;
##                         Pass  | Fail      |    Marginal
##   ----------    |-----------|-----------|----------
##   High School   |       P1    |      F1   |     50
##   College         |     P2    |      F2   |     50
##   Marginal      |     82    |      18   |     100
##
##   In addition, It was mentioned that Only 3 College Students have failed.
##   given,  F2 = 3; then P2 = 50 - 3 = 47; and F1 = 18-3 = 15;
##
##   if P2 = 47; then P1 = 82-47 = 35; We can also verify P1 + F1 = 35 + 15 = 50; P2 + F2 = 47 + 3 = 50
##
##                       Pass  | Fail  |   Marginal
##   ----------    |---------|-------|----------
##   High School   |      35    |    15   |   50
##   College          |   47    |  03   |     50
##   Marginal      |   82    |  18   |   100
##
##   Next step is to calculate probability of each events. So, simply divide the values
##   P1,P2, F1, F2 by 100. i.e total probability is 1 ( sum of edges)
##
##   Probablity Values of Students taken the Stat Exam,
##
##                     Pass  | Fail  |   Marginal
##   ----------    |---------|-------|----------
##   High School   |      .35   |   .15   |   .50
##   College          |   .47   | .03   |     .50
##   Marginal      |   .82   | .18   |   1
##
```

```
cat("\nWhat is the Pass rate of high School students?\n")
```

```
##
## What is the Pass rate of high School students?
```

```
cat("

from the above contingency tables, its clear that 35 high School students passed
the Stat Exam out of total 50 high school students (focusing only on High School Students)
So, the pass rate would be 35/50 *100 = 70%

")
```

```
##
##
## from the above contingency tables, its clear that 35 high School students passed
## the Stat Exam out of total 50 high school students (focusing only on High School Students)
## So, the pass rate would be 35/50 *100 = 70%
##
```

```
cat("\nContingency Table of Stat exam with Probablity & marginal results:\n")
```

```
##
## Contingency Table of Stat exam with Probablity & marginal results:
```

```
examMargin <- margin.table(stat_exam)
stat_examProbs <- stat_exam/examMargin
stat_examProbs <- addmargins(stat_examProbs)
stat_examProbs # Report probabilities to console
```

```
##            Pass Fail Sum
## HighSchool 0.35 0.15 0.5
## College    0.47 0.03 0.5
## Sum        0.82 0.18 1.0
```

```
tot_HS_Students <- margin.table(stat_exam,1)[1] # total number of High School Students
pass_HS_Students <- stat_exam[,1][1] # total number of High School Students Passed
HS_Pass_Percentage <- (pass_HS_Students/tot_HS_Students) * 100 # Calculate Success rate of High School
cat("\n Pass/Success rate of High School students with this trial:",HS_Pass_Percentage,"%")
```

```
##
##  Pass/Success rate of High School students with this trial: 70 %
```

```
cat("\n*************  Chapter 2, Exercise 6  - End **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 6  - End **************************************
```

Chapter 2, Exercise 7

In a typical year, 75 out of 100,000 homes in the United Kingdom is repossessed by the bank because of mortgage default (the owners did not pay their mortgage for many months). Barclays Bank has developed a screening test that they want to use to predict whether a mortgagee will default. The bank spends a year collecting test data (conveniently, also on 100,000 households): 93,954 households pass the test and 6,046 households fail the test. Interestingly, 5,997 of those who failed the test were actually households that were doing fine on their mortgage (i.e., they were not defaulting and did not get repossessed). Construct a complete contingency table from this information. (2 pts) Hint: The 5,997 is the only number that goes in a cell; the other numbers are marginal totals. What percentage of customers both pass the test and do not have their homes repossessed? (1 pt)

```
#  Chapter 2, Exercise 7

cat("\n*************  Chapter 2, Exercise 7   **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 7   **************************************
```

```
cat("
Below are the findings from the Barclays Bank Screning test
  * Total Number of households screened 100,000 (sum of all homes)
  * Marginal Column totals for households pass the test is  93,954
  * Marginal Column totals for households fail the test is  6,046
  * Marginal Row totals for households with mortgage default( true condition) is  75
  * Cell Value for True Negative Scenario of households that did not default but marked as default are

 Lets build a contingency table based on the observations;

 True condition   |    Test result      |
 (is Defaulter)   |  Postive | Negative |
------------------+----------+----------|
   True           |   True   |   False  | Sensitivity
                  | Positive | Negative |
------------------+----------+----------|
  False           |  False   |   True   | Specificity
                  | Positive | Negative |
------------------+----------+----------|

Populate the above table with values , based on our observation from the trials

In addition, It was mentioned that TrueN Negative value as 5,997.

With Marginal value of Negative result (column total) as 6,046 -- 5,997 would go into TN Cell;
Hence, 6046 - 5997 = 49 becomes FALSE Negative Value.
With total number of households as 100,000 and Actual defualts as 75 ;
actual non-defaulters would become 100,000- 75 = 999,25
With Above, False Positive on Non-Defualters result would become 939,28 ( i.e. 999,25 - 939,28)


 True condition   |    Test result      |
 (Defaulter)      |  Positve | Negative | Marginal Sum
------------------+----------+----------|-------------
  Yes             |   26     |   49     | 75
------------------+----------+----------|-------------
  No              |  939,28  |   5,997  | 999,25
------------------+----------+----------|-------------
  Marginal Sum    |  93,954  |   6,046  | 100,000

  Lets convert this into Probablities ( divide each cell by 100,000 -total number of samples/homes)

  True condition  |    Test result      |
 (is Defaulter)   |  PASS    |  FAIL    | Marginal Sum
------------------+----------+----------|-------------
  Yes             |   .00026 |   .00049 | .00075
------------------+----------+----------|-------------
  No              |   .93928 |   .05997 | .99925
------------------+----------+----------|-------------
  Marginal Sum    |   .93954 |   .06046 |    1

  *
  * What percentage of customers both pass the test and do not have their homes repossessed?

    In this case, critieria that fits both criteria as Non-Defaulter , and System result returing as Tru
    That would be False Positive ,here. As the earlier assumption to look for Is defualter.
    So, in this case - home owners are deafualters + test confirms too.

    Percentage will be (93928/99925) *100 = 93.99% ( ~94%) is the percentage of
    customers both pass the test and do not have thier homes re-possessed.
   ")
```

```
##
## Below are the findings from the Barclays Bank Screning test
##   * Total Number of households screened 100,000 (sum of all homes)
##   * Marginal Column totals for households pass the test is  93,954
##   * Marginal Column totals for households fail the test is  6,046
##   * Marginal Row totals for households with mortgage default( true condition) is  75
##   * Cell Value for True Negative Scenario of households that did not default but marked as default a
##
##  Lets build a contingency table based on the observations;
```

```
##
##  True condition   |    Test result      |
##  (is Defaulter)   |  Postive | Negative |
## ------------------+----------+----------|
##    True           |   True   |   False  | Sensitivity
##                   | Positive | Negative |
## ------------------+----------+----------|
##   False           |  False   |   True   | Specificity
##                   | Positive | Negative |
## ------------------+----------+----------|
##
## Populate the above table with values , based on our observation from the trials
##
## In addition, It was mentioned that TrueN Negative value as 5,997.
##
## With Marginal value of Negative result (column total) as 6,046 -- 5,997 would go into TN Cell;
## Hence, 6046 - 5997 = 49 becomes FALSE Negative Value.
## With total number of households as 100,000 and Actual defualts as 75 ;
## actual non-defaulters would become 100,000- 75 = 999,25
## With Above, False Positive on Non-Defualters result would become 939,28 ( i.e. 999,25 - 939,28)
##
##
##  True condition   |    Test result      |
##  (Defaulter)      |  Positve | Negative | Marginal Sum
## ------------------+----------+----------|-------------
##   Yes             |   26     |   49     | 75
## ------------------+----------+----------|-------------
##   No              |  939,28  |   5,997  | 999,25
## ------------------+----------+----------|-------------
##   Marginal Sum    |  93,954  |   6,046  | 100,000
##
##   Lets convert this into Probablities ( divide each cell by 100,000 -total number of samples/homes)
##
##   True condition  |    Test result      |
##  (is Defaulter)   |  PASS    |  FAIL    | Marginal Sum
## ------------------+----------+----------|-------------
##   Yes             |   .00026 |   .00049 | .00075
## ------------------+----------+----------|-------------
##   No              |   .93928 |   .05997 | .99925
## ------------------+----------+----------|-------------
##   Marginal Sum    |   .93954 |   .06046 |    1
##
##   *
##   * What percentage of customers both pass the test and do not have their homes repossessed?
##
##     In this case, critieria that fits both criteria as Non-Defaulter , and System result returing as
##     That would be False Positive ,here. As the earlier assumption to look for Is defualter.
##     So, in this case - home owners are deafualters + test confirms too.
##
##     Percentage will be (93928/99925) *100 = 93.99% ( ~94%) is the percentage of
##     customers both pass the test and do not have thier homes re-possessed.
##
```

```
cat("\n*************  Chapter 2, Exercise 7  - End ************************************\n")
```

```
##
## *************  Chapter 2, Exercise 7  - End ************************************
```

Chapter 2, Exercise 8

Imagine that Barclays Bank deploys the screening test from Exercise 7 on a new customer and the new customer fails the test. What is the probability that this customer will actually default on his or her mortgage? Show your work and especially show the tables that you set up to help with your reasoning. (1 pt)

```
#  Chapter 2, Exercise 8

cat("\n*************  Chapter 2, Exercise 8   **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 8   **************************************
```

```
cat(" Find out what is the probability that new customer will actually default on his/her mortgage

Lets look at both tables , again.
    True condition   |    Test result      |
 (Defaulter)      |  Positve | Negative | Marginal Sum
------------------+----------+----------|-------------
  Yes             |   26     |   49     | 75
------------------+----------+----------|-------------
  No              |  939,28  |   5,997  | 999,25
------------------+----------+----------|-------------
  Marginal Sum    |  93,954  |   6,046  | 100,000

  Lets convert this into Probablities ( divide each cell by 100,000 -total number of samples/homes)

  True condition  |    Test result      |
 (is Defaulter)   |  PASS    |  FAIL    | Marginal Sum
------------------+----------+----------|-------------
  Yes             |   .00026 |   .00049 | .00075
------------------+----------+----------|-------------
  No              |   .93928 |   .05997 | .99925
------------------+----------+----------|-------------
  Marginal Sum    |   .93954 |   .06046 |    1

  * Here the probability that new customer is both defaulter and
  trial predicts correctly , would be (26/75)*100 = 34.66%

  As you may see, total number of actual defaults is 75.
  Trial predicted 26 of them correctly. So, the probability is 26 over 75 as ~35%

    ")
```

```
##  Find out what is the probability that new customer will actually default on his/her mortgage
```

```
##
## Lets look at both tables , again.
##     True condition   |    Test result      |
##  (Defaulter)      |  Positve | Negative | Marginal Sum
## ------------------+----------+----------|-------------
##   Yes             |   26     |   49     | 75
## ------------------+----------+----------|-------------
##   No              |  939,28  |   5,997  | 999,25
## ------------------+----------+----------|-------------
##   Marginal Sum    |  93,954  |   6,046  | 100,000
##
##   Lets convert this into Probablities ( divide each cell by 100,000 -total number of samples/homes)
##
##   True condition  |    Test result      |
##  (is Defaulter)   |  PASS    |  FAIL    | Marginal Sum
## ------------------+----------+----------|-------------
##   Yes             |   .00026 |   .00049 | .00075
## ------------------+----------+----------|-------------
##   No              |   .93928 |   .05997 | .99925
## ------------------+----------+----------|-------------
##   Marginal Sum    |   .93954 |   .06046 |    1
##
##   * Here the probability that new customer is both defaulter and
##   trial predicts correctly , would be (26/75)*100 = 34.66%
##
##   As you may see, total number of actual defaults is 75.
##   Trial predicted 26 of them correctly. So, the probability is 26 over 75 as ~35%
##
##
```

```
cat("\n*************  Chapter 2, Exercise 8  - End **************************************\n")
```

```
##
## *************  Chapter 2, Exercise 8  - End **************************************
```