# IST772 Problem Set 2 Fall 2020

Shivani Sanjay Mahaddalkar

The homework for week two is based on exercises 1 and 2 on page 35, as well as problems 6, 7, and 8 on page 36, but with changes as noted in the text in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 1. I did this homework by myself, with help from the book and the professor 2. I did this homework with help from the book and the professor and these Internet sources: 3. I did this homework with help from  but did not cut and paste any code

Set the random number seed so that your results will match mine.

```
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).*

On flipping a physical fair coin by hand seven times, the number of heads obtained were 2.

```
flips <- table(rbinom(50000, 7, 0.5))
flips
```

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

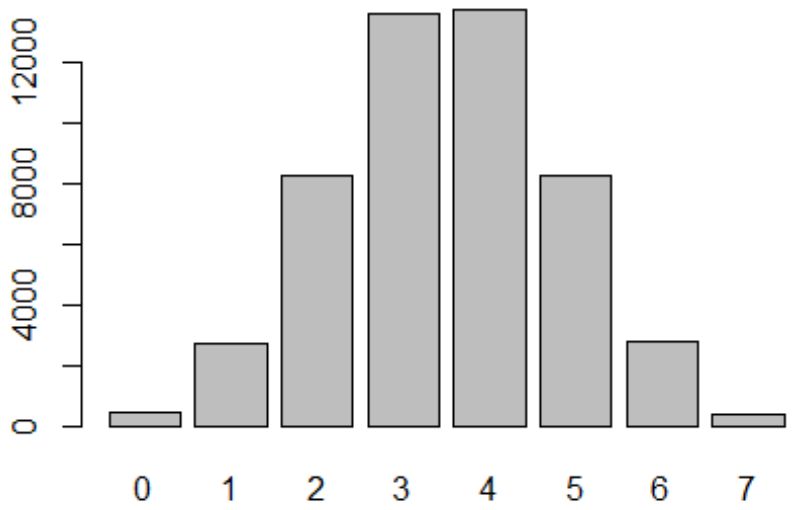
It can be seen from the table that the most frequently occurring number of heads was 3 and 4. It was expected to be a symmetrical binomial distribution function with higher frequency in the middle values. Because this is an experimental trial, the curve will be not be exactly symmetrical, but close to it. The extreme values appear with much lesser frequency.

## Chapter 2, Exercise 2

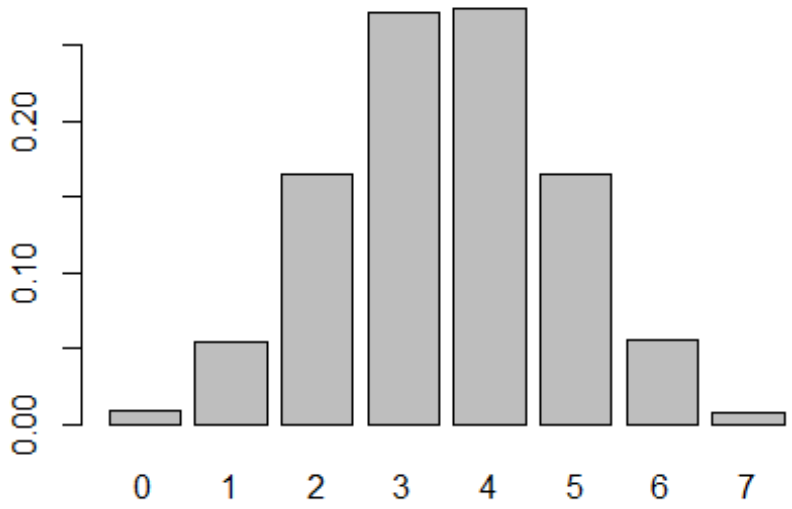
*Using the output from Exercise 1, summarize the results of your 50,000 trials of 7 flips each in a bar plot using the appropriate commands in R. Convert the results to probabilities and represent that in a bar plot as well (1 pt for the two bar plots). Write a brief interpretive analysis that describes 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).*

```
barplot(flips)
```




```
barplot(flips/sum(flips))
```




The two barplots are exactly similar in shape but different in scale. Since the second one is the probability of the occurrences it ranges from 0-1. The graph is almost symmetrical and shows that the probability of there being 3 heads or 4 heads, which is the centre, is higher than the rest of the possibilities. A binomial distribution is symmetrical and with the highest probabilities being around the centre and the extreme values have lowest probabilities of occurring. Since the number of possible outcomes is an even number, the two middle most values have highest frequencies of occurring.

## 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)*

| | HighSchool | College | Sum |
|---|---|---|---|
| Passed | 35 | 47 | 82 |
| Failed | 15 | 3 | 18 |
| Sum | 50 | 50 | 100 |

Since the row and column totals are known, on knowing one cell, the other value in its row and column can be known which leads to the last unknown value in the table.

```
studentMatrix <- matrix(data=c(35,47,15,3), nrow=2, byrow=T, dimnames =
list(c("Passed", "Failed"),c("HighSchool", "College")))
probStudentMatrix <- studentMatrix/sum(studentMatrix)
probStudentMatrix

##        HighSchool College
## Passed       0.35    0.47
## Failed       0.15    0.03

probPassHS <- probStudentMatrix[1,1]/sum(probStudentMatrix[,1])
probPassHS

## [1] 0.7
```

## 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)*

| | Home Not Repossessed | Home Repossessed | Sum |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Passed | 93,928 | 26 | 93,954 |
| Failed | 5,997 | 49 | 6,046 |
| Sum | 99,925 | 75 | 100000 |

```
barclaysMatrix <- matrix(data=c(93928,26,5997,49), nrow=2, byrow=T, dimnames
= list(c("Passed", "Failed"),c("Home Not Repossessed", "Home Repossessed")))
barclaysMatrix
```

```
##        Home Not Repossessed Home Repossessed
## Passed                93928               26
## Failed                 5997               49
```

```
(barclaysMatrix[1,1]/sum(barclaysMatrix))*100
```

```
## [1] 93.928
```

93.928 % of customers both pass the test and do not have their homes repossessed.

## 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)*

```
barclaysMatrix
```

```
##        Home Not Repossessed Home Repossessed
## Passed                93928               26
## Failed                 5997               49
```

```
probDefault <- barclaysMatrix[2,2]/sum(barclaysMatrix[2,])
```

The probability of a person defaulting given that they have failed the test is 0.008104532