IST772 Problem Set 2 Fall 2020

Abhijith Anil Vamadev

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) 3. I did this homework with help from Viktor 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).

```
#Number of heads and tails done by hand
#Heads(0) - 2; Tails(1) - 5
set.seed(772)
physical_coin <- table(rbinom(n = 50000, size = 7, prob = 0.5))
physical_coin
```

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

- So the function produces a random number from the binoamial series with 50,000 trials with two sizes 0 to 7 with each having a fair chance of landing on Heads or Tails. The numbers from 0 to 7 indicate the number of heads only, heads and tails mixed and so on, and the number of times it landed the respective number of heads.

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).

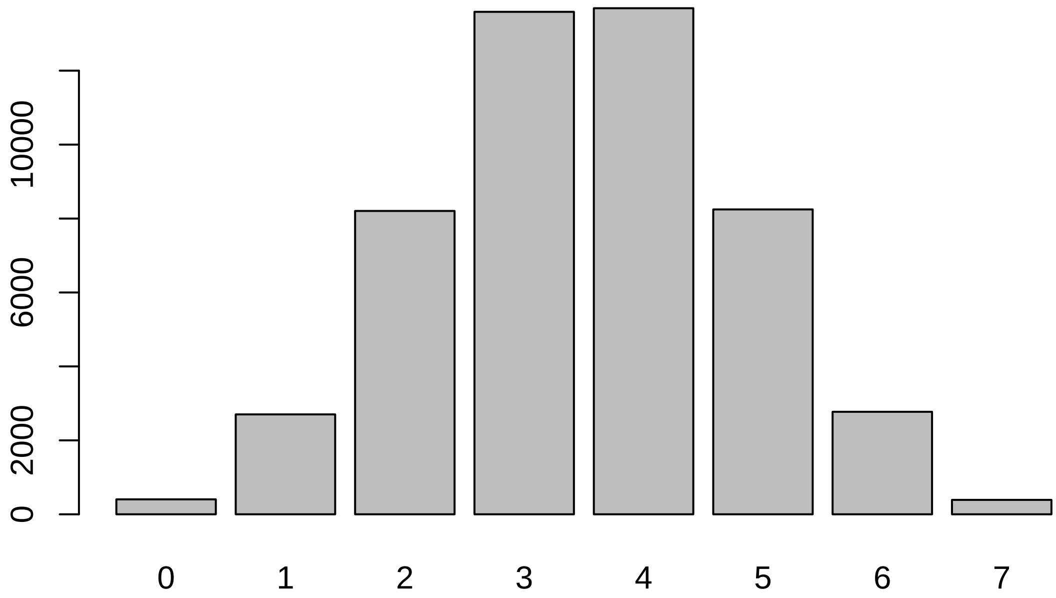
```
set.seed(772)
physical_coin <- table(rbinom(n = 50000, size = 7, prob = 0.5))
physical_coin_prob <- table(rbinom(n = 50000, size = 7, prob = 0.5))/50000
physical_coin
```

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

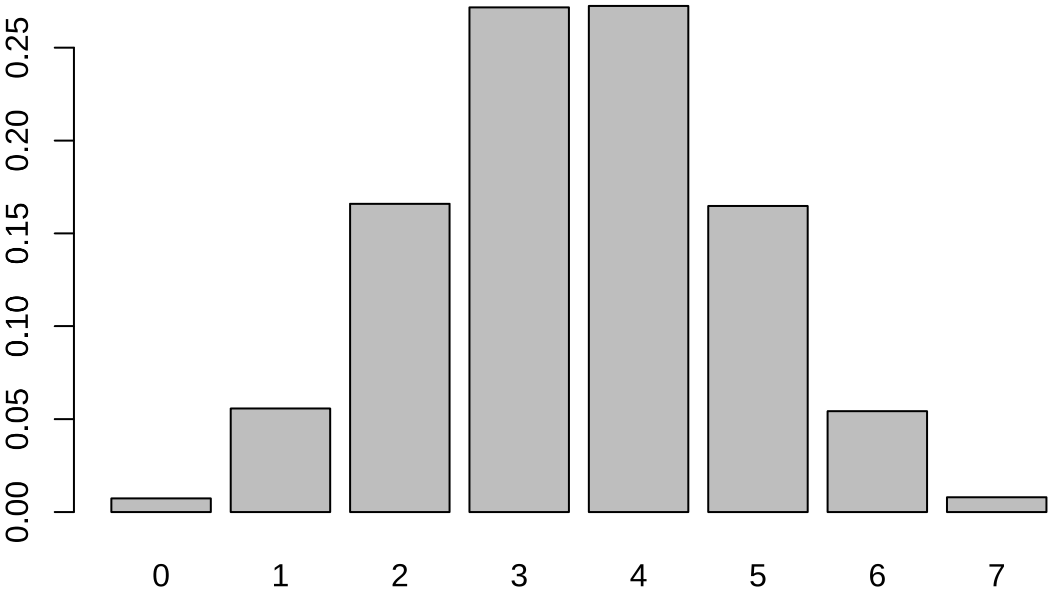
```
physical_coin_prob
```

```
## 
##       0       1       2       3       4       5       6       7 
## 0.00730 0.05572 0.16600 0.27168 0.27246 0.16470 0.05422 0.00792
```

```
barplot(physical_coin)
```



```
barplot(physical_coin_prob)
```



* The first bar plot indicates for each value on the x-axis the number of times the number fell for that respecive value in the x-axis after the coin toss.The second bar plot signify the same thing, expect on the y-axis the indicates the probablity of landing in the respective x-axis value. The first two plots are related as they indicate the exact numbers vs the probability, each of the value in the first graph is divided by 50,000. * The shape is almost like a bell shaped.The center is mostly in 3 and 4, because that's the most common number of types in the coins landed either 3 or 4 heads. As we keep running the number of times the coins are tossed, the shape of the graph will become a bell curve.

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) 100 Students 50 - High School 50 - College Students Passed - 82 Students Failed - 18 Passed HS - 47 Passed CO - 47 * If we only look on the passed students from Highschool the probability a student passes would be the marginal probabiity of Highschool students which is 0.62, and by taking the Passed HighSchool students which is 0.35/0.50 = 0.70

```
new_table <- matrix(data = c(35, 15, 47, 3), nrow = 2, byrow = F, dimnames = list(c("Passed", "Failed")
#new_table <- addmargins(new_table)
new_table_prob <- new_table/sum(new_table)
addmargins(new_table)
```

```
##        HighSchool College Students Sum
## Passed         35               47  82
## Failed         15                3  18
## Sum            50               50 100
```

```
new_table_prob
```

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

- With the new information we can fill in the Passed number of students which is 50-3 = 47, and likewise using the marginal totals we can deduce each of the answers, for each of the column and rows.

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)

```
mortgage <- matrix(data = c(26, 49, 93928, 5997), nrow = 2, byrow = F, dimnames = list(c("Passed", "Fail
addmargins(mortgage)
```

```
##        Default Non-defaulted    Sum
## Passed      26         93928  93954
## Failed      49          5997   6046
## Sum         75         99925 100000
```

```
mortgage_prob <- mortgage/sum(mortgage)
mortgage_prob
```

```
##        Default Non-defaulted
## Passed 0.00026       0.93928
## Failed 0.00049       0.05997
```

- Percentage of customers that passed and did not have their homes repossessed is 93928/1000000 = 0.93 * 100 = 93.92%.

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)

```
new_table_prob <- mortgage/sum(mortgage)
addmargins(new_table_prob)
```

```
##        Default Non-defaulted     Sum
## Passed 0.00026       0.93928 0.93954
## Failed 0.00049       0.05997 0.06046
## Sum    0.00075       0.99925 1.00000
```

- So here, we are only looking at the failed row. So we take the overall marginal total of that row which is 0.6046. Now the customer defaults, which would be 0.00049/0.6406 = 0.0081.