IST 772 Homework 2

Maya Mileva

due date: Oct 17th, 2019

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

```
## Run these three functions to get a clean test of homework code
## dev.off() # Clear the graph window
cat('\014')  # Clear the console
```

```
rm(list=ls()) # Clear user objects from the environment
```

Exercises

1. Flip a fair coin nine times and write down the number of heads obtained. Now repeat this process 100,000 times. Obviously you don't want to have to do that by hand, so cre- ate the necessary lines of R code to do it for you. Hint: You will need both the rbinom() function and the table() function. Write down the results and explain in your own words what they mean.

With the first task we have one trial of coin tosses where eacht trial contain 9 coin flips. After that we are doing 100 000 trials where each trial contain 9 coin flips.

```
## Outcome tables
table(rbinom(n=1, size = 9, prob = 0.5)) # 4 heads obtained
```

```
##
## 3
## 1
```

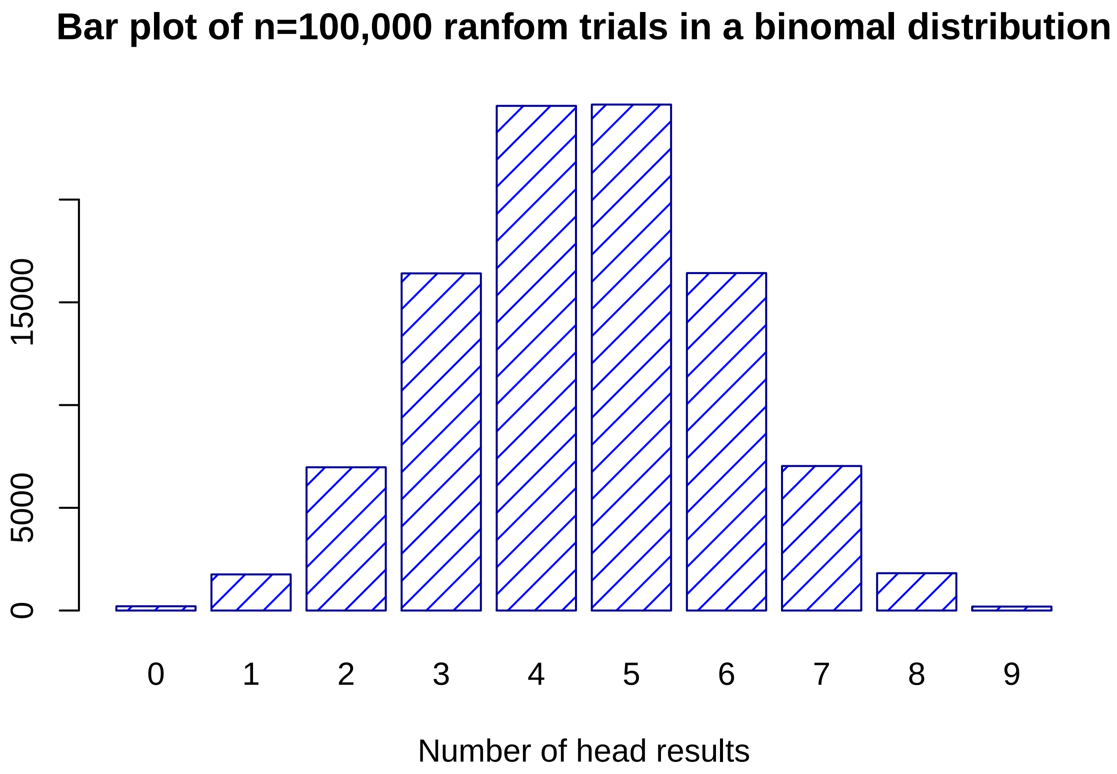
```
## produces random outcomes of binomal distribution with prob 50%
## n=9 - 9 coin flips events in each trial, 100 000 - trials
table(rbinom(n=100000, size = 9, prob = 0.5))# summarize the resulting list of trials
```

```
##
##     0     1     2     3     4     5     6     7     8     9
##   174  1777  6859 16556 24772 24504 16345  7049  1771   193
```

The top line of the table showsthe number of times we got head when we flipped the coin. We don't need to show the number of times we got tails, because is nine minus the number of the head. We can actually notice 10 scenarios. The second line summarizes what happened: how many of the trials come out with the head as a result. Out of the 100 000 trials only 203 had no head as a result, 1749 trials had 1 head result, 6991 had 2 head and so on.

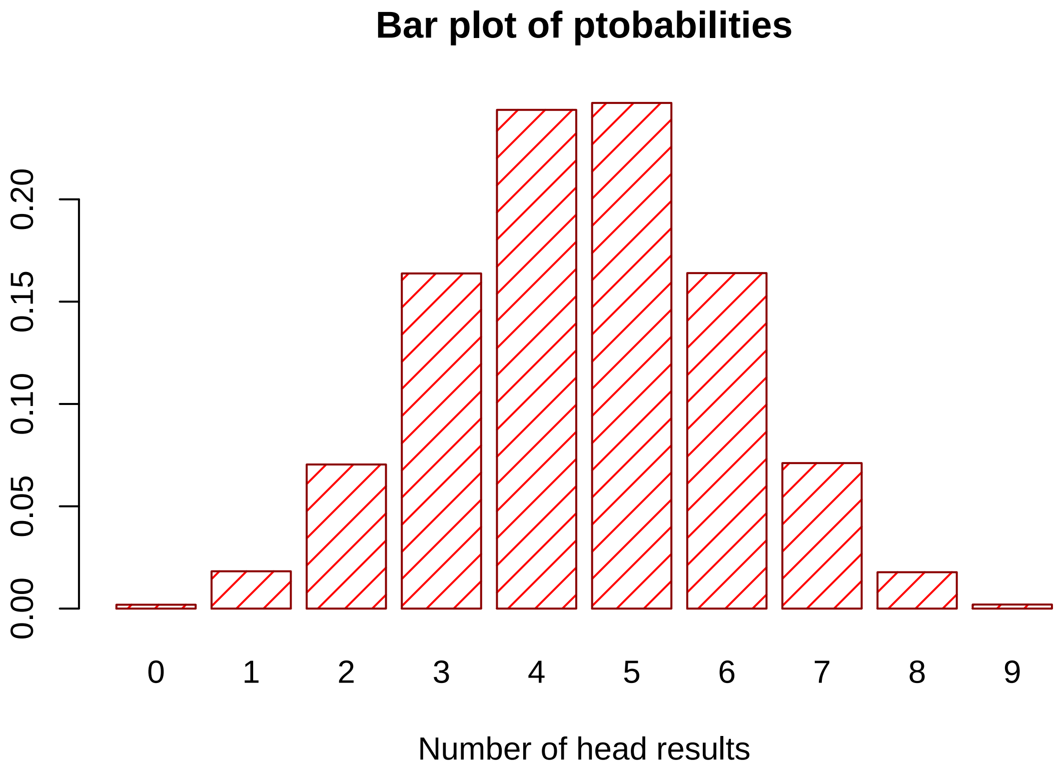
2. Using the output from Exercise 1, summarize the results of your 100,000 trials of nine flips each in a bar plot using the appropriate commands in R. Convert the results to probabilities and represent that in a bar plot as well. Write a brief interpretive analysis that describes what each of these bar plots signifies and how the two bar plots are related. 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.

```r
x <- table(rbinom(n=100000, size = 9, prob = 0.5)) # saving it into var x
barplot(x
        , main="Bar plot of n=100,000 ranfom trials in a binomal distribution"
        , xlab="Number of head results"
        , border="darkblue"
        , col="blue"
        , density=10)
```



```r
probTable <- table(rbinom(n=100000, size = 9, prob = 0.5))/100000 # devide each
# value of the table by total number ot trials go get the probabilities

barplot(probTable
        , main="Bar plot of ptobabilities"
        , xlab="Number of head results"
        , border="darkred"
        , col="red"
        , density=10)
```



The first bar plot is bar plot of n=100,000 ranfom trials in a binomal distribution. We can see that 4 and 5 are the most common. The x-axis are the different categories of outcome, and the y-axis are the event that count in each category.

The second bar plot is a bar plot of probabilities of each trial for n=100,000 random trials in a binomial distribution. All the probabilities add up to 1. The x-axis are the different categories of outcome, and the y-axis are the probabilities each category - 16% for the 3 category, 7% for the 7th category and so on.

The only thing that changed in the second graph(other than the hight of the bars, due to the new set of random trials) is the y-axis(now shows probabilities instead of row counts).

Both of the barplots are bell shaped. That‘s due to the random selection ot the random numbers generated by rbinom() function. The center of the bar plot is 4.5(mean=med) which is another sign for the normal distribution.

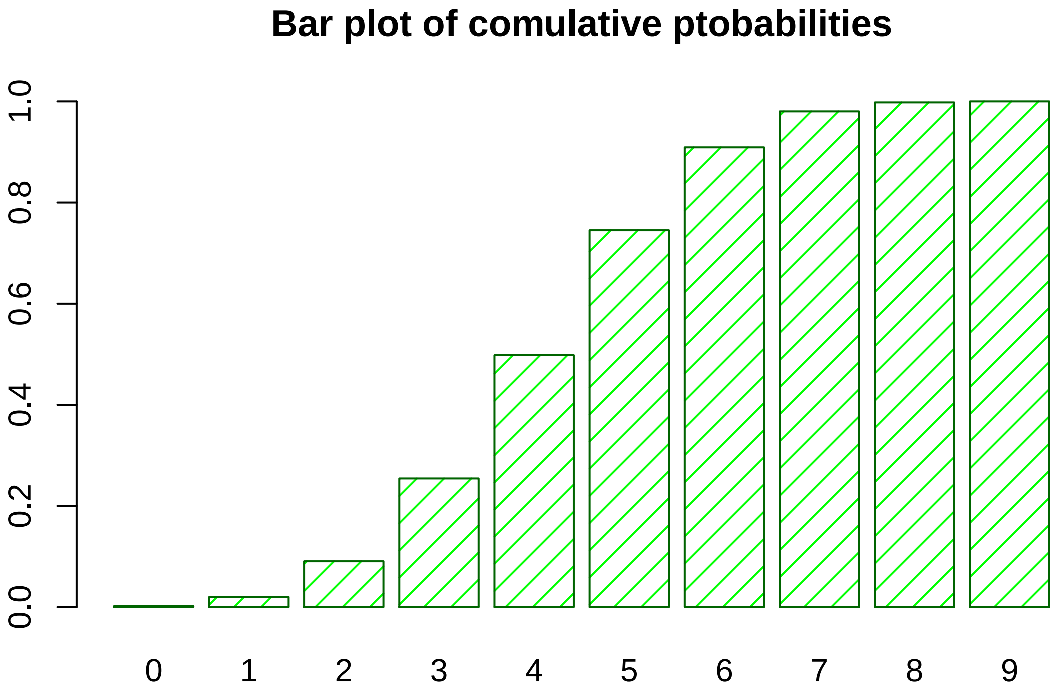
```
probTable # displaying the table
```

```
##
##       0       1       2       3       4       5       6       7       8
## 0.00190 0.01824 0.07043 0.16378 0.24372 0.24711 0.16395 0.07110 0.01778
##       9
## 0.00199
```

```
cumsum(probTable) # add each successive probability to the sum of the previous one
```

```
##       0       1       2       3       4       5       6       7       8
## 0.00190 0.02014 0.09057 0.25435 0.49807 0.74518 0.90913 0.98023 0.99801
##       9
## 1.00000
```

```
barplot(cumsum(probTable) # barplot of the comulative probabilities, 1 under the curve
         , main="Bar plot of comulative ptobabilities"
        , border="darkgreen"
        , col="green"
        , density=10)
```



3. (6)One hundred students took a statistics test. Fifty of them are high school students and 50 are college students. Eighty students passed and 20 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). Draw that table and write in the marginal totals. I'm now going to give you one additional piece of information that will fill in one of the four blank cells: only *three* 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. 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. Finally, create a second copy of the complete table, replacing the counts of students with probabilities. 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?

I am going to use R to build my contigency table. Total number of students took a statistic test is 100(50 hight school, 50 colledge). 80 pass, 20 fail(3 colledge students failed, so 47 passed). We know that total 50 colledge students tool the test, but had no idea how many of the passes 20 are from colledge and how many came from high school. We that information we can simply calculate: **hight school** - 33 passed, 17 failed. **colledge** - 47 passed, 3 failed. Now we are ready to build the table using R markdown.

```r
## Create 2 col. 2 row structure using the matrix() command
## Fill the cells row-by-row, (33, 17(first row) and  47,3(second row))
statTest <- matrix(c(33,17,47,3), ncol=2, byrow = TRUE) # creating the matrix

## Label the row and columns
colnames(statTest) <- c("Pass", "Fail")
rownames(statTest) <- c("High school students", "Colledge students")

## Convert the matrix to a table
# 34% fail rate <- as.table(statTest)
statTest # contigency table
```

```
##                      Pass Fail
## High school students   33   17
## Colledge students      47    3
```

```r
## Adding margins
margin.table(statTest) # grand total
```

```
## [1] 100
```

```r
margin.table(statTest, 1) # row`s total
```

```
## High school students    Colledge students
##                   50                   50
```

```r
margin.table(statTest, 2) # col`s total
```

```
## Pass Fail
##   80   20
```

```r
## Calculating probabilities
statTestPROB <- statTest/margin.table(statTest) # deviding each cell to grand total
statTestPROB # probaility table
```

```
##                      Pass Fail
## High school students 0.33 0.17
## Colledge students    0.47 0.03
```

In ordert to create a table replacing the count of the students with the probabilities we have to devide each cell count by the total number of events(100 in this case). (The pass rate for high school students is 33%(0.33) **when we look at all the students.**)

If we look only the high school students we have different picture. Normalizing the table by dividing each cell by the grand total will give us the probability we are looking for.

```r
statTest[1,1]/margin.table(statTest, 1)[1] # 66% pass rate
```

```
## High school students
##                 0.66
```

```
statTest[1,2]/margin.table(statTest, 1)[1] # 34% fail rate
```

```
## High school students
##                 0.34
```

```
#33/50 # 66% pass rate
#17/50 # 34% fail rate
```

The ratio pass:fail is 33:17 - roughly every second student pass. The probability a high school student pass the test is 66%(0.66)

4. (7)In a typical year, 71 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: 93,935 households pass the test and 6,065 households fail the test. Interestingly, 5,996 of those who failed the test were actually households that were doing fine on their mort- gage (i.e., they were not defaulting and did not get repossessed). Construct a complete contingency table from this information. Hint: The 5,996 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?

```
6065-5996 # 69 househols fail repossessed
```

```
## [1] 69
```

```
71-69 # 2 households of 100, 000 homes that passed got repossesed
```

```
## [1] 2
```

```
93935-2 # 93,933 households were passed not repossessed
```

```
## [1] 93933
```

```
93933+5996 # total not-repossessed households
```

```
## [1] 99929
```

```
## Create 2 col. 2 row structure using the matrix() command
## Fill the cells row-by-row
data <- matrix(c(2,69,93933,5996), ncol=2, byrow = TRUE) # creating the matrix

## Label the row and columns
colnames(data) <- c("Pass", "Fail")
rownames(data) <- c("Repossessed", "Not repossessed")

## Convert the matrix to a table
data <- as.table(data)
data # contigency table
```

```
##                    Pass  Fail
## Repossessed           2    69
## Not repossessed 93933  5996
```

```
## Adding margins
margin.table(data) # grand total
```

```
## [1] 1e+05
```

```
margin.table(data, 1) # row`s total
```

```
##     Repossessed Not repossessed
##              71           99929
```

```
margin.table(data, 2) # col`s total
```

```
##  Pass  Fail
## 93935  6065
```

```
## Calculating probabilities
dataPROB <- data/margin.table(data) # deviding each cell to grand total
dataPROB # probaility table
```

```
##                     Pass    Fail
## Repossessed     0.00002 0.00069
## Not repossessed 0.93933 0.05996
```

```
#dataPROB <- addmargins(dataPROB)

margin.table(dataPROB)
```

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

```
margin.table(dataPROB, 1)
```

```
##     Repossessed Not repossessed
##         0.00071         0.99929
```

```
margin.table(dataPROB, 2)
```

```
##    Pass    Fail
## 0.93935 0.06065
```

```
dataPROB[2,1]/margin.table(dataPROB, 2)[1] #99%
```

```
##      Pass
## 0.9999787
```

```
#0.93933/0.93935
```

99% of customers both pass the test and do not have their home repossessed.

5. (8) Imagine that Barclays deploys the screening test from Exercise 6 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.

We need to look only at the Fail col of the table and isolate it.

```
## Normalized probailities
dataPROB[1,2]/margin.table(dataPROB, 2)[2] # 1%
```

```
##       Fail
## 0.01137675
```

```
# 0.00069/0.06065

dataPROB[2,2]/margin.table(dataPROB, 2)[2] #99%
```

```
##      Fail
## 0.9886232
```

```
# 0.05996/0.06065
```

1% chance a customer dafault(reprocessed) on their mortgage if he or she fail the test. Tables bult in the previous excercise showed me the answer.