IST772 Problem Set 2 Fall 2024

Michael A. d'Amore

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

Be sure also to cite sources and AI usage as detailed in PS 1.

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 five times and write down the number of heads obtained (1 pt). Next you will repeat this process (flipping five coins and counting the heads) 30,000 times and summarize the results. 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 (i.e., how often your simulation came up with each number of heads) and explain in remarks in your own words what they mean (2 pts).

```
#I flipped 5 coins and 3 heads came out.
results <- rbinom(n = 30000, size = 5, prob = 0.5)
summary <- table(results)
print(summary)
```

```
## results
##    0    1    2    3    4    5
##  926 4590 9419 9418 4750  897
```

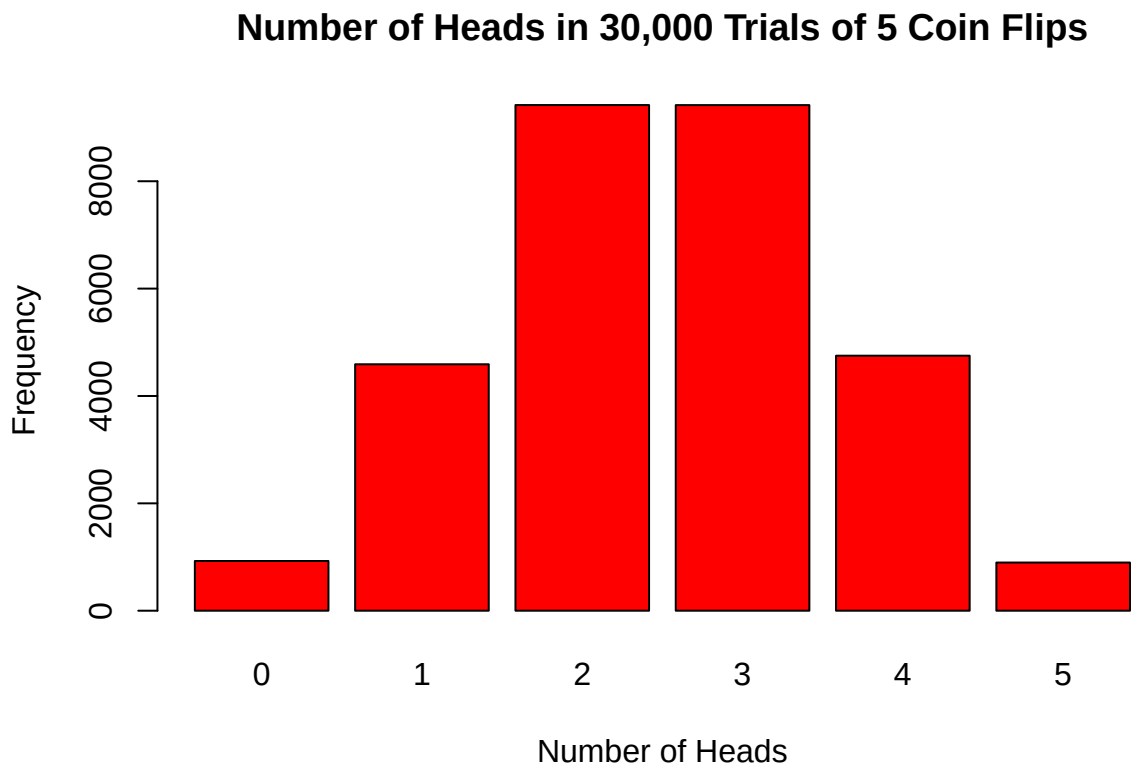
```
#More cases for for 2-3 heads because it is a 50% chance and 50% of 5 is 2.5
#so the most likely number should be either 2 or 3
```

Chapter 2, Exercise 2

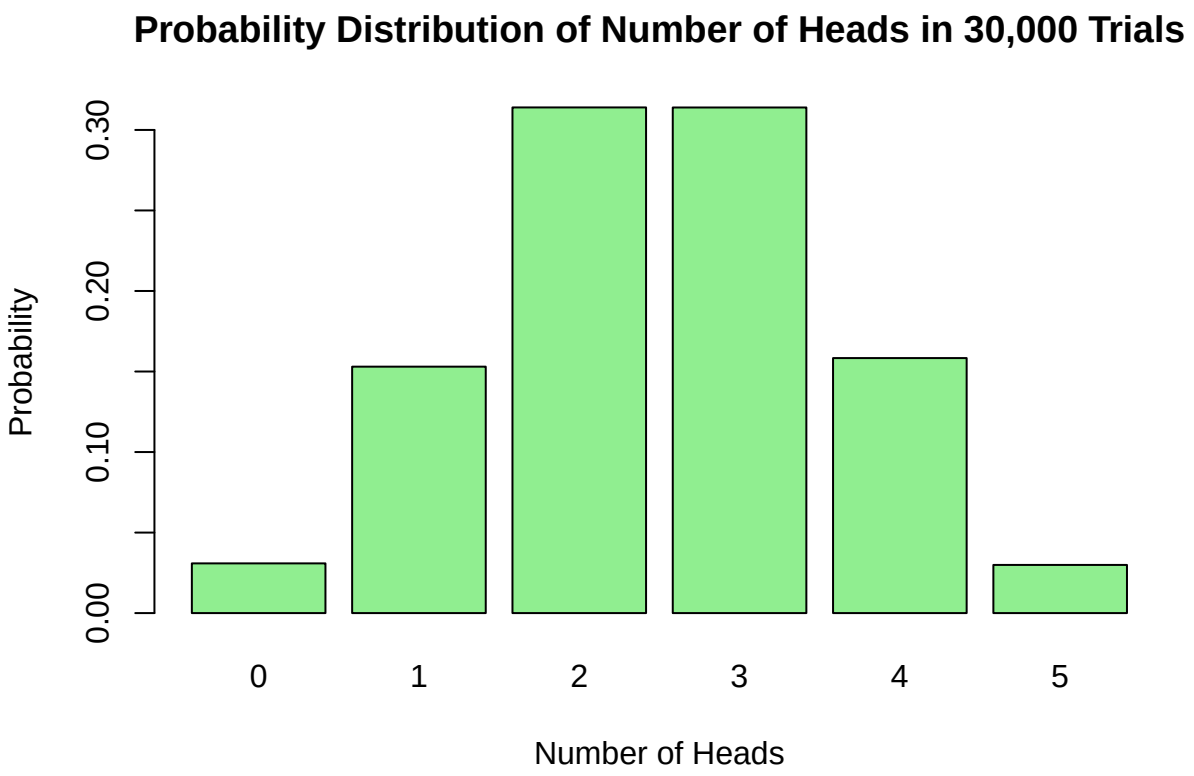
Using the output from Exercise 1, summarize the results of your 30,000 trials of 5 flips each in a bar plot using the appropriate commands in R. Convert the results to probabilities and represent that in a bar plot as well (1 pt for the two bar plots). Write a brief interpretive analysis that describes what each of these bar plots signifies and how the two bar plots are related (1 pt). Make sure to remark 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 explanation of shape; 1 pt for shape and centre).

```
barplot(summary,
        main = "Number of Heads in 30,000 Trials of 5 Coin Flips",
        xlab = "Number of Heads",
        ylab = "Frequency",
        col = "red")
```




```
probabilities <- summary / 30000
barplot(probabilities,
        main = "Probability Distribution of Number of Heads in 30,000 Trials",
        xlab = "Number of Heads",
        ylab = "Probability",
        col = "lightgreen")
```




```
##Plot 1 shows the frequency of each number the shape is a bell curve typical
#of binomial dist. the center is around the 2-3 as expected with a fair coin.
#Plot number 2 is exactly the same as the 1st graph but these are the
#probability of each case instead of the raw frequency
```

Chapter 2, Exercise 6

One hundred students took a statistics test. Fifty of them are high school students and 50 are college students. Eighty-three students passed and 17 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). (No need to explain the detailed computations, but you should show them, not just the completed 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)

```
contingency_table <- matrix(c(36, 14, 47, 3), nrow = 2, byrow = TRUE)

# Add row and column names
rownames(contingency_table) <- c("High School", "College")
```

```r
colnames(contingency_table) <- c("Passed", "Failed")

contingency_totals <- addmargins(contingency_table)

contingency_totals
```

```
##             Passed Failed Sum
## High School     36     14  50
## College         47      3  50
## Sum             83     17 100
```

```r
pass_HS <- contingency_table[1,1] / sum(contingency_table[1,])
cat("\nPass rate for High School students:", pass_HS, "\n")
```

```
##
## Pass rate for High School students: 0.72
```

Chapter 2, Exercise 7

In a typical year, 110 out of 100,000 homes in the United Kingdom are 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 (inconveniently, on only 10,000 households): 9,934 households pass the test and 66 households fail the test. Interestingly, 57 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, assuming that the year of the test is a typical year in terms of mortgage defaults. (2 pts) What percentage of customers both pass the test and do not have their homes repossessed? (1 pt)

```r
#(110/100,000) x 10,000 = 11
#total reposesed is 11
#Test Failed but fine: We know that 57 of the 66 households that failed
#the test did not have their homes repossessed, so they were false positives. #So 9 had them reposed. S

morgage_table <- matrix(c(2, 9, 9932, 57), nrow = 2, byrow = TRUE)
morgage_totals <- addmargins(morgage_table)

morgage_totals
```

```
##               Sum
##        2  9    11
##     9932 57  9989
## Sum 9934 66 10000
```

```r
#passed the test and did not have their homes repossessed
#(9,932/10,000) x 100 = 99.32%
```

Question

You can compute various conditional probabilities from the contingency table in Exercise 7. Which conditional probability do you think is of most importance to the bank and why? (1 pt) What is the numeric value of the selected conditional probability? (1 pt) What does that value suggest about the utility of the proposed test or about how it should be used? (1 pt)

```
#P(Repossessed|TestFailed)= 9/66 = 0.136
#The probability is approximately 13.6%, which means that even when if a
#house fails the test, the likelihood of repossession is relatively low.
#The test may not be highly reliable on its own and could lead to unnecessary
#concern or action on many households that are actually not at risk of
#defaulting.
```