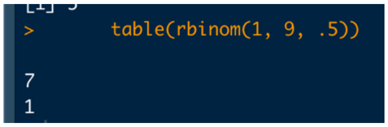
Matthew L. Pergolski
IST 772
Dr. Block
10/16/2021

Homework 2

Beginning Statement
"I produced the material below with no assistance [direct quote from IST 772 class syllabus]."
Note: Homework questions have been copied/pasted into the document for both the student and viewer's convenience.

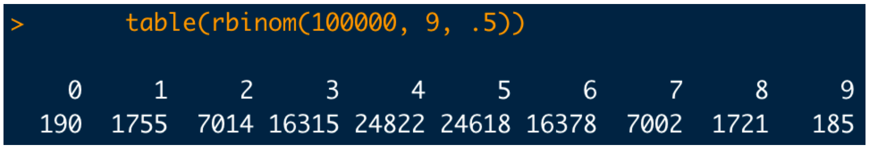
Homework Question 1

After performing one trial of nine occurrences of flipping a coin, I observed 'heads' 7 times.



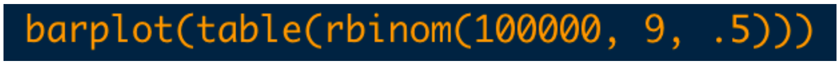
```
[1] 5
>       table(rbinom(1, 9, .5))

7
1
```

The following was derived from R after repeating the process 100,000 times:

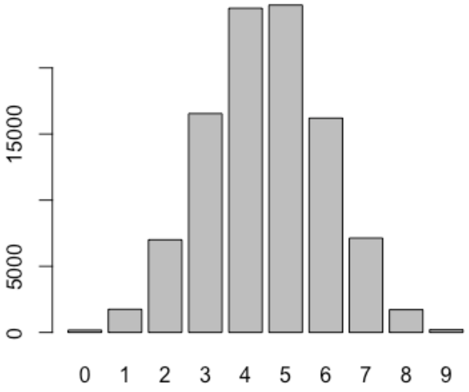


```
>       table(rbinom(100000, 9, .5))

    0     1     2     3     4     5     6     7     8     9 
  190  1755  7014 16315 24822 24618 16378  7002  1721   185 
```

In my own words, I would describe this as an example of a binomial distribution. If performing 100,000 trials of flipping a coin 9 times (with a specific outcome [e.g., heads] probability equaling 0.50, or 50%), we see the following graph (i.e., bar plot):



```
barplot(table(rbinom(100000, 9, .5)))
```

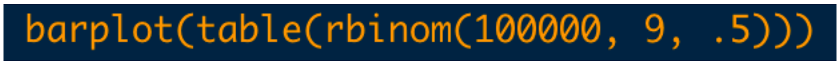


The most likely outcomes of this equation equate to observing a 'heads' 4 or 5 times (with a slight edge given to 4, based on the data table above, since its value is higher than that of 5). As we stray further from the middle of the graph, the outcome is less likely to happen (since these values are lower than that of 4 and 5).

The 'rbinom' function requires three parameters for input. First, it requests the number of trials that will be taking place; it then asks for how many observations will occur for each trial. Lastly, the probability of a specific outcome is assigned (which was set to 0.5, or 50%, to account for only two outcomes for the coinflip). The 'table' function is then applied to the result, which formats the data in a more 'user-friendly' manner, so to speak.

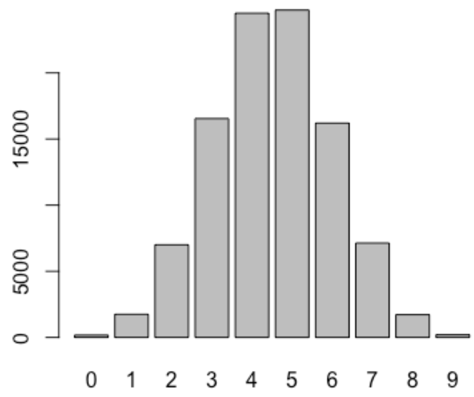
Homework Question 2

As mentioned, and seen above, the 'barplot' for the 100,000 trails looks like the below (I added it again for convenience of the reader).

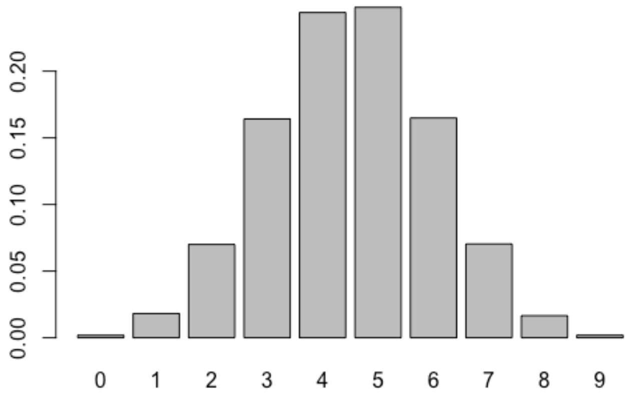


```
barplot(table(rbinom(100000, 9, .5)))
```

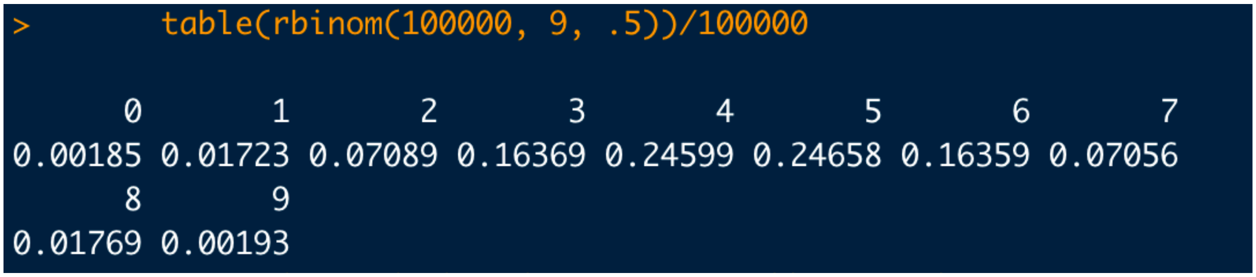
The 'barplot' function is the appropriate R command that allows us to look at the data, visually speaking, in a bar plot type format.



If we'd like to convert these values to probabilities, we can divide each 'cell,' or value, in the table by the total number of trials seen in the table. In this instance, we'll be dividing each value in the table by a value of 100,000. Once that is complete, we get the following result (the total of these probabilities will equal 1, or 100%):



Only differences in the data are attributed to randomness – since R generated new random numbers in the 'rbinom' function. Like the first graph, the shape is taking on a binomial distribution shape – similar to that of a normal distribution in which a bell-shaped curved is observed. The most likely outcomes appear to be 4 or 5, since the 'bars' are the highest for these values. Unlike the first graph that illustrated the frequency of each outcome, this table indicates the percentages of how likely each outcome will transpire. In this case, we see 4 ‘heads’ have a probability of ~24.5% while a 5 ‘head’ outcome is ~24.6%. Other outcomes are less likely, since their probabilities/percentages are lower.

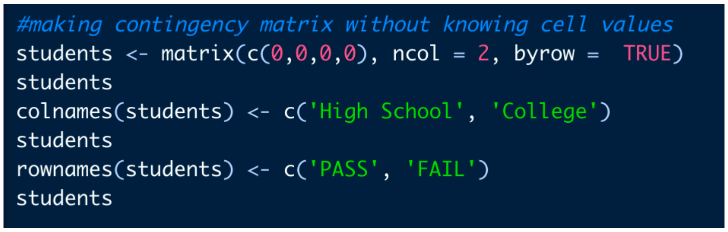


```
>       table(rbinom(100000, 9, .5))/100000

      0       1       2       3       4       5       6       7 
0.00185 0.01723 0.07089 0.16369 0.24599 0.24658 0.16359 0.07056 
      8       9 
0.01769 0.00193 
```

Since our ‘rbinom’ function contained a probability value of 0.5, it makes sense that we see that the middle values of 4 or 5 are the most common in our data tables.

Homework Question 6

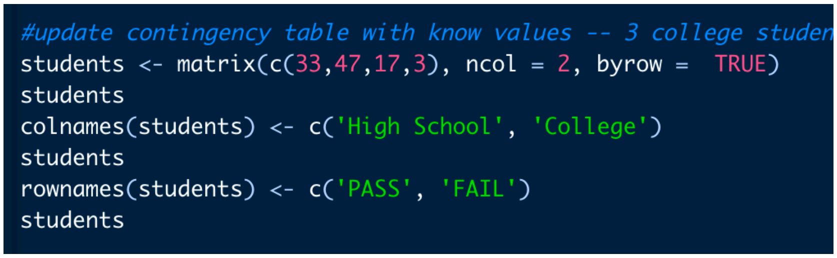
Before knowing all cell values, I placed the value of 0 as placeholders until more information was known in my contingency table:



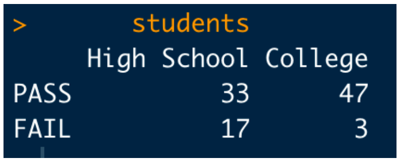
```
#making contingency matrix without knowing cell values
students <- matrix(c(0,0,0,0), ncol = 2, byrow =  TRUE)
students
colnames(students) <- c('High School', 'College')
students
rownames(students) <- c('PASS', 'FAIL')
students
```

After learning that only three college students failed the test, I was able to update the values in my table. Since 20 students in total failed, and only 3 were college students, I was able to infer that the remaining 17 students that failed were categorized as high school students.

Since only 3 college students failed, and there are 50 college students in total, I can also infer that 47 college students passed the exam. Lastly, since it was determined that 17 college students failed, we can subtract this number from the total number of high school students, 50. We calculate a value of 33. Based on the above, the contingency table now looks like this:

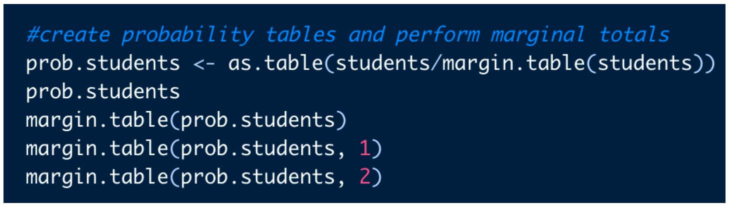


```
#update contingency table with know values -- 3 college studen
students <- matrix(c(33,47,17,3), ncol = 2, byrow =  TRUE)
students
colnames(students) <- c('High School', 'College')
students
rownames(students) <- c('PASS', 'FAIL')
students
```

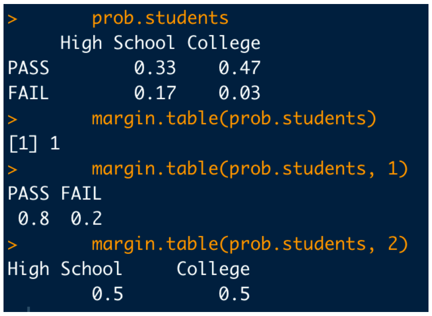


```
>        students
      High School College
PASS           33      47
FAIL           17       3
```

From here, we'd like to convert what we currently have to a probability table. This can be done by dividing each cell value by the total. From here, we can calculate the marginal totals by row and column:

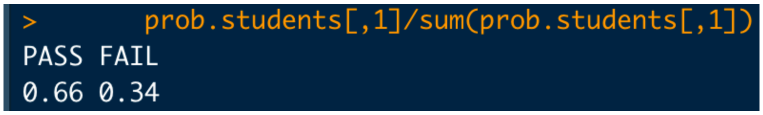


```
#create probability tables and perform marginal totals
prob.students <- as.table(students/margin.table(students))
prob.students
margin.table(prob.students)
margin.table(prob.students, 1)
margin.table(prob.students, 2)
```



```
>       prob.students
     High School College
PASS        0.33    0.47
FAIL        0.17    0.03
>       margin.table(prob.students)
[1] 1
>       margin.table(prob.students, 1)
PASS FAIL
 0.8  0.2
>       margin.table(prob.students, 2)
High School     College
        0.5         0.5
```

Lastly, if we were to focus only on high school students, we'd have to 'normalize' their respective probability values by dividing each by the total marginal value they produce. 0.33 + 0.17 = 0.5, so we would divide 0.33 and 0.17 by 0.5. Once this occurs, the following results are observed:



```
>       prob.students[,1]/sum(prob.students[,1])
PASS FAIL
0.66 0.34
```

0.66, or 66%, is the passing rate for the students.

<u>Homework Question 7</u>

Based on the information above, the following contingency table could be visualized before loading anything into R:

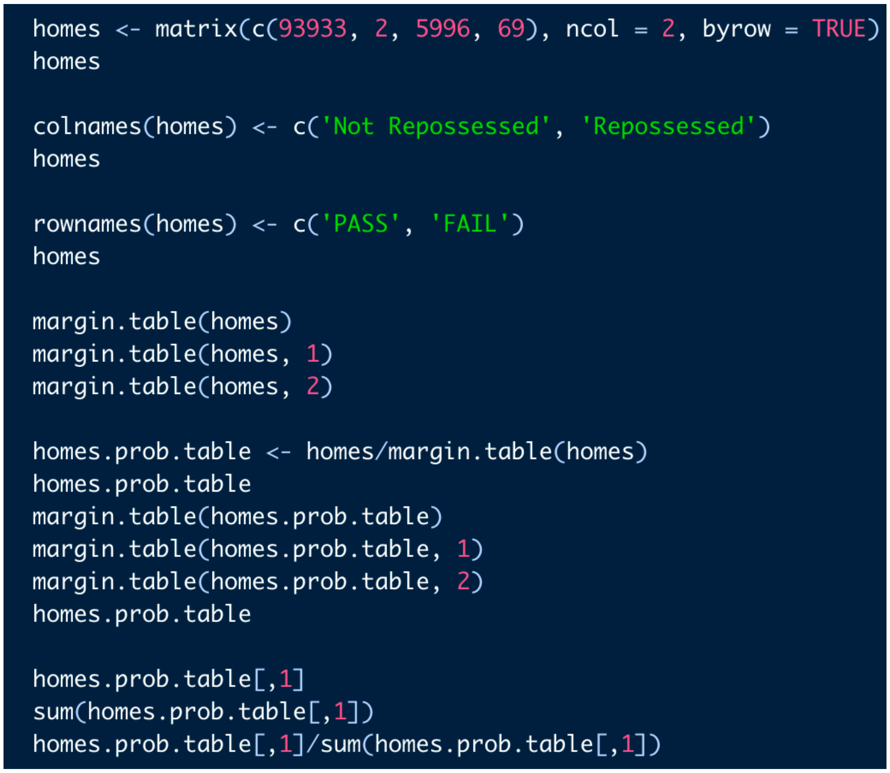
Cont. Table	Not Repossessed	Repossessed	Total
Pass	93,933	2	93,935
Fail	5,996	69	6,065

Total.	99,929	71	100,000

71 out of 100,000 homes are repossessed by the bank, meaning that the number of homes not repossessed is 99,929. We also learn that 93,935 homes pass the test while 6,065 do not. So far, all we've learned are marginal totals for our table. When we learn that 5,996 homes failed the test but did not get their homes repossessed, we can infer a few things.

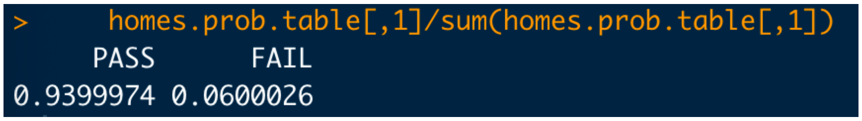
- 69 homes are repossessed that also fail the test (6,065 – 5,996)
- 2 homes that pass the test are actually repossessed (71 – 69)
- 93,933 homes pass the test and are not repossessed (99,929 – 5,996)

From here, we can create our contingency table in R



```
homes <- matrix(c(93933, 2, 5996, 69), ncol = 2, byrow = TRUE)
homes

colnames(homes) <- c('Not Repossessed', 'Repossessed')
homes

rownames(homes) <- c('PASS', 'FAIL')
homes

margin.table(homes)
margin.table(homes, 1)
margin.table(homes, 2)

homes.prob.table <- homes/margin.table(homes)
homes.prob.table
margin.table(homes.prob.table)
margin.table(homes.prob.table, 1)
margin.table(homes.prob.table, 2)
homes.prob.table

homes.prob.table[,1]
sum(homes.prob.table[,1])
homes.prob.table[,1]/sum(homes.prob.table[,1])
```

The final result yields the following:



```
>     homes.prob.table[,1]/sum(homes.prob.table[,1])
     PASS      FAIL
0.9399974 0.0600026
```

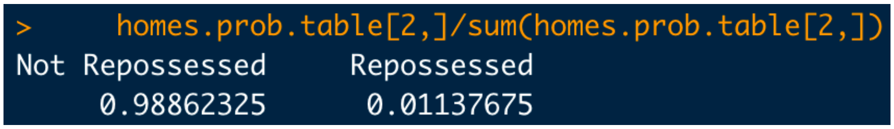
Almost 94% (.9399974) of homes passed the test and did not get repossessed.

Homework Question 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 actu- ally default on his or her mortgage? Show your work and especially show the tables that you set up to help with your reasoning. 9. The incidence of HIV in the U.S. population

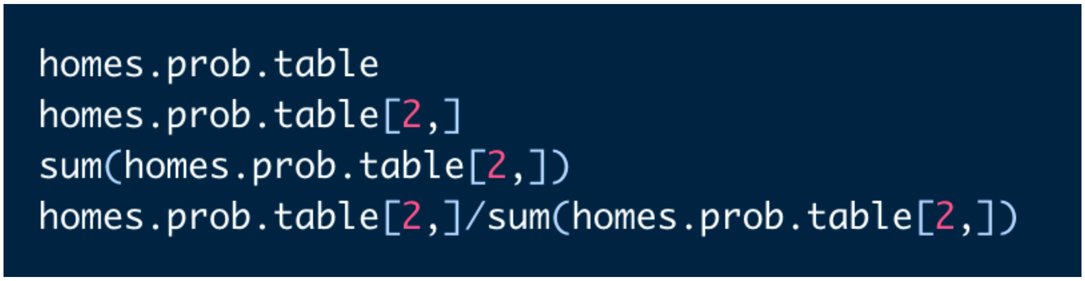
Stanton, Jeffrey M.. Reasoning with Data (p. 36). Guilford Publications. Kindle Edition.

For this question, we will choose to 'normalize' the 'FAIL' row (instead of the 'Not Repossessed' column, which is what we did for question #7). When the total probability for the 'FAIL' row is summed, a result of .06065 is observed. The remaining task now is to divide each 'FAIL' row value by this summed result. We get the following 'normalized' probabilities:



```
>     homes.prob.table[2,]/sum(homes.prob.table[2,])
Not Repossessed     Repossessed
     0.98862325      0.01137675
```

To answer the question, since this new customer failed the test, the probability of them having their home repossessed is just over 1%, or 0.1137675. The following code was used to determine this:



```
homes.prob.table
homes.prob.table[2,]
sum(homes.prob.table[2,])
homes.prob.table[2,]/sum(homes.prob.table[2,])
```