IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: September 1, 2020
Chapter Number: # 2
Title of Chapter: Reasoning with Probability

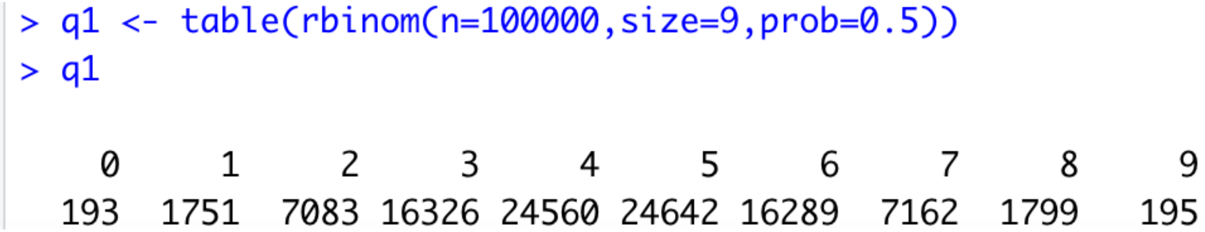
Probability and Tables

1. **Outcome Tables:** lists the various possible outcomes of a set of similar or related events.
    a. Example: Coin Tossing – whose outcome is **binomial** (only have two event).
    b. R function: using rbinom(n, size, prob) to build the binomial distribution
2. **Visualization functions in R**: barplot() and hist()
3. **Contingency Table:** is used in statistics to summarize the relationship between several categorical variables.
    a. R function: using table() function to summaries the frequency of events.
    b. One can add a **marginal row** to enhance the usefulness of contingency table.
4. **Normalization and prior**: divide each original data in the contingency table by the proportion that it takes in the whole population (scale the proportion from the original number to 100%).

Make Your Own Tables with R

1. Create a matrix by using **matrix()** function: toast < -matrix(c(2,1,3,4), ncol=2, byrow=TRUE)
2. Adding names for column and row: **colnames() and rownames()**
3. Using **as.table()** to convert data into table format: toast <- as.table(toast)
4. Using **margin.table()** to add margins for table: margin.table(toast, 1/2) specifically, here **1 represents row while 2 represents columns**
5. One can use the margin.table() function to calculate the probability: toastProbs <- toast/margin.table(toast)

Exercise Review

1. 

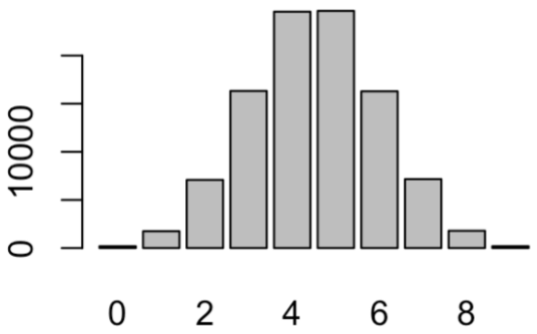
```
> q1 <- table(rbinom(n=100000,size=9,prob=0.5))
> q1


    0     1     2     3     4     5     6     7     8     9 
  193  1751  7083 16326 24560 24642 16289  7162  1799   195 
```

The question requires us to flip a coin 9 times in one experiment. Then, we need to repeat this experiment 100000 times. By using rbinom() and table() function, we can easily obtain the result. Normally, we will have two events (0 represents tails above while 1 represent head above). In this experiment, we will mark 1 for having a head above once we observed. Now we have 10 observations. Each observation represents a event for the experiment, Specifically, 0 in this table represents that there were 193 times that none of the coin had head above. There were 195 times that each time we flipped a coin, the coin had head above.

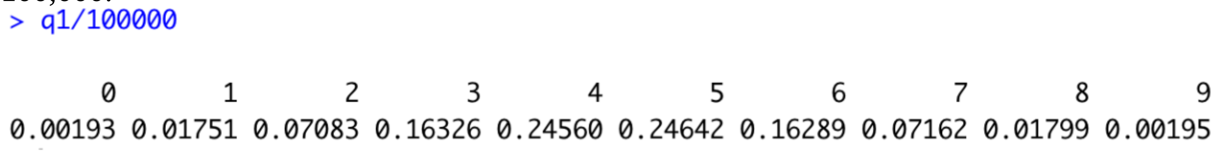
2. This bar plot shows the frequency of experiment we conducted in question 1. Here we can see that this plot is normally distributed since statistics has a theorem that if x is a random variable with distribution B(n,p), the following random variable has a standard normal distribution for sufficiently large n. In this case, 100,000 is large enough so that the plot has all the characteristics that normal distribution possesses. The reason why the events which has 4- or 5-times head above have the highest frequency is that, statistically speaking, extreme condition (such as event that has 0-time head above) is more unlikely to happen in real life.

barplot(q1, main=NULL)



Convert the result to probability:

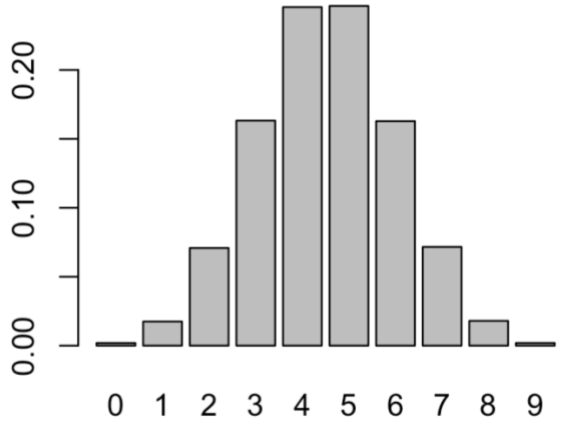
There is a slightly difference between the plot above and the plot below. The difference is the scale. The plot below shows the proportion of each event that occur in the experiment. Namely, plot below could be obtained by divided the plot above's y-axis by 100,000.

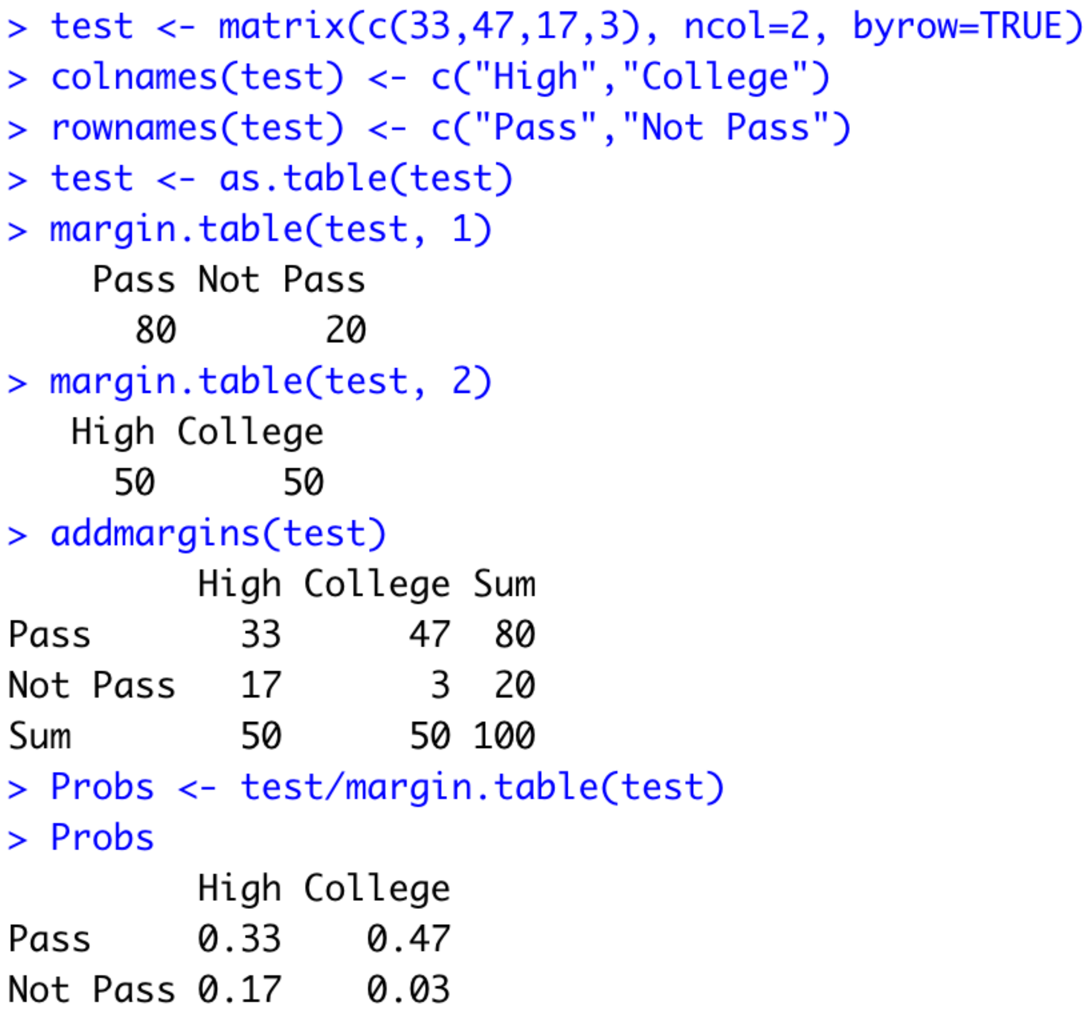


```
> q1/100000

      0       1       2       3       4       5       6       7       8       9 
0.00193 0.01751 0.07083 0.16326 0.24560 0.24642 0.16289 0.07162 0.01799 0.00195 
```

barplot(q1/100000, main=NULL)

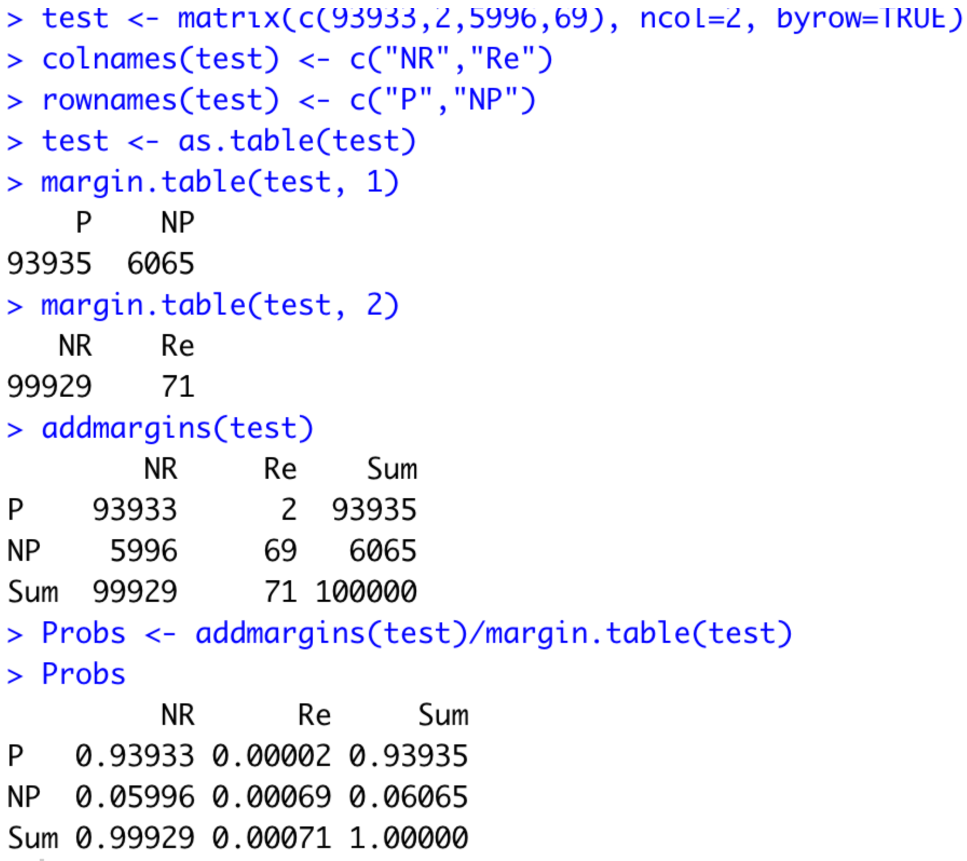




```
> test <- matrix(c(33,47,17,3), ncol=2, byrow=TRUE)
> colnames(test) <- c("High","College")
> rownames(test) <- c("Pass","Not Pass")
> test <- as.table(test)
> margin.table(test, 1)
    Pass Not Pass 
      80       20 
> margin.table(test, 2)
   High College 
     50      50 
> addmargins(test)
          High College Sum
Pass        33      47  80
Not Pass    17       3  20
Sum         50      50 100
> Probs <- test/margin.table(test)
> Probs
          High College
Pass      0.33    0.47
Not Pass 0.17    0.03
```

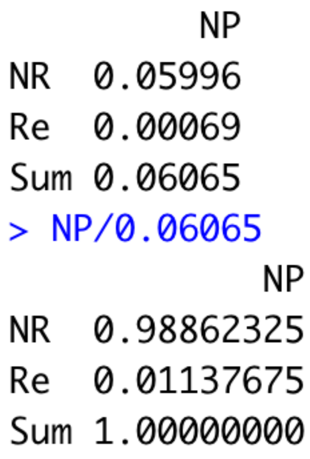
6. Since we have already known the total number of students, the number of high school student and that of college student (which is 50 respectively), the number of student who passed the test and that of student who did not pass the exam (which is 80 and 20 respectively), all we need is one condition of the four cells in the contingency table.

The pass rate for high school students is 0.33. However, if we only focus on high school student, the passing rate is 0.33/0.5=0.66.



```
> test <- matrix(c(93933,2,5996,69), ncol=2, byrow=TRUE)
> colnames(test) <- c("NR","Re")
> rownames(test) <- c("P","NP")
> test <- as.table(test)
> margin.table(test, 1)
    P    NP 
93935  6065 
> margin.table(test, 2)
   NR    Re 
99929    71 
> addmargins(test)
         NR     Re    Sum
P     93933      2  93935
NP     5996     69   6065
Sum   99929     71 100000
> Probs <- addmargins(test)/margin.table(test)
> Probs
         NR      Re     Sum
P   0.93933 0.00002 0.93935
NP  0.05996 0.00069 0.06065
Sum 0.99929 0.00071 1.00000
```

7. The percentage of customers both pass the test and do not have their homes repossessed is 0.93933.



```
           NP
NR  0.05996
Re  0.00069
Sum 0.06065
> NP/0.06065
             NP
NR  0.98862325
Re  0.01137675
Sum 1.00000000
```

8. From exercise 7, if the new customer is doomed to fail the test, there is a 0.01137675 chance he or she default his or her mortgage. Given that we had known that 0.06065 of the population will not pass the test. Among those who fail the test, 0.00069/0.06065 of them will default the mortgage.

R Code Fragment and Explanation

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

This week's most important R code is the rbinorm() function, which can be used to generate required number of random values of given probability from a given sample.

rbinom(n, size, prob)

- **n** is number of observations.
- **size** is the number of trials.
- **prob** is the probability of success of each trial.

Question for Class

1. What is the normalization of more complicated scenarios, such as a 3X3 contingency table?