MTH304 R Project

Suhaib Khan

2022-11-28

## Simulation of a Binomial Distribution

**Part A:**

A binomial distribution is defined by the following equation:

With that being said, The head shows up 12.5% of the time, meaning that p = 0.125. n would be the number of trials, in this case it would be 11. Therefore, the following pmf of this distribution would be:

**Part B:**

When using the pmf above, we can calculate these probabilities (using a calculator):

P(X=0) = 0.230

P(X=1) = 0.362

P(X=2) = 0.258

P(X=3) = 0.111

P(X=4) = 0.0316

P(X=5) = 0.00636

P(X=6) = 0.000904

P(X=7) = 0.0000922

P(X=8) = 0.00000659

P(X=9) = 0.000000313

P(X=10) = 0.00

P(X=11) = 0.00

**Part C:**

We have to create a code which will show the probability distribution of X to three decimal places.

#Function to calculate the random outcomes given the parameters  
dice\_dist <-rbinom(n=10000000,size=11,prob=0.125)  
  
#Function to calculate the probabilities of the dice roll  
dice\_probability = table(dice\_dist)\*(1/10000000)  
  
#List of probabilities rounded to 3. (NOTE: n=100000 was used because it displays the probabilities up to 9. 10 and 11 are not shown probably because they are a very small value).  
round(dice\_probability,3)

## dice\_dist  
## 0 1 2 3 4 5 6 7 8 9   
## 0.230 0.362 0.258 0.111 0.032 0.006 0.001 0.000 0.000 0.000

**Part D:**

We will now create a table of values representing the values of x and their probabilities

#Store x values and simulated probabilities in vectors  
x\_val <-c(0,1,2,3,4,5,6,7,8,9,10,11)  
x\_prob <-c(0.230,0.362,0.258,0.111,0.032,0.006,0.001,0.000,0.000,0.000,0.000,0.000)  
  
#Use table function to create the table of values  
table(x\_prob,x\_val)

## x\_val  
## x\_prob 0 1 2 3 4 5 6 7 8 9 10 11  
## 0 0 0 0 0 0 0 0 1 1 1 1 1  
## 0.001 0 0 0 0 0 0 1 0 0 0 0 0  
## 0.006 0 0 0 0 0 1 0 0 0 0 0 0  
## 0.032 0 0 0 0 1 0 0 0 0 0 0 0  
## 0.111 0 0 0 1 0 0 0 0 0 0 0 0  
## 0.23 1 0 0 0 0 0 0 0 0 0 0 0  
## 0.258 0 0 1 0 0 0 0 0 0 0 0 0  
## 0.362 0 1 0 0 0 0 0 0 0 0 0 0

#Note: Table shows a 1 when an X value corresponds to the probability and 0 otherwise.

Part E:

Now, We can compute the expectation of the function.

Expectation = (0\*0.230)+(1\*0.362)+(2\*0.258)+(3\*0.111)+(4\*0.032)+(5\*0.006)+(6\*0.001)  
Expectation

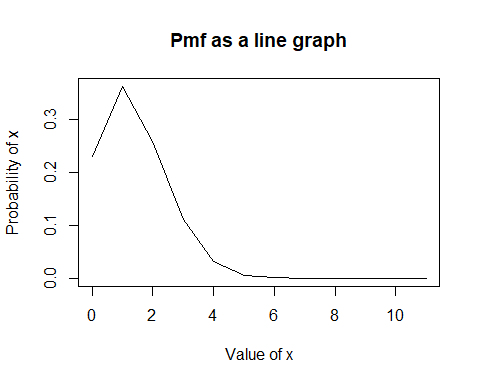
## [1] 1.375

With this calculation we can now see that the expectation of this function is 1.375.

**Part F:**

We will now plot a line graph of the simulated data.

#The line graph is created using the plot function. x\_val is in the x-axis because if it were to be in the y-axis the graph would not have been a function (vertical line test).  
pmf\_line<-plot(x\_val,x\_prob,xlab='Value of x',ylab='Probability of x',type='l',main='Pmf as a line graph')

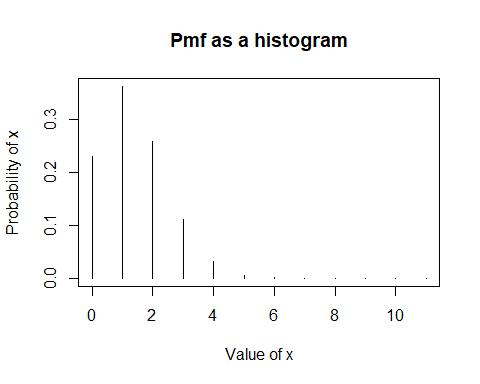


Above, we can see the graph of the pmf of f(x). We can see that the probability gradually goes down, except for the instance of x=1 where there is a sharp spike up.

**Part G:**

Now, we will graph the pmf as a histogram.

#This code is the same as the one in part F. Only difference is the type is changed to 'h' meaning it will be a histogram  
pmf\_hist<-plot(x\_val,x\_prob,xlab='Value of x',ylab='Probability of x',type='h',main='Pmf as a histogram')



We can see the same results in part F except in histogram form.

## Simulation of a Hypergeometric Distribution

**Part A:**

The denominator of this function represents the combinations you can get from a deck of cards when you pick 5. There are 52 cards in total and the problem says that only 5 are chosen.

We can say that:

A+B = 52

n = 5

Now, A would be 12 since there are 12 face cards in a deck of cards. B would be 40 since 52-12 = 40 which represents the rest of the cards available in the deck.

**Part B:**

When using the pmf above, we can calculate these probabilities (using a calculator):

P(x=0) = 0.253

P(x=1) = 0.422

P(x=2) = 0.251

P(x=3) = 0.0660

P(x=4) = 0.00762

P(x=5) = 0.000305

**Part C:**

#Place values of x into a vector  
x<-c(0,1,2,3,4,5)  
  
#Create a hypergemtric distribution function using the choose() function  
card\_dist <-(choose(12,x))\*(choose(40,5-x))\*((1/choose(52,5)))  
  
#Round the probabilities to 3 decimal places  
rounded\_card\_dist<-round(card\_dist,3)  
  
rounded\_card\_dist

## [1] 0.253 0.422 0.251 0.066 0.008 0.000

**Part D:** We will now create a table of values with the values of x corresponding to their probabilities

#Store x values and probabilities into vectors  
c\_val <-c(0,1,2,3,4,5)  
c\_prob <-c(0.253,0.422,0.251,0.066,0.008,0.000)  
  
#Create table using table() function  
table(c\_prob,c\_val)

## c\_val  
## c\_prob 0 1 2 3 4 5  
## 0 0 0 0 0 0 1  
## 0.008 0 0 0 0 1 0  
## 0.066 0 0 0 1 0 0  
## 0.251 0 0 1 0 0 0  
## 0.253 1 0 0 0 0 0  
## 0.422 0 1 0 0 0 0

#Note: Table shows a 1 when an X value corresponds to the probability and 0 otherwise.

**Part E:**

Now we will compute the expectation of the function:

Expectation\_card = (0\*0.253)+(1\*0.422)+(2\*0.251)+(3\*0.066)+(4\*0.008)+(5\*0.000)  
Expectation\_card

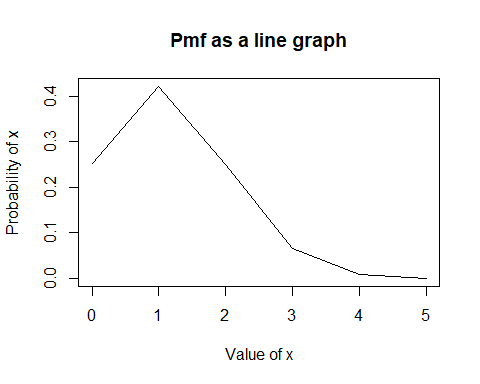
## [1] 1.154

As we can see from this calculation, the expectation of this function is 1.154

**Part F:**

Lets now plot a line graph of this function:

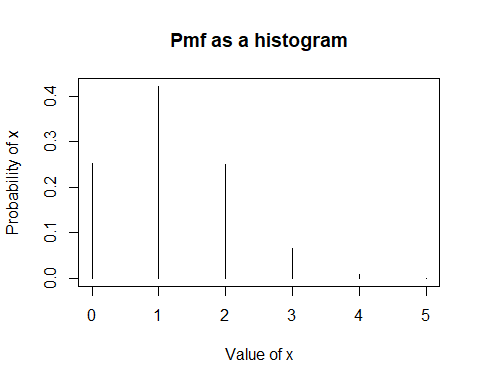
#The plot function was used to compute the line graph for this pmf  
pmf\_c\_line<-plot(c\_val,c\_prob,xlab='Value of x',ylab='Probability of x',type='l',main='Pmf as a line graph')



**Part G:**

Now, we will display the probability histogram for this pmf

#The plot function was used to compute the line graph for this pmf  
pmf\_c\_hist<-plot(c\_val,c\_prob,xlab='Value of x',ylab='Probability of x',type='h',main='Pmf as a histogram')



From these graphs, we can see that the probability decreases as x increases, except for when x=1 where there seems to be a spike.

## Dealing with a Small Dataset

**Part A:**

Exam Results are an ordinal level of measurement. This is because each grade is ranked from lowest to highest(0%-100%).

Study hours are a scale level of measurement. This is because it can be measured with an instrument(watch) and is numerical.

**Part B:**

Now, let’s store this data as vectors

hours\_studied<-c(5,4,8,7,10,6,10,4,0,0)  
  
#Use summary function to see descriptive statistics and sd function to view the standard deviation  
summary(hours\_studied)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 4.00 5.50 5.40 7.75 10.00

sd(hours\_studied)

## [1] 3.565265

We can see that the average hours studied was 5.4 hours. The lowest hours studied was 0 while the most was 10.

exam\_grade<-c(73,64,80,70,85,50,86,50,20,25)  
  
summary(exam\_grade)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 50.00 67.00 60.30 78.25 86.00

sd(exam\_grade)

## [1] 23.59402

We can see that the average grade was 60.30. The highest grade was 86 while the lowest was 20.

**Part C:**

Now, we have to find the correlation coefficient of this data set. This value ranges from -1 to 1. where if it is negative there is a negative correlation and vice versa. When the absolute value of the coefficient is closer to 1, it means there is a stronger correlation.

#cor function allows us to view the correlation coefficient without us doing a lot of math  
cor(exam\_grade,hours\_studied)

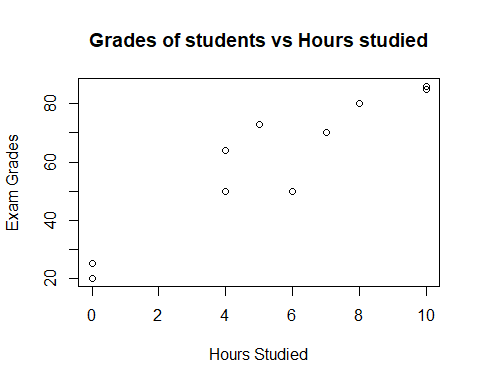
## [1] 0.9309571

As we can see, the correlation coefficient is 0.931 which means there is a strong positive correlation between the exam grades and hours studied. Meaning the more a student studied, the better they performed.

**Part D:**

Now, we have to create a scatter plot plotting the Exam results vs study hours.

plot(x=hours\_studied, y=exam\_grade, main='Grades of students vs Hours studied', xlab='Hours Studied', ylab='Exam Grades')



We can see that the more hours a student put into studying, the higher results they got. Both of the student who studied 0 hours have failed the exam.

## The End