Probability and Statistics with R Semester I 2021/22 Worksheet 6

https://www.isibang.ac.in/~athreya/Teaching/PaSwR/ Solutions

Prashant Sharma

```
> #import the given csv data set in to R
> data=read.csv(file="C://Users//shiva//Downloads//2data-dicetoss.csv")
> dicetoss=data.frame(data)
> dim(dicetoss)
```

```
[1] 104   8
```

Now, we will perform the data cleaning process for that start with omitting the NA values from data if any.

```
> dicetoss=na.omit(dicetoss)
> #Now, removing the errors where the number of heads is more than
> #the outcome of roll as the die is tossed exactly that number of times
> Hm=c()
> for (i in 1:dim.data.frame(dicetoss)[1]){
+ if(dicetoss$Outcome.of.Roll[i] < dicetoss$Y..Number.of.Heads[i])
+ {Hm = append(Hm,i)
+ }
+   }
> Hm #Here Hm gives us the indices of faulty rows which have more number of heads than tosses
```

```
NULL
```

Now, we will check the coin has been tossed exactly the number of times that is the outcome of roll and filter those rows out

```
> toss_check=c()
> for (i in 1:dim.data.frame(dicetoss)[1]){
+ if(dicetoss$Outcome.of.Roll[i] != 6 - length(which(dicetoss[i,] == "")))
+ {toss_check = append(toss_check,i)
+ }
+
+ }
> # here, toss_check gives the indices of faulty rows which have more entries in toss columns
> #than the number of tosses
> toss_check
```

```
[1]  28  57  58 103
```

Now, we will examine the faulty rows of the data set and try if they can rectified.

```
> dicetoss[toss_check,]
```

```
    Outcome.of.Roll Outcome.of.Toss Outcome.of.Toss.1 Outcome.of.Toss.2
28                2               H
57                4               T                 T                 T
58                1               H
104               6               H                 T                 T
    Outcome.of.Toss.3 Outcome.of.Toss.4 Outcome.of.Toss.5 Y..Number.of.Heads
28                                                                         1
57                  T                 0                                    0
58                                                      1                  1
104                 T                                                      1
```

As we can easily see that there are integers in the columns of toss outcomes for rows 57 and 58, which can be considered as errors. But, we have the required number of tosses and their outcomes in that rows, So we can remove those integer values

```
> dicetoss[57,6]=""
> dicetoss[58,6]=""
```

Since for the 28th and 104th rows there is not enough toss outcomes as the outcome of roll, therefore we remove these rows

```
> #Here, we use command dicetoss[-c(28,103),] becuase after removing 28th row
> #there will be 103 rows remaining
> new_dicetoss=dicetoss[-c(28,103),]
```

Problem: 1(a)

Now, for a given i, j, $(1 \leq i \leq 6; 0 \leq j \leq 6)$**;**

$$P(Y = j|X = i) = \frac{P(Y = j, X = i)}{P(X = i)}$$

```
> M1=matrix(data=NA,nrow=6,ncol=7)
> for (i in 1:6){
+ for (j in 1:7){
+ freq=0
+ for (k in 1:length(new_dicetoss[,1])){
+ if ((new_dicetoss[k,1]==i) & (new_dicetoss[k,8]==j-1)){
+ freq=freq+1}}
+ prob=freq/sum(new_dicetoss[,1]==i)
+ M1[i,j]=prob}}
> #Now we will save the matrix of the conditional distribution in a dataframe
> df1=data.frame(M1)
> colnames(df1)=c("P(Y=0|X)","P(Y=1|X)","P(Y=2|X)","P(Y=3|X)","P(Y=4|X)",
+  "P(Y=5|X)","P(Y=6|X)")
> rownames(df1)=c("X=1","X=2","X=3","X=4","X=5","X=6")
> df1
```

```
      P(Y=0|X)  P(Y=1|X)  P(Y=2|X)   P(Y=3|X)  P(Y=4|X)   P(Y=5|X)  P(Y=6|X)
X=1 0.28571429 0.7142857 0.0000000 0.00000000 0.0000000 0.00000000 0.0000000
X=2 0.12500000 0.6875000 0.1875000 0.00000000 0.0000000 0.00000000 0.0000000
X=3 0.07142857 0.3571429 0.3571429 0.21428571 0.0000000 0.00000000 0.0000000
X=4 0.23809524 0.3333333 0.3333333 0.09523810 0.0000000 0.00000000 0.0000000
X=5 0.00000000 0.1176471 0.2941176 0.41176471 0.1176471 0.05882353 0.0000000
X=6 0.00000000 0.0000000 0.4210526 0.05263158 0.2631579 0.15789474 0.1052632
```

```
> #Now, save the dataframe in a csv file
> write.csv(df1,"YgivenX.csv")
```

Problem: 1(b)
Now, calculating the conditional distribution of X given Y, for Y=1,2,3,4,5,6

```
> M2=matrix(data=NA,nrow=6,ncol=6)
> for (i in 1:6){
+ for (j in 1:6){
+ freq=0
+ for (k in 1:length(new_dicetoss[,8])){
+ if ((new_dicetoss[k,8]==i) & (new_dicetoss[k,1]==j)){
+ freq=freq+1}}
+ prob=freq/sum(new_dicetoss[,8]==i)
+ M2[i,j]=prob}}
> #Now we will save the matrix of the conditional distribution in a dataframe
> df2=data.frame(M2)
> colnames(df2)=c("P(X=1|Y)","P(X=2|Y)","P(X=3|Y)","P(X=4|Y)","P(X=5|Y)",
+ "P(X=6|Y)")
> rownames(df2)=c("Y=1","Y=2","Y=3","Y=4","Y=5","Y=6")
> df2
```

```
     P(X=1|Y)  P(X=2|Y)  P(X=3|Y)  P(X=4|Y)   P(X=5|Y)   P(X=6|Y)
Y=1 0.2857143 0.3142857 0.1428571 0.2000000 0.05714286 0.00000000
Y=2 0.0000000 0.1071429 0.1785714 0.2500000 0.17857143 0.28571429
Y=3 0.0000000 0.0000000 0.2307692 0.1538462 0.53846154 0.07692308
Y=4 0.0000000 0.0000000 0.0000000 0.0000000 0.28571429 0.71428571
Y=5 0.0000000 0.0000000 0.0000000 0.0000000 0.25000000 0.75000000
Y=6 0.0000000 0.0000000 0.0000000 0.0000000 0.00000000 1.00000000
```

```
> #Now, save the dataframe in a csv file
> write.csv(df2,"XgivenY.csv")
```