IST772

Parin Patel
Homework 2
10/13/2019

Attribution Statement:

1. Homework 2 by Parin Patel: I did this homework by myself, with help from the book and the professor. In addition I used the following websites to help with the r code for the bar plot:
    a. Bar Plot: https://www.statmethods.net/graphs/bar.html

Exercises:

1. **Flip a fair coin nine times and write down the number of heads obtained. Now repeat this process 100,000 times. 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. Write down the results and explain in your own words what they mean.**

When I first flipped the coin 9 times, I got 5 heads and 4 tails.Since the probability of flippoing coins is 50%, this outcome is likely. When creating the R code for 100,000 trials, first the parameter that tells the function to run 100 trials was set to n=100000. Next, the size=9 parameter was set to tell the function we want 9 coin flipping events in each trial. Finally, the probability of coin flipping, as mentioned above is 50%. Therefore, prob=0.5 was set since we expect the coin to land as heads-side up is equaly likely to land as tails-side up in any single event.

The results of our R code shown below first reiterates that the chance of having 4-5 heads is around 50%, or around 50,000 trials, out of 10 possible scenerios. The results also show that as you move away from the center of 4-5 head,s the probability decreases. So you are least likely to get 0 or 9 heads. However, you are more likely to get 3 or 6 heads than you are to get 0 or 9 heads. But the chance of getting 3 or 6 heads is still less likely than gettimg 4-5 heads.

```
> dt1<-table(rbinom(n=100000,size = 9,prob=0.5))
> dt1

    0     1     2     3     4     5     6     7     8     9 
  210  1668  7025 16530 24769 24430 16497  7013  1692   166 
```

IST772

Parin Patel
Homework 2
10/13/2019

2. Using the output from Exercise 1, summarize the results of your 100,000 trials of nine 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. Write a brief interpretive analysis that describes what each of these bar plots signifies and how the two bar plots are related. Make sure to comment 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.

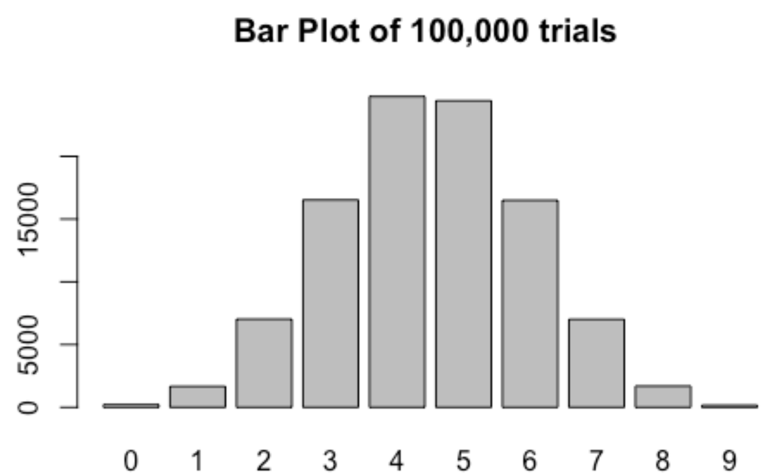

Figure 1 Bar plot of 100,000 Trials

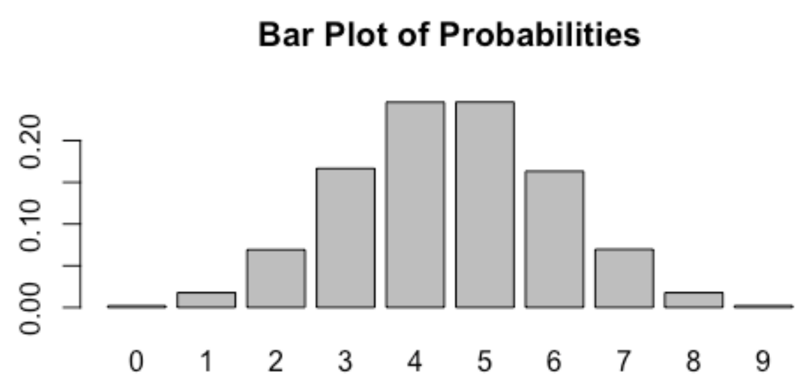

Figure 2 Bar Plot of Probabilities of the 100,000 trials

IST772

Parin Patel
Homework 2
10/13/2019

Both bar plots have a normal distribution where the center of the plots have the highest expected number of heads-side-up trials and the highest probabilities of achieving heads-side-up. Therefore, you can say that with some likelihood, you can anticipate the results of 9 events to have around 4-5 heads. The probability graph shows these two trials to have the highest probability, around 0.2. As you move away from the center, your probability of achieving those combination of heads-side-up events becomes more improbable. Therefore, you are extremely unlikely to have a 0 or 9 heads-side-up result.

```
#2
barplot(dt1, main = "Bar Plot of 100,000 trials ") #barplot of n=100,000
ProbTable_bp<-barplot(table(rbinom(n=100000,size = 9,
      prob=0.5))/100000, main = "Bar Plot of Probabilities") #barplot of probabilities
```

Figure 3 R code for barplots

6. One hundred students took a statistics test. Fifty of them are high school students and 50 are college students. Eighty students passed and 20 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). Draw that table and write in the marginal totals. I'm now going to give you one additional piece of information that will fill in one of the four blank cells: only three 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. 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. Finally, create a second copy of the complete table, replacing the counts of students with probabilities. 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?

```
> StatsExam
     College High School
Pass      47          33
Fail       3          17
```

The contingency table above shows the marginal total of students in college and high school that pass/failed the statistics test. In order to figure out all four of the table's main cells, we needed to first know the sample size (n=100 students). Then we were told that 50 students were in college and 50 were in high school, and that 80 had passed and 20 had failed. Since we were told that 3 of the failed students were in college, we could deduce that then that meant 17 of the high school students had failed the exam (20-3 = 17). Additionally, since 3 of the college students had failed,

IST772

Parin Patel
Homework 2
10/13/2019
this meant that 47 of them had passed (50-3 = 47). And finally, since we calculated that 17 high school students had failed ,that left 33 passing the statistics exam (50-17 = 33).

```
> ProbTable_statsExam
     College High School
Pass    0.47        0.33
Fail    0.03        0.17
```

The above table shows the second copy of the table where the counts of students are replaced with probabilties. The expected pass rate for high school students was found to be 66%, see the below table, after we normalized the dataset.Therefore, the 34% of high school students are likely going ot fail the statistics exam.

```
> ProbTable_statsExamFinal <- StatsExam/margin.table(StatsExam)
> ProbTable_statsExam[,2]/sum(ProbTable_statsExam[,2])
Pass Fail
0.66 0.34
```

7. In a typical year, 71 out of 100,000 homes in the United Kingdom is 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: 93,935 households pass the test and 6,065 households fail the test. Interestingly, 5,996 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. Hint: The 5,996 is the only number that goes in a cell; the other numbers are marginal totals. What percentage of customers both pass the test and do not have their homes repossessed?

```
> BarclayTest
     Repossessed Not Repossessed
Pass           0           93935
Fail          69            5996
```

The above contingency table shows the number of mortgages that passed the Barclay Banks screening test, and those that failed. Given that 6,065 households failed the test, but 5,996 that failed were not repossessed, the contingency table shows 69 (6065-5996 = 69) as the number of houses that were repossessed. Based on this we calculated that the percentage of customers that both passed the test and did not have their homes reposed. Assuming that the those who passed did not default on their loans, and therefore did not get their homes repossessed, the percentage is 1. See the table below for the outcome and code. However, logically, this result makes sense since those who pass the test are not likely to have their home repossessed.

IST772

Parin Patel
Homework 2
10/13/2019

```
BarclayTestProb <- BarclayTest/margin.table(BarclayTest)
BarclayTestProb[1,]/sum(BarclayTestProb[1,])
  Repossessed Not Repossessed
            0               1
```

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 actually default on his or her mortgage? Show your work and especially show the tables that you set up to help with your reasoning.

First, we reviewed the contingency table created in question 7 (copied code results shown below). It is important to remember that most of the people that failed the Barclay's screening test did not default on their mortgage, only 69 of 6065 people did. Therefore, if this new customer fails the test, the chance they would default on his or her mortgage, and consequently have their home repossessed, is 1.13%. They are more 98.86% more likely to have failed the exam and not have their home repossessed by the bank.

```
> #8:
>     #New customer from #6 failed the Barclay Test.
> # Use the samecontingency table as #7 (code copied below)
> BarclayTest <- matrix(c(0,93935,69,5996),ncol = 2, byrow = TRUE)
> colnames(BarclayTest) <-c("Repossessed","Not Repossessed")
> rownames(BarclayTest) <-c('Pass','Fail')
> BarclayTest <- as.table(BarclayTest)
> BarclayTest
     Repossessed Not Repossessed
Pass           0           93935
Fail          69            5996
>
> #prob new customer will default on mortgage = -prob of fail
> BarclayTestProb_fail <- BarclayTest/margin.table(BarclayTest)
> BarclayTestProb_fail[2,]/sum(BarclayTestProb_fail[2,])
    Repossessed Not Repossessed
     0.01137675      0.98862325
```

IST772

Parin Patel
Homework 2
10/13/2019

Final Code:

```
#1: 100,000 coin flips
dt1<-table(rbinom(n=100000,size = 9,prob=0.5))
dt1

#2:
barplot(dt1, main = "Bar Plot of 100,000 trials ") #barplot of n=100,000
ProbTable_bp<-barplot(table(rbinom(n=100000,size = 9,
    prob=0.5))/100000, main = "Bar Plot of Probabilities") #barplot of probabilities


 #Save Tables-
  #ProbTable
 probTable_prob<-data.frame(table(rbinom(n=100000,size = 9,prob=0.5))/100000)
  #100,000 events table
 probTable<-(table(rbinom(n=100000,size = 9,prob=0.5)))
  probTable
  View(probTable_prob)



#6:
    #Background:
       # 50 College, 50 High School
       # 80 Passed, 20 Failed
       # 3 college students failed -
       # Therefore:
          # 17 High school students failed.
          # 3 CS failed = 47 Passed:
          # 17 HSS failed = 33 Passed
#contingencyTable
StatsExam <- matrix(c(47,33,3,17),ncol = 2, byrow = TRUE)
colnames(StatsExam) <-c('College','High School')
rownames(StatsExam) <-c('Pass','Fail')
StatsExam <- as.table(StatsExam)
StatsExam
```

IST772

Parin Patel
Homework 2
10/13/2019

```
#replace counts of students with probabilites
ProbTable_statsExam <- StatsExam/margin.table(StatsExam)
ProbTable_statsExam

#normalize
ProbTable_statsExamFinal <- StatsExam/margin.table(StatsExam)
ProbTable_statsExam[,2]/sum(ProbTable_statsExam[,2])
```

```
#7.
        #Background:
          # 71 of 100,000 homes repossessed by bank
          #93,939 households pass
          # 6,065 failed
             #5996 of failed not Default/noRepos
             #Therefore, 6065-5996 = 69 actual Fails

BarclayTest <- matrix(c(0,93935,69,5996),ncol = 2, byrow = TRUE)
colnames(BarclayTest) <-c("Repossessed","Not Repossessed")
rownames(BarclayTest) <-c('Pass','Fail')
BarclayTest <- as.table(BarclayTest)
BarclayTest


#probability of passing test and having home repossessed
BarclayTestProb <- BarclayTest/margin.table(BarclayTest)
BarclayTestProb[1,]/sum(BarclayTestProb[1,])


#8:
   #New customer from #6 failed the Barclay Test.
# Use the samecontingency table as #7 (code copied below)
BarclayTest <- matrix(c(0,93935,69,5996),ncol = 2, byrow = TRUE)
colnames(BarclayTest) <-c("Repossessed","Not Repossessed")
rownames(BarclayTest) <-c('Pass','Fail')
BarclayTest <- as.table(BarclayTest)
BarclayTest

#prob new customer will default on mortgage = -prob of fail
BarclayTestProb_fail <- BarclayTest/margin.table(BarclayTest)
```

IST772

Parin Patel
Homework 2
10/13/2019
BarclayTestProb_fail[2,]/sum(BarclayTestProb_fail[2,])