HW2_Videtti

```
#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.

#Setting the seed will allow predictable results so that answers still make
sense after knitting.
set.seed(1)

#Flipping a fair coin 9 times.
rbinom(1,9,0.5)
```

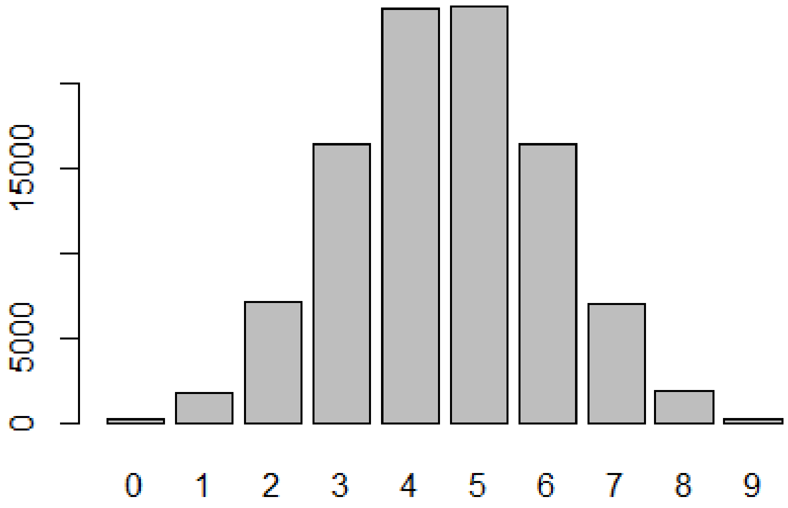
```
## [1] 4
```

```
#4 heads out of nine coin flips

#Repeating 100,000 times. Saving this into a variable for future use.
Results <- table(rbinom(100000,9,0.5))
Results
```

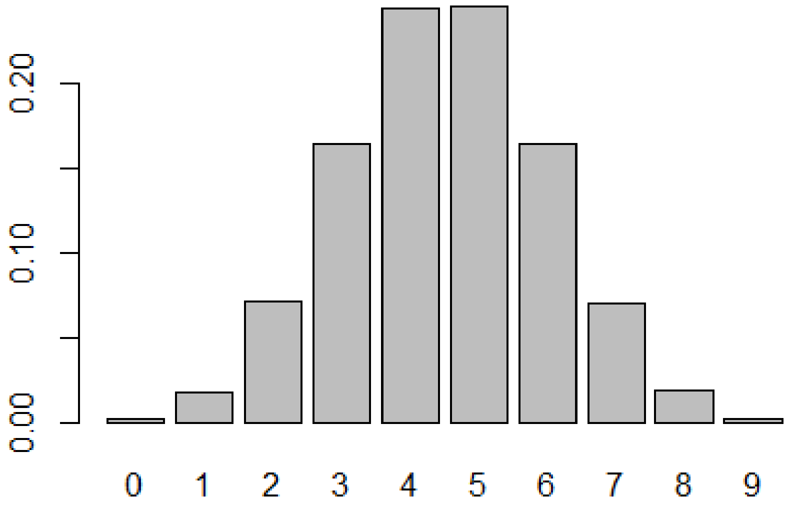
```
## 
##     0     1     2     3     4     5     6     7     8     9 
##   206  1785  7106 16447 24463 24555 16417  6991  1861   169
```

```
#There were 206 trials where there were 0 heads and 9 tails
#There were 1,785 trials where there was 1 heads and 8 tails
#There were 7,106 trials where there were 2 heads and 7 tails
#There were 16,447 trials where there were 3 heads and 6 tails
#There were 24,463 trials where there were 4 heads and 5 tails
#There were 24,555 trials where there were 5 heads and 4 tails
#There were 16,417 trials where there were 6 heads and 3 tails
#There were 6,991 trials where there were 7 heads and 2 tails
#There were 1,861 trials where there were 8 heads and 1 tails
#There were 169 trials where there were 9 heads and 0 tails








#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.

#Barplot of results from Exercise 1
barplot(Results)
```



```
#Probabilities of results from Exercise 1 (divide all by the number of trials
- 100,000)
Results/100000

##
##       0       1       2       3       4       5       6       7       8
9
## 0.00206 0.01785 0.07106 0.16447 0.24463 0.24555 0.16417 0.06991 0.01861
0.00169

#Barplot of probabilities from Exercise 1
barplot(Results/100000)
```



```
#Each bar in the barplots signify the number of heads that were in each trial
of 9 flips of a fair coin. Since both barplots are based off the same data,
and the only difference is that the probabilities barplot is the frequency
barplot divided by 100000, both should look the same at first glance. The
only difference should be that the y-axis labels change, specifically they
are 1/100000th in the probability barplot of what they are in the frequency
barplot. This is because we go from a scale of 0 to 100,000 to a scale of 0
to 1. Also, these barplots both resemble normal distribution. This is due to
the large number of trials that we have done. The larger the number of
trials, the closer we get to normal distribution. The center of the barplot
includes 4 and 5. This is the center becasue it is the middle two values of
integers 1 to 6, making this the median. We see this is also the mean since
we have a symmetrical distribution, and becasue we have approximately normal
distribution, where the center of the distribution is where the mean is.








#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?

#100 students total, 50 college and 50 high school
#80 passed, 20 failed
#3 college students failed, meaning that 50-3 = 47 college students passed
#3 college students failed, meaning 20-3 = 17 high school students failed
#17 high school students failed, meaning 50-17 = 33 high school students passed
#Let's build a contingency table!

Exercise6Contingency <- matrix(c(47,3,33,17),nrow = 2)
rownames(Exercise6Contingency) <- c('Passed','Failed')
colnames(Exercise6Contingency) <- c('College','High School')
Exercise6Contingency
```

```
##        College High School
## Passed      47          33
## Failed       3          17
```

```
#Now, let's add marginal totals
Exercise6Contingency <- matrix(c(Exercise6Contingency[,1],margin.table(Exercise6Contingency,2)[1],Exercise6Contingency[,2],margin.table(Exercise6Contingency,2)[2],margin.table(Exercise6Contingency,1)[1],margin.table(Exercise6Contingency,1)[2],margin.table(Exercise6Contingency)),nrow = 3)

rownames(Exercise6Contingency) <- c('Passed','Failed','Column Totals')
colnames(Exercise6Contingency) <- c('College','High School','Row Totals')
Exercise6Contingency
```

```
##               College High School Row Totals
## Passed             47          33         80
## Failed              3          17         20
## Column Totals      50          50        100
```

```
#Dividing by 100 (total number of students) will give us probabilities
Exercise6Probabilities <- Exercise6Contingency/100
Exercise6Probabilities
```

```
##                College High School Row Totals
## Passed            0.47        0.33        0.8
## Failed            0.03        0.17        0.2
## Column Totals     0.50        0.50        1.0
```

```
#We see that the pass rate for high school students is a little more
difficult. First, we will need to only pull the 2nd column in order to see
the High School probabilities, then we will need to normalize the column by
dividing by the probability of a student in this sample being a high school
student (0.5).

HighSchoolProbabilities <-
Exercise6Probabilities[1:2,2]/Exercise6Probabilities[3,2]
HighSchoolProbabilities
```

```
## Passed Failed
##   0.66   0.34
```

```
#We now see that given a student is a high school student, they are 66%
likely to pass the test.
```

```
#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?

#93,935 pass, 6,065 fail
#5,996 fail AND not repossessed, this means 6,065 - 5,996 = 69 fail AND
repossessed
#71 repossessed, this means that 100,000 - 71 = 99,929 not repossessed

#Lets fill in our table with what we know so far and make the rest 0's for
now. We will start with the marginal totals.
Exercise7 <- matrix(c(0,0,71,0,0,99929,93935,6065,100000),nrow=3)
colnames(Exercise7) <- c('Repossessed','Not Repossessed','Row Totals')
rownames(Exercise7) <- c('Pass','Fail','Column Totals')
Exercise7
```

```
##                 Repossessed Not Repossessed Row Totals
## Pass                    0               0      93935
## Fail                    0               0       6065
## Column Totals          71           99929     100000

#Now, we can add our information about how 5,996 fail AND not repossessed,
which means 6,065 - 5,996 = 69 fail AND repossessed
Exercise7[2,2] <- 5996
Exercise7[2,1] <- 69
Exercise7

##                 Repossessed Not Repossessed Row Totals
## Pass                    0               0      93935
## Fail                   69            5996       6065
## Column Totals          71           99929     100000

#Next, we subtract to find the last two values.
#71-69 = 2 passed and had their homes repossessed
#99929 - 5996 = 93933 passed and didn't have their homes repossessed
Exercise7[1,1] <- Exercise7[3,1] - Exercise7[2,1]
Exercise7[1,2] <- Exercise7[3,2] - Exercise7[2,2]
Exercise7

##                 Repossessed Not Repossessed Row Totals
## Pass                    2           93933      93935
## Fail                   69            5996       6065
## Column Totals          71           99929     100000

#Now, finally, lets make these probabilities!
Exercise7 <- Exercise7/100000
Exercise7

##                 Repossessed Not Repossessed Row Totals
## Pass              0.00002         0.93933    0.93935
## Fail              0.00069         0.05996    0.06065
## Column Totals     0.00071         0.99929    1.00000

#We see that 93.933% of customers both pass the test and do not have their
homes repossessed.




#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.

#To find probability of a customer defaulting on their mortgage, given the
new customer fails the test, we will need to isolate just the "Fail"
```

```
probabilities row, and we will also need to normalize those by dividing by
the probability of failing overall (0.06065).

Exercise8 <- Exercise7[2,1:2]/Exercise7[2,3]
Exercise8

##     Repossessed Not Repossessed
##      0.01137675      0.98862325

#We that given a new customer fails the test, there is approximately a 1.14%
chance they will actually default on their mortgage.
```