IST-772 Quantitative Reasoning in Data science

Week2/HW-2: Basic Probability

Reasoning with Probability (Page-35: Problems-1, 2, Page-36: Problems-6,7,8)

Sharat Sripada (vssripad@syr.edu)

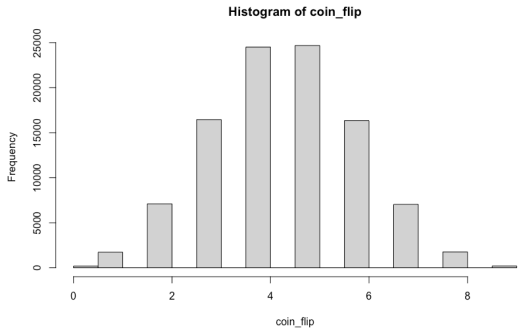
1. Flip a coin nine times and write down the number of heads obtained. Now repeat this process 100,000 times. Write down the results and explain in your own words what they mean

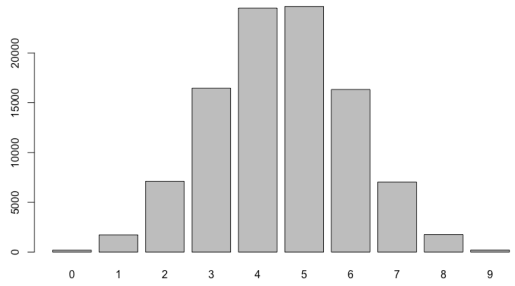
Used the following code to generate the simulation above:

```
# Simulate flipping a coin 100,000 times. Explain the results
coin_flip <- rbinom(n=100000, size=9, prob=0.5)
hist(coin_flip) # See gaps in the histogram since it is trying to plot intermediate
values like 1.5, 2.5 etc.

coin_flip_table <- table(coin_flip)
barplot(coin_flip_table)
```

See the corresponding histogram and bar plot:





Correspondingly, the table view of this as below:

```
> table(coin_flip)
coin_flip
    0     1     2     3     4     5     6     7     8     9
  189  1728  7103 16461 24512 24674 16335  7038  1762   198
```

This shows a normal distribution and can be summarized as follows:

100,000 coin-flip trials, 9 flips per trial										
HEADS count	0	1	2	3	4	5	6	7	8	9
Number of trials with that count	189	1728	7103	16461	24512	24674	16335	7038	1762	198

Interpretation: The number of trials where event outcomes had *NO* HEADS (up) were 189 and trials when outcomes were *ALL* HEADS (up) were 198. Event outcomes with 1 HEADS, 2 HEADS or 3 HEADS were 1728, 7103, 16461 respectively and so on.

With unbiased coin-toss (or any binomial event) we expect fairness, meaning 50% chance for HEADS or TAILS. This is corroborated with a high number of trials under columns 4 or 5 in row *HEADS count*. An aggregate or sum of trials under column 4 and 5 amounts to 49,186 which is approximately 50% of 100k trials.

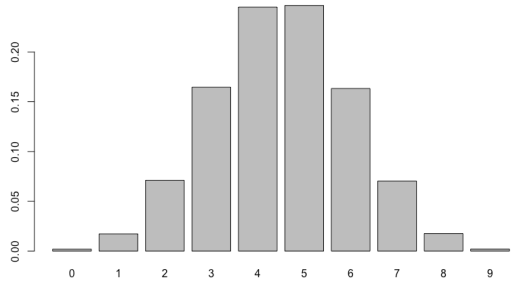
2. Using the output from Exercise 1, summarize the results of your 100,000 trials of nine flips each in a bar plot. Convert the results to probabilities and represent that in a bar plot as well.

Comment on the shape of each bar plot and why you believe that the bar plot has taken that shape.

Using the following R code, we can convert to outcomes to probabilities:

```
> coin_flip_prob_table <- coin_flip_table/100000
> barplot(coin_flip_prob_table)
> coin_flip_prob_table
coin_flip
      0       1       2       3       4       5       6       7       8       9
0.00189 0.01728 0.07103 0.16461 0.24512 0.24674 0.16335 0.07038 0.01762 0.00198
```

Correspondingly, the bar plot representing the probability table above:



Interpretation: Both bar plots (outcomes represented as counts or probability) are comparable in that they follow a bell-curve and are type *Normal Distribution*. Since a high density of outcomes are around the center with 4 or 5 HEADS amounting **~ 0.49** (0.24512 + 0.24674), the trails and outcomes seem fair and un-biased.

3. 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. Additional information: Only 3 college students failed the test.

Comment on why the additional information was required. Create a second copy of the table with probabilities and derive the probability of pass rate of high school students.

Without the additional information of college students who failed the test, the contingency table will comprise data in Marginal rows and columns only. With that alone it is not possible to answer granular questions of whether high school or college students, fared better at the statistics test.

The contingency table can be represented as follows with derived values represented in orange:

Contingency table for a statistics test			
	High school students	College students	Marginal
Passed	33	47	80
Failed	17	3	20
Marginal	50	50	100

Representing the table with probability numbers:

Contingency table *(probability)* for a statistics test			
	High school students	College students	Marginal
Passed	0.33	0.47	0.8
Failed	0.17	0.03	0.2
Marginal	0.50	0.50	1

Using this, we see that the probability of high school students passing the statistics test is 0.33 or 33%.

4. In a typical year 71 out of 100,000 homes in UK is repossessed by the bank because of mortgage default. Barclays developed a screening test and obtained the following: 93,935 households pass the test, and 6,065 households fail the test. 5,996 of those who failed the test were households that were doing fine on their mortgage.

Construct a complete contingency table from this information. What percent of customers both pass the test and do not have their homes repossessed?

Contingency table *(probability)* for screening test by Barclays Bank			
	Repossessed	Not repossessed	Marginal
Passed	2	93933	93935
Failed	69	5996	6065
Marginal	71	99929	100000

The derived values are represented in orange.

Using the table above, we can see that 93,933 out of 100,000 or 93.93% of people passed the screening test and do not have their homes repossessed.

5. Imagine that Barclays deploys the screening test on a new customer and the new customer fails the test. What is the probability that the customer will default on his or her mortgage?

Converting the contingency table to comprise probability numbers:

Contingency table *(probability)* for screening test by Barclays Bank			
	Repossessed	Not repossessed	Marginal
Passed	0.00002	0.93933	0.93935
Failed	0.00069	0.05996	0.06066
Marginal	0.00071	0.99929	1

From the table there is a very low probability 0.00069 or 0.07% that a customer who failed the test was likely to default on his or her mortgage.