Artificial Intelligence with Python

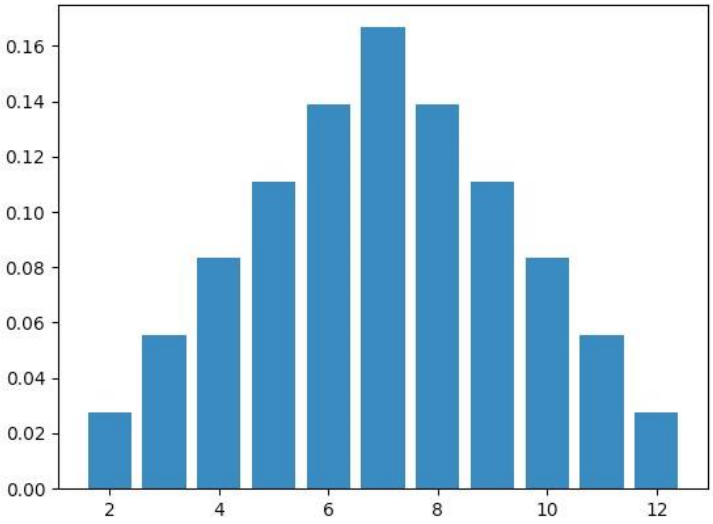
Assignment 4

Exercise 1: Regression to the mean

1. When throwing two fair dice the probabilites of possible values are

2 or 12	1/36 = 2.8%
3 or 11	2/36 = 5.6%
4 or 10	3/36 = 8.3%
5 or 9	4/36 = 11.1%
6 or 8	5/36 = 13.9%
7	6/36 = 16.7%

The histogram looks like this.



0) Write a for loop which repeats the steps 1)-3) below for values of n ranging as 500, 1000, 2000, 5000, 10000, 15000, 20000, 50000, 100000

1) Use numpy to simulate throwing of two dice n times. Compute the sum of the dice.

2) Use numpy's histogram() function to compute the frequencies as `h,h2 = np.histogram(s,range(2,14))`

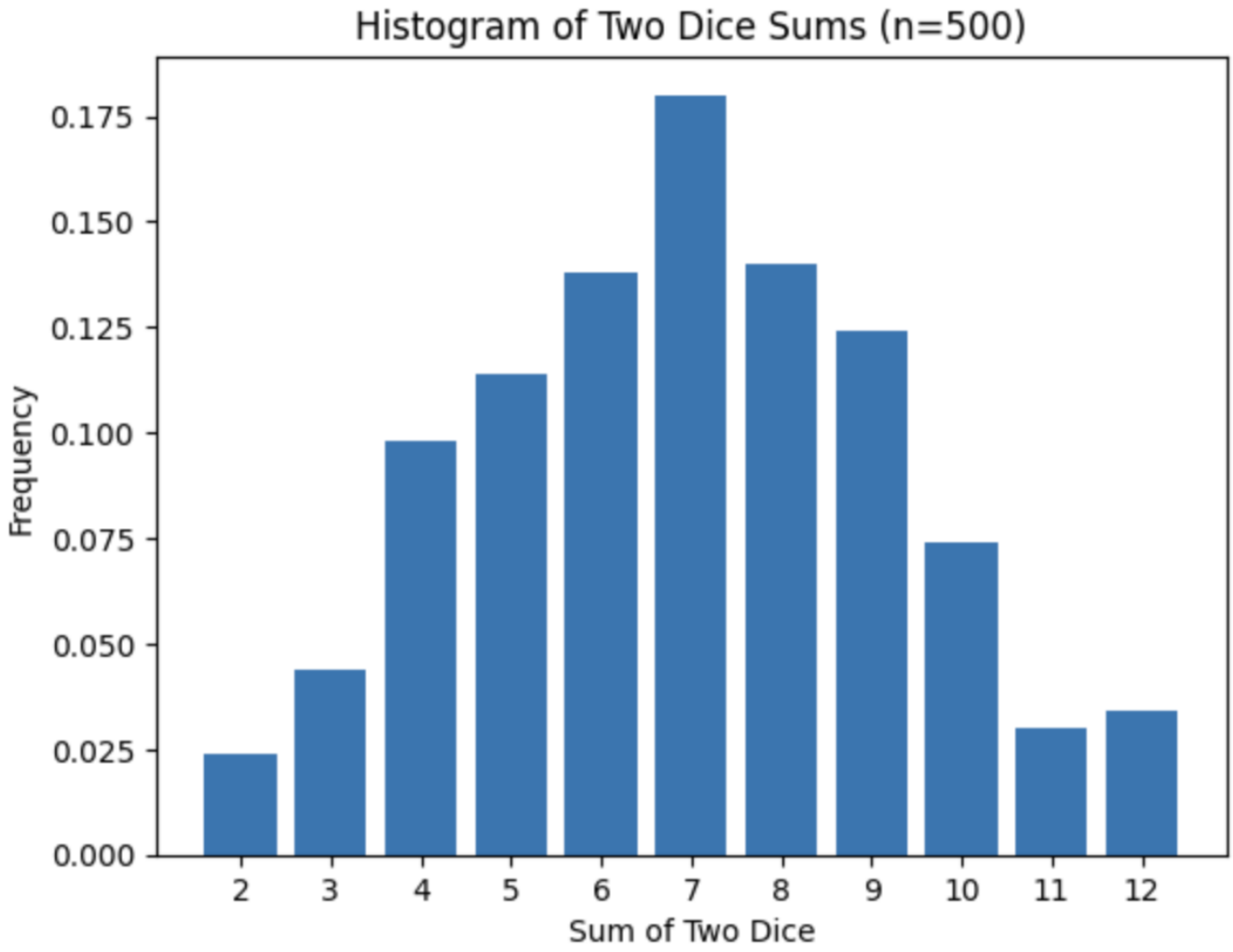
where s contains the sum.

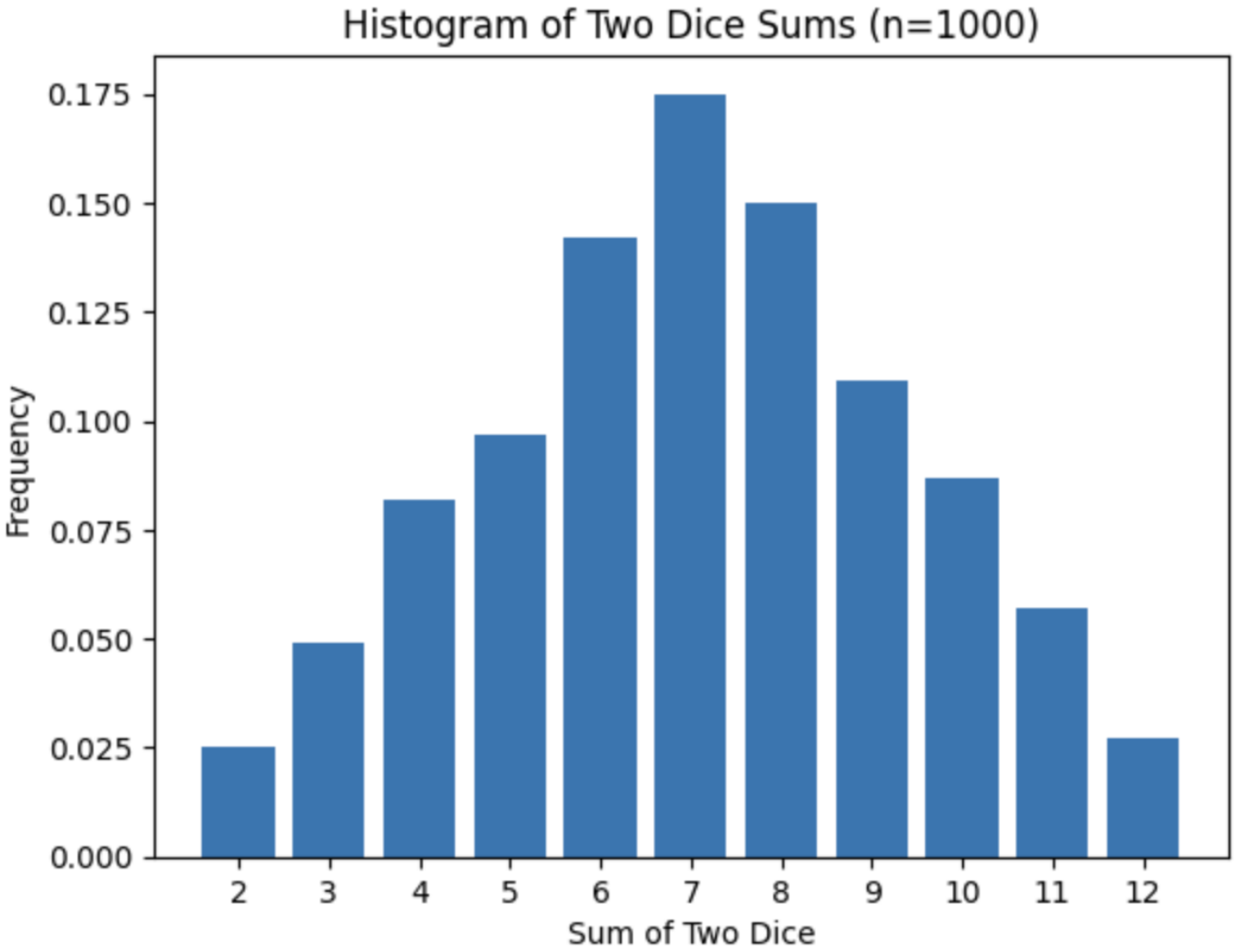
3) Use matplotlib's bar function to plot the histogram as `plt.bar(h2[:-1],h/n)` and show the value of n in the title.

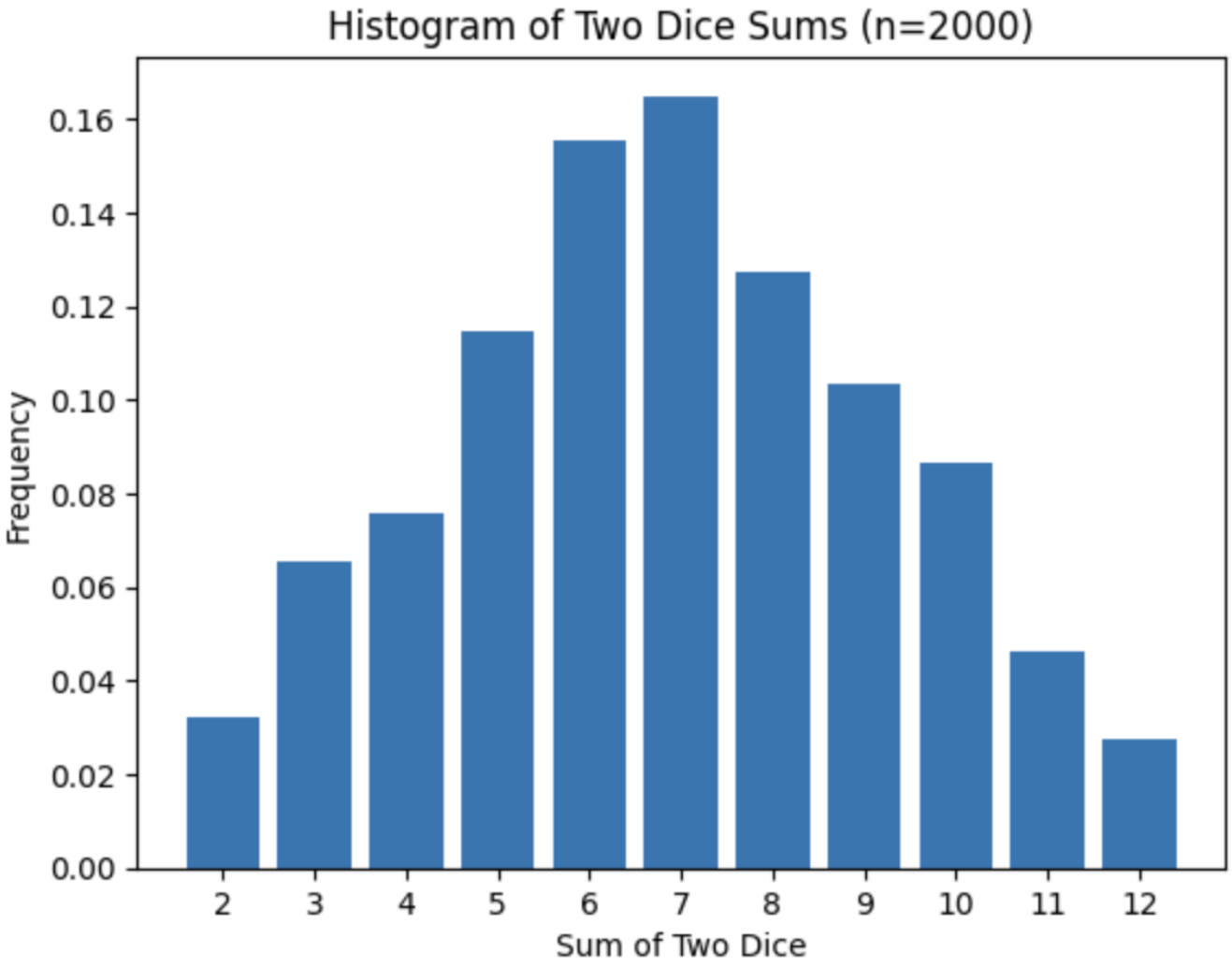
4) What do you observe? You may need to run the loop a few times to see it.

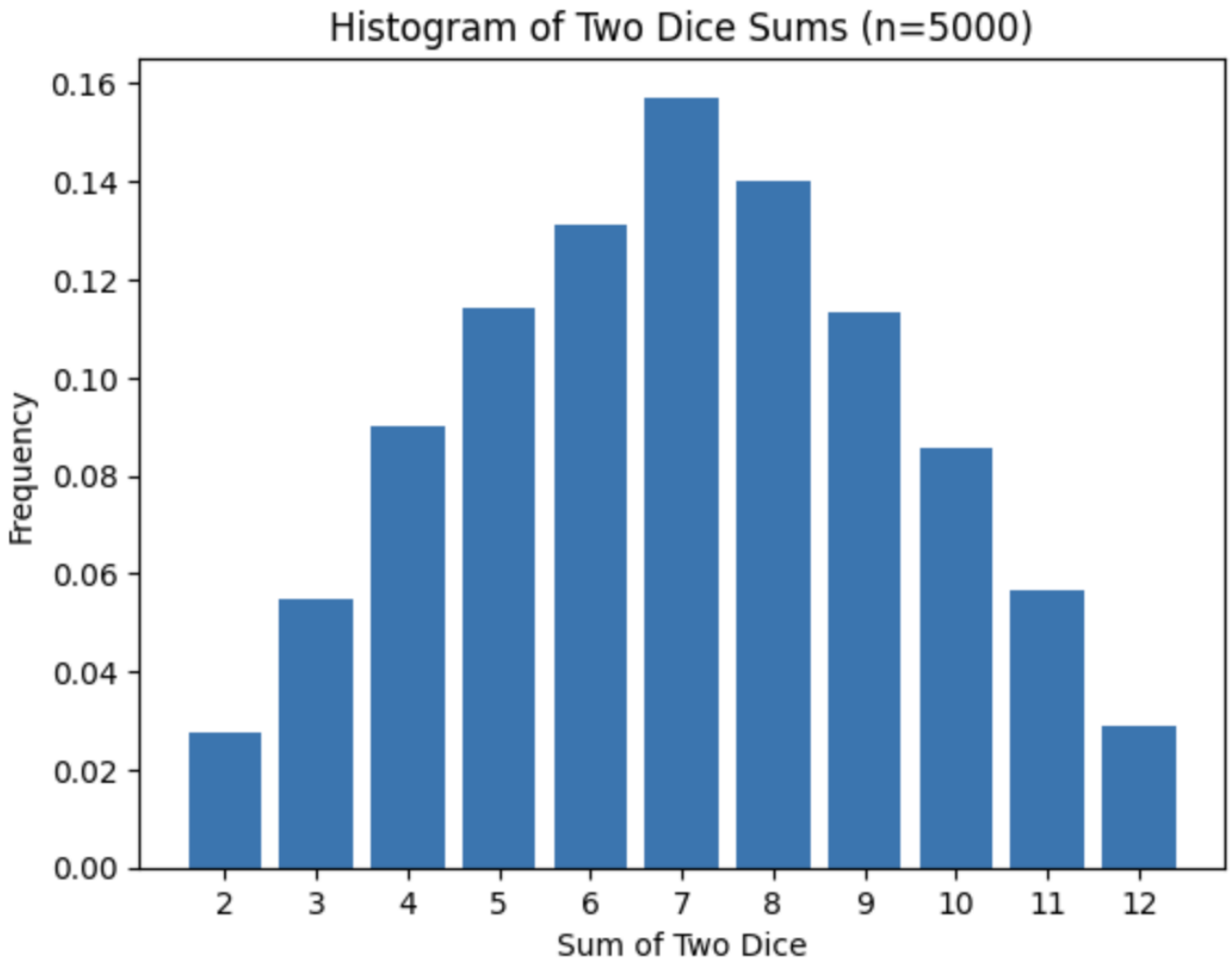
5) How is this related to "regression to the mean"?

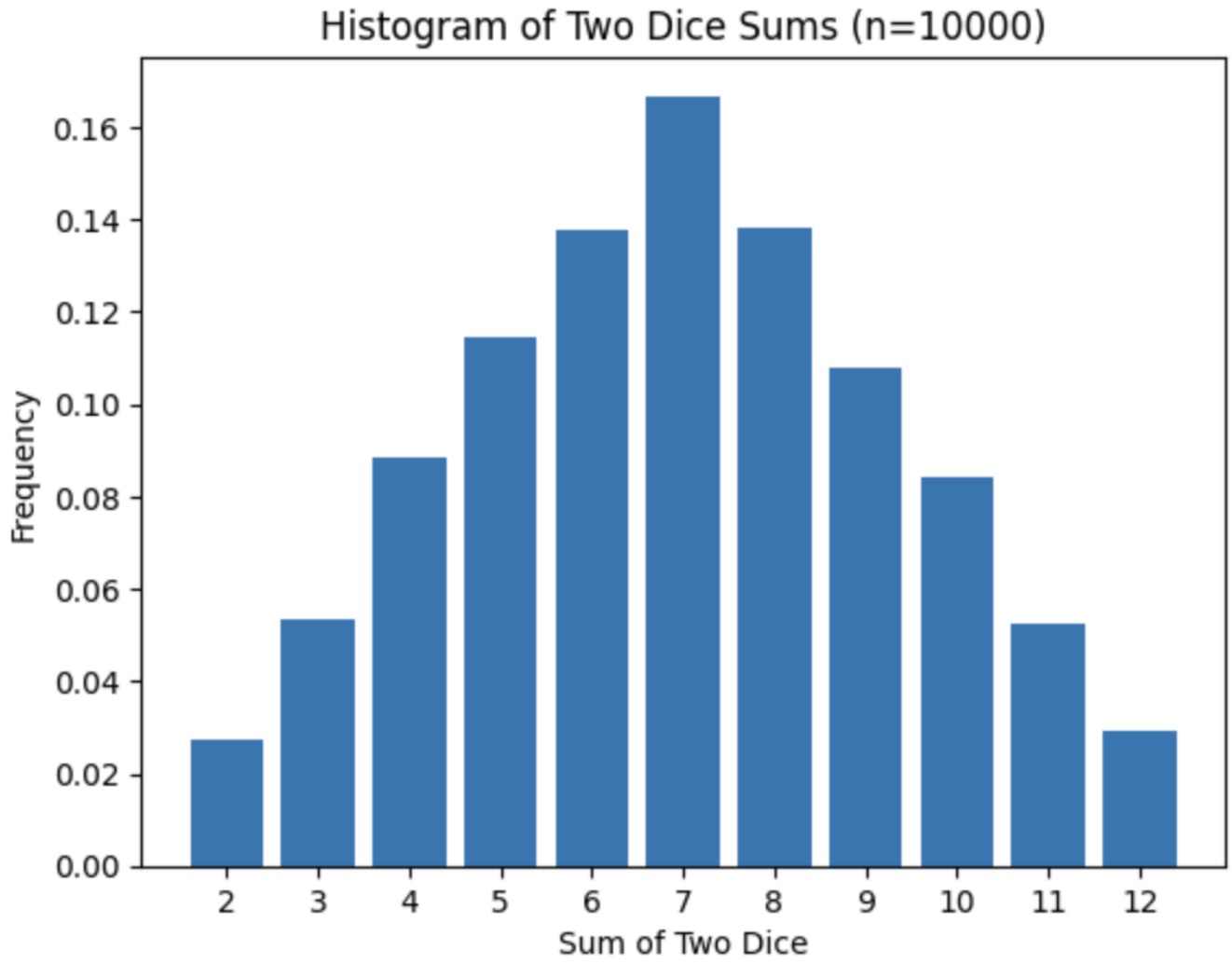
The "regression to the mean" refers to a statistical phenomenon where the extreme (i.e., very high or very low) values tend to move closer to the average. In the dice experiment, when the experiment size was low, the outcome of each possible result (i.e., 2 - 12) was far from the theoretical probability. As the size increased, the outcome moved closure to the theoretical probability i.e., the observed outcome/frequencies "regress" toward the theoretical probability. This is because that the initial extreme variations are balanced out and the outcomes "regress" toward the theoretical probabilities.

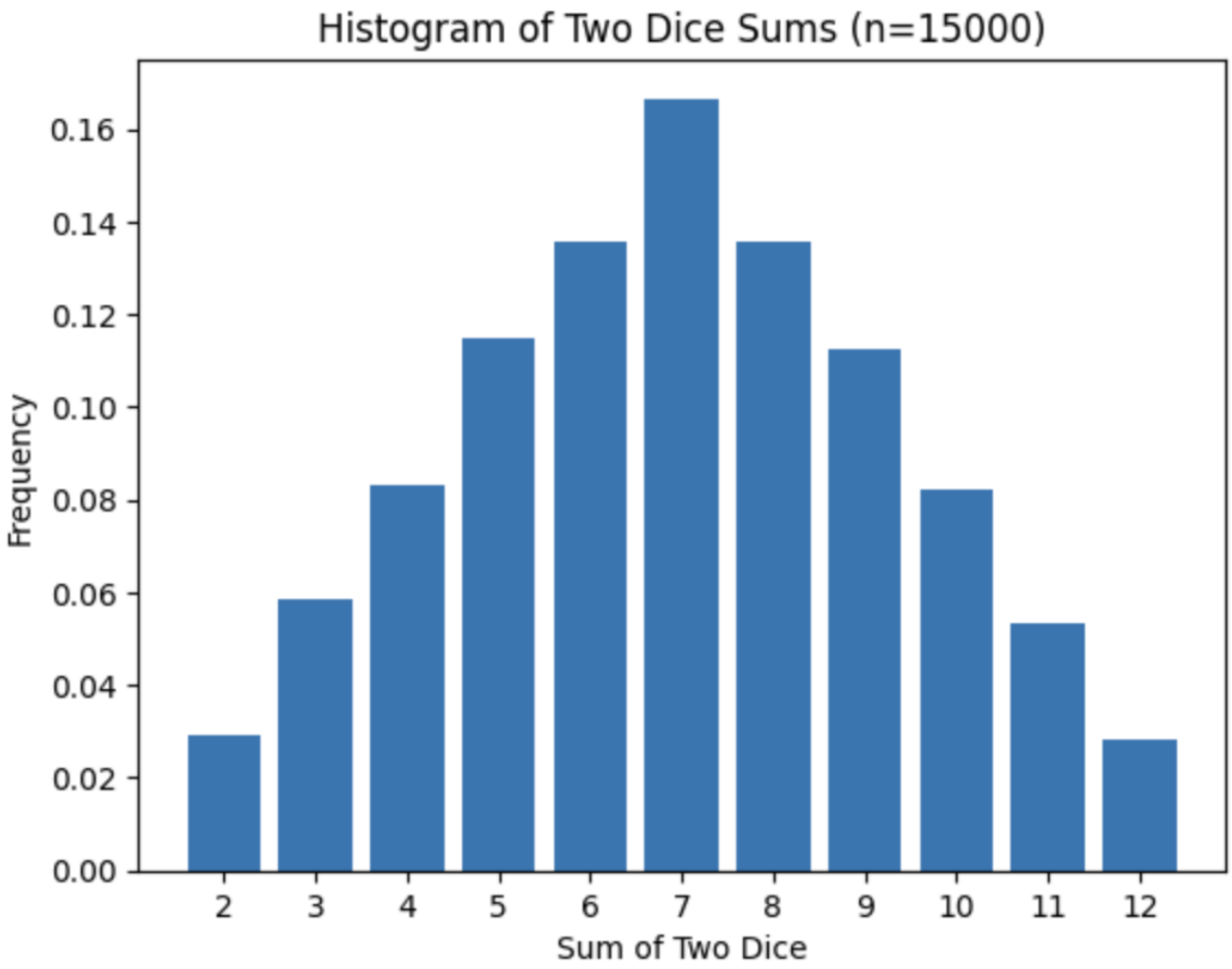


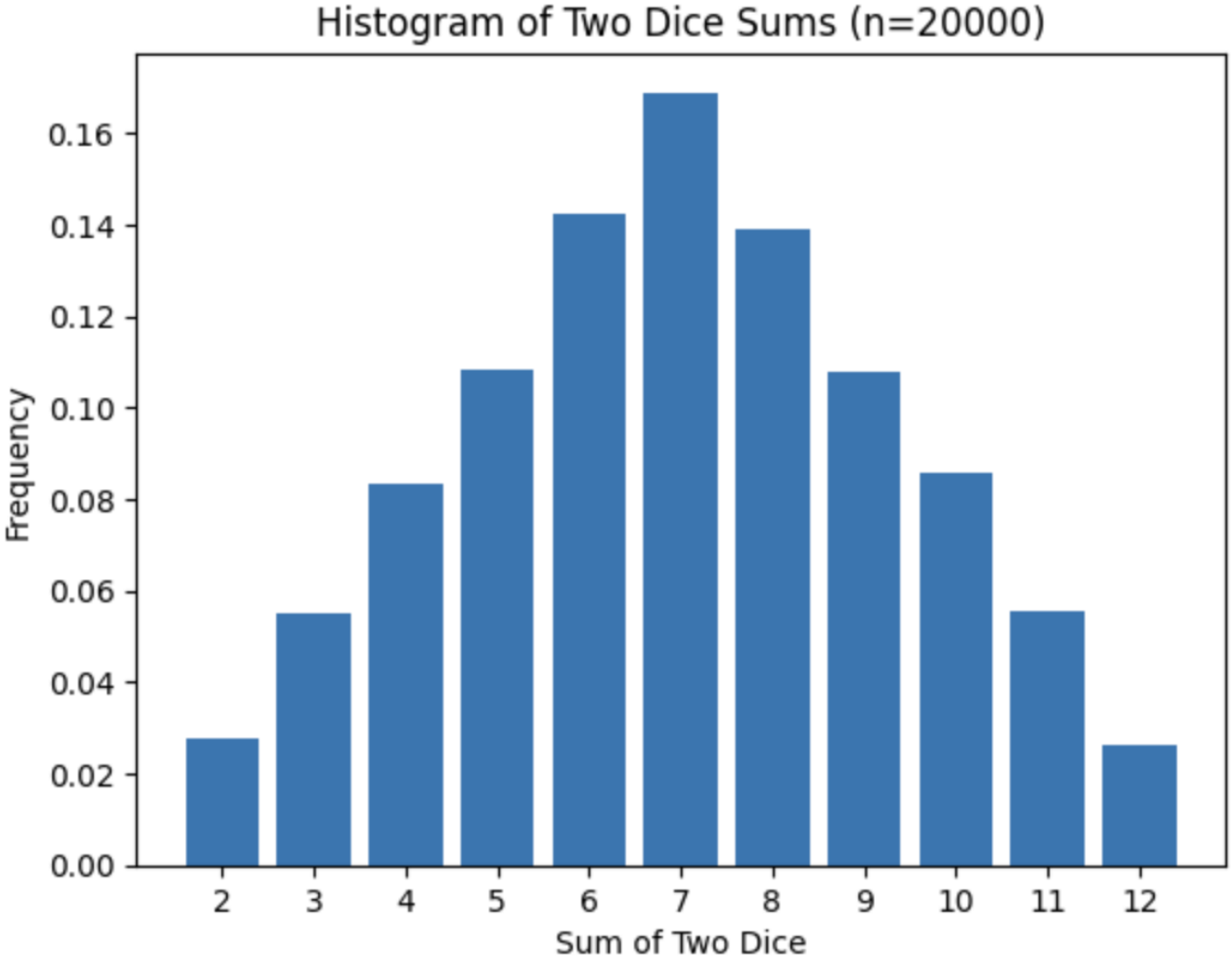

Histogram of Two Dice Sums (n=1000)
Frequency
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice

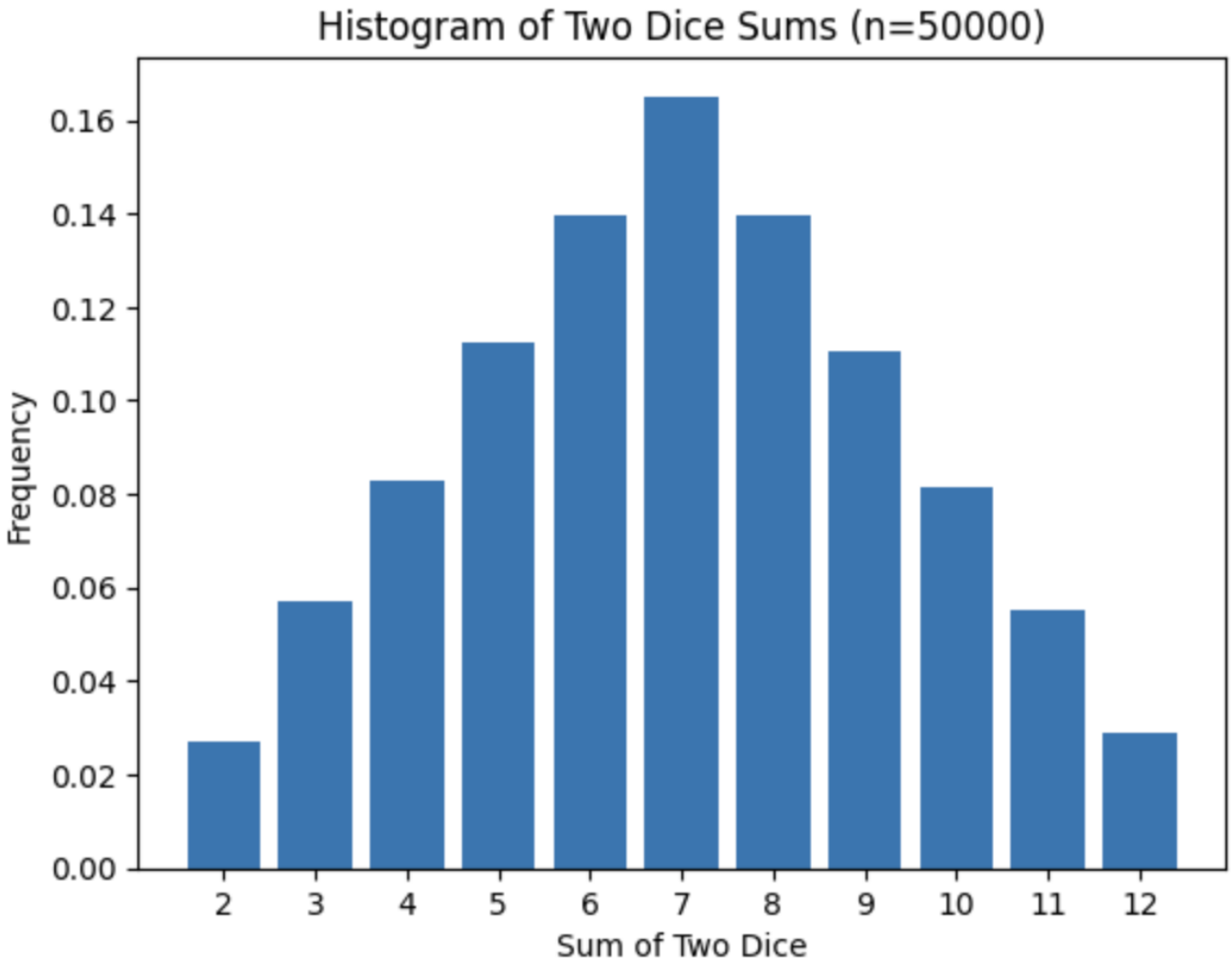

Histogram of Two Dice Sums (n=2000)
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice

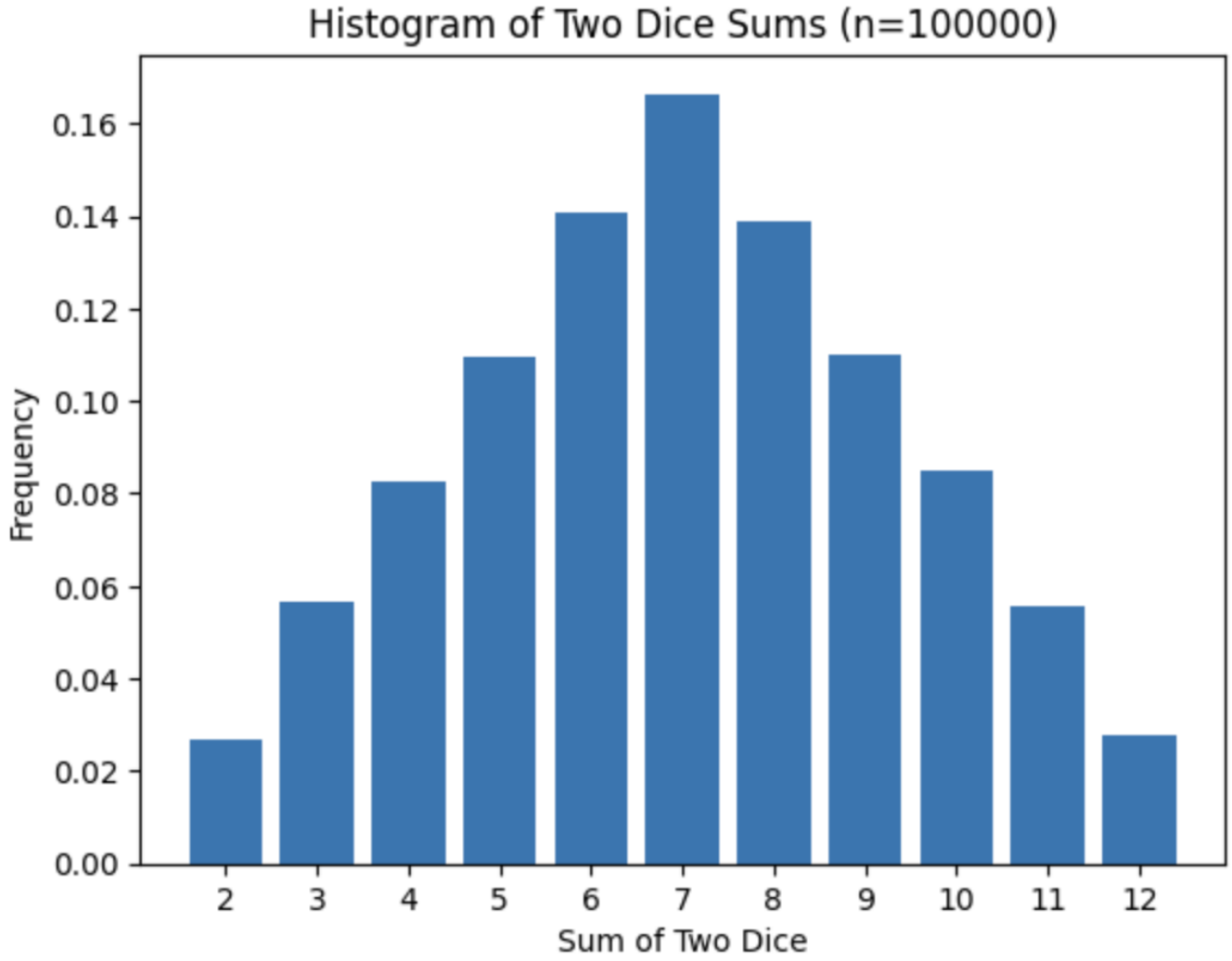

Histogram of Two Dice Sums (n=5000)
Frequency
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice


Histogram of Two Dice Sums (n=10000)
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice


Histogram of Two Dice Sums (n=15000)
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice


Histogram of Two Dice Sums (n=20000)
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice


Histogram of Two Dice Sums (n=50000)
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice


Histogram of Two Dice Sums (n=100000)
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
3
4
5
6
7
8
9
10
11
12
Sum of Two Dice

Exercise 2: Regression Model

Consider the data from the file weight-height.csv.

1) Inspect the dependence between height and weight using a scatter plot. You may use either of the variables as independent variable.

2) Choose appropriate model for the dependence

3) Perform regression on the data using your model of choice

4) Plot the results

5) Compute RMSE and R2 value

```
RMSE: 1.4641414538503257, R2: 0.8551742120609958
```

6) Assess the quality of the regression (visually and using numbers) in your own words.