Henry <hc7145@bard.edu>
9/10/2018
CMSC 201
lab1
Collaboration: I worked alone on this assignment.
Learned from: https://en.wikipedia.org/wiki/Benford%27s_law

Benford's law analysis:

I chose the first 82 numbers of the Fibonacci sequence, which has a strong relation to the growth of the biological world, as my data.txt. Both this one and the heights.txt dataset follow a trend that is similar to Benford's law, and the other two (r7 and r5) dataset are obviously far from following it. We can view the frequency distribution of first digits of each dataset below:

r5.txt

First Digit	Percent(%)	Times occurred	Graph
1	16.66	116627	*********
2	2.79	19525	*
3	5.58	39090	***
4	8.34	58363	*****
5	11.11	77738	******
6	13.77	96386	********
7	16.66	116633	*********
8	13.99	97928	********
9	11.10	77710	******

r7.txt

First Digit	Percent(%)	Times occurred	Graph
1	100	700000	***
2	0	0	
3	0	0	
4	0	0	
5	0	0	
6	0	0	
7	0	0	
8	0	0	
9	0	0	

heights.txt

First Digit	Percent(%)	Times occurred	Graph
1	47.06	24	****************************
2	7.84	4	****
3	15.69	8	*********
4	11.76	6	*******
5	1.96	1	*
6	5.88	3	***
7	5.88	3	***
8	3.92	2	**
9	0.00	0	

data.txt

First Digit	Percent(%)	Times occurred	Graph
1	29.27	24	*****************
2	18.29	15	**********
3	13.41	11	********
4	8.54	7	*****
5	8.54	7	*****
6	6.10	5	***
7	4.88	4	**
8	7.32	6	****
9	3.66	3	**

To measure how the frequency distributions of these datasets are similar to the Benford's frequency distribution, I compare them to a standard Benford's distribution, and measure the errors(%) calculated by: (P(first digit from dataset) – P(first digit of Benford's law))/ P(first digit of Benford's law). If an error is between +40 and -40, it will be highlighted in green to say the result is somehow similar to Benford's law.

First Digit	data	heights	r5	r7
1	-20.27	-20.27	232.23	-44.65
2	-14.77	-77.27	-100.00	-84.15
3	-12.00	-36.00	-100.00	-55.33
4	-27.84	-38.14	-100.00	-14.05
5	-11.39	-87.34	-100.00	40.58
6	-25.37	-55.22	-100.00	105.51
7	-31.03	-48.28	-100.00	187.27
8	17.65	-60.78	-100.00	174.31
9	-34.78	-100.00	-100.00	141.34

And then I use the Stats.java to break down the significant attributes of each dataset.

Dataset	r5.txt	r7.txt	heights.txt	Data.txt
N datapoints	700000	700000	51	82
Min	2	10000000	75	1
Max	12	10000001	829.8	6.1E+16
Mean	7	10000000.5	265.17	1.9E+15
Std. Dev.	2.42	approx. 0	171.69	8.4E+15

In this form that displays statistical results, it is clearly shown that the dataset r7.txt is impossible to follow Benford's law because it has only 1 as the starting digit for the numbers inside it.

In conclusion, the first 82 numbers of the Fibonacci sequence have a frequency distribution of leading digits that is the most similar to the Benford's law among the four datasets used in this analysis. The heights.txt dataset also has some similarity, and r5 and r7 has no significant similarity with the Benford's law.