Arnav Shirodkar

CMSC 201: Data Structures

Lab 1 Report – Benford's Law (07/02/2020)

Benford's Law - Findings

After first piping the data in the file "heights.txt" through "Benford.java", I retrieved a new file "startDigits.txt" and piped the values into "Hist.java". I did the same for the data in "r5.txt" and "r7.txt" – photos of all the results are included on the right.

First Digit	startDigits.txt	Percentage
1	24	47.06%
2	4	7.84%
3	8	15.69%
4	6	11.76%
5	1	1.96%
6	3	5.88%
7	3	5.88%
8	2	3.92%
9	0	0.00%

Benford's law states that in sets of naturally occuring numerical data, the first significant figure is likely to be small, with the digit 1 appearing at least 30% of the time and the successive numbers appearing less and less frequently.

First Digit	R5.txt	Percentage
1	116627	16.66%
2	19525	2.79%
3	39090	5.58%
4	58363	8.34%
5	77738	11.11%
6	96386	13.77%
7	116633	16.66%
8	97928	13.99%
9	77710	11.10%

This trend seems to generally hold true for the data from the heights.txt, as we can observe a general trend of decreasing occurrence from 1 – 9. However, this does not seem to hold true for the values in r5.txt and those in r7.txt. In the case of r7.txt, we have a unique situation where all the numbers are written in scientific notation and all have 1 as their first digit (r7.txt may not contain a "naturally occuring set of numbers").

First Digit	R7.txt	Percentage
1	700000	100.00%
2	0	0.00%
3	0	0.00%
4	0	0.00%
5	0	0.00%
6	0	0.00%
7	0	0.00%
8	0	0.00%
9	0	0.00%

I then explored a data set of my own, running the same set of operations on all the prime numbers within the range of 1-1000. Benford's law seems to generally hold true in this case as we can see the the occurrence of each digit has a generally decreasing trend.

First Digit	Primes.txt	Percentage
1	25	14.88%
2	19	11.31%
3	19	11.31%
4	20	11.90%
5	17	10.12%
6	18	10.71%
7	18	10.71%
8	17	10.12%
9	15	8.93%

Stats – Findings

Stats from r5.txt

N	min	max	mean	Standard Deviation
700000	2.000	12.000	7.000	2.417

Stats from r7.txt

(BONUS – The first method provided for calculating Std.Dev does not work with r7.txt and Welford's method must be used – refer to Statfinder.java for reworked code).

N	min	max	mean	Standard Deviation
700000	10000000.000	10000001.000	10000000.500	0.289

Stats from heights.txt

N	min	max	mean	Standard Deviation
51	75.000	829.800	265.172	124.084