**CIS 229 – Python Programming – Programming Project**

**Benford’s Law**

Overview

In this assignment, the student will write a script that heuristically analyzes a series of data files to generate a forgery coefficient based on the mathematical distribution described by Benford’s Law.

When completing this assignment, the student should demonstrate mastery of the following concepts:

* Data File Processing
* Heuristic Analysis Algorithms
* Data File Formatting
* Benford’s Law

Assignment

In this assignment, you will be writing a program that uses Benford’s Law to test a file of numbers for potential fraudulence. Frank Benford was a 20th century physicist who formalized a mathematical law to predict the tendency of the first digit on naturally occurring numeric data. Intuitive reasoning typically leads most people to believe that when looking at a naturally occurring set of numbers (i.e. street numbers, zip codes, phone numbers, etc...) the first digit assumes a uniform distribution. In other words, the chance that a number will start with a 9 (like 9,456) is the same as the chance that a number will start with a 3 (like 35,949). It turns out that this assumption is incorrect and Benford developed a formal mathematical algorithm to determine a discrete probability distribution for the occurrence of a number as the first digit:

P(d) = log­b(d + 1) - logb(d) = logb(1 + 1/d)

d = the value of the digit being speculated about

b = the base of the number system in which the numbers exist

Since humans typically work in a base-10 number system, we will use this equation to develop the following probability distribution that describes the first digit of naturally occurring data.

P(1) = log10(1 + 1/1) ≈ 0.301

P(2) = log10(1 + 1/2) ≈ 0.176

P(3) = log10(1 + 1/3) ≈ 0.125

P(4) = log10(1 + 1/4) ≈ 0.097

P(5) = log10(1 + 1/5) ≈ 0.079

P(6) = log10(1 + 1/6) ≈ 0.067

P(7) = log10(1 + 1/7) ≈ 0.058

P(8) = log10(1 + 1/8) ≈ 0.051

P(9) = log10(1 + 1/9) ≈ 0.046

This algorithm is often used to detect forged data. For example, the IRS examines the numbers on tax return requests because natural data will typically conform to Benford’s law. If someone cheats on their taxes by crafting numbers, Benford’s law will not hold and the IRS knows to take a closer look at that refund request to make sure everything checks out.

In this program, you will be writing a Python script that examines a series of ten data files populated with numbers. The frequency of the occurrences of each digit will be tallied and compared against the Benford probabilities to determine which of the data file contains forged data.

A script that generates a series of ten data files is provided along with this assignment, which is to be used to test your algorithm. The script can be run in two modes based on the values assigned to the CUSTOM\_FILE\_LAST global. If CUSTOM\_FILE\_LAST is set to True, the forged data file will always be the last file in the series. If CUSTOM\_FILE\_LAST is set to False, the forged data file will be hidden among the data files and your program can be tested by setting it loose on the files and having it determine which file is forged.

Your fraud detection algorithm will have to open each of the data files and read the numbers in. From each number, the first digit must be determined and a series of counters will be incremented to keep track of how often each first digit occurs. These frequencies can then be divided by the number of items in the file to determine to relative frequency for the occurrence of each digit. Once these relative frequencies have been determined, the differences between the empirical relative frequencies and the established Benford relative frequencies (see the beginning of this assignment description) can be summed to generate a statistic that acts as a forgery coefficient.

Consider the following example of a forgery coefficient calculation:

The data files has been read and it contains 10000 entries.

The number began with a 1 3153 times.

The number began with a 2 1737 times.

The number began with a 3 1266 times.

The number began with a 4 977 times.

The number began with a 5 755 times.

The number began with a 6 650 times.

The number began with a 7 537 times.

The number began with a 8 464 times.

The number began with a 9 461 times.

Therefore, the following empirical relative probabilities are established:

P(the leading digit is 1) = 3153/10000 ≈ 0.315

P(the leading digit is 2) = 1737/10000 ≈ 0.174

P(the leading digit is 3) = 1266/10000 ≈ 0.127

P(the leading digit is 4) = 977/10000 ≈ 0.098

P(the leading digit is 5) = 755/10000 ≈ 0.075

P(the leading digit is 6) = 650/10000 ≈ 0.065

P(the leading digit is 7) = 537/10000 ≈ 0.054

P(the leading digit is 8) = 464/10000 ≈ 0.046

P(the leading digit is 9) = 461/10000 ≈ 0.046

From here, we will compare the empirical probability to the established Benford probabilities and calculate the sum of the absolute value of each difference:

|EmpiricalProbability – BenfordProbability| = BefordOffset

|0.315 - 0.301| = 0.014

|0.174 - 0.176| = 0.002

|0.127 - 0.125| = 0.002

|0.098 - 0.097| = 0.001

|0.075 - 0.079| = 0.004

|0.065 - 0.067| = 0.002

|0.054 - 0.058| = 0.004

|0.046 - 0.051| = 0.005

|0.046 - 0.046| = 0.000

Taking the sum of the Benford Offsets yeilds the ForgeryCoeffecient.

0.014 + 0.002 + 0.002 + 0.001 + 0.004 + 0.002 + 0.004 + 0.005 + 0.000 = 0.033

In this case, the forgery coefficient is small. There is no exact threshold for detecting forgery, but your program will be calculating this coefficient for each of the data files. When the coefficients are examined in relation to one another, the location of the forged data file will become apparent since the forgery coefficient will be abnormally high when the distribution of the first digits does not conform to the Benford probabilities. A sample run of the solution may yield results like this:

FILE NAME | FORGERY COEFFICIENT

--------------------+--------------------

P02\_DataFile\_0.dat | 0.033

P02\_DataFile\_1.dat | 0.015

P02\_DataFile\_2.dat | 0.023

P02\_DataFile\_3.dat | 0.024

P02\_DataFile\_4.dat | 0.020

P02\_DataFile\_5.dat | 0.029

P02\_DataFile\_6.dat | 0.020

P02\_DataFile\_7.dat | 0.337 🡸 FORGED DATA!!!

P02\_DataFile\_8.dat | 0.016

P02\_DataFile\_9.dat | 0.020

When your program is run, display the detailed statistics for each file as it is processed. After each file has been processed, display a table like the one above comparing the forgery coefficients so a reasonably intelligent human being can interpret the results and pinpoint the forged file.

Sample Run Output

10000 integer elements were encountered and analyzed in: P02\_DataFile\_0.dat

FREQUENCY TOTALS

FREQ\_01 = 3099, FREQ\_02 = 1708, FREQ\_03 = 1220, FREQ\_04 = 972

FREQ\_05 = 799, FREQ\_06 = 653, FREQ\_07 = 589, FREQ\_08 = 503

FREQ\_09 = 457

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.310, RELA\_02 = 0.171, RELA\_03 = 0.122, RELA\_04 = 0.097

RELA\_05 = 0.080, RELA\_06 = 0.065, RELA\_07 = 0.059, RELA\_08 = 0.050

RELA\_09 = 0.046

BENFORD OFFSET VALUES

BOFF\_01 = 0.009, BOFF\_02 = 0.005, BOFF\_03 = 0.003, BOFF\_04 = 0.000

BOFF\_05 = 0.001, BOFF\_06 = 0.002, BOFF\_07 = 0.001, BOFF\_08 = 0.001

BOFF\_09 = 0.000

The sum of the Benford Offsets is: 0.022

10000 integer elements were encountered and analyzed in: P02\_DataFile\_1.dat

FREQUENCY TOTALS

FREQ\_01 = 2994, FREQ\_02 = 1761, FREQ\_03 = 1233, FREQ\_04 = 1000

FREQ\_05 = 753, FREQ\_06 = 633, FREQ\_07 = 590, FREQ\_08 = 536

FREQ\_09 = 500

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.299, RELA\_02 = 0.176, RELA\_03 = 0.123, RELA\_04 = 0.100

RELA\_05 = 0.075, RELA\_06 = 0.063, RELA\_07 = 0.059, RELA\_08 = 0.054

RELA\_09 = 0.050

BENFORD OFFSET VALUES

BOFF\_01 = 0.002, BOFF\_02 = 0.000, BOFF\_03 = 0.002, BOFF\_04 = 0.003

BOFF\_05 = 0.004, BOFF\_06 = 0.004, BOFF\_07 = 0.001, BOFF\_08 = 0.003

BOFF\_09 = 0.004

The sum of the Benford Offsets is: 0.021

10000 integer elements were encountered and analyzed in: P02\_DataFile\_2.dat

FREQUENCY TOTALS

FREQ\_01 = 2975, FREQ\_02 = 1725, FREQ\_03 = 1238, FREQ\_04 = 993

FREQ\_05 = 802, FREQ\_06 = 660, FREQ\_07 = 603, FREQ\_08 = 523

FREQ\_09 = 481

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.297, RELA\_02 = 0.172, RELA\_03 = 0.124, RELA\_04 = 0.099

RELA\_05 = 0.080, RELA\_06 = 0.066, RELA\_07 = 0.060, RELA\_08 = 0.052

RELA\_09 = 0.048

BENFORD OFFSET VALUES

BOFF\_01 = 0.004, BOFF\_02 = 0.004, BOFF\_03 = 0.001, BOFF\_04 = 0.002

BOFF\_05 = 0.001, BOFF\_06 = 0.001, BOFF\_07 = 0.002, BOFF\_08 = 0.001

BOFF\_09 = 0.002

The sum of the Benford Offsets is: 0.018

10000 integer elements were encountered and analyzed in: P02\_DataFile\_3.dat

FREQUENCY TOTALS

FREQ\_01 = 2998, FREQ\_02 = 1797, FREQ\_03 = 1307, FREQ\_04 = 947

FREQ\_05 = 803, FREQ\_06 = 635, FREQ\_07 = 562, FREQ\_08 = 500

FREQ\_09 = 451

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.300, RELA\_02 = 0.180, RELA\_03 = 0.131, RELA\_04 = 0.095

RELA\_05 = 0.080, RELA\_06 = 0.064, RELA\_07 = 0.056, RELA\_08 = 0.050

RELA\_09 = 0.045

BENFORD OFFSET VALUES

BOFF\_01 = 0.001, BOFF\_02 = 0.004, BOFF\_03 = 0.006, BOFF\_04 = 0.002

BOFF\_05 = 0.001, BOFF\_06 = 0.004, BOFF\_07 = 0.002, BOFF\_08 = 0.001

BOFF\_09 = 0.001

The sum of the Benford Offsets is: 0.021

10000 integer elements were encountered and analyzed in: P02\_DataFile\_4.dat

FREQUENCY TOTALS

FREQ\_01 = 3018, FREQ\_02 = 1730, FREQ\_03 = 1241, FREQ\_04 = 936

FREQ\_05 = 805, FREQ\_06 = 702, FREQ\_07 = 594, FREQ\_08 = 528

FREQ\_09 = 446

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.302, RELA\_02 = 0.173, RELA\_03 = 0.124, RELA\_04 = 0.094

RELA\_05 = 0.081, RELA\_06 = 0.070, RELA\_07 = 0.059, RELA\_08 = 0.053

RELA\_09 = 0.045

BENFORD OFFSET VALUES

BOFF\_01 = 0.001, BOFF\_02 = 0.003, BOFF\_03 = 0.001, BOFF\_04 = 0.003

BOFF\_05 = 0.002, BOFF\_06 = 0.003, BOFF\_07 = 0.001, BOFF\_08 = 0.002

BOFF\_09 = 0.001

The sum of the Benford Offsets is: 0.017

10000 integer elements were encountered and analyzed in: P02\_DataFile\_5.dat

FREQUENCY TOTALS

FREQ\_01 = 3113, FREQ\_02 = 1693, FREQ\_03 = 1238, FREQ\_04 = 946

FREQ\_05 = 816, FREQ\_06 = 666, FREQ\_07 = 533, FREQ\_08 = 531

FREQ\_09 = 464

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.311, RELA\_02 = 0.169, RELA\_03 = 0.124, RELA\_04 = 0.095

RELA\_05 = 0.082, RELA\_06 = 0.067, RELA\_07 = 0.053, RELA\_08 = 0.053

RELA\_09 = 0.046

BENFORD OFFSET VALUES

BOFF\_01 = 0.010, BOFF\_02 = 0.007, BOFF\_03 = 0.001, BOFF\_04 = 0.002

BOFF\_05 = 0.003, BOFF\_06 = 0.000, BOFF\_07 = 0.005, BOFF\_08 = 0.002

BOFF\_09 = 0.000

The sum of the Benford Offsets is: 0.031

10000 integer elements were encountered and analyzed in: P02\_DataFile\_6.dat

FREQUENCY TOTALS

FREQ\_01 = 2966, FREQ\_02 = 1741, FREQ\_03 = 1263, FREQ\_04 = 988

FREQ\_05 = 825, FREQ\_06 = 654, FREQ\_07 = 579, FREQ\_08 = 528

FREQ\_09 = 456

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.297, RELA\_02 = 0.174, RELA\_03 = 0.126, RELA\_04 = 0.099

RELA\_05 = 0.083, RELA\_06 = 0.065, RELA\_07 = 0.058, RELA\_08 = 0.053

RELA\_09 = 0.046

BENFORD OFFSET VALUES

BOFF\_01 = 0.004, BOFF\_02 = 0.002, BOFF\_03 = 0.001, BOFF\_04 = 0.002

BOFF\_05 = 0.004, BOFF\_06 = 0.002, BOFF\_07 = 0.000, BOFF\_08 = 0.002

BOFF\_09 = 0.000

The sum of the Benford Offsets is: 0.017

10000 integer elements were encountered and analyzed in: P02\_DataFile\_7.dat

FREQUENCY TOTALS

FREQ\_01 = 2999, FREQ\_02 = 1762, FREQ\_03 = 1226, FREQ\_04 = 969

FREQ\_05 = 794, FREQ\_06 = 641, FREQ\_07 = 623, FREQ\_08 = 510

FREQ\_09 = 476

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.300, RELA\_02 = 0.176, RELA\_03 = 0.123, RELA\_04 = 0.097

RELA\_05 = 0.079, RELA\_06 = 0.064, RELA\_07 = 0.062, RELA\_08 = 0.051

RELA\_09 = 0.048

BENFORD OFFSET VALUES

BOFF\_01 = 0.001, BOFF\_02 = 0.000, BOFF\_03 = 0.002, BOFF\_04 = 0.000

BOFF\_05 = 0.000, BOFF\_06 = 0.003, BOFF\_07 = 0.004, BOFF\_08 = 0.000

BOFF\_09 = 0.002

The sum of the Benford Offsets is: 0.013

10000 integer elements were encountered and analyzed in: P02\_DataFile\_8.dat

FREQUENCY TOTALS

FREQ\_01 = 1939, FREQ\_02 = 1997, FREQ\_03 = 1065, FREQ\_04 = 1015

FREQ\_05 = 503, FREQ\_06 = 473, FREQ\_07 = 1010, FREQ\_08 = 1998

FREQ\_09 = 0

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.194, RELA\_02 = 0.200, RELA\_03 = 0.107, RELA\_04 = 0.102

RELA\_05 = 0.050, RELA\_06 = 0.047, RELA\_07 = 0.101, RELA\_08 = 0.200

RELA\_09 = 0.000

BENFORD OFFSET VALUES

BOFF\_01 = 0.107, BOFF\_02 = 0.024, BOFF\_03 = 0.019, BOFF\_04 = 0.005

BOFF\_05 = 0.029, BOFF\_06 = 0.020, BOFF\_07 = 0.043, BOFF\_08 = 0.149

BOFF\_09 = 0.046

The sum of the Benford Offsets is: 0.440

10000 integer elements were encountered and analyzed in: P02\_DataFile\_9.dat

FREQUENCY TOTALS

FREQ\_01 = 3009, FREQ\_02 = 1745, FREQ\_03 = 1260, FREQ\_04 = 981

FREQ\_05 = 738, FREQ\_06 = 711, FREQ\_07 = 589, FREQ\_08 = 514

FREQ\_09 = 453

RELATIVE FREQUENCY TOTALS

RELA\_01 = 0.301, RELA\_02 = 0.174, RELA\_03 = 0.126, RELA\_04 = 0.098

RELA\_05 = 0.074, RELA\_06 = 0.071, RELA\_07 = 0.059, RELA\_08 = 0.051

RELA\_09 = 0.045

BENFORD OFFSET VALUES

BOFF\_01 = 0.000, BOFF\_02 = 0.002, BOFF\_03 = 0.001, BOFF\_04 = 0.001

BOFF\_05 = 0.005, BOFF\_06 = 0.004, BOFF\_07 = 0.001, BOFF\_08 = 0.000

BOFF\_09 = 0.001

The sum of the Benford Offsets is: 0.015

FILE NAME | FORGERY COEFFICIENT

--------------------+---------------------

P02\_DataFile\_0.dat | 0.022

P02\_DataFile\_1.dat | 0.021

P02\_DataFile\_2.dat | 0.018

P02\_DataFile\_3.dat | 0.021

P02\_DataFile\_4.dat | 0.017

P02\_DataFile\_5.dat | 0.031

P02\_DataFile\_6.dat | 0.017

P02\_DataFile\_7.dat | 0.013

P02\_DataFile\_8.dat | 0.440

P02\_DataFile\_9.dat | 0.015

Assessment

This assignment will be assessed based on the provided grading rubric.