Python for Tool Developers

Lab 2

11/11/2016

Lab 2. Advanced concepts in Python

Benford's Law, also known as the *first digit law* is an empirical observation that the frequency distribution of digits in naturally occuring sets of numbers isn't uniform. The law states that approximately 30 percent of the numbers in the set will have a '1' as their first number. The number '9' only appears about 5 per cent of the time as the most significant digit. One of the applications that uses this law is the detecting of money laundering.

One technique used by money launderers is to conduct transactions of less then 10000.00 (in currency amounts, i.e. USD, GBP or Euro). Therefore, if we see that we have a list of transactions that doesn't fit the expected output from Benford's Law, then we can raise an alert.

The formula for calculating the expected results of complying with Benford's Law is as follows:

$$Nlog_{10}(1+\frac{1}{d})$$

Where *N* is the number of elements in the data set and *d* is set of digits (i.e. one through ten).

How will we know if the observed data set fits the expected data set? Or to put it another way, are the transaction amounts suspicious enough to raise an alert for possible laundering of money?

We can apply a test called the *Pearson Chi Square* test. The formula for this is test is

$$X^2=\sum_{i=1}^{n}(\frac{(O_i-E_i)^2}{E_i})$$

We can then look this chi squared value up in a table and see if the result is greater than or less than the expected result. If the result is greater than the value in the chart, then we can raise an alert.

Exercise 1. You will be given two data sets, trans1.csv and trans2.csv

1. Calculate the expected distribution values for each data set.

2. Calculate the actual distribution. (Hint, use the counter module from the Python collections library).

3. Calculate the chi square value.

4. Print out the following:

The expected distribution

The observed distribution

The chi square value.