FACULTY FORUM

You Cannot Conceive The Many Without The One
-Plato-

Issue No. 21, Spring 2015 April 29, 2015

The Effective Use of Benford's Law to Assist in Detecting Fraud in U.S. Environmental Protection Agency (EPA) Toxics Release Inventory (TRI) Data

By
JEFF IRWIN
Professor of Accounting

Abstract

This paper analyzes data regarding toxic chemicals released into surface bodies of water based on Benford's law, an empirical law that describes the distribution of leading digits in a collection of numbers met in naturally occurring phenomena. The law is based on observations that certain digits appear more frequently than others in data sets. After discussing the background of the law and the development of its use in natural sciences, this paper analyzes how Benford's law can be applied to U.S. Environmental Protection Agency (EPA) Toxics Release Inventory (TRI) data. The theory advanced is that any type of deviation affecting TRI data, including fraud and data manipulation, can be detected by investigating the first-digit distributions of the TRI data. This premise is then supported by corroborative statistical tests that achieve encouraging results.

I. Introduction

Benford's law posits that certain digits appear more frequently than others in data sets. It has been used as an empirical law that describes the distribution of leading digits of a collection of numbers met in naturally occurring phenomena such as the drainage areas of rivers, stock market prices, census data, and the heat capacities of chemicals (Benford, 1938). Though experimental at the beginning, it is now established that it holds for various mathematical series as well (Wlodarski, 1971).

Benford's law has found its applications in natural sciences. Sambridge et al. tested its compliance on various geophysics data sets such as the length of time between geomagnetic reversals, depths of earthquakes, models of Earth's gravity, and geomagnetic and seismic structure, as well as other natural-science observables, such as the rotation frequencies of pulsars, greenhouse-gas emissions, and the masses of exoplanets (Sambridge et al., 2010).

The assumption is that Benford's law applies to the amounts of toxic chemicals discharged into surface bodies of water as reported by the U.S. Environmental Protection Agency (EPA). The experiments described in the succeeding sections clarify that such an assumption is in fact reasonable; i.e., Benford's law holds on Toxics Release Inventory (TRI) data. However, while there is remarkable conformity to Benford's law, this analysis uncovered deviations from Benford's law that were not systematic. Natural-science data should follow Benford's law and this nonconformity to Benford could be indicators of (a) an incomplete data set, (b) the sample

not being representative of the population, (c) excessive rounding of the data, or (d) data manipulation, fraud, or data errors.

The remainder of this paper is as follows. Section II gives a general overview and background on Benford's law and how it is used as a data-manipulation-detection concept. Section III provides background on Toxics Release Inventory (TRI) data and data-integrity methods currently employed by the EPA. Section IV outlines the data used, and section V describes the methods used in the experiment and the results of data analysis. Section VI summarizes the results and provides insights into further testing methods and research.

II. Background on Benford's Law

According to Benford's *law of anomalous numbers* (Benford, 1938) the frequency of the digit *d*, appearing as the first significant digit in a collection of numbers, is not uniform as expected intuitively. Instead, it follows closely the logarithmic relation:

$$P_d = \log_{10}\frac{(d+1)}{d}$$

Using this formula, the probability of the first digit being one is about 30 percent while the probability of the first digit being nine is only 4.6 percent. Table 1 shows the expected frequencies for all digits 1 through 9 for the leftmost or first-place integer in any number.

Table 1

Leftmost Integer	Min. Benfords Law
9	4.58%
8	5.12%
7	5.80%
6	6.69%
5	7.92%
4	9.69%
3	12.49%
2	17.61%
1	30.10%

Source: Nigrini, 1996

While this law may seem surprising at first, there are several references in literature explaining and justifying the law as well as defining conditions for data sets that do or do not follow the law.

Pinkham (1961) argued that if there is going to be a universal law expressing the frequency of the first digit of numbers, it should be invariant under scale change of the underlying distribution. He then proved that the only scale-invariant distribution for first significant digits is the logarithmic distribution.

Furthermore, Boyle (1994) shows that 1) the log distribution is the limiting distribution when random variables are repeatedly multiplied, divided, or raised to integer powers, and 2) once achieved, the log distribution persists under all further multiplications, divisions, and raisings to integer powers.

Hill (1995) and Hill (1998) present a rule for the first-digit frequency of numbers in bases other than 10, and show that Benford's law is the only *base-invariant* distribution for first-digit frequencies. Hill also generalized Benford's law for all significant digits in a number and presented a new statistical interpretation of this generalized law. Accordingly, he proved that if distributions are selected at random (in any unbiased way) and random samples are then taken from each of these distributions, the significant digits of the combined sample will converge to the logarithmic distribution. Based on the statistical formulation of Hill, other researchers began to study conditions for the distributions to satisfy Benford's law.

III. What is Toxics Release Inventory (TRI) Data?

The EPA tracks the management of certain toxic chemicals that may pose a threat to human health and the environment. U.S. facilities in different industry sectors must report annually how much of each chemical is released to the environment and/or managed through recycling, energy recovery, and treatment. A "release" of a chemical means that it is emitted to the air or water, or placed in some type of land disposal.

The information submitted by facilities is compiled in the Toxics Release Inventory (TRI). TRI helps support informed decision-making by industry, government, non-governmental organizations, and the public.

The EPA works continuously to ensure that TRI data are accurate and reliable. Steps taken to promote data quality include analyzing data for potential errors, contacting TRI facilities concerning potentially inaccurate submissions, providing guidance on reporting requirements, and, as necessary, taking enforcement actions against facilities that fail to comply with TRI requirements.

The EPA conducts an extensive data-quality analysis after TRI reporting forms are received. It first identifies TRI forms containing potential errors, then EPA staff contacts the facilities that submitted these reports to discuss the potential errors. If errors are found, the facilities then should submit a correct report to EPA and the appropriate state or tribe.

The EPA conducts many different analyses to identify errors in TRI reports. Examples of these analyses include: (a) Facilities that reported a large change in disposal or other release and/or other waste management quantities for certain chemicals of concern (with a focus on air and water releases); (b) Facilities that have potential errors in reporting dioxin and dioxin-like compounds; (c) Facilities that transmitted but failed to certify their reports; (d) Facilities that reported large quantities of volatile organic chemicals on-site but reported small quantities of air releases; (e) Facilities that reported the same quantities on multiple sections of the reporting Form R for more than 2 years; and (f) Facilities that reported significantly different data to other EPA programs.

These efforts help to ensure the quality and accuracy of the TRI data and of the annual National Analysis report, and makes TRI a more reliable starting point for understanding how communities and the environment may be exposed to toxic chemicals.

The United States Code authorizes civil and administrative penalties for noncompliance with TRI reporting requirements. Section 1101 of Title 18 of the U.S. Code makes it a criminal offense to falsify information given to the United States government (including intentionally false records maintained for inspection). The knowing failure to file an EPCRA Section 313 report may be prosecuted as concealment under the same section.

While the EPA uses many techniques for data-quality analysis with the authority to penalize violators, it does not utilize a Benford's law technique to audit the self-reported figures. This paper proposes that Benford's law can be used by the EPA Office of Inspector General to prevent and detect fraud, waste, and abuse.

IV. Data Analyzed

Data for the years 1987–2012 for 496 sites throughout South Carolina was downloaded from the EPA Envirofocts website (http://www.epa.gov/enviro/index.html). The data is specific for toxic chemicals released to surface bodies of water in SC. TRI data is recorded in pounds (lbs) of toxic chemicals discharged. Figure 1.1 shows the SC waterways affected by toxic chemicals.

Figure 1.1

Source: geology.com

South Carolina Surface Bodies of Water: Ashley River, Black River, Broad River, Catawba River, Cooper River, Edisto River, Enoree River, Great Pee Dee River, Little Pee Dee River, Lynches River, North Fork Edisto River, Pacolet River, Salkahatchie River, Saluda River, Santee River, Savannah River, South Fork Edisto River, Waccamaw River, Hartwell Reservoir, J. Strom Thurmond Lake, Lake Greenwood, Lake Jocassee, Lake Keowee, Lake Marion, Lake Moultrie, Lake Murray, Richard B. Russell Lake, and Wateree Lake

Null and zero pound records were not included in the analysis. A review of the 1202 null and zero pound records found that all years had more than one zero or null value. The null and zero entries are probably not data errors, but cases where the site is required to submit TRI data but had no data to report.

The number of usable records after the removal of the null values and zeros totaled 7390. The data set is particularly interesting because (a) the period covered is fifteen years and it is rare for a data set to cover such an extended period; (b) the number of records is relatively large compared with other data analyzed in Benford's law literature; (c) the range 1 – 87,400,000.00 shows that the sites covered everything from the smallest release to the largest emission into SC waterways; (d) the EPCRA Section 313 reports have been the same over the entire 15 years of reporting, which means that there are no distortions due to technical changes; and (e) the data is used for many important purposes, which means that data integrity is an important issue.

Table 1.1

TABLE 1.1	Descriptive Statistics of the Toxic Release Inventory Data		
Description	**Number**	**Unit**	
Records downloaded from Envirofacts EPA Website	8592		
Records equal to zero	125		
Records less than 1	0		
Null (blank) records	1077		
Valid usable records	7390		
Statistics for the Valid Records			
Duplicate records	0	records	
Sites with duplicate records	0	sites	
Number of records after deletion of duplicates	7390	records	
Number of unique sites	496	sites	
Latest year on record for any site	2012	calendar year	
Earliest year on record for any site	1987	calendar year	
Year with the highest record count	1987	calendar year	
Year with the lowest record count	2003	calendar year	
Minimum TRI for any single site year	1	pounds	
Quartile 1	13	pounds	
Quartile 2 (median)	171	pounds	
Quartile 3	1744	pounds	
Maximum TRI for any single site	87400000	pounds	
AverageTRI over all records	40179	pounds	

V. Data Analysis and Results

The digits of a large collection of TRI data over an extended period of time showed a remarkable conformity to Benford's law. This analysis demonstrates the use of the Chi-square goodness of fit (GOF) test to assess whether the deviations from Benford's law were systematic.

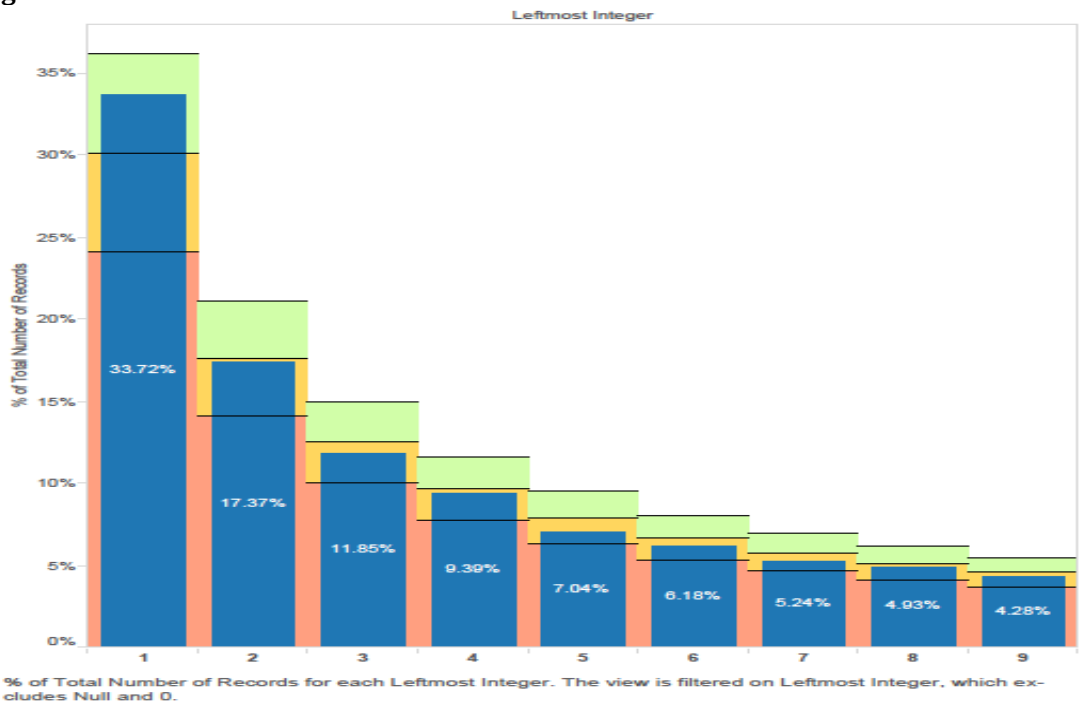
Table 1.2 shows the occurrence of the leftmost or first digit integer compared to expected Benford's law percentages. For example, the digit two was expected to occur 17.61 percent, or 1301 times out of the sample of 7390. The actual data shows the digit two was observed 1284 times or 17.37 percent of the time.

Table 1.2

Leftmost Integer	Min. Benfords Law	% of Total Number of Records
9	4.58%	4.28%
8	5.12%	4.93%
7	5.80%	5.24%
6	6.69%	6.18%
5	7.92%	7.04%
4	9.69%	9.39%
3	12.49%	11.85%
2	17.61%	17.37%
1	30.10%	33.72%

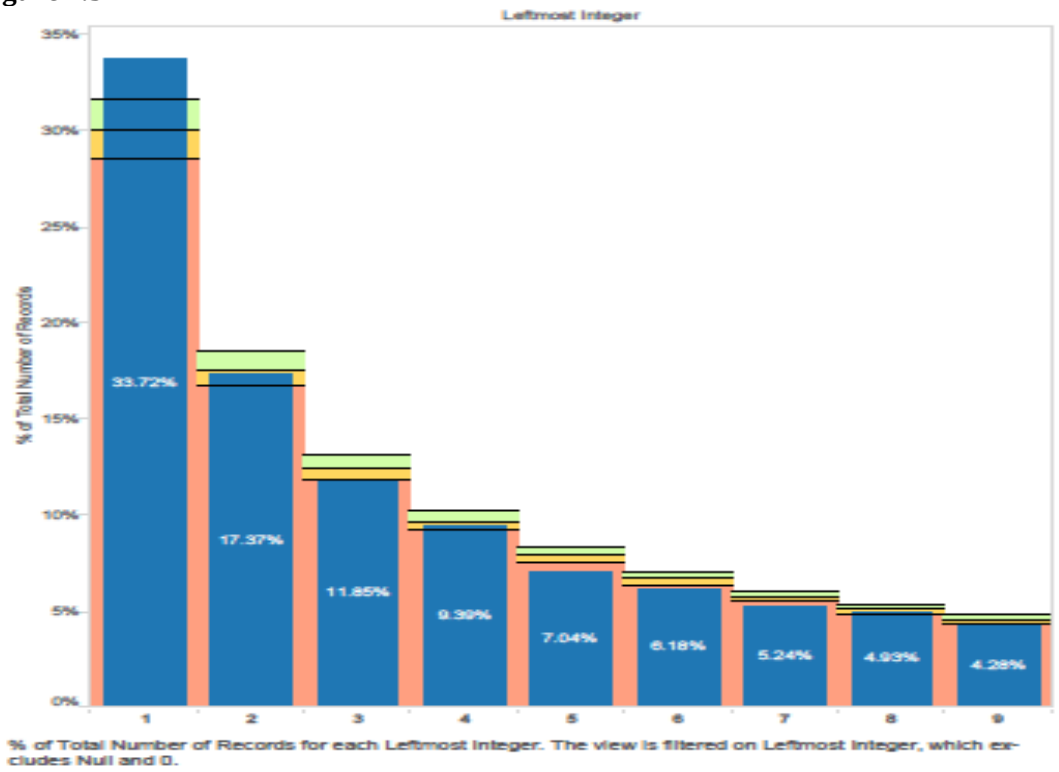
Figure 1.2 shows Benford's law as a decreasing line based on the Benford proportions, which range from a high of 2226 down to a low of 338, with actual counts shown as vertical bars. The Tableau graph includes upper and lower limits at 20% above and below the expected. This supports conformity to Benford's law.

Figure 1.2



Actual counts that exceed the upper limit or that are less than the lower limit are significant at 5 percent above or below the expected counts and are shown in Figure 1.3. The middle line represents the Benford's expected count while the upper and lower lines represent 5 percent deviations from the expected count. The blue bar represents the actual count. For example, the first count, the integer one, is significantly higher than the expected count.

Figure 1.3



% of Total Number of Records for each Leftmost Integer. The view is filtered on Leftmost Integer, which excludes Null and 0.

As in any statistical test, digital analysis compares the actual numbers of items observed to the expected and calculates the deviation. In a Benford distribution, for example, the expected proportion of numbers that feature the integer one in the first position is 30.10 percent. The actual proportion observed will most likely deviate from this expected amount due to random variation. While no data set can be expected to conform precisely, at what point is the deviation considered large enough to be significant? A Chi-squared test of goodness of fit (GOF) can be performed. The Chi-square test combines the results of testing each digit's expected frequency with actual frequency into one test statistic that indicates the probability of finding the result.

Table 1.3

	Gp 1	Gp 2	Gp 3	Gp 4	Gp 5	Gp 6	Gp 7	Gp 8	Gp 9	Gp 10	
Observed:	2492	1284	876	694	520	457	387	364	316		7390
Expected:	2226	1301	923	716	585	494	429	378	338		7390

Output:

Calculate | Reset all

Chi-square:	51.133
degrees of freedom:	8
p-value:	2e-8
Yates' chi-square:	50.537
Yates' *p*-value:	3e-8

Status: Status okay

http://quantpsy.org

A Chi-square test for goodness of fit (GOF) and association between two categorized variables was performed to examine the association between the expected counts according to Benford's law and the actual counts observed. In Table 1.3 the observed data counts are given in the first row and the expected counts are given in the second row. The Chi-square value of 51.13 is very high. The Chi-square test is significant and this means that the observed values are significantly different from the expected values. There is less than a 5 percent chance that the deviation of Group 1 is due to chance. For every group except the first, the observed value was lower than the expected. For the first group, the observed value was much greater than the expected.

VI. Conclusion

The close conformity to Benford's law with a high Chi-square due to a deviation in Group 1 makes this data a good candidate for further testing. The fact that Group 1 was quite a bit larger than expected leads to the conclusion that there might be reason why one might record a 1999 instead of a 2000 or 19,999 instead of 20,000 on the TRI report. More investigation of TRI reporting thresholds would need to be completed. While Benford analysis by itself might not be a conclusive indication of fraud, it can be a useful tool to help identify data for further testing and therefore should assist auditors such as the EPA Office of Inspector General in preventing and detecting fraud, waste, and abuse.

REFERENCES

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572.

Boyle, J. (1994). An application of Fourier series to the most significant digit problem. *American Mathematical Monthly*, 101(9), 879–886.

Hill, T.P. (1995). A statistical derivation of the significant digit law. *Statistical Science,* 10(4),354–363.

Hill T.P. (1998). The first digit phenomenon. *American Scientist*, 86(4),358–363.

Nigrini M. (1996). A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association*, 18 (1) (1996), 72–91.

Pinkham R.S. (1961). On the distribution of first significant digits. *The Annuals of Mathematical Statistics*, 32 (4) (1961), 1223–1230.

Preacher, K. J. (2001, April). *Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence.* Retrieved from http://www.quantpsy.org/chisq/chisq.htm

Sambridge, H., Tkalčić, H., & Jackson, A., 2010. Benford's Law in the natural sciences. *Geophysical Research Letters*, 37 (2010), L22301.

U.S. Environmental Protection Agency (2014) *Toxics Release Inventory TRI Program TRI Data Quality*. Washington, DC: Environmental Protection Agency.

U.S. Environmental Protection Agency (2014) *Toxics Release Inventory TRI Program TRI Compliance and Enforcement*. Washington, DC: Environmental Protection Agency.

U.S. Environmental Protection Agency (2014) *Toxics Release Inventory TRI Program Learn about Toxic release inventory*. Washington, DC: Environmental Protection Agency.

Wlodarski, J. (1971). Fibonacci and Lucas numbers tend to obey Benford's Law. *Fibonacci Quarterly,* 9(1)(1971), 87–88.

Professor Jeff Irwin is a forensic accountant and Certified Fraud Examiner (CFE) whose research focuses on counter-fraud analytics using statistical and predictive modeling techniques. He received his Bachelor of Arts degree in Economics. Through the University of South Carolina's Moore School of Business International MBA program he completed his MSc in International Management at the Vienna University of Economics and Business in Vienna, Austria, and completed his International MBA with a concentration in accounting at the University of South Carolina in Columbia. Prior to joining the USC Salkehatchie faculty in 2013, Professor Irwin worked in finance and audit roles for JP Morgan Chase Bank and GE Capital.

FACULTY FORUM
IS A NEWSLETTER PUBLISHED
ELECTRONICALLY ON OUR WEBSITE AT
http://uscsalkehatchie.sc.edu/
AND IN HARD COPY
BY THE UNIVERSITY OF SOUTH CAROLINA
SALKEHATCHIE CAMPUS
807 HAMPTON STREET (P.O. Box 1337)
WALTERBORO, SOUTH CAROLINA 29488

C. Bryan Love, Ph.D.
EDITOR-IN-CHIEF

David Hatch, Ph.D.
EDITOR