Benford's Law, data mining, and financial fraud: a case study in New York State Medicaid data

B. Little[1], R. Rejesus[2], M. Schucking[3] & R. Harris[4]
[1]*Department of Mathematics, Physics, and Engineering, Texas Data Mining Research Institute and Centre for Agribusiness Excellence, Tarleton State University, Stephenville, Texas, USA*
[2]*Department of Agricultural and Resource Economics, North Carolina State University, Raleigh, North Carolina, USA*
[3]*Data Mining And Information Sciences Division, Qinetiq North America - Planning Systems Inc, Stephenville, Texas, USA*
[4]*New York Comptroller's Office, Albany, New York, USA*

Abstract

Benford's Law was first described by an astronomer in 1881, but physicist Frank Benford lent his name to the property in a mathematical treatise published in 1938. Behaviour of numbers described by the Law defies intuition, demonstrating that one is the most frequent (30.1%), and nine is the least frequent (4.6%). The property holds for a wide variety of numbers, including but not limited to: stock indices, river lengths, road numbers, etc. Departures from the classic Benford distribution are linked to anomalies, specifically in financial data where the property has been successfully employed in financial audits. The limitation of Benford's Law is that it identifies a relatively large pool of "candidate" anomalies that must be manually evaluated. In the present analysis of Medicaid data, multivariate cluster analysis in multiple tandem analyses is used to winnow the number of anomalies to a pool of high probability anomalies for evaluation. This approach makes the application of Benford's Law more practical.

Keywords: Benford's Law, cluster analysis, ensemble multivariate technique.

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080191

1 Benford's Law

In lists of numbers from almost any source, the leading digit is 1 approximately 30% of the time, with progressively decreasing frequency until 9 as the leading digit occurs less than 5% of the time (Table 1). This property is termed Benford's Law, which is named for physicist Frank Benford who expounded on the property in 1938 [1]. However, the property was first noted by an astronomer, Simon Newcomb in 1881 [2]. The first mathematical treatment of Benford's Law was published in 1988 [3].

Table 1: Distribution of first digits according to Benford's Law.

Digit	Probability
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

A generalization that holds is that measurements in the practical world have a logarithmic distribution, and it follows that the logarithm of almost any given set of measurements has a uniform distribution. Although a counter-intuitive phenomenon, a wide variety of numbers conform to Benford's Law: phone bills, ledger entries, mileages from fleet vehicles, street addresses, stock prices, census numbers, death rates, distances between cities, mathematical constants, and processes described by power laws (Figure 1).

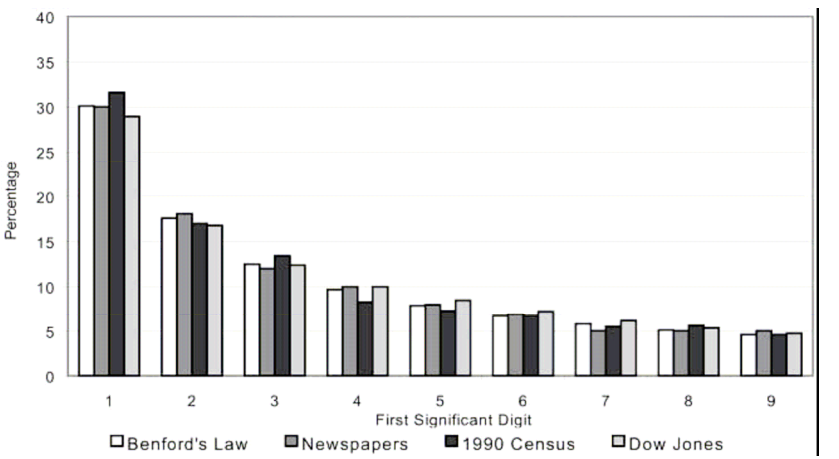

Figure 1: Distribution of first digits compared to Benford's Law.

An even less obvious property is that Benford's Law is true regardless of the base of the numbers, but the proportion of occurrence will of course differ. Benford's law states that the leading digit d where d is a member of the set {1, …, $b - 1$}, and base b ($b \geq 2$) occurs with probability proportional to:

$$\log_b (d + 1) - \log_b d = \log_b ((d + 1)/d).$$

A number has a given first significant digit d with probability Pr: Pr (first significant digit) = d = $\log_{10} (1 + d)^{-1}$ where d = 1,…., 9 [4]. Extension of probability to the general law is given by [5]:

$$\Pr(D_1 \cdots D_k = d_1 \cdots d_k) = \log_{10} (1 + (d_1 \ldots d_k)^{-1})$$

Thus, the probability of the first two significant digits in a distribution being 32 is: $P(D_1D_2) = 32 = \log_{10}(1 + (32)^{-1}) = 0.01336$ [6]. The first (non-zero) digit of the counts, lengths or distances of objects should have the same distribution whether the unit of measurement is inches, feet, yards, centimeters or meters. All existing or conceivable measurement scales will yield a logarithmic distribution and properties of logarithms (i.e., $\log_{10} (1) = 0$ and $\log_{10} (10) = 1$) results in a generalized Benford's law. For a distribution of initial digits the general property must apply to any set of data without regard to units of measure used, and that distribution of first digits fits the Benford Law. Therefore, for any specific distribution of first numbers complete independence of scale must hold (e.g., multiplication by a constant does not change the distribution and the only distribution for which this holds is a uniform logarithm distribution).

The objective in this investigation is to extend the Benford's Law to practical use to define a small set of highly anomalous observations.

2 Methods and materials

Data for three years of New York State Medicaid payments was provided by the Comptroller's Office to conduct a proof of concept for use of data mining to identify a small subset of anomalies in financial data. Analysts had no prior knowledge of the data. Tables were joined into single dataset, cleaned of anomalies and non-sense data values (e.g., negative values), de-duplicated, and data homogenized (e.g., subtotal rows were removed). In addition, spelling consistency checks were conducted and nulls dropped (names of cost centers and object codes had nulls). Finally, only values $\geq$ \$10.00 were included in the analysis. Analytical variables included: date paid, \$ amount, cost center name, and object code.

Benford's Law analyses were conducted using software by Nigrini [7] and Sherry Consulting (UK). Stepwise multi-stage cluster analyses were done using SPSS V.16 (SPSS, Inc., Chicago, Ill, 2007) and SAS v9.1 (SAS Institute, Cary, NC USA 2007).

3 Results

3.1 Benford's Law analysis

Following the data treatments described in Section 2, descriptive statistics for the data set (mean is greater than median, right skewed) indicate that the dataset is acceptable for a Benford's Law analysis because the basic moment conditions are satisfied (Table 2).

Table 2: Descriptive statistics for medicaid dataset.

N	Valid	60,969
	Missing	0
	Mean	14,337.896
	Median	1,093.780
	Mode	17.000
	Skewness	45.072
	Std. Error of Skewness	0.010
	Kurtosis	2944.776
	Std. Error of Kurtosis	0.020
	Sum	874,167,198.82

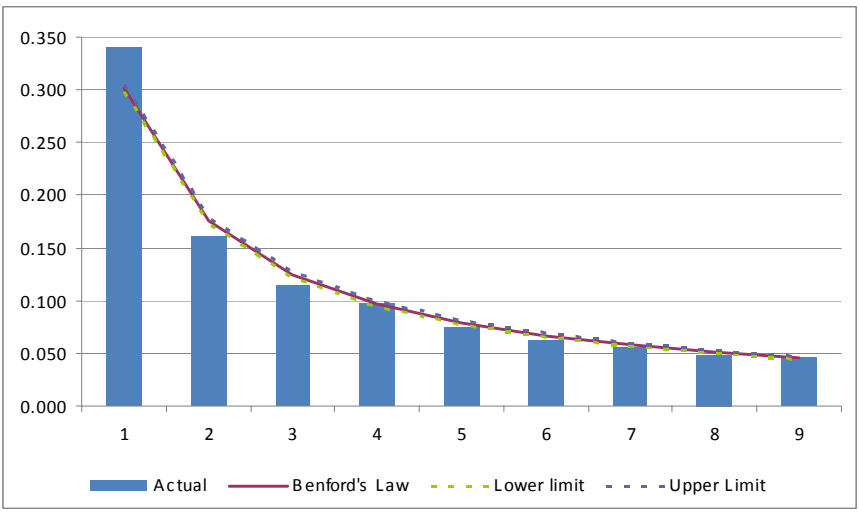

Figure 2: Distribution of first digits observed in analytical dataset.

Benford's analysis of the first digits indicate 1's occur more frequently than expected, 2's and 3's occur less frequently than predicted (Figure 2).

Among the second digits, there were too few 1's, an excess of 2's, and a deficit in the number of 3's (Figure 3).

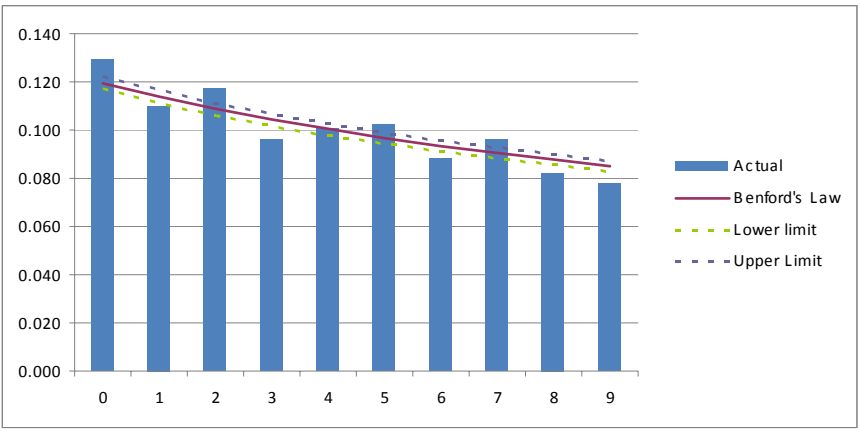

Figure 3: Distribution of second digits.

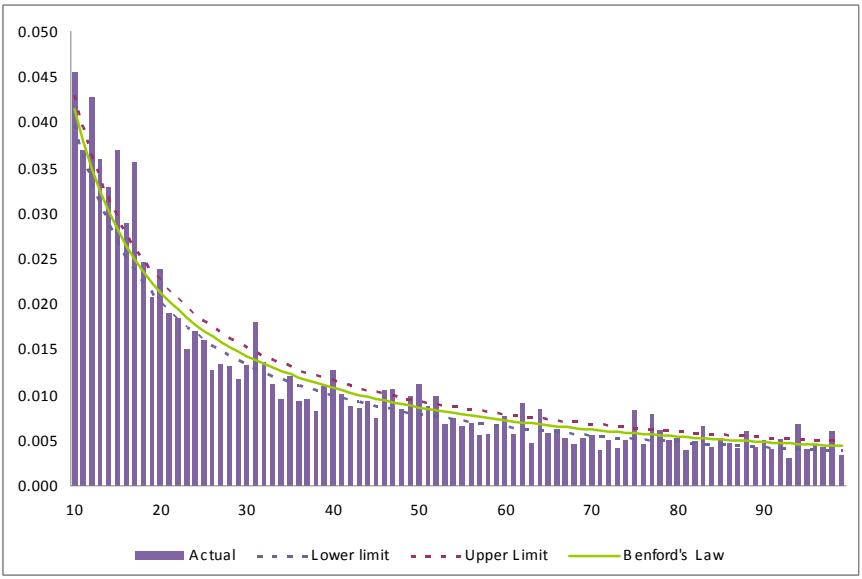

Figure 4: Distribution of first two digits.

Several "second digits" occurred more frequently than expected under the Benford distribution (10, 12, 13, 14, 15, 17, 20, 23) as shown in the spikes (Figure 4). Of these, 17 is the most anomalous occurrence (n = 2176, z-statistic = 17.233, an excess of 1.1%). The predicted number of anomalous rows is n=24 (0.011 * 2176). The problem is how to identify, among all the 17's, which ones are anomalous? Which 17's are the ones that occur normally as part of the distribution?

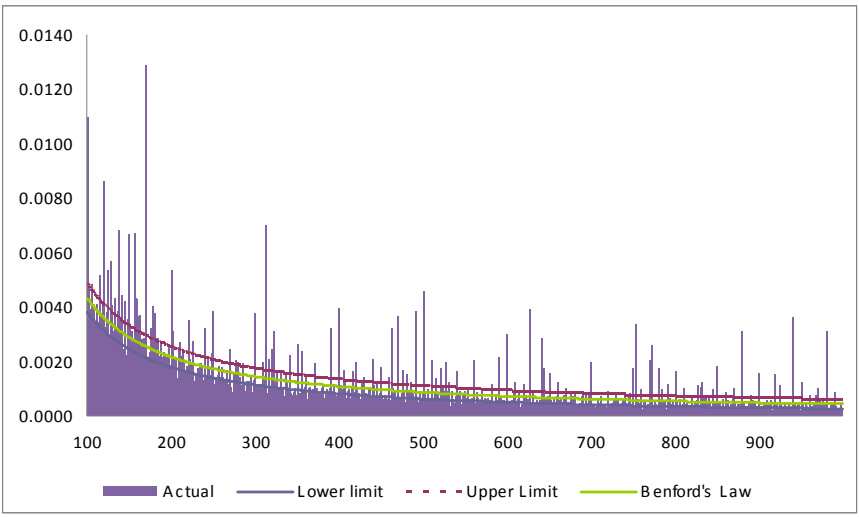

Figure 5: First three significant digits distribution.

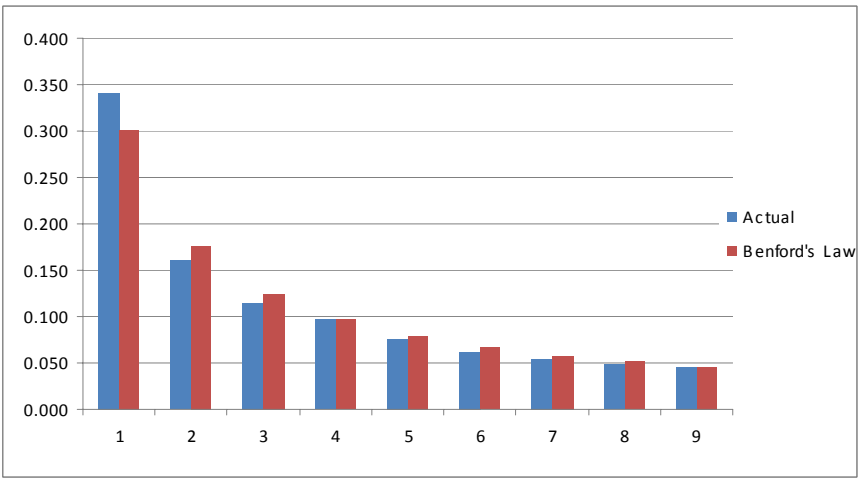

Figure 6: Distortion factor in Medicaid data.

Of the digits between 100 and 999 evaluated, values of 156 (n=409) and 170 (n=786) are most striking and highly significant (z-statistic = 18.42 and 50.64, respectively). In the present analysis, focus is on 170 because it has the highest z-statistic. For the 170's, there is an excess of 1.04% (expected = 0.25%, observed = 1.29%). The same problem remains – how to distinguish the anomalous 170's from those that are part of the expected distribution.

The Distortion Factor analysis of the whole dataset does not appear to be highly unusual (Figure 6), although some anomalous behavior was identified. The overall book of business in the Medicaid data analyzed in the present study

is not highly unusual, indicating that as a whole the transaction dataset is not egregious. Nonetheless, significant anomalies were detected in the dataset.

As observed earlier, the number of "candidate anomalies" with a Benford's Law analysis is usually large (i.e., includes all digits of the identified set or sequence) and does not provide a method for narrowing the number of candidates down to a reasonable list of suspect values. The next step in the traditional Benford's Law analysis is manual evaluation. In the present analysis, 2,176 different rows (Medicaid transactions) would necessarily be evaluated to fully utilize the list of anomalies identified by the Benford analysis.

3.2 Multivariate cluster analysis

Cluster analysis was chosen because it can analyze initial significant digits data, and other types. It can be used to analyze continuous and categorical data. The weakness of cluster analysis is that it will cluster ANYTHING – even non-sense. Therefore, cluster analysis results must be closely scrutinized.

The first stage cluster analysis begins at the top level with all observations that were included in the analysis, and results in two clusters and an outlier. An outlier cluster is one whose members are at least as distant from one another as they are from the two defined clusters. The outlier cluster contains 21% of the cases, which is an unusually high number of cases for an outlier of any variety, not just a cluster.

The outlier cluster was used as the "dataset" for further clustering because the anomalies that are the object of the analysis are contained among the outliers. Flags were created for the 17's and the 170's for analytical purposes, and the outlier designation was also retained as a flag.

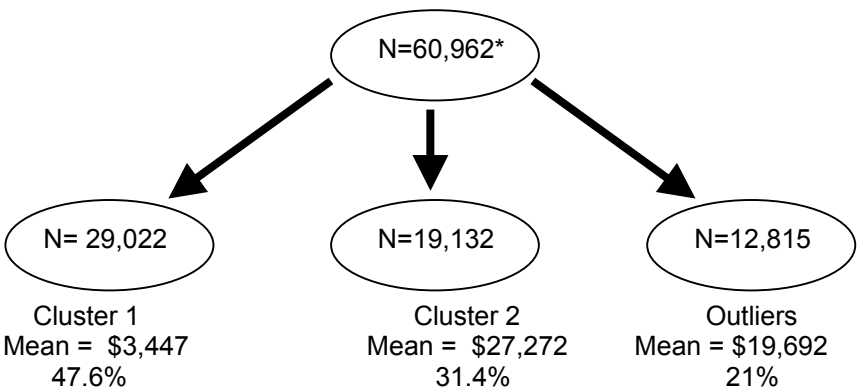

Figure 7: First cluster analysis.

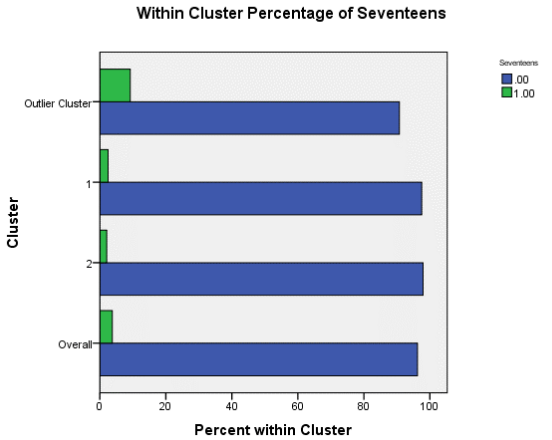

Figure 8: Within cluster occurrence of Benford anomalies.

As hypothesized, the Benford anomalies were concentrated in the outlier cluster, and continued clustering of the outlier cluster led to a small number of anomalies.

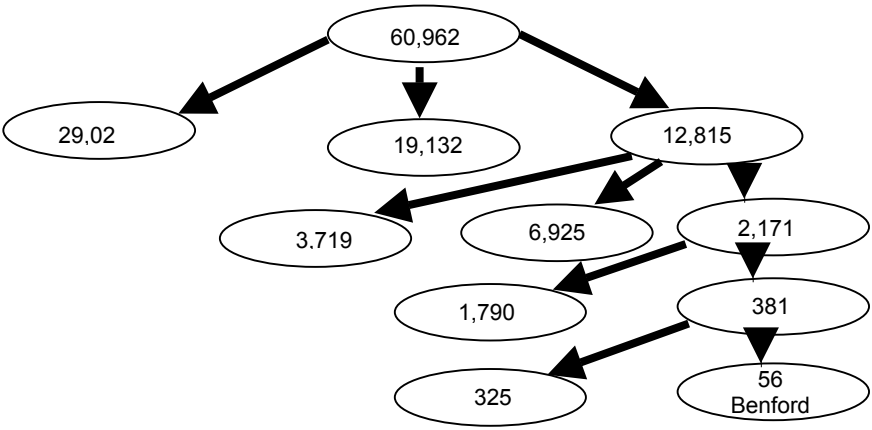

Figure 9: Overview of multi-cluster drill down to Benford anomalies to 56 candidates.

4 Discussion and conclusion

Benford's Law can be used in tandem with multivariate techniques to identify anomalous financial transactions. In this case cluster analysis was used, but other such scoring and distance related multivariate techniques could be used

also. Benford's law has been applied previously to large scale analyses of waste, fraud, and abuse [8]. However, the limitations were as discussed – the number of anomalies was too great to make the analysis of practical use.

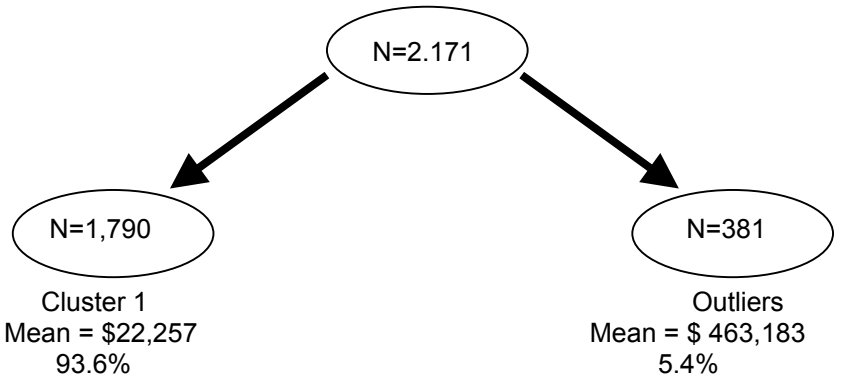

Figure 10: 381 anomalies of which 56 are Benford's.

In future applications, ensemble techniques that employ several analytical applications may be used to detect waste, fraud, and abuse. Ensemble techniques may include approaches such as Benford's Law, and use these findings in an integrated sequence of analyses to narrow down the number of suspect transactions or individuals to high probability, high value anomalies that can justify human evaluation of the anomalies. In this analysis, it was expected that 24 anomalies of 17's would be found, and 56 were identified.

In summary, this combination of anomaly detection techniques may add another tool to the methods available for analysis of large datasets for anomalous behaviour.

References

[1] Benford, Frank, "The law of anomalous numbers." *Proceedings of the American Philosophical Society* 78 (4): 551–572, 1938.
[2] Newcomb, Simon, "Note on the frequency of use of the different digits in natural numbers". *American Journal of Mathematics* 4 (1/4): 39–40, 1881.
[3] Hill, Theodore P. 1988 Theodore P. Hill (July–August 1998). "The first digit phenomenon". *American Scientist* 86: 358, 1988.
[4] Cohen, D., "An explanation of the first digit phenomenon," *J. Combin. Theory, Ser. A*, 20 (1976) 367–370, 1976.
[5] Hill, T.P., "Base-invariance implies Benford's law," *Proc. Amer. Math. Soc.*, 123:3 887–895, 1995.
[6] Geyer, C.L. and P.P. Williamson, 2004. Detecting fraud in data sets using Benford's Law. *Communications in Statistics* B 33, 229–246, 2004.
[7] Nigrini, M., "A taxpayer compliance application of Benford's law," *J. Amer. Taxation Assoc.*, 18: 72–91, 1996.

[8] Rejesus, R.M., B.B. Little, and M. Jaramillo, "Is there manipulation of yield data in crop insurance? An application of Benford's law. *J of Forensic Accounting* VII: 495–512, 2006.