Detecting Fraud In Financial Data Sets

Dominique Geyer, Audencia Nantes School of Management, France

ABSTRACT

An important neef of corporations for internal audits is the ability to detect fraudulently reported financial data. Benford's Law is a probability distribution which is useful to analyse patterns of digits in numbers sets. A history of the origins of Benford's Law is given and the types of data sets expected to follow Benford's Law is discussed. This paper examines how BA students falsify financial numbers. The paper shows that they fail to imitate Benford's law and that there are cheating behaviour patterns coherent with previous empirical studies.

Keywords: income statement manipulation, cheating behaviour and Benford's law

INTRODUCTION

In an article published in 1881, the mathematician and astronomer Simon Newcomb notes that the first pages of his logarithm table book are more worn than the others. He deduces that the searchers prefer to work on numbers starting with 1 rather than by 2; numbers starting with 2 being preferred with those starting with 3, etc. Intuitively this observation will appear strange insofar as one could think that there is a equiprobability of appearance of the various figures.

From this surprising discovery, the mathematician proposes the following formula indicating the probability that a number extract from a numbers set has C as first digit (C is an integer between 1 and 9): log10 [1 + (1/C)]. This discovery is passing unnoticed and it is only 57 years later that a physicist of General Electric, Franck Benford, makes the same observation as Newcomb (always with the logarithms tables). However, Benford will spend many years to collect data to validate this law. His article in 1938 registers twenty lists of numbers with 20,229 observations coming from varied sources, such as geographic, scientific and demographic data to test this law. Several empirical studies demonstrated the utility of this law to detect fraud (digital analysis).

The aim of this short paper is to examine how BA students falsify the loss of a balance sheet. The paper is organized into three sections. Section 1 describes Benford's law which is origin of digital analysis. Section 2 will present a synthesis of the empirical studies devoted to the application of this law in accounting. Section 3 presents the results of a laboratory study: a sample of 393 BA students translate a loss in a profit in a balance sheet.

1. THE BENFORD'S LAW

In a data set obeying Benford's law, approximately 30.1 % of numbers have 1 as first digit whereas this percentage falls to 4.6 % for the numbers having 9 as first digit. This law can be generalized with the second, third, etc digits. Once can formalize this law for numbers having two digits c1 c2; (for example, number 23 has two digits, the first digit c1 is 2 and the second digit c2 is 3); the generalization to N digits is immediate:

- probability of the event: the first digit of a number of a data set is c1 :
  $P(C1 = c1) = \log10 (1 + (1 / c1))$ with $c1 \in \{1;2;3;4;5;6;7;8;9\}$
- probability of the event: the second digit of a number of a data set is c2:

$$P(C_2 = c_2) = \sum_{c_1 =1}^{9} \log_{10} (1 + (1/ c_1 c_2)) \text{ with } c_2 \in \{0 ; 1; 2; 3; 4; 5; 6; 7; 8; 9\}$$

Thus the probability that a number of a data set obeying the Benford's law is 23 is log10 (1 + (1 / 23)) = 0.0184. For 3 digits, the formula becomes simply:

P(C1C2C3= c1c2c3) = log10 (1 + (1 / c1c2c3)). The following table shows the expected frequencies in the first fith positions.

Table 1: Benford's law: expected digital frequencies

	Position in Number				
Digit	**First**	**Second**	**Third**	**Fourth**	**Fifth**
0		0.11968	0.10178	0.10020	0.10000
1	0.30103	0.11389	0.10138	0.10010	0.10000
2	0.17609	0.10882	0.10097	0.10010	0.10000
3	0.12494	0.10433	0.10057	0.10010	0.10000
4	0.09691	0.10031	0.10018	0.10000	0.10000
5	0.07918	0.09668	0.09979	0.10000	0.10000
6	0.06695	0.09337	0.09940	0.09990	0.10000
7	0.05799	0.09035	0.09902	0.09990	0.10000
8	0.05115	0.08757	0.09864	0.09990	0.10000
9	0.04576	0.08500	0.09827	0.09980	0.10000

Note: the number 482 has 3 digits: 4 is the first digit, 8 the second and 2 the third. This table shows that under Benford's law the expected proportion of numbers with a first digit 4 is 9.69 % (8.75 % with 8 as second digit and 10.09 % with 2 as third digit).

The following example illustrates intuitively the Benford's law. Suppose that the size of a company is 10 000 employees the first year. This size grows 10 % each year. The first digit of the size's number will be one during eight years (one as first digit reappears in the 26th year, i.e. a size higher than 100 000 employees). Two as first digit appears four times. Nine appears only once for a size lower than 100 000 (25th year). This is an important property of the Benford's law. When the numbers of such data sets are ordered in an increasing way, they follow roughly a geometrical distribution (roughly because in a Benford's data set, two numbers can be identical). Three others conditions are necessary to have a Benford's data set:

- data set must constitute a homogeneous unit: populations of cities, surfaces of lake, value of shares, etc.
- data should not have of lower (except zero) or higher limit. Thus, for example, by studying the reimbursements of meal's expenses of a company, there will be strong probability that this data set do not obey a Benford's law because the firm will plan a upper limit for reimbursement.
- data should not be codified like the telephone numbers, the postal codes, the social security numbers, etc. It is obvious that such data set will not obey Benford's law.

Another fundamental property of the Benford's law is the scale invariance (Pinckam, 1961). In other words, if a data set is multiplied by a nonnull constant, the new data set will also obey the same law. Thus, if a data set of shares valued in euros obeys the Benford's law, this data set valued in dollars or yens will have the same property. This is a problem in cases of fraud by systematic under or overvaluation.

In 1993, in State of Arizona v. Xayne James Nelson, the accused was found guilty of trying to defraud the state of nearly 2 millions dollars.

Table 2: Check Fraud in Arizona

Date of Check	Amount in dollars
October 9, 1992	1,927.48
	27,902.31
October 14, 1992	86,241.90
	72,117.46
	81,321.75
	97,473.96
October 19, 1992	93,249.11
	89,658.17
	87,776.89
	92,105.83
	79,949.16
	87,602.93
	96,879.27
	91,806.47
	84,991.67
	90,831.83
	93,766.67
	88,338.72
	94,639,49
	83,709.28
	96,412.21
	88,432.86
	71,552.16
Total	**1,878,687.58**

Source : Nigrini M.J. (1999)

Several points are important:

- as is often the case in fraud, the embezzler started small and then increased dollar amount.
- the amounts of check fraud are lower than 100,000 dollars. Generally, there is a upper limit which requires an authorization. By not exceeding this limit, the evader does not want to draw attention;
- the frequency of ten digits is very different from those of Benford. More than 90 % of the amounts have 7, 8 or 9 as first digit.
- the manager repeats unconsciously certain digit's sequences. For the first two digits, 87, 88, 93 and 96 appear twice. For the last two digits, 16, 67 and 83 appear twice. There is a preference (comprehensible!) for high digits: 160 digits were used to draw the 23 cheques. From 0 to 9, the frequencies are respectively 7, 19, 16, 14, 12, 5, 17, 22, 22 and 26.

2. DIGITAL ANALYSIS AND FRAUD DETECTION

The empirical studies concerning Benford's law have often the same issue: insofar as a accounting data set follows the Benford's law, tests who shows significant variations between the observed frequencies and the theoretical frequencies can highlight fraud. The first application is due to Carslaw (1988). This author is interested in the second digit of positive earnings of a sample of New Zealand firms. He notes that for the second digit there is an excess of 0 and a lack of 9. The reason is simple: the managers will tend to round up the firm's profit in order to embellish the situation. Consider a profit of 4.98 millions euros. Rounding this number to 50 millions allows to reach a psychological influence whose importance will be greater whereas the second number is only marginally more important than the first.

This first study is followed by Thomas (1989) who studied U.S. firms. He studies Earnings before extraordinary items and discontinued operations at the quarterly and annual level. His study is finer because he distinguishes positive and negative earnings. He also notes an excess of 0 for the second digit of positive earnings.

In loss cases, one rounds down (less 0 and more 9) whereas in the profit level, one will round rather up. At the per share level, the author notes that multiples of 5 and 10 cents are observed considerably more often than others numbers.

The same rounding up behaviour have been observed by Finnish firms (Niskanen and Keloharju, 2000), English firms (Van Caneghem, 2002) and firms in 18 countries (Kinnumen and Koskela, 2003).

Note that this phenomenon didn't appear in a sample of U.S. public firms extracted from the Thompson One Banker database. The analysis includes data for positive earnings (i.e. net income available to common) for U.S. public active companies. Data concern the 2008 year period. Firms for which positive earning is composed of a single digit are eliminated because the tests concern the first and the second digit. So the sample concerns 3,379 public U.S. firms.

Table 3: first digit frequencies of positive earnings of 3,379 public U.S. firms

Digit	Expected frequency	Observed frequency	Bias	Z value	P- level
1	0.3140	0.3010	-	1.6434	0.1003
2	0.1758	0.1761	+	-0.0455	0.9637
3	0.1267	0.1249	-	0.3032	0.7617
4	0.0991	0.0969	-	0.4385	0.6610
5	0.0746	0.0792	+	-0.9906	0.3219
6	0.0627	0.0670	+	-0.9790	0.3276
7	0.0515	0.0580	+	-1.6155	0.1062
8	0.0542	0.0512	-	0.7937	0.4274
9	0.0414	0.0458	+	-1.2038	0.2286

Chi-square	Degrees of freedom	Level of significance
8.3807	8	0.3972

Note: the expected proportions are those of Benford's Law. The Bias column reads + if the actual proportions exceed those of Benford's Law, and – otherwise. The null assumption is the following one: the observed frequencies are not significantly different from those of the Benford's law. Z values and P-levels result from a proportions test comparing the observed frequency with the expected frequency.

Table 4: second digit frequencies of positive earnings of 3,379 public U.S. firms

Digit	Observed frequency	Expected frequency	Bias	Z value	P- level
0	0.1197	0.1243	+	0.8269	0.4083
1	0.1139	0.1062	-	-1.3990	0.1618
2	0.1088	0.1086	-	-0.0388	0.9690
3	0.1043	0.1083	+	0.7580	0.4485
4	0.1003	0.0971	-	-0.6269	0.5307
5	0.0967	0.0968	+	0.0185	0.9852
6	0.0934	0.0917	-	-0.3250	0.7452
7	0.0904	0.0885	-	-0.3776	0.7057
8	0.0876	0.0962	+	-0.4791	0.6319
9	0.0850	0.0823	-	-0.5684	0.5697

Chi-square	Degrees of freedom	Level of significance
6.5892	9	0.6487

In Tables 3 and 4 the theoretical frequencies of the Benford's law are compared with the observed frequencies. The observed frequencies are very close to the theoretical frequencies as well for the first digit as for the second. The Z value tests the null assumption "the observed proportion is equal to the theoretical proportion" does not reveal any significant difference for the two tables. The chi-square test does not detect any significant difference between the observed distribution and the theoretical distribution for the first and the second digit at one, five or ten percent level.

The preceding studies concerned company data sets. Nigrini (1996) analysed Tax returns on the U.S. Internal Revenue Service Individual Tax Model Files. The digital frequencies of Interest Received and Total Interest Paid (with 70,725 observations in 1985 and 54,737 observations in 1988) were analysed because the evasion distortion would be that interest paid numbers are overstated and that interest received numbers are understated. Nigrini distinguishes the unplanned evasion (UPE) from planned evasion (PE). In the first case, the taxpayer manipulates line items at filling time: the typical example is the taxpayer who will never declare interests received by a foreign bank. In the second case, there are planned actions to conceal an audit trail. The act to falsify a number is influenced by the manner of thinking the number. Rosch E (1975) showed that the manipulation of a number is generally done in the same row of this last. Thus, if a number is between 10 and 99, the invented number has very strong probabilities to be included in this interval. For the first digit of received interests, the observed frequencies are higher than the theoretical frequencies for the small figures (and conversely for the high figures). For the interests paid, the opposite phenomenon occurs: for the first digit, the observed frequencies of small figures are lower than the theoretical frequencies (and opposite for high figures). The excess of small figures for interest received suggests minoration by certain taxpayers whereas the excess of large figures for the interests paid suggests an increase by certain taxpayers.

3 LABORATORY STUDY: DISCUSSION AND RESULTS

In the experiment, 393 BA students played the role of evader. In a balance sheet with a very important loss, the students must transform the loss in a profit; the proposition must be included between 100,000 to 999,999 (six digits). The balance sheet was accompanied by the following text:

Accountant director in a multinational, you note that the last balance sheet reveals a catastrophic situation because the loss is higher than a billion Monetary Units. In order to preserve appearances, you decide to falsify this loss by putting a profit. Your accountant will preserve the balance sheet equilibrium. The suggested profit will be between 100,000 and 999,999 Monetary Units in order not to wake up suspicions ". The real loss is 1,255,663 Monetary Units. You must spontaneously propose another number of six digits.

If one asks people to generate series of numbers, those are absolutely not random. (for a good review of literature, see Tune (1964)). Indeed, people cannot generate randomly numbers. Ted Hill (1998) reports an instructive experiment:" I ask the students to make the following test. If the maiden name of their mother starts with a letter between A and L, they toss two hundred times a coin and note the result. In the other case, they propose themselves a two hundred numbers serie of heads or tails. The following day, I collect the results and separate in a general astonishment actual random series from the others with 95 % of success. Although the rigorous demonstration is difficult, I observe the following rule: in a two hundred tosses serie, six consecutive heads or tails appear with a very small probability. A person trying to imitate randomly numbers set seldom writes such long homogeneous series ". If people are not able to generate randomly series, it is of course possible to find artificial means (Neuringer, 1986).

The objective of the experiment is to study the relationship between Benford's law and unplanned fraud. Insofar as humans cannot imitate random and so evaders too, do the invented numbers follow the Benford's law? In others terms, knowing that the evader falsifies numbers among other numbers obeying the famous law, can one consider that there is a contamination effect? This experiment is relatively similar of T.P. Hill (1988) which had required of a sample of 742 students to propose a number of six digits. The null assumption was the following: the distribution of digit i (for i=1 to 6) obeys a Benford's law. The Chi-square test is used to validate the assumptions (the Kolmogorov-Smirnov test is not presented, but he confirms the Chi-square test). The results are summarized in the following tables.

Table 5: Descriptive statistics of the sample (393 students)

	Sample (393 students)
Minimum	100,000
Maximum	999,999
Mean	401,504.97
Standard deviation	251,842.73
First quartile	200,000
Median	327,466
Third quartile	561,200

Table 6: first digit frequencies of the sample

First digit	Expect	Actual	Bias	Z value	P-level
1	0.3010	0.2290	-	-3.113	0.002
2	0.1761	0.2239	+	2.489	0.013
3	0.1249	0.1298	+	0.290	0.772
4	0.0969	0.0941	-	-0.185	0.853
5	0.0792	0.0916	+	0.912	0.362
6	0.0670	0.0712	+	0.341	0.733
7	0.0580	0.0585	+	0.045	0.964
8	0.0512	0.0585	+	0.664	0.507
9	0.0458	0.0433	-	-0.237	0.812

Chi-square	Degrees of freedom	Level of significance
13.3297	8	0.1001

Table 7: second digit frequencies of the sample

Second digit	Expect	Actual	Bias	Z value	P-level
0	0.1197	0.1985	+	4.812	< 0.0001
1	0.1139	0.0611	-	-3.296	0.001
2	0.1088	0.1221	+	0.848	0.397
3	0.1043	0.0636	-	-2.641	0.008
4	0.1003	0.0865	-	-0.910	0.363
5	0.0967	0.2316	+	9.048	< 0.0001
6	0.0934	0.0611	-	-2.201	0.028
7	0.0904	0.0483	-	-2.905	0.004
8	0.0876	0.0662	-	-1.502	0.133
9	0.0850	0.0611	-	-1.701	0.089

Chi-square	Degrees of freedom	Level of significance
128.3609	9	<0.0001

Table 8: third digit frequencies of the sample

Third digit	Expect	Actual	Bias	Z value	P-level
0	0.1018	0.2595	+	10.344	< 0.0001
1	0.1014	0.0560	-	-2.982	0.003
2	0.1010	0.0712	-	-1.956	0.050
3	0.1006	0.1018	+	0.080	0.936
4	0.1002	0.0433	-	-3.759	0.000
5	0.0998	0.1883	+	5.854	< 0.0001
6	0.0994	0.0687	-	-2.034	0.042
7	0.0990	0.0611	-	-2.519	0.012
8	0.0986	0.0840	-	-0.975	0.329
9	0.0983	0.0662	-	-2.139	0.032

Chi-square	Degrees of freedom	P-value
165.5211	9	<0.0001

Table 9: fourth digit frequencies of the sample

Fourth digit	Expect	Actual	Bias	Z value	P-level
0	0.1002	0.2748	+	11.528	< 0.0001
1	0.1001	0.0483	-	-3.418	0.001
2	0.1001	0.0840	-	-1.065	0.287
3	0.1001	0.0967	-	-0.225	0.822
4	0.1000	0.0840	-	-1.059	0.289
5	0.1000	0.1120	+	0.790	0.429
6	0.0999	0.1043	+	0.293	0.770
7	0.0999	0.0560	-	-2.904	0.004
8	0.0999	0.0687	-	-2.062	0.039
9	0.0998	0.0712	-	-1.888	0.059

Chi-square	Degrees of freedom	P-value
147.5008	9	<0.0001

Table 10: fifth digit frequencies of the sample

Fith digit	Expect	Actual	Bias	Z value	P-level
0	0.1000	0.3181	+	14.410	< 0.0001
1	0.1000	0.0331	-	-4.422	< 0.0001
2	0.1000	0.0687	-	-2.068	0.039
3	0.1000	0.0814	-	-1.227	0.220
4	0.1000	0.0534	-	-3.077	0.002
5	0.1000	0.1145	+	0.958	0.338
6	0.1000	0.1272	+	1.799	0.072
7	0.1000	0.0611	-	-2.573	0.010
8	0.1000	0.0611	-	-2.573	0.010
9	0.1000	0.0814	-	-1.227	0.220

Chi-square	Degrees of freedom	Level of significance
235.2188	9	<0.0001

Table11: sixth digit frequencies of the sample

Sixth digit	Expect	Actual	Bias	Z value	P-level
0	0.1000	0.3308	+	15.251	< 0.0001
1	0.1000	0.0585	-	-2.741	0.006
2	0.1000	0.0738	-	-1.732	0.083
3	0.1000	0.1069	+	0.454	0.650
4	0.1000	0.0356	-	-4.254	< 0.0001
5	0.1000	0.0585	-	-2.741	0.006
6	0.1000	0.0941	-	-0.387	0.699
7	0.1000	0.0865	-	-0.891	0.373
8	0.1000	0.0840	-	-1.059	0.289
9	0.1000	0.0712	-	-1.900	0.057

Chi-square	Degrees of freedom	Level of significance
247.1272	9	<0.0001

The main findings are:

- for the six digits, the chi-square test does not detect any significant difference between the observed distribution and the theoretical distribution of Benford’s law for the first and the second digit at one, five or ten percent level.
- another significant result is the excess of zeros which is very important (significant at 1 % level) ;
- there is an excess of fives for the second, third and sixth digit (significant at 1 % level).

These findings are coherent with previous studies: although not conforming precisely to the predictions of the Benford's law, the results of the experiment indicate that the distributions of random numbers guessed by people share the following properties with the Benford distributions:

(i) the frequency of numbers with first significant digit 1 is much higher than expected;
(ii) the frequency of numbers with first significant digit 8 or 9 is much lower than expected.

Students proposed 2,358 digits (393 times 6). From 0 to 9, the frequencies are respectively 543, 191, 253, 228, 156, 313, 207, 146, 166 and 155.

These conclusions are consistent with Chernoff's (1981) findings that generally high numbers are less likely to be chosen in numbers games.

CONCLUSION

Benford's law is a logarithmic function discovered by Newcomb predicting the frequency of digits in certain data sets. If, for the first digit, the variations between theoretical frequencies with those of random are relatively important (30.1 % to 4.58 % versus 11.11 %), for the second digit, the variation is reduced until becoming almost null from the fifth digit. Through a laboratory study, the findings show students fail to imitate Benford's law. This result is similar to previous empirical studies. Tuture research could explore the mechanisms used to invent numbers.

AUTHOR INFORMATION

Dominique Geyer is Associate Professor at Nantes Audencia School of Management (France). He holds a Ph.D. from the University of Lille (France). His primary teaching areas are Financial Accounting and Management Accounting Information Systems. He has published articles in numerous journals, including Revue Francaise de Comptabilite, Revue Francaise de Gestion, Systeme d'Information et Management, European Management Journal, Journal of Computer Information Systems and Information and Management.

REFERENCES

1. Benford, F., March 1938, "The law of anomalous numbers", Proceedings of American Philosophical Society, 78,: 551-572.
2. Carlslaw, C., April 1988, "Anomalies in income numbers: Evidence of goal oriented behavior*", The Accounting Review*, 63: 321-327.
3. Chernoff, H., 1981, "How to beat the Massachusetts Numbers Game",*Math. Intel.*, 3, 166-172
4. Hill, T.P., 1988, "Random-number guessing and the first-digit phenomenon", *Psychological Reports* 62, 967-971.
5. Hill, T.P., July-August 1998, "The First Digit Phenomenon", *American Scientist*, Vol. 86, 358-363.
Kinnunen J. et Koskela M., 2003, « Who is Miss World in Cosmetic Earnings Management ? A Cross-National Comparison of Small Upward Rounding of Net Income Numbers among Eigtheen Countries », *Journal of International Accounting Research*, Vol. 2, p. 39-68.
6. Neuringer, A., 1986, "Can people behave randomly?: the role of feedback", *Journal of experimental Psychology (General)*, 115 : 62-75.
7. Newcomb, S., 1881, "Note on the Frequency of Use of the Different Digits in Natural Numbers", *The American Journal of Mathematics*, 4 : 39-40.
8. Nigrini, M.J., 1996, "A Taxpayer Compliance Application of Benford's Law", *Journal of the American Taxation Association*, Vol. 18, No. 1: 72-91.
9. Nigrini, M.J., 1999, "I've got your number", *Journal of Accountacy* : 79-83.
10. Niskanen J. et Keloharju M., 2000, « Earnings cosmetics in a tax-driven accounting environment: evidence from Finnish public firms », *European Accounting Review*, 9 : 3, p. 443-452.
11. Pinkham, R., 1961, "On the distribution of first significant digits", *Annals of Mathematical Statistics*, 32 : 1223-1230.

12. Rosh, E., October 1975, “Cognitive reference points”, *Cognitive Psychology*, 7: 532-547.
13. Thomas, J.K., October 1989, “Unusual patterns in reported earnings”, *The Accounting Review*, 64: 773-787.
14. Tune, G., 1964, “Responses preferences: a review of some relevant literature “, *Psychological Bulletin*, 61: 286-302.
15. Van Caneghem T., 2002, « Earnings management induced by cognitive reference points », *British Accounting Review*, 34, p. 167-178.

NOTES