Detecting fraud in financial data sets

Dominique Geyer

Detecting Fraud In Financial Data Sets

Dominique Geyer, (E-mail: dgeyer@audencia.com) Nantes Graduate 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 numerical phenomenon in which sets of data that are counting or measuring some event follow a certain distribution. 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 a sample of 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.

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 statistical 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 objective of this short paper is to examine how men and women falsify the loss of a balance sheet. Are there gender differences in cheating behavior? The paper is organized into three sections. Section 1 describes Benford's law who is origin of digital analysis. Section 2 will present a synthesis of the empirical studies devoted to the application of this law in fraud detection. Section 3 presents the results of a laboratory study: a sample of 393 students had to 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 :
-

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

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 three positions.

Table 1: Benford's law: expected digital frequencies

	Position in Number				
Digit	**First**	**Second**	**Third**	**Fourth**	**Fifth**
0		0.11968	0.10178	0.1002	0.1000
1	0.30103	0.11389	0.10138	0.1001	0.1000
2	0.17609	0.10882	0.10097	0.1001	0.1000
3	0.12494	0.10432	0.10057	0.1001	0.1000
4	0.09691	0.10031	0.10018	0.1000	0.1000
5	0.07928	0.09668	0.09979	0.1000	0.1000
6	0.06695	0.09337	0.09940	0.0999	0.1000
7	0.05799	0.09035	0.09902	0.0999	0.1000
8	0.05115	0.08757	0.09864	0.0999	0.1000
9	0.04576	0.08499	0.09827	0.0998	0.1000

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 allows us to understand 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 26^{th} 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 (25^{th} 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 continuation (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:

- the data set must constitute a homogeneous unit: populations of cities, surfaces of lake, value of shares, etc.
- the 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.
- the 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 such 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
	93766.67
	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 facts would have drawn attention to:

- 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 the 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 always 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 the profit 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 samples. He studies Earnings before extraordinary items and discontinued operations at the quarterly and annual level. His study is finer because he distinguishes the profits from the losses. He also notes an excess of 0 for the second digit of profits. 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.

Note that we didn't observe this phenomenon in a sample of French firms extrated from Diane (a French database). The studied variable is the profit (no losses). The accounting period is closed in 1998. The firms whose 1998 result was lower than 9 thousands of French francs were eliminated. This elimination is pertinant because in the database, variables are valued in thousand of French francs and tests concerns the first two digits. So the sample has 81 259 firms profits.

Table 3 : first digit frequencies of 81 259 French firms profits

First digit	Expect	Actual	Bias	Z
1	0.3010	0.2991	-	1.2228
2	0.1761	0.1742	-	1.4321
3	0.1249	0.1251	+	0.1072
4	0.0969	0.0984	+	1.4074
5	0.0793	0.0790	-	0.1649
6	0.0669	0.0672	+	0.2452
7	0.0580	0.0585	+	0.6324
8	0.0512	0.0523	+	1.4155
9	0.0458	0.0464	+	0.8275

Chi-square	Degrees of freedom	Level of significance
7.6291	8	0.4705

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. Variable Z is calculated as follows: $Z = (|pr - Pt| - 1/2N) / V (Pt X (1 - Pt)/n)$ where Pt is the theoretical frequency, Pr the real frequency and N the total number of observations. The term (1/2N) is a continuity correction term and is only used when it is smaller than the first term in the numerator. The null assumption will be rejected when variable Z is higher than 1,96 with a risk error of 5 % (2,51 for a risk error of 1 %).

Table 4 : second digit frequencies in 81 259 French firms profits

Second digit	Expect	Actual	Bias	Z
0	0.1197	0.1206	+	0.7726
1	0.1139	0.1150	+	1.0259
2	0.1088	0.1084	-	0.4081
3	0.1043	0.1040	-	0.3008
4	0.1003	0.1001	-	0.2387
5	0.0967	0.0962	-	0.4795
6	0.0934	0.0930	-	0.3622
7	0.0904	0.0915	+	1.1697
8	0.0876	0.0869	-	0.6497
9	0.0850	0.0843	-	0.6579

Chi-square	Degrees of freedom	Level of significance
4.1556	9	0.9009

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 variable Z 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 two digits.

The preceding studies concerns data sets companies. 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 this 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, one generally notes who the observed frequencies are higher than the theoretical frequencies for the small figures (and conversely for the high figures). For the interests paid, one notes the opposite phenomenon: for the first digit, the observed frequencies of the small figures are lower than the theoretical frequencies (and opposite for the 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 play 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 French francs. In order to preserve appearances, you decide to falsify this loss by putting a profit. Your accountant will preserve active total assets = total liability. The suggested profit will be between 100 000 and 999 999 thousands French francs in order not to wake up suspicions ". The students are in a situation of unplanned fraud: the loss is 1 255 663 thousands French francs. The students 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 inspired of T Varga [Revesz, 1978]: 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 the Benford's law and the unplanned fraud. Insofar as men 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? Our 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 used, 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
1	0,3010	0,2290	-	3,0576**
2	0,1761	0,2239	+	2,4230*
3	0,1249	0,1298	+	0,2134
4	0,0969	0,0941	-	0,0999
5	0,0792	0,0916	+	0,8187
6	0,0670	0,0712	+	0,2399
7	0,0580	0,0585	+	0,0453
8	0,0512	0,0585	+	0,5491
9	0,0458	0,0433	-	0,1168

Chi-square	Degrees of freedom	Level of significance
13.3297	8	0.1010

Table 7 : second digit frequencies of the sample

Second digit	Expect	Actual	Bias	Z
0	0,1197	0,1985	+	4,7346**
1	0,1139	0,0611	-	3,2168**
2	0,1088	0,1221	+	0,7668
3	0,1043	0,0636	-	2,5580**
4	0,1003	0,0865	-	0,8264
5	0,0967	0,2316	+	8,9621**
6	0,0934	0,0611	-	2,1141*
7	0,0904	0,0483	-	2,8166**
8	0,0876	0,0662	-	1,4125
9	0,0850	0,0611	-	1,6107

Chi-square	Degrees of freedom	Level of significance
128.3609	9	0.0000

Table 8: third digit frequencies of the sample

Third digit	Expect	Actual	Bias	Z
0	0,1018	0,2595	+	10,2602**
1	0,1014	0,0560	-	2,8983**
2	0,1010	0,0712	-	1,8720
3	0,1006	0,1018	+	0,0798
4	0,1002	0,0433	-	3,6745**
5	0,0998	0,1883	+	5,7698**
6	0,0994	0,0687	-	1,9497
7	0,0990	0,0611	-	2,4344*
8	0,0986	0,0840	-	0,8908
9	0,0983	0,0662	-	2,0538*

Chi-square	Degrees of freedom	Level of significance
165.5211	9	0.0000

Table 9: thourth digit frequencies of the sample

Thourth digit	Expect	Actual	Bias	Z
0	0,1002	0,2748	+	11,4464**
1	0,1001	0,0483	-	3,3364**
2	0,1001	0,0840	-	0,9814
3	0,1001	0,0967	-	0,1384
4	0,1000	0,0840	-	0,9765
5	0,1000	0,1120	+	0,7076
6	0,0999	0,1043	+	0,2058
7	0,0999	0,0560	-	2,8194**
8	0,0999	0,0687	-	1,9761*
9	0,0998	0,0712	-	1,8055

Chi-square	Degrees of freedom	Level of significance
147.5008	9	0.0005

Table 10: fifth digit frequencies of the sample

Fith digit	Expect	Actual	Bias	Z
0	0,1000	0,3181	+	14,3259**
1	0,1000	0,0331	-	4,3381**
2	0,1000	0,0687	-	1,9841*
3	0,1000	0,0814	-	1,1434
4	0,1000	0,0534	-	2,9929**
5	0,1000	0,1145	+	0,8744

6	0,1000	0,1272	+	1,7151
7	0,1000	0,0611	-	2,4885*
8	0,1000	0,0611	-	2,4885*
9	0,1000	0,0814	-	1,1434

Chi-square	Degrees of freedom	Level of significance
235.2188	9	0.0000

Table11: sixth digit frequencies of the sample

Sixth digit	Expect	Actual	Bias	Z
0	0,1000	0,3308	+	15,1666**
1	0,1000	0,0585	-	2,6566**
2	0,1000	0,0738	-	1,6478
3	0,1000	0,1069	+	0,3699
4	0,1000	0,0356	-	4,1699**
5	0,1000	0,0585	-	2,6566**
6	0,1000	0,0941	-	0,3027
7	0,1000	0,0865	-	0,8071
8	0,1000	0,0840	-	0,9752
9	0,1000	0,0712	-	1,8160

Chi-square	Degrees of freedom	Level of significance
247.1272	9	0.0000

Table13: digit sizes of the sample

Digits	First digit	Second digit	Third digit	Fourth digit	Fith digit	Sixth digit
0		78	102	108	125	130
1	90	24	22	19	13	23
2	88	48	28	33	27	29
3	51	25	40	38	32	42
4	37	34	17	33	21	14
5	36	91	74	44	45	23
6	28	24	27	41	50	37
7	23	19	24	22	24	34
8	23	26	33	27	24	33
9	17	24	26	28	32	28
Sum	393	393	393	393	393	393

Chi-square	Degrees of freedom	Level of significance
122.123	36	0.0000

(Note : the Chi-square test concerns the contingency table without the First Digit column because zero is missing)

The main findings are:

- for the first digit, the distribution suggested by the students obeys a Benford's law (see table 6) ;
- for the five other digits, the distributions suggested by the students are very different from a Benford's distribution ;
- another significant fact is the excess of zeros which is very important (significant at 1 %) ;
- there is en excess of fives for the second and third digit.

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.

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 stud, we find that the students imitate correctly Benford's law only for the first digit. This result is similar to previous empirical studies. An important future research question is the following: how the fraud affects the digital frequencies?

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. Croson, R and Buchan, N., Gender and Culture: International Experimental Evidence from Trust Games, Gender and Economic Transactions, 89, 2, 386-391
5. Dollar, D., Fishman, R., and Gatti, R., 2001, "Are women really the fairer sex? Corruption and women in government, Journal of Economic Behavior & Organization, Vol. 46, 423-429
6. Glover, S.H., Bumpus, M.A., Logan, J.E., Ciesla, J.R., 1997, "Reexamining the influence of invidual values on ethical decision-making", Journal of Business Ethics, 16 (12/13), 1319-1329
7. Hill, T.P., March 1995, "Base-invariance implies Benford's Law", Proceedings of the American Mathematical Society, 123: 887-895.
8. Hill, T.P., July-August 1998, "The First Digit Phenomenon", American Scientist, Vol. 86, 358-363.
9. Neuringer, A., 1986, "Can people behave randomly?: the role of feedback", Journal of experimental Psychology (General), 115 : 62-75.
10. Newcomb, S., 1881, "Note on the Frequency of Use of the Different Digits in Natural Numbers", The American Journal of Mathematics, 4 : 39-40.
11. Nigrini, M.J., 1992, "The detection of income tax evasion through an analysis of digital distributions", Ph. D. Dissertation, University of Cincinnaty.
12. Nigrini, M.J., 1996, "A Taxpayer Compliance Application of Benford's Law", Journal of the American Taxation Association, Vol. 18, No. 1 : 72-91.
13. Nigrini, M.J., 1999, "I've got your number", Journal of Accountacy : 79-83.

14. Pinkham, R., 1961, “On the distribution of first significant digits”, Annals of Mathematical Statistics, 32 : 1223-1230.
15. Rosh, E., October 1975, “Cognitive reference points”, Cognitive Psychology, 7: 532-547.
16. Thomas, J.K., October 1989, “Unusual patterns in reported earnings”, The Accounting Review, 64: 773-787.
17. Tune, G., 1964, “Responses preferences : a review of some relevant literature “, Psychological Bulletin, 61: 286-302