Autobiographical Note:

Bruce Busta, Ph. D., C.P.A. is a professor of accounting at St. Cloud State University, in St. Cloud, Minnesota. Dr. Busta has worked in both public and private accounting. Additionally, he has taught financial and managerial accounting at Richmond College (London), Huron University (London) and the University of Nebraska. Dr. Busta does research in the areas of auditing and instructional development.

Randy S. Weinberg, Ph.D., is Professor of Business Computer Information Systems at St. Cloud State University, St. Cloud Minnesota, USA. He is interested in interdisciplinary applications of artificial neural networks and genetic algorithms.

Using Benford's Law and Neural Networks as a Review Procedure

Bruce Busta, Ph.D., C.P.A.
Department of Accounting
College of Business
St. Cloud State University
720 Fourth Avenue South
St. Cloud, Minnesota 56301 U.S.A.
(320) 255 - 3967
Harv@stcloudstate.edu

And

Randy Weinberg, Ph. D.
Department of Business Computers and Information Systems
College of Business
St. Cloud State University
720 Fourth Avenue South
St. Cloud, Minnesota 56301 U.S.A.
(320) 255 - 3273
Weinberg@stcloudstate.edu

14 January 1998

Key Words: Fraud Detection, Review Procedure, Benford's Law, Neural Networks, Digit Distributions

Running Title: Benford's Law as a Review Procedure

Abstract

In 1938, Benford found that for many types of data, the digits of the numbers are distributed in a predictable pattern. That is, for certain data sets a specified percentage of the numbers should start with the digit 1, the digit 2 and so on. Accounting data is one of the many types of data which is expected to follow the pattern predicted by Benford (Carslaw 1988; Thomas 1989). It has also been discovered that manufactured, or manipulated, numbers do not have digit patterns which follow a Benford pattern (Hill 1988). Therefore, in theory, the more an observed digit pattern in certain accounting data sets, deviates from the expected Benford pattern, the more a data set is suspected to have been manipulated.

This paper introduces a new analytical review procedure that measures the degree to which a data set's digit distribution deviates from a Benford digit distribution. This deviation can indicate potential manipulation and can be used to signal the need for further audit testing. We use an artificial neural network to distinguish between "normal" and "manipulated" financial data. The results show that if data has been contaminated (at a 10 percent level or more) a Benford analytical review procedure will detect this 68 percent of the time. If the data is not contaminated, the test will indicate that the data is "clean" 67 percent of the time. Because analytical review procedures are not used in isolation, these results probably understate the effectiveness and potential of a digits based analytical review procedure.

This procedure's fraud detection results compare favorably to traditional analytical review procedures. Importantly, its unique analysis procedure allows it *to*

complement traditional analytical review procedures. A key limitation of this study is that it uses simulated data, rather than actual data. Such an enhancement will be a critical step in future research. This method appears to have potential merit and provides many opportunities for new research

Introduction

The role of the auditor is to assess the credibility of financial information. Auditors use many tools, including analytical review procedures, to complete this process. Effective and efficient analytical review procedures reduce the risk and increase the efficiency of the audit process. Most existing analytical review procedures investigate ratios and trends of financial data. This paper introduces a new analytical review procedure that examines the digit distribution of the numbers in the underlying financial information. In this paper, we use an artificial neural network to distinguish between "normal" and "manipulated" financial data.

Benford (1938) found that for many classes of data the digits of the numbers are distributed in a predictable pattern. It has also been discovered that manufactured, or manipulated, numbers do not have digit patterns that follow a Benford pattern (Hill 1988). In theory, the more an observed digit pattern deviates from the expected Benford pattern, the more the data set is suspect.

This expected phenomenon has been empirically established over the years by various researchers who have used digit distributions as an investigation tool. Varian (1972) tested the reasonableness of census data. Becker (1982) used digit analysis to test failure rates to see if they were following a systematic pattern. Nigrini (1996) uses Benford's Law to rate taxpayer compliance. In a similar vein, we propose the use of digit analysis to look for spurious financial data.

To establish the effectiveness of the digit analysis procedure, numerous data sets with known levels of deviation from the Benford distribution are needed for testing.

Because of the difficulty of accessing "live" corporate or financial data with fraudulent entries, simulated data sets are used in this paper. To operationalize the procedure, an artificial neural network (see Appendix for an overview of artificial neural networks) is deployed to distinguish between Benford and Benford-deviant data sets.

The remainder of the paper is organized as follows. In the next section, analytical review procedures are briefly introduced. Then, Benford distributions and the logic for using digit analysis as an analytical review procedure is discussed. The process used in this study, the data, the variables and the design of the neural network, are then explained. Lastly, results and conclusions are presented. We conclude that digit distribution analysis shows promise as an analytical review procedure. Further research is required, particularly with the use of actual or "live" financial data.

Analytical Review Procedures and Their Effectiveness

Analytical review procedures are techniques used to improve the efficiency of audits. Basically, analytical review procedures compare expected relationships among data items to actual observed relationships. If the actual relationships are not consistent with the expected relationships further audit investigation is required to explain the unexpected results. Analytical review procedures can be highly effective; Wright and Ashton (1989) found that of the errors found during an audit, 48 percent were detected by detail tests, 21 percent using experience with corrections from prior years, *16 percent by analytical review procedures*, 13 percent by client inquiry and 2 percent with general audit procedures. Importantly, the auditors in the Wright and Ashton study felt that analytical review procedures would have found more errors had a

different procedure failed. Analytical review procedures range from the naive to sophisticated statistical approaches; ratio, trend, percent change analysis and regression analysis are some typical analytical review procedures. Because analytical review procedures rely on assumptions about the underlying data and how the data should be related, they must be used cautiously and are most effective when used by an experienced auditor.

Analytical review procedures produce two different signals, with four different outcomes:

Signal 1 - More Investigation is Warranted.

Outcome 1 - This is a *correct* signal if, in fact, the data being audited has been manipulated.

Outcome 2 - This is an *incorrect* signal if, in fact, the data being audited is "clean" and does not require more audit effort. This "Over Auditing," or "false positive" outcome is called a Type I Error.

Signal 2 - No Further Investigation is Warranted.

Outcome 3 - This is a *correct* signal if, in fact, the data being audited is "clean."

Outcome 4 - This is an *incorrect* signal if, in fact, the underlying data being audited has been manipulated. This "Undetected Fraud," or "false negative" outcome is called a Type II Error. Undetected fraud can, in practice, result in audit failure with substantial consequences.

One method for determining the effectiveness of an analytical review procedure is to sum the Type I and II error rates. Of course, the lower the combined error rate, the greater the effectiveness of the analytical review procedure. Assuming that, over the

long-run, Type I and II errors are equally costly, the maximum combined error rate should not exceed 1.00. An error level of 1.00 implies that, on average, the analytical review procedure will be incorrect in 50 percent of the cases - just as effective as a coin toss.

Several studies have examined the effectiveness of analytical review procedures (Kinney, 1978; Knechel, 1986, 1988; Loebbecke and Steinbart, 1987; Wilson and Colbert, 1989; Wheeler and Pany, 1990; and Green and Choi, 1997). In four of these studies it is possible to summarize the findings in terms of combined Type I and Type II error rates (See Table 1). These studies reflect typical error rates of analytical review procedures currently in practice.

****

INSERT TABLE 1 ABOUT HERE

****

Table 1 shows that combined error rates (with the exception of Green & Choi) are surprisingly high. In fact, Loebbecke & Steinbart found a combined error rate in excess of 1.00. Notwithstanding these high error rates, Table 1 provides a benchmark to measure the effectiveness of the analytical review procedure introduced in this paper.

Benford Numbers and Benford's Law

In 1938, Frank Benford published a paper describing a numerical phenomena that has come to be known as Benford's Law.[1] In that paper, Benford demonstrated that the digits of naturally occurring numbers are distributed in a predictable and specific pattern. In this paper, we use the term "naturally occurring numbers" to define numbers which result from measuring or counting some phenomenon. These numbers can be thought of as "counting" numbers, in that they begin at zero and move in a positive or negative direction corresponding to a positive or negative change in the phenomenon. Numerous examples of Benford sets abound - invoice amounts, check numbers, heights, weights, distances (miles, kilometers, inches) are all examples of such counting numbers; large collections of these measures will follow Benford's distribution. Street addresses, death rates, areas of rivers, population of cities (Benford 1938), accounting measures of net income (Carslaw 1988; Thomas 1989), and dollar amounts on utilities bills have been shown to be Benford sets. Wlodarski (1971) and Sentance (1973) discovered that Fibonacci and Lucas numbers follow a Benford distribution. Burke and Kincanon (1991) observed that twenty constants commonly used in physics approximate a Benford set. Busta and Sundheim (1992) discovered that tax return data follow a Benford distribution. Numbers generated by a random process (lottery numbers) or a structured process (phone numbers, social security numbers) are not

[1] It was discovered that in 1881, Simon Newcomb described the same numerical phenomena that Benford described in his 1938 paper. As a consequence, Benford technically *rediscovered* the numerical phenomena which came to be known as Benford's Law.

considered counting numbers and are not Benford sets. Nigrini (1997 p. 15 - 16) lists four criteria for Benford data sets: 1) The data set should measure the same phenomena (that is, the data should be all dollar amounts or all measures of distance, etc.), 2) There should be no built-in minimum or maximum values, 3) The data set should not be assigned numbers - such as phone numbers, and 4) the data set should have more observations of small items than large items (for example, when measuring the area of rivers there are many more small rivers than large rivers; the same is generally true for financial transactions - the number of small financial events is greater than the number of large financial events.).

At this time, it is an open research question as to whether Benford's Law will apply to calendar dates. When observing dates, clearly one could not directly use calendar dates such as: 14 February 1999 or 6 October 1997. Such dates have built in maximums of 28, 30 and 31 days, which violates the second criteria. However, if the distance between dates is measured (for example: 10 days, 205 days, 424 days, etc.) such a data set *might* follow a Benford distribution. For dates, it seems that in some instances the second and fourth criteria may be violated. For example, if investigating invoice dates to see if they follow a Benford set, one would have to be certain that there is not a maximum created in the measurement by payment terms (i.e., "payment due in 30 days "). Additionally, and more troublesome is the fourth criteria, one would have to be confident that there are more small "time observations" than large "time observations." Consequently, the application of Benford's Law to dates will be dependent on the factors underlying the creation of the data set. Unfortunately, no

research has been done in this area to provide more guidance. Individuals who have data sets which would allow such research are encouraged to contact the authors.

For large groups of naturally occurring numbers (Benford sets), the leading digits tend to fall into a geometric series. In particular, the expected occurrence of each digit is based on a logarithmic distribution and can be calculated as follows:

The expected occurrence for the first digit is:

$$Probability\,(x\ is\ the\ first\ digit) = Log_{10}(x+1) - Log_{10}(x) \qquad (1$$

The expected occurrence for *x* as the first digit and *y* as the second digit is described as:

$$Log_{10}\left(x+\frac{y+1}{10}\right) - Log_{10}\left(x+\frac{y}{10}\right) \qquad (2)$$

Thus, the expected probability of occurrence for the second digit is calculated as:

$$\Pr obability\,(y\ is\ the\ second\ digit) = \sum_{x=1}^{9}\left(Log_{10}\left(x+\frac{y+1}{10}\right) - Log_{10}\left(x+\frac{y}{10}\right)\right) \qquad (3)$$

Benford's Law predicts that the first, second, third and fourth place digits (counting from the left) of the naturally occurring numbers will be distributed as shown on Table 2.

****

INSERT TABLE 2 ABOUT HERE

****

For example, if one examined a checkbook to determine how many checks begin with 1, 2, 3, etc, one would expect to find that approximately 30 percent (.30103) of the checks would begin with the digit 1, 18 percent (.17609) of the checks would begin with 2, 12 percent (.12494) would begin with 3, and so on. Similarly, approximately 12 percent (.11968) of the checks would have 0 for the second digit, approximately 11 percent (.11389) would have 1 as the second digit, and so on. For a comprehensive literature review of Benford's Law, see: Nigrini (1997).

"Manufactured" Numbers

Hill (1988) found that when individuals manufacture numbers "from their heads" these numbers do not follow a Benford set. In his study, Hill asked 742 undergraduate calculus students to write down on a slip of paper a six-digit random number "out of their heads." These manufactured numbers did not follow a Benford distribution. Table 3 shows Hill's observed frequencies for the first two digits.

****

INSERT TABLE 3 ABOUT HERE

****

This important empirical discovery, coupled with the fact that naturally occurring numbers follow a Benford set, provides the theoretical foundation for the concept that Benford's Law could be used to detect manipulated numbers.

Examining Benford's Law as an Analytical Review Procedure

The Method

Following Burke and Kincanon (1991), Thomas (1989), Carslaw (1988), Sentance (1973), Wlodarski (1971), and Benford (1938) it is assumed that in many cases non-manipulated accounting data will follow a Benford set. Coupling this with the work of Hill (1988), it is assumed that manipulated accounting data will not follow a Benford set. Thus, an analytical review procedure that measures the degree to which a data set's digit distribution deviates from a Benford digit distribution can indicate potential manipulation and can be used to signal the need for further audit testing.

To operationalize this concept, we investigate the ability of an artificial neural network to detect the degree to which various manipulated data sets deviate from a Benford distribution. We assume that any manipulated data will follow the Hill distribution. It seems a reasonable assumption that if an individual were to manipulate accounting data at random, the bogus data would be derived by a similar cognitive process that Hill had his students follow.[2]

[2] Of course, the actual distribution of manipulated data depends on the context. In looking at tax return information, for example, Nigrini (1996) reports that income figures near the tax table "breakpoints" are subject to the most downward manipulation. Previous research by the authors has found that, if contaminated numbers are assumed to follow a random pattern (uniform distribution) the effectiveness of the analytical review procedure proposed in this paper is greatly improved. However, we believe that this is an unlikely contamination pattern. Many researchers (Chernoff 1981; Tune 1964; Bakan 1960;

The Data

Simulated Benford and non-Benford data sets containing 200 2-digit numbers were generated. Each data set consisted of Benford numbers together with a known proportion of Hill numbers. Each data set was one of the following distributions: (1) a Benford only distribution, (2) a 90% Benford and 10% Hill distribution, (3) an 80% Benford and 20% Hill distribution, or (4) a 50% Benford and 50% Hill distribution. For example, if a data set was simulated to have 0 percent contamination all 200 numbers were selected at random from the Benford distribution. If a data set was simulated to have 20 percent contamination, 160 numbers were randomly selected from the Benford distribution while 40 numbers were randomly selected from the Hill distribution.

The Variables

When observing the degree to which a data set follows a Benford distribution the first, and second, digits can be examined individually. Additionally, because the combinations of the first and second digits follow a distinctive pattern they can be investigated jointly as a separate variable. For example, when observing the *combination* of the first and second digits of a stream of numbers, the numbers can range from 10, 11, 12 . . . 99. Thus, there are 90 combinations of numbers. For a Benford distribution, these two digit combinations follow a logarithmic distribution (formula (2) above and Table 4); or a Hill distribution, the two digit combinations follow joint probabilities (Table 5).

Chapanis 1953) have shown that individuals have great difficulty in generating random numbers "from their heads."

****

INSERT TABLES 4 AND 5 ABOUT HERE

****

To measure the amount the first digit, second digit and combined first and second digits deviate from a Benford distribution the observed versus expected frequencies can be compared. Additionally, summary statistics (e.g., mean, median, standard deviation, kurtosis[3], skewness[4]) can be used to measure the first digit, second digit and the two digit conformity with Benford's Law. Thus, for each data set 34 variables can be used to describe its contents - frequencies of the first digit (9) together with its mean, median, standard deviation, kurtosis and skewness, frequencies of the second digit (10), together with its mean, median, standard deviation, kurtosis and skewness, and for the two digit combinations, the mean, median, standard deviation, kurtosis, and skewness can be used.[5] Various experiments, described below, use different combinations of these variables in order to better understand their importance and impact.

---

[3] Kurtosis measures the peakness or flatness of a distribution. A positive value means the distribution is peaked. A negative value means the distribution is flatter.

[4] Skewness measures the degree to which data is distributed about its median value. A positive value indicates that the distribution is primarily to the left of the median. A negative value indicates that the distribution is primarily to the right of the median.

[5] We felt that the use of the 90 frequencies which exist for the combined two-digits would dramatically lower the efficiency of the neural network. Thus, only these five summary statistics were used to examine the two-digit number distribution in each data set.

The Neural Network

To train the neural network, 800 data sets of 200 2-digit numbers were generated. Known proportions of Benford and Hill numbers were used to generate the sample data. Of the 800 data sets, 20 percent (160 data sets) were completely derived from a Benford distribution; 40 percent (320 data sets) contained 90 percent Benford numbers with 10 percent Hill numbers; 25 percent (200 data sets) were 80 Benford and 20 percent Hill; 15 percent (120 data sets) were 50 percent Benford and 50 percent Hill. These proportions were selected so that the trained network would learn to recognize the features of contaminated data - ideally, reducing the frequency of Type II errors (undetected fraud) when deployed.

An independent test (or holdout) set of 800 data sets was created to test the network against previously unseen data. Of these 800 data sets, 50 percent (400 data sets) contained no Hill numbers; 35 percent (280 data sets) contained 10 percent Hill numbers; 10 percent (80 data sets) contained 20 percent Hill numbers; 5 percent (40 data sets) contained 50% Hill numbers. This combination of manipulated levels was selected for two reasons. (1) It seems likely that, in practice, the probability of encountering manipulated data in the course of an actual audit is low; thus, only 50 percent of the data sets were contaminated and, of these, most were contaminated at a relatively low level (10 percent). (2) To avoid overstating the power of the results, the contamination levels for the holdout set were purposely different than those in the training set.

The neural network analyzes the input variables and then generates an estimate of the degree of contamination in the data set. Six network designs were tested to determine the most effective model. In each design, the inputs to the neural network were a different subset of the 34 variables. The output from the network represents the network's estimate of the proportion of Hill numbers in the data set. A threshold value of .09 was established experimentally during network training as a reasonable cutoff point. That is, if the network-estimated proportion of Hill numbers in any data set exceeds .09, that data set is flagged as suspicious[6].

For model development, NeuroShell 2 for Windows by Ward Systems Group was used. In all cases, a three layered (input layer, hidden layer, output layer), feed forward, backpropagation model was used. The input layer consists of nodes – one for each independent variable of interest. All input values are scaled, by the software, to the range [-1,1], to remove effects of scale. Software defaults for learning rate (.1) and momentum (.1) were accepted. Initial weights between nodes were generated as random numbers. The transfer function employed by hidden layer processing nodes is the commonly used logistic function, which scales data to (0, 1) according to the following formula:

$$f(x) = \frac{1}{1 + e^{-((x-mean)/sd)}} \qquad (4)$$

[6] Selecting a reasonable cutoff point is necessary to balance the expected frequency and costs of Type I and Type II errors. A lower cutoff point will produce more Type I errors (over audit) and fewer Type II errors (undetected fraud); a higher cutoff point will produce fewer Type I errors and more Type II errors.

where the mean is the average of all of the values of that variable in the pattern file, and sd is the standard deviation of those values.

The network training continued until either of the following two events occurred: The average backpropagation error was less then .01, or 2000 epochs (complete passes through the training data set) passed with no improvement.

Summary of Network Designs:

Network Design	Variables Used	Results
Design 1	All 34 variables: for the first digit - frequencies for each digit (9) and mean, median, standard deviation, kurtosis and skewness, for the second digit - frequencies for each digit (10), and mean, median, standard deviation, kurtosis and skewness, for the two digit combinations - the mean, median, standard deviation, kurtosis, and skewness	Table 6
Design 2	24 variables: Frequencies for first digit (9), frequencies for second digit (10), summary statistics for combined first and second digits (5). Summary statistics for the first and second digit *omitted.*	Table 7
Design 3	15 variables: Mean, median, standard deviation, kurtosis, and skewness for first, second, and combined first two digits. Frequencies for the first and second digit *omitted*.	Table 8
Design 4	5 variables: Mean, median, standard deviation, kurtosis, and skewness for two digit combinations only. Frequencies and summary statistics *omitted* for the first and second digits.	Table 9
Design 5	1 variable: Mean for two-digit combinations only.	Table 10
Design 6	1 variable: Mean for two-digit combinations only. Z-test for significance rather than use of neural network.	Table 11

Results

The results show that the neural network was able to correctly classify 70.8% of the 800 data sets correctly[7]; however, the results are very sensitive to the level of contamination in each data set. Thus, it is more appropriate to examine the results for each individual contamination level. Table 6 reveals that if the data set is contaminated at the 50 percent level it is always correctly identified. If it is 20 percent contaminated, the neural network was correct in 83.8 percent of the cases. If a data set of 200 numbers was comprised of 10 percent manipulated numbers (Hill numbers) it was identified as manipulated 68.2 percent of the time. If a data set was uncontaminated, it was correctly classified at the 67.0 percent level.

Using a 10 percent contamination level, the combined Type I & Type II error rate is 64.8 percent. This compares favorably to the combined error rates found in earlier studies (See Table 1). These results indicate that if an auditor is examining a population of numbers, in which some maybe manipulated, a sample of 200 numbers would be expected to produce one of two signals.

---

[7] The independent test set had 800 data sets contaminated as follows:

Level of Contamination	Number of Data Sets	Percent Correctly Identified	Data Sets Correctly Identified
0% Contamination	400	67.0%	268
10% Contamination	280	68.2%	191
20% Contamination	80	83.8%	67
50% Contamination	40	100.0%	40
Total	800		566

Total Data Sets Correctly Identified = 70.8% (566/800)

Signal 1 - More Investigation is Warranted.

This will be the *correct* signal 68.2 percent of the time, assuming a contamination level of at least 10 percent. If the contamination level is higher the analytical review procedure will provide even a higher percentage of correct signals.

This will be an *incorrect* signal in 33.0 percent of the cases, when the numbers are not contaminated. Thus resulting in "over auditing."

Signal 2 - No Further Investigation is Warranted.

This will be the *correct* signal 67.0 percent of the time, when the numbers are not contaminated.

This will be an *incorrect* signal in 31.8 percent of the cases, assuming a contamination level of at least 10 percent. This outcome results in "Undetected Fraudulent Numbers" and can result in serious audit failure.

****

INSERT TABLES 6 - 9 ABOUT HERE

****

In order to determine the impact made by the variables, three neural networks designs were examined. Comparing designs 2 and 3 (Tables 7 and 8) reveals that by omitting summary statistics (mean, median, etc.) the network is less successful at detecting fraud. The omission of the frequencies for the first and second digits inhibits the detection of "clean" data sets. Thus, we conclude that the summary statistics help in the detection of fraud and minimize the Type II error, while the digital frequencies are important in the identification of uncontaminated data and minimize Type I errors. Design 4 (Table 9) utilizes only the summary statistics for the two digits combined. The

combined error rate of 71.8% (Table 9) is larger than the 64.8% combined error rate of Design 1; however, because this difference is relatively small, the results suggest that a considerable portion of the detection power of the analytical review procedure is contained in the two digit combination. Therefore, these variables play an important role in the effectiveness of the analytical review procedure.

Previous research of Benford's Law has used conventional statistical analysis (Nigrini 1996). This paper presents the first use of neural networks. In order to compare the effectiveness of neural networks to traditional statistical methods Designs 5 and 6 are presented (Tables 10 and 11, respectively). Design 5 has only 1 variable, that being the mean of the two digit combination. Using this as the single input variable, the neural network produced a combined error rate of 74.1%. One of the analysis that Nigrini (1996) presents is a z-test for the mean of the two digit combination. This method tests for significant deviation from an expected mean. Using a simple large sample test of hypothesis about the mean of a population, we can determine which data sets differ significantly from the expected mean of a pure Benford distribution. For each data set, we compute the statistic:

$$z = \frac{(\bar{x} - 39)}{\sigma_x} = \frac{(\bar{x} - 39)}{\sigma / \sqrt{N}}$$

where 39 is the expected mean of the Benford distribution (Nigrini 1996), σ_x is the standard deviation of the sample mean and n is the number of sample digits in each data set, 200. Design 6 parallels Nigrini's analysis. That is, if the mean of the data set

was significantly different (at the .05 level of significance) than the expected mean, the data set was identified as contaminated. Table 11 shows that this method has very few Type I errors but numerous Type II errors, with a resulting combined error rate of 90.4%. Therefore, it can be concluded that the neural network has an analytical advantage over traditional statistical analysis.

****

INSERT TABLES 10 - 11 ABOUT HERE

****

Conclusions

This paper introduces the concept of using digital analysis as a new analytical review procedure. A neural network procedure is used to distinguish between Benford and non-Benford data sets . Unlike traditional analytical review procedures that analyze the magnitudes, trends and relationships of accounting data, the neural network method analyzes the digits of the accounting data.

The efficacy of this technique compares favorably to traditional analytical review procedures. The results show that if data has been contaminated (at a 10 percent level or more) a Benford analytical review procedure will detect this 68 percent of the time. If the data is not contaminated, the test will indicate that the data is "clean" 67 percent of the time. Because analytical review procedures are not used in isolation, these results probably understate the effectiveness and potential of a digits based analytical review

procedure. Typically, the results of several analytical review procedures are observed as a group, compared to each other and examined for trends. As a result, an erroneous signal by one analytical review procedure is mitigated by the results of the other tests.

The analytical review procedure described in this paper has advantages over traditional analytical review procedures. Because the analysis is based on the distribution of the digits, it is independent of the magnitudes of the numbers under question. Traditional analytical review procedures make comparisons in such a way in that large errors are typically detected, but numerous small errors are less likely to be detected. For example, a large fraudulent sales accrual can be detected with traditional analytical review procedures by comparing sales to cost of goods sold. On the other hand, small but numerous false sales entries may not be detected by traditional analytical methods. A Benford based analytical review procedure can potentially find such manipulation even if the frequency (10 percent of the sample) of the bogus data is small, because the test is insensitive to the magnitude of the error.

This test is also independent of the number's relationship to other data. As described in the above example, a traditional analytical review procedure may typically analyze the relationship between sales and cost of goods sold. Relationships are not used with a Benford based procedure. Thus, providing a different "angle" from which to investigate the data.

Traditional analytical review procedures can also be sensitive to the distribution of the contamination. Knechel (1988, 1986) found that seeding errors in a "bunched" versus a smoothed pattern had a significant impact of the effectiveness of traditional

analytical review methods. Because a Benford test analyzes a data stream without regard to the time period of the data, the location of the manipulation, or compare the data to any cumulative totals, this method is less sensitive to the pattern of contamination.

The limitations of this study prompt the demand for future research. A key limitation of this study is that it uses simulated data, rather than actual data. Such an enhancement will be a critical step in future research. It is difficult however, to obtain real data of the type required. Detailed data from business transactions is generally confidential, especially when the data is known to have been manipulated. However, the investigation of such data is essential in order to test the assumptions that "clean" data follows a Benford distribution and contaminated data does not follow a Benford distribution.[8] Further, data set size may influence the usefulness of the analytical review procedure.[9]

Since the training and testing sets used in the neural network model development were contaminated at known levels, the overall generality of the technique is not known. If an auditor knew approximately the expected level of contamination, the training set could be developed based on the anticipated contamination. This

[8] In this case it was assumed that contaminated data would follow a Hill distribution. Theoretically, however as long as manipulated data deviates from a Benford distribution by a large enough degree, this analytical review procedure should be able to discover the contamination.

[9] Early experiments used data set sizes of 100 data points, by increasing the data set size to 200 the combined error rate was decreased on average by 18 percent.

refinement in the training of the network is important because the power of the analytical review procedure is increased when the training set emulates the data being tested.[10]

This paper has demonstrated a novel analytical review procedure using artificial neural networks to detect deviation from Benford's distribution. This procedure's fraud detection results compare favorably to traditional analytical review procedures. Importantly, its unique analysis procedure allows it *to complement* traditional analytical review procedures. This method appears to have potential merit and provides many opportunities for future research.

[10] In preliminary tests where the training set was contaminated at the same level as the testing set combined error rates as low as 50 percent were achieved.

REFERENCES

Bakan P. 1960. Response tendencies in attempts to generate random binary series. American Journal of Psychology 73: 127 - 131.

Becker, P. 1982. Patterns in listings of failure-rates and MTTF values and listings of other data. IEEE Transactions of reliability, R-31: 132-134.

Benford, F. 1938. The law of anomalous numbers. Proceedings of the American Philosophical Society Vol. 78 No. 4 (March 31): 551 - 572.

Burke, J., and E. Kincanon. 1991. Benford's law and physical constants: the distribution of initial digits. American Journal of Physics Vol. 59 No. 10 (October): 952.

Busta, B. and R. Sundheim. 1992. Tax return numbers tend to obey benford's law. Center for Business Research Working Paper No. W93-106-94, St. Cloud State University, St. Cloud, Minnesota.

Carslaw, C. 1988. Anomalies in income numbers: evidence of goal oriented behavior. The Accounting Review Vol. 63 No. 2 (April): 321 - 327.

Chapanis, A. 1953. Random-number guessing behavior. The American Psychologist 8:332.

Chernoff, H. 1981. How to beat the Massachusetts numbers game. Math. Intel. 3:166 - 172.

Green, B. P. and J. H. Choi. 1997. Assessing the risk of management fraud through neural network technology. Auditing: A Journal of Practice and Theory Vol. 16, No. 1 (Spring): 14 - 28.

Hill, T. 1988. Random-number guessing and the first digit phenomenon. Psychological Reports 62: 967-971.

Kinney, W.R., Jr. 1978. ARIMA and regression in analytical review: An empirical test. Accounting Review , Vol. 53, (January): 48-60.

Knechel, W. 1986. A simulation study of the relative effectiveness of alternative analytical review procedures. Decision Sciences Vol. 17: 376-394.

Knechel, W. 1988. The effectiveness of statistical analytical review as a substantive auditing procedure: A simulation analysis. The Accounting Review Vol. 63, No. 1 (January): 74 - 95.

Loebbecke, J.K. and P.J. Steinbart. 1987. An investigation of the use of preliminary analytical review to provide substantive audit evidence. Auditing: A Journal of Practice and Theory Vol. 6, No. 2 (Spring): 74 - 86.

Masters, T. 1993. Practical Neural Network Recipes in C++, Academic Press, Inc., New York.

Newcomb, S. 1881. Note of the frequency of use of the different digits in natural numbers. American Journal of Mathematics Vol. 4: 39 - 40.

Nigrini, M. 1996. A Taxpayer Compliance Analysis of Benford's Law. The Journal of the American Taxation Association Vol 18, No. 1 (Spring): 72 - 91.

Nigrini, M. J. 1997. *Digital Analysis Tests and Statistics*. Allen, Texas: The Nigrini Institute, Inc. Mark_Nigrini@classic.msn.com

Sentance, W. A. 1973. A further analysis of Benford's Law. Fibonacci Quarterly 11: 490 - 494.

Thomas, J. 1989. Unusual patterns in reported earnings. The Accounting Review Vol. 64 No. 4 (October): 773 - 787.

Tune, G. 1964. Response preferences: a review of some relevant literature. Psychological Bulletin 61 (4): 286 - 302.

Varian, H. 1972. Benford's Law. The American Statistician Vol 23 No. 3 (June): 65 - 66.

Wheeler, S. and K. Pany. 1990. Assess the performance of analytical procedures: A best case scenario. The Accounting Review Vol. 65. No. 3 (July): 557 - 577.

Wilson, A. and J. Colbert. 1989. An analysis of simple and rigorous decision models as analytical procedures. Accounting Horizons Vol. 3, No. 4 (December): 79 - 83.

Wlodarski, J. 1971. Fibonacci and lucas numbers tend to obey benford's law. Fibonacci Quarterly Vol. 9 No. 1: 87 - 88.

Wright, A. and R. H. Ashton. 1989. Identifying audit adjustments with attention-direction procedures. The Accounting Review Vol. 64 No.4 (October):710 - 728.

APPENDIX

NEURAL NETWORK TECHNOLOGY

An artificial neural network consists of a set of processing elements, or neurons, linked together via weighted directed arcs. Each neuron accepts input information, processes the input information, and produces an output. Input information to a neuron in a neural network can come from an external source (a database or a sensor, for example) or as the output from other processing elements. Output from a processing element can be the final result of the network or can be input to other neurons in the network to which there is a forward connection.

Neurons can be connected to each other in complex networks. The variety of theoretical network architectures is quite extensive, but typically in practice, neural networks are organized in layers so that information flows in one direction only (feedforward network) or may circulate in cyclic patterns (feedback network). A typical feedforward neural network design with three layers - an input layer, an intermediate, or hidden layer, and an output layer - is shown in Figure 1. This basic type of network architecture has been shown to be quite robust in a wide variety of applications and is the most commonly used in practice.

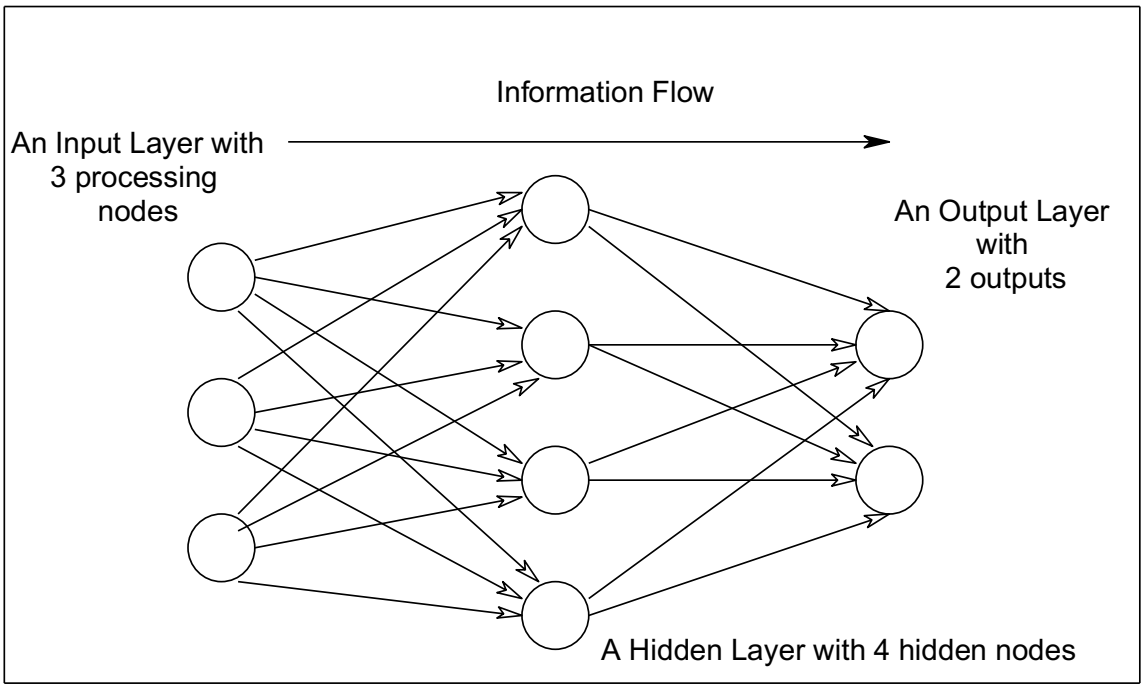


Figure 1. An Artificial Neural Network with One Hidden Layer

Each neuron, or node, in the input layer corresponds to some attribute in the problem domain. An attribute can be any information or fact that can be digitized. For example, in a neural network designed to appraise residential real estate values, each input node would correspond to one attribute of interest - such as age of the house, size of lot, square footage, number of bedrooms, and so on. The aggregate of input values is denoted by the vector **x**. In this feedforward, layered design, the neurons in the input layer are usually designed to scale raw input values into a uniform range, like [0,1]. This helps to reduce the overwhelming effect that variables with a large possible range of values can have on other variables with smaller ranges.

Output neurons provide the result of the network processing for final inspection. In the residential property appraising network, for instance, the final outputs of the network could be both the estimated selling price of the property and the anticipated time to sale. The aggregate of output values from a neural network is denoted by the vector **y**.

Each connection between nodes in a neural network is individually weighted. The strength of a connection between any two neurons in a network is given by the value of this weight. Since all the weights in a neural network can be different, each connection has its own strength, or relative importance. The effect of any neuron's output on another thus depends both on its level and on the strength of its connection to the other neuron.

The ability to "learn" is what distinguishes artificial neural networks from other types of traditional statistical or rule-based decision support systems. Neural networks use training examples to learn, iteratively adapting the interconnection weights to produce outputs that are reasonably close to what is expected. The "knowledge" or learning of a neural network, at any time, is stored as the value of the weights between nodes. A little more formally, to achieve its learning, the network is provided with a set of example input and output combinations (**x**,**y**) - a training set - and modifies its interconnection weights according to some algorithm to approximate the underlying relation from which these combinations have been selected. Contemporary neural networks employ many different learning algorithms to reach desired levels of learning in minimal time.

Ideally, a trained network reflects the features in the underlying training data set - its complex interactions, correlations, noise and inconsistencies. However, unlike traditional statistical methods, explanation or justification of a neural network's results is difficult at best. A network's knowledge is maintained internally in the form of its interconnection weights (there may be thousands). Since the weights do not represent any fact, rule, or parameter in the problem domain, they usually have no obvious interpretation - unlike, say, the coefficients in a regression equation.

Consider a typical node, node *i*, in a neural network and the other nodes to which node *i* is connected. Each value sent to node *i* is weighted by the strength of the

corresponding connection. Thus, if node *j* produces output value - or activation level - $x_j(t)$ at time t, the value of the signal reaching node *i* is equal to $w_{ji}x_j(t)$, where w_{ji} is the weight, or strength, of the connection between node *j* and node *i*. Since node *i* receives signals from many nodes, its actual input, $n_i(t)$, at time t is given by the expression:

$$n_i(t) = \sum_j w_{ji} x_j(t) + \beta_i,$$

where β_i is neuron i's bias value.

The actual relationship between a node's inputs and outputs is specified by its *transfer function.* Transfer functions can be constructed in a way that its inputs will excite the neuron, increasing its output, or inhibit the neuron, decreasing its output. A design issue in neural network applications is to select an appropriate transfer function for each neuron. While many transfer functions have been proposed, the sigmoid function that produces a continuous value in the range (0,1) is the most widely used:

$$x_i(t) = \frac{1}{1 + e^{-gain * n_i(t)}} \quad ,$$

where *gain* is a parameter determined by the network designer.

Backpropagation, the most popular training algorithm, is used extensively in practice and is described in [3]. The basic ideas of backpropagation are straightforward. During training, input vectors - $\mathbf{x}_i$ - are presented to the network and corresponding output vectors - $\mathbf{o}_i$ - derived. The performance of the network is typically measured as the Mean Squared Error (MSE) of the output values. (The Root Mean Squared Error (RMSE) - the square root of the Mean Squared Error is also commonly reported.) This is calculated summing the squared differences between the outputs of the network - $\mathbf{o}_i$ - and the expected values - $\mathbf{y}_i$ - and then dividing by the number of input/output pairs in the training set:

$$MSE = \frac{1}{n} \sum_n (y_i - o_i)^2 \ .$$

To attempt to minimize the error function, the interconnection weights are then adjusted by the training algorithm, typically employing a gradient descent method or a variation of it. The process of presenting the training set to the network, computing the

error function and adjusting the weights is repeated iteratively until some suitable stopping criterion is reached. When (and if) the performance of the trained network is deemed satisfactory - determined by its performance on previously unseen combinations of input and output vectors - the final interconnection weights are fixed and the network is deployed. Because neural networks are typically employed in contexts involving highly noisy, complex, and often conflicting data, perfect performance can never be achieved. Thus, it is the responsibility of the network designer to determine an acceptable level of training and to recognize the possibility, cost, and risk of errors when the network is deployed.

Table 1. Error rates of analytical review procedures in previous research.

Researchers	Loebbecke &Steinbart[11]	Wilson & Colbert[12]	Wheeler & Pany[13]	Green & Choi[14]
Type I Error Rate	34.57%	6.90%	58.40%	15.09%
Type II Error Rate	73.00%	85.63%	26.10%	21.95%
Combined Error Rate	107.57%	92.53%	84.50%	37.04%

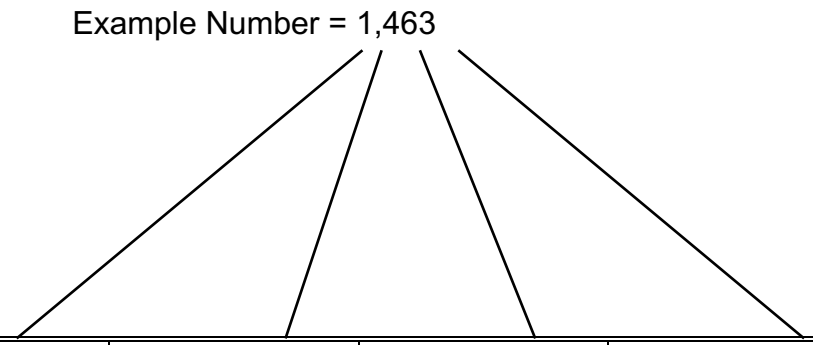
[11] Loebbecke and Steinbart (1987) page 79, Table 2, Using the 15% change rule, average of the last three columns.

[12] Wilson and Colbert (1989), page 82, Table 1, Using Statistical Model with Statistical Investigation Rule and One Error of Materiality.

[13] Wheeler and Pany (1990), page 568, Table 2. Using Investigation Rule - Statistical (0.33). The Type I and II errors are estimated as follows: (.637/.922) * .845 = .584, and (.285/.922) * .845 = .261.

[14] Green and Choi, 1997, page 24, Table 4. Using PSYDNN Model.

Table 2. Digit frequencies for Benford Numbers - First Four Digits.



Digit	First Place	Second Place	Third Place	Fourth Place
0	-	.11968	.10178	.10018
1	.30103	.11389	.10138	.10014
2	.17609	.10882	.10097	.10010
3	.12494	.10433	.10057	.10006
4	.09691	.10031	.10018	.10002
5	.07918	.09668	.09979	.09998
6	.06695	.09337	.09940	.09994
7	.05799	.09035	.09902	.09990
8	.05115	.08757	.09864	.09986
9	.04576	.08500	.09827	.09982
Total	1.00000	1.00000	1.00000	1.00000

This table is extracted from: McLaughlin, W. I., and S. A. Lundy. Digit Functions of Integer Sequences. Fibonacci Quarterly 22 (May): 109.

Table 3.

Digit frequencies for Manufactured Numbers (Hill Numbers) - First Two Digits.

Digit	First Place	Second Place
0	-	.058
1	.147	.106
2	.100	.117
3	.104	.109
4	.133	.105
5	.097	.100
6	.157	.112
7	.120	.128
8	.084	.073
9	.058	.092
Total	1.000	1.000

Table 4. Probability Distributions for the First and Second Digits - Benford Numbers.

	Second Digit										
First Digit	0	1	2	3	4	5	6	7	8	9	Total of First Digit
1	4.14	3.78	3.48	3.22	3.00	2.80	2.63	2.48	2.35	2.23	30.10
2	2.12	2.02	1.93	1.85	1.77	1.70	1.64	1.58	1.52	1.47	17.61
3	1.42	1.38	1.34	1.30	1.26	1.22	1.19	1.16	1.13	1.10	12.49
4	1.07	1.05	1.02	1.00	0.98	0.95	0.93	0.91	0.90	0.88	9.69
5	0.86	0.84	0.83	0.81	0.80	0.78	0.77	0.76	0.74	0.73	7.92
6	0.72	0.71	0.69	0.68	0.67	0.66	0.65	0.64	0.63	0.62	6.69
7	0.62	0.61	0.60	0.59	0.58	0.58	0.57	0.56	0.55	0.55	5.80
8	0.54	0.53	0.53	0.52	0.51	0.51	0.50	0.50	0.49	0.49	5.12
9	0.48	0.47	0.47	0.46	0.46	0.45	0.45	0.45	0.44	0.44	4.58
Total for Second Digit	11.97	11.39	10.88	10.43	10.03	9.67	9.34	9.04	8.76	8.50	100.00

Table 5. Joint Probability Distributions for the First and Second Digits - Hill Numbers.

	Second Digit										
First Digit	0	1	2	3	4	5	6	7	8	9	Total of First Digit
1	0.85	1.56	1.72	1.60	1.54	1.47	1.65	1.88	1.07	1.35	14.70
2	0.58	1.06	1.17	1.09	1.05	1.00	1.12	1.28	0.73	0.92	10.00
3	0.60	1.10	1.22	1.13	1.09	1.04	1.16	1.33	0.76	0.96	10.40
4	0.77	1.41	1.56	1.45	1.40	1.33	1.49	1.70	0.97	1.22	13.30
5	0.56	1.03	1.13	1.06	1.02	0.97	1.09	1.24	0.71	0.89	9.70
6	0.91	1.66	1.84	1.71	1.65	1.57	1.76	2.01	1.15	1.44	15.70
7	0.70	1.27	1.40	1.31	1.26	1.20	1.34	1.54	0.88	1.10	12.00
8	0.49	0.89	0.98	0.92	0.88	0.84	0.94	1.08	0.61	0.77	8.40
9	0.34	0.61	0.68	0.63	0.61	0.58	0.65	0.74	0.42	0.53	5.80
Total for Second Digit	5.80	10.60	11.70	10.90	10.50	10.00	11.20	12.80	7.30	9.20	100.00

Table 6
Design 1

Level of Contamination	Correctly Identified	Type I Error Over Audit	Type II Error Undetected Fraud
0% Contamination	67.0%	33.0%	NA
10% Contamination	68.2%	NA	31.8%
20% Contamination	83.8%	NA	16.2%
50% Contamination	100.0%	NA	0.0%

Independent Variables: All 34 variables.

Combined Error Rate: Type I and Type II error rate if data is contaminated at the 10 percent level: 64.8%.

Neural Network Settings: 4 hidden nodes.

NA = Not Applicable.

Table 7
Design 2

Level of Contamination	Correctly Identified	Type I Error Over Audit	Type II Error Undetected Fraud
0% Contamination	69.3%	30.7%	NA
10% Contamination	58.6%	NA	41.4%
20% Contamination	81.3%	NA	18.7%
50% Contamination	100.0%	NA	0.0%

Independent Variables: 24 variables.

Combined Error Rate: Type I and Type II error rate if data is contaminated at the 10 percent level: 72.1%.

Neural Network Settings: 6 hidden nodes.

NA = Not Applicable.

Table 8
Design 3

Level of Contamination	Correctly Identified	Type I Error Over Audit	Type II Error Undetected Fraud
0% Contamination	53.3%	46.7%	NA
10% Contamination	73.2%	NA	26.8%
20% Contamination	93.8%	NA	6.2%
50% Contamination	100.0%	NA	0.0%

Independent Variables: 15 variables.

Combined Error Rate: Type I and Type II error rate if data is contaminated at the 10 percent level: 73.5%.

Neural Network Settings: 6 hidden nodes.

NA = Not Applicable.

Table 9
Design 4

Level of Contamination	Correctly Identified	Type I Error Over Audit	Type II Error Undetected Fraud
0% Contamination	72.8%	27.2%	NA
10% Contamination	55.4%	NA	44.6%
20% Contamination	80.0%	NA	20.0%
50% Contamination	100.0%	NA	0.0%

Independent Variables: 5 variables.

Combined Error Rate: Type I and Type II error rate if data is contaminated at the 10 percent level: 71.8%.

Neural Network Settings: 6 hidden nodes.

NA = Not Applicable.

Table 10
Design 5

Level of Contamination	Correctly Identified	Type I Error Over Audit	Type II Error Undetected Fraud
0% Contamination	74.5%	25.5%	NA
10% Contamination	51.4%	NA	48.6%
20% Contamination	76.3%	NA	23.7%
50% Contamination	100.0%	NA	0.0%

Independent Variables: 1 variable, Mean for two digit combination.

Combined Error Rate: Type I and Type II error rate if data is contaminated at the 10 percent level: 74.1%.

Neural Network Settings: 4 hidden nodes.

NA = Not Applicable.

Table 11
Design 6

Z-Test of Data Set Mean (p<.05)

Level of Contamination	Correctly Identified	Type I Error Over Audit	Type II Error Undetected Fraud
0% Contamination	97.5%	2.5%	NA
10% Contamination	12.1%	NA	87.9%
20% Contamination	22.5%	NA	77.5%
50% Contamination	95.0%	NA	5.0%

Independent Variables: 1 variable, Mean for two digit combination.

Combined Error Rate: Type I and Type II error rate if data is contaminated at the 10 percent level: 90.4%.

NA = Not Applicable