A Guide to Benford's Law

A CaseWare IDEA Research Department Document

January 30, 2007

CaseWare IDEA Inc. is a privately-held software development and marketing company, with offices in Toronto and Ottawa, Canada, related companies in The Netherlands and China, and distributor partners serving over 90 countries. CaseWare IDEA Inc. is a subsidiary of CaseWare International Inc., the world leader in business-intelligence software for auditors, accountants, and systems and financial professionals (www.caseware-IDEA.com).

IDEA is distributed under an exclusive license by:

CaseWare IDEA Inc.
469 King Street West
Toronto, Canada
M5V 1K4

Table of Contents

Preface

Benford's Law is a method of analysis within Digital Analysis (DA). It is a procedure which analyzes digits in numerical data. This procedure helps to identify irregularities in a data supply. In this context, irregularities are defined as numbers, which, for example, may have been created through the (systematic) manipulation of data. An irregularity is measured based on the scale of digit distribution in a 'natural' population corresponding to the empirical legalities of Benford's Law.

History of Benford's Law

In 1920, Frank Benford, a physicist at the General Electric research laboratories, began to analyze the laws discovered by Simon Newcomb. The latter had noticed that the first pages of the books with logarithmic tables were more battered than the other pages. He concluded that, in reality, numbers starting with lower digits occur more often than numbers starting with higher digits. Therefore, the digit 1 at the beginning of a number should occur more frequently than the digit 2, the digit 2 more frequently than the digit 3 and so on. Thus, the least frequent number should be one starting with the digit 9. Principally, this conclusion contradicts the general (theoretical) idea that the frequency of all digits (0-9) at the beginning of numbers is equal.

Benford was convinced by the legality and started proving it empirically. He analyzed the frequency of the first digits of numbers from 20 different lists. The numbers on the lists were, for example, populations of different cities and countries, results of the American baseball league, and the electricity bills on the Solomon Islands. He conducted a total of 20,000 individual observations and noted that the digit 1 occurred most frequently as the first digit of numbers.

Using some mathematical assumptions and with the help of the integral calculus, he formulated equations to calculate the expected frequencies of initial digits within any lists of numbers (see also Benford's Law – Mathematical Basic Thoughts). These formulas are known as Benford's Law. Material results of his work are described in Section One.

For a long time it seemed as if Benford's Law was a freak of nature which did not meet with any general mathematical understanding. In 1961, however, Roger Pinkham managed to prove that Benford's Law was universally valid. He demonstrated this with many and diverse examples, for example, the geographical surface of different rivers. It complies with Benford's Law rules, no matter whether you measure the surface in square meters, hectare, Morgan or Ruten (scale invariance of numbers which comply with Benford's Law; so-called Benford Sets). Although the digits change through the conversion, the overall distribution does not change as it varies in accordance with the scale. In mathematical terms, this means that if a Benford Set is multiplied with a constant (unequal to 0) the newly created list of numbers is a Benford Set as well. Pinkham could even prove that Benford's Law is the only possible law on the frequency of digits which complies with this condition.

In 1920, Benford himself saw no practical benefit from his work. The conclusions and results of his numerical analyses, which took several years, were not applicable in those times.

It was not until 1988 that Benford's Law was cited in a survey by Charles Carslaw. Carslaw assumed that managers round off the company's earnings if these are slightly below a certain psychological threshold (for example, earnings of 19.9 million are rounded off to 20 million). Should such rounding occur, assumed Carslaw, the number 9 as the second digit in a list of company earnings would occur rather rarely, whereas the number 0 as the second digit would occur relatively frequently. In this case, Carslaw used the frequencies calculated by Benford as a benchmark for the results of his analyses. They resulted in the fact that in a list of company earnings the number 0 as the second digit would occur relatively often and the number 9 relatively seldom (compared to Benford's Law).

Also in 1988, Hill proved that numbers which were (randomly) invented by people do not regularly confirm with Benford's Law. The psychology of a swindler generally leads to number patterns which deviate from the distribution expected by Benford's Law. Furthermore, numbers which are based on an artificial system (for example, telephone numbers, account numbers, and check numbers) do not comply with the numerical distribution according to Benford's Law.

In 1993 Christian and Gupta discovered another interesting phenomenon with reference to the practice. They analyzed data of tax figures in order to discover signs for tax evasion. They assumed taxpayers intended to force their taxable income into the next highest graduated tax rate. Thus, the values of the graduated tax rates created the threshold values which were to be supervised in the income tax charts. Any reduction of the taxable income by a couple of US Dollars to below a certain graduated tax rate could possibly lead to substantial tax savings. According to analyses of income tax filings, comparably more taxpayers have an income which ends with the digits 40-49 and 90-99 than an income which ends with the digits 50-59 and 00-09. This indicates that US taxpayers intended (and surely still intend) to force their income below the next highest graduated tax rate of the US income tax charts.

Thus, the avoiding of threshold values, limits to authorization and so on can also be discovered by the help of the distribution of the leading digits.

In 1997, Nigrini and Mittermaier developed six numerical tests which were first applied by the worldwide operating auditing firm Ernst & Young as an inherent part of their audit method in order to discover irregularities in the numerical distribution of client data from different audit areas. The aim is to determine whether an analyzed data supply appears realistic (authentic) and to analyze possibly doubtful or irregularly appearing data more thoroughly. These tests are an integral part of the Benford Module and will be described in more detail in Section Two: The Fundamental Analyses.

In the course of years and decades, a multitude of analyses have been published, which have concentrated, on the one hand, on the purely mathematical aspects of Benford's Law and, on the other hand, on the practical applicability to examine the integrity of data. The section Fields of Application includes a selection of secondary publications on these topics.

SECTION 1: Method

- *Preliminary Thoughts*
- *Benford's Law - Mathematical Basic Thoughts*
- *Pre-Conditions for the Application of Benford's Law*
- *Independence Theorem According to Pinkham*

1.1 PRELIMINARY THOUGHTS

In 1938, the physicist Frank Benford laid the foundation for the empirical law named after him (Benford's Law). He analyzed the distribution of the first digit in a natural population of numbers and discovered that the number 1 as the first digit of every number occurs in 30.6 % of the cases compared to the number 9 as first digit in only 4.5 % of the analyzed cases. Thus, Benford's main statement is that the frequency of the first digit in a population's numbers decreases with the increasing value of the number in the first digit.

In the course of further analysis and with the help of some statistical assumptions, which are explained in the section Benford's Law – Mathematical Basic Thoughts, Benford was able to prove empirically that his discovery includes a legality which facilitates, in the form of mathematical formulas, the derivation of the probable frequency of occurrence of any digit or any numerical combination at the beginning of numbers from a number series.

The formula is: $P(d) = \log(1 + 1/d)$

P (d) stands for the probability that a number starts with the digit d.

The following chart states the frequencies calculated by Benford for each digit:

Digit	1st Position	2nd Position	3rd Position
0	N/A	0.11968	0.10178
1	0.30103	0.11389	0.10138
2	0.17609	0.10882	0.10097
3	0.12494	0.10433	0.10057
4	0.09691	0.10031	0.10018
5	0.07918	0.09668	0.09979
6	0.06695	0.09337	0.09940
7	0.05799	0.09035	0.09902
8	0.05115	0.08757	0.09864
9	0.04576	0.08500	0.09827

According to the above chart, the digit 2 as the first digit of a number within any list of any numbers occurs with a frequency of 17.609 %. Here it is of no importance how many digits the numbers have (invariance of scales, see Independence Theorem According to Pinkham). In other words, the absolute size of a number has no effect on the expected leading digit.

Based on the following number scale, which will be analyzed for the distribution of the first digits, the following example demonstrates how the algorithm calculates:

Number scale:
100
17.20
37
10259
251
1.8
97654
…

Each of the seven digits has a different numerator. In our case, based on the first number (100) the numerator would jump from the digit 1 to 0 to 1, as demonstrated in the following example:

1. Number = 100 Counter ‘Number 1’ of 0 up 1
2. Number = 17.20: Counter ‘Number 1’ of 1 up 2
3. Number = 37 Counter ‘Number 3’ of 0 up 1
4. Number = 10259: Counter ‘Number 1’ of 2 up 3
5. Number = 251 Counter ‘Number 2’ of 0 up 1
6. Number = 1.8 Counter ‘Number 1’ of 3 up 4
7. Number = 97654: Counter ‘Number 9’ of 0 up 1

The observed frequency of each digit is entered into a diagram. The x-axis (abscissa) includes the individual analyzed digit, and the y-axis (ordinate) the observed frequency per digit:

Analysis of the first digit

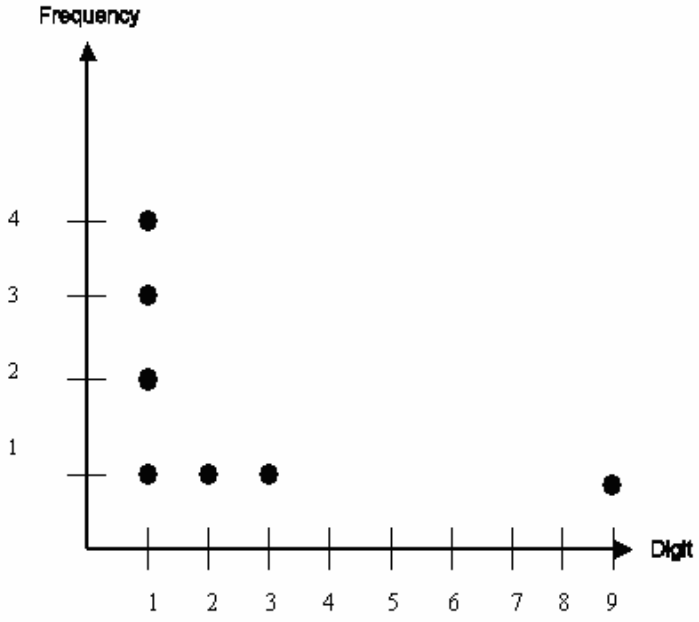

Figure 1: Visualized distribution of the leading digits

The frequency is entered into the diagram for all digits of the analyzed population and presented in graphic form. The comparison of the actual numerical distribution with the expected numerical distribution according to Benford's Law can be visualized as follows:

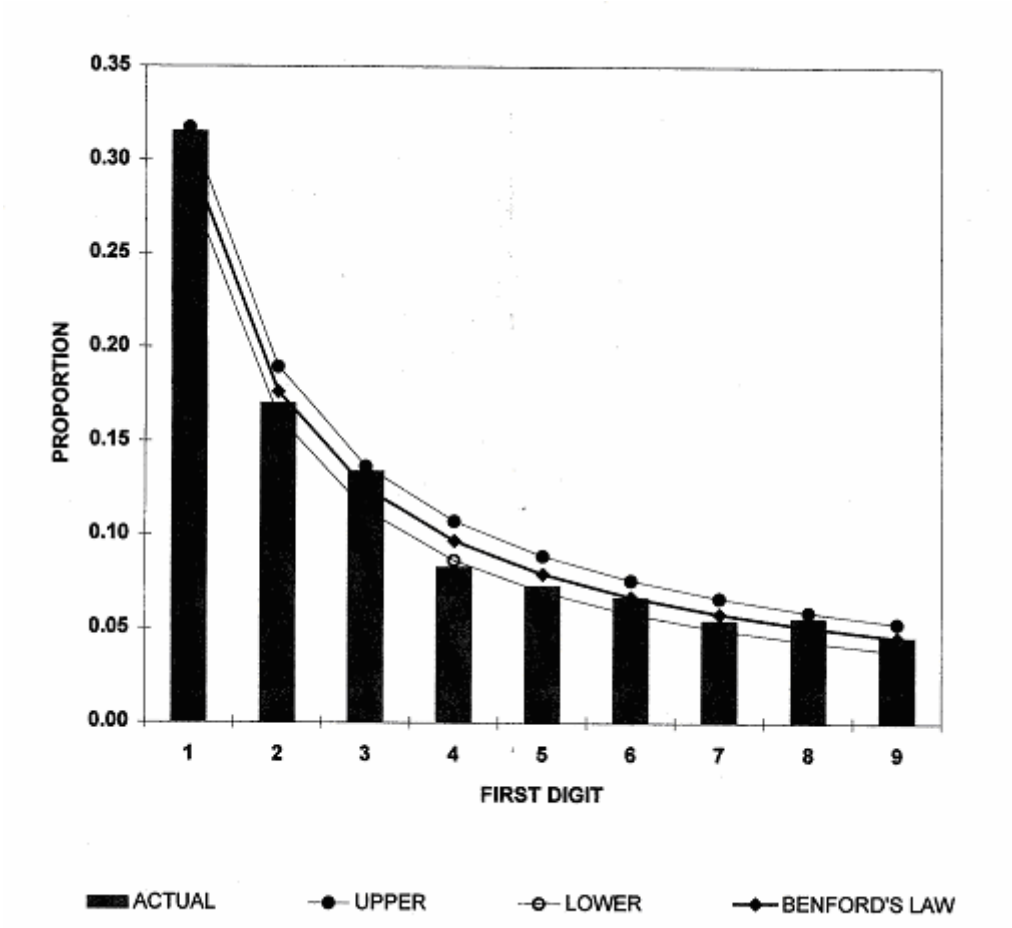

Figure 2: Visualized distribution of the leading digits measured by Benford's Law.

The black bars represent the actual (observed) frequencies of the digits within the analyzed data supply. Those will be evaluated according to Benford's expected frequencies (black, squared line). Thus, this line represents the curve of the reference value.

The reference value curve (EXPECTED-curve) is valid for the range, which is calculated on the basis of a pre-determined confidence level (for example, 95 %). This range is presented as an upper and lower bound. In case the actual frequencies are above the upper bound or below the lower bound, the individual actual frequency deviates in statistical terms significantly from the expected frequency. This is an indication of an irregularity in accordance with Bedford's Law.

The section Benford's Law – Mathematical Basic Thoughts explains the calculation and significance of the range in more detail.

The question why does the digit 1 as the first digit occur more frequently in a natural population than the digits 8 or 9 can be explained with a simple example. If someone starts counting upwards in whole numbers, starting with 1, the 1 will at any point in time occur proportionally (amount of the numbers starting with 1 divided by the amount of numbers starting with 9) more frequently or equally frequently as the 9. Furthermore, the phenomenon can be explained by the fact that in a natural series of numbers (age development of people, population development of a city) there is indeed a larger amount of smaller than bigger values.

1.2 BENFORD'S LAW - MATHEMATICAL BASIC THOUGHTS

Benford's Law is the expected distribution of digits at the beginning of numbers in any list of numbers. The number 0 as the first digit is not valid.

The basis of Benford's Law is the fact that a natural group of numbers creates a geometrical series if they are sorted from the smallest to the highest number.

The mathematical definition of a geometrical series is as follows:

$$S_n = ar^{n-1}$$

Where *a* represents the starting value of the series, *r* the ratio, *n* the nth value of the series and *S* the individual value of the nth element.

Thus, the resulting value is calculated in a geometrical series through a firm increase in percent to the preceding value (predecessor). Digits creating a geometrical series, and thus complying with Benford's Law, are also called a Benford Set.

Benford summarized formulas to calculate the expected digit frequencies of an "ideal" Benford Set, more or less as a benchmark or measure of the reference value (Expected) for actually observed values.

The resulting formulas for the frequencies of the first, second and the first two digits are as follows:

$$P(D_1=d_1) = \log(1+1/d_1); \qquad d_1 \in \{1,2,\ldots,9\} \qquad (1)$$

$$P(D_2=d_2) = \sum_{d1=1}^{9} \log(1+1/(d_1 d_2)); \qquad d_2 \in \{1,2,\ldots,9\} \qquad (2)$$

$$P(D_1 D_2=d_1 d_2) = \log(1+1/(d_1 d_2)); \qquad d_1 d_2 \in \{10,11,\ldots,99\} \qquad (3)$$

P represents the probability of the occurrence of the event set in parenthesis. In this case, it is the probability that in the number D_1 the first digit is = d_1

For example, the equation (3) for the calculation of the expected probability of the numerical combination 64 (the first and second digit) in a list of numbers is as follows:

$$P(D_1D_2 = 64) = \log(1+(1/64)) = \log(1{,}015625) = 0{,}006733$$

In this case, the first two digits in a number are equal with a probability of 0.67 equaling 64.

The strict condition that the sorted amount of data that is to be analyzed must create a geometrical series is now an obligatory condition for the application of Benford's Law. It is sufficient if the numerical frequency lies within a certain range (limited by an upper and lower bound). The size of the range depends on the statistical significance of the deviations. The statistical significance is calculated by the so-called Z-statistic that is calculated as follows:

$$Z <= (|p_o - p_e| - (1/2n)) / \sqrt{p_e * (1 - p_e)/n} \qquad (5)$$

p_e Represents the expected frequency, p_o the observed frequency and *n* the number of observations. The term (1/2n) is a factor, which is simply used to catch deviations in the calculation of the Z-statistic resulting from a large data supply. It has only a small effect on the calculation.

The Z-statistic depends on the determined confidence level. Assuming a confidence level of 95%, the value of the Z-statistic is 1.96.

If this Z-value is exceeded for an observed frequency this means that the actual distribution of the numerical combination deviates significantly in statistical terms from the expected distribution according to Benford's Law and that with a probability of 95% it is not within the range. Depending on the interpretation, such cases must be further analyzed. Within the Benford Module a general confidence level of 95% is assumed.

The determining size of this evaluation is firstly the difference between the observed and the expected frequency. If this value increases, thus the value of the Z-statistic increases as well. Secondly, the size of the data supply, i.e., the amount of the observed values, is significant for the value of the Z-statistic.

Example:

The digit 6, as the first digit, has an expected frequency of 6.7% (to be viewed in the chart in Preliminary Thoughts). Among 1,000 analyzed numbers (n = 1000) the observed frequency for the digit 6 was 7.7%. Thus, the Z-statistic has a value of 4.805. Based on a data supply of n = 2000 the resulting value for the Z-statistic is 6.98. Assuming a confidence level of 95 %, in this case both values exceed the comparative value of 1.96. Thus, both of the analyses result in a significant deviation between the actual (observed) frequency and the expected frequency. In the latter case, the value of the Z-statistic has not doubled compared to n.

Thus, the Z-statistic shows whether the actual frequency deviates significantly for a specific digit or numerical combination from the expected frequency (in statistical terms) (based on a pre-determined confidence level). This deviation may range within a certain interval. The upper and the lower bounds of this interval are calculated, based on a confidence level of 95 %, according to the equation (5) as follows:

$$\text{(Upper Bound)} = p_e + 1{,}96 * \sqrt{(p_e * (1 - p_e)/n)} + 1/(2n)$$

$$\text{(Lower Bound)} = p_e - 1{,}96 * \sqrt{(p_e * (1 - p_e)/n)} - 1/(2n)$$

If the observed frequency exceeds the upper or lower bound of this interval, the deviation must be classified as statistically significant. However, one should bear in mind that the statistically determined significance does not necessarily implicate significance regarding the content. If a deviation is classified as statistically significant, the auditor must verify or falsify this (based on random samples, analytical audit procedures or experience).

A statistical deviation simply indicates abnormalities in the data, based on Benford's Law. However, it is by no means equal to evidence for the manipulation of data supplies and so on. In fact, the deviation can have explainable, company-specific, logical reasons.

1.3 PRE-CONDITIONS FOR THE APPLICATION OF BENFORD'S LAW

1) Geometrical series

The mathematical pre-condition for the analysis of a data supply based on Benford's Law is that the data supply is based on a geometrical series (thus, that it is presented as a Benford Set). However, in reality this condition is rarely met. Experience shows that data must only meet this condition partially, i.e., the constant, increase percentage-wise of an element compared to the predecessor must only be met partially. Otherwise, this would mean that no number may occur twice which is improbable in business data supplies. However, the pre-condition is that there is at least a 'geometrical tendency'.

2) Description of the same object

The data must describe the same phenomenon. Examples are the population of cities, the surface of lakes, the height of mountains, the market value of companies quoted on the NYSE, the daily sales volume of companies quoted on the Stock Exchange, and the sales figures of companies.

3) Non-existence of minima and maxima

The data may not be limited by artificial minima or maxima. A limitation of exclusively positive numbers (excluding 0) is permissible as long as the figures to be analyzed do not move within a certain, limited range. This applies, for example, to price data (for example, the price of a can of soda will generally always range between 80 and 99 cents) or fluctuations in temperature between night and day.

4) Non-existence of a numeric system

The data may not consist of numbers following a pre-defined system, for example, account numbers, telephone numbers, social security numbers etc. Such numbers show numerical patterns which refer to the intentions of the producer of the system of that number, rather than to the actual size object represented by the number (e.g. a telephone number starting with a 9 does not mean that this person possesses a bigger telephone).

Basically, data comply with Benford's Law the best if they meet the rules mentioned above, i.e. if the data consists of large numbers with up to 4 digits and if the analysis is based on a sufficiently large data supply. A large data supply is necessary to come as close to the expected numerical frequencies as possible. For example, the expected frequency of the digit 9 in any data supply is 0.0457. If the data supply consists of only 100 numbers, there might be 5 % of the numbers that have a 9 as their first digit. Thus, the data supply of over-proportional deviations from the Benford's Law is too low. In large data supplies, the numerical distribution is increasingly closer to the expected frequencies.

1.4 INDEPENDENCE THEOREM ACCORDING TO PINKHAM

For a long time it seemed that Benford's Law was a freak of nature that was far away from the general mathematical understanding. In 1961 Roger Pinkham managed to prove that Benford's Law has universal validity. He demonstrated this with multiple examples, for example, the geographical surface of different rivers. Benford's Law follows them, no matter, whether the surface is measured in square meters, Hectare, Morgan or Ruten (scale invariance of series of numbers which comply with Benford's Law; the so-called Benford Sets).

Despite the fact that the numbers change through the transition, the general distribution does not change. It is scale invariant. In mathematical terms this means that if a Benford Set is multiplied with a constant (does not equal 0), the newly created list of numbers is also a Benford Set. Pinkham could even prove that Benford's Law is the only possible law on numerical frequencies that meets this condition.

This observation is relevant to the application of Benford's Law because, this allows a data series to be evaluated independently of units of currency or quantity (this would be intuitively illogical anyway). A Benford Set with amounts payable in USD should surely be characterized as a Benford Set as well if the amounts are stated in EURO.

The Independence Theorem also implies a limit to the application of Benford's Law as the systematic manipulation of the entire data supply by a constant factor will not result in any difference in the outcome. This factor can range within a random, limited interval (for example, between 0.94 and 0.99) without being recognized by Benford's Law. Manipulation of a data supply corresponding with Benford's Law would only be recognizable if this multiplication was used for one part of the data.

SECTION 2: The Fundamental Analyses

- *First Digit and Second Digit Analyses*
- *First Two Digits Analysis*
- *First Three Digits Analysis*
- *Rounded By Analysis*
- *Duplicates Analysis*

2.1 First Digit and Second Digit Analyses

In these first tests, the individual first or second digit of numbers in a data series will be analyzed. As a result of the analysis, the frequency of the digits 1-9 (when the first digit) or 0-9 (when the second digit) is presented in graphic and table form. Thus, it is a comparison of the reference value and the actual value to evaluate the plausibility of the underlying data material (actual value) according to the expected distribution (reference value) expected by Benford.

The expected output serves as rough check of the actual numerical distribution in the population. Statistically significant deviations may be questioned. The justifications can result from value limits in the data (for example, maximum amounts of payment) or numeric systems (for example, circles of numbers) or individual reasons leading to the explainable increase in the frequency of certain digits.

2.2 First Two Digits Analysis

This test examines the frequency of the numerical combinations 10 to 99 in the first two digits of a series of numbers. The test is presented in a graphic form that shows the expected frequencies according to Benford (reference value) and the actual frequencies of the analyzed data (actual value) on an abscissa divided into 10-99. The x-axis includes the expected and actual frequencies per numerical combination. Numerical combinations, which occur with a frequency exceeding the confidence interval, are marked as anomalies. In addition to the presentation in graphic form, the data, which the graphs are based on, are presented in table form.

In particular, the output serves for the analysis of avoided threshold values. Thus, these tests help to clearly visualize when order or permission limits have been systematically avoided.

2.3 First Three Digits Analysis

This test examines the frequency of the numerical combinations 100 to 999 in the first three digits of a series of numbers. The test is presented in a graphic form in which the expected frequencies of the analyzed data (actual value) are illustrated in an abscissa divided into 100 to 999. The x-axis includes the expected and actual frequencies per numerical combination. Numerical combinations, which occur with frequencies exceeding the confidence interval, are marked as anomalies. In addition to the presentation in graphic form, the data, which the graphs are based on, are also presented in table form.

In particular, the output serves for analysis after conspicuous rounding off operations. In general, this analysis will include many deviations because in order to receive a comparative distribution with Benford, there must be a large amount of observation values. The reason is that at least 899 observation values are needed so that every numerical combination occurs at least once (100-999). Therefore, this analysis usually does not lead to a meaningful result until it is based on a population of over 10,000 observation values. It seems advantageous that the degree of exactitude is higher and the business events to be questioned per numerical combination tend to be lower in this test than in the others.

2.4 Rounded By Analysis

This test is used to analyze the relative increasing frequency of rounded numbers. The determination comprises the frequency of the numbers that are divisible by 10, 25, 100, and 1,000 (as well as any user-defined value of whole numbers) without remainders.

The empirically observed frequency of the analyses conducted by Nigrini is used as measure of the reference value. According to this, values that are divisible by 10 are expected in a range of 10% of the observation values and values divisible by 25, 100 and 1,000 in a range of 4%, 1% and 0.1% as reference value. Here it is important that the decimal places of a number are considered as well. If they were included, the number 100.50 would no longer be a multiple of 25. In the opposite case, the places after the decimal separator are simply 'cut off', i.e., in our example the number 100.50 is a multiple of 25 as it is interpreted as a 100. Thus, values are treated like whole numbers.

2.5 Duplicates Analysis

The analysis of multiple duplicates includes all number values in the entire database that occur more than once. This test helps the user to recognize all existing duplicates in the data supply whereas the result table presents the duplicates sorted according to the descending frequency. The aim of the test is to identify certain numbers that occur more than once (for example, possible double payments). The difference from the other tests is that this test does not analyze any numerical combinations but the pure value of a number.

SECTION 3: Fields of Application

- *Forensic Audit*
- *Tax Audit*
- *Audit of Annual Financial Statements (External Audit)*
- *Internal Audit*
- *Corporate Finance/Company Evaluation*
- *Controlling*

3.1 Forensic Audit

The forensic audit sector is one of the classic fields of application for Digital Analysis. Digital Analysis can be applied where the occurrence of systematic suppression or fraud (for example, check fraud, avoiding permission limits, intentional double payments and/or multiple payments to possibly non-existing creditors) is suspected. This is generally where data is deliberately manipulated, because, according to Hill, the system of fraud leads to number patterns, which deviate from the natural expected distribution according to Benford's Law. Deceitful actions to deliberately manipulate a series of numbers causes the number patters to deviate from the legality according to Benford's Law.

3.2 Tax Audit

Digital Analysis can also be applied in the tax audit. For example, Benford's Law can help to examine whether taxpayers force their taxable income, when filing it, below a basis of measurement (see the example in section The History of Benford's Law). Similar problems can also occur in the corporate finance sector and can be examined based on the results of Digital Analysis.

3.3 Audit of Annual Financial Statements (External Audit)

Within the framework of the external audit, Digital Analysis can be applied in multiple ways. Digital Analysis enables the discovery and detailed examination of irregularities or abnormalities in payment transactions (for example, manipulation of checks, cash on hand, or accounts, double payments, and double invoice numbers). Results of Digital Analysis can also be used for the analysis of systematically incorrect valuations of inventories (see Example III in the section Application Examples). Within receivables for example, it can be used to determine whether there is an increased frequency of outstanding loans with an amount slightly below the next permission limit. This may indicate insufficient asset management by avoiding escalation stages. Furthermore, this could also indicate cases where partial receivables were deliberately remitted (written off). Moreover, it offers the possibility of conducting plausibility checks of data in the framework of roll-forward audit procedures by comparing the numerical distribution observed during the course of the year with the patterns which resulted for the data for the period between pre-audit and audit of the annual financial statements. Significant deviations in this case should be examined more thoroughly afterwards as they might indicate extraordinary activities within this period (for example, due to active window dressing).

3.4 Internal Audit

Basically the same fields of application from the external audit also apply to the internal audit. In the particularly important sector of materiality testing, the internal audit can conduct process optimizations to examine whether any economies of scale can be realized. These tests can determine, for example, if small orders are frequently placed with one supplier (possibly by different responsible persons), whether a concentration of resources will achieve economies of scale. If, for example, multiple orders of USD 200 are placed within one month, the digit 20 will occur comparably often in an order data supply and will possibly be classified as an anomaly.

3.5 Corporate Finance / Company Evaluation

In corporate finance, numerical analysis can be used to examine cash-flow-forecasts for profit centers. Numerical analysis can determine whether the forecasts are systematically adjusted upwards which would result in an excessive occurrence of higher digits.

3.6 Controlling

Digital Analysis can also be used within the controlling sector (besides the functions described for the audit which are also applicable to controlling) to discover possible conspicuous features in the data supply, which could be created through the deliberate manipulation of actual values. It is conceivable, for example, that managers conduct systematic changes to the data supply in order to achieve their pre-determined reference value target (excessive rounding). Digital Analysis can then help the controller to discover such conspicuous data in actual value data.

In every sector it is important that irregularities like acts of suppression or fraud that occur systematically in the data supply, are only classified as anomalies by the algorithm. Due to the scale invariance, Digital Analysis cannot discover sporadically occurring individual cases!

SECTION 4: Application Examples

- *Example I: Systematic undershooting of limits to authorization*
- *Example II: Check for fraud in refund claims for medical costs*
- *Example III: Errors in the valuation of inventories*

4.1 EXAMPLE I:

SYSTEMATIC UNDERSHOOTING OF LIMITS TO AUTHORIZATION

Within the framework of an unfinished prophylactic analysis of the data of a used car dealer, it was recognized that the Benford threshold value for the numerical combination 49 was exceeded significantly. The Benford report on the analysis of the first two digits of car purchases and sales data supply, classified the frequency of the numerical combination 49 at the beginning of the analyzed button/field as an anomaly. Without additional information, no definite interpretation could be made for the discovered frequency, as such a significant deviation from the Benford-EXPECTED could possibly be based on logically explainable and justifiable grounds. The following example will explain the possible different backgrounds for interpreting this anomaly:

The significant deviation in excess of the Benford threshold can be explained if the numerical combination 49 is included in the purchase price field of a data file related to the purchase of used cars by a car dealer. This may happen if data is for a period of time during which the car dealer had a special promotion where at least USD 4,900 was paid for the trade-in of a used car when a new car was sold in exchange. Then the unusual number of cars with a purchase price starting with the numerical combination 49 is not surprising. In such a case, many cars that would have been entered with a purchase price of USD 1,000, USD 1,500 or USD 3,000 under normal circumstances, were entered instead with a value of USD 4,900. The deviation is therefore logically justifiable based on the facts of the situation. Questions on the treatment of possibly hidden discounts will not be discussed here.

In contrast, a significant deviation in excess of the Benford threshold within the sales results data field could indicate a relevant auditing issue. Supposing that every used car sale, which results in a loss of USD 5,000 or more must be approved by the manager of the dealership, an unusually large number of the 49 numerical combination could indicate that this approval rule is being circumvented. In this case, the systematic sale of cars at a loss of USD – 4,900 could be the reason for the excess in the numerical combination 49.

In general, the numerical test or "Number test" allows the user to test the existence of so-called 'Salami tactics', whereby ordering limits, credit limits, and so on are circumvented by splitting up a larger amount into smaller amounts below the authorization threshold.

Accordingly, it is not sensible or necessary to apply the numerical distribution according to Bedford's Law as the absolute standard. It may be much more helpful to investigate specific numerical combinations for which the user expects a different distribution. For example, if the user knows that in a certain department for a particular time period, all out-going check numbers began with the numerical combination 10-19, the Benford module can be used to test whether a data record with a check number that begins with a different numerical combination exists in the accounts payable payments file for this department and period of time. In this manner, the user can investigate every numerical distribution as desired, completely independent of the reference numbers according to Benford.

4.2 EXAMPLE II:

CHECK FRAUD ON THE BASIS OF REFUND CLAIMS FOR MEDICAL COSTS

In 1995, Digital Analysis was conducted for an American company on the reimbursement of medical costs to employees for the previous year (1994). It was suspected that cases of fraud had occurred connected to this because employees charged the costs they had first borne themselves, to the company, although they were not justified.

The company's controls connected to the reimbursements of costs were extremely poor. Basically, it depended on the head of the department, who had been working with the company for 10 years.

Due to the suspected fraud, all claims for reimbursement exceeding USD 1,500 were analyzed using digital analysis. (For processing reasons and in order to reduce the amount of data to be processed, only reimbursements of more than USD 1,500 were registered as it was assumed that the reimbursements from any fraud would exceed this amount.)

The analysis showed that the starting digits 10 to 14 occurred relatively rarely. This could basically be referred back to the erasing of all reimbursements from USD 1,000 to USD 1,500. In addition, statistically significant deviations in the section with digits starting from 15 to 31 could be observed in relation to the expected values according to Benford. The reason for this is that the majority of numbers begin with low digits and because the starting digits from 10 to 14 were eliminated, the starting digits 15 to 31 occurred relatively frequently.

A different deviation, which was classified as an anomaly, resulted for the frequency of the starting digits 65 (thus, USD 65, 650 to 659, USD 6,500 to 6.599 etc.). After a thorough analysis it was found that 13 of the supervised checks and incidents of fraud were for 13 bottle/sleigh rides. Through the analysis it was found that these 13 cases were prepared and processed directly by the head of the department for the reimbursement of the rent and heating.

4.3 EXAMPLE III:

ERRORS IN THE VALUATION OF INVENTORIES

The assigned auditing company examined the valuation of the inventories of a company in the consumer goods sector. As the company had a wide range of products, with great differences in value and extremely different quantities of stock, it was assumed that the values of the inventory followed the natural distribution of Benford's Law. Hence, the inventory was analyzed to determine whether the starting digits of the inventories corresponded with the expected frequency distribution.

The auditors conducted Digital Analysis based on a data file with inventory data for all locations of the company. Based on the analysis, it was realized that the numerical combination 10 occurred relatively often as a leading figure. The analysis of the first three digits showed as well, that the numerical combination 100 had an extraordinarily high frequency.

Except for these distinctive features all other digits remained within the framework of expected frequencies. Further analysis of the deviations showed that an increasing number of articles, which were planned for advertising purposes, had been recorded in the inventory list with a wrong value of USD 0.01. According to the system of Benford's Law they were assigned to the groups of the starting digits 1, 10 and 100, as soon as the user chose the option Include digits after the decimal point. If these articles had been valued correctly their value would have been USD 200,000.

SECTION 5: Application Limits

- *Data-based Conditions*

5.1 DATA-BASED CONDITIONS

1) Geometrical Series

The mathematical pre-condition for the examination of a data supply based on Benford's Law is that the data supply is based on a geometrical series (thus, that it is presented as a Benford Set). In reality this condition is rarely met. Experience shows however, that data must only partially meet this condition, i.e., the constant increase, percentage-wise of an element compared to the predecessor must only be met partially. Otherwise, this would mean that no number may occur twice which is quite improbable in the case of business data supplies. However, the pre-condition is that there is at least a 'geometrical tendency'.

2) Description of the same object

The data must describe the same phenomenon. Some examples are the population of cities, the surface of lakes, the height of mountains, the market value of companies quoted on the NYSE, the daily sales volume of companies quoted on the Stock Exchange, and the sales figures of companies.

3) Unlimited data space (non-existence of minima and maxima)

The data must not be limited by artificial minima or maxima. A limitation to exclusively positive numbers (excluding 0) is permissible as long as the figures to be analyzed do not move within a certain, limited range. This applies, for example, to price data (for example, the price of a case of beer will generally always range between 15 and 20 dollars) or fluctuations in temperature between night and day.

4) No systematic data structure

The data must not consist of numbers following a pre-defined system, such as account numbers, telephone numbers, and social security numbers. Such numbers show numerical patterns that rather refer to the intentions of the producer of the number system than to the actual object size, represented by the number (for example, a telephone number starting with the number 9 does not mean that this person possesses a bigger telephone).

Basically, data complies best with Benford's Law if it meets the rules mentioned above, namely the data consists of large numbers with up to four digits and if the analysis is based on a sufficiently large data supply. A large data supply is necessary in order to come as close to the expected numerical frequencies as possible. For example, the expected frequency of the digit 9 in any data supply is 0.0457. If the data supply consists of only 100 numbers, the numbers which have a 9 as their first digit may be 5%. Thus, in the case of a small data supply, there may be an over-proportional deviation from Benford's Law. In large data supplies, the numerical distribution is increasingly closer to the expected frequencies.

If the data supply has, or just roughly has, the characteristics mentioned above, it can be analyzed based on Benford's Law. However, the results of the Benford analyses are not interpretable on the basis of Benford's Law. As stated before, the expected frequencies according to Benford's Law often represent, in the practical use, nothing more than a type of benchmark for the observed frequencies. Since the observed frequencies will only be compared with the legality discovered by Benford, not interpreted accordingly, it is not necessary that all conditions mentioned above be met. In fact, the analysis results will help the auditor interpret the personal expectation of the auditor, without including the reference value according to Benford in the argumentation. If, for example, the personal expectation of the user is that the starting digit 4 must occur twice as often in the analyzed data than the starting digit 2, the results of the analyzed values must not be compared with the expected frequencies according to Benford but with the individual expectation of the user.

The application of Digital Analysis and the Benford Module is also permissible in the framework of Data Mining when certain distinctive facts in a data supply are measured against the personal expectations of the user and interpreted according to them. In this case it is not necessary for the data that is to be analyzed, to create a Benford Set in a strict sense. In fact, it is permissible under these circumstances to analyze the numerical distribution of the leading digits of each data quantity and to interpret it independently of Benford's Law.