Benford.Analysis Package

Anuj Kumar|Chen Chen|Mihir Sanghvi|Mithra Chintha|Samyuktha

Contents

Executive Summary

The paper contains information about detecting fraud in naturally occurring numerical data using Benford's Law. Team used Benford.analysis package to explore the Benford's law and the process to detect fraud in in-built datasets, census 2000-2010 and corporate payments audit data.

Introduction to Benford's Law

Benford's Law also known as first digit law or significant digit law, is an observation about the frequency distribution of leading digits in many real time numerical data. Discovery of which was made when Benford observed that the copies of the logarithm books had their beginning pages more worn out than pages dealing with higher digit, indicating that fellow scientists dealt with numbers with lower leading digits than higher ones. According to this law, in naturally occurring numbers, the leading digits is most likely to be small. For example, in sets which obey the law the number 1 appears as the most significant digit about 30% of the time, while 9 appears as the most significant digit less than 5% of the time. By contrast, if the digits were distributed uniformly, they would each occur about 11.1% of the time. This law had been extended to second digit, third digit and so on.

To understand how this law work, let us consider the Increment rise from 1 to 2 and 5 to 6 is same. But percentage change from 1 to 2 is 100% whereas from 5-6 is just 20%.

An individual attempting to commit fraud will be very unlikely to manipulate numbers in such a way as to conform with the Benford distribution. Even if the person cleverly succeeds in making up "random" numbers, the resulting distribution will still deviate from the expected distribution.

Purpose of the Benford's Law should not be used as a decision-making tool by itself, it may prove to be a useful screening tool to indicate that a set of financial statements deserves a deeper analysis.

Benford's Formulae

- Probability of first digit d1= log10(1+1/d1)
    - Ex: P(d1=1) = log10(1+1/1) =log10(2) =0.30
- Probability of second digit d2 = $\sum_{d1=1}^{9} \log10(1+1/d1d2)$
    - Ex: P(d2=1) = $\sum_{d1=1}^{9} \log10(1+1/d11)$ = log10(1+1/11) + ….. +log10(1+1/91) =0.1138
- Probability of first two digits d1d2= log10(1+1/d1d2)
    - Ex: P(d1d2=31) = log10(1+1/31) =0.0137

Dataset

We have used the Corporate Payment Dataset (sample dataset) from the "benford.analysis" package

Corporate Payment Dataset: Corporate payments of a West Coast utility company - 2010

- Feature used in the data set is Amount

Data Profiling

Data profile is the exploratory data analysis that precedes the primary Benford's Law tests, it might detect serious issues that show it isn't a good idea to continue with the analysis.

Example of this could be the there's error in the data or the data is incomplete, therefore, there's no point working on data like these. The data profile splits the data into strata and shows the count and sum for each stratum.

Data profiling of corporate payment amount

From	To	Count	PercentTotal	SumTotal	PercentSum
10.00	Inf	177763	93.82118541	492913582.26	1.005376e+02
0.01	9.99	7320	3.86340846	40159.47	8.191169e-03
0.00	0.00	123	0.06491793	0.00	0.000000e+00
-9.99	-0.01	195	0.10291867	-1121.31	-2.287092e-04
-Inf	-10.00	4069	2.14756954	-2674995.52	-5.456083e-01

Lower Values (between 0.01 and 50)

From	To	Count	PercentTotal	SumTotal	PercentSum
0.01	50	43253	22.82842	1188603	0.002424347

Higher Values (greater than 100,000)

	From	To	Count	PercentTotal	SumTotal	PercentSum
7	1e+05	Inf	370	0.1952816	242946614	0.4955287

Primary Benford's Law Tests

First Digit Test

The first digit test is an extremely high-level test and will only identify obvious anomalies, it is useful in following scenarios,

Scenario#1

This statistical test for a data might fail if there are repetitive transactions for the same amount in case of a hospital inventory or any goods inventory. Therefore, not all accounts not conforming can be named fraudulent and would require a deeper analysis.

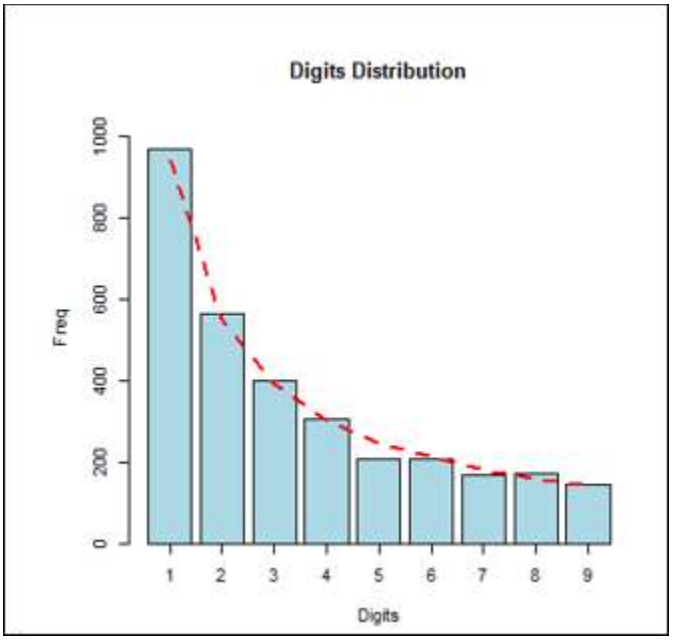
Scenario#2

The problem with the first digit test is that the first digits might show a conforming pattern even though the data has some serious issues that show that it doesn't conform to the spirit of Benford's Law (the uniform distribution of the mantissas). Biases can occur in marketing when supermarkets or discount chains set selling prices with ending digit 9s.

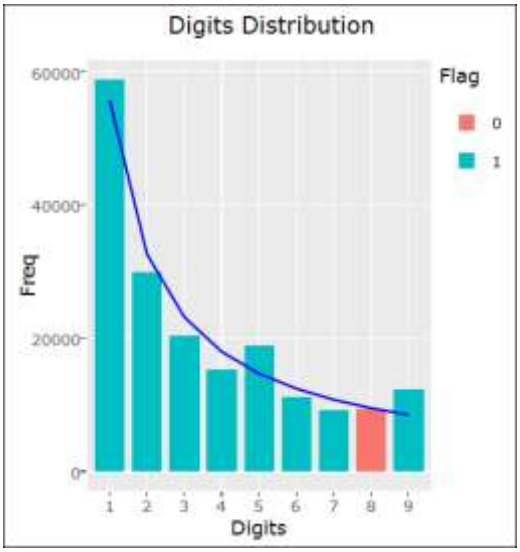
Test Result

In this example, we check for Benford's law conformity for corporate payment data.

The graph below shows the result of first digit test on the amount column of the corporate data set.



The package does not provide z-statistic, we have calculated the z-statistics and created the same plot as follows:



The results show that all the digits except digit 8 don't follow Benford's law. The graph show that some digits appear to have higher frequency such as '5', while some have lower than the normal threshold such as '4'. This means that if the dataset was really big it would lure us into sampling a major portion of the data without the knowledge of whether the other digits conform to Benford's law or not. Thus, first digit is not sufficient to deem an entire dataset as erroneous and will require further investigation.

Second Digit Test

The second digit test is the high-level test high level useful for spotting non-conformity or clustering around a threshold, it is useful in following scenarios

Scenario#1

The expected second digit proportions are less skewed than expected first digit proportions. The second digit test can be used to detect specific behaviors in corporate earnings reports such as rounding up the revenue as citied in the example.

Example - When controllers round-up their sales or net income numbers up from $998,000 to $1,002,000, there will be more second digit 0s than expected and fewer second-digit 9s and this change is noticeable using an analysis of second digits.

Scenario#2

This test also results in a large sample selection and cannot be used to select audit samples. However, it can be used to identify potential problems in a data set, especially if we assess conformity using the Z-statistic.

Example - A gasoline station that sells 3,000 gallons per day will usually have sales in the $10,000 to $12,000 range, and all first digits will equal 1. However, even with nonconforming first digits, the second digits should still conform to Benford's Law.

The other areas that this test might be useful in detecting issues are

- election counts
- inventory counts
- odometer readings
- daily sales numbers, etc.

The first and second digit graphs are highly aggregated. The first-two digits' test gives us more information than both of these tests.

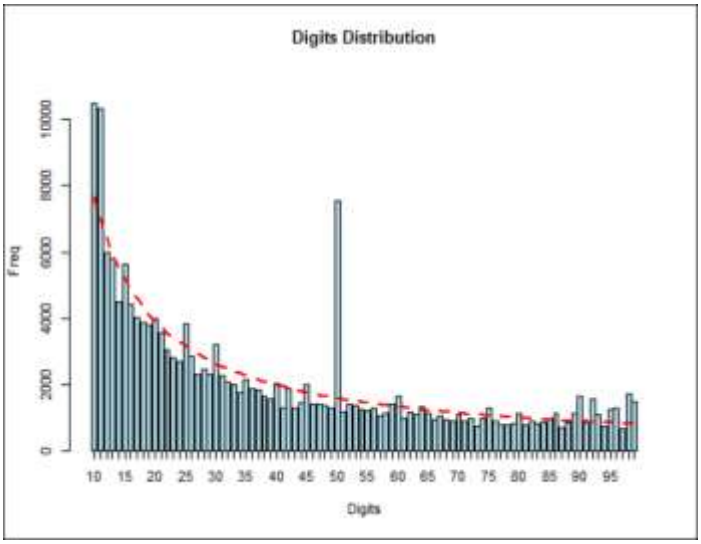
The package does not provide second digit test; however, we can populate the test using user defined function.

First Order Test

This test is a more focused test than the first digit's test and it is used to detect abnormal duplications of digits and possible biases in the data. Spikes that are just below a certain cutoff might be an indicator of potential fraud.

For example, if the non-taxable income limit is say 15,000$ for small businesses, and we see a spike at 14 in First order test, that may be due to self-employed taxpayers were managing their sales numbers just below 15,000$.

Test Results

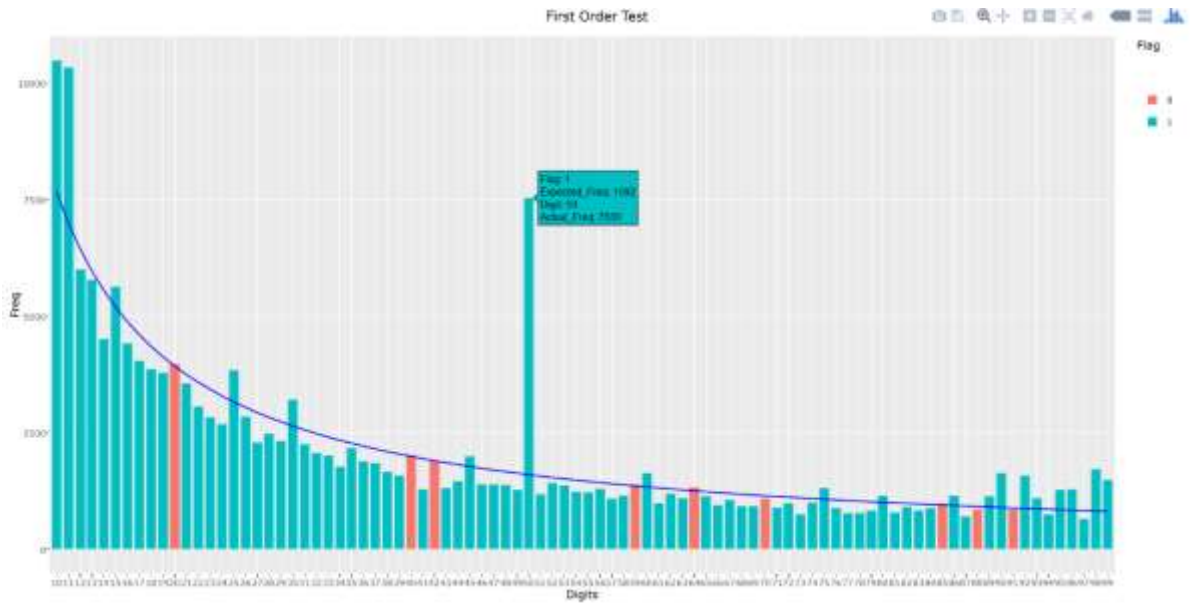


First Order test results for the corporate payment data shows spikes at 10, 11, 50, 98, and 99.

Z-Statistic for First Order Test

We can check the conformity of First Order test using Z-Statistic. By default, "benford.analysis" does not provide z-statistics. Thus, we have created our own function to create z-statistics for first order test.

It is difficult to check which digit has spikes from the default plots of the "benford.analysis" package. We have used functions from "ggplot" and "plotly" package to display same plot with tooltip and z-statistics.



From the plot, we can see that except few digits, all the digits have z-statistics that is more than 1.96 or less than -1.96 (significance level of 5 %) (Flag=1). It means that the First Order test does not confirm Benford's law.

In our z-stat calculation function we are returning data-frame with z-stat column and flag column that will be 1 if the z-stat is more than 1.96 or less than -1.96.

Advanced Benford's Law Tests

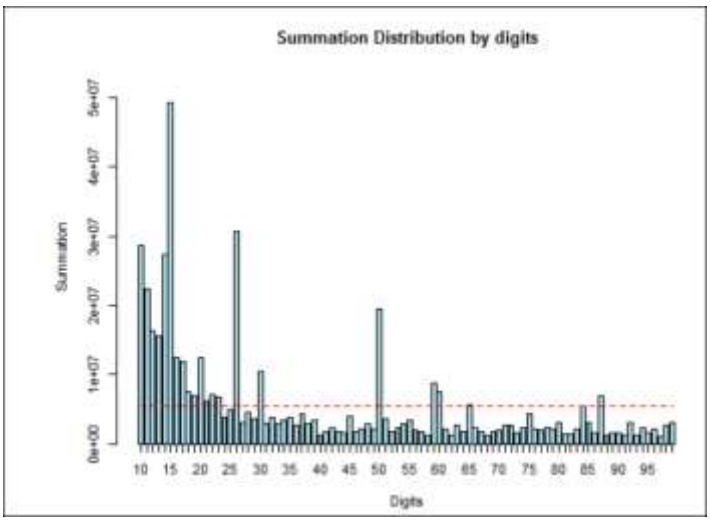
Summation Test

The summation test looks for excessively large numbers in a data field. The test identifies numbers that are large compared to the norm for that data.

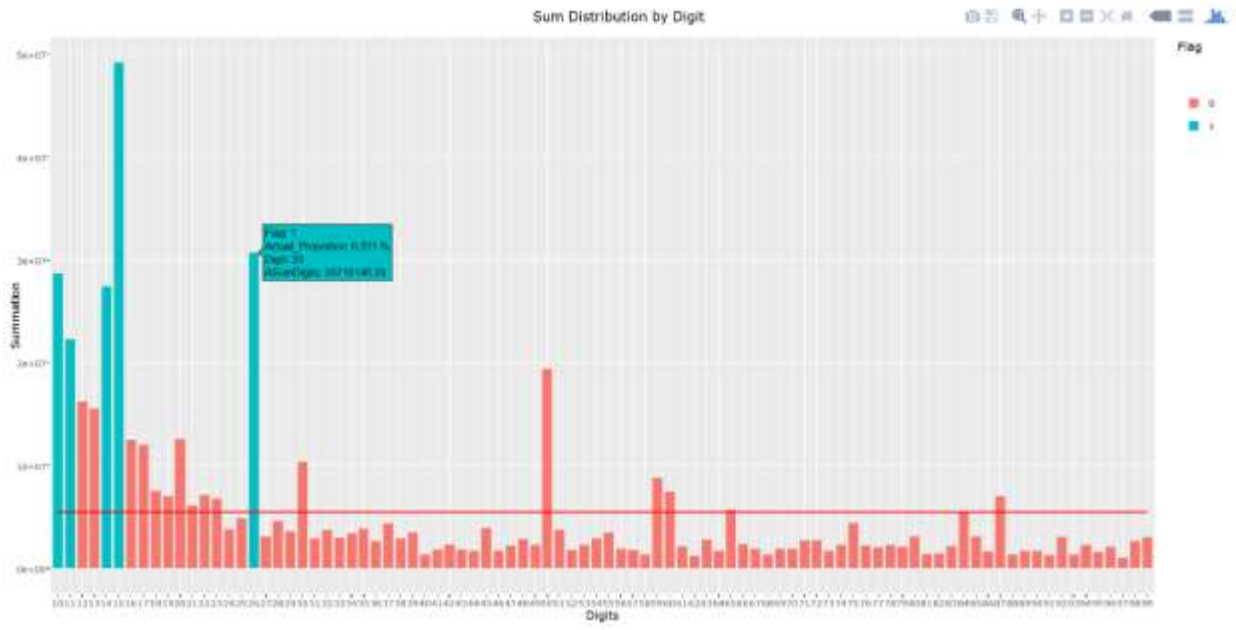
The summation test detects abnormally large transactions where abnormally large means large relative to the rest of the data.

In summation test we group the records of the first two digits and then compute the summation of each group to check if the uniform distribution is followed or not. In theory, the sums of numbers with the same first two digits should be equal.

Test Results



Test results for the corporate payment data shows spikes at 10, 11, 14, 15 and 26 and 50. Similar to the First Order Test, we have created a function that will plot summation test results, here we don't use Z-statistics. We have provided a cutoff percentage (4%), means that if value is more than 4% then the flag will be one.



First Order test showed some spikes at 98,99. Since we do not have spikes at the same places on the summation graph, it means that while the counts might have been higher than expected, these numbers are all relatively small and the sums are below average.

It is not clear from the spikes whether they were caused by one very large number or many medium-size numbers.

We have filtered payments starting with digit 15 and sorted it in the descending order and results were as follows:

```
library(dplyr)

data(corporate.payment)
bfa <- benford(corporate.payment$Amount,2)

DF <- getDigits(bfa,corporate.payment,c(15))
DF <- arrange(DF,desc(Amount))

head(DF,20)
```

```
> head(DF,20)
   VendorNum       Date                InvNum     Amount
1      16059 2010-02-18             000531002 15779215.2
2       7172 2010-11-24              112410NB  1500000.0
3      16721 2010-02-26                022610  1500000.0
4      16721 2010-03-05                030510  1500000.0
5      16721 2010-03-19                031910  1500000.0
6      16721 2010-04-09                040910  1500000.0
7      16721 2010-05-12                051210  1500000.0
8      16721 2010-06-18            1016900001  1500000.0
9      16721 2010-07-02            1018300001  1500000.0
10     16721 2010-07-15            1019600003  1500000.0
11     16721 2010-07-29            1021000001  1500000.0
12     16721 2010-08-19            1023100002  1500000.0
13     16721 2010-09-23 SEE ATTCH. BAL SHEET  1500000.0
14     16721 2010-10-22            1029400005  1500000.0
15     16721 2010-10-28            1030100003  1500000.0
16      2101 2010-12-30                 11543   159964.0
17     16906 2010-12-26              D9F8000   159600.0
18      5817 2010-12-27              V5101145  157471.7
19      2679 2010-11-11                155013  157197.1
20      5817 2010-02-25                022500  156231.0
```

There are large invoices of amount 1.5 million and one invoice of amount 15 million that caused the spike in digit 15. Similarly, we can analyze rest of the digits with large deviations from the expected values.

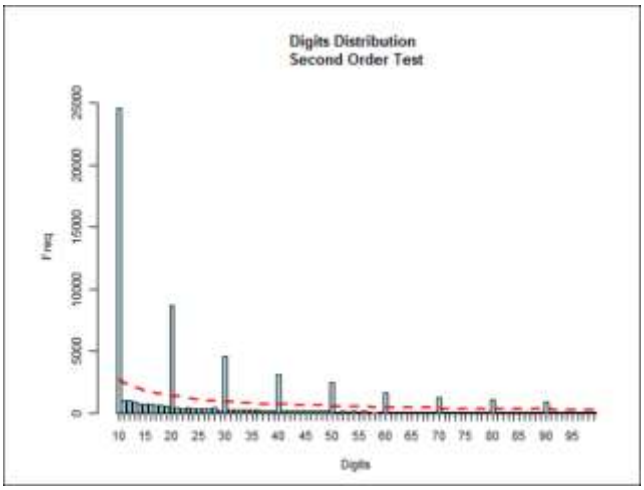
Second Order Test

This test is based on the first two digits in the data. To perform this test, we need to compute a new data column as follows:

1) The numeric data column is sorted from the smallest to largest
2) Difference between each pair of consecutive records is computed

Ideally the newly computed differences column should follow Benford's law, if the data analysis column follows the Benford's law.

Test Results

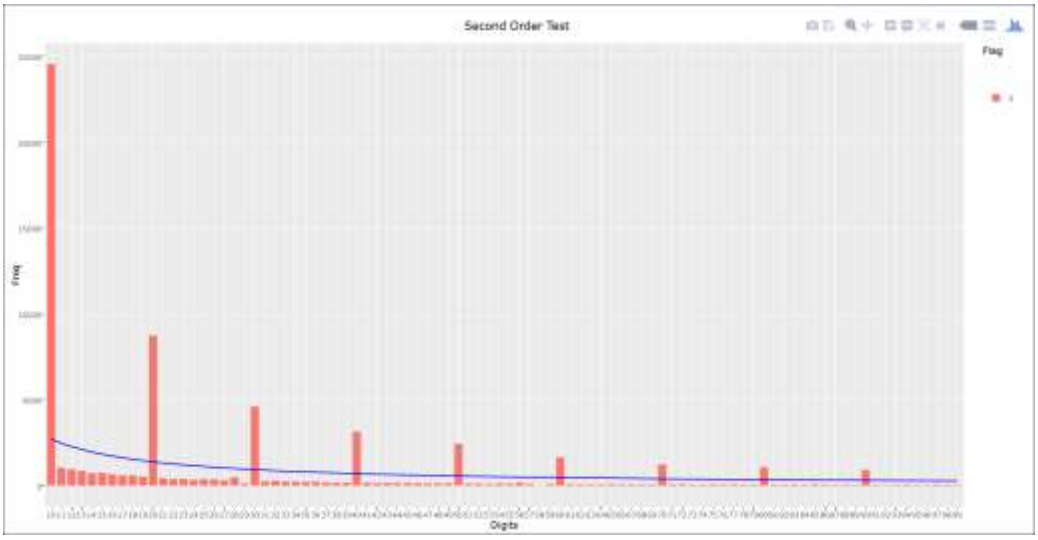


We can see spikes at 10, 20, 30, …, 90.

Z-Statistic for First Order Test

We can check the conformity of First Order test using Z-Statistic. By default, "benford.analysis" does not provide z-statistics. Thus, we have created our own function to create z-statistics for second order test.

Similar to the First Order Test, we have created a function that will plot summation test result plot with tooltip and z-statistics.



The results show non-conformity to Benford's law because all the digits are not following Benford's low.

Goodness-of-fit Statistics

We can use following statistics to assess the conformity to Benford's Law.

Z-Statistic

Z-Statistic is used to check if the actual proportion of a digit differs significantly from the expected digit as per the Benford's law.

Equation:

$$Z = \frac{|AP - EP| - \left(\frac{1}{2N}\right)}{\sqrt{\frac{EP(1-EP)}{N}}}$$

AP = Actual Proportion

EP = Expected proportion

N = total number of records

Chi Squared Test

Null hypothesis of the chi-square test in Benford's law is "Observed distribution of significant digits is same as the expected distribution"

Test results

Chi-square test for Corporate payment amounts

```
> benfordCorporate$`Chi-Squared Test`

        Pearson's Chi-squared test

data:  as.matrix(benfordDist[, c("ActualCounts", "BenfordsCount")])
X-squared = 1951.1, df = 8, p-value < 2.2e-16
```

As p value is less than 0.05, reject the hypothesis that the payment amounts distribution is following Benford's distribution for the first digit.

Chi-square test for USA census population data of 2010

```
> benfordCensus$`Chi-Squared Test`

        Pearson's Chi-squared test

data:  as.matrix(benfordDist[, c("ActualCounts", "BenfordsCount")])
X-squared = 5.5231, df = 8, p-value = 0.7005
```

As p value is greater than 0.05, failed to reject the hypothesis that the 2010 population distribution is following Benford's distribution for the first digit.

Mantissa Arc Test

Mantissa is the fractional part of the common logarithm

Important Stats related to Mantissa:
Mean of the Mantissa
Variance of the Mantissa
Kurtosis of the Mantissa
Skewness of the Mantissa

Null hypothesis of Mantissa arc test is mantissas of the numbers are uniformly distributed over the range [0,1).

Test Results

```
        Mantissa Arc Test

data:  corporate.payment$Amount
L2 = 0.0039958, df = 2, p-value < 2.2e-16
```

As p value is less than 0.05, reject the hypothesis that the mantissa of payment amounts are uniformly distributed.

MAD

The mean absolute deviation of a set of data is the average distance between each data value and the mean.
In the calculations, MAD test ignores N, thereby overcoming the problem related to large data sets. Like every coin has its two sides, the issue with this test is that there are no objective critical values.

The higher the MAD, the larger the average difference between the actual and expected proportions.

TABLE. Critical Values and Conclusions for Various MAD values (Nigrini–Assessing Conformity to Benford's Law)

Digits	Range	Conclusion
First Digits	0.000 to 0.006	Close conformity
	0.006 to 0.012	Acceptable conformity
	0.012 to 0.015	Marginally acceptable conformity
	Above 0.015	Nonconformity
Second Digits	0.000 to 0.008	Close conformity
	0.008 to 0.010	Acceptable conformity
	0.010 to 0.012	Marginally acceptable conformity
	Above 0.012	Nonconformity
First-Two Digits	0.0000 to 0.0012	Close conformity
	0.0012 to 0.0018	Acceptable conformity
	0.0018 to 0.0022	Marginally acceptable conformity
	Above 0.0022	Nonconformity
First-Three Digits	0.00000 to 0.00036	Close conformity
	0.00036 to 0.00044	Acceptable conformity
	0.00044 to 0.00050	Marginally acceptable conformity
	Above 0.00050	Nonconformity

Test Results

For first digits

```
> benfordCorporate1 <- benford(corporate.payment$Amount,1)
> MAD(benfordCorporate)
[1] 0.01321141
```

Results shows Nonconformity to Benford's law

For second digits

```
> benfordCorporate2 <- benford(corporate.payment$Amount,2)
> MAD(benfordCorporate2)
[1] 0.002336614
```

Results shows Nonconformity to Benford's law

Distortion factor

This statistic is used to show whether the data has an excess of lower digits or higher digits. It suggests that if numbers are overstated or understated.

Distortion Factor Equation is as follows,

DF = (AM – EM)/ (EM)

Where,
AM = Actual Mean
EM = Expected Mean

Mark Nigrini approximated standard deviation of the distortion factor as follows,

Standard Deviation (SD) = 0.638253 / SQRT(N)

To calculate the statistical significance of distortion factor, we need to calculate the Z-statistic as follows,

Z-Statistic = DF/ SD

Test Results

```
> bfa <- benford(corporate.payment$Amount,2)
> DF <- bfa$distortion.factor
> DF
[1] 0.5749073
```

The value of 0.5749 suggests that the numbers are overstated by 57.49%.

```
> SD <- 0.638253 / sqrt(nrow(corporate.payment))
> z_stat <- DF/SD
> z_stat
[1] 392.0805
```

The calculated Z-statistic of 392.0805 shows that the result is significant and there is en ough evidence to reject the null hypothesis that the DF equals 0.

Do's and Don'ts of Benford's Law

In some circumstances, a set of naturally occurring numbers may not always follow Benford's law due to human interventions.

The law applies only to some naturally occurring data such as,

- Purchase amounts, payment amounts, stock prices, accounts payable data, inventory prices and customer refunds
- Baseball statistics, areas of lakes, and the populations of towns—all of which Benford examined in his research but which are usually of less interest to accountants.

The Law does not apply to assigned values such as,

- telephone numbers, lottery tickets, sequential customer numbers or check numbers (all of which, by definition, cannot repeat).

It is advised to avoid using financial data that are not natural,

- Purchase amounts at a discount store because there often is a single price point per item.

- Values with upper limits, such as airline passenger counts per plane or employee days worked per year, do not lend themselves to such analyses.

Once we understand that the data is not following Benford's law we next proceed towards sampling the data to understand the reason, in which case it is necessary to sample "fairly" for further analysis

- Limiting a sample of invoices to values between US $100 and US $999 defeats the tests described here, because the data are limited to a narrow range.
- For small companies, using the complete data for an entire month or for a random day of each month is a better option.
- If the numbers are limited to a specific range, for instance, such as a price range of $7.99 to $9.99 that is set by the vendor, this set range will cause the pattern to be shifted or limited to a particular area.

Conclusion

Benford's Law is an excellent tool to predict the distribution of the first digit or first two' digits in a large population of data, given that the data has not been interfered with human interaction.

Given conformity to Benford's Law, one can use this as method of detecting possible fraudulent or errant transactions, insurance claims etc.

References

1. Nigrini, M. J. (2012). *Benford's Law Applications for Forensic Accounting, Auditing, and Fraud Detection*.
2. https://cran.r-project.org/web/packages/benford.analysis/benford.analysis.pdf
3. http://www.isaca.org/JOURNAL/ARCHIVES/2010/VOLUME-1/Pages/Using-Spreadsheets-and-Benford-s-Law-to-Test-Accounting-Data1.aspx
4. http://www.auditanalytics.com/blog/benfords-law-and-financial-statements/
5. https://en.wikipedia.org/wiki/Benford's_law
6. https://www.acfe.com/uploadedFiles/Shared_Content/Products/SelfStudy_CPE/Using BenfordsLawtoDetectFraud-2014Extract.pdf
7. http://www.isaca.org/Journal/Blog/Lists/Posts/Post.aspx?ID=69
8. file:///C:/Users/Mihir%20Sanghvi/Downloads/108-521-1-PB%20(1).pdf
9. https://blog.techandmate.com/benfords-law-and-algorithm-in-sas/

Appendix

Functions in 'Benford.analysis' Package

Function	Important Parameters	Usage
benford	data (a numeric vector) number.of.digits (how many first digits to analyze)	This function validates a dataset using Benford's Law. Its main purposes are to find out where the dataset deviates from Benford's Law and to identify suspicious data that need further verification. Returns Benford object
getSuspects	bfd (Benford object) data (original Data used for the Benford's analysis) by (method selecting how to order digits)	Gets the 'suspicious' observations in a data frame according to Benford's Law
suspectsTable	bfd (Benford object)	gives a data frame with the first digits and the differences from Benford's Law in decreasing order
plot	bfd (Benford object)	Plot the Benford object Which gives digits' distribution with respect to frequency, summation and chi-square

User Defined Functions

Function	Parameters	Usage
DataProfiling	column (numeric vector which is used for the Benford analysis	Returns the data frame of informative summary
benfordFrequencyDistribution	numericData (numeric vector which is used for Benford analysis) positive (if 1 then only positive values are considered , if 0 then both positive and negative)	Returns Benford object which contains number of observations used, Benford distribution counts, chi-square test results, mantissa distribution data
zScoring	Data Column Numeric Vector and Order (1 or 2)	Calculates Z-Stat for First order and second order results
zScoringFirstDigit	Data Column Numeric Vector	Calculates Z-Stat for first digit results
SumDist	Data Column Numeric Vector	Calculates Summation statistics that will be used to make summation test plots