Validity of Financial Statements: Benford's Law

Ivaan Shrestha

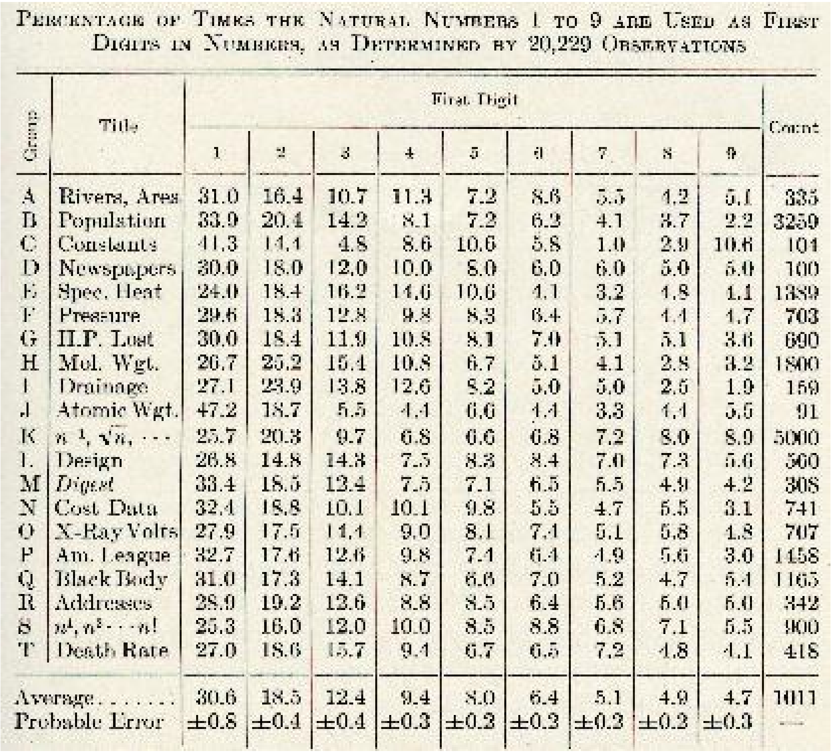
Indiana, United States

Abstract

Benford law states that the occurrence of digits from 0-9 in a large set of data is not uniformly distributed but instead in a decreasing logarithmic distribution with 1 occurring at most. Almost all set of data follows this trend however this law is widely used as a base for various fraud detection and forensic accounting. Benfords law is an observation that leading digits in data derived from measurements doesnt follow uniform distribution. Different financial statements such as cash flows, income statement and balance sheet of the 20 tech companies of the Fortune 500 are analyzed in this project. Cash flow is the net amount of cash and cash-equivalents moving into and out of a business. Income statement is a financial statement that measures a company's financial performance over a specific accounting period. Balance sheet is a financial statement that summarizes a company's assets, liabilities and shareholders equity at a specific point in time. All of these data of financial statements are extracted from Morning Star database and are analyzed by Python program written by me.I also wrote the Python program to calculate Benford's second digit and third digit probability using the formula. I would like to thank Prof. Erin Wagner and Dr. Courtney Taylor for helping in this research project.

1. Introduction

The first known reference to the logarithmic distribution of this phenomenon dates back to 1881, when the American astronomer Simon Newcomb noticed how much faster the first pages wear out than the last ones[1]. After some 50 years, physicist Frank Benford rediscovered this law and supported it with more than 20,000 phenomenon happenings such as heats of chemical compounds, baseball statistics, paper and newspapers. Some may have argued that the Benford manipulated round-off errors to obtain better fit, but the without the manipulation also data were really close. Later in 1938 this was published as is known as Benfords law. Some of the areas where the numerical data do not follow this trend are telephone numbers as it starts with particular digit, lotteries are distributed uniformly, heights of human adults, square root tables of integers and so forth [2]



PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Figure 1: Benfords original data from; reprinted courtesy of the American Philosophical Society

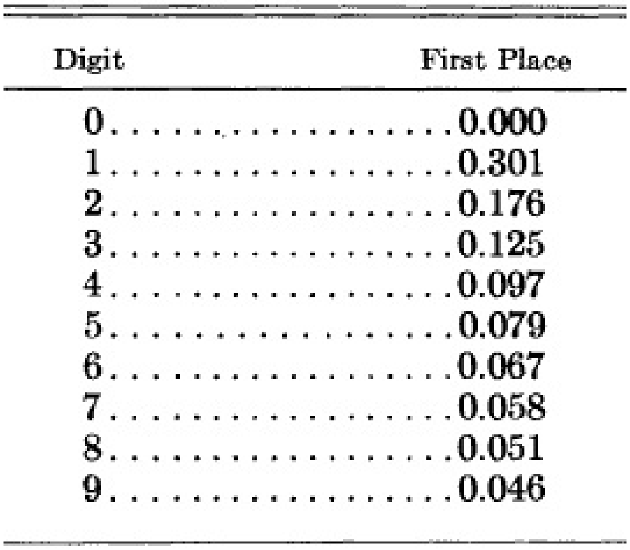
A set of numbers is said to satisfy Benford's law if the leading digit d occurs with probability

$$P(d) = log_{10}(d+1) - log_{10}(d) - log_{10}(\frac{d+1}{d}) = log_{10}(1+\frac{1}{d}) \qquad (1)$$

[3]

d is in (1,2,.....,9)

For example: In 65789, leading digit is 6 whereas in number 112489 leading digit is 1. The figure below helps to understand:



Digit	First Place
0	0.000
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

Figure 2: A table showing d as Digit and probability P(d) as First Place[1]

Here you can see that there is 0.301 probability of 1 being a first digit, 0.176 probability of 2 being a first digit and the probability decreases as the number increases where at the end probability of 9 being a first digit is 0.046.Here are the list of tech companies that were used in my research:

Company	**Company**	**Company**
Amazon	Apple	Western Digital
Cisco Systems	Computer Science Corps	Qualcomm
Danaher	eBay	Texas Instruments
EMC	Thermo Fisher Scientific	Xerox
Google	HPQ	Microsoft
IBM	Intel	Oracle
Jabil Circuit	Micron Technology [4]	

Table 1: Companies that were analyzed in this research

2. Method

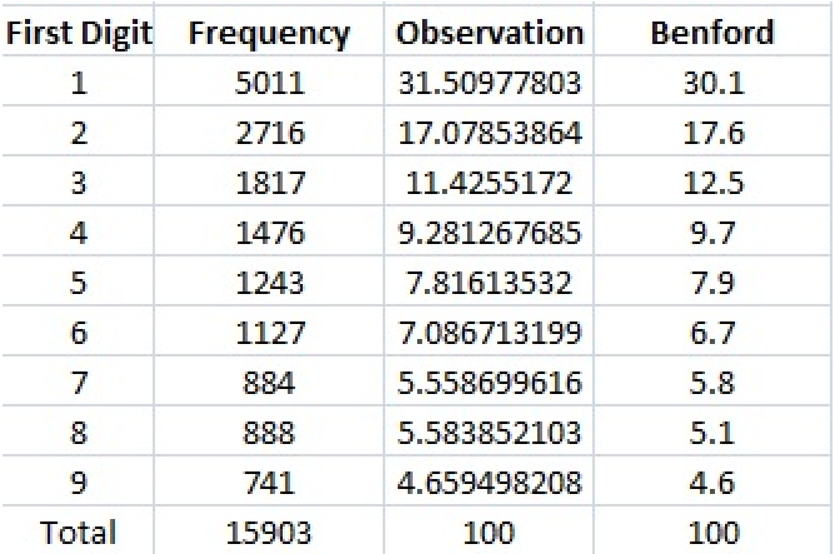
The top 20 tech companies from Fortune 500 are used in this project and the list can be found in the table above. The financial statements of these companies were retrieved from the Morningstar database provided by Investment Research Center.There are three kinds of financial statement that are included in this project. The Cash flow statement shows net amount of cash and cash-equivalents moving into and out of a business. Income statement is a financial statement that measures a company's financial performance over a specific accounting period. Balance sheet is a financial statement that summarizes a company's assets, liabilities and shareholders equity at a specific point in time. Using the Excel data from web each of these statements are extracted in the excel format. For each of these statements the data analyzed are from 2005 to 2014, and are in each column of the excel spreadsheet. The format of data is that each sheet will have particular kind of statement(income, balance sheet or cash flow statement) which will hold the data from 2005 to 2014, whereas each file will have all the statements for particular company. The Python program will extract each cell from each column from sheet and will convert it into string where it will read occurrence of each first digit, second digit and third digit and store the values in the variable. Once we have these values, each of these values are stored in excel sheet where all the company's first digit, second digit and third digit will be added and series of plot and histograms were made.

3. Results

The data of all the financial statements were analyzed by the method described in the method section.

3.1. First Digit

At first all the data from all the financial statements of each company were analyzed with first leading digit of the numbers. There were in total of 15,903 numbers and first digit of each number was analyzed and according to the number the parameter counter was increased.If the number was negative the Python program would replace the first string with the second string. Here is the result for the first digit for each number in statements.



First Digit	Frequency	Observation	Benford
1	5011	31.50977803	30.1
2	2716	17.07853864	17.6
3	1817	11.4255172	12.5
4	1476	9.281267685	9.7
5	1243	7.81613532	7.9
6	1127	7.086713199	6.7
7	884	5.558699616	5.8
8	888	5.583852103	5.1
9	741	4.659498208	4.6
Total	15903	100	100

Figure 3: Figure of table showing the observation for leading first digit [5]

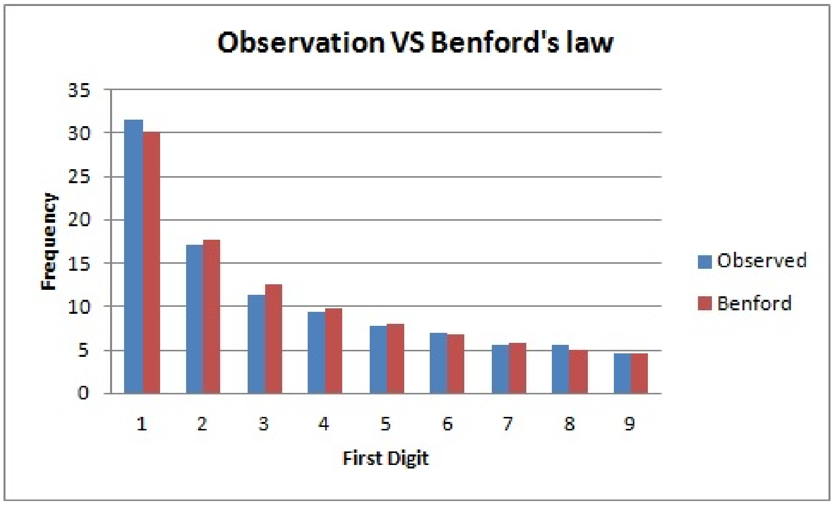


Figure 4: Histogram comparing Benford's probability against actual observation

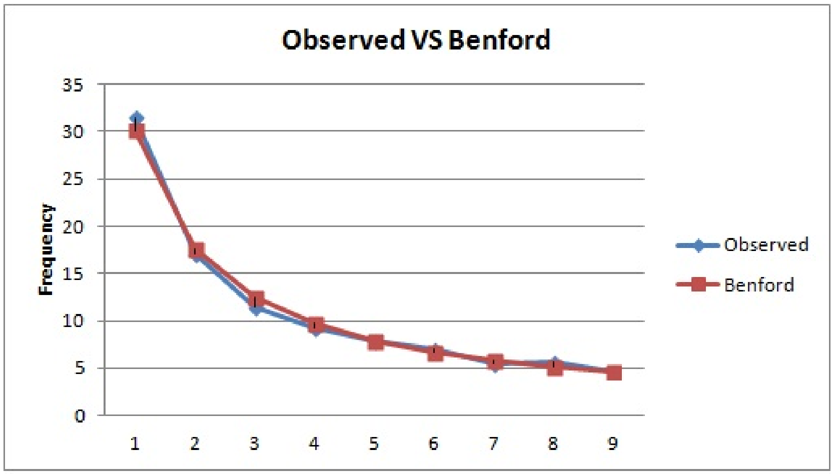


Figure 5: Graph comparing Benford's probability against actual Observation

The above graph will be able to provide exact differences between the observed and Benford's law.

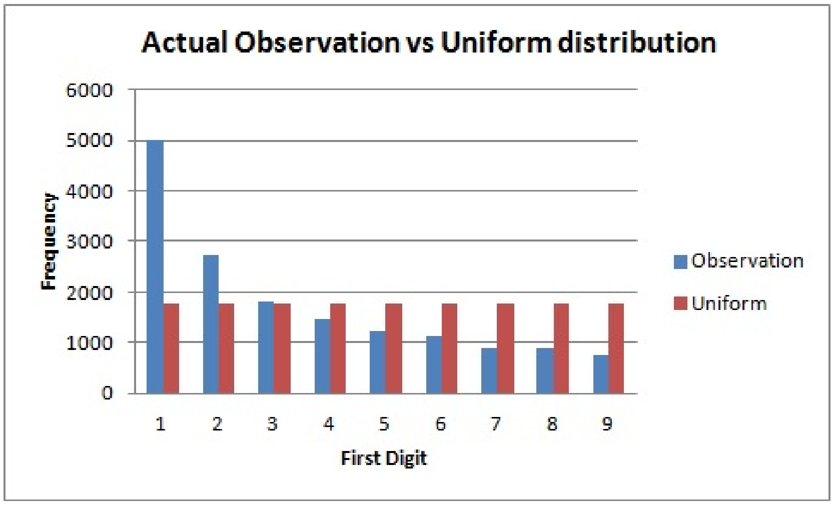
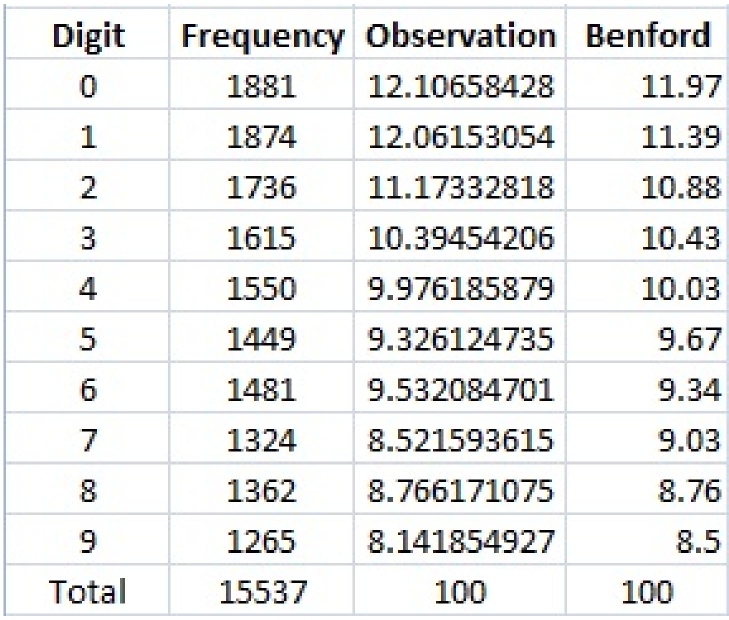


Figure 6: Histogram comparing uniform distribution against actual observation

3.2. Second Digit

There were total of 15,537 numbers analyzed for second leading digits in all financial statements of all company.



Digit	Frequency	Observation	Benford
0	1881	12.10658428	11.97
1	1874	12.06153054	11.39
2	1736	11.17332818	10.88
3	1615	10.39454206	10.43
4	1550	9.976185879	10.03
5	1449	9.326124735	9.67
6	1481	9.532084701	9.34
7	1324	8.521593615	9.03
8	1362	8.766171075	8.76
9	1265	8.141854927	8.5
Total	15537	100	100

Figure 7: Table showing the second leading digits observation

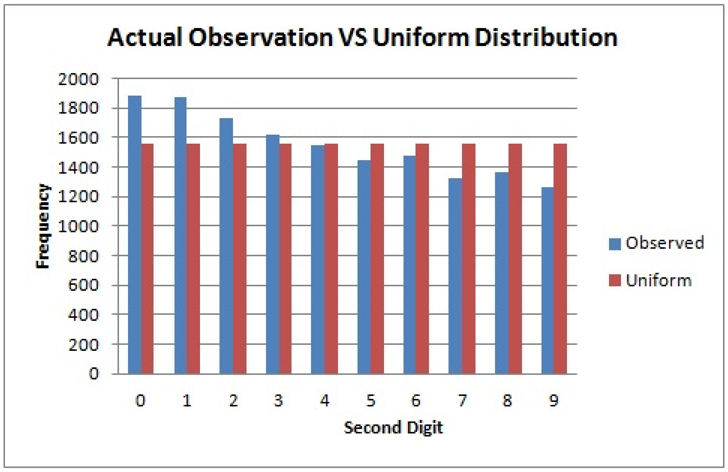


Figure 8: Histogram comparing uniform distribution against actual observation

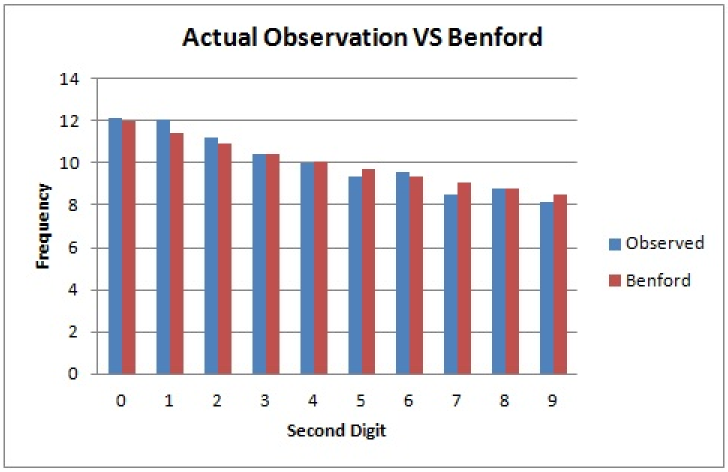


Figure 9: Histogram comparing Benford's law against actual observation

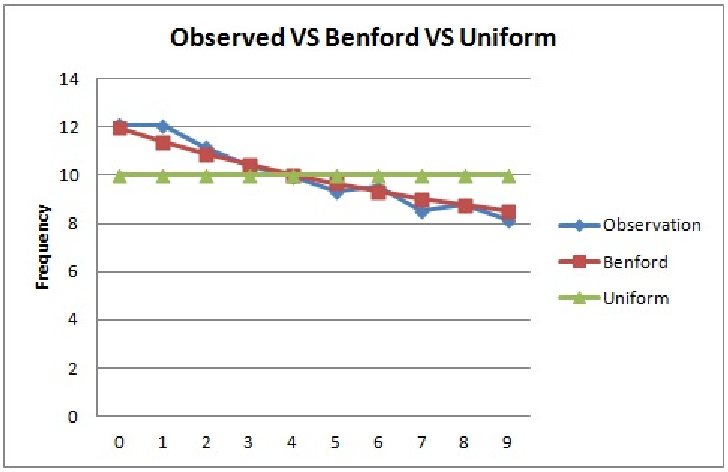
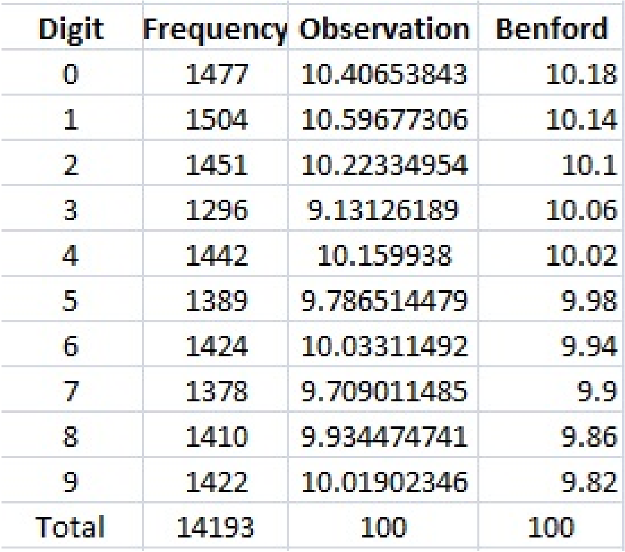


Figure 10: Graph comparing uniform distribution against actual Observation against Benford's law for leading second digit

We can observe in this section that the digit 0 has also been added in both Benford's law and also in my comparison. The observation almost follows the Benford's law.

3.3. Third Digit

There were total of 14,193 numbers analyzed for third leading digits in all financial statements of all company.



Digit	Frequency	Observation	Benford
0	1477	10.40653843	10.18
1	1504	10.59677306	10.14
2	1451	10.22334954	10.1
3	1296	9.13126189	10.06
4	1442	10.159938	10.02
5	1389	9.786514479	9.98
6	1424	10.03311492	9.94
7	1378	9.709011485	9.9
8	1410	9.934474741	9.86
9	1422	10.01902346	9.82
Total	14193	100	100

Figure 11: Table showing the Benford's probability and Observed frequency for third digit.

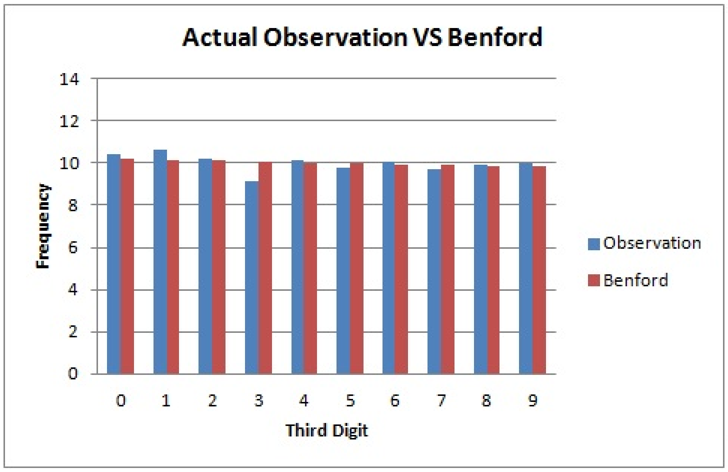


Figure 12: Graph comparing actual Observation against Benford's law

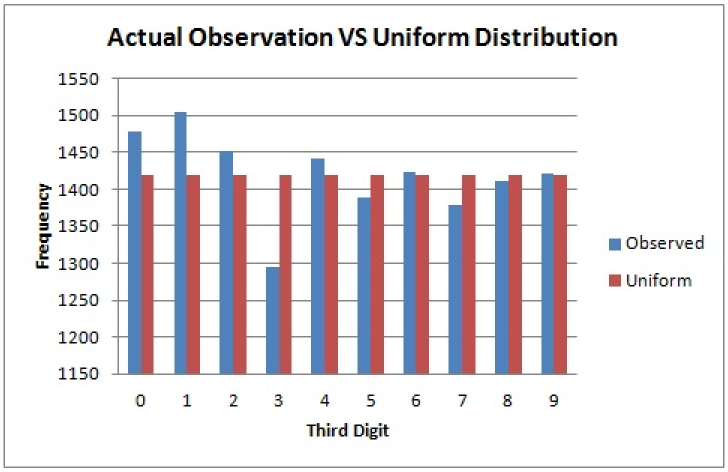


Figure 13: Histogram comparing uniform distribution against actual observation for third leading digit

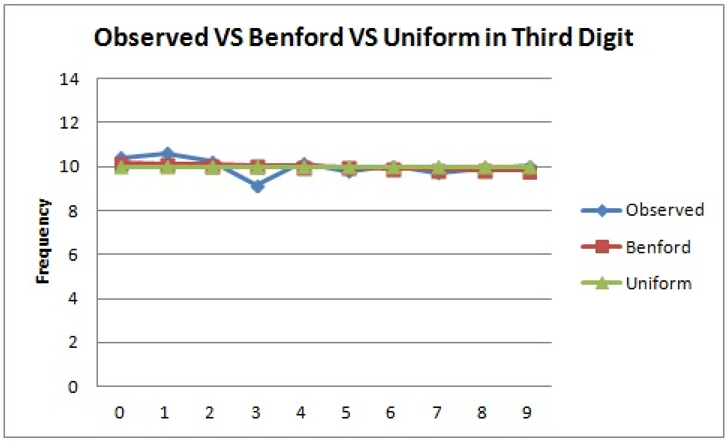


Figure 14: Graph comparing uniform distribution against actual Observation against Benford's law for third leading digit

The above table helps to closely analyze the differences between frequency probability of third leading digit for actual observation against Benford's law probability frequency against uniform distribution.

3.4. Small Data

I have added this section in my research to understand if Bendford's law holds for first leading digit if the data set gets smaller and smaller. Here are the results.

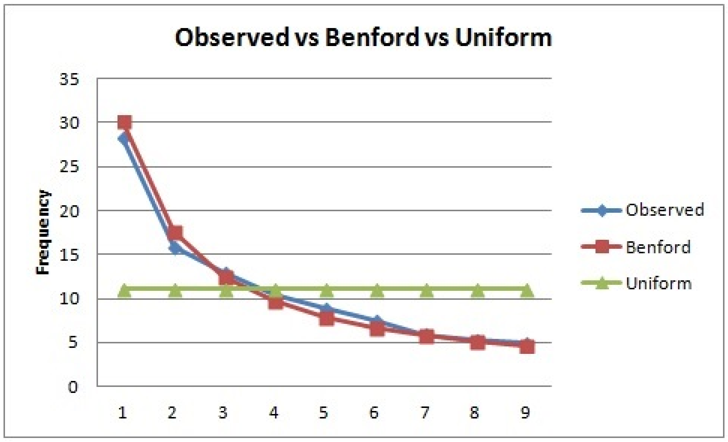


Figure 15: Graph comparing uniform distribution against actual Observation against Benford's law for first leading digits with 2,348 numbers from Amazon, Apple and Cisco

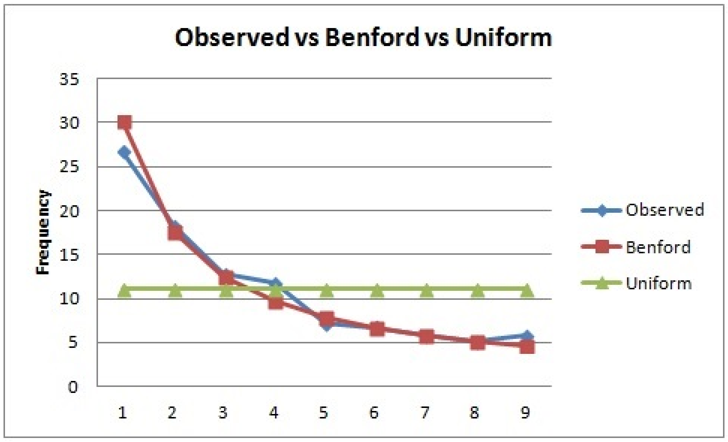


Figure 16: Graph comparing uniform distribution against actual Observation against Benford's law for first leading digits with 791 numbers from Amazon.

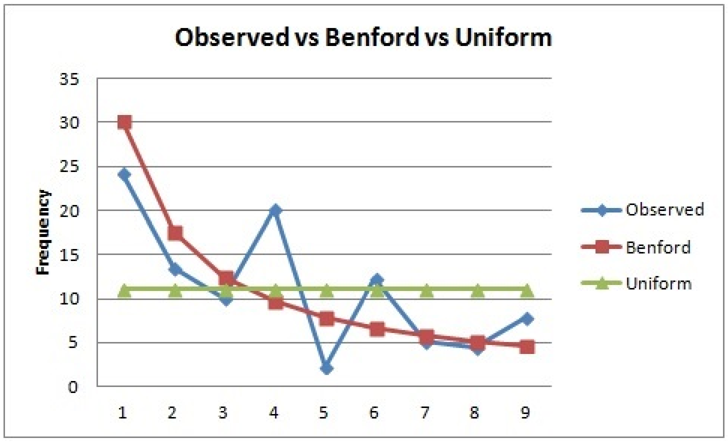


Figure 17: Graph comparing uniform distribution against actual Observation against Benford's law for first leading digits with 178 numbers from Amazon's income statement

It can be seen here from these graphs that as the numbers in the data set decreases it stops following the Benford's law neither are close to uniform distribution. In fact the values are just random.'

4. Analysis

The Chi-Square test was performed on all the observed values and the Benford's probability on all the three leading digits. The Chi-Square test was also done for small data set. The absolute value of maximum and minimum difference observed between the data set against Benford's probability is also analyzed for all the observations. The null hypothesis and alternative hypothesis for each data set will be:

- H0: Observed data set and Benford's probability are related.
- Ha: Observed data set and Benford's probability are not related.

The significance level will be 0.5 for all of the Chi-Square test.

4.1. Chi-Square Test

For first digit after the Chi-Square test the p value was 0.999987241, which means that the observed data almost followed the Benford's law. The p value

is more than 0.5, thus I accepted the null hypothesis and observed data set and Benford's probability are related.
For second digit after the Chi-Square test the p value was 0.999999962, which means that the observed data almost followed the Benford's law. The p value is more than 0.5, thus I accepted the null hypothesis and observed data set and Benford's probability are related. In fact the observed data for leading second digit follows Benford's law better than the observed data for leading first digit.
For third digit after the Chi-Square test the p value was 0.999999924, which means that the observed data almost followed the Benford's law. The p value is more than 0.5, thus I accepted the null hypothesis and observed data set and Benford's probability are related. The observed data for leading third digit follows the Benford's law better than any other observations.
In small data set of 2348, 791 and 178 numbers, on the first digit after the Chi-Square test the p value was 0.999728129, 0.99628304 and 0.001370532 respectively. This means that the observed data for first two data set almost followed the Benford's law. The p value is more than 0.5, thus I accepted the null hypothesis for both data set and observed data sets and Benford's probability are related. Whereas in the last data set, the p value was less 0.5 and thus I have to reject the null hypothesis and observed data set and Benford's law probability are not related.

4.2. Absolute Error

For the leading first digit, the absolute value of maximum error is 1.409 whereas minimum error is 0.059. The maximum error was in number 1 whereas minimum error was in number 9. For the leading second digit, the absolute value of maximum error is 0.671 whereas minimum error is 0.006. The maximum error was in number 1 whereas minimum error was in number 8. For the leading third digit, the absolute value of maximum error is 0.928 whereas minimum error is 0.074. The maximum error was in number 3 whereas minimum error was in number 8.
For leading first digit small data set with 2348 numbers, the absolute value of maximum error is 1.863 whereas minimum error is 0.077. The maximum error was in number 1 whereas minimum error was in number 7. For data set with 791 numbers, the absolute value of maximum error is 3.424 whereas minimum error is 0.0003. The maximum error was in number 1 whereas minimum error was in number 6. For data set with 178 numbers, the absolute value of maximum error is 10.525 whereas minimum error is 0.605. The

maximum error was in number 4 whereas minimum error was in number 8.

5. Conclusion

There were couple of things that I found was something that was not expected. The p value for Chi square test kept on getting close to 1 as digit placed was increased from first to third digit. The peculiar thing I noticed was that from graph in Figure 14, there is a slight noticeable drop in frequency of number 3, but yet this data set had the best p value. If you compare the graph of Fig 5 with Fig 14, it seems that p value of leading first digit should have the best p value, however, this is not the case. All the data set somewhat followed the Benford's law except the smallest data set that only had income statement of Amazon. The maximum error value increased as the data set got smaller. There was no particular pattern in the maximum and minimum error values between the leading digits from first to three. The financial statements did follow the Benford's law, therefore I can conclude that the financial statements of these company are authentic. We know that each of company financial statements are authentic because if it was not, there would be significant change in error values. From analyzing the error value, we can find out which number (0,1,2...,9) has the highest error and we look in each file of company to find out which company has highest or lowest frequency of that particular number. Thus the fraud can be detected. In this research I observed that mostly number 1 has maximum error values whereas the minimum error values had different numbers in each observation. Lastly, I can also conclude that the more the number of data, more accurate will be the observation as we saw in this research that when I decreased the number of data, the p value started decreasing and maximum error value started increasing.

6. Bibliography

[1] F. Benford, The law of anomalous numbers, Proceedings of the American Philosophical Society (1938) 551–572.

[2] T. W. Singleton, Understanding and applying benford's law, ISACA Journal 3 (2011) 1–4.

[3] C. Durtschi, W. Hillison, C. Pacini, The effective use of benfords law to assist in detecting fraud in accounting data, Journal of forensic accounting 5 (2004) 17–34.

[4] Fortune 500, http://fortune.com/2015/06/13/fortune-500-tech/, ????

[5] Morning star, http://library.morningstar.com/, ????