Empirical Evidence of Benford's Law in Small Cap Stock Markets and Its Applications

Jacob S. Mason

11/11/2013

In this paper, the leading digits of the prices of securities from the Russell 2000 index are examined and evidence is shown that the distribution of the leading digits of securities from the Russell 2000 index conform to the distribution proposed by Sam Benford for naturally occurring sets of numbers commonly known as Benford's Law. Also, transition Markov matrixes are formed and they suggest that, in the short run, the most likely leading digits transition will be from a leading digit 9 to a leading digit 1. This information is then used to examine whether profitable returns can be generated using this information.

Statement of the Problem

In 1881, an American astronomer named Simon Newcomb made a discovery while examining a book of logarithmic tables. He discovered that the beginning pages of the book that contained numbers beginning with 1 were more worn than the back pages of the book that contained numbers having higher leading digits. Based on this discovery, he then published a paper stating that the probability of a number N being the leading digit of a number is equal to log(N+1)-log(N). Frank Benford, for whom the law is named, later extended this idea 1938 to show that US population totals, death rates, the length of rivers, physical constants, molecular weights, etcetera all follow the distribution proposed by Sam Newcomb. Populations that have been shown to conform to Benford's Law are naturally occurring sets of numbers that span more than one magnitude of 10 (1-1000, not 1-10) and have no set limit or cut off point. The price of small cap stocks, represented in this paper by the Russell 2000 index, should theoretically conform to each of these criteria. They are naturally occurring, span more than one magnitude of 10, and have no strict maximum price. This being the case; a random sample of small cap stock price leading digits should conform somewhat to the distribution proposed by Benford's law. If the evidence suggest that they do in fact conform, then this conforming to Benford's law supplies investors with additional information regarding small cap stock markets; and any additional information can potentially be used to generate an above average return. This paper seeks to find empirical evidence that the distribution of leading digits of the securities of the Russell 2000 do in fact conform to Benford's law and that this information can be used to generate a positive return on investment.

Literary Review

Only a marginal amount of research has been done in applying Benford's Law to stock markets and no paper was found that examined the actual prices of the stocks within any market.

A paper published in 1994 by Eduardo Ley focused on a series of one day returns of both the S&P 500 and the Dow Jones Industrial Average. Eduardo found statistical evidence that a series of one day returns did conform to Benford's law.

Mario Zgela published a similar paper in 2011 which showed that the percent changes in the value of the Deutscher Aktien IndeX (DAX) for a 10 year period did not conform to Benford's Law. This, she speculated, could be the result of human influence on the DAX that would make the percent changes no longer a set of natural or non-influenced numbers.

No research was found that used the leading digits of actual securities instead of indexes or index returns. Also, no research could be found to show that knowledge of the conformity to Benford's Law could be used to generate returns.

Empirical Evidence

Why small cap stocks and the Russell 2000?

In order for a set of numbers to conform to Benford's law, they need to be naturally occurring, span several magnitudes of 10, and have no strict limitations. Given this criteria, it could be hypothesized that a distribution of the leading digits of every security traded in the US stock markets should conform to the Benford's law distribution. This could be the case but for the purposes of this paper the entire US stock market is too massive a population and would require an excessively large sampling frame to achieve a representative sample. Instead, it is simpler to select one index and use that as a sampling frame. The first index considered for this use was the S&P 500 which consists of the 500 largest market capitalization firms listed on either the NYSE or the NASDAQ. The stock prices of large cap companies generally do not span more than one magnitude of 10 and do not frequently change (low volatility = low risk). For the purposes of this paper; which are to see if the leading digits of some set of securities follows Benford's law and to see how those leading digits change over time, the S&P 500 index would not be appropriate. The next index that was selected was the Russell 2000 which contains approximately 2000 small capitalization firms that range from 150 million to a little over 3 billion in market capitalization. Small cap stocks have a higher likelihood of conforming to Benford's Law because their prices have a greater range (from a less than a dollar to more than 150 dollars) and they are generally riskier than large cap stocks. This additional risk is partly due to fact that their prices tend to have a greater variance than those of large cap stocks. Think of ships on a sea. Little ships will move more and be affected by smaller waves whereas the larger ships will tend to remain fairly level.

Since the sampling frame of this paper is limited to only stocks contained within the Russell 2000, any findings could potentially be applied to all small cap stocks but not necessarily to mid or large capitalization stocks.

Examination of Samples Taken From The Russell 2000

Four exploratory samples of randomly selected securities from the Russell 2000 were taken and examined to see if there was any evidence that a distribution of their leading digits would conform to the distributions hypothesized by Benford's Law.

Date	**08/27**	**08/28**	**08/29**	**09/02**	**Benford's Distribution**
Sample Size (n)	50	100	100	200	
1	0.320	0.260	0.260	0.320	0.301
2	0.280	0.240	0.230	0.175	0.176
3	0.140	0.130	0.120	0.110	0.125
4	0.100	0.060	0.050	0.120	0.097
5	0.060	0.060	0.130	0.075	0.079
6	0.020	0.100	0.070	0.060	0.067
7	0.040	0.050	0.040	0.060	0.058
8	0.040	0.040	0.040	0.060	0.051
9	0.000	0.060	0.060	0.020	0.046
χ^2	5.1434	7.1731	9.0404	5.1491	
Prob > χ^2	0.6425	0.5181	0.3389	0.7415	

A Pearson's χ^2 test was performed on each of the samples and the results indicated that the distribution of samples were not statistically different from the hypothesized Benford's law distribution. The leading sample of size 50 may lack sufficient sample size to conduct an accurate χ^2 test and should not be regarded to hold the same weight as the tests conducted on the other samples. Not taking into consideration the first sample of size 50, the other samples do seem to indicate that the leading digits of the Russell 2000 conform to the Benford's law distribution.

Since the initial samples taken of the Russell 2000 strongly suggested that the leading digits of all the securities in the Russell 2000 conform to Benford's Law, a final sample was taken of every stock that was included in the most recent composition of the Russell 2000 (composed on June 28 2013) and that was still trading in the US securities markets.

Date	**09/04**	**Benford's Distribution**
Sample Size (n)	1968	
1	0.309	0.301
2	0.219	0.176
3	0.131	0.125
4	0.082	0.097
5	0.076	0.079
6	0.062	0.067
7	0.048	0.058
8	0.038	0.051
9	0.035	0.046

χ^2	41.5437
Prob > χ^2	<.0001

While the χ^2 test performed on this sample indicates that the distribution of this sample differs statistically from the distribution proposed by Benford, this distribution does not appear to differ from Benford's by a practically significant amount. The distribution of the whole Russell 2000 may differ statistically from Benford's proposed distribution but it does not differ practically from it. The statistically significant difference could be due to the fact that we are dealing with a large and unbalanced data set which has been shown to cause an error in the χ^2 test. This sample was also taken on just one day. An average of multiple samples of the Russell 2000 could give a better idea as to the true distribution of the leading digits.

Even though the distribution of the leading digits of the stock prices found in the Russell 2000 do not appear to strictly follow Benford's distribution, it does appear that they follow a distribution somewhat similar to Benford's especial in the relative sizes of the number of leading digit 1 and leading digit 9. Since there appears to be a large percentage of leading digit 1 securities and a small percentage of leading digit 9 securities, these disproportionate distributions could affect the transitional probabilities of a leading digit 9 transitioning to a leading digit 1. If a greater number of leading digit 9 securities changed into leading digit 1, some type of return might be generated by buying a leading digit 9 and selling when it becomes a leading digit 1. In order to test this hypothesis, transition probabilities for these securities were examined.

Transitional Probabilities, Markov Chains, and Limiting Probabilities

Even though the leading digits of the Russell 2000 appear to conform somewhat to Benford's law, this information is only helpful to investors if they can use it to generate a return. If the conformity of the Russell 2000 to a Benford-like distribution affects the transition probabilities of these securities, then investors could use that information to generate a return by buying low (leading digit 9) and selling high (leading digit 1).

To examine the transitional probabilities of the securities with different leading digits in the Russell 2000, three Markov chains were created using 5, 10, and 15 day transition periods.

A Markov chain is a transitional probability matrix that shows the probability of a unit transitioning from state *i* to state *j* for a given time period having no other prior knowledge of the unit except that it began in state *i.* A Markov Chain is only appropriate to be used in situations in which the probability of moving from state *i* to state *j* is influenced only by the knowledge of that the unit started in state *i.*

In the large sample taken on September 4 2013, the smallest group of securities were those with a leading digit 9 which consisted of only 68 securities. Because of this, each of the samples of leading digits securities taken for the Markov chains contained 68

randomly chosen securities except for the leading digit 9 group which contained the entire sample of 68 securities.

5 day		*Ending State*								
		1	2	3	4	5	6	7	8	9
Beginning State	1	0.9323	0.0677	0	0	0	0	0	0	0
	2	0.0147	0.9706	0.0147	0	0	0	0	0	0
	3	0	0	0.9853	0.0147	0	0	0	0	0
	4	0	0	0.0147	0.8235	0.1618	0	0	0	0
	5	0	0	0	0.0441	0.897	0.0589	0	0	0
	6	0	0	0	0	0.0588	0.7941	0.1324	0.0147	0
	7	0	0	0	0	0	0.0294	0.7353	0.2353	0
	8	0	0	0	0	0	0	0.0441	0.7353	0.2206
	9	0.1765	0	0	0	0	0	0	0	0.8235

15 day		*Ending State*								
		1	2	3	4	5	6	7	8	9
Beginning State	1	0.8667	0.1333	0	0	0	0	0	0	0
	2	0.0441	0.8971	0.0588	0	0	0	0	0	0
	3	0	0.0441	0.8382	0.1177	0	0	0	0	0
	4	0	0	0.0294	0.603	0.3382	0.0294	0	0	0
	5	0	0	0.0147	0.0294	0.7206	0.2353	0	0	0
	6	0	0	0	0	0.1029	0.4265	0.4265	0.0294	0.0147
	7	0	0	0	0	0	0.103	0.3382	0.5147	0.0441
	8	0.0588	0	0	0	0.0147	0	0.0882	0.4118	0.4265
	9	0.5147	0	0	0	0	0	0	0.0441	0.4412

30 day		*Ending State*								
		1	2	3	4	5	6	7	8	9
Beginning State	1	0.8382	0.1324	0.0294	0	0	0	0	0	0
	2	0.0882	0.8088	0.103	0	0	0	0	0	0
	3	0	0.1029	0.6765	0.2206	0	0	0	0	0
	4	0	0	0.1618	0.5147	0.2647	0.0441	0.0147	0	0
	5	0	0	0.0147	0.1176	0.4853	0.3824	0	0	0
	6	0.0147	0	0	0	0.1324	0.3235	0.4853	0.0441	0
	7	0.0147	0	0	0	0	0.2059	0.2059	0.5441	0.0294
	8	0.087	0	0	0	0.0145	0.029	0.145	0.2899	0.4346
	9	0.6029	0	0	0	0	0.0147	0	0.103	0.2794

In each of these transition matrices, it appears that the two greatest probabilities of an increasing transition are 9 to 1 or 7 to 8. Between the 15 day and the 30 day time period, the 9 to 1 transitional probability grows whereas the 7 to 8 transition probability decreases. Also the chance of the a leading digit 7 regressing back into an leading digit 6 is significantly greater than a leading digit 9 regressing back into a leading digit 8 at each time period. These Markov chains suggest that if an investor is interested in a high probability of the leading digit (and also the price) of a security increasing and a low probability of the leading digit decreasing, he would focus on the transition from a leading digit 9 to a leading digit 1.

One interesting application of Markov chains is that a Markov chain based on a time period *t* raised to a power *x* will display transitional probabilities for a period of time *x*t.* So a Markov chain taken over a time period of 5 days raised to the third power would give the transitional probabilities of going from state *i* to state *j* within a 15 day period. If a Markov chain is raised to an infinite (or extremely high) power, then the transitional probabilities contained within that matrix are the probability of a unit beginning in state *i* and transitioning to state *j* given an infinite period of time. The transitional probabilities of a Markov chain converge to a limiting probability which is the amount of time each security spends in a given state regardless of the state that it began in. In other words the transitional probabilities of a Markov chain will converge to a stationary probability distribution. This is a way in which the transitional probabilities displayed in each Markov chain can be shown to match the distribution from which the original data was taken. If the limiting probabilities of the Markov chain look similar to the probability distribution of the original data; this is evidence that the transitional probabilities contained within each Markov Chain better reflex the expected transitional probabilities for the original distribution.

			Limiting Probabilities		
	Benford's	*Sept. 4 Sample*	*5 Day*	*15 Day*	*30 Day*
1	0.301	0.309	0.124	0.234	0.270
2	0.176	0.219	0.284	0.367	0.271
3	0.125	0.131	0.339	0.144	0.156
4	0.097	0.082	0.054	0.050	0.085
5	0.079	0.076	0.104	0.076	0.059
6	0.067	0.062	0.033	0.039	0.054
7	0.058	0.048	0.020	0.029	0.041
8	0.051	0.038	0.019	0.029	0.039
9	0.046	0.035	0.024	0.025	0.025

As can be seen in the table above, the limiting probability of the 30 day Markov chain most closely matches the original distribution of the sample taken on September 4th. This gives an indication that the most accurate Markov chain to examine and rely upon when making investing decisions would be the 30 day Markov chain. The limiting probabilities of the 30 day Markov chain being similar to the distribution of the original data does not prove that the transitional probabilities given in the 30 day chain are the expected transitional probabilities of the Russell 2000 but it does show that the 30 day Markov chain is a more accurate representation of the true transitional probabilities than the 5 day and 15 day chains. Since it appears that the 30 day Markov chain displays more accurate transitional probabilities than the other Markov chains, it was used as a basis for a hypothetical trading done with the 68 leading digit 9 securities found in the September 4th sample.

Thirty Day Trading

Since the 30 day Markov chain was found to contain the most reliable transitional probabilities, a thirty day trading scenario was conducted using the 68 original leading 9 securities from the September 4th sample.

The 68 securities were hypothetically purchased at the closing prices on September 4th. They were then sold two different ways. First, they were sold at the closing prices of October 4th and the return generated from the sale was calculated. Second, the stocks that closed with a leading digit 1 were sold immediately when they transitioned and the stocks that never changed to a leading digit 1 were sold on October 4th. The second trading scheme was to simulate an investor who put a limit order that would sell the security as soon as it transitioned into a leading digit one. Putting limit order on securities limits the risk and also the return since an investor is guaranteed to keep some of the profit but forgoes any additional profit above the initial leading digit 1 price.

These returns were compared to the expected return based on the 30 day Markov chain and the return of the overall Russell 2000 index. The expected return based on the 30 Day Markov Chain was calculated as follows:

$$Expected\ Return = [P(increase) * average\ percentage\ increase] - [P(decrease) * average\ \ perecent\ decrease]$$

$$Expeceted\ Return = \left[.6029 * \left(\frac{(10.5-9.5)}{9.5}\right)\right] - [\left\{.103 * \left(\frac{(9.5-8.5)}{9.5}\right)\right\} + \left\{.0147 * \left(\frac{(9.5-6.5)}{9.5}\right)\right\}]$$

$$Expeceted\ Return = .0635 - \ .0155 = .048 = 4.8\%$$

These are the returns that were generated from the various schemes.

	Percentage Returns
Expected Return	4.80%
Russell 2000 Index Return	5.18%
Sell in 30 days	4.28%
Sell after transition	3.43%

Holding the securities and then selling them all after 30 days yielded the higher return of the two trading schemes but it did not yield as great a return as purchasing the Russell 2000 index fund which is equally weighted between all the securities in the Russell 2000. One reason for this could be that by committing to a trading scheme based solely on the leading digit of a security, an investor is limiting his return by benefiting only from single leading digit transitions and forgoing any return from multiple leading digit transitions. Replications of this observational study could give some idea as to the risk or variance of a return generated by a leading digit trading scheme. It could be that this type of trading scheme has less variability of returns (less risk) than simply buying and selling an index fund of the Russell 2000 every 30 days but that knowledge can only come through replicating this study.

Conclusions and Further Work to Be Done

The leading digits of the securities that comprise the Russell 2000 index do appear to follow a Benford-like distribution. There is some evidence to suggest that the leading digit that is most likely to increase and has a low probability to decrease would be leading digit 9. Of the three Markov chains that were created, the limiting probabilities of the 30 day Markov chain most closely matched the distribution of the sample taken on September 4. For this reason, the trading scheme based on leading digits was set up on a thirty day time period. While both trading schemes generated a positive return, the Russell 2000 index generated a greater return over the same time period than either of them. This could be due to the fact that this type of leading digit trading scheme may limit the potential return of the investor since generally each security will only move one leading digit in either direction in the given time period.

This observational study should be seen as one replication of a much bigger study that still needs to be done. If the techniques used in this paper were replicated; much clearer findings would emerge. More dependable distributions and Markov chains could be created by taking an average of the replications. The risk or profit variability of the 30 day

trading schemes could be calculated and would give some indication of whether the return generate would be worth the risk.

The reason this observational study was not replicated more than once was due to time and technology constraints. Each price of each security had to be looked up one at a time for every period of time in question. This is another problem that would complicate a trading scheme based only on leading digits. In order to get an adequate sample of leading digit 9 securities; roughly ten to twelve hours of researching stock prices would be necessary. If some sort of computer program could be used to search for securities based only on their leading digit then a leading digit trading scheme could be worthwhile.

Another consideration for a trading scheme that involves buying and selling over a short period of time would be transaction costs. Every time a security is bought and sold a transaction cost is incurred. These transaction costs could potentially eliminate the profit generated of any thirty day trading scheme and should be taken into consideration. A large investment firm that had minimal transaction costs might be able to better use this type of short term trading scheme.

References

Current Composition of the Russell 2000. 28 June 2013. Raw data. Https://www.russell.com/indexes/documents/Membership/Russell2000_Membership_list.pdf, Seattle, Washington.

Ley, Eduardo. *On the Peculiar Distribution of the U.S. Stock Indeces' Digits*. Tech. Resources for the Future, Washington DC, 29 Nov. 1994. Web. 10 Oct. 2013. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.202.655&rep=rep1&type=pdf>.

Pinsky, Mark A., and Samuel Karlin. An Introduction to Stochastic Modeling. 4th ed. Amsterdam: Academic, 2011. Print.

Žgela, Mario. *Application of Benford's Law in Analysis of DAX Percentage Changes*. Tech. BULGARIAN ACADEMY OF SCIENCES, 01 Jan. 2011. Web. 10 Oct. 2013. <http://www.cit.iit.bas.bg/cit_2011/v11-4/zgela-53-70.pdf>.