Utilizing Benford's Law as a Predictor for Crashes in the Stock Market

Name: Prabhneet Sidhu
Coordinator: Mr. Webster
Email: prabhneet777@gmail.com
Additional Contacts: 587-914-3346

Benford's law describes a phenomenon in which numbers like 1 and 2 appear more frequently as leading digits rather than numbers like 8 and 9 (Fewester, 2008, 1-2). 1 should occur approximately 30.1% of the time as the leading digit, 2 should occur 17.6% of the time, and 9 should occur 9.6% of the time (Fewester, 2008, 1). In simpler terms, smaller values should naturally occur more often than larger values in raw data sets (Berger et al., 2017, 1). The law has a multitude of ways it can be applied, from financial fraud detection and forensics to even computer science and elections. It is commonly employed by accountants and forensic scientists when investigating sets of data to determine whether they have been manipulated or not (Kessel, 2020). While it is not a completely reliable method, Benford's law still has the potential to be expanded further and utilized in other ways like the Stock market.

The Stock market is infamous for its unpredictability and dramatic crashes, but can Benford's law be used to predict crashes? By applying Benford's law to data derived from the Stock market's pre-crashes and post-crashes we can determine if Benford's law is a viable method to foresee crashes if there are deviations in the leading digit frequencies. In the case Benford's law does significantly deviate from pre-crashes, then investors and economists can potentially use it as a tool to guide and navigate their investments and financial decisions.

Method:

1. Background Research
    a. Form an extensive understanding of the theory and origins of Benford's law
    b. Become proficient in implementing Benford's law to a data set and determining whether it follows Benford's law or not
2. Compiling Market Data Sets
    a. Use Google Finance, Yahoo Finance, QuantQuote, TickData, and Kaggle to find data
    b. Select specific market crashes to investigate and periods of stability to compare to
    c. Divide the cases being investigated into timelines: pre-crash, during, and post-crash recovery and compare to periods of stability
    d. Accumulate a variety of pieces of data from specified market crashes: trading volumes, daily closing prices, and market indices
3. Applying Benford's Law
    a. Extract leading digits from stock price and trading volume data
    b. Analyze frequency of each leading digit (1-9)
    c. Graph and compare distribution derived from each time period's data and Benford's expected distribution
4. Determine Deviations
    a. Employ Chi-squared, Kolmogorov-Smirnov, and Mean Absolute Deviation test
    b. Compare results to hypothesis
5. Differentiate conclusions from each market phase
    a. Compare and contrast deviations and findings during each phase
        i. Come up with potential explanations
6. Construct Predictive Models
    a. Utilize machine learning to detect patterns between deviations and crashes
        i. Logistic regression and decision trees
    b. Train the model using previous crash data
    c. Test out on current stock data

Background Research:

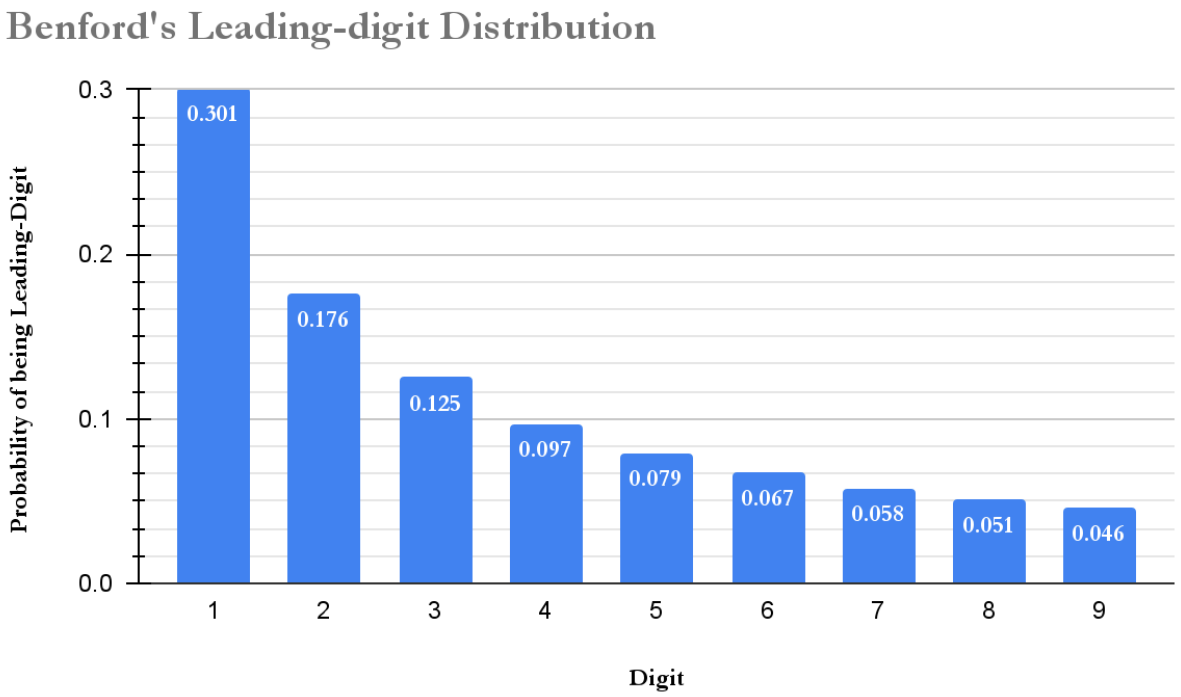
Initially discovered by Simon Newcomb (Canadian and American Astronomer / Mathematician) in 1881, Newcomb observed how dirty the early pages of the logarithm books were compared to their clean back pages (Berger et al., 2017, 1; Fewester, 2008, 1). From this observation Newcomb began investigating the frequency of leading-digits and mathematically determined a law that he expected these leading digits to follow (Fewester, 2008, 1).

$P(\text{leading digit} = d) = \log_{10}(1 + 1/d), d = 1, 2, \cdots, 8, 9.$

It wasn't until 1938 that Frank Benford truly tested Newcomb's law (Fewester, 2008, 1). Benford gathered 20 000 numbers from sources, such as *Reader's Digest* Articles, population sizes, river drainage rates, atomic weights, and numerous others. Benford's analysis and data revealed Newcomb's logarithmic rule to be true. While Frank Benford did coin a name for the law, he did not provide a persuasive explanation. Though he did state that the law only applied to anomalous numbers, and that orderly data like atomic weights were an exception to the law (Fewester, 2008, 1).

Intuitive reasoning and explanations for the law's effectiveness have been the scale-invariance and base-invariance arguments (Fewester, 2008, 2). Scale-invariance argues that if a potential universal law of nature is controlling the distribution, then the units should have no effect on the law's validity. This would mean that Benford's law is a universal law. The Base-invariance argument states that if Benford's law applies in any base number system whether its binary, octal, etc it would be a universal law. Though, neither of these reasons explain why it would even be a universal law of nature to begin with (Fewester, 2008, 2).

Benford's law is based on the mathematical theory that the probability of the leading-digit is $\log_{10}$ of 1 plus 1 over the leading digit (Fewester, 2008, 3). When 1 is the leading-digit, X (a positive number drawn from a probability distribution) is exactly $\log_{10}(X)$ between *n* and *n* + 0.301 (n is the sample size and integer) (Fewester, 2008, 3).



Definitions:

Leading-digit: the first number in a number.
Example: in 3678, 3 is the leading-digit. In 19045, 1 is the leading-digit.
Scale-invariance: The behaviour/structure is consistent no matter the scale.
Base-invariance: A mathematical law staying consistent no matter the base/coordinate system.
Market Index: a tool measuring the value of a specific stock/portfolio.

To find high quality, accurate stock data I will be using sources such as Yahoo Finance, Google Finance, Kaggle, and Quandl as advised by the *Caltech Quantitative Finance Group* and *Investopedia*. Through sources such as *Yahoo Finance* and *Investopedia* I have narrowed my choice of market indexes to investigate to S&P 500, Dow Jones Industrial, and Nasdaq Composite as they are some of the highest valued and popular indexes.

Data Collection:

Format for organizing data

Time Frames

Crash Name:	**Year:**	**Pre-Crash Period:**	**Crash Period:**	**Post-Crash Recovery:**
Great Depression	1929	1928-1929	October 28, 1929	1930-1932
Black Monday	1987	1985-1986	October 19, 1987	1988-1990
Dot-Com Bubble	2000	1998-1999	March 10, 2000	2001-2002
2008 Financial Crisis	2008	2006-2007	September 29, 2008	2009-2010
Covid-19 Crash	2020	2018-2020	February 20, 2020	2021-2022

First-Digits Extracted

Leading-digit	Expected Distribution (%)	Observed Distribution Pre-Crash (%)
1	30.1	-
2	17.6	-
3	12.5	-
4	9.7	-
5	7.9	-
6	6.7	-
7	5.8	-
8	5.1	-
9	4.6	-

Limitations:

Studying smaller scale crashes and a larger sample size to see if the result still holds.