Testing of Compliance of Benford's Law with Disaster Death Tolls

Abdullah Al Mahmud
Lecturer, Department of Statistics, Pabna Cadet College

Abstract

In 1938, the physicist Frank Benford published a counter-intuitive pattern among many naturally occurring collections of numbers, a pattern which later came to be known as Benford's law, also dubbed Newcomb-Benford's law, law of anomalous numbers, and first-digit law. The law simply states that the leading significant digit is expected to be small. For instance, in a set where the law fits, most numbers begin with digit 1, accounting for over 30% of total numbers, and the frequency of numbers starting with later digits falls almost exponentially (Benford, 1938). The law has been shown to apply to multitude of natural data sets, including stock and house prices, population numbers, death rates, length of large objects, and mathematical constants (Kvam, 2007), among others. In this paper, the law (for the first digit) has been tested for the number of deaths due to four different disasters: earthquakes, wars, accidents, and floods. A recapitulation on population data has also been accomplished, with a focus of time series data of population of countries. The law has been shown to apply to the instances of death tolls due to disaster and population data, while not being in conformity with time series data. The implications with the associated probability distributions of the data were also discussed. The findings were tested with Pearson's Chi-squared goodness of fit test.

Key words: *First digit law, Benford, Newcomb, disaster death toll, naturally occurring number.*

1. Introduction

The first digit of a randomly chosen number could be one of nine, out of one through nine, resulting to seem that the probability that any of the nine numbers will be on the first position of a number should be $\frac{1}{9}$. However, many real-world data sets differ from this orthodox estimation.

The law, though popularized by Frank Benford, actually traces its back to an 1881 work by the mathematician and astronomer Simon Newcomb, who showed that ten digits in the logarithmic table do not occur with equal frequency (Newcomb, 1881). He concluded that numbers in logarithmic tables obey the pattern, while anti-logarithms do not. He went on to propose a law that the probability of a single number N being the first digit of a number would be equal to $\log(N+1) - \log(N)$.

The law has numerously been studied in plenty of research papers, dealing with matters including theoretical developments, verifying for numbers of bases other than decimal, explanation of the pattern, testing compliance with real-world data, scale invariance (Pinkham, 1961), applications to various phenomena, including for legal and accounting purposes, price digit analysis, and detection of fraudulent scientific data, and developing of statistical tests to verify agreement with the law.

Benford (1937) showed several data sets to agree with the law, including surface areas of 335 rivers, 3259 US populations, 104 physical constants, 1389 specific heat values, 703 pressure data, 1165 Black Body data, and 418 death rates, among others. It has later been shown (Smith, 2012 & Fewster, 2009) that data sets containing uniform numbers of several orders of magnitude (e.g. populations) agree with the Benford's law, while data sets with numbers with mainly within only one order of magnitude (e.g. IQ scores) mostly, if not always, defy the law. Many well-known integer sequences have been shown to

comply with the law, among them being Fibonacci numbers (Washington, 1981), the powers of 2 (Raimi, 1976), and continuous growth processes, especially exponential growth or decay mechanism.
A number of testing procedures have been suggested to test for agreement with the Benford's law. Pearson's chi-squared test is the de facto test in this scenario, while Kolmogorovo-Smirnov and Kuiper tests are better for small sample sizes (Stephens, 1970).

2. Objectives of the Study

Objectives of the study are multifold: to show that the Benford's law fits well to the number of deaths due to disaster, to reexamine the compliance of the law with populations of countries (testing for 263 countries and territories, both with small and large populations), and to test for possible violation of Benford law by time series populations data.

3. Methods

A collection of numbers is said to obey the Benford's law if the first digit d (where d $\in$ {1,...,9} has the probability of occurring

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

According to the law, the distribution of the leading digits is given in Table 01.

Table 01: Expected Percentages of Occurrence of Digits

d	1	2	3	4	5	6	7	8	9
P (d)%	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

Where, P (d) is the percentage of numbers starting with the digit d.

The primary variable in the study is number of deaths from disasters, considering earthquakes, flood, accidents and wars. For all type of disasters, data were collected from the whole world, focusing primarily on the deadliest cases on record with respect to number of deaths. Data on earthquakes were collected from various relevant sources, including the Unites States Geological Survey (USGS), International Association of Engineering Geology, National Geophysical Data Center etc.; on accidents from various news sources including BBC (2006), the Age , and Brisbane Times etc.; on wars from *Darkest Hours* (Nash, 1976), *The Cambridge History of China: Alien Regimes And Border States* (Franke, 1994), and *World Christian Trends* (Joh, 2013), among others; and on flood from Global Active Archive of Large Flood Events (Brakenridge, 1985), and the Dawn etc. Whenever an interval of estimates were found instead of a single number, the lowest estimate was considered. Among the types of accidents considered were structural fires and collapses, road, aviation and maritime accidents, explosions, industrial disasters, and sporting events. Additionally, gathered were data of populations of 263 countries and dependencies since 1960 until 2016, totaling 57 years of data for 263 regions.

To test the acquired data, a function was written in R programming language, which produced a table providing observed frequencies and percentages of the digits one through nine, and a bar plot showing the relevant data, along with the p-value of the associated Pearson's chi-squared test.
Since the data size considered are not small, the chi-squared test of goodness of fit can be assumed to be perfect to test for the compliance with the law. The relevant test statistic is as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{1}^{n} \frac{\left(\frac{O_i}{N} - p_i\right)^2}{p_i}$$

Where, χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2distribution.
O_i = the number of observations of type i
E_i = the expected number of count of type i, as stated by the null hypothesis that the proportion of type i in the population is p_i.
n = number of cells in the table.

In simple term, the null hypothesis is that there is no difference between observed and expected frequencies of numbers. P-values less than 0.01 and 0.05 would imply that the observed frequencies are different from expected frequencies at 0.01 and 0.05 levels of significance, respectively i.e., if the test were conducted a hundred times, only once and five times, respectively, the observed frequencies would comply with expected frequencies. Thus, p-values, in this study, greater than 0.01 imply strong agreement of data with Benford's law.

4 Results and discussion

4.1 Testing of disaster data

A total of 193 severe floods were considered from all over the world spanning more than a millennium, while a total of 197 significant wars have been considered. Included are both World War I and II, Gallic wars, Crmbrian War, Crusades, Vietnam War, Napoleonic Wars, among others. The result is shown on table 3. As far as earthquakes are concerned, a total of 190 instances were taken into consideration. As stated earlier, accidents considered were of many different types, with a total of 1538 cases from all over the world.

The table below and the succeeding plot summarize the result.

Table 2: Summary of disaster death tolls with expected and observed frequencies of numbers starting with 1 through 9

Digit	Expected percentage (by Benford's Law, %)	Expected percentage (Floods, %)	Expected percentage (Earthquakes, %)	Expected percentage (Wars, %)	Expected percentage (Accidents, %)
1	30.1	26.42	26.32	35.53	35.7
2	17.61	22.8	20.53	16.75	17.88
3	12.49	15.03	14.21	9.64	10.4
4	9.69	10.36	8.95	9.64	6.89
5	7.92	6.74	8.42	7.11	7.09
6	6.69	4.15	7.37	4.57	7.15
7	5.8	3.63	5.26	5.58	5.53
8	5.12	6.74	5.26	5.58	4.62
9	4.58	4.15	3.68	5.58	4.75

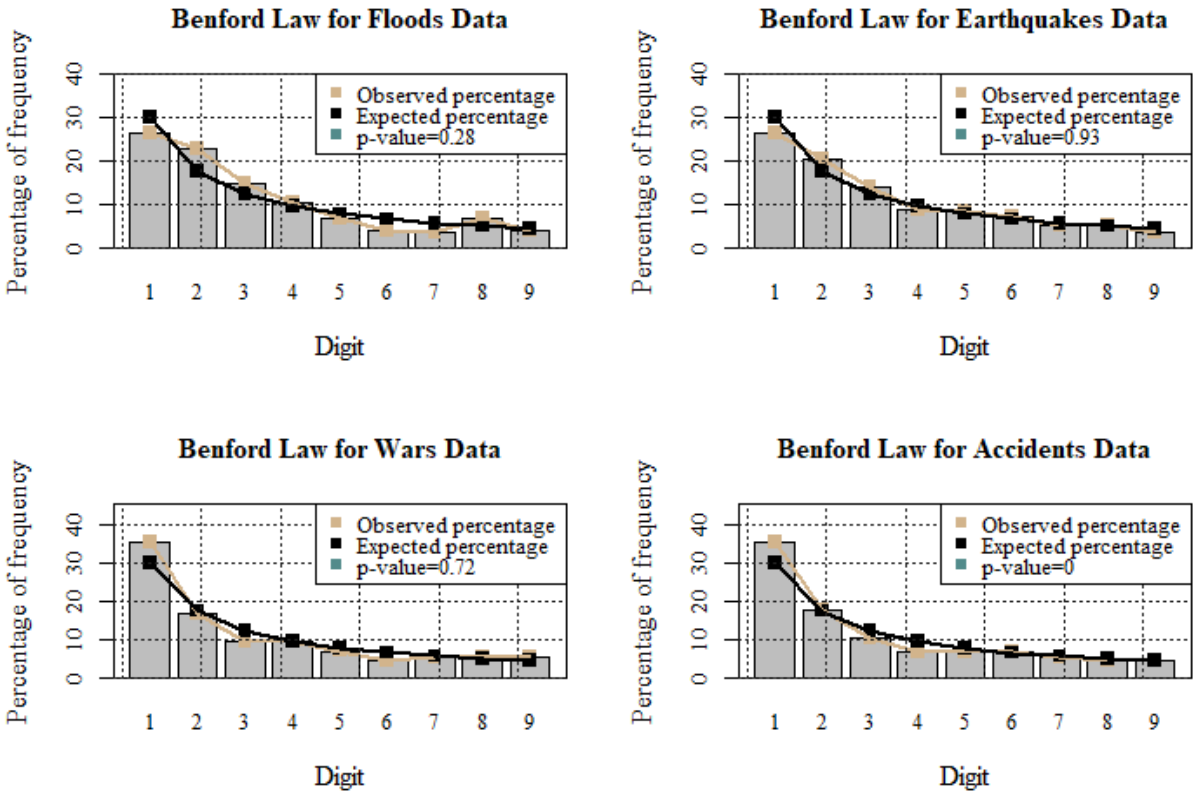

Figure 1. Expected and empirical percentages and frequencies of numbers starting with digits one through nine for disaster death estimates.

The bars represent the observed percentages of numbers with each digits, while x axis represent the digits and y axis refers to the percentages of frequencies of numbers. The tan-colored and black-colored lines represent the expected and observed percentages of frequencies of numbers, respectively. The associated Pearson's chi-squared test p-value has also been provided on the plot for each data set.

A careful looking on the plot suggests a strong agreement of the four studied data sets with Benford's law. While the percentages of the digit 1 for floods and earthquakes are 26.42 and 26.32, both being less than the corresponding expected frequencies, the percentages of wars and accidents data sets for the same digit (1) exceed what is stated by Benford's law, being 35.53 and 35.7, respectively i.e. the first digit percentages exceed expected percentage by over 5%. All the data sets, however, seem to, from the illustrations, satisfy the law. This is confirmed by the p-values for all data sets except for accidents, with the p-value almost 0. For this data set, Pearson's chi-squarest test and the graphical connotation differ. This contradiction might have resulted from the fact that although there have been covered many different types of accidents, many other types remain to be explored. Still, the pattern of distribution of number to different digits conforms to Benford's distribution. All other p-values are well above 0.01, being no less than 0.28 and as high as 0.93, indicating clear conformity with Benford's law.

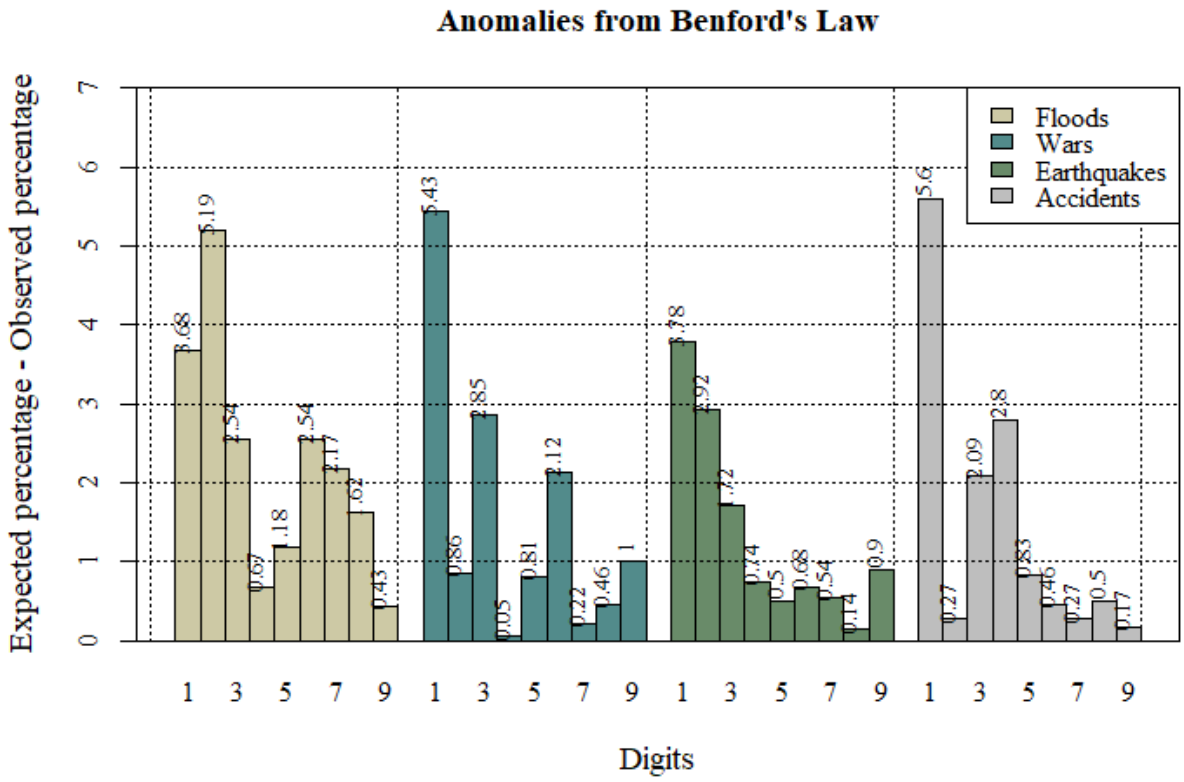

Figure 2. Anomalies of empirical percentages from expected percentages of numbers starting with each of the nine digits for all data sets

Clearly, the maximum anomalies are seen for digit 1, exceeding 5 for all data sets except for floods earthquakes. This implies the strong dominance of digit 1, taking away more numbers than what is expected by Benford's law. Digit 2 has maximum anomaly for floods data. No other digits significantly differ from the expected pattern. For earthquakes and accidents data, the anomalies decreased most consistently for larges digits.

4.2 Testing for populations data

In this instance, world populations comprising of 263 countries and territories were considered. Earlier in 2002, Sandron showed that populations of 198 countries or geopolitical entities as in the year 1997 followed Benford's law (Sandron, 2002).

In this analysis, we have taken into account world populations from 1960 to 2016, each year dealt separately, applying the function and carrying out Pearson's chi-squared test, and recording the associated p-values. The resulting p-values are shown in figure 3.

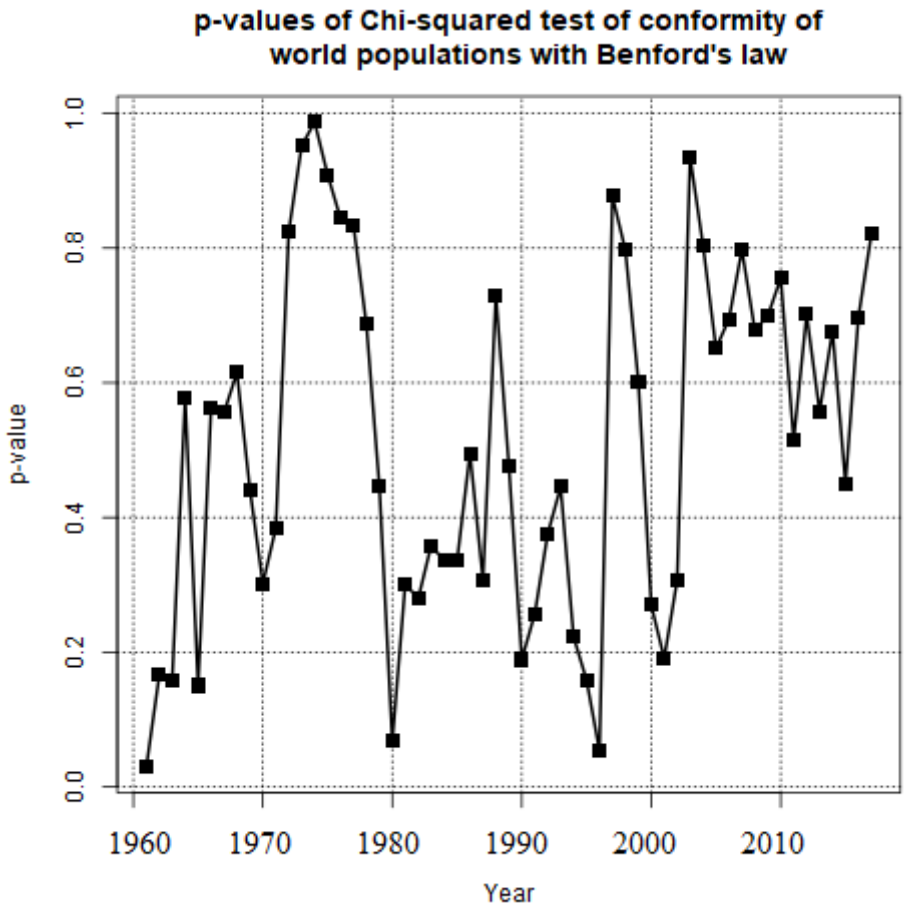

Figure 3. p-values of chi chi-squared tests of goodness-of-fit with regard to testing whether yearly world populations of 263 countries and territories conform to Benford's law, with each year dealt separately. A total of 57 years were analyzed.

A total of 57 years were used for this computation, none of which happened to be below 0.01, which implies that world population data in any year conform to Benford's law, at 1% level of statistical significance. The population in 1960 is the only instance which shows incongruity with the law, at 5% level of significance.

4.3 Testing of Benford's Law for Time Series Data

Benford's law for time series data were also tested, considering 57 years of yearly population data for each of the available 263 countries and territories. Testing the compliance with our function and carrying out the relevant chi-squared test of goodness of fit test, we have obtained 263 p-values, one for each territory.

The result is worth noting, as is evident in the following bar diagram:

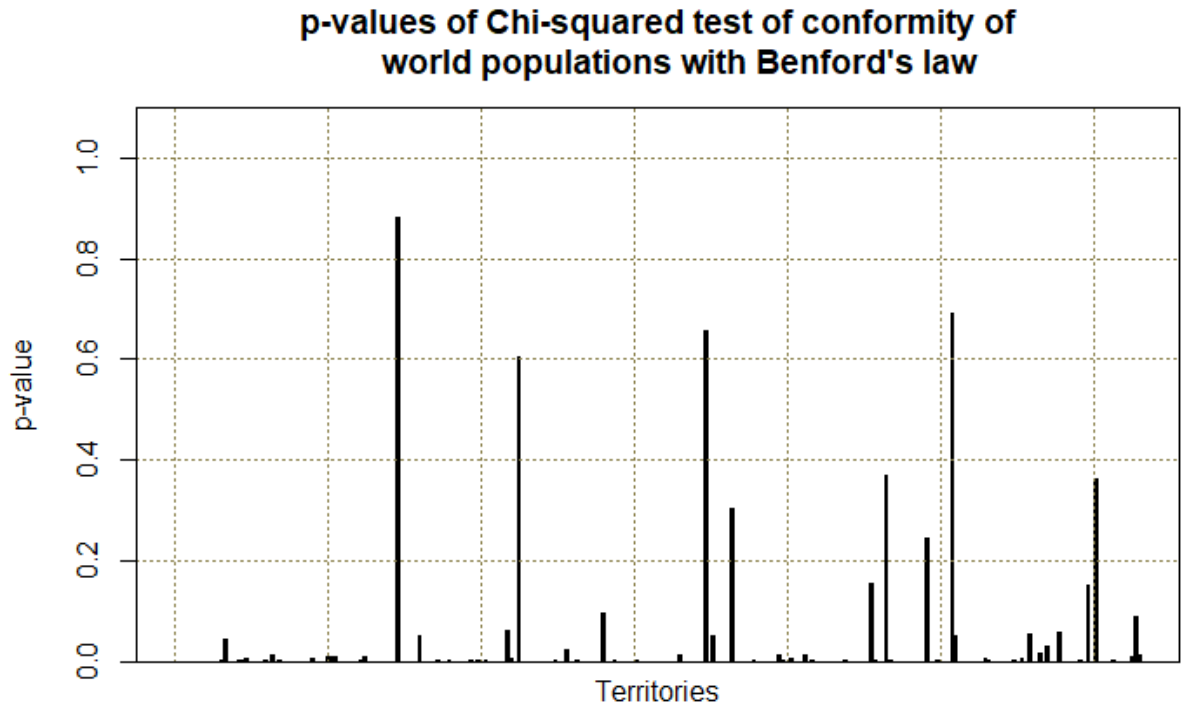

Figure 4. p-values of chi chi-squared tests of goodness-of-fit with regard to testing whether yearly time series data world populations of 263 countries and territories conform to Benford's law.

For the time series data, most p-values are indicative of insignificant compliance with the Benford's law. Out of a total of 263, p-values are less than 0.05 for 245 regions, while it is below 0.10 for 253 regions, which implies time series data of only a handful of regions (10, at 0.10 level of significance) conform to the Benford's law.

Conclusion

We showed that the number of people dying in disasters around the world conforms well to the Benford's law. The fact has been clearly illustrated by the plots as well as by Pearson's chi-squared goodness-of-fit tests. The tests have been performed for a total four types of disasters: floods, wars, earthquakes, and accidents. As far as the accidents are concerned, many different types were considered, including structural fires and collapses; road, aviation and maritime accidents. Yet, although the graph indicated conformity with the law, the p-values indicated noncompliance, a fact possibly resulting from the inadequate types of accidents considered. In all other cases, p-values were in complete harmony with the graphs in establishing compliance of data with the law. Analysis of population data were also shown to be in conformity with the law. Another significant finding was that time series data do not comply with the law, a fact due possibly to the fact that time series data, since the data are considered from an arbitrary year, start with a certain digit and continue with the same or a few next digits for many years. For example, digit 4 preceded the population of Bangladesh in 1960, and since then never could the digits 2 or 3 precede the population value, resulting in the deviation from the Benford's law. The findings show that the Benford's Law might be useful in real world scenario for validating data and to use them for model building to make predictions for data.

References

Age. (1979, June 11). Seven die at fun fair: Suddenly fire and smoke make the mock horror real. The Age.

BBC. (2006, September 13). Hope fades for water park victims. BBC News. http://news.bbc.co.uk/2/hi/europe/3489915.stm

Benford, F. (1938). The Law of Anomalous Numbers. Proceedings of the American Philosophical Society, 78(4), 551-572. Retrieved November 2, 2020, from http://www.jstor.org/stable/984802

Brakenridge, G. R. (1985). Global Active Archive of Large Flood Events. Dartmouth Flood Observatory, University of Colorado.

Fewster, R. M. (2009). A simple explanation of Benford's Law. The American Statistician, 63 (1): 26–32. doi:10.1198/tast.2009.0005.

Franke, H. & Twitchett, D. C. (Eds.). (1994). The Cambridge History of China: Alien Regimes And Border States. 907–1368, 1994, p.622

Joh, B. (2013). World Christian Trends. William Carey Library.

Kvam, P. H. & Vidakovic, Brani. (2007). Nonparametric Statistics with Applications to Science and Engineering. Wiley. p. 158

Nash, J. R. (1976). Darkest Hours. Rowman & Littlefield, p. 775.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics. American Journal of Mathematics, Vol. 4, No. 1. 4 (1/4): 39–40. doi:10.2307/2369148. JSTOR 2369148.

Pinkham, R. S. (1961). On the Distribution of First Significant Digits. Ann. Math. Statist. Volume 32, Number 4 (1961), 1223-1230.

Raimi, R. A. (1976). The First Digit Problem. American Mathematical Monthly, 83 (7): 521–538. doi:10.2307/2319349.

Sandron, F. (2002). Les populations suivent-elles la loi des nombres anomaux?. Population, vol. 57,(4), 753-761. doi:10.3917/popu.204.0761.

Smith, W, S. (2012) The Scientist and Engineer's Guide to Digital Signal Processing, chapter 34.

Stephens, M. A. (1970). Use of the Kolmogorov–Smirnov, Cramér–Von Mises and Related Statistics without Extensive Tables. Journal of the Royal Statistical Society, Series B. 32 (1): 115–122.

Washington, L. C. (1981). Benford's Law for Fibonacci and Lucas Numbers. The Fibonacci Quarterly, 19 (2): 175–177.

Appendix

Table 3: Observed Frequency of Leading Digits for Different Accidents

Digit	Floods	Wars	Earthquakes	Accidents
1	51	70	50	549
2	44	33	39	275
3	29	19	27	160
4	20	19	17	106
5	13	14	16	109
6	8	9	14	110
7	7	11	10	85
8	13	11	10	71
9	8	11	7	73
Total	193	197	190	1538