

Testing for the Benford Property

Daniel P. Pike

School of Mathematical Sciences, Rochester Institute of Technology, 85 Lomb Memorial Drive,
Rochester, New York 14623–6627 USA. E-mail: dpp1382@rit.edu

Sponsor

David L. Farnsworth, School of Mathematical Sciences, Rochester Institute of Technology, 85
Lomb Memorial Drive, Rochester, New York 14623–6627 USA. E-mail: dlfsma@rit.edu

Abstract

Benford's Law says that many naturally occurring sets of observations follow a certain logarithmic law. Relative frequencies of the first significant digits k are $\log(1 + 1/k)$ for $k = 1, 2, \dots, 9$, where the base of the logarithm is ten. Financial and other auditors routinely check data sets against this law in order to investigate for fraud. We present the principal underlying mechanism that produces sets of numbers with the Benford property. Examples in which each observation consists of a product of variables are given. Two standard statistical tests that are useful for testing compliance with Benford's Law are outlined. A new Minitab macro, which implements both statistical tests and produces a graphical output, is presented.

1. Introduction

Benford's Law is named for Frank Benford, a General Electric scientist whose paper (Benford, 1938) was the first modern article about it. He noticed that many data sets have a very unintuitive property. Taking the set consisting of the first significant digit of each value in the observations and counting the numbers of 1s, 2s, and so forth to 9s, there was about 30% 1s, 18% 2s, and in a descending fashion to 5% 9s. Without knowing about Benford's Law, we might think that there should be about the same number of 1s, 2s, and so forth. Specifically, the law says that the proportions of first significant digits k for $k = 1, 2, \dots, 9$ are $b(k) = \log_{10}(1 + 1/k)$. These proportions appear in Table 1.

<u>k</u>	<u>b(k)</u>
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	<u>0.0458</u>
Sum	1.0000

Table 1: First digits, k , and Benford's Law proportions, $b(k) = \log_{10}(1 + 1/k)$.

Since data sets that have been tampered with usually do not follow this law, Benford's Law is relatively well-known among auditors who examine quantitative observations. Data fakers might not think of this law. In practice it is used as a filter to flag data sets for financial fraud (Nigrini, 2000).

In Section 2, a way to see why the law has this logarithmic form is presented. This leads to a mechanism to simulate artificial data sets that follow the law. In Section 3, examples based on multiplication tables are presented. Section 4 contains brief discussions of the chi-square goodness-of-fit test and the test for proportions, which can be used to check observations for Benford's Law. Section 5 discusses a Minitab macro for implementing the tests and contains a link to the macro. Concluding comments are in Section 6.

2. A Mechanism for Benford's Law

Sets composed of numbers that are themselves products often follow Benford's Law. Exponential growth is an example of this, in which the amount present is the product of the amount just before times 1 plus a rate of increase. A discrete example is interest posted daily on a bank savings account.

Consider the random variable $Y = 10^X$ where X has the uniform distribution on the interval $(0, 1)$. If a value for Y has first significant digit $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, then Y must be in the interval $10^n k \leq Y < 10^n(k + 1)$ for some integer n . Taking the logarithm, obtain $n + \log_{10}k \leq X < n + \log_{10}(k + 1)$. The length of this interval is $\log_{10}(k + 1) - \log_{10}k = \log_{10}(1 + 1/k)$, which is the probability that Y has first digit k (Feller, 1971, p. 63).

This argument shows that if X is uniformly distributed, then the variable $Y = 10^X$ has the Benford property. One way to produce numbers that will closely follow Benford's Law is to take numbers that are uniformly distributed over the interval $(0, 1)$ from a random number generator and exponentiate them with 10^x . Such an artificial set of observations can be used to tryout the Minitab macro in Section 5.

Generally, sets of products have the Benford property. An example of a special case, which would not follow the law, is the set where all multiplicands are powers of 10, so that all of the first digits are 1.

3. Examples – Multiplicands of the First Nine Positive Integers

For illustrative purposes, consider all products of two and of three numbers where every possible combination of values for each multiplicand is selected from the set {1, 2, 3, 4, 5, 6, 7, 8, 9}. Table 2 contains all the products of two numbers from the set, and Table 3 illustrates the closeness to the Benford's Law proportions. The set of all products of three integers appears to be closer, as displayed in Table 4.

9	9	18	27	36	45	54	63	72	81
8	8	16	24	32	40	48	56	64	72
7	7	14	21	28	35	42	49	56	63
6	6	12	18	24	30	36	42	48	54
5	5	10	15	20	25	30	35	40	45
4	4	8	12	16	20	24	28	32	36
3	3	6	9	12	15	18	21	24	27
2	2	4	6	8	10	12	14	16	18
1	1	2	3	4	5	6	7	8	9
	1	2	3	4	5	6	7	8	9

Table 2: Multiplication table for the first nine positive integers.

First Digit k	Frequency f	Proportions $p(k) = f/81$	Benford's Law Proportions b(k)
1	18	0.2222	0.3010
2	15	0.1852	0.1761
3	11	0.1358	0.1249
4	12	0.1481	0.0969
5	6	0.0741	0.0792
6	7	0.0864	0.0669
7	4	0.0494	0.0580
8	5	0.0617	0.0512
9	<u>3</u>	<u>0.0371</u>	<u>0.0458</u>
Sums	81	1.0000	1.0000

Table 3: Proportions $p(k)$ for Table 2 of all products of two numbers from the set $\{1, 2, \dots, 9\}$ and $b(k)$ for Benford's Law.

First Digit k	Frequency f(k)	Proportions $p(k) = f(k)/729$	Benford's Law Proportions b(k)
1	218	0.2990	0.3010
2	137	0.1879	0.1761
3	94	0.1289	0.1249
4	81	0.1111	0.0969
5	46	0.0631	0.0792
6	43	0.0590	0.0669
7	37	0.0508	0.0580
8	37	0.0508	0.0512
9	<u>36</u>	<u>0.0494</u>	<u>0.0458</u>
Sums	729	1.0000	1.0000

Table 4: Proportions $p(k)$ for all products of three numbers from the set $\{1, 2, \dots, 9\}$ and $b(k)$ for Benford's Law.

In the next section the proportions from multiplications, appearing in Tables 3 and 4, are used to demonstrate two tests for compliance with Benford's Law.

4. Testing for Compliance with Benford's Law

Since these are categorical data being tested for goodness of fit, the Pearsonian chi-square test can be used. The chi-square goodness-of-fit statistic is calculated using

$$\chi_{\text{calc}}^2 = N \sum_{k=1}^9 \frac{(p(k) - b(k))^2}{b(k)}$$

where N is the sum of the frequencies, and $p(k)$ and $b(k)$ are the proportions from the data and the Benford's Law proportions, respectively. This is an intrinsically one-sided test. Small values of χ^2_{calc} , close to zero, reinforce that the data's proportions are close to Benford's Law, which is the null hypothesis. Large values would tell us that the data do not come from a process that produces Benford's Law, which is the alternative hypothesis.

The P-value is the probability $P(\chi > \chi^2_{\text{calc}})$, where χ is the chi-square random variable with degrees of freedom $9 - 1 = 8$. If the P-value is less than the level of significance, α , then the null hypothesis is rejected, since the observed χ^2_{calc} would be deemed to be large. Otherwise, we would say that there is not sufficient evidence to reject that the data arise from a process that produces first significant digits that follow Benford's Law. A common value for the level of significance is $\alpha = 0.05$, which we will use.

Table 5 contains the chi-square statistics and P-values for the proportions appearing in Tables 3 and 4. Both data sets pass the goodness-of-fit test, since the P-values are greater than the level of significance $\alpha = 0.05$; therefore, the decision is that both data sets come from mechanisms that produce Benford's Law proportions.

Number of Multiplicands	Chi-Square Statistic (χ^2_{calc})	P-Value for χ^2_{calc}
2	4.881	0.770
3	6.141	0.631

Table 5: Chi-square goodness-of-fit statistics χ^2_{calc} and their P-values for 2 and 3 multiplicands from the set $\{1, 2, \dots, 9\}$.

Another option is to test each of the nine proportions separately. The normal-distribution approximation to the binomial distribution is used. We find the Z-scores and their corresponding two-sided P-values for each observed proportion, against the Benford proportions, using

$$Z_k = \frac{p(k) - b(k)}{\sqrt{\frac{(b(k))(1 - b(k))}{N}}}$$

Table 6 gives the outcomes of these analyses for two and three multiplicands. The P-values should not be compared to the level of significance, $\alpha = 0.05$, since for each number of

multiplicands, there are nine comparisons. The process of reducing the level of significance used in each of the nine tests is based on Bonferroni's inequality (Hogg, McKean, and Craig, 2005, pp. 14, 481). Each P-value is compared to $\alpha/9 = 0.05/9 = 0.0056$, giving an approximate overall probability of rejection of 0.05. Since $P(|Z| > 2.77) = 0.0056$, any Z-score greater in absolute value than 2.77 implies rejection of the null hypothesis. Since all P-values in Table 6 are greater than 0.0056 (and all Z-scores are less than 2.77 in absolute value), we do not reject the null hypotheses that each proportion is the corresponding Benford proportion.

		Number of Multiplicands			
		2		3	
		Z-Score	P-value	Z-Score	P-value
First Significant Digit, k	1	-1.55	0.122	-0.12	0.907
	2	0.21	0.830	0.84	0.401
	3	0.30	0.767	0.33	0.744
	4	1.56	0.119	1.30	0.195
	5	-0.17	0.865	-1.61	0.108
	6	0.70	0.483	-0.86	0.390
	7	-0.33*	0.740	-0.84	0.403
	8	0.43*	0.666	-0.05	0.961
	9	-0.38*	0.707	0.47	0.640

Table 6: Z-scores and P-values for testing each significant digit against the corresponding Benford proportion as the null hypothesis. The observations' proportions are the proportions from the products and are taken from Tables 3 and 4. The criteria that the mean of the binomial distribution be at least 5, in order for the normal distribution approximation to be close, is not satisfied for the three z-scores marked "*".

5. The Minitab Macro

A computer program that determines whether a data set follows Benford's Law is particularly useful if the dataset is very large or the test must be done quickly. A Minitab macro could be a very good choice. Minitab is used in many statistics courses, so it is readily available, widely used, and familiar to many people. Although Microsoft Excel is even more readily available than Minitab, it has limitations. It can only hold up to 65,536 values. Minitab can hold over one million values in a single column. The Minitab macro Benford.txt performs the tests described in Section 4. The following documentation gives a step-by-step guide about how to implement the Minitab macro and what the output represents.

5.1 Instructions and Procedures

Before implementing this macro, it is important to make sure that no data value in your set is greater than 100 million or less than 0.000000001; otherwise, the first digit can't be obtained by the macro. These numbers refer to the absolute value of your data values. Negative numbers are allowed. If any data value is outside this range, divide any number greater than 100 million by 10 to some power in order to provide a number less than 100 million and multiply any number less than 0.000000001 by 10 to some power in order to provide a number greater 0.000000001. This macro must be run in Minitab Version 15 or later; any earlier version may not produce accurate results.

1. In order to run this macro, the text file containing the code must first be downloaded and saved into your Macros folder for Minitab. Click on [Benford.txt](#) and select "Save As..." under File at the top left part of your browser. Under "Save as type", make sure it's a text file (*.txt). Make sure the name of the document is "Benford" without the quotes and save the file in your Minitab folder under Macros by following the path: My Computer → C: Drive → Program Files → Minitab 15 → English → Macros. (This is the default passage to the Macros folder. In some cases, people may have installed Minitab in a different location on their computer. If Minitab was not installed on the default passage, then search for "Minitab 15" using the computer's search engine and proceed as shown.) Once the file is saved, the macro can be accessed through Minitab.
2. Open Minitab with a new worksheet. Enter your data set into C1. Make sure all data is in this column and only this column. The order does not matter. The data's format must be numeric. Any format other than numeric, such as date/time or text, will not work with the macro. Click on the Session Window, and then select "Editor" at the top of the screen. The option "Enable Commands" should be selectable. If it doesn't have a check mark next to it, click it. Otherwise, leave it alone.
3. In the session window, there will be "MTB >". Click to the right of it and type in "%Benford.txt" without the quotes and hit "Enter." The macro will run and take just a few seconds. For larger datasets, such as one million values, it could take up to fifteen seconds or more to complete.

5.2 Description of the Output

1. C2, "First Digit," is the first digit other than 0s of the data value in the corresponding cell of C1. For example, a data value equal to 0.00342 will result in a "First Digit" of 3.

2. C3 and C4, “Numbers” and “First Digit Counts,” represent the number of times that specific digit appears as the first digit in the data set. The last cell in the “First Digit Counts” column represents the sum of that column. It is equal to the total number, N, of data values in the data set.
3. C5, “First Digit Proportions,” is the proportions of the numbers that appear as the first digit. This is the observed proportions, $p(k)$, to be compared with the Benford proportions, $b(k)$, in C6.
4. C6, “Benfords Proportions,” contains the expected proportions, $b(k)$, from Benford’s Law.
5. C7, “Chi-Squared Test,” is the individual terms in the chi-square statistic for the observed proportions versus the Benford proportions of each corresponding first significant digit, that is, $N \frac{(p(k) - b(k))^2}{b(k)}$. The last cell in this column is the sum of all nine of these terms.

This is the chi-square test statistic for the goodness-of-fit test for Benford’s Law.
6. C8, “Chi-Square P-Value,” is the P-value associated with the chi-square test statistic. The *s in the cells above the P-value represent cell fillers, used to help format the output.
7. C9 and C10, “Individual Z-Scores” and “P-Values,” are the Z-scores for the first digit proportions, that is, the observed proportions, and their corresponding P-values. These P-values are used to determine if the observed proportions differ significantly from the Benford proportions for each first significant digit.
8. The graph is a visual representation of the first digit proportions versus the Benford proportions.

6. Concluding Comments

Testing for the Benford property when fraud is suspected is a very useful tool. If the first significant digit proportions come from a process that does not produce Benford’s proportions, then fraud may be present or at least we might want to investigate the data further. The Minitab macro is a quick way to check this using both the chi-square test and the test for proportions in one procedure, which also gives histograms for a visual aid.

References

F. BENFORD (1938), The law of anomalous numbers, Proc. Amer. Philos. Soc., 78, pp. 551-572.

W. FELLER (1971), An Introduction to Probability Theory and Its Applications, Volume II, Wiley, New York.

R. V. HOGG, J. W. McKEAN, and A. T. CRAIG (2005), Introduction to Mathematical Statistics, 6th ed., Pearson Prentice Hall, Upper Saddle River, New Jersey.

M. J. NIGRINI (2000), Digital Analysis Using Benford's Law: Tests & Statistics for Auditors, Global Audit Publications, Vancouver.