Formelsammlung Benford

V1.0, 26-Mar-2019, Herbert Feichtinger

Inhalt

Benford's Expected First Digit Frequencies

D_1 first digit
D_2 second digit
D_1D_2 first-two digits of a number
Prob Probability

The next stage of Benford's research was to derive the expected frequencies of the digits in lists of numbers. The formulas for the digit frequencies are shown next with D_1 representing the first digit, D_2 the second digit, and D_1D_2 the first-two digits of a number.

$$\text{Prob}(D_1 = d_1) = \log\left(1 + \frac{1}{d_1}\right); \qquad d_1 \in \{1, 2, \ldots, 9\} \tag{1.1}$$

$$\text{Prob}(D_2 = d_2) = \sum_{d_1=1}^{9} \log\left(1 + \frac{1}{d_1 d_2}\right); \qquad d_2 \in \{0, 1, \ldots, 9\} \tag{1.2}$$

$$\text{Prob}(D_1 D_2 = d_1 d_2) = \log\left(1 + \frac{1}{d_1 d_2}\right); \qquad d_1 d_2 \in \{10, 11, \ldots, 99\} \tag{1.3}$$

where Prob indicates the probability of observing the event in parentheses. The formula for the first digit proportions is shown in Equation 1.1. The formula for the second digit proportions is shown in Equation 1.2, and the formula for the first-two digit proportions is shown in Equation 1.3. For example, the probability of the first digit being equal to *1* is calculated as shown in Equation 1.4.

$$\text{Prob}(D_1 = 1) = \log\left(1 + \frac{1}{1}\right) = \log(2) = 0.30103 \tag{1.4}$$

The probability of the second digit being equal to *1* is calculated using Equation 1.2 and the steps in the calculation are shown in Equation 1.5.

$$\begin{aligned}\text{Prob}(D_2 = 1) &= \sum_{d_1=1}^{9} \log\left(1 + \frac{1}{d_1 d_2}\right) \\ &= \log\left(1 + \frac{1}{11}\right) + \log\left(1 + \frac{1}{21}\right) + \log\left(1 + \frac{1}{31}\right) \\ &\quad + \log\left(1 + \frac{1}{41}\right) + \log\left(1 + \frac{1}{51}\right) + \log\left(1 + \frac{1}{61}\right) \\ &\quad + \log\left(1 + \frac{1}{71}\right) + \log\left(1 + \frac{1}{81}\right) + \log\left(1 + \frac{1}{91}\right) \\ &= 0.11389\end{aligned} \tag{1.5}$$

The steps in Equation 1.5 are based on the fact that the second digit is equal to *1* if the first-two digits are either *11, 21, 31, 41, 51, 61, 71, 81*, or *91*. The probability of the second digit being *1* is the sum of the nine probabilities. The probability of the first-two digits being *11* is calculated as shown in Equation 1.6.

$$\text{Prob}(D_1D_2 = 11) = \log\left(1 + \frac{1}{11}\right) = \log\left(\frac{12}{11}\right) = 0.03779 \tag{1.6}$$

The Benford's Law proportions for the digits in the first, second, third, and fourth positions are shown in Table 1.2. The first digit proportions were calculated using Equation 1.1, and the second digit proportions were calculated using Equation 1.2. The third and fourth digit proportions were calculated using the logic in Equation 1.2. For example, a third digit *0* occurs in *100, 110, 120, 130, . . . , 990*. The third digit *0* probability is the sum of the *110, 120, 130, . . . , 990* probabilities. The table shows that as we move from left to right, the digits tend toward being evenly distributed. If we are dealing with numbers with three or more digits, for all practical purposes the ending digits (the rightmost ones) are expected to be evenly (uniformly) distributed.

The first few pages of this book were equation-free, but I'm afraid that we now need to do a little catching up in the equation department. In the next section we're going to develop a formal definition of what we mean by the first and second digits of a number and we're also going to show the general equation for calculating the expected proportion for any combination of digits.

Digit Proportions of Benford's Law

TABLE 1.2 First, Second, Third, and Fourth Digit Proportions of Benford's Law

	Position in Number			
Digit	**1st**	**2nd**	**3rd**	**4th**
0		.11968	.10178	.10018
1	.30103	.11389	.10138	.10014
2	.17609	.10882	.10097	.10010
3	.12494	.10433	.10057	.10006
4	.09691	.10031	.10018	.10002
5	.07918	.09668	.09979	.09998
6	.06695	.09337	.09940	.09994
7	.05799	.09035	.09902	.09990
8	.05115	.08757	.09864	.09986
9	.04576	.08500	.09827	.09982

Source: "A Taxpayer Compliance Application of Benford's Law," by M. Nigrini, 1996, *Journal of the American Taxation Association, 18*(1), page 74.

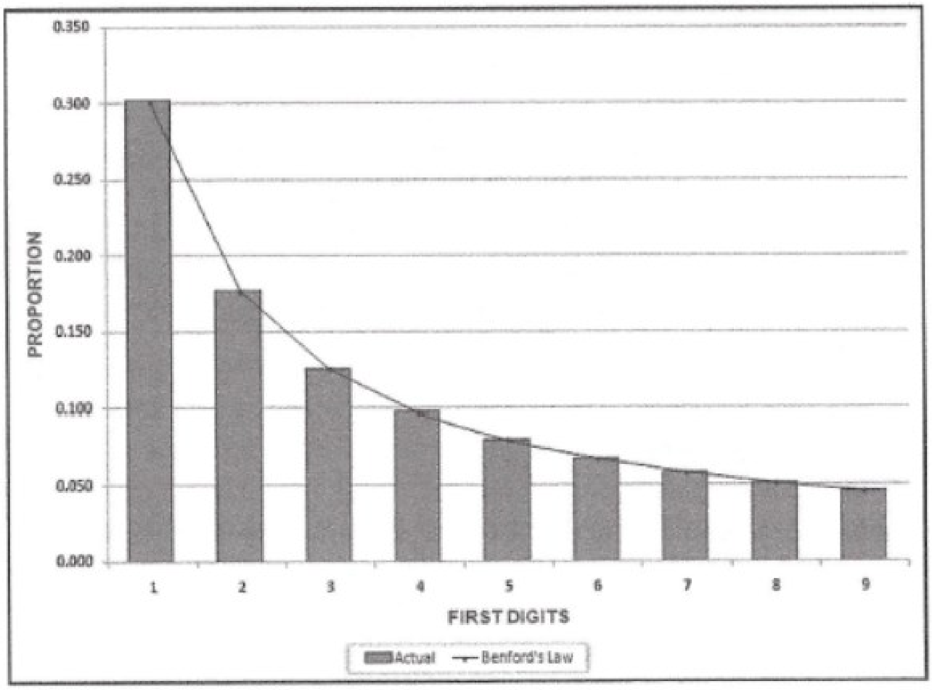

FIGURE 1.7 First Digit Graph of the Data in Table 1.3

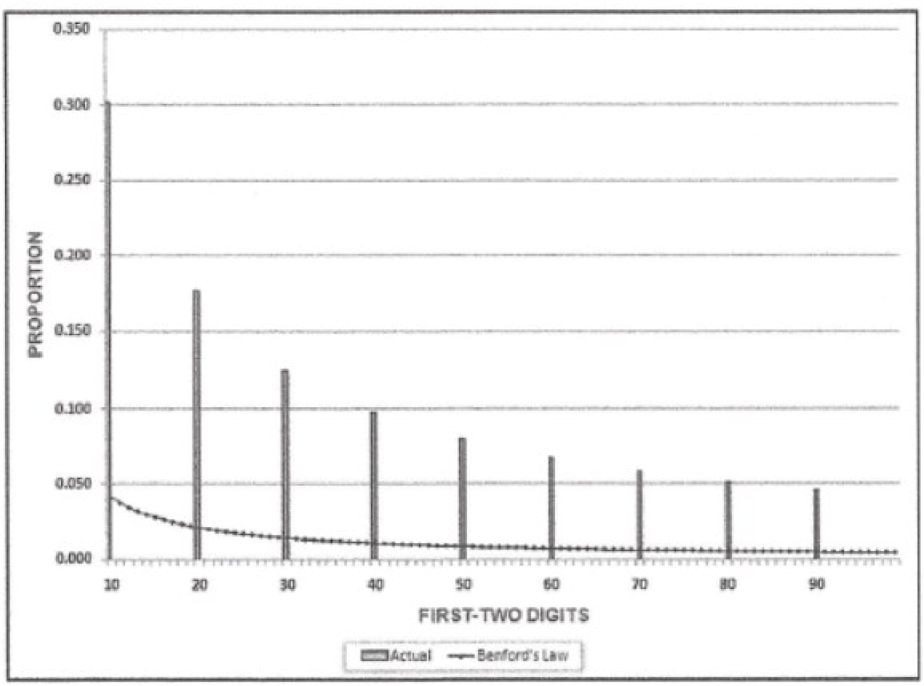

FIGURE 1.8 First-Two Digit Graph of the Data in Table 1.3

Assessing Conformity to Benford's Law

Test	Applied to	Limitations	Cutoff Value	
Mean absolute deviation (MAD)	Whole column		By experience; see Table 7.1	
z-statistic	Single data row For first-two digit combination, For first-three digit combination, For 1^{st} and 2^{nd} digit	The z-statistic becomes larger as the difference between the observed (actual) proportion and expected proportion becomes larger.	At a significance level of 5%, the cutoff score is 1,96. At a 1% significance level our cutoff score would be 2,57. ToDo. … videos analysieren … S.68	
Chi-square	Whole column	maximum N of 2.500 rows; suffers from the excess power problem in that when the data table becomes very large, the calculated chi-square will almost always be higher than the cutoff value	Use MS-Excel CHIINV(probability, degree_freedom) function to calculate cutoff value	
Kolmogorov-Smirnoff	For cumulative sum of the expected proportions of the first-two digits			
Mantissa Arc				

The higher the MAD, z-statistic or chi-square the larger the average difference between the actual and expected proportions.

Mean Absolute Deviation (MAD)

Achtung: Nigrini verwendet Mittelwert, nicht Median wie bei Wikipedia. Die Formel ist auch anders!
Die MAD Zahl wird für die gesamte Spalte berechnet.

The Mean Absolute Deviation (MAD) test ignores the number of records, N. The MAD is calculated using Equation 6.4:

$$\text{Mean Absolute Deviation} = \frac{\sum_{i=1}^{K} |AP - EP|}{K} \tag{6.4}$$

where EP denotes the expected proportion, AP the actual proportion, and K represents the number of bins (which equals 90 for the first-two digits).

AP … actual proportion
EP … expected proportion
K ….. number of bins; 9 für first digit test, 10 für second digit test, 90 für first-two digit test

Mean Absolute Deviation for First Digits

Table 6.1 on page 115 gives the MAD conclusions for the first-two digits. The MAD ranges for the first digits are shown below:

Digits	Range	Conclusion
First Digits:	0.000 to 0.006	Close conformity
	0.006 to 0.012	Acceptable conformity
	0.012 to 0.015	Marginally acceptable conformity
	Above 0.015	Nonconformity

TABLE 7.1 Critical Values and Conclusions for Various MAD Values

Digits	Range	Conclusion
First Digits	0.000 to 0.006	Close conformity
	0.006 to 0.012	Acceptable conformity
	0.012 to 0.015	Marginally acceptable conformity
	Above 0.015	Nonconformity
Second Digits	0.000 to 0.008	Close conformity
	0.008 to 0.010	Acceptable conformity
	0.010 to 0.012	Marginally acceptable conformity
	Above 0.012	Nonconformity
First-Two Digits	0.0000 to 0.0012	Close conformity
	0.0012 to 0.0018	Acceptable conformity
	0.0018 to 0.0022	Marginally acceptable conformity
	Above 0.0022	Nonconformity
First-Three Digits	0.00000 to 0.00036	Close conformity
	0.00036 to 0.00044	Acceptable conformity
	0.00044 to 0.00050	Marginally acceptable conformity
	Above 0.00050	Nonconformity

Z-Statistic

$$Z = \frac{|AP - EP| - \left(\frac{1}{2N}\right)}{\sqrt{\frac{EP(1 - EP)}{N}}} \qquad (6.1)$$

where EP denotes the expected proportion, AP the actual proportion, and N the number of records. The $(1/2N)$ term is a continuity correction term and is only used when it is smaller than the first term in the numerator.

The Z-statistic is used to test whether the actual proportion for a specific first-two digit combination differs significantly from the expectation of Benford's Law. The formula takes into account the absolute magnitude of the difference (the numeric distance from the actual to the expected), the size of the data set, and the magnitude of the expected proportion.

The Z-statistics cannot be added or combined in some other way to get an idea of the overall extent of nonconformity.

The z-statistic becomes larger as the difference between the observed (actual) proportion and expected proportion becomes larger.

Chi-Square Test

Wird für gesamte Spalte berechnet!

The chi-square test is often used to compare an actual set of results with an expected set of results. Our expected result is that the data follows Benford's Law.
The hypothesis is that the first two digits of the data follow Benford's Law. The chi-square statistic for the digits is calculated as shown:

$$\text{chi-square} = \sum_{i=1}^{K} \frac{(AC - EC)^2}{EC} \qquad (6.2)$$

where AC and EC represent the Actual Count and Expected Count respectively, and K represents the number of bins (which in our case is the number of different first-two digits). The summation sign indicates that the results for each bin (one of the 90 possible first-two digits) must be added together

AC .. Actual Count
EC .. Expected Count
K ….. number of bins; 9 für first digit test, 10 für second digit test, 90 für first-two digit test

The chi-square test is often used to compare an actual set of results with an expected set of results. In this case our expected result is that the data follows Benford's Law. The null hypothesis is that the first two digits of the data follow Benford's Law.
The calculated chi-square statistic is compared to a cutoff value which can be calculated in Excel by using the CHIINV function. For example, CHIINV(0.05,89) equals 112.02. If the calculated chi-square value exceeds 112.02 then the null hypothesis of conformity of the first-two digits must be rejected and we would conclude that the data does not conform to Benford's Law.

MS-Excel: CHIINV(probability, deg_freedom)
The CHIINV function syntax has the following arguments:
Probability Required. A probability associated with the chi-squared distribution.
Deg_freedom Required. The number of degrees of freedom.

- ⇨ Um chi-square in C# zu programmieren, muss man wissen, wie die Formel für die CHIINV Funktion lautet

15,50731306	=CHIINV(0,05; 8)
16,9189776	=CHIINV(0,05; 9)
112,0219857	=CHIINV(0,05; 89)
20,09023503	=CHIINV(0,01; 8)
21,66599433	=CHIINV(0,01; 9)
122,9422068	=CHIINV(0,01; 89)

The calculated chi-square statistic is compared to a cutoff value. A table of cutoff scores can be found in most statistics textbooks. These cutoff values can also be calculated in Excel by using the CHIINV function.
The higher the calculated chi-square statistic, the more the data deviate from Benford's Law.

The chi-square statistic also suffers from the excess power problem in that when the data table becomes very large, the calculated chi-square will almost always be higher than the cutoff value making us conclude that the data does not follow Benford's Law. This problem starts being noticeable for data tables with more than 5.000 records.
This means that small differences, with no practical value, will cause us to conclude that the data does not follow Benford's Law.
It was precisely this issue that caused the developers of IDEA to build a maximum N of 2.500 into their Benford's Law bounds.

<u>The chi-square test is also not really of much help in forensic analytics because we will usually be dealing with large data tables</u>.