Two Digit Testing for Benford's Law

Dieter W. Joenssen[1,2]
[1] University of Technology Ilmenau, Ilmenau, Germany
[2] Corresponding author: Dieter W. Joenssen, e-mail: Dieter.Joenssen@TU-Ilmenau.de

Abstract

Benford's law has been used by auditors to help reveal data manipulation not only in the context of tax audits and corporate accounting, but also election fraud. The principle idea behind Benford's law is that the frequency distribution of the first digits from numerical data of certain processes should conform to a discrete distribution known as the Benford distribution. Goodness-of-fit tests have been used to assess if the data's first digits conform to said distribution. When data should conform to Benford's law, a null-hypothesis rejection suggests that some form of data manipulation has taken place. Goodness-of-fit tests, like all tests of statistical significance, are prone not only to the type I error, which is limited by the chosen level of significance, but also to the type II error which decreases not only with sample size but is also inherently lower for some testing procedures than others. One possible procedural change is not to test the distribution of the data's first digit, as is the current standard, but to test the joint distribution of the data's first two digits. The gain in power would be due to an increased utilization of information, because, given that the null hypothesis is true, the distributions of the first and second digits are not independent. This paper describes how four goodness-of-fit tests can be extended to test the joint distribution of the first and second digit for conformity to Benford's law. Additionally, a comparison of power yielded by the original (one-digit) as well as the proposed (two-digit) analysis is provided.

Keywords: fraud detection, goodness-of-fit test, power comparison, type II error

1. Introduction

Auditors use Benford's law to help reveal data manipulation by testing whether or not the data's first digits conform to Benford's distribution. This law of anomalous numbers, as Benford (1938) called it, states that the logarithm's first digits, of certain data, is uniformly distributed. In other words, a random continuous variable X, with realization x, is said to follow Benford's law, if the application of the first-k-significant-digits function, $D_k(x)$, yields a Benford-distributed discreet variable. Specifically

$$p_{d_k} = P(D_k(X) = d_k) = log_{10}\left(1 + d_k^{-1}\right) \forall k \in \mathbb{N}^+ \tag{1}$$

with

$$D_k(x) = \left\lfloor |x| \cdot 10^{\left(-1 \cdot \lfloor log_{10}|x| \rfloor + k - 1\right)} \right\rfloor \tag{2}$$

where $d_k \in \{10^{k-1}, 10^{k-1}+1, \ldots, 10^k - 1\}$ (cf. Hill, 1995, p. 354).
These predicted probabilities, especially for the first significant digit, have been shown to hold not only for some theoretical distributions (cf. Leemis et al., 2000), but also for real data sets (for examples see (Hill, 1995, p. 355)). Given that data resulting from any geometric growth process, by definition, follow Benford's law, makes its usage for fraud detection, where exponential growth happens, natural and explains the widespread use to detect fraud in tax audits and corporate accounting (cf. Nigrini, 1996; Nigrini and Mittermaier, 1997; Swanson et al., 2003; Watrin et al., 2008; Rauch et al., 2011).
If it has been established beforehand that Benford's law must be true for un-manipulated data, the interpretation and usage of results from a statistical test is straightforward. The significant digits' observed frequencies are tested against the expected frequencies, p_{d_k},

using a goodness-of-fit test. If the null-hypothesis is rejected, it can be assumed that data manipulation has occurred and a more stringent audit should be conducted. As with any significance test, the chosen level of significance is the maximum proportion of false positives in the long run and thus represents those records recommended for a more thorough audit, even though no manipulation should be found. In this sense, the α-error quantifies the expected proportion of sunk costs due to auditing non-fraudulent data, and the β-error quantifies the expected proportion of records where manipulation is present, but no further auditing is scheduled. Accordingly, tests with a higher power will unequivocally lead to a more efficient detection of fraud.

Options for reducing the proportion of false negatives in statistical testing include increasing sample size, reducing the confidence level, or changing the approach to testing. Ceteris paribus, all tests' power increase monotonically with sample size, but the size of some data sets may not be increased and for others the cost of increasing sample size may be prohibitively large. Decreasing the level of confidence used in testing may offer a reduction of false negatives, if resources for auditing the increased amount of false negatives are available. Changing the approach to testing can involve using a different test that has higher inherent power or using more information provided by the same data. Certainly, new tests specific to the Benford distribution may be developed, but the power properties of all tests could be improved by using more information afforded by the same data.

Testing for Benford's law is usually performed using only the first significant digit (cf. any of Nigrini, 1996; Nigrini and Mittermaier, 1997; Leemis et al., 2000; Cho and Gaines, 2007; Morrow, 2010; Rauch et al., 2011). Relatively recently Diekmann (2007) proposed that the decision, whether or not data conform to Benford's law, should not focus on the first significant digit, but on the second significant digit. The frequencies of the second significant digit, a marginal distribution Benford's prediction for two digits (cf. table 1), may easily be derived using equation (1). Reasoning behind this, based on experimental results by Diekmann (2007), is that the deviations between observed and expected frequencies in the second significant digit are more pronounced than in the first, when data are manually falsified. This novel recommendation only considers different, but not more, information, again limiting fraud detection to those types of fabricated data that exhibit signs of fraud only in the second digit.

Table 1: First and second digits' joint and marginal distributions, in %, rounded

		Second significant digit										
		0	1	2	3	4	5	6	7	8	9	$\approx \Sigma$
First significant digit	1	4.14	3.78	3.48	3.22	3.00	2.80	2.63	2.48	2.35	2.23	30.10
	2	2.12	2.02	1.93	1.85	1.77	1.70	1.64	1.58	1.52	1.47	17.61
	3	1.42	1.38	1.34	1.30	1.26	1.22	1.19	1.16	1.13	1.10	12.49
	4	1.07	1.05	1.02	1.00	.98	.95	.93	.91	.90	.88	9.69
	5	.86	.84	.83	.81	.80	.78	.77	.76	.74	.73	7.92
	6	.72	.71	.69	.68	.67	.66	.65	.64	.63	.62	6.69
	7	.62	.61	.60	.59	.58	.58	.57	.56	.55	.55	5.80
	8	.54	.53	.53	.52	.51	.51	.50	.50	.49	.49	5.12
	9	.48	.47	.47	.46	.46	.45	.45	.45	.44	.44	4.58
$\approx \Sigma$		11.97	11.39	10.88	10.43	10.03	9.67	9.34	9.04	8.76	8.50	

Adjusting the significance level and then testing both digits' frequencies independently would mitigate this problem; but the expected frequencies of the first and second digits are not independent. This is shown by Hill (1995, p. 355) and may be readily tabulated using table 1. Testing both digits in this fashion, whether using the Bonferroni or Šidák

adjustment, will yield a conservative and thus less powerful test, even though more information is being utilized. Due to this, testing the first two digits' joint distribution should lead to increased power for currently available discrete distribution goodness-of-fit tests. Section 2 shows a comparison of the proposed new approach, and the conventional approach of testing only the first digit. The comparison is made via Monte Carlo simulation for four popular tests. Section 3 summarizes results, offers a critical assessment of the proposed approach, and highlights possible areas for further research.

2. A Monte Carlo Comparison of Approaches

The following section highlights possible differences in power if testing is extended from the first significant digit to the first two significant digits. This is achieved by a Monte Carlo simulation implemented in the R framework for statistical computing (R Core Team, 2012). To this end, section 2.1 briefly describes not only the considered test statistics, but also the five distributions for which the comparison is made, and further relevant procedural details. The results are reported in section 2.2.

2.1. Methodology

Four goodness-of-fit tests are selected to determine the magnitude of any differences in power between using only the first digit and using the first two digits when testing for Benford's law. The test statistics compared are Pearson's χ^2 (Pearson, 1900), Kolmogorov-Smirnov's D (Kolmogorov, 1933), Freedman's modification of Watson's U_n^2 for discrete distributions (Freedman, 1981) and the J_P^2 correlation statistic, a Shapiro-Francia type test (Shapiro and Francia, 1972). All tests are available in the R-package **BenfordTests** (Joenssen, 2013). Critical values for each of the four tests are pre-computed at an $\alpha = 5\%$ level of significance via one million simulated replications for the twelve sample sizes considered (cf. table 2).

Table 2: Critical values $\alpha = .05$, using 1 000 000 random samples, rounded

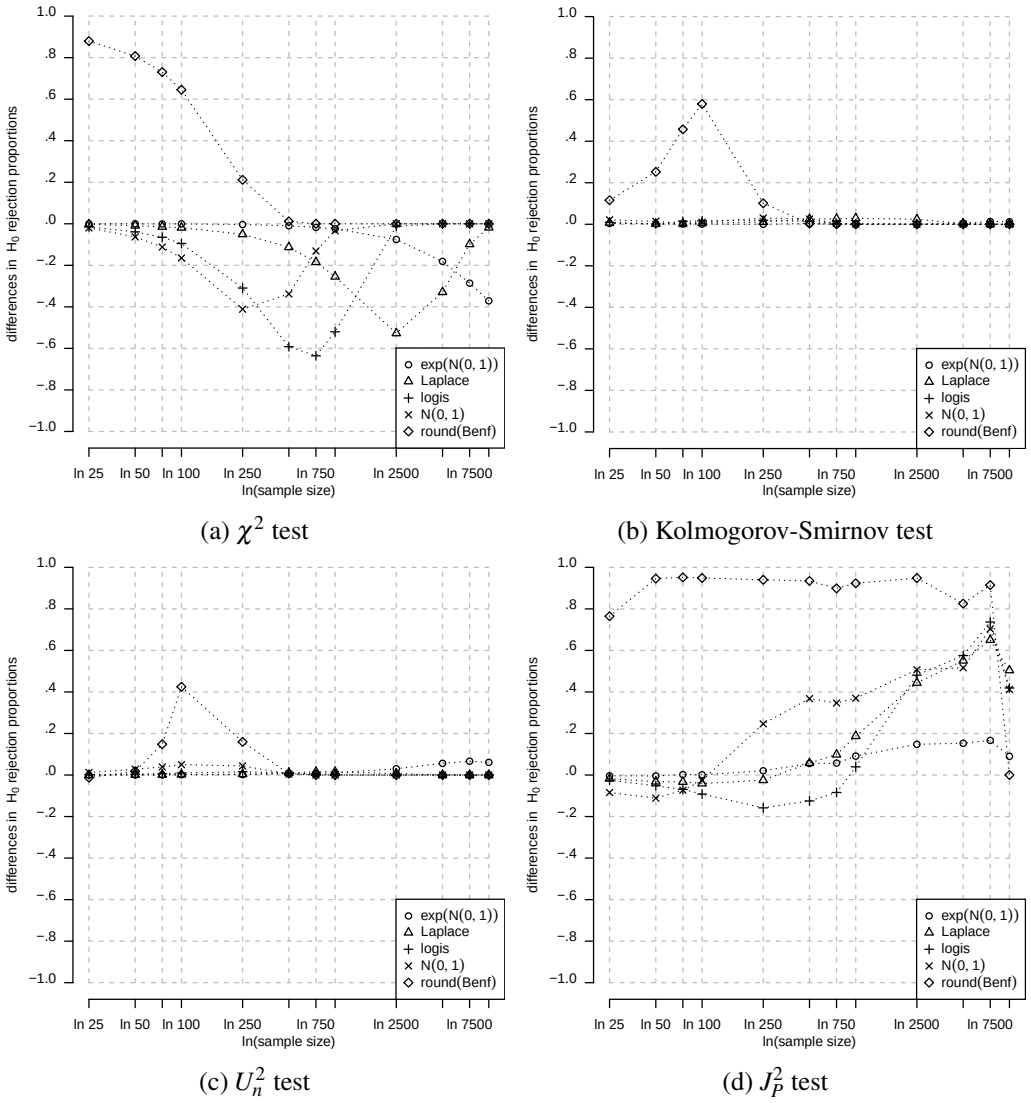
Statistic	χ^2		D		U_n^2		J_P^2	
Sig. digits	1	2	1	2	1	2	1	2
25	15.643	120.752	1.142	1.287	.177	.173	.222	.024
50	15.493	116.921	1.139	1.285	.178	.173	.515	.090
75	15.504	115.502	1.146	1.285	.178	.174	.659	.158
100	15.492	114.684	1.146	1.288	.178	.173	.740	.221
250	15.480	113.042	1.147	1.286	.178	.174	.894	.478
500	15.480	112.537	1.146	1.287	.178	.174	.947	.669
750	15.465	112.398	1.147	1.288	.179	.174	.964	.759
1 000	15.491	112.261	1.148	1.289	.178	.174	.973	.810
2 500	15.501	112.125	1.148	1.287	.179	.174	.989	.916
5 000	15.489	112.093	1.147	1.289	.178	.174	.994	.956
7 500	15.504	112.126	1.148	1.288	.178	.174	.996	.970
10 000	15.485	112.097	1.146	1.288	.178	.174	.996	.977

The appropriate, pre-computed critical values are used to test five distributions, all known not to follow Benford's law. Distributions tested include the standard normal, standard log-normal, standard Laplace and standard logistic. The selection of these first four alternative distributions is more or less arbitrary, while the last distribution may be of additional interest. This distribution, dubbed the rounded Benford-distribution, is constructed so that rounding to the nearest significant digit may influence its value. For example, 6.67 is rounded to 7 and 3.14 is rounded to 3 if the first significant digit is tested, while 6.67 is rounded to 6.7 and 3.14 is rounded to 3.1 if the second significant

digit is tested. All distributions are tested for sample sizes equaling 25, 50, 75, 100, 250, 500, 750, 1 000, 2 500, 5 000, 7 500, and 10 000. Null-hypothesis-rejections of each combination of number of digits tested, sample size, alternative distribution, and test statistic are counted for 1 000 000 replications. These rejections are compared between testing one and two significant digits to estimate the differences in power for the two approaches.

2.2. Results

The results presented in the following section are discussed on a test by test basis referencing sub-figures 1a through 1d. These figures all show the difference in power between testing, as proposed, with the first two significant digits and testing with only the first significant digit. Thus, positive values indicate that testing with two digits yields superior power, and negative values indicate that testing with only the first significand is dominant.



(a) χ^2 test

(b) Kolmogorov-Smirnov test

(c) U_n^2 test

(d) J_P^2 test

Figure 1: Power deviations between two digit and one digit testing.

The χ^2 test statistic offers a relatively homogeneous picture, as shown by figure 1a. All alternative distributions, with the exception of the rounded Benford, proved to favor the one digit testing approach. Neither approach was widely favored for small sample sizes, where power for both is relatively low. As sample sizes increase, so does the advantage

of testing only one digit. For large sample sizes, where power of both methods approach 100% power, differences recede. Thus, the advantages for testing only the first digit lie solely in the intermediate sample size range. An alternative picture is shown by the comparison in power against the rounded Benford distribution. Here, testing with two significant digits is clearly the superior option, especially for the small sample sizes. Again, as testing only the first significant digit increases in power, the deviation in power decreases. This last result is especially surprising, as it is common knowledge that the power of the χ^2 test should decrease as the number of classes increase.

When referring to figure 1b, it is apparent that testing the first two digits with the Kolmogorov-Smirnov test is never inferior in power relative to utilizing only the first digit with the same test. For most distributions the improvement peaks between 1 and 3 percentage points, a statistically significant deviation, albeit a small difference overall. Again, notable exception is the rounded Benford distribution. For this distribution, the advantages of testing the first two digits' joint distribution reaches up to 60 percentage points for the small sample sizes.

The results for Freedman's extension of Watson's U_n^2 statistic, as shown in figure 1c, exhibit deviations similar to those calculated for the Kolmogorov-Smirnov test. Involving two digits is clearly the superior strategy when testing with the U_n^2 statistic. Deviations peak at around 5 percentage points in favor of utilizing an additional digit for testing, considering four of the five distributions. Beyond this, the rounded Benford distribution achieves higher power more quickly when testing with two digits, with differences climaxing at about 40 percentage points.

The most discernible differences in power when testing the first two significant digits or only the first are shown by the results for the J_P^2 statistic (cf. figure 1d). For this test, testing only the first significant digit is beneficial for the smaller sample sizes of the normal, Laplace or logistic distribution. Nonetheless, it becomes advantageous to test two digits for all distributions. For the logistic distribution this change in preference happens at a sample size of about 1 000, for the Laplace at about 300 and for the normal distribution at about 100. The log-normal and rounded Benford distributions favor two digit testing for all inspected sample sizes. Advantages in either direction can be considered quite drastic with up to 20 percentage points in favor of testing only the first digit for the logistic distribution, and up to about 95 percentage points in favor of utilizing the first two digits for the rounded Benford distribution.

3. Conclusions

The conjecture that utilizing more information may translate into power gains for goodness-of-fit tests holds in the context of evaluating data's conformity to Benford's law. While the approach is not universally superior, independent of the statistical test, it is superior for at least one alternative distribution for every test. For two of the tests considered, the Kolmogorov-Smirnov and Freedman-Watson test, power could be gained in every instance by taking into consideration an additional significant digit. Admittedly, an expected improvement of 4 to 5 percentage points may be small in some contexts, but in others, for example when auditing tax returns, this advantage can translate into enormous gains. The results for the χ^2 and J_P^2 statistics show that a more varied procedure based on sample size and suspected deviation should be followed and may require the development of a mixed strategy approach.

Although these results are unambiguous, some questions remain unanswered. While the wide range of sample sizes simulated does make the study comprehensive in this aspect, other facets are considered on a more limited scale in the space allotted. Four tests and five distributions are hardly all-encompassing as more tests are available in literature and a plethora of alternative distributions may also be investigated. Further information of interest could relate to which combination of test and digit count presents the absolute

best procedure to determine if Benford's law is violated.

Further, these results raise questions and indicate possibilities for further research. First, there seem to exist certain classes of alternative distributions where multi-digit testing is always superior. Identifying which alternative distributions appear under certain conditions could lead to recommendations on the best test to use. Second, there may be an optimal number of significant digits to use in testing. The use of an additional, third or fourth significant digit may indeed lead to further power improvements. Then again, the ideal number may be context dependent. Lastly, using multiple digits makes the usage of statistics reserved for continuous data more plausible, paving the way for the development of further goodness-of-fit tests.

References

Benford, F. (1938) "The law of anomalous numbers," *Proceedings of the American Philosophical Society*, 78, 551–572.

Cho, W.K.T. and Gaines, B.J. (2007) "Breaking the (Benford) law: Statistical fraud detection in campaign finance," *The American Statistician*, 61, 218–223.

Diekmann, A. (2007) "Not the first digit! Using Benford's law to detect fraudulent scientific data," *Journal of Applied Statistics*, 34, 321–329.

Freedman, L.S. (1981) "Watson's U_n^2 statistic for a discrete distribution," *Biometrika*, 68, 708–711.

Hill, T.P. (1995) "A statistical derivation of the significant digit law," *Statistical Science*, 10, 354–363.

Joenssen, D.W. (2013) *BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law*, R package version 1.0.

Kolmogorov, A.N. (1933) "Sulla determinazione empirica di una legge di distibuzione," *Giornale dell'Istituto Italiano degli Attuari*, 4, 83–91.

Leemis, L.M., Schmeiser, B.W. and Evans, D.L. (2000) "Survival distributions satisfying Benford's law," *The American Statistician*, 54, 236–241.

Morrow, J. (2010) "Benford's law, families of distributions and a test basis," http://www.johnmorrow.info/projects/benford/benfordMain.pdf.

Nigrini, M.J. (1996) "A taxpayer compliance application of Benford's law: Tests and statistics for auditors," *Journal of the American Taxation Association*, 18, 72–91.

Nigrini, M.J. and Mittermaier, L.J. (1997) "The use of Benford's law as an aid in analytical procedures," *Auditing A Journal of Practice and Theory*, 16, 52–67.

Pearson, K. (1900) "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine Series 5*, 50, 157–175.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rauch, B., Göttsche, M., Brähler, G. and Engel, S. (2011) "Fact and fiction in EU-governmental economic data," *German Economic Review*, 12, 243–255.

Shapiro, S. and Francia, R. (1972) "An approximate analysis of variance test for normality," *Journal of the American Statistical Association*, 67, 215–216.

Swanson, D., Cho, M.J. and Eltinge, J. (2003) "Detecting possibly fraudulent or error-prone survey data: Using Benford's law," *Proceedings of the Section Research Methods, American Statistical Association*, 4172–4177.

Watrin, C., Struffert, R. and Ullmann, R. (2008) "Benford's law: An instrument for selecting tax audit targets?" *Review of Managerial Science*, 2, 219–237.