Journal of Applied Statistics
Vol. 34, No. 3, 321–329, April 2007

Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data

ANDREAS DIEKMANN

Swiss Federal Institute of Technology Zurich, Switzerland

ABSTRACT *Digits in statistical data produced by natural or social processes are often distributed in a manner described by 'Benford's law'. Recently, a test against this distribution was used to identify fraudulent accounting data. This test is based on the supposition that first, second, third, and other digits in real data follow the Benford distribution while the digits in fabricated data do not. Is it possible to apply Benford tests to detect fabricated or falsified scientific data as well as fraudulent financial data? We approached this question in two ways. First, we examined the use of the Benford distribution as a standard by checking the frequencies of the nine possible first and ten possible second digits in published statistical estimates. Second, we conducted experiments in which subjects were asked to fabricate statistical estimates (regression coefficients). The digits in these experimental data were scrutinized for possible deviations from the Benford distribution. There were two main findings. First, both digits of the published regression coefficients were approximately Benford distributed or at least followed a pattern of monotonic decline. Second, the experimental results yielded new insights into the strengths and weaknesses of Benford tests. Surprisingly, first digits of faked data also exhibited a pattern of monotonic decline, while second, third, and fourth digits were distributed less in accordance with Benford's law. At least in the case of regression coefficients, there were indications that checks for digit-preference anomalies should focus less on the first (i.e. leftmost) and more on later digits.*

KEY WORDS: Benford, first digit law, digital analysis, data fabrication, distribution of digits from regression coefficients

Introduction

The digits in numerical data produced by a large number of very different natural and social processes take the form of a logarithmic distribution described by Benford's law. Given the number and variety of processes that produce Benford-distributed data, it is often assumed that the first (i.e. the leftmost, regardless of the position of the decimal point), second and later significant digits in many kinds of real numerical data adhere to Benford's law. The additional assumption that fabricated or falsified data are detectable through the deviation of their digits from the Benford distribution has been tested recently in several contexts. For example, some studies have reported success in identifying fraudulent information with a check of digital frequencies in tax or other financial data against the Benford distribution (Carslow, 1988; Berton, 1995; Nigrini, 1996; Quick & Wolz, 2003). Similar results have been reported for fabricated survey interviews (Schraepel & Wagner, 2005;

Correspondence Address: Andreas Diekmann, Swiss Federal Institute of Technology, Zurich, Switzerland. Email: diekmann@soz.gess.ethz.ch

0266-4763 Print/1360-0532 Online/07/030321–9 © 2007 Taylor & Francis
DOI: 10.1080/02664760601004940

Schäfer *et al.*, 2005). It may well be that 'Benford tests' can also be used to identify fraudulent scientific data or results.

In the empirical sciences, publications often contain large tables with statistical estimates (such as regression coefficients) whose digits might fruitfully be compared with the Benford distribution. In this article, we will empirically investigate the application of the Benford test to regression coefficients and other statistics. Regression coefficients were chosen as an object of study because of their ubiquity in the scientific literature, and not only in fields such as sociology or psychology. Estimates for regression coefficients are, for example, frequently reported in econometrics. Biomedical researchers also use regression analysis or related techniques such as logistic regression.

However, before we can apply Benford tests to these data,[1] it must be demonstrated that the digits of regression coefficients or other statistical estimates are generally distributed in accordance with Benford's law. And, even if there is evidence for this conformance, employing the Benford test to identify fraudulent data also requires that the distribution of fraudulent data deviate from the distribution implied by Benford's law (Figure 3). Good evidence is required for both of these hypotheses before the Benford test can be accepted as a valid procedure for detecting anomalies in scientific publications. The first of the above hypotheses (that real data are Benford distributed) is tested in the next section. In an effort to learn more about the distributional properties of digits from estimated regression coefficients, we collected a large sample of regression coefficients from the published sociology literature. The second hypothesis (that the digits in fraudulent data deviate from the Benford distribution) is tested in the fourth section, where we report on the results of three experiments. In these experiments, students attending university-level statistics courses were asked to fabricate a table of regression coefficients that would support a certain hypothesis. The second hypothesis predicts that the first, second, and later digits in the fabricated data will deviate from Benford's law.

Benford's Law

The logarithmic distribution of the first digit d_1 of various naturally occurring quantities is described by 'Benford's law' or the 'first digit phenomenon' (Hill, 1998; Raimi; 1969, 1976):

$$P(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right) \tag{1}$$

According to the formula, the probability that a number's first digit is '1' is 0.301, while a '9' is expected with a much lower probability of 0.046 (see Table 1).

This phenomenon was discovered by Newcomb (1881), who observed that tables of logarithms were used more often for smaller first digits than for larger ones. Newcomb also derived the formulas for the first and second significant digit (Hill, 1995a, p. 354). (Newcomb's work was forgotten, but 'Newcomb's law' would be a more appropriate label.) Half a century later, Benford (1938) happened upon this regularity through the same observation (Hill, 1995a). Benford went further in computing frequency distributions for the first digits in a variety of data, such as the area of riverbeds, figures published in the newspaper, population statistics and other data. The distribution of first digits in all of these data could be closely approximated by the logarithmic distribution.

A generalized distribution describes the data's other digits. The joint distribution of both the first and later significant digits takes the following form (Hill, 1995a):

$$\mathrm{P}(D_1 = d_1, \ldots, D_k = d_k) = \log_{10}[1 + (\Sigma d_i 10^{k-i})^{-1}] \tag{2}$$

Table 1. Probabilities predicted by Benford's Law for the first and higher-order digits*

d_i	P(d_1)	P(d_2)	P(d_3)	P(d_4)
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.10882	0.10097	0.10010
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.09940	0.09994
7	0.05799	0.09035	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

*Figures are adapted from Nigrini (1996). Computational formulas follow from equation (2). For example, the marginal distribution for the second digit $d_2 = 0, 1, \ldots, 9$ is $\mathrm{P}(d_2) = \Sigma \log_{10}[1 + (10k + d_2)^{-1}]$ with summation $k = 1, 2, \ldots, 9$ (Hill, 1995a, p. 354).

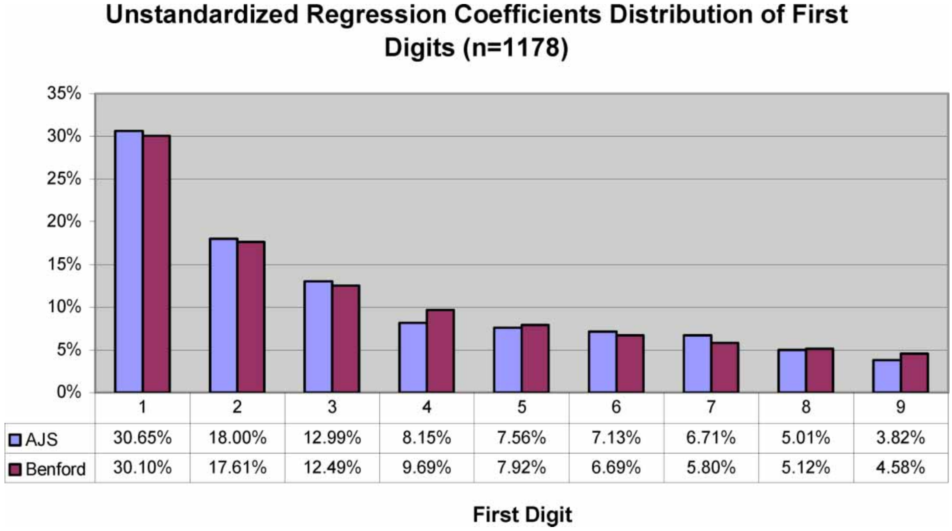
With random variables $D_1, D_2, \ldots, D_k$ containing the first (leftmost, independent of the location of the decimal point), second, $\ldots$, kth significant digits, $d_1 = 1, 2, \ldots, 9$ and $d_j = 0, 1, \ldots, 9$ $(j = 2, \ldots, k)$. For example, if digits are Benford distributed, the combination of significant digits 1028 (e.g. 1.028 or 0.001028) is expected with probability $\log_{10}[1 + 1/1028]$. This 'general significant-digit law' (Hill, 1995a) permits the derivation of the marginal distributions of second-order and *later* digits. Table 1 displays the probabilities for the first four significant digits.

It follows from the joint distribution described above that the distribution of higher-order digits increasingly approximates the uniform distribution.

Since Benford's publication, substantial progress has been made in explaining the mechanism behind the generation of Benford-distributed digits. Hill (1995a, 1998) proved a 'random samples from random distributions theorem'. If one first chooses a sample of distributions at random and then take random samples digits from those distributions, the resulting distribution will–under certain conditions–approximate Benford's law. Also, Hill (1995a) was able to prove the base and scale invariance of Benford's law rigorously. Hence, if Benford's law, for example, applies to the distribution of the digits in data on the area of lakes in units of acres it will (on average) also apply to the same data in units of square metres. Moreover, Hill (1995b) showed that Benford's logarithmic distribution is the only such scale-invariant distribution of significant digits.

Digit Distribution of Statistical Estimates

A necessary prerequisite for the application of Benford tests for the accuracy of any kind of data is that the real (i.e. not fabricated or falsified) data be Benford distributed.[2] Some information exists on the Benford conformity of the digits in raw data, but almost none exists on whether the digits in statistical estimates take the form of the Benford distribution. To our knowledge, with the exception of Becker's (1982) analysis of failure rates, there have been no published investigations of the typical distribution of digits for statistical estimates such as standard deviations or regression coefficients. To examine the use of the Benford distribution as a standard, we created a dataset of first digits from means, standard deviations, correlation coefficients, and standardized and unstandardized regression coefficients (including those from ordinary least squares and logistic regression models), including about 1000 digits for



	1	2	3	4	5	6	7	8	9
□ AJS	30.65%	18.00%	12.99%	8.15%	7.56%	7.13%	6.71%	5.01%	3.82%
■ Benford	30.10%	17.61%	12.49%	9.69%	7.92%	6.69%	5.80%	5.12%	4.58%

Figure 1. Relative frequencies of first digits of regression coefficients from articles published in the *American Journal of Sociology* (Sample 1, Volumes 101 and 102)

each type of statistic. These data were collected from tables published in two volumes of the *American Journal of Sociology* from January 1996 (Vol. 101) to May 1997 (Vol. 102).

The relative frequencies of the first digits of unstandardized regression coefficients closely approximate the Benford distribution. For example, a first digit of '1' has a relative frequency of 0.307 in our sample, while the value predicted by Benford's law is 0.310 (Figure 1). For a significance level of $\alpha = 0.05$, a comparison with the Benford distribution supports the null hypothesis of no difference between the predicted and observed distributions ($\chi^2 = 7.115$, $\text{df} = 8$, $p = 0.524$).

On the other hand, the fit between the distribution of the statistical estimates' first digits and the Benford distribution is much worse for means, standard deviations, correlations, and standardized coefficients (results not shown).

To explore the robustness of the above result and to gather information on the Benford conformity of the estimates' second digits, we inspected an additional sample of regression coefficients (since we had collected only first digits in the initial sample). The second sample was drawn from the same journal as the first and contained 1457 first and second digits from all the tables of (OLS) regression coefficients published in Volume 104, Issues 1–6 (1999) and Volume 105, Issues 1–5 (2000) of the same journal.

Although the χ^2 test results in the rejection of the null hypothesis that the first digits of the second set of regression coefficients are drawn from a Benford distribution ($\chi^2 = 21.07$, $\text{df} = 8$, $p = 0.007$), the approximation is not all that poor in descriptive terms. The significant deviation is caused mostly by the higher-than-expected occurrence of the digit '5', which has a relative frequency 0.101 in the sample of regression coefficients, as compared to an expected frequency of 0.079. Moreover, the second digits are distributed largely in accordance with the monotonic decline of digit frequencies predicted by Benford's law (Figure 2).

The observed distribution of second digits yields a better approximation of the Benford-predicted distribution ($\chi^2 = 7.12, \text{df} = 9, p = 0.524$). Note that the observed values exhibit the typical pattern of a monotonic decline and therefore deviate systematically from a uniform distribution.

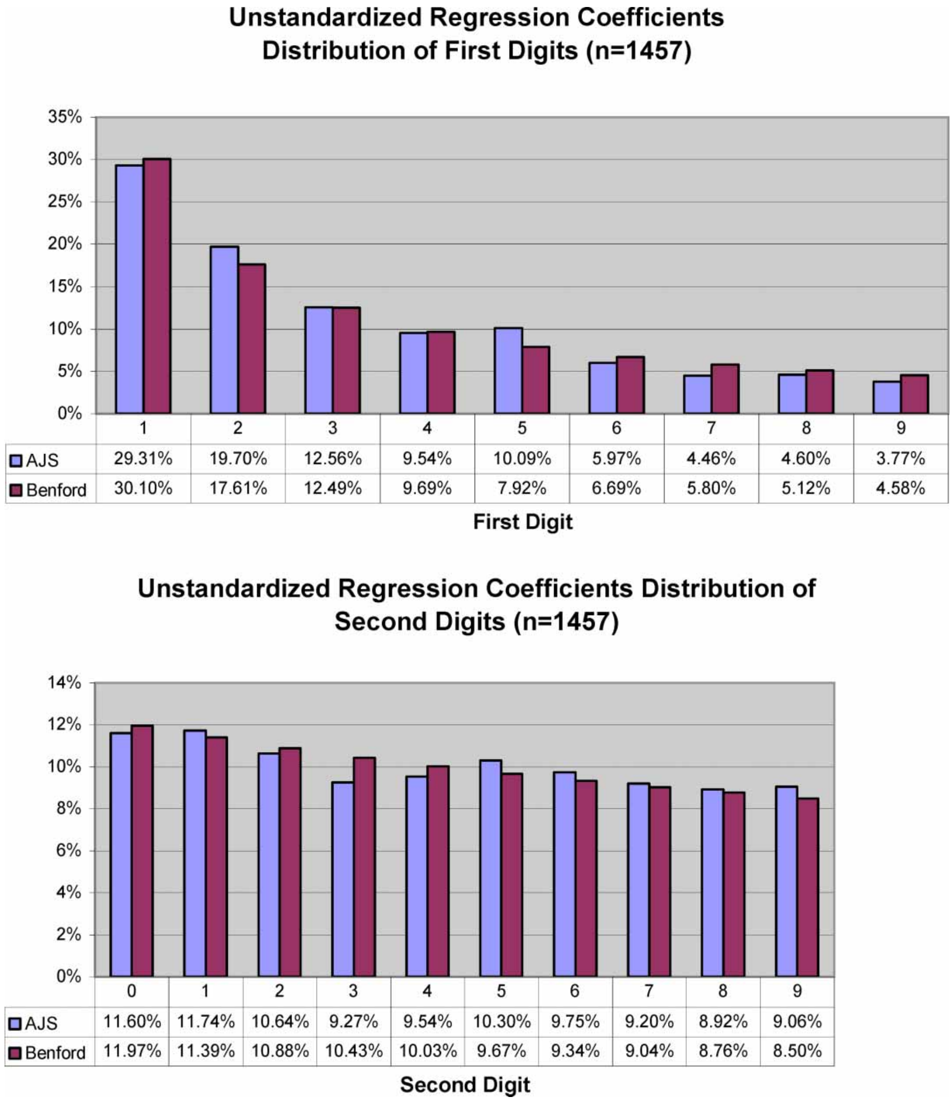
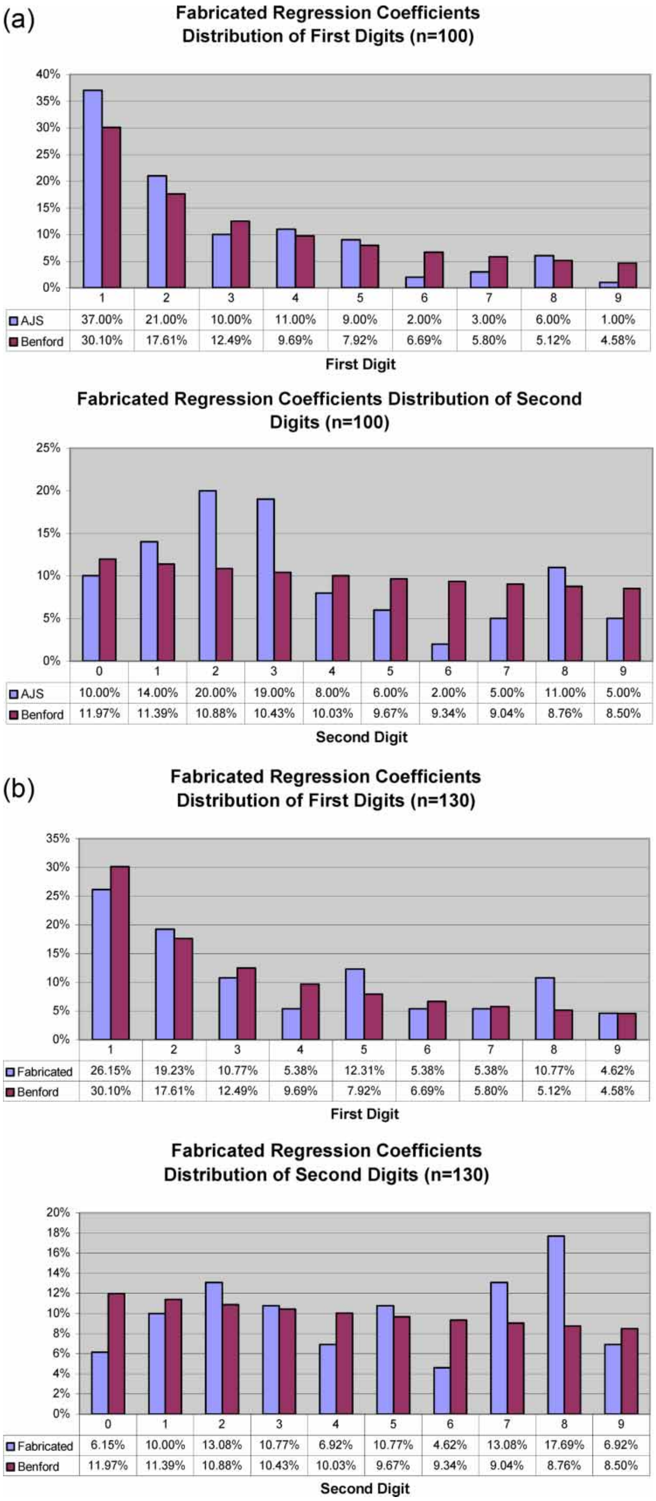


Figure 2. Relative frequencies of first and second digits of regression coefficients from articles published in the *American Journal of Sociology* (Sample 2, Volumes 104 and 105)

In summary, the largest discrepancy between the predicted and observed digit frequencies is 0.022 for a first digit of '5' in this second sample. Furthermore, all of the above tests on regression coefficients reveal the pattern of a monotonic decline in the digital frequencies. Hence, the conclusion that the digits of published unstandardized regression coefficients closely approximate Benford's law is justified.

Experiments with Fabricated Regression Coefficients

The fact that the distribution of digits of published regression coefficients roughly corresponds to the Benford distribution alone would not be enough to justify the use of a Benford test for fabricated data. Fabricated data would have to deviate from the Benford



	1	2	3	4	5	6	7	8	9
■ AJS	37.00%	21.00%	10.00%	11.00%	9.00%	2.00%	3.00%	6.00%	1.00%
■ Benford	30.10%	17.61%	12.49%	9.69%	7.92%	6.69%	5.80%	5.12%	4.58%

	0	1	2	3	4	5	6	7	8	9
■ AJS	10.00%	14.00%	20.00%	19.00%	8.00%	6.00%	2.00%	5.00%	11.00%	5.00%
■ Benford	11.97%	11.39%	10.88%	10.43%	10.03%	9.67%	9.34%	9.04%	8.76%	8.50%

	1	2	3	4	5	6	7	8	9
■ Fabricated	26.15%	19.23%	10.77%	5.38%	12.31%	5.38%	5.38%	10.77%	4.62%
■ Benford	30.10%	17.61%	12.49%	9.69%	7.92%	6.69%	5.80%	5.12%	4.58%

	0	1	2	3	4	5	6	7	8	9
■ Fabricated	6.15%	10.00%	13.08%	10.77%	6.92%	10.77%	4.62%	13.08%	17.69%	6.92%
■ Benford	11.97%	11.39%	10.88%	10.43%	10.03%	9.67%	9.34%	9.04%	8.76%	8.50%

Figure 3. Relative frequencies of first and second digits of fabricated regression coefficients (a) Experiment 1, $n = 100$ (b) Experiment 2, $n = 130$

standard in order for such a test to be meaningful. We therefore tested the Benford conformity of fabricated data. In three separate experiments, students participating in statistics courses at the University of Berne in Switzerland were asked to fabricate regression coefficients that would support a certain hypothesis.[3] Subjects were students, mainly from the departments of sociology (Experiment 1 in January 2001, and Experiment 3 in January 2004) and economics (Experiment 2 in October 2001). Subjects were asked to construct 'plausible values' of regression coefficients that would support a controversial hypothesis from neoclassical economics, and then record these values on a form provided by the researchers. The hypothesis was, 'The higher the unemployment benefits, the longer the duration of unemployment'. They were asked to generate four-digit coefficients for the unemployment benefit variable and nine other independent variables or controls, such as education in years, job experience, gender, and so on. In Experiments 1 and 2, each subject produced the ten coefficients detailed above. The task in Experiment 3 was the same, except that subjects were asked to fabricate ten sets of those ten coefficients, in other words to fabricate 100 four-digit regression coefficients.

A few students produced data that indicated they had either not understood the task or not followed instructions to any meaningful extent; their questionnaires were excluded from the analysis. Data from a total of 10 questionnaires were used from Experiment 1 ($n = 100$ coefficients), 13 questionnaires from Experiment 2 ($n = 130$), and 14 questionnaires from Experiment 3 ($n = 882$). Only four subjects completed the entire Experiment 3 questionnaire within the time allotted (about 35 minutes), while the other ten filled in the questionnaire at least partially. Data were aggregated for analysis across subjects in Experiments 1 and 2, while the large number of fabricated coefficients collected in Experiment 3 allowed for separate analysis of the data for every individual.

The distribution of first digits produced in both Experiments 1 and 2 exhibits a pattern similar to the one predicted by Benford's law (Figure 3). In both experiments, χ^2 tests for the equivalence of the expected and the observed distributions do not permit the rejection of the null hypothesis for $\alpha = 0.05$ (experiment 1: $\chi^2 = 10.64$, df $= 8$, $p = 0.223$; experiment 2: $\chi^2 = 15.30$, df $= 8$, $p = 0.054$), although the test statistic for Experiment 2 just failed to reach the level of statistical significance. More importantly, the shape of the frequency distribution mirrors the monotonic decline of the Benford distribution for both experiments. Thus, data from these experiments do not support the idea that the first digits in fabricated data deviate from Benford's law.

What about the second digit? In both experiments, the observed distributions of the second digits deviate significantly from the Benford distribution (Figure 3) (Experiment 1: $\chi^2 = 27.00$, df $= 9$, $p = 0.001$; Experiment 2: $\chi^2 = 23.57$, df $= 9$, $p = 0.005$). The hypothesis that 'true' regression coefficients follow Benford's law while fabricated data do not is supported by the analysis of the second digits, but not by the analysis of the first.

Of course, a weakness of these experiments is that they permit analysis of only the aggregated data. Assuming that there is individual variance in the falsification patterns, an individual-level analysis might be more informative. The third experiment was therefore conducted to collect enough data from each subject to permit an individual-level analysis.

In principle, the results from the third individual-level experiment are very much in line with those from aggregate-level experiments. Most subjects exhibit fabrication patterns that conform to Benford's law for the first digit, but not for the second or later digits. The pattern of the failure to reject the null hypothesis (Benford distribution) for the first digit and of the rejection of the null hypothesis for the second and later digits is supported by most of the individual-level significance tests conducted for these data: out of 14 tests, three are significant ($\alpha = 0.05$) for the first digit, while ten tests are significant for the second digit, 12 for the third digit, and 13 for the fourth digit (Table 2).[4]

Table 2. Analysis of fabricated data for individual subjects (Experiment 3)

	1st digit			2nd digit			3rd digit			4th digit		
Subject	χ^2	*p* value	*n*	χ^2	*p* value	*n*	χ^2	*p* value	*n*	χ^2	*p* value	*n*
1	**18.49**	**0.018**	100	**30.11**	**0.000**	100	**32.35**	**0.000**	99	**28.28**	**0.001**	98
2	14.14	0.078	100	**23.88**	**0.004**	100	**25.49**	**0.002**	100	**19.29**	**0.023**	100
3	9.08	0.336	100	12.58	0.182	100	**19.59**	**0.021**	100	**33.04**	**0.000**	100
4	7.90	0.443	100	**30.15**	**0.000**	99	**33.70**	**0.000**	93	**35.83**	**0.000**	85
5	5.60	0.692	26	14.75	0.098	26	11.72	0.229	26	12.44	0.190	26
6	9.19	0.326	20	9.44	0.398	20	**24.61**	**0.003**	20	**18.05**	**0.035**	20
7	3.12	0.926	24	**17.16**	**0.046**	24	**57.31**	**0.000**	24	**22.60**	**0.007**	23
8	**34.03**	**0.000**	45	**17.09**	**0.047**	45	**17.69**	**0.039**	45	**38.68**	**0.000**	45
9	7.85	0.448	68	**42.69**	**0.000**	68	**40.91**	**0.000**	68	**25.36**	**0.003**	67
10	6.42	0.601	60	**17.26**	**0.045**	60	**44.47**	**0.000**	60	**103.07**	**0.000**	56
11	9.13	0.331	63	**40.88**	**0.000**	63	**113.22**	**0.000**	62	**162.69**	**0.000**	52
12	13.64	0.092	46	**22.91**	**0.006**	46	**19.64**	**0.020**	46	**40.46**	**0.000**	44
13	**19.39**	**0.013**	50	**23.79**	**0.005**	49	8.33	0.502	47	**29.32**	**0.001**	42
14	5.49	0.705	80	13.40	0.145	80	**22.48**	**0.007**	79	**27.34**	**0.001**	75
All	12.26	0.140	882	**31.83**	**0.000**	880	**59.90**	**0.000**	869	**112.74**	**0.000**	833

Comparison of digit frequencies for faked data with the Benford distribution for the first, second, third, and fourth digits. Bold values indicate significant deviations for $\alpha = 0.05$ (χ^2 with df $= 8$ for first digit and df $= 9$ for second and later digits).

It is not the first digit that matters! This result fits well with the finding by Mosimann *et al.* (1995, 2002) that the inspection of the rightmost digits in fabricated data provides better clues to errors or data fabrication than does the inspection of the first digit. Quite interestingly, subjects favour smaller first digits in fabricating regression coefficients, resulting in a Benford-like pattern for the distribution of first-digits in fabricated data. So, a test for the fabrication of regression coefficients might most fruitfully focus on the second, third or later digits. If second and later digits deviate from the Benford distribution, this deviation may yield an indication that the data have been fabricated. At least for regression coefficients, it appears that using a Benford test of first digits for data fabrication would provide misleading results.

Acknowledgements

For data collection, the author would like to thank Regula Bieri (the third section), Ben Jann (Experiments 1 and 3), and Kurt Schmidheiny (Experiment 2). The author would also like to thank Niels Lepperhoff, Peter Preisendörfer, and an anonymous referee for valuable hints. The author is very much indebted to Elisabeth Coutts for helpful comments and assistance. This research was supported by a grant from the ETH Zurich.

Notes

[1] In the context of this analysis, ‘data’ are not raw data but statistical estimates usually published in tables of regression coefficients.

[2] Of course, the digits in many data sets do not follow the Benford distribution, for example data with upper or non-zero lower limits. For example, in most cases the first digit of measurements of the systolic blood pressure is a ‘1’ (example given by the anonymous referee of this paper), since death occurs above and below certain levels.

[3] For the questionnaire (in German) see: http://www.socio.ethz.ch/diekmann/index

[4] Graphs of all individual distributions can be found at: http://www.socio.ethz.ch/diekmann/index

References

Becker, P. (1982) Patterns in listings of failure-rate and MTTF values and listings of other data, *IEEE Transactions on Reliability*, 31, pp. 132–134.

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

Berton, L. (1995) He's got their number. Scholar uses math to foil financial fraud, *Wall Street Journal*, July 10.

Carslow, C. (1988) Anomalies in income numbers. Evidence of goal oriented behavior, *The Accounting Review*, 63, pp. 321–327.

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

Hill, T. P. (1995b) Base invariance implies Benford's Law, *Proceedings of the American Mathematical Society*, 123, pp. 887–895.

Hill, T. P. (1998) The first digit phenomenon, *American Scientist*, 86, pp. 358–363.

Mosimann, J. E., Wiseman, C. V. & Edelman, R. E. (1995) Data fabrication: can people generate random digits? *Accountability in Research*, 4, pp. 31–55.

Mosimann, J. E., Dahlberg, J. E., Davidian, N. M. & Krueger, J. W. (2002) Terminal digits and the examination of questioned data, *Accountability in Research*, 9, pp. 75–92.

Newcomb, S. (1881) Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics*, 4, pp. 39–40.

Nigrini, M. J. (1996) A taxpayer compliance application of Benford's Law, *The Journal of the American Taxpayer Association*, 18, pp. 72–91.

Quick, R. & Wolz, M. (2003) Benford's Law in deutschen Rechnungslegungsdaten, *Betriebswirtschaftliche Forschung und Praxis*, pp. 208–224.

Raimi, R. A. (1969) The peculiar distribution of first digits, *Scientific American*, 221, pp. 118–120.

Raimi, R. A. (1976) The first digit problem, *American Mathematical Monthly*, 83, pp. 521–538.

Schäfer, C., Schräpler, J.-P., Müller, K.-R. & Wagner, G. G. (2005) Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch –Journal of Applied Social Science Studies*, 125.

Schräpler, J.-P. & Wagner, G. G. (2005) Characteristics and impact of faked interviews in surveys, *Allgemeines Statistisches Archiv*, 89, pp. 7–20.