German Economic Review 11(3): 397–401

Benford's Law and Fraud Detection: Facts and Legends

Andreas Diekmann and Ben Jann
ETH Zurich

Abstract. *Is Benford's law a good instrument to detect fraud in reports of statistical and scientific data? For a valid test, the probability of 'false positives' and 'false negatives' has to be low. However, it is very doubtful whether the Benford distribution is an appropriate tool to discriminate between manipulated and non-manipulated estimates. Further research should focus more on the validity of the test and test results should be interpreted more carefully.*

JEL classification: *C*12, *C*16.

Keywords: Benford's law; fraud detection; false positive; false negative; regression coefficients.

In a recent article in the *German Economic Review*, Tödter (2009) describes the potential of 'Benford's law as an indicator of fraud in economics'. His far-reaching conclusions are based on an analysis of the distribution of the first digits of regression coefficients published in economic journals. His analysis did not include a control group with known faked material, nor an assessment of the validity of the Benford test for detecting fraud. Nevertheless, Tödter arrives at the conclusion that '... violations of Benford's law occurred in about 25% of the articles, far more often than could be expected in untampered samples' (p. 349). Accounting for the nominal α-error, he estimates a 'lower bound' of the proportion of articles containing manipulated estimates of 22%.[1] Although Tödter does not explicitly contend that this proportion is due to fraud, he nevertheless suggests that a high proportion of published articles contain manipulated estimates.

[1] A newspaper article in the *Handelsblatt* (30 November 2009) by Hans Christian Müller reports: 'Einige Ökonomen dürften mächtig kalte Füße bekommen haben, als sie vor einigen Wochen den German Economic Review aufschlugen. In der Fachzeitschrift fand sich ein Artikel mit delikatem Inhalt: Er liefert Hinweise darauf, dass manche Forscher bei ihrer Arbeit mogeln. (...) Bei jedem fünften Artikel fanden sich verdächtige Ungereimtheiten in nennenswertem Ausmaß'.

To ascertain the validity of the Benford test, one has to demonstrate that (1) true data are in accordance with the distribution proposed by Newcomb (1881) and Benford (1938), while (2) manipulated data follow a different distribution. Tödter's (2009) interpretation of the test results appears problematic because it neglects the possibility of a violation of assertion (1). There may be many reasons other than fraud for why regression coefficients deviate from Benford's law. Hence, the hypothesis that the data[2] are Benford distributed (the formal null hypothesis of the Benford test) and the hypothesis that the data were not manipulated may not coincide. This increases the probability of 'false positive' test results with respect to the latter.

Even if assertion (1) is true, observed digit distributions in non-manipulated articles will randomly deviate from Benford and there is a certain (predefined) probability of α (5%, say) that the Benford test will flag such differences as 'significant' (assuming the test has correct size). Hence, in a population of non-manipulated articles, the test would falsely identify 5% of the articles as manipulated. This is the well-known α-error of falsely rejecting the null hypothesis in a statistical test, which has been taken into account by Tödter in his computations.

However, translation of the α-error to the proportion of 'false positives' with respect to the question of whether data were manipulated or not hinges on the validity of assertion (1). If non-manipulated data do not follow Benford's law exactly, then the proportion of 'false positives' must be higher than the nominal α-error. So, what do we know about the distribution of digits of regression coefficients? A necessary condition is that the aggregate distribution from a sample of non-manipulated articles is in accordance with Benford's law. Based on data from published articles in different journals, Diekmann (2007; see also Diekmann, 2002) and Günnel and Tödter (2009) both conclude that the aggregate distribution of digits from unstandardized regression coefficients is very close to the Benford distribution. Nonetheless, one should keep in mind that the fit is not perfect. The analysis by Günnel and Tödter is an example. They investigate the four volumes of *Empirica* from 2003 to 2006. For three of the four volumes (2003, 2005 and 2006) the χ^2 value is highly significant, indicating a deviation from Benford. Only after combining the digits from all volumes is there no significant deviation (p. 279, Table 2). Of course one could argue that the observed deviations are due to samples being contaminated by manipulated data, but this would lead to circular reasoning.[3]

[2] By 'data' we mean raw data or statistics such as estimated regression coefficients.

[3] In fact, the analysis by Tödter (2009) is flawed by such circular reasoning. Tödter refers to the results in Günnel and Tödter (2009) as a justification for the validity of the Benford test (i.e. that non-manipulated data follow Benford's law), but then employs the test to show that a substantial fraction of the *same* data is manipulated. If the latter is true, then the data could not be used as a justification for the former.

Despite these difficulties, assume, for now, that the aggregate condition is satisfied. Unfortunately, this is not quite a sufficient condition for the validity of the Benford test. An aggregate distribution closely following Benford's law can be the result of a mixture of a wide variety of individual distributions that do not comply with the law at all [in fact, as noted by Tödter (2009, p. 342), Benford's law is motivated by Hill (1998) as the result of a mixture of individual distributions]. If subsets of non-manipulated data deviate from Benford, the proportion of false positives is again larger than the α-error. For example, some digits in the internet are Benford distributed (Humenberger, 2008). If you google '19' and '99', you will find that the ratio of the number of hits is almost perfectly in accordance with Benford. However, for '1,999' and '9,999', the ratio is much larger than expected. Of course, the reason for the systematic deviation is that the former number also denotes the year 1999. Likewise, Hungerbühler (2007) shows that digits from all numbers contained in the bible follow a Benford distribution except for an excess of digit '7'. Coming back to economics and regression coefficients, assume that a researcher reports on estimates of rates of return for education. For example, the tables may contain comparisons for gender, region and country. Estimates of rates of return for education are frequently in the range 0.06–0.08 and a Benford test would yield the false positive result of a highly significant χ^2 value, indicating that the estimates are manipulated.

The key message is that as soon as there is individual heterogeneity, that is, as soon as the individual distributions deviate from Benford's law, for example because they depend on topic and research questions, the proportion of false positives must increase. The bias can only go in one direction. For the Benford test to be valid, we have to assume that each single article is a draw from a distribution following Benford's law. In the most extreme case, the proportion of false positives could reach 100% due to individual heterogeneity, despite Benford's law being perfectly fulfilled at the aggregate. In such a situation, all articles would be identified as manipulated even if not a single one actually is.

Günnel and Tödter (2009) are right to be cautious. They do not draw conclusions concerning the proportion of doubtful articles. Because it seems likely that subsets of true data are not in accordance with Benford, the expected proportion of 'false positive' test results is larger than the nominal α-error. An estimate of the proportion of doubtful articles as given by Tödter (2009) is misleading. We do not deny that, presumably, there are many errors in empirical articles. Most errors are from sources such as publication bias, misspecified econometric models, false coding or data errors.[4] Replication is

[4] See the replication study by Dewald *et al.* (1986). Moreover, Ioannidis (2005) had good statistical reasons for his assertion that in biomedical research 'most published research findings are false'. His conclusion follows from α- and β-errors and the crucial assumption of a low *a priori* probability of the truth of a hypothesis. Only Bayes' formula, with no assumption on data manipulation, is necessary for his provocative result.

the most promising remedy to reduce erroneous results in science. However, it seems unlikely that regression coefficients in more than 20% of published articles had been manipulated.

Apart from the problem with false positives, it is questionable whether a test based on digits distributions is a powerful tool at all to discriminate between manipulated and non-manipulated articles. Given typical sample sizes of around 100 regression coefficients per article, a large fraction of 'false negatives' (i.e. a high β-error) is to be expected if distributional differences are only small. Indeed, Bauer and Gross (2009) and Diekmann (2007) provide evidence that the power of the Benford test to detect faked data may be low.[5] A case study gives another hint of the problem of β-error. In 2006, it was discovered that the Norwegian researcher Jon Sudbø falsified data on cancer research (Ekbom *et al.*, 2006). We analyzed the published data with Benford methods. Surprisingly, no significant deviation from the Benford distribution was found.[6] Hence, also with respect to 'false negatives' the Benford procedure does not seem to do very well.

Tests based on Benford's law are repeatedly put forth for fraud detection in the accounting literature and by articles from other disciplines. The charm of the method seems to dispel legitimate doubts of its practicability. Yet doubts arise concerning the discriminatory power of the method. Astonishingly, there is little concern about the validity of Benford tests.

ACKNOWLEDGEMENTS

The work was supported by ETH Research Grant Number TH-8/04-3 on 'Empirical studies in scientific misconduct: prevalence and methods for detecting fraud and errors using the Benford distribution'. We thank Stephanie Eckman and Walter Krämer for helpful comments.

Address for correspondence: Andreas Diekmann, ETH Zurich, Sociology, SEW E 21, CH-8092 Zurich. Tel.: +41 44 632 55 56; fax: +41 44 632 10 54; e-mail: diekmann@soz.gess.ethz.ch

[5] Diekmann's (2007) experiments with faked regression coefficients yield the following estimates for false negative test results: first digit 0.79, second digit 0.29, third digit 0.14, fourth digit 0.07 (Diekmann, 2007, Table 2). Here the sample size n is between 20 and 100. Including only 'larger' samples (n about 100) Diekmann finds false negative proportions of 0.75, 0.25, 0.0, 0.0. He therefore recommends the use of higher-order digits for the inspection of published regression coefficients. Bauer and Gross (2009) provide further evidence from experiments with invented regression coefficients. The proportions of false positive tests are 0.25 for the first digit, 0.20 for the second digit, 0.11 for the third digit and 0.11 for the fourth digit (n about 100). Further experiments with invented numbers are reported by Burns (2009).

[6] Results not yet published. The data were analyzed by Cyrill Bannwart and Goran Jevdjic for a student project supervised by the first author. However, Sudbø falsified the raw data. We do not know whether he also falsified statistical estimates or whether he reported 'correct' estimates based on faked data.

REFERENCES

Bauer, J. and J. Gross (2009), *Difficulties Detecting Fraud? The Use of Benford's Law on Regression Tables*, Mimeo, Institute of Sociology, LMU Munich.

Benford, F. (1938), 'The Law of Anomalous Numbers', *Proceedings of the American Philosophical Society* 78, 551–572.

Burns, B. D. (2009), 'Sensitivity to Statistical Regularities: People (Largely) Follow Benford's Law', in: N. Taatgen and H. van Rijn (eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, Cognitive Science Society, Austin, TX, pp. 2872–2877.

Dewald, W. G., J. G. Thursby and R. G. Anderson (1986), 'Replication in Empirical Economics: The Journal of Money, Credit and Banking Project', *American Economic Review* 76, 587–603.

Diekmann, A. (2002), 'Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung', ITA manuscript 02-04. Available at http://www.oeaw.ac.at/ita/pdf/ita_02_04.pdf (accessed on 27 January 2010).

Diekmann, A. (2007), 'Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data', *Journal of Applied Statistics* 34, 321–329.

Ekbom, A., G. E. M. Helgesen, A. Tverdal, T. Lunde and S. E. Vollset (2006), 'Report from the Investigation Commission appointed by Rikshospitalet, Radiumhospitalet MC and the University of Oslo January 18, 2006'. Available at http://www.rr-research.no/general/docs/ekbom/Report_Investigation_Commission.pdf (accessed on 27 January 2010).

Günnel, S. and K.-H. Tödter (2009), 'Does Benford's Law Hold in Economic Research and Forecasting?', *Empirica* 36, 273–292.

Hill, T. P. (1998), 'The First Digit Phenomenon', *American Scientist* 86, 358–363.

Humenberger, H. (2008), 'Eine elementarmathematische Begründung des Benford-Gesetzes', *Der Mathematikunterricht* 54, 24–34.

Hungerbühler, N. (2007), 'Benfords Gesetz über führende Ziffern. Wie die Mathematik Steuersündern das Fürchten lehrt', EducETH. Available at http://www.educ.ethz.ch/unt/um/mathe/ana/benford/Benford_Fuehrende_Ziffern.pdf (accessed on 27 January 2010).

Ioannidis, J. P. A. (2005), 'Why Most Published Research Findings are False', *PLoS Medicine* 2, 696–701.

Newcomb, S. (1881), 'Note on the Frequency of Use of the Different Digits in Natural Numbers', *American Journal of Mathematics* 4, 39–40.

Tödter, K.-H. (2009), 'Benford's Law as an Indicator of Fraud in Economics', *German Economic Review* 10, 339–351.