A combined test of the Benford hypothesis with anti-fraud applications

A COMBINED TEST OF THE BENFORD HYPOTHESIS WITH ANTI-FRAUD APPLICATIONS

Lucio Barabesi [1], Andrea Cerasa[2], Andrea Cerioli[3] and Domenico Perrotta[2]

[1] University of Siena, Department of Economics and Statistics, Siena, Italy, (e-mail: lucio.barabesi@unisi.it)

[2] European Commission, Joint Research Centre (JRC), Ispra, Italy, (e-mail: andrea.cerasa@ec.europa.eu, domenico.perrotta@ec.europa.eu)

[3] University of Parma, Department of Economics and Management, Parma, Italy, (e-mail: andrea.cerioli@unipr.it)

ABSTRACT: In this work we describe a combined test of the null hypothesis that the significant digits in a random sample of numbers follow Benford's law. We also show the potential of the method for the purpose of fraud detection in international trade.

KEYWORDS: Anomaly detection, Benford's law, sum-invariance, customs data.

1 Motivating framework of data analysis

Most unsupervised fraud detection methods look for anomalies in the data. Therefore, all of these techniques assume that the available data have been generated by an appropriate contamination model. Any parameter of the distribution that models the "genuine" part of the data, say F_0, must then be estimated in a robust way, in order to avoid the well-known masking and swamping effects due to the anomalies themselves (Cerioli *et al.*, 2019b). In the context of fraud detection in international trade, where the value of an individual import transaction X originates from the product of the traded amount v with the unit price β, the available anti-fraud tools are derived from the theory of outlier identification in robust regression; see, e.g., Perrotta *et al.*, 2020b. Under this approach it is then assumed that non-fraudulent transactions for a specific good are generated according to the distribution function

$$F_0(x) = \Phi\left(\frac{x-\beta v}{b}\right), \tag{1}$$

where Φ is the distribution function of a standard Normal random variable. In model (1), the regression slope β corresponds to unit price and $b > 0$ defines the (usually unknown) model variability, which is taken to be constant.

Robust and efficient estimation of β in model (1) may lead to the definition of a "fair" unit price for the good under consideration, against which individual or aggregate transaction prices can be contrasted. Transactions well below the "fair" price may correspond to revenue frauds leading to substantial undervaluation of goods imported into the European Union; see, e.g., European Anti-Fraud Office, 2018, p. 26. The normality assumption in model (1) has proven to be satisfactory in the case of monthly-aggregated trade data (Perrotta *et al.*, 2020b). However it may become less adequate when analyzing individual customs declarations, where multiple populations often occur and a skew distribution may seem more appropriate for the definition of F_0 (Perrotta *et al.*, 2020a). An alternative contamination model based on Benford's law then becomes very useful in such a framework: see Cerioli *et al.*, 2019a.

2 Benford's law

Benford's law (BL, for short) is a fascinating phenomenon which rules the pattern of the leading digits in many types of data. Informally speaking, the law states that the digits follow a logarithmic-type distribution in which the leading digit 1 is more likely to occur than the leading digit 2, the leading digit 2 is more likely than the leading digit 3, and so on. Indeed, the first-digit form of BL gives the probability that the first leading digit equals d, for $d = 1, \ldots, 9$, as

$$\log_{10}\left(1+\frac{1}{d}\right). \tag{2}$$

Another, perhaps even less intuitive, property of Benford's law concerns sum invariance. Given an absolutely-continuous random variable X, in the first digit setting of (2), this property states that, for $d = 1, \ldots, 9$,

$$\mathrm{E}[S(X)\mathrm{I}_{[d,d+1[}(S(X))] = C, \tag{3}$$

where I_E is the indicator function of the set E, while

$$S(x) = 10^{\langle \log_{10}|x|\rangle} \tag{4}$$

is the *significand* of the non-null real number x, $\langle x\rangle = x - \lfloor x \rfloor$ and $C = \log_{10}\mathrm{e}$. First-digit sum invariance thus means that the expected value (3) does not depend on d when X is a Benford random variable. Although (2) and (3) are not equivalent when only the first digit is concerned, they are both implied by the full form of BL, which states that

$$S(X) \stackrel{\mathcal{L}}{=} 10^U, \tag{5}$$

with U a Uniform random variable on $[0,1[$. We refer to Berger & Hill, 2020 for a recent survey of the mathematical properties of BL and to Barabesi *et al.*, 2021 for a thorough study of the relationship between (2) and (3).

3 Tests of the Benford hypothesis

In the motivating framework sketched in §1, Cerioli *et al.*, 2019a investigate the conditions under which Benford's law may yield a reasonable approximation for the first-digit distribution of customs declarations. If Benford's law is expected to hold for genuine transactions, then deviations from the law can be taken as evidence of possible data manipulation. Several exact tests of the Benford hypothesis exist according to which characterization is considered. Those that follow have proven to be useful under a variety of circumstances:

- The chi-square test of the first-digit distribution (2) considered by Barabesi *et al.*, 2018, say χ^2;
- The Hotelling-type test of the sum-invariance property (2) proposed by Barabesi *et al.*, 2021, say Q;
- The Kolmogorov-Smirnov test of the Benford property (4) described in Barabesi *et al.*, 2021, say KS.

Barabesi *et al.*, 2021 show that the combination of χ^2 and Q provides a test which is consistently close to the best solution provided by either χ^2 or Q. We further develop this strategy in two directions. First, we derive the asymptotic joint distribution of χ^2 and Q under Benford's law. This result gives theoretical substance to the observed empirical behavior of the combined test. We then extend our combination strategy to include KS. The proposed extension is extremely relevant in view of the motivating framework of §1, since the performance of the individual tests may vary considerably according to the actual digit generating process when Benford's law does not hold. Our combined test thus provides a powerful, yet robust, solution when the type of departure from Benford's law is unknown, as it happens in anti-fraud applications. Some preliminary simulation results for a sample size of $n = 100$ observations are shown in Table 1, where $L_{\chi^2,Q,KS}$ denotes the newly developed combined test. The alternative data generating models for X are a Lognormal random variable of scale parameter 1 and shape parameter 0.5, and a Generalized Benford random variable of parameter -0.6.

Table 1. *Estimated power of tests of the Benford hypothesis for sample size $n = 100$.*

Alternative	χ^2	Q	KS	$L_{\chi^2,Q,KS}$
Lognormal	0.903	0.926	0.899	0.940
Generalized Benford	0.446	0.466	0.853	0.785

References

BARABESI, L., CERASA, A., CERIOLI, A., & PERROTTA, D. 2018. Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud. *Journal of Business and Economic Statistics*, **36**, 346–358.

BARABESI, L., CERASA, A., CERIOLI, A., & PERROTTA, D. 2021. On Characterizations and Tests of Benford's Law. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2021.1891927.

BERGER, A., & HILL, T. P. 2020. The mathematics of Benford's law: a primer. *Statistical Methods and Applications*. https://doi.org/10.1007/s10260-020-00532-8.

CERIOLI, A., BARABESI, L., CERASA, A., MENEGATTI, M., & PERROTTA, D. 2019a. Newcomb-Benford law and the detection of frauds in international trade. *PNAS*, **116**, 106–115.

CERIOLI, A., FARCOMENI, A., & RIANI, M. 2019b. Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics*, **46**, 235–256.

EUROPEAN ANTI-FRAUD OFFICE. 2018. *The OLAF report 2017*. Tech. rept. Publications Office of the European Union, Luxembourg. https://doi.org/10.2784/652365.

PERROTTA, D., CHECCHI, E., TORTI, F., CERASA, A., & NOVAU, X. A. 2020a. *Addressing Price and Weight heterogeneity and Extreme Outliers in Surveillance Data: The Case of Face Masks*. Tech. rept. JRC121650, EUR 12345 EN. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/817681.

PERROTTA, D., CERASA, A., TORTI, F., & RIANI, M. 2020b. *The Robust Estimation of Monthly Prices of Goods Traded by the European Union*. Tech. rept. JRC120407, EUR 30188 EN. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/635844.