



UNIVERSITÀ DI PARMA

ARCHIVIO DELLA RICERCA

University of Parma Research Repository

Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud

This is the peer reviewed version of the following article:

Original

Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud / Barabesi, Lucio; Cerasa, Andrea; Cerioli, Andrea; Perrotta, Domenico. - In: JOURNAL OF BUSINESS & ECONOMIC STATISTICS. - ISSN 0735-0015. - 36:2(2018), pp. 346-358. [10.1080/07350015.2016.1172014]

Availability:

This version is available at: 11381/2820297 since: 2021-11-09T15:46:42Z

Publisher:

American Statistical Association

Published

DOI:10.1080/07350015.2016.1172014

Terms of use:

Anyone can freely access the full text of works made available as "Open Access". Works made available

Publisher copyright

note finali coverpage

(Article begins on next page)






Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud

Lucio Barabesi, Andrea Cerasa, Andrea Cerioli & Domenico Perrotta

To cite this article: Lucio Barabesi, Andrea Cerasa, Andrea Cerioli & Domenico Perrotta (2016): Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud, Journal of Business & Economic Statistics

To link to this article: <http://dx.doi.org/10.1080/07350015.2016.1172014>

 View supplementary material 

 Accepted author version posted online: 06 Apr 2016.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 

Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud

Lucio Barabesi*

Department of Economics and Statistics
University of Siena, Italy

Andrea Cerasa

Institute for the Protection and Security of the Citizen
European Commission, Joint Research Centre, Ispra, Italy

Andrea Cerioli

Department of Economics and Management
University of Parma, Italy

Domenico Perrotta

Institute for the Protection and Security of the Citizen
European Commission, Joint Research Centre, Ispra, Italy

Abstract

The Newcomb-Benford law for digit sequences has recently attracted interest in anti-fraud analysis. However, most of its applications rely either on diagnostic checks of the data, or on informal decision rules. We suggest a new way of testing the Newcomb-Benford law that turns out to be particularly attractive for the detection of frauds in customs data collected from international trade. Our approach has two major advantages. The first one is that we control the rate of false rejections at each stage of the procedure, as required in anti-fraud applications. The second improvement is that our testing procedure leads to exact significance levels and does not rely on large-sample approximations. Another contribution of our work is the derivation of a simple expression for the digit distribution when the Newcomb-Benford law is violated, and a bound for a chi-squared type of distance between the actual digit distribution and the Newcomb-Benford one.

*This research was partly supported by the project MIUR PRIN “MISURA – Multivariate models for risk assessment” and by the “Technology Transfer Office” project of the 2014-20120 Work Programme of the Joint Research Centre of the European Commission. We acknowledge the collaboration of the Italian Customs, initiated in 2011 with the JRC-OLAF project THESEUS. We thank in particular Marco de Andreis, Director of the Statistical Analysis Office of the Central Direction for Anti-fraud and Controls, and all of his team members, for their long-standing partnership and for providing the data analyzed in this work. We are also grateful to Alessio Farcomeni for discussion on a previous draft of the manuscript.

Keywords: anti-fraud, international trade, multiple testing, outlier detection, significant digit distribution

1 Introduction

The Newcomb-Benford (NB) law is a fascinating phenomenon which rules the pattern of the leading digits in many types of numerical data and mathematical sequences. Informally speaking, the law states that the digits are not uniformly scattered – as one may naively expect – but 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 simplest form of the NB law, dating back to the original discoveries by Newcomb (1881) and Benford (1938), gives the probability that the first leading digit equals d_1 as

$$\log_{10}\left(1 + \frac{1}{d_1}\right), \quad (1)$$

for $d_1 = 1, \dots, 9$. This probability is clearly decreasing with the value of d_1 , and it is higher than 30% for $d_1 = 1$.

In spite of its long history, the mathematical and statistical challenges of the NB law have been recognized only recently. From a mathematical perspective, appropriate versions of this law appear in number theory, such as in the Weyl’s Equidistribution Theorem (Havil, 2008, p. 186), or in integer sequences, such as the celebrated Fibonacci sequence or the factorial sequence (Diaconis, 1977). The NB law is considered as one of the 250 mathematical milestones by Pickover (2009) and is described at length in a book of the monumental series “The art of computer programming” (Knuth, 1997, p. 254). In a probabilistic setting, a deep analysis of the NB law was first carried out by Hill (1995), who proved a limit theorem for the significant-digit distribution; see also Berger and Hill (2011a,b, 2015) and Miller (2015).

Recently, scientists and practitioners have applied the NB law in diverse settings, fraud detection in business accounting being perhaps the most noticeable one (Nigrini, 2012). The anti-fraud rationale behind the use of the NB law is that producing empirical distributions of digits that conform to the law is difficult for non-experts. Fraudsters may thus be biased towards simpler and more intuitive distributions, such as the Uniform. However, most of the applications in this area rely ei-

ther on diagnostic checks of the data, with compliance to the digit distribution implied by the NB law evaluated through graphical displays, or on informal decision rules, when suitable statistics of conformity are applied. For instance, Nigrini (2012) suggests conclusions of “Close conformity”, “Acceptable conformity”, “Marginally acceptable conformity” and “Non conformity” to the law, based on the observed value of the chosen test statistic.

We argue that one problematic issue in formal goodness-of-fit testing of the NB law for anti-fraud purposes lies in the choice of an appropriate version of the null hypothesis. In fact, under a careful statistical approach, the statement that the data are distributed according to the NB law actually involves the distribution of all the significant digits and is not restricted to that of the leading ones. The global NB hypothesis is thus likely to be too stringent for practical use, also because of rounding errors and other non-fraudulent anomalies in few of the reported digits. Another disadvantage of testing the global NB null hypothesis is that rejection does not shed light on how many digits deviate from the law, nor on which digits are responsible for rejection. We are interested in anti-fraud analysis of customs data arising from international trade, where the goal is to detect illegal actions such as tax evasion and money laundering (Deng et al., 2009; Sudjianto et al., 2010). In this context, multiple-digit deviation from the NB law may seem more suspicious than single-digit non-conformity, as a signal of data fabrication. Similarly, non-conformities in the first significant digits might correspond to trade frauds which are economically more relevant, while deviations in the last digits may be likely attributed to the effect of rounding or to other market conditions under which compliance to the law should not be expected (see, e.g., Tam Cho and Gaines, 2007).

At the opposite side of the goodness-of-fit panorama is the idea of testing conformity of individual-digit frequencies to the one-dimensional marginal probabilities prescribed by the NB law. For instance, Tam Cho and Gaines (2007) and Pericchi and Torres (2011) consider individual tests for the first and for the first-two digits, respectively, while Diekmann (2007) suggests to extend scrutiny to later digits as well. Multiplicity issues may occur when several digits are involved, thus increasing the chance of incurring in false alarms. Controlling the number of false discoveries is a crucial problem in many anti-fraud exercises. This requirement is particularly compelling in the study of international trade data, since hundreds of traders must be investigated over thousands of markets and substantial inspection of possible frauds rapidly becomes impractical if the number

of false signal increases. We refer to Cerioli and Perrotta (2014) and to Barabesi et al. (2015) for a detailed description of the statistical challenges involved by such data, and for a precise statement of the anti-fraud tasks afforded by the Joint Research Centre of the European Commission. Simultaneous tests of conformity for many digits, although being less prone to liberality, are not attractive in this context, since they share the same lack of sensitivity to different alternatives as the global one. Furthermore, the corresponding number of degrees of freedom increases in powers of ten with the number of digits under scrutiny and it may easily become large in comparison to the available sample of transactions for each trader, thus invalidating the common confidence in large-sample results.

The aim of our work is to address the issues sketched above and to suggest a new and effective way of testing the NB law. Although our procedure is general, it turns out to be particularly attractive in the context of anti-fraud applications for international trade data. Our goal is reached through a hierarchical procedure where different varying-dimensional marginals of the NB distribution are tested in sequence, starting from a reasonable simplification of the global null and possibly ending with the one-digit marginals. Our approach has two major advantages. The first one is that we explicitly take into account the hierarchical nature of the sequence of tests that we implement and we base our inferences on the resulting conditional distributions. This leads to proper control of the rate of false rejections at each stage of the analysis, as required in our anti-fraud applications, without resorting to multiple-tests adjustments of P -values. The second improvement is that our testing procedure leads to exact significance levels and does not rely on large-sample approximations to the conditional distribution of the test statistics. Instead, we adopt a computationally efficient and easy-to-implement Monte Carlo approximation of the exact distribution of these test statistics. Another contribution of our work is the derivation of a simple and neat expression for the digit distribution under the alternative hypothesis that the NB law is violated. Correspondingly, we derive a bound for a chi-squared type of distance between the actual digit distribution and the NB one. These findings can be useful for comparing the power of different goodness-of-fit statistics under the alternative, and for the purpose of assessing the “degree” of discrepancy between the empirical digit distribution and the NB law.

Our approach is consistent with the view that in financial applications anomalous observations may correspond to fraudulent transactions (Sudjianto et al., 2010, p. 16). Therefore, we see our method as a statistically principled criterion for detecting possible fraudsters in international trade on the basis of the digit distribution of their transactions. Available outlier identification techniques cannot be applied in this context, since they assume that the “good” part of the data follows the Normal distribution (Cerioli, 2010). However, we adopt the same philosophy and take the NB law as our baseline model, to which each trader must be contrasted. The well-known masking and swamping problems (Hubert et al., 2008) do not affect our tests, because we are in the fortunate situation where no parameter has to be estimated under the NB model. We can thus base our anomaly inference on the full sample of transactions for each trader.

Also our strategy towards false discoveries resembles the attitude of some outlier detection tools that have proven to be effective in other anti-fraud exercises (Riani et al., 2009; Cerioli, 2010): we first test a simultaneous hypothesis of conformity in order to control the global test size, and then we move to more detailed investigations on specific digits only when the upper level test is rejected. Sequential multiple-testing techniques that follow related principles have received extensive research interest and have proven to be effective in application fields where a compromise between false discovery control and power enhancement is needed (see, e.g., Goeman and Solari, 2010; Finos and Farcomeni, 2011; Dmitrienko and Tamhane, 2013, and the references therein). However, we emphasize that our procedure does not involve P -values adjustments at each step of the testing hierarchy. These adjustments are not required since we are able to compute an arbitrarily good approximation to the conditional distribution of our test statistics, given the decisions made at the preceding steps. We argue that our improvements could make goodness-of-fit testing for the NB law an appealing tool for the purpose of contrasting illegal financial activities.

The rest of the paper is structured as follows. In §2 we review some theoretical features of the NB law which are important for the purpose of anti-fraud analysis. Our sequential testing procedure is proposed in §3, with the simplest first-two digit case in §3.1 and the general first- k digit situation in §3.2. In §4 we give Monte Carlo estimates of the exact quantiles of our test statistics and we provide simulation comparison with other methods. The k -digit distribution under the alternative hypothesis is derived in §5, together with a bound on a discrepancy measure. The paper

ends with an application to trade data in §6 and with some concluding remarks in §7. Technical details are given in the Appendix.

2 Some theory of the NB law in view of anti-fraud applications

For each $x \neq 0$, define the significant function $s : \mathbb{R} \mapsto [1, 10[$ as

$$s(x) = 10^{\log_{10} |x| - \lfloor \log_{10} |x| \rfloor},$$

where $\lfloor \cdot \rfloor$ denotes the “floor function”, i.e. the largest integer less than or equal to the argument value. When $x = 0$, we assume that $s(0) := 0$. Clearly, $\log_{10} s(x)$ represents the fractional part of $\log_{10} |x|$. In addition, the first significant digit of x , say $D_1(x)$, may be rephrased in terms of the significant function, i.e.

$$D_1(x) = \lfloor s(x) \rfloor,$$

while, for $k = 2, 3, \dots$, the k -th significant digit of x , say $D_k(x)$, may be rewritten as

$$D_k(x) = \lfloor 10^{k-1} s(x) \rfloor - 10 \lfloor 10^{k-2} s(x) \rfloor.$$

Let us consider the random variable (r.v.) X defined on the probability space (Ω, \mathcal{F}, P) . This r.v. follows the NB law (Berger and Hill, 2011a, p. 23) if

$$P(s(X) \leq t) = \log_{10} t \tag{2}$$

for $t \in [1, 10[$. Trivially, condition (2) is equivalent to assume that $P(\log_{10} s(X) \leq u) = u$ for $u \in [0, 1[$, i.e. that the fractional part of the r.v. $\log_{10} |X|$ is uniformly distributed in $[0, 1[$. In the following, X will always represent a NB r.v..

Equivalently, Berger and Hill (2011a) emphasize that condition (2) implies (and vice versa) the following joint probability function (p.f.) for all $k \in \mathbb{N}$:

$$P(D_1(X) = d_1, \dots, D_k(X) = d_k) = \log_{10} \left(1 + \frac{1}{c_{d_1, \dots, d_k}} \right), \tag{3}$$

where $d_1 \in \{1, \dots, 9\}$, $d_l \in \{0, \dots, 9\}$ for $l = 2, \dots, k$, while

$$c_{d_1, \dots, d_k} = \sum_{l=1}^k 10^{k-l} d_l.$$

Obviously, $c_{d_1, \dots, d_k} \in \mathbb{N}$ and $10^{k-1} \leq c_{d_1, \dots, d_k} < 10^k$. For $k = 2$, expression (3) reduces to

$$P(D_1(X) = d_1, D_2(X) = d_2) = \log_{10} \left(1 + \frac{1}{10d_1 + d_2} \right). \quad (4)$$

In addition, the marginal p.f. of $D_1(X)$ is given by

$$P(D_1(X) = d_1) = \log_{10} \left(1 + \frac{1}{d_1} \right), \quad (5)$$

which represents the celebrated result originally obtained by Newcomb (1881) and subsequently rediscovered by Benford (1938), already displayed in Equation (1). The marginal p.f. of $D_2(X)$ is given by

$$P(D_2(X) = d_2) = \sum_{d_1} \log_{10} \left(1 + \frac{1}{10d_1 + d_2} \right). \quad (6)$$

Similarly, the marginal p.f. of each $D_l(X)$, as well as the joint p.f. of any choice of m digits out of the first k , say $D_{l_1}(X), \dots, D_{l_m}(X)$, can be easily obtained on the basis of (3), even if they display cumbersome expressions as m increases. In particular,

$$P(D_{l_1}(X) = d_{l_1}, \dots, D_{l_m}(X) = d_{l_m}) = \sum_{d_{j_1, \dots, j_{k-m}}} \log_{10} \left(1 + \frac{1}{c_{d_1, \dots, d_k}} \right), \quad (7)$$

where $\{j_1, \dots, j_{k-m}\} = \{1, \dots, k\} \setminus \{l_1, \dots, l_m\}$.

The task of anti-fraud analysis is to check whether the observed sequences of digits conform to the NB law. Let Y be the r.v. under scrutiny. In this setting, Y typically represents the transaction value for a specific good and a given trader. Our inferential target thus consists in assessing if Y is NB. According to definition (2), this hypothesis can be written as

$$H_0 : \log_{10} s(Y) \stackrel{L}{=} U, \quad (8)$$

where U denotes a r.v. uniformly distributed on $[0, 1[$. We call (8) the global NB null hypothesis, because it involves the distribution of any k -ple of significant digits, as shown in (3). A goodness-of-fit test of uniformity may be carried out to test H_0 (see, e.g., Marhuenda et al., 2005, for a survey of such testing procedures). Actually, the Mantissa Arc test illustrated by Nigrini (2012) falls in this category.

Rather than on the global NB null (8), anti-fraud applications have mainly focused on the marginal null hypothesis

$$H_0^{(1)} : D_1(Y) \stackrel{\mathcal{L}}{=} D_1(X), \quad (9)$$

where the p.f. (5) is the basis for implementing a suitable test statistic. Nigrini (2012) provides a detailed account of several popular tests of $H_0^{(1)}$. In order to improve the testing procedure, Diekmann (2007) suggests to assess the four marginal null hypotheses

$$H_0^{(l)} : D_l(Y) \stackrel{\mathcal{L}}{=} D_l(X)$$

for $l = 1, \dots, 4$. However, this proposal does not address simultaneity issues, which are instead important in order to avoid an excess of falsely declared anomalies. Alternatively, Nigrini (2012) proposes to test a joint null hypothesis on the first-two significant digits, i.e.

$$H_0^{(1,2)} : (D_1(Y), D_2(Y)) \stackrel{\mathcal{L}}{=} (D_1(X), D_2(X)),$$

based on the joint p.f. (4). Clearly, it does not suffice to assess any $H_0^{(l)}$, $l = 1, \dots, k$, nor $H_0^{(1,2)}$ in order to show that the r.v. Y is NB, i.e. to accept the global null hypothesis H_0 .

In our opinion the implications of the probabilistic framework described in this section are often not adequately considered in anti-fraud testing procedures for the NB law. One major drawback is that the global null hypothesis H_0 is likely to be too general to be interesting in most applications. For example, owing to measurement limitations, the realizations of the r.v. Y are often recorded up to few significant digits. Therefore, the global NB null (8) may be systematically rejected simply because of these rounding errors. On the other hand, a simplified version of H_0 based on the first- k significant digits, for a reasonable choice of $k > 1$, appears to be more appealing. In this case also the marginal null hypotheses on each significant digit (or on suitable choices of $m < k$ significant digits) could be jointly assessed, thus adequately controlling the number of false alarms. A testing procedure which accomplishes this task is considered in detail in the next section. Other advantages of our proposal are both the ability to identify which digits are responsible for departure from the NB law, and the possibility to rely on the exact distribution of the test statistics, by means of (3).

3 Testing the first- k significant digits

3.1 The case of two-digits

We start by illustrating our procedure in the simplest case of the first-two significant digits. Our initial target consists in assessing the joint null hypothesis $H_0^{(1,2)}$. Recall that $H_0^{(1,2)} \Rightarrow H_0^{(1)}$, as well as $H_0^{(1,2)} \Rightarrow H_0^{(2)}$. However, if $H_0^{(1,2)}$ is rejected more insight on this decision may be achieved on the basis of the assessment of the marginal hypotheses $H_0^{(1)}$ and $H_0^{(2)}$. Inference is carried out by assuming a random sample of n copies obtained from Y , say Y_1, \dots, Y_n , giving rise to n copies of $(D_1(Y), D_2(Y))$, which are subsequently arranged as a frequency table given by the matrix $N = (N_{d_1, d_2})$, where

$$N_{d_1, d_2} = \sum_{i=1}^n I_{\{(d_1, d_2)\}}(D_1(Y_i), D_2(Y_i)),$$

$d_1 \in \{1, \dots, 9\}$, $d_2 \in \{0, \dots, 9\}$, and I_C is the indicator function of a given set C . Therefore, N_{d_1, d_2} represents the cardinality of pairs (d_1, d_2) in the digit sample. Obviously, $n = \sum_{d_1, d_2} N_{d_1, d_2}$. The sample space of such matrices, i.e. the matrices of order (9×10) with non-negative integer entries summing up to n , is denoted by Ψ . This space is finite, even if potentially very large, since $\text{Card}(\Psi) = 90^n$.

The two marginal null hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ are assessed by means of the test statistics $T_{\{1\}} = T_{\{1\}}(N)$ and $T_{\{2\}} = T_{\{2\}}(N)$, respectively, where both $T_{\{1\}} : \Psi \rightarrow \mathbb{R}$ and $T_{\{2\}} : \Psi \rightarrow \mathbb{R}$. A popular example is given by the χ^2 test statistics

$$T_{\{1\}} = \sum_{d_1} \frac{(N_1(d_1) - n\pi_1(d_1))^2}{n\pi_1(d_1)}, \quad T_{\{2\}} = \sum_{d_2} \frac{(N_2(d_2) - n\pi_2(d_2))^2}{n\pi_2(d_2)}, \quad (10)$$

where $N_1(d_1) = \sum_{d_2} N_{d_1, d_2}$, $N_2(d_2) = \sum_{d_1} N_{d_1, d_2}$, while $\pi_l(d_l) := P(D_l(X) = d_l)$, for $l = 1, 2$, and X a NB r.v.. Although we mainly base our applications on these χ^2 statistics, any test statistic for conformity could be adopted in their place. For instance, Nigrini (2012) suggests the Mean Absolute Deviation (MAD) as a suitable alternative, leading to

$$T_{\{1\}} = \frac{1}{9} \sum_{d_1} |N_1(d_1) - n\pi_1(d_1)|, \quad T_{\{2\}} = \frac{1}{10} \sum_{d_2} |N_2(d_2) - n\pi_2(d_2)|. \quad (11)$$

In a similar way, the joint null hypothesis $H_0^{(1,2)}$ is assessed by means of the test statistic $T_{\{1,2\}} =$

$T_{\{1,2\}}(N)$, where $T_{1,2} : \Psi \rightarrow \mathbb{R}$. The χ^2 option leads to

$$T_{\{1,2\}} = \sum_{d_1, d_2} \frac{(N_{d_1, d_2} - n\pi_{1,2}(d_1, d_2))^2}{n\pi_{1,2}(d_1, d_2)},$$

with $\pi_{1,2}(d_1, d_2) := P(D_1(X) = d_1, D_2(X) = d_2)$, but other choices are possible as in (11). In what follows, we assume that, whatever test is considered, it rejects the null hypothesis for large values of the test statistic, as it happens for χ^2 and MAD above.

Subsequently, let us consider the random vector $T = T(N) = (T_J)_{J \in \mathcal{J}}$, where $\mathcal{J} = \{\{1\}, \{2\}, \{1, 2\}\}$.

The joint survival function (s.f.) of T is given by

$$S_T(t) = P(T > t) = \frac{1}{\text{Card}(\Psi)} \sum_{M \in \Psi} I_{C_t}(T(M)), \quad (12)$$

where $t = (t_J)_{J \in \mathcal{J}}$ is a vector in \mathbb{R}^3 , $C_t = \bigotimes_{J \in \mathcal{J}}]t_J, \infty[$, the symbol \bigotimes denotes the Cartesian product and vector inequalities are to be interpreted in the lexicographical sense. The marginal s.f.'s corresponding to $T_{\{1\}}$, $T_{\{2\}}$ and $T_{\{1,2\}}$, and to each couple of such r.v.'s, can readily be obtained from (12). In the following, these marginal s.f.'s are indexed by the corresponding r.v.'s, e.g. we write $S_{T_{\{1,2\}}}$, $S_{T_{\{l\}}}$ and $S_{T_{\{l\}}, T_{\{1,2\}}}$, for $l = 1, 2$.

Our first inferential conclusion concerns the assessment of $H_0^{(1,2)}$. This choice is motivated by the requirement to have more stringent control on the rate of false rejections than implied by individual tests of $H_0^{(1)}$ and $H_0^{(2)}$. In fact, it is crucial in our anti-fraud applications that the number of false alarms is kept to the prescribed level, say 1%, over all the different-digit tests. We thus require that the (two-digit) NB hypothesis is rejected only when there is enough evidence against it in the joint distribution of $D_1(Y)$ and $D_2(Y)$. $H_0^{(1,2)}$ is tested by means of $T_{\{1,2\}}$, yielding the P -value $S_{T_{\{1,2\}}}(t_{0\{1,2\}})$, where $t_{0\{1,2\}}$ is the observed value of the test statistic. If $S_{T_{\{1,2\}}}(t_{0\{1,2\}})$ suggests rejection, it is then important to establish how many and which digits are responsible for the decision, since anti-fraud actions may vary according to the answer. Clearly, the required information cannot be provided by $T_{\{1,2\}}$. We thus propose to test $H_0^{(1)}$ and $H_0^{(2)}$ individually after rejection of $H_0^{(1,2)}$, based on the observed values $t_{0\{1\}}$ and $t_{0\{2\}}$ of $T_{\{1\}}$ and $T_{\{2\}}$, respectively. The individual tests should not be based on the marginal distributions of $T_{\{1\}}$ and $T_{\{2\}}$, but on the corresponding conditional distributions given that $T_{\{1,2\}}$ is larger than its observed realization. Hence we require the conditional

P -values

$$\frac{S_{T_{\{1\}}, T_{\{1,2\}}}(t_{0\{1\}}, t_{0\{1,2\}})}{S_{T_{\{1,2\}}}(t_{0\{1,2\}})}, \quad \frac{S_{T_{\{2\}}, T_{\{1,2\}}}(t_{0\{2\}}, t_{0\{1,2\}})}{S_{T_{\{1,2\}}}(t_{0\{1,2\}})} \quad (13)$$

from the joint s.f. (12), as our basis for inference on $H_0^{(1)}$ and $H_0^{(2)}$. An empirical assessment of the difference between the marginal distributions of $T_{\{1\}}$ and $T_{\{2\}}$, and those obtained after conditioning on large values of $T_{\{1,2\}}$, is provided in §4. There, we also compare our conditional approach with some multiple-testing adjustments for the marginal distributions of $T_{\{1\}}$ and $T_{\{2\}}$.

In principle, expression (12) could be computed exactly since Ψ is finite. However, this task is likely to be infeasible owing to the cardinality of Ψ . We thus suggest a Monte Carlo procedure in order to approximate the relevant probabilities in (13). Let $\Pi = (\pi_{1,2}(d_1, d_2))$ be the matrix of probabilities under the NB hypothesis $H_0^{(1,2)}$, with $d_1 \in \{1, \dots, 9\}$ and $d_2 \in \{0, \dots, 9\}$. We first randomly generate a frequency table N with probabilities provided by Π , in such a way that the total of the entries is fixed to n . Write $q = (q_1, \dots, q_{90})^\top = \text{vec}(\Pi)$ and $Z = (Z_1, \dots, Z_{90})^\top = \text{vec}(N)$. Since Z is a Multinomial random vector with parameters n and q , its components can be sequentially generated by means of the well-known properties of its conditional distributions. In fact, if the distribution of the r.v. Z_1 is Binomial with parameters n and q_1 , if the conditional distribution of the r.v. Z_2 given $Z_1 = z_1$ is Binomial with parameters $(n - z_1)$ and $q_2/(1 - q_1)$, and – generally – if the conditional distribution of the r.v. Z_l given $Z_1 = z_1, \dots, Z_{l-1} = z_{l-1}$ is Binomial with parameters $(n - \sum_{j=1}^{l-1} z_j)$ and $q_l/(1 - \sum_{j=1}^{l-1} q_j)$, then Z is a Multinomial random vector with parameters n and q . Note that the generation of one realization of Z with this algorithm requires the generation of 89 Binomial random variates with different parameters values. Therefore, a Binomial generator with a small set-up is suitable for our purpose; see, e.g., Hörmann et al. (2004) and the discussion in Barabesi and Pratelli (2014, 2015). Finally, the frequency table N is easily obtained from its one-to-one correspondence with Z .

We run our Monte Carlo procedure by simulating B frequency tables, say N_1^*, \dots, N_B^* , under the matrix of probabilities Π , and by computing the corresponding vectors $T(N_b^*)$, with $b = 1, \dots, B$. The Monte Carlo counterpart of (12) is then

$$S_T^*(t) = \frac{1}{B} \sum_{b=1}^B I_{C_t}(T(N_b^*)). \quad (14)$$

Obviously, on the basis of the standard Glivenko-Cantelli Theorem, uniform convergence of the Monte Carlo s.f. holds, i.e.

$$\|S_T^*(t) - S_T(t)\|_\infty \xrightarrow{a.s.} 0, \tag{15}$$

as $B \rightarrow \infty$. The Monte Carlo marginal s.f.'s are suitably indexed by the corresponding r.v.'s and can be obtained from (14). Therefore, the Monte Carlo P -value for $H_0^{(1,2)}$ is given by $S_{T_{(1,2)}}^*(t_{0(1,2)})$, while the conditional Monte Carlo P -values for $H_0^{(1)}$ and $H_0^{(2)}$ are

$$\frac{S_{T_{(1)}, T_{(1,2)}}^*(t_{0(1)}, t_{0(1,2)})}{S_{T_{(1,2)}}^*(t_{0(1,2)})}, \quad \frac{S_{T_{(2)}, T_{(1,2)}}^*(t_{0(2)}, t_{0(1,2)})}{S_{T_{(1,2)}}^*(t_{0(1,2)})},$$

respectively. Almost sure convergence of these Monte Carlo estimators to the conditional P -values is ensured by (15) and the continuous mapping theorem.

3.2 The general case

A plethora of null hypotheses could be potentially assessed when the first- k significant digits are considered. Indeed, for each $m = 1, \dots, k$ and for each choice of indexes $\{l_1, \dots, l_m\}$, we could state the $\binom{k}{m}$ null hypotheses

$$H_0^{(l_1, \dots, l_m)} : (D_{l_1}(Y), \dots, D_{l_m}(Y)) \stackrel{\mathcal{L}}{=} (D_{l_1}(X), \dots, D_{l_m}(X)), \tag{16}$$

ranging from the marginal nulls $H_0^{(1)}, \dots, H_0^{(k)}$, to the joint test of $H_0^{(1, \dots, k)}$. The latter may be seen as a practically sensible simplification of the global NB null H_0 , which corresponds to the case where (16) holds for all $m = k \in \mathbb{N}$. A total of $(2^k - 1)$ null hypotheses is thus available for a fixed value of k , even if a small selection of them is likely to be interesting in practice. Since $H_0^{(1, \dots, k)} \Rightarrow H_0^{(l_1, \dots, l_m)}$, we pursue the sequential scheme detailed in §3.1 and test $H_0^{(l_1, \dots, l_m)}$, for $m < k$, only when $H_0^{(1, \dots, k)}$ is rejected. In practice it will often suffice to test the marginal null hypotheses $H_0^{(1)}, \dots, H_0^{(k)}$, after rejection of the joint null $H_0^{(1, \dots, k)}$.

With straightforward extension of the notation in §3.1, our sample is now made of n copies of $(D_1(Y), \dots, D_k(Y))$, which are arranged as a k -way frequency table $N = (N_{d_1, \dots, d_k})$, where $d_1 \in \{1, \dots, 9\}$ and $d_l \in \{0, \dots, 9\}$, for $l = 2, \dots, k$. The sample space of such frequency tables is again denoted by Ψ , while the k -dimensional array of probabilities is $\Pi = (\pi_{1, \dots, k}(d_1, \dots, d_k))$. In

order to assess the null hypothesis (16) we consider the test statistic $T_{\{l_1, \dots, l_m\}} = T_{\{l_1, \dots, l_m\}}(N)$, where $T_{\{l_1, \dots, l_m\}} : \Psi \rightarrow \mathbb{R}$. If we follow the χ^2 option

$$T_{\{l_1, \dots, l_m\}} = \sum_{d_{l_1}, \dots, d_{l_m}} \frac{(N_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m}) - n\pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m}))^2}{n\pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m})},$$

where $\pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m}) := P(D_{l_1}(X) = d_{l_1}, \dots, D_{l_m}(X) = d_{l_m})$, while

$$N_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m}) = \sum_{d_{j_1}, \dots, d_{j_{k-m}}} N_{d_1, \dots, d_k}.$$

As in §3.1, we consider the random vector $T = T(N) = (T_J)_{J \in \mathcal{J}}$, where in this case \mathcal{J} represents the collection of the $(2^k - 1)$ index choices from $\{1, \dots, k\}$. The joint survival function (s.f.) of T has a similar expression as the one given in (12) and the marginal s.f.'s are suitably indexed by the corresponding r.v.'s. Let N_0 be the observed frequency table and $t_0 = T(N_0) = (t_{0J})_{J \in \mathcal{J}}$. We start our assessment from the joint k -digit null $H_0^{(1, \dots, k)}$, for which the P -value is given by $S_{T_{\{1, \dots, k\}}}(t_{0\{1, \dots, k\}})$.

The conditional P -value

$$\frac{S_{T_{\{l_1, \dots, l_m\}}, T_{\{1, \dots, k\}}}(t_{0\{l_1, \dots, l_m\}}, t_{0\{1, \dots, k\}})}{S_{T_{\{1, \dots, k\}}}(t_{0\{1, \dots, k\}})}$$

is then considered in order to assess if rejection of $H_0^{(1, \dots, k)}$ depends on the m -ple of digits l_1, \dots, l_m , with $m < k$. The simplest case $m = 1$ corresponds to test the k marginal hypotheses $H_0^{(1)}, \dots, H_0^{(k)}$, on the basis of their conditional P -values.

Since $S_T(t)$ is not practically computable, the Monte Carlo procedure introduced in §3.1 is generalized to the k -digit setting. Now the order of $q = \text{vec}(II)$ and $Z = \text{vec}(N)$ is 9×10^k , so that even this procedure might become computationally expensive when k is large. Our approach produces a set of simulated frequency tables N_1^*, \dots, N_B^* , from which the Monte Carlo s.f. $S_T^*(t)$ is computed and the relevant conditional P -values are estimated.

We can also exploit our simulation framework to obtain Monte Carlo estimates of the quantiles of the chosen conformity measure under the NB model. For $m = 1, \dots, k$ and $0 < \alpha < 1$, let

$$q_{T_{\{l_1, \dots, l_m\}}}(\alpha) = \inf_{t \in \mathbb{R}} \left\{ \frac{1}{\text{Card}(\Psi)} \sum_{M \in \Psi} I_{[-\infty, t]}(T_{\{l_1, \dots, l_m\}}(M)) \geq \alpha \right\} \quad (17)$$

be the α -quantile of $T_{\{l_1, \dots, l_m\}}$. In our sequential testing procedure we also require the conditional version of this α -quantile, given the event

$$\Psi_{\{1, \dots, k\}}(\beta) = \{M \in \Psi : T_{\{1, \dots, k\}}(M) \geq q_{T_{\{1, \dots, k\}}}(\beta)\}. \quad (18)$$

In (18), $0 < \beta < 1$ and $q_{T_{\{1,\dots,k\}}}(\beta)$ is defined as in (17). For $m = 1, \dots, k-1$, the conditional α -quantile is thus

$$q_{T_{\{l_1, \dots, l_m\}} | \Psi_{\{1, \dots, k\}}(\beta)}(\alpha) = \inf_{t \in \mathbb{R}} \left\{ \frac{1}{\text{Card}(\Psi_{\{1, \dots, k\}}(\beta))} \sum_{M \in \Psi_{\{1, \dots, k\}}(\beta)} I_{1-\infty, t}(T_{\{l_1, \dots, l_m\}}(M)) \geq \alpha \right\}. \quad (19)$$

The Monte Carlo estimators of (17) and (19) turn out to be, respectively,

$$q_{T_{\{l_1, \dots, l_m\}}}^*(\alpha) = \inf_{t \in \mathbb{R}} \left\{ \frac{1}{B} \sum_{b=1}^B I_{1-\infty, t}(T_{\{l_1, \dots, l_m\}}(N_b^*)) \geq \alpha \right\} \quad (20)$$

and

$$q_{T_{\{l_1, \dots, l_m\}} | \Psi_{\{1, \dots, k\}}(\beta)}^*(\alpha) = \inf_{t \in \mathbb{R}} \left\{ \frac{\sum_{b=1}^B I_{1-\infty, t} \times [q_{T_{\{1, \dots, k\}}}^*(\beta), \infty[(T_{\{l_1, \dots, l_m\}}(N_b^*), T_{\{1, \dots, k\}}(N_b^*))}{\sum_{b=1}^B I_{[q_{T_{\{1, \dots, k\}}}^*(\beta), \infty[(T_{\{1, \dots, k\}}(N_b^*))}]} \geq \alpha \right\}. \quad (21)$$

Again, almost sure convergence of these estimators follows by (15) and the continuous mapping theorem.

4 Experimental results for the χ^2 test statistic

Our first experiment aims at providing empirical estimates of the quantiles of the main conformity measures under the NB model, following the Monte Carlo approach described in §3 for a selection of different sample sizes. Extension to other values of n can be easily performed, thanks to the computational efficiency of our algorithm. For instance, even in the case of our largest sample size $n = 500$, full quantile computation with $k = 4$ takes less than 3 hours on a 2.80 GHz Dual Core Processor, using our Matlab routine (which is available upon request). A simple alternative would be interpolation of the estimated quantiles for the closest available values of n . Here we focus on the χ^2 test statistic, while the results for the MAD statistic are reported in the Supplementary Material. Any other suitable measure can be dealt with under the same approach. We take $B = 1,000,000$ in all the simulations that follow, although larger values may be required if β in (18) is chosen to be very close to 1.

We start our experimental analysis by computing estimates (20) and (21) of the quantiles of the χ^2 test statistic, for $\alpha = \beta = 0.99$, $k = 1, 2, 3, 4$ and all possible index selections l_1, \dots, l_m ($m = 1, \dots, k$). Table 1 displays our results and compares them to their χ^2 counterpart, which

provides the standard large-sample approximation to (17). We first note the importance of relying on the exact p.f. (3) if more than one digit is involved. Even in the simple two-digit case, we obtain that replacing the exact 0.99-quantile $q_{T_{\{1,2\}}}(0.99)$ with its large-sample version $\chi_{89}^2(0.99) = 122.94$ leads to increased Type I errors when testing $H_0^{(1,2)}$. Indeed, we find that the error rate is 0.02 with $n = 100$ and becomes as large as 0.048 when $n = 20$. The error rate grows even further when more digits are considered and the number of degrees of freedom increases, in spite of the fact that no parameter has to be estimated.

Whenever $l_1 = 1$, i.e. the first digit is involved in (16), the quantiles of the conditional distribution of $T_{\{l_1, \dots, l_m\}}$ are markedly different from those computed under the unconditional p.f. (7). In such a case the statistics for testing $H_0^{\{l_1, \dots, l_m\}}$ and $H_0^{\{1, \dots, k\}}$ are strongly dependent and neglecting the stochastic outcome of the uppermost comparison may considerably inflate the Type I error rate of the lower-level tests. A quantitative measure of this effect is provided in Table 2, which reports the proportion of false rejections made by the χ^2 goodness-of-fit statistic for testing $H_0^{\{l_1, \dots, l_m\}}$, given previous rejection of $H_0^{\{1, \dots, k\}}$, when $q_{T_{\{l_1, \dots, l_m\}}}^*(0.99)$ is used instead of $q_{T_{\{l_1, \dots, l_m\}} | \Psi_{\{1, \dots, k\}}(0.99)}^*(0.99)$. It is clearly seen that the Type I error rates of tests involving the first digit are always much larger than the nominal value $1 - \alpha = 0.01$, thus providing erroneous evidence in favor of falsification of the first digit. Furthermore, the empirical size of the test of the marginal null hypothesis (9) does not appear to depend crucially on the chosen value of k . We conclude that the naive approach which ignores the sequentiality of tests when scrutinizing the first-digit distribution can lead to a plethora of false signals and can thus have very harmful consequences for anti-fraud purposes. On the other hand, the NB law implies that the distribution of $D_l(X)$ approaches the Uniform as l increases. Our results show that the effect of this convergence on test statistics is very fast and already appreciable when $l_1 = 2$. In some sense, it is the first digit which “dominates” the decision on the joint null $H_0^{\{1, \dots, k\}}$. Nevertheless, marginally inspecting the subsequent digits might still be useful, since there is no guarantee that fraudsters will fabricate their data following the Uniform distribution.

We conclude this section by comparing our approach with the results obtained through some popular multiple-testing adjustments for the critical values of the test statistics $T_{\{l_1, \dots, l_m\}}$. These methods may be used when the hypotheses under scrutiny are not exchangeable and must be tested in a specified order, as in our context. In particular, we consider a sequential multiple-testing

procedure which is known as serial gatekeeping (see, e.g., Goeman and Solari, 2010, §6.1). It amounts to testing all the possible hypotheses (16), starting from the highest-level null $H_0^{(1,\dots,k)}$ and moving to the set of hypotheses $H_0^{(l_1,\dots,l_{m-1})}$ only when all the nulls $H_0^{(l_1,\dots,l_m)}$ have been rejected. At every step the required tests are based on the corresponding unconditional survival functions obtained from $S_T(t)$, with appropriate adjustments for multiplicity. We also implement the basic Bonferroni adjustment $(1-\alpha)/k$ when testing the marginal hypotheses $H_0^{(1)}, \dots, H_0^{(k)}$ at level $(1-\alpha)$. This adjustment also turns out to be the appropriate correction in a single step of a tree-structured testing procedure (Goeman and Solari, 2010, §6.3).

For simplicity, we restrict our experimental analysis to the simplest two-digit case, but similar findings have been obtained also when $k > 2$. We take $\alpha = 0.99$, as before, and we run a new set of $B = 1,000,000$ independent simulations, on which the empirical test sizes for the χ^2 test of $H_0^{(1)}$ and $H_0^{(2)}$ are computed. These test sizes are obtained by conditioning on rejection of $H_0^{(1,2)}$, based on the estimated quantile $q_{T_{(1,2)}}^*(0.99)$. The results are displayed in Table 3. It is apparent that multiple-testing adjustments are not able to provide proper control of Type I error rates of the one-digit hypotheses given that the joint null $H_0^{(1,2)}$ has been rejected, even if our Monte Carlo estimate of the exact s.f. (12) is adopted. As expected, the effect is larger for the first digit and when n is small, but it is still present in all the reported instances. This behavior is not a fault of multiple-testing procedures, which are constructed to control the familywise error rate over all the hypotheses under testing, but which are not designed to obtain the same performance in the tail of the distribution of $T_{(1,2)}$. We conclude that exploiting the hierarchical nature of successive tests of the NB law, through our conditional quantiles (19) and our conditional P -values (13), is to be recommended when the goal is to control the proportion of false detections at every step of the sequential testing procedure. This is precisely the framework of the anti-fraud applications of §6, where rejection of $H_0^{(1)}$ is likely to have different consequences from rejection of $H_0^{(l)}$, for $l = 2, \dots, k$, given that the NB law has been judged to be invalid by the test of the joint hypothesis $H_0^{(1,\dots,k)}$.

5 The significant digit distribution under the alternative hypothesis

We extend the distributional results for the k -digit p.f. to the case where the NB law does not hold. Our simple and neat expressions can be useful both to implement simulation studies under the alternative, in order to compare the power of different procedures, and to assess the discrepancy of a given r.v. Y from the NB r.v. X . Actually, some results on the global discrepancy – i.e. in terms of the Kolmogorov and Kuiper distances for the r.v.'s $s(Y)$ and $s(X)$ – are given by Dümbgen and Leuenberger (2008), but they do not involve the marginal discrepancies. Instead, through our results it is possible to construct a suitable and simple distance for the r.v.'s $D_1(Y)$ and $D_1(X)$.

Let F_Y denote the distribution function (d.f.) of the r.v. Y . Hence the d.f. of $|Y|$ is given by $F_{|Y|}(x) = F_Y(x) - F_Y(-x) + P(Y = x)$, for $x \in \mathbb{R}^+$. We first give a result which provides the joint distribution of $D_1(Y), \dots, D_k(Y)$.

Proposition 1 *For all $k \in \mathbb{N}$ it holds that*

$$\begin{aligned} P(D_1(Y) = d_1, \dots, D_k(Y) = d_k) &= \\ &= \sum_{m \in \mathbb{Z}} (F_{|Y|}(10^{m-k+1}(c_{d_1, \dots, d_k} + 1)) - F_{|Y|}(10^{m-k+1}c_{d_1, \dots, d_k})), \end{aligned}$$

where $d_1 \in \{1, \dots, 9\}$ and $d_l \in \{0, \dots, 9\}$ for $l = 2, \dots, k$.

Proof. See the Appendix. □

It follows from Proposition 1 that

$$\begin{aligned} P(D_1(Y) = d_1, D_2(Y) = d_2) &= \\ &= \sum_{m \in \mathbb{Z}} (F_{|Y|}(10^{m-1}(10d_1 + d_2 + 1)) - F_{|Y|}(10^{m-1}(10d_1 + d_2))), \end{aligned}$$

while

$$P(D_1(Y) = d_1) = \sum_{m \in \mathbb{Z}} (F_{|Y|}(10^m(d_1 + 1)) - F_{|Y|}(10^m d_1))$$

and

$$P(D_2(Y) = d_2) = \sum_{d_1} \sum_{m \in \mathbb{Z}} (F_{|Y|}(10^{m-1}(10d_1 + d_2 + 1)) - F_{|Y|}(10^{m-1}(10d_1 + d_2))).$$

Similar, even if more cumbersome, expressions can be obtained in general. Indeed, for each $m = 1, \dots, k$ and for each choice of indexes $\{l_1, \dots, l_m\}$, we have

$$\begin{aligned} P(D_{l_1}(Y) = d_{l_1}, \dots, D_{l_m}(Y) = d_{l_m}) &= \\ &= \sum_{d_{j_1}, \dots, d_{j_{k-m}}} \sum_{m \in \mathbb{Z}} (F_{|Y|}(10^{m-k+1}(c_{d_1, \dots, d_k} + 1)) - F_{|Y|}(10^{m-k+1}c_{d_1, \dots, d_k})), \end{aligned}$$

where $\{j_1, \dots, j_{k-m}\}$ is defined as in §3.2, while $d_l \in \{1, \dots, 9\}$ if $l = 1$ and $d_l \in \{0, \dots, 9\}$ otherwise.

Leemis et al. (2000) and Fewster (2009) considered a χ^2 -type index for measuring the distance between the distributions of $D_1(X)$ and $D_1(Y)$, i.e.

$$\rho_{\{1\}}^2 = \sum_{d_1=1}^9 \frac{(P(D_1(Y) = d_1) - \pi_1(d_1))^2}{\pi_1(d_1)}.$$

This distance may be generalized to each choice of indexes, yielding

$$\rho_{\{l_1, \dots, l_m\}}^2 = \sum_{d_{l_1}, \dots, d_{l_m}} \frac{(P(D_{l_1}(Y) = d_{l_1}, \dots, D_{l_m}(Y) = d_{l_m}) - \pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m}))^2}{\pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m})}.$$

Let $TV(g)$ be the total variation of a given function g , i.e.

$$TV(g) = \sup \left\{ \sum_{i=1}^K |g(a_{i+1}) - g(a_i)| : (a_1, \dots, a_{K+1}) \in \mathcal{P} \right\},$$

where \mathcal{P} is the collection of ordered $(K + 1)$ -ples (a_1, \dots, a_{K+1}) such that $a_i \in I \subset \mathbb{R}$ for each interval I , and $K \in \mathbb{N}$ is arbitrary. It follows that $TV(g) = \int |g'(x)|dx$ when g is differentiable a.e., while $TV(g) = 2 \max g(x)$ when g is increasing and then decreasing. The following result provides a bound on $\rho_{\{l_1, \dots, l_m\}}^2$, by assuming that the r.v. Y is absolutely continuous with probability density function (p.d.f.) f_Y with respect to the Lebesgue measure. In such a case, let $f_{\log_{10}|Y|}$ be the p.d.f. of the r.v. $\log_{10}|Y|$, i.e.

$$f_{\log_{10}|Y|}(x) = (f_Y(10^x) + f_Y(-10^x))10^x \ln 10.$$

Proposition 2 *Let the r.v. Y defined on the probability space (Ω, \mathcal{F}, P) be absolutely continuous. If $f_Y \in C^p(\mathbb{R})$, then for each $\{l_1, \dots, l_m\} \subset \{1, \dots, k\}$*

$$\rho_{\{l_1, \dots, l_m\}}^2 \leq \frac{\gamma_{l_1, \dots, l_m} TV(f_{\log_{10}|Y|}^{(p)})^2}{4 \cdot 6^{2p}},$$

where

$$\gamma_{l_1, \dots, l_m} = \sum_{d_1, \dots, d_{l_m}} \frac{(\sum_{d_{j_1}, \dots, d_{j_{k-m}}} \pi_{1, \dots, k}(d_1, \dots, d_k)(1 - \pi_{1, \dots, k}(d_1, \dots, d_k)))^2}{\pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m})}.$$

Proof. See the Appendix. □

Note that γ_{l_1, \dots, l_m} is a fixed constant which may be easily computed for the selected index choice. Therefore, the tightness of the inequality in Proposition 2 crucially depends on $TV(f_{\log_{10}|Y|}^{(p)})$, i.e. loosely speaking on the regularity of $f_{\log_{10}|Y|}$. Hence, similarly to the case of the global discrepancy measure, small values of $\rho_{\{l_1, \dots, l_m\}}^2$ are expected, e.g., if Y is a Log-Normal or a Weibull r.v.; see Dümbgen and Leuenberger (2008) for the evaluation of $TV(f_{\log_{10}|Y|}^{(p)})$ in such settings.

For the special case $\rho_{\{1\}}^2$, we have $\max \rho_{\{1\}}^2 = \log_{10} 9 / (1 - \log_{10} 9) \simeq 20.85$ when $P(D_1(Y) = 9) = 1$. Moreover, $\gamma_1 = \sum_{d_1=1}^9 \pi_1(d_1)(1 - \pi_1(d_1))^2 \simeq 0.7059$ and Proposition 2 shows that

$$\rho_{\{1\}}^2 \leq \frac{\gamma_1 TV(f_{\log_{10}|Y|}^{(p)})^2}{4 \cdot 6^{2p}} \simeq 0.1765 \frac{TV(f_{\log_{10}|Y|}^{(p)})^2}{6^{2p}}.$$

As an example, if the r.v. Y is distributed according a Log-Normal law with scale parameter σ , it holds that $TV(f_{\log_{10}|Y|}^{(p)}) \leq \sqrt{(p+1)!} / (\sigma / \ln 10)^{p+1}$ and hence

$$\rho_{\{1\}}^2 \leq \frac{9\gamma_1(p+1)!}{(36\sigma^2 / \ln^2 10)^{p+1}}.$$

The bound is minimal when $p+1 = \lfloor 36\sigma^2 / \ln^2 10 \rfloor$. As an example, for $\sigma = 1$ it holds that $\rho_{\{1\}}^2 \leq 0.0245$, while for $\sigma = 2$ it holds that $\rho_{\{1\}}^2 \leq 3.9 \times 10^{-11}$, i.e. $\rho_{\{1\}}^2$ decreases exponentially as $\sigma \rightarrow \infty$, similarly to the results in Dümbgen and Leuenberger (2008).

6 Application to international trade data

The EU legislation regulates the collection of customs declarations by the EU Member States authorities and imposes a common customs form for international trade, which is called Single Administrative Document (SAD). Traders use the SAD to declare their trade operations. SAD data are thus used by Customs authorities to monitor all goods arriving from third countries into the EU (imports), those exported outside the EU (exports) and the movement of non-EU goods within the

EU (transits). The collected SAD data contain detailed information on the goods commodity code, the movement of the goods, the customs procedure code that specifies how the Customs authority treats the entry, and of course the traded quantities and values. Many customs duties and the Value Added Tax (VAT) are calculated as a percentage of the declared values. Therefore, misdeclaring the value almost certainly implies fraud. For example, undervaluation is usually attempted to pay less duties or excises, or to evade import restrictions and certain anti-dumping measures. On the overvaluation side, there are money laundering schemes, attempts to obtain higher export refunds or duty compensations, to avoid anti-dumping duties and even to evade internal taxes (FATF-OECD, 2008, 2013).

Our benchmark data set is formed by about 7.5 million SAD import records collected in 2011 by the Italian Customs. Transactions involved more than 200,000 traders. In order to restrict the scope of our test to a set of traders of major operational interest for the Customs auditors, we have considered only traders with more than 10 SAD declarations. This reduced the focus to around 50,000 traders, which made on average 134 import transactions. In this subset of traders, the most interesting are those who operated on a certain variety of product types. We have thus additionally restricted the focus to importers of more than 10 different products, whose trade operations should comply with the assumptions of the NB law under the assumption of fair trade. The average number of traded products for this set of importers is 30. In the following, we illustrate our NB tests on SAD import values obtained from two representatives of these traders, labelled for obvious confidentiality reasons as Trader 1 and Trader 2, respectively.

The marginal distributions of the first-four digits for Trader 1 are presented in Figure 1. They refer to the traded values of $n = 100$ transactions, ranging from approximately 38 to 131,213 euros. Consideration of the first digit would suggest non-conformity to the NB law. The same conclusion is reached by comparing the observed χ^2 test statistic, $T_1 = 28.71$, to the unconditional quantile given in Table 1 for the corresponding sample size. Instead, very different findings are obtained from our hierarchical testing procedure, for which a selection of steps are reported in Table 4. It is clearly seen that no suitable simplification of the global NB hypothesis can be rejected at the selected nominal size. This result holds for the joint four-digit null $H_0^{(1,2,3,4)}$ and also for all possible index selections, up to the marginal hypotheses $H_0^{(1)}, \dots, H_0^{(4)}$. We thus conclude that rejection of

$H_0^{(1)}$ based on the marginal unconditional distribution of $T_{\{1\}}$ provides only very weak evidence of data fabrication and might be seen as an instance of false discovery, perhaps due to the somewhat limited range of traded values for this trader. Indeed, further investigation on Trader 1 reveals that its transactions are well in line with the market, from the point of view of both traded quantities and prices, thus supporting the idea of non-fraudulent behavior.

The picture that we present in our second example is somewhat different. Figure 2 displays the marginal distributions of the first-four digits in the traded values for Trader 2, a company with $n = 103$ transactions ranging from 5.49 to 231,963 euros. Again, both visual inspection and goodness-of-fit testing on the marginal distributions (Table 5) would suggest non-conformity in the first digit, using the quantiles obtained for the case $n = 100$ as a slight approximation to the required values. In this example, however, the evidence against the NB law is considerably stronger than for Trader 1 and leads to rejection of the joint null $H_0^{(1,2,3,4)}$. One advantage of our approach is that we can safely conclude, at the specified nominal level $1 - \alpha = 0.01$, both that the 4-digit NB hypothesis (16) is not likely to hold and that rejection of this hypothesis must be attributed to non-compliance in the first digit, which is the most harmful deviation from the point of view of tax evasion. Further inspection shows that indeed a few of the traded values and prices for Trader 2 might have suspicious features. Our analysis might thus open the door to more detailed substantive investigation.

7 Concluding remarks

Testing the Newcomb-Benford law for the first significant digits is often seen as a useful instrument for detecting frauds in financial data. However, most of the applications in this area rely either on diagnostic checks of the data, or on informal decision rules. Formal goodness-of-fit testing of the law poses some challenging statistical problems that include both the choice of the most appropriate version of the null hypothesis, and derivation of the exact distribution of the test statistic. Non-trivial solutions to these issues are required in order to satisfy a crucial requirement for many anti-fraud exercises, i.e. to ensure adequate control over the number of falsely discovered anomalies.

In this work we propose a hierarchical testing procedure where different varying-dimensional marginals of the Newcomb-Benford distribution are tested in sequence, starting from the k -digit distribution and possibly ending with the one-digit marginals. By explicitly taking into account the hierarchical nature of the sequence of tests that we implement, we base our inferences on the resulting conditional distributions and reach proper control of the rate of false rejections at each stage of the analysis. Furthermore, our testing procedure leads to exact significance levels and does not rely on possibly misleading large-sample approximations.

Although our approach is general, it turns out to be particularly attractive for the detection of frauds in customs data collected from international trade. We have seen two instances where our methodology could help to discriminate between false signals and potential frauds requiring further substantive investigation. This information is not provided at the same accuracy level by other existing techniques.

We emphasize that, for investigation purposes, Customs authorities must concentrate their human resources on a restricted number of potential signals of fraud. Therefore, a test of compliance to the NB law can become a practical working tool only if the number of false alarms is carefully controlled at a prescribed (low) level. We have used some supplementary information to corroborate our findings in the two reported empirical examples, where the true behavior of traders was unknown. However, this supplementary information requires complex data queries and is not available in routine analysis on thousands of traders, for which signals can be obtained in an automatic way only through a formal testing procedure. This is the reason why we see the availability of an accurate test of the NB law, like the one that we propose, as a pre-requisite for concrete application of the NB methodology in anti-fraud analysis.

We have not investigated the power of the suggested testing procedure, which is the subject of ongoing research. Nevertheless, we can anticipate that our strict control of the false alarm rate will considerably affect power only when the degree of deviation from the NB law is small or moderate, an instance of minor relevance for anti-fraud purposes. We instead expect to detect “serial” fraudsters with a probability which is almost as high as the one ensured by standard methods that evaluate each digit separately, without allowing for multiplicity of tests. A related issue concerns the conditions under which the NB law can be expected to provide a reliable approximation to

the digit distribution of regular non-fraudulent transactions in the specific context of international trade. Our preliminary simulation results seem to point to wide applicability conditions. Again, this is the subject of ongoing research and details will be reported elsewhere.

Appendix

Proof of Proposition 1

By the definition of s , the following equivalence of events holds

$$\{s(Y) \leq t\} = \{|Y| \in \bigcup_{m \in \mathbb{Z}}]10^m, 10^{m+1}t]\} \quad (22)$$

for $t \in [1, 10[$. Hence, since the intervals in (22) are disjoint, we have

$$P(s(Y) \leq t) = \sum_{m \in \mathbb{Z}} (F_{|Y|}(10^{m+1}t) - F_{|Y|}(10^m))$$

for $t \in [1, 10[$. In addition, since

$$s(x) = \sum_{j \in \mathbb{N}} 10^{-j+1} D_j(x),$$

the following equivalence holds

$$\{D_1(Y) = d_1, \dots, D_k(Y) = d_k\} = \{s(Y) \in]10^{-k+1}c_{d_1, \dots, d_k}, 10^{-k+1}(c_{d_1, \dots, d_k} + 1)]\}.$$

Therefore,

$$\begin{aligned} P(D_1(Y) = d_1, \dots, D_k(Y) = d_k) &= \\ &= P(s(Y) \leq 10^{-k+1}(c_{d_1, \dots, d_k} + 1)) - P(s(Y) \leq 10^{-k+1}c_{d_1, \dots, d_k}), \end{aligned}$$

which concludes the proof.

Proof of Proposition 2

Proof. First, note that

$$\begin{aligned} & |P(D_{l_1}(Y) = d_{l_1}, \dots, D_{l_m}(Y) = d_{l_m}) - \pi_{l_1, \dots, l_m}(d_{l_1}, \dots, d_{l_m})| = \\ & = | \sum_{d_{j_1}, \dots, d_{j_{k-m}}} (P(D_1(Y) = d_1, \dots, D_k(Y) = d_k) - \pi_{1, \dots, k}(d_1, \dots, d_k)) | \\ & \leq \sum_{d_{j_1}, \dots, d_{j_{k-m}}} |P(D_1(Y) = d_1, \dots, D_k(Y) = d_k) - \pi_{1, \dots, k}(d_1, \dots, d_k)|. \end{aligned}$$

By Corollary 4 by Dümbgen and Leuenberger (2008), the following inequality holds for $u < v$

$$|P(s(Y) \leq 10^v) - P(s(Y) \leq 10^u) - (v - u)| \leq \frac{TV(f_{\log_{10}|Y|}^{(p)})}{2 \cdot 6^p} (v - u)(1 - (v - u)).$$

By choosing $v = -k + 1 + \log_{10}(c_{d_1, \dots, d_k} + 1)$ and $u = -k + 1 + \log_{10} c_{d_1, \dots, d_k}$, we find (see the proof of Proposition 1)

$$P(s(Y) \leq 10^v) - P(s(Y) \leq 10^u) = P(D_1(Y) = d_1, \dots, D_k(Y) = d_k)$$

and (see expression (3))

$$v - u = \pi_{1, \dots, k}(d_1, \dots, d_k).$$

Therefore, the previous inequality provides

$$\begin{aligned} & |P(D_1(Y) = d_1, \dots, D_k(Y) = d_k) - \pi_{1, \dots, k}(d_1, \dots, d_k)| \leq \\ & \leq \frac{TV(f_{\log_{10}|Y|}^{(p)})}{2 \cdot 6^p} \pi_{1, \dots, k}(d_1, \dots, d_k)(1 - \pi_{1, \dots, k}(d_1, \dots, d_k)). \end{aligned}$$

The result then follows from the definition of $\rho_{\{l_1, \dots, l_m\}}^2$.

SUPPLEMENTARY MATERIAL

In the Supplementary Material we provide experimental results for the MAD conformity measure that complement those for the χ^2 goodness-of fit statistic given in §4.

References

- Barabesi, L. and Pratelli, L. (2014). A note on a universal random variate generator for integer-valued random variables. *Statistics and Computing* **24**, 589–596.
- Barabesi, L. and Pratelli, L. (2015). Universal methods for generating random variables with a given characteristic function. *Journal of Statistical Computation and Simulation* **85**, 1679–1691.
- Barabesi, L., Cerasa, A., Perrotta, D. and Cerioli A. (2016). Modelling international trade data with the Tweedie distribution for anti-fraud and policy support. *European Journal of Operational Research* **248**, 1031–1043.
- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society* **78**, 551–572.
- Berger, A. and Hill, T.P. (2011a). A basic theory of Benford’s law. *Probability Surveys* **8**, 1–126.
- Berger, A. and Hill, T.P. (2011b). Benford’s law strikes back: no simple explanation in sight for mathematical gem. *Mathematical Intelligencer* **33**, 85–91.
- Berger, A. and Hill, T.P. (2015). *An Introduction to Benford’s Law*. Princeton University Press, Princeton.
- Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association* **105**, 147–156.
- Cerioli, A. and Perrotta, D. (2014). Robust clustering around regression lines with high density regions. *Advances in Data Analysis and Classification* **8**, 5–26.
- Deng, X., Joseph, V.R. and Wu, C.F.J. (2009). Active learning through sequential design, with applications to the detection of money laundering. *Journal of the American Statistical Association* **104**, 969–981.
- Diaconis, P. (1977). The distribution of leading digits and uniform distribution mod 1. *Annals of Probability* **5**, 72–81.

- Diekmann, A. (2007). Not the first digit! Using Benford's law to detect fraudulent scientific data. *Journal of Applied Statistics* **34**, 321–329.
- Dmitrienko, A. and Tamhane, A.C. (2013). General theory of mixture procedures for gatekeeping. *Biometrical Journal* **55**, 402–419.
- Dümbgen, L. and Leuenberger, C. (2008). Explicit bounds for the approximate error in Benford's law. *Electronic Communications in Probability* **13**, 99–112.
- FATF-OECD, Financial Action Task Force (2008). Best Practices on Trade Based Money Laundering. Available through <http://www.fatf-gafi.org/>.
- FATF-OECD, Financial Action Task Force (2013). Money laundering and terrorist financing through trade in diamonds. Available through <http://www.fatf-gafi.org/>.
- Fewster, R.M. (2009). A simple explanation of Benford's law. *The American Statistician* **63**, 26–32.
- Finos, L. and Farcomeni, A. (2011). k -FWER control without p -value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics* **67**, 174–181.
- Goeman, J.J and Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* **38**, 3782–3810.
- Havil, J. (2008). *Impossible? Surprising Solutions to Counterintuitive Conundrums*. Princeton University Press, Princeton.
- Hill, T.P. (1995). A statistical derivation of the significant-digit law. *Statistical Science* **10**, 354–363.
- Hörmann, W., Leydold, J. and Derflinger, G. (2004). *Automatic Nonuniform Random Variate Generation*. Springer, Berlin.
- Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science* **23**, 92–119.

- Knuth, D.E. (1997). *The Art of Computer Programming, Seminumerical Algorithms, vol. 2, 3rd ed.*. Addison-Wesley, Reading.
- Leemis, L.M., Schmeiser B.W. and Evans, D.L. (2000). Survival distributions satisfying Benford's law. *The American Statistician* **54**, 236–241.
- Marhuenda, Y., Morales, D. and Pardo, M.C. (2005). A comparison of uniformity tests. *Statistics* **39**, 315–328.
- Miller, S.J. (Editor) (2015). *Benford's Law: Theory and Applications*. Princeton University Press, Princeton.
- Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* **4**, 39–40.
- Nigrini, M.J. (2012). *Benford's Law*. Wiley, Hoboken.
- Pericchi, L. and Torres, D. (2011). Quick anomaly detection by the Newcomb-Benford law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. *Statistical Science* **26**, 502–516.
- Pickover, C. (2009). *The Math Book*. Sterling Publishing, New York.
- Riani, M., Atkinson, A.C. and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B* **71**, 447–466.
- Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D. and Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics* **52**, 5–19.
- Tam Cho, W.K. and Gaines, B.J. (2007). Breaking the (Benford) law. *The American Statistician* **61**, 218–223.

Table 1: Monte Carlo estimates $q_{T_{\{l_1, \dots, l_m\}}}^*(0.99)$ (left) and $q_{T_{\{l_1, \dots, l_m\}}|\Psi_{\{1, \dots, k\}}(0.99)}^*(0.99)$ (right), as given in Equations (20) and (21), of the quantiles of the χ^2 goodness-of-fit test statistics under the NB model, for $k = 1, 2, 3, 4$ and all possible index selections l_1, \dots, l_m . $B = 1,000,000$ independent replications for each sample size. The last column gives the large-sample χ^2 approximation to $q_{T_{\{l_1, \dots, l_m\}}}(0.99)$.

k	$\{l_1, \dots, l_m\}$	$n = 20$		$n = 100$		$n = 500$		$\chi^2(0.99)$
		(20)	(21)	(20)	(21)	(20)	(21)	
2	{1}	21.52; 37.27		20.44; 31.09		20.13; 28.93		20.09
	{2}	21.73; 29.04		21.70; 29.23		21.68; 29.69		21.67
	{1, 2}	143.82; –		128.82; –		124.34; –		122.94
3	{1}	21.52; 37.27		20.44; 30.40		20.13; 26.17		20.09
	{2}	21.73; 23.78		21.70; 23.68		21.68; 23.99		21.67
	{3}	21.73; 23.39		21.62; 23.42		21.63; 23.92		21.67
	{1, 2}	143.82; 198.28		128.82; 158.52		124.34; 142.26		122.94
	{1, 3}	143.92; 197.55		128.70; 159.31		124.11; 141.11		122.94
	{2, 3}	140.87; 158.87		136.06; 148.88		134.90; 147.88		134.64
	{1, 2, 3}	1249.76; –		1083.71; –		1022.62; –		1000.57
4	{1}	21.52; 37.61		20.44; 31.84		20.13; 28.17		20.09
	{2}	21.73; 22.61		21.70; 22.73		21.68; 22.29		21.67
	{3}	21.73; 22.18		21.62; 22.22		21.63; 22.15		21.67
	{4}	21.98; 21.98		21.66; 21.66		21.66; 21.89		21.67
	{1, 2}	143.82; 196.67		128.82; 155.82		124.34; 139.77		122.94
	{1, 3}	143.92; 194.52		128.70; 155.97		124.11; 138.55		122.94
	{1, 4}	143.79; 196.34		128.65; 155.48		124.16; 138.94		122.94
	{2, 3}	140.87; 145.03		136.06; 141.32		134.90; 138.42		134.64
	{2, 4}	140.60; 146.14		136.12; 139.42		134.95; 138.37		134.64
	{3, 4}	140.17; 140.37		135.99; 137.51		134.88; 136.78		134.64
	{1, 2, 3}	1249.76; 1537.64		1083.71; 1223.95		1022.61; 1098.15		1000.58
	{1, 2, 4}	1249.41; 1541.06		1083.60; 1224.99		1022.83; 1096.64		1000.58
	{1, 3, 4}	1250.48; 1550.01		1082.92; 1217.56		1022.58; 1093.62		1000.58
	{2, 3, 4}	1162.62; 1219.30		1123.10; 1150.01		1109.28; 1134.71		1105.92
	{1, 2, 3, 4}	12164.36; –		10422.04; –		9698.41; –		9314.03

Table 2: Monte Carlo estimates of the proportion of false rejections in testing $H_0^{(l_1, \dots, l_m)}$ through the χ^2 goodness-of-fit statistic when $q_{T_{(l_1, \dots, l_m)}}^*(0.99)$ is used instead of $q_{T_{(l_1, \dots, l_m)} | \Psi_{(1, \dots, k)}}^*(0.99)$, for $k = 1, 2, 3, 4$ and all possible index selections l_1, \dots, l_m ($m = 1, \dots, k - 1$). For each sample size, $B = 1,000,000$ independent replications are run under the NB model and only those in which $H_0^{(1, \dots, k)}$ is rejected at level 0.01 are retained.

k	$\{l_1, \dots, l_m\}$	$n = 20$	$n = 100$	$n = 500$
2	{1}	0.201	0.121	0.085
	{2}	0.056	0.067	0.076
3	{1}	0.200	0.112	0.052
	{2}	0.017	0.019	0.021
	{3}	0.017	0.016	0.020
	{1, 2}	0.268	0.169	0.098
	{1, 3}	0.266	0.171	0.099
	{2, 3}	0.041	0.047	0.055
4	{1}	0.242	0.158	0.098
	{2}	0.014	0.014	0.012
	{3}	0.012	0.012	0.011
	{4}	0.010	0.010	0.011
	{1, 2}	0.235	0.143	0.071
	{1, 3}	0.235	0.135	0.068
	{1, 4}	0.233	0.134	0.070
	{2, 3}	0.014	0.016	0.016
	{2, 4}	0.016	0.014	0.016
	{3, 4}	0.011	0.014	0.013
	{1, 2, 3}	0.468	0.357	0.186
	{1, 2, 4}	0.473	0.351	0.185
	{1, 3, 4}	0.464	0.341	0.188
	{2, 3, 4}	0.025	0.027	0.031

Table 3: Monte Carlo estimates of the empirical sizes of our test using the conditional quantile (21), of the serial gatekeeping (Gate) and the Bonferroni-adjusted (Bonf) tests of $T_{(1)}$ and $T_{(2)}$. The χ^2 statistic is adopted and the nominal test size is 0.01. The empirical sizes are computed by conditioning on rejection of $H_0^{(1,2)}$, based on the estimated quantile $q_{T_{(1,2)}}^*(0.99)$. $B = 1,000,000$ independent replications are taken for each value of n .

	$n = 20$			$n = 100$			$n = 500$		
	(21)	Gate	Bonf	(21)	Gate	Bonf	(21)	Gate	Bonf
$H_0^{(1)}$	0.010	0.127	0.126	0.010	0.086	0.085	0.011	0.057	0.055
$H_0^{(2)}$	0.009	0.034	0.032	0.009	0.041	0.040	0.010	0.049	0.048

Table 4: Trader 1: Chi-squared statistics $T_{\{l_1, \dots, l_m\}}$ for testing (16), the corresponding unconditional quantiles (20) and conditional quantiles (21) from Table 1, with $\alpha = \beta = 0.99$. A selection of null hypotheses is displayed for the case $k = 4$. The instances where the test would lead to rejection are reported in bold.

$\{l_1, \dots, l_m\}$	$T_{\{l_1, \dots, l_m\}}$	(20)	(21)
{1,2,3,4}	8010.37	10422.04	10422.04
{1,2,3}	837.68	1083.71	1223.95
{1,2}	89.50	128.82	155.81
{1}	28.71	20.44	31.84
{2}	11.54	21.70	22.73
{3}	12.29	21.62	22.22
{4}	7.41	21.66	21.66

Table 5: Trader 2: Outcome of the hierarchical chi-squared testing procedure as in Table 4. The approximated sample size $n = 100$ is used for quantile computation.

$\{l_1, \dots, l_m\}$	$T_{\{l_1, \dots, l_m\}}$	(20)	(21)
{1,2,3,4}	11404.48	10422.04	10422.04
{1,2,3}	1184.40	1083.71	1223.95
{1,2}	149.98	128.82	155.81
{1}	37.91	20.44	31.84
{2}	7.19	21.70	22.73
{3}	14.77	21.62	22.22
{4}	12.86	21.66	21.66

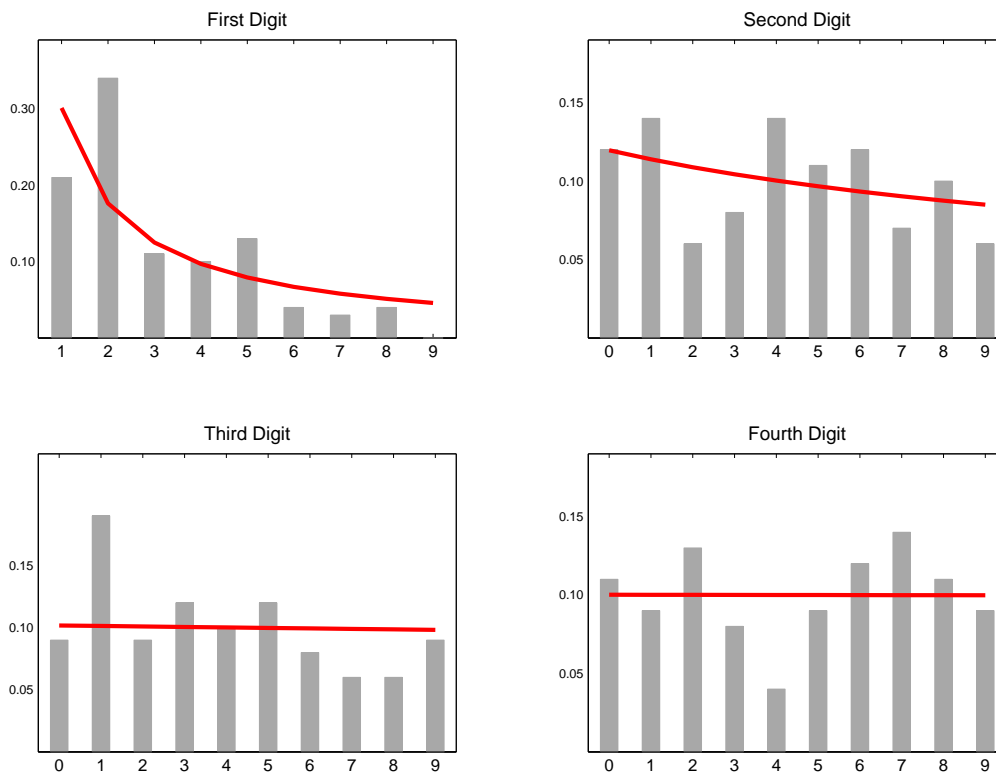


Figure 1: Trader 1: Empirical distribution of the first-four digits in the traded values and theoretical proportions (solid line) under the NB law.

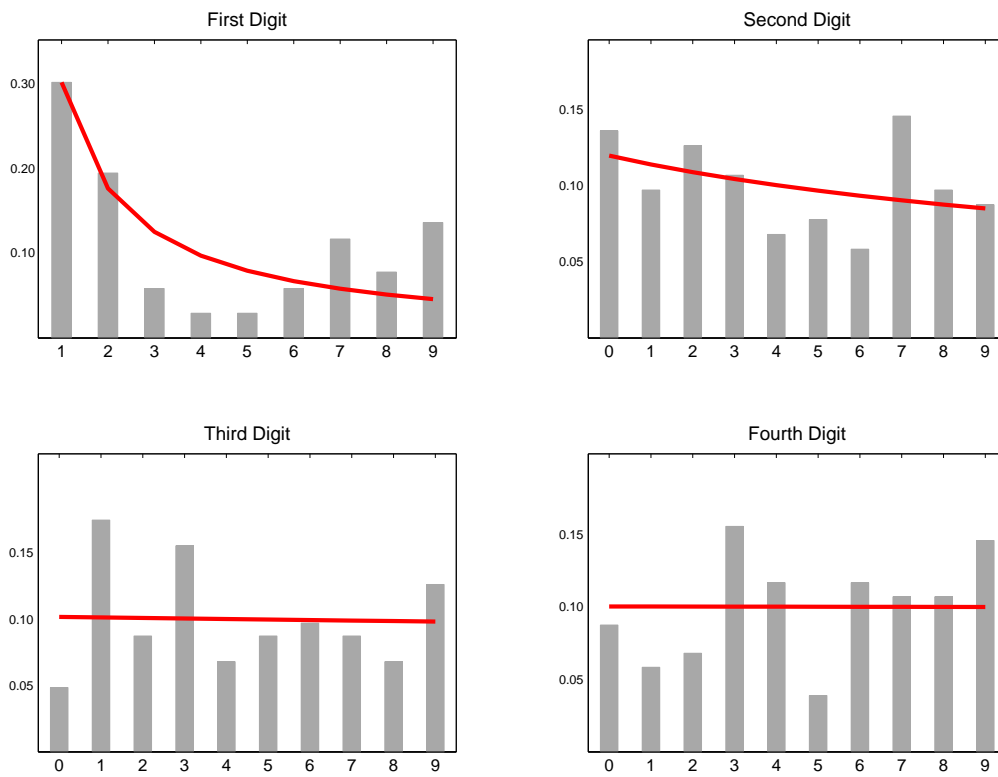


Figure 2: Trader 2: Empirical distribution of the first-four digits in the traded values and theoretical proportions (solid line) under the NB law.