Supplementary Material to the paper: On characterizations and tests of Benford's Law

Lucio Barabesi
University of Siena, Department of Economics and Statistics, Siena, Italy
and
Andrea Cerasa
European Commission, Joint Research Centre (JRC), Ispra, Italy
and
Andrea Cerioli
University of Parma, Department of Economics and Management
and University Centre "Robust Statistics Academy" (Ro.S.A.), Parma, Italy
and
Domenico Perrotta
European Commission, Joint Research Centre (JRC), Ispra, Italy

1 Testing the sum-invariance property for the first-k significant digits

We assume the results and the notation given in Section 2 of the paper. The generalization of the test statistics described in Section 5 of the paper is quite straightforward, even if more cumbersome notation must be adopted.

In the following, the superscript $\{k\}$ means that the hypothesis – or the test statistic – is referred to the first-k significant digits. We recall that $D(x) = (D_1(x), \ldots, D_k(x))$ and $d = (d_1, \ldots, d_k)$, where $d_1 \in \{1, \ldots, 9\}$ and $d_j \in \{0, \ldots, 9\}$ for $j = 2, \ldots, k$. In such a case, the null hypothesis on the first-k digits is

$$H_0^{\text{digit}\{k\}} : D(X) \stackrel{\mathcal{L}}{=} D(10^U).$$

In addition, the generalization of H_0^{sum} is

$$H_0^{\text{sum}\{k\}} : \mathrm{E}[Z_d(X)] = C, \quad \forall d.$$

Given the random sample $(X_1, \ldots, X_n)$, we consider the standardized sample mean

$$T_d = \frac{\bar{Z}_d - C}{\sqrt{C(c_d + 1/2 - C)/n}},$$

where in this case

$$\bar{Z}_d = \frac{1}{n} \sum_{i=1}^{n} Z_d(X_i).$$

By adopting the notation $a_k = 9 \times 10^{k-1}$, the generalization of the Hotelling-type test is obtained by considering the random vector $\bar{Z}^{\{k\}}$ of order a_k with elements $\bar{Z}_d$, which are suitably ordered for $d_1 \in \{1, \ldots, 9\}$, $d_j \in \{0, \ldots, 9\}$ and $j = 2, \ldots, k$. As an example, a possible order could be achieved by noting that the values of $d = (d_1, \ldots, d_k)$ are in a one-to-one relationship with the set of integers $\{10^{k-1}, \ldots, 10^k - 1\}$. Under $H_0^{\text{sum}\{k\}}$, the expectation of $\bar{Z}^{\{k\}}$ is $\mathrm{E}[\bar{Z}^{\{k\}}] = (C, \ldots, C)^{\mathrm{T}}$, while its variance-covariance matrix turns out to be $\mathrm{var}[\bar{Z}^{\{k\}}] = n^{-1}\Sigma^{\{k\}}$, where in this case $\Sigma^{\{k\}}$ is a $(a_k \times a_k)$ matrix with elements given by $C(c_d + 1/2 - C)$ on the main diagonal and $-C^2$ off the main diagonal. The order of the elements of $\Sigma^{\{k\}}$ on the main diagonal is congruent with the order of the elements of $\bar{Z}^{\{k\}}$. With this notation, the Hotelling-type test $Q^{\{k\}}$ has the same structure as given in Section 5.2, i.e.

$$Q^{\{k\}} = n(\bar{Z}^{\{k\}} - \mathrm{E}[\bar{Z}^{\{k\}}])^{\mathrm{T}}(\Sigma^{\{k\}})^{-1}(\bar{Z}^{\{k\}} - \mathrm{E}[\bar{Z}^{\{k\}}]).$$

If $F_{Q^{\{k\}}}$ represents the null distribution function of $Q^{\{k\}}$, for a realization q of $Q^{\{k\}}$, the p-value is $p_{Q^{\{k\}}}(q) = 1 - F_{Q^{\{k\}}}(q)$. In addition, $Q^{\{k\}} \stackrel{\mathcal{L}}{\to} \chi^2_{a_k}$ as $n \to \infty$.

As to the generalization of the sup-norm test considered in Section 5.3, the null hypothesis $H_0^{\text{sum}\{k\}}$ may be decomposed into a_k sub-hypotheses as

$$H_0^{\text{sum}\{k\}} = \bigcap_d H_{0,d}^{\text{sum}\{k\}},$$

where $H_{0,d}^{\text{sum}\{k\}} : \mathrm{E}[Z_d(X)] = C$. Hence, in this case, the sup-norm test statistic for assessing $H_0^{\text{sum}\{k\}}$ is

$$M^{\{k\}} = \max_d |T_d|,$$

while the test statistic $\bar{Z}_d$ may be adopted for assessing the marginal sub-hypothesis $H_{0,d}^{\text{sum}\{k\}}$. If $q_{M^{\{k\}},1-\gamma/2}$ represents the $(1-\gamma/2)$th quantile of $M^{\{k\}}$ under the null, simultaneous acceptance intervals at significance level γ for the marginal test statistics $\bar{Z}_d$ are

$$[C - q_{M^{\{k\}},1-\gamma/2}\sqrt{C(c_d + 1/2 - C)/n}, C + q_{M^{\{k\}},1-\gamma/2}\sqrt{C(c_d + 1/2 - C)/n}].$$

These simultaneous acceptance intervals could be adopted in the plot reporting the observed values of $\bar{Z}_d$ to emphasize the digits involved in the rejection of the null hypothesis.

To generalize the test considered in Section 5.4, we consider the marginal p-values corresponding to the test statistics $|T_d|$. Hence, if $F_{|T_d|}$ is the null distribution function of $|T_d|$, let $p_{|T_d|}(t_d) = 1 - F_{|T_d|}(t_d)$ be the p-value for a realization t_d of $|T_d|$. Moreover, let $\psi^{\{k\}} : [0,1]^{a_k} \to \mathbb{R}$ be a suitable combining function, whose arguments are suitably ordered for $d_1 = 1, \ldots, 9$, $d_j = 0, \ldots, 9$ and $j = 2, \ldots, k$. The test statistic $G^{\{k\}}$ is the function $\psi^{\{k\}}$ with arguments given by the values $p_{|T_d|}(|T_d|)$. If the test rejects $H_0^{\text{sum}\{k\}}$ for large values of $G^{\{k\}}$, the p-value is $p_{G^{\{k\}}}(g^{\{k\}}) = 1 - F_{G^{\{k\}}}(g^{\{k\}})$, where $g^{\{k\}}$ is a realization of $G^{\{k\}}$ and $F_{G^{\{k\}}}$ is the null distribution function of $G^{\{k\}}$. In turn, $g^{\{k\}}$ is the value of the function $\psi^{\{k\}}$ whose arguments are the values $p_{|T_d|}(T_d)$.

The generalization of the combined test introduced in Section 5.5 is tailored to the null hypothesis $H_0^{\text{digit}\{k\}} \cap H_0^{\text{sum}\{k\}}$. In order to assess the marginal null hypothesis $H_0^{\text{digit}\{k\}}$, the chi-squared test statistic could be adopted, i.e.

$$\chi^{2\{k\}} = \sum_d \frac{(N_d - np_{D(X)}(d))^2}{np_{D(X)}(d)},$$

where $N_d = \sum_{i=1}^n \mathrm{I}_d(D(X_i))$. Moreover, in order to assess $H_0^{\text{sum}\{k\}}$, the test statistic $Q_{\{k\}}$ could be considered. Thus, the null hypothesis $H_0^{\text{digit}\{k\}} \cap H_0^{\text{sum}\{k\}}$ could be assessed by defining a suitable combining function $\phi : [0,1]^2 \to \mathbb{R}$, which provides the test statistic $L^{\{k\}} = \phi(p_{\chi^{2\{k\}}}(\chi^{2\{k\}}), p_{Q^{\{k\}}}(Q^{\{k\}}))$, where $p_{\chi^{2\{k\}}}(x) = 1 - F_{\chi^{2\{k\}}}(x)$ and $F_{\chi^{2\{k\}}}$ is the null distribution function of $\chi^{2\{k\}}$. By assuming that the test rejects the null hypothesis for large values of $L^{\{k\}}$, the p-value is given by $p_{L^{\{k\}}}(l^{\{k\}}) = 1 - F_{L^{\{k\}}}(l^{\{k\}})$, where $l^{\{k\}} = \phi(p_{\chi^{2\{k\}}}(x^{\{k\}}), p_{Q^{\{k\}}}(q^{\{k\}}))$ is a realization of $L^{\{k\}}$ and $F_{L^{\{k\}}}$ represents the null distribution function of $L^{\{k\}}$, while $x^{\{k\}}$ and $q^{\{k\}}$ are the realizations of $\chi^{2\{k\}}$ and $Q^{\{k\}}$, respectively.

Finally, all the Monte Carlo procedures described in Section 5 of the paper, where B Monte Carlo replicates of the relevant test statistic are generated from

$$H_0 : S(X) \stackrel{\mathcal{L}}{=} 10^U,$$

are easily extended to the general case.

2 Additional simulation results

In this section we provide additional simulation results that complement those given in §6 of the paper.

Table 1: Deviance matrix $\Upsilon = n\text{var}[\bar{Z}]$ under model (b).

	d								
d'	1	2	3	4	5	6	7	8	9
1	0.3350	-0.1491	-0.1528	-0.1550	-0.1564	-0.1574	-0.1582	-0.1588	-0.1593
2	-0.1491	0.7382	-0.1609	-0.1632	-0.1647	-0.1658	-0.1666	-0.1672	-0.1677
3	-0.1528	-0.1609	1.1570	-0.1672	-0.1688	-0.1699	-0.1707	-0.1714	-0.1719
4	-0.1550	-0.1632	-0.1672	1.5830	-0.1712	-0.1723	-0.1732	-0.1738	-0.1743
5	-0.1564	-0.1647	-0.1688	-0.1712	2.0110	-0.1739	-0.1748	-0.1754	-0.1759
6	-0.1574	-0.1658	-0.1699	-0.1723	-0.1739	2.4410	-0.1759	-0.1766	-0.1771
7	-0.1582	-0.1666	-0.1707	-0.1732	-0.1748	-0.1759	2.8730	-0.1774	-0.1780
8	-0.1588	-0.1672	-0.1714	-0.1738	-0.1754	-0.1766	-0.1774	3.3050	-0.1786
9	-0.1593	-0.1677	-0.1719	-0.1743	-0.1759	-0.1771	-0.1780	-0.1786	3.7370

Table 2: Monte Carlo estimate of the rejection probability of each test under different data generating models, when $n = 50$. The nominal test size is 0.01.

Data generating model	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
Benford's law: H_0 true	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
(a): H_0^{digit} and H_0^{sum} true; H_0 false	0.010	0.011	0.010	0.010	0.010	0.010	0.010	0.010
(b): H_0^{digit} true; H_0^{sum} false	0.009	0.116	0.770	0.004	0.004	0.729	0.007	0.007
$(c1)$: H_0^{sum} true; H_0^{digit} false	0.010	0.011	0.010	0.010	0.010	0.010	0.010	0.010
$(c2)$: H_0^{sum} true; H_0^{digit} false	0.010	0.021	0.011	0.010	0.011	0.011	0.010	0.010
$(c3)$: H_0^{sum} true; H_0^{digit} false	0.009	0.245	0.017	0.012	0.013	0.015	0.012	0.013

Table 3: Monte Carlo estimate of the rejection probability of each test under different data generating models, when $n = 200$. The nominal test size is 0.01.

Data generating model	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
Benford's law: H_0 true	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
(a): H_0^{digit} and H_0^{sum} true; H_0 false	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
(b): H_0^{digit} true; H_0^{sum} false	0.010	0.700	1.000	0.009	0.011	1.000	0.010	0.011
$(c1)$: H_0^{sum} true; H_0^{digit} false	0.010	0.015	0.010	0.010	0.011	0.010	0.010	0.010
$(c2)$: H_0^{sum} true; H_0^{digit} false	0.012	0.080	0.011	0.010	0.010	0.012	0.011	0.011
$(c3)$: H_0^{sum} true; H_0^{digit} false	0.029	0.981	0.018	0.012	0.012	0.030	0.025	0.025

Table 4: Monte Carlo estimate of the rejection probability of each test under different data generating models, when $n = 500$. The nominal test size is 0.01.

Data generating model	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
Benford's law: H_0 true	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
(a): H_0^{digit} and H_0^{sum} true; H_0 false	0.011	0.011	0.010	0.010	0.010	0.010	0.011	0.010
(b): H_0^{digit} true; H_0^{sum} false	0.010	0.997	1.000	0.059	0.064	1.000	0.043	0.047
$(c1)$: H_0^{sum} true; H_0^{digit} false	0.011	0.026	0.010	0.010	0.010	0.010	0.010	0.010
$(c2)$: H_0^{sum} true; H_0^{digit} false	0.016	0.382	0.011	0.011	0.011	0.016	0.014	0.014
$(c3)$: H_0^{sum} true; H_0^{digit} false	0.105	1.000	0.017	0.012	0.013	0.095	0.080	0.080

Table 5: Monte Carlo estimate of the rejection probability of each test under different data generating models, when $n = 100$. The nominal test size is 0.05.

Data generating model	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
Benford's law: H_0 true	0.050	0.051	0.050	0.050	0.050	0.050	0.050	0.049
(a): H_0^{digit} and H_0^{sum} true; H_0 false	0.050	0.051	0.048	0.050	0.050	0.050	0.049	0.050
(b): H_0^{digit} true; H_0^{sum} false	0.049	0.538	1.000	0.037	0.034	1.000	0.040	0.042
$(c1)$: H_0^{sum} true; H_0^{digit} false	0.050	0.061	0.050	0.051	0.051	0.050	0.050	0.050
$(c2)$: H_0^{sum} true; H_0^{digit} false	0.052	0.133	0.053	0.050	0.052	0.054	0.052	0.053
$(c3)$: H_0^{sum} true; H_0^{digit} false	0.070	0.845	0.069	0.063	0.064	0.078	0.068	0.070

Table 6: Monte Carlo estimate of the rejection probability of each test under different data generating models, *when the null distribution function of each test statistic is estimated by Monte Carlo simulation from model (*c3*)*. The sample size is $n = 100$ and the nominal test size is 0.01, as in Table 1 of the manuscript.

Data generating model	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
Benford's law: H_0 true	0.007	0.000	0.005	0.008	0.010	0.006	0.007	0.010
(a): H_0^{digit} and H_0^{sum} true; H_0 false	0.007	0.000	0.005	0.006	0.009	0.005	0.006	0.009
(b): H_0^{digit} true; H_0^{sum} false	0.007	0.001	1.000	0.004	0.009	0.999	0.005	0.009
$(c1)$: H_0^{sum} true; H_0^{digit} false	0.008	0.000	0.006	0.008	0.009	0.006	0.007	0.009
$(c2)$: H_0^{sum} true; H_0^{digit} false	0.006	0.000	0.005	0.008	0.009	0.005	0.006	0.008
$(c3)$: H_0^{sum} true; H_0^{digit} false	0.010	0.011	0.009	0.009	0.010	0.010	0.009	0.010

Table 7: Estimated power when X is a Lognormal random variable with scale parameter 1, for different values of the shape parameter and $n = 50$. The nominal test size is 0.01.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
0.5	0.520	0.314	0.554	0.141	0.135	0.549	0.457	0.454
0.6	0.153	0.108	0.164	0.058	0.055	0.162	0.127	0.126
0.7	0.048	0.041	0.050	0.030	0.027	0.049	0.043	0.041
0.8	0.020	0.019	0.021	0.017	0.015	0.020	0.019	0.018

Table 8: Estimated power when X is a Lognormal random variable with scale parameter 1, for different values of the shape parameter and $n = 200$. The nominal test size is 0.01.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
0.5	1.000	0.996	1.000	0.974	0.981	1.000	1.000	1.000
0.6	0.897	0.629	0.908	0.460	0.484	0.910	0.865	0.865
0.7	0.311	0.169	0.321	0.113	0.116	0.324	0.264	0.264
0.8	0.065	0.044	0.066	0.036	0.035	0.066	0.055	0.055

Table 9: Estimated power when X is a Lognormal random variable with scale parameter 1, for different values of the shape parameter and $n = 500$. The nominal test size is 0.01.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.6	1.000	0.997	1.000	0.990	0.990	1.000	1.000	1.000
0.7	0.865	0.584	0.879	0.486	0.489	0.881	0.829	0.830
0.8	0.232	0.121	0.238	0.098	0.099	0.241	0.196	0.196

Table 10: Estimated power when X is a Weibull random variable with scale parameter 1, for different values of the shape parameter and $n = 50$. The nominal test size is 0.01.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
1.2	0.021	0.032	0.022	0.016	0.018	0.022	0.019	0.020
1.4	0.045	0.048	0.047	0.027	0.027	0.046	0.039	0.039
1.6	0.099	0.077	0.105	0.045	0.042	0.103	0.083	0.082
1.8	0.194	0.131	0.213	0.067	0.058	0.209	0.161	0.157
2.0	0.330	0.213	0.371	0.102	0.086	0.364	0.281	0.275
2.2	0.487	0.325	0.555	0.144	0.119	0.543	0.427	0.421

Table 11: Estimated power when X is a Weibull random variable with scale parameter 1, for different values of the shape parameter and $n = 200$. The nominal test size is 0.01.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
1.2	0.089	0.092	0.089	0.047	0.052	0.091	0.076	0.078
1.4	0.328	0.198	0.334	0.121	0.131	0.338	0.279	0.281
1.6	0.716	0.432	0.733	0.316	0.340	0.738	0.662	0.664
1.8	0.953	0.760	0.962	0.674	0.706	0.963	0.935	0.936
2.0	0.997	0.956	0.998	0.930	0.942	0.999	0.996	0.996
2.2	1.000	0.997	1.000	0.993	0.995	1.000	1.000	1.000

Table 12: Estimated power when X is a Weibull random variable with scale parameter 1, for different values of the shape parameter and $n = 500$. The nominal test size is 0.01.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
1.2	0.365	0.279	0.365	0.186	0.190	0.370	0.323	0.326
1.4	0.890	0.639	0.898	0.546	0.552	0.901	0.859	0.859
1.6	0.999	0.963	0.999	0.945	0.947	0.999	0.998	0.998
1.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 13: Estimated power under the Generalized Benford's law with parameter $\alpha \neq 0$, for $n = 50$. The nominal test size is 0.01.

α	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
-0.8	0.279	0.814	0.327	0.234	0.343	0.321	0.264	0.332
-0.6	0.089	0.531	0.114	0.086	0.148	0.110	0.089	0.133
-0.4	0.021	0.221	0.028	0.023	0.046	0.027	0.022	0.038
0.4	0.114	0.137	0.105	0.063	0.056	0.108	0.101	0.097
0.6	0.303	0.395	0.283	0.140	0.135	0.290	0.266	0.263
0.8	0.576	0.706	0.552	0.270	0.280	0.561	0.528	0.527

Table 14: Estimated power under the Generalized Benford's law with parameter $\alpha \neq 0$, for $n = 200$. The nominal test size is 0.01.

α	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
-0.8	0.999	1.000	0.999	0.985	0.989	0.999	0.999	0.999
-0.6	0.928	0.996	0.936	0.805	0.832	0.937	0.912	0.916
-0.4	0.425	0.821	0.441	0.312	0.349	0.443	0.402	0.416
0.4	0.577	0.773	0.556	0.344	0.357	0.567	0.532	0.536
0.6	0.964	0.993	0.960	0.814	0.828	0.962	0.952	0.953
0.8	1.000	1.000	1.000	0.987	0.989	1.000	1.000	1.000

Table 15: Estimated power under the Generalized Benford's law with parameter $\alpha \neq 0$, for $n = 500$. The nominal test size is 0.01.

α	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
-0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
-0.4	0.971	0.998	0.973	0.884	0.890	0.974	0.963	0.964
0.4	0.982	0.998	0.980	0.894	0.897	0.982	0.976	0.976
0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 16: Estimated power when X is a Lognormal random variable with scale parameter 1, for different values of the shape parameter and $n = 100$. The nominal test size is 0.05.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
0.5	0.992	0.954	0.993	0.869	0.872	0.993	0.986	0.986
0.6	0.711	0.563	0.724	0.426	0.431	0.727	0.659	0.657
0.7	0.287	0.231	0.291	0.180	0.183	0.293	0.256	0.252
0.8	0.118	0.107	0.118	0.091	0.091	0.118	0.110	0.107

Table 17: Estimated power when X is a Weibull random variable with scale parameter 1, for different values of the shape parameter and $n = 100$. The nominal test size is 0.05.

Shape parameter	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
1.2	0.132	0.156	0.132	0.102	0.105	0.133	0.123	0.122
1.4	0.289	0.253	0.292	0.183	0.186	0.294	0.258	0.256
1.6	0.544	0.427	0.556	0.325	0.330	0.559	0.493	0.491
1.8	0.793	0.656	0.810	0.540	0.550	0.813	0.750	0.751
2.0	0.938	0.857	0.950	0.768	0.776	0.950	0.918	0.919
2.2	0.988	0.963	0.993	0.914	0.918	0.992	0.983	0.983

Table 18: Estimated power under the Generalized Benford's law with parameter $\alpha \neq 0$, for $n = 100$. The nominal test size is 0.05.

α	χ^2	KS	Q	M	G	$L_{\chi^2,Q}$	$L_{\chi^2,M}$	$L_{\chi^2,G}$
-0.8	0.954	0.999	0.960	0.877	0.884	0.960	0.942	0.945
-0.6	0.714	0.963	0.730	0.594	0.607	0.732	0.684	0.697
-0.4	0.294	0.715	0.309	0.245	0.255	0.309	0.279	0.293
0.4	0.475	0.626	0.453	0.358	0.362	0.463	0.449	0.446
0.6	0.840	0.936	0.825	0.690	0.695	0.833	0.814	0.815
0.8	0.981	0.997	0.979	0.920	0.922	0.980	0.975	0.976

3 Contamination scheme of §7.2

In this section we detail the digit contamination scheme adopted in §7.2 of the paper, for the purpose of comparing the performance of Benford tests to that of outlier detection methods. Coherently with the attention that the European Commission pays on undervaluation[1], and consistently with some recently detected fraud cases[2], this contamination scheme aims at mimicking the behavior of a trader that modifies the declared value of his import transactions in order to reduce the amount of the due duties.

For each transaction i to be manipulated, we define x_i to be the non-contaminated value (i.e., the simulated market value of the import), while $S(x_i) = 10^{\langle \log_{10} |x_i| \rangle}$ and $K(x_i) = \lfloor \log_{10} |x_i| \rfloor$ denote its significand and its order of magnitude, respectively. The result of the digit manipulation algorithm is a new value $\tilde{x}_i$ which is obtained as follows.

First, we assume that the cheating trader modifies the first digit of transaction value i according to the rule

$$\tilde{D}_1(x_i) = \begin{cases} D_1(x_i) - \theta_i & \text{if } D_1(x_i) > \theta_i \\ 9 - |D_1(x_i) - \theta_i| & \text{if } D_1(x_i) \leq \theta_i, \end{cases}$$

where $\theta_i \in \{1, \ldots, 8\}$ represents the first-digit shift of transaction value i. The shift θ_i is a realization of the random variable Θ with probability function

$$p_\Theta(\theta) = \frac{9 - \theta}{36} \, \mathrm{I}_{\{1,\ldots,8\}}(\theta). \tag{1}$$

Being inversely related to the value of the shift, probability distribution (1) represents the common tendency of the manipulator to declare a value which is lower than, but anyway close to, the original one, in order to reduce the risk of being excessively "out of the market". To further emphasize this tendency, we also assume that $\tilde{D}_2(x_i) = \tilde{D}_3(x_i) = \tilde{D}_4(x_i) = 9$, as it often happens in real markets, whereas the remaining digits – if any – are all rounded to 0, i.e. $\tilde{D}_j(x_i) = 0$ for $j > 4$.

We obtain the significand of the fabricated value, with k digits, as

$$\tilde{S}(x_i) = \sum_{j=1}^{k} 10^{1-j} \tilde{D}_j(x_i).$$

The order of magnitude $\tilde{K}(x_i)$ of the manipulated value is instead defined as

$$\tilde{K}(x_i) = \begin{cases} K(x_i) & \text{if } D_1(x_i) > \theta_i \\ K(x_i) - 1 & \text{if } D_1(x_i) \leq \theta_i. \end{cases}$$

[1] see:
https://ec.europa.eu/jrc/en/research-topic/antifraud
https://ec.europa.eu/anti-fraud/investigations/eu-revenue/trade_customs_fraud_en

[2] see:
https://ec.europa.eu/commission/presscorner/detail/en/IP_14_1001
https://ec.europa.eu/anti-fraud/sites/antifraud/files/pif_report_2018_en.pdf

Finally, the fabricated value $\tilde{x}_i$ is

$$\tilde{x}_i = \tilde{S}(x_i)10^{\tilde{K}(x_i)}.$$

The following table shows how the manipulation scheme works in a concrete example, where we take $x_i = 43457.23$. For instance, if the most likely shift $\theta_i = 1$ is selected, the contaminated value becomes $\tilde{x}_i = 39990$, which is fairly close to the original one but still below the threshold 40000. On the other hand, $\tilde{x}_i = 5999$ when the unlikely shift $\theta_i = 8$ occurs, so that a much larger and risky undervaluation takes place.

θ_i	(1)	$\tilde{S}(x_i)$	$\tilde{K}(x_i)$	$\tilde{x}_i$
1	0.222	3.999	4	39990
2	0.194	2.999	4	29990
3	0.167	1.999	4	19990
4	0.139	9.999	3	9999
5	0.111	8.999	3	8999
6	0.083	7.999	3	7999
7	0.056	6.999	3	6999
8	0.028	5.999	3	5999