Statistical Models and the Benford Hypothesis: A Unified Framework

Appendix: Technical Proofs

Lucio Barabesi[1], Andrea Cerioli[2*] and Marco Di Marzio[3]

[1]Department of Economics and Statistics, University of Siena, Siena, Italy.
[2*]Department of Economics and Management, University of Parma, Parma, Italy.
[3]Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, "G. D'Annunzio" University, Pescara, Italy.

*Corresponding author(s). E-mail(s): andrea.cerioli@unipr.it;
Contributing authors: lucio.barabesi@unisi.it;
marco.dimarzio@unich.it;

Proof of Corollary 1

On the basis of Theorem 1, the distribution function of $s(X)$ may be expressed as

$$F_{s(X)}(u) = u + \frac{1}{2\pi} \sum_{k\in\mathbb{Z},k\neq 0} \varphi_k \phi_k(u), \qquad u \in [0,1[,$$

where for simplicity we assume that $\varphi_k = \varphi_Y(k)$ and $\phi_k(u) = \frac{1-e^{-2\pi \mathrm{i}ku}}{\mathrm{i}k}$. Since for $k \in \mathbb{N}$ it holds

$$\Re(\varphi_{-k}) = \Re(\varphi_k)\ ,\ \Im(\varphi_{-k}) = -\Im(\varphi_k),$$

and

$$\Re(\phi_{-k}(u)) = \Re(\phi_k(u))\ ,\ \Im(\phi_{-k}(u)) = -\Im(\phi_k(u)),$$

it reads

$$\varphi_k\phi_k(u) + \varphi_{-k}\phi_{-k}(u) = 2\Re(\varphi_k\phi_k(u)).$$

Hence, it follows

$$F_{s(X)}(u) = u + \frac{1}{\pi} \sum_{k\in\mathbb{N}} \Re(\varphi_k\phi_k(u)), \qquad u \in [0,1[.$$

Moreover, it holds

$$\begin{aligned}\Re(\varphi_k\phi_k(u)) &= |\varphi_k\phi_k(u)|\cos(\arg(\varphi_k\phi_k(u))) = |\varphi_k|\cdot|\phi_k(u)|\cos(\vartheta_k + \arg(\phi_k(u)))\\ &= |\varphi_k|\left(|\phi_k(u)|\cos(\vartheta_k)\cos(\arg(\phi_k(u))) - |\phi_k(u)|\sin(\vartheta_k)\sin(\arg(\phi_k(u)))\right)\\ &= |\varphi_k|\left(\Re(\phi_k(u))\cos(\vartheta_k) - \Im(\phi_k(u))\sin(\vartheta_k)\right).\end{aligned}$$

Finally, since

$$\Re(\phi_k(u)) = \frac{\sin(2\pi ku)}{k}\ ,\ \Im(\phi_k(u)) = -\frac{1-\cos(2\pi ku)}{k},$$

it follows that

$$\begin{aligned}\Re(\varphi_k\phi_k(u)) &= \frac{|\varphi_k|}{k}\left(\sin(2\pi ku)\cos(\vartheta_k) + (1-\cos(2\pi ku))\sin(\vartheta_k)\right)\\ &= \frac{|\varphi_k|}{k}\left(\sin(\vartheta_k) - \sin(\vartheta_k - 2\pi ku)\right),\end{aligned}$$

from which the result follows.

Proof of Corollary 2

On the basis of Corollary 1, $F_{s(X)}$ is expressed as a series of differentiable functions with continuous derivatives on $[0,1[$, and which converges for each

$u \in [0, 1[$. Moreover, the derived series

$$1 + 2 \sum_{k \in \mathbb{N}} |\varphi_Y(k)| \cos(\vartheta_k - 2\pi k u), \qquad u \in [0, 1[,$$

converges uniformly under the assumed condition on the basis of the Weierstrass M-test (Junghenn, 2015, Theorem 7.3.6), since

$$|\sum_{k \in \mathbb{N}} |\varphi_Y(k)| \cos(\vartheta_k - 2\pi k u)| \le \sum_{k \in \mathbb{N}} |\varphi_Y(k)| \cdot |\cos(\vartheta_k - 2\pi k u)| < \sum_{k \in \mathbb{N}} |\varphi_Y(k)| < \infty.$$

Hence, by means of Junghenn (2015, Theorem 7.3.13), we have that $F_{s(X)}$ is differentiable and

$$F'_{s(X)}(u) = 1 + 2 \sum_{k \in \mathbb{N}} |\varphi_Y(k)| \cos(\vartheta_k - 2\pi k u), \;\; u \in [0, 1[.$$

Proof of Proposition 1

For notational simplicity, we write $\delta = \frac{\mu}{\sqrt{2}\sigma}$. From the definition of the characteristic function it holds

$$\begin{aligned}
\varphi_Y(t) &= \int_{-\infty}^{\infty} e^{2\pi \mathrm{i} t \log_{10}|x|} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\
&= \int_{-\infty}^{\infty} |x|^{2\pi \mathrm{i} C t} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\
&= \frac{1}{\sqrt{\pi}} (2\sigma^2)^{\pi \mathrm{i} C t} \int_{-\infty}^{\infty} |x|^{2\pi \mathrm{i} C t} e^{-(x-\delta)^2} dx \\
&= \frac{1}{\sqrt{\pi}} (2\sigma^2)^{\pi \mathrm{i} C t} \int_{0}^{\infty} x^{2\pi \mathrm{i} C t} (e^{-(x-\delta)^2} + e^{-(x+\delta)^2}) dx.
\end{aligned}$$

On the basis of Gradshteyn and Ryzhik (2015, Formula 3.462.1), we obtain

$$\int_0^{\infty} x^{\upsilon - 1} e^{-(x+\gamma)^2} dx = 2^{-\frac{1}{2}\upsilon} e^{-\frac{1}{2}\gamma^2} \Gamma(\upsilon) U\left(\upsilon - \frac{1}{2}, \sqrt{2}\gamma\right),$$

where $\Re(\upsilon) > 0$, $\gamma \in \mathbb{R}$ and $U(a, z)$ is the parabolic cylinder function for $a, z \in \mathbb{C}$. Hence,

$$\begin{aligned}
\varphi_Y(t) &= \frac{\sigma^{2\pi \mathrm{i} C t} e^{-\frac{1}{2}\delta^2} \Gamma(2\pi \mathrm{i} C t + 1)}{\sqrt{2\pi}} \times \\
&\quad \times \left(U\left(2\pi \mathrm{i} C t + \frac{1}{2}, \sqrt{2}\delta\right) + U\left(2\pi \mathrm{i} C t + \frac{1}{2}, -\sqrt{2}\delta\right)\right).
\end{aligned}$$

Since

$$U(a,z)+U(a,-z)=\frac{\sqrt{\pi}2^{-\frac{1}{2}a+\frac{3}{4}}e^{-\frac{1}{4}z^2}}{\Gamma\left(\frac{1}{2}a+\frac{3}{4}\right)}\,M\left(\frac{a}{2}+\frac{1}{4},\frac{1}{2},\frac{z^2}{2}\right),$$

where $M(a,b,z)$ is the Kummer confluent hypergeometric function and $a,b,z\in\mathbb{C}$, we also have

$$\varphi_Y(t)=\frac{2^{-\pi\mathrm{i}Ct}\sigma^{2\pi\mathrm{i}Ct}e^{-\delta^2}\Gamma(2\pi\mathrm{i}Ct+1)}{\Gamma\left(\pi\mathrm{i}Ct+1\right)}\,M\left(\pi\mathrm{i}Ct+\frac{1}{2},\frac{1}{2},\delta^2\right). \tag{1}$$

By adopting the well-known Legendre duplication formula for the Gamma function, it holds

$$\frac{\Gamma(2\pi\mathrm{i}Ct+1)}{\Gamma\left(\pi\mathrm{i}Ct+1\right)}=\frac{1}{\sqrt{\pi}}\,2^{2\pi\mathrm{i}Ct}\Gamma\left(\pi\mathrm{i}Ct+\frac{1}{2}\right).$$

Moreover, by considering the Kummer's transformation (Olver et al, 2010, Formula 13.2.39)

$$M(a,b,z)=e^zM(b-a,b,-z),$$

we obtain

$$M\left(\pi\mathrm{i}Ct+\frac{1}{2},\frac{1}{2},\delta^2\right)=e^{\delta^2}M\left(-\pi\mathrm{i}Ct,\frac{1}{2},-\delta^2\right)$$

and the statement follows by substituting these results in formula (1). In addition, since for $c\in\mathbb{R}$ it holds

$$\left|\Gamma\left(\frac{1}{2}+\mathrm{i}c\right)\right|=\sqrt{\frac{\pi}{\cosh(\pi c)}},$$

the expression of $|\varphi_Y(t)|$ follows.

Proof of Proposition 2

The characteristic function φ_Y follows from the definition and the properties of the Gamma function, since

$$\begin{aligned}\varphi_Y(t)&=\int_0^\infty e^{2\pi\mathrm{i}t\log_{10}x}\,\frac{1}{\Gamma(\alpha)}\,x^{\alpha-1}e^{-x}dx=\frac{1}{\Gamma(\alpha)}\int_0^\infty x^{2\pi\mathrm{i}Ct+\alpha-1}e^{-x}dx\\&=\frac{\Gamma(\alpha+2\pi\mathrm{i}Ct)}{\Gamma(\alpha)}.\end{aligned}$$

Moreover, on the basis of the formula

$$\left|\frac{\Gamma(a+\mathrm{i}b)}{\Gamma(a)}\right|^2=\prod_{n=0}^{\infty}\frac{1}{1+\frac{b^2}{(a+n)^2}},$$

where $a \in \mathbb{R} \setminus \mathbb{Z}^-, b \in \mathbb{R}$ (Olver et al, 2010, Formula 5.8.3), the expression of $|\varphi_Y(t)|$ follows.

Proof of Proposition 3

The characteristic function φ_Y follows from the definition and the properties of the Beta function, since

$$\begin{aligned}
\varphi_Y(t) &= \int_0^1 e^{2\pi \mathrm{i} t \log_{10} x} \, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \, x^{\alpha-1}(1-x)^{\beta-1} dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{2\pi \mathrm{i} Ct+\alpha-1}(1-x)^{\beta-1} dx \\
&= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+2\pi \mathrm{i} Ct)}{\Gamma(\alpha)\Gamma(\alpha+\beta+2\pi \mathrm{i} Ct)}.
\end{aligned}$$

Moreover, the expression of $|\varphi_Y(t)|$ follows from the formula for $\left|\frac{\Gamma(a+\mathrm{i}b)}{\Gamma(a)}\right|$ already adopted in the proof of Proposition 2.

Proof of Proposition 4

We consider the Wright function (Mainardi et al, 2010) given by

$$W_{a,b}(z) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(ak+b)} \, z^k,$$

where $a \in]-1, \infty[$, $b \in \mathbb{C}$ and $z \in \mathbb{C}$. The probability density function of X can be expressed as

$$f_X(x) = \frac{1}{x} \, W_{-\alpha,0}(-x^{-\alpha}), \qquad x \in \mathbb{R}^+.$$

Thus, the Mellin transform of X is given by

$$\mathrm{E}[X^s] = \int_0^\infty x^{s-1} W_{-\alpha,0}(-x^{-\alpha}) dx = \frac{1}{\alpha} \int_0^\infty x^{-\frac{s}{\alpha}-1} \, W_{-\alpha,0}(-x) dx.$$

On the basis of Mainardi et al (2010, Formula (4.2)), it follows that

$$\int_0^\infty x^{\delta-1} \, W_{-\upsilon,0}(-x) dx = \frac{\upsilon\Gamma(1+\delta)}{\Gamma(1+\upsilon\delta)},$$

where $\Re(\delta) > -1$ and $v \in [0,1[$. Hence, by taking $\delta = -\frac{s}{\alpha}$ and $v = \alpha$ in the previous expression, it holds

$$\mathrm{E}[X^s] = \frac{\Gamma\left(1-\frac{s}{\alpha}\right)}{\Gamma(1-s)}$$

with $\Re(s) < \alpha$. Thus, the expression of φ_Y is obtained by letting $s = 2\pi \mathrm{i} Ct$. Moreover, $|\varphi_Y(t)|$ follows by noting that

$$|\Gamma(1+\mathrm{i}b)| = \sqrt{\frac{\pi b}{\sinh(\pi b)}}.$$

Proof of Proposition 5

If $V_l = 2\pi \log_{10}|Z_l|$, we have

$$\begin{aligned}
\varphi_Y(t) &= \sum_{l=1}^{L} w_l \varphi_{V_l}(t) = \sum_{l=1}^{L} w_l \int_{\mathbb{R}} e^{2\pi \mathrm{i} t \log_{10}|x|} \frac{1}{k_l\sigma} f_Z\left(\frac{x-k_l\mu}{k_l\sigma}\right) dx \\
&= \sum_{l=1}^{L} w_l e^{2\pi \mathrm{i} t \log_{10} k_l} \int_{\mathbb{R}} e^{2\pi \mathrm{i} t \log_{10}|x|} \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right) dx \\
&= \varphi_{Y_1}(t)\varphi_{Y_2}(t), \qquad t \in \mathbb{R}.
\end{aligned}$$

Proof of Proposition 6

First, since $c = c_R + \mathrm{i} c_I$ with c_R and c_I real vectors of order $(N+1)$, on the basis of the definition of $\mathcal{C}$, we take

$$c_R = \begin{pmatrix} c_0 \\ v \end{pmatrix}, \qquad c_I = \begin{pmatrix} 0 \\ z \end{pmatrix},$$

where v and z are real vectors of order N such that $c_0^2 + v^t v + z^t z = 1$. In addition, we note that $T_u = w_u w_u^*$, where $w_u = w_{R,u} + \mathrm{i} w_{I,u}$, while $w_{R,u} = \begin{pmatrix} 1 \\ \alpha_u \end{pmatrix}$ and $w_{I,u} = \begin{pmatrix} 0 \\ \beta_u \end{pmatrix}$ with

$$\alpha_u = \begin{pmatrix} \cos(2\pi u) \\ \vdots \\ \cos(2\pi N u) \end{pmatrix}, \qquad \beta_u = \begin{pmatrix} \sin(2\pi u) \\ \vdots \\ \sin(2\pi N u) \end{pmatrix}.$$

Hence, we obtain

$$\begin{aligned}
g(u) &= |(c_R^t w_{R,u} + c_I^t w_{I,u}) + \mathrm{i}(c_R^t w_{I,u} + c_I^t w_{R,u})|^2 \\
&= (c_0 + v^t \alpha_u + z^t \beta_u)^2 + (v^t \beta_u + z^t \alpha_u)^2.
\end{aligned}$$

It is well known that $\Lambda_{N,n}$ is asymptotically equivalent to the score test statistic under the null hypothesis, i.e. the two statistics coincide except for a stochastic order $o_P(1)$ as $n \to \infty$. In order to implement the score test statistic, we consider the unique parameter vector $\theta = \begin{pmatrix} v \\ z \end{pmatrix}$ of order $(2N)$ (since c_0 depends on v and z) and we introduce the notation $\ell_\theta(u) = \log g(u)$. In such a case, it follows that

$$\dot{\ell}_\theta(u) = \frac{\partial \ell_\theta}{\partial \theta^t}(u) = \frac{2}{g(u)} \begin{pmatrix} (c_0 + v^t\alpha_u + z^t\beta_u)(-\frac{v}{c_0} + \alpha_u) + (v^t\beta_u + z^t\alpha_u)\beta_u \\ (c_0 + v^t\alpha_u + z^t\beta_u)(-\frac{v}{c_0} + \beta_u) + (v^t\beta_u + z^t\alpha_u)\alpha_u \end{pmatrix}.$$

Under H_0, the parameter vector θ reduces to the zero vector or order $(2N)$, i.e. $\theta = 0$, while $c_0 = 1$. Hence, we obtain

$$\dot{\ell}_0(u) = 2 \begin{pmatrix} \alpha_u \\ \beta_u \end{pmatrix}.$$

In addition, the Fisher information matrix is given by $I_\theta = \mathrm{E}[\dot{\ell}_\theta(s(X))\dot{\ell}_\theta(s(X))^t]$. Under H_0, it holds $s(X) \overset{\mathcal{L}}{=} U$ where U is a Uniform random variable on $[0, 1[$. Since

$$\mathrm{E}[\cos(2\pi kU)\cos(2\pi jU)] = \mathrm{E}[\sin(2\pi kU)\sin(2\pi jU)] = \frac{1}{2}\,\delta_{kj},$$

and

$$\mathrm{E}[\cos(2\pi kU)\sin(2\pi jU)] = 0$$

for $k, j \in \{1, \ldots, N\}$, where δ_{kj} is the Kronecker delta function, it is easily assessed that $\frac{1}{2}I_0$ reduces to the identity matrix of order $(2N)$. Thus, under H_0 the score test statistic is given by

$$\begin{aligned} &\frac{1}{n}\left(\sum_{i=1}^{n} \dot{\ell}_0(s(X_i))\right)^t I_0^{-1} \left(\sum_{i=1}^{n} \dot{\ell}_0(s(X_i))\right) \\ &= \frac{2}{n}\begin{pmatrix} \sum_{i=1}^n \alpha_{s(X_i)} \\ \sum_{i=1}^n \beta_{s(X_i)} \end{pmatrix}^t \begin{pmatrix} \sum_{i=1}^n \alpha_{s(X_i)} \\ \sum_{i=1}^n \beta_{s(X_i)} \end{pmatrix} \\ &= Z_{N,n}^2 \end{aligned}$$

and the result follows.

Proof of Proposition 7

First, we remark that $e^{2\pi k i \langle x \rangle} = e^{2\pi k i x}$ for $k \in \mathbb{N}$, so that it follows $e^{2\pi k i s(\sigma x)} = e^{2\pi k i s(\sigma)} e^{2\pi k i s(x)}$. Write $Z_{N,n}^2(X_1, \ldots, X_n)$ and $R_{N,n}(X_1, \ldots, X_n)$ for $Z_{N,n}^2$ and

$R_{N,n}$. Thus, for $Z^2_{N,n}$ we obtain

$$\begin{aligned}
Z^2_{N,n}(\sigma X_1,\dots,\sigma X_n) &= 2n\sum_{k=1}^{N}\left|\frac{1}{n}\sum_{i=1}^{n}e^{2\pi k\mathrm{i}s(\sigma X_i)}\right|^2\\
&= 2n\sum_{k=1}^{N}|e^{2\pi k\mathrm{i}s(\sigma)}|^2\left|\frac{1}{n}\sum_{i=1}^{n}e^{2\pi k\mathrm{i}s(X_i)}\right|^2\\
&= 2n\sum_{k=1}^{N}\left|\frac{1}{n}\sum_{i=1}^{n}e^{2\pi k\mathrm{i}s(X_i)}\right|^2\\
&= Z^2_{N,n}(X_1,\dots,X_n).
\end{aligned}$$

For $R_{N,n}$, we instead have

$$g(s(\sigma x)) = \sum_{k=0}^{N}\sum_{j=0}^{N}c_k c_j^* e^{2\pi\mathrm{i}(k-j)s(\sigma x)} = \sum_{k=0}^{N}\sum_{j=0}^{N}h_k h_j^* e^{2\pi\mathrm{i}(k-j)s(x)} = h^* T_{s(x)} h,$$

where $h = (h_0, h_1, \dots, h_N)^{\mathrm{T}}$ and $h_k = e^{2\pi\mathrm{i}ks(\sigma)}c_k$. It holds $h_0 = c_0$, while $h^*h = 1$, and therefore h is in turn defined on the parameter space $\mathcal{C}$, i.e. h is actually a "rotation" of the original parameter vector c in $\mathcal{C}$. In such a case, it follows that

$$\max_{c\in\mathcal{C}}\prod_{i=1}^{n}c^* T_{s(\sigma X_i)}c = \max_{h\in\mathcal{C}}\prod_{i=1}^{n}h^* T_{s(X_i)}h = \max_{c\in\mathcal{C}}\prod_{i=1}^{n}c^* T_{s(X_i)}c,$$

since c and h are constrained on the same space $\mathcal{C}$. Therefore, from the definition of $R_{N,n}$, it follows that $R_{N,n}(\sigma X_1,\dots,\sigma X_n) = R_{N,n}(X_1,\dots,X_n)$.

References

Gradshteyn IS, Ryzhik IM (2015) Table of Integrals, Series, and Products, 8th ed. Academic Press, San Diego

Junghenn H (2015) A Course in Real Analysis. CRC Press, Boca Raton

Mainardi F, Mura A, Pagnini G (2010) The m-Wright function in time-fractional diffusion processes: a tutorial survey. International Journal of Differential Equations p Article: 104505

Olver FWJ, Lozier DW, Boisvert RF, et al (2010) NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge