View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universidad Industrial de Santander: Revistas Institucionales UIS

Revista de la Universidad Industrial de Santander.
***Salud UIS* vol. 52(2), abril-junio de 2020**

Nota metodológica

Benford´s Law with small sample sizes: A new exact test useful in health sciences during epidemics

Ley de Benford con muestras pequeñas: una prueba exacta nueva útil ciencias de la salud durante epidemias

José Moreno-Montoya[1]

Forma de citar: Moreno-Montoya J. Benford´s Law with small sample sizes: A new exact test useful in health sciences during epidemics. Salud UIS. 2020; 52(2): 161-163. doi: http://dx.doi.org/10.18273/revsal.v52n2-2020010

Abstract

Benford or "first digit" law has been used successfully to evaluate epidemiological surveillance systems, especially during epidemics. Conventional statistical methods for evaluation (x^2 and log-likelihood ratio) are controversial when the number of data is small (n <7). In this methodological note a new test is proposed to evaluate compliance with Benford's law with small samples, which can be used with biomedical, medical and public health data.

Keywords: Analysis of data; Epidemics; COVID-19 infection; Public health emerging infections.

Resumen

La ley de Benford o de los "primeros dígitos" ha sido usada exitosamente para evaluar los sistemas de vigilancia epidemiológica, en especial durante epidemias. Los métodos estadísticos convencionales para la evaluación (x^2 y razón de log-verosimilitud) son controversiales cuando los datos son poco (n<7). En esta nota metodológica se propone una nueva prueba para evaluar el cumplimiento de la ley de Benford con muestras pequeñas, que puede ser usada con datos de biomedicina, medicina y salud pública.

Palabras clave: Análisis de datos; Epidemia; Infección con COVID-19; Infecciones emergentes en salud pública.

In some articles published the Benford´s or first-digit Law was proposed as a cost effective tool to evaluate data in biomedicine, medicine and public health[1-6] (see **Figure 1**). Its use could be very important in sanitary emergencies as COVID-19 epidemic, when rapid evaluation of epidemiological surveillance systems require to be evaluated[7]. Since the use can be controversial when only few data are available (n<7), we developed a new exact test to screen the fulfilment of Benford distribution. Under this law, the expected number digits for sample sizes varying from n=1 to 6 are in **Table 1**. In this case we assume that data come from an independent sequence of events (i.e., the occurrence of any particular digit doesn't depend on the occurrence

1. Fundación Santa Fe de Bogotá. Bogotá, D.C., Colombia.
Correspondence: José Moreno Montoya. Address: calle 119 No. 7-74 - Piso 2, Phone number: (57 1) 6030303 ext 1130. Bogotá D.C. Colombia. Email: josemorenomontoya@gmail.com

Received: 28/02/2020 **Approved:** 04/03/2020 **Published online:** 19/03/2020

of any previous one). For this particular case, it can be used according with the known probabilities given by Benford, and using it as the scenario corresponding to the H_0: "the data are Benford's law distributed". It can be expressed with the equation:

$$H_0 : E[X_1,X_2,...,X_9|n] = \left(nLog_{10}\left(1+\frac{1}{1}\right), nLog_{10}\left(1+\frac{1}{2}\right),..., nLog_{10}\left(1+\frac{1}{9}\right)\right)$$

where

$$P[H_0] = \frac{n!}{X_1!X_2!...X_9!}\left(Log_{10}\left(1+\frac{1}{1}\right)\right)^{X_1} + \left(Log_{10}\left(1+\frac{1}{2}\right)\right)^{X_2},..., \left(Log_{10}\left(1+\frac{1}{9}\right)\right)^{X_9}$$

Naturally, data coming from no-Benford distribution are less probable to appear distributed like that, and consequently are more likely to reject the H_0. Thus, the analysis does not depend on the sample size.

To observe the performance of this test we used data with small sample sizes (n<7) from a previous publication[6]. This test was implemented in the R package, using the code: dmultinom(x = c(#1,#2,#3,#4,#5,#6,#7,#8,#9), size = NULL, prob = c(0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046)), where #i represents the absolute frequency of the correspondent digit i=1,.. 9. With this procedure we obtained exact probabilities. After, *p* values were calculated with the EMT package developed by Menzel[8] using the code: observed <- c(4,1,0,0,0,0,0,0,0) # observed data: underH0 <- c(0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046) # underH0 out <- multinomial.test(observed, underH0) # p.value, where #i represents the absolute frequency of the correspondent digit i=1,.. 9. Comparisons with Kuiper´s tests were realized[9].

With the new statistical test is possible to extend the use of Benford´s law to biomedical, medical and public health areas with small sample sizes.

References

1. Morag S, Salmon-Divon M. Characterizing human cell types and tissue origin using the Benford law. Cells. 2019;8(9): E1004. doi: http://10.3390/cells8091004
2. Pollach G, Brunkhorst F, Mipando M, Namboya F, Mndolo S, Luiz T. The "first digit law" - A hypothesis on its possible impact on medicine and development aid. Med Hypotheses. 2016; 97:102-106. doi: http://10.1016/j.mehy.2016.10.021
3. Pinilla J, López-Valcárcel BG, González-Martel C, Peiro S. Pinocchio testing in the forensic analysis of waiting lists: using public waiting list data from Finland and Spain for testing Newcomb-Benford's Law. BMJ Open. 2018;8(5):e022079. doi: http://10.1136/bmjopen-2018-022079
4. Manrique-Hernández EF, Fernández-Niño JA, Idrovo AJ. Global performance of epidemiologic surveillance of Zika virus: rapid assessment of an ongoing epidemic. Public Health. 2017;143:14-16. doi: http://10.1016/j.puhe.2016.10.023
5. Gómez-Camponovo M, Moreno J, Idrovo ÁJ, Páez M, Achkar M. Monitoring the Paraguayan epidemiological dengue surveillance system (2009-2011) using Benford's law. Biomedica. 2016;36(4):583-592. doi: http://10.7705/biomedica.v36i4.2731
6. Idrovo AJ, Fernández-Niño JA, Bojórquez-Chapela I, Moreno-Montoya J. Performance of public health surveillance systems during the influenza A(H1N1) pandemic in the Americas: testing a new method based on Benford's Law. Epidemiol Infect. 2011;139(12):1827-34. doi: http://10.1017/S095026881100015X
7. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382:727-733.
8. Menzel U. EMT. Exact multinomial test: goodness-of-fit test for discrete multivariate data. R package version 1.0; 2012.
9. Kuiper NH. Tests concerning random points on a circle. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A 1962;63:38–47.

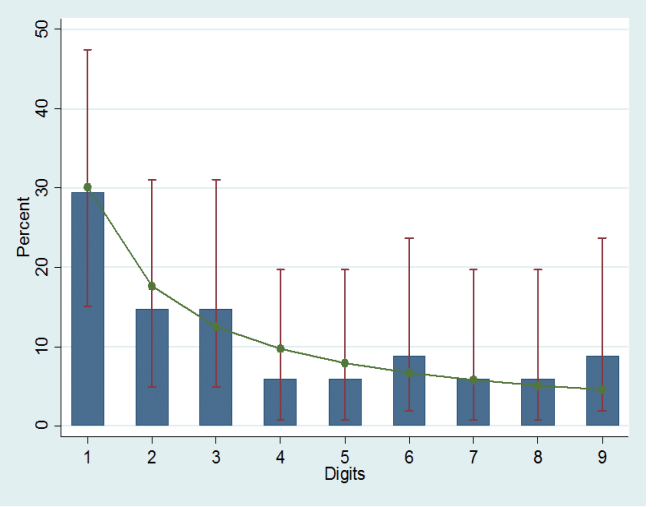

Figure 1. Fulfilment of Benford´s law of data on COVID-19 outbreak from Chinese provinces, regions and cities – situation report 17 (n=34).

Table 1. Expected occurrence of first digits following Benford distribution with small sample sizes.

Sample size	Expected first digit								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	0	0	0	0	0	0	0	0
2	1 (100%)	0	0	0	0	0	0	0	0
3	1 (50%)	1 (50%)	0	0	0	0	0	0	0
4	1 (33.33%)	1 (33.33%)	1 (33.33%)	0	0	0	0	0	0
5	2 (50%)	1 (25%)	1 (25%)	0	0	0	0	0	0
6	2 (40%)	1 (20%)	1 (20%)	1 (20%)	0	0	0	0	0

Table 2. Fulfillment of Benford distribution by American countries reporting few data during the influenza A(H1N1) outbreak (epidemiological weeks 13–47, 2009).

Country	Weeks*	First digit (%)									Kuiper	Exact multinomial†
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		
Antigua and Barbuda	3	2	1	0	0	0	0	0	0	0	<0.005	1
Saint Kitts and Nevis	3	1	2	0	0	0	0	0	0	0	<0.005	0.847
St. Vincent and Grenadines	3	3	0	0	0	0	0	0	0	0	<0.005	0.817
Surinam	4	2	1	0	0	0	0	1	0	0	<0.010	1
Belize	4	2	0	0	0	0	1	0	1	0	<0.010	0.824
Haiti	4	0	1	2	1	0	0	0	0	0	<0.010	0.344
Granada	4	4	0	0	0	0	0	0	0	0	<0.005	0.671
Santa Lucia	5	1	1	1	1	1	0	0	0	0	<0.010	0.847
Uruguay	5	3	1	1	0	0	0	0	0	0	<0.010	1
Dominica	5	2	1	0	0	0	1	0	1	0	<0.010	0.616
Guyana	5	0	2	1	0	2	0	0	0	0	<0.01	0.188

* Only weeks with report (one or more cases) to the Pan American Health Organization.
† estimated with EMT package.