Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Sobre la importancia del primer dígito en el análisis de datos

Félix Almendra Arao [a], María del Rocío Reyes Reyes [b], Marian Catherian Díaz Arias [c].
[a,b,c] *UPIITA del Instituto Politécnico Nacional, México.*

[a] falmendra@ipn.mx, [b] m.reyesr1306@gmail.com, [c] catherin.jakefin@gmail.com.

Resumen

En este trabajo se expone la importancia de conocer el comportamiento del primer dígito en bases de datos formadas por una gran cantidad de números aleatorios que usualmente incluyan datos de diferentes ordenes de magnitud, este comportamiento es conocido como Ley de Benford. Esta ley es especialmente relevante, debido a que una extensa variedad de datos ha mostrado ser consistente con ella y por ello esta ley ha sido aplicada en muy diversas áreas del conocimiento. Se dan algunos ejemplos de su uso y se describe cuando la base de datos no tiene por qué seguir dicha ley. Además, se analiza si hay un posible comportamiento irregular de tres conjuntos de tres bases de datos de México: Número de habitantes por municipio, número de casos diarios de COVID19 por estado y número de muertes diarias por COVID19 por estado. Dicho análisis se base en la conformidad o no de los datos a la Ley de Benford.

Palabra claves: Ley de Benford, desviación media absoluta, COVID19.

Introducción

Imagine que tiene un conjunto de números provenientes de algún fenómeno del mundo real, por ejemplo, el número de habitantes de todos los municipios de México (México tiene 2,469 municipios), en este conjunto de datos habrá valores muy pequeños (el más pequeño es 81, de acuerdo con el censo de 2020) y habrá otros valores muy grandes (el máximo es 1’922,523, de acuerdo con el censo mencionado).

Tales números no surgen mediante un proceso específico o sistemático, sino más bien aparecen por procesos demográficos y económicos. Debido a ello se esperaría que tales datos fueran aleatorios. Uno podría preguntarse ¿cuál proporción de tales números empiezan con el dígito 1 en comparación con cuántos empiezan con el dígito 2, 4, …, 9?

Probablemente el lector esté pensando que más o menos la misma proporción inician con 1 qué con cualquier otro dígito, sin embargo, estaría equivocado, ya que la respuesta es muy diferente: hay una mayor proporción de números cuyo primer dígito es 1, en realidad, aproximadamente 30.1% de dichos datos inician con el dígito 1 ¿sorprendente?

Dicho de otra forma, aunque uno podría esperar que en un conjunto de datos aleatorios cada uno de los dígitos del 1 al 9 apareciera como primer dígito 1/9 de las veces, es decir, aproximadamente 11.11% de las veces, frecuentemente esto no es así.

Antes de presentar la ley en cuestión, requerimos de alguna notación; primero especifiquemos lo que entenderemos por *primer dígito*, para ello recordemos la *notación científica*.

Todo número real positivo x puede representarse en la forma $x=S(x)\cdot 10^k$ donde $S(x)\in[1,10)$ y k es un número entero. La parte entera de $S(x)$ es llamada *primer dígito*.

Así por ejemplo si se tiene el número 315.7069, en notación científica se escribe como $3.157069\cdot 10^2$ y de acuerdo con lo anterior su primer dígito es 3.

La ley de Benford (LB) trata sobre la distribución del primer dígito en un conjunto de números, aunque es conocida como ley de Benford, fue Simon Newcomb (Newcomb (1881)) el primero en observar el comportamiento del primer dígito, 50 años antes de Benford.

Newcomb notó que las primeras páginas (es decir, las que empiezan con 1) de las tablas de logaritmos estaban mucho más desgastadas que las páginas al final (las que empiezan con 9).

Basado en esta observación Newcomb conjeturó que la distribución del primer dígito en una colección de números no es uniforme, las probabilidades de ocurrencia que él estableció para el primer dígito se muestran en la Tabla 1.

Tabla 1. Probabilidad de aparición del primer dígito conjeturada por Newcomb (Newcomb (1881)).

d	Probabilidad
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	0.0458

El siguiente estudio sobre la distribución del primer dígito en un conjunto de números se debe a Benford (Benford (1938)).

Una importante observación en el artículo de Benford es que, aunque algunos conjuntos de datos individuales puedan no cumplir la LB, la combinación de muchos conjuntos de datos diferentes conduce a un nuevo conjunto cuya distribución suele aproximarse a LB.

En dicho trabajo, Benford sugirió que la probabilidad de que el primer dígito sea *d* está dada por

$$P(D_1 = d) = log_{10}\left(\frac{1+d}{d}\right), d = 1,2,\ldots,9$$

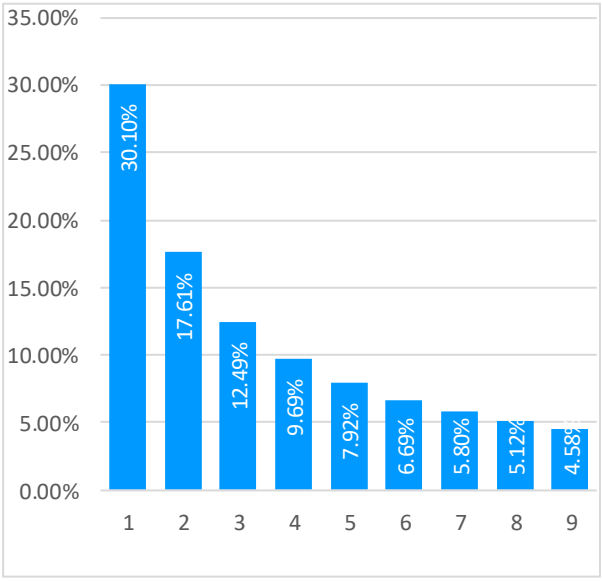

Figura 1.- Ley de Benford. Frecuencias del primer dígito.

Se ha observado que LB se cumple para conjuntos de datos que crecen exponencialmente (por ejemplo, cuando los valores se van duplicando), también parece cumplirse para muchos casos en los cuales un comportamiento exponencial no es obvio en apariencia como datos que tienen que ver con transacciones contables.

Esta ley aplica bastante bien a conjuntos de datos que incluyen diferentes ordenes de magnitud; ejemplos conocidos de datos que se sabe que siguen LB son: Transacciones de tarjetas de crédito, préstamos, poblaciones de ciudades o municipios, balances de clientes, números que aparecen en periódicos, precios de acciones, precios de inventarios, ingresos de las personas, etc. Una importante suposición para que se cumpla LB es que el conjunto de datos haya sido generado de manera aleatoria, sin alguna restricción real o artificial de ocurrencia.

Aunque se sabe que la distribución de Benford es aplicable a una gran variedad de datos, hay muchos otros conjuntos de datos que no se comportan de acuerdo con esta distribución; ejemplos de estos conjuntos de datos son aquéllos que se espera inicien con un conjunto determinado de dígitos. Por ejemplo,

alturas, pesos y temperaturas de personas, cocientes intelectuales, presiones sanguíneas. Además de aquellos datos que cubren solamente uno o dos órdenes de magnitud; otros ejemplos de conjuntos de datos que no siguen LB son los números telefónicos, números de cuentas bancarias, número de pasajeros en un avión.

En cuanto a la cantidad de datos en un análisis de LB, Nigrini (2012) sugiere que se consideren bases de datos con al menos 1000 registros. Si se analizan conjuntos de datos más pequeños deberá permitirse una mayor desviación de los datos respecto a LB.

Mediante el uso de LB se puede reconocer si un conjunto de datos podría presentar algún tipo de irregularidad, ya que un ajuste débil de los datos a LB podría ser una señal de alerta de que probablemente exista algún tipo de anomalía en ellos y en consecuencia generar sospecha de fraude.

Evaluar si los datos deben o no ajustarse a LB es un primer paso. Tal evaluación puede basarse en las consideraciones arriba mencionadas, en experiencias previas con datos de igual naturaleza o en el mismo tipo de datos de periodos previos.

Metodología

Aspectos generales

Se analizó el comportamiento de tres bases de datos:

- Número de habitantes de todos los municipios de México. INEGI, Censo de población y vivienda 2020.
- Número diario de casos de COVID-19 por estado en México entre el 28 de febrero de 2020 y el 15 de marzo de 2021. (https://datos.covid-19.conacyt.mx/#DownZCSV).
- Número diario de muertes por COVID-19 por estado en México entre 19 de marzo de 2020 y el 15 de marzo de 2021. (https://datos.covid-19.conacyt.mx/#DownZCSV).

Metodología Estadística

Para cada base de datos se realiza un primer análisis mediante graficas para comparar el comportamiento del primer dígito de cada una de las bases de datos con LB, posteriormente se realiza una valoración de conformidad a LB.

Un asunto de suma importancia, en el análisis de la conformidad de los datos a LB es seleccionar un método estadístico adecuado para valorar la significancia estadística entre los datos y los valores predichos por LB. Este no es un asunto trivial cuando se trabaja con conjuntos grandes de datos.

Por ejemplo, aunque es común que en investigaciones relativas al tema de LB se utilice la popular prueba ji-cuadrada, el uso de ésta no es recomendado cuando se trabaja con conjuntos grandes de datos, debido a su falta de robustez a grandes conjuntos de datos ya que es muy sensible a desviaciones pequeñas, ver por ejemplo Miller (2015) y Nigrini (2012), algo similar sucede con la prueba z.

Adicionalmente, Nigrini (2012) ha exhibido que la prueba de Kolmogorov-Smirnoff tampoco es adecuada en este contexto.

De esta forma, sucede que en la literatura pueden encontrarse conclusiones falsas acerca de la conformidad de datos a LB debido al uso indebido de las pruebas de ji-cuadrada, z y de Kolmogorov-Smirnoff.

En busca de una técnica adecuada para valorar la conformidad de los datos a LB, Nigrini (2012) sugiere utilizar la desviación media absoluta

$$DMA = \frac{1}{9}\sum_{i=1}^{9}|P_i - B_i|$$

donde P_i es la proporción observada y B_i es la proporción predicha por LB.

Adicionalmente, Kossovsky (2014) propone usar la suma de diferencias de cuadrados

$$SDC = 10^4 \sum_{i=1}^{9}(P_i - B_i)^2$$

Los rangos de conformidad sugeridos en el uso de la desviación media absoluta y de la

suma de diferencias de cuadrados se presentan en la Tabla 2.

Tabla 2. Rangos de conformidad para DMA y SDC correspondientes al primer dígito.

MAD	SSD	Conformidad
[0,0.006)	[0,2)	Estrecha
[0.006,0.012)	[2,25)	Aceptable
[0.012,0.015)	[25,100)	Marginal
[0.015,1]	[100,90000]	Disconformidad

Resultados y Discusión

En primer lugar, presentamos los resultados correspondientes a la base de datos del número de habitantes de los municipios de México.

En la Figura 2 se muestran las frecuencias del primer dígito en el número de habitantes de los municipios de México y de la ley de Benford para el primer dígito. En ella se observa que los porcentajes de aparición del primer dígito para los municipios se aproxima bastante bien a los predichos por la ley de Benford. Más adelante analizaremos de manera más precisa la conformidad de estos datos a LB.

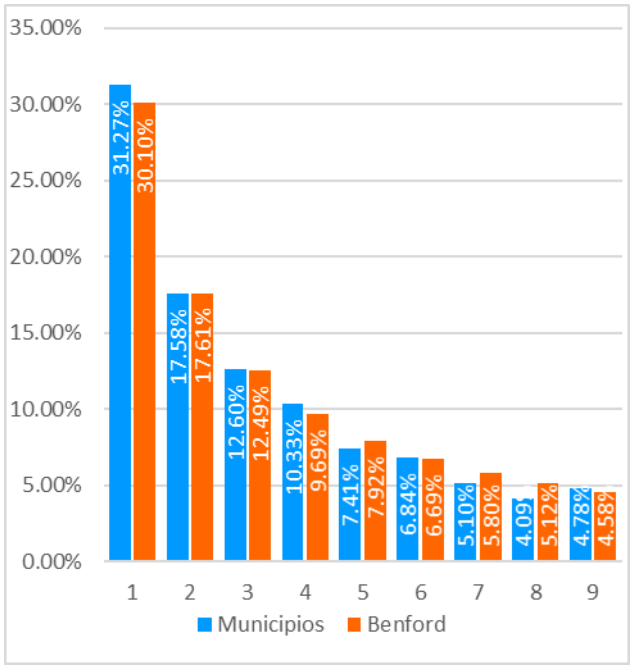

Figura 2.- Frecuencias del primer dígito en el número de habitantes de los municipios de México y de LB para el primer dígito.

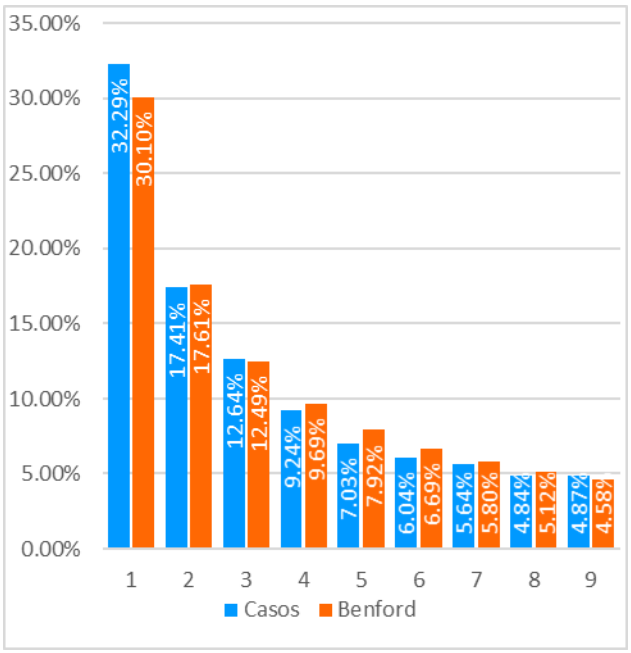

Figura 3.- Frecuencias del primer dígito en el número casos diarios de COVID19 por estados en México y LB para el primer dígito.

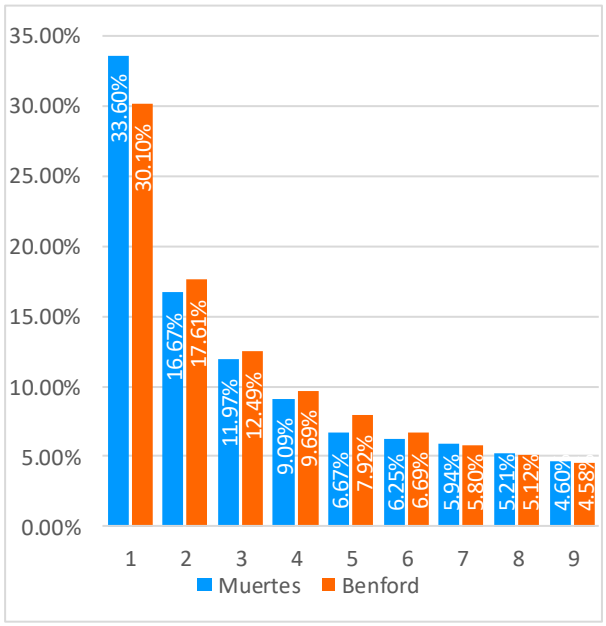

Figura 4.- Frecuencias del primer dígito en el número muertes diarias por COVID19 por estados en México y LB para el primer dígito.

En la Figura 3 se presentan las frecuencias correspondientes al número de casos diarios de

COVID19 por estados en México, en dicha figura se observa que el comportamiento de los datos es bastante cercano a LB.

En la Figura 4 se aprecia que el comportamiento de los datos de muertes diarias por COVID19 es similar a LB.

A continuación, se valora la conformidad de los datos a LB.

Tabla 3. Pruebas de conformidad para los datos de municipios de México, Casos y muertes por COVID19 en México.

Datos	**DMA** **SDC**	**Conformidad**
Municipios	0.0050 3.6280	*Estrecha* Aceptable
Casos COVID-19	0.0058 6.4412	*Estrecha* Aceptable
Muertes COVID-19	0.0083 15.5690	*Estrecha* Aceptable

De acuerdo con los resultados mostrados en la tabla 3, para las tres bases de datos analizadas se obtienen conclusiones similares. En los tres casos la prueba de conformidad DMA nos permite concluir que hay una estrecha conformidad entre los datos y LB, en tanto que la prueba SDC concluye que hay conformidad aceptable.

Así, una vez aplicadas las dos pruebas de conformidad mencionadas se puede afirmar que con ninguna de se logró descubrir algún tipo de anomalía en los datos analizados, por lo cual no hay sospecha de manipulación de datos.

Conclusiones

Se presentó la Ley de Benford y se comentaron condiciones generales bajo las cuales se espera que un conjunto de datos cumpla la dicha ley. LB puede usarse para detectar anomalías en un conjunto de datos.

Si un conjunto de datos cumple con los supuestos bajo los cuales se espera se ajusten a LB y la prueba de conformidad utilizada es adecuada para ser aplicada a LB, el hecho de que los datos no se ajusten a LB permite sospechar algún tipo de manipulación de los datos (incluidos errores de captura, duplicación de datos, etc.), situación que sugiere realizar un análisis meticuloso de los datos.

El análisis realizado en este trabajo no sugiere ningún tipo de anomalía en los datos de ninguna de las tres bases de datos examinadas, ya que, de acuerdo con el análisis de conformidad a LB realizado, en las tres bases de datos se obtuvo que tienen conformidad estrecha, de acuerdo con DMA o bien conformidad aceptable, de acuerdo con SDC.

LB es una herramienta útil la cual debe ser usada con las debidas precauciones además de que se debe tener cuidado y no utilizar pruebas de bondad de ajuste que no sean adecuadas para analizar este modelo pues ello podría conducir a conclusiones erróneas.

Referencias

Benford F. (1938). The Law of Anomalous Numbers. **Proceedings of the American Philosophical Society** 78, 551–572.

Kossovsky AE. (2014). **Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications.** World Scientific Pub. Co. Inc.

Miller S. (2015). Benford's Law. Theory and applications. Princeton University Press. Princeton, New Jersey.

Newcomb S. (1881). Note on the Frequency of Use of the Different Digits in Natural Numbers. **Amer. J. Math.** 4, 39–40.

Nigrini J.N. (2012). **Benford's Law.Applications for Forensic Accounting, Auditing, and Fraud Detection.** John Wiley & Sons, Inc. New Jersey.