Newcomb, Benford y la aristocracia del primer dígito

Andrea Burgos y Andrés Santos

Oculta en cualquier tabla de datos de la física o del mundo real rige una sutil aristocracia: los datos que empiezan por 1 resultan ser unas 7 veces más frecuentes que los que empiezan por 9. ¿Por qué las frecuencias no se reparten por igual entre todos los dígitos del 1 al 9?

La ley de Newcomb-Benford proporciona la distribución de probabilidad del primer dígito de una enorme variedad de datos (constantes físicas, medidas experimentales, poblaciones de ciudades, datos fiscales o electorales, etc.). Según esta poco intuitiva ley, el primer dígito significativo tiene una probabilidad del 30,1 % de ser 1 y del 4,6 % de ser 9. Lo que mejor caracteriza la ley de Newcomb-Benford es que debe aplicarse a datos con unidades (invariancia bajo cambio de escala). Sorprendentemente, también se aplica a datos no relacionados directamente con magnitudes físicas, como los números de Fibonacci o las potencias de 2.

Introducción

Finales del siglo XIX. Un astrónomo y matemático visita la biblioteca de su institución y consulta una tabla de logaritmos para realizar ciertos cálculos astronómicos. Como en ocasiones anteriores, le llama la atención el hecho de que las primeras páginas (las que corresponden a números que comienzan por 1) están mucho más gastadas que las últimas (correspondientes a números que comienzan por 9). Intrigado, esta vez decide no dejar pasar el asunto por alto. Cierra los ojos para concentrarse, esboza unos cuantos cálculos sobre un papel y finalmente sonríe. Ha encontrado la respuesta y es enormemente simple y elegante.

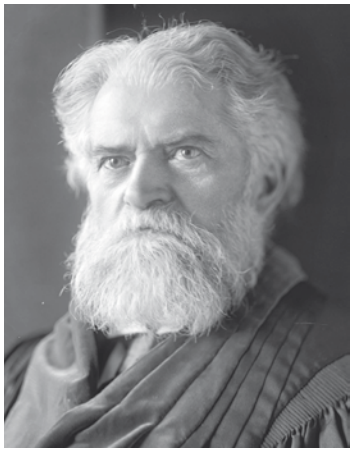

Figura 1. Simon Newcomb (1835-1909).

Figura 2. Frank Benford (1883-1948).

Algo más de medio siglo después, un físico e ingeniero eléctrico que ignora el descubrimiento de su predecesor, observa el mismo curioso fenómeno en las páginas de las tablas de logaritmos y llega a idéntica conclusión. Ambos han comprendido que, en una larga lista de registros obtenidos de la naturaleza, la fracción p_d de registros que comienzan por el dígito significativo $d = 1, 2, \ldots, 9$ no es $p_d = 1/9$, como ingenuamente podría esperarse, sino que sigue una ley logarítmica. Más concretamente,

$$p_d = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, \ldots, 9. \tag{1}$$

Los valores numéricos de p_d se muestran en la segunda columna de la Tabla I. Vemos que los registros que comienzan por 1, 2 o 3 acaparan en torno al 60 % del total, mientras que los demás 6 dígitos deben conformarse con el 40 % restante.

Nuestro personaje del siglo XIX se llama Simon Newcomb (Figura 1) y publicó su hallazgo en una modesta nota de dos páginas [1]. El segundo personaje es Frank Benford (Figura 2) y escribió un artículo de 22 páginas [2] en el que, además de justificar matemáticamente la Ec. (1), mostró su

Tabla I. Probabilidades para el primero, segundo, tercero y cuarto dígito significativo.

Dígito	Primero	Segundo	Tercero	Cuarto
d	p_d	$p_d^{(2)}$	$p_d^{(3)}$	$p_d^{(4)}$
0		0,11968	0,10178	0,10018
1	0,30103	0,11389	0,10138	0,10014
2	0,17609	0,10882	0,10097	0,10010
3	0,12494	0,10433	0,10057	0,10006
4	0,09691	0,10031	0,10018	0,10002
5	0,07918	0,09668	0,09979	0,09998
6	0,06695	0,09337	0,09940	0,09994
7	0,05799	0,09035	0,09902	0,09990
8	0,05115	0,08757	0,09864	0,09986
9	0,04576	0,08490	0,09827	0,09982

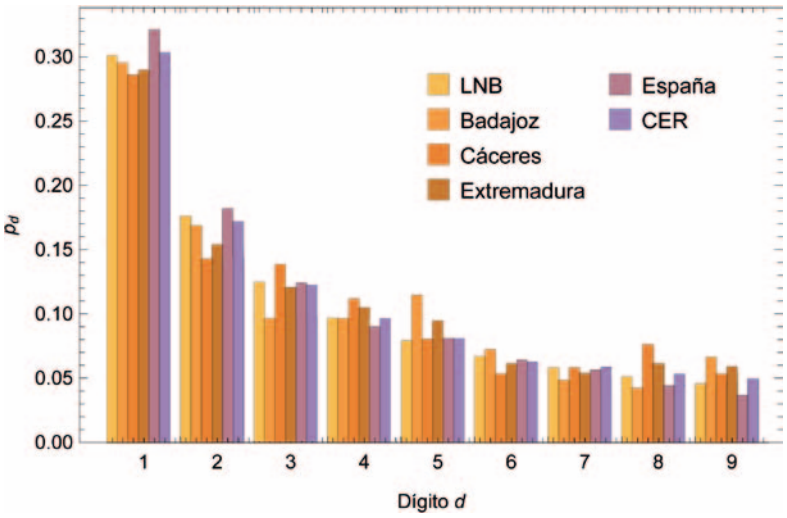

Figura 3. Comparación con la LNB de la distribución del primer dígito en las poblaciones de los municipios de las provincias de Badajoz y Cáceres, de la comunidad de Extremadura y de España, así como del censo electoral de españoles residentes en España (CER).

validez en el análisis de más de 20.000 primeros dígitos tomados de fuentes tan variadas como áreas de ríos, poblaciones de ciudades americanas, constantes físicas, pesos atómicos y moleculares, calores específicos, números extraídos de periódicos o del *Reader's Digest,* direcciones postales, etc., o las series n^{-1}, $\sqrt{n}$, n^2 o $n!$, entre otras, con n =1 – 100.

Con tan apabullante evidencia, no es de extrañar que la Ec. (1) se conozca como *ley de Benford* (o ley del primer dígito), a pesar de que fue encontrada casi sesenta años antes por Newcomb. Esta no es sino una manifestación más de la llamada ley de Stigler, según la cual ningún descubrimiento científico recibe el nombre de quien lo descubrió en primer lugar. De hecho, como el propio Stigler señala [3], la ley que lleva su nombre fue en realidad enunciada de forma parecida veintitrés años antes por el sociólogo estadounidense Robert K. Merton. A fin de no caer totalmente en la ley de Stigler, muchos autores se refieren a la Ec. (1) como *ley de Newcomb-Benford* y ese es el criterio (mediante las siglas LNB) que seguiremos en este artículo.

Invariancia bajo cambio de escala

Con frecuencia, cuando por primera vez le hablamos a un amigo, un familiar o incluso un colega acerca de la LNB, su primera reacción suele ser de escepticismo. ¿Por qué el primer dígito no se encuentra distribuido uniformemente entre los 9 valores posibles? Un argumento sencillo muestra que, de existir una ley de distribución robusta, esta no puede ser la distribución uniforme.

Imaginemos una larga lista de longitudes de ríos, alturas de montañas y superficies de países, por ejemplo. Es posible que las longitudes de los ríos estén en km, la altura de las montañas en m y las superficies de países en km^2, pero podrían también estar en millas, pies o acres, respectivamente. ¿Dependerá la distribución p_d de si utilizamos unas unidades u otras o incluso de si las mezclamos? Parece lógico que no, es decir, que la distribución p_d sea (estadísticamente) independiente de las unidades elegidas, o en otras palabras, que sea *invariante bajo cambio de escala.* La distribución uniforme $p_d = \frac{1}{9}$ obviamente no verifica esa propiedad de invariancia. Supongamos que partimos de una lista en la que todos los valores del primer dígito están igualmente representados. Si multiplicamos todos los registros de la lista por 2 [4], podemos ver que aquellos registros que empezaban antes por 1, luego empiezan por 2 o por 3; los que empezaban por 2, luego empiezan por 3 o por 4; los que empezaban por 3, luego empiezan por 6 o por 7; y los que empezaban por 4, luego empiezan por 8 o por 9. Por el contrario, todos los que empezaban por 5, 6, 7, 8 o 9 empezarán ahora por 1. Por consiguiente, si $p_d = \frac{1}{9}$ inicialmente, entonces $p_1 = \frac{5}{9}$ y $p_2 + p_3 = p_4 + p_5 = p_6 + p_7 = p_8 + p_9 = \frac{1}{9}$ tras multiplicar por 2 todos los registros, destruyéndose así la uniformidad inicial. Podemos continuar multiplicando por 2 y la distribución continuará evolucionando hasta alcanzarse una distribución estacionaria e invariante bajo ese cambio de escala [5].

Así pues, el sello más identificativo de la LNB es que debe ser de aplicación a registros que tienen unidades, de modo que la ley es invariante bajo cambio de escala, lo que permite deducir fácilmente la Ec. (1) [5]. Y no solo eso, sino que es posible generalizar la Ec. (1) más allá del primer dígito para así obtener la probabilidad $p_{d_1, d_2, \ldots, d_m}$ de que los m primeros dígitos coincidan con una cierta cadena ordenada $(d_1, d_2, \ldots, d_m)$, donde $d_1 \in \{1, 2, \ldots, 9\}$ y $d_i \in \{0, 1, 2, \ldots, 9\}$ si $i \geq 2$. Por ejemplo, la probabilidad de que los tres primeros dígitos de un registro formen precisamente la cadena (3, 1, 4) es $p_{3,1,4} = \log_{10}(1 + 1/314) = 0{,}00138$. A partir de $p_{d_1, d_2, \ldots, d_m}$, podemos calcular la probabilidad $p_d^{(m)}$ de que el m-simo dígito sea d, independientemente de los valores de los $m - 1$ dígitos anteriores, sumando para todos los valores posibles de esos $m - 1$ dígitos anteriores. En la Tabla I, la ley del primer dígito, p_d, está acompañada de las leyes para el segundo, tercero y cuarto dígitos. Como puede observarse, a medida que el dígito es más interno, la probabilidad se hace menos dispar.

Aplicaciones y ejemplos

Las aplicaciones y verificaciones de la LNB son numerosas y abarcan temas tan variados y prosaicos como el estudio del genoma, la vida media de los núcleos inestables, física de partículas, astronomía, fenómenos críticos cuánticos, emisiones tóxicas, auditorías fiscales, fraudes electorales o científicos, producto interior bruto, mercado bursátil, datos de inflación, *world wide web,* actividades religiosas, fechas de nacimiento, caudales de ríos, o incluso la COVID-19. Otros ejemplos pueden verse en el enlace [6]. En esta sección presentaremos algunos ejemplos adicionales.

Comencemos con una de las situaciones que el propio Benford estudió en su clásico artículo [2]: el de las poblaciones de ciudades. Utilizando los datos del Instituto Nacional de Estadística (INE), hemos considerado la población (en 2019) de los 165 municipios de la provincia de Badajoz (más la población total de la provincia de Badajoz), de los 223 municipios de la provincia de Cáceres (más la población total de la provincia de Cáceres) y el conjunto de los 388 municipios de la comunidad de Extremadura (más las poblaciones totales de las provincias de Badajoz y Cáceres). También hemos considerado la población (según el padrón de 2016) de los 8.110 municipios españoles, así como los 8.184 datos del censo electoral de españoles residentes en España (CER, agosto de 2020) correspondientes al número de electores por municipio de inscripción (más el total de los electores de cada comunidad y el total nacional). Con todas estas listas de registros hemos analizado la frecuencia de aquellos que empiezan por d = 1, 2, ..., 9 y los resultados se comparan en la Figura 3. Se observa un buen acuerdo general entre los datos de poblaciones (especialmente en el caso del CER) y las predicciones de la LNB. Esto es interesante, ya que no es obvio que la distribución de los significandos del número de habitantes de municipios deba ser invariante bajo cambio de escala.

Pasemos ahora a dos ejemplos relacionados con la astronomía. En el primero de ellos, tomamos la distancia a la Tierra (en años-luz y en párseс) de las 300 estrellas más brillantes [7]. En el segundo caso, los datos considerados corresponden al número diario de manchas solares desde 1818 hasta la actualidad [8]. Como se observa en la Figura 4, las distancias entre nuestro planeta y las estrellas siguen generalmente bien la LNB, a pesar de que la lista no es excesivamente extensa (solo 300 datos) y de que hay desviaciones "locales" (por ejemplo, $p_6 < p_7$ en las dos elecciones de unidades). Este buen acuerdo general era de esperar, ya que la distribución de dígitos en distancias (que se expresan en unidades) es un claro ejemplo de invariancia bajo cambio de escala. Sin embargo, en el caso del número diario de manchas solares se observan diferencias cuantitativas (aunque no cualitativas) con la LNB, sobre todo en los casos d =1, 3, 4 y 5. Conviene tener en cuenta que, aunque la serie es muy larga (más de 59.000 registros, una vez excluidos los días sin datos o con 0 manchas), cada registro solo tiene una, dos o tres cifras (el número máximo de manchas solares es de 528 y corresponde al 26 de agosto de 1870).

Por último, hemos analizado los precios de 1.016 artículos de una cadena de marca de moda [9] y de 1.373 productos de una red de hipermercados [10]. Los resultados se muestran en la Figura 5. En este caso, las discrepancias con la LNB son más acusadas. Aunque las frecuencias mayores se

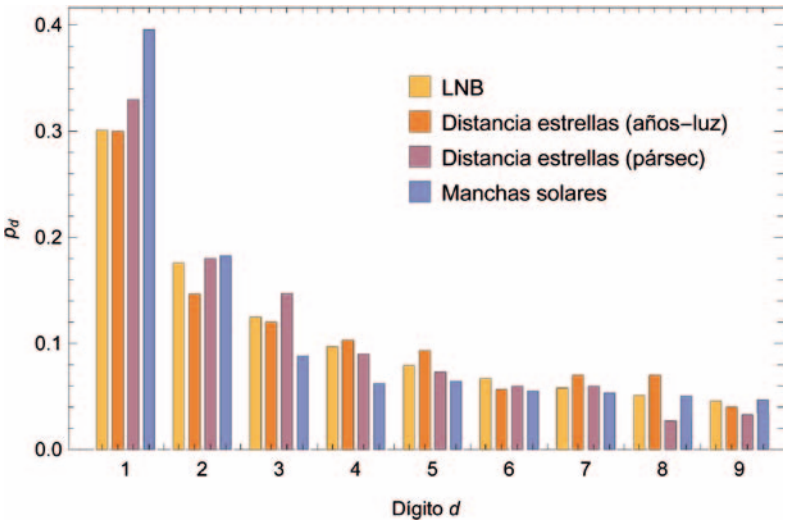

Figura 4. Comparación con la LNB de la distribución del primer dígito en la distancia a la Tierra (en años-luz y en párseс) de las 300 estrellas más brillantes y en el número diario de manchas solares.

presentan para d = 1 y d = 2, los valores observados de p_d no disminuyen monótonamente al aumentar d. En el caso de la cadena de marca de moda, se tiene que $p_4 > p_3$ y $p_9 > p_8 > p_6 > p_7$; en los precios de la red de hipermercados, $p_8 > p_9 > p_7 > p_6$. En principio, podría pensarse que, puesto que pueden expresarse en euros, dólares, rublos, yenes, etc., los precios debieran satisfacer la propiedad de invariancia bajo cambio de escala inherente a la LNB. Sin embargo, a esa invariancia hay que superponer las estrategias comerciales y artificiales de asignación de precios, lo que genera desviaciones relevantes respecto de la LNB.

Comentarios finales

Esperamos que este artículo haya contribuido a mostrar que, en contra de lo que en un primer momento pudiera pensarse, el primer dígito significativo de un conjunto de datos extraídos de la naturaleza o del mundo real no está distribuido de manera uniforme entre los nueve posibles valores (d = 1, 2, ..., 9), sino que típicamente la frecuencia es mayor para d = 1 y va disminuyen-

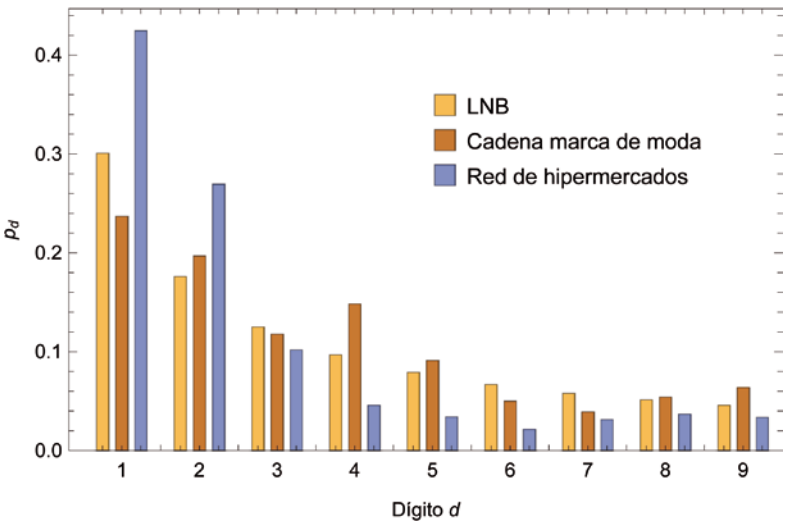

Figura 5. Comparación con la LNB de la distribución del primer dígito en los precios de artículos de una cadena de marca de moda y de una red de hipermercados.

do a medida que aumenta *d*. La LNB (1) da una expresión matemática a ese hecho empírico, aunque no siempre tiene por qué verificarse de modo riguroso. Sí es de esperar que, salvo las inevitables fluctuaciones estadísticas, la ley se cumpla en conjuntos de datos acompañados de unidades (como sucede generalmente en física), de modo que la distribución del primer dígito sea independiente de las unidades escogidas (invariancia bajo cambio de escala). Más en general, la LNB se satisface cuando la mantisa de los logaritmos (en cualquier base) de los datos considerados está distribuida uniformemente. Eso hace que listas tan poco relacionadas en principio con magnitudes físicas como la de los números de Fibonacci o las potencias de 2 también verifiquen la LNB. Además, si una lista inicial de datos no cumple la ley, la multiplicación iterada de los datos por una potencia irracional de 10 lleva a que la distribución del primer dígito en las listas resultantes converja hacia la LNB [5].

Hasta los años setenta del pasado siglo (que es cuando empezaron a usarse las calculadoras científicas de bolsillo) los físicos utilizaban las tablas de logaritmos (o su aplicación en las reglas de cálculo) para pequeños cálculos científicos cotidianos, aunque si los cálculos eran más complicados podían utilizar programas de ordenador en las computadoras de la época. En la actualidad, esos cálculos se realizan en calculadoras de bolsillo, en teléfonos móviles o en los ordenadores personales con la amplia variedad de programas matemáticos existentes. Como los datos que se manipulan en física están extraídos de situaciones "reales", tales como experimentos, modelos, constantes físicas, etc., podemos concluir, como homenaje a Newcomb y Benford y sus tablas de logaritmos, que la tecla del 1 será la que presente un mayor desgaste y la del 9 será la menos utilizada.

Referencias

[1] S. Newcomb, "Note on the Frecuency of Use of the Different Digits in Natural Numbers", *Am. J. Math.* **4**, 39-40(1881).

[2] F. Benford,"The Law of the Anomalous Numbers", *Proc. Am. Philos. Soc.* **78**, 551-572 (1938).

[3] S. M. Stigler, "Stigler's Law of Eponimy", *Trans. N. Y. Acad. Sci.* **39**, 147-158 (1989).

[4] J. M. R. Parrondo, "La misteriosa ley del primer dígito", *Investigación y Ciencia* 315 (diciembre), 84-85 (2002).

[5] A. Burgos y A. Santos,"The Newcomb–Benford Law: Do Physicists Use More Frequently the Key 1 than the Key 9?", https://arxiv.org/abs/2101.12068

[6] "Testing Benford's Law", https://testingbenfordslaw.com

[7] "The brightest stars", http://www.atlasoftheuniverse.com/stars.html

[8] "Sunspot number", http://sidc.be/silso/datafiles.

[9] https://cortefiel.com/es/es/mujer?srule=price-high-to-low&

[10] https://www.hipercor.es/supermercado/alimentacion/

Andrea Burgos
Graduada en Física,
Universidad de Extremadura

Andrés Santos
Dpto. de Física, Universidad
de Extremadura