Publicaciones Matemáticas del Uruguay
Volumen XX, Xxxx XXXX, Páginas xxx–xxx
S 0797-1443

LEY DE BENFORD

PROF. MARÍA CAPUTI ZUNINI

1. Resumen

La Ley de Benford establece, contrariamente a la intuición, que, en algunos conjuntos de datos numéricos, la frecuencia de aparición del primer dígito significativo no es uniforme. La frecuencia con la que aparece cada dígito sigue una proporción particular que se explicita en la denominada por Benford [1] en 1938 como "Ley de los números anómalos". El 1 aparece como primer dígito significativo un 30,1 % de las veces, el 2 un 17,6 %, el 3 un 12,5 %, el 4 un 9,7 %, el 5 un 7,9 %, el 6 un 6,7 %, el 7 un 5,8 %, el 8 un 5,1 % y el 9 un 4,6 %, aproximadamente. Esta ley permaneció como una mera "curiosidad estadística" por varias décadas. En 1992 fue catapultada a la luz e interés público por el contador norteamericano Mark Nigrini, quien en su tesis de doctorado la utilizó para detectar fraudes en las declaraciones fiscales.

En este trabajo se presenta muy sintéticamente la Ley y se resumen algunos métodos estadísticos desarrollados en mi trabajo final del Diploma en Matemática mención Aplicaciones (ANEP-UdelaR)[2] para analizar si un conjunto de datos sigue o no la Ley de Benford. Se analiza el cumplimiento de la Ley de Benford en los resultados del censo realizado en Uruguay en 2011.

2. Palabras clave

Ley de Benford; Primer dígito significativo.

3. Antecedentes históricos

La primera publicación sobre la que posteriormente sería llamada "Ley de Benford" fue autoría del astrónomo norteamericano Simon Newcomb y data de 1881. En un artículo de dos páginas titulado "Note on the Frequency of Use of the Different Digits in Natural Numbers" publicado en American Journal of Mathematics el autor describe sus observaciones en el uso de las tablas de logaritmos. Newcomb detecta que las hojas más desgastadas eran las primeras. El artículo de Newcomb pasa desapercibido por la comunidad matemática de la época, quedando como una mera curiosidad.

Frank Benford en 1938, aparentemente sin conocer el artículo de Newcomb, realiza la misma observación en las tablas de logaritmos. Éstas eran muy utilizadas para realizar rápidamente multiplicaciones y divisiones antes de la era de las calculadoras. Benford tenía contacto frecuente con dichas tablas, ya que fue un físico que trabajó en el centro de investigaciones de General Electric en Nueva York. Para comprobar empíricamente la llamada por Benford como "Ley de los números anómalos" recolectó 20.229 números extraídos de las más diversas fuentes (entre

otros: periódicos, revistas, estadísticas de béisbol, longitudes de ríos, tablas de constantes matemáticas y físicas) y sin ayuda de calculadoras o procesadores de datos realizó los cálculos de las proporciones en las que aparecían los primeros 9 números enteros positivos. Esos datos los recogió en una tabla que fue retomada en Nigrini ([6], pág. 4) quien afirma que la misma contiene algunos errores, entre otros de redondeo. La publicación de Benford en Proceedings of the American Philosophical Society tuvo una repercusión mucho mayor que la de Newcomb, según Hill ([4], pág. 359), debido a estar previo al artículo de H. A. Bethe, M. E. Rose y L. P. Smith, titulado "The Multiple Scattering of Electrons", que luego sería muy renombrado. Habiendo pasado desapercibida la publicación de Newcomb, la bautizada por Benford como "Ley de los números anómalos" sería conocida como "Ley de Benford".

Mark Nigrini, un contador estadounidense, se encuentra por primera vez con la Ley de Benford en un curso de doctorado en 1989 en la Universidad de Cincinnati (EEUU). Según Nigrini [6] interesado por el tópico y tras leer muchos de los artículos que había publicados en esa materia decide contactarse con Ralph A Raimi, quien era el autor más citado en la temática del momento. Raimi consideraba que la ley no tenía fines prácticos. En ese momento, Nigrini [6] si bien vislumbraba que podía usarse la ley para la detección de anomalías en las declaraciones de impuestos, descarta la aplicación práctica por no tener los medios informáticos necesarios para poder procesar "gran cantidad" de datos. Sin embargo, a principios de 1991, Nigrini con ayuda de John Byant decide buscar un modelo matemático de cómo los contribuyentes engañan al fisco y de cómo la Ley de Benford puede ayudar a detectarlos. Esta idea sería la semilla de la tesis de doctorado de Nigrini publicada en 1992 y de muchas otras publicaciones del autor sobre el tópico. La tesis de Nigrini es la que lanza a la fama a la Ley de Benford. A partir de allí según Joseph T. Wells (autor del prólogo de Nigrini [6]) se elaboraron otras aplicaciones de la Ley de Benford.

Desde la publicación de Benford fueron numerosos los intentos por dar una prueba matemática rigurosa de la Ley. Pero, ésta recién fue conseguida por Hill en 1995.

4. Ley de Benford o ley del primer dígito significativo

4.1. La Ley de Benford como distribución de probabilidad discreta. Una primera aproximación. La Ley de Benford establece una distribución para los primeros dígitos significativos de un número real positivo. Siendo el primer dígito significativo de un real positivo el dígito distinto de cero que aparece más a la izquierda en su expresión decimal. Por ejemplo, el primer dígito significativo de 73,15 es 7, el de 0,045 es 4 y el de e es 2. Se observa que 0 no es considerado como primer dígito significativo. Al primer dígito significativo de un número real lo notaremos D_1.

Una primera descripción de la Ley de Benford como una ley de probabilidad podría tabularse como sigue.

Dígito d	1	2	3	4	5	6	7	8	9
P($D_1 = d$)	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Con más formalidad podemos enunciar la Ley: $P(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right)$ con $d = 1 \ldots 9$.

Así definida la Ley es una distribución de probabilidad discreta. La Ley verifica las siguientes proposiciones: $P(D_1 = 1) = P(D_1 = 2) + P(D_1 = 3)$ y $P(D_1 = 1) = \sum_{i=5}^{9} P(D_1 = i)$.

5. Ley general de probabilidad para los primeros k dígitos significativos de un número en base 10

Hill [3] propone una Ley general de probabilidad para los primeros k dígitos significativos de un número. Análogamente a como se definió el primer dígito significativo de un real positivo pueden definirse las funciones $D_1{:}R^+ \longrightarrow \{1,2,3,4,5,6,7,8,9\}$ y $D_i{:}R^+ \longrightarrow \{0,1,2,3,4,5,6,7,8,9\}$ para $i \geq 2$ tal que D_i hace corresponder a un real positivo su i-ésimo dígito significativo en base 10.

Definición 5.1. Ley general para los primeros k dígitos significativos de un número real positivo en base 10. Dado $k \in Z^+$, para todo d_1 con $d_1 \in \{1,2,3,4,5,6,7,8,9\}$, y para todos d_j con $d_j \in \{0,1,2,3,4,5,6,7,8,9\}$, $j = 2 \ldots k$ se define:

$$P\left(\bigcap_{i=1}^{i=k}\{D_i = d_i\}\right) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} d_i 10^{k-i}}\right)$$

.

Se observa que esta Ley incluye como caso particular a la Ley de Benford para el primer dígito significativo. Se puede probar que a partir de esta Ley que dado un natural n, $n \geq 2$ y siendo D_n el enésimo dígito significativo de un real positivo que $P(\{D_n = d\}) = \sum_{i=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(1 + \frac{1}{10i+d}\right)$ con $d = 0 \ldots 9$.

Otro corolario que se puede deducir de la Ley general de probabilidad para los primeros k dígitos significativos es que existe una dependencia entre los dígitos significativos de un real positivo. Por ejemplo, la probabilidad de que el segundo dígito significativo sea 2 ($P(D_2 = 2) \cong 0,109$), es diferente a la probabilidad de que el segundo dígito significativo sea 2 sabiendo que el primer dígito significativo es 1 ($P(D_1 = 1, D_2 = 2/D_1 = 1) \cong 0,115$), y es a su vez diferente a la probabilidad de que el segundo dígito significativo sea 2 sabiendo que el primer dígito significativo es 2 ($P(D_1 = 2, D_2 = 2/D_1 = 2) \cong 0,110$). Esta dependencia se hace más débil a medida que crece la distancia entre los dígitos. Es decir, puede demostrarse, por ejemplo, que $(a_n) : a_n = P(\{D_1 = 1, D_n = 1\})$ es decreciente para $n \to +\infty$.

Otras implicancias son que $P(\{D_1 = i, D_n = j\})$ tiende a $P(\{D_n = j\})$, para $n \to +\infty$, con $i \in \{1,2,3,4,5,6,7,8,9\}$, y $j \in \{0,1,2,3,4,5,6,7,8,9\}$. Además, se verifica que la distribución del enésimo dígito tiende a la distribución uniforme a medida que n crece. Se ofrecerá al lector una prueba de la última afirmación (para el caso $d = 0$) que fuera enunciada por Hill en varias oportunidades pero cuya demostración no pudo ser encontrada en la bibliografía. Nigrini ([6], pág.6) plantea que en conjuntos de datos con tres o más cifras significativas, a partir del cuarto dígito significativo el error con la distribución uniforme es despreciable a fines prácticos.

Teorema 5.2. *Distribución del enésimo dígito tiene a la distribución uniforme para $n \to +\infty$. Dado un natural n, $n \geq 2$ y siendo D_n el enésimo dígito significativo de un real positivo. Entonces, $P(\{D_n = 0\}) \to \frac{1}{10}$, $n \to +\infty$.*

Demostración. $P(\{D_n = 0\}) = \sum_{i=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(1 + \frac{1}{10i}\right) =$

$= \frac{1}{\log_e 10} \sum_{i=10^{n-2}}^{10^{n-1}-1} \log_e\left(1 + \frac{1}{10i}\right).$

No es difícil ver que $\sum_{i=10^{n-2}}^{10^{n-1}-1} \log_e\left(1 + \frac{1}{10i}\right) \sim \sum_{i=10^{n-2}}^{10^{n-1}-1} \frac{1}{10i}$ para $n \to +\infty$.

Sea $\mu = \frac{1}{\log_e 10}$. Entonces, $P(\{D_n = 0\}) \sim \mu \sum_{i=10^{n-2}}^{10^{n-1}-1} \frac{1}{10i} =$

$= \mu \sum_{i=10^{n-2}}^{10^{n-1}-1} \frac{1}{i}$ para $n \to +\infty$.

Como $\log_e m \sim \sum_{i=1}^{m} \frac{1}{i}$, para $m \to +\infty$.

Entonces, $P(\{D_n = 0\}) \sim \mu\left(\sum_{i=1}^{10^{n-1}-1} \frac{1}{i} - \sum_{i=1}^{10^{n-2}-1} \frac{1}{i}\right) \sim$

$\sim \mu\left[\log_e(10^{n-1} - 1) - \log_e(10^{n-2} - 1)\right] = \frac{\mu}{10}\left[\log_e \frac{10^{n-1} - 1}{10^{n-2} - 1}\right] \sim$

$\sim \frac{\mu}{10}\left[\log_e \frac{10^{n-1}}{10^{n-2}}\right] = \frac{\mu}{10} \log_e 10 = \frac{1}{10}.$ □

6. Aplicaciones de la Ley de Benford

El uso de la Ley de Benford para detectar fraude o falsificación de datos en las declaraciones de impuestos fue abordado por el contador norteamericano Mark Nigrini en su tesis de doctorado. Nigrini ha recogido una extensa cantidad de evidencia empírica de la ocurrencia de la Ley de Benford en muchas áreas de la contabilidad y de la demografía y concluyó que en una amplia y variada cantidad de situaciones contables los datos reales siguen la Ley de Benford. Sin embargo, al falsear datos estos deben inventarse, lo que puede hacerse tomando números uniformemente generados o mediante una elección personal del falseador. Nigrini ha diseñado varios tests para medir la conformidad de un conjunto de datos con la Ley de Benford.

6.1. Ley de Benford para los primeros dos dígitos. Nigrini ([6], pág.15-19) bajo el título "Love at first sight" explica porqué usar solo la Ley de Benford para el primer dígito a veces no es suficiente al analizar un conjunto de datos de dos o más dígitos sigue o no la Ley. Allí, construye un contraejemplo en el que podría concluirse que el conjunto de datos sigue la Ley de Benford para el primer dígito, pero en cambio se evidencia la no conformidad al considerar el análisis con los primeros dos dígitos.

Sin embargo, Nigrini ([6], pág. 20) sugiere no comparar una distribución con la Ley de Benford para los primeros dos dígitos en conjuntos de menos de trescientos datos. Para esos casos sugiere utilizar la Ley de Benford para el primer dígito.

6.2. ¿Cómo comprobar que un conjunto de datos sigue la Ley de Benford?

6.2.1. Chi Cuadrado. El test de bondad de ajuste de Chi Cuadrado χ^2 permite comprobar si ciertos datos siguen una cierta distribución de probabilidad con un cierto error α .

El estadístico Chi Cuadrado depende de la cantidad de datos. Nigrini ([6], pág 154) afirma que el test de Chi Cuadrado sufre de un "exceso de poder" cuando la cantidad de datos es muy grande. Es decir, que casi siempre será mayor que

su valor crítico, lo que implicaría que los datos no conforman la Ley de Benford aunque las diferencias sean mínimas. Afirma que se nota este problema cuando los conjuntos tienen más de 5000 datos y que con más de 25000 datos se requeriría casi la perfección para que pasaran dicho test.

6.2.2. Mean Absolute Deviation Test (MAD Test). Nigrini ([6], pág. 158) afirma que para desentenderse del problema de la cantidad de datos (que en las auditorías de fraude fiscal suelen ser muchos) se debe utilizar un estadístico que no dependa de ese parámetro. Propone utilizar el Mean Absolute Deviation Test (MAD Test), definiendo el estadístico $MAD = \dfrac{\sum_{j=1}^{K} |o_j - p_j|}{K}$, siendo K la cantidad de categorías que toma la variable, o_j la frecuencia relativa observada y p_j la probabilidad esperada para la distribución de probabilidad. Cuanto mayor es MAD, mayor es la diferencia promedio entre la frecuencia relativa observada y la probabilidad. En el caso de comparar una distribución de datos con la Ley de Benford para el primer dígito K tomará el valor 9, con la Ley de Benford para el primer y segundo dígito K será 90.

A diferencia del Test de Chi Cuadrado no hay una grilla con valores críticos objetivos según un cierto error u otros parámetros. En Nigrini ([6], pág. 160) aparece una tabla con ciertos valores críticos para decidir la conformidad o no conformidad de un conjunto de datos con la Ley de Benford. La misma fue desarrollada en base a las experiencias empíricas de Nigrini en el trabajo diario con este tópico en diferentes áreas.

Dicha tabla proporciona valores críticos para la comparación con la Ley de Benford para el primer dígito significativo, para el segundo dígito significativo, y para los primeros dos dígitos significativos. En algunas ocasiones las conclusiones pueden diferir según el Test MAD utilizado sea el del primer dígito, el segundo o el de los primeros dos dígitos. Allí, será necesario entonces, buscar explicaciones de acuerdo a las características particulares del conjunto de datos. Con respecto a esto, Nigrini sugiere una posición más bien pesimista en casos de resultados encontrados.

Nigrini ([6], pág.170) concluye que el test de Chi Cuadrado funcionará bien en conjuntos con una cantidad pequeña de datos. Para otros conjuntos con mayor cantidad de datos sugiere el uso del Test MAD siguiendo los parámetros dados por los valores críticos que aparecen en la tabla citada anteriormente.

6.2.3. Últimos dos dígitos. Como se demostró en el teorema 5.2 la frecuencia de los dígitos tiende a uniformizarse a medida que avanzamos hacia la derecha en la posición del dígito. Tomando esto como base, Nigrini ([6], pág.129) afirma para propósitos prácticos a partir del tercer dígito significativo la probabilidad de los dígitos es uniforme. Utilizando esto crea el Test de los últimos dos dígitos que servirá para detectar invención de números en determinadas situaciones. El Test de los últimos dos dígitos consiste en comparar la distribución empírica con la distribución uniforme utilizando el Test de bondad de ajuste de Chi Cuadrado.

¿Cuáles son los últimos dos dígitos de un número? En general, esta pregunta carece de sentido, pero lo que importa en el análisis forense de resultados es qué dígitos serían los apropiados para este análisis. Según Nigrini ([6], pág.129) para cuestiones que involucren dinero los centavos serían los apropiados candidatos a últimos dos dígitos, y que, para cuestiones en donde puede darse invención o fraude con números naturales las decenas y las unidades serán los apropiados.

El Test de los dos últimos dígitos es utilizado en datos donde se buscan signos de invención de datos. Por ejemplo, censos poblacionales, resultados electorales, inventarios, números en las deducciones de impuestos.

7. Análisis del cumplimiento de la Ley de Benford: censo de la población uruguaya en 2011

En esta sección se propone estudiar la distribución de la población total de las localidades uruguayas con más de 1.000 habitantes según el censo realizado por el Instituto Nacional de Estadística de Uruguay [5] en 2011. Dicho censo tuvo la particularidad que, a diferencia de los anteriores, no fue realizado todo un mismo día si no que llevó varios meses en concluirse. Cabe preguntarse si los resultados del mismo serán válidos. La Ley de Benford provee una estrategia para analizar los datos recogidos.

A continuación, aparece una tabla en donde se registran las frecuencias del primer dígito significativo del total de población de las 176 diferentes localidades uruguayas de más de 1.000 habitantes.

Dígito	1	2	3	4	5	6	7	8	9
Frecuencia	70	37	21	9	7	10	9	6	7

Al aplicar el test de Bondad de ajuste de Chi-cuadrado, con 8 grados de libertad y $\alpha = 0,05$, se obtiene $\chi^2 = 15,48 < 15,51$. Por lo tanto, es aceptada la hipótesis de que este conjunto de datos satisface la Ley de Benford para el primer dígito significativo.

Al aplicar el Test MAD para el primer dígito significativo obtenemos el valor 0,029 lo que describe una no conformidad de la cantidad de habitantes de las localidades uruguayas con más de mil habitantes con la Ley de Benford para el primer dígito significativo. Esto podría explicarse debido a que el conjunto de datos es pequeño o tal vez se deba a otras razones como invención de datos.

Podemos utilizar el Test de los últimos dos dígitos para analizar si hay signos de invención en los datos. Consideraremos los últimos dos dígitos como las decenas y unidades de los datos. Al aplicar el test de Bondad de ajuste de Chi-cuadrado, con 99 grados de libertad y $\alpha = 0,05$, se obtiene $\chi^2 = 98,12,48 < 124,34$. Por lo tanto, es aceptada la hipótesis de que la distribución de los últimos dos dígitos satisface la distribución uniforme.

8. Bibliografía y referencias

[1] Benford, F. (1938) The law of anomalous numbers, Proc. Amer. Phil. Soc. 78, 551-572.

[2] Caputi, M (2016) Ley de Benford (Tesis de Diploma). ANEP-UdelaR, Montevideo, Uruguay.

[3] Hill, TP (1995). A Statistical Derivation of the Significant-Digit Law. Statistical Science 10(4), 354-363.

[4] Hill, TP (1998). The First-Digit Phenomenon. American Scientist 86 (4), 358-363.

[5] Instituto Nacional de Estadística. (s.f.). Instituto Nacional de Estadística Uruguay. Recuperado el 14 de Septiembre de 2015, de http://www.ine.gub.uy/

[6] Nigrini, M. (2012). Benford's Law Applications for Forensic Accounting, Auditing, and Fraud Detection. Wiley.