La Ley de Newcomb-Benford y sus aplicaciones al Referendum Revocatorio en Venezuela.
Reporte Técnico no-definitivo 2a. versión: Octubre 01,2004

Luis Raúl Pericchi(1,2) y David Torres(1)
(1) Universidad de Puerto Rico, (2) Universidad Simón Bolívar.

Introducción El siguiente vector es una pequeña muestra delos VOTOS NO en 10 máquinas de votación escogidas al azar, en el Referendum Revocatorio del 15 de Agosto del 2004 (RR-15AGO):

$$[303, 451, 213, 456, 234, 123, 222, 411, 122, 176]$$

El correspondiente vector de primeros dígitos significativos del vector es

$$[3, 4, 2, 4, 2, 1, 2, 4, 1, 1]$$

Entonces la frecuencia con que aparece como primer dígito el 1 en el vector escogido al azar es de 3/10, del 2 es 3/10, del 3 es 1/10 etc.
Que ocurre cuando en lugar de tener el vector un tamaño de $N = 10$, tiene un $N = 19064$ máquinas? Es sorprendente que la frecuencia con la que aparecen los digitos en los numeros que típicamente nos encontramos en la realidad NO siguen la distribución Uniforme como esperaríamos.
Por ejemplo la frecuencia con la que aparecen los primeros dígitos del 1 al 9 no es el esperado

$$Prob(1er\, digito = d) = 1/9; \quad d = 1, \ldots, 9.$$

Las frecuencias del segundo dígito significativo tampoco son

$$Prob(2do\, digito\, significativo = d) = 1/10; \quad d = 0, 1, \ldots, 9.$$

Las fórmulas para las frecuencias del primer y el segundo dígito significativo son:

$$Prob(1er\, digito = d) = \log_{10}(1 + d^{-1}); \quad d = 1, \ldots, 9 \tag{1}$$

$$Prob(2do.digito = d) = \sum_{k=1}^{9} \log_{10}(1 + (10k + d)^{-1}); \quad d = 0, 1, \ldots, 9. \tag{2}$$

Estas expresiones fueron propuestas por el matemático y astrónomo norteamericano Newcomb y generalizadas y popularizadas por Benford.
Se cumple esta Ley?. Se sabe ha sido observada en contextos tan diversas como áreas de ríos, estadíisticas de Baseball, constantes físicas, poblaciones en diferentes municipios o pagos de impuestos sobre la renta. Esto no quiere decir que cualquier conjunto de números debe cumplirlo (por ejemplo los números de telefono de un municipio), pero una amplia gama de casos reales se han encontrado que se ajustan admirablemente bien. Se

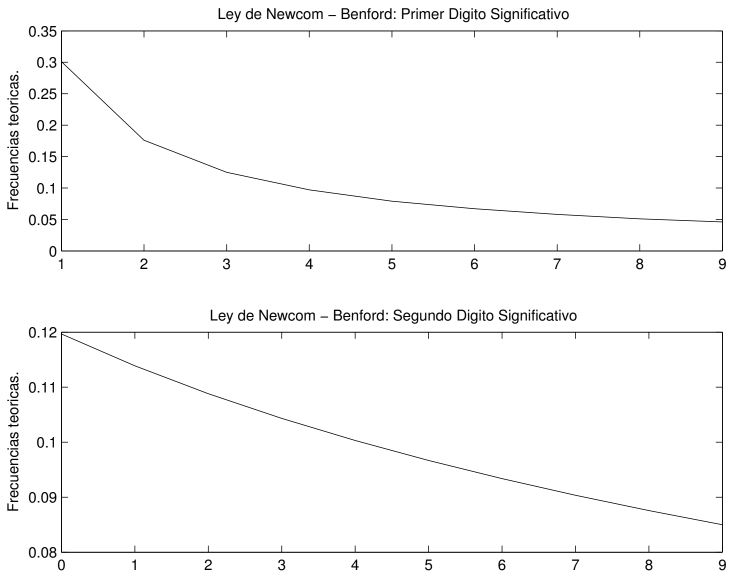

Figure 1: Frecuencias Teóricas de la Ley Newcomb-Benford para el Primer y Segundo dígito significativo.

ha utilizado esta "Ley" para poner en duda, y justificar un escrutinio más cuidadoso, de datos que vayan a ser auditados, particularmente en pago de impuestos.
Porque debe cumplirse esta Ley? Recientemente se ha conseguido una derivación estadística bastante general: *Hill (1995)*.
El resultado de Hill establece que se obtiene la Ley cuando la variable medida es el resultado de una **"mezcla insesgada de distribuciones"**, y por ello ha sido propuesto que se realicen test de Bondad de Ajuste contra la "Ley" para detectar la posibilidad de alteración de números. Enfatizamos, esto NO es una prueba de alteración, pero si es un indicio.
Los gráficos que siguen se refieren al **Segundo Dígito Significativo**, de la Ley de Newcomb-Benford. Es preferible revisar el segundo dígito ya que el primer dígito estará afectado por el hecho de que las diferentes máquinas tienen números pre-especificados y diseñados de electores.
Revisamos la Ley del Segundo Dígito con respecto al **Número de Votantes** tanto para los cuadernos manuales, como para los automatizados. Encontramos que la Ley se cumple en los manuales pero **NO** en los automatizados.

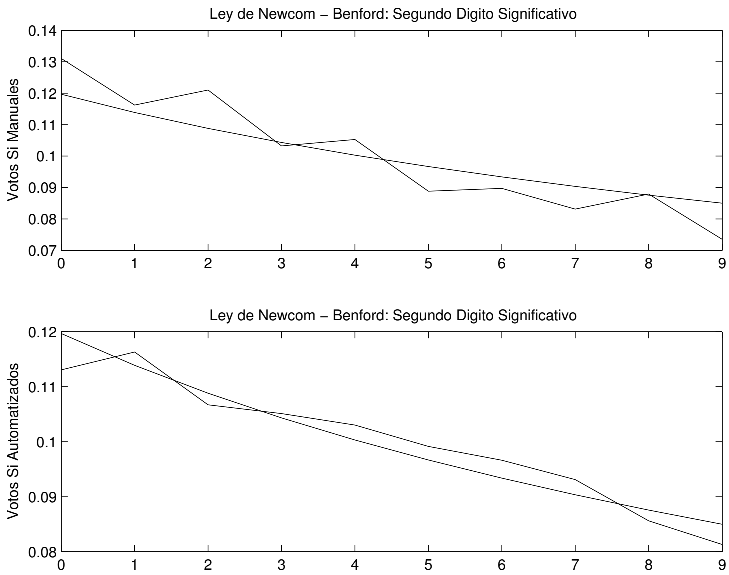

Figure 2: Frecuencias del Segundo Dígito de los VOTOS SI, Manuales (arriba), Automatizados (debajo)

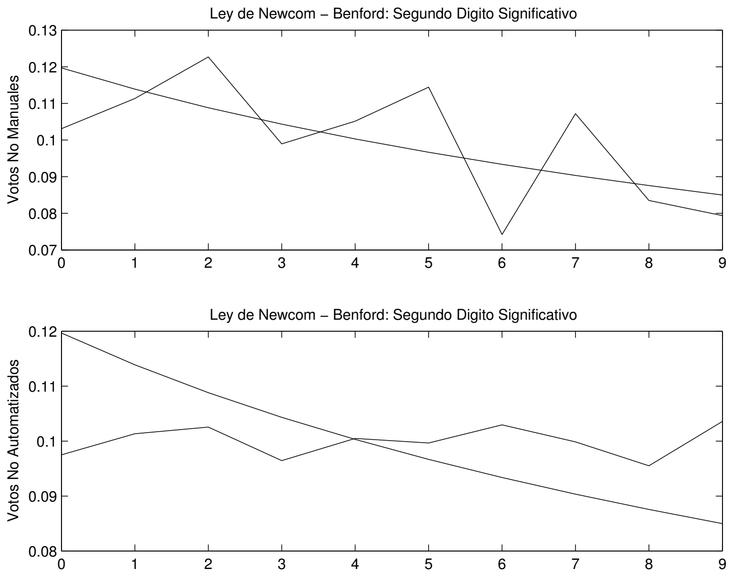

Figure 3: Frecuencias del Segundo Dígito de los VOTOS NO, Manuales (arriba), Automatizados (debajo)

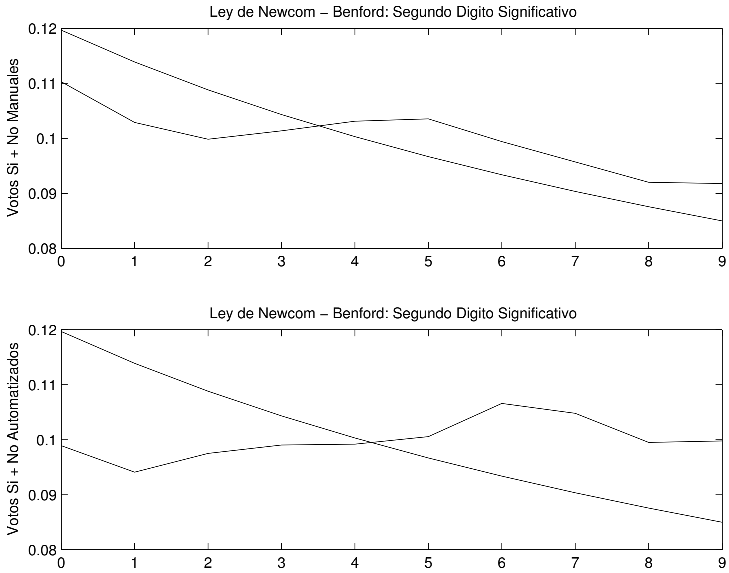

Figure 4: Frecuencias del Segundo Dígito de los VOTOS SI+NO, Manuales (arriba), Automatizados (debajo)

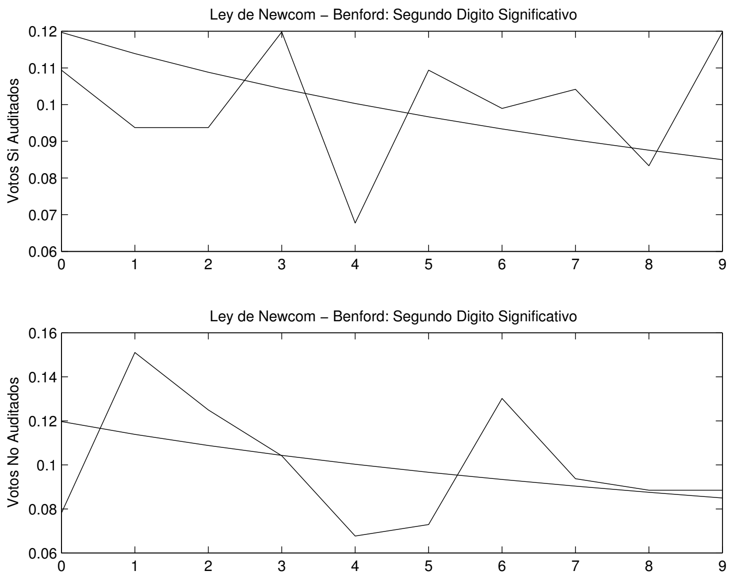

Figure 5: Frecuencias del Segundo Dígito de las máquinas Auditadas por el CNE-Centro Carter: VOTOS SI (arriba), VOTOS NO (debajo)

Table 1: P-Valores convertidos en Probabilidades de Hipotesis.

P_{val}	$\underline{P}(H_0\|datos)$
0.05	0.29
0.01	0.11
0.001	0.0184

1 Convirtiendo P-Valores en Probabilidades de la Hipótesis Nula H_0

Antes que nada aclaramos que, dadas las dudas razonables que se han levantado respecto del RR-15AGO, suponemos iguales probabilidades de que no haya habido intervención (H_0), como que haya habido intervención (H_1). Es importante medir la evidencia que se tiene en contra de Hipótesis Nulas, H_0. En nuestro caso H_0 es la hipótesis de que los datos obedecen la Ley de Newcomb-Benford, pero más generalmente, en los estudios que se han venido realizando sobre el RR-15AGO, H_0 es la hipótesis nula de no-intervención sobre la voluntad de los electores.
Usualmente se confunde la probabilidad en contra de H_0 con su p-valor.
Para una Hipótesis Nula El P-Valor es:

$$P_{val} = Prob(Resultado\ igual\ o\ mas\ extremo$$

$$que\ los\ datos|HipotesisNula)$$

P_{val} pequeño (Ej. $P_{val} < 0.05$ o menor) indica observación SORPRESIVA....
PERO: P_{val} NO son probabilidades de la Hipótesis Nula.
Si $P(H_0) = P(H_1)$, entonces , Sellke, Bayarri y Berger (2001)

$$\frac{P(H_0|data)}{P(H_1|data)} \geq -e.P_{val}.\log_e[P_{val}] \Rightarrow$$

$$P(H_0|P_{val}) \geq \underline{P}(H_0|datos) = 1/(1 + [-e.P_{val}.\log_e(P_{val})]^{-1})$$

Como se ve de la Tabla 1, la corrección de p-valores a $\underline{P}$ es importante.
Es natural calcular el p-valor, respecto a un test de bondad de ajuste de las proporciones observadas vs las especificadoas por la Ley de Newcomb-Benford, y es de allí que obtenemos los p-valores.
Ahora bien el resultado de arriba apenas nos da una cota inferior, que en algunos casos puede ser algo conservadora. Podemos calcular probabilidades aproximadas en base al criterio BIC (Bayesian Information Criterion) o criterio de Schwarz, ver por ejemplo Berger y Pericchi (2001), el cual toma en cuenta de manera explícita el tamaño de la muestra utilizada:

$$\log[P(H_0|datos)/P(H_1|datos)] \approx \log(TasadeVerosimilitudes) + (k_1 - k_0)/2\log(N), \quad (3)$$

Tipo de voto	p-valor de H_0	Cota Inferior $\underline{P}$ de H_0	Probabilidad Aproximada de H_0	Número de cuadernos: N
NO Automatizado	0.0	0.0	1.34×10^{-36}	19.064
NO Manual	0.1553	0.44	1.0	4.556
Sí Automatizado	0.024	0.1957	1.0	19.063
Sí Manual	0.0032	0.0476	1.0	4.379
No Auditado	0.2393	0.4819	1.0	192

Table 2: Evidencia en contra de la Hipótesis Nula H_0: Los datos siguen la Ley de Newcomb-Benford. Claramente los VOTOS NO automatizados violan H_0, pero los VOTOS NO manuales y los VOTOS NO automatizados de la Auditoría respetan la Ley

donde k_i es el número de parámetros ajustables para H_i y N el tamaño muestral. Las tasas de verosimilitudes las calculamos de la densidad multinomial: en el numerador con las proporciones especificadas por H_0 y en el denominador por proporciones libres.
En lo que sigue aplicamos tanto la cota inferior como la aproximación BIC para medir la evidencia en contra de la hipótesis nula de que los votos siguen la Ley de Newcomb-Benford.

2 Conclusiones

Las conclusiones son bastante claras. Los votos NO, en las mesas automatizadas violan la Ley de Newcom-Benford con Probabilidad virtualmente 1 (la Probabilidad de la Hipotesis complementaria, o sea la hipótesis de que cumple la Ley, es de 1.34×10^{-36}, virtualmente 0). Ello hace relevante indagar sobre mecanismos de intervención que hayan alterado los votos NO automatizados, de tal manera que violen de una forma clara la Ley Newcomb-Benford. Por otra parte, los votos NO de la auditoria no violan la Ley (lo cual es sumamente interesante ya que estos votos auditados son recogidos en forma automatizada tambien y sugiere que las mesas auditadas tienen otra estructura y por ello no son represntativos del total de votos automatizados, y ponen en duda que sean una muestra aleatoria), ni tampoco los votos NO manuales violan la Ley. Enfáticamente aclaramos que esto último no demuestra, que en particular, los votos manuales no hayan sido manipulados, solo que si se han manipulado no lo han sido en forma tal que sea detectada por las proporciones de Newcomb-Benford. De hecho en otro trabajo (Huerta, Pericchi, en preparación) se estudia la intervención en los centros con votos manuales.

3 References

Berger, J.O. and Pericchi, L.R. (2001) "Objective Bayesian methods for model selection:Introduction and comparison (with discussion)." In Model Selection (P.Lahiri ed.) pp. 135-207. Institute of Mathematical Statistics, Monographs. Beachwood, OH.

Hill, T.P "A Statistical Derivation of the Significant-Digit Law", 1995, "Statistical Science", Vol. 10, No. 4, pp. 354-363.

Sellke, T., Bayarri, M.J., and Berger, J.O. (2001). Calibration of p-values for precise null hypothesis. "The American Statistician", (55), pp. 62-71.