La ley de Benford aplicada al tamaño de las tablas de una base de datos y como indicador del riesgo inherente de la información contenida

Héctor Rubén Morales, Cecilia Beatriz Díaz,
Ricardo Justo Castello

Ponencia presentada en 47 JAIIO Jornadas Argentinas de Informática. Simposio Argentino de GRANdes Datos realizado en 2018 en la Facultad de Ingeniería - Universidad de Palermo. Buenos Aires, Argentina

47º JORNADAS ARGENTINAS DE INFORMÁTICA - JAIIO
Simposio Argentino de GRANdes DAtos (AGRANDA 2018)

Tema: Aplicaciones de Big Data para la Ciencia, la Empresa, el Gobierno y la Sociedad

Título del trabajo:

"LA LEY DE BENFORD APLICADA A UNA BASE DE DATOS. POSIBLE INDICADOR DE RIESGO INHERENTE DE LA INFORMACIÓN CONTENIDA"

AUTORES:

Lic. Héctor Rubén MORALES
(rmorales@eco.unc.edu.ar)

Dra. Cecilia Beatriz DÍAZ
(cdiaz@eco.unc.edu.ar)

Dr. Ricardo Justo CASTELLO
(castello@eco.unc.edu.ar)

FACULTAD CIENCIAS ECONOMICAS – UNIVERSIDAD NACIONAL DE CORDOBA

Abril 2018

Resumen

La Ley de Benford considera que en un conjunto determinado de números, más del 30% de estos empiezan con el dígito 1, con el dígito 2 inician casi el 18%, y desciende sucesivamente hasta el 9 con menos del 5%. Este comportamiento ha sido verificado para conjuntos de números que son objeto de estudio en distintos ámbitos científicos. El objetivo de este trabajo es verificar si la distribución estadística de Benford se aplica a los números representados por el tamaño (cantidad de registros) que contienen las distintas tablas que conforman una base de datos relacional.

Los resultados alcanzados confirman esa hipótesis, para lo cual se recurre al análisis estadístico de pruebas de bondad de ajuste.

El estudio pretende servir de base para su uso como posible indicador del riesgo inherente de la información que el auditor utiliza para su tarea de control.

Palabras clave: Benford – registros- tablas - base datos – riesgo

1- Introducción

La ley surgida hace 80 años del físico Frank Benford se encuentra en pleno estudio y debate sobre su aplicación. Es promovida en distintos ámbitos científicos con la finalidad de verificar si el comportamiento de un conjunto de números, que representan el objeto de estudio, se apega a la misma. Esta ley considera que ciertos dígitos aparecen más frecuentemente que otros en un conjunto determinado de datos. Observa que más del 30% de los números empiezan con el dígito uno, con el dígito dos inician cerca del 18% de los números y, desciende sucesivamente hasta el nueve, a menos del 5% la incidencia como primer dígito de una cifra. También destaca observaciones similares, aunque con frecuencias más estrechas, cuando el análisis se efectúa sobre el segundo dígito del compendio de números estudiados.

Distintas posturas tratan de explicar este patrón de comportamiento que, según se verifica, responde a una función logarítmica (Benford, 1938). Los menos rigurosos sostienen que este fenómeno es sólo la forma en que escribimos los números, como la tesis de Goudsmit y Furry (1944), y hasta la teoría de que ello refleja "la verdad de la naturaleza" (Furlan, 1948).

En los últimos treinta años, algunos estudios analizaron su aplicación sobre datos de la contabilidad con fines de auditoría (Hill, 1995). Consideran que si el comportamiento del conjunto representativo de datos no cumple con la distribución de Benford, puede entenderse la presencia de posibles riesgos de irregularidades o fraudes (Nigrini, 1999).

El objetivo de nuestro trabajo es verificar si la distribución estadística de Benford se aplica al conjunto de números representados por la cantidad de registros (tamaño) que contienen las distintas tablas que conforman una base de datos relacional.

La finalidad primera es la comprobación en sí misma, dado que de la revisión bibliográfica no se advierte un estudio similar. En segundo lugar, se pretende atender –en parte- la problemática o incertidumbre que experimenta el auditor que actúa sobre un sistema de información en un contexto computarizado, respecto a la confianza o posible riesgo inherente en la información contenida en la base de datos que utilizará para su control. Al respecto las Normas Internacionales de Auditoría (NIA) propician que los auditores empleen procedimientos analíticos durante la fase de planificación y ejecución de la auditoría con el objetivo de identificar, entre otras, la existencia de tendencias inusuales (NIA 300, 400).

Para ello se analizan ocho módulos informáticos que conforman la base de datos bajo estudio, correspondiente a una empresa de envergadura del sector energético de Argentina. Cada módulo contiene datos de entre 3 y 12 años de operatoria. La metodología aplicada es probar el cumplimiento de la ley de Benford a partir de un análisis puntual (en un momento determinado), tanto sobre la cantidad de registros (tamaño) de las tablas de cada módulo, como sobre el conjunto de todos estos módulos informáticos, que engloban más de 1.900 tablas con 4.500 millones de registros y que constituyen la base de datos relacional considerada.

Los resultados obtenidos muestran que la ley de Benford se ajusta total o parcialmente a cada módulo sometido bajo análisis, dependiendo ello de la cantidad de tablas que lo conforman, mientras se

ciñe de manera absoluta cuando el objeto analizado abarca al conjunto de todos los módulos que constituyen la base de datos. Para ello se recurre al análisis de bondad de ajuste mediante la prueba de chi cuadrado y desviación absoluta media.

El presente trabajo resulta novedoso como posible aporte para contribuir a mitigar la incertidumbre del auditor. En ese sentido, es un estudio empírico que puede servir de base, para un análisis más profundo. Se pretende generar un aporte en pos al convencimiento acerca de que el resultado alcanzado pueda ser interpretado como un indicador sobre la confianza o alerta del posible riesgo inherente o preexistente de los datos informatizados que son puestos a disposición del auditor al iniciar su tarea de contralor. En consecuencia, este estudio constituye una primera etapa de investigación, pudiendo ser ampliado y/o comparado con otras bases de datos.

2- Revisión bibliográfica

2.1 La Ley de Benford

En 1938 el físico Frank Benford advirtió que su libro de tablas de logaritmos tenía más desgastadas las primeras páginas que las últimas. Dedujo que trabajó más con números cuyas cifras iniciales eran bajas (1, 2 o 3) y menos con aquellas que empiezan con dígitos mayores (7, 8 o 9). Concluyó que la primera cifra de los números no se distribuía de manera uniforme (como podría pensarse).

Benford realizó una comprobación empírica sobre un total de 20.229 números agrupados en 20 muestras muy diversas, entre ellas: longitud de más de 300 ríos, cantidad de habitantes de más de 3.200 ciudades, constantes y magnitudes físicas y químicas (como el peso atómico de los elementos), funciones matemáticas e incluso números de direcciones de calles. A partir de los resultados Benford postuló la "ley de los números anómalos" para la probabilidad de que el *primer dígito* sea d. Esta ley logarítmica se conoce como "ley de Benford" que se describe como sigue (Benford, 1938):

$$\text{Prob}\,(d_1) = \log_{10}\left(1+\frac{1}{d_1}\right), \quad d_1 = 1, 2, 3, \ldots\ldots, 9 \qquad (1)$$

Si la atención es verificar como se distribuye según la ley de Benford el *segundo dígito*, está dada por la siguiente expresión:

$$\text{Prob}(d_2) = \sum_{k=1}^{9} \log_{10}\left(1+\frac{1}{10k+d_2}\right), \quad d_2 = 0, 1, 2, \ldots\ldots, 9 \qquad (2)$$

De igual manera, matemáticamente, se deducen las fórmulas para la ubicación del *tercer dígito* y siguientes.

Los resultados que arroja esta ley respecto a la probabilidad de ocurrencia de los primeros dígitos, se describen en el **Cuadro 1.**

Cuadro 1 – Probabilidad de ocurrencia para cada dígito de acuerdo a la posición que ocupa en un número

Dígito/Posición	Primera	Segunda	Tercera	Cuarta	Quinta o superior
0		11,97%	10,18%	10,02%	10,00%
1	30,10%	11,39%	10,14%	10,01%	10,00%
2	17,61%	10,88%	10,10%	10,01%	10,00%
3	12,49%	10,43%	10,06%	10,01%	10,00%
4	9,69%	10,03%	10,02%	10,00%	10,00%
5	7,92%	9,67%	9,98%	9,99%	10,00%
6	6,69%	9,34%	9,94%	9,99%	10,00%
7	5,80%	9,04%	9,90%	9,99%	10,00%
8	5,12%	8,76%	9,86%	9,99%	10,00%
9	4,58%	8,50%	9,83%	9,98%	10,00%

Fuente: elaboración propia a partir de la Ley de Benford.

Es decir, según la ley el 30,1% de las veces, la primera cifra significativa (no incluye el 0) será un 1, mientras (en el otro extremo) sólo un 4,6% de las veces será 9.

Esta ley presenta una propiedad matemática que la hace exclusiva. Es la única ley de probabilidad invariante frente a cambios de escala. Se aplica independientemente de la escala de medición. Arribamos al mismo resultado tanto si trabajamos o convertimos la información multiplicando o dividiendo por una constante, o bien usando datos en kilómetros, millas o metros, o en términos financieros daría lo mismo utilizar importes en Pesos o Dólares.

Mientras como limitantes principales se cuenta que no es aplicable a un conjunto de números aleatorios como la lotería, o de números asignados, como los números de teléfono celular, las cédulas de identidad o números de cheques, ya que comienzan con números correlativos; o números que fluyen sólo en un rango determinado, como puede ser la estatura de las personas.

2.2 Antecedentes de aplicaciones de la Ley de Benford

La aplicabilidad de la ley de Benford es investigada en distintos ámbitos científicos. En el bagaje de estudios realizados, y sólo a título ilustrativo se cuentan: la demostración del ajuste a la ley de las constantes físicas (Burke y Kincanon, 1991), su aplicación como indicador de fiabilidad para evaluar riesgo de toxicidad (Pepijn de Vries et al., 2013), el uso para la posible detección de fraudes electorales (Roukema, 2009 y Castañeda, 2011), y el cumplimiento de este postulado, sobre la cantidad de seguidores de distintas redes sociales (Golbeck, 2015).

Los desarrollos antes citados emanaron luego de que fuera estudiada su aplicación en las ciencias económicas, y encontrara espacio concreto para su utilidad. Varian (1972), economista, sugiere que la ley de Benford puede usarse como una prueba de la honestidad o validez de datos científicos supuestamente aleatorios en un contexto de ciencias sociales. Esto recién fue recogido por los contables a fines de los años ochenta, cuando dos estudios se basaron en el análisis digital para detectar la manipulación de los ingresos. Carslaw (1988) encontró que el número de ganancias de las empresas neozelandesas no se ajustaba a la distribución esperada, y a su vez Thomas (1989) descubrió un patrón similar en las ganancias de las empresas estadounidenses.

Hill (1995), aportó una prueba para la ley de Benford, y demostró cómo se aplicaba a los datos bursátiles, las estadísticas del censo y ciertos datos contables. Señaló que la distribución de Benford, como la distribución normal, es un fenómeno observable empíricamente.

Nigrini parece ser el primer investigador en aplicar la ley de Benford de manera amplia a los números de la contabilidad y con el objetivo de detectar posibles fraudes. Para su tesis tomó el trabajo sobre la manipulación de ganancias por Carslaw y Thomas, y añadió el de Benford. Su estudió se basó en el análisis digital para ayudar a identificar a evasores de impuestos (Nigrini 1996). Posteriormente, publicó distintos artículos sobre aplicaciones prácticas con fines de auditoría a partir de pruebas en conjuntos de números de contabilidad (Nigrini y Mittermaier 1997, Nigrini 1999).

Sin embargo, la literatura académica es algo cautelosa al hacer afirmaciones sobre la efectividad de los procedimientos basados en la ley de Benford para detectar el fraude. En general, se sostiene que si al someter a prueba un conjunto de datos, este no se ajusta a la ley de Benford, sólo puede mostrar ineficiencias operativas o fallas en sistemas, en lugar de precisar un fraude (Etteridge y Srivastava, 1999). Bajo esa óptica se la entiende, entonces, como una orientación concreta a dónde o en qué dirección fijar el énfasis del control.

A su vez, la normativa internacional insta al auditor que actúa en entornos informatizados a evaluar la confiabilidad de los datos con los que desarrolla su tarea y a aplicar pruebas de funcionamiento de controles y pruebas sustantivas sobre datos a nivel de transacciones (NIA 315, 330, 401). En tal sentido, es variada la literatura que propicia para tal fin, entre otras pruebas de consistencia de datos, al análisis de Benford.

Este trabajo trata de complementar esa iniciativa aplicando Benford en una etapa preliminar a aquella del control a nivel de datos puros. Nos referimos a considerar este análisis como control previo, es decir un aporte a dilucidar el denominado (por la normativa, NIA 400) riesgo inherente de la información que es puesta a disposición del auditor para su contralor.

El riesgo inherente está dado por las características de la entidad y del sistema de información bajo análisis. Este riesgo no puede ser cambiado por el auditor pero si conocerlo, es innato de la empresa. Es una medida apriorística del riesgo e independiente a los controles que se estén aplicando, tratando de indagar sobre posibles errores o situaciones de importancia, antes de la evaluación del control interno (NIA 400).

La posibilidad de verificar si el perfil de la base de datos se ajusta a la distribución de Benford, a partir de analizar el tamaño (cantidad de registros) de las tablas que la componen, permitiría esa mirada anticipada o previa, superficial pero a la vez casi radiográfica de la base de datos. Este diagnóstico ayudaría a mitigar la incertidumbre del auditor, como posible indicador de fiabilidad, utilidad esta ya aplicada en otras disciplinas.

3. Objetivo del estudio:

El objetivo principal es verificar si la distribución estadística de Benford se aplica al conjunto de números representados por el tamaño (cantidad de registros) que contienen las distintas tablas que conforman una base de datos relacional.

El objetivo secundario es servir de base para profundizar el análisis, a fin de determinar su uso como posible indicador de riesgo inherente de la información que el auditor utiliza para su control.

4. Metodología

4.1 Fuente de datos

Para verificar empíricamente si la ley de Benford se ajusta al tamaño de las tablas de una base datos, se recurrió a los datos de una empresa de envergadura que opera en el sector energético del país, prestando servicio a más de 1,1 millones de clientes. La base de datos es relacional bajo Oracle, conformada de distintos módulos funcionales según la especificidad del proceso (comercial, contabilidad, sueldos, etc.).

Los datos para el análisis es el conjunto de números que indican la cantidad de registros (tamaño) de las distintas tablas que integran cada módulo que, a su vez, conforman la base de datos. Estos valores son obtenidos a un momento determinado, cuando se ejecuta un back up, y entendemos constituyen un tipo de perfil de la base de datos. Se excluyeron del análisis las tablas nulas, que no cuentan con registros.

La base de datos se conformó seleccionando ocho módulos informáticos, tratando que la composición resulte heterogénea. La heterogeneidad surge al considerar módulos de características diferentes en cuanto al tipo de procesos que realizan como al volumen de información que almacenan. De este modo, se genera mayor complejidad en el análisis y en consecuencia, mayor robustez en los resultados que se alcancen.

El **Cuadro 2** describe la composición de cada uno de los módulos (volumen en cantidad de tablas y registros de estas) que integran la base de datos. La base de datos contiene un total de 1.923 tablas con casi 4.500 millones de registros. El módulo de Gestión Comercial (GC) es el de mayor trascendencia con 808 tablas (42%) que alcanza al 91% de los datos con 4.093 millones de registros, lo que implica un promedio por tabla superior a 5 millones de registros. En relación inversa se observa que el módulo Recursos Humanos (RH) tiene 340 tablas (17,7%) con el 0,03% de los datos (1,4 millones de registros), es decir, un promedio de sólo 4.365 registros por tabla.

Cuadro 2 -Módulos informáticos que forman la Base de Datos

Módulo informático	Cantidad tablas	Tablas Módulo sobre Total	Registro del Módulo	Registros Módulo sobre Total	Promedio Registros por Módulo
Contabilidad (CO)	78	4,1%	225.245.009	5,02%	2.887.757
Recursos Humanos (RH)	340	17,7%	1.483.944	0,03%	4.365
Inventario (IN)	71	3,7%	14.916.459	0,33%	210.091
Gestión Obras (GO)	58	3,0%	380.572	0,01%	6.562
Solicitudes Int. (SI)	87	4,5%	1.832.473	0,04%	21.063
Gestión Comercial (GC)	808	42,0%	4.093.412.030	91,25%	5.066.104
Liquidación sueldos (SU)	381	19,8%	135.126.395	3,01%	354.662
Adm. Expedientes (EX)	100	5,2%	13.416.811	0,30%	134.168
Total Base de Datos	1.923	100%	4.485.813.693	100,00%	2.332.716

Fuente: elaboración propia

4.2 Clasificación de los datos

Para calcular la distribución de Benford, los datos son clasificados mediante el recuento del primer dígito (inicial) del número referido a la cantidad de registros que contiene cada tabla. Esta tarea se puede realizar con distintas herramientas informáticas de planillas de cálculo. En esta oportunidad se recurre al uso del software de auditoria ACL que cuenta con un comando específico para el análisis de Benford. A su vez, para una mejor ilustración se utilizan distintos cuadros que muestran los pasos sucesivos del análisis.

El **Cuadro 3** describe por cada módulo el conteo o frecuencia de tablas cuyas cantidades de registros inician con el dígito 1 y, en ese orden, hasta el 9. Ejemplo: el módulo Contabilidad (CO) tiene 78 tablas, de estas hay 22 tablas cuya cantidad de registros inician con el dígito 1, otras 11 tablas inician con el dígito 2, y así sucesivamente.

Cuadro 3 - Discriminación en función al dígito con que inician las cantidades de registros de las tablas de cada módulo

DIGITO	CO	RH	IN	GO	SI	GC	SU	EX	Total
1	22	94	21	12	23	254	114	26	**566**
2	11	74	17	13	17	135	70	24	**361**
3	11	35	11	12	10	89	62	13	**243**
4	6	38	6	2	10	89	39	12	**202**
5	5	34	7	4	9	54	33	6	**152**
6	10	20	2	6	2	60	23	7	**130**
7	4	16	3	5	5	39	14	7	**93**
8	5	15	3	3	5	45	14	3	**93**
9	4	14	1	1	6	43	12	2	**83**
Total	**78**	**340**	**71**	**58**	**87**	**808**	**381**	**100**	**1.923**

Fuente: elaboración propia.

4.3 Análisis de los datos

Previo a iniciar el análisis propiamente dicho, corresponde revisar los posibles limitantes que rigen para aplicar Benford a un conjunto de números.

El tema abordado, en general, cumple con todos los condicionantes que considera la literatura. El conjunto de datos está formado por magnitudes medibles de un mismo fenómeno. Los datos no son números asignados o aleatorios. La distribución de la variable es ligeramente asimétrica positiva, es decir tiene un mayor número de valores pequeños que grandes, lo que es consecuencia natural del fenómeno analizado (cantidad de registros). Los datos están generados en periodos de tiempo prolongados, dado que entre los módulos el rango de la información oscila entre 3 y 12 años.

Por otra parte, para la prueba del primer dígito la bibliografía recomienda que el conjunto de datos analizados sea mayor a 1.000 números. Sobre este aspecto, se cumple respecto a la cantidad de tablas que conforman la base de datos en su conjunto. Pero si el análisis se efectúa por cada módulo, al reunir una cantidad menor de tablas podría no cumplir con la distribución de Benford. No obstante, según Nigrini en el estudio de Wallace (2002) se verifica el cumplimiento de Benford usando cuatro conjuntos de datos con sólo 67 observaciones cada uno (Nigrini et al., 2007).

En el **Cuadro 4** se muestra para cada módulo en términos porcentuales la frecuencia apuntada en el Cuadro 3, con el agregado de la columna que indica la distribución esperada para el primer dígito por la Ley de Benford.

Cuadro 4 - Detalle porcentual por dígito con que inician las cantidades de registros de las tablas de cada módulo

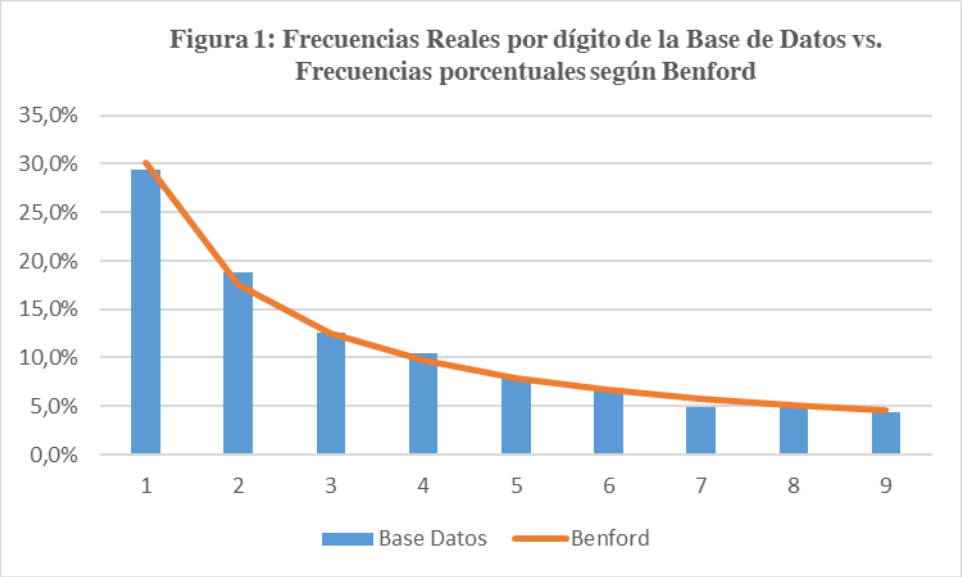
DIGITO	CO	RH	IN	GO	SI	GC	SU	EX	Total	Benford
1	28,2%	27,6%	29,6%	20,7%	26,4%	31,4%	29,9%	26,0%	**29,4%**	**30,1%**
2	14,1%	21,8%	23,9%	22,4%	19,5%	16,7%	18,4%	24,0%	**18,8%**	**17,6%**
3	14,1%	10,3%	15,5%	20,7%	11,5%	11,0%	16,3%	13,0%	**12,6%**	**12,5%**
4	7,7%	11,2%	8,5%	3,4%	11,5%	11,0%	10,2%	12,0%	**10,5%**	**9,7%**
5	6,4%	10,0%	9,9%	6,9%	10,3%	6,7%	8,7%	6,0%	**7,9%**	**7,9%**
6	12,8%	5,9%	2,8%	10,3%	2,3%	7,4%	6,0%	7,0%	**6,8%**	**6,7%**
7	5,1%	4,7%	4,2%	8,6%	5,7%	4,8%	3,7%	7,0%	**4,8%**	**5,8%**
8	6,4%	4,4%	4,2%	5,2%	5,7%	5,6%	3,7%	3,0%	**4,8%**	**5,1%**
9	5,1%	4,1%	1,4%	1,7%	6,9%	5,3%	3,1%	2,0%	**4,3%**	**4,6%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Fuente: elaboración propia.

Si se compara la distribución de frecuencias porcentuales de cada módulo con la de Benford (última columna del Cuadro 4), a simple vista surgen coincidencias y desvíos. Entre los desvíos, algunos aparentan ser significativos. Un ejemplo es el módulo CO, donde el dígito 6 como primera cifra de la cantidad de registros, está en el 12,8% de las tablas que lo componen, cuando lo esperado según Benford es 6,7%. Esta cuestión puede ser consecuencia de lo comentado en párrafos anteriores, respecto a que una

reducida cantidad de datos analizados puede ser una limitante, como es el caso del módulo CO que solo comprende 78 tablas en total.

Finalmente, la comparación visual de los valores obtenidos para el total de la base de datos con la distribución de Benford (dos últimas columnas del Cuadro 4), muestra diferencias aunque no significativas. Gráficamente se puede apreciar en **Figura 1**.



4.4 Pruebas de bondad de ajuste

Comenzamos analizando los valores de la base de datos en su conjunto, aplicando la prueba de bondad de ajuste Chi Cuadrado. La hipótesis nula (Ho) es que los datos reales u observaciones siguen la distribución de probabilidad esperada por la Ley de Benford. La fórmula de Chi Cuadrado (χ^2) considerada es la siguiente:

$$\chi^2 = \sum_{d=m}^{9} \frac{(P_{obs}(d) - P_t(d))^2}{P_t(d)} \qquad (3)$$

donde: - *Pt(d)* es la frecuencia esperada según Benford

- *Pobs(d)* es la frecuencia observada
- *m* es el dígito analizado. En este estudio es sólo el primer dígito (*m*=1)

Para su aplicación recurrimos a las frecuencias reales observadas sobre la base de datos en su conjunto (ya conocidas, última columna Cuadro 3), y a las frecuencias esperadas, que surgen de considerar la distribución de Benford para cada dígito. Estas frecuencias esperadas se calculan sobre el número total de 1.923 tablas. Así, por ejemplo, la frecuencia esperada para el dígito 1, es 578,88 (1.923 * 0,301) y para el dígito 2 es 338,62 (1.923 * 0,176).

En función a lo expresado, conformamos el **Cuadro 5**, y obtenemos los componentes de χ^2:

Cuadro 5 - Obtención de χ^2 a partir de las frecuencias reales y las esperadas de la base de datos

DIGITO	Observ.	Benford	*Pobs(d) - Pt(d)*	*(Pobs(d) - Pt(d))²*	*(Pobs(d) - Pt(d))²*
	Pobs(d)	*Pt(d)*			*Pt(d)*
1	566	578,88	-12,88	165,91	0,29
2	361	338,62	22,38	500,71	1,48
3	243	240,26	2,74	7,52	0,03
4	202	186,36	15,64	244,67	1,31
5	152	152,27	-0,27	0,07	0,00
6	130	128,74	1,26	1,59	0,01
7	93	111,52	-18,52	342,94	3,08
8	93	98,37	-5,37	28,80	0,29
9	83	87,99	-4,99	24,92	0,28
Total	1.923	1.923	**Valor de Chi Cuadrado (χ^2) ->**		**6,77**

Fuente: elaboración propia.

Obtenemos el valor crítico para la distribución de χ^2 para un $\alpha = 0{,}05$ (confianza del 95%) y 8 grados de libertad ((9 filas – 1) x 1 columna), que resulta $\chi^2_{(0{,}95;8)} = 15{,}51$. Al ser el estadístico obtenido menor al valor crítico de 15,51 se acepta que el conjunto de la base de datos se ajusta a la ley de Benford.

Luego, evaluamos la prueba de ajuste χ^2 para cada uno de los módulos. De igual manera se conforma el **Cuadro 6**, con las frecuencias reales (obtenidas en el Cuadro 3) y las esperadas según la distribución de Benford. Por ejemplo, para el módulo CO, en el dígito 1 la frecuencia esperada es 23,5 (78 * 0,301).

Cuadro 6 - Frecuencias reales y esperadas según Benford para cada módulo

Dig.	CO		RH		IN		GO		SI		GC		SU		EX	
	Real	Esper	Real	Esper.	Real	Esper	Real	Esper.	Real	Esper	Real	Esper.	Real	Esper	Real	Esper.
1	22	23,5	94	102,3	21	21,4	12	17,5	23	26,2	254	243,2	114	114,7	26	30,1
2	11	13,7	74	59,9	17	12,5	13	10,2	17	15,3	135	142,3	70	67,1	24	17,6
3	11	9,7	35	42,5	11	8,9	12	7,2	10	10,9	89	100,9	62	47,6	13	12,5
4	6	7,6	38	32,9	6	6,9	2	5,6	10	8,4	89	78,3	39	36,9	12	9,7
5	5	6,2	34	26,9	7	5,6	4	4,6	9	6,9	54	64,0	33	30,2	6	7,9
6	10	5,2	20	22,7	2	4,7	6	3,9	2	5,8	60	54,1	23	25,5	7	6,7
7	4	4,5	16	19,7	3	4,1	5	3,4	5	5,0	39	46,9	14	22,1	7	5,8
8	5	4,0	15	17,4	3	3,6	3	3,0	5	4,5	45	41,4	14	19,5	3	5,1
9	4	3,6	14	15,6	1	3,3	1	2,7	6	4,0	43	37,0	12	17,4	2	4,6
Total	78	78	340	340	71	71	58	58	87	87	808	808	381	381	100	100
χ^2	**6,09**		**9,48**		**6,15**		**10,98**		**5,17**		**8,55**		**11,34**		**6,51**	

Fuente: elaboración propia.

Los valores obtenidos de χ^2 son menores al valor crítico de 15,51 por lo que todos los módulos se ajustan a la Ley de Benford.

Cabe aclarar que otra alternativa para ejecutar χ^2 para el total de la base de datos es la suma de todos los χ^2 obtenidos para cada módulo. Esto arroja para toda la base de datos un χ^2=64.28. A su vez, el valor crítico χ^2 será para 56 grados de libertad ((9 filas – 1) x (8 columnas – 1)), que es de 74,47. Al ser el valor crítico mayor que el estadístico obtenido, también se confirma que el comportamiento de las frecuencias reales para toda la base de datos se ajusta a la Ley de Benford.

Nigrini (et al. 2012) considera que corresponde también aplicar el análisis de bondad mediante el test de la desviación absoluta media (MAD). La fórmula es:

$$MAD = \frac{1}{9}\sum_{d=1}^{9} |P_{obs}(d) - P_t(d)| \qquad (4)$$

donde: - *Pt(d)* es la proporción esperada según Benford

- *Pobs(d)* es la proporción observada

Cuando se utiliza este estadístico para la Ley de Benford, Nigrini (et al. 2012) sostiene que se puede determinar el nivel de conformidad según el rango donde se encuentren los valores obtenidos. Estos se muestran en **Cuadro 7.**

Cuadro 7 – Rangos de Conformidad para MAD

Rango	Nivel de Conformidad
0.000 a 0.006	Alta
0.006 a 0.012	Acepta
0.012 a 0.016	Media
Más de 0.016	Baja

En función a ello, y partiendo de los datos del Cuadro 3, efectuamos los cálculos de la MAD y calificamos siguiendo el criterio de Nigrini. Esto se expone en **Cuadro 8**.

Cuadro 8 - Cálculo del MAD para cada módulo y para la base de datos

DIGITO	CO	RH	IN	GO	SI	GC	SU	EX	Total	Benford
1	28,2%	27,6%	29,6%	20,7%	26,4%	31,4%	29,9%	26,0%	**29,4%**	**30,1%**
2	14,1%	21,8%	23,9%	22,4%	19,5%	16,7%	18,4%	24,0%	**18,8%**	**17,6%**
3	14,1%	10,3%	15,5%	20,7%	11,5%	11,0%	16,3%	13,0%	**12,6%**	**12,5%**
4	7,7%	11,2%	8,5%	3,4%	11,5%	11,0%	10,2%	12,0%	**10,5%**	**9,7%**
5	6,4%	10,0%	9,9%	6,9%	10,3%	6,7%	8,7%	6,0%	**7,9%**	**7,9%**
6	12,8%	5,9%	2,8%	10,3%	2,3%	7,4%	6,0%	7,0%	**6,8%**	**6,7%**
7	5,1%	4,7%	4,2%	8,6%	5,7%	4,8%	3,7%	7,0%	**4,8%**	**5,8%**
8	6,4%	4,4%	4,2%	5,2%	5,7%	5,6%	3,7%	3,0%	**4,8%**	**5,1%**
9	5,1%	4,1%	1,4%	1,7%	6,9%	5,3%	3,1%	2,0%	**4,3%**	**4,6%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
MAD	**0,0213**	**0,0172**	**0,0251**	**0,0434**	**0,0203**	**0,0102**	**0,0130**	**0,0238**	**0,0049**	
Califica	***Baja***	***Baja***	***Baja***	***Baja***	***Baja***	***Acepta***	***Media***	***Baja***	***ALTA***	

Fuente: Elaboración propia.

5. Resultados:

Los resultados obtenidos por el test de la χ^2 permiten afirmar que, considerando el primer dígito de la distribución de las cantidades de registros que contiene las tablas de la base de datos analizada, siguen la Ley de Benford. También las distribuciones de las observaciones de cada módulo informático permiten confirmar que se asemejan a lo esperado por la Ley de Benford. Es decir, en todos los casos a través de χ^2 se acepta la hipótesis nula.

El test de la Desviación Absoluta Media (MAD) presenta resultado muy favorable para la base de datos en su conjunto, cumpliendo con la Ley de Benford. Cuando el MAD se realiza a nivel de cada módulo, surge que solo los módulos GC (Gestión Comercial) y SU (Sueldos) permiten afirmar la hipótesis de que la distribución de sus datos se asemejan a la Ley de Benford. En tanto en los seis módulos restantes se rechaza la hipótesis nula.

Profundizando el resultado del MAD, se observa que, además de la base de datos en su conjunto, los módulos GC y SU, que cumplen con la Ley de Benford, son los de mayor cantidad de tablas y están entre los de mayor volumen de información. Esto se puede apreciar en el **Cuadro 9**.

Cuadro 9 – Módulos informáticos ordenados en función a menor MAD y cantidad de tablas que contienen

Módulo informático	Cantidad tablas	Total registros módulo	MAD Valor	MAD Conformidad
Total BD	1.923	4.485.813.693	**0,0049**	**ALTA**
GC	808	4.093.412.030	**0,0102**	**BUENA**
SU	381	135.126.395	**0,0130**	**MEDIA**
RH	340	1.483.944	**0,0172**	**BAJA**
SI	87	1.832.473	**0,0203**	**BAJA**
CO	78	225.245.009	**0,0213**	**BAJA**
EX	100	13.416.811	**0,0238**	**BAJA**
IN	71	14.916.459	**0,0250**	**BAJA**
GO	58	380.572	**0,0434**	**BAJA**

Fuente: elaboración propia.

En función a lo analizado, se observa en general que a menor cantidad de tablas del módulo, el valor del MAD es mayor, lo que lleva a la no conformidad y rechazar la hipótesis nula. Esta cuestión, como se dijo anteriormente, se puede corresponder con los posibles limitantes para verificar el comportamiento de Benford ante una baja cantidad de datos (número de tablas del módulo).

6. Conclusión:

De acuerdo con el análisis realizado y los resultados alcanzados, se verifica que la ley de Benford se ajusta a la distribución del tamaño (en cantidad de registros) de las tablas que integran una base de datos.

Si el análisis se parcializa a nivel de los módulos que conforman la base de datos, el ajuste a la distribución de Benford se verifica con la prueba de bondad de Chi Cuadrado. No ocurre lo mismo con la prueba de ajuste mediante la prueba de Desviación Absoluta Media (MAD). En este último, la prueba del MAD muestra que el ajuste con Benford no se cumple cuando la cantidad de tablas del módulo es menor a 350 tablas. El MAD alcanza una conformidad media con casi 400 tablas y valores aceptables con 800. La extrapolación indica alrededor de 700 tablas como la cantidad mínima para alcanzar un ajuste aceptable.

Dado que la cantidad de tablas de un módulo es de baja elasticidad, consideramos que el estudio podría ampliarse tratando de adecuarse para salvar esos limitantes. En tal sentido, la propuesta, ya testeada en parte, es considerar no solo un análisis en un momento determinado, sino sumar a ese punto de partida otros momentos (otros back up). La suma de cada momento duplicaría la cantidad de tablas pero con distintas cantidades de registros por el propio crecimiento. Esto permitiría revisar la distribución de Benford según la evolución temporal de la base de datos.

De lograrse resultados aceptables, torna posible indagar en la factibilidad de establecer un orden de prioridad sobre los módulos en que se debe orientar la atención de la auditoría o área de control.

Finalmente, consideramos que el presente trabajo es novedoso y puede ser un aporte para contribuir a mitigar la incertidumbre del auditor. Lo expresado se basa en la experiencia alcanzada en otras disciplinas, que también han demostrado que el conjunto de números analizados cumplen con la distribución de Benford. Estas infieren y algunas lo demuestran, que si la distribución sometida a examen no se ajusta a Benford, existen indicios de posibles irregularidades. Consideramos que este tipo de estudio empírico puede servir de base, para un análisis más profundo. El mismo deberá generar el convencimiento acerca de que el resultado alcanzado pueda ser interpretado como un indicador sobre la confianza o alerta del posible riesgo inherente o preexistente de los datos informatizados que son puestos a disposición del auditor al iniciar su tarea de contralor. En consecuencia, este estudio constituye una primera etapa de investigación, pudiendo ser ampliado y/o comparado con otras bases de datos, además de indagar sobre el perfil estadístico de los datos para permitir gestar conclusiones con otras aristas.

7. Referencias bibliográficas:

Benford, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society*. 78(4):551-572.

Burke, J. and E. Kincanon (1991). Benford's law and physical constants: the distribution of initial digits. *American Journal of Physics* 59, 952

Carslaw, C. A. P. N. 1988. Anomalies in income numbers: Evidence of goal oriented behavior. *The Accounting Review*. LXIII (2):321-327.

Castañeda, G. 2011. La ley de Benford y su aplicabilidad en el análisis forense de resultados electorales. *Scielo. Política y Gobierno*. Vol. 18 nº 2 Mexico

Etteridge M. L. and R. P. Srivastava. 1999. Using digital analysis to enhance data integrity. *Issues in Accounting Education*. 14(4):675-690.

Furlan, L. 1948. Das Harmoniegesetz der Statistik: Eine Untersuchung uber die metrische *Interdependenz der sozialen Erscheinungen, Basel, Switzerland* -G xiii: 504.

Golbeck J, 2015 "Benford's Law Applies to Online Social Networks," *PLOS ONE*, v. 10, nº 8

Hill, T. P. 1995. A statistical derivation of the significant digit law. *Statistical Science*. 10(4):354-363.

Nigrini, M. J. 1996. Taxpayer compliance application of Benford's law. *Journal of the American Taxation Association*. 18(1):72-92.

Nigrini, M. J. and L. J. Mittermaier. 1997. The use of Benford's law as an aid in analytical procedures. *Auditing: A Journal of Practice & Theory*. 16(2):52-67.

Nigrini, M. J. 1999. Adding value with digital analysis. *The Internal Auditor*. 56(1):21-23.

Nigrini, Mark and Miller Steven J., 2007. Benford's Law Applied to Hydrology Data—Results and Relevance to Other Geophysical Data. *International Association for Mathematical Geology*. Math Geol (2007) 39: 469–490

Nigrini, Mark Benford's Law: Applications for forensic accounting, auditing, and fraud detection, vol. 586. John Wiley & Sons, 2012

Normas Internacionales de Auditoría (NIA) (2014) 300 a 500.

Pepijn de Vries, Albertinaka J. 2013 Compliance of LC50 and NOEC data with Benford's Law: An indication of reliability? *Elsevier Ecotoxicology and Environmental Safety* 201 3.

Roukema, B. F. "Benford's Law anomalies in the 2009 Iranian presidential election," *Unpublished manuscript*, 2009.

Thomas, J. K. 1989. Unusual patterns in reported earnings. *The Accounting Review*. LXIV (4):773-787.

Varian, H. R. 1972. Benford's law. *The American Statistician*. 26:65-66.

Wallace, W. A. 2002. Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management*. (Fall) 51 (3):6-22.