IV CONGRESO INTERNACIONAL DE CONTABILIDAD, AUDITORÍA Y FINANZAS

“Pruebas de integridad de conjuntos de datos de alto volumen en empresas estatales argentinas: uso de la ley de Benford”

Autores:

Lic. Héctor Rubén Morales

Universidad Nacional de Córdoba- Argentina

Dra. Marcela Porporato

MSc. Nicolas Epelbaum

York University – Canadá

Setiembre 2020

Frank Benford 1938

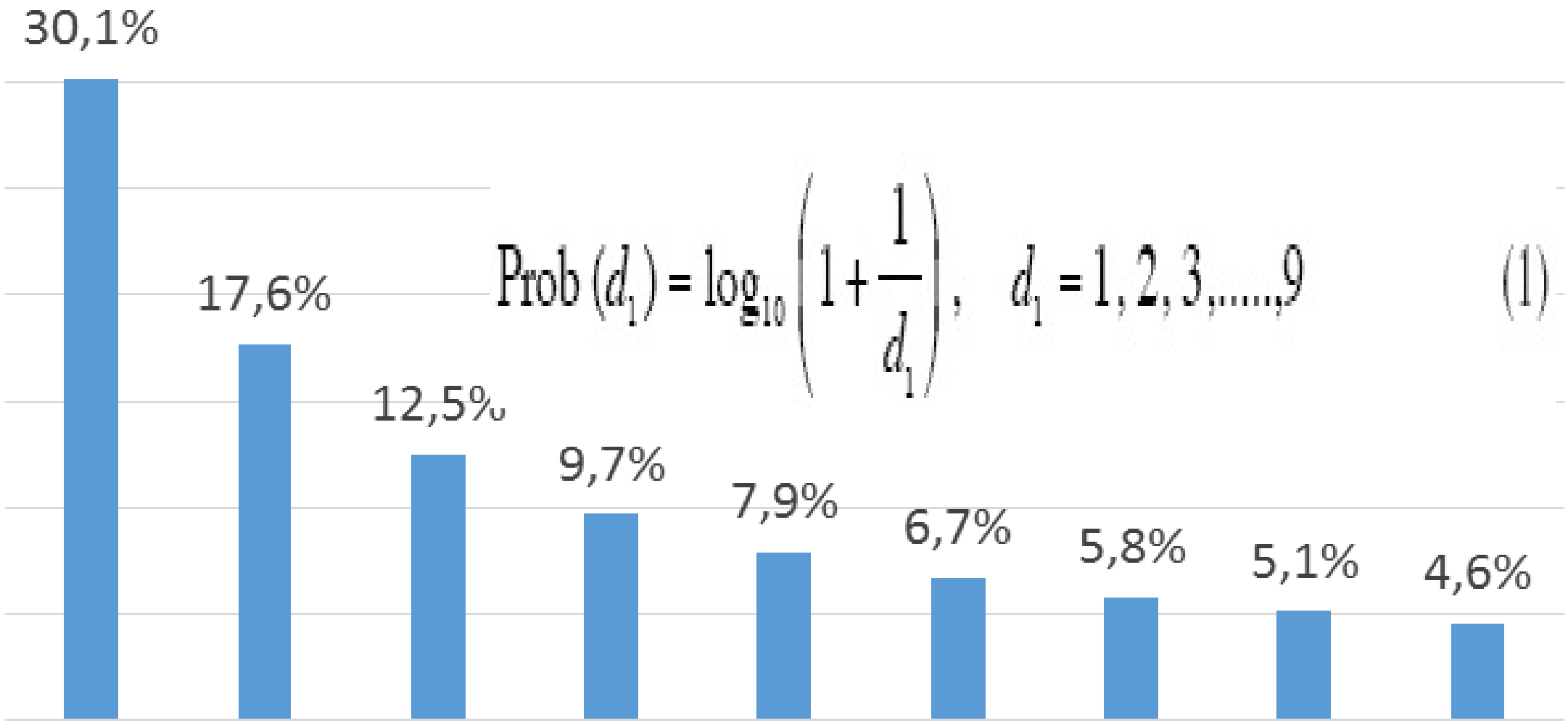
Benford 1938

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Ley de Benford - Primer dígito



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

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%

LEY DE BENFORD

LIMITANTES Y VENTAJAS DE LEY DE BENFORD

LIMITANTES

- MAGNITUDES MEDIBLES DE *UN MISMO FENÓMENO*
- LOS NÚMEROS *NO DEBEN ESTAR SUJETOS A UN RANGO*
- *NO* DEBEN SER *NÚMEROS ASIGNADOS O ALEATORIOS*
- EL ANÁLISIS SE AJUSTA PARA *CANTIDAD DE DATOS MAYORES A 1.000, AUNQUE ALGUNOS ESTUDIOS TRATAN A PARTIR DE 100 DATOS*

VENTAJAS

- ES *INVARIANTE A CAMBIOS DE ESCALAS* (EJ: $ a u$s, kms a millas)

Año	Autor	Aplicación Benford a otras disciplinas
1972	Varian	Datos bursátiles
1991	Burke y Kincanon	Constantes físicas
1996	Nigrini	Contabilidad - Datos Fiscales
2000	Tolle, Budzien	Dinámicas moleculares
2007	Jolion, Abdallhet	Imágnes digitales
2013	Pepijn de Vries	Indicador para evaluar riesgo toxicidad
2015	Golbeck	Seguidores redes sociales

FIBONACCI

7
14
21
35
56
91
147
238
385
623
1008
1631
2639
4270
6909
11179
18088
29267
47355
76622
123977
200599
324576
525175

AÑO	AUTOR	OPINION SOBRE LA LEY DE BENFORD
1942	Furlan	"Es la verdad de la naturaleza"
1972	Varian	"Puede servir de prueba de honestidad o validez de datos"
1995	Hill	"Fenómeno empírico, como lo es la dsitribución normal"
1996	Nigrini	"La falta de cumplimiento puede denotar fraude"
1999	Etteridge	"No fraude, sino ineficiencias operativas o fallas"

OBJETIVOS

- VERIFICAR SI EL PERFIL DE UNA BASE DE DATOS CUMPLE CON LA LEY DE BENFORD
- VERIFICAR SI CADA UNO DE LOS MODULOS DE LA BD CUMPLEN CON LA LEY DE BENFORD
- REVISAR LA POSIBLE UTILIDAD COMO INDICADOR DE RIESGO DE LA INFORMACIÓN

BASE DE DATOS BAJO ESTUDIO

- Empresa Provincial de Energía de Córdoba (EPEC)
- Actividad: Genera, distribuye y comercializa energía eléctrica
- Cuenta con 1,1 millones de clientes directos
- Abastece energía a 3,5 millones de habitantes
- Planta de personal de 3.100 empleados

Base de datos y sus modulos

- RRHH
- SUELDOS
- INVENTARIOS
- EXPEDIENTES
- SOLICITUDES INTERNAS
- SISTEMA COMERCIAL
- CONTABILIDAD
- GESTION OBRAS

Table	Schema	Last Analyzed	Num Rows	1º Dig
GL_BALANCES	GL	27/01/2020 8:29	132.772.203	1
GL_JE_LINES	GL	27/01/2020 8:31	85.699.120	8
GL_IMPORT_REFERENCES	GL	27/01/2020 8:29	77.868.343	7
GL_ACCOUNT_HIERARCHIES	GL	27/01/2020 8:27	8.438.293	8
GL_POSTING_INTERIM_206892	GL	27/01/2020 8:31	6.338.540	6
GL_INTERFACE	GL	27/01/2020 8:30	1.393.220	1
GL_POSTING_INTERIM_108398	GL	27/01/2020 8:31	876.783	8
GL_CODE_COMBINATIONS	GL	27/01/2020 8:29	870.200	8
GL_JE_HEADERS	GL	27/01/2020 8:30	447.370	4
GL_JE_BATCHES	GL	27/01/2020 8:30	422.840	4
GL_JOURNAL_REPORTS_ITF	GL	27/01/2020 8:31	324.553	3
GL_POSTING_INTERIM_174652	GL	27/01/2020 8:31	304.287	3
GL_DYNAMIC_SUMM_COMBINATIONS	GL	27/01/2020 8:29	201.807	2
GL_BC_PACKETS	GL	27/01/2020 8:29	174.760	1
GL_POSTING_INTERIM_2646	GL	27/01/2020 8:31	159.877	1
GL_POSTING_INTERIM_199732	GL	27/01/2020 8:31	128.137	1
GL_POSTING_INTERIM_209292	GL	27/01/2020 8:31	66.497	6
PRUEBA_636060	GL	27/01/2020 8:31	23.580	2
LINES_636060	GL	27/01/2020 8:31	23.287	2
GL_SEGMENT_FREQUENCIES	GL	27/01/2020 8:31	18.989	1
GL_BALANCES_DELTA	GL	27/01/2020 8:29	13.423	1
GL_DAILY_RATES	GL	27/01/2020 8:29	10.970	1
GL_PERIOD_STATUSES	GL	27/01/2020 8:31	8.316	8
GL_POSTING_INTERIM_170022	GL	27/01/2020 8:31	8.046	8
GL_DATE_PERIOD_MAP	GL	27/01/2020 8:29	7.305	7
GL_BC_PACKET_ARRIVAL_ORDER	GL	27/01/2020 8:29	6.775	6
GL_BC_PERIOD_MAP	GL	27/01/2020 8:29	6.143	6
GL_RECURRING_LINE_CALC_RULES	GL	27/01/2020 8:31	6.047	6
GL_ALLOC_HISTORY	GL	27/01/2020 8:27	3.554	3

MODULOS INFORMÁTICOS QUE CONFORMAN LA BASE DE DATOS

Módulo informático	Cantidad tablas	Total registros módulo
Contabilidad (CO)	78	225.245.009
Recursos Humanos (RH)	340	1.483.944
Inventario (IN)	71	14.916.459
Gestión Obras (GO)	58	380.572
Solicitudes Int. (SI)	87	1.832.473
Gestión Comercial (GC)	808	4.093.412.030
Liquidación sueldos (SU)	381	135.126.395
Adm. Expedientes (EX)	100	13.416.811
Total Base de Datos	1.923	4.485.813.693

MODULOS INFORMÁTICOS QUE CONFORMAN LA BASE DE DATOS

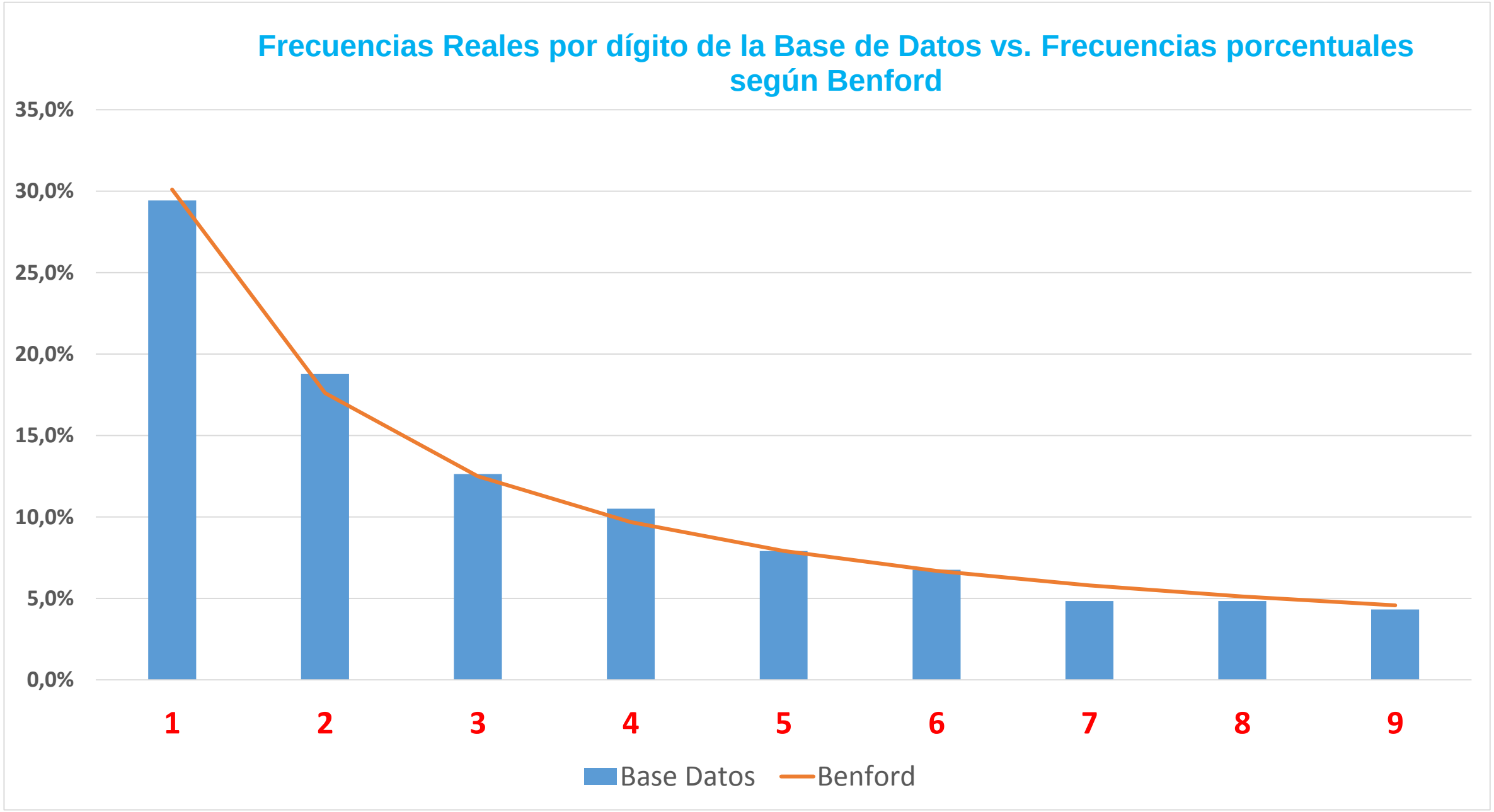
Módulo informático	Cantidad tablas	Tablas Módulo sobre Total	Registro del Módulo	Registros Módulo sobre Total	Promedio Registros por Módulo
Contabilidad (CO)	78	4,1%	225.245.009	5,02%	2.887.757
Recursos Humanos (RH)	340	17,7%	1.483.944	0,03%	4.365
Inventario (IN)	71	3,7%	14.916.459	0,33%	210.091
Gestión 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

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

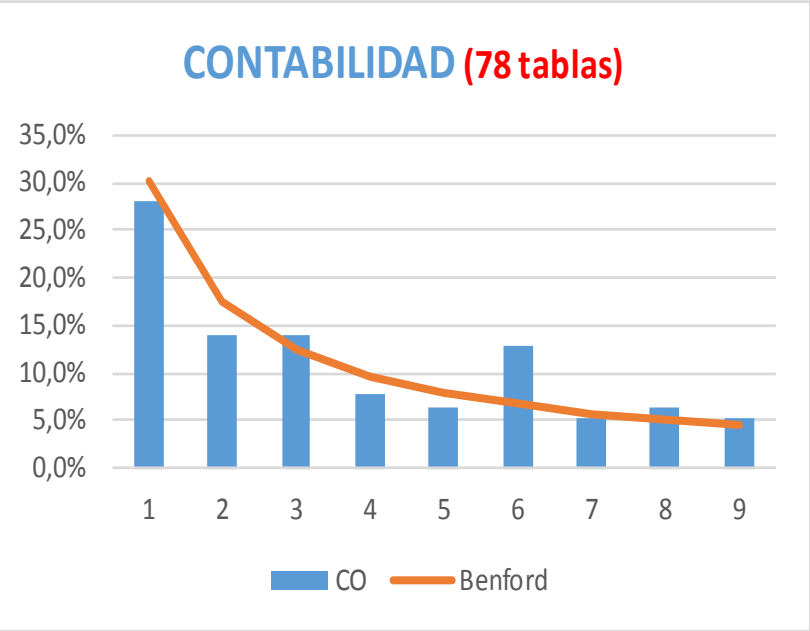
DIGITO	CONTABLE	RRHH	INVENTARIO	GESTION OBRAS	SOLICITUD INTERNAS	COMERCIAL	SUELDOS	EXPTES	Total
1	22	94	21	12	23	254	114	26	**566**
2	11	74	17	13	17	135	70	24	**361**
3	11	35	11	12	10	89	62	13	**243**
4	6	38	6	2	10	89	39	12	**202**
5	5	34	7	4	9	54	33	6	**152**
6	10	20	2	6	2	60	23	7	**130**
7	4	16	3	5	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**

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

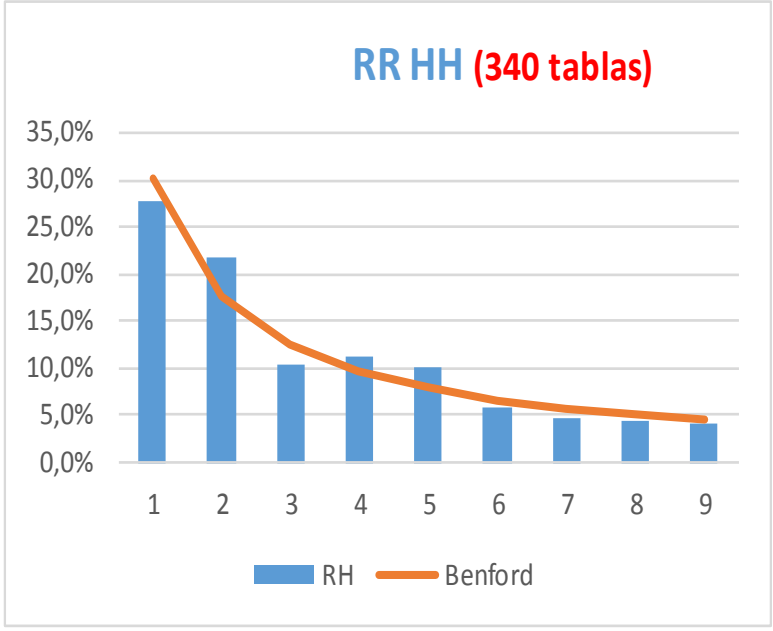
DIGITO	CO	RH	IN	GO	SI	GC	SU	EX	Total	Benford
1	28,2%	27,6%	29,6%	20,7%	26,4%	31,4%	29,9%	26,0%	29,4%	30,1%
2	14,1%	21,8%	23,9%	22,4%	19,5%	16,7%	18,4%	24,0%	18,8%	17,6%
3	14,1%	10,3%	15,5%	20,7%	11,5%	11,0%	16,3%	13,0%	12,6%	12,5%
4	7,7%	11,2%	8,5%	3,4%	11,5%	11,0%	10,2%	12,0%	10,5%	9,7%
5	6,4%	10,0%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


Frecuencias Reales por dígito de la Base de Datos vs. Frecuencias porcentuales según Benford
35,0%
30,0%
25,0%
20,0%
15,0%
10,0%
5,0%
0,0%
1
2
3
4
5
6
7
8
9
Base Datos
Benford

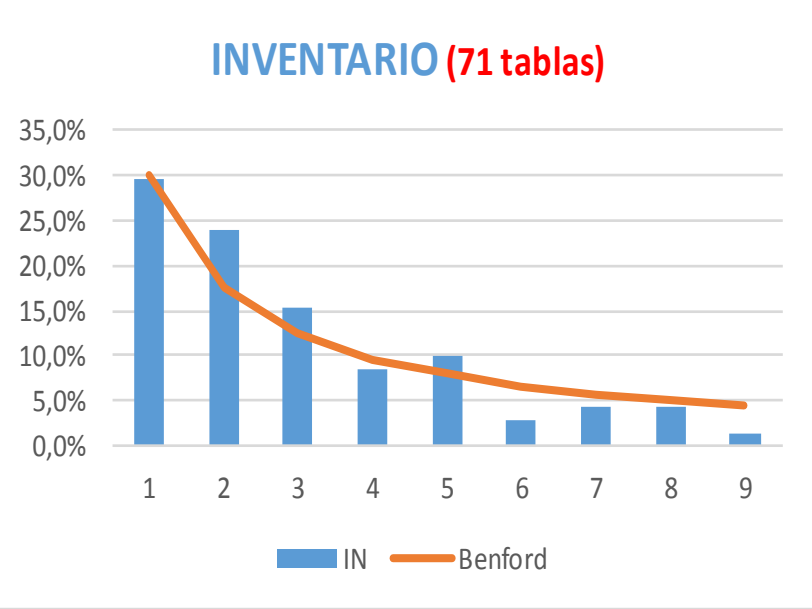
CO	Benford
28,2%	30,1%
14,1%	17,6%
14,1%	12,5%
7,7%	9,7%
6,4%	7,9%
12,8%	6,7%
5,1%	5,8%
6,4%	5,1%
5,1%	4,6%



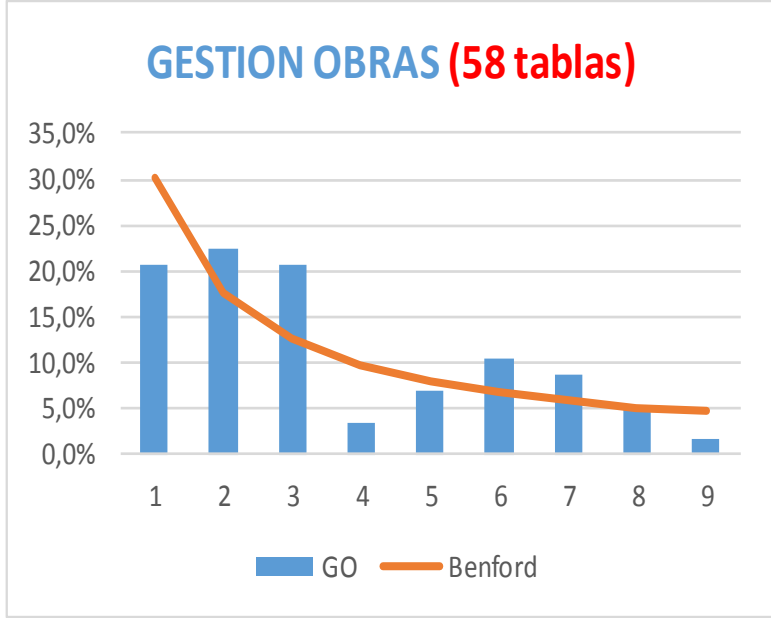
RH	Benford
27,6%	30,1%
21,8%	17,6%
10,3%	12,5%
11,2%	9,7%
10,0%	7,9%
5,9%	6,7%
4,7%	5,8%
4,4%	5,1%
4,1%	4,6%



IN	Benford
29,6%	30,1%
23,9%	17,6%
15,5%	12,5%
8,5%	9,7%
9,9%	7,9%
2,8%	6,7%
4,2%	5,8%
4,2%	5,1%
1,4%	4,6%

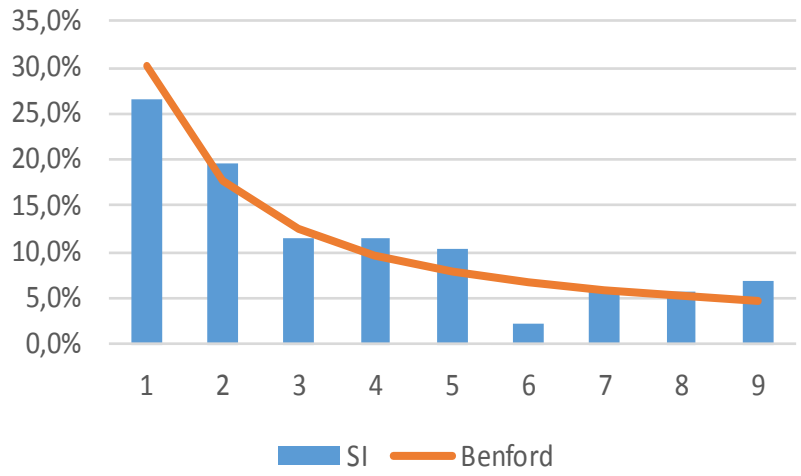


GO	Benford
20,7%	30,1%
22,4%	17,6%
20,7%	12,5%
3,4%	9,7%
6,9%	7,9%
10,3%	6,7%
8,6%	5,8%
5,2%	5,1%
1,7%	4,6%



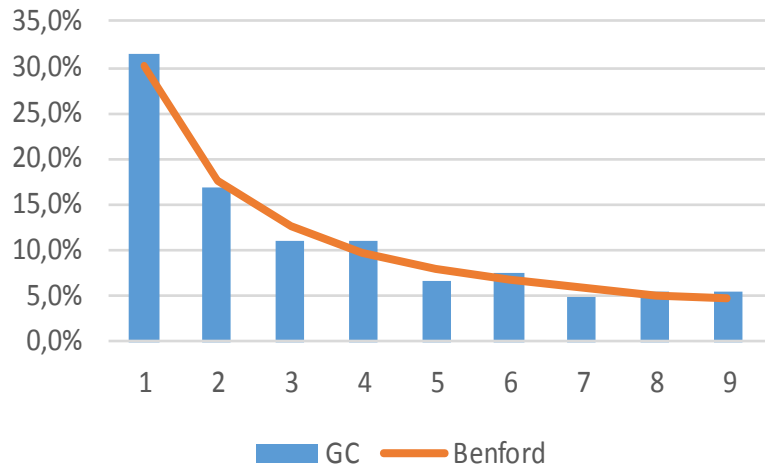
SI	Benford
26,4%	30,1%
19,5%	17,6%
11,5%	12,5%
11,5%	9,7%
10,3%	7,9%
2,3%	6,7%
5,7%	5,8%
5,7%	5,1%
6,9%	4,6%

SOLICITUDES INTERNAS (87 tablas)



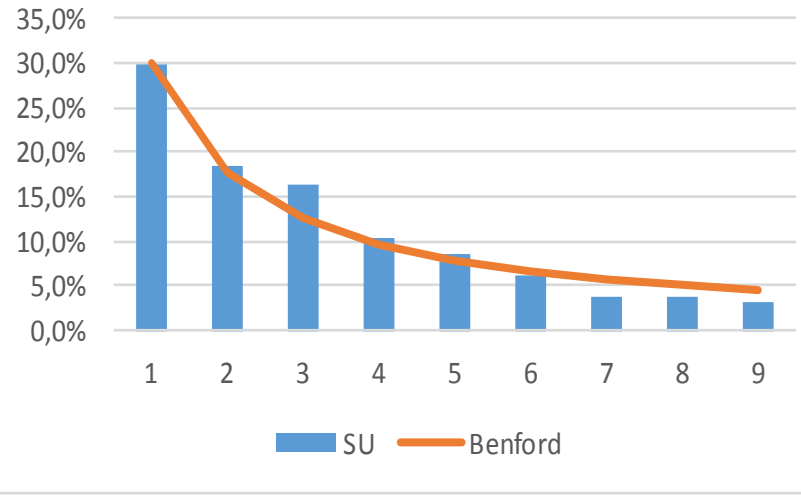
GC	Benford
31,4%	30,1%
16,7%	17,6%
11,0%	12,5%
11,0%	9,7%
6,7%	7,9%
7,4%	6,7%
4,8%	5,8%
5,6%	5,1%
5,3%	4,6%

GESTION COMERCIAL (808 tablas)



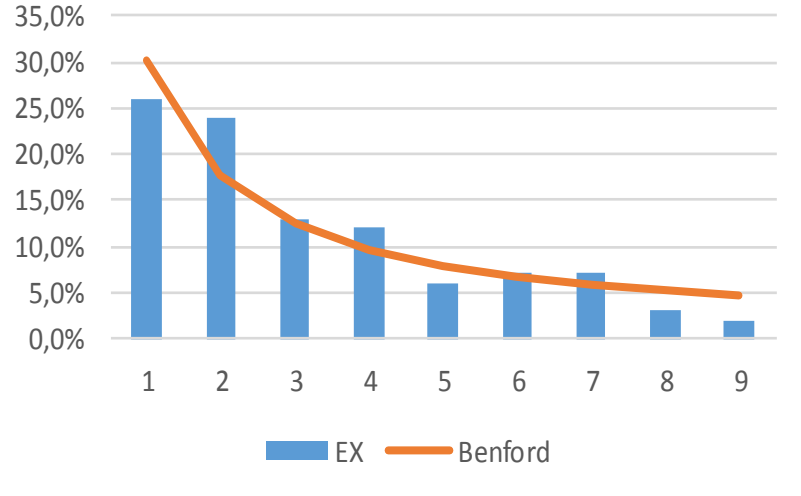
SU	Benford
29,9%	30,1%
18,4%	17,6%
16,3%	12,5%
10,2%	9,7%
8,7%	7,9%
6,0%	6,7%
3,7%	5,8%
3,7%	5,1%
3,1%	4,6%

SUELDOS (381 tablas)



EX	Benford
26,0%	30,1%
24,0%	17,6%
13,0%	12,5%
12,0%	9,7%
6,0%	7,9%
7,0%	6,7%
7,0%	5,8%
3,0%	5,1%
2,0%	4,6%

EXPEDIENTES (100 tablas)



PRUEBA BONDAD DE AJUSTE CHI CUADRADO

(verificar si los datos observados se ajustan con algún nivel de significancia a determinada distribución)

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

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)

	Prueba χ^2 (Chi cuadrado)								
	CO	RH	IN	GO	SI	GC	SU	EX	Total
χ^2	6,09	9,48	6,15	10,98	5,17	8,55	11,34	6,51	6,77
	Prueba χ^2 (Chi cuadrado): < 15,51 Acepta --> cumple distribución de Benford								

PRUEBA DE AJUSTE DESVIACIÓN MEDIA ABSOLUTA

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

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

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

CALCULO DEL MAD PARA LA BASE DE DATOS Y PARA 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%
MAD	0,0213	0,0172	0,0251	0,0434	0,0203	0,0102	0,0130	0,0238	0,0049	
Califica	*Bajo*	*Bajo*	*Bajo*	*Bajo*	*Bajo*	*Bueno*	*medio*	*Bajo*	*ALTO*	

Módulos informáticos ordenados en función a menor MAD y cantidad tablas

Módulo informático	Cantidad tablas	Total registros módulo	MAD Valor	MAD Conformidad
Total BD	1.923	4.485.813.693	0,0049	ALTA
G. COMERCIAL	808	4.093.412.030	0,0102	BUENA
SUELDOS	381	135.126.395	0,0130	MEDIA
RRHH	340	1.483.944	0,0172	BAJA
SOL. INTERNAS	87	1.832.473	0,0203	BAJA
CONTABILIDAD	78	225.245.009	0,0213	BAJA
EXPEDIENTES	100	13.416.811	0,0238	BAJA
INVENTARIO	71	14.916.459	0,0250	BAJA
G. OBRAS	58	380.572	0,0434	BAJA

POSIBLE INDICADOR DE RIESGO INHERENTE DE LOS DATOS

Conclusiones

	Verifica Benford?	Verifica χ^2	Verifica MAD	Causa	Solución Posible
Base Datos	SI	SI	SI		
Módulos	SI	SI	PARCIAL	Pocas tablas o datos	?

Deben ser +400 tablas

Ej: CONTABILIDAD

Mo	Cant. Datos
0	78
1	78
n	78

> 400 tablas

PROPUESTAS

- CONSIDERAR AL MAD COMO MEDIDA DEL RIESGO INHERENTE DE LOS DATOS

- ESTABLECER UN ORDEN ENTRE LOS MODULOS EN FUNCION AL RIESGO (MAD), QUE INDICA LA PRIORIDAD EN EL CONTROL DEL AUDITOR

Conclusiones e implicancias para la práctica profesional

Linea de Investigación	• mitigar o reducir la incertidumbre del auditor cuando se enfrenta a bases de datos de gran volumen
Contribución principal	• Ofrece un test simple y barato de ejecutar • genera un indicador que captura el riesgo inherente o preexistente en la información incluida en la bases de datos
Contribución especial	• El estudio se basa en una empresa pública donde la incidencia política es importante

IV CONGRESO INTERNACIONAL DE CONTABILIDAD, AUDITORÍA Y FINANZAS

Autores:

Lic. Héctor Rubén Morales

Universidad Nacional de Córdoba- Argentina

Dra. Marcela Porporato

MSc. Nicolas Epelbaum

York University - Canadá