Ley de Benford

Pablo Lorite Budiño

Ley de Benford

Pablo Lorite Budiño

Memoria presentada como parte de los requisitos para la obtención del título de Grado en Matemáticas por la Universidad de Sevilla.

Tutorizada por

Prof. José Fernando López Blázquez

Agradecimientos

A Fernando, mi tutor, por brindarme la oportunidad de hacer este trabajo con él y por contar con sus indicaciones, sugerencias y correcciones que han sido imprescindibles para la elaboración de este trabajo.

A mi familia, por su apoyo incondicional en mis decisiones y a lo largo de mi trayecto en este grado.

A todos mis amigos, a los de siempre y a los que se han ido incorporando a mi vida a lo largo de este trayecto, por siempre creer en mi y acompañarme en mi etapa universitaria.

Índice general

English Abstract

Benford's law is a statistical phenomenon that was documented for first time in the nineteenth century. This, also known as the Law of Significant Digits, is compared by many authors with Newton's Law of Gravitation because it is rather an observation of reality than a purely mathematical result. Due to this fact, in this work we will try to explain the most significant results in the most rigorous way possible and we will see some applications to data.

After the first time this phenomenon was observed, Benford's Law fell into oblivion due to its apparent lack of application in real life. Today, this phenomenon is being implemented every time in more diverse areas such as its most transcendental application, the detection of tax fraud, and in other fields such as data mining or creating models that can anticipate, for example, population growth.

Resumen

La ley de Benford es un fenómeno estadístico que se documentó por primera vez en el siglo XIX. Esta, también conocida por la Ley de los dígitos significativos, es comparada por muchos autores con la Ley de la Gravitación de Newton debido a que es más bien una observación de la realidad que puramente un resultado matemático. Debido a este hecho, en este trabajo intentaremos explicar los resultados más significativos de la manera más rigurosa posible y veremos algunas aplicaciones a datos reales de este curioso fenómeno.

Posteriormente a la primera vez que se observó este fenómeno, la Ley de Benford cayó en el olvido debido a su aparente poca aplicación en la vida real. A día de hoy, este fenómeno se está implementando cada vez en áreas más diversas como lo es su aplicación más trascendental, la detección de fraudes fiscales, y en otros campos como pueden ser la minería de datos o la creación de modelos que puedan prever, por ejemplo, el crecimiento demográfico.

1 El descubrimiento de la Ley de Benford

1.1 El artículo de Newcomb

Simon Newcomb (1835-1909) es considerado la primera persona en observar el fenómeno que posteriormente se llamará Ley de Benford o al menos la primera que publicó un artículo haciendo referencia a ello. Era un astrónomo americano altamente valorado en la época que destacó por su trabajo sobre teoría planetaria.

Newcomb no era matemático en el sentido estricto del término aunque esto no debería ser un hecho sorprendente debido a que una Ley como la de Benford, que tiene un carácter muy intuitivo llamaría más la atención en físicos o en científicos que trabajen en campos de matemáticas aplicadas y no tan teóricas.

En 1881, Newcomb, publica un breve artículo en el *American Journal of Mathematics* (ver [12]) en el cual se da cuenta del fenómeno a tratar: "*Los diez dígitos no ocurren con la misma frecuencia*", algo que podría ser evidente para alguien que use repetidamente las tablas de logaritmos (como era el caso de Newcomb). Además, se dio cuenta de que los dígitos más pequeños aparecían más repetidamente que dígitos mayores.

De acuerdo a sus observaciones, los números que tenían como primera cifra el uno se usaban mucho más que los demás, pero, ¿por qué? ¿no deberían aparecer todos equiprobablemente? Este fenómeno empieza a cobrar sentido cuando nos planteamos la naturaleza de estos números, los números utilizados por científicos como Newcomb o sus compañeros no son escogidos de una manera puramente aleatoria, son números que provienen de constantes físicas en la mayoría de las veces y que a se obtienen mediante procesos de derivación y/o computación.

Utilizando algunos argumentos heurísticos sobre los números, Newcomb llegó a la siguiente conclusión: "*La ley de probabilidad de la ocurrencia de los números es tal que la mantisa de sus logaritmos es equiprobable*"

Así llegamos a la siguiente conclusión: La probabilidad de que el primer dígito D_1 sea igual a d es:

$$\Pr[D_1 = d] = \log\left(1 + \frac{1}{d}\right)^{1}.$$

Aunque Newcomb no llegó explícitamente a esta formula era perfectamente consciente de ella ya que dio una tabla de probabilidades que daba detalladamente las probabilidades de los primeros dos dígitos (notados por D_1 y D_2 respectivamente) que obtuvo calculando a partir de las tablas de logaritmos:

Dig.	First Digit.	Second Digit.
0	. . .	0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

Figura 1.1: Tabla de probabilidades propuesta por Newcomb (ver [12]).

Mostramos en forma de diagrama de barras las probabilidades de los primeros y segundos dígitos dadas por Newcomb, vemos así de forma más visual que las probabilidades del segundo dígito están mucho más igualadas que las del primero. ¿Será esta la tendencia general al ir aumentando la posición del dígito en su mantisa? Newcomb responde también a esta cuestión de manera positiva, afirma así que de acuerdo con su ley de distribución de los dígitos, cuanto más aumente la posición del dígito en la mantisa la distribución tenderá a una distribución uniforme, es decir, todos los números son equiprobables.

[1] A partir de ahora cuando no especifiquemos la base del logaritmo nos referiremos al logaritmo decimal

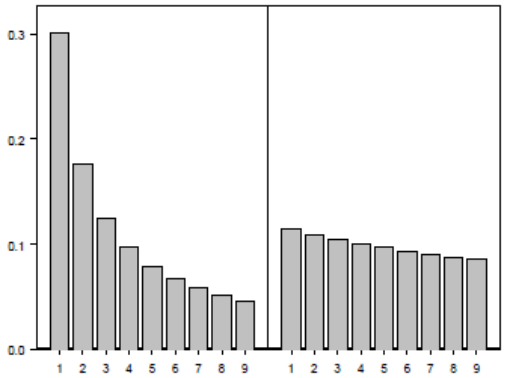


Figura 1.2: Diagrama de barras de probabilidades dadas por Newcomb correspondientes al primer y segundo dígito respectivamente.

1.2 El experimento de Benford

Cinco décadas después de la publicación del artículo de Newcomb el físico Frank Benford redescubrió este fenómeno en una publicación (ver [1]) en la cual apoyaba el hecho descubierto por Newcomb con mas de 20.000 datos de fuentes tan diversas como el área que cubren ríos en EE.UU., constantes físicas, los factoriales de los números naturales o estadísticas de la liga americana de béisbol (ver Figura 1.3).

A partir de estos datos Benford se dio cuenta de un hecho destacable: *los datos que mejor se ajustan a la ley logarítmica son los que tienen una naturaleza puramente aleatoria como por ejemplo los datos provenientes de periódicos o direcciones y no en cambio los que tienen una naturaleza más determinista como por ejemplo las raíces cuadradas o factoriales de números naturales.*

En la tabla publicada por Benford se muestra en la última fila el porcentaje medio de los datos de los primeros dígitos y su error probable del porcentaje para así ver como se ajustaban los datos a la ley logarítmica.

Aunque varios autores afirmaban que Benford manipuló los errores para que los datos se ajustasen mejor a la ley (ver [3, pág 363]), incluso los datos sin manipular se ajustaban de manera destacable lo cual llamó mucho la atención, tanto fue así que el autor pasó a darle nombre a la ley que nos atañe.

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	—

Figura 1.3: Tabla de Benford en su publicación *"The law of anomalous numbers"*.

1.3 Intentos de explicación de la Ley de Benford

Hubo varios intentos de explicar este curioso fenómeno matemático, el propio Benford intentó hacerlo estudiando el conjunto de los números naturales e intentando demostrar esta ley como una propiedad de los propios número naturales. Su demostración empezó por intentar demostrar que el subconjunto de los números naturales que tenían a 1 como primer dígito tienen una probabilidad de ocurrir de $log(2)$ sobre el conjunto de los naturales. Benford (y muchos matemáticos que trabajaron en esta demostración posteriormente) encontraron el mismo problema:

$$\nexists \lim \frac{1}{n} \#\{i \in \mathbb{N} | i \leq n \wedge d_1(i) = 1\}.$$

Para ver más claramente el comportamiento de esta sucesión (extrapolando entre naturales) representémoslo gráficamente (ver Figura 1.4). Vemos que la sucesión es oscilante y tanto su límite superior como inferior no son constantes aunque el límite inferior tiende a $\frac{1}{9}$ y el superior a $\frac{5}{9}$.

Hay muchas maneras de definir el límite de una sucesión para llegar al deseado $log(2)$ pero no todas nos llevan a la Ley de Benford, por lo tanto es más que evidente que no podemos llegar a una demostración rigurosa de que el límite buscado cumpla la ley logarítmica. Además, queremos probar que la ley es válida para el sistema de números completo, es decir que es completamente universal. Esto nos llevará a razonamientos un tanto dudosos ya que hay muchísimos subconjuntos de números "naturales"[2] que no siguen la la ley a demostrar.

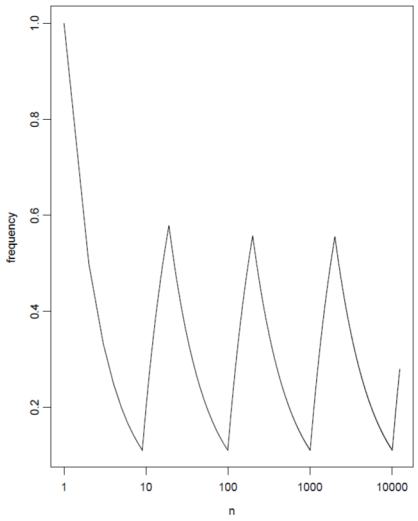


Figura 1.4: Frecuencia aparición de números con primer dígito uno ([8, pág 9]).

Uno de los mayores descubrimientos se atribuye al matemático Pinkham (ver [14]) el cual observó que si la Ley de Benford es una ley universal, los conjuntos de datos que la cumplan deberán satisfacerla también mediante una variación de escala, así se llegó a una de las propiedades más importantes de estos conjuntos de números, la invarianza de escala (tratada en capítulos posteriores).

Hubo además muchos problemas (como por ejemplo la no invarianza de escala de probabilidad en los Borels) en los razonamientos para hacer una demostración sobre esta "ley universal" lo que hizo que no se llegase a demostrar rigurosamente. Aún así, si era conocido que había muchos conjuntos de datos que la seguían.

[2]Refiriéndonos al sentido que le dio Newcomb a esta palabra y no al conjunto de los números naturales, es decir, números que ocurren en la naturaleza.

Las últimas aportaciones sobre la ley las hizo Theodore Hill (ver [5], [6], [7]), el cual estableció un marco probabilístico correcto, en la década de los 90, en el cual la ley era aplicable. La idea que sugirió era que los datos que cumplían la ley eran una mezcla de distintas distribuciones, lo que concuerda con el experimento de Benford y explica rigurosamente la invarianza de escala y no solo eso, sino que extiende esta propiedad a invarianza de base.

2 Resultados matemáticos

2.1 Distribuciones Benford

2.1.1 La función de distribución del k-ésimo dígito

En primer lugar, estudiemos la distribución del primer dígito como caso base y posteriormente extendámosla al caso genérico del k-ésimo dígito.

Definición 2.1. *La función de probabilidad (discreta) logarítmica para el primer dígito D_1 está definida por:*

$$\Pr[D_1 = d_1] = \log\left(1 + \frac{1}{d_1}\right) \tag{2.1}$$

donde $d_1 \in \{1, ..., 9\}$.

La función de distribución será fácilmente calculable utilizando (2.1):

$$\begin{aligned}
\Pr[D_1 \leq d_1] &= \sum_{1 \leq d_1' \leq d_1} \Pr[D_1 \leq d_1'] \\
&= \sum_{1 \leq d_1' \leq d_1} \log\left(1 + \frac{1}{d_1'}\right) \\
&= \log\left(\prod_{1 \leq d_1' \leq d_1} \left(1 + \frac{1}{d_1'}\right)\right) \\
&= \log\left(\left(1 + \frac{1}{1}\right)\left(1 + \frac{1}{2}\right) \cdots \left(1 + \frac{1}{d_1}\right)\right) \\
&= \log\left(\frac{2}{1} \cdot \frac{3}{2} \cdots \frac{d+1}{d}\right) \\
&= \log(d+1)
\end{aligned} \tag{2.2}$$

donde $d_1 \in \{1, ..., 9\}$.

Veamos ahora la función de distribución conjunta de los k primeros dígitos:

Definición 2.2. *La función de probabilidad conjunta de la distribución logarítmica de los primeros k dígitos significativos $D_1, D_2, \ldots, D_k$ (para todo $k \in \mathbb{N}$) queda definida como:*

$$\Pr[D_1 = d_1, \ldots, D_k = d_k] = \log\left(1 + \left(\sum_{i=1}^{k} 10^{k-i} d_i\right)^{-1}\right) \tag{2.3}$$

donde $d_1 \in \{1, \ldots, 9\}$ y los demás $d_j \in \{0, \ldots, 9\}$.

La función de probabilidad discreta para el k-ésimo dígito D_k queda calculada utilizando 2.3

$$\begin{aligned}
\Pr[D_k = d_k] &= \sum_{\substack{1 \leq d_1 \leq 9 \\ 0 \leq d_2 \leq 9 \\ \cdots \\ 0 \leq d_{k-1} \leq 9}} \Pr[D_1 = d_1, D_2 = d_2, \ldots, D_k = d_k] \\
&= \sum_{\substack{1 \leq d_1 \leq 9 \\ 0 \leq d_2 \leq 9 \\ \cdots \\ 0 \leq d_{k-1} \leq 9}} \log\left(1 + \left(\sum_{i=1}^{k} 10^{k-i} d_i\right)^{-1}\right).
\end{aligned} \tag{2.4}$$

Es de notable importancia ver que el si tenemos una familia de variables aleatorias que siguen la ley logarítmica, estas no serán mutuamente independientes, ya que, por ejemplo:

$$\begin{aligned} 0.035 = \Pr[D_1 = 1, D_2 = 2] &\neq \Pr[D_1 = 1] \cdot \Pr[D_2 = 2] \\ &= \log(2) \cdot (\log(\frac{13}{12}) + \log(\frac{23}{22}) + \cdots + \log(\frac{93}{92})) = 0.033. \end{aligned}$$

2.1.2 La distribución mantisa

En esta sección generalizaremos la función de distribución logarítmica de los k primeros dígitos significativos a una versión continua para la mantisa[1] M de la siguiente forma:

Lema 2.1. *La función de distribución de la mantisa de una variable continua que siga la Ley de Benford tiene la siguiente forma:*

$$\Pr[M \leq m] = \log(m)$$

donde $m \in [1, 10)$.

Demostración. Distingamos distintos casos:

- Primero consideremos el caso en el que $m = d_1$, es decir, m solo tiene una cifra significativa

$$\Pr[M \leq m] = \begin{cases} 0 & \text{si } d_1 \leq 1 \\ \Pr[D_1 \leq d_1 - 1] \underset{(2.2)}{=} \log(d_1) = \log(m) & \text{si } 1 < d_1 \leq 10 \end{cases}$$

- En segundo lugar supongamos $m = d_1 d_2 \dots d_k$ en notación científica, es decir, $m = \sum_{i=1}^{k} 10^{-(i-1)} d_i$ con $d_1 > 1, d_2 > 0, \dots, d_k > 0$:

[1] Definimos la mantisa como el significando de un número en notación científica

$$
\begin{aligned}
\Pr[M \leq m] &= \Pr[D_1 \leq d_1 - 1] \\
&\quad + \Pr[D_1 = d_1, D_2 \leq d_2 - 1] \\
&\quad + \cdots \\
&\quad + \Pr[D_1 = d_1, D_2 = d_2, \ldots, D_k \leq d_k] \\
&= \Pr[D_1 \leq d_1 - 1] \\
&\quad + \sum_{0 \leq d_2' \leq d_2 - 1} \Pr[D_1 = d_1, D_2 = d_2'] \\
&\quad + \ldots \\
&\quad + \sum_{0 \leq d_k' \leq d_k - 1} \Pr[D_1 = d_1, D_2 = d_2 - 1, \cdots, D_k = d_k'] \qquad (2.5) \\
&= \log(d_1) \\
&\quad + \sum_{0 \leq d_2' \leq d_2 - 1} \log\left(1 + \frac{1}{10d_1 + d_2'}\right) \\
&\quad + \ldots \\
&\quad + \sum_{0 \leq d_k' \leq d_k - 1} \log\left(1 + \left(\sum_{i=1}^{k-1} 10^{k-i} d_i + d_k'\right)^{-1}\right).
\end{aligned}
$$

Utilizando el razonamiento de (2.2) llegamos a:

$$
\begin{aligned}
\Pr[M \leq m] &= \log(d_1) + \log\left(\frac{10d_1 + d_2}{10d_1}\right) + \cdots + \log\left(\frac{\sum_{i=1}^{k} 10^{k-i} d_i}{\sum_{i=1}^{k-1} 10^{k-i} d_i}\right) \\
&= \log\left(\frac{\sum_{i=1}^{k} 10^{k-i} d_i}{10^{k-1}}\right) = \log\left(\sum_{i=1}^{k} 10^{-(i-1)} d_i\right) = \log(m).
\end{aligned}
$$

- Ahora, estudiemos el caso en el que uno o más de los d_j sean nulos ($j > 0$) o $d_1 = 1$:
  Supongamos entonces $m = d_1 d_2 \ldots d_{j-1} 0 d_{j+1} \ldots d_k$. Entonces el término j-ésimo de la primera igualdad de (2.5) es nulo. Llegamos así a la siguiente expresión:

$$\begin{aligned}
\Pr[M \leq m] &= \log(d_1) + \cdots + \log\left(\frac{\sum_{i=1}^{j-1} 10^{j-1-i} d_i}{\sum_{i=1}^{j-2} 10^{j-1-i} d_i}\right) + 0 \\
&\quad + \log\left(\frac{\sum_{i=1}^{j+1} 10^{j-1-i} d_i}{\sum_{i=1}^{j} 10^{j-1-i} d_i}\right) + \cdots + \log\left(\frac{\sum_{i=1}^{k} 10^{k-i} d_i}{\sum_{i=1}^{k-1} 10^{k-i} d_i}\right) \\
&= \log\left(\frac{1}{10^{j-2}} \cdot \frac{\sum_{i=1}^{j-1} 10^{j-1-i} d_i}{\sum_{i=1}^{j} 10^{j+1-i} d_i} \cdot \frac{1}{10^{k-j-1}} \cdot \sum_{i=1}^{k} 10^{k-i} d_i\right) \\
&= \log\left(\frac{1}{10^{j-2}} \cdot \frac{1}{10^2} \cdot \frac{1}{10^{k-j-1}} \cdot \sum_{i=1}^{k} 10^{k-i} d_i\right) \\
&= \log\left(\frac{1}{10^{k-1}} \cdot \sum_{i=1}^{k} 10^{k-i} d_i\right) = \log(m).
\end{aligned}$$

Esto puede ser generalizado fácilmente para el caso en el que varios d_j sean nulos simultáneamente o que $d_1 = 1$.

Observación 2.1. De hecho, habiendo demostrado el lema anterior podemos llegar como consecuencia a la fórmula (2.3):

$$\begin{aligned}
\Pr[D_1 = d_1, D_2 = d_2, \ldots, D_k = d_k] &= \Pr[d_1 + d_2 10^{-1} + \ldots + d_k 10^{-k+1} \leq M < \\
&\qquad d_1 + d_2 10^{-1} + \ldots + (d_k + 1) 10^{-k+1}] \\
&= \log(d_1 + d_2 10^{-1} + \ldots + (d_k + 1) 10^{-k+1}) \\
&\quad - \log(d_1 + d_2 10^{-1} + \ldots + d_k 10^{-k+1}) \\
&= \log\left(\frac{d_1 + d_2 10^{-1} + \ldots + (d_k + 1) 10^{-k+1}}{d_1 + d_2 10^{-1} + \ldots + d_k 10^{-k+1}}\right) \\
&= \log\left[1 + \left(\sum_{i=1}^{k} 10^{k-i} d_i\right)^{-1}\right].
\end{aligned}$$

2.1.3 Generación de variables aleatorias Benford

Definición 2.3. *Una variable aleatoria X satisface la Ley de Benford para la mantisa si $M = M(X)$ sigue la distribución logarítmica de la mantisa.*

Definición 2.4. *Una variable aleatoria X satisface la Ley de Benford para el k-ésimo dígito si $D_k = D_k(X)$ sigue la distribución logarítmica del k-ésimo dígito.*

Es tarea fácil generar variables aleatorias que sigan distribuciones Benford haciendo uso de la función de distribución para la mantisa M:

$$\Pr[M \leq m] = \log(m); \qquad m \in [1, 10).$$

Entonces, podemos generar una variable aleatoria Benford con:

$$M \leftarrow 10^U \quad \text{con} \quad U \backsim \mathcal{U}(0, 1).$$

Esto es evidente ya que:

$$\Pr[M \leq m] = \Pr[10^U \leq m] = \Pr[U \leq \log(m)] = \log(m).$$

Como consecuencia, tenemos que la variable aleatoria D_1 que sigue la distribución Benford del primer dígito puede ser generada con:

$$D_1 \leftarrow \lfloor 10^U \rfloor.$$

Los métodos de generación de variables anteriores justifican el hecho que descubrió Newcomb, es decir, que la mantisa de los logaritmos de los números están uniformemente distribuidas.

Podemos concluir entonces lo siguiente: Sea X la variable aleatoria que represente el número cuyo logaritmo estamos buscando, sea M su mantisa y por último sea S la variable aleatoria entera tal que $10^S \leq X < 10^{S+1}$.

Por lo tanto $M = 10^{\log X - S}$, y de acuerdo con el método de generación de variables Benford, si $\log X - S \backsim \mathcal{U}$, entonces M seguirá la distribución Benford de la mantisa y viceversa.

Observación 2.2. Concluimos así que una variable aleatoria X satisface la ley de Benford si y solo si $\log X - \lfloor \log X \rfloor \backsim \mathcal{U}(0, 1)$.

2.1.4 Distribuciones que siguen exactamente la Ley de Benford

Una distribución se dice que cumple la Ley de Benford si su función de distribución correspondiente lo hace. Es relativamente fácil encontrar "distribuciones Benford":

Supongamos, por ejemplo, que buscamos una distribución X en $[1, 10)$ que satisfaga la ley para el primer dígito y supongamos f_X su función de densidad. Entonces para $d \in \{1, \dots, 9\}$ tenemos que imponer lo siguiente:

$$\Pr[D_1 = d] = \Pr[d \leq X < d+1] = \int_d^{d+1} f_X(x)\, dx = \log\left(\frac{d+1}{d}\right).$$

Por lo tanto, una manera lógica de escoger f_X es $f_X(x) = \frac{1}{x \cdot \ln 10}$ lo cual concuerda con el método de generación de la sección anterior porque en este caso tenemos $X = 10^U$ con $U \backsim \mathcal{U}(0, 1)$. Esto nos da una pista de como encontrar más distribuciones Benford, por ejemplo: buscamos una distribución en $[10^a, 10^b)$, es natural pensar en la distribución 10^U con $U \backsim \mathcal{U}(a, b)$ que tiene como función de densidad $f_X(x) = \frac{1}{(b-a) \cdot x \cdot \ln 10}$. Tenemos entonces:

$$\begin{aligned}
\Pr[D_1 = d] &= \Pr[10^a \cdot d \leq X < 10^a \cdot (d+1)] \\
&\quad + \Pr[10^{a+1} \cdot d \leq X < 10^{a+1} \cdot (d+1)] \\
&\quad + \cdots \\
&\quad + \Pr[10^{b-1} \cdot d \leq X < 10^{b-1} \cdot (d+1)] \\
&= \frac{1}{b-a} \cdot \left(\log\left(\frac{10^a \cdot (d+1)}{10^a \cdot d}\right) \right. \\
&\quad + \log\left(\frac{10^{a+1} \cdot (d+1)}{10^{a+1} \cdot d}\right) \\
&\quad + \dots \\
&\quad \left. + \log\left(\frac{10^{b-1} \cdot (d+1)}{10^{b-1} \cdot d}\right) \right) \\
&= \frac{1}{b-a} \cdot \left((b-a) \cdot \log\left(\frac{d+1}{d}\right) \right) \\
&= \log\left(\frac{d+1}{d}\right).
\end{aligned}$$

De esta manera podemos encontrar varios ejemplos de distribuciones Benford. Estas son de la forma 10^W con W una variable aleatoria con soporte entre números enteros. Veamos los siguientes ejemplos:

1. Sea $W \backsim \text{TRIANGULAR}(0, 1, 2)$ a la cual le corresponde la siguiente función de densidad:

$$f_W(w) = \begin{cases} w & \text{si} \quad 0 \leq w \leq 1 \\ 2 - w & \text{si} \quad 1 < w \leq 2 \end{cases}$$

Utilizando la observación 2.2, verifiquemos que $Z = W - \lfloor W \rfloor \backsim \mathcal{U}(0, 1)$:

$$\begin{aligned} F_Z(z) &= \Pr[0 \leq W < 1] \cdot \Pr[W \leq z \,|0 \leq W < 1] \\ &\quad + \Pr[1 \leq W < 2] \cdot \Pr[W - 1 \leq z \,|1 \leq W < 2] \\ &= \int_0^z w \, dw + \int_1^{z+1} (2 - w) \, dw \\ &= \frac{1}{2}z^2 + \frac{1}{2}(2z - z^2) = z. \end{aligned}$$

Por lo tanto, $Z \backsim \mathcal{U}(0, 1)$, luego $X = 10^W$ es Benford.
Podemos generalizar esto para el caso con $W \backsim \text{TRIANGULAR}(a, b, c)$ con $a < b < c$ y $a, b, c \in \mathbb{Z}$ (ver [10]).

2. Sea ahora W con la siguiente función de densidad:

$$f_W(w) = \begin{cases} 1 - w^2 & \text{si} \quad -1 < w < 0 \\ (w - 1)^2 & \text{si} \quad 0 \leq w < 1 \end{cases}$$

Veamos que la distribución $X = 10^W$ es Benford utilizando el método anterior:

$$\begin{aligned} F_Z(z) &= \Pr[-1 \leq W < 0] \cdot \Pr[W + 1 \leq z \mid -1 \leq W < 0] \\ &\quad + \Pr[0 \leq W < 1] \cdot \Pr[W \leq z \,|0 \leq W < 1] \\ &= \int_{-1}^{z-1} (1 - w^2) \, dw + \int_0^z (w - 1)^2 \, dw \\ &= \left(-\frac{z^3}{3} + z^2\right) + \left(\frac{z^3}{3} - z^2 + z\right) \\ &= z. \end{aligned}$$

De nuevo, $Z \backsim \mathcal{U}(0, 1)$, luego $X = 10^W$ es Benford.

Esto también es generalizable para W con la siguiente función de densidad:

$$f_W(w) = \begin{cases} 1 - w^n & \text{si} \quad -1 < w < 0 \\ (w-1)^n & \text{si} \quad 0 \leq w < 1 \end{cases}$$

Para $n \in \mathbb{Z}^+$ par (ver [10]).

2.2 Convergencia a la distribución uniforme

La razón por la que la Ley de Benford es útil solo para los primeros dígitos es porque la distribución del k-ésimo dígito tiende a la distribución uniforme discreta en $\{0, \ldots, 9\}$ a velocidad exponencial cuando k crece. Este fenómeno ya fue observado por Newcomb en su artículo (ver sección 1.1) aunque fue Hill el que conjeturó más formalmente acerca de ello.

Conjetura 2.1. *Sea X una variable aleatoria continua con función de densidad acotada por partes infinitamente derivable. Entonces, la distribución $D_k(X)$ se aproxima a la distribución uniforme cuando k tiende a infinito.*

Demostración. Aunque no existe una demostración formal como tal, si podemos ofrecer al lector una idea de demostración informal.

Sin perdida de generalidad, supongamos que tratamos con números en base 2.

Escogemos un intervalo, sea este $[1, 2]$ y dibujemos en este una función continua acotada, por ejemplo, una exponencial en ese intervalo (corresponderá a la función de densidad de nuestra variable aleatoria). En base 2 nuestro intervalo es $[1, 10]$. Sobre todos los números reales en $[1, 2]$ aquellos que tienen como segundo dígito el 0 (en binario) son los que están entre 1.00 y 1.10 (1 y 1.5 en notación decimal). Aquellos que tienen como tercer dígito el 0 (en binario) son los que están entre 1.00 y 1.01 (1 y 1.25 en notación decimal) y entre 1 y 1.11 (1 y 1.75 en notación decimal), etc...

Cuando el proceso va hasta el infinito, se dividirá en la mitad la integral completa, y por lo tanto, la probabilidad de observar un 0 será la misma que la probabilidad de observar un 1 (distribución uniforme discreta en $\{0, 1\}$).

Gráficamente vemos este proceso de división en las siguientes imágenes y vemos como cuanto más vayamos haciendo crecer la k, la división será más cercana a 1/2 (ver Figura 2.1).

Hemos escogido números en base 2 y este intervalo por simplicidad, aunque de manera análoga se haría el razonamiento para números en base 10 aunque en este caso tendríamos 10 particiones.

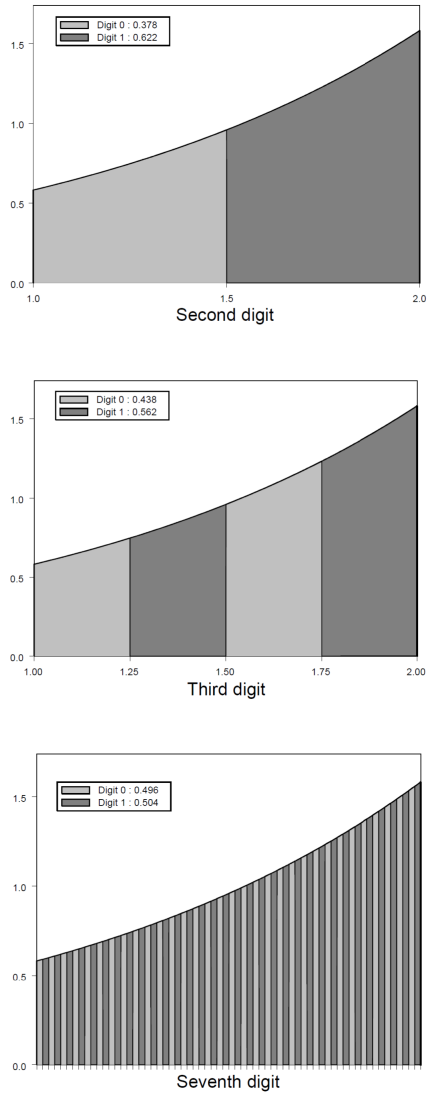


Figura 2.1: Convergencia distribución uniforme (ver [8]).

Aun teniendo esta demostración informal, hemos comprobado numéricamente el comportamiento de ocho distribuciones diferentes para ver las probabilidades de aparición de un cero en la k-ésima posición (distribuciones condicionadas entre $[1, 2]$ con dígitos binarios, es decir, en base 2):

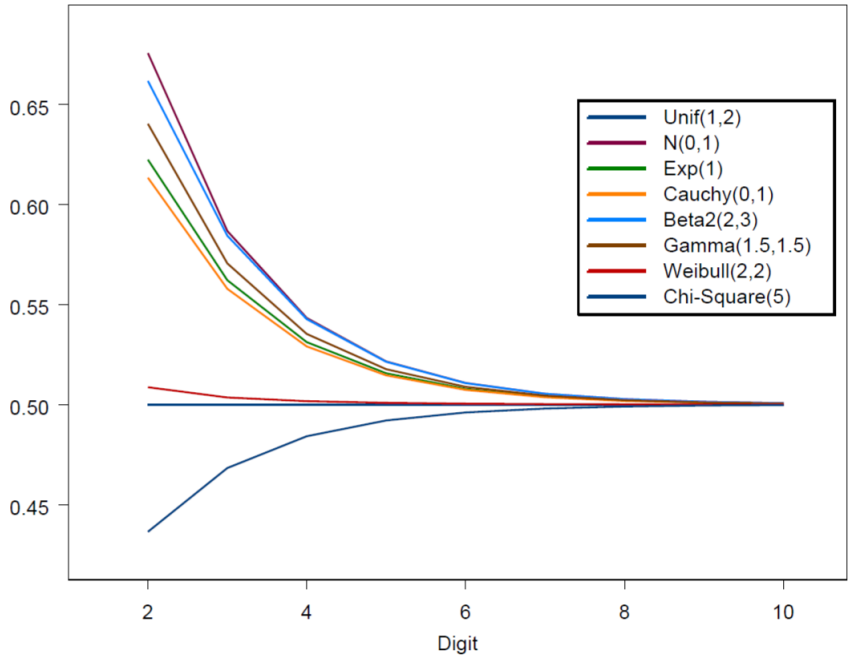


Figura 2.2: Gráfica de aparición de ceros en la k-ésima posición en distribuciones condicionadas en $[1, 2]$ con dígitos binarios (ver [8]).

2.3 Propiedades de las distribuciones Benford

2.3.1 Definición y propiedades del espacio de probabilidad

Nuestra primera tarea para hacer un estudio más exhaustivo de las propiedades de las distribuciones Benford es describir apropiadamente el espacio probabilístico en el que vamos a trabajar. Esta tarea es necesaria ya que en el espacio en el que trabajamos normalmente, es decir, el conjunto de los borelianos en $\mathbb{R}$ nos da ciertos problemas. Tenemos por ejemplo que cada número en $[10, 20)$, $[100, 200)$ o $[1000, 2000)$ tiene su mantisa en $[1, 2)$, por lo tanto no podemos asignar una probabilidad a un único boreliano, cuando todo un conjunto de ellos tienen la misma mantisa.

Es fácil ver entonces que el conjunto de números reales $\bigcup_{n=-\infty}^{\infty} B \cdot 10^n$ contiene todos los números reales positivos cuya mantisa pertenece a B con B boreliano en $[1, 10)$. Asignaremos así las probabilidades a estos borelianos.

Definición 2.5. *La función (decimal) mantisa $S : \mathbb{R} \to [1, 10)$ se define como sigue: Para todo $x \neq 0$ se tiene $S(x) = t$, con t el único número real en $[1, 10)$ que cumpla $|x| = 10^k \cdot t$, para algún (único) $k \in \mathbb{Z}$. Además, $S(0) = 0$ por convenio.*

Explicitamente,

$$S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor}$$

para $x \neq 0$.

Sea una función dada $f : \Omega \longrightarrow \mathbb{R}$. Recordamos que para todo subconjunto $C \in \mathbb{R}$, el conjunto $f^{-1}(C) \subset \Omega$, llamado preimagen de C por f, se define como:

$$f^{-1}(C) = \{w \in \Omega : f(w) \in C\}.$$

El σ-álgebra generado por la colección

$$\mathcal{E} = \{f^{-1}(J) : J \subset \mathbb{R}, J \text{un intervalo}\}$$

se referirá también al σ-álgebra generado por f, $\sigma(f)$.

Definición 2.6. *Dada una colección de funciones $\mathcal{F}$ con sus elementos $f : \Omega \longrightarrow \mathbb{R}$, se define el σ-álgebra de la familia de funciones como:*

$$\sigma(\mathcal{F}) = \sigma\left(\bigcup_{f \in \mathcal{F}} \sigma(f)\right) = \sigma(\{f^{-1}(J) : J \subset \mathbb{R}, J \textit{ un intervalo}, f \in \mathcal{F}\}).$$

Definición 2.7. *Definimos así el espacio de eventos $\mathcal{M}$, llamado álgebra mantisa, como:*

$$\mathcal{M} = \left\{ \bigcup_{n=-\infty}^{\infty} B \cdot 10^n \textit{ para todo Borel } B \subseteq [1, 10) \right\}.$$

Puede comprobarse fácilmente que $\mathcal{M}$ es una σ-álgebra.

Observación 2.3. Definimos entonces el espacio medible apropiado para nuestro estudio como la tupla $(\mathbb{R}^+, \mathcal{M})$ siendo $\mathcal{M}$ la álgebra mantisa definida anteriormente.

Observación 2.4. Notamos así la álgebra mantisa $\mathcal{M}$ como el σ-álgebra en $\mathbb{R}^+$ generado por la función mantisa, i.e. $\mathcal{M} = \mathbb{R}^+ \cap \sigma(S)$.

Teorema 2.1. *Para todo $A \in \mathcal{M}$,*

$$A = \bigcup_{k\in\mathbb{Z}} 10^k S(A)$$

donde $S(A) = \{S(x) : x \in A\} \subset [1, 10)$. Además,

$$\mathcal{M} = \mathbb{R}^+ \cap \sigma(D_1, D_2, \ldots) = \left\{ \bigcup_{k\in\mathbb{Z}} 10^k B : B \in \mathcal{B}[1, 10) \right\}$$

con $\sigma(D_1, D_2, \ldots)$ el sigma álgebra generado por las funciones $D_1, D_2, \ldots$

Demostración. Por definición,

$$\mathcal{M} = \mathbb{R}^+ \cap \sigma(S) = \mathbb{R}^+ \cap \{S^{-1}(B) : B \in \mathcal{B}\} = \mathbb{R}^+ \cap \{S^{-1}(B) : B \in \mathcal{B}[1, 10)\}.$$

Entonces, dado $A \in \mathcal{M}$, existe un conjunto $B \in \mathcal{B}[1, 10)$ con

$$A = \mathbb{R}^+ \cap S^{-1}(B) = \bigcup_{k\in\mathbb{Z}} 10^k B.$$

Así, utilizando que $S(A) = B$, llegamos al primer resultado.

Para probar la segunda igualdad, primero observamos que como podemos expresar $S(x) = \sum_{m\in\mathbb{N}} 10^{1-m} D_m(x)$, la función mantisa S queda completamente determinada por sus cifras significativas $\{D_1, D_2, \ldots\}$ y por lo tanto $\sigma(S) \subset \sigma(D_1, D_2, \ldots)$. Esto implica entonces que $\mathcal{M} \subset \mathbb{R}^+ \cap \sigma(D_1, D_2, \ldots)$.

Veamos ahora el recíproco: En primer lugar tenemos la siguiente expresión que nos define unívocamente cada $D_m(x)$:

$$D_m(x) = \lfloor 10^{m-1} S(x) \rfloor - 10 \lfloor 10^{m-2} S(x) \rfloor$$

para cada $m \in \mathbb{N}$. Utilizando esto tenemos que $\sigma(D_m) \subset \sigma(S)$ para todo $m \in \mathbb{N}$ y por lo tanto que $\sigma(D_1, D_2, \ldots) \subset \sigma(S)$. Para finalizar basta notar que para cada $A \in \mathcal{M}$, $S(A) \in \mathcal{B}[1, 10)$ y por lo tanto $A = \bigcup_{k\in\mathbb{Z}} 10^k B$ para $B = S(A)$ (utilizando el primer resultado de esta demostración). Por lo tanto, todo conjunto de la forma $\bigcup_{k\in\mathbb{Z}} 10^k B = \mathbb{R}^+ \cap S^{-1}(B)$ con $B \in \mathcal{B}[1, 10)$ claramente pertenece a $\mathcal{M}$.

El siguiente lema establece algunas propiedades básicas de cierre del álgebra mantisa $\mathcal{M}$ que serán esenciales para el posterior estudio de características de la Ley de Benford como la invarianza de escala o de base. Para formular estas propiedades definimos: $\forall C \subset \mathbb{R}^+$ y $n \in \mathbb{N}$, sea $C^{1/n} = \{x > 0 \,|\, x^n \in C\}$.

Lema 2.2. *(i) $\mathcal{M}$ es invariante con respecto a la multiplicación por potencias enteras de 10, es decir,*

$$\forall A \in \mathcal{M},\ k \in \mathbb{Z}: \quad 10^k A = A.$$

(ii) $\mathcal{M}$ es cerrado bajo multiplicación por escalares, es decir,

$$\forall A \in \mathcal{M},\ a > 0: \quad aA \in \mathcal{M}.$$

(iii) $\mathcal{M}$ es cerrado bajo raíces enteras, es decir,

$$\forall A \in \mathcal{M},\ n \in \mathbb{N}: \quad A^{1/n} \in \mathcal{M}.$$

Demostración. (i) Esto es claro debido al primer resultado del teorema anterior teniendo en cuenta que $S(10^k A) = S(A)$.
(ii) Dado $A \in \mathcal{M}$, utilizando el segundo resultado del teorema anterior, existe $B \in \mathcal{B}[1, 10)$ tal que $A = \bigcup_{k\in\mathbb{Z}} 10^k B$. En vista de (i) asumamos sin pérdida de generalidad que $1 < a < 10$. Entonces:

$$aA = \bigcup_{k\in\mathbb{Z}} 10^k aB = \bigcup_{k\in\mathbb{Z}} 10^k \left((aB \cap [a, 10)) \cup \left(\frac{1}{10} aB \cap [1, a)\right)\right) = \bigcup_{k\in\mathbb{Z}} 10^k C$$

con $C = \left((aB \cap [a, 10)) \cup \left(\frac{1}{10} aB \cap [1, a)\right)\right) \in \mathcal{B}[1, 10)$.
Lo que demuestra que $aA \in \mathcal{M}$.
(iii) Como los intervalos de la forma $[1, 10^s]$ con $0 < s < 1$ generan $\mathcal{B}[1, 10)$, es suficiente verificar el resultado para el caso $A = \bigcup_{k\in\mathbb{Z}} 10^k [1, 10^s]$ para todo $0 < s < 1$. En este caso:

$$A^{1/n} = \bigcup_{k\in\mathbb{Z}} 10^{k/n} [1, 10^{s/n}] = \bigcup_{k\in\mathbb{Z}} 10^k \bigcup_{j=0}^{n-1} [10^{j/n}, 10^{(j+s)/n}] = \bigcup_{k\in\mathbb{Z}} 10^k C$$

con $C = \bigcup_{j=0}^{n-1} [10^{j/n}, 10^{(j+s)/n}] \in \mathcal{B}[1, 10)$. Lo que implica que $A^{1/n} \in \mathcal{M}$.

2.3.2 Invarianza de escala

La invarianza de escala es una de las más simples hipótesis que surgen al pensar en la "universalidad" de la ley de Benford. Supongamos una ley que aparece de alguna manera en conjuntos de números "naturales" (en el sentido que le dio Newcomb a esta palabra). Lo primero que se nos viene a la cabeza es pensar que la ley se cumple independientemente de las unidades en las que trabajemos, es decir, que la multiplicación por una constante no afectará a la medida de probabilidad. Esta propiedad es el objeto de estudio de esta sección.

Definición 2.8. *Una medida de probabilidad P sobre el álgebra mantisa $\mathcal{M}$ es invariante en escala si cumple:*

$$\forall s \in \mathbb{R}^+, \forall S \in \mathcal{M}, P(S) = P(sS).$$

El siguiente teorema nos muestra que que la invarianza de escala es una caracterización de la ley de Benford, lo cual lo hace una propiedad muy peculiar.

Teorema 2.2. *Una medida de probabilidad P en $(\mathbb{R}^+, \mathcal{M})$ es invariante de escala si y solo si P sigue la ley de Benford.*

Demostración. Sea P una medida de probabilidad en $(\mathbb{R}^+, \mathcal{M})$, y sea $S_a = \bigcup_{n=-\infty}^{\infty} [1, 10^a) \cdot 10^n$ para un a arbitrario en $[0, 1)$ (una medida de probabilidad en $\mathcal{M}$ está definida completamente por sus valores en dichos conjuntos).

Sean $\bar{\mathrm{P}}$ y $\hat{\mathrm{P}}$ dos medidas de probabilidad definidas respectivamente en los espacios medibles $([0, 1), \mathcal{B}[0, 1])$ y $([1, 10), \mathcal{B}[1, 10])$ por:

$$\forall a \in [0, 1),\ \bar{\mathrm{P}}[0, a) = \hat{\mathrm{P}}[1, 10^a) = P(S_a) \tag{2.6}$$

Esta definición nos da una una correspondencia muy útil entre los espacios medibles de P, $\bar{\mathrm{P}}$ y $\hat{\mathrm{P}}$ respectivamente.

Ahora bien, en el espacio medible $(\mathbb{R}^+, \mathcal{M})$, P satisface la ley de Benford si y solo si $\forall a \in [0, 1)\ \mathrm{P}(S_a) = a$ (de acuerdo con la relación anterior, si y solo si $\bar{\mathrm{P}}$ es uniforme en $[0, 1]$ y si y solo si $\forall a \in [0, 1)\ \hat{\mathrm{P}}[1, 10^a) = a$).

Ahora supongamos que P satisface la ley de Benford y probemos que $\forall s \in \mathbb{R}$:

$$P(sS_a) = P\left(\bigcup_{n=-\infty}^{\infty} [s, s \cdot 10^a) \cdot 10^n\right) = P(S_a).$$

Sin perdida de generalidad, supongamos $s \in [1, 10)$ (sino tomaremos s mod 10)[2]. Distingamos dos casos:

- Si $s \cdot 10^a \leq 10$:

$$\begin{aligned} \mathrm{P}(sS_a) &= \hat{\mathrm{P}}[s, s \cdot 10^a) \\ &= \hat{\mathrm{P}}[1, s \cdot 10^a) - \hat{\mathrm{P}}[1, s) \\ &= \log(s \cdot 10^a) - \log(s) \\ &= a \\ &= \mathrm{P}(S_a). \end{aligned}$$

- Si $s \cdot 10^a > 10$ (como $s \leq 10$, $s \cdot 10^a \in [10, 100)$):

$$\begin{aligned} \mathrm{P}(sS_a) &= \hat{\mathrm{P}}[s, 10) + \hat{\mathrm{P}}([10, s \cdot 10^a) \bmod 10) \\ &= (1 - \log(s)) + \hat{\mathrm{P}}\left[1, \frac{s \cdot 10^a}{10}\right) \\ &= 1 - \log(s) + \log\left(\frac{s \cdot 10^a}{10}\right) \\ &= 1 - \log(s) + \log(s) + a - 1 \\ &= a \\ &= \mathrm{P}(S_a). \end{aligned}$$

De manera recíproca, supongamos ahora que P tiene invarianza de escala y demostremos que P satisface la ley de Benford.

Sea α un número arbitrario irracional. Tenemos entonces que se cumple que $\forall a \in [0, 1)\ \mathrm{P}(S_a) = \mathrm{P}(10^\alpha S_a)$ por hipótesis Sin perdida de generalidad, podemos suponer que $10^\alpha \in [1, 10)$ (de otra manera tomaremos $10^\alpha \mod 10$).

El isomorfismo definido en (2.6) implica entonces:

$$\hat{\mathrm{P}}[1, 10^a) = \hat{\mathrm{P}}([10^\alpha, 10^{a+\alpha}) \bmod 10) \quad \forall a \in [0, 1)$$

y como consecuencia:

$$\bar{\mathrm{P}}[0, a) = \bar{\mathrm{P}}([\alpha, a + \alpha) \bmod 1) \quad \forall a \in [0, 1).$$

[2] A partir de aquí la operación mod 10 notará el cociente de dividir entre 10

Esta igualdad implica que $\bar{P}$ es invariante bajo traslaciones irracionales en el intervalo unidad. Es conocido desde hace años que la única distribución en $[0, 1)$ que cumple esta propiedad es la uniforme (ver [16]). Por lo tanto P satisface la ley de Benford.

Observación 2.5. En esta prueba vemos como la definición de invarianza de escala es muy fuerte. De hecho, las hipótesis de la demostración de la invarianza de escala pueden ser reducidas a las hipótesis de invarianza de escala por un número arbitrario que no sea una potencia racional de la base (aquí α es irracional y la prueba solo usa la invarianza de escala por 10^{α}).

2.3.3 Invarianza de base

La invarianza de base es una hipótesis un tanto más fina que nos lleva a la ley de Benford. El concepto proviene de que si suponemos que en un conjunto de datos "naturales" en base 10 (por ejemplo) cumplen la citada ley, entonces este mismo conjunto de datos con distintas bases deberían también tener que cumplir la ley de Benford.

En este caso, el álgebra mantisa $\mathcal{M}$ con la que hemos tratado es un caso específico de $\mathcal{M}_b$ donde esto último denota el álgebra mantisa en base b ($\mathcal{M}_{10} = \mathcal{M}$). Todas las definiciones, propiedades y teoremas son esencialmente los mismos, exceptuando que b reemplaza 10 (por ejemplo $\log_b$ sustituye en este caso a log en las funciones de distribución y probabilidad).

Definición 2.9. *Una medida de probabilidad P en* $(\mathbb{R}^+, \mathcal{M}_b)$ *es invariante de base si*

$$\forall m \in \mathbb{N},\ \forall S \in \mathcal{M}_b,\ P(S) = P(S^{1/m}).$$

Entender porque esta definición es importante para la invarianza de base es un poco difícil a primera vista. Para motivar este entendimiento, consideramos:

$$S = \bigcup_{q=-\infty}^{\infty} [b^x, b^y) \cdot b^q$$

con $x, y \in [0, 1)$ (cualquier intervalo de la forma $[1, b)$ se puede expresar de la forma $[b^x, b^y)$).

Sea ahora:

$$B_m = \bigcup_{k=0}^{m-1} \left[b^{x/m}, b^{y/m}\right) \cdot b^{k/m}.$$

Tenemos entonces:

$$\begin{aligned} S^{1/m} &= \bigcup_{q=-\infty}^{\infty} B_m \cdot b^q \\ &= \bigcup_{q=-\infty}^{\infty} \left(\bigcup_{k=0}^{m-1} \left[b^{x/m}, b^{y/m}\right) \cdot b^{k/m} \right) b^q. \end{aligned} \tag{2.7}$$

Reemplazando b por b^m, nos queda:

$$\begin{aligned} S^{1/m} &= \bigcup_{q=-\infty}^{\infty} \left(\bigcup_{k=0}^{m-1} [b^x, b^y) \cdot b^k \right) b^{mq} \\ &= \bigcup_{q=-\infty}^{\infty} \left(\bigcup_{k=0}^{m-1} [b^x, b^y) \cdot b^{k+mq} \right) \\ &= \bigcup_{n=-\infty}^{\infty} [b^x, b^y) \cdot b^n \\ &= S. \end{aligned}$$

Luego, si una medida de probabilidad tiene invarianza de base debería ser invariante al menos para las potencias de la base inicial y por lo tanto la probabilidad de S debería coincidir con la de $S^{1/m}$. Aunque esta definición es un tanto débil ya que solo contaría con las bases que son potencias de la base inicial, un teorema que demostraremos a final de esta sección nos mostrará que es suficiente para cualquier base.

Sea ahora P_b la medida de probabilidad logarítmica en $(\mathbb{R}^+, \mathcal{M}_b)$:

$$\forall t \in [1, b) \quad P_b\left(\bigcup_{n=-\infty}^{\infty} [1, t] b^n \right) = \log_b(t).$$

Definición 2.10. *Una medida μ en $(\Omega, \mathcal{F})$ es invariante bajo la función medible $T: \Omega \to \Omega$ si:*

$$\mu(E) = \mu(T^{-1}(E)) \quad \forall E \in \mathcal{F}.$$

Ahora sea n un entero positivo arbitrario, consideremos el espacio medible Borel en $[0,1)$ y la función T_n definida en $[0,1)$ como $T_n(x) = nx \mod 1$.

Lema 2.3. *Una medida de probabilidad $\bar{P}$ en $([0,1), \mathcal{B}[0,1))$ es invariante bajo T_n si y solo si*

$$\bar{P}[0,a) = \sum_{k=0}^{n-1} \bar{P}\left[\frac{k}{n}, \frac{k+a}{n}\right) \qquad , \forall n \in \mathbb{N}.$$

Demostración. Basta probar que:

$$T_n\left(\bigcup_{k=0}^{n-1}\left[\frac{k}{n}, \frac{k+a}{n}\right)\right) = [0,a).$$

- En primer lugar, sea $x \in \bigcup_{k=0}^{n-1}\left[\frac{k}{n}, \frac{k+a}{n}\right)$
  $\Rightarrow \exists k' \in \{0,\dots,n-1\}$ tal que $k' \leq nx < k'+a$
  $\Rightarrow nx \mod 1 \in [0,a)$.
- Recíprocamente, escojamos ahora $y \in [0,a)$
  $\Rightarrow \forall k \in \mathbb{N}, \exists x' \in [k, k+a)$ tal que $x' = k + y$
  $\Rightarrow \forall k \in \mathbb{N}, \exists x \in \left[\frac{k}{n}, \frac{k+a}{n}\right)$ tal que $nx = k + y$ (o bien $y = nx \mod 1$)
  Ahora, para que x pertenezca a $[0,1)$, k deberá pertenecer a $\{0,\dots,n-1\}$
  $\Rightarrow \forall k \in \{0,\dots,n-1\}, \exists x \in \left[\frac{k}{n}, \frac{k+a}{n}\right)$ tal que $y = nx \mod 1$.

Ahora estamos en disposición de demostrar el siguiente lema. Denotamos λ a la medida de Lebesgue en $[0,1)$ y δ_0 la medida de probabilidad de Dirac en 0.

Lema 2.4. *Una medida de probabilidad $\bar{P}$ definida en $\mathcal{B}([0,1))$ es invariante bajo T_n si y solo si*

$$\bar{P} = q\delta_0 + (1-q)\lambda$$

para algún $q \in [0,1]$.

Demostración. En primer lugar, supongamos $\bar{\mathrm{P}} = q\delta_0 + (1-q)\lambda$ para algún $q \in [0,1]$. Entonces $\forall a \in [0,1)$:

$$\begin{aligned}
\sum_{k=0}^{n-1} \bar{\mathrm{P}}\left[\frac{k}{n}, \frac{k+a}{n}\right) &= q\sum_{k=0}^{n-1} \delta_0\left[\frac{k}{n}, \frac{k+a}{n}\right) + (1-q)\sum_{k=0}^{n-1} \lambda\left[\frac{k}{n}, \frac{k+a}{n}\right) \\
&= q + (1-q)\sum_{k=0}^{n-1} \lambda\left[\frac{k}{n}, \frac{k+a}{n}\right) \\
&= q + (1-q)\sum_{k=0}^{n-1} \frac{a}{n} \\
&= q + (1-q)a \\
&= q\delta_0[0,a) + (1-q)\lambda[0,a) \\
&= \bar{\mathrm{P}}[0,a).
\end{aligned}$$

Luego, haciendo uso del lema anterior, la medida de probabilidad $\bar{\mathrm{P}}$ en $\mathcal{B}([0,1))$ es invariante bajo T_n.

Recíprocamente supongamos una medida de probabilidad $\bar{\mathrm{P}}$ arbitraria en $\mathcal{B}([0,1))$ que sea invariante bajo T_n.
Recordamos que una medida de probabilidad $\bar{\mathrm{P}}$ en $[0,1)$ está unívocamente determinada por sus coeficientes de Fourier:

$$\phi_n = \int_0^1 e^{2i\pi nx} d\bar{\mathrm{P}}(x) \quad , n \in \mathbb{N}.$$

La invarianza de $\bar{\mathrm{P}}$ bajo T_n implica que sus coeficientes de fourier ϕ_n se mantienen constantes para todo $n \in \mathbb{N}$. Para verlo, utilicemos el cambio de variable $x' = T_n(x) = nx \bmod 1$ y teniendo en cuenta que $e^{2i\pi x'} = e^{2i\pi nx}$:

$$\begin{aligned}\phi_n &= \int_0^1 e^{2i\pi nx} d\bar{\mathrm{P}}(x) \\ &= \int_0^1 e^{2i\pi x'} d\bar{\mathrm{P}}(x') \quad \text{(debido a la hipótesis de invarianza de } \bar{\mathrm{P}} \text{ bajo } T_n) \\ &= \phi_1 \quad , \forall n \in \mathbb{N}.\end{aligned}$$

Ahora sea $\phi_n = q$ con $q \in \mathbb{C}$. Veamos que realmente $q \in [0,1) \subset \mathbb{R}$. Consideramos:

$$\bar{\mathrm{P}}(\{0\}) = \int_0^1 \lim_{x\to\infty} \left\{ \frac{1}{N} \sum_{n=1}^{N} e^{2i\pi nx} \right\} d\bar{\mathrm{P}}(x).$$

Esa igualdad viene del hecho de que dado $a \in \mathbb{C}$ con $|a| \leq 1$ (en este caso $a = e^{2i\pi x}$), $\frac{1}{N}\sum_{n=1}^{N} a^n \to 0$ así como $N \to \infty$, exceptuando el caso $a = 1$ en el que tendremos la igualdad a 1 para todo N.

Ahora, utilizando el teorema de convergencia dominada de Lebesgue (ya que se cumple $\left|\frac{1}{N}\sum_{n=1}^{N} e^{2i\pi nx}\right| \leq 1$):

$$\begin{aligned}\bar{\mathrm{P}}(\{0\}) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \phi_n \\ &= q \quad \text{(ya que } \phi_n \text{ es constante e igual a } q).\end{aligned}$$

Por lo tanto q es una probabilidad y $q \in [0,1) \subset \mathbb{R}$ como queríamos demostrar.

Ahora, derivando los coeficientes de Fourier ϕ'_n para $q\delta_0 + (1-q)\lambda$:

$$\begin{aligned}\phi'_n &= q \cdot e^{2i\pi n0} + (1-q) \int_0^1 e^{2i\pi nx} dx \\ &= q \\ &= \phi_n \text{ para todo } n.\end{aligned}$$

Como los coeficientes de fourier de una medida la determinan completamente:

$$\bar{\mathrm{P}} = q\delta_0 + (1-q)\lambda.$$

Sea ahora Δ_1 la medida delta de Dirac del conjunto $S_1 = \bigcup_{n=-\infty}^{\infty} \{1\} \cdot b^n$. Más precisamente, Δ_1 está definida para $S \in \mathcal{M}_b$ como $\Delta_1(S) = 1$ si $S_1 \subseteq S$ y 0 en otro caso (esta definición es valida ya que S_1 no contiene ningún subconjunto $\mathcal{M}_b$-medible no vacío).

Ahora sí, estamos en disposición de probar el resultado central de esta sección: El teorema que une la invarianza de base con estas medidas de probabilidad, y por lo tanto con la ley de Benford.

Teorema 2.3. *Una medida de probabilidad P en $(\mathbb{R}^+, \mathcal{M}_b)$ tiene invarianza de base si existe un $q \in [0,1]$ tal que:*

$$P = (1-q)P_b + q\Delta_1.$$

Demostración. En primer lugar veamos que P_b y Δ_1 tienen invarianza de base.

En el caso de Δ_1 es evidente ya que el dígito 1 sigue inalterado como 1 aunque apliquemos cualquier cambio de base. Para la demostración de P_b, sean x,y números arbitrarios en $[0,1)$ y sea $S = \bigcup_{q=-\infty}^{\infty} [b^x, b^y) \cdot b^q$:

$$\begin{aligned}
P_b(S^{1/m}) &\underset{(2.7)}{=} P_b\left(\bigcup_{q=-\infty}^{\infty}\left(\bigcup_{k=0}^{m-1}\left[b^{x/m}, b^{y/m}\right)\cdot b^{k/m}\right) b^q\right) \\
&= \sum_{k=0}^{m-1} P_b\left(\bigcup_{q=-\infty}^{\infty}\left[b^{\frac{x+k}{m}}, b^{\frac{y+k}{m}}\right) b^q\right) \\
&= \sum_{k=0}^{m-1} \log_b\left(\frac{b^{\frac{y+k}{m}}}{b^{\frac{x+k}{m}}}\right) \\
&= \sum_{k=0}^{m-1} \log_b\left(b^{\frac{y-x}{m}}\right) \\
&= y - x \\
&= P_b(S).
\end{aligned}$$

Por lo tanto P_b y Δ_1 tienen invarianza de base, y por lo tanto la combinación lineal $q\Delta_1 + (1-q)P_b$ también.

Ahora, supongamos que P tiene invarianza de base en $(\mathbb{R}, \mathcal{M})$. Sea $S = \bigcup_{q=-\infty}^{\infty} [1, b^a) b^q$ para algún $a \in [0,1)$, y utilicemos la expresión de $S^{1/m}$ para obtener:

$$P(S^{1/m}) = P\left(\bigcup_{q=-\infty}^{\infty}\bigcup_{k=0}^{m-1}\left[b^{\frac{k}{m}}, b^{\frac{k+a}{m}}\right) b^q\right).$$

De nuevo, haciendo uso del isomorfismo (2.6), la igualdad $P(S) = P(S^{1/m})$ (hipótesis de invarianza de base de P) implica

$$\hat{\mathrm{P}}[1, b^a) = \hat{\mathrm{P}}\left(\bigcup_{k=0}^{m-1}\left[b^{\frac{k}{m}}, b^{\frac{k+a}{m}}\right)\right)$$

y consecuentemente

$$\bar{\mathrm{P}}[0, a) = \sum_{k=0}^{n-1} \bar{\mathrm{P}}\left[\frac{k}{n}, \frac{k+a}{n}\right).$$

Utilizando los Lemas 2.3 y 2.4, esta ecuación nos asegura que $\exists q \in [0, 1)$ tal que $\bar{\mathrm{P}} = q\delta_0 + (1-q)\lambda$. Es fácilmente visto que $\Delta_1 = \delta_0$ y $P_b = \lambda$.

Por lo tanto, $\exists q \in [0, 1)$ tal que $P = q\Delta_1 + (1-q)P_b$.

Observación 2.6. La prueba de este Teorema explica que la hipótesis "$P(S) = P(S^{1/m})$" es suficiente para la invarianza en cualquier base.

Observación 2.7. Una consecuencia inmediata de este Teorema y del Teorema 2.2. es que la invarianza de escala implica la invarianza de base (aunque el recíproco, en general, no es cierto).

2.3.4 Invarianza de suma

En su tesis doctoral (ver [13]), Nigrini observó que las tablas de datos no manipulados se ajustaban bastante bien a las probabilidades de la ley de Benford y además notó que "La suma de todas las entradas[3] con primer dígito $d_1 = d$ es constante para varios d". Con esto, el matemático, se refería a que en una lista de datos, si sumamos todos los datos que empiecen por 1, todos los que empiecen por dos, y así recursivamente hasta el nueve, tendremos que las nueve sumas nos darán una constante.

Este hecho tiene sentido si lo pensamos un poco ya que: por ejemplo en un conjunto de enteros que sigan la ley logarítmica del primer dígito habrá una mayoría de números que empiecen por 1, un poco de menos números que empiecen por 2 y así sucesivamente hasta llegar a los menos probables, los que empiezan por 9 (ver

[3] hablando de las entradas al aplicarles la función decimal mantisa S definida en la sección 2.3.1.

Figura 1.1). Por lo tanto, debido a estas probabilidades, es de esperar que las sumas aproximadamente sean las mismas.

Se observó además que este hecho se generalizaba no solo cuando sumábamos las entradas en las que fijábamos una cifra, sino que si sumábamos las entradas de forma $d_1, d_2 \dots d_k$ tendremos una suma aproximada a la suma de las entradas que tenían como primeros k dígitos $d_1', d_2' \dots d_k'$. Por ejemplo, en un conjunto de datos que sigan la ley de Benford, la suma de todos los números que empiecen con $1,234$ será aproximada a la suma de los números que empiecen por $6,893$.

Para demostrar que la invarianza de suma es una caracterización de las distribuciones Benford, hagamos un par de notaciones:

Dado $k \in \mathbb{N}$, sea $d_1 \in \{1, \dots, 9\}$ y $d_2, \dots, d_k \in \{0, 1, \dots, 9\}$:

- Sea $\mathrm{A}(d_1, \dots, d_k)$ el conjunto de números reales cuyos primeros dígitos son $d_1, \dots, d_k$.
- Sea $\bar{\mathrm{A}}(d_1, \dots, d_k)$ el conjunto de números reales en $[1, 10)$ cuyos primeros dígitos son $d_1, \dots, d_k$.

Definición 2.11. *Dada una medida de probabilidad P en $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$, definimos P_S en $([1, 10), \mathcal{B}[1, 10))$ como la medida de probabilidad de su mantisa, es decir, al aplicarle la función decimal mantisa S:*

$$P_S[1, t) = P\left(\bigcup_{n=-\infty}^{\infty} [1, t) \cdot 10^n\right), \quad \forall t \in [1, 10).$$

Ahora es momento de anotar que para ver la invarianza de suma no tenemos que hacer la suma en sí, lo que nos interesará será la media. Así como hizo Nigrini en su tesis, la suma es solo una manera de simular la media.

Sea X una variable aleatoria arbitraria que venga de una distribución con invarianza de suma y sea $k \in \mathbb{N}$. Lo que realmente significa la invarianza de suma es que la esperanza de $S(X)$ condicionada a que $S(X)$ empiece por $d_1, d_2 \dots d_k$ es la misma independientemente de la k-tupla $(d_1, \dots, d_k)$ escogida.

Veamos una definición formal:

Definición 2.12. *Una medida de probabilidad P en $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ es suma-invariante si, para cualquier variable aleatoria X con distribución P, y para todo $k \in \mathbb{N}$ fijo, la esperanza $E[S(X)I_{A(d_1,\dots,d_k)}(X)]$ es constante independientemente de la k-tupla $(d_1, \dots, d_k)$ escogida (con $d_1 \in \{1, \dots, 9\}$ y $d_2, \dots, d_k \in \{0, 1, \dots, 9\}$).*

Teorema 2.4. *Una medida de probabilidad P en* $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ *es suma-invariante si y solo si* $P_S[1,t) = \log(t)$.

Demostración. En primer lugar observamos que para todos los dígitos $d_1, \dots, d_k$:

$$\begin{aligned} E[S(X)I_{A(d_1,\dots,d_k)}(X)] &= E[S(X)I_{\bar{A}(d_1,\dots,d_k)}(X)] \\ &= \int_{\bar{A}(d_1,\dots,d_k)} x d P_S(x). \end{aligned} \tag{2.8}$$

Observamos que todos los conjuntos de $\bar{A}(d_1, \dots, d_k)$ son de hecho intervalos que tienen la misma longitud, denotémosla como $\lambda(A)$ (donde λ denota la medida de Lebesgue en $[1, 10)$, y forman una partición de $[1, 10)$.

Supongamos en primer lugar que la medida de probabilidad P es suma-invariante, por lo tanto las esperanzas de (2.8) son constantes (sea su valor $\int_A x d P_S(x)$). Sumemos sobre todos los conjuntos de $\bar{A}(d_1, \dots, d_k)$:

$$\begin{aligned} \sum_{\substack{1\le d_1\le 9\\ 0\le d_2\le 9\\ \cdots\\ 0\le d_k\le 9}} \int_{\bar{A}(d_1,\dots,d_k)} x d P_S(x) &= \int_1^{10} x d P_S(x) \\ \underset{(*)}{\Rightarrow} \frac{9}{\lambda(A)} \int_A x d P_S(x) &= \int_1^{10} x d P_S(x) \\ \Rightarrow \int_A x d P_S(x) &= \frac{\lambda(A)}{9} \int_1^{10} x d P_S(x). \end{aligned} \tag{2.9}$$

(*)Ya que las esperanzas son constantes y hay $\dfrac{9}{\lambda(A)}$ particiones de longitud $\lambda(A)$ en el intervalo $[1, 10)$.

Evidentemente, la última igualdad de (2.9) es una condición necesaria y suficiente para la suma-invarianza. Supongamos ahora que P_S sigue la ley logarítmica ($d P_S(x) = \frac{1}{x \ln 10} d x$) y sustituyamos en ambos miembros de esta igualdad:

$$\int_A x d P_S(x) = \frac{1}{\ln 10} \lambda(A)$$

$$\frac{\lambda(A)}{9} \int_1^{10} x d P_S(x) = \frac{\lambda(A)}{9} \cdot \frac{9}{\ln 10}$$

Por lo tanto, vemos que se verifica la igualdad.

Por otra parte, supongamos que la condición de (2.9) se tiene para todos los A de la forma $\bar{A}(d_1, \dots, d_k)$. Cada intervalo de $[1, 10)$ se puede representar como una unión numerable de dichos intervalos, luego, sumando las integrales, (2.9) se tiene para todo intervalo.

Ambos miembros de (2.9) definen una medida de probabilidad del conjunto A, y dichas medidas coinciden en los conjuntos de intervalos de $[1, 10)$. Por el teorema de extensión de Carathéodory (ver [11] pág. 30-31), esto es suficiente para asegurar la igualdad en los borelianos de $[1, 10)$.

Por lo tanto, $x dP_S(x)$ es proporcional a dx, y como consecuencia $dP_S(x)$ es proporcional a $\frac{1}{x}dx$. Normalizando en $[1, 10)$ llegamos a que $dP_S(x) = \frac{1}{x \ln 10}dx$, es decir, P_S sigue la ley logarítmica.

2.4 Otras propiedades de las distribuciones Benford

2.4.1 Invarianza de la inversa

Proposición 2.1. Si X es una variable aleatoria la cual $S(X)$ cumpla la ley logarítmica, entonces $S(X^{-1})$ también sigue la ley logarítmica.

Demostración. Para todo $t \in [1, 10)$:

$$\begin{aligned}
\Pr[S(X^{-1}) \in [1, t)] &= \Pr\left[X^{-1} \in \bigcup_{n=-\infty}^{\infty} [1, t)10^n\right] \\
&= \Pr\left[X \in \bigcup_{n=-\infty}^{\infty} [\frac{1}{t}, 1)10^n\right] \\
&= \Pr\left[S(X) \in \left[\frac{10}{t}, 10\right)\right] \\
&= 1 - \log\left(\frac{10}{t}\right) \\
&= \log(t).
\end{aligned}$$

2.4.2 Multiplicación y división

Proposición 2.2. Sean X e Y dos variables aleatorias, y sean f, g y h las respectivas funciones de densidad de $S_b(X)$, $S_b(Y)$ y $S_b(XY)$. Entonces, para todo $z \in [1, b)$:

$$h(z) = \int_1^z \frac{1}{x} g\left(\frac{z}{x}\right) f(x)dx + \int_z^b \frac{b}{x} g\left(\frac{bz}{x}\right) f(x)dx. \tag{2.10}$$

Demostración. Sean F, G y H las funciones de distribución respectivas de $S_b(X)$, $S_b(Y)$ y $S_b(XY)$.

Ahora, sean $x, y \in \mathbb{R}$. Sin pérdida de generalidad, supongamos $x, y \in [1, b)$. La mantisa de xy está dada por:

$$S_b(xy) = \begin{cases} xy & \text{si} \quad 1 \leq xy < b \\ \frac{xy}{b} & \text{si} \quad b \leq xy < b^2 \end{cases}$$

Entonces, para todo $z \in [1, b)$:

$$\begin{aligned} S_b(xy) \leq z &\iff \begin{cases} xy \leq z \\ o \\ \dfrac{xy}{b} \leq z \quad \text{y} \quad b \leq xy \end{cases} \\ &\iff \begin{cases} y \leq \dfrac{z}{x} \\ o \\ \dfrac{b}{x} \leq y \leq \dfrac{zb}{x} \end{cases} \end{aligned}$$

Dividiendo la zona sombreada de la Figura 2.3 en tres e integrando:

$$\begin{aligned} H(z) &= \int_1^z \int_1^{\frac{z}{x}} f(x)g(y)dydx + \int_1^z \int_{\frac{b}{x}}^b f(x)g(y)dydx + \int_z^b \int_{\frac{b}{x}}^{\frac{zb}{x}} f(x)g(y)dydx \\ &= \int_1^z \left[G\left(\frac{z}{x}\right) - G(1) + G(b) - G\left(\frac{b}{x}\right)\right] f(x)dx \\ &\quad + \int_z^b \left[G\left(\frac{bz}{x}\right) - G\left(\frac{b}{x}\right)\right] f(x)dx. \end{aligned}$$

Por último, derivando con respecto a z:

$$
\begin{aligned}
h(z) &= f(z)\left[G(1) - G(1) + G(b) - G\left(\frac{b}{z}\right) - G(b) + G\left(\frac{b}{z}\right)\right] \\
&\quad + \int_1^z \frac{1}{x} g\left(\frac{z}{x}\right) f(x)dx + \int_z^b \frac{b}{x} g\left(\frac{bz}{x}\right) f(x)dx \\
&= \int_1^z \frac{1}{x} g\left(\frac{z}{x}\right) f(x)dx + \int_z^b \frac{b}{x} g\left(\frac{bz}{x}\right) f(x)dx.
\end{aligned}
$$

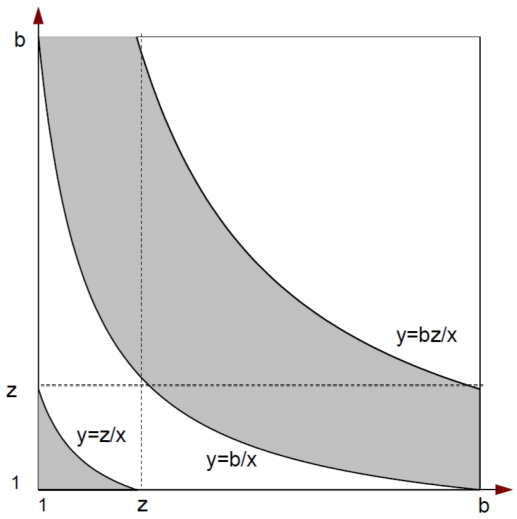


Figura 2.3: Función de distribución $S_b(xy) \leq z$

La invarianza de la multiplicación es una propiedad sorprendente que se cumple para cualquier variable aleatoria continua:

Proposición 2.3. Sea Y una variable aleatoria continua que satisfaga la ley de Benford. Sea X una variable aleatoria arbitraria con función de densidad continua. Entonces XY satisface también la ley de Benford.

Demostración. Utilizando la notación de la proposición anterior y teniendo en cuenta que Y es Benford y por lo tanto $g(y) = \dfrac{1}{y \ln b}$, para $y \in [1, b)$:

$$
\begin{aligned}
h(z) &= \int_1^z \frac{1}{x} \frac{x}{z \ln b} f(x)dx + \int_z^b \frac{b}{x} \frac{x}{bz \ln b} f(x)dx \\
&= \frac{1}{z \ln b} \int_1^b f(x)dx \\
&= \frac{1}{z \ln b} \quad \text{con } z \in [1, b).
\end{aligned}
$$

Por lo tanto XY satisface la ley de Benford.

Corolario 2.1. Sea Y una variable aleatoria no nula que satisfaga la ley de Benford. Sea X una variable aleatoria no nula con función de densidad continua. Entonces $\frac{X}{Y}$ y $\frac{Y}{X}$ satisfacen la ley de Benford.

Demostración. Consecuencia directa de las proposiciones 2.1. y 2.3.

2.4.3 Convergencia a distribuciones Benford tras multiplicaciones

Posteriormente a los resultados presentados en la subsección anterior, Hamming encontró un resultado aún mas impresionante resultado de la multiplicación (ver [4]) que explicaría porqué en grandes computaciones la ley de Benford podría aparecer "de la nada". El hecho remarcable es que el producto de variables aleatorias satisface mejor la ley de Benford que las variables antes de la multiplicación.

Definición 2.13. *Definimos la distancia relativa de la función de densidad f a la ley de Benford como:*

$$D(f) = \max_{1 \leq z < b} \left| \frac{f(z) - r(z)}{r(z)} \right|$$

donde $r(z) = \frac{1}{z \ln b}$.

Proposición 2.4. Sean X e Y dos variables aleatorias, y sean f, g y h las respectivas funciones de densidad de $S_b(X)$, $S_b(Y)$ y $S_b(XY)$.

$$D(h) \leq \min(D(f), D(g)).$$

Demostración. Utilizando la ecuación (2.10) y teniendo en cuenta que si r es Benford su multiplicación por f también lo será, para cualquier función de densidad f continua:

$$r(z) = \int_1^z \frac{1}{x} r\left(\frac{z}{x}\right) f(x)dx + \int_z^b \frac{b}{x} r\left(\frac{bz}{x}\right) f(x)dx$$

restándole esto a la expresión de (2.10) y dividiendo entre $r(z)$:

$$\begin{aligned}\frac{h(z)-r(z)}{r(z)} &= \int_1^z \frac{f(x)}{x}\frac{g\left(\frac{z}{x}\right)-r\left(\frac{z}{x}\right)}{r(z)}dx \\ &\quad + \int_z^b \frac{bf(x)}{x}\frac{g\left(\frac{bz}{x}\right)-r\left(\frac{bz}{x}\right)}{r(z)}dx \\ &\underset{(*)}{=} \int_1^z f(x)\frac{g\left(\frac{z}{x}\right)-r\left(\frac{z}{x}\right)}{r(\frac{z}{x})}dx \\ &\quad + \int_z^b f(x)\frac{g\left(\frac{bz}{x}\right)-r\left(\frac{bz}{x}\right)}{r(\frac{bz}{x})}dx.\end{aligned}$$

(*) Teniendo en cuenta que $xr(z) = r\left(\frac{z}{x}\right)$ y que $\frac{xr(z)}{b} = r\left(\frac{bz}{x}\right)$.

Como $f(x) \geq 0$ en $[1, b)$:

$$\begin{aligned}\left|\frac{f(z)-r(z)}{r(z)}\right| &\leq \int_1^z f(x)D(g)dx + \int_z^b f(x)D(g)dx \\ &\leq D(g).\end{aligned}$$

Por lo tanto, $D(h) \leq D(g)$, y como f y g son claramente intercambiables:

$$D(h) \leq \text{mín}(D(f), D(g)).$$

Evidentemente, y al no ser una desigualdad estricta, la convergencia no está asegurada (por ejemplo si tomásemos como f la función de Dirac, δ_1, $D(h) = D(g)$) pero en muchos de los casos si se dará la desigualdad estricta y llegaremos a la convergencia a la ley de Benford.

Observación 2.8. Podemos también demostrar un resultado análogo para la división de variables aleatorias aunque tendríamos que hallar una expresión para la división análoga a (2.10) y seguir el proceso de la demostración anterior.

En la práctica, Hamming demuestra (ver [4]) como hay una convergencia bastante rápida a una distribución Benford al multiplicar n distribuciones uniformes. Presenta esta tabla donde se relaciona el número de variables aleatorias uniformes que se multiplican con la distancia relativa con la ley de Benford (Figura 2.4).

Número de operadores	Distancia a la ley de Benford
1	1.558
2	0.3454
3	0.0980
4	0.0289

Figura 2.4: Tabla de Hamming en relación a la distancia de un producto de uniformes con la ley de Benford.

2.5 Ley de Benford en sucesiones

2.5.1 Definición y teoremas de interés

Definición 2.14. *Una sucesión real* $\{a_n\}_{n\in\mathbb{N}}$ *se dice que es una sucesión Benford si*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I_{S(a_n)\leq t} = \log(t).$$

Observación 2.9. La definición anterior es una definición clásica de sucesión Benford, quiere decir que la probabilidad de que $S(a_n) \leq t$ tiende a la ley logarítmica cuando hacemos tender $N \to \infty$. Esta definición es la que siguen las sucesiones "fuertemente" Benford. Podemos dar versiones de esta definición que nos darán como resultado las "débilmente" Benford aunque no serán objeto de nuestro estudio.

Las sucesiones Benford están fuertemente relacionadas con las famosas sucesiones uniformemente distribuidas módulo 1 (ver [16]). Todos los resultados trascendentales acerca de ellas están recopilados por Kuipers y Niederreiter (ver [9]). El siguiente teorema nos muestra la relación explicita:

Teorema 2.5. *La sucesión real* $\{a_n\}_{n\in\mathbb{N}}$ *es Benford si y solo si la sucesión* $\{\log(a_n)\}_{n\in\mathbb{N}}$ *está uniformemente distribuida módulo 1.*

Demostración. La prueba dada en este trabajo es una versión más corta (y quizás algo menos rigurosa) que la dada por Diaconis (ver [2]).

Si $\{\log(a_n)\}$ es uniformemente distribuida módulo 1, esta sucesión se puede considerar como una muestra de una distribución uniforme. De acuerdo con el método

de generación de variables aleatorias Benford descrito en la sección 2.1.3, podemos considerar la sucesión $\{a_n\}$ como una muestra de una variable aleatoria que satisface la ley de Benford. Utilizando este hecho y la definición de sucesión Benford, $\{a_n\}$ es una sucesión Benford. El recíproco es completamente análogo.

Para proseguir con el estudio de algunas sucesiones Benford, es necesario citar algunos resultados previos sobre las sucesiones uniformemente distribuidas módulo 1 cuyas demostraciones se pueden encontrar en [9]:

Teorema 2.6. *Si $\{x_n\}$ está uniformemente distribuida módulo 1 e $\{y_n\}$ es tal que $\lim_{n\to\infty}(x_n - y_n) = \alpha \in \mathbb{R}$ constante, entonces $\{y_n\}$ está uniformemente distribuida módulo 1.*

Teorema 2.7 (Criterio de Weyl). *$\{x_n\}$ está uniformemente distribuida módulo 1 si y solo si para todo $h \in \mathbb{N}^*$:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{2i\pi h x_n} = 0.$$

Teorema 2.8 (Estimación de Van der Corput para sumas trigonométricas). *Sean $a, b \in \mathbb{Z}$, $a < b$, y f una función dos veces diferenciable en $[a, b]$ con $f''(x) \geq \rho > 0$ o $f''(x) \leq -\rho < 0$ para $x \in [a, b]$. Entonces:*

$$\left| \sum_{n=a}^{b} e^{2i\pi f(n)} \right| \leq (|f'(a) - f'(b)| + 2)\left(\frac{4}{\sqrt{\rho}} + 3\right). \tag{2.11}$$

Teorema 2.9 (Teorema de Fejer). *Si $\{f(n)\}$ está uniformemente distribuida módulo 1, entonces:*

$$\limsup n|f(n+1) - f(n)| = \infty.$$

2.5.2 Sucesión geométrica

Proposición 2.5. Una sucesión geométrica $\{a^n\}$ es Benford si y solo si $\log(a)$ es irracional.

Demostración. Por el Teorema 2.5, $\{a^n\}$ es Benford si y solo si $\{n\log(a)\}$ está uniformemente distribuida módulo 1. Es conocido que las únicas sucesiones de la forma $\{n\alpha\}$ que están uniformemente distribuidas modulo 1 son aquellas con α irracional (ver [16]).

Ejemplo 2.1. Evidentemente, muchas sucesiones clásicas como $\{2^n\}$, $\{3^n\}$ o $\{5^n\}$ cumplen esta condición y por lo tanto son Benford.

2.5.3 $\{n!\}$ y $\{n^n\}$

Ambas sucesiones son Benford. Mostramos aquí la demostración utilizando tanto el criterio de Weyl como la estimación de Van der Corput:

- Para $\{n^n\}$ la demostración es bastante directa: Sean $h \in \mathbb{Z}^*$, $N \in \mathbb{N}^*$, y sustituyamos $a = 1, b = N$ y $f(n) = hn\log(n)$ en (2.11):

$$\begin{aligned} f'(n) &= h\log(n) + h \\ f''(n) &= \frac{h}{n} \\ \Rightarrow \rho &= \frac{|h|}{N} \end{aligned}$$

Sustituyendo lo anterior en la estimación de Van der Corput:

$$\begin{aligned} \frac{1}{N}\left|\sum_{n=1}^{N} e^{2i\pi hn\log(n)}\right| &\leq \frac{1}{N}(|h|\log(N)+2)\left(4\sqrt{\frac{N}{|h|}}+3\right) \\ &= O\left(\frac{\log(N)}{N^{1/2}}\right) \\ &\to 0 \quad \text{cuando} \quad N \to \infty. \end{aligned}$$

Por lo tanto, aplicando el criterio de Weyl, la sucesión $\{n\log(N)\}$ está uniformemente distribuida módulo 1 y por lo tanto por el Teorema 2.5 $\{n^n\}$ es Benford.

- La demostración para $\{n!\}$ es un poco más difícil: Recordamos en primer lugar la fórmula de Stirling:

$$n! \backsim \frac{1}{\sqrt{2\pi}} n^{(n+1/2)} e^{-n}$$

Haciendo uso de esta observamos que

$$\log(n!) - \left(\left(n + \frac{1}{2}\right)\log(n) - n\log(e)\right)$$

tiende hacia una constante cuando $n \to \infty$. Por el Teorema 2.6. es suficiente probar que $\left\{\left(n + \frac{1}{2}\right)\log(n) + kn\right\}$ con $k = -\log(e)$ está uniformemente distribuida módulo 1.
Sean entonces $h \in \mathbb{Z}^*$, $N \in \mathbb{N}^*$, y sustituyamos $a = 1, b = N$ y $f(n) = h\left(n + \frac{1}{2}\right)\log(n) + hkn$ en (2.11):

$$\begin{aligned} f'(n) &= h\log(n) + h + \frac{h}{2n} + hk \\ f''(n) &= \frac{h}{n} - \frac{h}{2n^2} \\ \Rightarrow \rho &= |h|\left(\frac{h}{N} - \frac{h}{2N^2}\right) \end{aligned}$$

Sustituyendo lo anterior en la estimación de Van der Corput:

$$\begin{aligned} \frac{1}{N}\left|\sum_{n=1}^{N} e^{2i\pi f(n)}\right| &\leq \frac{1}{N}\left(|h|\log(N) + \left|\frac{h}{2}\right|\left(\frac{1}{N} - 1\right) + 2\right)\left(\frac{4\sqrt{2}N}{\sqrt{|h|(2N-1)}} + 3\right) \\ &= O\left(\frac{\log(N)}{\sqrt{N}}\right) \\ &\to 0 \quad \text{cuando} \quad N \to \infty. \end{aligned}$$

Por lo tanto, utilizando el criterio de Weyl y el Teorema 2.5 llegamos al resultado.

2.5.4 Otras sucesiones

Veamos un corolario evidente del Teorema 2.5. utilizando $f(n) = \ln(a_n)$ que nos da una condición necesaria para ver si una sucesión es Benford.

Corolario 2.2. Si $\{a_n\}$ es una sucesión Benford entonces $\lim_{n\to\infty} n \ln\left(\frac{a_{n+1}}{a_n}\right) \longrightarrow \infty$.

Utilizando este corolario se comprueba fácilmente que para cualquier b escogido las sucesiones $\{n^b\}$, $\{bn\}$ y $\{\log_b n\}$ no son Benford, ya que al hacer los límites obtenemos b, 1 y 0 respectivamente.

Finalmente, encontramos algunas sucesiones más que son Benford como por ejemplo $\left\{\binom{n}{k}\right\}$ (ver [2]), o las sucesiones de Lucas y Fibonacci (ver [15]).

3 Ley de Benford en bases de datos reales

En este capítulo trabajaremos con tres bases de datos distintas:

- **Censo municipios españoles**: Base de datos de los censos de todos los municipios españoles en 2010 y 2021 con 8114 entradas (www.ine.es).
- **Censo municipios Comunidad de Madrid**: Base de datos de los censos de todos los municipios de la Comunidad de Madrid en 2021 con 180 entradas (www.ine.es).
- **Censo ciudades EE.UU.**: Base de datos de los censos de las ciudades de todos los estados de EE.UU. con 19494 entradas (www.census.gov).

3.1 Test de bondad de ajuste χ^2

En este primer apartado recordaremos las características principales de este test estadístico que usaremos para ver si una cierta base de datos se ajusta o no a la Ley de Benford.

Cuantificar la distancia entre una cierta base de datos y la ley de Benford puede ser un tanto complicado ya que para un conjunto de datos de este tamaño el test Chi-cuadrado no acepta la hipótesis nula para niveles de confianza razonables.

Recordemos que el estadístico Chi-cuadrado, s, se define como:

$$s = N \cdot \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

donde N es el tamaño de la muestra de datos que estamos tratando, n el número de clases, f_i la frecuencia observada y e_i la frecuencia esperada de la clase i.

La hipótesis nula, H_0, de este test es: "los datos provienen de una distribución con función de densidad $\{f_i\}_{i=1,\dots,n}$" que será rechazada para valores de s grandes. Bajo esta hipótesis nula se tiene $S \backsim \chi^2_{n-1}$ y por lo tanto el p-valor del test es $1 - F_{n-1}(s)$ con F_{n-1} la función de distribución de χ^2_{n-1}.

Tenemos que tener también en cuenta que para bases de datos con una cantidad considerable de entradas (N grande), el estimador s tenderá a ser grande y aunque haya muchas entradas quizás la muestra no se acerque lo suficiente a la distribución esperada para disminuir s (y por lo tanto para aceptar la hipótesis nula). Por este motivo, durante el análisis de nuestras bases de datos haremos hincapié tanto en el estadístico chi-cuadrado, s, como en la distancia entre la base de datos y la distribución esperada, es decir, $\sum_{i=1}^{n}(f_i - e_i)^2$.

Consideraremos así que el orden decimal de la distancia nos da una buena idea de como nuestra base de datos se ajusta a la Ley de Benford. En particular, para el primer dígito en base 10, una distancia del orden de 10^{-4} se considerará un valor medio de conformidad: Esto significa que de media $|f_i - e_i|$ es de orden entre 10^{-2} y 10^{-3}, es decir, que la diferencia absoluta entre la frecuencia esperada y la observada es de entre un 0.1 % y un 1 %.

El estadístico chi-cuadrado y el p-valor deberán ser considerados entonces con bastante tolerancia.

3.2 Ajuste de la Ley de Benford a bases de datos

En esta sección veremos como se ajustan los datos de nuestras tres bases de datos a la Ley de Benford, estableciendo así una relación evidente entre el número de entradas de la muestra y el orden de las distancias con la Ley de Benford. Además, también estudiaremos las diferencias absolutas y relativas de los datos de los censos de municipios españoles de los año 2010 y 2021.

Este estudio lo veremos como una de las aplicaciones de la Ley de Benford: La creación de modelos que puedan prever el crecimiento demográfico.

Para el primer dígito de los 8114 municipios españoles, los datos del test χ^2 de bondad de ajuste descrito en la sección anterior son los dados en la siguiente tabla.

La explicación de que los datos se ajusten bien a la ley es bastante sencilla. Tiene sentido suponer que el crecimiento demográfico en un municipio es una sucesión geométrica con respecto al tiempo ya que la velocidad de crecimiento es prácticamente constante, luego, los censos de cada municipio se ajustaran bastante bien a la Ley de Benford con respecto al tiempo por lo estudiado en la subsección 2.5.2, y por lo tanto, nuestra base de datos tanto en 2010 como en 2021 tienen un valor medio de conformidad a la Ley de Benford por lo dicho en la sección anterior. Podemos verlo gráficamente en las figuras 3.1 y 3.2.

Año	Distancia	Chi-cuadrado	p-valor
2010	2.62924e-04	14.04939	0.08048
2021	6.78587e-05	7.60323	0.47315

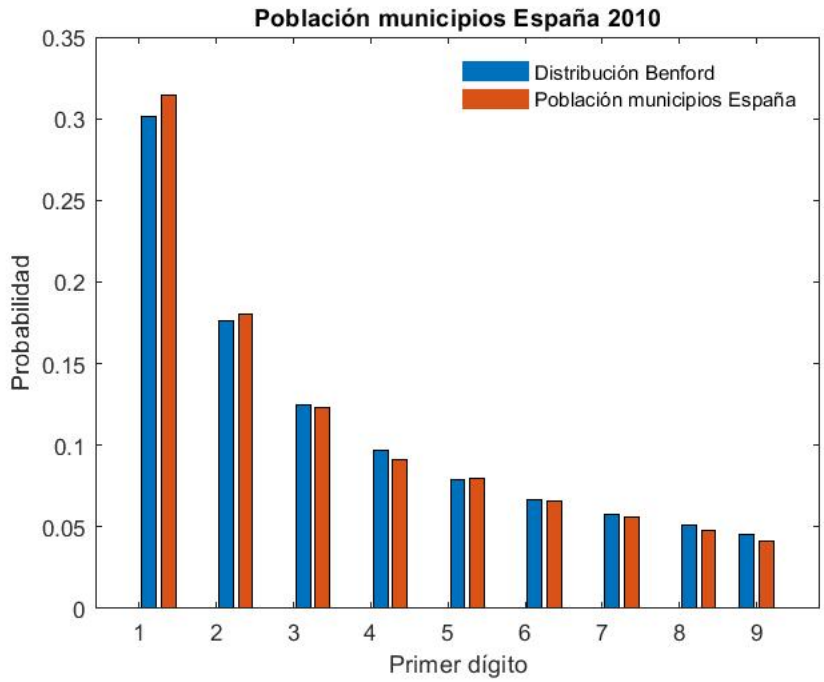


Figura 3.1: Histograma del censo de 2010 de municipios españoles.

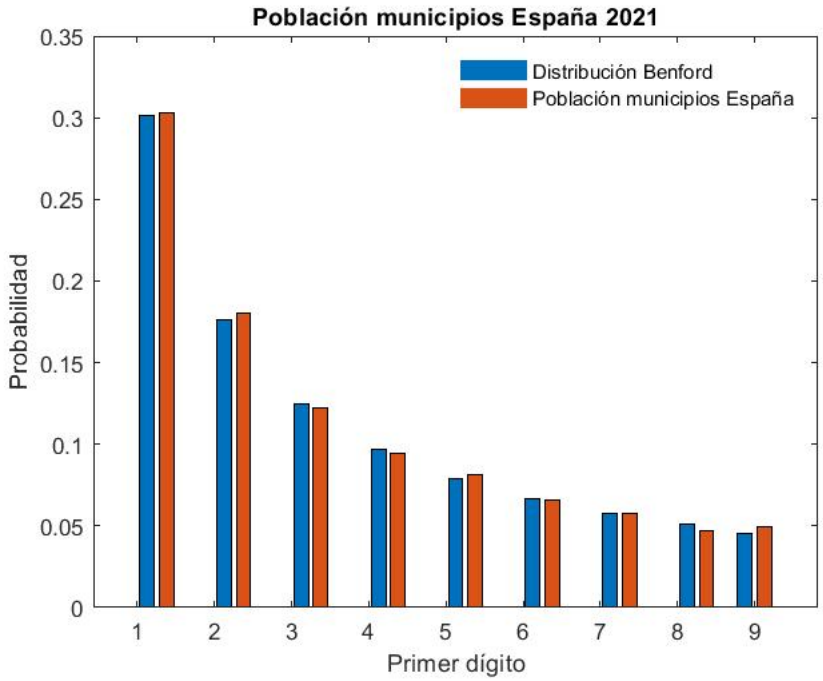


Figura 3.2: Histograma del censo de 2021 de municipios españoles.

Los cálculos para el primer dígito de los censos de las 19494 ciudades de EE.UU. son según el test χ^2:

Año	Distancia	Chi-cuadrado	p-valor
2020	7.51254e-05	17.32042	0.02694

Vemos como al haber aumentado el número de entradas la distancia se reduce y alcanza un orden de 10^{-5} (muy buen ajuste a la Ley del primer dígito) debido al carácter asintótico de la ley. Pese a esto, notamos que el tamaño de nuestra base de datos hace que el estadístico chi-cuadrado se vea penalizado y aumente, como comentamos que pasaría anteriormente. Gráficamente:

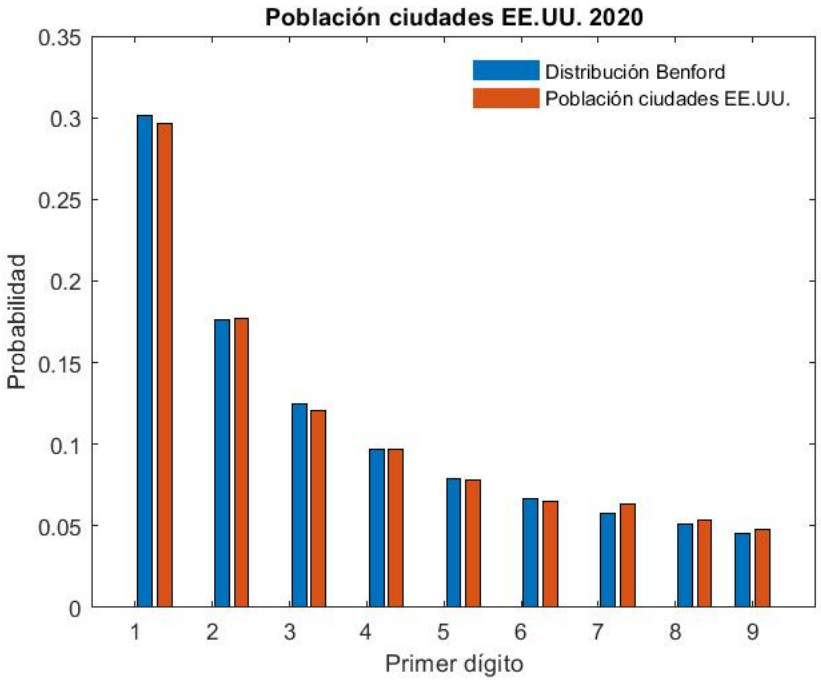


Figura 3.3: Histograma censo ciudades EE.UU.

Análogamente, pero utilizando una base de datos de menor tamaño (censo municipios Madrid 2021), podemos comprobar que la distancia es del orden de 10^{-3} lo cual nos dice que está diez veces más alejada de la Ley de Benford que lo que hemos considerado un valor medio de conformidad. La explicación de esto es análoga a la de EE.UU. pero teniendo en cuenta que a menos entradas en la base de datos más difícil será ver un ajuste a la Ley de Benford (ver Figura 3.4).

Año	Distancia	Chi-cuadrado	p-valor
2021	4.75476e-03	11.99719	0.15132

Retomemos ahora la base de datos inicial (censo municipios españoles 2010 y 2021) y hagamos un estudio de sus diferencia absoluta y relativa.

La diferencia absoluta puede ser vista como una muestra de $X - Y$ con X e Y dos distribuciones que satisfacen la Ley de Benford. Los datos mostrados en la siguiente tabla sobre la diferencia absoluta de nuestras bases de datos nos muestran que aunque en la teoría no hemos llegado a ningún resultado acerca de la invarianza por resta de distribuciones Benford, algunas veces se cumple.

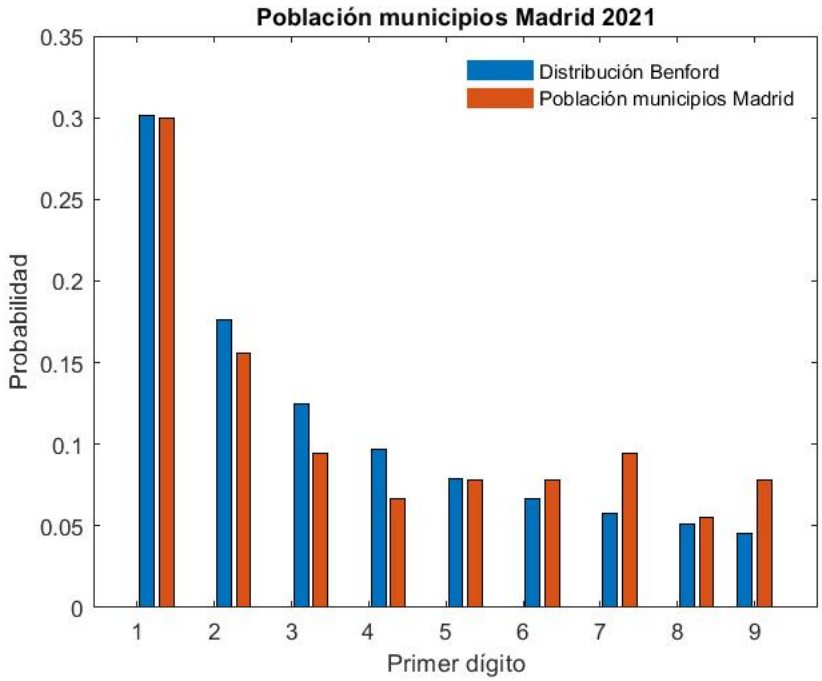
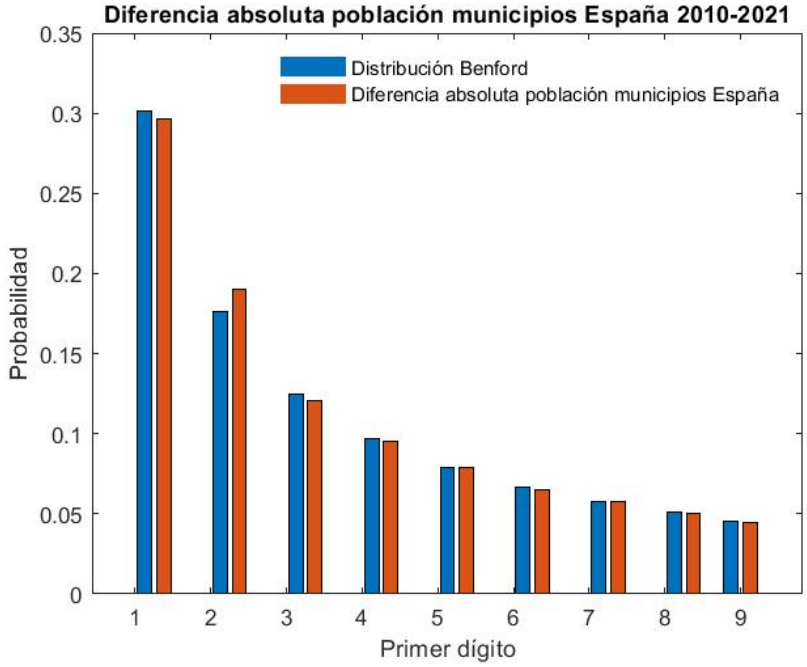
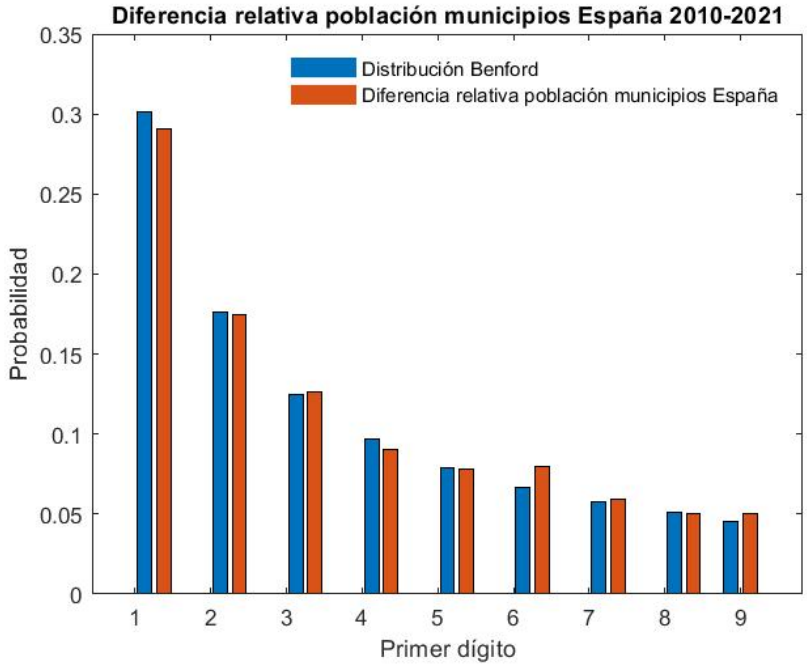


Figura 3.4: Histograma censo municipios Madrid.

La diferencia relativa puede ser vista como una muestra de $\frac{X-Y}{X}$, como $X - Y$ hemos visto que sigue bastante bien la Ley de Benford, la diferencia relativa no es más que la multiplicación (por la inversa de X, que también es Benford) de distribuciones Benford, que según la sección 2.4.2 se mantendrá siendo Benford, lo cual concuerda con los datos mostrados a continuación.

Diferencia	Distancia	Chi-cuadrado	p-valor
Absoluta	2.37335e-04	11.46134	0.17690
Relativa	3.25525e-04	29.18169	2.94637e-04


Diferencia absoluta población municipios España 2010-2021
Distribución Benford
Diferencia absoluta población municipios España
Probabilidad
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
Primer dígito


Diferencia relativa población municipios España 2010-2021
Distribución Benford
Diferencia relativa población municipios España
Probabilidad
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
Primer dígito

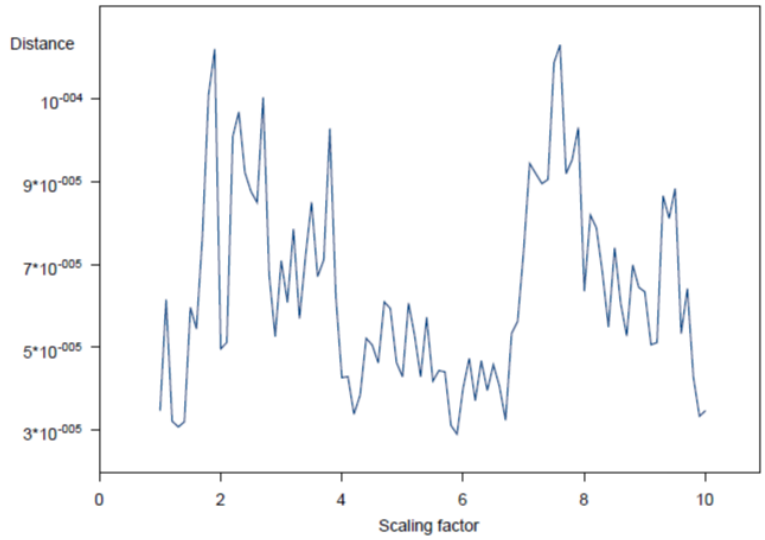
3.3 Propiedades Ley de Benford en bases de datos

En esta sección volvemos a utilizar la base de datos de los censos de las ciudades de EE.UU. la cual llamaremos a partir de ahora D, una base de datos de 19494 entradas con una mezcla de datos de censos de ciudades y estados de EE.UU. La procedencia de estos datos a partir de ahora no tendrá gran importancia ya que se verá como una base de datos Benford de un tamaño considerable con un poco de ruido "natural" (la cual es más conveniente que una base de datos que se ajuste a la perfección a la ley de Benford). La conformidad con la Ley de Benford recordamos que es muy alta y está calculada numéricamente en la sección anterior.

3.3.1 Invarianza de escala

Se ha multiplicado la base de datos D por 91 constantes espaciadas regularmente entre 1 y 10. La siguiente gráfica muestra la distancia a la Ley de Benford para el primer dígito de cada una de las bases de datos resultado de los cambios de escala.

Podemos observar que las distancias a la Ley de Benford varían con una media de $6.4 \cdot 10^{-5}$ y una varianza de $5.6 \cdot 10^{-10}$. La invarianza de escala por lo tanto es evidentemente observada ya que el orden decimal de la distancia no supera 10^{-4} en el peor de los casos. Notamos que, afortunadamente, la base de datos original es una de las mejores escalas.



3.3.2 Invarianza de base

En esta subsección expresamos la base de datos D en ocho bases diferentes, desde 3 hasta 10. No se estudiará en este trabajo las bases mayores de 10 ya que tienen un coste computacional mucho mayor. La base 2 tampoco será tratada ya que el primer dígito de un número en base 2 siempre será el 1.

La siguiente tabla muestra el estadístico chi-cuadrado, la distancia y el p-valor para el primer dígito de cada una de las bases de datos resultado del cambio de base mencionado. Aquí, el estadístico y la distancia no tienen el mismo significado en cada base (en base b existen $b-1$ clases para el primer dígito, luego tenderán a ser más pequeños cuando estudiemos las bases más pequeñas). Teniendo esto en cuenta, la distancia media $|f_i - e_i|$ en todas las bases ha resultado ser del orden de 10^{-3} según nuestro análisis numérico, luego se cumple la invarianza de base como era de esperar según los resultados teóricos del capítulo 2.

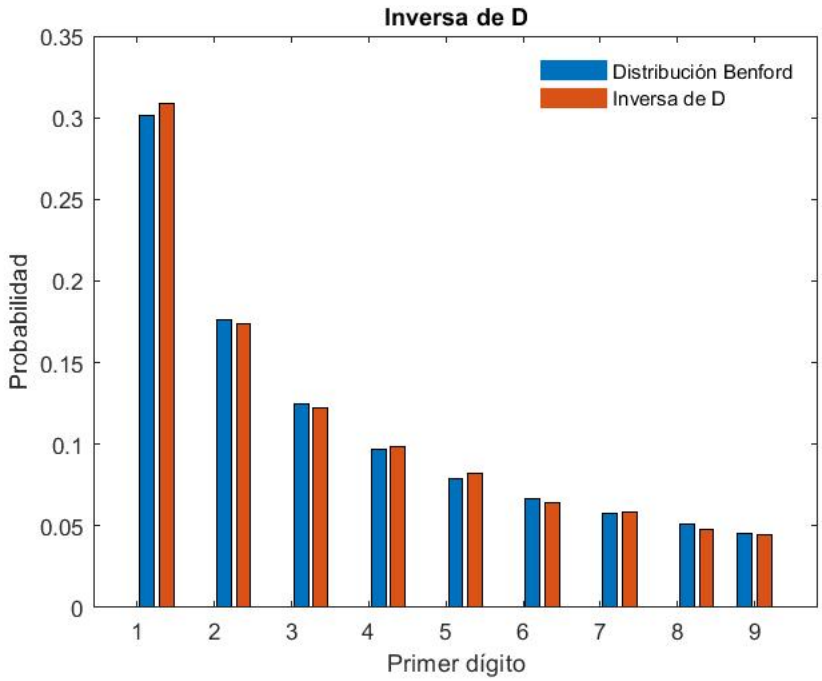
En el estudio de esta propiedad el estimador chi-cuadrado parece no ser relevante para este análisis y los p-valores hay que cogerlos con mucha tolerancia, pero fijándonos sobre todo en la distancia llegamos a la conclusión de que la propiedad se cumple.

Base	Chi-cuadrado	Distancia	p-valor
3	0.1	2.7e-06	0.74
4	5.7	7.1e-05	0.06
5	13.4	1.4e-04	0.00
6	13.9	1.3e-04	0.01
7	14.1	1.1e-04	0.01
8	20.4	1.9e-04	0.00
9	12.9	8.1e-05	0.08
10	6.1	3.5e-05	0.64

3.3.3 Inversa de distribución Benford

Los datos obtenidos numéricamente a partir de nuestra base de datos D nos muestran que se satisface que la inversa de una distribución Benford también es Benford.

Distribución	Distancia	Chi-cuadrado	p-valor
Inversa de D	1.08640e-04	16.18600	0.03979



3.3.4 Multiplicación de distribución Benford

Las figuras 3.5, 3.6 y 3.7 son histogramas del primer dígito del producto de D por una muestra (de tamaño 19494) de la distribución $U(0,1)$, $N(0,1)$ y $Exp(1)$. Como era de esperar por el resultado teórico que se probó en la sección 2.4.2, los tres productos siguen la Ley de Benford.

Distribución	Distancia	Chi-cuadrado	p-valor
D.*U(0,1)	6.36817e-05	12.94858	0.11363
D.*Exp(1)	2.47373e-05	5.31857	0.72304
D.*N(0,1)	3.25305e-05	8.55802	0.38093

Pese a que los p-valores tienen que escogerse con una gran tolerancia, es evidente que si el nivel medio de conformidad con la Ley de Benford se da cuando la distancia es del orden de 10^{-4}, con los resultados obtenidos sobre la distancia de los tres productos que son del orden de 10^{-5}, podemos asegurar un muy buen ajuste de dichas bases de datos con la Ley de Benford. Gráficamente:

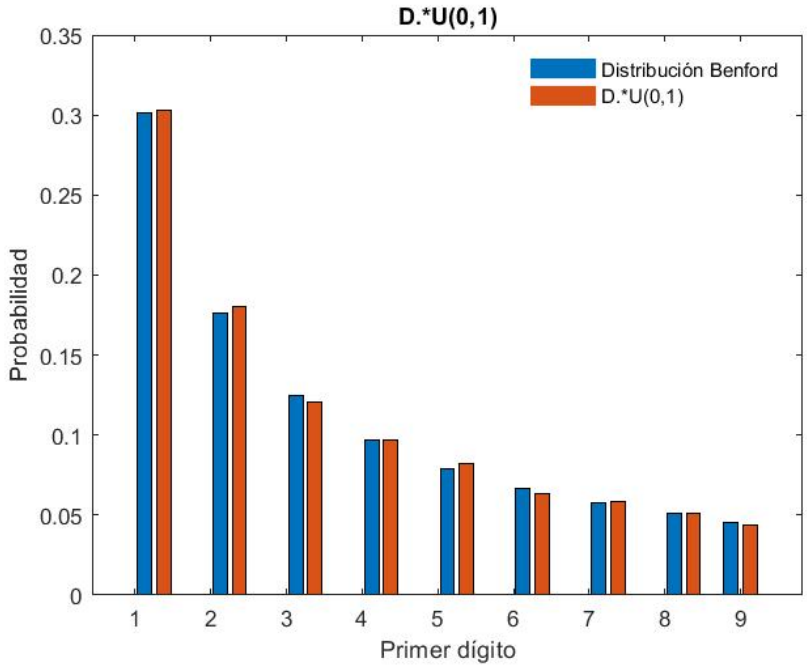


Figura 3.5: Histograma D.*$U(0,1)$.

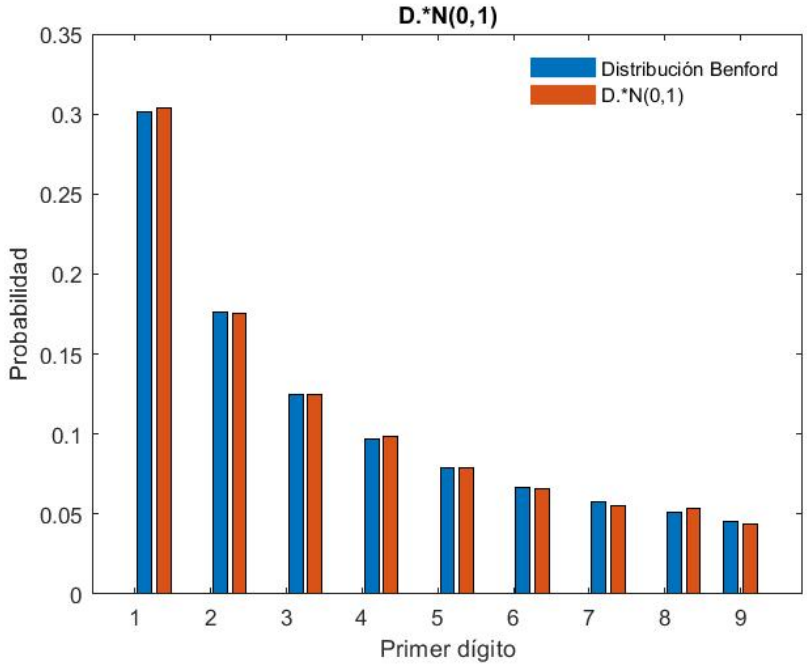


Figura 3.6: Histograma D.*$N(0,1)$.

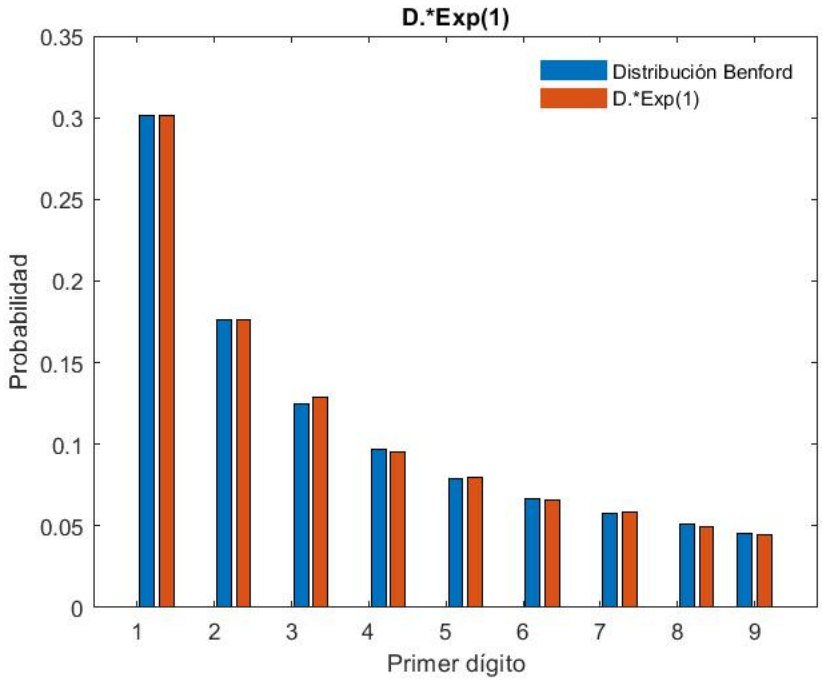


Figura 3.7: Histograma D.*$Exp(1)$.

3.4 Detección de fraudes

La Ley de Benford es utilizada por contables para detectar el fraude o la presencia de datos duplicados. Es una *regla parcialmente negativa*, es decir, si la Ley de Benford no se satisface entonces es probable que los datos hayan sido manipulados, aunque una base de datos que satisfaga la ley puede haber sido manipulada.

Realmente, este es el caso de la base de datos D. Esta base de datos ha sido obtenida mezclando censos de ciudades y estados. Lo que la hace "fraudulenta" es que algunas entradas son suma de otras (ya que los estados se subdividen en muchas ciudades). Por lo tanto tenemos que muchos de los datos de D están duplicados (no en el sentido común sino en el sentido de que hay datos que se pueden expresar como sumas de otros).

Realizamos otro experimento para estudiar la relevancia de la Ley de Benford en detección de fraudes. Contaminamos nuestra base de datos D con una muestra de una distribución normal con la misma media y varianza que la de nuestra base de

datos original. Reemplazamos aleatoriamente en D los datos de esta muestra por los datos fraudulentos que contenía D desde el principio (reemplazamos exactamente el número de datos fraudulentos que hay, por lo que el tamaño de la base de datos D sigue siendo el mismo con estos cambios). Este modelo ha sido escogido como un modelo para realizar un fraude inteligentemente, aunque hay muchos más que se podrían utilizar.

La figura 3.8. muestra la evolución de la distancia de la Ley de Benford de acuerdo con el tamaño de la muestra fraudulenta (expresado en porcentaje con respecto al total de datos de la base). El rango de contaminación escogido está entre el 0 % y 10 % para que sea realista.

Debemos hacer algunos comentarios acerca de esto:

- Encontramos una curva exponencial (en la gráfica) como un buen ajuste a esta distancia, lo que significa que la distancia de la Ley de Benford aumenta exponencialmente cuando el tamaño de la muestra aumenta. Esto hace que la distancia con la ley sea una herramienta buena para la detección de fraudes, pero entonces los problemas se reducen a encontrar un buen nivel de alerta.
- La varianza en la distancia es grande, especialmente para niveles altos de contaminación de la base de datos, por lo que este nivel de alerta no será del todo preciso. Por ejemplo, encontramos un valor de la distancia parecido para un nivel de contaminación del 3.5 % y del 8.5 %. Entonces vemos que separar el ruido "natural" de la base de datos con el fraude puede ser un tanto complicado.
- El orden decimal de la distancia en si de alguna manera contradice lo dicho en la primera sección de este capítulo (que 10^{-4} es un valor medio de conformidad). Aquí, un orden de 10^{-4} es bastante malo ya que coincide con un nivel de contaminación del 10 %. Este tipo de análisis ayudan a cuantificar que debe ser esperado de la distancia.

Evidentemente, todos los cálculos hechos aquí están hechos con una base de datos demográfica que nadie tiene interés de falsear. Además, el tipo de fraude es solo un ejemplo, hay otros muchos casos donde el fraude consiste en duplicar los datos directamente.

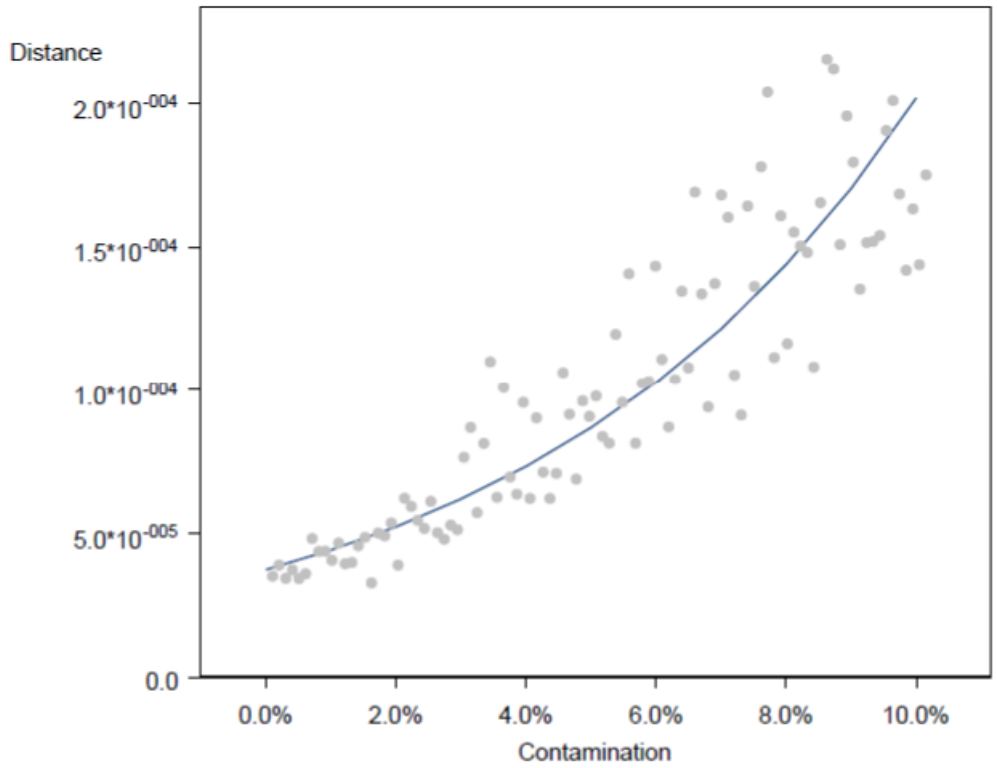


Figura 3.8: Detección de datos fraudulentos en D (ver [8]).

Conclusiones

Los resultados teóricos principales sobre el estudio de la Ley de Benford quedan resumidos de la siguiente forma:

1. Una variable aleatoria X satisface la Ley de Benford si y solo si $\log X - \lfloor \log X \rfloor \backsim U(0,1)$ (sección 2.1.3).
2. Una variable aleatoria X tiene invarianza de escala si y solo si X se ajusta a la Ley de Benford (sección 2.3.2).
3. Una variable aleatoria X es invariante ante cambios de base si y solo si se ajusta a la Ley de Benford (sección 2.3.3).
4. Una variable aleatoria X tiene invarianza de suma si y solo si se ajusta a la Ley de Benford (sección 2.3.4).
5. Si una variable aleatoria X sigue la Ley de Benford, entonces X^{-1} también la seguirá (sección 2.4.1).
6. Si una variable aleatoria continua Y satisface la Ley de Benford, para toda variable aleatoria con f.d.d. continua X, XY satisface la Ley de Benford. Además, $\frac{X}{Y}$ también la satisfacerá si X e Y son no nulas (sección 2.4.2).
7. Dadas dos distribuciones cualesquiera, su producto se acerca a la Ley de Benford manteniéndose en el peor de los casos la distancia del producto como la menor de las dos distancias (hablando de distancias a la Ley de Benford). Como consecuencia de esto se comprueba como el producto de varias distribuciones como el de n distribuciones $U(0,1)$ se acerca a la Ley de Benford con una velocidad sorprendente (sección 2.4.3).
8. Una sucesión real $\{a_n\}_{n\in\mathbb{N}}$ es Benford si y solo si $\{\log(a_n)\}_{n\in\mathbb{N}}$ está uniformemente distribuida módulo 1 (sección 2.5.1).

9. Las sucesiones geométricas $\{a^n\}$ son Benford si y solo si $\log(a)$ es irracional. Además, se demostró que las sucesiones $\{n!\}$ y $\{n^n\}$ son Benford y se dio una condición necesaria para sucesiones Benford que nos servía para llegar a sucesiones que no eran Benford como: $\{n^b\}$, $\{bn\}$ o $\{\log_b(n)\}$ (para cualquier b) (secciones 2.5.2-2.5.4).

En cuanto a los resultados numéricos el autor cree que hay también hay algunos hechos a destacar:

1. La ley de Benford aparece en bases de datos con un gran número de entradas, ejemplo de ello son los cálculos realizados para las bases de datos de censos de ciudades de EE.UU. y municipios de Madrid. Ante esta situación encontramos un problema: el estadístico chi-cuadrado pierde utilidad ya que tiende a ser grande cuando el tamaño de la base de datos es grande. La solución que se aplica es utilizar como estadístico la distancia absoluta, que nos dará una buena idea acerca del ajuste de los datos a la ley (sección 3.2).
2. Se comprueban las distintas propiedades citadas como resultados teóricos para bases de datos grandes (censo ciudades de EE.UU. con 19494 entradas): Invarianza de escala, de base, inversa de una distribución Benford o multiplicación de distribuciones Benford con $\boldsymbol{U}(0,1)$, $\boldsymbol{N}(0,1)$ y $\boldsymbol{Exp}(1)$ (sección 3.3)
3. También vemos que se comprueba el resultado de la división de una distribución Benford en el apartado en el que se hace el estudio de la diferencia relativa entre el censo de municipios españoles de 2010 y 2021 (sección 3.2).
4. También hacemos referencia a la aplicación principal de la Ley Benford, la detección de fraudes. Es evidente que la distancia con la ley es un motivo de alarma en cuanto al posible falseamiento de los datos. La tendencia de la distancia a la ley en relación al porcentaje de datos falseados es exponencial y ofrecemos en la sección 3.4 una manera de falsear datos de manera que se siga cumpliendo la ley. La tarea principal a desarrollar en esta aplicación es encontrar un estadístico (nivel de alerta) eficiente para la detección del falseamiento de datos (sección 3.4).

Como líneas de investigación que se pueden seguir posteriores a los resultados demostrados en este trabajo el autor cree que destacan tres de ellas: La demostración de la conjetura 2.1, el diseño de un método preciso para el estudio de la conformidad de grandes bases de datos con la Ley de Benford y por último, la búsqueda de otras distribuciones para los dígitos ya que la Ley de Benford no tendría que ser la única interesante.

Bibliografía

[1] F. Benford. «The law of anomalous numbers». En: *Proceedings of the American Philosophical Society* 78(4) (1938), págs. 551-572.

[2] P. Diaconis. «The distribution of leading digits and uniform distribution mod 1». En: *Annals of probability* 5 (1977), págs. 72-81.

[3] P. Diaconis y D Freedman. «On rounding percentages». En: *J, American Statistics Association* 74(366) (1979), págs. 359-364.

[4] R. Hamming. «On the distribution of numbers». En: *Bell system Technical Journal* 49 (1970), págs. 1609-1625.

[5] Theodore Hill. «Base-invariance implies Benford´s law». En: *Proceedings of the American Mathematical Society* 123 (1995), págs. 887-895.

[6] Theodore Hill. «The significant digit phenomenon». En: *American Mathematical Monthly* 103 (1995), págs. 322-327.

[7] Theodore Hill. «The significant-digit law». En: *Statistical Science* 10(4) (1995), págs. 354-363.

[8] Adrien Jamain. «Benford´s Law». Imperial College of London, 2001.

[9] L. Kuipers y H. Niederreiter. *Uniform distribution of sequences*. Wiley-New York, 1974.

[10] L.M. Leemis y B.W. Schmeiser. «Survival distributions satisfying Benford's Law». En: *The American Statician* 54(4) (2000), págs. 236-241.

[11] M. Muñoz de Ȯzak y L. Blanco Catañeda. *Introducción a la teoría avanzada de la probabilidad*. 1.ª ed. Universidad nacional de Colombia, 2002.

[12] S. Newcomb. «Note on the frequency of use of the different digits in natural numbers». En: *American Journal of Mathematics* 4 (1881), págs. 39-40.

[13] Nigrini.M. «The detection of income evasion through an analysis of digital distributions». University of Cincinatti, 1992.

[14] R.S. Pinkham. «On the distribution of first significant digits». En: *Annals of Mathematical Statistics* 32 (1961), págs. 1223-1230.

[15] P. Schatte. «On mantissa distributions in computing and Benford's Law». En: *Journal of Information Processing and Cybernetics* 24 (1988), págs. 443-455.

[16] H. Weyl. «Über die Gleichverteilung von Zahlen mod Eins». En: *Mathematische Annalen* 77 (1916), págs. 313-352.