Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

# LA LEY DE BENFORD, DEL PRIMER DÍGITO SIGNIFICATIVO

*Autor: Álvaro Villameriel Cuenca*

*Tutor: Carlos Matrán Bea*

# Introducción

Vivimos rodeados de números: nuestras casas están numeradas, todos tenemos un carnet de identidad y de seguridad social. Cada coche tiene una matrícula, llamamos a alguien marcando su número y en el supermercado, cada artículo tiene un precio. Cabe preguntarse si hay patrones en los números que vemos a diario. Pensemos en los dígitos de esos números, desde el 0 hasta el 9, teniendo en cuenta la posición que ocupan. Por ejemplo, examinemos la frecuencia de los dígitos en la primera posición, los más a la izquierda. Ahora, examinemos la frecuencia de los dígitos que se encuentran en la segunda posición. Observamos que las proporciones son diferentes, pero, ¿los dígitos no aparecen aleatoriamente?, ¿por qué sería más frecuente el número 7 en la segunda posición que en la primera?, ¿hay entonces alguna ley distribucional universal que gobierna la frecuencia con la que aparecen los dígitos?

Es un hecho observado que en muchas tablas de datos numéricos el primer dígito significativo no está uniformemente distribuido, como podría esperarse. Muchas tablas dan una frecuencia aproximadamente igual a $\log_{10}\left(\frac{p+1}{p}\right)$, en otras palabras, el número 1, como primer dígito significativo, aparece aproximadamente el 30 % de las veces, el número 2 aparece aproximadamente el 18 % de las veces, y esta disminución de la frecuencia relativa continúa hasta el número 9, que aparece menos del 5 % de las veces.

Esta peculiar distribución logarítmica ha dado lugar a una abundante literatura. La primera referencia conocida vino de la mano de Simon Newcomb [18], en 1881. Newcomb se percató de que en los libros de tablas logarítmicas las primeras páginas estaban más desgastadas que las últimas y vía un argumento heurístico concluye que la aparición del número 1 como primer dígito significativo es muy frecuente, y que la aparición del número 9 como primer dígito significativo es poco frecuente. Más específicamente sugirió que

$$\mathrm{Prob}(D_1 = d_1) = \log_{10}\left(\frac{1+d_1}{d_1}\right) \quad d_1 = 1, 2, \ldots, 9. \tag{1}$$

No fue hasta 1938 cuando el físico Frank Benford popularizó el problema. En su artículo, La Ley de los números anómalos [1], Benford recoge una gran cantidad de datos provenientes de numerosos campos, más de 20,000 observaciones, entre las que se encuentran: áreas de ríos, estadísticas de la Liga Americana de béisbol, números aleatorios en revistas, términos de la sucesión armónica, datos sobre el índice de mortalidad, facturas de la luz, direcciones postales,... Concluyó que la frecuencia del primer dígito significativo de estos datos, en conjunto, se ajustaba a (1).

La Ley de Benford (LB) aparece de forma natural en un amplio espectro de las matemáticas: soluciones de ecuaciones diferenciales, algoritmos iterativos, sistemas dinámicos, teoría de juegos, cadenas de Markov, teoría de números,... En el desarrollo teórico nos centraremos en los resultados más relevantes.

Este trabajo consta de dos partes bien diferenciadas. La primera tiene como objetivo un desarrollo de las principales propiedades teóricas de la LB. La segunda parte está dedicada a las aplicaciones prácticas de la ley.

Un aspecto relevante será explicar qué significa seguir la LB. Hablaremos de sucesiones que siguen la LB, al igual que de distribuciones que siguen la LB. Será por tanto necesario definir la ley para diferentes objetos matemáticos, después de esto tendrá sentido decir que una sucesión siga la LB o que una variable aleatoria lo haga.

La parte central del desarrollo teórico de la LB tiene como objetivo aclarar cuándo esperar que un conjunto de datos presente dígitos significativos distribuidos como en (1), para ello utilizaremos inicialmente los datos que Benford recoge en su artículo. Las secciones están ordenadas de manera constructiva, es decir, se empieza con lo más básico, el espacio medible en el que se va a trabajar; después, se presentan las propiedades de Invariancia por cambio de escala (IE) e Invariancia por cambio de base (IB) para distribuciones de probabilidad, que como veremos caracterizan a la distribución (1); finalmente, se generalizan estas nociones a las medidas de probabilidad aleatorias.

La forma de ordenar los contenidos, necesaria para una adecuada redacción en términos matemáticos, oscurece a veces el porqué de su inclusión. Para justificar las secciones expuestas es preferible empezar por el final. Los datos que Benford aporta en su artículo fueron recogidos de manera independiente, de numerosos campos. Esto sugiere que no provienen de una única distribución sino de varias, en otras palabras, estos datos son una muestra de una muestra de distribuciones, lo que llamaremos una muestra de una medida de probabilidad aleatoria (m.p.a). Sin embargo, no parece razonable pensar que todas las muestras de medidas de probabilidad aleatorias vayan a tener dígitos significativos que sigan la distribución (1), en tal caso, todas las distribuciones sobre la recta real la tendrían, sin más que considerar la m.p.a constante, igual a esa distribución; es decir, hay que pedir alguna restricción sobre las medidas de probabilidad aleatorias. Así surgen de manera natural las propiedades de IE e IB. La relación entre la distribución de dígitos significativos (1) y estas propiedades de invariancia por cambio de escala y base se probarán primero para distribuciones de probabilidad. Al trabajar con probabilidades, y con estas propiedades de invariancia sobre los dígitos significativos solamente, se tendrá que precisar primero el espacio muestral, para después seleccionar adecuadamente los eventos de los cuales tendrá sentido hablar de su probabilidad.

Otra razón por la que esperar que esta ley aparezca en numerosas tablas de datos es debido a que los procesos multiplicativos generan datos distribuidos acorde a la LB. Siendo más específicos, los productos de variables aleatorias independientes e igualmente distribuidas convergen en distribución hacia la LB. Este resultado se demostrará también en el desarrollo teórico. Además, se incluye una propiedad curiosa de la distribución logarítmica: si $X$ es una variable aleatoria que sigue la LB e $Y$ es otra variable aleatoria independiente de $X$, entonces $XY$ sigue la LB.

En la parte práctica el objetivo es emplear la LB como un método sencillo para examinar anomalías numéricas. Esta idea fue comentada por primera vez por Varian [17], quien sugirió utilizar la Ley de Benford para validar la razonabilidad de los datos: si en un conjunto de datos se espera que los dígitos significativos sigan la distribución logarítmica, por ejemplo, porque esos datos provienen de diversas distribuciones o son consecuencia de un proceso multiplicativo, y tras una inspección no siguen la LB, esto suscita sospechas de que ha habido manipulación.

Hoy en día las aplicaciones de la LB se multiplican. Es ampliamente conocido, siendo Nigrini [19] uno de los pioneros, que la Ley de Benford se puede utilizar como una herramienta de contabilidad forense, esto es, detectar fraude en situaciones que involucran datos financieros, tales como: cuentas corrientes, declaraciones de impuestos, mercados bursátiles, o en economía. Cuando los valores exactos y reales se ajustan a la LB, alterar esos valores por otros inventados típicamente resulta en una distribución de dígitos significativos que se desvía de la LB.

Si bien es cierto que la Ley de Benford no es un método infalible para detectar alteraciones intencionadas en los datos, la desviación de la ley sirve como pista para futuras investigaciones. Pongamos el caso de Grecia: en 2001 Grecia se une a la Euro-zona, obligándose a cumplir con las pautas establecidas en el tratado de Maastricht. Estas pautas protegen la estabilidad monetaria de toda la unión. La única forma que Grecia tuvo para cumplir con estas pautas fue, como más tarde fue descubierto, gracias al falseamiento de su déficit. De esto se extraen dos conclusiones, la primera, el falseamiento de los datos, en este caso en el terreno macroeconómico, no resulta descabellado y debe contemplarse con verdadera preocupación. La segunda, el desarrollo de técnicas estadísticas puede ser de gran ayuda para detectar este tipo de falseamientos.

Así como el uso de la LB en contabilidad es notorio gracias al trabajo de Nigrini, también es conocida a nivel mediático la implementación de la LB en la detección de fraude electoral gracias al trabajo de Mebane [16] y otros. A modo de referencia, en los artículos [21] y [12] se utiliza, junto con otras herramientas estadísticas, la LB para medir la verosimilitud de los resultados electorales del Referéndum revocatorio de Venezuela de 2004, en el cual hubo alegatos de fraude por parte de la oposición.

Además de la utilidad de la LB a nivel financiero, económico y electoral, orientada a la detección de fraudes, la LB ha encontrado y sigue encontrando utilidad en otras muchas áreas, por ejemplo, para validar modelos matemáticos sobre procesos físicos: si es sabido que las cantidades asociadas con estos procesos satisfacen la LB, entonces las simulaciones deberán hacerlo también. También se está investigando sobre las aplicaciones de la LB para detectar señales sobre ruido de fondo, por ejemplo, en series temporales.

Como modelo la LB es útil ya que de manera natural aparece en estos conjuntos de datos, equivalentemente, el proceso que genera estos datos sigue la LB; por ello, una desviación de esta distribución suscita sospechas, pero, ¿cómo inferir esta desviación? La manera estadística de proceder es mediante un test de ajuste, esto es, se supone que la distribución de dígitos significativos de los datos, sin manipulación, se ajusta a la LB y en caso de rechazar la hipótesis se concluye, con un nivel de confianza previamente fijado, que los datos han sido alterados. Este procedimiento será precisado en la parte práctica.

En conclusión, este trabajo busca primero dar una explicación matemática que justifique la aparición de la LB en tantas tablas de datos, para después utilizar la ley con el fin de detectar anomalías en tales conjuntos de datos. Hemos seleccionado varios conjuntos de datos que nos permitirán ilustrar a nivel práctico estos objetivos.

# Nomenclatura

$\mathbb{N}$ Conjunto de los números naturales. 1,2,...

$\mathbb{Z}$ Conjunto de los números enteros

$\mathbb{Q}$ Conjunto de los números racionales

$\mathbb{I}$ Conjunto de los números irracionales

$\mathbb{R}$ Conjunto de los números reales

$\mathbb{R}^+$ Conjunto de los números reales positivos

$\emptyset$ Conjunto vacío

$\mathrm{M}(x) : \mathbb{R}^+ \to [1, 10)$ Función mantisa (base 10)

$\mathrm{M}_b(x) : \mathbb{R}^+ \to [1, b)$ Función mantisa en base $b \in \mathbb{N}$

$x \mathrm{mod}\ 1$ Parte fraccionaria de $x$, $x \in \mathbb{R}^+$

$\lfloor x \rfloor$ Mayor entero no mayor que $x$

$\chi_A$, $I_A$ Función indicadora del conjunto $A$

$\sigma$ Sigma álgebra

$\mathscr{B}^+$ $\sigma$-álgebra de los conjuntos Borel en $\mathbb{R}^+$

$\mathscr{B}_{[1,b)}$ $\sigma$-álgebra de los conjuntos Borel en $[1, b)$

$\mathscr{M}$ $\sigma$-álgebra mantisa

$\mathscr{M}_b$ $\sigma$-álgebra mantisa en base $b$

$\Omega$ Espacio muestral

$(\Omega, \sigma)$ Espacio medible

$P$, $\mathbb{P}$ Probabilidades

$(\Omega, \sigma, P)$ Espacio probabilístico

$X : (\Omega, \sigma, P) \longrightarrow (\mathbb{R}, \mathscr{B})$ Variable aleatoria real

$\sigma(X)$ Mínima $\sigma$-álgebra que hace medible a $X$

$P_X$ Distribución de $X$ en $(\mathbb{R}, \mathscr{B})$

$\lambda_{0,1}$ Medida de Lebesgue en $(0,1)$

$\left(\widehat{P}(k)\right)_{k\in\mathbb{Z}}$ Coeficientes de Fourier de una probabilidad $P$

$\mathscr{P}(A)$ Conjunto de las partes del conjunto $A$

$\mathfrak{M}$ Conjunto de las probabilidades en $(\mathbb{R},\mathscr{B})$

$\stackrel{\mathrm{d}}{=}$ Igualdad en distribución

c.s Casi seguro

Prob Distribución de Benford

$D_n$ Dígito significativo n-ésimo (base 10)

$D_n^{(b)}$ Dígito significativo n-ésimo base $b \in \mathbb{N}$

$i$ Unidad imaginaria

$N(a,b)$ Distribución normal de media $a$ y varianza $b$

$U(0,1)$ $v.a$ uniforme en $(0,1)$

$v.a.$ Variable aleatoria

$v.a.i.i.d.$ Variables aleatorias independientes e igualmente distribuidas

# Índice general

# Índice de figuras

# Índice de tablas

# Capítulo 1

# Desarrollo teórico de la Ley de Benford

## 1.1. Forma general de la Ley de Benford

Comenzaremos definiendo de forma precisa en qué consiste la ley de probabilidad de Benford. En (1) se ha dado la versión original y más simple, que atañe solamente al primer dígito significativo $D_1$. Desde ahora, en este capítulo, cuando nos refiramos a la Ley de Benford estaremos hablando no solamente de la distribución del primer dígito significativo, sino de todos los dígitos significativos $D_1, D_2, \ldots, D_n, \ldots$.

Llamaremos forma general de la LB a

$$\operatorname{Prob}(D_1 = d_1, D_2 = d_2, \ldots, D_k = d_k) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}\right) \quad k \in \mathbb{N}. \tag{1.1}$$

Por ejemplo, $\operatorname{Prob}(D_1 = 1, D_2 = 2) = \log_{10} \frac{13}{12}$. Esta distribución de dígitos significativos lleva implícita la dependencia entre los dígitos significativos. Por ejemplo, la aparición del número dos en segunda posición afecta a la aparición del número uno en primera posición, veámoslo

$$\begin{aligned}\operatorname{Prob}(D_1 = 1 | D_2 = 2) = \frac{\operatorname{Prob}(D_1 = 1, D_2 = 2)}{\operatorname{Prob}(D_2 = 2)} = \frac{\log_{10}\frac{13}{12}}{\log_{10}(\frac{3}{2})} \approx 0,1974 \\ \neq \log_{10}(2) \approx 0,3 = \operatorname{Prob}(D_1 = 1).\end{aligned}$$

La forma general de la LB también se puede escribir de la forma

$$\operatorname{Prob}(\mathrm{M} \le t) = \log_{10}(t) \quad t \in [1, 10), \tag{1.2}$$

siendo $\mathrm{M}(x) : \mathbb{R}^+ \to [1, 10)$ la función que asigna a cada número real positivo $x$ su mantisa[1].

[1] La mantisa, que por ahora supondremos en base 10, de un número real positivo $x$, es el único número $r \in [1, 10)$ tal que $x = r \cdot 10^n$, para algún $n \in \mathbb{Z}$.

## 1.2. Marco probabilístico para la Ley de Benford

Un aspecto crucial es darle una interpretación rigurosa a Prob, y en consecuencia a (1.1) y a (1.2), construyendo un marco probabilístico adecuado. La forma más ingenua de proceder es establecer (1.1) solamente para $\mathbb{N}$, comenzando por ejemplo con el conjunto

$$(D_1 = 1) = \{1, 10, 11, 12, 13, 14, 15, ..., 100, 101, ..., 1000, ...\}.$$

Inmediatamente surgen dos cuestiones: ¿cómo se interpreta la frecuencia de un conjunto sobre $\mathbb{N}$?, y a continuación, ¿la frecuencia del conjunto $(D_1 = 1)$ es $\log_{10}(2)$? La manera razonable de interpretar la frecuencia de un conjunto sobre $\mathbb{N}$ es vía lo que se conoce como densidad natural[2]. Ahora, ¿es efectivamente la densidad natural de $(D_1 = 1)$ igual a $\log_{10}(2)$?, la respuesta es negativa, aún más, la densidad natural de este conjunto no existe. Así como la densidad natural de los números pares o impares se prueba fácilmente que es $\frac{1}{2}$, $(D_1 = 1)$ no tiene.

**Proposición 1.1.** *El conjunto $(D_1 = 1)$ no tiene densidad natural; es decir,*

$$\lim_{n\to\infty} \frac{1}{n}(D_1 = 1) \cap (1, 2, ..., n)$$

*no existe.*

DEMOSTRACIÓN. Consideramos la sucesión $(a_n)$, $a_n = \frac{1}{n}\#\{i \in \mathbb{N}/i \leq n \text{ , } D_1(i) = 1\}$. Veamos algunos términos de la sucesión:

$$\begin{array}{llllll} a_1 = 1 & a_2 = \dfrac{1}{2} & a_3 = \dfrac{1}{3} & \dots & a_9 = \dfrac{1}{9} & a_{10} = \dfrac{2}{10} \quad \dots \\ a_{19} = \dfrac{11}{19} & a_{20} = \dfrac{11}{20} & \dots & a_{99} = \dfrac{11}{99} & a_{100} = \dfrac{12}{100} & \dots . \end{array}$$

Consideramos la subsucesión

$$a_{n_k} = a_{10^k} = \frac{\left(\sum_{j=1}^{k-1} 10^j\right) + 2}{10^k}.$$

Veamos algunos términos de la subsucesión

$$a_{n_1} = \frac{2}{10} \quad a_{n_2} = \frac{12}{100} \quad a_{n_3} = \frac{112}{1000}.$$

Se tiene que

$$\lim_{k\to\infty} a_{n_k} = \lim_{k\to\infty} \frac{\sum_{j=1}^{k-1} 10^j}{10^k} + \lim_{k\to\infty} \frac{2}{10^k} = \lim_{k\to\infty} \frac{10 - 10^k}{-9 \cdot 10^k} = \frac{1}{9}.$$

---

[2] Sea $A \subseteq \mathbb{N}$. La densidad natural de $A$ se define por

$$d(A) = \lim_{n\to\infty} \frac{\#\{A \cap [1, n]\}}{n}.$$

Consideramos ahora otra subsucesión, la dada por

$$a_{n_k} = a_{2\cdot 10^k} = \frac{\sum_{j=0}^{k} 10^j}{2\cdot 10^k}.$$

Veamos algunos términos de la subsucesión

$$a_{n_1} = \frac{11}{20} \quad a_{n_2} = \frac{111}{200} \quad a_{n_3} = \frac{1111}{2000}.$$

Se tiene que

$$\lim_{k\to\infty} a_{n_k} = \lim_{k\to\infty} \frac{\frac{1-10^{k+1}}{-9}}{2\cdot 10^k} = \lim_{k\to\infty} \frac{10^{k+1}-1}{2\cdot 9\cdot 10^k} = \frac{5}{9};$$

por lo tanto, concluimos que

$$\nexists \lim_{n\to\infty} \frac{1}{n} \#\{i\in\mathbb{N} \quad / i\leq n \quad D_1(i)=1\},$$

ya que existen dos subsucesiones con límites diferentes. $\square$

Dado que la densidad natural de $(D_1 = 1)$ oscila entre $\frac{1}{9}$ y $\frac{5}{9}$, teóricamente es posible asignarle cualquier número del intervalo $\left[\frac{1}{9}, \frac{5}{9}\right]$ como probabilidad. Al fracasar en el intento de definir Prob sobre $\mathbb{N}$, se intentará definir sobre el espacio muestral $\Omega = \mathbb{R}^+$. En este caso

$$(D_1 = 1) = \bigcup_{n=-\infty}^{\infty} [1,2)\cdot 10^n.$$

El objetivo es poner la LB en una estructura de medibilidad adecuada. Esto se traduce en trabajar en una $\sigma$-álgebra adecuada. Al estar trabajando en $\mathbb{R}^+$, se puede pensar en considerar $\mathscr{B}^+$ como $\sigma$-álgebra de sucesos; no obstante, si los primeros dígitos obedecen alguna ley distribucional universal, esta ley deberá ser independiente de las unidades escogidas, por ejemplo, metros o pulgadas. Esta propiedad, la cual se desarrollará más adelante, es la de IE. La cuestión es que no hay medidas de probabilidad invariantes por escala en $\mathscr{B}^+$, ya que de ser así, por definición, la probabilidad de $(0,1)\in\mathscr{B}^+$ debería ser igual a la probabilidad de todo intervalo $(0,b)\in\mathscr{B}^+$, $b\in\mathbb{R}^+$; entonces, como consecuencia de las propiedades de continuidad desde abajo y desde arriba que son intrínsecas a cualquier probabilidad, se tendría que

$$\begin{aligned} &\operatorname{Prob}((0,1)) = \operatorname{Prob}((0,n)) \quad \forall n\in\mathbb{N} \\ \Longrightarrow\ &\operatorname{Prob}((0,1)) = \lim_{n\to\infty}\operatorname{Prob}((0,1)) = \lim_{n\to\infty}\operatorname{Prob}((0,n)) = \operatorname{Prob}\left(\mathbb{R}^+\right) = 1 \\ &= \lim_{n\to\infty}\operatorname{Prob}((0,1)) = \lim_{n\to\infty}\operatorname{Prob}\left(\left(0,\frac{1}{n}\right)\right) = \operatorname{Prob}(\emptyset) = 0, \end{aligned}$$

llegando al absurdo.

Como (1.1) involucra a las variables aleatorias $D_1, D_2, \ldots, D_n, \ldots$ lo más sensato es considerar la mínima $\sigma$-álgebra que las hace medibles, $\sigma(D_1, D_2, \ldots, D_n, \ldots)$. También podríamos haber considerado la mínima $\sigma$-álgebra que hace medible a M, $\sigma(\mathrm{M})$, ya que (1.2) involucra a la variable aleatoria M. Veamos que, como es de esperar,

$$\sigma(\mathrm{M}) = \sigma(D_1, D_2, \ldots, D_n, \ldots).$$

**Proposición 1.2.** $\sigma(\mathrm{M}) = \sigma(D_1, D_2, ..., D_n, ...)$.

DEMOSTRACIÓN.

Sea $S = \{0,1,2,3,4,5,6,7,8,9\}$. Consideremos las variables aleatorias

$$\mathrm{M} : \left(\mathbb{R}^+, \mathscr{B}^+\right) \longrightarrow \left([1,10), \mathscr{B}_{[1,10)}\right),$$
$$D_n : \left(\mathbb{R}^+, \mathscr{B}^+\right) \longrightarrow (S, \mathscr{P}(S)) \quad \forall n \in \mathbb{N}.$$

Para la inclusión $\sigma(\mathrm{M}) \subseteq \sigma(D_1, D_2, ..., D_n, ...)$ basta tener en cuenta que

$$\mathrm{M}(x) = \sum_{n=1}^{\infty} D_n(x) \cdot 10^{1-n} = g(\mathbf{D})(x) \ , x \in \mathbb{R},$$

siendo $g : S^\infty \longrightarrow [1,10)$, dada por $g\left((x_n)_{n=1}^{\infty}\right) = \sum_{n=1}^{\infty} x_n \cdot 10^{1-n}$ , $g$ medible $\Rightarrow \sigma(\mathrm{M}) \subseteq \sigma(\mathbf{D})$.

Para la inclusión $\sigma(\mathrm{M}) \supseteq \sigma(D_1, D_2, ..., D_n, ...)$ basta tener en cuenta que

$$D_n(x) = \lfloor 10^{m-1}\mathrm{M}(x) \rfloor - 10\lfloor 10^{m-2}\mathrm{M}(x) \rfloor = h(\mathrm{M})(x) \ , x \in \mathbb{R} \ , n \in \mathbb{N},$$

con $h : [1,10) \longrightarrow S$, dada por $h(a) = \lfloor 10^{m-1}a \rfloor - 10\lfloor 10^{m-2}a \rfloor$ , $h$ medible $\Rightarrow \sigma(\mathbf{D}) \subseteq \sigma(\mathrm{M})$.

En definitiva

$$\sigma(\mathrm{M}) = \sigma(\mathbf{D}).$$

□

**Definición 1.3.** *Llamaremos $\mathscr{M}$ a $\sigma(\mathrm{M}) = \sigma(D_1, D_2, ..., D_n, ...)$.*

**Lema 1.4.**

$$S \in \mathscr{M} \Longleftrightarrow S = \bigcup_{n=-\infty}^{\infty} B \cdot 10^n \quad B \subseteq [1,10) \text{ Borel.}$$

DEMOSTRACIÓN. Sea $S \in \mathscr{M}$. Entonces, $S = \mathrm{M}^{-1}(B)$ , $B \subseteq [1,10)$ Borel. Veamos que $\mathrm{M}^{-1}(B) = \bigcup_{n=-\infty}^{\infty} B \cdot 10^n$. Sea $a \in \mathrm{M}^{-1}(B) \Longrightarrow \mathrm{M}(a) \in B$.

Por definición de mantisa, $\exists! n_0 \in \mathbb{Z}$ tal que $a = \mathrm{M}(a) \cdot 10^{n_0}$. Por tanto, $a \in B \cdot 10^{n_0} \Longrightarrow a \in \bigcup_{n=-\infty}^{\infty} B \cdot 10^n$.

Sea $a \in \bigcup_{n=-\infty}^{\infty} B \cdot 10^n$. Entonces, $\exists n_0 \in \mathbb{Z}$ tal que $a \in B \cdot 10^{n_0}$ $\Longrightarrow a \cdot 10^{-n_0} \in B \subseteq [1,10)$. Necesariamente, $\mathrm{M}(a) = a \cdot 10^{-n_0}$, por definición de mantisa. En definitiva, $\mathrm{M}(a) \in B \Longrightarrow a \in \mathrm{M}^{-1}(B)$.

□

De aquí en adelante trabajaremos en el espacio medible $(\mathbb{R}^+, \mathscr{M})$. Recapitulando, nuestro propósito es definir Prob correctamente, para ello, nos hemos visto obligados a trabajar en el espacio muestral $\mathbb{R}^+$, en vez de $\mathbb{N}$. Además, por el propio problema con el que estamos tratando, como calcular probabilidades de sucesos del tipo $(D_1 = 1)$, ha sido necesario restringirnos a $\mathscr{M}$, en vez de considerar $\mathscr{B}^+$. Esto significa que no tiene sentido el cálculo de la probabilidad de sucesos como $[1,2) \in \mathscr{B}^+ \setminus \mathscr{M}$, ya que $1 \in [1,2)$ pero $10 \notin [1,10)$, sin embargo, $\mathrm{M}(1) = \mathrm{M}(10)$.

El siguiente teorema establece algunas propiedades básicas de $\mathscr{M}$, las cuales serán esenciales en el estudio de los aspectos característicos de la LB, tales como las propiedades de IE e IB.

**Teorema 1.5.** *$\mathscr{M}$ cumple las siguientes propiedades:*

**i)** *Cualquier conjunto no vacío $S \in \mathscr{M}$ no es acotado y, además, el $0$ es un punto de acumulación de $S$.*

**ii)** *$\mathscr{M}$ es cerrado para la multiplicación por escalares.*

**iii)** *$\mathscr{M}$ es cerrado por raíces enteras.*

**iv)** *$\mathscr{M}$ es auto-similar.*

DEMOSTRACIÓN.

**i)** *Cualquier conjunto no vacío $S \in \mathscr{M}$ no es acotado y, además, el $0$ es un punto de acumulación de $S$.*

Sea $S \in \mathscr{M}$, $S \neq \emptyset$.

$$S \in \mathscr{M} \Longrightarrow S = \bigcup_{n=-\infty}^{\infty} B \cdot 10^n.$$

para algún $B \subseteq [1,10)$ Borel.

Como $S \neq \emptyset \Longrightarrow B \neq \emptyset \Longrightarrow \exists a \in B$.

Consideramos la sucesión $\{a_n\}_{n=0}^{\infty}$, $a_n = a \cdot 10^n$. Se tiene que

$$\lim_{n\to\infty} a_n = +\infty$$

con $a_n \in S\ \forall n \in \mathbb{N} \cup \{0\}$. Entonces, $S$ no es acotado.

Consideramos la sucesión $\{b_n\}_{n=0}^{\infty}$, $b_n = a \cdot 10^{-n}$. Se tiene que

$$\lim_{n\to\infty} b_n = 0$$

con $b_n \in S\ \forall n \in \mathbb{N} \cup \{0\}$. Entonces, el 0 es un punto de acumulación de $S$.

**ii)** *$\mathscr{M}$ es cerrado para la multiplicación por escalares.*

Sea $S \in \mathscr{M}$, queremos ver que $a \cdot S \in \mathscr{M} \quad \forall a > 0$.

Fijamos $a > 0$.

$a \cdot S \in \mathscr{M} \Longleftrightarrow \exists \widetilde{B} \in \mathscr{B}_{[1,10)}$ tal que $a \cdot S = \mathrm{M}^{-1}(\widetilde{B})$.

Por hipótesis $\exists B \in \mathscr{B}_{[1,10)}$ tal que $S = \mathrm{M}^{-1}(B)$.

Veamos que $a \cdot S = \mathrm{M}^{-1}(\mathrm{M}(a \cdot B))$, donde $\mathrm{M}(C)$, $C$ conjunto contenido en $\mathbb{R}^+$, es $\{\mathrm{M}(a) : a \in C\}$. Lo primero es ver que $\mathrm{M}(C) \in \mathscr{B}_{[1,10)}\ \forall C \in \mathscr{B}^+$. Sea $C \in \mathscr{B}^+$. Consideramos la partición $\mathscr{P}$ de $\mathbb{R}^+$, $\mathscr{P} = \{[10^n, 10^{n+1}), n \in \mathbb{Z}\}$. Denotamos por $A_n$ al intervalo $[10^n, 10^{n+1})$. Tenemos entonces que

$$C = \bigcup_{-\infty}^{\infty} (C \cap A_n).$$

Veamos que
$$\mathrm{M}(C) = \bigcup_{n=-\infty}^{\infty} (C \cap A_n) \cdot 10^{-n}.$$

Sea $x \in \mathrm{M}(C)$. Entonces $\exists a \in C$ tal que $\mathrm{M}(a) = x$. Como $a \in C \Longrightarrow \exists n_0 \in \mathbb{Z}$ tal que $a \in A_{n_0} \cap C$. Como $a \in A_n \Longrightarrow a \in [10^{n_0}, 10^{n_0+1}) \Longrightarrow a \cdot 10^{-n_0} \in [1,10)$ y necesariamente $\mathrm{M}(a) = a \cdot 10^{-n_0} = x$. En definitiva $x \in (C \cap A_{n_0}) \cdot 10^{-n_0}$. Y se tiene que
$$\mathrm{M}(C) \subseteq \bigcup_{n=-\infty}^{\infty} (C \cap A_n) \cdot 10^{-n}.$$

Veamos la otra inclusión. Sea $x \in \bigcup_{n=-\infty}^{\infty} (C \cap A_n) \cdot 10^{-n} \Longrightarrow \exists n_1 \in \mathbb{Z}$ tal que $x = h \cdot 10^{-n_1}$, con $h \in (C \cap A_{n_1})$. Entonces, $\mathrm{M}(h) = x$, y por tanto
$$\mathrm{M}(C) = \bigcup_{n=-\infty}^{\infty} (C \cap A_n) \cdot 10^{-n}.$$

Como el escalado de conjuntos Borel es Borel, y tanto la intersección numerable, en este caso finita, como la unión numerable de conjuntos Borel es Borel, concluimos que
$$\mathrm{M}(C) \in \mathscr{B}_{[1,10)} \quad \forall C \in \mathscr{B}^+.$$

Por tanto $\mathrm{M}^{-1}(\mathrm{M}(a \cdot B))$ está bien definido.

Sea $x \in a \cdot S$, entonces $\exists s \in S$ tal que $x = a \cdot s$. Como $s \in S \Longrightarrow M(s) \in B \Longrightarrow \exists n_0 \in \mathbb{Z}$ tal que $s \cdot 10^{n_0} \in B$.
$$\begin{aligned} x \in \mathrm{M}^{-1}(\mathrm{M}(a \cdot B)) &\Longleftrightarrow \mathrm{M}(x) \in \mathrm{M}(a \cdot B) \\ &\Longleftrightarrow \exists n \in \mathbb{Z} \quad / x \cdot 10^n \in a \cdot B \end{aligned}$$

Por tanto,
$$a \cdot S \subseteq \mathrm{M}^{-1}(\mathrm{M}(a \cdot B)).$$

Sea $x \in \mathrm{M}^{-1}(\mathrm{M}(a \cdot B))$. Entonces, $\exists n_1 \in \mathbb{Z}$ tal que $x \cdot 10^{n_1} \in a \cdot B$. Se tiene que
$$x \cdot 10^{n_1} \in a \cdot B \Longrightarrow \frac{x}{a} \cdot 10^{n_1} \in B \Longrightarrow \frac{x}{a} \in S \Longrightarrow x \in a \cdot S.$$

Concluimos que
$$a \cdot S = \mathrm{M}^{-1}(\mathrm{M}(a \cdot B)) \Longrightarrow a \cdot S \in \mathrm{M} \quad \forall a > 0.$$

**iii)** *$\mathscr{M}$ es cerrado por raíces enteras.*

Sea $S \in \mathscr{M}$, queremos ver que $S^{\frac{1}{m}} \in \mathscr{M}$.

$S = M^{-1}(B)$, $B \in \mathscr{B}_{[1,10)}$.
$$S^{\frac{1}{m}} \in \mathscr{M} \Longleftrightarrow \exists \widetilde{B} \in \mathscr{B}_{[1,10)} \quad / S^{\frac{1}{m}} = \mathrm{M}^{-1}(\widetilde{B})$$

Veamos que $\widetilde{B} = \bigcup_{j=0}^{m-1}(B^{\frac{1}{m}} \cdot 10^{\frac{j}{m}})$ sirve.

Lo primero es ver que $\widetilde{B}$ definido así está en $\mathscr{B}_{[1,10)}$.

Sea $x \in \widetilde{B} \Longrightarrow \exists k \quad ,0 \leq k \leq m-1, \quad /x \in B^{\frac{1}{m}} \cdot 10^{\frac{k}{m}} \Longrightarrow x = 10^{\frac{k}{m}} \cdot b \quad ,b \in B^{\frac{1}{m}}$. Ahora, $b^m \in B \subseteq [1,10) \Longrightarrow b \in [1, 10^{\frac{1}{m}}) \Longrightarrow x \in [10^{\frac{k}{m}}, 10^{\frac{k+1}{m}}) \subseteq [1,10)$.

Por tanto,

$$\widetilde{B} \subseteq [1,10).$$

Veamos que es un conjunto Borel.

Sea $t_m : \mathbb{R}^+ \to \mathbb{R}^+$, dada por $t_m(x) = x^m \quad ,x \in \mathbb{R}, m \in \mathbb{N}$.

$t_m$ es continua $\forall m \in \mathbb{N}$, por tanto $t_m$ es $\mathscr{B}^+|\mathscr{B}^+$ medible. Por lo cual, $B^{\frac{1}{m}} = t_m^{-1}(B) \in \mathscr{B}^+ \quad \forall B \in \mathscr{B}^+$. En definitiva

$$\widetilde{B} \in \mathscr{B}_{[1,10)}.$$

Sea $a \in S^{\frac{1}{m}} \Longrightarrow a^m \in S \Longrightarrow \exists b_0 \in B, k_1 \in \mathbb{Z} \quad /a^m = b_0 \cdot 10^{k_1} \Longrightarrow a = b_0^{\frac{1}{m}} \cdot 10^{\frac{k_1}{m}}$.
Por el algoritmo de división euclídea, $\exists q, l \in \mathbb{Z} \quad /k_1 = mq + l$, con $l \in \{0, 1, ..., m-1\}$. Entonces

$$a = b_0^{\frac{1}{m}} \cdot 10^{\frac{mq+l}{m}} \Longleftrightarrow a = b_0^{\frac{1}{m}} \cdot 10^q \cdot 10^{\frac{l}{m}} \Longrightarrow \mathrm{M}(a) = b_0^{\frac{1}{m}} \cdot 10^{\frac{l}{m}}.$$

Tenemos que

$$a \in \mathrm{M}^{-1}\left(\bigcup_{j=0}^{m-1} B^{\frac{1}{m}} \cdot 10^{\frac{j}{m}}\right) \Longleftrightarrow \mathrm{M}(a) \in \left(\bigcup_{j=0}^{m-1} B^{\frac{1}{m}} \cdot 10^{\frac{j}{m}}\right).$$

Como $\mathrm{M}(a) \in B^{\frac{1}{m}} \cdot 10^{\frac{l}{m}}$ concluimos que

$$S^{\frac{1}{m}} \subseteq \bigcup_{j=0}^{m-1} B^{\frac{1}{m}} \cdot 10^{\frac{j}{m}}.$$

Veamos la otra inclusión. Sea $a \in \mathrm{M}^{-1}\left(\bigcup_{j=0}^{m-1} B^{\frac{1}{m}} \cdot 10^{\frac{j}{m}}\right) \Longrightarrow \exists l \in \{0, 1, ..., m-1\}$ tal que $M(a) \in B^{\frac{1}{m}} \cdot 10^{\frac{l}{m}} \Longrightarrow a \in B^{\frac{1}{m}} \cdot 10^{\frac{l}{m}} \cdot 10^q$, para algún $q \in \mathbb{Z}$. Entonces $a^m \in B \cdot 10^l \cdot 10^{qm} \Longrightarrow a^m \in S \Longrightarrow a \in S^{\frac{1}{m}}$. En definitiva

$$S^{\frac{1}{m}} = \mathrm{M}^{-1}\left(\bigcup_{j=0}^{m-1} B^{\frac{1}{m}} \cdot 10^{\frac{j}{m}}\right).$$

**iv)** *$\mathscr{M}$ es auto-similar.*

Sea $S \in \mathscr{M}$. Queremos ver que $10^m S = S \quad \forall m \in \mathbb{Z}$.

$$S \in \mathscr{M} \Longleftrightarrow S = \bigcup_{n=-\infty}^{\infty} B \cdot 10^n.$$

con $B \in \mathscr{B}_{[1,10)}$. Entonces

$$10^m S = \bigcup_{n=-\infty}^{\infty} B \cdot 10^{n+m} = \bigcup_{k=-\infty}^{\infty} B \cdot 10^k = S.$$

□

Con este Teorema finalizamos la descripción de $\mathscr{M}$.

# 1.3. Propiedad de Benford

Necesitamos especificar exactamente qué significa seguir la LB para diferentes objetos matemáticos si queremos dar una versión formal de (1.1) y a (1.2). Los objetos de interés serán: distribuciones de probabilidad, variables aleatorias positivas y sucesiones. Recordemos que el objetivo central del desarrollo teórico es explicar la aparición de la LB en numerosas tablas de datos, así pues, es natural considerar el estudio sobre variables aleatorias y distribuciones de probabilidad; no obstante, también es importante tener clara la noción de seguir la LB para sucesiones, ya que trabajaremos con medidas de probabilidad aleatorias.

## 1.3.1. Distribuciones de probabilidad

**Definición 1.6.** *Una distribución de probabilidad $P$ en $(\mathbb{R}^+, \mathscr{B}^+)$ sigue la LB si $P_{\mathrm{M}}$ sigue la distribución logarítmica en $\big([1,10), \mathscr{B}_{[1,10)}\big)$; es decir, si la variable mantisa,* M*, induce a partir de la $P$ la distribución logarítmica*

$$P_{\mathrm{M}}([1,t)) = \log_{10} t \quad \forall t \in [1,10). \tag{1.3}$$

**Ejemplo 1.7.** *Para todo $k \in \mathbb{Z}$, la distribución $P_k$, con densidad $f_k(x) = \frac{1}{x\ln(10)}$ en $[10^k, 10^{k+1})$ sigue la LB.*

DEMOSTRACIÓN.

$$\begin{aligned}
P_k(\mathrm{M} \le t) &= P_k\left(\bigcup_{n\in\mathbb{Z}} [10^n, t10^n)\right) \\
&= \sum_{n\in\mathbb{Z}} P_k[10^n, t10^n) = P_k[10^k, t10^k) \\
&= \int_{10^k}^{t10^k} f_k(x)dx = \int_{10^k}^{t10^k} \frac{1}{x\ln(10)}dx = \frac{\ln(t)}{\ln(10)} = \log_{10}(t).
\end{aligned}$$

□

## 1.3.2. Variables aleatorias

**Definición 1.8.** *Una variable aleatoria positiva y real $X$ sigue la LB si $P_X$ es Benford; es decir, si $P_{\mathrm{M}(X)}$ sigue la distribución logarítmica en $\big([1,10), \mathscr{B}_{[1,10)}\big)$, o equivalentemente si*

$$(P_X)_{\log_{10}(\mathrm{M})} \stackrel{\mathrm{d}}{=} U(0,1).$$

En otras palabras, una variable aleatoria $X$ es Benford si la distribución de sus dígitos significativos sigue la LB. Esto quiere decir que

$$\begin{aligned} &X \text{ Benford} \\ &\quad \Longrightarrow P(D_1(X)=1)=\log_{10}2 \\ &\dots \\ &\quad \Longrightarrow P(D_1(X)=9)=\log_{10}\frac{10}{9}. \end{aligned}$$

**Ejemplo 1.9.** *Consideremos la variable aleatoria $X = 10^U$, donde $U \stackrel{\mathrm{d}}{=} U(0,1)$, $X$ es Benford.*

DEMOSTRACIÓN.

$$\begin{aligned} (P_X)_{\log_{10}(\mathrm{M})} \stackrel{\mathrm{d}}{=} U(0,1) &\Longleftrightarrow P_{\log_{10}(\mathrm{M}(X))} \stackrel{\mathrm{d}}{=} U(0,1) \\ &\Longleftrightarrow P_{\log_{10}(\mathrm{M}(10^U))} \stackrel{\mathrm{d}}{=} U(0,1) \\ &\Longleftrightarrow P_U \stackrel{\mathrm{d}}{=} U(0,1). \end{aligned}$$

□

**Ejemplo 1.10.** *Sea $X \stackrel{\mathrm{d}}{=} U(0,1)$, $X$ no es Benford.*

DEMOSTRACIÓN.

$$\begin{aligned} P(\mathrm{M}(X)\leq t) &= \lambda_{0,1}\left(\bigcup_{k\in\mathbb{Z}} 10^k[1,t]\right) = \sum_{n=1}^{\infty}(1-t)\cdot 10^{-n} \\ &= \frac{t-1}{9} \neq \log_{10}t \qquad \text{para por ejemplo } t=2. \end{aligned}$$

□

Como era de esperar, los dígitos significativos de una variable uniforme en $(0,1)$ serán uniformes en $[1,10)$.

### 1.3.3. Sucesiones

**Definición 1.11.** *Una sucesión de números reales positivos $(x_n)$ sigue la LB si*

$$\lim_{N\to\infty} \frac{\#\{1\leq n\leq N : \mathrm{M}(x_n)\leq t\}}{N} = \log_{10}(t) \quad t\in[1,10). \tag{1.4}$$

**Definición 1.12.** *Una sucesión de variables aleatorias reales positivas $(X_n)$ sigue la LB si*

$$\lim_{N\to\infty} \frac{\#\{1\leq n\leq N : \mathrm{M}(X_n)\leq t\}}{N} \stackrel{\mathrm{c.s}}{=} \log_{10}(t) \quad t\in[1,10). \tag{1.5}$$

En la siguiente sección se darán ejemplos de sucesiones que siguen la LB y de sucesiones que no, para ello, será necesario recurrir a la teoría de uniformidad módulo 1. Los resultados de la siguiente sección se harán para sucesiones de números reales positivos por comodidad, siendo fácilmente transferibles a variables aleatorias reales y positivas.

## 1.4. Teoría de la distribución uniforme módulo uno

La teoría de la distribución uniforme módulo uno se ocupa de la distribución de las partes fraccionarias de los números reales en el intervalo $[0,1)$. El **Teorema 1.18** muestra cómo podemos aplicar esta teoría matemática, ampliamente desarrollada, para deducir propiedades de la LB. En primer lugar se ha de especificar qué significa estar distribuido uniformemente módulo uno, para los objetos matemáticos con los que venimos trabajando.

**Definición 1.13.** *Sea $x$ un número real. Diremos que* $x \bmod 1$ *es su parte fraccionaria; es decir,* $x \bmod 1 = x - \lfloor x \rfloor$.

**Definición 1.14.** *Una distribución de probabilidad $P$ en $(\mathbb{R},\mathscr{B})$ es $u.d.$* mod 1 *si*

$$P(\{x\colon x \bmod 1 \leq s\}) = P\left(\bigcup_{k\in\mathbb{Z}}[k,k+s]\right) = s \quad \forall s \in [0,1).$$

**Definición 1.15.** *Una variable aleatoria $X$ en un espacio probabilístico $(\Omega,\sigma,P)$ es $u.d.$* mod 1 *si*

$$P(X \bmod 1 \leq s) = s \quad \forall s \in [0,1). \tag{1.6}$$

*Es decir, si* $X \bmod 1 \overset{\mathrm{d}}{=} U(0,1)$.

**Definición 1.16.** *Una sucesión $(x_n)$ de números reales está uniformemente distribuida módulo 1, abreviado como $u.d.$* mod 1, *si*

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : x_n \bmod 1 \leq s\}}{N} = s \quad \forall s \in [0,1). \tag{1.7}$$

El siguiente Lema relaciona la mantisa de un número real positivo con su parte fraccionaria. Utilizaremos este Lema para demostrar el **Teorema 1.18**.

**Lema 1.17.** $\mathrm{M}(x) = 10^{\log_{10}(x) \bmod 1} \quad x \in \mathbb{R}^+$.

DEMOSTRACIÓN. Sea $x \in \mathbb{R}^+$. Sea $m \in \mathbb{Z}$ tal que $\log_{10}(x) + m \in [0,1)$.

$$\begin{aligned}&\log_{10}(x) + m \in [0,1) \Longrightarrow \log_{10}(x) \in [-m, 1-m)\\ &\Longrightarrow x \in \left[10^{-m}, 10^{1-m}\right).\end{aligned}$$

Como $\mathrm{M}(x) = x \cdot 10^l$, con $l \in \mathbb{Z}$, siendo $l$ único, $\mathrm{M}(x) \in [1,10)$, necesariamente $l = m$, y por tanto

$$\begin{aligned}10^{\log_{10}(x) \bmod 1} = 10^{\log_{10}(x)+m} &= x \cdot 10^m\\ &= \mathrm{M}(x).\end{aligned}$$

□

El siguiente Teorema, aunque sencillo, es una de las principales herramientas en la teoría de LB, ya que permite la aplicación de la teoría de uniformidad mod 1 a la Ley de Benford.

**Teorema 1.18.** *Una variable aleatoria positiva* c.s *(respectivamente, una sucesión de números reales positivos* $(x_n)$*, una distribución de probabilidad en* $(\mathbb{R}^+, \mathscr{B}^+)$*) sigue la* LB *si, y solo si,* $\log_{10}(X)$ *es* $u.d.$ mod 1 *(respectivamente,* $(\log_{10}(x_n))$*,* $P_{\log_{10}}$ *en* $(\mathbb{R}, \mathscr{B})$*)).*

DEMOSTRACIÓN.

- Caso variable aleatoria:

  Supongamos que $X$ sigue la LB. Entonces

$$\begin{aligned} P\left(\log_{10}(X)\text{mod } 1 \leq s\right) &= P\left(10^{\log_{10}(X)\text{mod } 1} \leq 10^s\right) \\ &= P\left(\mathrm{M}(x) \leq 10^s\right) = s. \end{aligned}$$

  Si $\log_{10}(X)$ es $u.d.$ mod 1

$$\begin{aligned} P\left(\mathrm{M}(x) \leq s\right) &= P\left(10^{\log_{10}(X)\text{mod } 1} \leq s\right) \\ &= P\left(\log_{10}(X)\text{mod } 1 \leq \log_{10}(s)\right) = \log_{10}(s). \end{aligned}$$

- Caso sucesión:

  Supongamos que $(x_n)$ sigue la LB. Entonces

$$\begin{aligned} &\lim_{N\to\infty} \frac{1}{N}\#\{1 \leq n \leq N\colon \log_{10}(x_n)\text{mod } 1 \leq s\} \\ &= \lim_{N\to\infty} \frac{1}{N}\#\{1 \leq n \leq N\colon 10^{[\log_{10}(x_n)\text{mod } 1]} \leq 10^s\} \\ &= \lim_{N\to\infty} \frac{1}{N}\#\{1 \leq n \leq N\colon \mathrm{M}(x_n) \leq 10^s\} = s. \end{aligned}$$

  Análogo el recíproco.

- Caso probabilidad:

  Supongamos que $P$ sigue la LB. Entonces

$$\begin{aligned} &P_{\log 10}\left(x \in \mathbb{R} : x\text{mod } 1 \leq s\right) = P\left(x \in \mathbb{R}^+ : \log_{10}(x)\text{mod } 1 \leq s\right) \\ &= P\left(x \in \mathbb{R}^+ : 10^{[\log_{10}(x)\text{mod } 1]} \leq 10^s\right) = P\left(x \in \mathbb{R}^+ : \mathrm{M}(x) \leq 10^s\right) \\ &= P_M\left([1, 10^s)\right) = s. \end{aligned}$$

  Análogo el recíproco.

□

Tras las caracterizaciones que nos permitirán aplicar los resultados relativos a la teoría de la uniformidad módulo uno al estudio sobre la LB, veremos qué resultados son los que pretendemos aplicar. Nos centraremos en los resultados ligados a sucesiones y a variables aleatorias.

### 1.4.1. Sucesiones

El fin de esta sección es demostrar el criterio de Weyl. La demostración se hará utilizando las propiedades de los coeficientes de Fourier presentadas y demostradas en el **Apéndice A**. Puesto que es la primera vez que se mencionan los coeficientes de Fourier en el trabajo, subrayar aquí que se utilizarán en gran parte de las demostraciones venideras.

**Teorema 1.19** (Criterio de Weyl)**.** *Una sucesión de números reales* $(x_n)$ *es* $u.d.$ mod 1 *si, y solo si,*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h x_n} = 0 \quad \forall h \in \mathbb{N}. \tag{1.8}$$

DEMOSTRACIÓN.
Sea $(x_n)$ una sucesión de números reales. Definamos para cada $N \in \mathbb{N}$ una probabilidad $\varphi_N$ en$\left([0,1)\,,\mathscr{B}_{[0,1)}\right)$ de la siguiente forma

$$\varphi_N(A) = \frac{\#\{1 \le n \le N : x_n \text{mod } 1 \in A\}}{N} \quad A \in \mathscr{B}_{[0,1)}.$$

Es claro que $(x_n)$ es $u.d.$ mod 1 si, y solo si, $(\varphi_N)$ converge en distribución hacia $\lambda_{0,1}$, ya que la definición de uniformidad módulo 1 para sucesiones se puede reescribir

$$\begin{aligned} &\lim_{N\to\infty} \frac{\#\{1 \le n \le N : x_n \text{mod } 1 \le s\}}{N} = s \\ \Longleftrightarrow\ &\lim_{N\to\infty} \varphi_N([0,s]) = \lambda_{0,1}([0,s]) \quad s \in [0,1). \end{aligned}$$

Por (II) en el Teorema A.3, $(\varphi_N)$ converge en distribución hacia $\lambda_{0,1}$ si, y solo si, $\left(\widehat{\varphi_N}(k)\right)$ converge hacia $\widehat{\lambda_{0,1}}(k)$ para todo $k \in \mathbb{Z}$. Por el **Ejemplo A.2** se sabe que $\widehat{\lambda_{0,1}}(0) = 1$ y $\widehat{\lambda_{0,1}}(h) = 0$ si $h \in \mathbb{Z} \setminus \{0\}$. Entonces $(x_n)$ es $u.d.$ mod 1 si, y solo si, se cumple que

$$\lim_{N\to\infty} \int_0^{\to 1} e^{-2\pi i h s} d\varphi_N(s) = 0 \quad \forall h \ne 0$$

$$\Longleftrightarrow \lim_{N\to\infty} \frac{1}{N}\sum_{p=1}^{N} e^{-2\pi i (x_n \text{mod } 1)} = 0 \quad \forall h \ne 0 \tag{1.9}$$

$$\Longleftrightarrow \lim_{N\to\infty} \frac{1}{N}\sum_{p=1}^{N} e^{-2\pi i x_n} = 0 \quad \forall h \ne 0. \tag{1.10}$$

(1.9) es consecuencia de que, por definición, $\varphi_N$ es la distribución uniforme discreta en los primeros $N$ números de la sucesión $(x_n \text{mod } 1)$. (1.10) es por la 1-periodicidad de la función $e^{-2\pi i h s}$.

□

**Corolario 1.20.** *La sucesión* $(n\theta)$ *es* $u.d.$ mod 1 *si, y solo si,* $\theta \in \mathbb{I}$.

DEMOSTRACIÓN. Sea $\theta \in \mathbb{I}$.

$$\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h n\theta}\right| = \frac{|e^{2\pi i h N\theta}-1|}{N|e^{2\pi i h\theta}-1|} \leq \frac{1}{N|\operatorname{sen}(\pi h\theta)|} \quad \forall h \in \mathbb{N}.$$

El último término está bien definido ya que, al ser $\theta$ irracional, $\operatorname{sen}(\pi h\theta) \neq 0$, $\forall h \in \mathbb{N}$.

En definitiva, aplicando el **Teorema 1.19**, concluimos que $(n\theta)$ es $u.d.$ mod 1.

Suponemos ahora $\theta \in \mathbb{Q} \Longrightarrow \exists p,q \in \mathbb{N}$ con $mcd(p,q) = 1$ tales que $\theta = \frac{p}{q}$. Con la notación utilizada, consideramos $h = q$. Entonces

$$\begin{aligned} e^{2\pi i h n\theta} &= \cos(2\pi h n\theta) + i \operatorname{sen}(2\pi h n\theta) \\ &= \cos(2\pi p n) + i \operatorname{sen}(2\pi p n) \\ &= 1 \quad \forall n \in \mathbb{N}, \end{aligned}$$

y en consecuencia

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h x_n} = 1 \qquad \text{con } h = q.$$

Aplicando el **Teorema 1.19**, concluimos que $(n\theta)$ no es $u.d.$ mod 1.

□

**Corolario 1.21.** *La sucesión* $(\alpha^n)$ *sigue la LB si, y solo si,* $\log_{10}(\alpha) \in \mathbb{I}$.

DEMOSTRACIÓN. Inmediata aplicando el **Teorema 1.18** y el **Corolario 1.20**.

Supongamos que $(\alpha^n)$ sigue la LB. Entonces: $(\log_{10}(\alpha^n)) = (n\log_{10}(\alpha))$ es $u.d.$ mod 1. Por tanto: $\log_{10}(\alpha) \in \mathbb{I}$.

Análogamente, si $\log_{10}(\alpha) \in \mathbb{I} \Longrightarrow (n\log_{10}(\alpha)) = (\log_{10}(\alpha^n))$ es $u.d.$ mod 1, lo cual implica que $(\alpha^n)$ sigue la LB.

□

**Ejemplo 1.22.** *La sucesión* $(2^n)$ *sigue la LB.*

DEMOSTRACIÓN. Por el **Corolario 1.21**, es suficiente con probar que $\log_{10}(2) \in \mathbb{I}$.

Supongamos que $\log_{10}(2) \in \mathbb{Q} \Longrightarrow \exists p,q \in \mathbb{N}$ con $mcd(p,q) = 1$ tales que $\log_{10}(2) = \frac{p}{q} \Longrightarrow 2^q = 10^p \Longrightarrow 2^q \equiv 0 \pmod{5}$. Absurdo.

□

**Ejemplo 1.23.** *La sucesión* $(0{,}01^n)$ *no sigue la LB.*

DEMOSTRACIÓN.

Consecuencia del **Corolario 1.21**, ya que $\log_{10}(0{,}01) = -2 \in \mathbb{Q}$.

□

### 1.4.2. Variables aleatorias

En esta sección nos centraremos en los resultados en relación a variables aleatorias. En el **Corolario 1.26** se demuestra que productos suficientemente grandes de variables aleatorias independientes e igualmente distribuidas siguen aproximadamente la LB; esto es, los procesos multiplicativos generan datos acorde a la LB. Además, en el **Corolario 1.28** se demuestra una propiedad curiosa de la Ley de Benford: si $X$ tiene la distribución logarítmica e $Y$ es una *v.a*, con cualquier distribución, independiente de $X$, entonces el producto tiene la distribución logarítmica.

**Definición 1.24.** *Sea $X$ una v.a. Diremos que $X$ no es puramente átomica si $P(X \in C) < 1$ para todo $C \subset \mathbb{R}$ numerable.*

**Teorema 1.25.** *Si $(X_n)$ es una sucesión de variables aleatorias i.i.d y $X_1$ no es puramente atómica, entonces*

$$\lim_{n\to\infty} P\left(\left(\sum_{j=1}^{n} X_j\right) \mod 1 \leq s\right) = s \quad \forall s \in [0,1).$$

*Es decir,* $\left(\left(\sum_{j=1}^{n} X_j\right) \mod 1\right)_n$ *converge en distribución hacia una $U(0,1)$.*

DEMOSTRACIÓN. Por (**II**) en el Teorema A.3, basta con ver que para todo $k \neq 0$

$$\begin{aligned} &\lim_{n\to\infty} \widehat{P_{(\sum_{j=1}^{n} X_j) \bmod 1}}(k) = 0 \\ \Longleftrightarrow &\lim_{n\to\infty} \left(\widehat{P_{X_1 \bmod 1}}(k)\right)^n = 0. \end{aligned} \tag{1.11}$$

(1.11) es consecuencia de (**II**) en el Teorema A.3 y de que las variables aleatorias están igualmente distribuidas.

Como $\left|\widehat{P_{X_1 \bmod 1}}(k)\right| \leq 1$, distingamos dos casos:

- $\left|\widehat{P_{X_1 \bmod 1}}(k)\right| < 1 \ \forall k \neq 0$. Entonces

$$\lim_{n\to\infty} \left(\widehat{P_{X_1 \bmod 1}}(k)\right)^n = 0.$$

- $\exists k_0 \neq 0$ tal que $\left|\widehat{P_{X_1 \bmod 1}}(k_0)\right| = 1$. Denotamos por $\Phi_k(s) = e^{-2\pi i k s}$. Entonces

$$\left|\int_0^{\to 1} \Phi_{k_0}(s)\, dP_{X_1 \bmod 1}(s)\right| = 1 \Longleftrightarrow \left|\int_0^{\to 1} \Phi_{k_0}(X_1 \bmod 1(\omega))\, dP(\omega)\right| = 1 \tag{1.12}$$

$$\Longleftrightarrow \int_0^{\to 1} \Phi_{k_0}(X_1 \bmod 1(\omega))\, dP(\omega) = \pm 1$$

$$\Longleftrightarrow \int_0^{\to 1} \left(\Phi_{k_0}(X_1 \bmod 1(\omega)) \mp 1\right) dP(\omega) = 0$$

$$\Longrightarrow \Phi_{k_0}(X_1 \bmod 1(\omega)) = \pm 1 \Longleftrightarrow e^{-2\pi i k_0 (X_1 \bmod 1)} \overset{\text{c.s}}{=} \pm 1. \tag{1.13}$$

En (1.12) se ha aplicado el teorema del cambio de variable y en (1.13) el teorema de anulación, dado que $\left|e^{2\pi i k_0 (X_1 \text{mod } 1)}\right| \leq 1$ implica que $e^{2\pi i k_0 (X_1 \text{mod } 1)} \mp 1$ tiene signo constante.

Ahora

$$
\begin{aligned}
& e^{2\pi i k_0 (X_1 \text{mod } 1)} \overset{\text{c.s}}{=} \pm 1 \\
& \Longleftrightarrow \cos\left(2\pi k_0 \left(X_1 \text{mod } 1\right)\right) + i\,\text{sen}(2\pi k_0 \left(X_1 \text{mod } 1\right)) \overset{\text{c.s}}{=} \pm 1 \\
& \Longrightarrow 2\pi k_0 \left(X_1 \text{mod } 1\right) \overset{\text{c.s}}{=} l\pi \qquad \text{para un } l \in \mathbb{Z} \text{ fijo} \\
& \Longrightarrow \left(X_1 \text{mod } 1\right) \overset{\text{c.s}}{=} \frac{l}{2k_0} \\
& \Longrightarrow P\left(\left(X_1 \text{mod } 1\right) \in \left\{\frac{l}{2k_0}\right\}\right) = 1 \\
& \Longrightarrow P\left(X_1 \in \left\{\frac{l}{2k_0} + b : b \in \mathbb{Z}\right\}\right) = 1.
\end{aligned}
$$

Absurdo ya que $\left\{\frac{l}{2k_0} + b : b \in \mathbb{Z}\right\}$ es numerable y $X_1$ no es puramente atómica. En definitiva, no se da este caso.

Concluimos que

$$
\begin{aligned}
& \lim_{n\to\infty} \widehat{P_{\left(\sum_{j=1}^{n} X_j\right) \text{ mod } 1}}(k) = 0 \quad \forall k \neq 0 \\
& \Longrightarrow \left(\left(\sum_{j=1}^{n} X_j\right) \text{ mod } 1\right)_n \text{ converge en distribución hacia una } U\left(0,1\right).
\end{aligned}
$$

□

**Corolario 1.26.** *Sea $(X_n)$ una sucesión de v.a.i.i.d positivas que no son puramente atómicas. Entonces, $\left(\prod_{j=1}^{n} X_j\right)_n$ converge en distribución hacia LB.*

DEMOSTRACIÓN. La sucesión $(\log_{10} X_n)$ es de v.a.i.i.d que no son puramente atómicas, ya que $\log_{10}$ es biyectiva en $\mathbb{R}^+$. Por tanto, $\left(\left(\sum_{j=1}^{n} \log_{10} X_n\right) \text{ mod } 1\right)_n$ converge en distribución hacia una $U(0,1) \Longrightarrow \left(\left(\log_{10}\left(\prod_{j=1}^{n} X_j\right)\right) \text{ mod } 1\right)_n$ converge en distribución hacia una $U(0,1) \Longrightarrow \left(\prod_{j=1}^{n} X_j\right)_n$ converge en distribución hacia LB.

□

**Teorema 1.27.** *Si $X$ es $u.d.$ mod 1 e $Y$ es independiente de $X$, entonces $X+Y$ es $u.d.$ mod 1.*

DEMOSTRACIÓN.

$X+Y$ es $u.d.$ mod 1 si, y sólo si, $(X+Y)\text{mod } 1 \overset{\text{d}}{=} U(0,1)$. Utilizando la tercera propiedad de los coeficientes de Fourier, $\widehat{P_{(X+Y)\text{mod } 1}}(k) = \widehat{P_{X\text{mod } 1}}(k) \cdot \widehat{P_{Y\text{mod } 1}}(k) = 0$ si $k \neq 0$

y si $k = 0$ $\widehat{P_{(X+Y)\text{mod } 1}}(k)(0) = 1$. En definitiva $(X+Y)\text{mod } 1 \stackrel{d}{=} U(0,1)$ y por tanto $X+Y$ es $u.d.$ mod 1.

□

Por los **Teoremas 1.18 y 1.27**, se sigue el siguiente corolario:

**Corolario 1.28.** *Si $X$ es una v.a positiva c.s que sigue la LB e $Y$ es una v.a positiva c.s independiente de $X$, entonces $X \cdot Y$ sigue la LB.*

DEMOSTRACIÓN. Sabemos que $\log_{10}(X)$ es $u.d.$ mod 1. Además, por ser $X$ independiente de $Y$, $\log_{10}(X)$ es independiente de $\log_{10}(Y)$. Aplicando el **Teorema 1.27**, $\log_{10}(X) + \log_{10}(Y)$ es $u.d.$ mod 1 $\Longrightarrow \log_{10}(X \cdot Y)$ es $u.d.$ mod 1 $\Longrightarrow X \cdot Y$ sigue la LB. □

## 1.5. Invariancia por cambio de escala

**Definición 1.29.** *Una probabilidad $P$ en $(\mathbb{R}^+, \mathscr{M})$ es invariante por cambio de escala si*

$$P(S) = P(aS) \quad \forall a > 0,\ \forall S \in \mathscr{M} \text{ fijo.}$$

Una propiedad característica de la LB es la de invariancia por cambio de escala. Este hecho fue sugerido por Roger Pinkham [22], quien intentó probar que la Ley de Benford es la única distribución invariante por cambio de escala. El argumento de Pinkham ha sido utilizado por muchos autores con el fin de explicar la aparición de la LB en abundantes tablas de datos, ya que si se presume que la distribución de los datos en cuestión son invariantes por cambio de escala, entonces necesariamente seguirán la Ley de Benford.

No olvidemos qué representa la LB. La Ley de Benford representa la peculiar distribución sobre los dígitos significativos intrínseca a una gran cantidad de procesos naturales. Si una tabla de constantes físicas, o una tabla de superficies de lagos o de países, es reescrita en otro sistema de unidades, el resultado será una tabla reescalada donde cada entrada es el mismo múltiplo de la correspondiente entrada en la tabla original.

Para fijar ideas, pensemos en una tabla con constantes físicas. Las entradas de esta tabla se pueden considerar como una muestra de una distribución desconocida de constantes físicas. Esta distribución desconocida tendrá asociada una distribución, también desconocida, de primeros dígitos significativos. Imaginemos que todas las constantes físicas fueran multiplicadas por un número fijo, ¿qué pasaría con la distribución intrínseca a las constantes físicas?, sería de esperar que fuera la misma.

Como veremos ahora, la propiedad de IE es suficiente para caracterizar la distribución completamente. El siguiente teorema prueba que la LB es la única distribución de dígitos significativos que cumple la propiedad de invariancia por cambio de escala.

**Teorema 1.30.** *Una probabilidad $P$ en $(\mathbb{R}^+, \mathscr{M})$ es invariante por cambio de escala, si, y sólo si, $P$ sigue la LB. Es decir, si $P(\mathrm{M} \leq t) = \log_{10}(t) \quad t \in [1, 10)$.*

DEMOSTRACIÓN. Veamos que si $P$ sigue la Ley de Benford entonces es invariante por cambio de escala. Queremos ver que

$$P(S) = P(aS) \quad \forall a > 0 \quad S \in \mathscr{M}.$$

Por la propiedad **iv)** del **Teorema 1.5**, asumimos que $1 < a < 10$ sin pérdida de generalidad. Como los intervalos de la forma $[1, 10^t]$, $0 < t < 1$, son una $\pi$ clase generadora de $\mathscr{B}_{[1,10)}$[3], es suficiente ver la propiedad de invariancia por escala para conjuntos de la forma

$$S_t = \bigcup_{n=-\infty}^{\infty} 10^n[1, 10^t] \quad 0 < t < 1.$$

Entonces

$$\begin{aligned}
&P(S_t) = P(aS_t) \\
&\iff P_M\left(\left[1, 10^t\right]\right) = P_M\left(\left(\left[a, a\cdot 10^t\right] \cap [a, 10)\right) \cup \left(\left[\frac{a}{10}, a\cdot 10^{t-1}\right] \cap [1, a)\right)\right) \\
&\iff P_M\left(\left[1, 10^t\right]\right) = P_M\left(\left[a, a\cdot 10^t\right] \cap [a, 10)\right) + P_M\left(\left[\frac{a}{10}, a\cdot 10^{t-1}\right] \cap [1, a)\right) \\
&\iff P_M\left(\left[1, 10^t\right]\right) = P_M\left(\left[a, \text{mín}\left(a\cdot 10^t, 10\right)\right)\right) + P_M\left(\left[1, a\cdot 10^{t-1}\right]\right).
\end{aligned} \tag{1.14}$$

En el segundo sumando de (1.14) se ha tenido en cuenta que $\frac{a}{10} < 1$ y que $a\cdot 10^{t-1} < a$ ya que $10^{t-1} < 1 \iff t < 1$.

Distingamos dos casos:

- Primer caso: $\text{mín}(a\cdot 10^t, 10) = a\cdot 10^t$

$$\begin{aligned}
&P_M\left(\left[1, 10^t\right]\right) = P_M\left(\left[a, \text{mín}\left(a\cdot 10^t, 10\right)\right)\right) + P_M\left(\left[1, a\cdot 10^{t-1}\right]\right) \\
&\iff P_M(\left[1, 10^t\right]) = P_M(\left[a, a\cdot 10^t\right]) + P_M\left(\left[1, a\cdot 10^{t-1}\right]\right) \\
&\iff (P_M)_{\log_{10}}([0, t]) = (P_M)_{\log_{10}}\left(\left[\log_{10}(a), \log_{10}\left(a\cdot 10^t\right)\right]\right) && (1.15) \\
&\iff t = \log_{10}\left(\frac{a\cdot 10^t}{a}\right) && (1.16) \\
&\iff t = t.
\end{aligned}$$

En (1.15) se ha tenido en cuenta que

$$\begin{aligned}
a\cdot 10^{t-1} < 1 &\iff a\cdot 10^t < 10 \\
&\implies P_M\left([1, a\cdot 10^{t-1}]\right) = 0,
\end{aligned}$$

y en (1.16) que $(P_M)_{\log_{10}} \stackrel{d}{=} U(0,1)$.

- Segundo caso: $\text{mín}(a\cdot 10^t, 10) = 10$

$$\begin{aligned}
&P_M\left(\left[1, 10^t\right]\right) = P_M\left(\left[a, \text{mín}\left(a\cdot 10^t, 10\right)\right)\right) + P_M\left(\left[1, a\cdot 10^{t-1}\right]\right) \\
&\iff P_M\left(\left[1, 10^t\right]\right) = P_M([a, 10]) + P_M\left(\left[1, a\cdot 10^{t-1}\right]\right) \\
&\iff t = (P_M)_{\log_{10}}([\log_{10}(a), 1)) + (P_M)_{\log_{10}}\left(\left[0, \log_{10}\left(a\cdot 10^{t-1}\right)\right)\right) \\
&\iff t = 1 - \log_{10}(a) + \log_{10}(a) + t - 1 \\
&\iff t = t.
\end{aligned}$$

[3] Sea $\pi$ el conjunto de los intervalos de la forma $[1, 10^t]$, $0 < t < 1$.
Sea $\Gamma = \left\{S \in \mathscr{B}_{[1,10)} : P(S) = P(aS) \text{ con } a > 0 \text{ fijo }\right\}$. Si IE se cumple para $\pi \implies \pi \subseteq \Gamma \subseteq \sigma(\pi) = \mathscr{B}_{[1,10)}$.
Ahora, es inmediato probar que $\Gamma$ es una $\lambda$-clase; por tanto, aplicando el Teorema $\pi$-$\lambda$ de Dynkin [4] $\mathscr{B}_{[1,10)} \subseteq \Gamma \implies \mathscr{B}_{[1,10)} = \Gamma$.

En definitiva, si $P$ sigue la Ley de Benford entonces es invariante por cambio de escala.

Sea ahora una probabilidad $P$ en $(\mathbb{R}^+, \mathscr{M})$ invariante por cambio de escala. Veamos que $P$ sigue la LB. Lo primero recordar que $P$ es Benford en el espacio medible $(\mathbb{R}^+, \mathscr{M})$ si, y sólo si, $P_{\mathrm{M}}$ sigue la distribución logarítmica en el espacio medible $\big([1,10), \mathscr{B}_{[1,10)}\big)$; es decir,

$$\begin{aligned}
&(P_{\mathrm{M}})_{\log_{10}} \overset{\mathrm{d}}{=} \lambda_{0,1} \\
&\iff P_{\log_{10}(\mathrm{M})} \overset{\mathrm{d}}{=} \lambda_{0,1}.
\end{aligned}$$

Por tanto, bastará con ver que $P_{\log_{10}(\mathrm{M})} \overset{\mathrm{d}}{=} \lambda_{0,1}$ en $\big([0,1), \mathscr{B}_{[0,1)}\big)$:

$$\begin{aligned}
&P_{\mathrm{M}}\left(\left[1,10^t\right)\right) = P_{\mathrm{M}}\left(\mathrm{M}\left(\left[10^{\alpha}, 10^{\alpha}10^t\right)\right)\right) \qquad \alpha \in \mathbb{I} \text{ fijo y } t \in (0,1) \\
&\iff P_{\mathrm{M}}\left(\left[1,10^t\right)\right) = P_{\mathrm{M}}\left(10^{[\log_{10}([10^{\alpha},10^{\alpha}10^t)) \bmod 1]}\right) \qquad (1.17) \\
&\iff P_{\log_{10}(\mathrm{M})}\left([0,t)\right) = P_{\log_{10}(\mathrm{M})}\left(\log_{10}\left(\left[10^{\alpha}, 10^{\alpha}10^t\right)\right) \bmod 1\right) \\
&\iff P_{\log_{10}(\mathrm{M})}\left([0,t)\right) = P_{\log_{10}(\mathrm{M})}\left([\alpha, \alpha+t) \bmod 1\right). \qquad (1.18)
\end{aligned}$$

En (1.17) se ha utilizado el **Lema 1.17**. Sea $T(x) = (x-\alpha) \bmod 1$, 1.18 se puede reescribir de la siguiente forma

$$P_{\log_{10}(\mathrm{M})}\left([0,t)\right) = P_{\log_{10}(\mathrm{M})}\left(T^{-1}\left([0,t)\right)\right) \quad \forall t \in [0,1).$$

Por tanto, $P_{\log_{10}(\mathrm{M})} \overset{\mathrm{d}}{=} \left(P_{\log_{10}(\mathrm{M})}\right)_T$, ya que los intervalos de la forma $[0,t)$, $t \in [0,1)$, son una $\pi$-clase generadora de $\mathscr{B}_{[0,1)}$. Por el **Teorema B.3**, concluimos que

$$P_{\log_{10}(\mathrm{M})} \overset{\mathrm{d}}{=} \lambda_{0,1}.$$

□

Hemos probado que la única distribución de dígitos significativos consistente con la propiedad de IE es la LB.

## 1.6. Invariancia por cambio de base

Al principio del trabajo se definió la mantisa de un número real positivo en base 10, y en consecuencia la definición de la LB involucra a $\log_{10}$. Esta elección puede considerarse arbitraria. En realidad, la LB se puede definir de manera más genérica, para cualquier base $b > 1$, $b \in \mathbb{N}$.

**Definición 1.31.** *Sea $b \in \mathbb{N}$, $b > 1$. La mantisa en base $b$ de un número real positivo $x$, es el único número $r \in [1,b)$ tal que $x = r \cdot b^n$, para algún $n \in \mathbb{Z}$.*

**Definición 1.32.** *Sea $b \in \mathbb{N}$, $b > 1$. Llamaremos $\mathrm{M}_b(x) : \mathbb{R}^+ \to [1,b)$ a la función mantisa base $b$. Si $x \in \mathbb{R}^+$, entonces $\mathrm{M}(x)$ es su mantisa base $b$.*

Llamaremos forma general de la LB en base $b$ a

$$\mathrm{Prob}(\mathrm{M}_b \le t) = \log_b(t) \quad t \in [1,b). \qquad (1.19)$$

La siguiente definición generaliza la **Definición 1.6**

**Definición 1.33.** *Una distribución de probabilidad $P$ en $(\mathbb{R}^+, \mathscr{B})$ sigue la LB en base $b$, si $P_{\mathrm{M}_b}$ sigue la distribución logarítmica base $b$ en $\left([1,b), \mathscr{B}_{[1,b)}\right)$; es decir, si*

$$P_{\mathrm{M}_b}([1,t)) = \log_b t \quad \forall t \in [1,b). \tag{1.20}$$

**Nota 1.34.** *Si no se especifica la base, se supondrá que $b = 10$, como hemos estado haciendo hasta ahora.*

**Definición 1.35.** *Llamaremos $\mathscr{M}_b$ a $\sigma(M_b)$.*

**Lema 1.36.**

$$S \in \mathscr{M}_b \Longleftrightarrow S = \bigcup_{n=-\infty}^{\infty} B \cdot b^n \quad B \subseteq [1,b) \text{ Borel.}$$

La demostración es idéntica a la del **Lema 1.4**.

**Teorema 1.37.** *$\mathscr{M}_b$ cumple las siguientes propiedades:*

**i)** *Cualquier conjunto no vacío $S \in \mathscr{M}_b$ no es acotado y, además, el $0$ es un punto de acumulación de $S$.*

**ii)** *$\mathscr{M}_b$ es cerrado para la multiplicación por escalares.*

**iii)** *$\mathscr{M}_b$ es cerrado por raíces enteras.*

**iv)** *$\mathscr{M}_b$ es auto-similar.*

La demostración es idéntica a la del **Teorema 1.5**. En particular, el apartado **iii)** del **Teorema 1.37** nos dice que si $S \in \mathscr{M}_b \Longrightarrow S^{\frac{1}{m}} \in \mathscr{M}_b$. Recordemos la forma general de $S^{\frac{1}{m}}$, a partir de la de $S$, demostrada en el apartado **iii)** del **Teorema 1.5**, la cual se generaliza a base $b$

$$\begin{aligned} S \in \mathscr{M}_b &\Longleftrightarrow S = \mathrm{M}_b^{-1}(B) \text{ para algún Borel en } [1,b) \\ &\Longrightarrow S^{\frac{1}{m}} = \mathrm{M}_b^{-1}\left(\bigcup_{j=0}^{m-1}\left(B^{\frac{1}{m}} \cdot b^{\frac{j}{m}}\right)\right). \end{aligned} \tag{1.21}$$

El siguiente Lema muestra cómo se expresa un conjunto originalmente en base $b$, en base $b^n$.

**Lema 1.38.** *Para toda base $b$,*

$$\mathrm{M}_b^{-1}(B) = \bigcup_{k=0}^{n-1} \mathrm{M}_{b^n}^{-1}\left(b^k B\right) \quad \forall n \in \mathbb{N} \text{ y } B \subseteq [1,b) \text{ Borel}$$

DEMOSTRACIÓN. Fijamos $n \in \mathbb{N}$ y $B \subseteq [1,b)$ y $x \in \mathbb{R}^+$. Se tiene que

$$\begin{aligned} x &= \mathrm{M}_b(x) b^{n_0} \qquad n_0 \in \mathbb{N} \\ x &= \mathrm{M}_{b^n}(x) b^{n_1 n} \quad n_1 \in \mathbb{N} \end{aligned}$$

$$\Longrightarrow \mathrm{M}_b(x) b^{n_0} = \mathrm{M}_{b^n}(x) b^{n_1 n}.$$

Veamos la inclusión $\mathrm{M}_b^{-1}(B) \subseteq \bigcup_{k=0}^{n-1} \mathrm{M}_{b^n}^{-1}\left(b^k B\right)$:

$$\mathrm{M}_b(x) \in B \Longrightarrow \mathrm{M}_{b^n}(x) \in B \cdot b^{n_0 - n_1 n},$$

si $n_0 - n_1 n \in \{0, 1, ..., n-1\}$ hemos acabado:

$$\begin{aligned} B \subseteq [1, b) &\Longrightarrow x \in \left[b^{n_0}, b^{n_0+1}\right) \\ &\Longrightarrow \mathrm{M}_{b^n}(x) \in \left[b^{n_0 - n_1 n}, b^{n_0+1-n_1 n}\right) \subseteq [1, b^n) \\ &\Longrightarrow n_0 - n_1 n \geq 1 \text{ y } n_0 - n_1 n \leq n-1 \\ &\Longrightarrow n_0 - n_1 n \in \{0, 1, ..., n-1\}. \end{aligned}$$

Veamos la otra inclusión:

$$\begin{aligned} &\mathrm{M}_{b^n} \in B \cdot b^k, \text{ con } k \in \{0, 1, ..., n-1\} \\ &\Longrightarrow \mathrm{M}_b(x) \in B \cdot b^{k + n_1 n - n_0} \\ &\Longrightarrow \mathrm{M}_b(x) \in B. \end{aligned}$$

□

Por el momento, la propiedad más destacada de la LB ha sido la de invariancia por cambio de escala. Hemos demostrado en el **Teorema 1.30** que la LB es la única distribución invariante por cambio de escala en $(\mathbb{R}^+, \mathscr{M})$. Uno de los inconvenientes de la IE es que no permite que un punto tenga probabilidad positiva, ya que todos los puntos tendrían esa misma probabilidad positiva, llegando al absurdo.

Para fijar ideas, volvamos a pensar en una tabla de constantes físicas, la cual puede considerarse como una muestra de una distribución desconocida de constantes físicas. Por ejemplo, la constante $c$, la velocidad de la luz, es una constante física, ya que aparece en la fórmula $E = mc^2$. Pero la constante 1 también es una constante física, ya que aparece en la fórmula $f = ma$. La cuestión es que la constante 1 no se suele registrar como una constante fundamental, de incluirse, es plausible que esta constante, este punto, tuviera probabilidad positiva considerando la distribución de constantes físicas.

En lugar de suponer que los datos son IE, supongamos que son IB, es decir, la distribución de dígitos significativos se mantiene al cambiar de base los datos. Como veremos ahora, la hipótesis de IB caracteriza las mezclas de la LB con la medida de Dirac concentrada en el 1[4].

Para motivar la idea de IB, y con ello la **Definición 1.39**, pensemos en un ejemplo concreto. Sea

$$S = (D_1 = 1) = \bigcup_{n \in \mathbb{Z}} 10^n [1, 2),$$

por 1.21

$$S^{\frac{1}{2}} = \bigcup_{n \in \mathbb{Z}} 10^n \left( [1, \sqrt{2}) \cup [\sqrt{10}, \sqrt{20}) \right).$$

[4]Llamaremos medida de Dirac concentrada en $a$ ($\delta_a$) a la medida de probabilidad definida en $(\Omega, \sigma)$, con $\Omega \subseteq \mathbb{R}^+$, $a \in \Omega$, por:

$$\delta_a(B) = I_B(a) \quad B \in \sigma$$

Hemos expresado $S$ en base $b=10$. Expresémoslo ahora en base $b=100$.

Por el **Lema 1.38**
$$S=\bigcup_{n\in\mathbb{Z}} 100^n\left([1,2)\cup[10,20)\right),$$
por tanto, el conjunto
$$C=\left\{x>0:\mathrm{M}_b(x)\in\left[1,b^{\frac{a}{2}}\right)\cup\left[b^{\frac{1}{2}},b^{\frac{(1+a)}{2}}\right)\right\}\qquad \text{con } a=\log_{10}2,$$
es igual a $S^{\frac{1}{2}}$ en base $b=10$, y es igual a $S$ en base $b=100$. En conclusión, si una probabilidad $P$ en $(\mathbb{R}^+,\mathscr{M}_b)$ cumple la propiedad de IB, $P(S)$ y $P(S^{\frac{1}{2}})$ deberían valer lo mismo, y análogamente para todas las raíces $n$-ésimas.

**Definición 1.39.** *Una probabilidad $P$ en $(\mathbb{R}^+,\mathscr{M}_b)$ es invariante por cambio de base si*
$$P(S)=P\left(S^{\frac{1}{n}}\right)\quad \forall n\in\mathbb{N}\,,\forall S\in\mathscr{M}_b \text{ fijo.}$$

Como ya hemos anticipado, en el siguiente teorema se demostrará que las únicas probabilidades en $(\mathbb{R}^+,\mathscr{M}_b)$ invariantes por cambio de base son combinaciones convexas de dos probabilidades: una que sigue la LB en base $b$ y la medida de Dirac concentrada en el 1.

**Teorema 1.40.** *Una probabilidad $P$ en en $(\mathbb{R}^+,\mathscr{M}_b)$ es invariante por cambio de base si, y solo si,*
$$P\overset{\mathrm{d}}{=}qP_b+(1-q)\,\delta_1\quad \text{para algún } q\in[0,1].$$
*Con $P_b$ siguiendo la LB en base $b$.*

DEMOSTRACIÓN. Para mayor simplicidad, la demostración se hará utilizando la base $b=10$. La demostración para cualquier base $b\in\mathbb{N}$ es análoga.

Denotemos por $P_L$ una probabilidad que sigue la LB en base 10. Sea $P\overset{\mathrm{d}}{=}qP_L+(1-q)\,\delta_1$ para algún $q\in[0,1]$. Veamos que tanto $P_L$ como $\delta_1$ tienen dígitos significativos invariantes por base, y en consecuencia, cualquier combinación convexa los tendrá.

Sea $S\in\mathscr{M}$ y $m\in\mathbb{N}$. Como $1\in S\iff 1\in S^{\frac{1}{m}}\implies \delta_1$ tiene dígitos significativos invariantes por base.

Consideremos la $\pi$-clase generadora de $\mathscr{B}_{[1,10)}$ formada por intervalos de la forma $[1,10^t]$, $t\in(0,1)$. Entonces
$$\begin{aligned}P_{L_{\mathrm{M}}}\left([1,10^t]^{\frac{1}{m}}\right)&=P_{L_{\mathrm{M}}}\left(\bigcup_{j=0}^{m-1}\left([1,10^t]^{\frac{1}{m}}\cdot 10^{\frac{j}{m}}\right)\right)=\sum_{j=0}^{m-1}P_{L_{\mathrm{M}}}\left(\left[10^{\frac{j}{m}},10^{\frac{j+t}{m}}\right]\right)\\&=\sum_{j=0}^{m-1}P_{L_{\log(\mathrm{M})}}\left(\left[\frac{j}{m},\frac{j+t}{m}\right]\right)=\sum_{j=0}^{m-1}\lambda_{0,1}\left(\left[\frac{j}{m},\frac{j+t}{m}\right]\right)\\&=\sum_{j=0}^{m-1}\frac{t}{m}=t=P_{L_{\mathrm{M}}}\left([1,10^t]\right)\end{aligned}$$

En la primera igualdad se ha tenido en cuenta 1.21. En definitiva, $P_L$ tiene dígitos significativos invariantes por cambio de base.

Sea ahora una probabilidad $P$ en $(\mathbb{R}^+,\mathscr{M})$ invariante por cambio de base. Veamos que $P \overset{d}{=} qP_L + (1-q)\,\delta_1$ para algún $q \in [0,1]$. Notar que $P \overset{d}{=} qP_L + (1-q)\,\delta_1$ en $(\mathbb{R}^+,\mathscr{M})$ si, y sólo si, $P_{\mathrm{M}} \overset{d}{=} qP_{L_{\mathrm{M}}} + (1-q)\,\delta_1$ en $\left([1,10),\mathscr{B}_{[1,10)}\right)$, es decir,

$$\begin{aligned}
&(P_{\mathrm{M}})_{\log_{10}} \overset{d}{=} q\lambda_{0,1} + (1-q)\,\delta_0\\
&\Longleftrightarrow P_{\log_{10}(\mathrm{M})} \overset{d}{=} q\lambda_{0,1} + (1-q)\,\delta_0.
\end{aligned}$$

Por tanto bastará con ver que $P_{\log_{10}(\mathrm{M})} \overset{d}{=} q\lambda_{0,1} + (1-q)\,\delta_0$ en $\left([1,10),\mathscr{B}_{[1,10)}\right)$ para ver que $P \overset{d}{=} qP_L + (1-q)\,\delta_1$

$$\begin{aligned}
&P_{\mathrm{M}}\left([1,10^t]\right) = P_{\mathrm{M}}\left(\bigcup_{j=0}^{m-1}\left[10^{\frac{j}{m}},10^{\frac{t+j}{m}}\right]\right) \qquad t\in(0,1)\,,\, m\in\mathbb{N}\\
&\Longleftrightarrow P_{\log_{10}(\mathrm{M})}\left([0,t]\right) = P_{\log_{10}(\mathrm{M})}\left(\bigcup_{j=0}^{m-1}\left[\frac{j}{m},\frac{t+j}{m}\right]\right).
\end{aligned} \tag{1.22}$$

Sea $T_m(x) = (mx) \bmod 1$, 1.22 se puede reescribir de la siguiente forma

$$P_{\log_{10}(\mathrm{M})}\left([0,t]\right) = P_{\log_{10}(\mathrm{M})}\left(T_m^{-1}\left([0,t]\right)\right) \quad t\in[0,1)\,,\, m\in\mathbb{N},$$

por tanto, $P_{\log_{10}(\mathrm{M})} \overset{d}{=} \left(P_{\log_{10}(\mathrm{M})}\right)_{T_m}$ $\forall m\in\mathbb{N}$, ya que los intervalos de la forma $[0,t]$, $t\in[0,1)$ son una $\pi$-clase generadora de $\mathscr{B}_{[0,1)}$. Por el **Teorema B.2**, concluimos que

$$(P_{\mathrm{M}})_{\log_{10}} \overset{d}{=} q\lambda_{0,1} + (1-q)\,\delta_0.$$

□

**Corolario 1.41.** *Sea $P$ una probabilidad en $(\mathbb{R}^+,\mathscr{M}_b)$ , con $P\left(\{\pm 10^k : k\in\mathbb{Z}\}\right) = 0$; entonces, $P$ sigue la LB en base b si, y solo si, es invariante por cambio de base.*

DEMOSTRACIÓN. Supongamos que $P$ es invariante por cambio de base. Por el Teorema anterior, $\exists q\in[0,1]$ tal que

$$P \overset{d}{=} qP_b + (1-q)\,\delta_1.$$

Ahora, si $q<1 \Longrightarrow P\left(\{\pm 10^k : k\in\mathbb{Z}\}\right) > (1-q)\,\delta_1\left(\{\pm 10^k : k\in\mathbb{Z}\}\right) > 0$, absurdo. Luego $q=1$; y en definitiva, $P \overset{d}{=} P_b$.

Si $P$ sigue la LB en base b, es invariante por cambio de base. Además

$$P\left(\left\{\pm 10^k : k\in\mathbb{Z}\right\}\right) = P_{\mathrm{M}}(1) = \log_b(1) = 0.$$

□

Habiendo expuesto las dos propiedades fundamentales de la LB, concluiremos esta sección teórica con la conexión entre la LB y las mezclas de distribuciones.

## 1.7. Mezclas de distribuciones

En la introducción se dijo que la LB aparece en numerosas tablas de datos. No obstante, ¿por qué deberían los dígitos significativos en una tabla de datos presentar una distribución logarítmica?, o equivalentemente, ¿por qué deberían ser invariantes por cambio de escala o base? Pensemos en un ejemplo concreto, precisamente en los datos que Frank Benford aportó en su artículo. Benford recogió una gran cantidad de datos de numerosos campos. Lo que parece natural por tanto, es pensar que en general, los datos provienen de muchas distribuciones diferentes.

El objetivo de esta sección es demostrar que si se toman muestras aleatorias de diferentes distribuciones y se combinan los resultados, entonces las muestras combinadas convergen a la LB, a condición de que el muestreo sea invariante por cambio de escala o base.

De manera informal, una m.p.a es una probabilidad $\xi$ elegida aleatoriamente en un espacio medible $(\mathfrak{M},\mathfrak{A})$, siendo $\mathfrak{M}$ el conjunto de todas las probabilidades en $(\mathbb{R},\mathscr{B})$. Desde este punto de vista, $\xi$ es simplemente una probabilidad que depende de un parámetro $\omega$, el cual pertenece a un cierto espacio muestral $\Omega$. De manera más formal, una m.p.a es una función $\xi:\Omega\longrightarrow\mathfrak{M}$ definida en el espacio probabilístico $(\Omega,\sigma,\mathbb{P})$ tal que para todo $B\in\mathscr{B}$, la función $\omega\longrightarrow\xi(\omega)(B)$ es una *v.a.*. Uno de los ejemplos más sencillos para ilustrar la noción de m.p.a es el siguiente:

**Ejemplo 1.42.** *Sea $F$ una distribución y sea $x_1,\ldots,x_n$ un muestreo independiente de $F$. Sea $\xi$ la aplicación que para cada muestreo, es la distribución empírica (asociada a esa muestra), siendo la distribución empírica, la que toma el valor $x_i$, $i=1,\ldots,n$ con probabilidad $\frac{1}{n}$. Entonces, $\xi$ es una m.p.a.*

Otro ejemplo sencillo :

**Ejemplo 1.43.** *Sea $\xi$ una m.p.a que es $U(0,1)$ con probabilidad $\frac{1}{2}$ o si no $N(0,1)$. Es decir, $\mathbb{P}(\xi=U(0,1))=\frac{1}{2}$ y $\mathbb{P}(\xi=N(0,1)))=\frac{1}{2}$. Para una realización de $\xi$; se lanza una moneda, si es cara $\xi(\omega)$ es una distribución $U(0,1)$, si es cruz, $\xi(\omega)$ es una $N(0,1)$.*

La siguiente definición formaliza la noción de combinar datos provenientes de diferentes distribuciones. Se basa en utilizar una m.p.a para generar una sucesión aleatoria de *v.a.*, para después generar muestras de esas variables aleatorias.

**Definición 1.44.** *Sea $m\in\mathbb{N}$ y $\xi$ una m.p.a. Una sucesión de m-muestras $\xi$-aleatorias es una sucesión $(X_n)$ de v.a. en $(\Omega,\sigma,\mathbb{P})$ tal que $\forall j\in\mathbb{N}$ y alguna sucesión de m.p.a $(\xi_n)$ i.i.d con $\xi_1=\xi$, se cumplen las dos propiedades siguientes:*

- *Dado $\xi_j\stackrel{d}{=}P$, las variables aleatorias*

$$X_{(j-1)m+1}\,,X_{(j-1)m+2}\,,\ldots,X_{jm}$$

  *son i.i.d con la distribución de P.*

- *Las variables*

$$X_{(j-1)m+1}\,,X_{(j-1)m+2}\,,\ldots,X_{jm}$$

*son independientes de*

$$\xi_i, X_{(i-1)m+1}, X_{(i-1)m+2}, ..., X_{im} \quad \text{si } i \neq j.$$

Ilustremos esta definición con un ejemplo:

**Ejemplo 1.45.** *Sea $\xi$ la m.p.a del* ***Ejemplo 1.43****. Veamos cómo es una sucesión de 5-muestras $\xi$-aleatorias; es decir, cómo es la sucesión $(X_n)$:*

*$X_1, X_2, X_3, X_4, X_5$ serán v.a.i.d, cuya distribución puede ser, o bien $U(0,1)$ con probabilidad $\frac{1}{2}$, o bien $N(0,1)$ con probabilidad $\frac{1}{2}$.*

*Las siguientes 5 v.a.i.d, $X_6, X_7, X_8, X_9, X_{10}$ serán independientes de las 5 primeras y cumplirán lo mismo: su distribución será $U(0,1)$ con probabilidad $\frac{1}{2}$ o $N(0,1)$ con probabilidad $\frac{1}{2}$.*

**Nota 1.46.** *Es importante decir que las v.a en cada 'bloque' de 5 no son independientes entre sí. Es decir, $X_6$ y $X_1$ sí son independientes por construcción, porque están en 'bloques' diferentes. Pero, por ejemplo, $X_2$ no es independiente de $X_3$. Si $X_2 > 1$, necesariamente $X_3$ será una v.a. con distribución $N(0,1)$, mientras que sin la condición $X_2 > 1$, $X_3$ es $U(0,1)$ con probabilidad $\frac{1}{2}$ ó $N(0,1)$ con probabilidad $\frac{1}{2}$.*

Resumiendo, estamos interesados en muestras de distribuciones, que a su vez son muestras de lo que hemos llamado una m.p.a. Para ello, se construye una sucesión de *v.a.* $(X_n)$, cumpliendo ciertas propiedades. En estas sucesiones centraremos nuestra atención. Ya hemos comentado que, en general, las *v.a.* de la sucesión no son independientes. En el **Lema 1.48** se probará que la sucesión $(X_n)$ es de *v.a.* igualmente distribuidas casi seguro, siendo su distribución la expuesta en la **Proposición 1.47**

**Proposición 1.47.** *Sea $\xi$ una m.p.a. Entonces, $\mathbb{E}\xi$, definida como:*

$$(\mathbb{E}\xi)(B) := \mathbb{E}(\xi(B)) \quad B \in \mathscr{B},$$

*es una probabilidad en $(\mathbb{R}, \mathscr{B})$.*

DEMOSTRACIÓN.

- $0 \leq (\mathbb{E}\xi)(B) \leq 1 \ \forall B \in \mathscr{B}$

$$\begin{aligned} &\text{Como } \forall \omega \in \Omega,\ 0 \leq \xi(\omega)(B) \leq 1 \\ &\Longrightarrow 0 = \int_\Omega 0 d\mathbb{P} \leq \int_\Omega \xi(\omega)(B)\, d\mathbb{P} \\ &= \mathbb{E}(\xi(B)) \leq \int_\Omega 1 d\mathbb{P} = 1. \end{aligned}$$

- $(\mathbb{E}\xi)(\emptyset) = 0$, $(\mathbb{E}\xi)(\Omega) = 1$

$$(\mathbb{E}\xi)(\emptyset) = \int_\Omega \xi(\omega)(\emptyset) d\mathbb{P} = \int_\Omega 0 d\mathbb{P} = 0.$$
$$(\mathbb{E}\xi)(\Omega) = \int_\Omega \xi(\omega)(\Omega) d\mathbb{P} = \int_\Omega 1 d\mathbb{P} = 1.$$

- Si $B_1, B_2, \ldots$ es una sucesión de elementos de $\mathscr{B}$ disjuntos y $\bigcup_{k=1}^{\infty} B_k \in \mathscr{B}$; entonces $(\mathbb{E}\xi)\left(\bigcup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} (\mathbb{E}\xi)(B_k)$

$$
\begin{aligned}
(\mathbb{E}\xi)\left(\bigcup_{k=1}^{\infty} B_k\right) &= \int_{\Omega} \xi(\omega)\left(\bigcup_{k=1}^{\infty} B_k\right) d\mathbb{P} = \int_{\Omega} \left(\sum_{k=1}^{\infty} \xi(\omega)(B_k)\right) d\mathbb{P} \\
&= \sum_{k=1}^{\infty} \left(\int_{\Omega} \xi(\omega)(B_k)\, d\mathbb{P}\right) \qquad (1.23) \\
&= \sum_{k=1}^{\infty} (\mathbb{E}\xi)(B_k)
\end{aligned}
$$

En (1.23) se ha utilizado el Teorema de la Convergencia Dominada, factible ya que

$$
\begin{aligned}
&0 \leq \sum_{k=1}^{\infty} \xi(\omega)(B_k) \leq 1 \quad \forall \omega \in \Omega \\
&\mathbb{P}(\Omega) = 1.
\end{aligned}
$$

□

**Lema 1.48.** *Sea $\xi$ una m.p.a, $(\xi_n)$ una sucesión de m.p.a i.i.d, con $\xi_1 = \xi$. Sea $(X_n)$ una sucesión de $k$-muestras $\xi$-aleatorias. Entonces, $(X_n)$ son v.a.i.i.d casi seguro, con distribución $\mathbb{E}\xi$.*

DEMOSTRACIÓN. Sea $B \in \mathscr{B}$, $n \in \mathbb{N}$. Sea $Y_n = I(X_n \in B)$

$$
\begin{aligned}
\mathbb{P}(X_n \in B) &= \mathbb{E}Y_n = \mathbb{E}(\mathbb{E}(Y_n|\xi_n)) \qquad (1.24) \\
&= \mathbb{E}(\xi_n(B)) = \mathbb{E}(\xi(B)) \qquad (1.25) \\
&= (\mathbb{E}\xi)(B).
\end{aligned}
$$

En (1.24) se ha utilizado la ley de esperanzas iteradas y en (1.25) que $\xi_n \stackrel{\mathrm{d}}{=} \xi$. □

La siguiente Proposición muestra que la proporción límite de veces que una sucesión de $m$-muestras $\xi$-aleatorias está en un conjunto $B \in \mathscr{B}$ es, con probabilidad 1, igual a $\mathbb{E}\xi(B)$. Si la sucesión $(X_n)$ fuera de *v.a.i.i.d* el resultado sería inmediato tras aplicar la Ley Fuerte de los Grandes Números. Dado que en general no son independientes, se considerará la sucesión de 'bloques', los cuales sí son independientes. Aún así no podremos aplicar la *L.F.G.N.* a estos bloques, ya que aunque sí sean independientes, no están igualmente distribuidos; sin embargo, sí se darán las condiciones para aplicar la Ley Fuerte de los Grandes Números de Kolmogorov. Este resultado, junto con su demostración, se puede encontrar en [5].

**Proposición 1.49.** *Sea $\xi$ una m.p.a, $(\xi_n)$ una sucesión de m.p.a i.i.d, con $\xi_1 = \xi$. Sea $(X_n)$ una sucesión de $m$-muestras $\xi$-aleatorias, con $m \in \mathbb{N}$. Entonces, para todo $B \in \mathscr{B}$*

$$
\frac{\#\{1 \leq n \leq N : X_n \in B\}}{N} \xrightarrow{\text{c.s}} \mathbb{E}\xi(B) \text{ cuando } N \to \infty.
$$

DEMOSTRACIÓN. Fijamos un $B \in \mathscr{B}$ y $j \in \mathbb{N}$.

Lo primero que haremos será distinguir la sucesión por bloques. Sea

$$Y_j = \#\left\{1 \leq i \leq m : X_{(j-1)m+i} \in B\right\},$$

se tiene que,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Y_j = m \lim_{N\to\infty} \left(\frac{\#\{1 \leq n \leq N : X_n \in B\}}{N}\right).$$

$Y_j$ es una *v.a* binomial, de parámetros $m$ y $p = \mathbb{P}\left(X_j \in B\right) = \mathbb{E}\xi B$, por el **Lema 1.48**.

Entonces, $(Y_n)$ es una sucesión de *v.a* independientes, con la misma media $m\mathbb{E}\xi B$. Como

$$0 \leq Y_j \leq m \Longrightarrow 0 \leq Y_j^2 \leq m^2$$
$$\Longrightarrow 0 \leq \mathbb{E}Y_j^2 \leq m^2 \Longrightarrow \sum_{j=1}^{\infty} \frac{\mathbb{E}Y_j^2}{j^2} < \infty,$$

por la Ley Fuerte de los Grandes Números de Kolmogorov, concluimos que

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Y_j \overset{\text{c.s}}{=} m\mathbb{E}\xi\,(B),$$

y por tanto

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : X_n \in B\}}{N} \overset{\text{c.s}}{=} \mathbb{E}\xi\,(B).$$

□

Hasta ahora se han probado propiedades genéricas de las m.p.a. Como se ha dicho, el objetivo es demostrar que hay convergencia hacia la LB de muestras de ciertas m.p.a. Está claro que un resultado de este tipo no va a ser cierto para cualquier tipo de m.p.a, por ejemplo, si consideramos una m.p.a que sea una $U(0,1)$ *c.s.*, la sucesión $(X_n)$ convergerá hacia una $U(0,1)$ trivialmente, y sabemos que una *v.a.* $U(0,1)$ no sigue la LB (**Ejemplo 1.10**).

Por tanto, hay que imponer ciertas restricciones sobre las m.p.a si queremos convergencia hacia la LB. Dadas las propiedades que se han probado de IE y IB de la LB, no es de extrañar que haya que pedir invariancia por cambio de escala o base de las m.p.a, para lo cual deberemos definir adecuadamente qué significa que una m.p.a sea invariante por cambio de escala o base.

**Definición 1.50.** *Una m.p.a $\xi$ es invariante por cambio de escala, si la probabilidad $\mathbb{E}\xi$ en $(\mathbb{R}^+, \mathscr{M})$ es invariante por cambio de escala.*

**Definición 1.51.** *Una m.p.a $\xi$ es invariante por cambio de base, si la probabilidad $\mathbb{E}\xi$ en $(\mathbb{R}^+, \mathscr{M})$ es invariante por cambio de base.*

Subrayar que el que una m.p.a tenga dígitos significativos invariantes por cambio de escala o base no implica que las distribuciones asociadas a la m.p.a los tengan. Expliquemos esta idea con los datos aportados por Benford. Benford eligió aleatoriamente numerosos datos de diversos campos: áreas de ríos, estadísticas de la Liga Americana de béisbol, números aleatorios en revistas, términos de la sucesión armónica, datos sobre el índice de mortalidad, facturas de la luz, direcciones postales,... Cada campo representa una distribución. Los datos son una muestra de este proceso. Por aislado, los datos de los campos no son invariantes por cambio de escala o base, como ya dijo Benford en su artículo, pero la unión de los datos de todos los campos sí sigue la distribución logarítmica.

La siguiente proposición relaciona las propiedades de invariancia por cambio de escala o base de las m.p.a con la LB. La demostración va a ser inmediata gracias al trabajo ya hecho para probabilidades.

**Proposición 1.52.** *Sea $\xi$ una m.p.a. Las siguientes propiedades son equivalentes:*

**1)** *$\xi$ es invariante por cambio de escala.*

**2)** $\mathbb{E}\xi\left(\left\{\pm 10^k : k \in \mathbb{Z}\right\}\right) = 0$ *y $\xi$ es invariante por cambio de base.*

**3)** $\mathbb{E}\xi$ *sigue la LB.*

DEMOSTRACIÓN.

**1)$\Longleftrightarrow$3)**

$\xi$ invariante por cambio de escala $\Longleftrightarrow$ $\mathbb{E}\xi$ es invariante por cambio de escala $\Longleftrightarrow$ $\mathbb{E}\xi$ sigue la LB por el **Teorema 1.30**.

**2)$\Longleftrightarrow$3)**

Consecuencia del **Corolario 1.41**.

$\square$

Este último Teorema recoge lo que se ha ido exponiendo a lo largo de la sección. La razón de que la LB aparezca en numerosas tablas de datos se justifica cuando estos datos provienen de diversas distribuciones independientes. Puede que estas distribuciones por sí solas no tengan dígitos significativos invariantes por cambio de escala o base, pero sí su mezcla.

**Teorema 1.53.** *Sea $\xi$ una m.p.a. Suponemos que $\xi$ es invariante por cambio de escala o por cambio de base, además de cumplir que $\mathbb{E}\xi\left(\left\{\pm 10^k : k \in \mathbb{Z}\right\}\right) = 0$. Entonces, para todo $m \in \mathbb{N}$, toda sucesión $(X_n)$ de $m$-muestras $\xi$-aleatorias sigue la LB con probabilidad $1$; es decir,*

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : \mathrm{M}(X_n) \leq t\}}{N} \overset{\text{c.s}}{=} \log_{10}(t) \quad t \in [1,10).$$

DEMOSTRACIÓN.

Por la **Proposición 1.52**, $\mathbb{E}\xi$ sigue la LB. Ahora, consideramos $S_t = \mathrm{M}^{-1}([1,t))$, $t \in [1,10)$. Entonces

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : X_n \in S_t\}}{N} \overset{\text{c.s}}{=} \mathbb{E}\xi\,(S_t) \quad t \in [1,10).$$

Como

$$\begin{aligned} X_n \in S_t &\Longleftrightarrow X_n \in \mathrm{M}^{-1}([1,t)) \Longleftrightarrow \mathrm{M}(X_n) \in [1,t) \\ &\Longleftrightarrow \mathrm{M}(X_n) \leq t \end{aligned}$$

y

$$\mathbb{E}\xi\,(S_t) = (\mathbb{E}\xi)_{\mathrm{M}}\,([1,t)) = \log_{10}(t),$$

ya que $\mathbb{E}\xi$ sigue la LB, se concluye.

$\square$

Aquí termina la exposición de las principales propiedades teóricas de la Ley de Benford. A lo largo de este capítulo se han dado diversas justificaciones que motivan que la LB aparezca en numerosas tablas de datos, en particular, en los datos aportados por Frank Benford en su artículo. De manera sucinta, es el combinar números de diferentes fuentes lo que genera una distribución de distribuciones, una ley de verdadera aleatoriedad que es universal.

Téngase en cuenta que, por supuesto, esta peculiar distribución no aparece en todas las tablas de datos, ni tampoco las razones expuestas son el único motivo de la aparición de la LB en conjuntos de datos. De hecho, muchos expertos están de acuerdo en que la ubicuidad de la Ley de Benford sigue siendo un misterio [2].

Es importante destacar la diferencia entre la teoría y la práctica. El desarrollo teórico tiene como fin justificar matemáticamente el interés del modelo asociado a la ley logarítmica, en cambio, con el fin de aprovechar la LB para detectar anomalías en un conjunto de datos basta con tener evidencia empírica de su aparición, es decir, que la distribución empírica - el histograma - se ajuste a la LB. Esto motiva la primera sección del siguiente capítulo, donde se aportan tres conjuntos de datos: datos financieros, datos electorales, datos sobre las ciencias naturales, donde la LB aparece. En la segunda sección sacaremos partido de la aparición de la LB para detectar fraude en un caso conocido, el falseamiento del déficit de Grecia a principios de este siglo.

# Capítulo 2

# Aplicaciones prácticas de la Ley de Benford

## 2.1. Sobre la aparición de la LB en numerosas tablas de datos

Como se comentó al final del primer capítulo, en esta sección se busca mostrar que la Ley de Benford es una característica común a un gran número de procesos naturales. Para ello, utilizaremos tres ejemplos pertenecientes a diversos ámbitos. En cada uno de ellos hemos seleccionado un conjunto de datos que se presume no manipulado, o al menos no significativamente. Con estos ejemplos se busca motivar la idea de que en numerosas tablas de datos la distribución de dígitos significativos no es uniforme, sino que es más frecuente, por ejemplo, la aparición del 1 que del 9; por este motivo, no nos preocuparemos de las pequeñas discrepancias entre las distribuciones empíricas y la LB, en cambio, este será el principal objetivo en la siguiente sección.

En el caso de querer detectar fraude en transacciones internacionales vía este método, es vital que sin fraude el proceso generador de precios de esas transacciones se ajuste a la LB. En el caso de querer detectar fraude electoral vía este método, es vital que sin fraude el proceso generador de votos se ajuste a la LB. En el caso de querer validar un modelo físico o biológico vía este método, es vital que la LB sea inherente a tal proceso físico o biológico. La exposición de los tres ejemplos seguirá este esquema: primero, se presentará el conjunto de datos, indicando el área a la que pertenece. En segundo lugar, se hará un pequeño comentario, basado en lo explicado en el desarrollo teórico, del porqué se espera la aparición de la LB en este conjunto de datos. En tercer lugar, se comprobará visualmente que la distribución del primer dígito significativo y del primer y segundo dígitos significativos sigue la LB, mediante el uso de histogramas. Por último, se reflexionará acerca de las implicaciones de la aparición de la LB en conjuntos de datos pertenecientes a la correspondiente área. El código creado para generar las tablas y los histogramas de los tres ejemplos se encuentra en la **Sección C.1** del **Apéndice C**.

### 2.1.1. Ejemplo 1

En este primer ejemplo se verá cómo la LB aparece de manera natural en los datos contenidos en Estados Financieros de diferentes empresas.

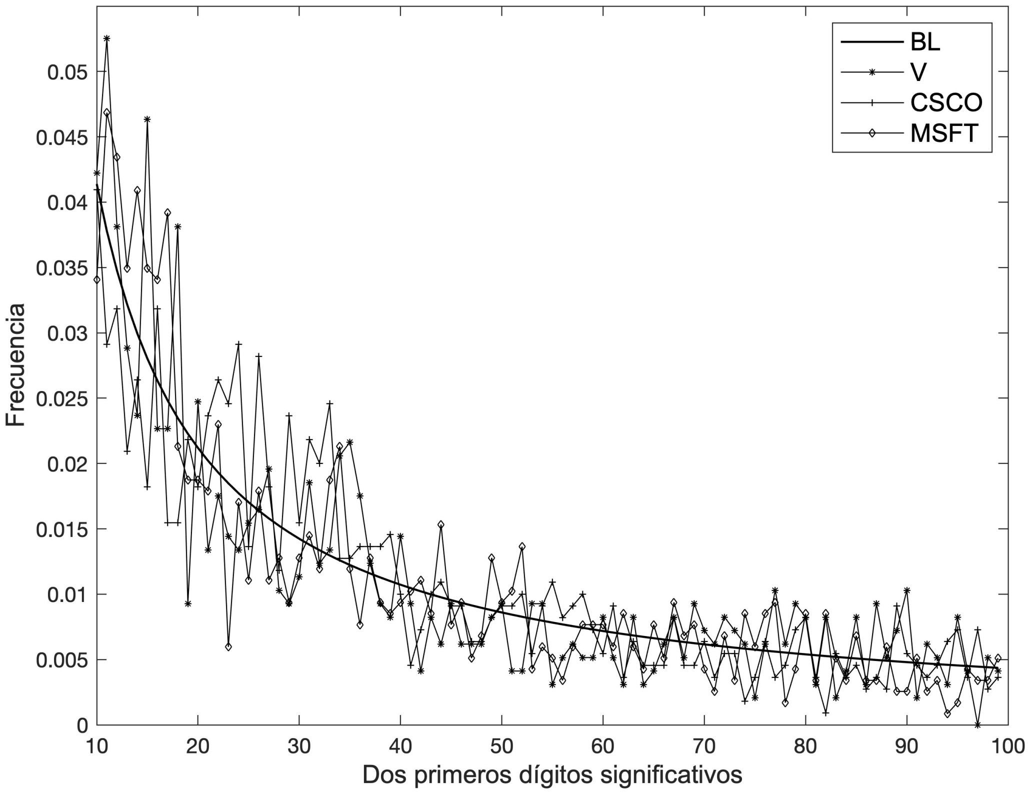
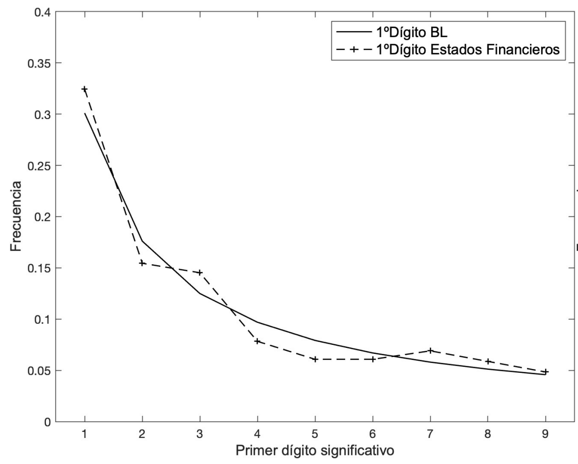


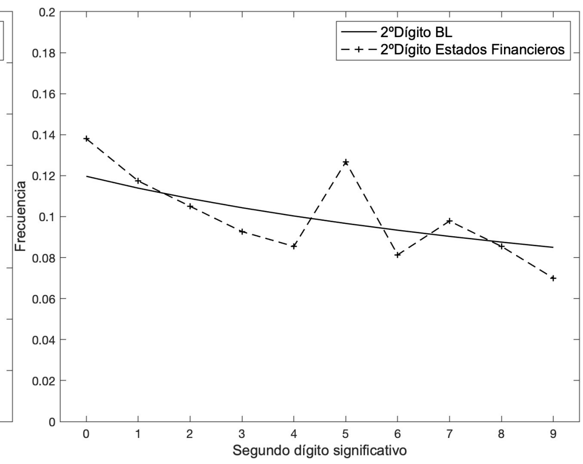

Figura 2.1: Distribuciones empíricas de los dos primeros dígitos significativos de los estados financieros de las empresas estadounidenses: Visa Inc., Cisco Systems Inc., Microsoft Corporation

| Categoría | Estado Financiero |
|---|---|
| Cuentas por Pagar a Corto Plazo | Balance General |
| Cuentas por Cobrar Netas a Corto Plazo | Balance General |
| Gastos Integrales Acumulados Netos de Impuestos | Balance General |
| Activos Corrientes | Balance General |
| Obligaciones Laborales a Corto Plazo | Balance General |
| Gastos Generales y Administrativos | Estado de Resultados |
| Plusvalía Mercantil | Balance General |
| Impuestos sobre la Renta Gastos/Beneficios | Estado de Resultados |
| Pasivos y Patrimonio Neto | Balance General |
| Pasivos Corrientes | Balance General |
| Resultado Neto | Estado de Resultados |
| Ingresos y Gastos no Operativos | Estado de Resultados |
| Resultado Operativo | Estado de Resultados |
| Otros Activos no Corrientes | Balance General |
| Otras Obligaciones no Corrientes | Balance General |
| Ingresos y Gastos no Operativos | Estado de Resultados |
| Propiedad, Planta y Equipo Netos | Balance General |

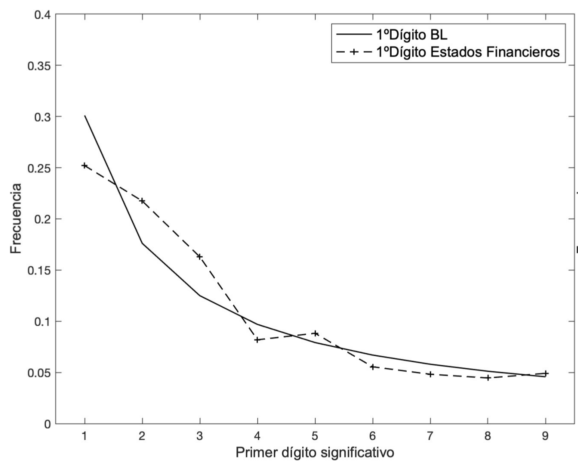
Tabla 2.1: Elementos en los Estados Financieros de las empresas Visa Inc., Cisco Systems, Inc. y Microsoft Corporation considerados en el análisis, junto al Estado Financiero al que pertenecen.



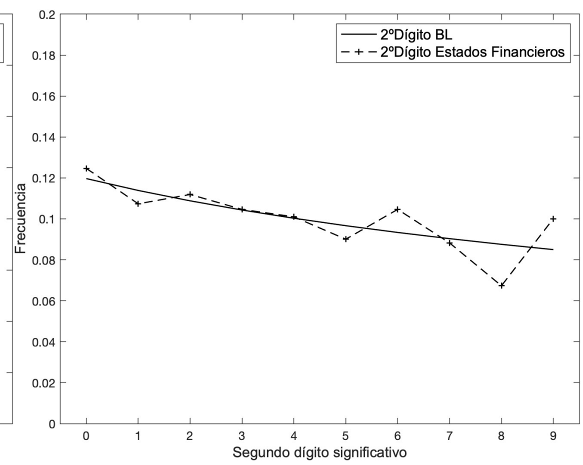

(a) Visa Inc.



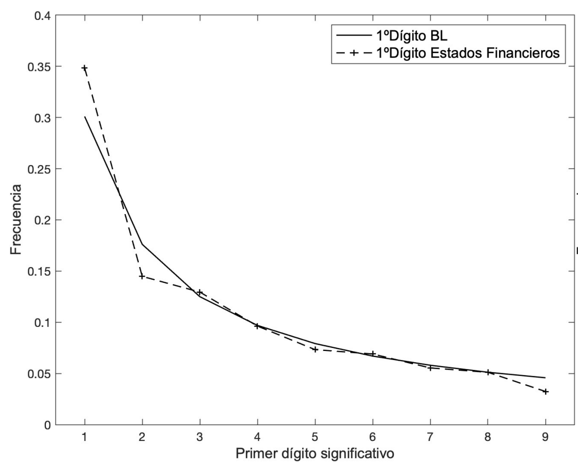

(b) Visa Inc.



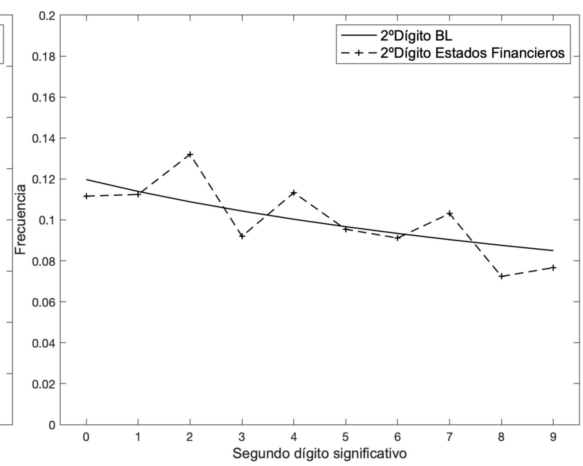

(c) Cisco Systems, Inc.




(d) Cisco Systems, Inc.




(e) Microsoft Corporation




(f) Microsoft Corporation

Figura 2.2: En la columna de la izquierda, la comparación de la distribución empírica del primer dígito significativo de las tres empresas estadounidenses con la Ley de Benford. En la columna de la derecha, la comparación de la distribución empírica del segundo dígito significativo de las tres empresas estadounidenses con la Ley de Benford.

Los Estados Financieros son tres documentos que recogen las actividades económicas de las empresas en un tiempo determinado. Estos tres documentos son: el Balance General, el Estado de Resultados y el Estado de flujos de Efectivo. Disponemos de tres conjuntos de datos, obtenidos de la base de datos EDGAR(SEC), con los valores de los elementos indicados en la tabla 2.1, de las empresas estadounidenses: Visa Inc., Cisco Systems, Inc. y Microsoft Corporation, listadas en el índice *Dow Jones Industrial Average*, uno de los más antiguos y seguidos.

Denotamos por $\xi_1$ al proceso generador de Estados Financieros. Sea $\xi_1(\omega_1)$ la muestra de tamaño $n_1 = 971$ correspondiente a Visa Inc., $\xi_1(\omega_2)$ la muestra de tamaño $n_2 = 1099$ correspondiente a Cisco Systems, Inc. y $\xi_1(\omega_3)$ la correspondiente a Microsoft Corporation., de tamaño $n_3 = 1174$. Un estado financiero refleja el rendimiento de una compañía, la cual realiza transacciones en varios mercados, los cuales están afectados por impredecibles y variados procesos económicos. Cada una de estas compañías comercia con diferentes productos, en diferentes países, cada cual con su peculiar distribución generadora de precios y cantidades. Además, los procesos multiplicativos son inherentes a gran cantidad de datos financieros. Estas razones motivarían *a priori* la aparición de la LB en los Estados Financieros.

La figura 2.1 muestra la distribución empírica conjunta del primer y segundo dígitos significativos de los Estados Financieros, en relación a la distribución esperada, la LB. Se observa un comportamiento similar, no uniforme, por parte de las distribuciones empíricas de los datos pertenecientes a las tres empresas, semejante a la LB. La figura 2.2 muestra las distribuciones empíricas marginales del primer y segundo dígitos significativos, en relación con la LB, la esperada. La distribución empírica tanto del primer como del segundo dígito significativo de los datos de las tres empresas parece aproximarse a la LB, sin apreciarse un patrón común en las desviaciones. En conclusión, nuestra premisa: la distribución de dígitos significativos de $\xi_1$ se ajusta a la LB, parece razonable.

Gracias al trabajo de Nigrini [19] es ampliamente conocido el uso de la LB entre auditores. Numerosos programas de *software* de auditoría utilizan la LB para detectar irregularidades, no solo en Estados Financieros sino también en declaraciones de impuestos.

### 2.1.2. Ejemplo 2

En este ejemplo nos centraremos en un proceso electoral. Disponemos de tres conjuntos de datos, en los cuales se encuentran el número de votos válidos emitidos a favor de las candidaturas, es decir, el número de votos válidos menos el número de votos en blanco, en todos los municipios españoles, correspondientes a las tres últimas elecciones generales de España: noviembre 2019, abril 2019, junio 2016; obtenidos de la página del Ministerio del Interior.

Denotamos por $\xi_2$ al proceso generador de votos válidos emitidos a favor de las candidaturas. Sea $\xi_2(\omega_1)$ la muestra de tamaño $n_1 = 8215$ correspondiente a las elecciones generales de noviembre de 2019, $\xi_2(\omega_2)$ la muestra de tamaño $n_2 = 8215$ correspondiente a las de abril de 2019 y $\xi_2(\omega_3)$ la muestra de tamaño $n_3 = 8209$ correspondiente a las de junio de 2016.

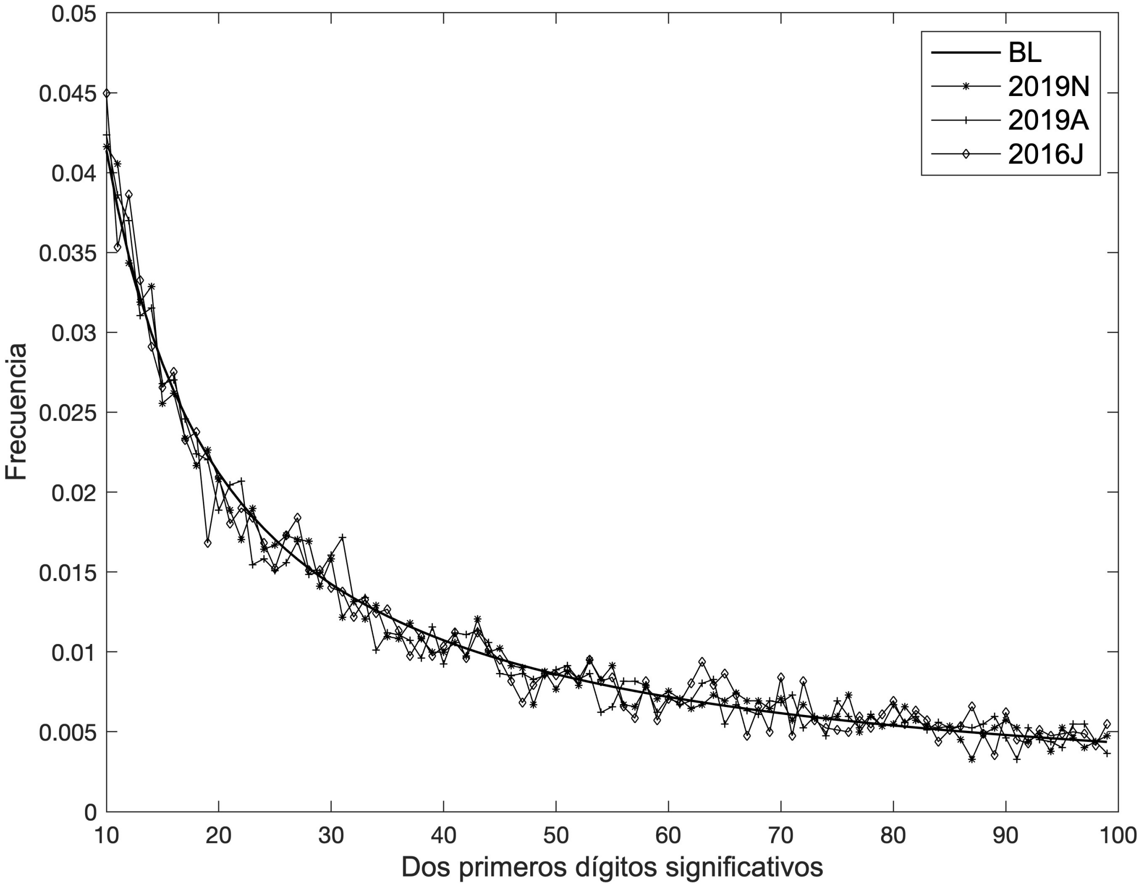
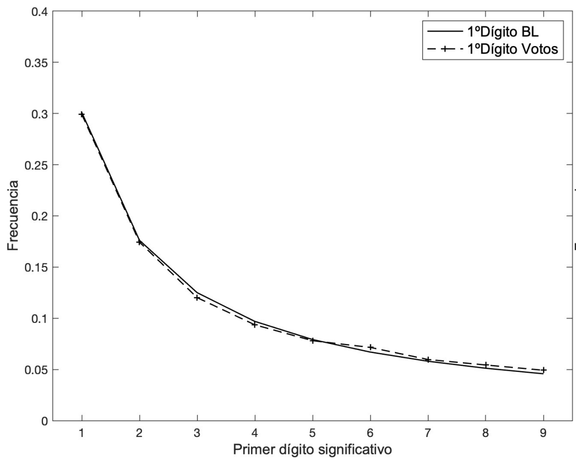


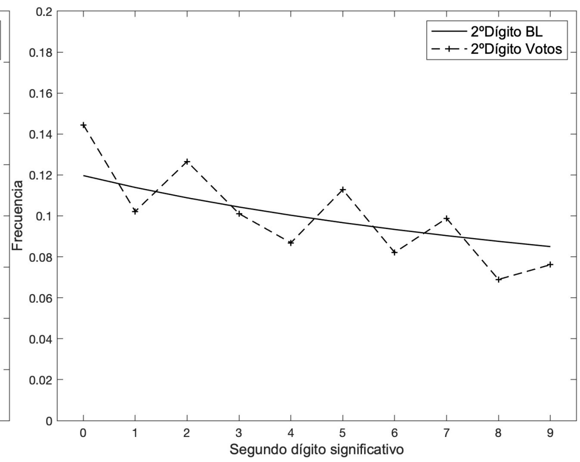

Figura 2.3: Distribuciones empíricas de los dos primeros dígitos significativos de los votos válidos emitidos a favor de las candidaturas en las últimas tres elecciones generales de España.

| **Municipio** | **Votos válidos emitidos a favor de las candidaturas** |
|---|---|
| Cuarte de Huerva | 6314 |
| Negueira de Muñiz | 107 |
| Torrubia | 16 |
| San Cristóbal de la Vega | 60 |
| Godojos | 41 |
| Pereruela | 280 |
| Graja de Iniesta | 216 |
| Guadalaviar | 142 |
| Oteiza | 482 |
| Guardo | 3126 |
| Villanueva de la Concepción | 1518 |
| Vega de Liébana | 449 |
| Garganta de los Montes | 227 |
| Almazán | 2651 |
| Teresa de Cofrentes | 351 |
| Torrecilla de la Orden | 197 |

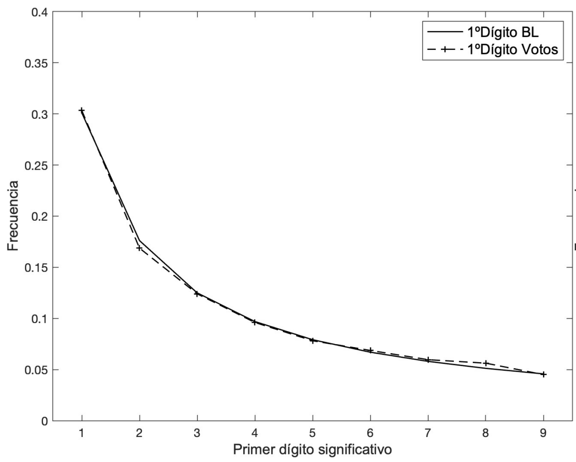
Tabla 2.2: 16 municipios españoles escogidos al azar, junto con su correspondiente número de votos válidos emitidos a favor de las candidaturas, en las elecciones generales de España de noviembre 2019.



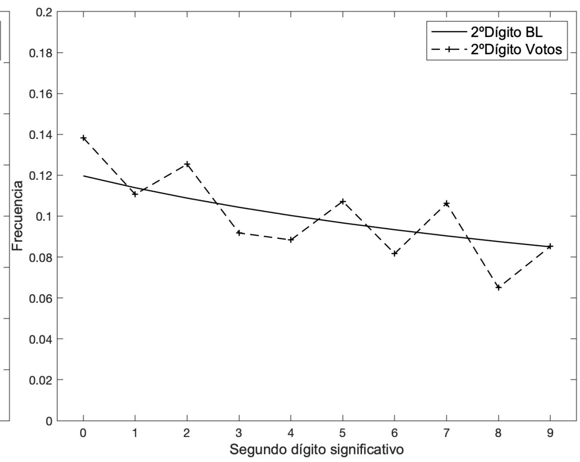

(a) Elecciones generales junio 2016



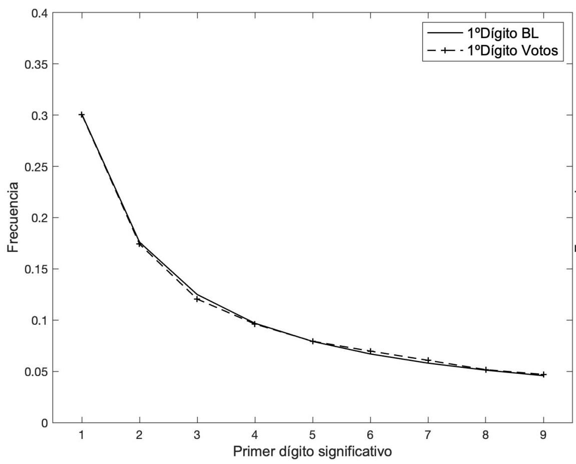

(b) Elecciones generales junio 2016



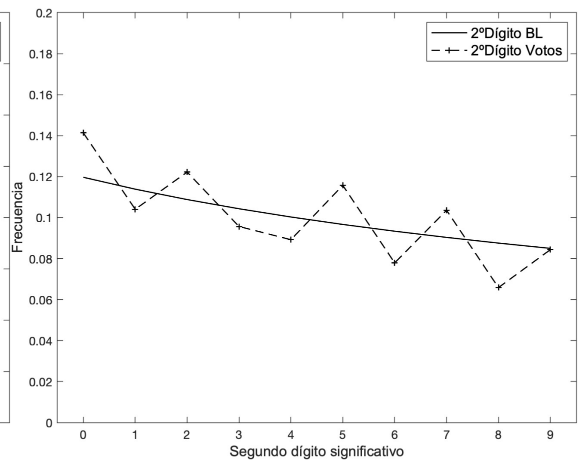

(c) Elecciones generales abril 2019




(d) Elecciones generales abril 2019




(e) Elecciones generales noviembre 2019




(f) Elecciones generales noviembre 2019

Figura 2.4: En la columna de la izquierda, la comparación de la distribución empírica del primer dígito significativo de las tres elecciones generales con la Ley de Benford. En la columna de la derecha, la comparación de la distribución empírica del segundo dígito significativo de las tres elecciones generales con la Ley de Benford.

En la tabla 2.2 se muestran 16 municipios españoles, escogidos al azar, y su correspondiente número de votos válidos emitidos a favor de las candidaturas en las elecciones generales de noviembre de 2019; esto es, una muestra sin reemplazamiento de tamaño $n = 16$ de $\xi_2(\omega_1)$. El proceso generador de votos es una mezcla de diferentes distribuciones: las que que determinan el número de población censada en cada municipio, las que determinan la participación electoral, las que determinan la proporción de votos que van al partido ganador,... luego se espera que $\xi_2$ siga la LB.

La figura 2.3 muestra la distribución empírica conjunta del primer y segundo dígitos significativos de los votos válidos, en relación a la distribución esperada, la LB. Se observa que la distribución empírica de las tres elecciones es prácticamente idéntica, muy próxima a la esperada. La figura 2.4 muestra las distribuciones empíricas marginales del primer y segundo dígitos significativos, en relación con la LB, la esperada. Se aprecia una muy buena adecuación por parte de la distribución del primer dígito significativo de los votos de las tres elecciones y unas oscilaciones similares en torno a la esperada en la tres distribuciones relativas al segundo dígito significativo. En conclusión, nuestra premisa: la distribución de dígitos significativos de $\xi_2$ se ajusta a la LB, parece aceptable. Esto implícitamente muestra cómo sería la distribución de dígitos significativos para un partido político que obtuviera el 100 % de los votos. Por ello es de esperar que, si se considera un partido político y su correspondiente número de votos totales, este conjunto de datos siga la LB, y por tanto una desviación significativa de la distribución logarítmica suscitaría sospechas. En el caso de procesos electorales, hay que considerar que anomalías en la distribución de dígitos significativos pueden ser debidas, no solo a la manipulación intencionada de los datos (los votos no corresponden fielmente a las intenciones de la población) sino a un comportamiento estratégico u otras políticas, como por por ejemplo el 'voto útil'. Con esto se quiere recalcar que el hecho de no ajustarse a la LB no implica que haya fraude, hay que entender en contexto los conjuntos de datos que se están analizando. Además, la LB se ha de usar como indicio, solamente una auditoría completa va a mostrar si hay un verdadero fraude o no. Este tipo de procedimientos estadísticos son útiles debido a que es inviable llevar a cabo una auditoría completa para todas las elecciones, por el elevado coste de recursos.

### 2.1.3. Ejemplo 3

En este ejemplo nos centramos en un proceso fisiológico. Disponemos de tres conjuntos de datos. Cada uno corresponde a un electrocardiograma fetal, obtenido directamente de la cabeza fetal, discretizado a 1000 entradas por segundo, de una duración total de 5 minutos. Los registros han sido obtenidos de tres mujeres diferentes, entre 38 y 41 semanas de gestación. Los datos se han obtenido de Physionet [8]. Un electrocardiógrafo es un aparato que mide la actividad eléctrica del corazón, gracias al uso de pequeños electrodos. Los electrodos miden las señales eléctricas que el corazón produce al latir, y son estas señales las que se registran. El gráfico resultante es el electrocardiograma. En la figura 2.5 se muestran los primeros 5 segundos de los tres electrocardiogramas fetales.

Denotamos por $\xi_3$ al proceso generador de electrocardiogramas fetales. Sean $\xi_3(\omega_1)$, $\xi_3(\omega_2)$, $\xi_3(\omega_3)$ las muestras de tamaño $n = 300000$ correspondientes al primer, segundo y tercer electrocardiograma fetal, respectivamente.

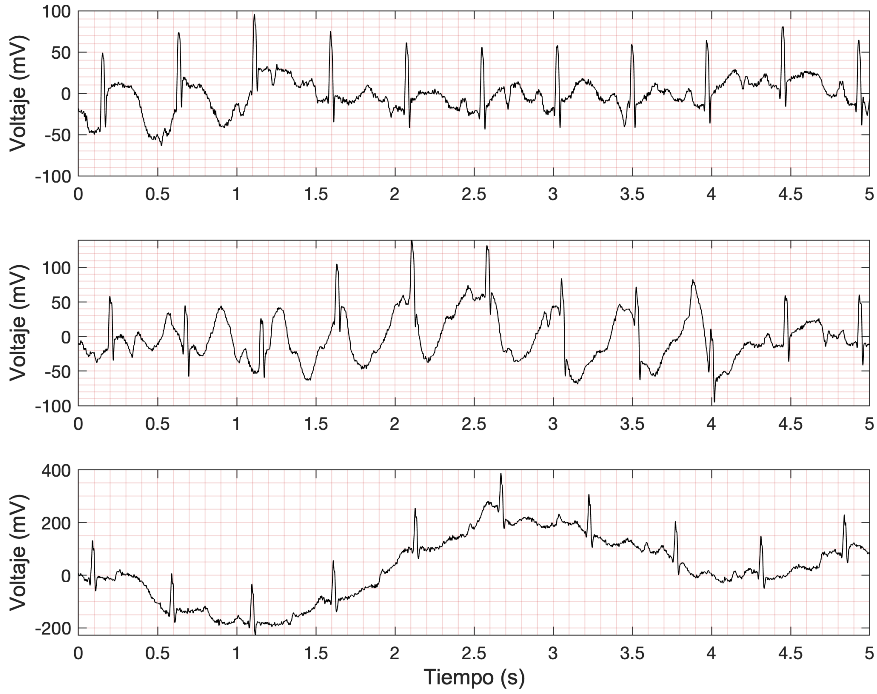



Figura 2.5: Primeros 5 segundos de los tres electrocardiogramas fetales, en orden descendente.

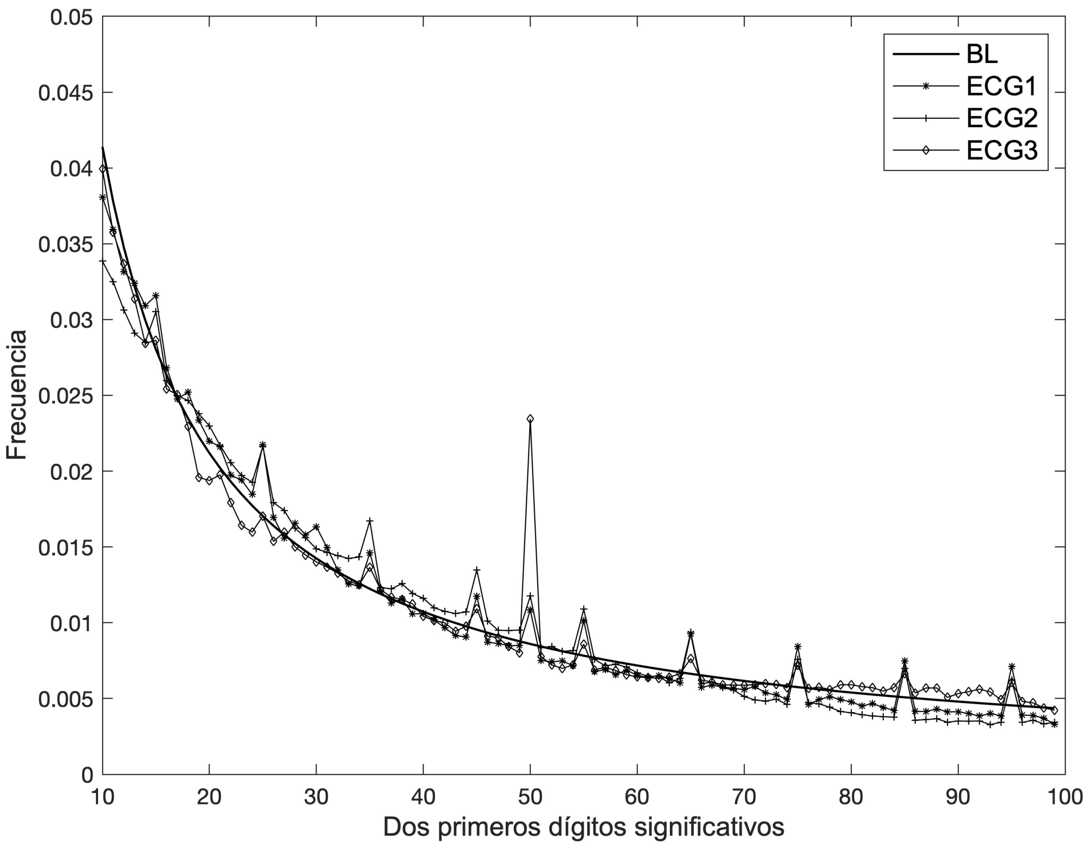
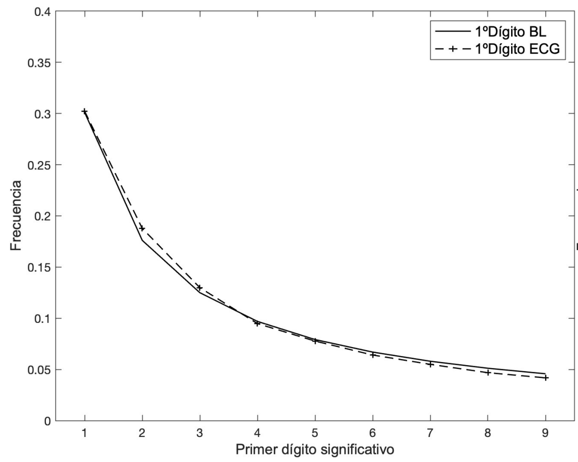
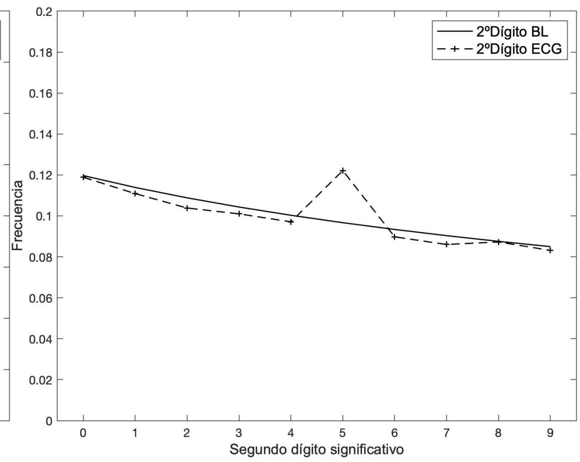


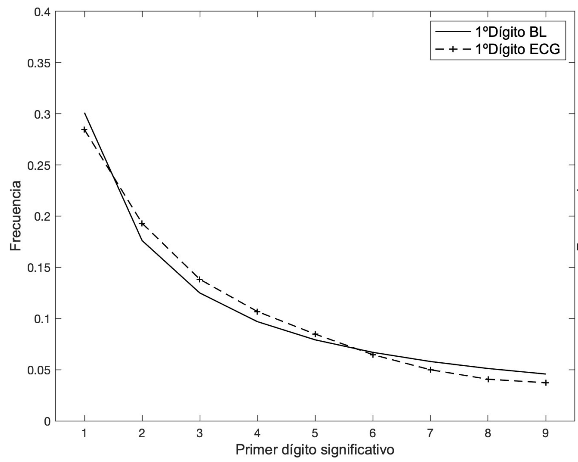

Figura 2.6: Distribuciones empíricas de los dos primeros dígitos significativos de los tres electrocardiogramas fetales discretizados.



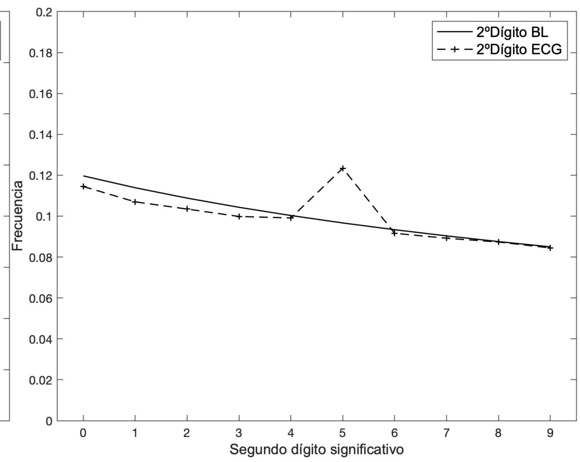

(a) Primer electrocardiograma fetal



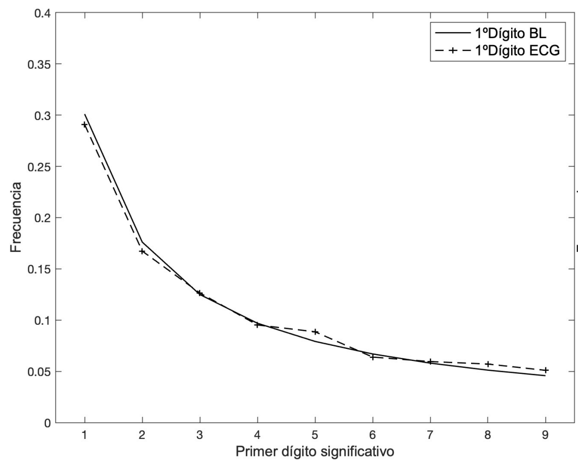

(b) Primer electrocardiograma fetal



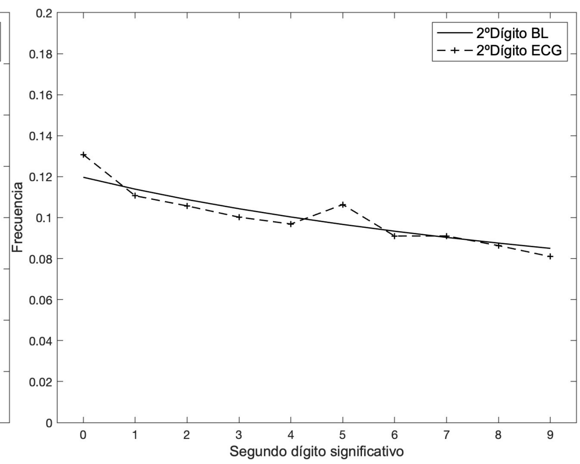

(c) Segundo electrocardiograma fetal




(d) Segundo electrocardiograma fetal




(e) Tercer electrocardiograma fetal




(f) Tercer electrocardiograma fetal

Figura 2.7: En la columna de la izquierda, la comparación de la distribución empírica del primer dígito significativo de los tres electrocardiogramas fetales con la Ley de Benford. En la columna de la derecha, la comparación de la distribución empírica del segundo dígito significativo de los tres electrocardiogramas fetales con la Ley de Benford.

En procesos relativos a las ciencias naturales, y en particular en $\xi_3$, se espera la aparición de la LB porque es razonable que su distribución de dígitos significativos sea invariante por escala, es decir, no es esperable que haya una métrica intrínseca característica de este tipo de procesos.

La figura 2.6 muestra la distribución empírica conjunta del primer y segundo dígitos significativos de los electrocardiogramas fetales, en relación a la distribución esperada, la LB. Se observa que la distribución empírica de los tres electrocardiogramas fetales es prácticamente idéntica, próxima a la LB y con pequeños saltos en los mismos puntos. La figura 2.7 muestra las distribuciones empíricas marginales del primer y segundo dígitos significativos, en relación con la LB, la esperada. Se aprecia una buena adecuación por parte de la distribución del primer dígito significativo a la LB y salvo una frecuencia un poco mayor de la esperada en el 5, la distribución del segundo dígito significativo confirma nuestra premisa: la distribución de dígitos significativos de $\xi_3$ parece ajustarse a la LB.

Desde el finales del siglo XIX es sabido que las deceleraciones en el ritmo cardíaco de un feto están asociadas con el sufrimiento fetal. Esta es la razón por la que se siguen investigando métodos para medir el ritmo cardíaco fetal; en particular, métodos no invasivos. Entre los procedimientos no invasivos destaca el que mide el ritmo cardíaco fetal junto con el ritmo cardíaco materno, a través de electrodos ubicados en el abdomen de la madre. El inconveniente de utilizar este método no invasivo reside en separar el electrocardiograma materno del fetal. Es en este aspecto donde entran en juego el desarrollo de algoritmos para obtener el electrocardiograma fetal con la mayor precisión posible; y el hecho de que los electrocardiogramas sigan la LB puede ser de utilidad, por ejemplo, para validar la calidad de los resultados obtenidos.

## 2.2. El caso de Grecia

Tras haberse comprobado empíricamente cómo la LB aparece en numerosas y variadas tablas de datos, en esta sección pondremos el foco en sacarle provecho, ¿en qué sentido?, en comprobar la calidad de ciertos datos macroeconómicos de 25 países pertenecientes a la Unión Europea, de los que disponemos datos, relevantes para conocer su posición económica. En particular, dado que Grecia quedó en evidencia ante la Comisión Europea el 12 de Enero de 2010 por falsificar sus finanzas públicas, pretendemos valorar si hubiera sido posible detectar estas actividades fraudulentas utilizando la LB como única herramienta estadística.

El pretender aplicar la LB al caso de Grecia fue tras leer [24]. El enfoque será diferente al propuesto en este artículo; a nivel económico, en [24] se consideran datos agregados correspondientes a 5 categorías relacionadas con el déficit público, la deuda pública y el producto nacional bruto. En esta sección se analizará únicamente el proceso generador de precios en transacciones internacionales de productos. *Grosso modo*, en [24] se analizan datos relativos al déficit fiscal y deuda mientras que aquí nos centraremos en los datos relativos al déficit comercial. Además, en [24] analizan datos comprendidos entre 1999 y 2009, mientras que aquí nos centramos en un año concreto: 2004, pudiéndose realizar este análisis para todos los años, por separado, comprendidos entre 1999 y 2009, lo cual aporta información acerca de la calidad de los datos macroeconómicos a lo largo del tiempo. En [24] se menciona que las autoridades griegas revisaron repetidamente las cifras de deuda entre 2005 y 2008, esto ha motivado escoger el año previo a este comienzo: 2004.

El objetivo en [24] y en esta sección es el mismo: determinar el grado de anomalía numérica de los datos macroeconómicos que Grecia aportó a la Unión Europea a principios de este siglo. La ventaja de analizar los precios de las transacciones de productos es doble: disponer de muestras de mayor tamaño en periodos de tiempo localizados, y poder deducir bajo qué condiciones, la manipulación de los datos conduce a que la distribución de dígitos significativos de los mismos no se ajuste a la LB. Esto último será ampliamente detallado.

En nuestro conjunto de datos disponemos de información sobre 50,049 transacciones, más específicamente, sobre 50,049 importaciones de 379 productos de 25 países pertenecientes a la Unión Europea, relativas al año 2004. Se han considerado productos con un número mínimo de 50 transacciones. Con información nos referimos a la cantidad de unidades de ese bien importado, medido en masa (1unidad $:=100kg$) y al precio por unidad (1unidad := 1€). Estos datos han sido obtenidos de Eurostat, la Oficina Estadística de la Unión Europea.

Denotemos por $\mathcal{M}$ al mercado formado por los 25 países y por los productos de los que disponemos datos, $\mathcal{G}=\{1,...,379\}$. Denotemos por $t_1,...,t_{25}\in\mathcal{M}$ a los países, $n_j$ al número de importaciones del país $t_j$ y $m_j$ a la cantidad de productos importados por $t_j$, $j=1,...,25$. Nuestro interés es en los procesos generadores de precios, $\xi_j$, $j=1,...,25$, definidos en el espacio producto

$$\xi_j=U_j\cdot Q_j \quad j=1,...,25, \tag{2.1}$$

donde $U$ representa el precio unidad y $Q$ la cantidad, siendo ambas variables aleatorias no negativas. El interés es determinar si $\xi_j$ ha sido manipulado o es 'natural', esto es, si el país $t_j$ ha falsificado precios de sus importaciones o no; para ello, se va a considerar un modelo de contaminación para cada país en $\sigma\left(D_1\circ\xi_j\right)$, es decir, analizaremos solamente el primer dígito significativo de $\xi_j$, por ser el que menos afectado está por el redondeo y por ser el que tiene un mayor impacto en el resultado. El modelo es el siguiente:

Para $j=1,...,25$, la forma general de nuestro modelo de contaminación es

$$\pi_j(d_1)=\left(1-\tau_j\right)\psi_j(d_1)+\tau_j\theta_j(d_1) \quad d_1=1,...,9, \tag{2.2}$$

donde $\pi_j$ es la distribución de $D_1\circ\xi_j$, $\psi_j$ es la distribución de $D_1\circ\xi_j$ en la ausencia de fraude y $\theta_j$ es la distribución de $D_1\circ\xi_j$ con fraude y $\tau_j\in[0,1]$ es la probabilidad de fraude de $t_j$; por tanto, el país $t_j$ es sospechoso de fraude si $\tau_j>0$. Reformulando nuestro problema, pretendemos decidir entre la hipótesis nula $H_0^j:\tau_j=0$ y la hipótesis alternativa $H_a^j:\tau_j>0$ basándonos en los datos que disponemos. Esta decisión la basaremos en el estadístico $\chi_j^2$

$$V_j:=\sum_{k=1}^{9}\frac{\left(N_j(d_1)-n_j\pi_j(d_1)\right)^2}{n_j\pi_j(d_1)} \quad j=1,...,9, \tag{2.3}$$

donde $N_j(d_1)$ es la frecuencia de $d_1$ en $D_1\circ\xi_j(\omega)$, el conjunto de los primeros dígitos significativos de los datos que disponemos del país $t_j$.

Es conocido que $V_j\xrightarrow{d}\chi_8^2$ cuando $n_j\to\infty$, bajo $H_0^j$, y que como consecuencia el test que rechaza $H_0^j$ si $V_j>\chi_8^2(1-\alpha)$ es un test de nivel aproximadamente $\alpha$. Ahora bien, en el cálculo de $V_j$ bajo $H_0^j$ necesitamos conocer $\psi_j$, la distribución del primer dígito significativo de $\xi_j$ sin fraude. A estas alturas del trabajo la respuesta parece inmediata: la LB, pero hay que proceder con cautela. El principal motivo por el que pensar en la LB como distribución 'natural' en $D_1 \circ \xi_j$ es por la intuición que nos da la el **Teorema 1.53**: los precios de las importaciones dependen del país de origen, del tipo de producto que se está importando, de las cantidades de estos productos importadas, e incluso del propio país que está importando, además, estos procesos no mantienen dependencia; es decir, con la notación del **Teorema 1.53** es sensato pensar en $\xi_j(\omega)$ como una $m_j$-muestra $\xi$-aleatoria de tamaño $n_j$, siendo $\xi_j$ una m.p.a invariante. Esto motiva considerar la siguiente restricción: $\psi_j = \psi_{(m_j,n_j)}$, es decir, que la distribución de $D_1 \circ \xi_j$ en la ausencia de fraude, por lo expuesto, la LB, depende tan solo del número de productos importados y de la cantidad de importaciones. Con esta consideración, el modelo de contaminación con el que trabajaremos a partir de ahora será el siguiente

$$\pi_j(d_1) = (1-\tau_j)\,\psi_{(m_j,n_j)}(d_1) + \tau_j\theta_j(d_1) \quad d_1 = 1,\ldots,9, \tag{2.4}$$

para $j = 1,\ldots,25$.

Recalcar que aunque (2.4) sea un modelo aproximado, es coherente con los elementos económicos que sugieren a la LB como distribución de dígitos significativos 'natural'. La razón detrás de esta restricción es poder deducir en qué muestras $\xi_j(\omega)$ se espera la aparición de la LB. El **Teorema 1.53** es un resultado asintótico del que a priori no se conoce su velocidad de convergencia; con esta restricción, deducir en qué muestras $\xi_j(\omega)$ se espera la aparición de la LB se traduce en examinar bajo qué combinaciones de $(m_j,n_j)$ la velocidad de convergencia es 'suficiente' para que el modelo (2.4) con $\psi_{(m_j,n_j)}$ siguiendo la LB sea apropiado. Esto implica la posibilidad de no poder analizar, con nuestro modelo basado en la LB y con los datos que disponemos, algún país $t_j$, pero también implica que las conclusiones que saquemos sean válidas para los países que sí se puedan analizar. Con este comentario se pretende recordar que hay que proceder con cautela al utilizar la LB como herramienta de detección de fraude, concretamente, verificar no solo que el proceso generador *per se* se ajuste a la LB, sino también las muestras que se tienen.

En resumen, el objetivo es examinar bajo qué parámetros $(m_j,n_j)$ la velocidad de convergencia es suficiente para que la distribución de $\psi_{(m_j,n_j)}$ se ajuste a la LB. De ser así, necesariamente $V_j \xrightarrow{d} \chi^2_8$ cuando $n_j \to \infty$ y por tanto necesariamente el suceso $V_j > \chi^2_8(1-\alpha)$ ocurre con una proporción de aproximadamente $\alpha$. Suponiendo que nos es factible generar muestras de $\psi_{(m_j,n_j)}$, como a fin de cuentas nuestra decisión se basa en el estadístico $\chi^2_8$, mediremos la discrepancia entre la distribución de $\psi_{(m_j,n_j)}$ y la LB en tanto en cuanto el estimador $\widehat{\alpha}_j$ dado por

$$\widehat{\alpha}_j = \frac{1}{T}\sum_{k=1}^{T} \chi_{[\zeta_{1-\alpha},+\infty]}\left(V_j\left(\psi_{(m_j,n_j)}(\omega_k)\right)\right) \tag{2.5}$$

aproxime bien a $\alpha$, con $\alpha$ en el rango usual de niveles de significación, siendo $\zeta_\gamma$ el $\gamma$ cuantil de $\chi^2_8$ y $T$ la cantidad de réplicas Monte Carlo, cuya generación estamos suponiendo factible.

Así, hemos reducido el problema de examinar bajo qué combinaciones de $(m_j,n_j)$ la velocidad de convergencia es suficiente para que el modelo (2.4) con $\psi_{(m_j,n_j)}$ siguiendo la LB sea adecuado, a obtener muestras de $\psi_{(m_j,n_j)}$. Para esto, recurriremos al algoritmo expuesto en [6]. El algoritmo es el siguiente

Tabla 2.3: Simulación de precios no contaminados

**Requiere:** $\mathscr{G}$ – el conjunto de productos en el mercado; $m_j$ – cantidad de productos importados por el país $t_j$; $n_j$ – número de importaciones del país $t_j$; $Q_k$ – conjunto de cantidades importadas del producto $k$ en todo el mercado; $U_k$ – conjunto de precios unidad de las importaciones del producto $k$ en todo el mercado

**Devuelve:** $X_j$ – vector con $n_j$ transacciones de $m_j$ productos

1: $X_j \leftarrow$ vector vacío

2: Seleccionar aleatoriamente $m_j$ elementos $g_{j,1},\dots,g_{j,m_j}$ de $G$ sin reemplazamiento, siendo la probabilidad de selección proporcional al número de transacciones que involucran a $g_{j,k}$ en todo el mercado

3: Seleccionar $m_j$ enteros positivos $n_{j,1},\dots,n_{j,m_j}$ aleatoriamente tal que $\sum_{k=1}^{m_j} n_{j,k} = n_j$

4: **Para** $k = 1$ hasta $m_j$ **Hacer**

5: Seleccionar $n_{j,k}$ precios unitarios $u_{j,k,1},\dots,u_{j,k,n_{j,k}}$ aleatoriamente del conjunto $U_k$

6: Seleccionar $n_{j,k}$ cantidades $q_{j,k,1},\dots,q_{j,k,n_{j,k}}$ aleatoriamente del conjunto $Q_k$

7: Calcular el vector $x_k^j = (x_{k,1}^j,\dots,x_{k,n_{j,k}}^j)$ donde $x_{k,i}^j = u_{j,k,i}q_{j,k,i}$ para $i = 1,\dots,n_{j,k}$

8: Añadir $x_k^j$ a $X_j$

9: **Fin Para**

10: **Devolver** $X_j$

La idea subyacente a este algoritmo es simple: dado un par $(m_j, n_j)$ correspondiente a un país $t_j$, se muestrea con reemplazamiento de $m_j \leq |\mathscr{G}|$ espacios producto (2.6), un total de $n_j$ veces.

$$\mathscr{U}_k \times \mathscr{Q}_k \quad k = 1,\dots,|\mathscr{G}|, \tag{2.6}$$

con $\mathscr{U}_k = \{u_1, u_2, \dots, u_{n_k}\}$ y $\mathscr{Q}_k = \{q_1, u_2, \dots, q_{n_k}\}$ representando los conjuntos de precios unidad y cantidades importadas, respectivamente, del producto $k$, siendo $n_k$ el número total de transacciones de ese producto en todo el mercado.

Desde un punto de vista económico, muestrear de $\mathscr{U}_k \times \mathscr{Q}_k$ implica asumir que no hay una relación sistemática entre precios y cantidades vinculadas al producto $k$. La cuestión es si esta suposición es realista; es decir, si las variables aleatorias $U$ y $Q$, definidas en (2.1) son independientes o no. Hay cuatro tipos de estructuras de mercado: competencia perfecta, competencia monopolística, oligopolio y monopolio; en consecuencia, $\forall k \in \{1, 2, \dots, |\mathscr{G}|\}$ existe una estructura de mercado tal que $k$ pertenece a esta estructura de mercado. Denotemos por $U_k$ y $Q_k$ a las variables aleatorias generadoras de precios unidad y cantidades en las importaciones del producto $k$, respectivamente. Si $k$ pertenece a la estructura de mercado: competencia perfecta, dado que el precio está determinado únicamente por la oferta y la demanda; en otras palabras, ningún proveedor puede influenciar en el precio de los productos, hay independencia entre $U_k$ y $Q_k$. Si $k$ pertenece a una de las otras tres estructuras de mercado, donde sí existe en mayor o menor grado poder de mercado, $U_k$ y $Q_k$ sí que pueden dependen del poder relativo de los proveedores y del resultado de los procesos de negociación entre ellos. Esto implica que la misma cantidad pueda ser comprada por el mismo país en diferentes transacciones a diferentes precios, estableciendo de nuevo independencia (aproximada) entre $U_k$ y $Q_k$.

Una vez presentado y explicado el procedimiento, es hora de aplicarlo a los datos que tenemos y sacar conclusiones; los resultados están recogidos en la tabla 2.4. En esta tabla se ha incluido para cada país: el número de productos importados, la cantidad de importaciones, el cociente entre productos importados y cantidad de importaciones; el estimador dado por (2.5), con $\alpha = 0.01$, basado en $T = 10{,}000$ réplicas Monte Carlo; el valor del estadístico $\chi^2$ calculado como se indicó en (2.3), con $\psi_j$ siguiendo la LB, y finalmente el p-valor de este problema de contraste. En la **Sección C.2** del **Apéndice C** se encuentra el código utilizado para generar los valores de la tabla 2.4.

Un rasgo destacado de los valores $\widehat{\alpha}_j$ presentados en la tabla 2.4 es que varían dependiendo de la relación $m_j/n_j$, haciendo patentes a $m_j$ y a $n_j$ en la determinación de la velocidad de convergencia hacia la LB en procesos generadores de precios ausentes de manipulación, y con ello apoyando la coherencia del modelo de contaminación restringido (2.4). En general, el estimador $\widehat{\alpha}_j$ mejora en tanto en cuanto $m_j/n_j$ crece; en otras palabras, en general la velocidad de convergencia es mayor cuando las transacciones involucran a una gran cantidad relativa de productos.

Analizaremos en conjunto los parámetros: el estimador $\widehat{\alpha}$ y el p-valor de contraste, para comprobar la calidad de los datos presentados por los 25 países considerados acerca de sus importaciones en el año 2004. Recordemos el significado de $\widehat{\alpha}_j$: estimar el valor 0.01$=\alpha$ vía el procedimiento (2.5). La cuestión ahora es determinar el umbral máximo para el cual vamos a considerar a $\widehat{\alpha}_j$ un buen estimador de $\alpha$=0.01. A la vista de la tabla 2.4 parece sensato considerar los valores mayores que 0.013 malas aproximaciones de $\alpha$, ya que a partir de este punto los valores del estadístico $\chi^2$ son muy elevados en todos los casos. De fijarnos solamente en el p-valor, rechazaríamos con probabilidad 0.99 la hipótesis de no contaminación en la distribución del primer dígito significativo de los países: Austria, Alemania, Dinamarca, Reino Unido, Grecia, Países Bajos y Eslovenia. Sin embargo, el modelo (2.4) con $\psi_{(m_j,n_j)}$ siguiendo la LB no es apropiado para los países: Alemania ($\widehat{\alpha}_6 = 0.0134$), Dinamarca ($\widehat{\alpha}_7 = 0.0133$), Reino Unido ($\widehat{\alpha}_{12} = 0.0142$) y Países bajos ( $\widehat{\alpha}_{20} = 0.0146$). En cambio, los estimadores $\widehat{\alpha}_1 = 0.0115$, $\widehat{\alpha}_{13} = 0.0108$ y $\widehat{\alpha}_{24} = 0.0109$, correspondientes a Austria, Grecia y Eslovenia respectivamente, pueden considerarse suficientemente próximos a 0.01 para dar por válido el modelo (2.4) con $\psi_{(m_j,n_j)}$ siguiendo la LB. En definitiva, nuestro procedimiento señala una elevada contaminación en la distribución del primer dígito significativo de los precios de las importaciones de Austria, Grecia y Eslovenia en el año 2004. Reiterar que el uso de la LB para detectar anomalías intencionadas en los datos, en este caso para detectar manipulaciones en datos macroeconómicos, ha de verse como otra herramienta más para medir la razonabilidad de los datos, y no como un método infalible.

| **j** | $\mathbf{t_j}$ | $\mathbf{m_j}$ | $\mathbf{n_j}$ | $\mathbf{m_j/n_j}$ | $\widehat{\alpha}_j$ | $\mathbf{V_j}$ | **p-valor** |
|---|---|---|---|---|---|---|---|
| 1 | Austria | 304 | 2243 | 0.1355 | 0.0115 | 24.6278 | 0.00179 |
| 2 | Bélgica | 375 | 3946 | 0.0950 | 0.0125 | 13.4473 | 0.09735 |
| 3 | Bulgaria | 71 | 472 | 0.1504 | 0.0117 | 10.2232 | 0.24970 |
| 4 | Chipre | 102 | 439 | 0.2323 | 0.0108 | 13.2076 | 0.10490 |
| 5 | República Checa | 279 | 1687 | 0.1654 | 0.0112 | 4.3029 | 0.82881 |
| 6 | Alemania | 379 | 4116 | 0.0921 | 0.0134 | 38.4049 | $6.3361 \cdot 10^{-6}$ |
| 7 | Dinamarca | 347 | 3335 | 0.1040 | 0.0133 | 40.7667 | $2.3047 \cdot 10^{-6}$ |
| 8 | Estonia | 169 | 898 | 0.1882 | 0.0112 | 18.5464 | 0.01748 |
| 9 | España | 374 | 4015 | 0.0932 | 0.0129 | 11.3131 | 0.18458 |
| 10 | Finlandia | 189 | 1172 | 0.1613 | 0.0115 | 15.1732 | 0.05586 |
| 11 | Francia | 378 | 4326 | 0.0874 | 0.0134 | 10.3933 | 0.23849 |
| 12 | Reino Unido | 360 | 3584 | 0.1004 | 0.0142 | 33.3776 | $5.2648 \cdot 10^{-5}$ |
| 13 | Grecia | 231 | 1386 | 0.1667 | 0.0108 | 23.3950 | 0.00289 |
| 14 | Croacia | 97 | 497 | 0.1952 | 0.0114 | 14.0787 | 0.07973 |
| 15 | Hungría | 255 | 1869 | 0.1364 | 0.0118 | 12.4631 | 0.13170 |
| 16 | Irlanda | 294 | 2434 | 0.1208 | 0.0104 | 9.9830 | 0.26622 |
| 17 | Lituania | 193 | 1044 | 0.1849 | 0.0109 | 17.0736 | 0.02935 |
| 18 | Luxemburgo | 249 | 1569 | 0.1587 | 0.0118 | 9.8970 | 0.27232 |
| 19 | Malta | 68 | 150 | 0.4533 | 0.0118 | 7.1038 | 0.52547 |
| 20 | Países Bajos | 378 | 4251 | 0.0889 | 0.0146 | 23.2104 | 0.00310 |
| 21 | Polonia | 298 | 2148 | 0.1387 | 0.0108 | 10.5636 | 0.22767 |
| 22 | Rumanía | 80 | 492 | 0.1626 | 0.0122 | 16.4306 | 0.03661 |
| 23 | Suecia | 277 | 1961 | 0.1413 | 0.0116 | 10.5390 | 0.22921 |
| 24 | Eslovenia | 166 | 927 | 0.1791 | 0.0109 | 19.9908 | 0.01037 |
| 25 | Eslovaquia | 194 | 1088 | 0.1783 | 0.0111 | 16.0920 | 0.04108 |

Tabla 2.4: Análisis del primer dígito significativo de los precios de las importaciones de 25 países de la Unión Europea relativas al año 2004.

# Apéndice A

# Coeficientes de Fourier

**Definición A.1.** *Sea $P$ una probabilidad en $\left([0,1)\,,\mathscr{B}_{[0,1)}\right)$. Los coeficientes de Fourier de $P$ son:*

$$\widehat{P}(m) = \int_{0}^{\to 1} e^{-2\pi i m s} dP(s) \quad m = 0, \pm 1, \pm 2, ...$$

*La relación entre $P$ y sus coeficientes de Fourier se expresa de manera formal por:*

$$dP(s) \sim \sum_{m=-\infty}^{\infty} \widehat{P}(m)\, e^{2\pi i m s} \quad s \in [0,1)\,. \tag{A.1}$$

*Al término de la derecha de (A.1), se le llama serie de Fourier de $P$.*

**Ejemplo A.2.** *Coeficientes de Fourier de $\lambda_{0,1}$.*

- $m = 0$

$$\widehat{\lambda_{0,1}}\,(0) = \int_{0}^{\to 1} e^{-2\pi i 0 s} ds = 1.$$

- $m \neq 0$

$$\widehat{\lambda_{0,1}}\,(m) = \int_{0}^{\to 1} e^{-2\pi i m s} ds = -\frac{e^{-2\pi i m} - 1}{2\pi i m} = -\frac{\cos(2\pi m) - i \operatorname{sen}(2\pi m) - 1}{2\pi i m} = 0.$$

El objetivo de este apéndice es demostrar tres propiedades de los coeficientes de Fourier, recogidas en el **Teorema** **A.3**. La primera de estas propiedades, la que garantiza la unicidad, será clave en las dos demostraciones del siguiente apéndice.

**Teorema A.3** (Propiedades de los coeficientes de Fourier)**.**

I *Los coeficientes de Fourier determinan la probabilidad; es decir, si $P$ y $Q$ son dos probabilidades en $\left([0,1),\mathscr{B}_{[0,1)}\right)$ tal que se cumple: $\widehat{P}(m)=\widehat{Q}(m)\,\forall m\in\mathbb{Z}$, entonces $P\overset{\mathrm{d}}{=}Q$.*

II *Una sucesión de probabilidades $(P_n)_{n=1}^{\infty}$ en $\left([0,1),\mathscr{B}_{[0,1)}\right)$ converge en distribución hacia una probabilidad $P$ si, y solo si, $\lim_{n\to\infty}\widehat{P_n}(m)=\widehat{P}(m)\,\forall m\in\mathbb{Z}$.*

III *Sean $X$ e $Y$ variables aleatorias independientes. Entonces:*

$$\widehat{P_{(X+Y)\bmod 1}}(m)=\widehat{P_{X\bmod 1}}(m)\cdot\widehat{P_{Y\bmod 1}}(m)\quad\forall m\in\mathbb{Z}.$$

DEMOSTRACIÓN. Sea $\varphi$ una probabilidad en $\left([0,1),\mathscr{B}_{[0,1)}\right)$. Sea $m\in\mathbb{N}$, denotemos por $s_m^{\varphi}(t):[0,1)\longrightarrow\mathbb{R}$ a la suma parcial $m$-ésima de la serie de Fourier de $\varphi$ en el punto $t$

$$s_m^{\varphi}(t)=\sum_{k=-m}^{m}\widehat{\varphi}(k)\,e^{2\pi ikt}\quad t\in[0,1). \tag{A.2}$$

Denotemos por $\sigma_m^{\varphi}(t):[0,1)\longrightarrow\mathbb{R}$ a la media aritmética de las primeras $m$ sumas parciales de la serie Fourier de $\varphi$ en el punto $t$

$$\sigma_m^{\varphi}(t)=\frac{1}{m}\sum_{k=0}^{m-1}s_k^{\varphi}(t)\quad t\in[0,1). \tag{A.3}$$

Teniendo en cuenta (A.2)

$$\sigma_m^{\varphi}(t)=\frac{1}{m}\sum_{k=0}^{m-1}\sum_{l=-k}^{k}\widehat{\varphi}(l)\,e^{2\pi ilt}=\frac{1}{m}\sum_{k=0}^{m-1}\sum_{l=-k}^{k}\left(\int_0^{\to 1}e^{-2\pi ils}d\varphi(s)\right)e^{2\pi ilt}$$
$$=\frac{1}{m}\int_0^{\to 1}\left(\sum_{k=0}^{m-1}\sum_{l=-k}^{k}e^{2\pi il(t-s)}\right)d\varphi(s).$$

Sea $\theta=e^{2\pi i(t-s)}$. Desarrollemos ahora el término que está dentro del paréntesis:

- Caso $\theta\neq 1\iff t\neq s$

$$\sum_{k=0}^{m-1}\sum_{l=-k}^{k}\theta^l=\sum_{k=0}^{m-1}\left(\sum_{l=-k}^{0}\theta^l+\sum_{l=1}^{k}\theta^l\right)=\sum_{k=0}^{m-1}\left(\frac{1-\theta^{-l-1}}{1-\frac{1}{\theta}}+\frac{\theta-\theta^{l+1}}{1-\theta}\right)$$
$$=\sum_{k=0}^{m-1}\frac{\theta^{-l}-\theta^{l+1}}{1-\theta}=\frac{1}{1-\theta}\left(\sum_{k=0}^{m-1}\theta^{-l}-\sum_{k=0}^{m-1}\theta^{l+1}\right)$$
$$=\frac{1}{1-\theta}\left(\frac{1-\theta^{-m}}{1-\frac{1}{\theta}}-\frac{\theta-\theta^{m+1}}{1-\theta}\right)=\frac{1}{1-\theta}\left(\frac{-1+\theta^{-m}}{\frac{1-\theta}{\theta}}-\frac{1-\theta^m}{\frac{1-\theta}{\theta}}\right)$$
$$=\frac{-2+\theta^{-m}+\theta^m}{\left(\theta^{-\frac{1}{2}}-\theta^{\frac{1}{2}}\right)^2}=\frac{\left(\theta^{\frac{m}{2}}-\theta^{-\frac{m}{2}}\right)^2}{\left(\theta^{\frac{1}{2}}-\theta^{-\frac{1}{2}}\right)^2}=\left(\frac{\frac{e^{\pi i(t-s)m}-e^{-\pi i(t-s)m}}{2i}}{\frac{e^{\pi i(t-s)}-e^{-\pi i(t-s)}}{2i}}\right)^2$$
$$=\left(\frac{\operatorname{sen}(\pi(t-s)m)}{\operatorname{sen}(\pi(t-s))}\right)^2.$$

- Caso $\theta = 1 \iff t = s$

$$\sum_{k=0}^{m-1} \sum_{l=-k}^{k} \theta^l = \sum_{k=0}^{m-1} 2k+1 = m + m(m-1) = m^2.$$

Dado que para todo $t \in [0,1)$ fijo

$$\lim_{s \to t} \left( \frac{\operatorname{sen}(\pi(t-s)m)}{\operatorname{sen}(\pi(t-s))} \right)^2 = m^2,$$

hemos llegado a que $\forall t \in [0,1)$:

$$\sigma_m^{\varphi}(t) = \frac{1}{m} \int_0^{\to 1} \left( \frac{\operatorname{sen}(\pi(t-s)m)}{\operatorname{sen}(\pi(t-s))} \right)^2 d\varphi(s).$$

Considerando $\lambda_{0,1}$, por el **Ejemplo A.2** sabemos que

$$\begin{aligned} \widehat{\lambda_{0,1}}(m) &= 1 \quad \text{si } m = 0 \\ \widehat{\lambda_{0,1}}(m) &= 0 \quad \text{si } m \neq 0, \end{aligned}$$

esto implica que

$$\begin{aligned} s_m^{\lambda_{0,1}}(t) = 1 \quad \forall m \in \mathbb{N} \\ &\Longrightarrow \sigma_m^{\lambda_{0,1}}(t) = 1 \quad \forall m \in \mathbb{N} \\ &\Longrightarrow \frac{1}{m} \int_0^{\to 1} \left( \frac{\operatorname{sen}(\pi(t-s)m)}{\operatorname{sen}(\pi(t-s))} \right)^2 ds = 1 \quad \forall m \in \mathbb{N} \\ &\Longrightarrow \frac{1}{m} \int_0^{1} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx = 1 \quad \forall m \in \mathbb{N}. \end{aligned} \tag{A.4}$$

En la última implicación se ha considerado el cambio de variable $x = s - t$, se ha tenido en cuenta que las funciones $\operatorname{sen}^2(\pi x m)$ y $\operatorname{sen}^2(\pi x)$ son 1-periódicas y también que $\{1\}$ es un conjunto de medida Lebesgue nula.

Sean $a, b \in (0,1)$ tales que $0 < a < b < 1$, con $\varphi(\{a\}) = \varphi(\{b\}) = 0$. Consideramos el espacio medible $\Upsilon = \left([a,b] \times [0,1), \mathscr{B}_{[a,b]\times[0,1)}\right)$ y las vectores aleatorios con llegada a este:

- $\left(Id_{[a,b]}, Id_{[0,1)}\right) : \left([a,b] \times [0,1), \mathscr{B}_{[a,b]\times[0,1)}, \lambda_{a,b} \times \varphi\right) \to \Upsilon$.
- $\left(Id_{[a,b]}, F^{-1}\right) : \left([a,b] \times (0,1), \mathscr{B}_{[a,b]\times(0,1)}, \lambda_{a,b} \times \lambda_{0,1}\right) \to \Upsilon$.

Siendo $F^{-1}$ la función cuantil asociada a la función de distribución de $\varphi$.

Consideramos también las aplicaciones $g_m(t,s) : \Upsilon \to (\mathbb{R}, \mathscr{B})$, dadas por

$$g_m(t,s) = \frac{1}{m} \left( \frac{\operatorname{sen}(\pi(t-s)m)}{\operatorname{sen}(\pi(t-s))} \right)^2 \quad m \in \mathbb{N}.$$

Como $g_m$ es continua $\forall m \in \mathbb{N}$ es medible. Ahora, como

$$\left(Id_{[a,b]}, Id_{[0,1)}\right) \overset{\mathrm{d}}{=} \left(Id_{[a,b]}, F^{-1}\right),$$

por el teorema del cambio de variable

$$\int\limits_{[a,b]\times[0,1)} g_m(t,s)\,(dt \times d\varphi(s)) = \int\limits_{[a,b]\times(0,1)} g_m\left(t, F^{-1}(x)\right) dt dx.$$

Como $g_m$ es una función medible y no negativa $\forall m \in \mathbb{N}_0$, por el Teorema de Tonelli

$$\begin{aligned}\int\limits_{[a,b]\times(0,1)} g_m\left(t, F^{-1}(x)\right) dt dx &= \int_a^b \left( \int_0^1 g_m\left(t, F^{-1}(x)\right) dx \right) dt \\ &= \int_0^1 \left( \int_a^b g_m\left(t, F^{-1}(x)\right) dt \right) dx.\end{aligned}$$

Entonces, volviendo a aplicar el teorema del cambio de variable

$$\begin{aligned}&\int_0^1 \left( \int_a^b g_m\left(t, F^{-1}(x)\right) dt \right) dx = \int_a^b \left( \int_0^1 g_m\left(t, F^{-1}(x)\right) dx \right) dt \\ &\iff \int_0^{\to 1} \left( \int_a^b g_m(t,s)\, dt \right) d\varphi(s) = \int_a^b \left( \int_0^{\to 1} g_m(t,s)\, d\varphi(s) \right) dt \\ &\iff \int_0^{\to 1} \left( \frac{1}{m} \int_a^b \left( \frac{\operatorname{sen}(\pi(t-s)m)}{\operatorname{sen}(\pi(t-s))} \right)^2 dt \right) d\varphi(s) = \int_a^b \sigma_m^{\varphi}(t)\, dt \\ &\iff \int_0^{\to 1} \left( \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) d\varphi(s) = \int_a^b \sigma_m^{\varphi}(t)\, dt \quad m \in \mathbb{N}_0 \\ &\implies \lim_{m\to\infty} \int_0^{\to 1} \left( \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) d\varphi(s) = \lim_{m\to\infty} \int_a^b \sigma_m^{\varphi}(t)\, dt. \end{aligned} \tag{A.5}$$

Dado que

$$\begin{aligned}&\left| \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right| \le \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \\ &\le \frac{1}{m} \int_{b-s-1}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx = \frac{1}{m} \int_0^1 \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx = 1,\end{aligned}$$

donde en la segunda desigualdad se ha tenido en cuenta que $b-a<1 \iff a-s > b-s-1$, y en la última igualdad, la 1-periodicidad del integrando junto con (A.4); se dan las condiciones para aplicar el teorema de convergencia dominada. Por tanto (A.5) es equivalente a:

$$\int_{0}^{\to 1} \lim_{m\to\infty} \left( \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) d\varphi(s) = \lim_{m\to\infty} \int_{a}^{b} \sigma_m^{\varphi}(t)\, dt. \tag{A.6}$$

En el estudio del término a integrar respecto de $\varphi$ a la izquierda de (A.6), distinguimos los casos:

- $s \in [0,a)$

  $s \in [0,a) \Longrightarrow b-s > a-s > 0$. Entonces

$$\begin{aligned} &\int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \leq \int_{a-s}^{b-s} \frac{dx}{\operatorname{sen}^2(\pi x)} \leq \int_{a-s}^{b-s} \frac{dx}{\operatorname{sen}^2(\pi(a-s))} \\ &= \frac{1}{\operatorname{sen}^2(\pi(a-s))}(b-a) < \infty \quad \forall m \in \mathbb{N} \\ &\Longrightarrow \lim_{m\to\infty} \left( \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) = 0. \end{aligned}$$

- $s \in [b,1)$

  $s \in [b,1) \Longrightarrow 0 > b-s > a-s$. Razonando como en el caso anterior

$$\lim_{m\to\infty} \left( \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) = \lim_{m\to\infty} \left( -\frac{1}{m} \int_{b-s}^{a-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) = 0.$$

- $s \in \{a,b\}$

  En este caso no nos importa el valor del límite ya que estamos suponiendo que $\varphi(\{a\}) = \varphi(\{b\}) = 0$.

- $s \in (a,b)$

  $s \in (a,b) \Longrightarrow a-s < 0 < b-s$. Entonces

$$\begin{aligned} 1 &= \frac{1}{m} \int_{0}^{1} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx = \frac{1}{m} \int_{b-s-1}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \\ &= \frac{1}{m} \int_{b-s-1}^{a-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx + \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx. \end{aligned}$$

  Tomando límite $m \to \infty$

$$1 = \lim_{m\to\infty} \left( \frac{1}{m} \int_{b-s-1}^{a-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right) + \lim_{m\to\infty} \left( \frac{1}{m} \int_{a-s}^{b-s} \left( \frac{\operatorname{sen}(\pi x m)}{\operatorname{sen}(\pi x)} \right)^2 dx \right).$$

Ahora, como $b-s-1<a-s<0$, razonando como en el caso segundo

$$\lim_{m\to\infty}\left(\frac{1}{m}\int\limits_{b-s-1}^{a-s}\left(\frac{\operatorname{sen}(\pi xm)}{\operatorname{sen}(\pi x)}\right)^2 dx\right)=0$$
$$\Longrightarrow 1=\lim_{m\to\infty}\left(\frac{1}{m}\int\limits_{a-s}^{b-s}\left(\frac{\operatorname{sen}(\pi xm)}{\operatorname{sen}(\pi x)}\right)^2 dx\right).$$

Hemos llegado por tanto a que

$$\lim_{m\to\infty}\left(\frac{1}{m}\int\limits_{a-s}^{b-s}\left(\frac{\operatorname{sen}(\pi xm)}{\operatorname{sen}(\pi x)}\right)^2 dx\right)=\chi_{(a,b)}(s)\quad s\in[0,1).$$

Y en definitiva, (A.6) es equivalente a

$$\int\limits_{0}^{\to 1}\chi_{(a,b)}(s)\,d\varphi(s) \tag{A.7}$$

$$=\varphi((a,b))=\lim_{m\to\infty}\int\limits_{a}^{b}\sigma_m^{\varphi}(t)\,dt. \tag{A.8}$$

En resumen, si $\varphi$ es una probabilidad en $\left([0,1),\mathscr{B}_{[0,1)}\right)$ y consideramos $a,b\in(0,1)$ tales que $0<a<b<1$, con $\varphi(\{a\})=\varphi(\{b\})=0$, se tiene la igualdad (A.8).

Sean entonces $P$ y $Q$ dos probabilidades en $\left([0,1),\mathscr{B}_{[0,1)}\right)$ cumpliendo que $\widehat{P}(k)=\widehat{Q}(k)\,\forall k\in\mathbb{Z}$, entonces $\sigma_m^P=\sigma_m^Q\ \ \forall m\in\mathbb{N}$.

Sea $\mathscr{C}$ el conjunto de los intervalos de la forma $(a,b)$, con $0<a<b<1$, cumpliendo además que $P(\{a\})=P(\{b\})=Q(\{a\})=Q(\{b\})=0$ (téngase en cuenta que el conjunto de puntos con probabilidad $P$ ó $Q$ positiva es a lo sumo numerable, por consiguiente $\mathscr{C}$ genera $\mathscr{B}_{[0,1)}$, siendo además una $\pi$-clase). Sea $\Lambda=\left\{S\in\mathscr{B}_{[0,1)}:P(S)=Q(S)\right\}$. Por (A.8), $\mathscr{C}\subseteq\Lambda$ ya que si $(a,b)\in\mathscr{C}$ entonces

$$P((a,b))=\lim_{m\to\infty}\int\limits_{a}^{b}\sigma_m^P(t)\,dt=\lim_{m\to\infty}\int\limits_{a}^{b}\sigma_m^Q(t)\,dt=Q((a,b)).$$

Ahora, es inmediato probar que $\Lambda$ es una $\lambda$-clase; por tanto, aplicando el Teorema $\pi$-$\lambda$ de Dynkin, $\mathscr{B}_{[0,1)}=\Lambda$. En definitiva, $P\overset{\mathrm{d}}{=}Q$. Con esto queda probada la primera propiedad de los coeficientes de Fourier.

Supongamos que la sucesión de probabilidades $(P_n)_{n=1}^{\infty}$ converge en distribución hacia $P$. Sea $k\in\mathbb{Z}$ fijo. Consideramos la sucesión $\left(\widehat{P_n}(k)\right)_{n=1}^{\infty}$. Por el Teorema Portmanteau, ya que $e^{-2\pi ims}$ es continua y acotada en $[0,1)$

$$\lim_{n\to\infty}\widehat{P_n}(k)=\lim_{n\to\infty}\int\limits_{0}^{\to 1}e^{-2\pi iks}\,dP_n(s)=\int\limits_{0}^{\to 1}e^{-2\pi iks}\,dP(s)$$
$$=\widehat{P}(k).$$

Queda probada entonces la condición necesaria de la segunda propiedad de los coeficientes de Fourier. Para probar la suficiente, consideramos la sucesión $(P_n)_{n=1}^{\infty}$ y la probabilidad $P$, cumpliendo que $\lim_{n\to\infty}\widehat{P_n}(k) = \widehat{P}(k)\ \ \forall k \in \mathbb{Z}$. Como $\forall n \in \mathbb{N}$ se tiene que $P_n([0,1)) = 1$, $(P_n)_{n=1}^{\infty}$ es ajustada. Sea $(P_{n_l})_{l=1}^{\infty}$ una subsucesión de $(P_n)_{n=1}^{\infty}$ que converge en distribución hacia una probabilidad, digamos, $\mu$. Entonces, razonando como antes

$$\lim_{l\to\infty}\widehat{P_{n_l}}(k) = \widehat{\mu}(k) \quad k \in \mathbb{Z}.$$

Por tanto, se tiene que

$$\begin{aligned}
&\lim_{n\to\infty}\widehat{P_n}(k) = \widehat{P}(k) \ \ k \in \mathbb{Z} \\
&\Longrightarrow \lim_{l\to\infty}\widehat{P_{n_l}}(k) = \widehat{P}(k) \quad k \in \mathbb{Z} \\
&\Longrightarrow \widehat{\mu}(k) = \widehat{P}(k) \quad k \in \mathbb{Z}.
\end{aligned}$$

Por la unicidad de los coeficientes de Fourier, $\mu \overset{\mathrm{d}}{=} P$. En definitiva, como cualquier subsucesión de $(P_n)_{n=1}^{\infty}$ que converge en distribución lo hace hacia $P$, concluimos que $(P_n)_{n=1}^{\infty} \xrightarrow{\mathrm{d}} P$.

Demostremos ahora la última de las tres propiedades de los coeficientes de Fourier. Sean $X$ e $Y$ variables aleatorias independientes, $k \in \mathbb{Z}$ fijo. Sea $\Phi_k(s) = e^{-2\pi iks}$

$$\begin{aligned}
&\widehat{P_{(X+Y)\text{mod }1}}(k) = \int_0^{\to 1} \Phi_k(s)dP_{(X+Y)\text{mod }1}(s) = \int_\Omega \Phi_k((X(\omega)+Y(\omega)) \text{ mod } 1)dP(\omega) \\
&= \int_\Omega \Phi_k(X(\omega)+Y(\omega))\,dP(\omega) = \int_\Omega \Phi_k(X(\omega)) \cdot \Phi_k(Y(\omega))\,dP(\omega) \\
&= \int_\Omega \Phi_k(X(\omega))\,dP(\omega) \cdot \int_\Omega \Phi_k(Y(\omega))\,dP(\omega) \\
&= \int_\Omega \Phi_k(X\text{mod } 1(\omega))\,dP(\omega) \cdot \int_\Omega \Phi_k(Y\text{mod } 1(\omega))\,dP(\omega) = \widehat{P_{X\text{mod }1}}(k) \cdot \widehat{P_{Y\text{mod }1}}(k).
\end{aligned}$$

En la segunda igualdad se ha utilizado el teorema del cambio de variable, en la tercera la 1-periodicidad de $\Phi_k$; en la cuarta, una de las propiedades de la función exponencial, en la quinta la independencia de las variables y en la sexta, otra vez la 1-periodicidad de $\Phi_k$. $\square$

# Apéndice B

# Ergodicidad

**Definición B.1.** *Sea* $(\Omega, \sigma)$ *un espacio medible. Sea* $T : \Omega \longrightarrow \Omega$ *una aplicación medible. Diremos que una probabilidad* $P$ *en* $(\Omega, \sigma)$ *es* $T$*-invariante si* $P_T \overset{\mathrm{d}}{=} P$.

El objetivo de este apéndice es demostrar qué probabilidades preservan la medida por $T(x) = (x + a) \mod 1$, $a \in \mathbb{I}$ y por $T_n(x) = (nx) \mod 1$, $n \in \mathbb{N}$, con $x \in [0, 1)$.

Estos resultados serán clave para caracterizar las distribuciones *IE* y *IB* en $(\mathbb{R}^+, \mathscr{M})$

**Teorema B.2.** *Una probabilidad* $P$ *en* $\left([0,1), \mathscr{B}_{[0,1)]}\right)$ *es* $T_n$*-invariante* $\forall n \in \mathbb{N}$*, con* $T_n(x) = (nx) \mod 1$ *si, y solo si,*

$$P \overset{\mathrm{d}}{=} q\delta_0 + (1-q)\,\lambda_{0,1} \qquad \text{para algún } q \in [0,1].$$

DEMOSTRACIÓN. La demostración se hará recurriendo a la unicidad que garantizan los coeficientes de Fourier. Denotemos por $\Phi_k(s) = e^{-2\pi iks}$

$$\widehat{P}(k) = \int_0^1 e^{-2\pi iks} dP(s) = \int_0^1 \Phi_k(s) dP(s) \quad k \in \mathbb{Z}.$$

$$\widehat{P_{T_n}}(k) = \int_0^1 \Phi_k(s) dP_{T_n}(s) = \int_0^1 \Phi_k(T_n(s)) dP(s) \tag{B.1}$$

$$= \int_0^1 \Phi_k((ns) \mod 1) dP(s) = \int_0^1 \Phi_k(ns) dP(s) \tag{B.2}$$

$$= \int_0^1 \Phi_{nk}(s) dP(s) = \widehat{P}(nk) \quad \forall n \in \mathbb{N}. \tag{B.3}$$

En (B.1) se ha aplicado el teorema del cambio de variable, (B.2) es consecuencia de la 1-periodicidad de $\Phi_k$ y (B.3) es consecuencia de las propiedades de la exponencial.

Veamos primero la condición suficiente. Sea $P \overset{\mathrm{d}}{=} q\delta_0 + (1-q)\,\lambda_{0,1}$ para algún $q \in [0,1]$. Como

$$\widehat{\delta_0}(k) = \int_0^1 e^{-2\pi ik\delta_0(s)} dP(s) = \int_{\{0\}} e^{-2\pi iks} dP(s) = 1 \quad \forall k \in \mathbb{Z}.$$

y por el **Ejemplo A.2**, $\widehat{\lambda_{0,1}}(0) = 1$ y $\widehat{\lambda_{0,1}}(k) = 0 \; \forall k \neq 0$, concluimos que

$$\begin{aligned} \widehat{P}(0) &= 1 \\ \widehat{P}(k) &= q \quad \forall k \neq 0. \end{aligned}$$

Ahora bien, $\widehat{P_{T_n}}(k) = \widehat{P}(nk) = q$, si $k \neq 0$ y $\widehat{P_{T_n}}(0) = 1$. Por ello

$$\begin{aligned} &\widehat{P_{T_n}}(k) = \widehat{P}(k) \quad \forall k \in \mathbb{Z} \\ &\Longrightarrow P \overset{\mathrm{d}}{=} P_{T_n} \Longrightarrow P \text{ es } T_n \text{ -invariante .} \end{aligned}$$

Veamos ahora la condición necesaria. Supongamos que $P$ es $T_n$-invariante

$$P \overset{\mathrm{d}}{=} P_{T_n} \Longrightarrow \widehat{P}(1) = \widehat{P_{T_n}}(1) \quad \forall n \in \mathbb{N}.$$

Pero como $\widehat{P}(nk) = \widehat{P_{T_n}}(k) \forall k \in \mathbb{Z} \Longrightarrow \widehat{P}(n) = \widehat{P_{T_n}}(1) = \widehat{P}(1) \forall n \in \mathbb{N}$.
Además

$$\begin{aligned} \widehat{P}(-n) &= \int_0^1 e^{2\pi n s} dP(s) = \int_0^1 \overline{e^{-2\pi n s} dP(s)} = \overline{\int_0^1 e^{-2\pi n s} dP(s)} \\ &= \overline{\widehat{P}(n)}. \end{aligned}$$

Entonces, $\exists q \in \mathbb{C}$ tal que

$$\widehat{P}(k) = \begin{cases} q & \text{si } k > 0 \\ 1 & \text{si } k = 0 \\ \overline{q} & \text{si } k < 0 \end{cases} \tag{B.4}$$

Veamos que, efectivamente, $q \in [0,1]$

$$P(\{0\}) = \int_{\{0\}} 1 dP(s) = \int_{\{0\}} \left( \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} e^{-2\pi i s j} \right) dP(s) \tag{B.5}$$

$$= \int_0^1 \left( \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} e^{-2\pi i s j} \right) dP(s) \tag{B.6}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \int_0^1 e^{-2\pi i s j} dP(s) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \widehat{P}(j) \tag{B.7}$$

$$= q \in [0,1].$$

En (B.5) se ha tenido en cuenta que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} e^{-2\pi i t j} = \begin{cases} 1 & \text{si } t \in \mathbb{Z} \\ 0 & \text{si } t \notin \mathbb{Z} \end{cases}$$

y en (B.7) se ha aplicado el teorema de convergencia dominada, factible dado que

$$\left| \frac{1}{n} \sum_{j=1}^{n} e^{-2\pi i s j} \right| \leq \frac{1}{n} \sum_{j=1}^{n} \left| e^{-2\pi i s j} \right| = 1 \quad \forall n \in \mathbb{N}.$$

En definitiva, B.4 queda

$$\widehat{P}(k) = \begin{cases} q & \text{si } k > 0 \\ 1 & \text{si } k = 0 \\ q & \text{si } k < 0 \end{cases}$$

Por tanto, los coeficientes de Fourier de $P$ son los de $q\delta_0 + (1-q)\lambda_{0,1}$, y por la unicidad de los mismos

$$P \stackrel{\mathrm{d}}{=} q\delta_0 + (1-q)\lambda_{0,1} \quad q \in [0,1].$$

□

**Teorema B.3.** *Una probabilidad $P$ en $\left([0,1), \mathscr{B}_{[0,1)]}\right)$ es $T$-invariante, con $T(x) = (x+a) \mod 1$, $a \in \mathbb{I}$; si, y solo si,*

$$P \stackrel{\mathrm{d}}{=} \lambda_{0,1}.$$

DEMOSTRACIÓN. El argumento es análogo al del teorema previo.

Como antes, $\Phi_k(s) = e^{-2\pi iks}$

$$\widehat{P}(k) = \int_0^1 e^{-2\pi iks} dP(s) = \int_0^1 \Phi_k(s) dP(s) \quad k \in \mathbb{Z}.$$

$$\begin{aligned} \widehat{P_T}(k) &= \int_0^1 \Phi_k(s) dP_T(s) = \int_0^1 \Phi_k(T(s))\, dP(s) \\ &= \int_0^1 \Phi_k((s+a) \mod 1)\, dP(s) = \int_0^1 \Phi_k(s+a)\, dP(s) \\ &= \Phi_k(a) \int_0^1 \Phi_k(s)\, dP(s) = \Phi_k(a)\widehat{P}(k). \end{aligned}$$

Veamos primero la condición suficiente. Si $P \stackrel{\mathrm{d}}{=} \lambda_{0,1} \Longrightarrow \widehat{P}(0) = 1$ y $\widehat{P}(k) = 0 \;\; \forall k \neq 0$. Ahora bien, $\widehat{P_T}(0) = 1$ y $\widehat{P_T}(k) = \Phi_k(a)\widehat{P}(k) = 0 \;\; \forall k \neq 0$. En definitiva, $P \stackrel{\mathrm{d}}{=} P_T$.

Veamos la condición necesaria. Supongamos que $P \stackrel{\mathrm{d}}{=} P_T$

$$\Longrightarrow \widehat{P}(k) = \widehat{P_T}(k) = \Phi_k(a)\widehat{P}(k) \quad \forall k \in \mathbb{Z}.$$

Sea $l \in \mathbb{Z} \setminus \{0\}$

$$\begin{aligned} &\text{Si } \widehat{P}(l) \neq 0 \Longrightarrow \Phi_l(a) = 1 \Longrightarrow e^{-2\pi ila} = 1 \\ &\Longrightarrow \cos(2\pi la) - i \operatorname{sen}(2\pi la) = 1 \\ &\Longrightarrow 2\pi la = 2\pi k \Longrightarrow a = \frac{k}{l} \in \mathbb{Q}. \end{aligned}$$

Hemos llegado a un absurdo; por tanto, $\widehat{P}(k) = 0 \;\; \forall k \neq 0$. En definitiva, $P \stackrel{\mathrm{d}}{=} \lambda_{0,1}$. □

# Apéndice C

# Código Matlab

## C.1. Sección 2.1.

### C.1.1. Código común a los tres ejemplos

```
% Crear un histograma correspondiente a la distribucion marginal
% del primer y/o segundo digito significativo.

classdef Histogramasmarginales
    properties
        soporteBL1 = 1:9;
        soporteBL2 = 0:9;
        soporteconjunta = 10:99;
        datos % vector de datos.
    end

    methods

        function obj = Histogramasmarginales(datos)
            obj.datos = datos;
        end

        % Distribucion marginal BL primer digito.
        function benford1 = benford1(obj)
            benford1 = log10(1 + 1./obj.soporteBL1);
        end

         % Distribucion marginal BL segundo digito.
        function benford2 = benford2(obj)
            benford2 = zeros(1, 10);
            for i = 0:9
                benford2(i+1) = sum(log10(1+1./
                (10.*obj.soporteBL1+i)));
            end
        end

        % Distribucion conjunta BL.
        function conjunta = conjunta(obj)
            conjunta = log10(1+1./obj.soporteconjunta);
        end

        % Distribucion marginal BL primer digito
        % empirica (del vector de datos).
```

```
        function BL1_emp = emp1(obj)
            BL1_emp = floor(obj.datos ./
            (10 .^ floor(log10(obj.datos))));
        end

        % Histograma correspondiente a la distribucion
        % marginal del primer digito significativo.
        function HistogramaBL1(obj)
            freq_data = histcounts(obj.emp1, 'BinLimits', [1 9],
                'BinMethod', 'integers',
                'Normalization', 'probability');
            figure
            plot(obj.soporteBL1, obj.benford1,
            'k-', 'LineWidth', 1, 'MarkerSize', 3);
            hold on
            plot(obj.soporteBL1, freq_data, 'k--+',
            'LineWidth', 1, 'MarkerSize', 5);
            hold off
            xlim([0.5 9.5])
            ylim([0 0.4])
            xlabel('Primer dgito significativo', 'FontSize',
            12,
                'FontName', 'Arial')
            ylabel('Frecuencia', 'FontSize', 12,
            'FontName', 'Arial')
            legend('1Dgito BL', '1Dgito datos', 'Location', ...
                'northeast', 'FontSize', 12, 'FontName', 'Arial')
        end

        % Histograma correspondiente a la distribucion
        % marginal del segundo digito significativo.
        function HistogramaBL2(obj)

        second_digits = floor((obj.datos ./ (10 .^
        floor(log10(obj.datos))) ...
            - obj.emp1)*10);
        freq_data = histcounts(second_digits,  'BinLimits', [0 9],
            'BinMethod', 'integers', 'Normalization',
            'probability');

        figure
        plot(obj.soporteBL2, obj.benford2, 'k-', 'LineWidth',
        1, 'MarkerSize', 3);
        hold on
        plot(obj.soporteBL2, freq_data, 'k--+', 'LineWidth',
        1, 'MarkerSize', 5);
        hold off
        xlim([-0.5 9.5])
        ylim([0 0.2])

        xlabel('Segundo dgito significativo', 'FontSize',
        12, 'FontName', 'Arial')
        ylabel('Frecuencia', 'FontSize', 12, 'FontName', 'Arial')
        legend('2Dgito BL', '2Dgito datos', 'Location',
        'northeast', ...
            'FontSize', 12, 'FontName', 'Arial')

        end
```

```
    end
end



%% Crear el histograma de la distribucion conjunta.

function conjunta(datos1, datos2, datos3)
    soporteBLconjunta = 10:99;
    % Distribucion conjunta del primer y segundo digitos
    % significativos BL.
    benford12 = log10(1+1./soporteBLconjunta);

    w1 = floor(datos1./10.^floor(log10(datos1) - 1));
    FS_digits1 = 10*floor(w1/10) + rem(w1,10);

    w2 = floor(datos2./10.^floor(log10(datos2) - 1));
    FS_digits2 = 10*floor(w2/10) + rem(w2,10);

    w3 = floor(datos3./10.^floor(log10(datos3) - 1));
    FS_digits3 = 10*floor(w3/10) + rem(w3,10);


    freq_data1 = histcounts(FS_digits1, 'BinLimits', [10 99],
    'BinMethod', ...
        'integers', 'Normalization', 'probability');
    freq_data2 = histcounts(FS_digits2, 'BinLimits', [10 99],
    'BinMethod', ...
        'integers', 'Normalization', 'probability');
    freq_data3 = histcounts(FS_digits3, 'BinLimits', [10 99],
    'BinMethod', ...
        'integers', 'Normalization', 'probability');

    figure
    plot(soporteBLconjunta, benford12, 'k-', 'LineWidth',
    1,
    'MarkerSize', 3);
    hold on
    plot(soporteBLconjunta, freq_data1, 'k-*', 'LineWidth',
    0.5,
        'MarkerSize', 3);
    plot(soporteBLconjunta, freq_data2, 'k-+', 'LineWidth',
    0.5,
        'MarkerSize', 3);
    plot(soporteBLconjunta, freq_data3, 'k-d', 'LineWidth',
    0.5,
        'MarkerSize', 3);
    hold off

    ylim([0 0.05])

    xlabel('Dos primeros dgitos significativos', 'FontSize',
    12,
        'FontName', 'Arial')
    ylabel('Frecuencia', 'FontSize', 12, 'FontName', 'Arial')
    legend('BL', '1','2', '3', 'Location', 'northeast', ...
        'FontSize', 12, 'FontName', 'Arial')
end
```

### C.1.2. Código del ejemplo 1

```
classdef OrganizarDatos
    methods

        % Guardamos todos los campos contenidos en
        % el estado financiero de la empresa.
        function allcampos = allcampos(~, filename)
            leer = fileread(filename);
            json_data = jsondecode(leer);
            allcampos = json_data.facts.us_gaap;
        end

        % Nos quedamos con los campos que satisfacen las
        % condiciones: tener mas de 100 registros y cuyo ultimo
        % registro sea de 2023.
        function selectedcampos = selectedcampos(obj, filename)
            todos = obj.allcampos(filename);

        % inicializamos el array donde vamos a almacenar
        % los campos seleccionados.
        selectedcampos = {};
        campos = fieldnames(todos);

        for i = 1:numel(campos)
            nombrecampo = campos{i};
            % comprobamos si el campo existe (evitar errores).
            if isfield(todos, nombrecampo)
                SubStruct = todos.(nombrecampo);
                % evitas los fieldNames vacios o 'raros'.
                if isfield(SubStruct.units, 'USD')
                % las que son de la forma struct (y no array)
                %  son entradas antiguas.
                    if ~isstruct(SubStruct.units.USD)
                        % elegimos campos con mas de 100 registros.
                        if numel(SubStruct.units.USD) > 100
                            % elegimos campos cuyo ultimo
                            %registro haya  sido en 2023.
                            % (evitamos campos obsoletos).
                            lastDateStr =
                            SubStruct.units.USD{end}.filed;
                            lastDate =
                            datetime(lastDateStr, 'InputFormat',
                            'yyyy-MM-dd');
                            if lastDate.Year >= 2023
                                selectedcampos =
                                [selectedcampos, nombrecampo];
                            end
                        end
                    end
                end
            end
        end
        end

        % Almacenamos el vector numerico final
        % (precios contenidos en los campos
        % finalmente seleccionados).
```

```
        function precios = precios(obj, filename, selectedcampos)
            todos = obj.allcampos(filename);
        % Inicializamos el vector de precios.
        precios = [];

        for i= 1:numel(selectedcampos)
            SubStruct = todos.(selectedcampos{i});
            for j = 1:numel(SubStruct.units.USD)
                precios = [precios, SubStruct.units.USD{j}.val];
            end
        end

        precios = precios(precios~=0);
        % Eliminamos los datos repetidos.
        precios = unique(precios);
        % la BL se aplica sobre los numeros
        % reales  y positivos.
        precios = abs(precios);
        end
    end
end


%% Abrir los archivos de datos.

% Visa Inc.
filename1 = 'CIK0001403161.json';
% Cisco Systems, Inc.
filename2 = 'CIK0000858877.json';
% Microsoft Corporation.
filename3 = 'CIK0000789019.json';

%% Extraer campos a analizar (Tabla 2.1)

organizador = OrganizarDatos();

% Nos quedamos con los campos que satisfacen las condiciones:
% tener mas de 100 registros y cuyo ultimo sea de 2023.
campos1 = organizador.selectedcampos(filename1);
campos2 = organizador.selectedcampos(filename2);
campos3 = organizador.selectedcampos(filename3);

% Nos quedamos con los campos comunes a las tres empresas.
% Estos son los que estan representados en la tabla 2.1.
camposfin = intersect(campos3,intersect(campos1, campos2));

%% Extraemos los los vectores de datos 'finales'
%% de las tres empresas.

% Visa Inc.
precios1 = organizador.precios(filename1, camposfin);
% Cisco Systems, Inc.
precios2 = organizador.precios(filename2, camposfin);
% Microsoft Corporation.
precios3 = organizador.precios(filename3, camposfin);

%% Histogramas
histogramas1 = Histogramasmarginales(precios1);
```

```
histogramas2 = Histogramasmarginales(precios2);
histogramas3 = Histogramasmarginales(precios3);

% Figura 2.2.
histogramas1.HistogramaBL1
histogramas1.HistogramaBL2
histogramas2.HistogramaBL1
histogramas2.HistogramaBL2
histogramas3.HistogramaBL1
histogramas3.HistogramaBL2

% Figura 2.1.
conjunta(precios1, precios2, precios3)
```

### C.1.3. Código del ejemplo 2

```
classdef OrganizarDatos
    properties
        sep = [2, 4, 2, 1, 2, 2, 3, 2, 100, ...
            1, 3, 3, 3, 8, 5, 8, 8, 8, 8, ...
            8, 8, 8, 8, 8, 3, 8, 8, 1]; % Separar los datos
        % (poblacion censada, votos validos, ...)
    end

    methods

        function data = abrirdatos(~, filename)
            fileID = fopen(filename, 'r');
            formatSpec = '%s';
            data = textscan(fileID, formatSpec, 'Delimiter',
            '\n');
            data = data{1};
            fclose(fileID);
        end

        function votos = obtenervotos(obj, filename)
            data = obj.abrirdatos(filename);
            datos = cell(length(data), length(obj.sep));
            for i = 1:length(data)
                fila = data{i};
                    for j = 1:length(obj.sep)
                        datos{i,j} = fila(1:obj.sep(j));
                        fila = fila(obj.sep(j)+1:end);
                    end
            end

            % aqui estan contenidos los votos
            % validos emitidos a favor de las candidaturas
            votos = datos(:,24);
            votos = cellfun(@(s) str2double(s), votos);
        end
    end
end



%% Abrir los archivos de datos.
```

```
% Noviembre 2019.
filename1 = '05021911.txt';
% Abril 2019.
filename2 = '05021904.txt';
% Junio 2016.
filename3 = '05021606.txt';


%% Extraer los votos de los datos.
organizador = OrganizarDatos();

votos1 = organizador.obtenervotos(filename1);
votos2 = organizador.obtenervotos(filename2);
votos3 = organizador.obtenervotos(filename3);

%% Histogramas
histogramas1 = Histogramasmarginales(votos1);
histogramas2 = Histogramasmarginales(votos2);
histogramas3 = Histogramasmarginales(votos3);

% Figura 2.4.
histogramas1.HistogramaBL1
histogramas1.HistogramaBL2
histogramas2.HistogramaBL1
histogramas2.HistogramaBL2
histogramas3.HistogramaBL1
histogramas3.HistogramaBL2

% Figura 2.3.
conjunta(votos1, votos2, votos3)
```

### C.1.4. Código del ejemplo 3

```

classdef OrganizarDatos
    methods

        function datos = obtenerelectro(~, filename)
            [data, ~] = edfread(filename);
            allCells = vertcat(data{:,1});
            datos = cell2mat(allCells);
        end
    end
end


%% Abrir los archivos de datos.

% Primer electrocardiograma fetal.
filename1 = 'r04.edf';
% Segundo electrocardiograma fetal.
filename2 = 'r07.edf';
% Tercer electrocardiograma fetal.
filename3 = 'r10.edf';

%% Extraer los electrocardiogramas discretizados de los datos.

```

```
organizador = OrganizarDatos();

electro1 = organizador.obtenerelectro(filename1);
electro2 = organizador.obtenerelectro(filename2);
electro3 = organizador.obtenerelectro(filename3);

%% Histogramas
histogramas1 = Histogramasmarginales(abs(electro1));
histogramas2 = Histogramasmarginales(abs(electro2));
histogramas3 = Histogramasmarginales(abs(electro3));
% Recordar que el analisis de la BL es sobre
% numeros reales y positivos.

% Figura 2.7.
histogramas1.HistogramaBL1
histogramas1.HistogramaBL2
histogramas2.HistogramaBL1
histogramas2.HistogramaBL2
histogramas3.HistogramaBL1
histogramas3.HistogramaBL2

% Figura 2.6.
conjunta(electro1, electro2, electro3)


%% Figura 2.5.

% Eje temporal.
t = linspace(0, 5, length(allCells1{1}));
% Primera onda.
y1 = electro1;
% Segunda onda.
y2 = electro2;
% Tercera onda.
y3 = electro3;

figure;
% dividimos la pantalla en tres.

% dibujo de la primera onda.
subplot(3,1,1);
plot(t, y1, 'k', 'LineWidth', 0.5);
ylabel('Voltaje (mV)', 'FontSize', 12, 'FontName', 'Arial');


set(gca, 'Color', '#fad1d0'); % fondo malla roja.
set(gca, 'XMinorGrid', 'on', 'YMinorGrid', 'on');
set(gca, 'MinorGridLineStyle', '-', 'MinorGridColor',
[0.8 0.2 0.2],
'MinorGridAlpha', 0.2, 'GridAlpha', 0);
set(gca, 'Box', 'on', 'LineWidth', 0.5);

% dibujo de la segunda onda.
subplot(3,1,2);
plot(t, y2, 'k', 'LineWidth', 0.5);
ylabel('Voltaje (mV)', 'FontSize', 12, 'FontName', 'Arial');

set(gca, 'Color', '#fad1d0'); % fondo malla roja.
```

```
set(gca, 'XMinorGrid', 'on', 'YMinorGrid', 'on');
set(gca, 'MinorGridLineStyle', '-', 'MinorGridColor',
[0.8 0.2 0.2],
'MinorGridAlpha', 0.2, 'GridAlpha', 0);
set(gca, 'Box', 'on', 'LineWidth', 0.5);

% dibujo de la tercera onda.
subplot(3,1,3);
plot(t, y3, 'k', 'LineWidth', 0.5);
ylabel('Voltaje (mV)', 'FontSize', 12, 'FontName', 'Arial');

set(gca, 'Color', '#fad1d0'); % fondo malla roja.
set(gca, 'XMinorGrid', 'on', 'YMinorGrid', 'on');
set(gca, 'MinorGridLineStyle', '-', 'MinorGridColor',
[0.8 0.2 0.2],
'MinorGridAlpha', 0.2, 'GridAlpha', 0);
set(gca, 'Box', 'on', 'LineWidth', 0.5);
xlabel('Tiempo (s)', 'FontSize', 12, 'FontName', 'Arial');
```

## C.2. Sección 2.2.

```
classdef OrganizarDatos
    properties
        datos
        uniquegoods %productos sin repetir.
        allgoods %todos los productos (incluyendo las veces
        % que se repiten).
        uniquecountries %paises sin repetir.
        allcountries %todos los paises (incluyendo las veces
        % que se repiten).
    end

    methods

        function obj = OrganizarDatos(datos)
            obj.datos = datos;
            % en filas estan todos los codigos (separados por
            % comas).
            filas = datos(:,1);
            splitted = cellfun(@(x) split(x, ','), filas,
            'UniformOutput', false);
            producto_agregado = cellfun(@(x) x{4}, splitted,
            'UniformOutput', false);
            pais_agregado = cellfun(@(x) x{3}, splitted,
            'UniformOutput', false);

            obj.uniquegoods = unique(producto_agregado,
            'stable');
            obj.allgoods = producto_agregado;
            obj.uniquecountries = unique(pais_agregado,
            'stable');
            obj.allcountries = pais_agregado;
        end

        % cambiamos el nombre a los bienes (p.e '02') a Goodi.
        function renamegood = renamegood(obj, goodcode)
```

```
            goods = obj.uniquegoods;
            pos = find(strcmp(goods, goodcode));
            renamegood = ['Good' num2str(pos)];
        end

        function extract = extract(obj, p)
            % si p=3 se extrae el country_code.
            % si p=4 se extrae el good_code.
            data = obj.datos;
            extract = struct();

            for i = 1:2:size(data, 1)
            % extraemos el codigo del producto/pais.
                splitted = strsplit(data{i, 1}, ',');
                code = splitted{p};
                if p==4
                % Utilizamos el nombre 'Good i' (renombramos).
                    code = obj.renamegood(code);
                end
                % vemos si ya se ha registrado alguna transaccion
                % de ese
                % producto/pais.
                if isfield(extract, code)
                % en caso afirmativo, anadimos esta transaccion
                % a las ya
                % registradas para este producto.
                    extract.(code){end+1} = data(i:i+1, 2:end);
                else
            % si no se ha registrado ninguna todavia, anadimos este
            % producto por primera vez.
                    extract.(code) = {data(i:i+1, 2:end)};
                end
            end
        end

        function [unit_price, quantity] = transacciones(obj, W)
            % El numero maximo de transacciones posibles para
            % un producto
            % o pais. Depende del input W
            % (extractgoods o extractcountry).
            max_transactions = max(cellfun(@numel,
            struct2cell(W)));

            t = size(obj.datos,2)-1; % numero de meses.
            nrow = max_transactions*t;

            % En la columna j se van a guardar todos las
            % cantidades de las
            % transacciones del producto j-esimo. Si las
            % transacciones son
            % nulas o no existen, se guardara un NaN.

            quantity = zeros(nrow, numel(fieldnames(W)));
            % En la columna j se van a guardar todos los precios
            % unidad de
            % las transacciones del producto j-esimo. Si las
            % transacciones
            % son nulas o no existen, se guardara un NaN.
```

```

        unit_price = zeros(nrow, numel(fieldnames(W)));

        elementos = fieldnames(W);

        for i = 1:numel(elementos)
        % fijamos un producto.
            current_elemento = W.(elementos{i});
            for j = 1:numel(current_elemento)
                % valores de la transaccion fijada
                % (a lo largo del tiempo).
                current_transaction = current_elemento{j};
                for k=1:t
                   precio = current_transaction{2,k};
                   % la transaccion puede estar guardada de
                   % dos formas.
                   % o en caracter o en numerico.
                   if ischar(precio)
                    precio_num = str2double(precio);
                   else
                       precio_num = precio;
                   end
                   cantidad = current_transaction{1,k};
                   if ischar(cantidad)
                    cantidad_num = str2double(cantidad);
                   else
                       cantidad_num = cantidad;
                   end
                   % si los valores son efectivamente
                   % transacciones 'reales', guardamos
                   % la cantidad
                   % y el precio unidad de la transaccion.
                   if ~isnan(precio_num) &&
                   ~isnan(cantidad_num) &&
                   cantidad_num~=0 ...
                           && precio_num~=0
                    unit_price(t*(j-1) + k, i) =
                    precio_num/cantidad_num;
                    quantity(t*(j-1) + k, i) = cantidad_num;
                   else
                   % si no ha habido transaccion, guardamos
                   % un NaN.
                    unit_price(t*(j-1) + k, i) = NaN;
                    quantity(t*(j-1) + k, i) = NaN;
                   end
                end
            end
        end
    end

    function contador = contartransacciones(~, matriz)
        contador = zeros(1, size(matriz, 2));
        for i = 1:size(matriz, 2)
            columna = matriz(:,i);
            % Contamos las entradas validas dentro de las
            % no nulas.
            contador(i) = nnz(columna(~isnan(columna)));
        end
```

```
        end

        % con esta funcion se busca quedarnos solamente con
        % los productos de los cuales haya habido mas de
        % 50 transacciones.
        function new_data = datoslimpios(obj)
            G = obj.extract(4);
            data = obj.datos;
            [~, matriz] = obj.transacciones(G);
            prop_goods = obj.contartransacciones(matriz);
            idx_eliminar = find(prop_goods<50);
            productos_eliminar = obj.uniquegoods(idx_eliminar);
            total_productos = obj.allgoods;

            % posicion de los productos a eliminar.
            pos = ismember(total_productos, productos_eliminar);
            data = data(~pos,:);
            % el conjunto de datos sin estos productos.
            new_data = data;

        end
    end
end


classdef MonteCarlo
    properties
        mt % numero de productos.
        nt % numero de transacciones.
        G % el elemento i-esimo representa a Good i.
        prop_transacciones % se almacenan las frecuencias
        % relativas de las transacciones de cada productos.
    end

    methods

        function obj = MonteCarlo(mt, nt, cant_transacciones)
            obj.G = 1:length(cant_transacciones);
            obj.mt = mt;
            obj.nt = nt;
            obj.prop_transacciones = cant_transacciones/...
            sum(cant_transacciones);
        end

         % Elegir mt productos SIN reemplazamiento, con
         % probabilidad de elecccion dada por su frecuencia.
        function selectgoods = selectgoods(obj)
            selectgoods = datasample(obj.G, obj.mt, 'Weights',
                obj.prop_transacciones, 'Replace', false);

        end

        % Conseguir mt numeros naturales no negativos que sumen
        % nt.
        % Por combinatoria: (nt-1 | mt-1) posibilidades.
        function trans_producto = trans_producto(obj)
            % nt-1 'separadores'.
            v = 1:obj.nt-1;
```

```
            % elegimos mt - 1 'separadores' sin reemplazamiento.
            idx = randperm(obj.nt-1);
            selected_idx = idx(1:obj.mt-1);
            selected_elements = sort(v(selected_idx));

            % los numeros seran la distancia entre 'separadores'.
            trans_producto =
            diff([0 selected_elements length(v)+1]);
        end


        % obtener transacciones de un trader (para los valores
        % de mt y nt).
        % la cantidad de cada producto elegido viene dado por
        % trans_producto.
        function untrader = untrader(obj, unit_price, quantity)
            % aqui se van a guardar la transaccion (los precios).
            untrader = [];

            T = obj.trans_producto;
            goods = obj.selectgoods;

            for j = 1:obj.mt
                % Fijas un producto de los elegidos.
                good = goods(j);
                % en las columnas de las matrices quantity y
                % unit_price
                % estan los valores de las cantidades y precios
                % unidad de los productos,
                % pero hay que eliminar las
                % entradas invalidas y/o nulas.

                columnaQ = quantity(:,good);
                columnaQ = columnaQ(~isnan(columnaQ));
                columnaQ = columnaQ(columnaQ~=0);
                columnaU = unit_price(:,good);
                columnaU = columnaU(~isnan(columnaU));
                columnaU = columnaU(columnaU~=0);

                % se obtienen tantos precios del producto fijado
                % como esta indicado en T.
                Qselected = datasample(columnaQ, T(j), 'Replace',
                true);
                Uselected = datasample(columnaU, T(j), 'Replace',
                true);

                % precio = precio_unidad * cantidad, lo anadimos.
                a = Qselected .* Uselected;
                untrader = [untrader a'];
            end
        end
    end
end


classdef ChiCuadrado
    properties
        observado % contiene la distribucion empirica.
```

```
        esperado % contiene la distribucion teorica (BL1).
        cuantil099 = 20.09023503; % para alpha = 0.01.
    end
    methods

        function obj = ChiCuadrado(transaccion)
            % se extrae el primer digito significativo de los
            % datos.
            todos = floor(transaccion ./ ...
                (10 .^ floor(log10(transaccion))));

            % se calcula la frecuencia absoluta de cada digito.
            obj.observado = histcounts(todos, 'BinMethod', ...
                'integers', 'BinLimits', [1 9]);
            soporteBL1 = 1:9;
            N = sum(obj.observado);

            obj.esperado = N.*log10(1 + 1./soporteBL1);
        end
        function valortest = valortest(obj)
            % test chi cuadrado (8 grados de libertad).
            [~, ~, stats] = chi2gof(1:9, 'Frequency',
            obj.observado,
                'Expected', obj.esperado);
            valortest = stats.chi2stat;
        end

        % calcula la potencia del test chi cuadrado por
        % simulacion.
        function dummy = indicador(obj)
            if obj.valortest > obj.cuantil099
                dummy = true;
            else
                dummy = false;
            end
        end
    end
end


%% Abrir el archivo de datos.

filename = 'Eurostat.tsv';
delimiter = '\t';

T = readtable(filename, 'FileType', 'text',
'Delimiter', delimiter, 'ReadVariableNames', true);
data = table2cell(T);

%% Ordenar los datos por productos y por paises.
organizador = OrganizarDatos(data);

% Seleccionamos productos con mas de 50
% transacciones.
new_datos = organizador.datoslimpios();
organizador = OrganizarDatos(new_datos);

% Dos matrices para guardar los precios unidad y
```

```
% cantidades de las transacciones. Cada columna es un producto.
G = organizador.extract(4);
C = organizador.extract(3);
[unit_priceG, quantityG] = organizador.transacciones(G);
[unit_priceC, quantityC] = organizador.transacciones(C);

cant_transacciones_prod =
organizador.contartransacciones(quantityG);
cant_transacciones_pais =
organizador.contartransacciones(quantityC);

%% Analisis de los paises (Tabla 2.4).

nt = organizador.contartransacciones(quantityC);
mt = zeros(25,1);
coc_mtnt = zeros(25,1);
alphapicos = zeros(25,1);
valoreschicuadrado = zeros(25,1);

for i=1:25
    arr = organizador.uniquecountries;
    target = arr{i};

    mt(i) = numel(C.(target));
    coc_mtnt(i) = mt(i)/nt(i);

    % Seleccionamos las transacciones del pais en cuestion:
    X1 = quantityC(:,i);
    X2 = unit_priceC(:,i);

    % todos los precios validos de las transacciones del pais.
    qq = X1.*X2;
    qq = qq(~isnan(qq));
    qq = qq(qq~=0);

    % Algoritmo MonteCarlo:
    montecarlo = MonteCarlo(mt(i), nt(i),
    cant_transacciones_prod);

    % 10,000 replicas Monte Carlo.
    realizaciones = 10000;
    cont = zeros(realizaciones,1);
    for j=1:realizaciones
    % simulacion de las transacciones
        transacciones = montecarlo.untrader(unit_priceG,
        quantityG);
        % valor del estadistico en estas transacciones
        chicuadrado = ChiCuadrado(transacciones);
        % indica si el valor es mayor que el cuantil 0.99
        dummy = chicuadrado.indicador();
        cont(j) = dummy;
    end
    % estimador de alpha (=0.01).
    alphapicos(i) = mean(cont);
    % valor del estadistico chi cuadrado.
    chicuadrado = ChiCuadrado(qq);
    valoreschicuadrado(i) = chicuadrado.valortest();
end
```

# Bibliografía


[1] Frank Benford. The law of anomalous numbers. *Proceedings of the American philosophical society*, pages 551–572, 1938.

[2] Arno Berger and Theodore P Hill. Benford's law strikes back: No simple explanation in sight for mathematical gem. *The Mathematical Intelligencer*, 33(1):85, 2011.

[3] Arno Berger and Theodore P Hill. *An introduction to Benford's law*. Princeton University Press, 2015.

[4] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[5] Leo Breiman. *Probability*. Addison Wesley, 1968.

[6] Andrea Cerioli, Lucio Barabesi, Andrea Cerasa, Mario Menegatti, and Domenico Perrotta. Newcomb–benford law and the detection of frauds in international trade. *Proceedings of the National Academy of Sciences*, 116(1):106–115, 2019.

[7] Rachel M Fewster. A simple explanation of benford's law. *The American Statistician*, 63(1):26–32, 2009.

[8] Ary L Goldberger, Luís AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000.

[9] Theodore P Hill. Base-invariance implies benford's law. *Proceedings of the American Mathematical Society*, 123(3):887–895, 1995.

[10] Theodore P Hill. The significant-digit phenomenon. *The American Mathematical Monthly*, 102(4):322–327, 1995.

[11] Theodore P Hill. A statistical derivation of the significant-digit law. *Statistical science*, pages 354–363, 1995.

[12] Raúl Jiménez. Forensic analysis of the venezuelan recall referendum. *Statistical science*, 2011.

[13] Alex Ely Kossovsky. *Benford's law: theory, the general law of relative quantities, and forensic fraud detection applications*, volume 3. World Scientific, 2014.

[14] Matthias Kreuzer, Denis Jordan, Bernd Antkowiak, Berthold Drexler, Eberhard F Kochs, and Gerhard Schneider. Brain electrical activity obeys benford's law. *Anesthesia & Analgesia*, 118(1):183–191, 2014.

[15] Lauwerens Kuipers and Harald Niederreiter. *Uniform distribution of sequences*. Courier Corporation, 2012.

[16] Walter R Mebane, R Michael Alvarez, Thad E Hall, and Susan D Hyde. Election forensics: The second-digit benford's law test and recent american presidential elections. *Election fraud: Detecting and deterring electoral manipulation*, pages 162–181, 2008.

[17] J. A. Morgan, A. S. Deaton, E. M. Cramer, J. Bibby, Wiorkowski, O'Neill Moore, and Varian. Letters to the editor. *The American Statistician*, pages 62–66, 1972.

[18] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of mathematics*, 4(1):39–40, 1881.

[19] Mark J Nigrini. *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*, volume 586. John Wiley & Sons, 2012.

[20] Mark J Nigrini. *Forensic analytics: Methods and techniques for forensic accounting investigations*. John Wiley & Sons, 2020.

[21] Luis Pericchi and David Torres. Quick anomaly detection by the newcomb—benford law, with applications to electoral processes data from the usa, puerto rico and venezuela. *Statistical science*, pages 502–516, 2011.

[22] Roger S Pinkham. On the distribution of first significant digits. *The Annals of Mathematical Statistics*, 32(4):1223–1230, 1961.

[23] Ralph A Raimi. The first digit problem. *The American Mathematical Monthly*, 83(7):521–538, 1976.

[24] Bernhard Rauch, Max Göttsche, Stefan Engel, and Gernot Brähler. Fact and fiction in eu-governmental economic data. *German Economic Review*, 12(3):243–255, 2011.

[25] Hermann Weyl. Über die gleichverteilung von zahlen mod. eins. *Mathematische Annalen*, 77(3):313–352, 1916.

# Lista de Acrónimos

**IB** Invariancia por cambio de base 2, 12, 28, 29, 34, 59

**IE** Invariancia por cambio de escala 2, 11, 12, 24–26, 28, 34, 59

**LB** Ley de Benford 1–3, 9, 11, 12, 16–19, 21–24, 26–31, 34–37, 40, 43, 46–48, 50

**m.p.a** medida de probabilidad aleatoria 2, 31–35, 48