Carmen Mª García Tardío

Clara Gómez García

Mª Jesús Macías Castillo

4. Supongamos que tenemos una larga lista de longitudes de ríos, alturas de montañas, superficies de países, precios de artículos, etc. y que nos fijamos tan sólo en el primer dígito. ¿Cuál es la probabilidad Pn de que el primer dígito sea n (n=1, 2, …., 9)?

Tras trabajar con los datos extraídos de distintas tablas, hemos confeccionado una tabla resumen con los datos de las probabilidades obtenidas.

	$P_{n=1}$	$P_{n=2}$	$P_{n=3}$	$P_{n=4}$	$P_{n=5}$	$P_{n=6}$	$P_{n=7}$	$P_{n=8}$	$P_{n=9}$
	(%)								
Países por superficies en km^2	28,45	20,50	11,30	11,30	7,95	5,02	6,69	4,18	4,60
Países por superficies en $millas^2$	33,47	16,02	12,13	12,97	4,18	5,02	5,86	4,18	5,86
Precios de un catálogo de moda	51,40	20,90	1,90	3,80	1,90	4,70	5,70	0,95	8,00
Calorías de los alimentos	17,40	22,72	26,84	8,85	9,43	4,42	3,83	3,54	2,65
Grasas de los alimentos	27,06	27,39	14,85	8,91	6,60	4,29	2,44	3,41	1,95
Proteínas de los alimentos	36,22	23,72	10,90	4,81	4,48	5,77	2,88	5,77	4,17

Si nos fijamos en las probabilidades de las dos primeras listas, se puede apreciar que son tanto menores cuanto mayor es el dígito. Podemos decir que estas dos listas siguen la llamada ***ley del primer dígito*** o ***ley de Benford***. Apreciaremos además que se trata de la misma lista de datos, difiriendo sólo en una constante de proporcionalidad, dada por el cambio de unidades.

Nuestra tarea, será deducir dicha ley de manera matemática, y demostrar que esta es invariante bajo cambio de escala. También daremos una posible explicación de por qué las probabilidades de las otras dos listas no siguen dicha ley.

Carmen Mª García Tardío

Física Estadística *Clara Gómez García*

Mª Jesús Macías Castillo

DEMOSTRACIÓN DE LA LEY DE BENFORD

La Ley de Benford se aplica a los datos que *no* son adimensionales, por lo que los valores numéricos de los datos dependen de las unidades. Si existe una distribución de probabilidad universal de $P(x)$ más de ese número, entonces debe ser invariante bajo un cambio de escala, de modo que:

$$P(kx) = f(k)P(x)$$

Si tenemos en cuenta la normalización,

$$\int P(x)dx = 1$$

De manera que,

$$\int P(kx)dx = \frac{1}{k}$$

Tenemos que la normalización implica lo siguiente:

$$f(k) = \frac{1}{k}$$

Derivando con respecto a *k* y estableciendo que *k=1*, tenemos

$$xP'(x) = -P(x)$$

Cuya solución es

$$\boxed{\boldsymbol{P(x) = \frac{1}{x}}}$$

Si muchas potencias de 10 se encuentran entre los puntos de corte, entonces la probabilidad de que el primer (decimal) dígito sea *D* está dada por una distribución logarítmica.

$$P_D = \frac{\int_D^{D+1} P(x)dx}{\int_1^{10} P(x)dx} = log_{10}\left(1 + \frac{1}{D}\right) \quad con\ D = 1, 2 \ldots., 9$$

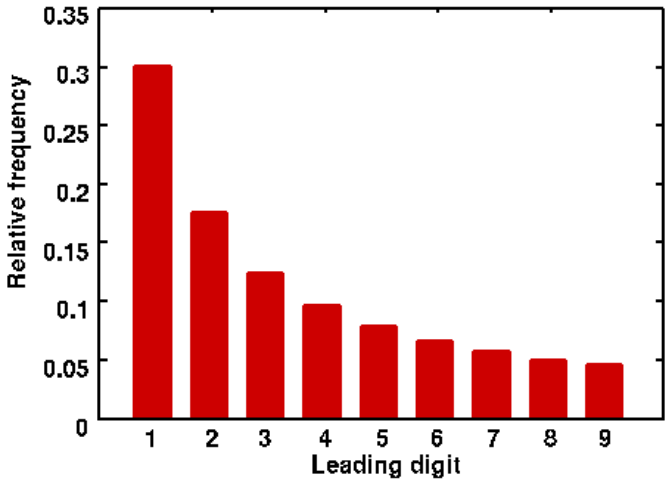
Carmen Mª García Tardío

Clara Gómez García

Mª Jesús Macías Castillo

A continuación, se recogen en una tabla las probabilidades teóricas para los distintos dígitos, según la ley anterior.

D	P_n
1	0.30103
2	0.176091
3	0.124939
4	0.09691
5	0.0791812
6	0.0669468
7	0.0579919
8	0.0511525
9	0.0457575



Para una lista de números que siga una distribución de probabilidad en forma de ley de potencias N^{-1}, tendremos que la probabilidad del primer dígito significativo es independiente de la década y sigue la ley de Benford:

$$\int_{10^{kd}}^{10^{k(d+1)}} N^{-1}dN = L_n\left(10^k(d+1)\right) - L_N\left(10^k d\right) = L_n\left(\frac{10^k(d+1)}{10^k d}\right) = L_n\left(\frac{d+1}{d}\right)$$

Normalizando,

$$\boxed{\boldsymbol{P(d) = log\left(\frac{d+1}{d}\right)}}$$

OBSERVACIONES Y COMENTARIOS

Tal y como comentábamos al comienzo del ejercicio, se ve claramente que las probabilidades $\boldsymbol{P_n}$ de las dos primeras tablas de datos, coinciden bastante bien con los datos

teóricos arriba expuestos. Esto pone de manifiesto el hecho de que esta ley sea invariante bajo el cambio de escala.

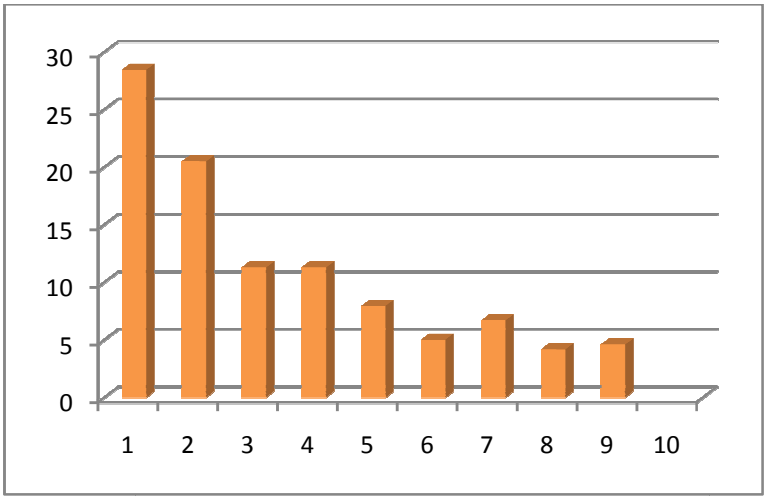
El hecho de que la lista de precios no lo siga, puede tener su origen, en que muchos de los datos se encuentran "enmascarados", ya que, al sólo fijarnos en la primera cifra, olvidamos que ésta podría perfectamente pertenecer al dígito inmediatamente anterior o posterior; es decir, que en el dato ***9,99***, contaríamos que pertenece al grupo ***n=9***, sin embargo, este dato enmascara al ***10,00***, de modo que debería estar en el grupo ***n=1***.

En ocasiones esta ley ha sido utilizada para detectar fraudes en las listas de datos. Este hecho resulta interesante, por los resultados obtenidos en la última lista (calorías de los alimentos) ya que por el nº de datos, y el tipo de magnitud, debería seguir dicha ley.

REPRESENTACIÓN GRÁFICA

A continuación, exponemos los histogramas correspondientes a cada una de las listas de datos recogidos.

Países por superficie (km^2)

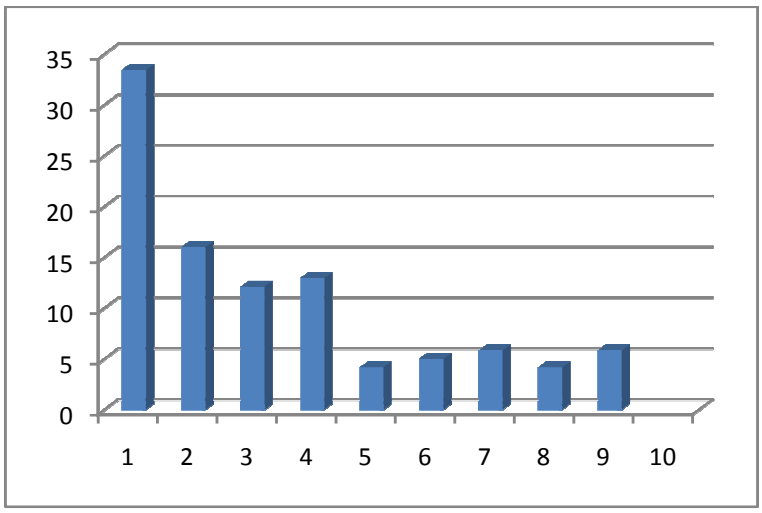


Carmen Mª García Tardío

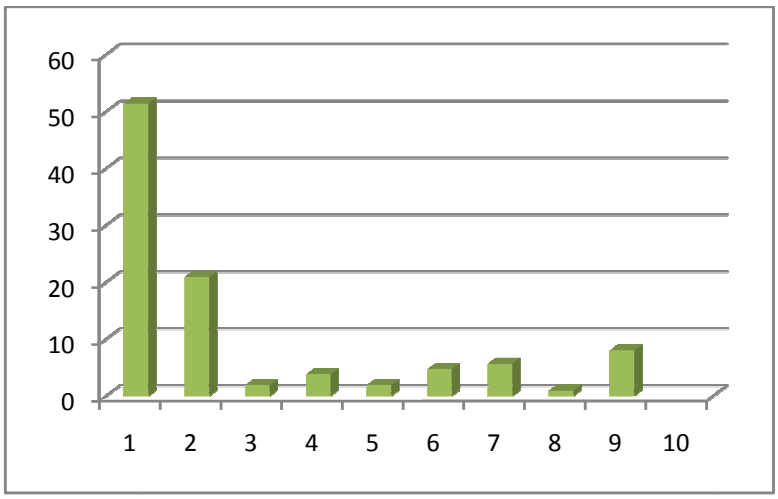
Física Estadística *Clara Gómez García*

Mª Jesús Macías Castillo

Países por superficie (Millas²)



Precios catálogo de moda

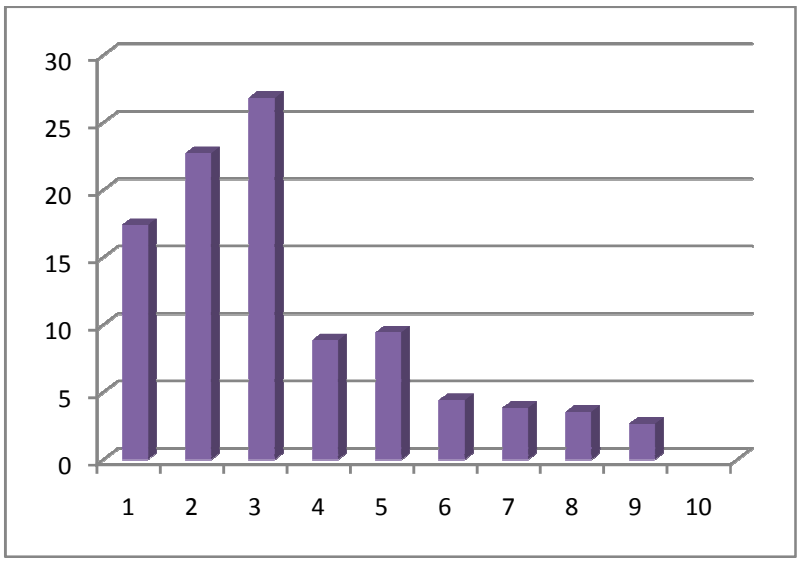


Carmen Mª García Tardío

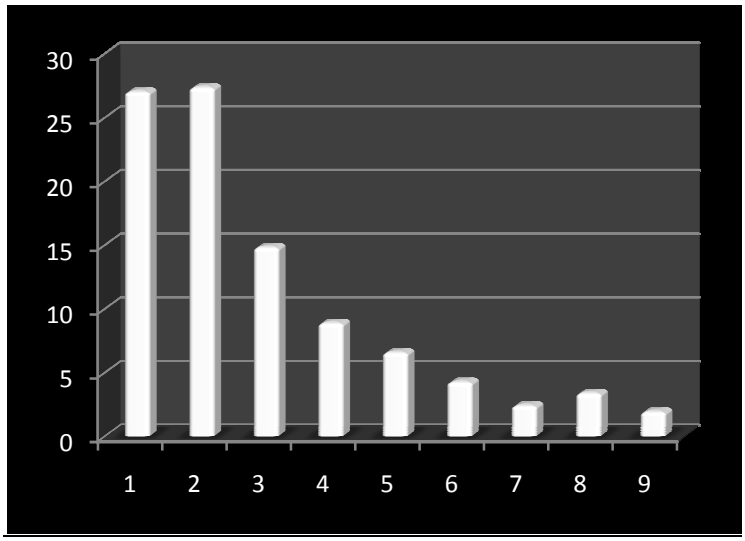
Física Estadística *Clara Gómez García*

Mª Jesús Macías Castillo

Calorías de los alimentos



Grasas de los alimentos



Carmen Mª García Tardío

Física Estadística *Clara Gómez García*

Mª Jesús Macías Castillo

Proteínas de los alimentos