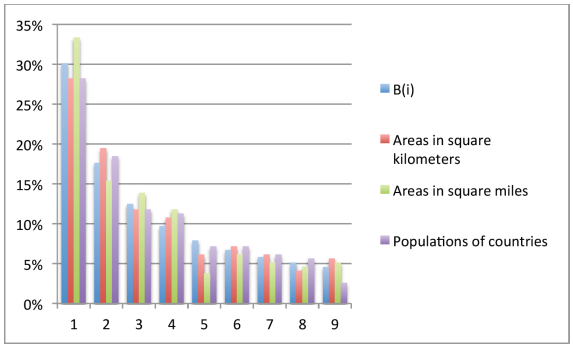
La ley de Benford: ¿aprender a defraudar o a detectar fraudes?

Autor original: Christiane Rousseau.



Cambiar demasiados números en documentos financieros puede resultar arriesgado si uno no conoce ciertas matemáticas. Muy a menudo, los números que aparecen en este tipo de documentos siguen cierta regla matemática, llamada ley de Benford o ley del primer dígito significativo. Si uno se olvida de seguir la regla, entonces los números no pasarán ciertos tests estadísticos y es probable que sean examinados con detenimiento por un hipotético agente fiscal. La ley de Benford afirma que si se toman números aleatoriamente y se calculan las frecuencias de sus primeros dígitos significativos, los números con primer dígito significante 1 representarían el 30%, mientras que los números con primer dígito significante 9 representarían el 4.5%. Esta regla se observa en otros muchos conjuntos de números, como las potencias de 2 o los números de Fibonacci.

¿Por qué?

A día de hoy se tienen explicaciones satisfactorias para este hecho y vamos a compartirlas con el lector.

La ley de Benford tiene que ver con la distribución de los primeros dígitos significativos de los números. El primer dígito significativo de un número positivo es el dígito no nulo que aparece más a la izquierda en su expresión decimal. Por ejemplo, el primer dígito significativo de π es 3, el de 2371.5 es 2 y el de 0.00563 es 5. Otra manera de definirlo que será útil en nuestra discusión matemática es escribir un número real positivo x como un número $m \in [1, 9)$ multiplicado por una potencia de 10:

$$x = m10^n \ , \quad n \in \mathbb{Z}.$$

Entonces el primer dígito significativo de x es la parte entera de m, que se denota por $\lfloor m \rfloor$. El número m se llama *mantisa* de x. Afirmamos que si tomamos una colección de números aleatorios y calculamos la frecuencia $B(i)$ del primer dígito significativo i, entonces $B(i)$ es aproximadamente $\log_{10}(1 + \frac{1}{i})$. Esta fórmula proporciona la siguiente tabla de frecuencias:

i	1	2	3	4	5	6	7	8	9
$B(i)$	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0511	0.0458

Tabla 1: Frecuencias en la ley de Benford.

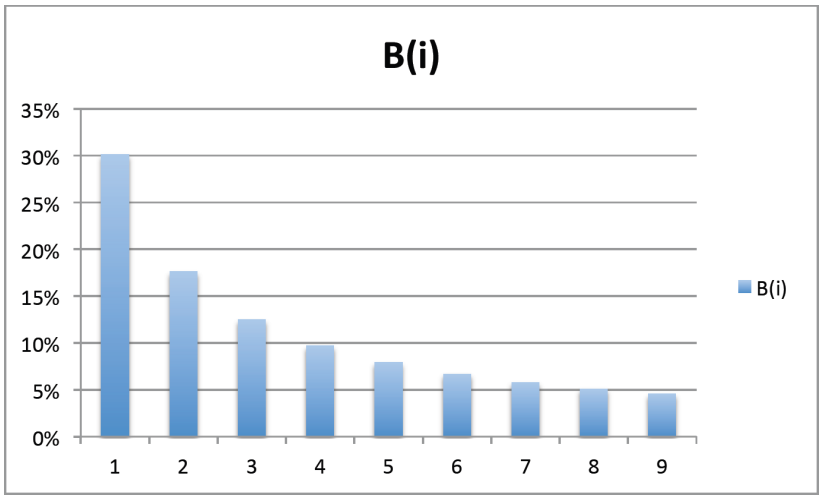


Figura 1: Frecuencias B(i) en la ley de Benford.

Demos ahora una breve reseña histórica. El fenómeno fue descubierto por primera vez por el astrónomo Simon Newcombe (1835-1909), quien se dio cuenta de que las primeras páginas de las tablas logarítmicas (correspondientes a dígitos significativos pequeños) aparecían mucho más desgastadas que las últimas páginas. Su descubrimiento fue olvidado y esta ley fue redescubierta por Frank Benford (1883-1948) hacia 1938. Frank Benford reunió decenas de miles de números de distintos orígenes que seguían su ley. La moderna base de datos de Simon Plouffe, que contiene 215 millones de constantes matemáticas también sigue la ley de Benford.

Muchos conjuntos de números que no son aleatorios también siguen la ley de Benford. Este es el caso de la población o la superficie de los países, la longitud de los ríos, etc. Quizá el lector quiera interrumpir la enumeración y empezar a ser escéptico... ¿En qué unidades se miden estas longitudes y estas áreas? ¿Las longitues vienen dadas en millas o en kilómetros? **Esto no importa...** Si las longitudes de los ríos en kilómetros siguen la ley de Benford entonces ¡las longitudes en millas también siguen la ley de Benford! Un cambio de unidades se corresponde con un cambio de escala. Veremos que la ley de Benford es **invariante frente a cambios de escala**. Más aún, es la única ley de probabilidad invariante frente a cambios de escala.

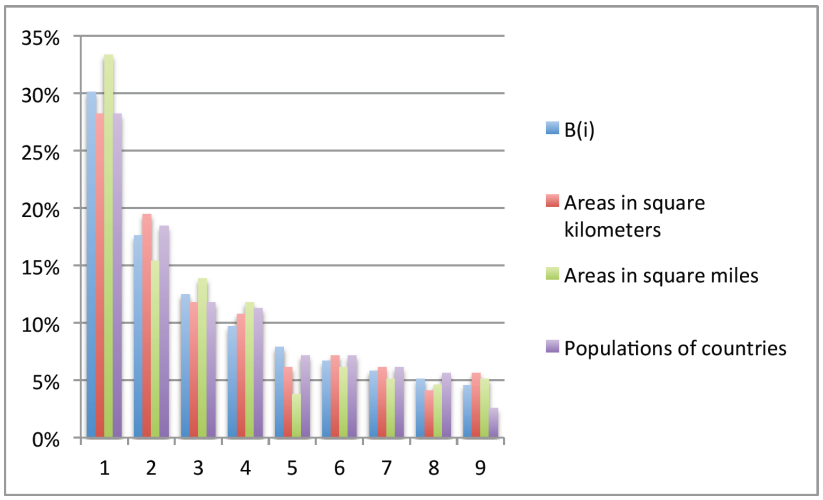


Figura 2: Algunos datos que siguen aproximadamente la ley de Benford: superficies de países en kilómetros cuadrados, áreas de países en millas cuadradas y poblaciones de países.

En la introdución se ha mencionado que los números de Fibonacci también siguen la ley de Benford. En cierto sentido, la ley de Benford es subjetiva, ya que depende de la base 10 en la que expresamos los números. En una base b con $b \neq 10$ los dígitos no nulos son los elementos del conjunto $\{1; ...; b-1\}$ y la ley de Benford en base b dice que la frecuencia del primer dígito significativo i es $B_b(i) = \log_b(1 + \frac{1}{i})$. Pues bien: ¡los números de Fibonacci

siguen la ley de Benford en cualquier base b! La ley de Benford es **invariante frente a cambios de base**.

Ya es tiempo de comenzar a dar explicaciones. Para ello se requiere al lector que recuerde sus cursos de probabilidad. O a lo mejor prefiere experimentar por sí mismo antes de leer matemáticas más serias.

1. Invarianza frente a cambios de escala

Consideremos un cambio de escala simple obtenido multiplicando todos los números por 2. Si consideramos los números con dígito significativo 1, todos ellos pasarán a tener como dígito significativo 2 o 3. Es fácil ver que $B(1) = B(2) + B(3)$. De hecho,

$$\begin{array}{rl} B(2) + B(3) & = \log_{10}(1 + \frac{1}{2}) + \log_{10}(1 + \frac{1}{3}) \\ & = \log_{10}\frac{3}{2} + \log_{10}\frac{4}{3} = \log_{10}\frac{3}{2}\frac{4}{3} \\ & = \log_{10} 2 = B(1) \end{array}$$

De manera similar se puede comprobar que $B(2) = B(4) + B(5)$, etc. Pero, ¿cómo arreglárselas al cambiar de kilómetros a millas, es decir, multiplicar números por $1,6$? Como se ha dicho anteriomente, la ley de Benford es demasiado restrictiva y necesitamos generalizarla. ¿Qué significa que el primer dígito significativo sea i? Significa que su mantisa m pertenece al intervalo $[i, i+1)$. Por tanto, la ley de Benford es una distribución de probabilidad parcial sobre la mantisa. La ley de Benford generalizada (que llamaremos ley de Benford haciendo abuso del lenguaje) en la mantisa viene dada por una función de densidad en el intervalo $[1, 10)$. Cuando elegimos un número al azar y calculamos su mantisa, obtenemos una variable aleatoria M que toma valores en $[1, 10)$. Podemos decir que sigue la ley de Benford si la función de densidad viene dada por

$$f(x) = \begin{cases} \frac{1}{x \log 10}, & x \in [1, 10), \\ 0, & \text{en otro caso.} \end{cases}$$

Si $P(a \leqslant M < b)$ es la probabilidad de que $a \leqslant M < b$ entonces se tiene que tener que

$$P(a \leqslant M < b) = \int_a^b f(x)dx.$$

Esto es una generalización de la ley de Benford, ya que

$$\begin{array}{rl} B(i) & = P(i \leqslant M < i+1) = \int_i^{i+1} \frac{1}{x \log 10} dx \\ & = \frac{1}{\log 10}(\log(i+1) - \log(i)) = \frac{1}{\log 10}(\log \frac{i+1}{i}) \\ & = \frac{\log(1+\frac{1}{i})}{\log 10} = \log_{10}(1 + \frac{1}{i}) \end{array}$$

¿Qué significa que una variable aleatoria X en $[1, 10)$ es invariante frente a cambios de escala? Significa que si c es un número real positivo y tomamos la variable aleatoria $Y = cX$ entonces la mantisa M de la variable aleatoria Y tiene la misma función de densidad que la de X. Esto no es difícil de probar en el caso en que X proviene de la ley de Benford, pero hay que distinguir casos en función del tamaño de c. Lo haremos para uno de los casos y dejaremos el resto al lector. Podemos escribir $c = m10^r$, donde $m \in [1, 10)$ es la mantisa de c. Como la mantisa de cX es la misma que la de mX, basta considerar el caso $c \in [1, 10)$. ¿Cuál es la herramienta necesaria para probar esto? Puede que el lector recuerde de sus cursos de probabilidad que la función de distribución (acumulada) es muchas veces más útil que la función de densidad para variables aleatorias continuas. La función de distribución de una variable aleatoria M se define como

$$F(x) = P(M \leqslant x).$$

Si X sigue la ley de Benford entonces su función de distribución viene dada por

$$F(x) = \begin{cases} 0, & x < 1, \\ \log_{10} x, & x \in [1, 10), \\ 1, & x \geqslant 10. \end{cases} \tag{1}$$

Por tanto, debemos probar que si X sigue la ley de Benford y M es la matisa de cX, para $c \in [1, 10)$, entonces la función de distribución de M viene dada por (1).

Para ello necesitamos calcular $P(M \leqslant z)$ para $z \in [1, 10]$. M es la mantisa de cX, que toma valores en $[c, 10c)$. Por tanto $M = cX$, si $cX < 10$ y $cX/10$ si $cX \geqslant 10$. El primer caso se da cuando $z < c$. La única posibilidad de que la mantisa de cX esté en $[1, c)$ es que $cX \in [10, 10c]$.Entonces la mantisa de cX es igual a $cX/10$. Por tanto,

$$\begin{aligned} P(M \leqslant z) &= P(1 \leqslant cX/10 \leqslant z) \\ &= P(\tfrac{10}{c} \leqslant X \leqslant \tfrac{10z}{c}) \\ &= F(\tfrac{10z}{c}) - F(\tfrac{10}{c}) \\ &= \log_{10} \tfrac{10z}{c} - \log_{10} \tfrac{10}{c} \\ &= \log_{10} z + \log_{10} \tfrac{10}{c} - \log_{10} \tfrac{10}{c} \\ &= \log_{10} z, \end{aligned}$$

como se buscaba. Los otros casos se resuelven de la misma manera.

El recíproco es más interesante...

2. La ley de Benford es la única ley de probabilidad sobre la mantisa invariante frente a cambios de escala

Esta es una afirmación impresionante. Sin embargo, veremos que la demostración no es mucho más complicada que el argumento anterior. Sea X la variable aleatoria que representa la mantisa y toma valores en $[1, 10)$. Busquemos su función de distribución $F(x)$ bajo la hipótesis de que X es invariante frente a cambios de escala; necesitamos calcular

$$F(x) = P(X \leqslant x) = P(1 \leqslant X \leqslant x).$$

Por tanto, tenemos que $F(0) = 0$ y $F(10) = 1$. La mayor dificultad de la demostración radica en interpretar qué significa que X es invariante frente a cambios de escala. Como $1 \leqslant X \leqslant x$ y $c \leqslant cX \leqslant cx$ son el mismo suceso, se tiene que

$$P(1 \leqslant X \leqslant x) = P(c \leqslant cX \leqslant cx) = F(x). \tag{2}$$

Como antes, consideramos el caso $c \in [1, 10)$, por lo que $cx < 10$ (c depende de x). Así, para $c \leqslant cX \leqslant cx$, cX es igual a su mantisa. Como X es invariante frente a cambios de escala, la mantisa de cX tiene la misma función de distribución que X. Por tanto,

$$P(c \leqslant cX \leqslant cx) = F(cx) - F(c).$$

Combinando con (2) se tiene que $F(x)$ verifica

$$F(x) = F(cx) - F(c), \quad F(1) = 0, \quad F(10) = 1. \tag{3}$$

siempre que $c \in [1, 10)$ no sea demasaido grande. Debemos hallar F en la *ecuación funtional* (3). Veamos cómo hacer esto. Si $c = 1 + \varepsilon$, entonces

$$F(x) = F(x(1 + \varepsilon)) - F(1 + \varepsilon),$$

que puede ser expresado como

$$\frac{F(x(1+\varepsilon))-F(x)}{x\varepsilon}=\frac{F(1+\varepsilon)-F(1)}{x\varepsilon},$$

ya que $F(1)=0$. Si tomamos el límite cuando $\varepsilon \longrightarrow 0$, reconocemos en cada lado de la ecuación un cociente cuyo límite es una derivada. En el lado izquierdo es $\frac{F(x+x\varepsilon)-F(x)}{x\varepsilon}$, cuo límite es $F'(x)$, y en el lado derecho $\frac{F(1+\varepsilon)-F(1)}{\varepsilon}$, que tiende a $F'(1)$. Por tanto, se tiene la siguiente ecuación diferencial en variables separables:

$$F'(x)=\frac{F'(1)}{x},$$

cuya solución es $F(x)=F'(1)\ln x+C$. Como $F(1)=0$, tenemos que $C=0$, y como $F(10)=1$, entonces $F'(1)=\frac{1}{\ln 10}$. Así, $F(x)=\frac{\ln x}{\ln 10}=\log_{10} x$ y con ello hemos terminado.

3. ¿Por qué números de todo tipo de procedencia siguen la ley de Benford?

Theodore Hill dio una respuesta en 1995. Discutamos brevemente su idea. Por supuesto, no todos los conjuntos de números siguen la ley de Benford. Por ejemplo, si se considera la altura en metros de las personas entonces los únicos dígitos significativos que aparecen son, salvo unos pocos casos, 1 y 2. Si se convierten estas medidas a pies (un pie equivale aproximadamente a 30 cm) entonces la ley de distribución de los dígitos significativos varía. Por tanto, este conjunto no es invariante frente a cambios de escala. Supongamos que tenemos un conjunto de números de diversa procedencia y le cambiamos la escala. En este conjunto existen subconjuntos de números con diferente escala. Como este conjunto es grande y los números tienen diferentes orígenes, lo más probable es que diferentes escalas estén presentes. Multiplicar todos los números del conjunto por una constante positiva induce una permutación de las escalas en el nuevo conjunto. Por tanto, podemos esperar que el conjunto se comporte como si no tuviera ninguna escala en particular, luego seguirá la ley de Benford.

Esta explicación es buena para conjuntos de números provenientes de orígenes diversos, pero no explica por qué las superficies de los países o sus poblaciones o las longitudes de los ríos siguen la ley de Benford. Comentaremos explicaciones recientes (2008) para estos casos dadas por Gauvrit, Delahaye y Fewster. Su razonamiento es válido también para conjuntos grendes de números de toda procedencia.

4. Es probable que los conjuntos de números que abarcan diferentes órdenes de magnitud sigan la ley de Benford

Trabajando en base 10 hemos visto que los números positivos pueden ser escritos como $x=m10^n$, donde $m\in[1,10)$ y $n\in\mathbb{Z}$. Podemos considerar n como el orden de magnitud de x. Decimos que hay diferentes órdenes de magnitud en un conjunto si aparecen diferentes valores de n para sus elementos. Notar que esta propiedad es invariante frente a cambios de escala. Para simplificar la explicación, supongamos que los números están en el intervalo $[1,10^6)$. En este caso, los números con dígito significativo 1 son los pertenecientes al conjunto

$$S_1=[1,2)\cup[10,20)\cup[100,200)\cup[1000,2000)\cup[10^4,2\times10^4)\cup[10^5,2\times10^5).$$

De manera similar definimos los conjuntos S_i para los otros dígitos. Es mejor trabajar con el logaritmo en base 10 de estos números: $y=\log_{10}x$; así, $y=\log_{10}m+n$. Probemos ahora que si una variable aleatoria M en $[1,10)$ sigue la ley de Benford entonces la variable aleatoria $Z=\log_{10}M$ es uniforme en $[0,1)$. Para ello, basta ver que la funcion de distribución de Z es la de una variable aleatoria uniforme en $[0,1)$, es decir,

$$F(z)=\begin{cases}0, & z<0,\\ z, & z\in[0,1),\\ 1, & z\geqslant 1.\end{cases}$$

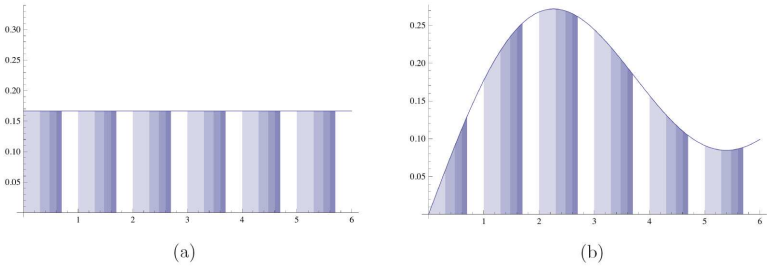
De hecho, si $z \in [0,1)$,

$$P(Z \leqslant z) = P(0 \leqslant \log_{10} M \leqslant z) = P(1 \leqslant M \leqslant 10^z) = \log_{10} 10^z = z.$$

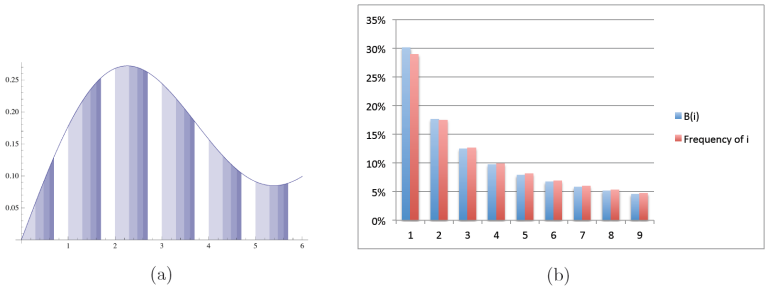
Si X pertenece al conjunto S_1, entonces Y está en $T_1 = \log_{10} S_1$:

$$\begin{aligned} T_1 \quad &= [0, \log_{10} 2) \cup [1, 1 + \log_{10} 2) \cup [2, 2 + \log_{10} 2) \\ &\cup [3, 3 + \log_{10} 2) \cup [4, 4 + \log_{10} 2) \cup [5, 5 + \log_{10} 2), \end{aligned}$$

y de manera similar para los demás dígitos. Supongamos que tomar un número aleatorio de nuestro conjunto es una variable aleatoria X que toma valores en $[1, 10^6)$. Entonces $Y = \log_{10} X$ toma valores en $[0, 6)$. Notar que la probabilidad de que una variable aleatoria pertenezca a determinado conjunto es igual al área bajo la gráfica de la función de densidad sobre el conjunto.Si la función de densidad f de Y sobre $[0, 6)$ fuera uniforme, como en la Figura 3 (a), obtendríamos lo que queríamos probar. Sin embargo, en la mayoría de los casos no es así, como en la Figura 3 (b). **Por eso es tan importante que el conjunto original de números abarque diferentes órdenes de magnitud.** Las diferentes partes correspondientes a un dígito significativo dado i se extienden horizontalmente a lo largo de varios segmentos, cuya suma de longitudes es del orden de $\log_{10}(1 + \frac{1}{i})$ de la anchura total. Por tanto, incluso si la altura de $f(x)$ no es la misma de un segemento a otro, se puede esperar que la altura media sea del mismo orden de magnitud para diferentes dígitos. Cuando esto sucede, los datos siguen la ley de Benford.



(a) función de densidad f uniforme
(b) función de densidad f no uniforme
Figura 3: Las áreas correspodientes a las frecuencias de los primeros dígitos significativos 1, 2, 3 y 4 para diferentes funciones de densidad de Y. Los valores de las correspondientes áreas están reflejadas en la Figura 4.



(a) función de densidad de f
(b) Áreas bajo la curva para los dígitos significativos de f y para la función uniforme
Figura 4: Las áreas correspondientes a las frecuencias de los primeros dígitos significativos 1, 2, 3 y 4 para la función de densidad de la Figura 3(b). A la derecha se puede ver que estos valores están muy cercanos a los obtenidos mediante la ley de Benford en el caso en que Y tenga una función de densidad uniforme.

5. ¿Cómo comprobar si un conjunto de números sigue la ley de Benford?

Si el lector ha tomado cursos de estadística, probablemente haya estudiado el test de bondad de ajuste chi

cuadrado. Este test permite comprobar si ciertos datos siguen cierta distribución de probabilidad. Supongamos que que se quiere hacer este test a un conjunto de n números. Necesitaremos construir una tabla, en la que n_i representa el número de números qdel conjunto que tienen como primer dígito significativo i. Por supuesto, $n = n_1 + ... + n_9$. N_i representa el número de números del conjunto que tendrían primer dígito significativo i si el conjunto siguiera la ley de Benford, es decir, $N_i = nB(i)$.

i	1	2	3	4	5	6	7	8	9
n_i	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
N_i	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9

Tabla 2: La tabla para el test de bondad de ajuste χ^2.

Se calcula

$$\chi^2 = \sum_{i=1}^{9} \frac{(n_i - N_i)^2}{N_i},$$

y se busca en la tabla de la χ^2 la línea que corresponde a 8 grados de libertad. Si se va a hacer un test con un error del 5%, entonces se acepta que los datos se ajustan a la ley de Benford si $\chi^2 < 15.51$ y se rechaza en otro caso. Este es un método sencillo, pero si se van a hacer tests con estudiantes es conveniente que se familiaricen con los detalles del test y su significado.

6. Invarianza de la ley de Banford frente a cambios de base

Este caso se modela de manera similar a la invarianza frente a cambios de escala, aunque es un poco más complicado, ya que no podemos limitar el trabajo únicamente a la mantisa. De hecho, si $x = m10^n$ entonces la parte 10^n también debe se convertida a la nueva base. La mayor dificultad radica en expresar en términos matemáticos qué significa que una variable aleatoria sea independiente frente a cambios de base. Omitimos los detalles de este caso.

7. Conclusión

La ley de Benford es fascinante: desafía la intuición, se puede comprobar por uno mismo y también adaptar para una actividad de aula. Lo que solía ser una mera curiosidad es ahora una herramienta estándar para detectar fraudes. Por supuesto, cada vez más evasores de impuestos saben de ella. Pero hay que prestar atención: el primer dígito significativo no es lo único a tener en cuenta. La ley de Benford generalizada nos permite derivar leyes para el segundo dígito significativo, el tercero, etc. El lector uede tratar de encontrarlas por sí mismo: basta pensar en qué uniones de intervalos debe encontrarse la mantisa de un número para que su segundo (tercer, etc.) dígito significativo sea i.