por Bartolo Luque

Bartolo Luque es físico y profesor de matemáticas en la Universidad Politécnica de Madrid. Sus investigaciones se centran en la teoría de sistemas complejos.

El problema de los tanques alemanes

Cuántos taxis hay en Barcelona y cuánta gente correrá la San Silvestre Vallecana

Las matemáticas desempeñaron un papel decisivo durante la Segunda Guerra Mundial. Quizás el caso más conocido, debido a su popularización en la literatura y en el cine, sea la participación de Alan Turing en el centro de criptografía británico de Bletchley Park para descifrar los códigos secretos nazis de las máquinas Enigma. Numerosos expertos consideran que, gracias a ello, la guerra se acortó unos dos años.

Una historia menos conocida tal vez sea que los Aliados usaron una sencilla fórmula estadística para estimar la producción alemana de armamento; en particular, la de sus temidos tanques. Aquel cálculo funcionó muchísimo mejor que las hipótesis de las fuentes de inteligencia, basadas en el recuento de tanques en el campo de batalla y en el espionaje de las fábricas germanas.

Fecha	Estimación estadística	Servicios de inteligencia	Valor real
Junio de 1940	169	1000	122
Junio de 1941	244	1550	271
Agosto de 1942	327	1550	342

Esta tabla, extraída de un artículo de R. Ruggles y H. Brodie de 1947, muestra tres valores reales de producción mensual (conocidos después de la guerra), las estimaciones de los estadísticos y las de los servicios de inteligencia. Mientras que estos últimos concluyeron que, entre junio de 1940 y septiembre de 1942, los alemanes estaban fabricando una media de 1400 tanques al mes, los estadísticos dedujeron que ese número debía ser considerablemente menor: 246. Al finalizar la guerra se tuvo acceso al valor exacto, que resultó ser... ¡245!

¿Cómo lograron los matemáticos acercarse tanto? En 1991, el coronel estadounidense Trevor Dupuy contaba la siguiente anécdota: «Hace pocos años, en Oriente Medio, obtuve permiso del Ejército israelí para visitar su línea de producción de tanques Merkava. En cierto momento pregunté cuántos habían producido, pero se me dijo que se trataba de información clasificada. Me pareció divertido, porque había un número de serie en el chasis de cada tanque». En efecto, los Aliados estimaron el número de tanques alemanes a partir de los números de serie de los vehículos nazis capturados o destruidos.

¡Taxi!

Pero ¿cómo deducir el número total de tanques a partir de una pequeña muestra de números de serie? Pere Grima, profesor de estadística de la Universidad Politécnica de Cataluña, nos propone en su libro *La certeza absoluta y otras ficciones* el siguiente experimento, de corte menos belicista.

En una ciudad como Barcelona, los taxis se encuentran numerados correlativamente por licencias que van de 1 a N. Dicho número figura en la puerta de cada vehículo y puede leerse a distancia con facilidad. Supongamos que deseamos estimar el número total de taxis, N, a partir de una muestra de n licencias observadas en la calle al azar. En jerga estadística, diremos que se trata de estimar el tamaño de una población a partir de una muestra aleatoria sin reposición.

Para tomar valores numéricos concretos, supongamos que el número de taxis es $N = 41$ y que el tamaño de nuestra muestra asciende a $n = 5$ licencias. Imaginemos que, ordenadas de menor a mayor, estas resultan ser 8, 14, 22, 27 y 35.

Si conociéramos la licencia m que ocupa la posición media exacta en la población de taxis, el número total de vehículos sería $N = (m - 1) + 1 + (m - 1) = 2m - 1$. En nuestro ejemplo, con una población de $N = 41$, m vale 21. Y, en efecto, $N = 2 \cdot 21 - 1 = 41$.

Pero en un experimento con taxis reales no podremos saber el valor de m. Sin embargo, parece razonable aproximar dicha cantidad a través de la *mediana muestral*, $\tilde{X}$, el dato de nuestra muestra ordenada que deja tantos valores a su izquierda como a su derecha. En nuestro caso, $\tilde{X} = 22$; es decir, el valor que se encuentra por encima de 8 y 14, y por debajo de 27 y 35. A partir de aquí, podemos construir el siguiente estimador puntual: $\hat{N}_1 = 2\tilde{X} - 1$. En nuestro ejemplo, este arroja una población de $\hat{N}_1 = 2 \cdot 22 - 1 = 43$. ¡Nada mal!

Pero disponemos de otras opciones. Una de ellas consiste en tomar la *media* muestral, $\bar{X}$, en lugar de la mediana. Como siempre, esta se calcula sumando todos los datos disponibles y dividiendo el resultado entre el tamaño de la muestra. En nuestro caso, obtenemos $\bar{X} = (8 + 14 + 22 + 27 + 35)/5 = 21{,}2$. Por tanto, el estimador asociado sería $\hat{N}_2 = 2 \cdot 21{,}2 - 1 = 41{,}4$.

Ambos estimadores parecen muy razonables. Pero ¿puede el lector encontrarles una sencilla pega? Supongamos que nuestras licencias hubieran sido 2, 10, 15, 25 y 40. La mediana y la media muestral serían ahora 15 y 18,4, respectivamente, y las estimaciones correspondientes, $\hat{N}_1 = 29$ y $\hat{N}_2 = 35{,}8$. ¡Ambas menores que 40, uno de los datos de la muestra! Una flagrante contradicción.

Estimadores fiables

¿De qué alternativas disponemos? Denotemos los datos de la muestra ordenados de menor a mayor como $X_1, X_2, \ldots, X_n$. Por simetría, tal vez podamos suponer que la cantidad de taxis con un número de licencia superior a X_n (el valor más alto de la muestra) es igual a la cantidad de vehículos con una licencia inferior a X_1 (el más bajo). De modo que podemos aventurar que $N - X_n = X_1 - 1$, lo que nos lleva a definir un tercer estimador: $\hat{N}_3 = X_n + X_1 - 1$. Aplicado a los datos de nuestro ejemplo, obtendríamos $\hat{N}_3 = 35 + 8 - 1 = 42$.

Podemos refinar algo más el argumento anterior. ¿Por qué limitarnos a tomar el valor más bajo como referencia? Parece más adecuado suponer que la cantidad de taxis que poseen un número de licencia superior al más elevado de la muestra debería ser igual al promedio de las «distancias» entre nuestras observaciones:

$$N - X_n = \frac{(X_1 - 1) + (X_2 - X_1 - 1) + \cdots + (X_n - X_{n-1} - 1)}{n} = \frac{X_n}{n} - 1.$$

Ello nos permite definir un cuarto estimador: $\hat{N}_4 = X_n + X_n/n - 1$, igual al máximo de la muestra más el hueco medio en la muestra. En nuestro caso, este vale $\hat{N}_4 = 41$.

Observemos que nuestros nuevos estimadores jamás podrán arrojar un resultado menor que ninguno de los datos de la muestra. Dado que $\hat{N}_3 = X_n + X_1 - 1$, el valor mínimo que puede tomar es $\hat{N}_3 = X_n$ (cuando $X_1 = 1$). Y en el caso de $\hat{N}_4 = X_n + X_n/n - 1$, su cota inferior viene dada por el valor más bajo posible de X_n. Dicho valor es n, en cuyo caso obtendremos $\hat{N}_4 = n = X_n$.

Ambos estimadores parecen funcionar bastante bien. ¿Con cuál nos quedamos? Para compararlos, podemos efectuar simulaciones de Monte Carlo, un ejercicio instructivo. Fijamos valores para N y n, y realizamos el siguiente experimento: tomamos una muestra aleatoria sin reposición de n datos de la población N, estimamos la población a través de las dos fórmulas anteriores y calculamos el error asociado a cada una.

Al repetir el experimento un gran número de veces, obtendremos las distribuciones de probabilidad para $\hat{N}_3$ y $\hat{N}_4$, a partir de las cuales podremos calcular sus respectivos valores medios, sus errores medios y sus varianzas. Al final, el mejor estimador será aquel que muestre un error medio y una varianza menores.

Puede demostrarse con facilidad que los valores medios de $\hat{N}_3$ y $\hat{N}_4$ coinciden con N; es decir, que sus errores medios son cero (en términos técnicos, decimos que son estimadores *no sesgados*). Por tanto, para desempatar habremos de recurrir a las varianzas. Estas también pueden calcularse de manera exacta. Al hacerlo, obtenemos:

$$\mathrm{Var}\,(\hat{N}_3) = \frac{2}{n+1}\,\frac{(N-n)(N+1)}{n+2},$$

$$\mathrm{Var}\,(\hat{N}_4) = \frac{1}{n}\,\frac{(N-n)(N+1)}{n+2}.$$

Vemos entonces que $\mathrm{Var}(\hat{N}_3) \geq \mathrm{Var}(\hat{N}_4)$ (observemos que coinciden cuando $n = 1$ y que ambas se hacen cero cuando $n = N$). Así pues, $\hat{N}_4 = X_n + X_n/n - 1$ se perfila como nuestro mejor estimador. Sin ir más lejos, esta fue la fórmula empleada por los Aliados para hacerse una idea de la cantidad de tanques alemanes a partir de sus números de serie.

iPhones y corredores de maratón

La tarea de estimar el tamaño de una población numerada a partir de una muestra sin reemplazo se conoce en el mundo anglosajón como problema de los tanques alemanes. Durante la Segunda Guerra Mundial, dicho método se aplicó a otros pertrechos militares, como los temidos cohetes V2.

En la práctica, sin embargo, el problema puede resultar mucho más complejo. Puede ocurrir que el valor mínimo de los números de serie no sea 1, sino una cantidad desconocida. O que los elementos de la población no cuenten con la misma probabilidad de ser extraídos de la muestra, como seguramente ocurrió con los tanques, cuya posibilidad de ser observados aumentaba a medida que transcurría la contienda.

SAN SILVESTRE VALLECANA del 31 de diciembre de 2005. ¿Sabría estimar el número total de corredores a partir del número de dorsal de una veintena de ellos?

Con todo, la idea básica para llevar a cabo dichas estimaciones es siempre la misma. Hoy en día, las empresas emplean este método para calcular la producción de sus competidores. Hace unos años, a partir de números de serie obtenidos en foros de Internet, los analistas estimaron que el número de iPhones vendidos desde su lanzamiento hasta finales de septiembre de 2008 era de 9,1 millones.

Por cierto: la estimación del número de taxis en Barcelona que hizo Pere Grima a partir de 20 licencias tomadas al azar fue de 10.989. El valor real en aquel momento, consultado en Internet, ascendía a 10.481.

Por mi parte, para no agraviar a los lectores madrileños y como homenaje final a 2013, Año Internacional de la Estadística, me comprometo a estimar el número de corredores que participarán en la San Silvestre Vallecana, la popular maratón que se celebra cada 31 de diciembre en Madrid, a partir de los dorsales de unos cuantos participantes escogidos al azar. Lo suyo sería correrla, pero ya no estoy para esos trotes.

PARA SABER MÁS

An empirical approach to economic intelligence in World War II. R. Ruggles y H. Brodie en *Journal of the American Statistical Association*, vol. 42, págs. 72-91, 1947.

Estimating the size of a population. Roger Johnson en *Teaching Statistics*, vol. 16, n.º 2, págs. 50-52, 1994.

La certeza absoluta y otras ficciones: Los secretos de la estadística. Pere Grima. RBA, colección *El mundo es matemático*, 2010.