Paradoja de Simpson

Solemos valorar el éxito de un tratamiento médico en la proporción de los pacientes con resultados positivos entre todos que lo reciben. Al parecer, las cifras calculadas de esta manera deberían ser válidas porque reflejan, en cierto sentido, la eficacia del tratamiento. ¿Pero has pensado alguna vez que podría haber un problema?

Vamos a investigar un caso real de dos tratamientos de cálculos reñales (Charig y col., 1986). Los tratamientos usados son la cirugía abierta (se denota por Tratamiento A) y la nefrolitotomía percutánea (Tratamiento B, que es menos invasivo). Los pacientes tratados se dividen en cuatro grupos según dos criterios: el tamaño de los cálculos (grandes o pequeños) y el tipo de tratamiento que reciben (A or B). Los resultados se muestran en Tabla 1.

	Tratamiento A	Tratamiento B
Cálculos pequeños	$\underline{93\,\%}$ $(81/87)$	$87\,\%$ $(234/270)$
Cálculos grandes	$\underline{73\,\%}$ $(192/263)$	$69\,\%$ $(55/80)$
Global	$78\,\%$ $(273/350)$	$\underline{83\,\%}$ $(289/350)$

Tabla 1. *Los porcentajes de recuperación calculados por cada grupo y por tratamiento (Global). Los datos en bruto que generan los porcentajes se muestran entre paréntesis. Se subraya el porcentaje más alto en cada fila.*

Por el momento, enfocamos en los efectos globales, que sugieren que el Tratamiento B sería mejor porque el $83\,\%$ es obviamente mayor que el $78\,\%$. Es posible afirmar, a través de una básica prueba estadística, que la diferencia es significativa. Sólo comparando estas cifras, es muy tentador concluir que el Tratamiento B conduce a un índice mejor de recuperación que el A, pero así perderíamos la otra parte de la historia.

Ahora nos limitamos el foco en los porcentajes entre el grupo con cálculos pequeños. Esta vez, el Tratamiento A parece superior al B. Pasa exactamente lo mismo con los números en el otro grupo con cálculos más grandes. ¿Las conclusiones se contradicen entre sí? O peor, ¿hemos cometido un error en alguna parte de nuestro análisis?

Sí y no. Este fenómeno es conocido como la *Paradoja de Simpson*. Se presenta frecuentemente cuando existe un factor de confusión (o una variable oculta) que es la causa común de las dos variables cuya relación queremos estudiar. En nuestro caso, se puede asumir un gráfico causal así en Figura 1.

Puesto que un paciente ya tiene cálculos grandes, lo más probable es que solicite Tratamiento A que sería más efectivo en casos graves. Por otro lado, es también más susceptible de beneficiarse menos del tratamiento y eso se refleja en un índice de recuperación más baja. Los pacientes con cálculos pe-

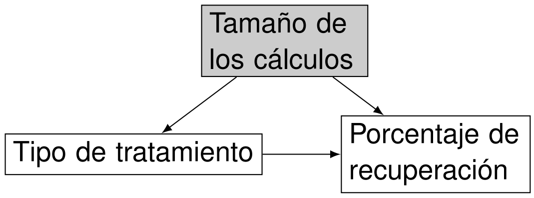


Figura 1. *Gráfico causal entre el tipo de tratamiento, el porcentaje de recuperación y el tamaño de los cálculos. La dirección de las flechas va de la causa al efecto.*

queños suelen recibir Tratamiento B, ya que no es necesario ser sometidos a una intervención quirúrgica, pero tienden a recuperarse mejor en general.

No es justificable asegurar, basado en los efectos globales, que Tratamiento B hace que aumente el porcentaje. Será más bien porque los pacientes que lo reciben no sufren tanto los cálculos de todos modos. Cada vez que intentamos sacar una conclusión causal, tenemos que pensar en los factores potenciales de confusión. Para calcular el verdadero efecto causal de los tratamientos en el índice de recuperación, hay que tener en cuenta (o controlar, o ajustar) el tamaño de los cálculos. (En este caso, es simplemente hacer la comparación fila por fila.)

Aunque sea clave tener mucho cuidado trabajando con causalidad, eso no quiere decir que no valga nada la asociación. Las asociaciones estadísticamente significativas, por lo general, ayudan a pronosticar las cantidades que nos interesan. Por ejemplo, si queremos saber algo del estado de recuperación de un paciente con Tratamiento A, se puede decir que es menos posible que se haya recuperado, en comparación con un paciente con Tratamiento B. Incluso podríamos utilizar la cifra $78\,\%$ para un cálculo numérico.

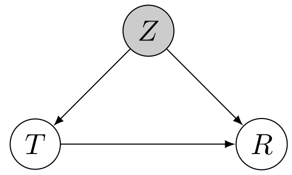
El mensaje final es que la relación causal (o el efecto de variable escondida) no cabe en el marco de la asociación o la correlación. A veces las conclusiones falsas son bastante graciosas, como «se producen más galardonados del Premio Nobel en países con más consumo de chocolate» (Messerli, 2012) (¿Qué podrá ser el factor de confusión aquí?). Pero estas conclusiones erróneas pueden llegar a ser perjudiciales o peligrosos, en caso de que se publiquen en diarios científicos o en informes gubernamentales. Así que, pensar lo bien antes de darlo por hecho, que la Paradoja de Simpson puede estar a todas partes.

Lectura Adicional

A los que se inclinan a matemáticas os puede interesar esta sección opcional.

Si nos diagnostican la existencia de los cálculos *sin conocimiento de su tamaño*, ¿cuál de los dos tratamientos deberíamos escoger? En otras palabras, con los datos observados en Tabla 1, ¿cómo podemos alcanzar el máximo de la probabilidad de recuperación?

Dibujamos el gráfico causal otra vez con las variables aleatorias binarias.



Aquí, Z es el tamaño de los cálculos (0 pequeño, 1 grande), T el tratamiento y R la recuperación (0 significa que no se recupere). Introducimos el símbolo $\Pr(E|do(I))$ que denota la probabilidad del evento E si hacemos la intervención I. Tened en cuenta que esta probabilidad no es necesariamente el mismo que la probabilidad condicional $\Pr(E|I)$. Resulta que la comparación entre $\Pr(R=1|do(T=\mathsf{A}))$ y $\Pr(R=1|do(T=\mathsf{B}))$ permita una decisión fidedigna.

El Ejemplo 6.37 en Peters y col. (2017) elabora el método correcto para calcular estos efectos. Más precisamente,

$$\begin{aligned}
\Pr(R=1|do(T=\mathsf{A})) &= \sum_{z=0}^{1} \Pr(R=1|T=\mathsf{A}, Z=z) \Pr(Z=z) \\
&\approx 0{,}93 \times \frac{357}{700} + 0{,}73 \times \frac{343}{700} = 0{,}832, \\
\Pr(R=1|do(T=\mathsf{B})) &= \sum_{z=0}^{1} \Pr(R=1|T=\mathsf{B}, Z=z) \Pr(Z=z) \\
&\approx 0{,}87 \times \frac{357}{700} + 0{,}69 \times \frac{343}{700} = 0{,}782.
\end{aligned}$$

Así concluimos que se prefiere el tratamiento A.

La derivación de esta fórmula supone la sabiduría de los conceptos en inferencia causal, incluyendo el modelado de ecuaciones estructurales (SEM), la do-intervención, los criterios de ajuste, etc. El libro de Peters y col. (2017) sirve para una introducción preliminar a los estudios de causalidad.

Referencias

Charig, C. R., Webb, D. R., Payne, S. R. & Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, *292*(6524), 879-882.

Messerli, F. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *The New England journal of medicine*, *367*, 1562-4.

Peters, J., Janzing, D. & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*.