

¿Cómo puede ser erróneo un resultado positivo?

Juan Segurola · Todos los derechos reservados

QUE · Artículo 047 · Preguntas sencillas, respuestas profundas · Versión 0.1 ES

Puede obtenerse un resultado positivo cuando la condición que se busca está ausente. Incluso una prueba que clasifica correctamente una gran proporción de casos puede generar muchos resultados así cuando la condición es poco frecuente en la población examinada. Para entender un resultado positivo, necesitamos contar tanto los verdaderos positivos como los falsos positivos que podrían haberlo producido.

Lo que sigue es un caso numérico inventado, no una prueba clínica ni un consejo sobre el diagnóstico de nadie. Usa frecuencias naturales para que los grupos pertinentes sigan siendo visibles. La investigación sobre razonamiento ha examinado cómo esta forma de representación puede facilitar las tareas de inferencia condicional, pero la aritmética aquí es independiente de cualquier afirmación sobre lo bien que vaya a rendir un lector concreto. [1]

1. Empecemos por la población, no por el resultado

Imaginemos 10 000 objetos que se examinan para detectar un defecto especificado. Un procedimiento de referencia independiente establece que 100 son defectuosos y 9 900 no lo son. Por tanto, la prevalencia del defecto es del 1% en esta población. Supongamos para este ejercicio que la clasificación de referencia es correcta.

Supongamos ahora que la prueba de cribado tiene una sensibilidad del 90%: da positivo en 90 de los 100 objetos defectuosos. No detecta los otros 10. Supongamos que su especificidad es del 95%: da negativo en 9 405 de los 9 900 objetos no defectuosos. Los 495 objetos no defectuosos restantes dan resultados positivos.

No hemos afirmado que el 90% o el 95% de los resultados positivos sean correctos. Esos porcentajes tienen denominadores diferentes. La sensibilidad parte de los objetos defectuosos; la especificidad parte de los no defectuosos. La pregunta que plantea un resultado positivo parte de los objetos que dieron positivo.

2. Cambiemos el denominador

Hay $90 + 495 = 585$ resultados positivos. Solo 90 corresponden al defecto. Por tanto, el valor predictivo positivo en esta población hipotética es $90/585$, aproximadamente el 15,38%.

Los 495 falsos positivos no constituyen evidencia de fraude ni de que la prueba no sirva para nada. Surgen de la tasa supuesta de falsos positivos del 5% aplicada a un grupo no defectuoso muy grande. Los 90 verdaderos positivos proceden de una tasa alta de detección aplicada a un grupo defectuoso mucho menor. Ambos recuentos deben entrar en el cálculo.

Esta es la trampa conceptual: «la prueba detecta el 90% de los defectos» responde a una pregunta sobre objetos que se sabe que son defectuosos. No responde a «dado un resultado positivo, ¿qué proporción de objetos son defectuosos?». Invertir el condicionamiento cambia la pregunta, aunque ambas frases contengan las mismas palabras: positivo y defectuoso.

El cálculo es exacto para los recuentos estipulados. En un estudio real, la sensibilidad, la especificidad y la prevalencia se estimarían, con incertidumbre y posible sesgo. Los enteros sencillos de aquí no deben confundirse con ese grado de certeza empírica.

3. Cambiemos un supuesto cada vez

Mantengamos fijas la sensibilidad y la especificidad de la prueba, pero dejemos que la prevalencia aumente al 10%. Entre 10 000 objetos, 1 000 son ahora defectuosos y 9 000 no lo son. Esperamos 900 verdaderos positivos y 450 falsos positivos, de modo que el valor predictivo positivo pasa a ser $900/1\,350$, aproximadamente el 66,67%.

Nada ha mejorado en la prueba supuesta. Lo que cambió fue la composición de la población examinada. Un resultado interpretado en una población no puede heredar sin más el valor predictivo medido en otra. Mantener constantes las características de la prueba es, en sí mismo, un supuesto de modelización; en la práctica, esas características también pueden diferir entre poblaciones y condiciones de funcionamiento.

Volvamos a una prevalencia del 1%, pero aumentemos ahora la especificidad al 99%, manteniendo la sensibilidad en el 90%. Sigue habiendo 90 verdaderos positivos, pero ahora solo 99 falsos positivos. El valor predictivo positivo pasa a ser $90/189$, aproximadamente el 47,62%. Esta mejora sustancial sigue sin hacer seguro un resultado positivo.

Estas dos variaciones separan dos factores sobre los que actuar: cambiar a quién examinamos y cambiar la tasa de falsos positivos. No son intervenciones intercambiables, y ninguno de los cálculos nos dice qué umbral de funcionamiento o política de pruebas es adecuado para una aplicación real.

4. Por qué repetir la prueba no es automáticamente una solución

Supongamos que una segunda prueba tiene las mismas tasas y que, condicionado al estado verdadero, sus errores son independientes de los de la primera prueba. Con prevalencia del 1%, el número esperado de objetos defectuosos positivos en ambas es $100 \times 0,9 \times 0,9 = 81$. El número esperado de objetos no defectuosos positivos en ambas es $9\,900 \times 0,05 \times 0,05 = 24,75$. La proporción correspondiente es $81/105,75$, aproximadamente el 76,60%.

El número no entero 24,75 es un recuento esperado a lo largo de repeticiones del proceso hipotético de muestreo, no una fracción de un objeto físico. Más importante aún, la independencia desempeña un papel crucial. Si la misma interferencia hace que ambas pruebas sean positivas, multiplicar las tasas de error puede exagerar mucho el beneficio de repetirlas.

En el extremo opuesto, si la segunda prueba simplemente reproduce todos los resultados de la primera, dos positivos no contienen más información que uno. El cálculo sigue siendo $90/585$. Los procedimientos reales de seguimiento requieren evidencia sobre la dependencia, el manejo de las muestras y el estándar de referencia; la repetición por sí sola no proporciona esa evidencia.

5. Lo que queda fuera de esta respuesta

Este ejemplo trata un estado binario y un resultado binario. No decide cómo definir una condición real, dónde fijar un umbral en una medición continua ni cómo ponderar los casos no detectados frente a las falsas alarmas. Esas decisiones requieren evidencia adicional y consecuencias explícitas.

Un falso positivo aquí tampoco es un valor p . Es una clasificación positiva incorrecta respecto al estado de referencia estipulado. Confundir ese suceso con una prueba de hipótesis de investigación o con una fabricación deliberada cambiaría de tema.

La lección intelectual práctica es preguntar por la población y los denominadores pertinentes antes de interpretar el resultado. Un resultado positivo es evidencia producida por un proceso con tasas de error especificadas, no una garantía. En este caso resuelto, la respuesta se aclara cuando contamos las vías alternativas por las que un objeto puede entrar en el grupo positivo.

Lectura relacionada

El artículo 046, «¿Por qué añadir una carretera puede empeorar el tráfico?», también muestra por qué un hecho local aparentemente favorable no determina el resultado final sin las condiciones que lo rodean.

Referencias

[1] Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. "Natural frequencies improve Bayesian reasoning in simple and complex inference tasks." *Frontiers in Psychology* 6, 1473 (2015). <https://doi.org/10.3389/fpsyg.2015.01473>