P-VALORES, α y β: ¿CUÁNTOS NOS EQUIVOCAMOS EN BIOLOGÍA?

P-values, α, and β: How wrong are we in Biology?

Autor: *Román-Palacios, C*[1].
[1]*Grupo de Investigaciones Entomológicas. Universidad del Valle, Facultad de Ciencias Exactas y Naturales, Departamento de Biología, Cali, Colombia.*

Correo correspondencia:
cristian.roman@correounivalle.edu.co

"…the function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable".
Colquhoun D. -1971

Encontrar un p-valor inferior a 0.05 no solo significa la descripción de un patrón "significativo" o no aleatorio (…y la culminación exitosa de muchos trabajos de investigación), implica en principio aceptar un 5% de falsos positivos, pero más aún, errar al menos un 30% de las veces en las conclusiones que de los análisis derivan.

Desde el desarrollo de la teoría tras los valores p por Ronald Fisher y su aplicación en pruebas de hipótesis, a cargo de Jerzy Neyman y Egon Pearson, la popularidad creciente de la utilidad de estos métodos cuantitativos para soportar ó refutar hipótesis, ha hecho poco conspicuos los notables problemas en los cuales se incurre durante uso. El valor p ha sido usualmente definido como "la probabilidad de hallar un efecto tan o más extremos que el observado, asumiendo como verdadera la hipótesis nula (Ho: no efecto)". Es decir, representa una medida de la intensidad de la evidencia en contra de la hipótesis nula sin dar indicio alguno sobre Ha.

Usualmente un p = 0.05 es considerado como una medida aceptable de la certeza que se tiene sobre los resultados y subsecuentes conclusiones. De esta forma se reconoce que en un 5% de los casos Ho será rechazada, aún siendo verdadera (Error tipo I; α). Por otro lado, los falsos negativos representan el escenario opuesto en donde no se rechaza Ho siendo falsa (Error tipo II; β). Controlar el efecto que ambos tipos de error tienen sobre la estimación de los valores p representaría un escenario ideal que daría soporte a resultados estadísticamente "significativos". Desafortunadamente, el efecto en la estructura de las hipótesis frecuentistas limita al investigador a controlar únicamente el Error tipo I (α= 0.05), dejando a un lado el calculo de la probabilidad del Error tipo II. La hipótesis nula (Ho) se escribe en términos de un valor puntual (eg. σ=0), pero la alterna (Ha) recoge

un conjunto de posibilidades que hacen de esta una interpretación general del patrón (eg. $\sigma \neq 0$). ¿Existe entonces "significancia" estadística bajo este paradigma incompleto? Hasta ahora, el único error claro corresponde al 5% de falsos positivos que se designan con el valor α, lo cual no parece excesivamente alto. Si evita la asunción de selección aleatoria de los sujetos a uno u otro grupo (para no invalidar un gran numero de trabajos) y se encuentra un valor p suficientemente bajo ($p<0.05$), parecería razonable indicar que existen un efecto no aleatorio en los datos. Esto, en principio solo implicaría errar 1 de cada 20 test, pero no es lo que sucede. Los valores p hacen exactamente lo que se plantea: son una propuesta sobre que sucedería si no existiera efecto verdadero, pero no permiten una generalización de todo el panorama.

El tipo de ejemplos clásicos como tratamientos experimentales son valiosos recursos para ejemplificar el problema. En un protocolo de laboratorio se establecen 1.000 acuarios con la misma especie de pez, pero variando la especie vegetal en cada uno. Tras un par de meses se encontró que el 10% de los acuarios contenía aún el pez vivo. Esto deriva a la obvia conclusión de sobrevivencia en 100 acuarios y el efecto contrario en los 900 restantes. En este sentido, los 900 acuarios representarían el caso en el cual no hay efecto real (Ho es realmente cierta), pero si asumimos un nivel de significancia del 5%, 45 de estos acuarios serian falsos positivos. 855 acuarios (95% de los 900) serian verdaderos negativos.

Hasta ahora, se ha resumido la mitad de la historia: solamente se ha hablado de la significancia haciendo énfasis únicamente a los 900 eventos donde no se encontró efecto real. ¿Qué sucede entonces con los casos en los que si se encontró un efecto?, ¿no se analiza ese 10% restante?. Para estos 100 acuarios no aplica la "significancia" estadística pues la hipótesis nula es falsa, pero si lo hace el poder de la prueba. Este depende parámetro del tamaño muestral y del tamaño del efecto que se pretende analizar, indicando la sensibilidad para detectar un efecto correcto cuando en realidad existe. Un poder de 0.8 es una medida "aceptablemente buena" que indica correctamente un efecto "significativo" en el 80% de los casos donde hay un efecto real. En los acuarios, 80 de estos (80% de los 100 acuarios) serán verdaderos positivos, es decir, un efecto directo de la planta sobre la sobrevivencia del pez, pero un 20% serán resultado de otro efecto diferente (20% de los 100 acuarios).

Lo importante entonces es como analizar ambas partes para dar una medida de confidencia: 45 falsos positivos y 80 verdaderos positivos. La tasa de falso descubrimiento supone una forma eficiente de establecer una relación clara entre el numero de falsos positivos y el total de eventos positivos. En este sentido, el numero de test positivos seria 125 (80+45), lo cual implicaría un 36% (45/125) de error en los cálculos. No es por lo tanto equivalente al 5% que se espera usualmente (α), sino 620% más, al menos para este experimento.

Es importante por lo tanto procurar también incluir análisis sobre el poder estadístico de la prueba, y analizar el error global en el cual se incurre durante los análisis. Trabajar con un 5% de significancia no implica un 5% de error, haciendo la palabra "significativo" invalida en las conclusiones que derivan de un $p<0.05$. En muchas casos, áreas enteras han procurado como alternativa a este efecto trabajar con valores p mucho más bajos, que controlan el efecto que los falsos positivos tienen sobre la estimación de la tasa de falso descubrimiento. Un $p=0.001$ tiene una tasa de falso descubrimiento ~1.8% y $p=0.0027$ corresponde a una tasa de falso descubrimiento del 4.2%, es decir, es al menos 15 veces menos probable descubrir algo que no existe. Hay diferentes alternativas que pueden ser usadas como alternativas al 0.05 y dan una mejor descripción de los patrones que se desean evaluar.