5.- El valor P está obsoleto.

The P value is out of date

José Antonio González[1],
jose.a.gonzalez@upc.edu
1 Departamento de Estadística e Investigación Operativa, Barcelona-Tech, UPC

En esta píldora seguimos las guías de publicación[1] y la declaración de la American Statistical Association (ASA)[2]. Pretendemos convencerle de que no use el valor de P (o *P-value*). No es tarea fácil. Pero seamos sinceros: tampoco habíamos entendido bien por qué el valor de P era tan importante. Esta discusión dura ya casi 100 años. Proponemos no resistirnos más al cambio. Puede ser un buen momento para intentar reflejar la incertidumbre de otra manera.

Probabilidad condicionada.

Para entender el valor de P necesitamos entender primero el concepto de probabilidad condicionada. No es lo mismo, dentro del grupo de enfermos, la probabilidad de dar positivo (sensibilidad) que, dentro del grupo de positivos, la probabilidad de estar realmente enfermo (valor predictivo de un resultado positivo). Son tan diferentes que los simbolizamos de forma distinta: la probabilidad de que un enfermo resulte positivo la escribimos P(+|E), que leemos «probabilidad de positivo entre los enfermos»; y la probabilidad de que un positivo esté realmente enfermo la escribimos P(E|+), «probabilidad de enfermo entre los positivos». A la derecha de la barra, lo que sabemos («conocido»); y a la izquierda, lo que es incierto.

		Diagnóstico		
		Positivo	Negativo	
Realidad	Enfermo	90	10	100
	Sano	110	790	900
		200	800	1000

Tabla 1. Situación real del paciente (enfermo o sano) y resultado del indicador diagnóstico (positivo o negativo). Los datos del ejemplo son puramente ilustrativos.

Imaginemos los resultados de la tabla 1. P(+|E) es la proporción de positivos sobre el total de la fila *Enfermo*; y P(E|+), la proporción de enfermos sobre el total de la columna *Positivo*. Veámoslo con detalle. En estos datos, 100 enfermos y 900 sanos. Del total de 100 enfermos, 90 dieron positivo: 90% de sensibilidad o P(+|E). Bien. Del total de 900 sanos, 790 dieron negativo: 88% de especificidad o P(-|S). Bien también. Además, de los 800 negativos, 790 estaban sanos: P(S|-) = 99%, genial. Pero de los 200 positivos, solo 90 estaban enfermos: P(E|+) = 45%, no tan bien. A pesar de que comparten numerador (90 positivos enfermos), como hay más positivos que enfermos, P(E|+) = 45% es mucho menor que P(+|E) = 90%.

Stephen Senn lo ejemplifica genialmente: la probabilidad de que un católico sea Papa es muy baja, pero la de que el Papa sea católico es del 100%.

Ojalá el valor de P nos dijera la probabilidad de que la hipótesis H sea cierta dados los datos, P(H|datos). Pero el valor de P no responde esta pregunta, sino una entelequia más sofisticada: «asumiendo cierta una hipótesis nula contraria a lo que deseo demostrar, el valor de P es la probabilidad de observar unos resultados como estos, o más extremos». Hemos dicho entelequia porque: (1) falta definir qué se entiende por «más extremos»; (2) no hemos observado estos resultados «más extremos»; y (3) no hemos aclarado qué creemos sobre esta «hipótesis nula».

Importancia de la potencia y el soporte previo.

En 2016, Cortés, González, Langohr y Casals[3] explicaron a los lectores de Medicina Clínica la poca credibilidad del valor de P en ausencia de: (1) soporte previo para la hipótesis; o (2) de un diseño con suficiente potencia. Volvamos a la analogía del diagnóstico. Para tener un alto valor predictivo positivo, el proceso diagnóstico requiere: (1) sensibilidad; (2) especificidad; y (3) un soporte clínico previo de la verosimilitud del diagnóstico sospechado. Si se cumplen estos 3 requisitos, podremos confiar en que el resultado positivo proviene de un caso enfermo. De la misma forma, confiar en que una P corresponde a una hipótesis cierta, requiere también de: (1) potencia; (2) una P cuanto más pequeña mejor; y (3) soporte de la hipótesis. Por ejemplo, el desarrollo de un fármaco tiene su cumbre en el ensayo clínico

confirmatorio, pivote o de fase III. Con un desarrollo previo para dar soporte a sus efectos. Y con un cálculo del tamaño requerido para garantizar la potencia del estudio.

Diferencia entre el valor de P y el riesgo alfa.

En 2003, Hubbard y Bayarri[4] alertaron de la confusión frecuente entre P y alfa, entre la prueba de significación de Fisher y el contraste de hipótesis de Neyman-Pearson. Fisher propuso su prueba de significación para resolver un problema de evidencia, de conocimiento, de inferencia estadística. En una palabra, de Ciencia. En cambio, Neyman y Pearson propusieron su contraste de hipótesis para decidir entre dos acciones alternativas en situación de incertidumbre. Por ejemplo, ¿deben las autoridades autorizar o no un nuevo fármaco? La naturaleza dicotómica de la pregunta (autorizar: ¿sí o no?) requiere un límite, un umbral, a partir del cual sí se autorizará la intervención. O, si no se alcanza este umbral, no se autoriza. Para poder tomar una decisión, este contraste de hipótesis requiere limitar sendos riesgos de adoptar decisiones erróneas, denominados riesgos alfa y beta, correspondientes con los llamados errores de tipo I y II, respectivamente autorizar el fármaco cuando no tiene el efecto hipotético; y no autorizarlo cuando sí lo tiene. Para delimitar el riesgo beta, el contraste de hipótesis ha de definir un valor para la hipótesis alternativa. En el ejemplo del fármaco, un tamaño del efecto llamado delta, que centra la hipótesis alternativa en un único valor, lo que permite fijar el tamaño del estudio para cierta potencia deseada. Así, un ensayo clínico pivote o decisorio, (1) fija un efecto de la intervención «delta»; (2) basa su cálculo del tamaño en este efecto delta; y (3) explica los estudios previos que dan soporte a delta. En todo caso, recuerde que un valor de P, por muy pequeño que sea, no dice nada sobre la potencia.

Posiciones en contra del valor P

La declaración[2] de la Asociación de Estadísticos Americanos se pronunció en 2016 en contra del valor de P. Pidió no interpretarla como P(H) o P(H|datos): «Los p-valores no miden la probabilidad de que la hipótesis estudiada sea verdadera».

En 2019, Nature pidió dejar de usar el término «estadísticamente significativo»[5]. Y NEJM comentó que ya habían empezado a usarlo de forma más parsimoniosa[6], y solo para el análisis principal que estaba justificado por el cálculo del tamaño del estudio.

Tanto el valor de P como «estadísticamente significativo» han contribuido al gran avance de la Ciencia en los últimos 100 años. Pero ha llegado el momento de jubilarlos. Algo no hacemos bien cuando el 85% de la inversión en salud no termina en resultados reproducibles[7]. La confusión de las probabilidades condicionadas que explicábamos al inicio puede facilitar la creencia de que, si $P<0.05$, la hipótesis contrastada era cierta, y, por tanto, resultados similares serían reproducidos por autores futuros. Razonamiento falso, originado en una interpretación errónea del valor de P; y también por la confusión entre P −prueba de significación− y alfa −contraste de hipótesis. La dicotomía, según sea $P<0.05$ o no, es irracional. La ciencia, a diferencia de las decisiones, no tiene puntos de corte.

Para obtener más información puede acceder a la presentación que mi compañero Erik Cobo realizó en un seminario conjunto[8] de las Sociedades de Estadística e Investigación Operativa (SEIO) y Española de Bioestadística (SEB)[9], celebrado en abril del año 2021.

Conclusión: propuesta de las guías de publicación

Así, ¿qué debemos reportar en lugar del valor de P? CONSORT apareció en 1996, hace un cuarto de siglo. Pide reportar el tamaño del efecto, junto con medidas de su incertidumbre, por delante de los valores de P. Propone usar los clásicos intervalos de confianza. A fin de cuentas, en un ensayo clínico, los investigadores desean contestar a la pregunta: «si cambio la forma de hacer las cosas, si cambio la intervención de referencia por la intervención en estudio, ¿cuánto mejorarán mis pacientes?». Esta diferencia de la evolución entre los tratados y los controles es el tamaño del efecto («effect size»). Al que añadiremos un intervalo que refleje la incertidumbre: estos resultados observados en esta muestra, ¿con qué valores poblacionales son probabilísticamente compatibles?

Estimar este tamaño del efecto es el objetivo de un ensayo clínico. Otros diseños y otras guías de publicación aconsejan otras medidas, de acuerdo con su objetivo. Pero siempre nos pedirán acompañarlas con un intervalo de compatibilidad entre parámetro y datos.

Referencias

[1] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D,Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996 Aug 28;276(8):637-9.

[2] The ASA Statement on p-Values: Context, Process, and Purpose", https://doi.org/10.1080/00031305.2016.1154108)

[3] Cortés J, González JA, Langohr K y Casals M. Importancia de la potencia y la hipótesis en el valor p, Med Clin Vol. 146. Núm. 4.; páginas 178-181.

[4] Hubbard R & Bayarri M J (2003) Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing, The American Statistician, 57:3, 171-178, DOI: 10.1198/0003130031856

[5] Amrhein V, Greenland S and McShane B. Scientists rise up against statistical significance. Nature 567, 305-307 (2019). doi: 10.1038/d41586-019-00857-9

[6] New Guidelines for Statistical Reporting in the Journal, N Engl J Med 2019; 381:285-286; DOI: 10.1056/NEJMe1906559

[7] Chalmers I1, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009 Jul 4;374(9683):86-9. doi: 10.1016/S0140-6736(09)60329-9. Epub 2009 Jun 12.

[8] https://www.youtube.com/watch?v=66alZO532tk&t=1s

[9] https://www.youtube.com/watch?v=p4TqFerAaA4