Abandonar la significación estadística

Valentin Amrhein, Sander Greenland, Blake McShane y más de 800 firmantes

Nature vol. 567, pp. 305-307, 21 marzo 2019

¿Cuándo fue la última vez que escuchó a un orador de un seminario afirmar que "no había diferencia" entre dos grupos porque la diferencia era "estadísticamente no significativa"? Si su experiencia coincide con la nuestra, es muy probable que esto sucediera en la última charla a la que asistió. Esperamos que al menos alguien en la audiencia se quedara perplejo si, como sucede con frecuencia, una gráfica o una tabla mostraran que realmente había una diferencia.

¿Cómo es que las estadísticas llevan tan a menudo a los científicos a negar diferencias que aquellos que no están educados en estadística pueden ver claramente? Durante varias generaciones, los investigadores han sido advertidos de que un resultado estadísticamente no significativo no "prueba" la hipótesis nula (la hipótesis de que no hay diferencia entre los grupos o ningún efecto de un tratamiento sobre algún resultado medido)[1]. Los resultados estadísticamente significativos tampoco "prueban" alguna otra hipótesis. Tales conceptos erróneos han deformado la literatura con afirmaciones exageradas y, de manera menos célebre, han llevado a afirmar conflictos entre estudios donde no había ninguno. Tenemos algunas propuestas para evitar que los científicos sean víctimas de estos conceptos erróneos.

Un problema omnipresente

Seamos claros sobre lo que debemos detenernos: nunca tenemos que concluir que "no hay diferencia" o "no hay asociación" solo porque un valor de P es mayor que un umbral como 0,05 o, de manera equivalente, porque un intervalo de confianza incluye cero. Tampoco debemos concluir que dos estudios entran en conflicto porque uno tuvo un resultado estadísticamente significativo y el otro no. Estos errores desperdician los esfuerzos de investigación y desinforman las decisiones políticas. Por ejemplo, considere una serie de análisis de los efectos no deseados de los medicamentos antiinflamatorios[2]. Debido a que sus resultados no fueron estadísticamente significativos, un grupo de investigadores concluyó que la exposición a los medicamentos "no se asoció" con la fibrilación auricular de nueva aparición (la alteración más común del ritmo cardíaco) y que los resultados contrastaban con los de una estudio anterior con un resultado estadísticamente significativo.

[1] Fisher, R. A. Nature **136**, 474 (1935).
[2] Schmidt, M. & Rothman, K. J. Int. J. Cardiol. **177**, 1089–1090 (2014).

Ahora, veamos los datos reales. Los investigadores que describieron sus resultados estadísticamente no significativos encontraron un índice de riesgo de 1.2 (es decir, un riesgo 20% mayor en los pacientes expuestos en comparación con los no expuestos). También encontraron un intervalo de confianza del 95% que abarcaba todo, desde una disminución insignificante del riesgo del 3% hasta un aumento considerable del riesgo del 48% (P = 0,091; cálculo nuestro). Los investigadores del estudio anterior, estadísticamente significativo, encontraron exactamente la misma razón de riesgo de 1,2. Ese estudio fue simplemente más preciso, con un intervalo que va desde un 9% a un 33% más de riesgo (P = 0,0003; cálculo nuestro).

Es ridículo concluir que los resultados estadísticamente no significativos mostraron "ninguna asociación", cuando la estimación del intervalo incluyó aumentos graves del riesgo; es igualmente absurdo afirmar que estos resultados contrastaban con los resultados anteriores que mostraban un efecto observado idéntico. Sin embargo, estas prácticas comunes muestran cómo la confianza en los umbrales de significación estadística puede inducirnos a error (ver "Cuidado con las conclusiones falsas").

Estos y otros errores similares están muy extendidos. Encuestas de cientos de artículos han encontrado que los resultados estadísticamente no significativos se interpretaban como una indicación de "ninguna diferencia" o "ningún efecto" en aproximadamente la mitad (consulte "Interpretaciones incorrectas" e Información complementaria). En 2016, la Asociación Estadounidense de Estadística publicó una declaración en The American Statistician advirtiendo contra el uso indebido de la significación estadística y los valores P. El número también incluyó muchos comentarios sobre el tema. Este mes, un número especial de la misma revista intenta impulsar estas reformas. Presenta más de 40 artículos sobre "Inferencia estadística en el siglo XXI: un mundo más allá de P <0,05". Los editores presentan la colección con la advertencia "no digas 'estadísticamente significativo'"[3].

Otro artículo[4] con docenas de firmantes también pide a los autores y editores de revistas que rechacen esos términos Estamos de acuerdo y piden que se abandone todo el concepto de significación estadística. Estamos lejos de estar solos. Cuando invitamos a otros a leer un borrador de este comentario y firmar sus nombres si estaban de acuerdo con nuestro mensaje, 250 lo hicieron dentro de las primeras 24 horas. Una semana después, teníamos más de 800 firmantes, todos revisados por una afiliación académica u otra indicación de trabajo actual o pasado en un campo que depende de modelos estadísticos (consulte la lista y el recuento final de signatarios en la Información complementaria). Estos incluyen estadísticos, investigadores clínicos y médicos, biólogos y psicólogos de más de 50 países y de todos los continentes excepto la Antártida. Un defensor lo calificó como un "ataque quirúrgico contra las pruebas irreflexivas de significación estadística" y "una oportunidad para hacer oír su voz a favor de mejores prácticas científicas".

[3] Wasserstein, R. L., Schirm, A. & Lazar, N. A. Am. Stat. https://doi.org/10.1080/00031305.20 19.1583913 (2019).
[4] Hurlbert, S. H., Levine, R. A. & Utts, J. Am. Stat. https://doi.org/10.1080/00031305.2018.1543 616 (2019).

No estamos pidiendo la prohibición de los valores P. Tampoco estamos diciendo que no puedan utilizarse como criterio de decisión en determinadas aplicaciones especializadas (como determinar si un proceso de fabricación cumple con algún estándar de control de calidad). Y tampoco estamos abogando por una situación de todo vale, en la que la evidencia débil de repente se vuelve creíble. Más bien, y en línea con muchos otros a lo largo de las décadas, estamos pidiendo que se detenga el uso de los valores P de la manera convencional y dicotómica, para decidir si un resultado refuta o apoya una hipótesis científica[5].

Salir de la categorización

El problema es más humano y cognitivo que estadístico: clasificar los resultados en "estadísticamente significativos" y "estadísticamente no significativos" hace que la gente piense que los elementos asignados de esa manera son categóricamente diferentes[6,7,8]. Es probable que surjan los mismos problemas bajo cualquier alternativa estadística propuesta que implique dicotomización, ya sea frecuentista, bayesiana o de otro tipo.

Desafortunadamente, la falsa creencia de que cruzar el umbral de la significación estadística es suficiente para demostrar que un resultado es "real" ha llevado a los científicos y editores de revistas a privilegiar tales resultados, distorsionando así la literatura. Las estimaciones estadísticamente significativas están sesgadas hacia arriba en magnitud y potencialmente en gran medida, mientras que las estimaciones estadísticamente no significativas están sesgadas hacia abajo en magnitud. En consecuencia, cualquier discusión que se centre en las estimaciones elegidas por su importancia estará sesgada. Además de esto, el enfoque rígido en la significación estadística anima a los investigadores a elegir datos y métodos que produzcan significación estadística para algún resultado deseado (o simplemente publicable), o que produzcan una no significación estadística para un resultado no deseado, como los posibles efectos secundarios de drogas, invalidando así las conclusiones.

El prerregistro de estudios y el compromiso de publicar todos los resultados de todos los análisis pueden contribuir en gran medida a mitigar estos problemas. Sin embargo, incluso los resultados de estudios prerregistrados pueden estar sesgados por decisiones que invariablemente se dejan abiertas en el plan de análisis[9]. Esto ocurre incluso con la mejor de las intenciones.

Nuevamente, no abogamos por la prohibición de los valores P, los intervalos de confianza u otras medidas estadísticas, solo que no debemos tratarlos de manera categórica. Esto

[5] Lehmann, E. L. Testing Statistical Hypotheses 2nd edn 70–71 (Springer, 1986).
[6] Gigerenzer, G. Adv. Meth. Pract. Psychol. Sci. **1**, 198–218 (2018).
[7] Greenland, S. Am. J. Epidemiol. **186**, 639–645 (2017)
[8] McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Am. Stat. https://doi.org/10.1080/0 0031305.2018.1527253 (2019)
[9] Gelman, A. & Loken, E. Am. Sci. **102**, 460–465 (2014)

incluye la dicotomización como estadísticamente significativa o no, así como la categorización basada en otras medidas estadísticas como los factores de Bayes.

Una razón para evitar esta "dicotomanía" es que todas las estadísticas, incluidos los valores de P y los intervalos de confianza, varían naturalmente de un estudio a otro y, a menudo, lo hacen en un grado sorprendente. De hecho, la variación aleatoria por sí sola puede conducir fácilmente a grandes disparidades en los valores de P, mucho más allá de caer a ambos lados del umbral de 0,05. Por ejemplo, incluso si los investigadores pudieran realizar dos estudios de replicación perfecta de algún efecto genuino, cada uno con un 80% de poder (probabilidad) de lograr P <0,05, no sería muy sorprendente que uno obtuviera P <0,01 y el otro P> 0,30. Ya sea que el valor de P sea pequeño o grande, se debe tener precaución.

Debemos aprender a aceptar la incertidumbre. Una forma práctica de hacerlo es cambiar el nombre de los intervalos de confianza como "intervalos de compatibilidad" e interpretarlos de una manera que evite el exceso de confianza. Específicamente, recomendamos que los autores describan las implicaciones prácticas de todos los valores dentro del intervalo, especialmente el efecto observado (o estimación puntual) y los límites. Al hacerlo, deben recordar que todos los valores entre los límites del intervalo son razonablemente compatibles con los datos, dados los supuestos estadísticos utilizados para calcular el intervalo[10,11]. Por lo tanto, señalar un valor en particular (como el valor nulo) en el intervalo como "mostrado" no tiene sentido.

Francamente, estamos hartos de ver "pruebas de la nulidad" tan absurdas y afirmaciones de no asociación en presentaciones, artículos de investigación, reseñas y materiales de instrucción. Un intervalo que contiene el valor nulo a menudo también contendrá valores no nulos de gran importancia práctica. Dicho esto, si considera que todos los valores dentro del intervalo son prácticamente insignificantes, entonces podría decir algo como "nuestros resultados son más compatibles sin efectos importantes".

Cuando se habla de intervalos de compatibilidad, tenga en cuenta cuatro cosas. Primero, el hecho de que el intervalo proporcione los valores más compatibles con los datos, dados los supuestos, no significa que los valores fuera de él sean incompatibles; simplemente son menos compatibles. De hecho, los valores que se encuentran justo fuera del intervalo no difieren sustancialmente de los que están justo dentro del intervalo. Por tanto, es incorrecto afirmar que un intervalo muestra todos los valores posibles.

En segundo lugar, no todos los valores del interior son igualmente compatibles con los datos, dados los supuestos. La estimación puntual es la más compatible y los valores cercanos a ella son más compatibles que los que están cerca de los límites. Es por eso que instamos a los autores a discutir la estimación puntual, incluso cuando tengan un valor P grande o un intervalo amplio, así como a discutir los límites de ese intervalo. Por ejemplo, los autores anteriores podrían haber escrito: "Al igual que en un estudio anterior, nuestros

[10] Cf. nota 7.
[11] Amrhein, V., Trafimow, D. & Greenland, S. Am. Stat. https://doi.org/10.1080/00031305.2018.1543137 (2019).

resultados sugieren un aumento del 20% en el riesgo de fibrilación auricular de nueva aparición en pacientes que recibieron medicamentos antiinflamatorios. No obstante, una diferencia de riesgo que va desde una disminución del 3%, una pequeña asociación negativa, hasta un aumento del 48%, una asociación positiva sustancial, también es razonablemente compatible con nuestros datos, dadas nuestras suposiciones." Interpretar la estimación puntual, reconociendo su incertidumbre , evitará que haga declaraciones falsas de "no diferencia" y que haga afirmaciones excesivamente confiadas.

En tercer lugar, al igual que el umbral de 0,05 del que procede, el 95% predeterminado que se utiliza para calcular los intervalos es en sí mismo una convención arbitraria. Se basa en la idea falsa de que existe un 95% de probabilidad de que el intervalo calculado contenga el valor verdadero, junto con la vaga sensación de que esto es una base para una decisión segura. Se puede justificar un nivel diferente, dependiendo de la aplicación. Y, como en el ejemplo de los fármacos antiinflamatorios, las estimaciones de intervalo pueden perpetuar los problemas de significación estadística cuando la dicotomización que imponen se trata como un estándar científico.

Por último, y lo más importante de todo, sea humilde: las evaluaciones de compatibilidad dependen de la exactitud de los supuestos estadísticos utilizados para calcular el intervalo. En la práctica, estos supuestos están sujetos, en el mejor de los casos, a una considerable incertidumbre[12]. Haga estas suposiciones lo más claras posible y pruebe las que pueda, por ejemplo, trazando sus datos y ajustando modelos alternativos, y luego informando de todos los resultados.

Independientemente de lo que muestren las estadísticas, está bien sugerir razones para sus resultados, pero discuta una variedad de posibles explicaciones, no solo las favorecidas. Las inferencias deben ser científicas, y eso va mucho más allá de lo meramente estadístico. Los factores como la evidencia de antecedentes, el diseño del estudio, la calidad de los datos y la comprensión de los mecanismos subyacentes suelen ser más importantes que las medidas estadísticas como los valores o intervalos de P. La objeción que más escuchamos contra la retirada de la significación estadística es que es necesaria para tomar decisiones de sí o no. Pero para las elecciones que a menudo se requieren en entornos regulatorios, políticos y comerciales, las decisiones basadas en los costos, beneficios y probabilidades de todas las consecuencias potenciales siempre superan a las que se toman basándose únicamente en la importancia estadística. Además, para las decisiones sobre si se debe seguir adelante con una idea de investigación, no existe una conexión simple entre un valor P y los resultados probables de estudios posteriores.

¿Cómo será la retirada de la significación estadística? Esperamos que las secciones de métodos y la tabulación de datos sean más detalladas y matizadas. Los autores enfatizarán sus estimaciones y la incertidumbre en ellas, por ejemplo, al discutir explícitamente los límites inferior y superior de sus intervalos. No dependerán de pruebas de significación. Cuando se informan los valores de P, se darán con precisión sensible (por ejemplo, P = 0,021 o P = 0,13), sin adornos como estrellas o letras para denotar significancia estadística

[12] Cf. notas 7, 8 y 11.

y no como desigualdades binarias (P <0,05 o P> 0,05). Las decisiones de interpretar o publicar resultados no se basarán en umbrales estadísticos. La gente dedicará menos tiempo al software estadístico y más tiempo a pensar.

Nuestra apelación a retirar la significación estadística y utilizar intervalos de confianza como intervalos de compatibilidad no es una panacea. Aunque eliminará muchas malas prácticas, bien podría introducir otras nuevas. Por lo tanto, el seguimiento de la literatura en busca de abusos estadísticos debe ser una prioridad permanente para la comunidad científica. Pero la erradicación de la categorización ayudará a detener las afirmaciones excesivamente seguras, las declaraciones injustificadas de "no hay diferencia" y las declaraciones absurdas sobre "fallas en la replicación" cuando los resultados de los estudios originales y de replicación son altamente compatibles. El mal uso de la significación estadística ha causado mucho daño a la comunidad científica y a quienes confían en el asesoramiento científico. Los valores P, los intervalos y otras medidas estadísticas tienen su lugar, pero es hora de que desaparezca la significación estadística. ■

Valentin Amrhein es profesor de zoología en la Universidad de Basilea, Suiza.

Sander Greenland es profesor de epidemiología y estadística en la Universidad de California, Los Angeles.

Blake McShane es metodólogo estadístico y profesor de marketing en la Universidad Northwestern en Evanston, Illinois.

Para obtener una lista completa de firmantes, consulte: go.nature.com/2tc5nkm

correo electrónico: v.amrhein@unibas.ch

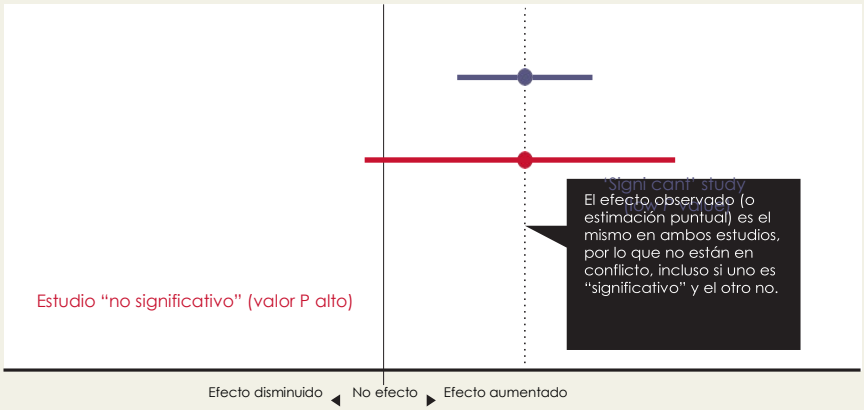
URL permanente:

https://www.researchgate.net/profile/Valentin_Amrhein/publication/331908769_Scientists_rise_up_against_statistical_significance/links/5c9685a9299bf11169438d0a/Scientists-rise-up-against-statistical-significance.pdf?origin=publication_detail

Recuadros

CUIDADO CON LAS FALSAS CONCLUSIONES

Los estudios que actualmente se denominan “estadísticamente significativos” y “estadísticamente no significativos” no tienen por qué ser contradictorios, y tales designaciones pueden hacer que se descarten efectos genuinos.



INTERPRETACIONES INCORRECTAS

Un análisis de 791 artículos en 5 revistas * descubrió que alrededor de la mitad asume erróneamente que la ausencia de significación significa que no hay efecto.

Datos tomados de: P. Schatz *et al. Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* **33**, 1–15 (2017).

Traducción: Francesc J. Hernàndez (Universitat de València)