

Declaración de la Asociación de Estadística Americana Sobre la Significancia Estadística y Los Valores- P

Editado por Ronald L. WASSERSTEIN

Introducción

El aumento del número de investigaciones científicas y la proliferación de largos y complejos conjuntos de datos en los años recientes han expandido el alcance en las aplicaciones de métodos estadísticos. Esto ha creado nuevos caminos para el progreso científico, pero también trae aparejadas preocupaciones acerca de algunas conclusiones obtenidas a partir de datos de investigación. La validez de las conclusiones científicas, incluyendo su reproducibilidad, no dependen únicamente de los métodos estadísticos utilizados. La elección de las técnicas apropiadas, la conducción de los análisis de forma adecuada y la correcta interpretación de los resultados estadísticos también juegan un papel importante en garantizar que las conclusiones obtenidas sean confiables y que la incertidumbre asociada a ellas esté representada apropiadamente.

Detrás de muchas conclusiones de publicaciones científicas subyace el concepto de “significancia estadística,” típicamente evaluada con el índice llamado valor- p . Si bien el valor- p puede ser una medida estadística útil, generalmente es utilizado e interpretado incorrectamente. Esto ha llevado a que algunas revistas científicas desalienten el uso de los valores- p , y a que algunos científicos y estadísticos recomienden su abandono, con ciertos argumentos que prácticamente no han cambiado desde que los valores- p fueron introducidos por primera vez.

En este contexto, la Asociación de Estadística Americana (ASA por sus siglas en Inglés) cree que la comunidad científica podría beneficiarse de una declaración formal que clarifique varios principios ampliamente aceptados subyacentes al adecuado uso e interpretación de los valores- p . Los temas tratados brevemente aquí afectan no sólo a la investigación, sino también al financiamiento de la misma, a las prácticas de escritura científica, al progreso profesional, a la educación científica, a las políticas públicas, al periodismo y a la ley. Esta declaración no busca resolver todos los problemas relacionados con la práctica estadística sensata, ni tampoco las controversias fundamentales; sino que, articula en términos no técnicos algunos principios que pueden mejorar la conducta o interpretación de la ciencia cuantitativa, de acuerdo con el consenso extendido en la comunidad científica.

Ronald L. Wasserstein, Director Ejecutivo, En representación de la Junta Directiva de la Asociación de Estadística Americana. Traducido al Español por: Fiorella Laco Mazzone, Maria Grampa, Matías Goldenberg, Francisco Aristimuño, Facundo Oddi, y Lucas A. Garibaldi. Instituto de Investigaciones en Recursos Naturales, Agroecología y Desarrollo Rural (IRNAD), Sede Andina, Universidad Nacional de Río Negro (UNRN) and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Mitre 630, CP 8400, San Carlos de Bariloche, Río Negro, Argentina.

¿Qué es un valor- p ?

Informalmente, un valor- p es la probabilidad bajo un modelo estadístico específico de que un resumen estadístico de los datos (por ejemplo, la diferencia de la media muestral entre dos grupos comparados) sea igual que o más extremo que su valor observado.

Principios

1. *Los valores- p pueden indicar qué tan incompatibles son los datos con un modelo estadístico específico.*

Un valor- p proporciona un enfoque para resumir la incompatibilidad entre un conjunto de datos particular y un modelo propuesto para dichos datos. El contexto más común es un modelo, construido bajo un conjunto de supuestos, junto con una presunta “hipótesis nula.” Comúnmente, la hipótesis nula postula la ausencia de un efecto, por ejemplo que no haya diferencia entre dos grupos o la ausencia de una relación entre un factor y un resultado. Cuanto menor sea el valor- p , mayor será la incompatibilidad estadística de los datos con la hipótesis nula, siempre y cuando los supuestos subyacentes utilizados para calcular el valor- p sean adecuados. Esta incompatibilidad puede ser interpretada como un elemento que genera duda sobre la veracidad, o proveer evidencia en contra, de la hipótesis nula o los supuestos subyacentes.

2. *Los valores- p no miden la probabilidad de que la hipótesis estudiada sea verdadera o la probabilidad de que el conjunto de datos haya sido generado por pura aleatoriedad.*

Los investigadores habitualmente desean convertir un valor- p en una declaración sobre la veracidad de una hipótesis nula o sobre la probabilidad de que los datos observados sean producidos por pura aleatoriedad. El valor- p no es ninguno de éstos; es una declaración sobre los datos en relación con una explicación hipotética específica y no sobre la explicación en sí misma.

3. *Las conclusiones científicas y las decisiones empresariales o políticas no deben ser basadas únicamente en si un valor- p dado supera un determinado umbral.*

Las prácticas que reducen el análisis de datos o la inferencia científica a reglas mecánicas (tales como “ $p < 0.05$ ”) pueden llevar a conclusiones y decisiones incorrectas. Una conclusión no se convierte inmediatamente en “verdadera” en un lado del umbral y “falsa” en el otro. Los investigadores deberían contextualizar sus inferencias científicas en función de muchos factores, incluyendo: el diseño del estudio, la calidad de las mediciones, la evidencia externa para el fenómeno bajo estudio y la validez de los supuestos que subyacen al análisis de los datos.

Las consideraciones pragmáticas generalmente requieren decisiones binarias de “sí-no,” pero esto no significa que los valores- p por sí solos puedan asegurar que la decisión sea correcta o incorrecta. El amplio uso de la “significancia estadística” (generalmente, interpretada como “ $p \leq 0.05$ ”) como una licencia para realizar una afirmación de un descubrimiento científico (o verdad implícita) lleva a una distorsión considerable del proceso científico.

4. Inferencias adecuadas requieren informes completos y transparentes.

Los valores- p y los análisis relacionados no deberían ser reportados en forma selectiva. Conducir múltiples análisis de los datos y reportar únicamente aquellos con ciertos valores- p (típicamente aquellos que pasan un umbral de significancia dado) dejan a los valores- p reportados esencialmente sin interpretación. Los descubrimientos prometedores que surgen de suprimir pruebas (“cherry-picking”), también conocidos con términos como: limpiar datos, persecución o búsqueda de significancia, inferencia selectiva y “ p -hacking,” llevan a un exceso espurio de resultados estadísticamente significativos en la literatura y deben ser enfáticamente evitados. No se necesita llevar a cabo múltiples pruebas estadísticas para que aparezca este problema: cada vez que un investigador elige que presentar basándose en resultados estadísticos y no comunica la naturaleza y las bases de dicha elección, se compromete severamente la posibilidad de que el lector haga una interpretación acertada de los resultados. Los investigadores deben informar el número de hipótesis exploradas durante el estudio, todas las decisiones acerca de la recolección de datos, todos los análisis estadísticos conducidos y todos los valores- p computados. Las conclusiones científicas basadas en valores- p y estadísticos relacionados no pueden ser formuladas sin saber al menos cuántos y cuáles análisis fueron conducidos, y cómo esos análisis (incluidos los valores- p) fueron seleccionados para reportar.

5. Un valor- p , o la significancia estadística, no da una medida del tamaño del efecto o de la importancia de un resultado.

La significancia estadística no es equivalente a la significancia científica, humana o económica. Valores- p más pequeños no necesariamente implican la presencia de efectos mayores o más importantes, y valores- p más grandes no implican la falta de importancia o incluso la falta de efecto. Cualquier efecto, sin importar cuán pequeño sea puede producir un valor- p pequeño si el tamaño de la muestra o la precisión de la medición es

lo suficientemente grande, y efectos grandes pueden producir valores- p poco importantes si el tamaño de la muestra es pequeño o si las mediciones son imprecisas. En forma similar, efectos idénticos van a tener valores- p diferentes si la precisión de las estimaciones difiere.

6. Por sí sólo, un valor- p no proporciona una buena medida de la evidencia con respecto a un modelo o a una hipótesis.

Los investigadores deben reconocer que un valor- p fuera de contexto o sin otra evidencia brinda información limitada. Por ejemplo, un valor- p cercano a 0.05 tomado por sí solo ofrece evidencia débil en contra de la hipótesis nula. De igual forma, un valor- p relativamente grande no implica evidencia a favor de la hipótesis nula; muchas otras hipótesis pueden ser igual o más consistentes con los datos observados. Por estas razones, el análisis de los datos no debe terminar con el cálculo de un valor- p cuando otros enfoques son apropiados y factibles.

Otros Enfoques

En vista de los prevalecientes malos usos y de las concepciones equivocadas concernientes a los valores- p , algunos estadísticos prefieren complementar o, incluso, reemplazar los valores- p con otros enfoques. Éstos incluyen métodos que enfatizan la estimación sobre el contraste de hipótesis, tales como los intervalos de confianza, credibilidad o predicción; métodos Bayesianos; mediciones alternativas de la evidencia, tales como cocientes de verosimilitud, o los factores de Bayes; y otros enfoques como teoría de la información y tasas de descubrimientos falsos. Todas estas medidas y enfoques también dependen de supuestos, pero podrían abordar más directamente el tamaño de un efecto (y su incertidumbre asociada) o si la hipótesis es correcta.

Conclusión

La buena práctica estadística, como un componente esencial de la buena práctica científica, hace énfasis en los principios del buen diseño y conducción de los estudios, una variedad de resúmenes numéricos y gráficos de los datos, la comprensión del fenómeno bajo estudio, la interpretación de los resultados en su contexto, un informe completo y la adecuada comprensión lógica y cuantitativa de lo que lo que significan los resúmenes de los datos (estadísticos). Ningún índice debería sustituir el razonamiento científico.