Ideas equivocadas frecuentes en estadística explicadas sencillamente

Common statistical misconceptions-plainly explained

María S. Navarrete [a*]

[a] Escuela de Medicina, Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile

*Autor corresponsal marisol.navarrete@usach.cl

Citación Navarrete MS. Common statistical misconceptions-plainly explained. *Medwave* 2019;19(6):7660

Doi 10.5867/medwave.2019.06.7660

Fecha de envío: 13/3/2019

Fecha de aceptación: 10/5/2019

Fecha de publicación: 2/7/2019

Origen: solicitado

Tipo de revisión: con revisión por tres pares revisores externos, a doble ciego

Palabras clave statistics, biomedical research/methods, censuses

Este artículo se propone tratar el problema bien conocido y muy extendido de la mala calidad de los análisis estadísticos presentes en la literatura científica[1-6]. El mejor uso que los lectores pueden hacer de este artículo es tomar las referencias y leerlas en extenso. Sin embargo, estos excelentes artículos no los leen aquellos que obtendrían el mayor beneficio. Quizás no lo hacen porque son muy largos o muy académicos o muy matemáticos. Cualquiera sea la razón, el resultado final es que estos artículos no llegan al público al que apuntaban. Personalmente, estoy convencida de que la formación en inferencia estadística no debería ser impartida por matemáticos, quienes estarán más sensibilizados con las dificultades que enfrentan comúnmente los estudiantes que no tienen facilidad con los números. En este artículo explicaré algunos errores o malentendidos frecuentes en lenguaje sencillo, exponiendo por qué son errores, por qué se cometen y qué se podría hacer al respecto.

Censo versus muestreo

Comenzaré tratando la más simple, y tal vez la más engañosa, idea equivocada con la que me he visto enfrentada actuando como revisora por pares para revistas biomédicas. Se trata del mal uso de las herramientas de inferencia estadística (test de hipótesis o intervalos de confianza) cuando se analizan datos provenientes de un procedimiento censal. De acuerdo con el diccionario Merriam-Webster censo es "la enumeración usualmente completa de la población". El diccionario además da un ejemplo: "de acuerdo con el último censo de los Estados Unidos, 16% de la población es de origen Hispánico o Latino". Por lo tanto, nadie debiera sentir la necesidad de calcular un intervalo de confianza en torno de 16%, puesto que sabemos exactamente el verdadero valor del parámetro de interés en la población.

El fin último de la inferencia estadística es descubrir alguna propiedad o un patrón general existente en un grupo grande a través del estudio de un grupo más pequeño, en el entendido que los resultados podrán extrapolarse al grupo mayor. Aplicamos la inferencia estadística con justa razón porque tomamos una muestra al azar y derivamos de ella una estimación del parámetro verdadero, la cual puede estar lejos o cerca del parámetro en cuestión. Acto seguido, aplicamos técnicas estadísticas que nos permiten tomar en consideración esta imprecisión, pudiendo así generalizar el resultado a la población de origen, con cierta confianza.

A pesar de este simple razonamiento, los autores se resisten a restringir sus análisis a resultados descriptivos, tal como manda el diseño del estudio realizado. Me atrevería a decir que esta resistencia proviene, en cierta medida, de la dificultad que los investigadores enfrentan cuando tienen que interpretar sus resultados. Infelizmente, muchos investigadores se apoyan en el test de hipótesis para sacar las conclusiones finales. Sin embargo, imaginemos que hacemos un estudio en la población total con el fin de evaluar una intervención y el efecto observado es 30%. Eso es todo. En seguida, todo lo que hay que hacer es discutir sobre posibles sesgos que puedan haberse introducido en el diseño y conducción del estudio. Para terminar, hay que elaborar respecto a si 30% es un buen resultado o cómo se compara con la intervención de control o cualquier otra consideración referida al impacto de los resultados en el conocimiento actualizado del tópico de interés. Infelizmente, el hecho de que no haya valores de p que ayuden en la discusión y a sacar conclusiones para el manuscrito, deja a muchos autores incómodos; teniendo que enfrentar la verdadera cuestión ¿qué significan los resultados?

Curiosamente, no he encontrado muchas referencias respecto a este asunto. He consultado profesores de conocidas universidades americanas y francesas que me han confirmado lo que expongo. Esto es, admiten que hay estadísticos que se sienten perplejos cuando se enfrentan con la situación donde "no hay muestreo, no hay incertidumbre, luego no hay inferencia, no hay intervalos de confianza, no hay valores de p".

Valores de p por separado no es la manera de comparar grupos

La segunda idea equivocada que elegí es el error común que se comete cuando luego de observar un cambio estadísticamente significativo en la variable X, cuando A está presente y no observar un cambio estadísticamente significativo en la variable X ante la presencia de B; se llega a la conclusión equivocada de que los efectos de A y B son diferentes. Este error ha sobrevivido por décadas. Douglas Altman escribió sobre ello en su libro en 1991[7]. En una lista de "errores en el análisis", él recomendaba lo siguiente "ejecutar análisis intragrupos y luego contrastar los grupos comparando los valores de p o los intervalos de confianza."

En 2009, Watson y colaboradores publicaron un artículo sobre un ensayo clínico que realizaron para evaluar la eficacia de un producto cosmético "antiarrugas"[8]. En una carta al editor, Martin Bland señaló varios defectos que identificó en el artículo, uno de ellos es el que explicamos a continuación. Él señaló que "en relación con las arrugas a los seis meses, los autores entregan resultados de tests comparando el puntaje respecto a la línea de base para cada grupo separadamente, encontrando que un grupo presentaba una diferencia significativa y el otro grupo no. Este es un error estadístico clásico. Lo que se debe hacer es comparar los dos grupos directamente"[9].

Asociado a esto, está la práctica común del diseño llamado "antes-después". Este diseño consiste en medir la variable de interés en un solo grupo de manera prospectiva: primero, la línea de base y más tarde en un punto definido en el tiempo. Con estos datos se puede calcular, para cada individuo, la diferencia observada entre los dos puntos en el tiempo (valor en línea de base menos valor en el punto de seguimiento), obteniéndose así el promedio de todas estas diferencias que representa el cambio promedio observado en el tiempo. Se puede efectuar un test de hipótesis que compare el promedio de las diferencias contra cero, con el fin de estimar qué tan probable sería observar una diferencia así cuando la hipótesis nula es verdadera. Sin embargo, el diseño "antes-después" no incluye un grupo control y muchos libros de texto no advierten a los lectores sobre los múltiples sesgos que este diseño conlleva. A veces, estudiantes graduados comienzan utilizando este diseño simple, barato y fácil de realizar. Tal vez todo el problema se origina en el hecho de enseñar el método estadístico separadamente, en circunstancia que debiera ir de la mano de los principios metodológicos de los diseños de investigación.

Presentar comparaciones estadísticas de las características iniciales en los informes de ensayos clínicos

Efectuar una prueba de hipótesis en cada variable presentada en la clásica Tabla 1 de un manuscrito, resumiendo las características iniciales de los pacientes de un ensayo clínico aleatorizado es innecesario. Primero, porque este análisis no aborda la pregunta de investigación y segundo, porque si se utilizó un método aleatorio para asignar los pacientes a cada grupo de tratamiento, entonces la hipótesis nula es verdadera para todas las características iniciales por definición[5].

Los estándares consolidados para el reporte de ensayos clínicos (CONSORT) establecen claramente esto en su ítem 15 Características demográficas y clínicas de cada grupo. La guía CONSORT es muy clara sobre cómo reportar las características iniciales: "desgraciadamente, tests de significancia de las diferencias entre las características iniciales son aún comunes (…) Pruebas de diferencias en las características iniciales no son necesariamente erróneas, son simplemente ilógicas. Tales tests de hipótesis son superfluos y pueden inducir a los investigadores y sus lectores al error. Más bien, las comparaciones entre los grupos al inicio deben basarse en consideraciones respecto a la fuerza pronóstica de las variables medidas y al tamaño del desequilibrio provocado por el azar que haya podido ocurrir"[10].

¿Qué se puede hacer?

¿Por qué son tan comunes los errores estadísticos en la literatura biomédica publicada? Una razón puede ser el hecho de que hay escasez de estadísticos para las revisiones por pares. En consecuencia, publicaciones de mala calidad aquejadas de errores estadísticos se publican continuamente[6] y mientras más ideas equivocadas aparecen, los lectores más los incorporan pues los asumen como correctos científica y estadísticamente. Es improbable que esta situación cambie en el corto plazo. En lo que respecta a la revisión por pares, las revistas debieran acudir a personas expertas conocedoras de la especialidad clínica, así como conocedoras de la estadística inferencial básica.

Notas

Conflictos de intereses
No hay.

Nota de la editora
Este comentario fue enviado en inglés y en español por la autora.

Referencias

1. von Elm E, Egger M. The scandal of poor epidemiological research. BMJ. 2004 Oct 16;329(7471):868-9. | PubMed |
2. Greenwood DC, Freeman JV. How to spot a statistical problem: advice for a non-statistical reviewer. BMC Med. 2015 Nov 2;13:270. | CrossRef | PubMed |
3. Nuzzo R. Scientific method: statistical errors. Nature. 2014 Feb 13;506(7487):150-2. | CrossRef | PubMed |
4. Zinsmeister AR, Connor JT. Ten common statistical errors and how to avoid them. Am J Gastroenterol. 2008 Feb;103(2):262-6. | CrossRef | PubMed |
5. Lang T. Twenty statistical errors even you can find in biomedical research articles. Croat Med J. 2004 Aug;45(4):361-70. | PubMed |
6. Altman DG. The scandal of poor medical research. BMJ. 1994 Jan 29;308(6924):283-4. | PubMed |
7. Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall; 1991.
8. Watson RE, Ogden S, Cotterell LF, Bowden JJ, Bastrilles JY, Long SP, et al. Effects of a cosmetic 'anti-ageing' product improves photoaged skin [corrected]. Br J Dermatol. 2009 Aug;161(2):419-26. | CrossRef | PubMed |
9. Bland JM. Evidence for an 'anti-ageing' product may not be so clear as it appears. Br J Dermatol. 2009 Nov;161(5):1207-8; author reply 1208-9. | CrossRef | PubMed |
10. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1):28-55. | CrossRef | PubMed |

Correspondencia a:
Avenida Libertador Bernardo O'Higgins n°3363
Estación Central
Santiago
Chile