Enfermería Intensiva 32 (2021) 42–44

SEEIUC
Sociedad Española de Enfermería Intensiva y Unidades Coronarias

Enfermería *Intensiva*

www.elsevier.es/ei

COMENTARIOS A ARTÍCULOS DE INVESTIGACIÓN

Diez errores estadísticos frecuentes que tener en cuenta al escribir o revisar un artículo

Ten common statistical mistakes to watch out for when writing or reviewing a manuscript

Makin TR, Orban de Xivry JJ. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife. 2019;8:e48175. doi:10.7554/eLife.48175

Resumen

Inspirados por esfuerzos más amplios para hacer más sólidas las conclusiones de la investigación científica, hemos recopilado una lista de algunos de los errores estadísticos más comunes que aparecen en la literatura científica. Los errores tienen su origen en diseños experimentales ineficaces, análisis inapropiados y/o razonamientos erróneos. Proporcionamos asesoramiento sobre la forma en que los autores, revisores y lectores pueden identificar y resolver estos errores y esperamos evitarlos en el futuro. Todos los errores pueden ser identificados en los distintos apartados de una publicación principalmente en material y métodos, resultados o conclusiones.

Ausencia de un grupo o una condición de control adecuada

El problema. En ocasiones en investigación clínica se pretende evaluar el efecto de una intervención sin disponer de un grupo control. Si se realiza el seguimiento de un grupo de pacientes después de una intervención y se evalúa la variable desenlace antes y después de la intervención, el cambio en esta variable podría suponerse que es debido principalmente al efecto de la intervención sin tener en cuenta el efecto del tiempo y esto no es asumible habitualmente. Por ello, es importante la inclusión adecuada de un grupo control.

Cómo detectarlo. Cuando observamos en un artículo que los datos son de un único grupo, o con varios grupos, pero sin un control adecuado.

Soluciones para las y los investigadores. El diseño experimental es imprescindible para evitar estos sesgos, seleccionando ambos grupos en el mismo momento, asignando a los participantes, desarrollando manipulaciones idénticas y favoreciendo que los estudios sean ciegos tanto para los participantes como para los investigadores. Si el diseño experimental no permite la separación del efecto del tiempo del efecto de la intervención, entonces las conclusiones con respecto al impacto de la intervención deben ser presentadas como preliminares.

Interpretar comparaciones entre 2 efectos sin que estén directamente comparados

El problema. A menudo las conclusiones sobre el impacto de una intervención se basan nada más que en el efecto estadísticamente significativo del tratamiento en el grupo experimental comparado con un efecto en el grupo control no significativo, basándose en 2 pruebas estadísticas realizadas por separado. Realmente, la correlación entre 2 variables en un grupo puede ser estadísticamente significativa y no serlo en otro teniendo un coeficiente de correlación similar. Esto puede ocurrir incluso si la relación entre las 2 variables es virtualmente idéntica entre los 2 grupos, por lo que no se debe inferir que una correlación es mejor que la otra sin utilizar una única prueba estadística para comparar los 2 efectos.

Cómo detectarlo. Se observa cuando la conclusión que se extrae con respecto a la diferencia entre 2 efectos se da sin haber sido comparados estadísticamente entre ellos.

Solución para las y los investigadores. La correlación entre 2 grupos puede ser comparada con simulación de Monte Carlo, con una prueba ANOVA, e incluso con pruebas estadísticas no paramétricas. Los procedimientos *network* metaanálisis permiten comparar múltiples tratamientos simultáneamente en un solo análisis, combinando pruebas directas e indirectas dentro de una revisión sistemática de ensayos clínicos aleatorizados.

https://doi.org/10.1016/j.enfi.2020.09.002
1130-2399/

Exagerar las unidades del análisis

El problema. La unidad experimental se define como la observación más pequeña que puede ser asignada de forma aleatoria e independiente. Sin una clara identificación de la unidad apropiada para evaluar un efecto puede resultar en un número elevado y adulterado de unidades experimentales que producen una inferencia estadística errónea.

Cómo detectarlo. En el apartado de metodología debe describirse la unidad de análisis adecuada, es decir, si el estudio tiene como objetivo entender los efectos en un grupo, la unidad de análisis debe reflejar la varianza entre los sujetos. A menudo en un mismo paciente se realizan varias medidas, por ejemplo, cuando se evalúan órganos pares (ojos, riñones o pulmones), cuando se evalúa al mismo sujeto en varias mediciones a lo largo del tiempo o cuando se evalúa el efecto de una intervención a nivel clúster, por ejemplo, cuando se aleatorizan controles de enfermería, pero se recogen datos en pacientes.

Solución para las y los investigadores. La mejor solución disponible es la utilización de modelos lineales de efectos mixtos, de forma que se pueden incluir todos los datos en el modelo sin quebrantar el supuesto de independencia. Sin embargo, requiere conocimientos estadísticos avanzados y los resultados deben ser interpretados con cautela.

Correlaciones engañosas

El problema. La correlación es una herramienta muy importante en términos de evaluar la magnitud de una asociación entre 2 variables, pero las correlaciones paramétricas (por ejemplo, r de Pearson) tienen una serie de supuestos que, de incumplirlos, pueden dar lugar a correlaciones engañosas. Las correlaciones engañosas más comunes se dan cuando uno o muchos *outliers* (valores fuera rango) están presentes en una de las 2 variables, ya que un único valor alejado puede inflar el coeficiente de correlación. Los valores *outliers* pueden aportarnos observaciones extremas que obedecen al fenómeno que se está estudiando, por lo que eliminar datos extremos debe considerarse con cautela.

Cómo detectarlo. Se puede detectar en el apartado de resultados, prestando particularmente atención a las correlaciones que no estén acompañadas de un gráfico de dispersión y considerar si se ha aportado justificación suficiente cuando algún dato extremo haya sido eliminado.

Solución para las y los investigadores. Los métodos de correlación robustos (por ejemplo, *bootstrapping*) son menos sensibles a los valores extremos, ya que toman en consideración la estructura de los datos.

Utilización de tamaños muestrales pequeños

El problema. Un tamaño muestral pequeño solo puede detectar efectos importantes y son también más susceptibles de no encontrar el efecto real que está presente en los datos (error de tipo II). Además, la distribución de una muestra pequeña tiende a desviarse de una distribución normal, y el limitado tamaño hace a menudo imposible probar con rigor el supuesto de normalidad.

Cómo detectarlo. Los revisores deben examinar críticamente el tamaño muestral utilizado en el artículo y juzgar si se tiene suficiente potencia estadística como para concluir los distintos resultados expuestos.

Solución para las y los investigadores. La mejor forma de solucionar estos problemas es desarrollar *a priori* un análisis de poder estadístico. La estadística bayesiana ofrece posibilidades para determinar el poder estadístico suficiente para identificar efectos *post hoc*. En casos en los que el tamaño muestral no pueda ampliarse, es necesario aportar replicaciones o incluir controles suficientes (por ejemplo, estableciendo intervalos de confianza).

Análisis circular

El problema. Se basa en seleccionar los datos que caracterizan las variables dependientes y utilizar los mismos datos para una primera caracterización de las variables de estudio y, posteriormente, realizar con ellos inferencias estadísticas. Formas habituales de análisis circular se muestran en la búsqueda del efecto de un tratamiento en subgrupos creados no con criterio explicito antes de realizar el estudio, sino basado en los resultados del propio estudio

Cómo detectarlo. En principio ocurre siempre que las pruebas estadísticas están sesgadas por la selección de un criterio a favor de la hipótesis que se está evaluando. Los revisores deben estar alerta ante elevados efectos imposibles que no son teóricamente plausibles, y/o están basados en medidas relativamente poco fiables. En estos casos, los autores deben justificar de la independencia entre el criterio de selección y el efecto de interés.

Solución para las y los investigadores. Definir el criterio de análisis con anterioridad e independientemente de los datos, siendo la solución más directa utilizar diferentes bases de datos para especificar los parámetros del análisis y probar las predicciones. Esta división puede hacerse a nivel de participante (utilizando un grupo diferente) o a nivel de prueba (utilizando distintas pruebas para todos los participantes). Esto puede conseguirse sin perder poder estadístico utilizando aproximaciones mediante *bootstrapping*.

Flexibilidad del análisis: cazadores de valores p significativos

El problema. El *p-hacking* se refiere a la práctica de manipular los datos (reemplazar parámetros de respuesta, añadir covariables, excluir sujetos, etc.) hasta que el resultado pase el umbral del error estadístico. Estimar un valor p en una base de datos no es necesariamente complicado y cualquier investigador puede aportar una explicación plausible para cualquier efecto. Por ello, es importante definir con anterioridad los análisis estadísticos que se van a realizar, el diseño de la experimentación o realizar posteriormente una replicación del estudio.

Cómo detectarlo. El *p-hacking* es difícil de detectar ya que raramente se desglosa toda la información necesaria. Se puede considerar si todas las elecciones de análisis no están bien justificadas, si el mismo planteamiento analítico no fue utilizado en anteriores publicaciones, si los investigadores presentan una nueva variable que no es habitual o si recogen un amplio número de variables presentando únicamente algunas significativas en los resultados.

Solución para las y los investigadores. Los análisis exploratorios que se basan en un estudio preliminar de datos flexibles son correctos si se reporta e interpreta de una forma transparente y, especialmente, si sirve como base de replicación. Estos análisis pueden ser una justificación válida para investigación adicional pero nunca puede ser el fundamento de fuertes conclusiones. Quizás la mejor forma de prevenir el *p-hacking* es mostrar cierta tolerancia a los resultados no significativos: si el experimento está bien diseñado, ejecutado y analizado, los revisores no pueden «castigar» a los investigadores por sus datos.

No corregir comparaciones múltiples

El problema. A menudo se explora un efecto en múltiples variables, normalmente con una hipótesis *a priori* indeterminada, lo que se conoce como análisis exploratorio. En estadística frecuentista, realizar múltiples comparaciones durante el análisis exploratorio puede incrementar la probabilidad de detectar un efecto significativo incluso si este efecto no existe (falso positivo, error de tipo I) debido al uso repetido de pruebas estadísticas.

Cómo detectarlo. Se puede detectar en el apartado de la metodología y de resultados; cuando se llevan a cabo análisis exploratorios con múltiples variables, es inaceptable interpretar los resultados que no han superado las correcciones de comparaciones múltiples sin justificación. Incluso aunque se ofrezca una predicción robusta, si esta predicción no puede ser probada en múltiples comparaciones independientes, se requiere una corrección para múltiples comparaciones.

Soluciones para las y los investigadores. Los investigadores deben revelar todas las variables medidas e implementar adecuadamente el uso de las correcciones de las comparaciones múltiples, justificando el porqué de la utilización de una determinada prueba.

Sobre interpretar resultados no significativos

El problema. Un valor p no significativo puede significar que un resultado es realmente nulo, que es un efecto que no tiene poder estadístico suficiente para su evaluación o que es un efecto ambiguo. Para interpretar un resultado no significativo como una evidencia en contra de la hipótesis, se necesitaría demostrar que esa evidencia es significativa. Esto supone que resultados que se encuentren cercanos al 0,05 no deban asumirse como no satisfactorios cuando realmente proporcionan evidencia preliminar que requiere atención adicional.

Cómo detectarlo. En apartado de resultado o conclusiones, se puede interpretar o describir un valor p no significativo como indicativo de que el efecto no está en absoluto presente. Este error es muy común y debe ser señalado como problemático.

Solución para las y los investigadores. Un primer paso importante es reportar el tamaño del efecto junto con el valor p en orden de proporcionar información sobre la magnitud del efecto, lo que es igualmente importante por cualquier metaanálisis futuro. Por ejemplo, si un efecto no significativo en un estudio con un amplio tamaño muestral es también muy pequeño en magnitud, es improbable que sea teóricamente significativo mientras que uno con un tamaño moderado del efecto puede potencialmente justificar más investigación. Por otro lado, los investigadores podrían tener ya determinado *a priori* si tienen suficiente poder estadístico para identificar el efecto deseado, o para determinar si los intervalos de confianza de este efecto previo contienen el cero. De lo contrario, los investigadores no deben sobreinterpretar resultados no significativos y solo describirlos como no significativos.

Correlación y causalidad

El problema. La correlación es usada frecuentemente para explorar la relación entre 2 variables, habitualmente asumiéndose que una es causa de la otra. Sin embargo, solo porque 2 variables parezcan ocurrir de forma lineal no necesariamente significa que exista una relación causal entre ellas, incluso si esta asociación es plausible. Correlaciones por separado no pueden utilizarse como evidencia de una relación de causa y efecto.

Cómo detectarlo. Los investigadores deben utilizar solo el lenguaje causal si la relación entre 2 o más variables se debe a un análisis adecuado desde el punto de vista metodológico y estadístico, e incluso entonces deben ser cautelosos sobre el papel de una tercera variable o de factores de confusión.

Solución para las y los investigadores. Se debe intentar explorar la relación con una tercera variable para aportar más apoyo en las interpretaciones, por ejemplo, utilizando análisis de mediación o índice de propensión Desde el punto de vista del diseño de la investigación, el único estudio, para la mayoría de los autores, que puede contestar preguntas de causalidad es un ensayo clínico aleatorizado y controlado cuando es posible realizarlo. De otra forma, el lenguaje causal debe ser evitado cuando la evidencia es de correlación.

L. del Campo-Albendea (MsC)[a]
y A. Muriel-García (MsC, PhD)[b,*]
[a] *Graduada en Biología, Universidad Complutense de Madrid. Estudiante Máster de Bioestadística, Facultad de Estudios Estadísticos. Universidad Complutense de Madrid, Madrid, España*
[b] *Doctor por la Universidad Autónoma de Madrid, Unidad de Bioestadística Clínica, Hospital Ramón y Cajal, IRYCIS, CIBERESP, Madrid, España*

[*] Autor para correspondencia.
Correo electrónico: alfonso.muriel@hrc.es
(A. Muriel-García).