¿Tienes poder estadístico? Considerando el error tipo II en la educación médica*

Gail Sullivan y Richard Feinn

Resumen: ¿Se puede abordar con sensatez la cuestión del poder estadístico para detectar las diferencias? En la mayoría de los casos, se puede, incluso en la investigación sobre educación médica donde habitualmente las muestras suelen ser pequeñas. Este breve ensayo está pensado para ayudar a los investigadores, tutores y lectores principiantes a adentrarse en las turbias aguas de los cálculos de potencia estadística y los tamaños de muestra, y salir ilesos.

Do You Have Power? Considering Type II Error in Medical Education Abstract: Can we sensibly address the question of power to detect differences? In most instances, we can, even in medical education research with typically small sample sizes. This brief introduction is designed to help beginning researchers and readers wade into the murky waters of power calculations and sample sizes?and emerge unscathed.

Los tamaños de las muestras en la educación médica, y en particular en los proyectos de educación médica normalmente son pequeños, lo que supone un problema. Si realmente existe una diferencia entre los grupos, es posible que los educadores no la descubran con un tamaño de muestra pequeño. Esto puede llevar a los investigadores a opinar en la Discusión lo siguiente: "Aunque no encontramos diferencias estadísticamente significativas entre nuestra propuesta y la habitual, esto probablemente se deba a que el tamaño de la muestra fue demasiado pequeño". En otras palabras, los autores creen que hay una diferencia, pero el estudio no tenía poder estadístico (PE). Este tipo de ?no hallazgo? o ?no conclusión? se ve frecuentemente en los estudios de educación médica, pero no es muy útil. De hecho, algunos expertos creen que no considerar el poder de un estudio de antemano es una irresponsabilidad. 1 Sin embargo, hay ocasiones, como en la narrativa o el trabajo exploratorio, en que los problemas con el PE del estudio pueden ser inevitables.

Revisión: errores de tipo I frente a errores de tipo II

Los errores de tipo I pueden producir resultados falsos positivos; si rechazamos la hipótesis nula (que no hay diferencia entre los grupos que estamos comparando) cuando en realidad es cierta, es decir, no hay diferencia. Por lo general, elegimos un número

pequeño, como 0,05 o menos, para el nivel de error de tipo I, o alfa, para comparaciones únicas o unas pocas. 2 Cuanto más bajo sea el nivel alfa, menos probable es que lleguemos a conclusiones falsas positivas: menos probable que sean las palabras operativas. 3

Por el contrario, los errores de tipo II pueden producir resultados falsos negativos: no rechazar la hipótesis nula (que no hay diferencia entre los grupos) cuando no es cierta: hay una diferencia. Cuando planificamos un estudio, generalmente elegimos 0.20 como el nivel de error tipo II (beta). Sin embargo, al elegir alfa y beta, también se debe considerar la pregunta de investigación y los efectos del mundo real de pasar por alto una diferencia real frente a afirmar diferencias que no existen.

El PE de un estudio para encontrar diferencias es 1 - beta, que es 0,80 u 80%, si se elige beta en 0,20. El poder es la probabilidad de rechazar correctamente la hipótesis nula (que no hay diferencia) cuando no es cierta. El PE responde a la pregunta: si realmente ocurre un efecto, de una magnitud específica, ¿cuál es la probabilidad de que una prueba, de un cierto tamaño de muestra, encuentre un resultado estadísticamente significativo dado el nivel alfa elegido? Cuanto mayor sea el poder de una prueba, más confianza tendremos en que seremos capaces de detectar una diferencia entre los grupos. 3 Una potencia de estudio establecida en 80 % acepta una posibilidad de 1 en 5 (20 %) de pasar por alto una diferencia que realmente existe. Los investigadores pueden establecer la potencia al 90 % para reducir la posibilidad de perder una diferencia real de 1 en 10.

Si existe una diferencia y cuán grande es esa diferencia no está bajo nuestro control; estas son características de la intervención, el entorno y los sujetos en estudio. Sin embargo, podemos controlar el tamaño de la muestra del ensayo: cuántos alumnos se incluyen o cuántas evaluaciones se examinan.

En investigaciones sobre educación médica, generalmente buscamos diferencias moderadas o grandes. Por ejemplo, puede haber una diferencia real en la satisfacción de los residentes por un programa nuevo frente al programa existente, en una escala tipo Likert de 1 a 5. Aunque si las medias son 3,16 frente a 3,27, esta diferencia no es significativa en un sentido educativo sin importar si es estadísticamente significativo. 4 (Hay que tener en cuenta que se necesita una muestra grande para demostrar una diferencia real, pero pequeña, como en este ejemplo). O un estudio nacional puede encontrar que una nueva iniciativa de bienestar muestra una disminución en las medidas de burnout del 28 % al 27 %. Esta diferencia parece real, ya que el valor de la P es más bajo que nuestro límite de nivel alfa elegido. ¿Nos importa? No. A diferencia de lo que pasa en medicina clínica, en educación médica generalmente queremos diferencias más grandes para justificar decisiones.

Cuándo pensar en la potencia y los tamaños de muestra

Es mejor realizar un análisis de potencia estadística antes de hacer un estudio para maximizar la capacidad de detectar las diferencias que existen. Como mínimo, las consideraciones de PE deben preceder a cualquier consideración sobre los datos o análisis de datos.

El poder depende del tamaño real de la diferencia (es decir, el tamaño del efecto), la variabilidad o varianza en las variables que estamos midiendo (por ejemplo, la desviación estándar), el nivel de significación que elegimos (alfa) y el tamaño de la muestra. Solo los dos últimos están bajo nuestro control: el nivel de significación (alfa) y el tamaño de la muestra. A medida que aumenta el tamaño de la muestra, disminuye la beta y, por lo tanto, aumenta el poder para encontrar una diferencia real. La mayoría de la gente acepta que una potencia del 80% es razonable, lo que significa seleccionar un error beta o tipo II de 0,20. Idealmente, la elección del nivel de potencia, o la otra cara, el error de tipo II, depende de la gravedad de las consecuencias de cometer un error de tipo II (hallazgo falso negativo), que se relaciona con las decisiones posteriores que se basarán en los hallazgos. Por ejemplo, ¿Afectarán los hallazgos a una evaluación que es de alto riesgo? ¿Conducirán a eliminar una rotación de residentes? Las consecuencias también deben incluir cuánto tiempo y esfuerzo se invierte en realizar el estudio para evitar desperdiciar estos valiosos recursos. Por lo general no queremos gastar mucho tiempo y recursos realizando un estudio para darnos cuenta, más tarde, de que no podemos llegar a una conclusión definitiva porque el estudio no tuvo suficiente poder estadístico.

Además, existe una compensación en el sentido de que, a medida que aumenta alfa, disminuye beta, lo que debe tenerse en cuenta en los planes de estudio. Considere qué es más crítico para informar las decisiones posteriores: evitar resultados falsos positivos (errores de tipo I) o evitar resultados falsos negativos (errores de tipo II).

Cálculo de tamaños de muestra

Para calcular el tamaño de la muestra, necesitamos los niveles de error alfa y beta elegidos, el tamaño del efecto mínimo esperado (magnitud de la diferencia), así como la variabilidad esperada en la variable resultado. 5 Los investigadores a menudo se preguntan cómo determinar el tamaño del efecto, cuando las comparaciones realizadas en el estudio no se han hecho antes o no de la misma manera. De hecho, los investigadores de educación médica rara vez pueden buscar en la literatura y encontrar números probables para las diferencias esperadas. Una estrategia es preguntar a los expertos, educadores médicos bien informados: ¿Cuál es la diferencia mínima, que les convencería de que un enfoque es mejor que otro? Esta estrategia también se utiliza en la investigación clínica. Por ejemplo, ¿cuán grande es la diferencia clínicamente significativa, en una escala de 0 a 70, para una escala cognitiva? En un estudio, los médicos eligieron una diferencia de 4 unidades como significativa. 6

A continuación, será necesario determinar o estimar la variabilidad esperada. Si se ha realizado un trabajo piloto, esto puede generar una estimación de la variabilidad. Una revisión de la literatura puede revelar la desviación estándar de una escala o, si una búsqueda de la literatura no arroja información, los expertos podrían opinar. Finalmente, debe decidir el nivel alfa (generalmente 0.05) y el poder (generalmente 0.80) y meter esta información en un programa para calcular el tamaño de muestra 7 (ver Cuadro).

Listado Cálculo del tamaño de la muestra - Determinar o estimar el promedio del resultado inicial (p. ej., grupo de control o de comparación).- A partir de la literatura o de expertos, estime la diferencia mínima que sea significativa desde el punto de vista educativo para el contexto.- Estimar la variabilidad en los resultados esperados (a partir de la historia pasada o expertos).- Elija el error de tipo II (beta) (como 0,20) o la potencia (como 0,80), según la importancia de las decisiones posteriores.- Elija el error de tipo I (alfa), a menudo 0,05. 2- Considere las pérdidas potenciales de sujetos (aprendices, profesores) u otras pérdidas durante el estudio.- Use una calculadora de tamaño de muestra 7 o consulte a un bioestadístico amigo.

Con un tamaño de muestra extremadamente grande, el poder es grande para encontrar una diferencia muy pequeña, que puede no ser significativa desde el punto de vista educativo. Por lo tanto, siempre hay que buscar un equilibrio entre la realidad (cuántas materias o pruebas puede permitirse incluir) y cuál es la diferencia significativa en un contexto educativo determinado.

Considere un ejemplo: el Jefe de Estudios de un Hospital Docente desea determinar si una conferencia interactiva de un día completo sobre profesionalismo resultará en que menos residentes infringen aspectos importantes del profesionalismo en su hospital, donde el número de incidentes de este tipo es creciente. El programa actual utiliza varios videos en línea requeridos basados ??en casos. Cincuenta incidentes fueron reportados el año pasado (5% de 1000 aprendices totales), con un promedio de 4% en general durante los últimos 5 años (25, 30, 45, 50 y 50 anualmente, respectivamente). La Comisión de Docencia del Hospital decide que sería significativa una diferencia de 15 incidentes menos (es decir, 3,5 %) en comparación con el informe del año pasado de 5 % de incidentes, dado el costo y el esfuerzo realizado con la nueva estrategia. En este caso, se realizará una comparación de proporciones. Las proporciones, junto con los niveles alfa y beta elegidos, se ingresan en una calculadora de tamaño de muestra para una prueba de proporciones. Resulta que, incluso con esta gran diferencia "mínima", el tamaño de muestra necesario, usando un alfa de 0,05 y una beta de 0,20 (con una potencia de 0,80), es 1505 7 ; el estudio requeriría 2 años. Esto se debe a que la incidencia de la medida de resultado (aquí los informes de profesionalismo) no es común en esta población. La reducción relativa del 30 % en los informes (del 5 % al 3,5 %) corresponde a una diferencia absoluta del 1,5 %.

Otro ejemplo se refiere a un jefe de estudios que está interesado en determinar si una nueva rotación de subespecialidad requerida aumentará la sub-puntuación promedio del examen de capacitación, para los 100 residentes en el programa, en al menos 5 puntos (a 55); la sub-puntuación promedio actual es 50. ¿Será suficiente 1 año, con 100 materias en la nueva rotación, para comparar con el promedio anterior? Usando una desviación estándar estimada de 10, obtenida de los datos del año anterior, el jefe de estudios necesitará de 60 a 70 residentes para determinar si la rotación mejoró la puntuación promedio al menos tanto; por lo tanto, 1 año es factible. 7

Hacer más de una inferencia a partir de sus datos requiere diferentes cálculos de potencia, que no se discutirán aquí.

¿Qué pasa con los intervalos de confianza?

El intervalo de confianza (IC), con un nivel que suele establecerse en el 95 %, estima que la verdadera diferencia media se encuentra dentro de este rango de intervalo. Al probar las diferencias entre los grupos, cuando el IC del 95 % excluye el 0, uno tiene más confianza en que el hallazgo no se debe al azar (un error de tipo I). Los IC dependen de los cálculos de PE, ya que el poder estadístico afecta el ancho (o la precisión) del IC: a medida que aumenta el PE (tamaño de la muestra), el ancho del IC se reduce alrededor de la diferencia del tamaño del efecto.

Informes de cálculos de potencia

Aunque el informe preciso de los cálculos del tamaño de la muestra ha mejorado en la literatura clínica (hasta un 34 % en una revisión de 2009 de revistas médicas generales de alto impacto 8), en nuestra experiencia sigue siendo bajo en los estudios de educación médica. Por esta razón, algunas revistas, dudan en considerar estudios con números pequeños de participantes o resultados, como menos de 40. El método preciso y los números utilizados para calcular el tamaño de la muestra se deben describir en la sección Métodos. Si esto requiere demasiadas palabras, los autores pueden agregar esta información en los datos complementarios. Reconocemos la dificultad de hacer suposiciones sobre el tamaño del efecto y la variabilidad de la muestra (desviación estándar), pero el razonamiento en torno a sus suposiciones debe resumirse para los lectores. La transparencia mejorará la credibilidad y ayudará en la posible replicación futura de su trabajo.

Si no se realizó un cálculo de potencia estadística antes del estudio, lo que puede ocurrir con nuevas experiencias educativas que deben iniciarse antes de las consideraciones del estudio, los autores deben indicar la falta de un cálculo de potencia en la sección Métodos: una sola oración será suficiente. No es posible calcular el tamaño de la muestra después de recopilar y analizar los datos del estudio debido al riesgo de sesgo. Un cálculo post hoc puede ser útil para su próximo proyecto, pero no servirá como evidencia al discutir la falta de diferencia observada en el estudio actual.

Conclusiones

Damos la bienvenida a los estudios cuantitativos y esperamos que los cálculos de potencia generalmente se realicen antes de ensamblar o analizar los datos, preferiblemente incluso antes, al comienzo de los proyectos. A menudo, no existen estudios que puedan proporcionar una estimación de la diferencia de tamaño que puede esperar para calcular el tamaño de una muestra. ¡Pero esto no debe ser un problema! Reúna a algunos expertos locales o nacionales, y pregúnteles cómo de grande debería ser la diferencia necesaria para determinar que una diferencia, en este contexto, es significativa desde el punto de vista educativo. ¿Cuál es la diferencia mínima, con las medidas de resultado que planea usar, que sería convincente con respecto al valor de la intervención en estudio? Utilice este número para calcular el tamaño de muestra que probablemente necesite. Proporcione detalles específicos de su razonamiento y proceso en la sección Métodos. Dado el pequeño tamaño de muchos programas educativos, es posible que deba repetir las intervenciones o evaluaciones, o agregar sitios, para obtener un número suficiente; esto también aumentará la generalización de sus resultados a otros entornos y sujetos. En la sección Discusión, sea transparente sobre el error y nunca, jamás, presente un cálculo de potencia post hoc como justificación de por qué no encontró una diferencia.

Referencias

1.? Shreffler J, Huecker MR. Type I and type II errors and statistical power. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2021. Google Scholar

2.? Sullivan GM, Feinn RS. Facts and fictions about multiple comparisons. J Grad Med Educ. 2021; 13 (4): 457? 460. doi:10.4300/JGME-D-21-00599.1 Google Scholar

3.? Greenland S, Senn SJ, Rothman, KJ, et al Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31 (4): 337? 350. doi:10.1007/s10654-016-0149-3 Google Scholar

4.? Sullivan GM, Feinn RS. Using effect size?or why the P value is not enough. J Grad Med Educ. 2012; 4 (3): 279? 282. doi:10.4300/JGME-D-12-00156.1 Google Scholar

5.? Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. Med J Aust. 2003; 177 (7): 256? 257. doi:10.5694/j.1326-5377.2003.tb05240.x Google Scholar

6.? Raina P, Santaguida P, Ismaila A, et al Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. Ann Intern Med. 2008; 148 (5): 379? 397. doi:10.7326/0003-4819-148-5-200803040-00009 Google Scholar

7.? ClinCalc. Sample size calculator. https://clincalc.com/Stats/SampleSize.aspx. Accessed September 30, 2021.

8.? Charles P, Giraudeau B, Dechartres A, et al Reporting of sample size calculation in randomised controlled trials: review. BMJ. 2009; 338:b1732. doi:10.1136/bmj.b1732

* Este artículo es una Traducción del original titulado: **Do You Have Power? Considering Type II Error in Medical Education** y publicado en: J Grad Med Educ (2021) 13 (6): 753?756.**https://doi.org/10.4300/JGME-D-21-00964.1**