Mitos y realidad en el cálculo del tamaño muestral

Luis Prieto-Valiente, Carmen Carazo-Díaz

Resumen. Cuando decidimos hacer un estudio, una de las primeras cuestiones que se plantea es ¿qué número de individuos debo incluir en la muestra para que sea 'representativa' y el estudio sea 'válido'? Como en otros ámbitos de la vida, hay muchas cuestiones para las que no hay una cantidad 'adecuada' y son válidas diferentes cantidades. Aquí ocurre lo mismo. La pregunta '¿cuántos euros costó esta bicicleta?' tiene como respuesta un número concreto. Pero la pregunta '¿cuántos euros necesito para comprar una bicicleta?' admite muchas cifras distintas como respuesta, dependiendo del tamaño y otras características de la bicicleta. Los libros de estadística contienen fórmulas que relacionan el tamaño de la muestra con ciertos parámetros y la mayoría de los médicos cree que una de ellas les dará el tamaño 'adecuado' para su investigación, y que usándolas queda 'justificado el tamaño de la muestra' ante posibles revisores. En este documento se hace una reflexión sobre el verdadero peso que tienen dichas fórmulas y cuál debe ser el uso adecuado que el investigador haga de ellas. Es necesario mostrar errores y simulaciones que no benefician a nadie y perjudican a muchos restando tiempo y energía.

Palabras clave. Cálculo. Efecto real. Investigación médica. Potencia estadística. Tamaño de muestra. Valor *p*.

Universidad Católica San Antonio de Murcia. Guadalupe de Maciascoque, Murcia (L. Prieto-Valiente, C. Carazo-Díaz). Sociedad Científica de Investigación Biomédica. Palma de Mallorca, España (L. Prieto-Valiente).

Correspondencia:
Dra. Carmen Carazo-Díaz. Universidad Católica San Antonio de Murcia. Avenida de los Jerónimos, 135. E-30107 Guadalupe de Maciascoque, Murcia.

E-mail:
ccarazo@ucam.edu

ORCID:
0000-0003-3782-2772 (C.C.D.).

Aceptado tras revisión externa:
12.06.23.

Conflicto de intereses:
Los autores declaran no tener conflictos de interés.

Cómo citar este artículo:
Prieto-Valiente L, Carazo-Díaz C. Mitos y realidad en el cálculo del tamaño muestral. Rev Neurol 2023; 77: 31-3. doi: 10.33588/rn.7701.2023133.

Introducción

Cuando se plantea hacer un estudio, una de las primeras cuestiones que hay que decidir es el número de individuos que se van a incluir en él. Los libros de estadística contienen fórmulas que relacionan el tamaño de la muestra con ciertos parámetros y la mayoría de los médicos cree que una de ellas les dará el tamaño 'adecuado' para su investigación. Sin embargo, no saben cómo interpretarlas ni aplicarlas y, por ello, en muchos casos, acaban tomando el número de individuos que sus recursos les permiten, aunque creen –equivocadamente– que eso es incorrecto desde el punto científico y que deberían justificar 'con rigor' el tamaño elegido [1,2].

Lamentablemente, comparten ese error muchos evaluadores de artículos científicos, tesis y proyectos. Si no aparecen en el texto cierto tipo de fórmulas o frases, entienden que falta ese aval y piden al autor que justifique 'científicamente' la elección del tamaño usado. En algunas ocasiones, esta exigencia es razonable y el autor la subsana correctamente. Pero, en la mayoría de los casos, puesto que en realidad el médico eligió el tamaño basándose en sus posibilidades de tiempo y recursos, para responder a la exigencia del evaluador, copia y pega algún párrafo de otro proyecto previamente aprobado en el que aparece una de esas fórmulas o alguna frase alusiva a ellas. En el proceso de adaptarlo a su caso particular, puesto que el médico no entiende bien el contenido de lo que copia, suele cometer errores que se van acumulando en los sucesivos copia y pega, lo que llega a hacer el texto totalmente ininteligible. Pese a esto, el evaluador asume que con dicho párrafo queda avalado 'rigurosamente' el tamaño de la muestra usada, pues en muchos casos tampoco entiende bien el tema.

Es un proceso absurdo, carente de sentido, que perjudica a todos sin beneficiar a nadie. Acabar de una vez con esta cadena de equívocos y sustituirla por los modos correctos ahorraría tiempo, esfuerzo e incomodidad a muchos investigadores. Para ello no necesitan aprender más matemáticas, porque el proceder correcto es más simple y directo que el equivocado.

Determinación del tamaño muestral

Consideramos como ejemplo un estudio destinado a indagar si cierto fármaco, 'F', incrementa los niveles en sangre de cierto neurotransmisor, 'N12', en mayor medida que el placebo. Mediremos el nivel en dos muestras, tratamiento y placebo, de tamaño *N*, y compararemos las medias de ambos grupos. En cualquier libro encontraremos que el *N* que podríamos usar para hacer el estudio depende de cuatro parámetros, según la expresión:

Tabla. Tamaño muestral, *N*, correspondiente a dos valores distintos de los parámetros β, α, σ y D utilizando la fórmula (1).

		σ = 12		σ = 15	
Valor *p*	Potencia	D = 12	D = 6	D = 12	D = 6
0,05	80%	*N* = 17	64	26	100
0,05	95%	27	105	42	164
0,01	80%	26	96	39	148
0,01	95%	38	145	77	303

$$N = 2(Z_\alpha + Z_\beta)^2 \sigma^2/D^2 \ (1)$$

donde 1 − β indica potencia del estudio, que es la confianza en que el test arroje un valor *p* menor de cierto límite α acordado por el investigador como suficientemente significativo, σ es la desviación estándar de la variable y D es la diferencia real de las medias poblacionales, que mide el efecto real del fármaco [3]. La mayoría de los médicos no sabe cómo buscar las cantidades Z de la fórmula, pero superan esa dificultad usando un *software* estadístico que calcula *N* aplicando esa fórmula o una versión muy próxima a ella, cuyos detalles no interesan aquí.

Cabría pensar que, dando los valores 'correctos' a los cuatro parámetros β, α, σ y D, se obtiene el valor de *N* 'adecuado', pero la realidad es que no hay valor 'correcto' para ninguno de esos cuatro parámetros y, por tanto, no existe 'el *N* adecuado'. El médico elije β y α dentro de una amplia horquilla de márgenes razonables, mientras que σ y D se estiman basándose en indicios generalmente no muy precisos. Por ello, ninguno de los cuatro parámetros tiene valor exacto en ninguna investigación real [4,5].

Por ejemplo, si queremos tener confianza al 95% en obtener $p < 0{,}01$ y estimamos la desviación estándar en 15, si realmente F aumenta la media del neurotransmisor seis unidades más que el placebo, necesitamos *N* = 303. Pero, si elegimos tener confianza al 80% en tener $p < 0{,}05$ y estimamos la desviación en 12, para un efecto real 12, necesitamos *N* = 17. Son dos tamaños muestrales muy diferentes: ¿cuál de ellos es 'el correcto' o 'el adecuado'? La respuesta es: ambos tamaños de muestra (y otros muchos) son 'adecuados' y ninguno es 'el adecuado'.

En la Tabla vemos que, a igualdad de otros parámetros, *N* es mayor cuanto menor sea el efecto real del fármaco, D. Así, con desviación estándar 12, para tener una probabilidad del 95% de obtener valor $p \leq 0{,}01$, necesitamos *N* = 38, si realmente F aumenta la media del neurotransmisor 12 unidades más que el placebo. Pero si es D = 6, necesitamos *N* = 145. De la misma manera, vemos que, a igualdad de otros parámetros, *N* es mayor cuanto menor es el valor *p* del test acordado, que *N* es mayor cuanto mayor es la potencia que el investigador elige y que *N* es mayor cuanto mayor es la desviación estándar de la variable.

Lo importante es entender que ninguno de los cuatro parámetros tiene un valor exacto e inamovible. La desviación raramente se conoce con exactitud y el investigador debe estimarla mediante bibliografía o con estudios piloto que nunca indicarán un valor inamovible. Aunque lo deseable es tener potencia lo más alta posible para conseguir un valor *p* lo menor posible, cualquier cantidad concreta que se elija puede cambiarse por otra no muy lejana sin pérdida de validez. La fórmula mostrará como resultado casi cualquier *N* usando valores razonables de los cuatro parámetros.

Consideremos, por ejemplo, el *N* = 100 de la tabla que nos proporciona probabilidad del 80% de encontrar $p \leq 0{,}05$, si la desviación es 15 y la D es 6. Modificando solamente una unidad cada parámetro, encontramos (dato no recogido en la Tabla) que, para tener una probabilidad del 81% de encontrar $p \leq 0{,}04$, si la desviación es 16 y la D es 5, necesitamos *N* = 178. Y, modificando cada parámetro una unidad en sentido contrario, encontramos que para tener potencia del 79%, $p \leq 0{,}06$, desviación 14 y D = 7, necesitamos *N* = 59.

Vemos ahora un ejemplo con variable dicotómica en investigación básica. El 80% de las ratas de una cepa genéticamente modificada desarrolla cáncer el tercer mes de vida. Para ver si el alimento A es cancerígeno, se le administrará desde el nacimiento a una muestra de *N* ratas de esa cepa. La fórmula para *N* en este caso es:

$$N = [Z_\alpha \sqrt{p_1 q_1} + Z_\beta \sqrt{p_2 q_2})/(p_1 - p_2)]^2 \ (2)$$

la cual el médico no entiende ni tiene obligación de entender, porque no es su especialidad. Los programas informáticos calculan la *N*, pidiéndonos cuatro valores: la potencia (β), el valor *p* que elegimos como barrera (α), la proporción de ratas con cáncer en la población modificada sin A (p_1), en nuestro ejemplo 0,8, y la proporción de ratas con cáncer en la población modificada que ha recibido A (p_2). Nótese que $q_i = 1 - p_i$ (i = 1,2).

Aplicando la fórmula (2), encontramos que si queremos, por ejemplo, tener una probabilidad del

80% de encontrar el valor $p \leq 0,05$ y con esa dieta desarrollan realmente cáncer el 85%, necesitamos $N = 491$ ratas tratadas con A. Pero si con A desarrollan cáncer el 90%, necesitamos 118 ratas. Y si con A desarrollan cáncer el 95%, necesitamos 49 ratas.

Aprovechamos este último supuesto para mostrar que, aunque estas fórmulas no calculan el *N* adecuado, en ocasiones son muy útiles para detectar valores de *N* inadecuados. Si el médico dispone sólo de $N = 20$ ratas para tratar con A, le diremos que no procede hacer el estudio, porque tiene potencia del 15%, es decir, aunque con A desarrollen cáncer realmente el 95%, es muy improbable que en el test aparezca $p \leq 0,05$. Por otra parte, si el médico propone usar $N = 150$, diremos que ese *N* es innecesariamente grande, porque con $N = 120$ tendrá potencia del 99,99% (aunque podría estar justificado si se desea estimar la p_2 poblacional con un intervalo de confianza muy estrecho).

Todos los razonamientos expuestos son puramente lógicos y equivalentes al caso en que nos pregunten, por ejemplo, '¿cuántos euros necesito para comparar una bicicleta?' Cuanto más dinero invierta, tendré una bicicleta con mejores prestaciones, pero ninguna cantidad es 'la adecuada'. Con 50 euros compraré una bicicleta de niño, con 200 euros, una de adulto y con 3.000, una de profesional. Del mismo modo, es obvio que 5 euros es una cantidad insuficiente para comprar una bicicleta y 50.000 es innecesariamente grande.

Conclusión

El médico debe tener claro que, cuanto mayor sea *N*, tendrá más probabilidad de encontrar un valor *p* menor, pero hay muy distintos tamaños que le darán información útil. Las fórmulas calculan el *N* necesario para tener ciertas prestaciones, pero no 'el *N* adecuado'. Lo razonable sería que el médico usara, entre los *N* que aportan prestaciones razonables para su estudio, el que más se ajustara a sus posibilidades reales.

Bibliografía

1. Abellán-Huerta J, Prieto-Valiente L. El mito del tamaño de la muestra. Rev Esp Cardiol 2020; 73: 785-6.
2. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. Hoboken: John Wiley & Sons; 2008.
3. Valiente LP, Tejedor IH. Bioestadística sin dificultades matemáticas. Madrid: Ediciones Díaz de Santos; 2010.
4. Martínez-Sellés M, Prieto L, Herranz I. Frequent mistakes in the statistical inference of biomedical data. Ital Heart J 2005; 6: 90-5.
5. Prieto L, Herranz I, Martinez-Selles M, Alonso R. Tests of significance vs tests of hypothesis. Far East Journal of Theoretical Statistics 2007; 21: 97.

Myths and reality about calculating sample size

Abstract. When we decide to conduct a study, one of the first questions that arises is what number of individuals should be included in the sample for it to be 'representative' and for the study to be 'valid'? As in other areas of life, there are many matters for which there is no 'right' amount and different quantities are valid. The same applies here. When asked the question 'How many euros did this bicycle cost?', the answer is a definite number. But the question 'How many euros do I need to buy a bicycle?' can be answered in many different ways, depending on the size and other characteristics of the bicycle. Statistics textbooks contain formulas relating sample size to certain parameters and most doctors believe that one of these will give them the 'right' size for their research, and that by using them their choice of sample size will be justified in the eyes of potential reviewers. This document reflects on the true value of these formulas and how researchers should make proper use of them. It is necessary to show errors and simulations that benefit no one and hinder many by taking up large amounts of time and energy.

Key words. Calculation. Medical research. *p* value. Real effect. Sample size. Statistical power.