Cir Pediatr. 2017; 30: 119-120

Mentiras, malditas mentiras y estadísticas

I. Tuduri Limousin

Servicio de Cirugía Pediátrica. Hospital Universitario Cruces. Barakaldo.

RESUMEN

Con estas líneas, queremos reflejar las limitaciones que tiene el uso de la estadística con pequeños volúmenes de muestra, algo habitual en nuestras series y los errores más habituales que se cometen por interpretación errónea de la *p*.

Planteamos como conclusión una serie de recomendaciones en base a un estudio simulado.

PALABRAS CLAVE: Bioestadística.

LIES, DAMN LIES AND STATISTICS

ABSTRACT

With this article, we want to emphasize the limits of statistics with little samples, a condition very common in our speciality, focusing in the most usual mistakes in the interpretation of the p-value.

Finally, as conclusion, we simulate a clinical study to look what may be a more apropiate boarding.

KEY WORDS: Biostatistics.

INTRODUCCIÓN

Según la definición de la Real Academia de la Lengua Española, la estadística es una "Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades".

Congreso tras congreso, se ve una hiperinflación de variables estadísticas con series ínfimas, que en general podríamos clasificar bajo el aforismo de Mark Twain que he utilizado como título a estas líneas.

Correspondencia: Dr. I. Tuduri Limousin. Servicio de Cirugía Pediátrica. Hospital Universitario Cruces. Pza. Cruces s/n. 48903 Barakaldo.
E-mail: tuduri@yahoo.com

Recibido: Julio 2017 *Aceptado:* Agosto 2017

Nuestro objetivo es plantear las limitaciones de la estadística y tratar de reflexionar acerca de un uso adecuado a nuestras casuísticas.

LA ESTADÍSTICA SOLO ES UNA HERRAMIENTA

La *American Statistical Association* define la p como la probabilidad de que con nuestro experimento cometamos un falso positivo, esto es, que digamos que existe una diferencia en la muestra que no se corresponda con la población. En Medicina, por consenso, si esa probabilidad es menor al 5%, la despreciamos. Es un parámetro probabilístico, nunca algo que sirva para tomar una decisión clínica o de la relevancia del hallazgo detectado[1].

En esa línea, el premio Nobel de Química Ernest Rutherford decía que si tu experimento necesitaba estadística, era necesario diseñar un estudio mejor. Así, la última gran revolución en el campo de la Cirugía Pediátrica podría ser el uso del propranolol para tratar los hemangiomas y su publicación fue una simple carta al director en el *New England*. Si la leemos encontraremos que se citan once pacientes, cero estadísticas y su impacto ha sido infinito, cambiando radicalmente el tratamiento de esta patología. No sería un trabajo mejor con una $p< 0,05$.

Pero además la p tiene dos grandes problemas:

MÚLTIPLES ANÁLISIS

Al repetir los estudios bivariantes lo que haremos será multiplicar las probabilidades. Con un análisis único, tendremos un 5% de probabilidades de que veamos una diferencia y en realidad no lo sea. Para análisis múltiples la fórmula para calcular el riesgo de falso positivo será $1-0,95^n$. Así, con 12 análisis, la probabilidad de un falso positivo sube a 0,45; con 30 ya estaríamos cercanos al 80%[2]. Y, ¿qué podríamos hacer? Un análisis multivariante.

Para lograr un adecuado estudio multivariante necesitamos una buena n, un buen tamaño muestral. Por cada variable que

queramos introducir en el modelo, deberemos reclutar un mínimo de 10-20 pacientes para que los números salgan con fiabilidad, y lo que es más grave, con menos de 10 eventos por variable correremos el riesgo de que los resultados se inviertan[3,4].

Cuando tengamos el volumen necesario, un estadístico nos podrá hacer un análisis multivariante pero el clínico deberá controlar muy bien lo que se incluye en el modelo. Los sesgos de confusión no se pueden controlar con un análisis de este tipo y si hemos incluido un confusor, lo normal es que todo el modelo salte por los aires, aunque nuestro estadístico afirme que matemáticamente es impecable.

El sesgo de confusión debe ser controlado y previsto durante el diseño del estudio y hay que temerlo.

AUSENCIA DE DIFERENCIA ESTADÍSTICA NO QUIERE DECIR IGUALDAD

Este es el error más clásico que tenemos en nuestros estudios. Desgraciadamente nuestros estudios tienen tamaños muestrales pequeños y solo son capaces de detectar diferencias muy grandes. Y eso, si queríamos demostrar la equivalencia del método A con el B nos lleva a una falsa alegría. Una ausencia de diferencia estadística la podemos interpretar como equivalencia cuando en realidad el método B puede que sea superior al A por un 5% y nuestro estudio solo lo detecte a partir del 10%[5].

Cuando hacemos un estudio, normalmente optamos por un retrospectivo y tomamos una muestra que es el total de los pacientes disponibles. Con ello planteamos nuestra estadística y ni se nos pasa por la cabeza pensar en la potencia que tiene el estudio, porque no podemos aumentarla.

Pero el método científico es diferente. Debemos definir *a priori* qué efecto buscamos detectar, p. ej., que tras la cirugía laparoscópica tendremos la mitad de infección de herida que por la técnica abierta. A continuación se define el error β o su inverso, la potencia, que es el error que vamos a tener si no encontramos diferencia o probabilidad de falsos negativos. Con una potencia del 80% (β= 0,20), en caso de no encontrar diferencia, defenderemos que, en este estudio, la probabilidad de que no haya diferencia en la población es de un 80%.

Con el efecto buscado y la β se calcula el tamaño muestral necesario que, desgraciadamente, suele ser enorme.

Y ANTE ESTE PANORAMA, ¿QUÉ PUEDO HACER?

Si tenemos una serie de 30 atresias esofágicas, no tenemos por qué desesperarnos. Basta con cambiar el enfoque:

- Planteémonos qué es lo más destacado y hagamos un estudio con una sola variable. La estadística será débil por tener una n= 30, pero al menos el riesgo de un falso positivo será solo del 5%.
- Si tenemos 6 o 7 variables interesantes y poca muestra, no insistamos en compararlas; mejor los datos crudos. Una tasa de un 20% de dehiscencia mayor con un hilo reabsorbible frente a otro irreabsorbible es lo suficientemente interesante como para no necesitar una mala estadística que la soporte.
- Si la vía toracoscópica ha tenido una tasa de dehiscencia del 18% y la toracotómica del 15% no intentemos vender la moto de que son iguales porque la p sea mayor de 0,05. No hay diferencia porque la potencia del estudio es ínfima.

Para poder afirmar que una diferencia del 3% es relevante con una seguridad del 80% de lo que decimos, necesitamos 1.892 pacientes en cada rama. Es más elegante no estropear un buen resultado con una estadística mal entendida.

Para terminar, volvamos a la definición de la RAE y recordemos que la estadística es una rama de la matemática que emplea *grandes conjuntos de datos*. Por lo que, con volúmenes pequeños, quizás en vez de estadística, deberíamos hablar de torturas a las matemáticas.

BIBLIOGRAFÍA

1. Wasserstein R, Lazar N. The ASA's statement of p-values: context, process and purpose. Am Stat. 2016; 70: 129-33.
2. Molina Arias M. El problema de las comparaciones múltiples. Rev Pediatr Aten Primaria. 2014; 16: 367-70.
3. Ortega Calvo M, Cayuela Domínguez A. Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. Rev Esp Salud Pública. 2002; 76: 85-93.
4. Peduzzi P, Concato J, Feinstein A, Holford T. Importance of events per independent variable in proportional hazards regression analysis. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995; 48: 1503-10.
5. Altman D, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995; 311: 485.