ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Octubre 2020

ORIGINAL

¿Rioja o Ribera? Estadística frecuentista vs bayesiana.

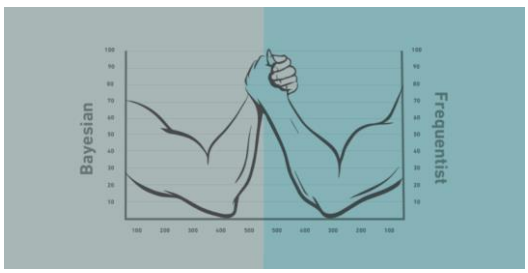
Molina Arias M.

Hospital Infantil Universitario La Paz. Madrid. España.

Resumen

La estadística frecuentista, la más conocida y a la que estamos más acostumbrados, es la que se desarrolla según los conceptos clásicos de probabilidad y contraste de hipótesis. Por su parte, la estadística bayesiana incorpora información externa al estudio que se está realizando, de forma que la probabilidad de un determinado suceso puede verse modificada por la información previa de que dispongamos sobre el suceso en cuestión. Se revisan algunos de los argumentos en contra del abordaje frecuentista, sustentados sobre un mal uso de su metodología más que sobre debilidades propias del método.

Introducción



La estadística frecuentista, la más conocida y a la que estamos más acostumbrados, es la que se desarrolla según los conceptos clásicos de probabilidad y contraste de hipótesis. Por su parte, la estadística bayesiana incorpora información externa al estudio que se está realizando, de forma que la probabilidad de un determinado suceso puede verse modificada por la información previa de que dispongamos sobre el suceso en cuestión. Se revisan algunos de los argumentos en contra del abordaje frecuentista, sustentados sobre un mal uso de su metodología más que sobre debilidades propias del método.

Este es uno de los debates típicos que uno puede mantener con un cuñado durante una cena familiar: si el vino de Ribera es mejor que el de Rioja, o viceversa. Al final, como siempre, tendrá (o querrá tener) razón el cuñado, lo que no impedirá que nosotros nos empeñemos en llevarle la contraria. Eso sí, deberemos plantearle buenos argumentos para no caer en el error, en mi humilde opinión, en que caen algunos al participar en otro debate clásico, este del campo menos lúdico de la epidemiología: ¿estadística frecuentista vs bayesiana?

Y es que estos son los dos abordajes que podemos utilizar a la hora de enfrentarnos con un problema de investigación.

La estadística frecuentista, la más conocida y a la que estamos más acostumbrados, es la que se desarrolla según los conceptos clásicos de probabilidad y contraste de hipótesis. Así, se trata de llegar a una conclusión basándose en el nivel de significación estadística y de la aceptación o rechazo de una hipótesis de trabajo, siempre dentro del marco del estudio que se esté realizando. Esta metodología obliga a estabilizar los parámetros de decisión a

priori, lo que evita subjetividades respecto a los mismos.

El otro enfoque para resolver los problemas es el de la estadística bayesiana, cada vez más de moda y que, como su nombre indica, se basa en el concepto probabilístico del teorema de Bayes. Su característica diferenciadora es que incorpora información externa al estudio que se está realizando, de forma que la probabilidad de un determinado suceso puede verse modificada por la información previa de que dispongamos sobre el suceso en cuestión. Así, la información obtenida a priori se utiliza para establecer una probabilidad a posteriori que nos permita realizar la inferencia y llegar a una conclusión sobre el problema que estemos estudiando.

Esta es otra de las diferencias entre los dos abordajes: mientras que la estadística frecuentista evita la subjetividad, la bayesiana introduce una definición subjetiva (que no caprichosa) de la probabilidad, basada en la convicción del investigador, para emitir juicios sobre una hipótesis.

En realidad, la estadística bayesiana no es nueva. La teoría de la probabilidad de Thomas Bayes se publicó en 1763, pero experimenta un resurgir a partir del último tercio del pasado siglo XX. Y como suele ocurrir en estos casos en que hay dos alternativas, aparecen partidarios y detractores de ambos métodos, que se emplean a fondo para demostrar las bondades de su método de preferencia, a veces buscando más las debilidades del contrario que las fortalezas propias.

Y de esto es de los que vamos a hablar en esta entrada, de algunos argumentos que los bayesianos esgrimen en alguna ocasión que, otra vez en mi humilde opinión, se aprovechan más de un mal uso de la estadística frecuentista por muchos autores, que de defectos intrínsecos de esta metodología.

Comenzaremos con un poco de historia.

La historia del contraste de hipótesis comienza allá por los años 20 del siglo pasado, cuando el gran Ronald Fisher propuso valorar la hipótesis de trabajo (de ausencia de efecto) a través de una observación concreta y la probabilidad de observar un valor mayor o igual al encontrado. Esta probabilidad es el valor p, tan sacralizado y tan malinterpretado, que no significa más que eso: la probabilidad de encontrar un valor igual o más extremo que el encontrado si la hipótesis de trabajo fuese cierta.

En resumen, la p que propuso Fisher no es, ni más ni menos, que una medida de la discrepancia que podía existir entre los datos encontrados y la hipótesis de trabajo planteada, la hipótesis nula (H0).

Casi una década después se introduce el concepto de hipótesis alternativa (H1), que no existía en el planteamiento original de Fisher, y el razonamiento se modifica en función de dos tasas de error de fasos positivos y negativos:

1. Error alfa (error de tipo 1): probabilidad de rechazar la hipótesis nula cuando, en realidad, es cierta. Sería el falso positivo: creemos detectar un efecto que, en realidad, no existe.

2. Error beta (error de tipo 2): es la probabilidad de aceptar la hipótesis nula cuando, en realidad, es falsa. Es el falso negativo: fracasamos en detectar un efecto que, en realidad, existe.

Así, fijamos un valor máximo para el que nos parece el peor de los escenarios, que es el detectar un efecto falso, y escogemos un valor “pequeño”. ¿Cuánto es pequeño? Pues, por

convenio, 0,05 (a veces, 0,01). Pero, repito, es un valor elegido por convenio (y hay quien dice que caprichoso, porque el 5% recuerda el número de los dedos de la mano, que suelen ser 5).

De este modo, si $p < 0,05$, rechazamos H0 en favor de H1. De lo contrario, nos quedamos con H0, la hipótesis de no efecto. Es importante destacar que solo podemos rechazar H0, nunca demostrarla de forma positiva. Podemos demostrar el efecto, pero no su ausencia.

Todo lo dicho hasta ahora parece sencillo de comprender: el método frecuentista trata de cuantificar el nivel de incertidumbre de nuestra estimación para tratar de extraer una conclusión de los resultados. El problema es que la p, que no es más que una forma de cuantificar esa incertidumbre, se sacraliza y malinterpreta con excesiva frecuencia, lo que es aprovechado (si se me permite la expresión) por los detractores del método para intentar poner en evidencia sus debilidades.

Uno de los grandes defectos que se atribuyen al método frecuentista es la dependencia que tiene el valor de p del tamaño de la muestra. En efecto, el valor de la p puede ser el mismo con un tamaño de efecto pequeño en una muestra grande que con un tamaño de efecto grande en una muestra pequeña. Y esto es más importante de lo que pueda parecer en un primer momento, ya que el valor que nos va a permitir llegar a una conclusión va a depender de una decisión exógena al problema que estamos examinando: el tamaño de muestra elegida.

Aquí estaría la ventaja del método bayesiano, en el que muestras más grandes servirían para proporcionar cada vez más información sobre el fenómeno de estudio. Pero yo pienso que este argumento se sustenta sobre una mala comprensión sobre lo que es una muestra adecuada. Estoy convencido, más no siempre es mejor.

Ya otro grande, David Sackett, dijo que "las muestras demasiado pequeñas pueden servir para no probar nada; las muestras demasiado grandes pueden servir para no probar nada". El problema es que, en mi opinión, una muestra no es ni grande ni pequeña, sino suficiente o insuficiente para demostrar la existencia (o no) de un tamaño de efecto que se considere clínicamente importante.

Y esta es la clave del asunto. Cuando queremos estudiar el efecto de una intervención debemos, a priori, definir qué tamaño de efecto queremos detectar y calcular el tamaño muestral necesario para poder hacerlo, siempre que el efecto exista (algo que deseamos cuando planteamos el experimento, pero que desconocemos a priori). Cuando hacemos un ensayo clínico estamos gastando tiempo y dinero, además de sometiendo a los participantes a un riesgo potencial, por lo que es importante incluir solo a aquellos necesarios para tratar de probar el efecto clínicamente importante. Incluir los participantes necesarios para llegar a la ansiada $p<0,05$, además de poco económico y nada ético, demuestra un desconocimiento sobre el verdadero significado del valor de p y del tamaño muestral.

Esta mala interpretación del valor de p es también la causa de que muchos autores que no alcanzan la deseada significación estadística se permitan afirmar que con un tamaño muestral mayor lo habrían logrado. Y tienen razón, hubiesen alcanzado la deseada $p<0,05$, pero vuelven a obviar la importancia de la significación clínica frente a la significación estadística.

Cuando se calcula, a priori, el tamaño de la muestra para detectar el efecto clínicamente importante, se calcula también la potencia del estudio, que es la probabilidad de detectar el efecto si este, en realidad, existe. Si la potencia es superior al 80-90%, los valores admitidos por convenio, no parece correcto decir que no tienes muestra suficiente. Y, claro está, si no has calculado antes la potencia del estudio, deberías hacerlo antes de afirmar que no tienes resultados por falta de muestra.

Otro de los argumentos en contra del método frecuentista y a favor del bayesiano dice que el contraste de hipótesis es un proceso de decisión dicotómica, en el cual se acepta o rechaza una hipótesis como el que rechaza o acepta una invitación para la boda de un primo lejano que hace años que no ves.

Pues bien, si antes se olvidaban de la significación clínica, los que afirman este hecho se olvidan de nuestros queridos intervalos de confianza. Los resultados de un estudio no pueden interpretarse únicamente en base al valor de p. Debemos fijarnos en los intervalos de confianza, que nos informan de la precisión del resultado y de los valores posibles que puede tener el efecto observado y que no podemos concretar más por el efecto del azar. Como ya vimos en una entrada anterior, el análisis de los intervalos de confianza puede darnos información importante desde el punto de vista clínico, a veces, aunque la p no sea estadísticamente significativa.

Por último, dicen algunos detractores del método frecuentista que el contraste de hipótesis adopta decisiones sin considerar la información externa al experimento. Una vez más, una mala interpretación del valor de p.

Como ya contamos en una entrada anterior un valor de $p<0,05$ no significa que H0 sea falsa, ni que el estudio sea más fiable, ni que el resultado sea importante (aunque la p tenga seis ceros). Pero, lo más importante para lo que estamos discutiendo ahora, es falso que el valor de p represente la probabilidad de que H0 sea falsa (la probabilidad de que el efecto sea real).

Una vez que nuestros resultados nos permiten afirmar, con un pequeño margen de error, que el efecto detectado es real y no aleatorio (dicho de otra forma, cuando la p es estadísticamente significativa), podemos calcular la probabilidad de que el efecto sea "real". Y para ello, ¡Oh, sorpresa! tendremos que calibrar el valor de p con el valor de la probabilidad basal de H0, que será asignada por el investigador en base a su conocimiento o a datos previos disponibles (lo cual no deja de ser un enfoque bayesiano).

Como podéis ver, la valoración de la credibilidad o verosimilitud de la hipótesis, una de las características diferenciadoras del enfoque bayesiano, puede también emplearse si utilizamos métodos frecuentistas.

Y aquí lo vamos a ir dejando por hoy. Pero antes de terminar me gustaría hacer un par de consideraciones.

La primera, en España tenemos muchos vinos estupendos por toda nuestra geografía, no solo Ribera o Rioja. Que nadie se me ofenda, he elegido estos dos porque suelen ser los que te piden los cuñados cuando vienen a comer a casa.

La segunda, no me malinterpretéis si os ha podido parecer que soy defensor de la estadística frecuentista frente a la bayesiana. Lo mismo que cuando voy al supermercado me siento contento de poder comprar vino de varias denominaciones de origen, en

metodología de investigación me parece muy bueno tener diferentes formas de abordar un problema. Si quiero saber si mi equipo va a ganar un partido, no parece muy práctico repetir el partido 200 veces para ver qué media de resultados sale. Igual sería mejor tratar de hacer una inferencia teniendo en cuenta los resultados previos.

Y esto es todo. No hemos entrado en profundidad en lo que hemos comentado al final sobre la probabilidad real del efecto, mezclando de alguna manera ambos abordajes, frecuentista y bayesiano. La forma más sencilla, como ya vimos en una entrada previa, es utilizar un nomograma de Held. Pero esa es otra historia…

Bibliografía

– Silva LC, Muñoz A. Debate sobre métodos frecuentistas vs bayesianos. Gac Sanit.2000;14:482-94. (PDF)

– Silva LC, Benavides A. El enfoque bayesiano: otra manera de inferir. Gac Sanit.2001;15:341-6. (PDF)

– Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol.2016;31:337-50. (PubMed)

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en mayo de 2020