Localizador web
Artículo 89.323

Entendiendo la "p < 0,001"

Pere Rebasa
Servicio de Cirugía General. Hospital de Sant Pau. Barcelona. España.

Resumen

El objetivo de este trabajo es dar una orientación básica para el cirujano sobre términos como "grado de significación", "nivel de significación" y "valor de p". Se ofrece una pequeña introducción histórica que nos sitúa en lo que representa la p y para qué sirve. Más adelante explicamos las conclusiones que pueden obtenerse con una prueba de significación, y remarcamos especialmente las precauciones que deben tenerse en cuenta para interpretar correctamente el resultado de una prueba de significación. Finalmente, se dan unas recomendaciones prácticas para tener en cuenta antes y después de hacer una prueba de significación. A lo largo de todo el artículo procuramos evitar la terminología estadística en aras de una mayor claridad en la exposición.

Palabras clave: *Estadística. Bioestadística. Metodología.*

UNDERSTANDING P-VALUES

This article aims to provide surgeons with a basic guide to terms such as significance level and p-value. A short historical introduction explaining the meaning of p-values and what they are used for is provided. Subsequently, we explain the conclusions that can be drawn from a significance test and highlight the precautions that should be borne in mind to correctly interpret the results of these tests. Finally, practical recommendations that should be considered before and after applying a significance test are made. For the sake of clarity we try to avoid the use of statistical terminology throughout the article.

Key words: *Statistics. Biostatistics. Methodology.*

Introducción

Este trabajo pretende dar una serie de criterios sencillos para orientar al lector de revistas científicas y al cirujano que lleva a cabo trabajos de investigación en términos como "grado de significación", "nivel de significación" y "valor de la p". Expondremos para qué sirven, lo que pueden y no pueden demostrar y también qué representa el grado de significación p.

Sólo un poco de perspectiva histórica

El uso de la p en medicina se basa en dos grandes grupos de pruebas estadísticas que hoy en día se hallan totalmente confundidas por la mayoría de nosotros. Por una parte tenemos las pruebas de significación de la hipótesis nula, desarrolladas por Fisher en 1922[1]. A Fisher le preocupaba cómo podíamos alcanzar conclusiones válidas a partir de los resultados de un experimento o unas observaciones. Con este objetivo propuso tres formas diferentes para conseguir un método científico que fuera una buena herramienta para obtener dichas conclusiones a partir de observaciones, uno de los cuales ha llegado hasta nuestros días en la forma de la famosa p. No fue el primero en utilizarla, pero sí lo fue en formalizar la lógica que subyace tras dichas pruebas, y también fue el primero en dar las reglas para su cálculo en múltiples situaciones diferentes.

Por otra parte, tenemos las pruebas de contraste de hipótesis, desarrolladas a partir de las ideas de Fisher por Neyman y Pearson en 1928[2,3]. Hoy en día existe una notable confusión acerca del valor p, que, tal como fue concebido por Fisher, no es compatible con las pruebas de contraste de hipótesis de Neyman-Pearson, con las cuales se ha confundido[4].

¿Pero qué es la p para Fisher?

El valor de p fue diseñado como un método para obtener conclusiones válidas a partir de unos datos (inferen-

Correspondencia: Dr. P. Rebasa Caldera.
Bartomeu Serret Argemí, 35. 08400 Granollers. Barcelona. España.
Correo electrónico: prebasa@hsp.santpau.es

Aceptado para su publicación en abril de 2003.

cia). Y ¿qué significa “inferencia”? Veamos un caso práctico. Tenemos unos datos que hemos obtenido de un estudio. Pero en realidad estos datos no nos interesan en absoluto. Lo que de verdad nos interesa saber es si esos datos son extrapolables, es decir, los podemos aplicar a todos los pacientes que vamos a ver a partir de ese momento. Esto es lo que significa “inferencia”: poder extrapolar en función de nuestros resultados una conclusión aplicable a todos los pacientes. Además, Fisher pretendía que su método no fuera rígido, sino flexible. Y es “flexible” porque, para Fisher, el famoso criterio 0,05 no es un valor inamovible. Puede modificarse como se quiera y se puede usar 0,01, o 0,1, si así se considera. Esta famosa “p” lo que nos permite es dar un criterio “objetivo” a las conclusiones. En cambio, para Neyman-Pearson la p es una regla de decisión. Es decir: si sale cara es blanco, si sale cruz es negro, o, dicho de manera más elegante y científica: si $p \leq 0,05$, la hipótesis verdadera es A, y si $p > 0,05$, la hipótesis verdadera es B.

No es necesario ser un investigador experto para darse cuenta de que los resultados obtenidos en una investigación o en un experimento se basan en mediciones realizadas en una de las infinitas muestras que podrían haberse obtenido de la población de referencia. Para decirlo más claramente, los valores que hemos obtenido pueden presentar fluctuaciones debidas puramente al azar. Es precisamente el análisis estadístico el que nos permite *cuantificar* esta variabilidad, y una vez nos la ha cuantificado (usando para ello la p) nos permite tomar una decisión sobre la verdad o no de la hipótesis.

Veámoslo con un ejemplo. Existen razones teóricas para suponer que la utilización del abordaje laparoscópico disminuye la respuesta inflamatoria del organismo a la agresión quirúrgica. O sea, creemos que los pacientes operados por vía laparoscópica tienen menos respuesta inflamatoria que los que se operan por vía abierta. Aquí, como en casi todos los estudios que nos planteemos, existe una hipótesis nula que, de manera consciente o no, nos planteamos. Y la hipótesis nula de este estudio es: los pacientes operados por vía laparoscópica tienen la misma respuesta inflamatoria que los operados por vía convencional. Es muy importante que entendamos el concepto de hipótesis nula, porque toda la base teórica de la p parte de él, y porque nos limita claramente a la hora de diseñar algún tipo de estudio que no puede basarse en este concepto, como por ejemplo los estudios de equivalencia, recientemente comentados en nuestra revista[5].

Lo siguiente, una vez tenemos bien definida la hipótesis nula, es realizar el estudio. Se toman 40 pacientes operados por laparoscopia y 40 pacientes operados por vía convencional (y vamos a obviar por hoy si es necesario que sean 40 por cada lado, cómo los escogemos y si hace falta pedirles permiso: ésa es otra historia). Miramos cómo tienen la proteína C reactiva (PCR) y obtenemos que la distribución de los niveles de PCR en estos pacientes presenta una media de 16 mg/dl con una desviación estándar de 12 mg/dl. Y además, los que se han operado por vía abierta tienen una PCR media de 25 mg/dl con una desviación estándar de 17 mg/dl. Y ahora hay que preguntarse: ¿la media de 16 que he obtenido en el abordaje laparoscópico es igual a la media de 25 del abordaje abierto? O hemos tenido muy mala suerte y por casualidad (azar) nos ha salido diferente, o realmente estos pacientes han sufrido menos agresión que los otros.

Los pasos que debemos seguir a partir de ahora son sencillos, especialmente si tenemos SPSS al lado y alguien que sepa manejarlo: SPSS nos dirá que la diferencia entre ambas medias tiene una p que es igual a 0,00041. Fisher asigna el valor de esta probabilidad al grado de significación p de nuestro experimento (en notación habitual: nuestro experimento ha sido significativo con una $p = 0,00041$). Explicado claramente: en 4 de cada 10.000 veces nuestros resultados se deben al azar.

Todavía falta el tercer y más importante paso. Se debe evaluar el grado de significación p que hemos hallado. En nuestro ejemplo, ante un valor tan pequeño (4 por 10.000) no dudaremos en rechazar la hipótesis nula, es decir, rechazamos que nuestra muestra pueda proceder de la población general, y, por tanto, concluiremos que la diferencia entre la media experimental y la media teórica es estadísticamente muy significativa ($p = 0,00041$). Si el valor de la p en nuestro estudio no hubiera sido tan pequeño, declararíamos que la diferencia hallada es estadísticamente *no* significativa. Así pues, Fisher considera que el valor p tan pequeño asociado a nuestro experimento es un dato más a favor de la hipótesis de que la laparoscopia disminuye la respuesta inflamatoria del organismo.

¿Qué conclusiones podemos sacar de una prueba de significación?

La primera y la más importante es entender qué nos ha dicho el valor de p. La p es la probabilidad de obtener (asumiendo que la hipótesis nula es cierta, es decir, en nuestro ejemplo: que los valores de PCR en pacientes postoperados de apendicitis es igual por vía laparoscópica que por vía abierta) resultados con la discrepancia igual o superior a la que hemos obtenido con nuestros datos después de realizar el estudio[6]. O, dicho de otra manera: ¿nuestros resultados son “raros” y “diferentes” a lo esperado porque el azar nos ha jugado una mala pasada o son “raros” porque algo hace que sean así? Una hipótesis con una buena base teórica, por ejemplo que la laparoscopia disminuye la agresión quirúrgica, unida a un valor p pequeño nos dan argumentos suficientes para seguir manteniendo la hipótesis que hemos planteado. En caso contrario (p grande), hay que replantearse el problema: ¿la agresión de la cirugía convencional no es lo suficientemente grande? ¿El número de sujetos del experimento es insuficiente para detectar diferencias? ¿Nos hemos equivocado, y la laparoscopia no disminuye la respuesta del organismo a la agresión? ¿La PCR no mide exactamente la respuesta del organismo a la agresión? Es entonces cuando se vuelve a activar el método científico y se puede incluir una modificación (por ejemplo, medir la interleucina 6) y realizar un nuevo experimento.

¿Por qué hemos dicho una hipótesis “con una buena base teórica”? Pues porque para Fisher este punto es indispensable. Los estadísticos jamás se cansan de insis-

tirnos a los clínicos sobre el hecho de que nuestras hipótesis deben tener una buena base teórica. No sirve de nada tener una p significativa sin el apoyo teórico detrás. O, dicho más claramente, no sirve de nada recoger 40 variables diferentes, agitarlas en la coctelera de un SPSS y ver qué es lo que sale significativo para publicarlo después.

Precauciones para interpretar el resultado de una prueba de significación

El rechazo de la hipótesis nula no sugiere causalidad

Esto es fundamental. La obtención de una relación de causa-efecto, que es al fin y al cabo el objetivo de nuestro estudio, sólo puede conseguirse a partir del *diseño* del estudio. Si se ha efectuado un diseño experimental, asignando aleatoriamente a los enfermos a los grupos de cirugía abierta y laparoscópica, entonces sí puede afirmarse que existe esa relación de causalidad. Pero si el diseño del estudio no asegura la comparabilidad de los grupos, entonces no puede establecerse ningún juicio de causalidad, ya que la diferencia puede deberse, por ejemplo, a que todos los pacientes con cirugía abierta tenían una apendicitis gangrenosa y perforada. Es un error grave, y lamentablemente muy frecuente en toda la literatura médica, asumir que una prueba estadísticamente significativa lleva asociada una relación de causa-efecto[7]. En este sentido nos gustaría recomendar el excelente artículo publicado en nuestra Revista en 1996, a cargo de Doménech y Serra, en el cual se dan amplias explicaciones sobre diversos aspectos de la estructura de una investigación científica, haciendo especial hincapié en la importancia del diseño previo al uso de la estadística[7]. Sólo desde un sólido diseño del estudio puede descartarse con absoluta seguridad que llevar calcetines amarillos tenga relación con los niveles de PCR tras la apendicectomía.

Un resultado no significativo no demuestra que la hipótesis nula sea cierta

Es también fundamental entender esto. Un resultado no significativo sólo indica que es compatible con la hipótesis nula porque la discrepancia es pequeña. Es un error también muy frecuente interpretar que el resultado negativo es sinónimo de hipótesis nula demostrada. Nada más lejos de la verdad: los resultados no significativos sólo indican que los datos no consiguen aportar suficientes pruebas para dudar de la credibilidad de la hipótesis nula. De hecho, es perfectamente posible que la hipótesis nula sea falsa, pero no lo hemos detectado porque hemos realizado el estudio con una muestra demasiado pequeña para descubrir el efecto esperado. En nuestro mismo ejemplo, si en lugar de hacer el estudio con 40 pacientes lo hubiéramos hecho con 8, con la misma media y la misma desviación estándar, el resultado de la p sería de 0,067. Es decir, el estudio no sería significativo. Así pues, no significativo es equivalente a no demostrado o no concluyente, pero nunca a ausencia de relación de causa-efecto. Esto nos lleva al siguiente punto:

El resultado estadísticamente significativo no tiene nada que ver con la clínica

La expresión "muy significativo" es un término estadístico que se utiliza para indicar que la hipótesis nula es poco creíble y nada, pero absolutamente nada, tiene que ver con la importancia clínica o biológica de la hipótesis. Un resultado puede ser muy significativo y carecer en absoluto de la menor relevancia clínica. Acudamos de nuevo a nuestro ejemplo y supongamos esta vez que hemos obtenido en 500 pacientes una media ± DE de 17 ± 3 mg/dl. Recordemos que la media del grupo operado por vía abierta es de 16 mg/dl. Esto nos da una $p = 4{,}6 \times 10^{-14}$ (0,000000000000046), resultado con una contundencia estadística ciertamente impresionante. Sin embargo, como clínico me cuesta bastante creer que la diferencia entre una media de 16 mg/dl y otra de 17 mg/dl pueda tener alguna importancia en el curso clínico del paciente… Y esto nos lleva también al siguiente punto:

La p no es una medida de la magnitud del efecto

La expresión "muy significativo" no tiene nada que ver con la magnitud del efecto ni con la intensidad de la relación entre las variables. La significación estadística depende tanto de la magnitud del efecto investigado como del número de sujetos incluidos en el estudio[8], una trampa muy habitual para intentar demostrar cualquier cosa. Es lógico que los estudios realizados con muestras demasiado pequeñas tiendan a dar resultados estadísticamente no significativos a pesar de que el efecto investigado tenga tamaño suficiente para ser considerado clínicamente interesante. De la misma manera, estudios con muestras demasiado grandes tienden a dar resultados estadísticamente muy significativos, aunque el tamaño del efecto investigado sea irrelevante y carezca de interés clínico, hecho que la potente industria farmacéutica conoce perfectamente. Veámoslo con otro ejemplo, tomado de Porta[9]. Hemos estudiado la infección de herida quirúrgica en las apendicectomías gangrenosas tras la utilización de antibióticos profilácticos. En las 200 apendicectomías por cirugía abierta hemos obtenido una infección de herida del 30% en las que no hemos tratado con antibióticos y del 14% en las que sí hemos tratado. Esta diferencia (del 16%) es claramente significativa utilizando una χ^2, con una $p < 0{,}01$ ($\chi^2 = 7{,}46$). Por tanto, en nuestro trabajo concluiremos que la utilización de antibióticos disminuye significativamente la infección quirúrgica en nuestros pacientes. Simultáneamente hemos investigado también la infección de herida quirúrgica en 100 apendicectomías por vía laparoscópica, obteniendo un 30% de infecciones sin utilizar antibiótico y un 14% de infecciones con la utilización de antibióticos. Es decir, otro 16% de diferencia, Ahora bien, el test de χ^2 es igual a 3,73, con una $p > 0{,}05$. Por tanto, concluiremos que la utilización de antibióticos en cirugía laparoscópica no disminuye el porcentaje de infecciones de herida tras la apendicectomía… No, algo no cuadra… La diferencia es exactamente la misma, del 16%; es el número de pacientes lo que varía. Por eso es tan importante entender que la p mide la magnitud del efecto *más* el número de pa-

cientes analizados. Al presentar los resultados de un estudio es muy importante no quedar deslumbrados por el grado de significación estadística, y aprender a describir y valorar nuestros resultados antes de aplicar ningún test estadístico.

¿Y qué son los intervalos de confianza?

Un intervalo de confianza es una forma de expresar la precisión estadística de una forma clínicamente útil. A menudo pensamos que un intervalo de confianza es la probabilidad de que el verdadero parámetro que hemos estimado esté situado dentro de ese intervalo. Nada más lejos de la verdad. De hecho, los intervalos de confianza están completamente determinados por los resultados del estudio que los ha generado. Los intervalos de confianza son útiles porque nos definen un límite superior y un límite inferior que es consistente con los datos de nuestro estudio, pero no nos da ninguna probabilidad de saber dónde está el verdadero parámetro que intentamos determinar. Entones, ¿para qué sirven? Una de las ventajas de usar intervalos de confianza al expresar nuestros resultados es que no los reducen a “blanco o negro”, como por desgracia ocurre cuando el resultado sólo se expresa con una p (que suele expresarse como “significativa” o “no significativa”)[9,10]. En nuestro ejemplo, el intervalo de confianza para nuestra muestra de media ± DE de 16 ± 12 está entre 12,3 y 19,7 mg/dl. Incluso podemos presentar los datos de una manera todavía más elegante y muy clínica, que es dar el intervalo de confianza de la diferencia. En nuestro ejemplo, podemos afirmar que los niveles de PCR después de la apendicectomía laparoscópica se sitúan entre 2,5 y 15,6 mg/dl por debajo de los niveles tras la cirugía convencional (o quedarnos con la frase de siempre: los niveles de PCR tras la laparoscopia son más bajos que la convencional con una p = 0,000041…).

Recomendaciones para tener en cuenta antes y después de hacer pruebas de significación estadística

Del excelente artículo de Porta[9] hemos extraído las siguientes recomendaciones para cualquier cirujano que pretenda utilizar en sus estudios las pruebas de significación estadística. Se dividen en recomendaciones *antes* de utilizarlas y *después* de hacerlo:

Antes

– Perdamos tiempo en revisar críticamente la literatura. Antes de iniciar el estudio hay que pensar en lo que nosotros, nuestros colegas y los pacientes necesitan saber.

– Limitémonos a una o, como mucho, dos hipótesis operativas. Los trabajos que pretenden averiguar un montón de cosas acaban por no averiguar nada útil.

– Escojamos el diseño más adecuado y realista para el estudio. Es aquí cuando se debe consultar al estadístico experimentado, y si conoce el ámbito médico, mucho mejor. Un buen estudio con un mal análisis estadístico puede arreglarse. Un estudio mal diseñado de entrada no puede arreglarse con ninguna prueba estadística, por sofisticada que sea.

– Definamos concretamente qué variables se medirán y qué relaciones entre ellas se considerarán. No vale medirlo todo, porque introducirá mucho “ruido” en el análisis, y no vale cruzar todas las variables entre sí una vez las hayamos recogidos “para ver qué sale”.

– Escojamos las pruebas estadísticas que serán necesarias, y si han de ser unilaterales o bilaterales.

– Decidamos un riesgo de significación estadística (alfa) teniendo en cuenta el número de pruebas que pensamos aplicar cuando tengamos los datos. Si aplicamos muchas pruebas estadísticas, hay que reducir el riesgo alfa.

– Escojamos un riesgo beta y calculemos el número de pacientes necesarios para extraer conclusiones. Nunca debemos empezar un estudio en el que los resultados puedan ser ininterpretables por falta de pacientes: es una pérdida de tiempo y de dinero.

– Recojamos únicamente los datos esenciales y relevantes para poner a prueba nuestra hipótesis, asegurándonos de su veracidad. La tentación de mirar también esa variable “por si acaso” o “porque ya que estamos no cuesta nada” acaba siempre en unos protocolos con decenas de variables inútiles y farragosos de rellenar, lo que acaba llevando a errores y abandonos prematuros del estudio.

Después

– No nos dejemos fascinar por la p. Hay que describir los resultados con independencia de la significación estadística, preferiblemente con tablas y figuras. Hubo una época en que los trabajos en medicina consistían en una detallada explicación de cómo se había llevado a cabo, una presentación de los datos muy precisa y la interpretación que el autor hacía de esos datos, dejando que el lector sacara sus conclusiones. En los últimos años los trabajos son una explicación detallada de cómo se ha llevado a cabo, la aplicación de 20 supermega-tests estadísticos impresionantes y la interpretación del autor sobre lo que quieren decir esos tests, dejando al lector con la única opción de “créetelo, lo dice la p”.

– Calculemos la *magnitud* de la asociación, la diferencia o el riesgo, no sólo su *dirección*.

– Calculemos el grado de significación p y los intervalos de confianza.

– Si la diferencia no ha resultado ser estadísticamente significativa, calculemos la potencia estadística. Al menos sabremos si podemos arreglarlo reclutando más casos.

– Juzguemos la relevancia clínica de nuestros resultados, tanto si son estadísticamente significativos como si no lo son.

– Hay que discutir los posibles errores en el estudio y las explicaciones alternativas a los resultados obtenidos.

– Propongamos otros estudios concretos para ir más allá de donde nuestro estudio ha llegado.

Conclusiones

De todo lo que hemos explicado en este artículo, pueden extraerse unas pocas conclusiones:

1. Las pruebas estadísticas jamás pueden sustituir el juicio clínico. Es igual lo significativa que sea la p, lo importante es lo significativa que es la clínica.

2. Las hipótesis deben formularse antes del estudio, y se ponen a prueba con las pruebas estadísticas. No al revés (primero hacemos el estudio y luego introducimos los datos en la batidora SPSS a ver qué sale).

3. Y lo más importante de todo. A lo largo del artículo hemos asumido que los datos son válidos, es decir, el estudio está libre de sesgos. La estadística no puede corregir un mal diseño. Un estudio con grupos no comparable, con datos clínicos erróneos, o que no tiene en cuenta factores de confusión, no mejorará por muchas p significativas que obtengamos.

Agradecimientos

Queremos agradecer a los Dres. Eduardo Targarona y Helena Vallverdú sus consejos durante la escritura de este artículo. Sin ellos, hubiera sido ininteligible.

Bibliografía

1. Fisher RA. On the mathematical foundations of theoretical statistics. Philosophical transactions of the Royal Society of London 1922; 222A:309-68.
2. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference (Part I). Biometrika 1928;20A:175-240.
3. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference (Part II). Biometrika 1928;20A:263-94.
4. Goodman S. P values, hypothesis tests and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993;137:485-96.
5. Escrig VJ. Comparaciones entre cirugía convencional y laparoscópica. Ha llegado el momento de los estudios de equivalencia. Cir Esp 2003;73:75-7.
6. Browner WS, Newman TB. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987;257:2459-63.
7. Doménech JA, Serra J. Diseño y estadística (I): ¿Cómo plantear un estudio de investigación quirúrgica? Cir Esp 1996;60:307-18.
8. Domènech JM. Comprobación de hipótesis. Pruebas de significación y pruebas de hipótesis. En: Métodos estadísticos en ciencias de la salud. Barcelona: Signo, 1998.
9. Porta M, Plasencia A, Sanz F. La calidad de la información clínica (y III): ¿estadísticamente significativo o clínicamente importante? Med Clin 1998;90:463-8.
10. Bulpitt CJ. Confidence intervals. Lancet 1987;1:494-7.