Rev. Chilena de Cirugía. Vol 60 - Nº 1, Febrero 2008; págs. 86-89

DOCUMENTOS

El valor de "p" y la "significación estadística". Aspectos generales y su valor en la práctica clínica*

Interpretation of medical statistics

Drs. CARLOS MANTEROLA D.,[1,2] VIVIANA PINEDA N.[1], GRUPO MINCIR

[1]Departamento de Cirugía, Facultad de Medicina, [2]CIGES (Capacitación, Investigación y Gestión para la Salud Basada en Evidencia), Universidad de La Frontera, Temuco, Chile.

INTRODUCCIÓN

Desarrollar cualquier estudio clínico tiene como objetivo poner de manifiesto la existencia de asociación entre variables. Sin embargo, esta asociación puede ser real o ficticia, ya sea producto del azar, la existencia de sesgos, presencia de confundentes, etc.

Es quizás por esto que los clínicos, nos basamos habitualmente en la "significación estadística" para nuestra toma de decisiones. Este artículo, intenta poner una nota de alerta al respecto en relación a que, si bien es cierto que es una herramienta útil, no necesariamente va ligada a la relevancia clínica del fenómeno en estudio; esto se debe a que la "significación estadística" puede no resolver la incertidumbre clínica ante un escenario puntual, dado que es un concepto exclusivamente matemático y no de garantía de calidad.

El concepto "significación estadística" se relaciona con la necesidad de "probar hipótesis", situación a la cual los clínicos no estamos habituados y, quizás, sea ésta una de las razones por las que confiamos tanto en el concepto de "significación estadística" y nos dejamos llevar por el "valor de p".

Antes de valorar el "valor de p", es relevante tener en cuenta que este concepto depende de dos elementos esenciales: la magnitud de la diferencia que queremos probar y el tamaño de la muestra; si estos elementos no están adecuadamente considerados en el estudio permitirán la generación de resultados espurios, que pueden finalmente llevar a la toma incorrecta de decisiones, ya sea por errores de tipo I ó II.

Existen formas más apropiadas de representar los resultados en investigación clínica como la razón de odds, el riesgo relativo, el número necesario de pacientes a tratar para reducir un evento, entre otras, que se asocian a la significación clínica y permiten dilucidar de mejor forma la incertidumbre existente frente a una situación clínica puntual.

Desarrollar cualquier estudio clínico tiene como objetivo poner de manifiesto la existencia o no de asociación entre diversas variables. La asociación encontrada puede ser real; sin embargo, con mayor frecuencia de la que uno se imagina ésta es producto del azar, de la existencia de sesgos, de la presencia de variables de confusión o de la variabilidad biológica del fenómeno en estudio. Para dilucidar este problema existen una serie de pasos fundamentales al momento de diseñar y conducir una investigación; y, posteriormente, al momento del análisis de los datos, que es donde aparece recién la utilización de herramientas estadísticas tanto de carácter descriptivo como analítico. Y es la utilización de estas últimas la que permite

*Recibido el 1 de Julio de 2007 y aceptado para publicación el 28 de Agosto de 2007.

Correspondencia: Dr. Carlos Manterola D.
Casilla 54-D, Temuco, Chile.
Fax: 56-45-325761
e-mail: cmantero@ufro.cl

generalizar resultados, o inferir los resultados obtenidos de la muestra estudiada a la población blanco que la generó[1].

Por todo lo anteriormente expuesto es que resulta fundamental el cuidadoso diseño del estudio, tomar en consideración los criterios de selección y la estimación del tamaño de la muestra, puesto que mientras más grande es el tamaño de la muestra, mayor es la precisión; y por ende, la variabilidad secundaria al azar se reduce. De todos modos, el rol que siempre jugará el azar debe tenerse en cuenta, evaluarse y medirse, por ejemplo considerando los intervalos de confianza que nos permiten conocer la precisión de la estimación dentro de un margen de error previamente establecido[2,3].

Es por todo esto que, desde la perspectiva clínica, el concepto de "significación estadística" no es relevante, pues no resuelve la incertidumbre. Se debe tener en cuenta que estamos hablando de un concepto matemático, por lo que una asociación estadísticamente significativa puede no ser clínicamente relevante; una asociación estadísticamente significativa puede no ser causal; y una asociación estadísticamente no significativa puede deberse a un problema de tamaño de muestra insuficiente. Es decir, podemos encontrar asociaciones "estadísticamente significativas y conceptualmente espurias"[4]; por ello, hay que tener siempre presente que el término "estadísticamente significativo" no es "garantía de calidad".

El concepto "significación estadística" se relaciona con la necesidad de "probar hipótesis". Este proceso se realiza utilizando "pruebas de hipótesis", las que permiten cuantificar hasta que punto la variabilidad de la muestra en estudio es responsable de los resultados obtenidos en el estudio. Es así como H_0 o hipótesis nula, representa la afirmación de que no hay asociación entre las dos variables; y H_a, o hipótesis alternativa, afirma que existe asociación entre las dos variables. Entonces, la estadística nos permite decidir sobre que hipótesis debemos elegir, lo que será con el nivel de seguridad que previamente se haya establecido por el equipo de investigación (habitualmente en clínica es 95%).

Las pruebas estadísticas funcionan entonces de la siguiente forma: se verifica la magnitud de la diferencia existente entre los grupos a comparar (A y B). Si esta magnitud es mayor que un error estándar definido multiplicado por una seguridad definida, concluimos que la diferencia entre A y B es significativa; por ende, "se rechaza la hipótesis nula" y se "acepta la hipótesis alternativa". Por ejemplo, en un estudio en el que se compararon los resultados obtenidos en 641 pacientes colecistectomizados por vía laparoscópica (199 disecados con KTP/532 láser y 442 disecados con electrocirugía monopolar), se observó desarrollo de complicaciones graves en 16 pacientes disecados con KTP/532 láser y 11 con electrocirugía monopolar. ¿Existe diferencia significativa respecto del porcentaje de complicaciones graves entre ambas técnicas de disección?[5].

H_0 (hipótesis nula)= No hay diferencia entre ambas técnicas de disección.

H_a (hipótesis alternativa)= Sí existe diferencia entre ambas técnicas de disección.

Tratamiento	*Nº pacientes*	*Respuesta*	*p*
KTP/532 láser	199	16/199 = 0,080	p1
Electrocirugía monopolar	422	11/422 = 0,026	p2

Si $[p_1 - p_2]$ es mayor que el producto de 1,96 ($Z\alpha$-0,05) multiplicado por el error estándar, concluimos que la diferencia es significativa. Por lo tanto, hemos de calcular el error estándar para luego compararlo con la diferencia observada en los grupos en estudio.

$[p_1 - p_2] = [0,080 - 0,026] = 0,054$

$p= [p_1 + p_2] / 2 = [0,080 + 0,026] / 2= 0,053$

El error estándar se calcula de la siguiente forma:

$$\text{Error estándar} = \sqrt{p\,(1-p)\,(1/n_1 + 1/n_2)} = \sqrt{0,053\,(1-0,053)\,(1/199 + 1/422)} = 0,00035$$

Error estándar multiplicado por $Z\alpha$-0,05 = 0,00035 * 1,96= 0,00069

Entonces, si la diferencia de [p1 - p2] = 0,054 supera al error estándar multiplicado por $Z\alpha$-0,05 (0,00069) concluimos que existe una diferencia estadísticamente significativa entre los grupos en estudio; razón por la cual rechazamos H_0, por ende, aceptamos la H_a.

No obstante ello, se ha de tener en cuenta que el rechazo de H_0 tiene implícito el riesgo de cuantificar el "valor de p", que representa la probabilidad de aceptar la H_a, cuando en realidad la hipótesis correcta podría ser H_0.

El "valor de p" que indica que la asociación es estadísticamente significativa ha sido arbitrariamente aceptado por consenso; y, en clínica, se admite 0,05. Dicho en otros términos, esto representa una seguridad del 95% que la asociación que estamos estudiando no sea por el azar; por lo que si queremos trabajar con un margen de seguridad de 99%, éste lleva implícito un valor de p inferior a 0,01.

Pero ¿qué significa que el "valor de p" sea superior a 0,05? Entonces hemos de plantearnos

que los resultados pueden estar influidos por el azar y entonces no podemos rechazar H_0, que avala que las variables no están asociadas[6].

Sin embargo, es relevante tener en cuenta que el concepto de "significación estadística" depende de dos elementos esenciales: la magnitud de la diferencia que queremos probar y el tamaño de la muestra. Con respecto a la magnitud de la diferencia, es importante comprender que a mayor diferencia entre las variables en estudio, más fácil será poder demostrar que la diferencia es significativa; al revés, si la diferencia es pequeña las posibilidades de detectar diferencias se minimizan. Ahora, respecto del tamaño de la muestra, es fácil comprender que mientras mayor sea éste, más fácil será detectar diferencias entre las variables en estudio; entonces, cuando las diferencias son pequeñas se requiere de muestras de gran tamaño; al revés, cuando las diferencias son grandes se necesita de muestras pequeñas para conducir el estudio. Así, el tamaño de la muestra afecta la significación estadística a través del error estándar que se hace más pequeño cuantos más pacientes tenga el estudio. En resumen, cualquier diferencia entre las variables en estudio puede ser "estadísticamente significativa" si se dispone del número suficiente de pacientes.

Por ejemplo, en un estudio referente a tratamiento de cáncer gástrico, al calcular la muestra necesaria con una supervivencia de 50% para el grupo de cirugía y quimiorradioterapia y 41% para el grupo con cirugía exclusiva (resultados reportados en el artículo), un alfa de 0,05 y una potencia de 80%, la muestra necesaria para la conducción del estudio es de 960 sujetos (480 por grupo); y no 556 en total (281 para el grupo de cirugía y quimiorradioterapia; 275 para el grupo de cirugía exclusiva)[7,8].

Con este ejemplo, se hace patente la aparición de dos conceptos: el de error tipo I o alfa y el de error tipo II o beta. El error tipo I corresponde a un "falso positivo", es decir rechazar la H_0 cuando en realidad es verdadera; en términos más sencillos, creer que hay una asociación estadísticamente significativa cuando no la hay, que es lo que ocurrió en el estudio de MacDonald[7,8]. Éste es un claro ejemplo que el "valor de p" no es un indicador de fuerza de una asociación, como tampoco de su importancia (para ello existen la razón de Odds, el riesgo relativo, etc.). Por otra parte, el error tipo II corresponde a un "falso negativo", es decir, consiste en aceptar H_0 cuando es falsa; en términos más sencillos, creer que no existe una asociación estadísticamente significativa cuando quizás la hay[9].

¿Cómo reducir la probabilidad de cometer un error tipo I? Realizar un adecuado diseño y planificación del estudio de forma tal de evitar buscar asociación entre variables "por si resulta" o "disparar a la bandada esperando que caiga algo"; reducir el número de pruebas estadísticas a utilizar, sólo a las necesarias, evitando sobreutilizar herramientas estadísticas; limpiar la base de datos para evitar errores de valores extremos que puedan producir hallazgos falsamente significativos; recurrir al uso de valores de alfa más pequeños o reducir los intervalos de confianza (0,01 ó 0,001); observar si los resultados del estudio se pueden reproducir.

¿Cómo reducir la probabilidad de cometer un error tipo II? Incrementar el tamaño de la muestra, evaluar el poder estadístico del estudio, aumentar el tamaño del efecto a detectar, elevar el valor de alfa y utilizar pruebas estadísticas más robustas como las denominadas pruebas paramétricas (t-test, ANOVA, etc.).

Por todas las razones antes expuestas, más relevante que hablar de "significación estadística" es utilizar el concepto de "relevancia clínica"; esto, debido a que la relevancia clínica de un fenómeno va más allá de cálculos matemáticos y depende de la gravedad del problema, la morbilidad y mortalidad generada por el mismo, la magnitud de la diferencia, la vulnerabilidad, los costes involucrados, etc.

De este modo, las formas más apropiadas de representar los resultados en investigación clínica son la razón de odds, el riesgo absoluto (RA), el riesgo relativo (RR), la reducción relativa del riesgo (RRR), la reducción absoluta del riesgo (RAR), el número necesario de pacientes a tratar para reducir un evento (NNT) y el número necesario de pacientes a dañar (NND)[10-12]; y la significación estadística no es nada más que eso, "la significación estadística", que en ocasiones puede ser positiva y clínicamente irrelevante, o negativa, sin que eso signifique necesariamente que no hay diferencias reales entre las variables en estudio.

Por ello es que el "valor de p", debe ser observado con cautela y siempre tomado en cuenta en el contexto del estudio, su diseño, las características de la muestra o la población en estudio, de los potenciales sesgos, etc. Y no como una cifra mágica que nos seduzca de tal forma, que nos invite o autorice a tomar decisiones o cambiar conductas relacionadas con la práctica clínica cotidiana.

Por último, antes de tomar decisiones o cambiar conductas basadas en un "valor de p", se ha de considerar además la validez externa o generalización de los resultados obtenidos en ese estudio respecto de la población blanco y, particularmente, respecto de nuestros pacientes o nuestra realidad

laboral, que pueden ser no necesariamente equivalentes a los utilizadas en el estudio valorado por nosotros.

REFERENCIAS

1. Manterola C. El proceso que conduce al desarrollo de la investigación científica. Su aplicación en cirugía. Rev Chil Cir 2001; 53: 104-109.
2. Altman DG. Confidence intervals for the nember needed to treat. BMJ 1998; 317: 1309-1312.
3. Daly LE. Confidence limits made easy: interval estimation using a subsitution method. Am J Epidemiol 1998; 147: 783-790.
4. Silva Ayçaguer LC. Cultura estadística e investigación científica en el campo de la salud: una mirada crítica. Editorial Díaz de Santos, Madrid, 1997.
5. Lane GE, Lathrop JC. Comparison of results of KTP/532 laser versus monopolar electrosurgical dissection in laparoscopic cholecystectomy. J Laparoendosc Surg 1993; 3: 209-214.
6. Jekel JF, Elmore JG, Katz DL. Epidemiology Biostatistics and Preventive Medicine. WB Saunders Company, Philadelphia, 1996.
7. MacDonald JS, Smalley SR, Benedetti J, Hundahl SA, Estes NC, Stemmermann GN *et al.* Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or the gastroesophageal junction. New Engl J Med 2001; 345: 725-730.
8. Manterola C, Torres R, Burgos L, Vial M, Pineda V. Methodological quality of an article on the treatment of gastric cancer adopted as protocol by some Chilean hospitals. Rev Med Chil 2006; 134: 920-926.
9. Daly LE, Bourke GJ. Interpretation and uses of medical statistics. Blackwell science, Oxford, 5th ed, 2000.
10. Cook RJ, Sackett DL. The number needed to treat: a clincally useful measure of treatment effect. BMJ 1995; 310: 452-454.
11. Laupacis A, Sackett DL, Roberts RS: An assesment of clinically useful measures of treatment. New Engl J Med 1988; 318: 1728-1733.
12. Sackett DL, Richarson WS, Rosenberg W, Hynes RB. Evidence-based medicine: how to practice and teach EBM. Churchill-livingstone; London, 2nd ed. 2000.