Contra la sumisión estadística: un apunte sobre las pruebas de significación

Autores:

[1]Alina Benavides Rodríguez [2]Luis Carlos Silva Ayçaguer

[1] Especialista de Primer Grado de Bioestadística, Dirección Provincial de Salud, Villa Clara.
[2] Investigador Titular, Vicerrectoría de Investigación y Postgrado, ISCM/H.

Dirección de Contacto:
Luis Carlos Silva Ayçaguer. Vicerrectoría de Investigación y Postgrado.
Instituto Superior de Ciencias Médicas de la Habana (ISCM/H).
Edificio Ramón Paz, 6º piso, C/ G y 25, Municipio Plaza, Ciudad de La Habana (Cuba).

Resumen / *Abstract*

• Las pruebas de significación surgen en la década de los años 40 a partir de la unión de dos teorías en buena medida incompatibles y, a pesar de sus limitaciones conceptuales y prácticas, son consideradas como un único y coherente enfoque de inferencia estadística. Este artículo ilustra a través de un ejemplo artificial, aunque típico de la investigación sanitaria, una limitante esencial de las pruebas de significación: básicamente, que el rechazo de la hipótesis nula queda asegurado con un tamaño de muestra suficientemente grande. Se bosquejan asimismo otras endebleces tales como que no toman en cuenta la información proveniente de estudios previos o de la experiencia empírica informalmente acumulada, así que no manejan el resultado como un modo de tomar decisiones clínicas. En consonancia con esto, se fundamenta que es posible prescindir de este método en la investigación, máxime cuando se cuenta con alternativas, una de ellas incluso muy sencilla, como la descripción valorativa de los resultados a través de intervalos de confianza. Se destacan los métodos bayesianos como una posibilidad que, si bien tiene cierta complejidad, se vislumbra como un recurso en desarrollo y altamente promisorio.

Palabras claves:
Estadística; pruebas de significación; hipótesis; intervalos de confianza

Against the statistical submission: a note on the significance tests

• *The significance tests, as they are presently known, appeared in the 1940s; they are the result of the combination of two rather incompatible theories. In spite of their conceptual and practical limitations, they have been generally considered as a single and coherent approach to make statistical inferences. By means of an artificial example, which at the same time is quite typical in the frame of medical research, this article illustrates one of the most important limitations of the significance tests. Basically, that the rejection of the null hypothesis could always be possible, provided that a large enough sample size is used. Other drawbacks are also mentioned: not to take into account the background knowledge and to use the findings as a way to take clinical decisions. It is explained that it is possible to avoid such an approach in the investigation. Some alternatives are presented; the most simple of them consist in the description of the findings by means of confidence intervals. The most complex is the bayesian theory which seems a very promising approach.*

Key words:
Statistic; the significance tests; hypothesis; confidence intervals

Introducción

La mayoría de los investigadores sanitarios de la primera mitad del siglo XX no poseían los conocimientos necesarios ni la debida asesoría para aplicar adecuadamente los métodos bioestadísticos. La ausencia de una herramienta que aquilatara cuantitativamente el significado de los resultados produjo que las anécdotas clínicas poblaran las revistas médicas. Ello condujo a que los editores comenzaran a exigir el uso de procedimientos que cuantificaran la evidencia y complementaran los razonamientos verbales; desde entonces, los investigadores se interesan vivamente por dominar métodos para conseguirlo (1).

Este fue el caldo de cultivo para lo que hoy conocemos como prueba de hipótesis, procedimiento que ha alcanzado un lugar prominente en la investigación sanitaria, considerado incluso por muchos como un árbitro llamado a decir la última palabra (2).

Sin duda, el acceso universal que se tiene en la actualidad a los ordenadores ha venido a consolidar tal tendencia por parte de los investigadores, amén de la sostenida presión de muchos editores en esa dirección.

Pero lo cierto es que las pruebas de hipótesis están congénitamente cuestionadas en tanto procedimiento idóneo para cumplir la tarea que les dio origen. En este artículo nos proponemos bosquejar parte de sus limitaciones y comentar algunas alternativas que pueden permitir, con tanto o mayor éxito, un análisis adecuado de los datos.

El procedimiento de prueba de hipótesis utilizado en la actualidad está lejos de ser imprescindible en la investigación sanitaria, dado que existen alternativas mucho más sólidas y cercanas al sentido común

Breve reseña histórica

Las pruebas de significación, tal y como las conocemos hoy en día, son el resultado de la fusión de dos métodos: uno propuesto por Ronald Fisher en la década de los 20 para valorar el grado de incompatibilidad de los datos con una hipótesis, y otro centrado en la elección entre dos hipótesis, llamado prueba de hipótesis y desarrollado en los primeros años de la década de los 30 por Jerzy Neyman y Egon Pearson (3).

El enfoque de Fisher comienza por establecer una hipótesis nula. Para el caso frecuente en que se trata de dos grupos que han de ser comparados, ella afirma que éstos son esencialmente equivalentes; la diferencia observada en los dos grupos (d_0) y el error estándar de la diferencia se usan para construir el estadístico por conducto del cual se realizará la comparación. En ese punto se puede calcular p; es decir, la probabilidad, bajo el supuesto de que la hipótesis nula es cierta, de encontrar una diferencia tanto o más grande que el valor observado d_0; si este valor p es pequeño, se rechaza la hipótesis nula, y el resultado observado se considera significativo (4).

Por su parte, el enfoque de Neyman y Pearson plantea dos hipótesis: una nula, como antes, y una hipótesis alternativa, la cual se opone a la hipótesis nula (por ejemplo, afirma que el efecto debido a cierto tratamiento que distingue a los grupos no es cero). La idea de la hipótesis alternativa aparece así por primera vez. Este método fue concebido para tomar decisiones: rechazar una hipótesis y aceptar la otra, solamente sobre la base de los datos. Por tanto, los investigadores enfrentan el riesgo de cometer dos tipos de error, aceptar que los efectos de las dos terapias difieren cuando realmente son equivalentes (error tipo I), o rechazar que sean iguales cuando en realidad difieren (error tipo II) (3).

Alrededor de la Segunda Guerra Mundial, ambos métodos se unen de manera anónima como un intento de conciliar estas perspectivas originalmente contrapuestas, y dan lugar a la prueba estadística de hipótesis actual. El método toma de Fisher su valor p para ser usado como un índice que mide la fuerza de la evidencia, y de Neyman y Pearson el propósito de adoptar una decisión, consistente en "rechazar la hipótesis nula" si el valor de p es pequeño (típicamente, cuando $p<0,05$) y en "no rechazar" cuando p no es tan pequeño. Nótese que esto último no equivale a "aceptar" la hipótesis nula, ya que según este método híbrido, dicha hipótesis no se acepta nunca.

Sin duda, este recurso constituyó un notable paso de avance tanto matemático como conceptual. De hecho, se convirtió en una estrategia ampliamente promovida y aceptada en el mundo investigativo, probablemente porque, tanto para los investigadores como para los editores de revistas y responsables administrativos, resulta muy atractivo contar con procedimientos cuantitativos que generen conclusiones independientemente de las personas que realizan el estudio.

Su funcionamiento interno

Veamos un ejemplo muy simple que nos permitirá comprender mejor el funcionamiento de esta técnica y, más adelante, sus limitaciones.

Supongamos que hay motivos teóricos e indicios empíricos nacidos del trabajo de Enfermería que hacen pensar que los pacientes afectados por quemaduras se recuperan más rápidamente cuando el tratamiento combina cierta crema antiséptica con un apósito hidrocoloide que cuando se utiliza la crema antiséptica solamente.

Se diseña, entonces, un experimento con la esperanza de rechazar la hipótesis nula que afirma que el tratamiento simple es tan efectivo como el combinado. Imaginemos que se tienen 80 pacientes; aleatoriamente se eligen 40 que son atendidos con el tratamiento experimental (combinación: crema antiséptica y apósito hidrocoloide), en tanto que a los 40 restantes se les aplica el tratamiento convencional (crema únicamente).

Una vez obtenido el dato (porcentajes de recuperación en uno y otro grupo, p_1 y p_2, y su diferencia, $d_0=p_1-p_2$), se calcula la probabilidad asociada a ese resultado bajo Ho. Supongamos que el 75% $(p_1=0,75)$ de los pacientes bajo el tratamiento experimental mejora apreciablemente a los 5 días, mientras que para los pacientes tratados de manera convencional, esta tasa de recuperación fue del 60% $(p_2=0,60)$. La Tabla 1 recoge la información relevante de este ejemplo. Según la práctica regular, ahora sólo resta aplicar la prueba estadística más usada en estos casos para valorar la diferencia de porcentajes: la prueba Ji-cuadrado. Es fácil corroborar que $x^2_{obs}=2,05$ y que el valor p que corresponde es igual a 0,15.

Tabla 1. Distribución de una muestra de 80 pacientes según tratamiento asignado y según se recuperarán o no

Tratamiento	Recuperación		Total
	Sí	No	
Experimental	30	10	40
Convencional	24	16	40
Total	54	26	80

Puesto que tal valor de p obtenido no es lo suficientemente pequeño como para ser considerado significativo a ninguno de los niveles habituales (0,10; 0,05 y 0,01), según la práctica al uso, el investigador tiene que concluir (aunque quizás lo haga a regañadientes) que no tiene suficiente evidencia muestral como para afirmar que el tratamiento con crema y apósito sea más efectivo que el tratamiento con crema solamente, de manera que no rechazará la hipótesis nula.

Limitaciones

Prácticamente todos los textos sobre estadística inferencial dan por sentado que las pruebas de significación constituyen un procedimiento sin fisuras, con un sólido respaldo matemático, ignorando sus limitaciones conceptuales y prácticas. Si bien algunas fuentes bibliográficas hacen consideraciones críticas, en la inmensa mayoría de los textos aplicados y en las revistas del mundo sanitario, el método se presenta como una verdad, con escasísimas alusiones a cualquier controversia.

De manera que ni la literatura general ni los programas docentes de estadística informan a sus usuarios de sus contradicciones (5), ni del intenso debate desarrollado durante casi 70 años por muchos estadísticos en relación con la "solidez" de las pruebas de significación (6), lo que ha provocado que los profesionales sanitarios lo desconozcan y contribuido a abonar en los investigadores la errónea convicción de que sus resultados tendrán más rigor científico por el sólo hecho de que vengan acompañados de un valor *p*.

Las observaciones críticas que se han venido acumulando desde la creación de las pruebas de significación conforman hoy un reclamo metodológico de tal magnitud que cada día se tornan más difícil soslayar. Las objeciones más obvias son las siguientes.

Factores tales como la plausibilidad biológica de la hipótesis alternativa y la fuerza de los resultados precedentes no se articulan formalmente al proceso inferencial; el método como tal no los toma en cuenta, hecho bastante sorprendente y contrario a la intuición, hasta el punto de llegar a ser, para algunos, anticientífico (3).

Si el valor *p* es pequeño, se "rechaza" la posible validez de la hipótesis nula; en caso contrario, no se toma decisión alguna (7). El valor de *p* se interpreta en función de un umbral mágico por encima del cual un resultado es demostrativo de algo y, por debajo del cual, no nos dice nada. También esto es chocante, pues lo lógico sería que cualquier desenlace nos dijera algo, en uno u otra dirección, con más o con menos fuerza.

Con mucha frecuencia se sabe que la hipótesis nula es falsa, incluso antes de recoger los datos, lo cual es lógico, pues no existe razón alguna, por ejemplo, para que un coeficiente poblacional sea exactamente igual a cero, ni para que tratamientos como los del ejemplo produzcan exactamente el mismo efecto.

Sin embargo, la objeción más seria que se le hace a este método quizás sea que, dada la naturaleza de los valores *p,* el rechazo o la aceptación de una hipótesis resulta ser entonces, simplemente un reflejo del tamaño de la muestra. Esto nos conduce a una paradoja: si valoramos una parte muy pequeña de la realidad (una muestra muy reducida) no podemos obtener conclusión alguna, como es lógico e intuitivo, y conduce a que muchos investigadores cuyos resultados no alcanzan la esperada significación estadística proclamen que con un tamaño de muestra mayor lo hubieran logrado; pero, y esto es lo grave, tampoco se puede sacar nada en claro cuando se trabaja con una muestra muy grande, puesto que en tal caso el rechazo de la hipótesis nula queda virtualmente asegurado (1).

Para ver esto con más claridad, volvamos al experimento que evalúa si el tratamiento con crema y apósito es más efectivo que el tratamiento con crema solamente. Imaginemos por un momento que se sabe que las verdaderas tasas de recuperación son 75% y 60% (y, por ende, que la diferencia entre ambos es 15%). Es obvio que, cuanto mayor sea el tamaño de muestra, es más probable que la diferencia estimada se acerque a ese valor verdadero. Pero ¿qué pasaría con la prueba de hipótesis de Ji-cuadrado? Si en lugar de 40 pacientes en cada grupo se hubieran tomado 60, entonces los datos del estudio, suponiendo que las estimaciones p_1 y p_2 fueran exactas, hubieran sido los que recoge la Tabla 2.

Tabla 2. Distribución de una muestra de 120 pacientes según tratamiento asignado y según se recuperarán o no

Tratamiento	Recuperación		Total
	Sí	No	
Experimental	45	15	60
Convencional	36	24	60
Total	81	39	120

En tal caso se obtiene *p=0,08,* de manera que podría declararse el hallazgo de significación estadística sin escandalizar a nadie, pues el valor de *p* es menor que 0,1, uno de los valores que se usan como referencia para hacer esta afirmación.

Si se hubieran tomado 200 pacientes en total (Tabla 3), el resultado de la prueba Ji-cuadrado arroja un valor de *p* "significativo" incluso para el sacralizado $\alpha=0,05$.

Tabla 3. Distribución de una muestra de 200 pacientes según tratamiento asignado y según se recuperarán o no

Tratamiento	Recuperación		Total
	Sí	No	
Experimental	75	25	100
Convencional	60	40	100
Total	135	65	200

Mientras que, cuando el tamaño de muestra asciende a 280, se obtiene *p=0,007,* un resultado

significativo incluso ante el más exigente umbral del 1% (Tabla 4).

Tabla 4. Distribución de una muestra de 280 pacientes según tratamiento asignado y según se recuperaran o no

Tratamiento	Recuperación		Total
	Sí	No	
Experimental	105	35	140
Convencional	84	56	140
Total	189	91	280

Este patrón es válido cualquiera que sea la diferencia verdadera: siempre se puede conseguir una *p* tan pequeña como uno quiera con sólo tomar un tamaño de muestra suficientemente grande.

Se ha argumentado que la decisión no debe basarse exclusivamente en la significación estadística, sino que ha de atenderse también a la significación clínica. En este sentido, Feinstein apuntaba (8):

> *"Si la demanda crítica hubiera sido que la investigación produjese ambos tipos de significación (la que concierne al área estocástica y la que se vincula con los atributos cualitativos) entonces la alienación intelectual de hoy no hubiera ocurrido. Desafortunadamente, sin embargo, la palabra "significación" fue reservada sólo en su connotación estocástica, y la palabra estadística le fue adjuntada para crear la "significación estadística" como paradigma de calidad e importancia en la investigación médica…"*

La valoración acerca de si una diferencia (por ejemplo, de 15%) es o no cualitativamente importante ha de hacerse independientemente del valor de *p;* de modo que si *p* es pequeña, pero también lo es la diferencia estimada, entonces no se le concederá trascendencia al tratamiento experimental. Ahora bien, si la diferencia (grande o no) siempre puede llegar a ser significativa, lo único que hace falta es conocer el valor de la diferencia; claro, poder decir que, efectivamente, "se conoce" tal valor equivale a contar con una estimación adecuada, precisa, lo cual también se consigue con un tamaño muestral suficientemente grande y se concreta en la obtención de un intervalo de confianza estrecho, como se discute en la próxima sección.

Este método supuso un cambio sutil en la balanza de la autoridad médica, que pasó de estar en manos de aquellos con conocimiento de las bases biológicas de la medicina, a las de aquellos que dominan los métodos cuantitativos, o hacia los resultados cuantitativos solamente, como si los números pudieran hablar por sí solos.

Los textos sobre estadística inferencial dan por sentado que las pruebas de significación constituyen un procedimiento sin fisuras, ignorando sus limitaciones conceptuales y prácticas

Lo cierto es que las pruebas de significación están lejos de ser una panacea; un recurso con numerosas endebleces, del que como mínimo puede decirse que no es imprescindible en la investigación. Tanto es así, que hay revistas punteras de la producción científica internacional que no aceptan actualmente trabajos en los cuales sólo aparezcan pruebas de este tipo. Por ejemplo, British Heart Journal anuncia en un editorial de 1988 que se une a la demanda que ya desde 1986 había aparecido en British Medical Journal cuando exhortaba a los autores a calcular intervalos de confianza, fuesen acompañados o no de pruebas de significación. Tal postura ya era compartida por revistas tan importantes como Lancet, Annals of Internal Medicine y American Journal of Public Health (9).

Alternativas

Se manejan varias alternativas que pueden suplir la supuesta horfandad en que nos dejaría el abandono de las pruebas de significación. Una de ellas, sin dudas la más sencilla, consiste en poner el énfasis en la construcción de intervalos de confianza. Si bien ellos se inscriben en la órbita de las mismas matemáticas frecuentistas que las pruebas de significación, se apartan de la interpretación automática de los valores *p* y constituyen un recurso para medir la diferencia entre los efectos y el grado en que el conocimiento de la verdadera diferencia es adecuado.

La Tabla 5 muestra lo que sucede cuando se calculan intervalos al 95% de confianza para

las distintas situaciones relacionadas con el ejemplo arriba considerado.

Tabla 5. Intervalos de confianza para la diferencia entre tratamientos calculados con los diferentes tamaños de muestra

Tamaño de muestra	Intervalo de confianza	
	Límite inferior	Límite superior
80	-5,3%	35,3%
120	-1,5%	31,5%
200	2,2%	27,8%
280	4,2%	26%

Obsérvese que, incluso con un tamaño de muestra suficientemente grande (por ejemplo de 200) como para producir una clarísima significación y una diferencia en las tasas de recuperación (15%) que probablemente se considere cualitativamente importante a efectos prácticos, el intervalo de confianza brinda un amplísimo espectro de posibles diferencias compatibles con los datos. Este hecho sale a la luz cuando se utilizan intervalos de confianza. Si sólo se usara la prueba de significación, tendríamos una d_0 muy significativamente diferente de cero y a la vez clínicamente muy relevante, pero no nos enteraríamos de que el grado de incertidumbre que la envuelve es de tal magnitud que, a pesar de todo, seguimos sin poder sacar conclusiones definitivas, ya que 2,2% es un valor perfectamente posible de la diferencia y acaso carezca de sentido práctico aplicar el nuevo recurso terapéutico para conseguir tan magros dividendos. De ahí que muchas revistas insistan en el uso de los intervalos de confianza, y que actualmente sean cada vez más empleados en las investigaciones sanitarias.

Sin embargo, es necesario aclarar que entrañan, aunque de una forma más sutil, algunos de los problemas que afectan a los métodos habituales; el más importante es que no brindan un mecanismo para la integración de la evidencia externa o previa con la proporcionada por el estudio actual.

Finalmente, cabe mencionar otro enfoque: los métodos bayesianos. No nos detendremos aquí a explicarlos, puesto que su exposición exige un nivel teórico que desborda el que consideramos oportuno para estas reflexiones. Sin embargo, procede destacar que se trata de una aproximación metodológica que está exenta de las impugnaciones que se le hacen a las pruebas de significación y que goza del atractivo de incorporar las evidencias aportadas por experiencias previas dentro del proceso analítico y las contempla, por ende, en las conclusiones (10).

Aunque las bases de este enfoque datan de hace más de dos siglos, es ahora cuando empieza a asistirse a un uso apreciable del mismo en la investigación biomédica. Una de las razones que explican tal realidad y que a la vez augura un prominente futuro, es que algunos de los problemas de cierta complejidad que posee este método exigen el uso de recursos computacionales accesibles sólo ahora para el común de los investigadores.

En cualquier caso, lo más importante es que se comprenda que el procedimiento de prueba de hipótesis utilizado en la actualidad no sólo sufre de serias limitaciones, sino que está lejos de ser imprescindible en la investigación sanitaria, dado que existen alternativas mucho más sólidas y cercanas al sentido común.

BIBLIOGRAFÍA

1. Silva LC. La crisis de las pruebas de significación y la alternativa bayesiana. Memorias del XI Congreso de la Sociedad Gallega de Estadística e Investigación Operativa, Santiago de Compostela. 1999.
2. Silva LC. Cultura estadística e investigaciones en el campo de la salud. Madrid: Díaz de Santos; 1997.
3. Goodman SN. Toward evidence-based medical statistics (I): The p value fallacy. Annals of Internal Medicine. 1999;130:995-1004.
4. Feinstein AR. P-values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin. Journal Clinical of Epidemiology. 1998;51(4):355-360.
5. Royal RM. Statistical evidence: a likelihood paradigm. Boca Raton: Chapman & Hall/CRC; 1997.
6. Morrison DE, Henkel RE. The Significance Test Controversy –A Reader. Chicago: Aldine Publishing Company; 1970.
7. Goodman SN. Valores p, pruebas de hipótesis y verosimilitud: las consecuencias para la epidemiología de un debate histórico ignorado. Boletín Oficina Sanitaria Panamericana. 1995;118(2):141-155.
8. Feinstein AR. Clinical Epidemiology: The architecture of clinical research. Philadelphia: W.B. Saunders Company; 1985.
9. Evans SJW, Mills P, Dawson J. The end of the p value? British Heart Journal. 1988;60:177-180.
10. Silva LC, Suárez P. ¿Qué es la inferencia bayesiana? JANO, Medicina y Humanidades. 2000;58(1338):65-66.