ACTUALIZACION

Utilidad y limitaciones de las pruebas de hipótesis en la epidemiología nutricional ¿Cómo proceder frente a un problema?

PERSPECTIVAS EN NUTRICIÓN HUMANA
ISSN 0124-4108 Número 9 junio de 2003
Universidad de Antioquia. Medellín. Colombia pags. 72-87

Pedro A. Monterrey Gutiérrez Doctor en Ciencias Matemáticas (Ph. D).
Instituto de Nutrición e
Higiene de los Alimentos. Cuba.
E-mail: pmonterrey@guay.com

Lilia Yadyra Cortés Sanabria Master en Nutrición Clínica y normal.
Departamento de Nutrición Bioquímica.
Pontificia Universidad Javeriana. Bogotá.
E-mail: ycortes@javeriana.edu.co

María Elena Díaz Hernández. Doctora en Ciencias Biológicas,
Ms C. en Antropometría .
Instituto de Nutrición e
Higiene de los Alimentos. Cuba
E-mail: meds49@yahoo.es

Resumen

Las pruebas de hipótesis constituyen un elemento central en los análisis de datos en la investigación biomédica. Su uso ha conducido, en ocasiones, a inconsistencias y a veces a errores en las conclusiones que se han derivado de su aplicación lo que ha conducido a que su uso sea desestimulado por algunas publicaciones. En el presente trabajo se analizan las causas de esas inconsistencias, recorriéndose en el análisis un espectro de argumentos que van desde la carencia de un sólido sustrato teórico para los valores P, errores en la interpretación de sus resultados, pasando por serios problemas

PALABRAS CLAVE:
Pruebas de hipótesis, valores p, *intervalos de confianza, sesgos,* significación estadística, significación clínica.

en el diseño de los estudios, lo que incluye, como elemento muy importante, un deficiente control de los sesgos, resaltándose cómo se presentan estos problemas en el análisis de datos de la epidemiología nutricional. El uso adicional de los intervalos de confianza, con el análisis de la significación clínica de los hallazgos, se presentan como elementos metodológicos centrales, junto a un buen diseño, en la aplicación exitosa de las pruebas de hipótesis. A partir de un ejemplo se aplica la metodología propuesta.

Usefulness and limitations of the hypothesis tests in the nutritional epidemiology. How to proceed when facing a problem?

Abstract

Hypothesis tests constitute a central element in the analyses of data in biomedical research. Its use has driven, in occasions, to inconsistencies and sometimes to errors in the conclusions that have been derived of its application, which has led some publications to discourage its implementation . In the present work the causes of those inconsistencies are analyzed, traveling through a spectrum in the analysis of arguments that go from the lack of a solid theoretical substance for the P values to errors in the interpretation of their results, as well as serious problems in the design of the studies, which includes, as a very important element, a faulty control of manipulated facts; highlighting how these problems are present in the analysis of nutritional epidemiology data. The additional use of the intervals of trust, paired with the analysis of the clinical significance of the discoveries, are presented as central methodological elements, along with a good design, in the successful application of the hypothesis tests.

An example is presented in which the proposed methodology is applied.

KEY WORD:
Tests of hypothesis, p values, *intervals of trust*, manipulated facts, statistical significance, clinical significance.

INTRODUCCIÓN

Las pruebas, o como las denominan algunos autores en los textos de estadística las dócimas de hipótesis estadísticas, constituyen el elemento de análisis de datos más profusamente utilizado por los

investigadores no sólo en el área de la salud sino también en otras áreas. Una simple observación de las revistas biomédicas basta para afirmar que, a pesar de la gran cantidad de métodos estadísticos para el análisis de datos, una pequeña cantidad de pruebas de hipótesis estadísticas se repiten sistemáticamente. Este uso extendido ha propiciado una pérdida de rigor en su uso y, hablando de manera general, se podría llamar un abuso de las mismas. Ese uso generalizado, extensivo, fundamentalmente por parte de profesionales no especializados en el tema y por tanto desconocedores de los fundamentos mismos ha conducido, en momentos, a la obtención de conclusiones absurdas, ilógicas e inconsistentes y han puesto en tela de juicio la pertinencia e incluso, la validez de la técnica.

Existe una tendencia, en algunos círculos de epidemiólogos, a relegar las pruebas de hipótesis a un segundo plano

Existe una tendencia, en algunos círculos de epidemiólogos, a relegar las pruebas de hipótesis a un segundo plano en el que "sólo son útiles para investigar asuntos tangenciales y características poco importantes de los datos; por lo demás están fuera de lugar"(1). Esta tendencia, muy bien sintetizada en el trabajo publicado por Walker en 1993(2), ha sido reflejada, en diversos grados, por diferentes publicaciones del área de la biomedicina en un espectro de regulaciones que van desde la no aceptación de la aplicación de esta técnica, como en el British Heart Journal y en el British Medical Journal, hasta el desestimular su uso como es el caso de Lancet y Annals of Internal Medicine. La crisis en que han caído las pruebas de hipótesis se refleja incluso en las normas que publica el Comité Internacional de Directores de Revistas Médicas (Grupo de Vancouver) en las que se consigna "se evitará la dependencia exclusiva de las pruebas estadísticas de verificación de hipótesis, tal como el uso de los valores de P, que no aportan ninguna información cuantitativa importante"(3).

Con el presente trabajo se pretende analizar los fundamentos de las pruebas de hipótesis haciendo énfasis en el origen, significado y limitaciones de los valores P, caracterizar los diferentes errores que se cometen al utilizarlos y sus fuentes, asimismo, dar consejos prácticos a los investigadores, fundamentalmente en el área de la epidemiología nutricional, sobre cómo analizar los datos para evaluar la validez de sus hipótesis para lo cual se propone y fundamenta el uso adicional de otra técnica de la estadística elemental o básica: los intervalos de confianza, como parte de una estrategia de análisis de datos en la que las conclusiones surgen de una interacción con los mismos datos y no de una regla de decisión impersonal en la que el problema se reduce a un acepto o rechazo.

LAS PRUEBAS DE HIPÓTESIS. SUS FUNDAMENTOS E HISTORIA

En los primeros años del siglo pasado R A Fisher, denominado el padre de la estadística, abordó el problema de la realización de inferencias a partir de los datos. Para la solución de este problema él construyó lo que llamó **valores P** para evaluar el grado de discrepancias de los datos respecto a **una hipótesis** que él llamó nula: su valor de P, al igual que los que se usan actualmente, eran la probabilidad del valor observado más la de otros más extremos bajo la hipótesis nula. Su significado, sin embargo, era completamente diferente al de aquellos que se usan en la actualidad(4). Para Fisher los valores P eran "una medida del carácter probatorio del experimento, que debía emplearse para reflexionar sobre la credibilidad de la hipótesis nula en función de los datos. **Medida que debía combinarse con otras fuentes de información sobre el fenómeno en estudio**"(4).

Para Fisher los valores P eran "una medida del carácter probatorio del experimento, que debía emplearse para reflexionar sobre la credibilidad de la hipótesis nula en función de los datos

El fundamento teórico de Fisher fue fuertemente criticado en los medios especializados, múltiples fueron los argumentos utilizados; **la ausencia de una hipótesis "alternativa" en el proceso de análisis,** fue uno de ellos. En ese contexto en 1928 Neyman y Pearson publicaron un procedimiento para realizar inferencias a partir de los datos que llamaron "pruebas de hipótesis"(5).

Las pruebas de hipótesis, en el sentido de Neyman y Pearson, son reglas de decisión diseñadas para eliminar de los datos el efecto del azar y poder evaluar la validez de las hipótesis de interés sobre la base de las evidencias que arrojan las observaciones. En la teoría de Neyman y Pearson se consideran **dos posibles hipótesis** que representan las alternativas de decisión en un problema: la hipótesis nula (H_0) y la hipótesis alternativa (H_A); asociados a ellas se analizan dos posibles errores: rechazar H_0 siendo cierta H_A (error de primer tipo o de tipo I) y rechazar H_A siendo cierta H_0 (error de segundo tipo o de tipo II). Ambos errores se caracterizan por las correspondientes probabilidades de ser cometidos.

Múltiples situaciones experimentales conducen a la necesidad de evaluar un par de hipótesis, pero en todos los casos esas hipótesis representan los intereses o expectativas del investigador en el contexto del problema que aborda; así por ejemplo, si en una encuesta para evaluar el estado nutricional de una comunidad se obtuvo, hace unos años, que el valor promedio del índice de masa corporal (IMC) de los adultos era de 21.5 Kg/m^2 y hace poco se hizo otro estudio en el que el valor medio observado fue de 18.5 Kg/m^2, los resultados de la nueva encuesta evidencian la posibilidad del deterioro del estado nutricional de esa población. Pero ¿es ésto realmente cierto? en otras palabras ¿es **significativa** la diferencia observada en las estimaciones hechas en los dos momentos? El problema radica entonces, si se sigue enfoque de Neyman y

Contra los sesgos poco se puede hacer en la etapa de análisis de datos

Pearson, en decidir, sobre la base de los datos, cuál de las **dos hipótesis** posibles en el problema, H_0 (no cambio en el estado nutricional de la comunidad; la diferencia observada no es significativa) y H_A (cambio en el estado nutricional de la comunidad; la diferencia observada es significativa) debe ser aceptada. Si fuera a emplearse la formulación de Fisher el problema sería evaluar el grado en que los datos discrepan de **una hipótesis**: H_0.

Un elemento que hay que tener bien presente para entender el alcance y las posibilidades de las pruebas de hipótesis es que, la diferencia numérica entre los valores del indicador del estado nutricional en los dos momentos de recogida de la información en la comunidad puede deberse a tres factores: el efecto de los sesgos, el azar o la presencia real de cambios en la población. El problema es dilucidar si **realmente** hay cambios en la población o si la diferencia es debida a otros factores y por tanto no es digna de ser tenida en consideración.

Contra los sesgos poco se puede hacer en la etapa de análisis de datos, que es el momento en que entran a jugar un papel importante las pruebas de hipótesis. Los sesgos pueden ocurrir en la selección de los individuos, en la recogida de la información y su análisis. Ellos pueden ser introducidos, en etapas tan iniciales como es en el diseño del estudio, mediante el uso de modelos o concepciones de muestreo inadecuadas. Los sesgos y la importancia de su control es un aspecto que algunos investigadores, que utilizan profusamente las pruebas de hipótesis, no tienen en consideración en su justa medida; su prevención y control se debe convertir en uno de los elementos fundamentales en todas las etapas de una investigación, es decir en el diseño, la ejecución y el análisis de los datos de una investigación.

Al aplicar las pruebas de hipótesis los datos deben estar lo más libres de sesgos que sea posible; éste es un punto de partida necesario para lograr su correcta aplicación.

La función de las pruebas es despojar los datos del efecto del azar y ver si entonces la diferencia es aún digna de ser tomada en consideración, por eso se dice que la diferencia es significativa. Con esa terminología se resalta el que hay una diferencia numérica que no es debida al azar y que por tanto debe ser considerada. Si los datos están muy sesgados, el efecto de los sesgos se mantiene una vez removido el azar y ellos actúan como un factor de distorsión en las inferencias que se realicen con las pruebas de hipótesis, por lo que pueden desvirtuar las conclusiones a que se llegue. Este es un punto que debe estar muy claro en las mentes de todos aquellos que aspiren a aplicar las pruebas.

La teoría de Neyman y Pearson se basa en la determinación de un estadígrafo (estadístico) con el que se comparan los valores cuya

significación se desea establecer; con él se construye una denominada **región crítica o región de rechazo** que permite determinar cuándo las diferencias observadas son lo suficientemente grandes como para que sean debidas, además de al azar, a patrones reales de cambio o de diferencia, en tal caso se rechaza H_0 (la no diferencia) y se dice entonces, siguiendo el lenguaje del ejemplo que se analiza, que las diferencias observadas en el comportamiento del IMC en los dos momentos en que se evaluó la comunidad son significativas, lo que se traduce en aceptar el cambio en el estado nutricional de la población.

La región crítica, o región de rechazo de H_0, se construye buscando un valor que marque un umbral a partir del cual se pueda decir que los valores del estadígrafo de la prueba son apreciablemente grandes, lo suficiente como para decir que las discrepancias observadas no pueden ser debidas sólo al azar sino, presumiblemente, a la existencia de un patrón de cambio o de diferencia. Para la construcción de ese valor de umbral se tiene en cuenta la probabilidad de cometer un error de tipo I, que se identifica con la letra griega alfa (α). En el ejemplo que se viene desarrollando, en el que se comparan las medias de dos poblaciones, la región crítica sería $E > z$; denotando mediante E el estadígrafo que permitiría comparar los valores medios del IMC en los dos momentos y mediante z el punto de umbral que identifica a partir de qué valores esas diferencias son lo suficientemente grandes como para poder decir que hay un cambio. El error de tipo I queda controlado al determinar z de forma tal que la probabilidad de que $E > z$, sea igual a α. La probabilidad mencionada anteriormente se calcula suponiendo H_0 cierta, es decir, como una probabilidad condicional. La expresión completa del estadígrafo E puede ser consultada en cualquier libro de estadística básica.

En la construcción de la prueba de hipótesis en el sentido de Neyman y Pearson, que no es más que una regla de decisión para determinar si se acepta una u otra de las hipótesis, no se utiliza el error de Tipo II. Este error es complicado de manejar, mucho más complicado que el tratamiento del error de tipo I. Para su manejo se utiliza la denominada función de potencia de la prueba. La importancia del error de tipo II es que él se controla, en la etapa de diseño, mediante la determinación de un tamaño de muestra que permita que esté en un rango predeterminado a priori y aceptable en dependencia de los riesgos que se cometan con una mala decisión. Éste es otro elemento que escapa de las manos de muchos de los que aplican las pruebas de hipótesis y que, en muchos casos, no tienen la menor idea de cuál es el error de tipo II que están cometiendo en sus decisiones. Esto introduce un manto de dudas sobre la calidad de las inferencias que realizan cada vez que, como resultado de la aplicación de la prueba, se decide aceptar H_0.

LOS VALORES P, SUS FUNDAMENTOS Y USO.

En la década de los años 40 del siglo pasado se unieron, de forma anónima y pragmática, los valores P de Fisher y las reglas de decisión de Neyman-Pearson(3). De esta forma, huérfanos de un sustrato teórico sólido, se dio lugar a los valores P como se usan en la actualidad para realizar las inferencias estadísticas en las publicaciones científicas.

Goodman analiza diferentes definiciones en uso del valor de P, una de ellas establece: "el valor P es la verosimilitud de que un estudio sea positivo (*rechazar* H_0) cuando la hipótesis nula es verdadera: es análogo a la tasa de resultados positivos falsos...de una prueba diagnóstica"(4). Las definiciones señaladas por él identifican el valor P como una tasa de error, pero esto no es exactamente así. Tal vez la expresión de Fisher de que son "una medida racional y bien definida de la renuencia a aceptar la hipótesis sometida a prueba"(4) dé un mejor marco de comprensión aunque no reproduce exactamente el sustrato conceptual que envuelve los valores P actuales. Lo cierto es que "es erróneo considerar que los valores P puedan ser a la vez, una tasa de error de falsos positivos, en el sentido de Neyman y Pearson y una medida de la evidencia contra la hipótesis nula en el sentido de Fisher"(3).

Huérfanos de un sustrato teórico sólido, se dio lugar a los valores P como se usan en la actualidad para realizar las inferencias estadísticas en las publicaciones científicas

¿En qué se basan los valores P? Para comprenderlos es necesario, primero, una pequeña digresión práctica. Si se supone que se tiene una bolsa con 10 000 bolas en la que se sospecha, pero no se está seguro, hay 5 bolas blancas y 9995 bolas rojas, la probabilidad de una bola blanca es 5/10 000 = 0.0005 mientras que la probabilidad de una bola roja es 0.9995. Si se extrae al azar una bola de la bolsa y ésta es de color blanco, se puede pensar en dos posibilidades: o sucedió algo muy poco probable (lo cual, valga la redundancia, es poco probable) o la idea que se tenía sobre la composición de la bolsa es falsa. A todas luces, en el caso de tener que tomar una decisión, la segunda es menos ingenua y por tanto marcaría un proceder lógico. Este argumento es el que se replica, a la hora de fundamentar las decisiones a tomar en el problema, utilizando el valor P.

Si se designa mediante E al estadígrafo de la prueba de hipótesis y e_0 representa su valor observado en la muestra. El valor P no es más que la probabilidad de que E tome el valor e_0 ó valores más extremos en la dirección que establece el rechazo de H_0 (según la región crítica). Esta probabilidad se calcula, como una probabilidad condicional, con la premisa de que H_0 sea cierta. En el ejemplo que se viene desarrollando en el que se comparan las medias de dos poblaciones sería P = Probabilidad $(E > e_0 / H_0)$.

Para la interpretación de P se utiliza un razonamiento por reducción al absurdo semejante al que se

explicó en el ejemplo de la bolsa, si P es pequeño eso significa que ha ocurrido, como resultado del muestreo realizado, un resultado poco probable; ésto es contradictorio, tal y como ocurrió con la bola blanca en el ejemplo, y por eso el fundamento sobre el cual se calculó P, es decir H_0 cierta, no es válido. En ese caso se decide rechazar H_0. Por el contrario si P es grande entonces ocurrió algo posible, como la bola roja, y entonces no hay razón para rechazar H_0. Un valor P pequeño indica que el azar es una explicación poco probable de las discrepancias encontradas y por eso se rechaza H_0.

El problema con los valores P es que en su formulación no se tiene en cuenta qué significa rechazar H_0

El problema con los valores P es que en su formulación no se tiene en cuenta qué significa rechazar H_0, no se hace mención a ninguna alternativa; no queda claro en él, incluso, cómo manejar el efecto que pudiera tener sobre su proceso de decisión el disponer de diferentes tamaños de muestra, lo cual es crucial pues los niveles de información de la muestra son mayores en la medida en que el tamaño de muestra lo es y viceversa, pero ellas son muy inestables si los tamaños de muestra son pequeños porque el azar tiene un peso muy importante. Todo lo anterior, entre otros elementos, introduce un componente teórico dudoso en su formulación.

Al analizar los valores P se acostumbra compararlos con 0.05, valor que se considera usualmente para el error de tipo I, para decidir si P es pequeño o grande y en correspondencia aceptar o rechazar H_0. Pueden emplearse otros valores (0.01 o 0.1 por ejemplo) en correspondencia con el nivel de error que se esté dispuesto a trabajar, pero lamentablemente este tipo de consideración casi nunca se hace y los artículos deciden con el mismo valor (0.05). En la literatura biomédica como no existe un estándar para la presentación del resultado de esa comparación, se utilizan múltiples formas: escribiendo el propio valor P, tal y como fue calculado, o utilizando siglas para representar el resultado de la comparación (S, NS, *, **, ***, entre otras). Las formas de presentación, que excluyen la presentación del valor exacto de P, son bastante inespecíficas y no arrojan claridad pues se priva al lector, que no tiene acceso a los datos, de la posibilidad de saber cuán diferente es el valor P del valor de comparación respecto al que se dice si es o no pequeño. Esta información, como se verá posteriormente, es crucial para establecer la fortaleza o no del efecto encontrado.

¿CÓMO USAR LAS PRUEBAS ESTADÍSTICAS Y LOS VALORES P? ¿CÓMO PRESENTAR LOS RESULTADOS?

El problema no es desechar las pruebas estadísticas sino darle un sentido más teórico y racional a su uso en el análisis de datos.

La Tabla 1 presenta una situación hipotética que tal vez pudiera desconcertar a quienes usan las pruebas de hipótesis. En ella se presenta

el resultado de aplicar la prueba de comparación de medias de dos poblaciones para comparar las medias de los pesos al nacer de dos grupos de niños, aquellos cuyas madres tuvieron una ganancia de peso adecuada con aquellos cuyas madres tuvieron una ganancia de peso deficiente. En la tabla se analiza qué pasa con la inferencia al analizar el comportamiento de las mismas estimaciones de los valores medios en el caso de tamaños de muestra muy diferentes. Para los cálculos se supuso, de manera hipotética, que para las madres con ganancia de peso adecuada el peso promedio de los niños fue de 3249.46 g, mientras que entre aquellas que tuvieron una ganancia deficiente fue de 3028.85g. Por comodidad se supone que las dos muestras tienen una misma desviación estándar igual a 619.55 g. ¿Cómo es posible que los mismos valores medios lleven a valores P muy diferentes, que conducen a conclusiones muy distintas?

TABLA 1

Efecto del tamaño de muestra sobre los valores P

Tamaño de muestra	Valor P	Intervalo de confianza			
		Población 1		Población 2	
		Límite inferior	Límite superior	Límite inferior	Límite superior
665	0.0000	2982.00	3075.70	3202.61	3296.31
300	0.0000	2959.10	3098.60	3179.71	3319.21
150	0.0022	2930.21	3127.49	3160.82	3349.10
75	0.0307	2889.35	3168.35	3109.96	3388.96
35	0.1401	2824.64	3233.06	3045.25	3453.67
10	0.4344	2646.81	3410.89	2867.42	3631.50

La primera cuestión a tener en cuenta al usar los valores P es que en la presentación de los resultados es necesario indicar su valor exacto. No significa lo mismo decir, por ejemplo, $P < 0.05$ en el caso del tamaño de muestra $n = 665$ (tabla 1) que en el caso $n = 75$. Aunque la conclusión puede ser la misma en los dos casos, la cautela con que ésta debe ser tomada es diferente pues la evidencia no es igualmente fuerte en ambos casos.

Cuando el tamaño de muestra es grande las medias estiman mejor el valor poblacional ya que los errores de muestreo son menores

Respecto a la pregunta central, referida a llegar a conclusiones distintas con los mismos valores medios, la observación de los intervalos de confianza que aparecen en la misma tabla 1 puede dar la clave de la respuesta. El diámetro (amplitud) del intervalo para estimar la media de cada población es inversamente proporcional al tamaño de muestra; ésto no es casual sino que es la consecuencia de que los errores de estimación (errores de muestreo, error estándar de estimación) son mayores mientras menor sea el tamaño de muestra. En consecuencia, los mismos valores medios presentados para cada grupo, que son la base de la prueba estadística de comparación de medias de dos poblaciones, tienen un significado diferente para cada tipo de tamaño de muestra.

Cuando el tamaño de muestra es grande las medias estiman mejor el valor poblacional ya que los errores de muestreo son menores; en otras palabras el azar está menos presente en la estimación y en consecuencia, el intervalo de confianza es más pequeño; eso significa que los valores calculados a partir de la muestra, de manera muy estable, aproximan el valor poblacional con mucha calidad o exactitud. Para tamaños de muestra pequeños sucede todo lo contrario. La conclusión es inmediata, lo que hacen los valores P no es más que reflejar el comportamiento de los estimadores a partir de los que se calcula y por eso las conclusiones a que se llega en la tabla 1 son diferentes según el tamaño de muestra a que se refiera.

Un hecho que tiene que estar muy claro, para todo aquel que utiliza las pruebas de hipótesis, es que ellas son muy propensas a aceptar como significativas diferencias pequeñas en la medida que el tamaño de muestra es mayor. Este comportamiento es consecuencia de la pérdida del efecto del azar sobre las estimaciones, pues cuando las muestras son grandes es poco el azar que hay que remover en las comparaciones y por tanto el problema, más que un problema de prueba de hipótesis, es un problema de estimación. En el caso de tamaños muestrales pequeños sucede todo lo contrario.

Una conclusión importante es la necesidad de introducir intervalos de confianza como elemento adicional en el análisis de las pruebas de hipótesis(6). De esta forma se evalúan los datos desde varios ángulos al incorporar en el análisis las variaciones producidas en las estimaciones por los errores de muestreo. Los intervalos de confianza tienen el atributo de que en ellos los límites de la incertidumbre están claramente establecidos, por eso son una innegable ayuda a la hora de interpretar lo que dicen los datos respecto a la validez de las hipótesis, "ellos indican cuándo los datos son tan limitados que no son consistentes con H_0 y cuándo marcan desviaciones de H_0 que no son de importancia científica"(7). Los intervalos de confianza complementan la visión del problema pero no dan una visión completa del mismo. Su utilidad, en el caso de las inferencias, radica en

la posibilidad de ser utilizados conjuntamente con las prueba de hipótesis. La idea es que ambas técnicas se complementen, pues muestran desde dos ángulos diferentes cuánto H_0 es contradicha por los datos.

Un consejo importante a la hora de analizar los datos es entonces combinar las pruebas de hipótesis estadísticas y con los correspondientes intervalos de confianza. La decisión se toma conjugando las dos técnicas, pero considerando, como elemento teórico concomitante, que para muestras grandes el problema es principalmente un problema de estimación, pues el azar tiene poco peso, y por ello la evidencia más fuerte la da el intervalo de confianza. Para muestras pequeñas la situación es inversa.

El uso del intervalo de confianza obliga al investigador a no perder contacto con lo que dicen sus datos y a no tomar decisiones observando solamente un valor, P

El uso del intervalo de confianza obliga al investigador a no perder contacto con lo que dicen sus datos y a no tomar decisiones observando solamente un valor, P, del que en ocasiones sólo conoce que se compara con un valor de referencia y se decide en correspondencia con esta comparación.

Volviendo a la tabla 1 se puede mostrar este proceder. Para los tamaños de muestra mayores de 150 el valor P es pequeño y evidencia la necesidad de rechazar la hipótesis nula; los intervalos de confianza con los que se estima la media de cada población son disjuntos y ésto corrobora la diferencia de las medias, es decir, confirma el rechazo. En el caso $n = 75$, $P = 0.0377 < 0.05$ indica el rechazo de H_0, pero la evidencia no es tan fuerte como en los casos anteriores pues de haberse comparado con 0.01 la decisión hubiera sido otra; claro el nivel de exigencia en el segundo caso es mayor y por tanto se es más cauteloso en el rechazo de H_0 si se analizan los intervalos de confianza; buscando un argumento adicional, se observa que los valores comunes para las estimaciones de las medias del peso de los niños al nacer en ambos grupos cubren el intervalo 3109.96g a 3168.35g , el diámetro de este intervalo es 58g, que es bastante pequeño, tal vez irrelevante desde el punto de vista clínico, por eso es lógico aceptar como consistente la decisión de rechazar H_0 pues los datos no arrojan muchas evidencias a favor de la semejanza de medias. Los dos casos restantes se razonan de igual forma.

Una vez aclarado el efecto del tamaño de muestra sobre las estimaciones y las decisiones, tal y como fue discutido en la tabla 1, es válido comentar que en esa tabla fueron utilizados intervalos de confianza para la media de cada población con la intención de mostrar cómo se comportaba la variabilidad de las estimaciones frente a cambios en el tamaño de muestra y su efecto sobre el valor P. Sin embargo, existe una forma alternativa, y "más tradicional", de abordar la solución del problema, ésta consiste en utilizar intervalos de confianza para la diferencia de medias de dos poblaciones, pero en este caso los efectos sobre las estimaciones de la media

Una vez determinado el papel de los posibles sesgos sobre los resultados de la aplicación de las pruebas estadísticas queda evaluar la lógica de los mismos en el contexto de su relevancia o lógica biológica

no se ven tan claramente como con la forma analizada, aunque el comportamiento de las estimaciones y los valores P es el mismo.

¿CÓMO TOMAR LAS DECISIONES?

La aplicación de las pruebas de hipótesis conduce a establecer la significación estadística de las diferencias o asociaciones en estudio una vez eliminado el efecto del azar en los datos. Evidentemente los diferentes sesgos cometidos en las etapas de la investigación influyen de forma negativa en esas decisiones.

Al estudiar el efecto de la dieta sobre diferentes eventos de salud, esta evaluación se ve fuertemente influida por los sesgos que pueden hacer indetectables ciertas asociaciones. El uso de instrumentos con baja confiabilidad o validez puede ser una de las causas de sesgos en la evaluación del riesgo debido a la dieta. Estos sesgos pueden operar en la dirección de que asociaciones fuertes aparecen como débiles y asociaciones débiles pueden resultar indetectables. Es poco probable pensar que los sesgos debidos al uso de instrumentos inadecuados, para la recogida de la información dietaria, operen en la dirección de hacer aparecer un riesgo en el caso en que no exista(8).

Una vez determinado el papel de los posibles sesgos sobre los resultados de la aplicación de las pruebas estadísticas queda evaluar la lógica de los mismos en el contexto de su relevancia o lógica biológica. En el análisis e interpretación de los resultados del procesamiento estadístico es obligatorio ser muy cautelosos a la hora de traducir la significación estadística como significación clínica o su ausencia como ausencia de efecto.

Existe una marcada diferencia entre significación estadística y significación clínica. La significación estadística es el resultado del análisis de las evidencias que arrojan los resultados del estudio, despojados del azar pero bajo los efectos de los sesgos en el caso en que éstos no hayan sido controlados.

Las pruebas estadísticas facilitan la interpretación de los datos, pero no arrojan claridad sobre los mecanismos que se desencadenan en los procesos estudiados; éstos deben buscarse o fundamentarse en los procesos biológicos, fisiológicos, bioquímicos etc. que se desencadenan en cada caso. Para que una evidencia de significación estadística sea considerada como relevante o no, debe articular en la lógica de los procesos que se estudian. La significación clínica parte de la significación estadística, pero conjuga la plausibilidad biológica y la consistencia de la evidencia dentro y entre estudios. Sólo a partir de consideraciones de este tipo es que es posible asumir como válidos, desde el punto de vista clínico, la significación de los efectos o las diferencias obtenidas a partir de la aplicación de las pruebas de hipótesis a los datos. Esta forma de proceder se corresponde con la

lógica de las disciplinas y está en concordancia con los nueve criterios de causalidad propuestos por Hill (9) para establecer factores de riesgo para una entidad.

La ausencia de significación estadística no puede ser considerada directamente como no existencia de significación clínica. En los estudios epidemiológicos y en especial en la epidemiología nutricional son múltiples las causas por las que una significación estadística que existe realmente puede no ser observada. Entre otras causas se pueden citar: poca variación en la dieta de la población, falta de precisión del método dietético, poca potencia estadística debida al pequeño tamaño de muestra, efecto de factores de confusión y de los sesgos(10). Esto significa que lo funesto en los estudios del efecto de la dieta sobre un evento de salud es la acción de los errores de Tipo II, es decir aceptar el no efecto de la dieta sobre la entidad (H_0) existiendo. Esta evidencia tiene necesariamente que marcar una exigencia a la etapa de diseño del estudio y marca la relevancia de la necesidad de ajustar el tamaño de muestra, en correspondencia, con esta posibilidad de error.

La ausencia de significación estadística no puede ser considerada directamente como no existencia de significación clínica

"Si los resultados de diferentes estudios sobre el efecto de un factor nutricional sobre la entidad consistentemente detectan un riesgo positivo con un valor numérico pequeño, tiene sentido sospechar que existe un riesgo verdadero que no se detecta con fuerza por la presencia de errores en la evaluación de los datos de la dieta y los diferentes sesgos que pueden presentarse en un estudio epidemiológico. Cuando la mayoría de los hallazgos evidencian que la exposición a la dieta aumenta el riesgo del evento de salud bajo estudio, los resultados que den un débil aporte a esta hipótesis deben ser mirados con reserva, pues ellos no aportan elementos que debilitan la fuerza de la hipótesis de existencia de un efecto sino que, presumiblemente, sus resultados deben estar influidos por el efecto negativo de los sesgos. En el caso en que en algunos estudios se detecten riesgos que con consistencia indican la presencia de una asociación negativa (o de tipo protector) de la dieta sobre la entidad estudiada, es poco probable que estos resultados sean debidos a los sesgos, en tal caso ellos deben ser interpretados como fuertes evidencias contra la hipótesis del efecto positivo de la dieta sobre la entidad"(8).

UN EJEMPLO DE APLICACIÓN DE LA METODOLOGÍA PROPUESTA

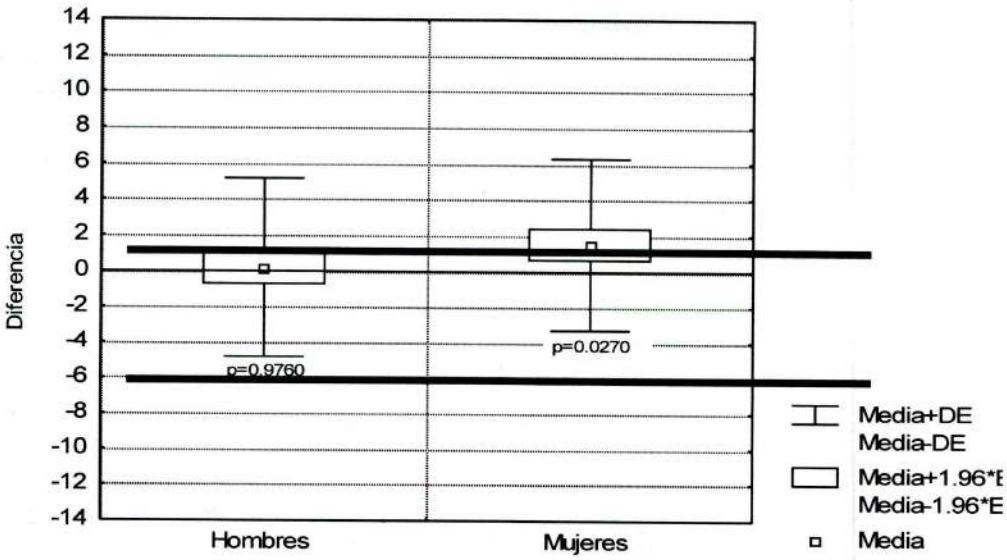
La Figura 1 muestra los resultados de la validación, para cada sexo, de la estimación del porcentaje de grasa corporal mediante mediciones antropométricas a partir de las ecuaciones de Durnin y Womersly y las ecuaciones de Siri. En la validación se utilizó como estándar de oro el método del deuterio(11).

Para decidir acerca de la concordancia entre mediciones se aplicó la prueba de Wilcoxon para datos

pareados. Como consecuencia de comparar los valores P con 0.05 se obtuvo que las discordancias entre el porcentaje de grasa corporal estimado por ambas pruebas no son significativas entre los hombres y sí para las mujeres. La Figura 1 presenta los diagramas de tipo boxplot para describir, utilizando intervalos de confianza del 95%, el comportamiento de las diferencias entre ambos puntajes. Para los hombres el intervalo está centrado en 0 y para las mujeres no lo contiene, aunque no está muy distante. Este comportamiento confirma lo observado en la prueba de Wilcoxon y reafirma, mediante una cuantificación, que la discrepancia entre antropometría y deuterio, en el caso de las mujeres, existe pero no es muy grande.

FIGURA 1

Validación del método antropométrico para estimar el porcentaje de grasa corporal utilizando el método del deuterio como estándar de oro



En este punto el análisis estadístico se puede dar por concluído y el problema cerrado, pero ¿cuál es el sentido biológico de la conclusión a que se ha llegado? ¿qué significa que en las mujeres las pruebas no coinciden pero sus diferencias, aunque significativas, no son muy grandes y, como máximo, no rebasan un 3% del porcentaje real de su grasa corporal?

En el caso de replicaciones de una misma medición los puntajes, de manera general, no se reproducen exactamente. Existe un rango en

el que las discrepancias son tolerables. Cuando se estima la grasa corporal por el método del deuterio una discrepancia del 3% es tolerable y está en el rango de error de la técnica (12). Si se aplica este criterio al análisis del significado de la discrepancia observada entre antropometría y el método del deuterio en las mujeres la interpretación de los resultados tiene una nueva dimensión.

La franja dibujada en la Figura 1 muestra el rango en el que las diferencias entre antropometría y deuterio se encuentran en el rango de error tolerable para el deuterio. Los intervalos de confianza para las medias de las diferencias entre los dos métodos en ambos sexos se encuentran en ese rango. Eso significa que las diferencias observadas en el caso de las mujeres, a pesar de ser estadísticamente significativas, no son **biológicamente relevantes** para los fines de estimar el porcentaje de grasa corporal.

Como consecuencia del análisis anterior se concluye que en los hombres no existen diferencias significativas entre el porcentaje de grasa corporal estimado por el deuterio y el estimado por la antropometría. Para las mujeres sí hay diferencias significativas, pero estas discrepancias entre el valor obtenido por cada una de las dos técnicas están en un rango de error aceptable. En consecuencia el estudio puede concluir que, con sus datos, es aconsejable utilizar la técnica antropométrica, más simple y menos costosa que el método del deuterio, para estimar el porcentaje de grasa corporal en los estudios poblacionales.

Referencias

1. Fleiss J. Las pruebas de significación tienen una función en la investigación epidemiológica: Respuesta a AM Walter. Bol Of Sanit Panam 1993; 115(2):155-58.

2. Walter A. Como presentar los resultados en los estudios epidemiológicos. Bol Of Sanit Panam 1993; 115(2): 148-154.

3. Benavides A. Insuficiencias del paradigma frecuentista y el enfoque bayesiano como alternativa. J Finlay; 2002

4. Goodman SN, Valores P. Pruebas de hipótesis y verosimilitud: las consecuencias para la epidemiología de un debate histórico ignorado. Bol Of Sanit Panam 1995; 118(2): 141-155.

5. Neyman J, Pearson E. On the use and interpretation of certain test criteria for pourposes of statistical inference. Biometrika 1928; 20(1):175-240.

6. Gardner MJ, Altman DG. Intervalos de confianza y no valores P: estimación en ves de pruebas de hipótesis. Bol Of Sanit Panam 1993; 114(6): 536-549.

7. Schlesselman JJ. Case control studies: design, conduct, analysis. New York: University Press; 1981.

8. Marshall JR. The reliability and validity of dietary data as used in epidemiology. Can Surveys 1987; 6(4): 673-683.

9. Hill AB. El medio ambiente y la enfermedad: ¿asociación o causalidad? Bol Of Sanit Panam 1992; 113(3): 233-242.

10. Willet W. Nutricional Epidemiology, 2nd. ed. Oxford: University Press; 1998.

11. Díaz ME, Monterrey P, Hernández M, Sánchez V, Wong I, Moreno V, Toledo E, Matos D. Análisis de la concordancia entre métodos de la composición corporal. En: VIII Simposio de Antropología Física "Luis Montené". Cuba: Universidad de la Habana; 2003.

12. Lukaski HC. Methods for assessment of human body composition: traditional and new. AJCN 1987; 46:537-56.

FECHA DE INGRESO: 23 de Enero del 2003

FECHA DE ACEPTACIÓN : 30 de Mayo del 2003