Metodología de investigación y lectura crítica de estudios

Más allá del valor *p*

Pedro Agustín Monterrey Gutiérrez[1]
Carlos Gómez-Restrepo[2]

Resumen

Introducción: En medicina se ha privilegiado el valor *p* y lo que éste aporta. No obstante, cada día se usan otros criterios, como el intervalo de confianza, y nuevas formulaciones de las pruebas de hipótesis que pueden proveer más profundidad en la identificación de resultados clínicamente relevantes. *Objetivos:* Exponer criterios y pruebas de hipótesis que vayan más allá del valor *p*. *Resultados:* Se da una explicación a los intervalos de confianza y a diferentes pruebas de hipótesis para identificar, en el análisis de los datos de la investigación, los valores clínicamente relevantes. *Conclusión:* El valor *p*, los intervalos de confianza y la identificación de diferencias clínicamente relevantes por medio del uso de hipótesis de superioridad, de no inferioridad y de equivalencia son fundamentales para la investigación clínica.

Palabras clave: valor *p*, intervalos de confianza, diferencia clínica relevante, significación estadística.

Title: Beyond *p* value

Introduction: In medicine the p value has had an important place because of its contribution. In addition the confidence intervals and new formulations of significant test are used every day as a way to identify clinically relevant results. *Objective:* To describe the criteria and the significant test beyond the p value. *Results:* Confidence intervals and significant tests are review to identify in data analysis clinically relevant findings. *Conclusion:* The p value, confidence intervals and the identification of clinically relevant findings by means of superiority, non-inferiority and equivalence hypothesis are fundamentals in clinical research.

Key words: P value, confidence intervals, relevant clinical difference, statistical signification

Introducción

En un artículo previo (1) analizamos las críticas a las pruebas de hipótesis en el marco de su desarrollo histórico. En él se mostró cómo diferentes aspectos en la historia de las pruebas de significación han condicionado las deficiencias que son reportadas actualmente en su uso; deficiencias que se han tradu-

[1] Ph. D. en Matemáticas. Profesor asociado del Departamento de Epidemiología Clínica y Bioestadística, Pontificia Universidad Javeriana, Bogotá Colombia.
[2] Psiquiatra, psicoanalista, MSc en Epidemiología Clínica. Profesor del Departamento de Psiquiatría y Salud Mental y del Departamento de Epidemiología Clínica y Bioestadística. Coordinador de la Especialidad en Psiquiatría de Enlace. Pontificia Universidad Javeriana, Bogotá, Colombia.

cido en limitaciones y restricciones como las reflejadas en la Normas de Vancouver (2).

Las dificultades con las pruebas de hipótesis se han derivado de inconsistencias teóricas de la técnica (1) y de un uso deficiente de esos procedimientos, aunque matizado ello por el mal hábito de reducir su aplicación a la ejecución de un algoritmo de conducta que dicotomiza el análisis de los datos que lo reduce a dos categorías: *acepto/rechazo*.

Como parte de los mecanismos de análisis con los que se ha pretendido suplir los problemas en tal sentido, ha comenzado a introducirse una cierta carga de subjetivismo en el análisis de datos de las diferentes áreas de la biomedicina (1,3). Estas nuevas limitaciones tienen un eje central en el uso inadecuado de la plausibilidad biológica, a lo cual se adiciona un manejo deficiente de la información que brindan los intervalos de confianza, y que permitiría, en caso de ser utilizados adecuadamente, complementar la información del valor *p* obtenido en la prueba de hipótesis.

Fleiss (4) señaló que las pruebas de hipótesis tienen una función importante en el análisis de datos; sin embargo, es difícil encontrar en las revistas científicas del tema artículos que las defiendan: ciertamente, existen, pero lo habitual en la literatura es criticarlas y promover que no sean utilizadas. Esta situación genera un contrasentido, pues, por una parte, se siguen utilizando en la inmensa mayoría de las publicaciones donde se dan resultados de las investigaciones, y, por la otra, la inmensa mayoría de los artículos metodológicos estimula su no utilización. Este artículo pretende mostrar cómo se logra un uso eficiente de las pruebas de hipótesis, para poder seguir utilizándolas y no desechar la objetividad que ellas podrían introducir en los análisis. El punto central está en la identificación de cómo la relevancia biológica del resultado puede vincularse con la metodología de las pruebas de significación.

El punto de partida de esta propuesta metodológica fue que el uso eficiente de las pruebas de hipótesis queda determinado por el sustrato clínico y estadístico de su aplicación, así como por la comprensión del alcance y el significado de las conclusiones que se deriven de su uso: las pruebas de significación brindan una base objetiva para los análisis al presentar y cuantificar la evidencia contenida en los datos contra la hipótesis nula; es decir, contra la hipótesis que interesa evaluar.

La relevancia del planteamiento metodológico que se pretende hacer en este artículo está en que llena un vacío en la literatura docente, pues, lamentablemente, muchos de los libros de texto con los que se enseña la estadística, tanto en el pregrado como en el postgrado de las ciencias biomédicas, tienen serias limitaciones metodológicas al presentar las pruebas de hipótesis (5), por lo que han contribuido, y continúan contribuyendo, a los

problemas que se han presentado en su aplicación.

Este artículo pretende mostrar cómo debe plantearse un problema, de qué manera identificar las hipótesis estadísticas más pertinentes e indicar cómo la relevancia biológica del resultado puede formar parte del planteamiento del problema; eso aseguraría que el propio análisis de los datos se derive de aquella, con lo que se ganaría en objetividad al realizar el análisis de los datos de las investigaciones. Adicionalmente se analizará la manera más adecuada de utilizar los intervalos de confianza como complemento a las pruebas de hipótesis, al resaltar qué nueva información podría introducir su aplicación en el proceso de análisis utilizando las pruebas de hipótesis.

Las hipótesis estadísticas en el centro del problema

La metodología del planteamiento de un problema de investigación establece la necesidad de identificar una pregunta de investigación; en muchos casos ella es una pregunta clínica que se deriva de un interés asistencial específico. Esta pregunta de investigación viene acompañada por una hipótesis de investigación, la que se construye a partir de una de las posibles respuestas a la pregunta de investigación (6). La respuesta a la pregunta de investigación se basa en determinar la veracidad de la hipótesis de investigación, decisión que se toma a partir de los datos obtenidos en la investigación que se diseñe para tal fin. Los datos contienen la información que permitirá validar la hipótesis de investigación; para enjuiciar esa información y orientar el proceso de decisión se construyen las hipótesis estadísticas.

Las hipótesis estadísticas representan un punto medio entre los datos y la hipótesis de investigación. Estas se formulan en términos de elementos de las distribuciones de probabilidad, distribuciones que, a su vez, son entes teóricos que representan las frecuencias observadas en los datos de las variables del estudio. Así, si en un ensayo clínico se pretende medir la eficacia de dos intervenciones psicoterapéuticas para el manejo de la distimia, entrevista motivacional *vs.* terapia cognitiva, la pregunta de investigación se referiría a la semejanza de las intervenciones o a que alguna de ellas pudiera resultar más efectiva que la otra.

De la pregunta elegida se derivaría la hipótesis de investigación, la cual establecería un planteamiento general en la comparación de las dos intervenciones bajo estudio. Para tomar una posición acerca de su validez se realizaría el ensayo clínico, en el que diferentes mediciones brindarían la información necesaria.

La combinación de las evidencias brindadas por esas mediciones, sea a favor o en contra de las intervenciones, sería la evidencia experimental sobre la base de la cual se debe decidir acerca de la validez o no de la hipótesis de investigación;

por ejemplo, como medida de desenlace, se podría optar por una medición del nivel de depresión mediante la escala de Zung: la respuesta a la pregunta de investigación tendría como componente la comparación del comportamiento de los valores del puntaje de Zung en ambos grupos, y esta comparación podría consistir en comparar los promedios de puntaje de Zung para ambos grupos, o en comprar los porcentajes de pacientes con mejoría en cada grupo de tratamiento, y definir como mejoría si se obtenía una reducción del 30 % de los valores iniciales de depresión, de acuerdo con la escala. Como se aprecia a raíz del ejemplo, diferentes criterios de construcción de hipótesis estadísticas pueden ser utilizados alternativamente para dar respuesta a una misma pregunta de investigación.

Lo usual en la construcción de las hipótesis estadísticas es identificar una hipótesis nula o de no cambio, hipótesis que se identifica como H_0. Para decidir si, sobre la base de los datos, se puede rechazar $H_{0,}$ se calcula la probabilidad de obtener la muestra observada o muestras "más extremas" en la dirección en que H_0 no sea cierta; es decir, la probabilidad de muestras que determinen el rechazo de H_0. Este valor se conoce con el nombre de valor *p*, y es calculado por los sistemas computacionales para el análisis estadístico.

Por ejemplo, si se comparan las medias de dos poblaciones, como se podría hacer en el ensayo clínico, la hipótesis nula para considerar sería H_0: $\mu_{grupo\ 1} = \mu_{grupo\ 2}$; la probabilidad buscada sería la correspondiente a las muestras en las que sus medias tendrían una diferencia mucho mayor que la observada en los datos que se analizan, y donde serían las diferencias de interés las que correspondan con la dirección de las diferencias que identifica H_A. Esta probabilidad, el valor *p*, es el centro del proceso de decisión en el que se pretende aceptar o rechazar H_0.

Los valores *p* cuantifican la discrepancia entre los datos y la hipótesis nula (7); a medida que el valor *p* sea más pequeño se considera más fuerte la evidencia contra H_0, y más factible, la posibilidad de que el cambio o diferencia indicado por la hipótesis alternativa, si existe alternativa, sea cierto; si no existe una alternativa plausible, la decisión sería que no se encontraron evidencias de que H_0 sea cierta. En el caso contrario, es decir *p* elevado, se acepta H_0; es decir, se acepta que no hay cambio.

Por ejemplo, en el caso de la distimia, si la diferencia entre ambos grupos, entrevista motivacional y terapia cognoscitiva fuera a favor de la terapia cognitiva al comparar las medias con un valor $p = 0{,}00001$, se concluiría que la terapia cognoscitiva es más eficaz para el manejo de la distimia, en tanto que si el valor *p* fuese igual a 0,9999 se concluiría que no se puede rechazar la hipótesis nula de igualdad de los dos tipos de intervención. Valores tales como $p = 0{,}061$ indican una

débil evidencia contra la hipótesis de nulidad, y, en correspondencia, serían portadores de una cierta incertidumbre en la decisión (8).

Existen algunas creencias erróneas al interpretar el valor *p*: por ejemplo, es falsa la afirmación de que el valor *p* representa la probabilidad de que H_0 sea cierta, y también es falso que *p* mida o caracterice la magnitud de las diferencias entre los grupos que se comparan, o de los efectos bajo estudio, y también es falso que, como consecuencia, mientras *p* sea más pequeño mayor es la diferencia o el efecto. El valor *p* no mide ni cuantifica efectos: sólo indica la evidencia de los datos a favor o en contra de la validez de H_0. Es importante recordar el carácter aleatorio del valor *p*.

Un error muy difundido en el uso de los valores *p* es tratar de dicotomizar la decisión: los libros de estadística recomiendan tomar un valor α, usualmente 0,05, como valor de umbral para esa dicotomización, y enfocar el análisis según el siguiente algoritmo: si $p < 0,05$ rechazar H_0, y en caso contrario, aceptarla (9,10). Esta forma de proceder es incorrecta, como fue argumentado en el artículo previo (1). Por ejemplo, si un estudio obtiene $p = 0,04999$ rechaza H_0 porque $p < 0,05$; pero si hubiera obtenido $p = 0,05001$ en ese caso acepta H_0, porque $p > 0,05$. La inconsistencia salta a la vista: los dos valores *p* que se utilizan se diferencian en 0,00002; es decir, son prácticamente iguales, y, sin embargo, conducen a dos conclusiones absolutamente distintas. Lo relevante sería publicar el valor *p* en la cuantía en que fue observado, y analizarlo siguiendo las recomendaciones de Sterne y Smith (1,8).

En ningún caso los datos deben ser analizados mediante un algoritmo de conducta que dicotomice el proceso de análisis en un acepto/rechazo, proceso que no tiene en cuenta ninguno de los matices de los datos, ni el propio significado de las mediciones. Esta forma de conducta en los análisis necesita cambios conceptuales en la forma de ver las pruebas de hipótesis: la cuestión primaria es que no interesa saber si las diferencias son significativas o no: lo relevante en realidad es saber de cuánta evidencia disponemos para afirmar que la hipótesis nula es cierta o no, y esto lo cuantifica el valor *p*.

Los intervalos de confianza como parte del análisis

Los libros de texto de estadística (9,10) recomiendan utilizar los intervalos de confianza como un sustituto del proceso de decisión con las pruebas de hipótesis. Para ello, indican construir un intervalo de confianza que estime el parámetro que se analiza en la hipótesis, y utilizar esa estimación como una pauta para decidir.

Ejemplo de ello sería un experimento clínico controlado en el cual se compare la eficacia de dos tratamientos para la reducción de deli-

rios en pacientes con esquizofrenia; como medida de interés, se podría considerar el número de actividades delirantes, y la hipótesis estadística podría ser H_0: $\mu_{tratamiento1} = \mu_{tratamiento2}$, que compara las medias de los dos grupos y establece su semejanza; como resultado del ensayo clínico, la estimación mediante un intervalo de confianza del 95% de $\mu_{tratamiento1} - \mu_{tratamiento2}$ bien podría ser que esa diferencia sea un valor entre -4 y 5, lo cual indicaría que H_0 no debe ser rechazada; sin embargo, si la estimación obtenida hubiera sido entre 3 y 5 indicaría que el Tratamiento 1 tiene valores superiores, por lo que se rechazaría H_0.

Esta forma de proceder es parcialmente correcta; su limitación está en que también se traduce en un algoritmo que esquematiza el proceso de análisis. Como fue referido en (1), eso equivale a dicotomizar la decisión utilizando el valor *p* y el umbral α (5,8). Pretender eliminar los problemas de las pruebas de hipótesis con este procedimiento es una falacia, se traduce en hacer lo mismo de una forma diferente; lo grave es que se piensa que se está haciendo algo distinto, y que, por tanto, se eliminaron los problemas.

El verdadero aporte del intervalo de confianza, aquello que lo diferencia de las pruebas de hipótesis en el proceso de decisión, está en el significado biológico de sus valores como estimación de los parámetros de interés y en su diámetro; es decir, su amplitud, entendida ésta como la diferencia o separación entre sus valores. El diámetro es una medida de la incertidumbre en la estimación, incertidumbre que se presenta como consecuencia del efecto del azar; ella es consecuencia directa de limitaciones en el tamaño de muestra y del propio esquema de muestreo. El diámetro del intervalo de confianza es directamente proporcional a la variabilidad del estimador que se utiliza en su construcción, estimador que, a su vez, aproxima al parámetro que se pretende analizar. Con poca información en la muestra no es posible emitir juicios seguros, pues la variabilidad sería grande.

Información en la muestra y exactitud en las conclusiones son dos categorías inversamente proporcionales. Por ejemplo, en un estudio de casos y controles que pretende estudiar los factores de riesgo asociados a la presencia de un cuadro de hipocondría se tendrían los siguientes factores de riesgo y sus respectivos OR (*odds ratios*) e intervalos de confianza:

- Maltrato durante la infancia. OR= 1,2 (0,4 – 84) es un intervalo amplio, que contiene 1; esto es, la no diferencia. Por otra parte, el intervalo muestra en un extremo que es protector, y en el otro tiene 84 *odds* más de presentar hipocondría. La amplitud del intervalo es muy grande, como consecuencia de una elevada variabilidad en la estimación del OR. Eso significaría que los niveles de incertidumbre de la

estimación son grandes, y eso se refleja en la estimación y, consecuentemente, en la calidad de la inferencia.

- Hospitalizaciones en la primera infancia. OR=2,5 (1,5-3,4) es un intervalo bastante estrecho, lo que sería un reflejo de una variabilidad baja de la estimación del OR; es decir, un efecto del azar no muy elevado, y, en correspondencia, la estimación y los juicios que de ella se deriven serían más confiables. Muestra cómo las internaciones durante la primera infancia pueden estar asociadas a un mayor riesgo de presentar hipocondría. En el caso de este estudio, el OR oscilaría entre 1,5 y 3,4.
- Herencia de hipocondría. OR=4 (3, 267) es un intervalo bastante amplio y, por ello, poco preciso. Ello puede obedecer a la falta de poder del estudio para detectar y determinar esta variable.

¿Cómo utilizar la relevancia clínica de un resultado en los análisis? La superioridad, inferioridad y equivalencia como su reflejo

Ante las limitaciones de las pruebas de hipótesis se ha generalizado la recomendación de identificar la *plausibilidad biológica (clínica) de los resultados*. El momento de confusión se genera al determinar qué se entiende por *introducir la plausibilidad biológica (clínica) en el análisis*.

El procedimiento más socorrido es aplicar la prueba de hipótesis, y si su conclusión no es aceptable para el investigador, éste no asume el resultado, sino que comenta un argumento estadístico, como "lo que está observando es reflejo de insuficiencias en el tamaño muestral", o algo semejante, y, en consecuencia, concluye que no es relevante. De esta manera se introduce un subjetivismo en los análisis, lo cual es contrario a la objetividad que se pretende lograr con las pruebas de hipótesis: objetividad, que es el fundamento de su uso. En pocas palabras, si el investigador al final del estudio decide lo que es relevante y lo que no, ¿para qué le sirven los datos? Esta situación es contradictoria e introduce un subjetivismo inaceptable en el marco de la imparcialidad y objetividad que se deben exigir a un criterio de análisis de datos. La tarea a este respecto es poner en concordancia el análisis estadístico con lo que se entiende es relevante según la biología del problema, e introducir ésta en los análisis de manera independiente, que las conclusiones se obtengan independientemente de los juicios "personales" del investigador.

El tamaño de muestra debe ser establecido *a priori*, es uno de los componentes centrales del diseño del estudio y está directamente relacionado con su alcance y sus limitaciones: medición que no pueda ser tratada con la calidad necesaria no debe ser analizada; sencillamente, el estudio no fue diseñado para

eso. Ésta es una decisión que se toma en el momento del diseño de la investigación; es decir, se toma *a priori*, no *a posteriori*, si el investigador descubre que los resultados no son como deseaba.

Cuando se introduce la hipótesis nula como hipótesis de no cambio o de semejanza, y se pretende con ello analizar si existen o no "diferencias significativas", no siempre se está siendo consecuente con lo que es biológica o clínicamente relevante en el problema como cambio.

Por ejemplo: en estudios con antidepresivos, se podría determinar cuál es el nivel mínimo de diferencia de los valores encontrados en la escala de Hamilton para medir síntomas depresivos entre los grupos comparados; valdría la pena preguntarse: ¿cuál es la diferencia mínima relevante para considerar que un tratamiento antidepresivo es mejor que otro?

Una diferencia de 2 o 3 puntos en la escala de Hamilton podría ser estadísticamente significativa, pero, ¿es clínicamente relevante? ¿Cuál es el valor (delta) mínimo para considerar una diferencia clínicamente relevante en este caso? Siendo consecuentes con la definición de la escala, la respuesta serían 5 o 6 puntos, o incluso más. Esta consideración es de gran importancia al momento de determinar qué diferencia se espera encontrar, y cómo ella determina cambios relevantes en los valores de la escala.

Dicha situación fue identificada por Cohen (11), quien en 1994 denominó *hipótesis nada* a la hipótesis nula de no cambio (*null hypothesis*). Es un hecho reconocido entre los profesionales de la estadística que estas hipótesis, en muchos casos, son irrelevantes en el proceso de decisión. Tuckey comentó en 1991, según cita Cohen (11): "Es tonto preguntar: '¿Los efectos A y B son diferentes?'. Éstos son siempre diferentes en algún lugar decimal" y siempre es posible encontrar un tamaño de muestra lo suficientemente grande, que haga significativa esa diferencia.

Supóngase que se desea comparar la talla de dos grupos de personas (A y B). En cada grupo se toman 15 personas al azar: en el A se obtiene una media de 178 cm, con una desviación estándar de 2 cm, y en el B, una media de 180 cm, con la misma desviación estándar. Aplicando la prueba de comparación de medias se obtiene p = 0,0053; aplicando el proceso usual de decisión con un umbral de 0,05 habría que concluir que se rechaza la igualdad de las medias de los dos grupos; se aceptaría, según el análisis estadístico, que las dos medias son "significativamente diferentes", pero el investigador, que tiene algunas posiciones en el problema, tiene otras ideas respecto al resultado estadístico:

Observando críticamente el resultado obtenido, la pregunta sería: ¿hasta qué punto 2 cm identifican, en promedio, una diferencia de estatura relevante desde el punto de vista no sólo numérico, que es lo que se obtuvo con el análisis

anterior, sino también biológico? Pensando lógicamente, se esperaría que el análisis considerara, de alguna forma, el hecho de que 2 cm no marcan una diferencia relevante en las tallas, pero el procedimiento de análisis empleado fue genérico, no tuvo en cuenta las peculiaridades de la medición analizada; de hecho, es el mismo que los textos de estadística sugieren para todas las áreas del conocimiento, sin excepciones y sin tener en cuenta sus especificidades. Luego, el análisis no incorporó ningún elemento de la biología del problema que se está abordando. En esta generalidad está, justamente, la insuficiencia del análisis que se hizo. Este punto será discutido posteriormente.

Supóngase que al discutir los resultados de ese análisis, usted, como investigador(a), tiene antecedentes de otros estudios, acerca de que en este problema no se espera que la estatura cambie al comparar los grupos; por otra parte, usted ve que 2 cm no marcan diferencias entre los grupos, aunque se lo diga la estadística, y, por último, diseñó el estudio con la fuerte convicción de que no iba a encontrar diferencias.

Si se fueran a interpretar las conclusiones obtenidas, tal como las discuten algunos investigadores, utilizando la "plausibilidad biológica", se diría: "esta diferencia, significativa según la estadística, no es relevante clínicamente, está demostrado que lo que se comprara no cambia la estatura pues se sabe que …". En consecuencia, no se asume la diferencia entre las tallas de los grupos, encontrada con los métodos estadísticos; para justificar esta arbitrariedad se utilizan argumentos pseudoestadísticos, como: "si se hubiera dispuesto de un mayor tamaño de muestra se hubiera tenido un menor efecto del azar y no se hubieran encontrado diferencias".

Pero esta forma de proceder, por subjetiva, es absolutamente incorrecta, un análisis de este tipo no sería objetivo, aunque se presente muy frecuentemente. Proceder así no es deseable. Cuando se introdujeron las pruebas de hipótesis, a principios del siglo XX, la necesidad de su uso se derivó del imperativo de eliminar el subjetivismo y los matices anecdóticos, que predominaban en las publicaciones científicas; la forma de proceder que se ilustra constituye un retroceso a ese momento. Sin embargo, la inconsistencia de la estadística es evidente; ¿qué es lo que no funciona bien?

La búsqueda de un camino para un análisis de datos que sea metodológicamente correcto debe tener en cuenta la causa del error, y aquel estuvo en el procedimiento de comparación que se empleó: la metodología utilizada para el análisis de los datos no es razonable en el marco del significado de la medición, porque no responde al interés del análisis visto este en el contexto de las características biológicas de la variación de la medición: estadística y significación biológica no aparecieron integrando conjuntamente el procedimiento de

análisis utilizado, fueron utilizadas como dos componentes aislados. ¿Cómo proceder?, ¿cómo integrar dos elementos tan diferentes?

Lo que se propone para recuperar la objetividad es introducir hipótesis que combinen la formulación estadística y la biología o clínica del problema. Para ello, las hipótesis estadísticas deberían permitir que se identifiquen cambios sólo si estos son biológicamente relevantes: la respuesta se encuentra en los ensayos de equivalencia, superioridad e inferioridad; concretamente, en las hipótesis que se someten a evaluación en ellos. Por supuesto, dicha propuesta será útil en la medida en que el diseño del estudio garantice que el tamaño de muestra permita abordar las comparaciones de interés con la potencia considerada *a priori*.

Para formular hipótesis que tengan en cuenta la biología del problema o aspectos de costo, beneficio u oportunidad, es necesario identificar un valor umbral, δ, que identifique un rango que marque la aceptabilidad clínica o biológica de las diferencias o las semejanzas. Es decir, identificar un valor que marque la diferencia mínima que sería clínicamente relevante. En ese caso, el valor de umbral identificaría diferencias relevantes solamente si son mayores o menores que él.

Por ejemplo, en un estudio para evaluar la efectividad de un nuevo tratamiento, el X sobre la W como antipsicóticos, se realizó la comparación entre los tratamientos, de manera que para aceptar el nuevo producto bastaría con que sus porcentaje de éxitos no fueran inferiores a los de la Y en un 2%; es decir, aunque Y es superior al W, si esta superioridad no sobrepasa un 2% se aceptaría como relevante el nuevo producto por sus ventajas.

Visto el ensayo clínico de la manera usual, las hipótesis serían: H_0: $P_X = P_W$ contra la alternativa H_A: $P_X \neq P_W$, donde P_X denota el porcentaje de éxitos terapéuticos con X, y P_W, el porcentaje de éxitos con W. Esta forma de ver el ensayo, aunque es la usual, no tiene en cuenta que el nuevo producto tiene tantas ventajas adicionales que es deseable, aun si es ligeramente menos eficiente en términos terapéuticos que el tradicional.

Considerando el umbral del 2% como punto central de la aceptación del nuevo producto, las hipótesis para el análisis de los datos que integrarían todos los intereses serían: H_0: $P_W - P_X \geq 2$ contra la alternativa H_A: $P_W - P_X < 2$. Estas hipótesis, que se denominan *de no inferioridad*, establecen la aceptación del nuevo producto incluso si es sobrepasado en su efectividad por la W, pero si esta superioridad no es mayor que el 2%, así: aceptar H_0 representaría que la W es superior al X en un 2%; en este caso no es recomendable el cambio según lo planteado; aceptar H_A representaría que la W no sobrepasa al X al menos en un 2%, y esto significaría que la W puede ser superior, pero con superioridad por debajo del 2%; o sea, que el X es superior, lo que sería representado por los valores negativos de la diferencia:

en ambos casos, según lo planteado, se aconseja sustituir el tratamiento usual por el nuevo producto, y eso es lo que establece H_A: aceptar el nuevo producto si es mejor, o si no es "muy inferior" al usual, según el valor de umbral aceptado *a priori*.

Como alternativa al planteamiento de las hipótesis de semejanza, que son las que se analizan usualmente en los ensayos clínicos y comúnmente en los análisis estadísticos de los diferentes problemas, se plantea la necesidad de sustituirlas por hipótesis de equivalencia, no de inferioridad o de superioridad (12). Esta nueva formulación de las hipótesis, que realmente no es nueva, pues son conocidas por los estadísticos desde hace tiempo, permitirá introducir componentes adicionales en los análisis al vincular en una sola pieza los aspectos estadísticos y metodológicos que se identifican a partir de la plausibilidad biológica o de otras consideraciones, como serían el costo-beneficio o la oportunidad. La estructura de tales hipótesis se discute, por ejemplo, en el texto de Chow (12).

La búsqueda de los valores que marcan los umbrales que definen las hipótesis podría ser compleja, pues tales valores no siempre están establecidos (12). En el caso de los ensayos clínicos, existen algunas áreas donde están normados (13), su determinación debe ser vista como una componente fundamental en la investigación clínica y todo parece indicar que es la única forma de dar sentido clínico al uso de las pruebas de hipótesis.

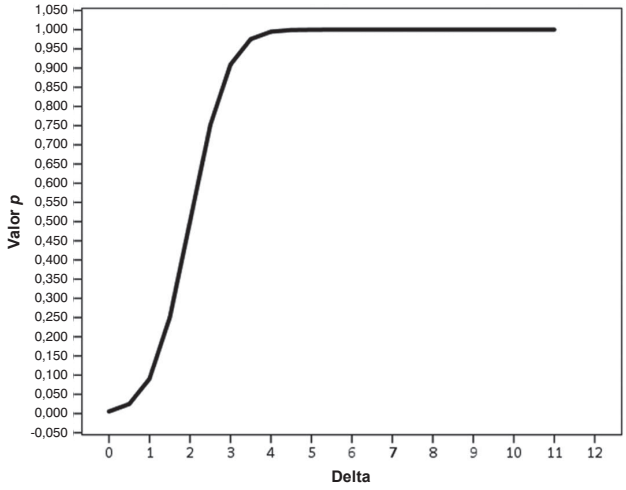
La no existencia de valores de umbral que cuenten con un consenso bien podría ser una limitación para su uso; aunque dichos valores no siempre están determinados, sí se ha establecido que ellos deben cuantificar el peor caso de pérdida en eficacia que es aceptable clínicamente, considerando la seguridad potencial, la conveniencia o las ventajas del nuevo tratamiento, en el caso de los ensayos clínicos (13), o los niveles mínimos del comportamiento de las mediciones que son biológicamente relevantes para detectar diferencias.

La Federal Drug Administration de Estados Unidos ha establecido algunas reglas para la identificación de los valores de umbral en los ensayos de equivalencia o no inferioridad (14,15); ellas combinan consideraciones estadísticas y clínicas.

En el caso del ejemplo donde se comparó la talla de dos grupos de personas, el valor δ sería determinado por una diferencia en la talla que identifique una discrepancia en las tallas que sea relevante al comparar los grupos en el contexto del problema abordado. Esto determina que el valor umbral es contextual; es decir, relativo al problema: si lo que se pretende en este caso es ver si las diferencias de tallas entre los grupos marca una diferencia en el estado nutricional, el valor de δ debe detectar ese umbral de cambio relevante en la talla.

Según este criterio, un candidato para umbral podría ser 8 cm, pues es la variación en la talla que

Figura 1. Valor p como función de δ



identifica un cambio de una unidad en el índice de masa corporal para un individuo con un peso promedio de 70 kg, suponiendo que este fuera el peso promedio en la población. Ello significaría que son relevantes, en promedio, diferencias de tallas por encima de δ = 8 cm. En ese caso, la hipótesis nula sería H_0: $\mu_B - \mu_A \leq 8$, lo cual significa que diferencias que no sean significativamente superiores a 8 cm no son relevantes en el problema. En ese caso, rechazar H_0 sí significaría una decisión de cambio nutricionalmente relevante.

Ante la dificultad de identificar el valor de umbral, se propone una alternativa gráfica para el análisis. Esta alternativa parte de considerar diferentes valores de δ, realizar la prueba de hipótesis correspondiente y calcular el valor p correspondiente. La Figura 1 parte de considerar la variación del valor p como función de δ. En los casos en que no esté determinado un valor de umbral para la identificación de cambios biológicamente relevantes, es recomendable construir un gráfico de ese tipo para identificar a partir de qué valor de las diferencias, o sea, para qué umbral, se aceptaría o rechazaría la hipótesis del cambio: de esta manera se podría mejorar la calidad del análisis de los datos y obtener conclusiones más objetivas y con un componente biológico intrínseco.

La Figura 1 muestra la relevancia que tendría el valor δ en el proceso de inferencia del ejemplo. Para valores de $\delta \geq 1$ cm el valor *p* que se observaría sería mayor que 0,01, lo que indicaría el no rechazo de H_o: $\mu_B - \mu_A \leq \delta$. Bastaría con considerar un valor mínimo de 1cm como umbral para la relevancia de las diferencias de talla, para que cambie la conclusión a que se arribó en el ejemplo con el cual se mostró, anteriormente, la forma de proceder en el análisis estadístico habitual.

De esta forma queda clara la irrelevancia de las diferencias que se encontraron entre los grupos, y se hace evidente lo inadecuado de buscar diferencias significativas como criterio de análisis de los datos en algunos problemas.

Referencias

1. Monterrey P, Gómez-Restrepo C. Aplicación de las pruebas de hipótesis en la investigación en salud. ¿Estamos en lo correcto? Universitas Médica. 2007;48(3):193-206.
2. International Comitee of Medical Journal Editors. Uniform requirements for manuscript submitted to biomedical journals. Br Med J. 1997;336(4):309-15.
3. Jonson D. The insignificance of statistical significance testing. Journal of Wildlife management. 1999;63(3):763-72.
4. Fleiss JL. Significance test have a role in epidemiologic research: reactions to AM Walker. Am J Pub Health. 1986;76(5):559-60.
5. Gliner J, Leech N, Morgan G. Problems with null hypothesis significance testing (NHST): what do the textbooks say? The Journal of the Experimental Education. 2002;71(1):83-92.
6. Polit DF, Hungler BP. Investigación científica en ciencias de la salud. Principios y métodos. Mexico: McGraw-Hill Interamericana, HealthCare Group; 2000.
7. Weinberg C. It´s time to rehabilitate the P-Value. Epidemiology. 2001;12(3):288-90.
8. Sterne JA, Davie Smith GD. Sifting the evidence-what´s wrong with significance tests? Br Med J. 2001;322(7280):226-31.
9. Daniell W. Bioestadística. Base para el análisis de las ciencias de la salud. México: Limusa; 2002.
10. Pagano M, Gauvreau K. Fundamentos de bioestadística. México: Thomson Learning; 2001.
11. Cohen J. The earth is Round (p<.05) American Psychologist. 1994;49(12):997-1003.
12. Chow S, Shao J, Wang H. Sample size calculations in clinical research. New York: Marcel Dekker; 2003.
13. Kaul S, Diamon G, Weintraub W. Trials and tribulations of non-inferiority: The ximelagatran experience [Internet]. San Diego: J Am Coll Cardiol; 2005 Nov; [2009 jun 20]. Disponible en: http://content.onlinejacc.org/cgi/reprint/j.jacc.2005.07.062v1.pdf.
14. International Conference on Harmonisation. Statistical principles for clinical trials (ICH E9). 1998, Feb 5 [2009 jun 20]. Disponible en: http://www.ich.org/LOB/media/MEDIA485.pdf.
15. International Conference on Harmonisation. Guidance on choice of control group and related design and conduct issues in clinical trials (ICH E 10). Federal Register, 2000 Jul 20 [2009 jun 20]. Disponible en: http://www.ich.org/LOB/media/MEDIA486.pdf.

Conflictos de interés: los autores manifestamos que no tenemos ningún conflicto de interés en este artículo.

Recibido para evaluación: 7 de abril del 2009
Aprobado para publicación: 17 de julio del 2009

Correspondencia
Pedro Monterrey
Hospital Universitario San Ignacio
Departamento de Epidemiología Clínica y Bioestadística.
Carrera 7ª No. 40-62, 2º piso
Bogotá, Colombia
pmonterrey@javeriana.edu.co