Evaluación de resultados clínicos I: Entre la significación estadística y la relevancia clínica

Ioseba Iraurgi
Psicólogo clínico
Módulo de Asistencia Psicosocial de Rekalde. Bilbao

Hay quien utiliza la estadística como el borracho la farola: más para apoyarse que para iluminarse

Resumen: Resulta frecuente encontrar en las publicaciones biomédicas y psicosociales estudios orientados a valorar efectos entre tratamientos o intervenciones que presentan el grado de significación *p* como el principal argumento científico sobre la existencia o ausencia de tales efectos. Esta tendencia a centrar las conclusiones estadísticas en el grado de significación conlleva un grave problema: confundir la 'significación o importancia clínica o científica' con la 'significación estadística'. Inicialmente, a partir de datos simulados, se presenta evidencia de que ambas significaciones corresponden a procesos diferentes, los cuales han de considerarse en conjunto y no de manera excluyente. Se repasan los conceptos de significación estadística, el proceso de investigación e inferencia estadística y, finalmente, el concepto de significación o importancia clínica.

Palabras clave: Significación estadística, Significación clínica, Contraste de hipótesis, Inferencia estadística.

Summary: In the biomedical and psychosocial publications, result frequent to find studies guided to value effects between treatments or interventions that present the *p* value as the principal scientific argument on the existence or absence of such effects. This trend to center the statistics conclusions in the significance value involve a serious problem: to confuse the 'clinic/scientific significance or importance' with the 'statistical significance'. Initially, on the basis of simulated data, is presented evidence of the fact that both significances correspond from different processes, those which have of be considered by and large and not of excluding way. The concepts of statistical significance, investigation process and statistical inference and, finally, the clinical significance or clinical importance are reviewed.

Key words: Statistical significance, clinic significance, Hypothesis contrast, Statistical inference.

La cita que sigue al título de este artículo —una adaptación de la ofrecida por Cantervill y recogida en la editorial a este número de la revista *Norte de salud mental*—, pretende ser un a priori de la idea que se aspira a transmitir en las páginas siguientes: la estadística es un medio, no un fin para dar respuesta a las preguntas científicas. Es una herramienta útil y válida utilizada en el método científico, pero se espera que quien la utilice lo haga con rigor y

no de forma inapropiada, sea de forma intencionada porque conoce su utilidad y pervierte su uso, o bien de forma ingenua, por desconocimiento de las normas que la rigen y ante la necesidad de publicar unos resultados. A esto último ha contribuido enormemente la disponibilidad de sofisticados programas informáticos que ponen a nuestro alcance todo un abanico de pruebas y algoritmos estadísticos, hasta el punto que muchos usuarios de estos programas creen ser expertos en estadística. Pero no basta con cruzar unas variables y observar los valores de significación estadística que nos dan estos programas para elaborar un adecuado informe de investigación, no por tener más datos y pruebas de contraste complicadas se ofrece un mejor informe. Hacer investigación no es manejar la estadística; se puede ser un buen investigador sin tener ni idea de estadística. Dominar el método científico es lo que hace a uno buen investigador, la estadística sólo es una de sus herramientas. Pero, precisamente por ser una utilidad para la ciencia, es ineludible que un buen investigador conozca qué es lo que subyace a la estadística y a su interpretación.

En el ámbito clínico, donde muchos profesionales sanitarios muestran interés en los avances diagnósticos y terapéuticos y siguen periódicamente las novedades en los medios de divulgación científica, o incluso colaboran en ellos, es de especial importancia tener unos conocimientos básicos de lo que nos ofrece la estadística. Muy a menudo se malinterpretan los resultados de las pruebas de contraste de hipótesis para descubrir las diferencias entre dos tipos de terapias, considerando que un resultado estadísticamente significativo implica un resultado clínicamente concluyente. Y resulta de enorme importancia diferenciar entre lo que es estadísticamente significativo de lo que es clínicamente importante o sustantivamente relevante. Como bien plantean Pita y Pértega (2001), desde el punto de vista clínico la significación estadística no resuelve todos los interrogantes que hay que responder ya que la asociación estadísticamente significativa puede no ser clínicamente relevante y, además, la significación estadística puede no ser causal. Es decir, podemos encontrar asociaciones estadísticamente posible y conceptualmente estériles (Silva, 1997).

Veamos qué queremos decir mediante un ejemplo práctico, aunque para el cálculo se hallan utilizado datos simulados. Supongamos que disponemos de dos tratamientos (A y B) para el manejo de una enfermedad, dolencia o síntoma (p.ej.: un trastorno del sueño), y deseamos conocer cuál de ellos es mejor. Las intervenciones pueden ser tanto farmacológicas, como psicoterapéuticas, como de otra índole. El tratamiento A, al que llamaremos experimental (Exp), es una modalidad de intervención novedosa que se espera tenga más éxito que el tratamiento B estándar, al que llamaremos control (Ctrol). Para valorar el éxito del tratamiento se consideran dos criterios: una inducción del sueño en menos de 20 minutos y un mínimo de duración del sueño de 6 horas. Por otra parte, para considerar mejor un tratamiento sobre el otro se espera una diferencia de mejoría de al menos un 10% en el porcentaje de enfermos que logran el objetivo. Con ello, estaríamos definiendo lo que entendemos por diferencia clínicamente importante, que en nuestro caso es de un 10%. Para llevar a cabo el experimento se diseña un ensayo clínico aleatorio con dos grupos de intervención realizado con condiciones óptimas de ejecución y control exhaustivo de los sesgos. Este ensayo es realizado en cuatro contextos diferentes: en un Centro de Salud Mental (CSM) de una localidad pequeña (Estudio 1), en un CSM de una gran ciudad (Estudio 2), y a través de la colaboración multicéntrica de 20 CSMs (Estudios 3 y 4). Los resultados obtenidos se presentan en la tabla 1 bien a través de una medida basada en un indicador dicotómico (obtiene o no, mejoría), por lo que estaríamos comparando una medida basada en proporciones, o a través de una escala de medida continua, donde la comparación sería de medias.

TABLA 1

Comparación de proporciones y medias para la valoración de los efectos clínicamente importantes y estadísticamente significativos. Datos simulados.

Contraste de Proporciones	Estudio 1		Estudio 2		Estudio 3		Estudio 4	
Tto	Exp	Ctrol	Exp	Ctrol	Exp	Ctrol	Exp	Ctrol
Mejoría								
Si	7 (70%)	5 (50%)	70 (70%)	50 (50%)	700 (70%)	700 (70%)	550 (55%)	500 (50%)
No	3 (30%)	5 (50%)	30 (30%)	50 (50%)	300 (30%)	300 (30%)	450 (45%)	500 (50%)
Total	10	10	100	100	1000	1000	1000	1000
Magnitud Efecto	20%		20%		20%		5%	
DP	1,55 (0,56 a 4,29)		1,55 (1,13 a 2,14)		1,55 (1,40 a 1,72)		1,10 (1,01 a 1,20)	
RR (IC 95%)								
Prueba contraste y Significación	χ^2= 0,83333 p= 0,361		χ^2= 8,333 p< 0,005		χ^2= 83,33 p< 0,00001		χ^2= 5,01 p< 0,025	

Contraste de Medias	Estudio 1		Estudio 2		Estudio 3		Estudio 4	
Tto	Exp	Ctrol	Exp	Ctrol	Exp	Ctrol	Exp	Ctrol
Mejoría								
Media	5	4	50	40	5	4	4,25	4
Desviación	3,03	1,61	3,04	1,62	3,04	1,62	3,03	1,61
Total	10	10	100	100	1000	1000	1000	1000
Magnitud Efecto								
DM (IC 95%)	1 (-1,46 a 3,46)		1 (0,32 a 1,68)		1 (0,78 a 1,21)		0,25 (0,03 a 0,46)	
Prueba contraste, Grados d libertad y Significación	t= 0,87 g.l.= 13,71 p= 0,398		t= 2,90 g.l.= 150,82 p< 0,005		t= 9,20 g.l.= 1521,89 p< 0,0001		t= 2,30 g.l.= 1521,89 p= 0,021	

DP: Diferencia de Proporciones /
RR (IC 95%): Riesgo Relativo (Intervalo de Confianza del 95% del RR)
DM (IC95%): Diferencia de Medias (Intervalo de Confianza del 95% de la DM)

χ^2: Ji cuadrado;
t: *t* de Student

En el Estudio 1 han participado 10 personas por grupo y observamos como con el tratamiento A (Exp) se han curado 7 personas (7/10*100= 70%) mientras que con el B (Ctrol) han mejorado 5 (5/10*100= 50%). Como podemos ver la diferencia de curaciones observada entre uno y otro es del 20% (70%–50%) y resulta bastante superior al 10% que previamente nos habíamos fijado como importante. Utilizando la prueba adecuada para la comparación de proporciones, la Ji–cuadrado de Pearson, obtenemos un valor de χ^2= 0,83 y un valor de probabilidad asociado de p=0,361. Para nuestra desgracia es un resultado no significativo, si tomamos como umbral de significación el famoso valor de p≤0,05. Nos encontramos ante un ejemplo de una diferencia clínicamente importante pero estadísticamente no significativa.

Por su parte, en el Estudio 2 intervienen 100 personas en cada grupo y se observan las mismas proporciones de éxito en ambos grupos que en el Estudio 1, de modo que la diferencia de mejoría del grupo Exp frente al Ctrol sigue siendo de un 20%. Ahora bien, en este caso se obtiene una prueba de χ^2= 8,33 con un valor de probabilidad estadísticamente muy significativo (p≤0,005). Incluso en el Estudio 3, con 1000 casos en cada grupo, se sigue manteniendo la misma diferencia de mejoría del 20%, siendo la χ^2= 83,33, pero en este caso alcanzando un valor de probabilidad de 0,00001. Y aquí encontramos una primera observación obvia: la significación estadística aumenta a medida que incrementamos el tamaño de la muestra, sin que se modifique en absoluto el efecto de mejoría del tratamiento.

Pero centrémonos ahora en el Estudio 4 respecto al 3; en los cuales participan 1000 personas en cada grupo de tratamiento. Como hemos podido observar en el Estudio 3 se ha logrado una diferencia clínicamente importante (un 20% de mejoría, superior al 10% planteado como diferencia mínima importante) y estadísticamente significativa (p≤0,05; concretamente p≤0,00001). En el Estudio 4 también se obtiene una diferencia estadísticamente significativa (p≤0,025), pero la importancia clínica lograda (de un 5%) no alcanza la diferencia clínica mínimamente importante del 10%. Es decir, el estudio arroja datos estadísticamente significativos pero clínicamente irrelevantes.

Es aquí donde resulta apropiada una reflexión de Guttman que, en una obra ya clásica de la estadística —'What is not what statistics', de 1977— proponía lo siguiente: *"Una prueba estadística 'no es una prueba de relevancia científica'. Bajo ningún concepto la relevancia del dato se encuentra en la técnica de análisis, sino en la repercusión del mismo en el conocimiento del objeto de estudio. Es perfectamente compatible encontrar diferencias estadísticamente significativas que no suponen relevancia clínica, y viceversa"*.

Volviendo a nuestro ejemplo, ¿qué podemos concluir de estos datos?. Pues que al aumentar

el tamaño muestral conseguimos aumentar la precisión de nuestras mediciones y disminuir la variabilidad explicada por el azar. Por eso, ante la misma diferencia pero con un mayor tamaño muestral hemos conseguido reducir el valor de la p de 0,361 a 0,005, e incluso a un valor menor de 0,0001. Como vemos el valor de p depende no solo de la diferencia de los grupos de estudio, sino del tamaño muestral. Siempre podemos encontrar diferencias estadísticamente significativas con un tamaño muestral lo suficientemente grande aunque las diferencias sean muy pequeñas e irrelevantes desde un punto de vista clínico o científico, como es el caso del Estudio 4. En definitiva, la p no es una medida de asociación, sino la probabilidad de que el resultado observado se deba al azar, y por ello se ve muy influenciada por el tamaño muestral ya que con ello se aumenta la precisión al reducirse el error de medida. Por tanto, las pruebas estadísticas, y por ende los valores p, son función de dos características fundamentales, el tamaño de las muestras y del tamaño del efecto encontrado; si ambos o alguno de ello aumenta considerablemente, los valores de p cada vez se hacen más pequeños. De este modo, lograr resultados estadísticamente significativos es relativamente fácil, aunque costoso, simplemente aumentando el número de efectivos participantes en un estudio; aumentar el tamaño de los efectos es precisamente lo que se pretenden con la investigación, pero este logro resulta normalmente ser más arduo de lograr.

La investigación y el proceso de contraste de hipótesis

Hasta el momento hemos venido hablando de la significación estadística, la cual no puede desligarse del proceso de inferencia a partir del contraste de hipótesis, el cual conlleva toda una conceptualización que a menudo es el causante de las miradas de perplejidad hacia la estadística. Como venimos diciendo reiteradamente, la estadística es una disciplina fundamentada en la matemática que tiene su principal base en la teoría de la probabilidad. La inferencia es el proceso de toma de decisiones, sobre la base de unos datos parciales pertenecientes a una población, para extraer conclusiones acerca de esa población. En el ámbito clínico trabajamos con personas que presentan una determinada enfermedad, trastorno o síntoma, por ejemplo una depresión, a las cuales tratamos con una(s) determinada(s) terapéutica(s). Acumulamos cierta experiencia y obtenemos resultados, pero el acceso al número de pacientes con esa patología, por muy larga que sea nuestra vida profesional, rara vez se aproxima al total de individuos que constituyen la población de personas con diagnóstico de depresión. En definitiva, en nuestro estudio contamos con una muestra limitada que forma parte de las infinitas posibles muestras de una población de referencia, lo cual implica que los resultados obtenidos pueden presentar fluctuaciones debidas puramente al azar. No obstante, trabajamos con unos pocos e inferimos que su respuesta es la que tendría el total de las personas con esa afección que fueran tratadas con esa determinada terapéutica. Pero para hacer esa extrapolación nos estaríamos manejando con un cierto grado de incertidumbre, y es ahí donde la estadística nos permitiría estimar el grado o probabilidad de acertar o equivocarnos. Si pudiéramos tratar al total de la población, no sería necesaria la estadística para llegar a conclusiones, ya que tendríamos la certeza absoluta. El ejemplo más evidente de este principio se halla en las elecciones de un país: antes del plebiscito los sondeos ofrecen una estimación del voto que perfila la posible configuración de los electos, con un nivel de confianza del 95% y un margen de error del 3% —¿les suena?—; una vez celebradas las elecciones la configuración de electos es de tantos y cuantos, sin horquillas ni variaciones posibles.

Pero cuál es la implicación de la estadística en el proceso de investigación. Pongamos un caso y vayamos desarrollándolo. En la situación más sencilla de investigación, por ejemplo, evaluar el efecto de un tratamiento sobre la mejo-

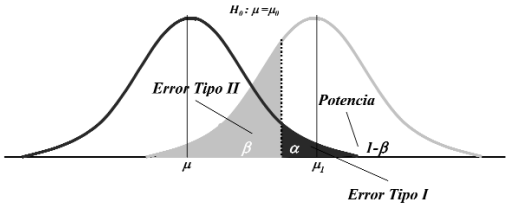
ría de un trastorno, nos encontramos con dos variables: una independiente —la primera— con al menos dos niveles de variación (ofrecer un tratamiento experimental —Exp— frente a otro alternativo o no ofrecer nada, denominado usualmente control —Ctrol—), y otra dependiente —la segunda— que puede ser medida también a través de dos niveles (éxito vs fracaso, existe o no sintomatología, etc.) o mediante una medida de tipo continuo (p.ej.: un inventario de síntomas como la BPRS) —recordemos los datos de la Tabla 1—. La tarea del investigador consiste en determinar el grado en que los datos de la investigación reflejan una relación entre las variables independiente y dependiente; o dicho de otro modo, el análisis estadístico de los datos persigue determinar si dos grupos que difieren en el lugar que ocupan en la variable independiente (exp vs Ctrol), difieren también en la variable dependiente (éxito vs fracaso). Pues bien, tanto en la relación entre las variables como la ausencia de relación entre las mismas, pueden resultar enmascaradas por dos clases de errores: 1) los errores constantes, también llamados sesgos, y 2) los errores variables o error estocástico o debido al azar. Los sesgos son producidos por variables extrañas que afectan de manera constante los resultados de una investigación; por ejemplo afectan siempre de la misma manera la relación (o la falta de relación) entre las dos variables (independiente y dependiente), o afectan siempre de la misma manera (favorable o desfavorablemente) a los grupos de la investigación, es decir, a los niveles de la variable independiente. Ejemplos de sesgos sería ofrecer una atención especial al grupo Exp más allá de lo estipulado en el protocolo, utilizar instrumentos mal calibrados, que los participantes del estudio reciban tratamiento simultáneo en otro servicio sin conocerlo, etc.). Por su parte, los errores estocásticos son producidos por variables extrañas que afectan de manera variable los resultados de la investigación; se trata de variables ajenas a la investigación que actúan aleatoriamente sobre todos los sujetos, y que por esa misma razón, sus efectos tienden, a largo plazo, a anularse mutuamente.

Los sesgos son controlados usualmente a través del diseño de investigación. Un buen protocolo, una adecuada implementación y un correcto seguimiento permiten controlar el efecto de tales sesgos. La experiencia del investigador, la evidencia recogida en estudios similares ya realizados, la supervisión de expertos permiten detectar muchos de los sesgos usualmente comunes. Ahora bien, nunca se tiene garantía de que algún sesgo pudiera influir en la investigación, y en el caso de que actuase afectaría a la validez de los datos recogidos impidiendo la correcta interpretación de los resultados (Iraurgi, 2000; Argimón y Jiménez, 2004).

En cuanto a los errores variables, además de los esfuerzos que puede hacer el investigador para minimizar la varianza de error a través del diseño de la investigación, una forma de afrontar dichos errores es mediante la inferencia estadística. Como hemos comentado más arriba, esta expresión se aplica a un conjunto de procedimientos utilizados para determinar el grado en que la relación observada entre dos variables puede explicarse como resultado del azar o, por el contrario, son producto del efecto del tratamiento investigado. Para esta determinación, históricamente se han venido realizando dos tipos de procedimientos (Doménech, 1998; Argimón y Jiménez, 2004): el basado en la prueba de significación estadística o prueba de la hipótesis nula, originaria de Fisher (1922), y la prueba de contraste de hipótesis de Neyman–Pearson (1928a, 1928b).

En el enfoque de Fisher se parte de la especificación a priori de una hipótesis nula, la cual plantea la no asociación entre dos variables; que para el caso de un estudio clínico en el que se comparan dos tratamientos, la hipótesis nula establece que no existen diferencias entre los tratamientos evaluados. Del mismo, en la prueba de significación estadística, Fisher propone un valor de 'p' como criterio utilizado para eva-

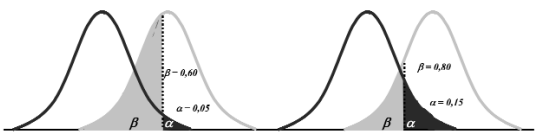
luar la hipótesis nula; es la probabilidad, bajo el supuesto de que la hipótesis nula es correcta, de encontrar un resultado igual o más extremo que el observado en el estudio —en el supuesto también de no hay fuentes de sesgo en la recolección de los datos o en el proceso de análisis—. En la prueba de significación estadística fisheriana valores de 'p' muy pequeños indican bajo grado de compatibilidad entre la hipótesis nula y los datos del estudio, por tanto este valor de 'p' es un índice que mide la fuerza de evidencia contra la hipótesis nula. Así, Fisher propuso que valores de 'p' menores a 0,05 fueran tomados como criterios de evidencia en contra de la hipótesis nula, pero no como criterio absoluto. Es más, un valor de 'p' alrededor de 0,05 no podría llevar ni al rechazo ni a la aceptación de la hipótesis nula, sino a la decisión de realizar otro experimento. Es decir, Fisher pretendía ofrecer un criterio objetivo a las conclusiones, pero sin pretender que fuera un método rígido.



Tipos de error al contrastar hipótesis

		Realidad	
		H_0 cierta	H_0 falsa
Resultado de la prueba estadística	No significativa (p> 0,05) Decisión: No se rechaza H_0	**Decisión Correcta (1-α)** Se falla que el tratamiento NO tiene efecto y en realidad NO lo tiene Verdadero Negativo	**Error Tipo II (β)** El tratamiento SÍ tiene efecto pero no lo percibimos. Falso Negativo
	Significativa (p≤ 0,05) Decisión: Se rechaza H_0 y se acepta H_1	**Error Tipo I (α)** El tratamiento no tiene efecto pero se decide que sí. Falso Positivo	**Decisión Correcta (1-β)** Se falla que el tratamiento SÍ tiene efecto y realmente SÍ lo tiene Verdadero Positivo

H_0 – Hipótesis Nula (No diferencias)
H_1 Hipótesis Alternativa (Existen diferencias)



Relación entre los errores tipo I (α) y II (β)
Si reducimos α aumenta β, y viceversa
Por lo general, α es elegido a priori, y resulta recomendable, a posteriori, el cálculo de la potencia (1 β)

Figura 1.- Inferencia Estadística: Contraste de hipótesis y tipos de error

Por su parte, en el enfoque de Neyman–Pearson —o contraste de hipótesis— se establecen unas reglas de decisión para interpretar los resultados del estudio con antelación a la realización de éste, y el resultado del análisis es sencillamente el rechazo o la aceptación de la hipótesis nula. La primera de ellas fija un punto de corte, llamado nivel de significación o nivel Alfa (α), usualmente de 0,05 ó 0,01, para juzgar al valor de 'p' y este criterio se usa para rechazar o no la hipótesis nula (si $p \leq 0,05$ se rechaza la hipótesis nula, si $p > 0,05$ se acepta). En contraste con el enfoque de Fisher, bajo la perspectiva de Neyman–Pearson los valores de 'p' no son interpretados, sino que el valor 'p' es valorado con respecto al nivel alfa preestablecido. Es decir, el nivel alfa se especifica en el diseño y planificación del estudio y el valor de 'p' es la cantidad derivada del estudio una vez concluidos y analizados los datos de éste. Por ello, el nivel de significación alfa es una de las incorporaciones distintivas de estos autores. Otra aportación interesante del enfoque de Neyman–Pearson es que, además de la hipótesis nula (H_0), se debe especificar también a priori una hipótesis alternativa (H_1) lo más precisa posible. Es decir, no basta con sólo señalar que no hay diferencias entre los tratamientos evaluados, sino que hay que indicar cuánto es mejor el tratamiento a prueba respecto al de comparación. El contraste de hipótesis planteado por Neyman–Pearson no establece la verdad de la hipótesis, sino un criterio que nos permite decidir si una hipótesis se acepta o se rechaza, o el determinar si las muestras observadas difieren significativamente de los resultados esperados. En este proceso podemos incurrir en dos tipos de errores según sea la situación real o falsa y la decisión que tomemos (Figura 1), siendo esta la tercera gran innovación de estos autores al proceso de inferencia estadística. Estos errores consisten en lo siguiente: 1) Si la hipótesis para probar (H_0) es realmente verdadera y ésta se rechaza de manera errónea a favor de la alternativa (H_1), es decir, no existen diferencias y se asume que sí las hay, se estaría cometiendo un error Tipo

I, lo cual supondría asumir un falso positivo; y 2) si la hipótesis para probar (H_0) es realmente falsa y ésta no se rechaza, es decir, existen diferencias pero se asume que no las hay, se cometería un error Tipo II, que implicaría una decisión falsa negativa. El riesgo de cometer el error tipo II se designa con una probabilidad: llamada probabilidad Beta (β), siendo su complemento ($1-\beta$) la llamada 'potencia' o poder del estudio, el cual se puede definir como la probabilidad que tiene un estudio para detectar diferencias entre tratamientos cuando realmente existen. En otras palabras, la potencia de un estudio es la probabilidad de rechazar la hipótesis nula cuando ésta es falsa, o lo que es lo mismo, la probabilidad de aceptar la hipótesis alternativa cuando ésta es cierta. Por otro lado, minimizar los errores no es una cuestión sencilla, un tipo suele ser más grave que otro y los intentos de disminuir uno suelen producir el aumento del otro. En la Tabla 2 se proponen algunas recomendaciones para disminuir ambos tipos de errores.

Estas serían las bases de ambas posiciones, enfrentadas por otra parte, y que aún hoy su debate está sin resolver. Un desarrollo histórico de la inferencia estadística y del papel que jugaron estos padres de la estadística es descrita de forma ejemplar por Rodríguez–Arias (2005) y al cual seguiremos en el apartado siguiente al abordar los pasos del ritual de las pruebas de significación estadística.

Pasos del ritual de la prueba de significación estadística

El texto que sigue en este apartado ha sido tomado de Rodríguez–Arias (2005) prácticamente de forma textual, en la convicción de que es una de las mejores descripciones del proceso de inferencia que conozco y que difícilmente podría mejorar. Discúlpenme esta licencia, sin duda abusiva, pero con ello quiero rendir tributo a su autor, al cual se debe citar sin duda cuando se quiera referenciar algo de lo que a continuación se detalle.

Rodríguez–Arias, plantea cuatro pasos en el ritual de la prueba de significación estadística; a saber:

«1. El investigador formula la hipótesis nula. En términos generales, la hipótesis nula afirma

TABLA 2

Recomendaciones para disminuir los errores de tipo I y II

Para el error de tipo I	Para el error de tipo II
• Disponer de una teoría que guíe la investigación, y permita definir a priori las hipótesis de investigación. • Realizar un adecuado diseño y planificación del estudio de forma que se evite 'ir de pesca' una vez obtenidos los datos. • Disminuir el número de pruebas estadísticas llevados a cabo en el estudio, solo las necesarias y mayormente guiadas por la hipótesis a priori • Depurar la base de datos para evitar errores de valores extremos que puedan alterar los hallazgos hacia niveles significativos. • Utilizar niveles de significación o valores de alfa más reducidos del convencional $p<0,05$ (0,01 ó 0,001). • Reproducir el estudio. Si al reproducir el estudio se obtienen resultados similares, estaremos más seguros de no estar cometiendo el error de tipo I.	• Incrementar el tamaño de la muestra. • Estimar el poder estadístico del estudio. • Incrementar el tamaño del efecto a detectar. • Incrementar el valor de alfa. • Utilizar pruebas estadísticas más robustas y potentes, las llamadas paramétricas (t de Student, F de Fisher, pruebas basadas en el Modelo Lineal General -MLG-), en lugar de las pruebas no paramétricas (McNemar, Friedman, Wilcoxson, etc.).

Adaptado de: Pita y Pértega, 2001; Manterola y Viviana, 2008

que no existe ninguna relación real o verdadera entre las variables independiente y dependiente de una investigación, y que, por tanto, si alguna relación es observada entre dichas variables en los datos de la investigación, la misma podría explicarse como resultado del azar. Es por eso que a la hipótesis nula se le llama la hipótesis del azar. Dicho de otra manera, la hipótesis nula expresa que si se repitiera la investigación un número suficiente de veces, siempre con una muestra distinta extraída aleatoriamente de la misma población, las diferencias en la variable dependiente entre los grupos de la investigación tenderían a neutralizarse y terminarían siendo cero. El razonamiento implícito en la hipótesis nula es el siguiente: suponiendo que el resultado de una investigación particular constituye una selección al azar de entre una multitud de resultados posibles, el investigador se pregunta cuál sería la probabilidad de obtener por azar la diferencia que él ha encontrado entre los grupos de su investigación. Si esa probabilidad es igual o menor que un nivel de probabilidad convencional previamente establecido ($p \leq 0,05$), entonces el investigador concluye que los resultados por él observados no se deben al azar y, por tanto, rechaza la hipótesis nula. Si, en cambio, la probabilidad de que la diferencia observada entre los grupos se pueda explicar como resultado del azar es superior al nivel de probabilidad convencional previamente establecido ($p > 0,05$), entonces no se puede descartar el azar, es decir, no se rechaza la hipótesis nula. Esta formulación es puramente fisheriana.»

«2. Es obvio que la decisión sobre la hipótesis nula requiere de que se haya establecido previamente un nivel de significación estadística, es decir, un criterio que sirva de base a la decisión de rechazar o no rechazar la hipótesis nula. Al establecer un criterio de decisión sobre la hipótesis nula, el investigador puede valorar los errores que podría cometer en su decisión sobre la hipótesis nula. Una primera forma de error (se conoce como el error tipo I) consiste en rechazar una hipótesis nula verdadera, es decir, descartar el azar como explicación cuando los resultados podrían explicarse razonablemente con base en el mismo. Este es el error que comete el investigador que ve más de lo que hay en los datos; es decir, el investigador concluye que existe una relación real o verdadera entre las variables independiente y dependiente de la investigación, cuando en realidad la relación observada se puede explicar razonablemente como resultado del azar. El llamado error tipo I es el error del investigador que se apresura a concluir a favor de su hipótesis de investigación. Fisher no habló de ningún otro error, pues la prueba de la hipótesis nula para él no era otra cosa que un freno a la tendencia natural de un investigador a creer que la hipótesis ha sido confirmada por el simple hecho de que los resultados de la investigación siguen la misma dirección de la hipótesis.»

«En la estrategia de Fisher sólo hay un error posible: rechazar una hipótesis nula verdadera. Una segunda forma de error (se conoce como el error tipo II), introducida por Egon Pearson y Jerzy Neyman consiste en no rechazar una hipótesis nula falsa, es decir, no descartar el azar aun cuando éste no constituye una explicación razonable de los datos. Este es el error que comete el investigador que ve menos que lo que hay en los datos; por miedo a rechazar incorrectamente el azar, el investigador puede exponerse al riesgo de pasar por alto una relación real o verdadera entre las variables de su investigación. Fueron Pearson y Neyman los que, al introducir un segundo tipo de error, bautizaron como error tipo I al error del que había hablado Fisher.»

«En la perspectiva fisheriana el nivel de significación estadística es el punto que separa las probabilidades que nos conducen a rechazar la posibilidad de que la relación observada entre las variables de una investigación se deba completamente a errores variables (errores de azar) de aquellas probabilidades que nos conducen a no rechazar esa posibilidad. Según Fisher, el nivel de significación estadística equivale a la magni-

tud del riesgo que está dispuesto a correr el investigador de cometer el error de rechazar una hipótesis nula verdadera (el llamado error tipo I). Para la mayoría de los propósitos, el nivel de significación previamente establecido suele ser de 0,05, aunque en áreas de investigación más rigurosas se trabaja con un nivel de significación de 0,01. Suponiendo que se trabaja con un nivel de significación de 0,05, se rechazaría la hipótesis nula siempre que la probabilidad de explicar los resultados obtenidos en una investigación, como si fueran obra del azar, sea igual o menor que 0,05.»

«En la perspectiva de Pearson y Neyman, para establecer el nivel de significación estadística habría que atender al impacto de cada tipo de error en el objetivo del investigador, y a partir de ahí se decidiría cuál de ellos es preferible minimizar. Pearson y Neyman llamaron alfa al error tipo I y beta al error tipo II; a partir de este último tipo de error, introdujeron el concepto de "poder de una prueba estadística", el cual se refiere a su capacidad para evitar el error tipo II, y está definido por 1–beta, y en estrecha relación con éste se ha desarrollado el concepto de "tamaño del efecto" que algunos han propuesto como sustituto de los valores p en los informes de investigación científica». (Cohen, 1990, 1994; Kraemer y Thiemann, 1987; Murphy y Myors, 2004).

«3. El tercer paso del llamado ritual de la prueba de significación estadística consiste en la elección de la prueba estadística que se utilizará para someter a prueba la hipótesis nula. Hay dos clases de pruebas estadísticas: las paramétricas y las no paramétricas. Se llama paramétricas a aquellas pruebas estadísticas que exigen que los datos a los que se aplican cumplan con los siguientes requisitos: que los valores de la variable dependiente sigan la distribución de la curva normal, por lo menos en la población a la que pertenezca la muestra en la que se hizo la investigación; que las varianzas de los grupos que se comparan en una variable dependiente sean aproximadamente iguales (homoscedasticidad, u homogeneidad de las varianzas); y que la variable dependiente esté medida en una escala que sea por lo menos de intervalo, aunque este último requisito no es compartido por todos los estadísticos (McGuigan, 1993; Siegel, 1956). Cuando los datos cumplen con los requisitos indicados, especialmente con los dos primeros, las pruebas estadísticas paramétricas exhiben su máximo poder, es decir, su máxima capacidad para detectar una relación real o verdadera entre dos variables, si es que la misma existe. Las pruebas paramétricas más conocidas y usadas son la prueba t de Student, la prueba F, llamada así en honor a Fisher, y el coeficiente de correlación de Pearson, simbolizado por r. Cuando estas pruebas estadísticas se aplican a datos que violan los dos primeros de los requisitos señalados, pierden parte de su poder. Las pruebas estadísticas no paramétricas, en cambio, no hacen a los datos ninguna de las exigencias que les hacen las pruebas estadísticas paramétricas, por eso se les denomina 'pruebas estadísticas libres de distribución'. Las más conocidas y usadas de estas pruebas son la Ji cuadrada de Pearson, la prueba de la probabilidad exacta de Fisher, los coeficientes de contingencia de Pearson y Cramer, la prueba U de Mann–Whitney, el coeficiente de correlación de rangos de Spearman, y el coeficiente de asociación ordinal de Goodman–Kruskal (coeficiente gamma), (Conover, 1999; Leach, 1979; Siegel, 1956). Todas estas pruebas poseen menos poder que las pruebas paramétricas correspondientes, pero han demostrado ser muy útiles como alternativas cuando no se considera apropiado el uso de pruebas paramétricas.»

«4. El último paso del ritual de la prueba de significación estadística consiste en comparar el valor arrojado por la prueba estadística aplicada a los datos con el valor que, en circunstancias comparables, puede ocurrir por azar con una probabilidad de 0,05 ó 0,01, según el valor de la probabilidad que se haya adoptado como nivel de significación estadística. Si, al consultar la tabla de los resultados de la prue-

ba estadística que pueden ocurrir por azar con diferentes niveles de probabilidad, se observa que el resultado de la investigación tiene una probabilidad de ocurrir por azar igual o menor que la probabilidad adoptada como nivel de significación estadística, entonces no se rechaza la hipótesis nula. Si, en cambio, el resultado de la investigación tiene una probabilidad de ocurrir por azar mayor que la probabilidad adoptada como nivel de significación estadística, entonces no se rechaza la hipótesis nula. Esto es todo cuanto diría Fisher al terminar la prueba de la hipótesis nula. Pearson y Neyman, en cambio, incorporaron la idea de simetría entre el rechazo y la confirmación de la hipótesis nula; es a partir de ellos que los libros de texto de estadística han incorporado la expresión 'se acepta la hipótesis nula', pues para Fisher sólo era posible rechazar o no rechazar la hipótesis nula».

Esta lógica de decisiones sobre la aceptación y/o rechazo de hipótesis o de la significación estadística es una de las bases de la inferencia estadística que articulan y justifican el uso de esta herramienta. Usualmente prescindimos de está lógica y tendemos a interpretar directamente los valores–p de las pruebas estadísticas haciendo atribuciones muchas veces inciertas. En la Tabla 3 se presentan algunos de los equívocos más usuales y de las precauciones que debiéramos tener en cuenta cuando interpretamos el resultado de una prueba de significación.

De forma sintética, y a sabiendas de que se trata de una simplificación, cabe concluir que las pruebas de significación estadística nos proporcionan un valor 'p' que nos permiten conocer la probabilidad de que nuestros resultados hayan sido producto del azar, o bien efecto de nuestra intervención. Obtener un valor de 'p' pequeño (inferior al nivel de significación elegido; $p \leq 0{,}05$ ó $p \leq 0{,}01$ ó $p \leq 0{,}001$) implica que existe una probabilidad pequeña —y por tanto asumible— de que los resultados obtenidos se deban al azar y en este caso admitimos que las diferencias o asociación entre las variables son reales (recordemos que podemos estar incurriendo en un error en la decisión). Ahora bien, que podamos asumir que los resultados encontrados se hayan producido verdaderamente, no

TABLA 3
Precauciones a la hora de interpretar el resultado de una prueba de significación

1.	El rechazo de la hipótesis nula no sugiere causalidad. Es un error grave asumir que una prueba estadísticamente significativa lleva asociado una relación de causa-efecto
2.	Un resultado no significativo no demuestra que la hipótesis nula sea cierta, simplemente advierte de que los datos no aportan suficientes pruebas para dudar de la credibilidad de la hipótesis nula.
3.	De otro modo, No significativo es equivalente a No demostrado o No concluyente, pero nunca a ausencia de relación causa-efecto. La relación causa efecto y la validez de un estudio dependen del diseño realizado, no de las pruebas estadísticas. La estadística no puede corregir las insuficiencias de un mal diseño.
4.	El resultado estadísticamente significativo no tiene nada que ver con la clínica. La expresión 'muy significativo' es un término estadístico que se utiliza para indicar que la hipótesis nula es poco creíble. La relevancia clínica tiene que ver más con la magnitud del efecto que con la significación estadística.
5.	Por tanto, la p no es una medida de la magnitud del efecto ni de la intensidad de la relación entre las variables, ni del grado de eficacia de un tratamiento.
6.	Por tanto, al interpretar los resultados de un estudio es importante no quedarse deslumbrado por el grado de significación estadística y aprender a valorar la importancia de los hallazgos y si estos son poco probables que se hallan producido por azar.

Adaptado de: Porta, Plasencia y Sanz, 1988; Rebasa, 2003

nos dicen nada de la magnitud del efecto logrado ni de su precisión (Porta, elt al, 1988; Pita–Fernández y Pertega, 2001; Rebasa, 2003).

A este respecto, desde hace ya tiempo se recomienda, con acierto, acompañar a los valores 'p' con el uso de los intervalos de confianza —e incluso sustituyendo a los propios valores 'p'—, ya que esta herramienta sí que nos aporta información sobre la magnitud y la precisión del efecto (Gardner y Altman, 1986; Clark, 2004). El Intervalo de Confianza (IC) construido a partir de una muestra, es un rango de valores mínimo y máximo entre los cuales esperamos que se encuentre el verdadero valor del parámetro que tratamos de estimar. En las distribuciones normales los intervalos de confianza se construyen sumando y restando al estimador del efecto (la media, la razón de riesgos, etc.) su error estándar [EE=DT/√n] multiplicado por el valor de z=1,96 para obtener intervalos de confianza del 95%, o por el valor de z= 2,58 si se pretende obtener un IC del 99%. Por tanto, la amplitud de los intervalos dependerá de la variabilidad o desviación estándar (numerador en la fórmula del EE) y de los efectivos utilizados en la muestra (denominador de la fórmula), de forma que si disminuye el numerador (menor variabilidad) o se incrementa el denominador (aumento del tamaño de la muestra) se reduce el error de medida y, por tanto, aumenta la precisión. Un intervalo de confianza del 95% quiere decir que si se toman 100 muestras de un mismo tamaño y se utiliza cada muestra para construir un IC del 95%, se podría esperar que en promedio 95 de los intervalos incluirían el verdadero efecto de la terapia y cinco no lo hicieran. Una de las características del IC es que existe una relación entre éste y la prueba de hipótesis: cuando el IC del 95% no contiene el valor '0' (en el caso de diferencias de medias) o el valor '1' (en el caso de la razón de riesgos) se presenta una diferencia estadísticamente significativa ($p<0,05$), mientras que si el IC contiene el '0', o el '1' según el caso, entonces no existirían efectos significativos estadísticamente ($p>0,05$). Por ello, el IC a demás de utilizarse como estimador de la magnitud y precisión, también es válido como medida de significación estadística.

Veámoslo con los ejemplos de la Tabla I. En el estudio 1 se ha obtenido una diferencia de proporciones del 20% que arroja un Riesgo Relativo (RR) de 1,55 a favor del grupo Exp respecto al Ctrol. La prueba de significación estadística a ofrecido una p=0,361 y el IC del 95% ha resultado de (0,56 a 4,29). Como vemos, este intervalo de confianza contiene el valor '1', lo cual quiere decir que en alguna de las muestras aleatorias que pudiéramos obtener ambos grupos presentarían la misma proporción de éxitos del tratamiento (numerador y denominador serían iguales). En el estudio 2 se ha observado la misma diferencia de proporciones (20%) y el mismo RR (1,55), pero en este caso el IC del 95% es de (1,13 a 2,14). Como podemos apreciar, el valor 1 esta fuera del recorrido observado, por lo que podríamos concluir que en el 95% de las muestras aleatorias encontraríamos RR de la magnitud del encontrado en nuestro estudio. Observemos como en este caso el valor de la prueba de significación es de $p<0,005$. También ha que destacar como el IC del 95% del Estudio 2 es más estrecho que el del estudio 1, es decir, resulta más preciso: en el Estudio 1 el verdadero valor de la RR en la población oscilaría entre valores de 0,56 a 4,29, mientras que en el Estudio 2 lo haría entre valores de 1,13 a 2,14. En nuestro ejemplo, esta mayor precisión se debe al tamaño de la muestra, 20 casos en el Estudio 1 y 200 en el Estudio 2.

Los intervalos de confianza, a diferencia de los valores 'p', no reducen los resultados a un simple 'blanco o negro', 'estadísticamente significativo o no significativo', sino que nos ofrecen un límite inferior y otro superior entre los cuales se sitúa el verdadero efecto en la población, es decir, nos ofrece una aproximación a la estimación del efecto. En el ámbito clínico, y en el científico en general, es preciso conocer si los resultados obtenidos en nuestra investigación pueden ser extrapolables a la

población con un riesgo mínimo de equivoco —con una probabilidad inferior a un 5%—, pero también nos interesa sobre manera cuál es la magnitud del efecto logrado o la importancia de un resultado.

Clínicamente significativo, relevante o importante

Ernest Rutherford, considerado el padre de la física nuclear, planteaba que "si tu experimento necesita estadística deberías haber hecho uno mejor". La cita no es un descrédito a las potencialidades de la estadística, sino que descubre dos cuestiones desde mi punto de vista fundamentales: 1) que el éxito de una investigación está en su diseño y planificación (Iraurgi, 2000), y 2) que la magnitud de los resultados deberían ser tan evidente que estuviera más allá de la más mínima probabilidad de error. Lo que buscamos al investigar, en definitiva, no es tanto la significación estadística sino la significación práctica (Kirt, 1996), o lo que Levy (1967) denominó significación sustantiva y que dentro del área asistencial se conoce como significación o relevancia clínica (Jacobson, Follette y Revenstorf, 1984; Kendall, Flannery y Ford, 1999).

Cuando se trata de analizar el denominado cambio terapéutico el efecto formulado en la hipótesis científica recoge el cambio hacia los valores de normalidad dentro del área clínica sometida a intervención, por ejemplo una reducción de la sintomatología depresiva hacia niveles no patológicos. Como hemos podido apreciar en apartados anteriores, en estos casos puede suceder que el cambio estadísticamente significativo no indique el verdadero valor terapéutico, es decir, como en el Estudio 4 del ejemplo (Tabla 1), el efecto logrado no alcance la diferencia clínicamente significativa. Entonces, ¿qué se entiende por clínicamente significativo?. Veamos algunas definiciones encontradas en la literatura:

- "La relevancia clínica de un fenómeno va más allá de cálculos aritméticos y está determinada por el juicio clínico. La relevancia depende de la magnitud de la diferencia, la gravedad del problema a investigar, la vulnerabilidad, la morbimortalidad generada por el mismo, su coste y por su frecuencia entre otros elementos" (Pita–Fernández y Pértega, 2001).
- "Se entiende por significación clínica la magnitud del cambio atribuida al tratamiento terapéutico que permite que el funcionamiento del sujeto pueda ser considerado normal" (Kendall, Flannery y Ford, 1999).
- Hollon y Flick sugieren que (1988) "la unidad mínima de importancia clínica debe ser definida en términos del más pequeño cambio fiable de algún interés, pero no necesariamente a todas las partes interesadas".
- El grupo de Lindgren proponen que "cuando se comparan dos métodos del tratamiento, la diferencia más pequeña entre las terapias con respecto a una variable del resultado importante que llevaría a la decisión de modificar el tratamiento denotaría un resultado de importancia clínica" (Lindgren, Wielinsky, Finkelstein y Warwick, 1993).
- LeFort plantea que el término 'clínicamente significativo' refleja "la magnitud de cambio, si verdaderamente el cambio representa una diferencia real para conservar la vida, una duración en el tiempo de los efectos, la aceptación del paciente, la relación costo–efectividad y la facilidad de aplicación".
- Kingman (1992) propone que la significación estadística deviene en una condición necesaria para la importancia clínica y que ambas, la importancia estadística y la clínica, deben coincidir. Para ello, se requiere de una definición de lo que se entiende por importancia clínica, y para lograr este requisito es necesario un consenso entre expertos reconocidos.
- Killoy (2002) sugiere que la significación clínica es una valoración subjetiva realizada por el clínico y que antes de que pueda hallarse un hallazgo clínicamente significativo, debería haberse logrado la significación estadística.

Varias ideas afloran en el conjunto de estas propuestas. Siguiendo a Greenstein (2003), la importancia clínica de los datos necesita ser interpretada por el médico antes de tomar las decisiones terapéuticas. Sin embargo, no existe una manera precisa de definir la relevancia clínica ya que es cada situación la que especificará cuán pequeño ha de ser la mejora mínima necesaria. Por consiguiente, teniendo en la cuenta las definiciones anteriores, se podría plantear la siguiente propuesta sobre lo que comporta un cambio clínicamente significativo o relevante: "la importancia clínica implica la existencia de un cambio que puede influir en la decisión de un clínico sobre cómo tratar a un paciente. Para llegar a la conclusión de que un resultado es clínicamente relevante, el hallazgo ha de ser, simultáneamente, clínicamente y estadísticamente significativo. No obstante, este criterio supone un juicio de valor que varia en función de la situación clínica de cada caso. Médicos, pacientes, investigadores, representantes de la salud pública, la industria farmacéutica, y otros protagonistas del escenario socio–sanitario (Tabla 4) pueden interpretar la relevancia clínica de forma diferente, en tanto que cada uno de ellos pueden poner su(s) objetivo(s) de resultado(s) en opciones diversas (p.ej.: el tamaño de efecto, el alivio de la dolencia, los costes, la duración del tratamiento, la comodidad de la implementación, el mantenimiento de la mejora de salud y aceptación del tratamiento por el paciente, etc.). De este modo, un resultado puede ser estadística y clínicamente significativo pero tener poca relevancia médica porque el beneficio no supera el riesgo o el coste del tratamiento o porque el beneficio sólo se observa en un grupo de pacientes muy pequeño.

TABLA 4
Grupos de interés en la definición de lo 'Clínicamente Significativo o Relevante' en función de sus intereses

Clínico	Necesita relacionar los indicadores de monitorización clínica con los objetivos de la terapia. También puede estar interesado en valorar el tamaño de los efectos, el tiempo necesario para la terapia, la facilidad de aplicación del tratamiento, los efectos adversos, la duración de la mejoría lograda por el tratamiento, el coste económico, la aceptación del tratamiento por los pacientes, la satisfacción con el tratamiento.
Enfermo	Desde la perspectiva del paciente, el éxito del tratamiento, esto es, 'curarse', sería lo principal. La reducción de síntomas específicos que pudieran interesar al clínico no son, sin embargo, demasiado importantes para el paciente. Para éste, un resultado clínico tendría que ver con la reducción de la dolencia, del malestar físico o emocional, de la ausencia de efectos colaterales, de los cambios en su calidad de vida y satisfacción.
Investigador	Le interesa probar que los cambios producidos en los estudios de investigación son de magnitud importante y que no se deben a la casualidad, es decir, puedan admitirse con fundamentación estadística. Logros pequeños, no obstante, podrían considerarse como hipótesis etiológicas de nuevos estudios. Ahora bien, para poder determinar un investigador la magnitud de la significación clínica, previamente debe determinarse el propósito global del estudio, su planificación, el tamaño del efecto mínimo a lograr con el tratamiento en un paciente tipo.
Agencias Sanitarias	Son organismos nacionales o supranacionales que se interesan por la seguridad y efectividad de los tratamientos y los protocolos terapéuticos. Supervisan los resultados de los ensayos clínicos para la aprobación de nuevos tratamientos, y controlan los efectos colaterales de los ya aprobados para restringirlos si fuera necesario.
Sistemas de Dispensación de Salud	Su objetivo es proporcionar el acceso adecuado y dispensación de calidad de unos servicios de salud a un costo razonable. Dependiendo de la política sanitaria de un país, velarán por la equidad y universalidad de los servicios de salud ofreciendo las terapias necesarias a las necesidades específicas, si bien restringirá determinados servicios que serán bajo pago particular. Sus decisiones se basan en los análisis de coste-efectividad.
Industria Farmacéutica	Son empresas que compiten por un mercado y su interés es ofrecer buenos productos. Hacen importantes inversiones en investigación, por lo que se adscribirían a lo planteado en el epígrafe del investigador, pero también hacen importantes esfuerzos en marketing.
Mutuas y Seguros	En la medida que cubren los gastos devenidos de una baja por enfermedad y su tratamiento, están interesados en terapias eficaces que reduzcan el tiempo de la dolencia y su capacidad para prevenir nuevos episodios de recidiva. Les interesa la protocolización de los procesos terapéuticos para ajustar sus prestaciones al curso esperado de una determinada enfermedad.

Adaptado de: Greenstein, 2003

Conclusiones

Varias son las conclusiones básicas que quisiéramos destacar de esta revisión:

1. La estadística debe entenderse como una humilde herramienta para ayudarnos en la comprensión, organización o estructuración de los datos de la realidad obtenidos en nuestros estudios. Es importante no dejarse seducir por ella hasta el punto de convertirla en la base de la investigación. Investigar consiste en conocer bien las bases teóricas que fundamentan el estudio, realizar un buen diseño y una correcta planificación y seguimiento, la utilización adecuada de la estadística para el análisis de los datos, y la competencia intelectual para asociar los resultados a las evidencias existentes y a los retos de investigación futuros. La estadística, por tanto, es un medio, no un fin.
2. La utilización de la estadística en general, y de la inferencia, en particular, han de entenderse y utilizarse dentro del marco de sus competencias: la estadística nos informa del nivel de probabilidad que el estudio nos ofrece de que el efecto encontrado se haya producido por casualidad. Si la probabilidad hallada, esto es el valor–p, es menor que el que a priori se ha establecido como criterio (puede ser del 0,05, ó bien del 0,01, o incluso podría llegar a plantearse el del 0,1), entonces podremos asumir que los hallazgos son efectos de nuestra intervención (asumiendo que nos podemos equivocar en un 5%, 1% ó 10% de las veces, respectivamente).
3. Pero no olvidemos que el objetivo de la investigación no es sólo verificar la plausibilidad de nuestros hallazgos, esto es, encontrar que los efectos son estadísticamente significativos, sino también determinar la magnitud del factor, evento o relación objeto de estudio. Debemos, por tanto, diferenciar claramente lo que es estadísticamente significativo y lo que es clínicamente relevante, pues la conjunción de ambas cuestiones son las que dan sentido a la investigación. Un resultado estadísticamente significativo sin relevancia clínica no deja de ser una anécdota; un hallazgo clínicamente importante sin significación clínica no puede ser asumido como concluyente, ya que no podemos atribuir con seguridad el hallazgo clínico a la intervención realizada.
4. La significación, relevancia o importancia clínica o terapéutica de una intervención, o si se prefiere, la significación práctica o sustantiva de un efecto, supone el establecimiento a priori de un criterio que puede venir dado por la evidencia recogida en estudios previos sobre el área de interés, o bien por juicio arbitrario de expertos. De este modo, no existen criterios fijos y en su definición sería conveniente tener en cuenta los intereses de todos los protagonistas que directa o indirectamente se hallasen implicados en los objetivos del estudio. Solo a nivel practico, y tomando como referencia el ámbito clínico, propondríamos el siguiente principio general: "efecto clínicamente significativo es aquel efecto mínimo entre subgrupos de pacientes a partir del cual merece la pena modificar una actitud terapéutica, diagnóstica o preventiva a favor de la que se haya mostrado más beneficiosa".

En definitiva, se hace necesario evitar el uso indiscriminado de las pruebas de significación y, cuando éstas se utilicen, deberían complementarse acompañándolas con procedimientos estadísticos que informen acerca de la magnitud, la dirección y la importancia real de los resultados obtenidos con las pruebas de significación. Estos procedimientos deben pasar por la aplicación de índices del tamaño del efecto, ya que son capaces de proporcionar información acerca de la relevancia clínica, social, práctica o sustancial de un resultado empírico. Repasar, describir e interpretar estos índices del tamaño del efecto será el objetivo del siguiente artículo sobre esta serie que acabamos de iniciar sobre la evaluación de resultados terapéuticos.

Contacto:
Ioseba Iraurgi
MAPS Rekalde. Villabaso 24, bajo. Bilbao
IRAURGI@telefonica.net

BIBLIOGRAFÍA

- Argimon JM y Jiménez-Villa J (2004). Métodos de investigación clínica y epidemiológica. Tercera edición. Barcelona: Elsevier. Accesible en: http://books.google.es/books?id=_BLemLvp9XAC&pg=PA257&lpg=PA257&dq=clinicamente+significativo+relevante&source=web&ots=k6vBs5TQCL&sig=u32SczY0z6CCIRa9zS2CMQJxbn0&hl=es&sa=X&oi=book_result&resnum=9&ct=result#PPP1,M1
- Barrera M. (2008). Diferencias estadísticamente significativas vs relevancia clínica. *Rev CES Med*; 22 (1): 89-96. Accesible en: www.ces.edu.co/Descargas/CES%20Diferencias%20estadisticamente%20 significativas%20Vol22N1.pdf
- Cohen J. (1990). Things I have learned (so far). *American Psychologist*; 45(12): 1304-1312. Traducción: Cohen J. (1992). Cosas que he aprendido (hasta ahora). Anales de Psicología, 8, 1-2, 3-17. Accesible en: http://www.um.es/analesps/v08/02-08.pdf
- Cohen J. (1994). The earth is round (p<.05). *American Psychologist;* 49(12): 997-1003.
- Conover WJ. (1999). *Practical Nomparametric Statistics* (3rd Ed.). New York: John Wiley & Sons, Inc.
- Clark ML. (2004). Los valores de P y los intervalos de confianza. *Rev Panam Salud Publica*; 15(5): 293-296. Accesible en: http://www.scielosp.org/pdf/rpsp/v15n5/21999.pdf
- Domènech JM. (1998). Comprobación de hipótesis. Pruebas de significación y pruebas de hipótesis. En Doménech JM. *Métodos estadísticos en ciencias de la salud. Unidad didáctica 6*. Barcelona: Signo.
- Fisher RA. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London*; 222A: 309-368.
- Frías MD, Pascual J, García JF. (2002). La hipótesis nula y la significación práctica. *Metodología de las Ciencias del Comportamiento*; 181-185. Accesible en: http://www.uv.es/garpe/C_/A_/C_A_0020.pdf
- Gardner MJ, Altman DG. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*; 292: 746-750. Accesible en: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1339793
- Greenstein G. (2003). Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *JADA*; 134: 583-591. Accesible en: http://jada.ada.org/cgi/content/full/134/5/583
- Guttman L. (1977). What is not what statistics. The Statistician 26, 81-107. Traducido al castellano: Guttman L. (1977). Malos usos en estadística, *REIS*; 6: 101-117. Accesible en: www.dialnet.unirioja.es/servler/ fichero_articulo?codigo=665680&orden=80913
- Hollon SD, Flick SN. (1988). On the meaning if clinical significance. *Behav Assess*; 10: 197-206.
- Iraurgi I. (2000). Cuestiones metodológicas en la evaluación de programas terapéuticos. *Trastornos Adictivos*; 2(2): 99-113. Accesible en: http://db2.doyma.es/pdf/182/182v2n2a10017604pdf001.pdf.
- Jacobson NS, Follette WC, Revenstorf D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Journal of Consulting and Clinical Psychology*; 15: 336-352.
- Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application and alternatives. *Journal of Consulting and Clinical Psychology*; 67(3): 300-307.
- Kendall PC, Grove WM. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment*; 10: 147-158.
- Kendall PC, Flannery-Schroeder EC, Ford JD. (1999). Therapy outcome research methods. En PC Kendall, JN Butcher y GN Holmbeck (eds.). *Handbook of research methods in clinical psychology.* New York: Wiley and Sons.
- Killoy WJ. (2002). The clinical significance of local chemotherapies. *J Clin Periodontol*; Supplement 2: 22-29.
- Kingman A. (1992). Statistical vs clinical significance in product testing: can they be designed to satisfy equivalence?. *J Public Health Dent*; 52: 353-360.
- Kirk RE. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*; 56: 746-759.
- Kraemer HC, Thiesman S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications, Inc.
- Leach C. (1979). Introduction to Statistics: *A Nomparametric Approach for Social Sciences*. New York: John Wile y & Sons, Inc.
- LeFort SM. (1993). The statistical versus clinical significance debate. *Image J Nurs Surg*; 25: 57-62.
- Levy P. (1967). Substantive significance of significant differences between two groups. *Psychological Bulletin*, 67, 37-40.
- Lindgren BR, Wielinskyi CL, Finkelstein SM, Warwick WJ. (1993). Contrasting clinical and statistical significance within the research setting. *Pediatr Pulmonol*; 16: 336-340.
- McGuigan FJ. (1993). Experimental Psychology: *Methods of Research* (6th Ed.). New York: Prentice-Hall.
- Murphy KR, Myors B. (2004). *Statistical Power Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Neyman J, Pearson ES. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference (Part I). *Biométrica*; 20A: 175-240.
- Neyman J, Pearson ES. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference (Part II). *Biométrica*; 20A: 263-294.
- Pita-Fernández S, Pértega S. (2001). Significación estadística y relevancia clínica. *Cad Aten Primaria*, 8: 191-195. Accesible en: http://www.fisterra.com/mbe/investiga/signi_estadi/signi_estadi.asp
- Porta M, Plasencia A, Sanz F. (1998). La calidad de la información clínica (y III): ¿estadísticamente significativo o clínicamente importante? *Medicina Clínica*; 90: 463-468. Accesible en: http://clon.uab.es/recursos/listatipos.asp?tipo=PDF&page=82
- Rebasa P. (2003). Entendiendo la 'p<0,001'. *Cir Esp*; 73(6): 361-365. Accesible en: www.aecirujanos.es/ revisiones_cirugia/2003/Junio2.pdf
- Rodríguez-Arias E. (2005). Estadística y psicología: análisis histórico de la inferencia estadística. *Perspectivas Psicológicas*; 5: 96-102. Accesible en: www.psicologiacientifica.com/publicaciones/ biblioteca/articulos/ar-rodriguez01.htm
- Shaver, W.D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education*; 61: 383-387.
- Siegel S. (1956). *Nomparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill Book Company.
- Silva-AyÇaguer LC (1997). *Cultura estadística e investigación científica en el campo de la salud:* Una mirada crítica. Madrid: Díaz de Santos. Accesible en: http://books.google.com/books?id=hi7pxRZGa4C&pg=PA284&lpg=PA284&dq=What+is+not+what+statistics+Guttman+1977&source=web&ots=1x5ZR04H9S&sig=COVIWprHlpgxr9-NoUHQJStL6Xo&hl=es&sa=X&oi=book_result&resnum= 3&ct=result#PPR11, M1
- Thompson B, Snyder PA. (1997). Statistical significance testing practices in the Journal of Experimental Education. *Journal of Experimental Education*; 66: 75-83.