Metodología de Encuestas
Vol 1, Num 1, 1999, 69-84

COMENTARIOS Y RESPUESTA A "LA PRUEBA DE SIGNIFICACIÓN DE LA «HIPÓTESIS CERO» EN LAS INVESTIGACIONES POR ENCUESTA" DE VISO M. ARES

África Borges del Rosal

Vicente Manzano Arrondo

Alfonso Sánchez Bruno

África Borges del Rosal — Universidad de la Laguna

Para empezar, me ha parecido un verdadero hallazgo el ejercicio de simulación que presenta. Evidentemente, existe el error de que las muestras cuanto más grandes mejor, olvidando que el aspecto fundamental de una muestra es que haya sido extraída aleatoriamente y sea representativa de la población a la que pertenece. Con el ejemplo se muestra de forma bien clara como la relación más absurda entre variables puede resultar significativa siempre que se obtenga un número suficientemente grande de observaciones.

Me parece muy interesante el índice *nu* sugerido. No obstante, va a resultar un poco difícil que los investigadores se atengan a usarlo. Fundamentalmente, la dificultad del uso de otro índice (como es el caso del sugerido v) estriba en el cambio de hábitos. Los investigadores suelen ser reacios a ello. La prueba está en que desde la literatura se sigue insistiendo en la conveniencia de incluir cuestiones tales como tamaño del efecto, intervalos de confianza (temas que, si bien no con demasiada profundidad, forman parte de los temarios básicos de las asignaturas de análisis de datos) y resulta arduo que se ponga en práctica

Echo en falta en el trabajo algo que yo considero fundamental. Actualmente, después de décadas de "dejar hablar al dato", existe una proliferación tal de produc-

ción de trabajos científicos (Yela, 1991), que cabría asumir que tenemos suficiente información como para saber no sólo qué variables se relacionan entre sí, sino, incluso, en muchas ocasiones, cuál será la cuantía esperable de tal relación. Entonces, lo que se debe pedir a los investigadores es la realización de estudios más sólidos, como forma de superar las incongruencias derivadas de las dificultades que conllevan los contrastes de hipótesis. Ello supone dos tipos de actuaciones. De una parte, un mayor peso a los valores obtenidos en las investigaciones, es decir, dar un mayor énfasis a la cuantificación. Se ha propuesto insistentemente aspectos como incluir intervalos de confianza (Serlin, 1987; Cohen, 1990, 1994), o cuantificaciones del tamaño del efecto (Cohen, 1988).

También resulta interesante, frente a la hipótesis cero, el procedimiento propuesto por Serlin y Laspley (1985) del principio suficientemente bueno. Básicamente: una H_0 afirma que el valor de un determinado parámetro δ es δ_0. Ahora bien, aún en el caso de ser cierta H_0, dada la inexactitud de los procedimientos experimentales en general, obtendremos un valor muestral δ^* que diferirá de δ_0. Dado un tamaño muestral lo suficientemente grande, dicha diferencia (espúrea) se hará significativa. Lo que este procedimiento sugiere es incluir en la hipótesis nula unos márgenes que determinen *a priori* la magnitud de error Δ aceptable; así, el intervalo $\delta_0 \pm \Delta$ sería el intervalo de valores "suficientemente buenos" y H_0 quedaría expresada como $|\delta - \delta_0| \leq \Delta$. Un posible procedimiento de contraste estadístico de esta hipótesis se puede encontrar en el artículo original de Serlin y Lapsley (1985).

Por otra parte, creo que hay que insistir en la importancia de la teoría a la hora de dirigir las investigaciones. Muchas voces abogan por una realización de un quehacer científico más sólido, con menor interés en estudios puntuales, enfatizando más en la profundización en temas de interés (Rosnow y Rosenthal, 1989; Serlin, 1987; Gigerenzer, 1993; Yela, 1994). Parece especialmente oportuna la cita de Bakan que incluí en mi trabajo sobre el contraste de hipótesis (Borges, 1997), y que, pese a que no es justamente reciente, no deja de tener vigencia (yo diría que más cada día): "Cuando llegamos a un punto en que los procedimientos estadísticos se convierten en sustitutos del pensamiento en lugar de ayudas a él, ha llegado el momento de regresar al camino del sentido común" (Bakan, 1966).

Por todo ello, creo que una revisión congruente de la literatura sobre el tema puede llevar a que el investigador, en la planificación de su estudio, sepa qué variables se encuentran relacionadas, en el caso concreto que nos ocupa de estudios correlacionales, y también los valores que se esperan obtener, basándose en los trabajos previos. Por tanto, el procedimiento del índice *nu*, desde mi punto de vista, estaría especialmente indicado para trabajos exploratorios, donde no se pudiera establecer de antemano el interés que pueda tener el resultado obtenido.

Vicente Manzano Arrondo Universidad de Sevilla

Si bien considero que el trabajo del profesor Ares tiene un interés general, mis comentarios se circunscriben a su propuesta sobre el índice *nu*, ya que constituye el aspecto más innovador del conjunto y centra en un artilugio útil el esfuerzo de la obra.

Comparto la opinión de que existe una clara separación entre los trabajos de los profesionales de la metodología de investigación y el quehacer cotidiano de los investigadores aplicados. Sin embargo, creo que se trata de algo deseable. El tiempo que transcurre entre la génesis de la idea y su aplicación generalizada sirve de colchón que amortigua las pasiones del creador y permite que se establezca una discusión interna entre especialistas en metodología. Durante un tiempo, además, los primeros ensayos en la práctica muestran el nivel de adecuación de las nuevas herramientas. Cuando finalmente los útiles llegan en masa a los investigadores aplicados las creaciones iniciales han sufrido modificaciones y se han estudiado sus limitaciones y poder fáctico. Esta reflexión no tiene por objetivo restar importancia a las hipótesis de comodidad ni minimizar el perjuicio que causan las investigaciones *descuidadas* que no se adecuan a los principios metodológicos. Me mueve la intención de mostrar la otra cara de la moneda: la resistencia para adoptar nuevas estrategias permite evaluar éstas con más dedicación.

Así pues, ignoro la repercusión práctica que puede tener el índice ν o cualquier otra concreción de la idea que lo sustenta, pero aplaudo la iniciativa y la percibo como una alternativa ingeniosa que puede ayudar a tomar decisiones más aceptables.

Sin embargo, desde el punto de vista práctico no espero que ν tenga un efecto inmediato. Otros artilugios con objetivos parcialmente solapados con la idea que sustenta a *nu*, como los tamaños de efecto, llevan más tiempo en discusión y cuentan con el respaldo de los especialistas en metodología. Los sistemas informatizados de análisis de datos, cuando se hacen eco de estas herramientas, las arrinconan en módulos u opciones independientes. Para potenciar el uso de ν es imprescindible que el programa de ordenador suministre su valor automáticamente o *por defecto*, a la vez que incluso lleve a tomar decisiones como ocurre actualmente con el grado de significación. Y esto no ocurrirá (si lo hace alguna vez) hasta que exista un clima en extremo favorable, lo que no tiene lugar ni con rapidez ni con frecuencia.

En esta línea, quizá sería recomendable estudiar el comportamiento de ν tras restar importancia al elemento *potencia*. La idea central de ν es considerar la potencia de la prueba para interpretar el valor de significación y evaluar su importancia. Una alternativa más tímida a este objetivo es recurrir a la misma estrategia pero minimizando el papel de β, sustituyendo este elemento por β^m, con lo que la expresión quedaría como sigue:

$$\nu = \left[\left(1 - \frac{p\left(D/H_0\right)}{\beta^m}\right)^k - \frac{1}{2}\right]^h + \frac{1}{2} \qquad (1)$$

Dado que $0 \leq (1-\beta) \leq 1$, entonces

$$\lim_{m \to 0} \beta^m = 1 \qquad \lim_{m \to \infty} \beta^m = 0$$

Luego, para minimizar la importancia de la potencia debe escogerse un valor de $m < 1$ que cumpla además la condición:

$$\beta^m > p\left(D/H_0\right) \quad \Rightarrow \quad m < \frac{\log p\left(D/H_0\right)}{\log \beta} \qquad (2)$$

Cabe esperar, no obstante, que la condición expresada en (2) se cumpla siempre, puesto que es habitual que $\beta > p(D/H_0)$ y ocurre que si $m<1$ entonces $\beta^m > \beta$, con lo que

$$\beta^m > \beta > p\left(D/H_0\right)$$

Tabla 1: *Valores de k y h asociados a m*

m	k	h
,1	11,45200539	1,40741503
,2	9,69687462	1,40939450
,3	8,20244884	1,41173875
,4	6,92993641	1,41452587
,5	5,84627581	1,41784108
,6	4,92334652	1,42179775
,7	4,13616936	1,42653060
,8	3,46731067	1,43221700
,9	2,89635205	1,43907893
1,0	2.40942097	1.44740665

La tabla 1 muestra los efectos en los valores de *k* y de *h* según *m* para que se siga cumpliendo que con β=0,2 y α=0,01 ó 0,05 v' siga suministrando 75 ó 50 respectivamente. La tabla 2 muestra una alternativa a la tabla 4 de Ares, bajo el supuesto de tomar *m*=0,01. Se ha tomado un valor intencionadamente bajo para *m* con el objeto de mostrar el efecto: se ha disminuido la importancia de β de tal manera que se minimizan las distancias entre filas.

Tabla 2: *Valores de significación relativa en el caso de m=0,01*

		Grado de significación $p(D/H_0)$						
		0,1	*0,075*	*0,05*	*0,025*	*0,01*	*0,005*	*0,001*
Error tipo II (β)	*0.05*	34,59	42,62	49,94	60,91	74,84	80,89	86,30
	0,075	34,71	42,72	49,96	60,99	74,89	80,92	86,31
	0,1	34,80	42,79	49,98	61,04	74,92	80,94	86,31
	0,2	35,00	42,97	50,00	61,17	75,00	80,98	86,32
	0,3	35,12	43,07	50,01	61,24	75,05	81,01	86,33
	0,4	35,21	43,14	50,02	61,30	75,08	81,02	86,33
	0,5	35,27	43,19	50,03	61,34	75,10	81,04	86,33

Alfonso Sánchez Bruno Universidad de la Laguna

Creo que la primera parte resume bastante bien las críticas que se hacen en este momento a los contrastes de significación, si bien se olvida de la que, desde mi punto de vista, resulta fundamental: es una técnica poco informativa, ya que en el mejor de los casos nos informa de relaciones ordinales entre los parámetros, mientras que su alternativa, la determinación de intervalos de confianza, nos aporta toda la información del contraste de hipótesis y, además, nos permite estimar los verdaderos valores de los parámetros.

Enlazando con lo anterior, debo dejar claro, por si aún no lo estuviese, que creo que existen muy pocos casos en los que esté justificado llevar a cabo un contraste de significación, pese a lo cual creo que es fundamental comprender perfectamente todas sus características.

Desgraciadamente, creo que en la segunda parte del trabajo el autor cae en muchos de los errores que critica en la primera y, en mi opinión, lo hace por una deficiente comprensión de lo que son las probabilidades de error de tipo I y II. Creo que el siguiente párrafo, extraído del artículo, sirve perfectamente como ejemplo:

"...En definitiva, pues, en la medida en que la prueba sea más potente será más fácil rechazar la hipótesis nula que, ya sabemos, es falsa y, por tanto, el grado de significación perderá valor por sí mismo. En otros términos, $p(D/H_0)$ no sólo recoge la fuerza de la relación, sino también la influencia de la potencia de la prueba"

Para empezar, la potencia de la prueba no hace más "fácil" el rechazo de la hipótesis nula falsa, sino más "probable". Esto no es una mera disquisición gratuita, sino que en mi opinión puede llevar a errores de bulto. La probabilidad en el marco del contraste de hipótesis tiene una interpretación puramente frecuentista y, por consiguiente, una potencia de, por ejemplo, 0,70, nos dice única y exclusivamente que, como promedio, rechazaremos dicha hipótesis nula (falsa) el 70% de las veces que llevemos a cabo el contraste (evidentemente, con muestras diferentes). Los términos "fácil" o "difícil", por contra, son inexactos y carentes de interpretación objetiva.

El grado de significación, no tiene ningún valor por sí mismo más que el de ser comparado con el nivel α seleccionado previamente. Esta deficiente interpretación de $p(D/H_0)$ es uno de los errores más frecuentes en el uso de los contrastes de hipótesis, tal como se ha puesto de manifiesto en muchas publicaciones[1] y tal como dice el propio autor de este artículo unas páginas antes.

Por otra parte, $p(D/H_0)$ no recoge la fuerza de la relación[2], sino sólo la probabilidad de obtener unos resultados tan discrepantes como los que se han obtenido, bajo el supuesto de que la hipótesis nula es cierta. El mismo autor en párrafos anteriores cita a Cohen (1994) para aclarar que es posible obtener un valor bajo de $p(D/H_0)$ siendo alto el valor de $p(H_0/D)$. En cualquier caso, no olvidemos que $p(D/H_0)$ se ha obtenido a partir de una muestra, por lo que para extraer de él inferencias acerca de la relación entre variables poblacionales, tendríamos que deducir de alguna forma su distribución muestral y llevar a cabo un nuevo contraste o una estimación de intervalo de confianza, proceso que podría repetirse eternamente.

Finalmente, $p(D/H_0)$ simplemente no depende de la potencia de la prueba, sino de la distribución muestral del estadístico de contraste. Otra cosa es que D pueda verse afectado, al igual que la potencia, por el tamaño del efecto, pero eso tampoco creo que nos conduzca a nada y, en cualquier caso, no es este el camino seguido por el autor.

En cuanto al nuevo índice propuesto, una vez más, se trata de un índice obtenido a partir de unos datos muestrales ¿Cuál es su distribución muestral?.

---

[1] ver, por ejemplo, Borges (1997), a la que el propio autor cita unas páginas antes.

[2] lo que, en el ejemplo que se usa en el artículo, hace el Coeficiente de Correlación de Pearson.

En cuanto a su interpretación, se habla de "resultados poco significativos", grados de significación "altamente meritorios", etc. ¿Qué relación guarda esto con la verdad o falsedad de la hipótesis estadística o con la decisión a tomar? ¿Está defendiendo el autor una alternativa a los criterios de decisión habituales?

Finalmente, todo podría disculparse si el nuevo índice fuese útil para algo, pero la finalidad es, al parecer, obtener una medida estándar del grado de asociación entre variables. ¿No tenemos ya los diferentes índices de efecto y en el caso que nos ocupa el coeficiente de correlación de Pearson?. ¿Que es necesario apoyarse en una teoría previa para interpretarlos adecuadamente y eso los hace difíciles de usar? ¡Claro! La Ciencia es difícil de construir y supone pequeños pasos o, en palabras de Ramón y Cajal, un 10% de inspiración y un 90% de transpiración.

Resumiendo, creo que el índice propuesto es innecesario, confuso en su interpretación y basado en una interpretación errónea de las diferentes probabilidades implicadas en la metodología del contraste de hipótesis.

Respuesta: Viso M. Ares

En primer lugar, agradezco sinceramente los comentarios que ha recibido mi trabajo no sólo por su contenido, sino también por la oportunidad que representan para aclarar algunos aspectos que, precisamente por haber sido tocados, muestran cierta importancia. Para facilitar la lectura de mi respuesta, organizaré ésta no en función de la procedencia de los comentarios sino de su contenido.

Algunas cuestiones hacen referencia a la suficiencia o independencia de los artilugios del análisis de datos para recoger información específica. En este sentido, se ha mencionado el tamaño de la muestra, el tamaño del efecto y el grado de significación. Otros aspectos se refieren a la dicotomía entre las soluciones ideales y las soluciones prácticas, frecuentemente incompatibles. Y, por último, otros comentarios se han centrado directamente en el índice *nu*. Por este mismo orden serán abordados a continuación. Algunos comentarios se justifican como argumentos en contra del uso de la PSHN y a favor del tamaño del efecto y los intervalos de confianza. Considerando que las soluciones perfectas no existen, es comprensible encontrar también críticas a estas alternativas (por ejemplo, Lipsey, 1990; Frick, 1996; o Cortina y Dunlap, 1997). No es objetivo ni de estas respuestas ni del trabajo original el entrar en esta discusión que ya de por sí genera una gran cantidad de páginas de texto y de vivas discusiones. Sólo mencionaré algún aspecto relacionado cuando las respuestas a algunos comentarios así lo exijan.

En relación a la suficiencia[3] e independencia de la medidas estadísticas

Un problema constante en estadística es encontrar índices que sean suficientes para mostrar una información específica. Este ideal es difícil de conseguir y, como consecuencia, todas las medidas adolecen de algo. Un ejemplo básico lo constituye la media aritmética: sin al menos la presencia de una medida de variación no podemos interpretar la bondad de la media aritmética como representación del conjunto. Otro aspecto, muy ligado al anterior, es que las informaciones utilizadas por los índices se solapan entre sí. De esta forma, tamaño de muestra, tamaño de efecto, potencia de la prueba y grado de significación se encuentran indirectamente relacionados.

1. El principio "el tamaño de la muestra cuanto más grande mejor" es cierto en términos generales. Pero el índice *n* es insuficiente. Como señala la profesora Borges, lo importante no es el tamaño sino la representatividad de la muestra. Sin embargo, el concepto de muestra representativa es harto escurridizo (Chou, 1972; Kruskal y Mosteller, 1979a, 1979b, 1979c, 1980; Silva, 1993; Martínez, 1995) y tiene relación con multitud de aspectos. Si *todo permanece constante* el tamaño de la muestra sí es un criterio válido. Si dos encuestas cuentan con el mismo grado de control durante el trabajo de campo, obtienen la muestra a partir del mismo marco y con el mismo método de muestreo y se utilizan para aplicar una misma herramienta de inferencia a la misma variable, el resultado obtenido con la muestra más grande es más creíble que el que proviene de la muestra más pequeña. Ésta obedece a una distribución muestral más variable, es decir, donde es más fácil (o en palabras del profesor Sánchez Bruno, más probable) encontrar resultados más alejados del parámetro. En la situación ideal límite, si es igualmente accesible la población que una muestra, con el mismo error de medida y el mismo grado de control, es preferible trabajar con la población (la muestra más grande posible). El trabajo de investigación, en la práctica, es complicado y muestra retos continuamente porque los principios aislados suelen resultar incompatibles entre sí. Por ejemplo, junto con el principio "la muestra cuanto más grande mejor", se encuentran los principios "cuanto mayor control mejor" o "cuanto menos coste mejor" y estos segundos son habitualmente incompatibles con el primero.
2. El tamaño del efecto es un índice excelente para medir el efecto en la muestra, se estandarice o no. Dependiendo de la medida utilizada para medir el efecto, existe más o menos suficiencia o más o menos independencia de otros elementos. El indice de correlación que he utilizado en la exposición del trabajo, es

[3] El sentido con el que se utiliza *suficiencia* aquí no coincide enteramente con el significado estadístico como propiedad deseable en un estimador. Se está utilizando la suficiencia en su sentido popular o cotidiano sin una medida concreta.

sensible a varios elementos. Uno de ellos es la variabilidad de las variables que se relacionan. Luego, r_{xy} no es suficiente. El panorama se complica cuando es utilizado como estimador. Los tamaños de efecto estandarizados se utilizan como estimaciones puntuales de efectos en la población. En este sentido son sensibles a la distribución muestral subyacente, que es desconocida, pero que se le supone cierta identidad si se cumplen determinadas condiciones (que comienzan en la aleatoriedad de la muestra, omnipresente en las encuestas, pero una rareza en otros tipos de investigación). Esto hace que la variabilidad esperable en los tamaños de efectos obtenidos en las investigaciones dependa, entre otras cuestiones, del tamaño de la muestra.

3. El grado de significación es como una esponja que absorve de todo y depende de ello. En este sentido es poco suficiente. Por ejemplo, con el mismo valor para el estimador, el grado de significación disminuye conforme aumenta el tamaño de la muestra. Pero ante la constancia de las variables que definen el contexto, el grado de significación sí recoge la fuerza de la relación, puesto que conforme ésta es mayor, $p(D/H_0)$ disminuye. El problema en el uso del grado de significación es, precisamente, que no sólo recoge la fuerza de la relación, sino más aspectos. Pero este inconveniente puede resultar también un aspecto positivo, cuando el valor del estimador sea claramente insuficiente. Pongamos por caso a dos investigaciones con un mismo objetivo: la relación entre la edad y el grado o nivel de convencimiento en la decisión de voto. Unos resultados obtenidos se muestran en la tabla 1.

Tabla 1: *Resultados de dos investigaciones*

Investigación	r_{xy}	$p(D/H_0)$
A	0,4	0,062
B	0,2	0,001

En el ejemplo, utilizar el tamaño del efecto o el grado de significación lleva a comportamientos paradójicos. Según la investigación A el efecto es mayor que según B. Pero si el investigador recurre a la decisión según la prueba de significación de la hipótesis cero y utiliza el habitual α=0,05 concluirá que las variables están relacionadas en B, pero no en A.

Lamentablemente, ninguna de las dos estrategias es escrupulosamente correcta, porque ninguna es *suficiente*. En A, la muestra indica una mayor relación entre variables que en B. Pero en la inferencia, el valor obtenido en B es menos probable suponiendo cierta H_0 que en A. La diferencia entre ambos resultados puede explicarse por un tamaño de muestra muy desigual y la circunstancia de que los resultados extremos son más probables en las muestras pe-

queñas (Timothy, 1989), por lo que los efectos muestrales son menos fiables (¿Queda justificado en este ejemplo el uso aislado de r_{xy} ?)

En relación a las soluciones ideales y prácticas

En el trabajo original, ya manifesté mi desacuerdo con respecto al uso de la PSHN, muy especialmente en el sentido de que su uso habitual lleva a contemplar la hipótesis de no relación estricta. Pero es innegable que la utilización de esta estrategia se encuentra muy fuertemente arraigada. El sentido del trabajo no es indicar que la utilización de *nu* es metodológicamente lo más indicado, sino que obedece a un intento de mitigar los inconvenientes ocasionados con motivo del uso de PSHN con hipótesis cero. Para aclarar más este punto es necesario entrar en: el dilema entre estimaciones puntuales y por intervalo, el tratamiento de los índices y el de las distribuciones de probabilidad, el dilema entre la información y la decisión, y el dilema entre soluciones correctas y soluciones prácticas.

El dilema entre las estimaciones puntuales y por intervalo

En las investigaciones mediante encuestas, los objetivos de inferencia se sacian en su mayoría mediante la estimación con intervalos de confianza. Son estos intervalos los que protagonizan la información de resultados en los medios de comunicación, diversos bancos de datos y multitud de informes. En este contexto, cuando se realiza una estimación de cualquier índice, se observa que se requieren informaciones previas sobre la población. Esto es lo que se ha venido a llamar la paradoja de Friedman (Azorín y Sánchez Crespo, 1986): para estimar algo en la población es necesario saber ya sobre ella. En el ejemplo básico de la estimación de una media en el contexto de un muestreo aleatorio simple, el objetivo es conocer la media aritmética de una variable en la población, pero para ello es necesario conocer previamente la varianza de la variable también en la población. En este sentido es útil distinguir entre medidas objetivo y medidas procedimentales (Manzano, 1996). Las primeras se refieren a los índices que se pretenden estimar porque forman parte de los objetivos de la investigación. Las segundas se corresponden a los índices cuya estimación es exigida por el procedimiento. En el ejemplo sobre la media aritmética, ésta constituye una medida objetivo, mientras que la varianza es una medida procedimental. En la práctica se estiman por intervalo las medidas objetivo y por punto las procedimentales. Si no fuera así, se pondría en marcha un ciclo infinito puesto que la construcción de intervalos de confianza para las medidas procedimentales también exige la estimación de otras medidas.

Las situaciones prácticas requieren soluciones imperfectas. Romper con el ciclo implica utilizar estimaciones puntuales quizá poco acertadas. En este mismo sentido,

los tamaños de efecto deben ser objeto del mismo tratamiento, según se consideren medidas objetivo o procedimentales.

Tratamiento de los índices y de las distribuciones de probabilidad

Partimos de la hipótesis de que las dos caras de una moneda son equiprobables. Al lanzar la moneda diez veces, se obtiene el resultado:

c, x, x, x, x, c, x, x, c, x

La probabilidad de obtener este resultado o situaciones más extremas suponiendo cierta la hipótesis de partida es:

$$p(D/H_0) = 2\sum_{i=0}^{3}\binom{10}{i}0{,}5^{10} = 0{,}3438$$

Este valor no es propiamente hablando una medida o un índice de la muestra, es una traducción de un resultado obtenido en la muestra según una distribución de probabilidad asociada a la distribución muestral que se le supone si fuera cierta la hipótesis de partida. ¿Qué sentido tiene la cuestión "cuál es la probabilidad de que la probabilidad de obtener 3 caras al lanzar la moneda 10 veces (o situaciones más alejadas de obtener 5 caras) sea 0,3438"? $p(D/H_0)$ no está sujeto a una distribución muestral, puesto que ya es el resultado de una distribución muestral que parte de suposiciones (por ejemplo, una hipótesis y una distribución de probabilidad concretas)

Así pues, en la práctica, sólo interesan las estimaciones de índices utilizados como estimadores y no de funciones de probabilidad.

El dilema entre la información y la decisión

Como ya se ha mencionado, uno de los inconvenientes asociados a PSHN es que obliga a dicotomizar un continuo y tomar una decisión en función del resultado de esta *rotura*. Este aspecto está ausente en los intervalos de confianza, salvo que éstos se utilicen también para tomar una decisión, realizando la dicotomía en función de que el valor que defiende la hipótesis nula se encuentre o no en el intervalo. El inconveniente aquí no es tanto el hecho de que se utilice el valor de probabilidad $p(D/H_0)$ o un intervalo de confianza, sino que el artilugio sirva para tomar una decisión basada en una dicotomía artificial. $p(D/H_0)$ suministra información más rica que la utilizada para tomar la decisión. El intervalo de confianza es aún más idóneo, puesto que no está basado en la suposición de la hipótesis nula y, por tanto, no depende de ésta. Entonces ¿Por qué dicotomizar?

La dicotomía es muy útil en la práctica, puesto que permite tomar una decisión automática. Una vez definido el proceso, el investigador sólo debe observar si el valor de la hipótesis nula se encuentra o no en el intervalo de confianza o si $p(D/H_0)$ es o no mayor que α. La discusión subyacente es de gran trascendencia pero muy escurridiza: ¿Es necesario tomar decisiones cualitativas en forma de SI/NO?

En la práctica, las decisiones son inevitables. El político decide utilizar una estrategia electoral u otra en función de la cercanía de su estimación de voto con respecto a otras fuerzas políticas. El empresario toma una decisión u otra en función de la estimación de éxito de una estrategia comercial. Cuando se pone en marcha una encuesta, existe una necesidad de tomar decisiones más o menos explícitas. Incluso nuestras decisiones cotidianas se sustentan en el mismo procedimiento: antes de pedir un aumento de sueldo (pedir o no pedir, es una dicotomía) el empleado debe estar *suficientemente* convencido o desesperado o encontrar *suficientes* indicios contextuales de que su petición será exitosa.

La profesora Borges señala un punto de innegable trascendencia para la investigación: lo ideal es llegar a leyes que se expresen en modelos cuantitativos. Si el momento histórico lo permite, porque se cuenta ya con suficientes indicios empíricos y modelos teóricos sólidos, hay que aspirar a establecer conclusiones cuantitativas y superar las operaciones ordinales. Supongamos que se consigue establecer tamaños de efecto producto del meta-análisis a través de una ingente cantidad de investigaciones sólidas. Finalmente se requieren resultados cuantitativos concretos e individuales. Una distribución de resultados probables no es suficiente. Tal vez se estima que la relación entre dos variables es, aproximadamente, $\rho_{xy}=0{,}31$ ¿Y qué? ¿Qué utilidad tiene este resultado? ¿Qué hacer después? Si se trata de la relación entre la edad y el grado de resistencia a variar la intención de voto ¿Para qué sirve $\rho_{xy}=0{,}31$? Finalmente, se tomarán decisiones. Las tomarán los partidos políticos, los asesores de imagen, las empresas de encuestas, los medios de comunicación...

Pero el hecho de que las decisiones suelen reducirse a contextos dicotómicos no justifica el uso habitual de las PSHN cuando contemplan la hipótesis cero, puesto que se debería definir cuantitativamente el punto de rotura del continuo. Esto es incómodo, subjetivo y discutible. Por ello, el investigador utiliza los puntos clásicos o habituales, de tal forma que evita la reflexión.

El dilema entre las soluciones correctas y las soluciones prácticas

Pensar en el investigador aplicado como un metodólogo impecable que, además, domina un campo de conocimiento sustantivo, es ingenuo. El especialista en conductas de salud infantiles, conoce bien ese campo, sabe qué variables están en juego y qué implicaciones se derivan de actuaciones concretas. Puede poner en marcha una encuesta con el objetivo de identificar las percepciones de la población infantil con respecto a determinadas cuestiones de salud. Este proceso debe ser muy elaborado:

los cuestionarios utilizados para la población infantil deben ser especiales, quizá sólo contendrán dibujos y el niño deberá poner un dedo en el que más se identifique con su percepción acerca del objeto sobre el que se le pregunta; el marco para la investigación tendrá innumerables problemas si no se ciñe a la población escolarizada; deberá contar con la colaboración de las direcciones de los centros escolares o con un respaldo oficial suficiente; se enfrentará también a problemas de no respuesta derivados de la relación entre el encuestador y el niño; etc. Pero estos problemas serán llevaderos si cuenta con un equipo de especialistas en encuestas.

¿Qué ocurrirá en la etapa de análisis y obtención de conclusiones? Este investigador puede estar interesado, por ejemplo, en la dependencia existente entre el nivel de ingresos de la unidad familiar a la que pertenece el niño (variable X) y si consume o no algún tipo de bebida alcohólica (variable Y). Para saciar esta curiosidad de investigación puede utilizar la diferencia entre las medias de la variable X según el nivel de Y. En la práctica el investigador estará interesado en concluir si X e Y están relacionadas o no. No es que este comportamiento se derive necesariamente del problema de estudio, es que sabemos que lo hará, porque es la costumbre o la tradición y varias décadas de críticas metodológicas no la han modificado sensiblemente. Es más, para concluir utilizará una PSHN y, además, contemplará un efecto estrictamente nulo, es decir, partirá de

$$H_0 : |\mu_{Y1} - \mu_{Y2}| = 0$$

La alternativa a este comportamiento, sin salir del contexto de decisión dicotómica, es contemplar otro valor para H_0, otro punto de corte en el continuo de posibles diferencias observables entre $\bar{x}_{Y1}$ y $\bar{x}_{Y2}$. Si se utilizara la estrategia de la *magnitud suficiente* subyacente al principio "suficientemente bueno" de Serlin y Lapsley (1985), se debería también escoger un punto de corte para la magnitud en el continuo que parte de cero. Pero cualquiera de estas alternativas exigirá un proceso reflexivo que el investigador no está dispuesto a realizar (y probablemente tampoco esté preparado para ello). Ni las instituciones que sustentan económicamente el estudio, ni el resto de sus colegas, ni las revistas que pueden publicar su trabajo, le exigirán otro comportamiento que el que está realizando. Es más, si utilizara intervalos de confianza sin tomar decisiones dicotómicas, si facilitara tamaños de efecto en lugar de comparaciones entre $p(D/H_0)$ y α o, incluso, si utilizara sus resultados como meras evidencias empíricas sin motivación de inferencia estadística, encontraría sin duda más problemas en su camino.

Luego, el comportamiento más adaptativo para el investigador aplicado, aún suponiendo que domina a la perfección cualquier aspecto metodológico, es hacer las cosas exactamente como se están haciendo.

En este contexto caben dos salidas: seguir insistiendo en que el camino no es el correcto o disminuir en algo la incorrección del proceso. Particularmente creo que ambas salidas son aconsejables, en paralelo. En el fondo se trata de una mezcla entre

la idoneidad metodológica de los procedimientos y la práctica de la investigación a partir de los comportamientos de los investigadores. Y lo que parece claro es que los metodólogos todavía no han dado con la llave que permite controlar el comportamiento de los investigadores. Si éste se basa, en cierta medida, en la comodidad o en el automatismo, parece fundado buscar alternativas más adecuadas que sigan siendo cómodas y automáticas.

En relación directa al índice v

Dos son los comentarios relacionados con este índice. Por un lado, desconfianza acerca de su viabilidad. Por otro, la posibilidad de establecer modificaciones. Por último, indicaré algunos aspectos a modo de conclusión, de tal forma que se unifiquen los argumentos esgrimidos hasta el momento.

En referencia a su viabilidad. En efecto, quizá sea conveniente que las sugerencias de cambio, modificación, variación o innovación se retarden un tiempo prudencial hasta que la comunidad de usuarios pruebe las sugerencias y los especialistas en metodología evalúen los postulados previos, el comportamiento intermedio y los resultados obtenidos. Sin embargo, en los aspectos relacionados con la inferencia estadística en general y con las PSHN en particular, parece claro que el tiempo excede lo prudente y que existe una resistencia que no es reflejo de una actitud reflexiva. Por otro lado, creo que v es más viable que las alternativas apuntadas hasta la fecha. Éstas pueden ser (o son) más correctas, pero exigen del investigador una dosis de reflexión (tal vez más de un 90% de transpiración) que de momento las hacen poco utilizadas en la práctica. Además, los intereses comerciales de las empresas del sector informático se ciñen más a las posibilidades de venta y expansión que a la corrección de los procedimientos que implementan. Así, tendrá un mayor éxito comercial el programa de ordenador que no exija casi nada al usuario, frente al sistema informatizado que requiera la toma de decisiones comprometidas.

Con respecto a las modificaciones del índice, el profesor Manzano sugiere una variación encaminada a disminuir la influencia de β, de tal forma que v se acerque más a α. Me parece juiciosa. Sin embargo creo que la interpretación de *nu* se complica. Y aún me parece que se podrían realizar modificaciones más drásticas. El objetivo es aumentar la suficiencia del índice utilizado para tomar decisiones, más allá de α, siempre que la medida resultante permita mantener la comodidad para el investigador.

Conclusiones

Según lo mencionado hasta el momento, se puede concluir que:

1. El objetivo del índice *nu* es precisamente buscar una solución (aún imperfecta) en el campo de la comodidad y el automatismo. Su implementación informáti-

ca es fácil y su interpretación puede estar sujeta a los mismos convenios que β o α.

2. ν no es un resultado de la muestra, sino una razón modificada entre probabilidades. Luego, plantearse una distribución muestral asociada a ν tiene los mismos problemas conceptuales que una distribución muestral asociada a $p(D/H_0)$
3. Su lógica se basa en que uno de los elementos de los que *bebe* $p(D/H_0)$ es la potencia de la prueba. Existen otras alternativas similares. Se podría utilizar directamente el tamaño de la muestra (que condiciona tanto $p(D/H_0)$ como β), pero β no sólo es sensible a éste sino a otros elementos como al tipo de prueba. Por otro lado, tanto β como $p(D/H_0)$ son dos medidas de probabilidad, lo que permite controlar los efectos de escala.

Esta salida es, incluso, incómoda de aconsejar puesto que, en el fondo, se está proponiendo una incorrección como sustituta de otra. Creo, no obstante, que existe un abismo entre el uso de $p(D/H_0)$ y de ν, por cuanto que éste es más suficiente que el primero, al controlar mejor el contexto de decisión. Salvando las distancias y utilizando un ejemplo quizá no muy grato, la justificación para ν comparte ciertas similitudes con algunas decisiones en torno al consumo de drogas y los riesgos de padecer SIDA. El criterio ideal es favorecer la desaparición de las adicciones a las drogas y erradicar el SIDA tanto como cualquier otra de las consecuencias nefastas derivadas de las drogadicciones. No obstante, esto no ha ocurrido. Se trabaja en ello y existen confluencias de voluntades en este sentido. Pero no es suficiente. Una salida paralela ha consistido en suministrar gratuitamente jeringuillas esterilizadas. Esta solución llega a escandalizar a algunos puesto que parece que las instituciones públicas fomentan el consumo de drogas. No obstante, constituyen realmente una salida, no ideal, pero sí práctica.

Referencias

Azorín, F. y Sánchez Crespo, J.L. (1986) *Métodos y aplicaciones del muestreo*. Madrid: Alianza Editorial.

Bakan, D. (1966) The test of significance in psychological research. *Psychological Bulletin*, 66, 1-29.

Borges, A. (1997) Algunos problemas frecuentes en la interpretación de los contrastes de hipótesis estadísticas en psicología. *Iberpsicología*, 2:3:7. (http://fs-morente.filol. ucm.es/publicaciones/iberpsicologia/iberpsicologia.htm)

Chou, Y.L. (1972) *Análisis estadístico*. México: Nueva Editorial Interamericana.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. Hilsdale, New Jersey: Erlbaum.

Cohen, J. (1990) Things I have learned (so far). *American Psychologist*, 45.1304-1312.

Cohen, J. (1994) The earth is round (p<.05). *American Psychologist*, 49, 1304-1312.

Cortina, J.M. y Dunlap, W.P. (1997) On the logic and purpose of significance testing. *Psychological Methods*, 2, 161-172.

Frick, R.W. (1996) The appropiate use of null hypothesis testing. *Psychological Methods*, 1, 379-390.

Gigerenzer, G. (1993) The Superego, the Ego, and the Id in statistical reasoning. En Keren y Lewis (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, New Jersey: Erlbaum.

Kruskal, W. y Mosteller, F. (1979a). Representative Sampling I: Non-scientific literature. *International Statistical Review*. 47,13-24.

Kruskal, W. y Mosteller, F. (1979b). Representative Sampling II: Scientific Literature. *International Statistical Review*. 47, 111-127.

Kruskal, W. y Mosteller, F. (1979c). Representative Sampling III: the current statistical literature. *International Statistical Review*. 47, 245-265.

Kruskal, W. y Mosteller, F. (1980). Representative Sampling IV: the history of the concept in statistics. *International Statistical Review*. 48, 169-195.

Lipsey, M.W. (1990) *Desing sensitivity: statistical power for experimental research*. London: Sage.

Manzano, V. (1996) *Tamaño de muestra óptimo en investigaciones mediante encuesta. Fundamentos e implementación de un sistema de ayuda a la decisión*. Tesis doctoral. Sevilla: Facultad de Psicología, Universidad de Sevilla.

Martínez Arias, R. (1995). El método de encuestas por muestreo: conceptos básicos. En M.T. Anguera, J. Arnau, M. Ato, R. Martínez, J. Pascual y G. Vallejo (Ed.) *Métodos y técnicas de investigación en psicología*. Madrid: Síntesis, 385-431.

Rosnow, R.L. y Rosenthal, R. (1989) Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.

Serlin, R.C. (1987) Hypothesis testing, theory building and the Philosophy of Science. *Journal of Counseling Psychology*, 34, 365-371.

Serlin, R.C. y Laspley, D.K. (1985). The good-enough principle. *American Psychology*, 40, 73-83.

Silva, L.C. (1993) *Muestreo para la investigación en ciencias de la salud*. Madrid: Díaz de Santos.

Timothy, R. (1989). Variations on a seminal demonstration of people's insensivity to sample size. *Organizational behavior and human decision processes*, 43, 52-57.

Yela, M. (1991) Unidad y diversidad en Psicología. En J. Mayor y J.L. Pinillos, *Tratado de Psicología General. Historia, teoría y método*. Madrid: Alhambra.

Yela, M. (1994) El problema del método científico en psicología. *Anuario de Psicología*, 60, 3-12.