Metodología de Encuestas
Vol 1, Num 1, 1999, 47-68

LA PRUEBA DE SIGNIFICACIÓN DE LA «HIPÓTESIS CERO» EN LAS INVESTIGACIONES POR ENCUESTA

Viso M. Ares
SIPIE

RESUMEN

Hay muchos trabajos que abordan el papel de la prueba de significación de la hipótesis nula (PSHN) en la construcción del conocimiento científico. La mayoría de tales estudios encuentran algunos problemas tanto en la lógica como en el uso de la PSHN. Uno de estos inconvenientes tiene un efecto importante en la interpretación de los resultados. Se trata del hábito de la hipótesis-cero. En este trabajo se abordan los efectos de la hipótesis-cero, de la potencia y del tamaño de la muestra en la interpretación del grado de significación. Dado que la investigación por encuestas usualmente considera tamaños grandes de muestra, el efecto de la hipótesis-cero resulta especialmente negativo. Por otro lado, con frecuencia, los investigadores utilizan mecánicamente el análisis de los datos, por lo que es muy difícil modificar su comportamiento mediante llamadas de atención en las revistas especializadas. Por esto, las soluciones deben ser muy prácticas y fáciles de implementar en un programa de ordenador. De acuerdo con todo ello, se propone una estrategia concreta para facilitar la interpretación del grado de significación en función de la potencia de la prueba: el índice de significación relativa.

Palabras clave: prueba de significación, hipótesis nula, significación relativa, potencia de la prueba, encuestas

Pruebas de significación de la hipótesis nula

La prueba de significación de la hipótesis nula (PSHN) constituye una de las herramientas esenciales de la inferencia estadística, desde su creación por Fisher (1956, 1966). Esquemáticamente consiste en:

1. Se enuncia una hipótesis (hipótesis nula, H_0) sobre el valor de un parámetro. Por ejemplo, que $\rho_{xy} = 0$.
2. Se escoge una función muestral (estadístico) como estimador de la función poblacional. Es indispensable que se cuente con una distribución muestral asociada al estimador. En el ejemplo, el estimador puede ser r_{xy}, que sigue la distribución t de Student con n-2 grados de libertad y con $t = r\,[(n-2)/(1-r^2)]^{1/2}$.
3. Se obtiene una muestra aleatoria de tamaño n y se calcula en ésta el valor del estimador.
4. Se calcula $p(D/H_0)$, es decir, la probabilidad de obtener resultados al menos tan extremos como el que se ha obtenido (D) suponiendo que es cierta H_0
5. Si $p(D/H_0)$, también llamado *grado de significación*, tiene un valor muy pequeño, se considera que H_0 es poco creíble a la luz de los resultados y se rechaza. Si, por el contrario, el grado de significación no es pequeño, se mantiene H_0

La estrategia para decidir si el grado de significación es o no suficientemente pequeño como para rechazar H_0 consiste en comparar $p(D/H_0)$ con un referente teórico, un umbral que define cuál es el máximo riesgo que se está dispuesto a cometer al rechazar H_0 siendo ésta cierta. Este valor de comparación, denominado *nivel de significación*, es simbolizado con la letra griega α. Desde el trabajo de Fisher (1966, cuya primera edición fue en 1935), α ha tomado casi exclusivamente el valor 0,05 con la consiguiente crítica de numerosos autores (por ejemplo, Rosnow y Rosenthal, 1989).

PSHN ha cumplido un papel crucial en la generación de conocimiento contrastado empíricamente (Wainer, 1999), desde su creación hasta nuestros días. Aceptando que su lógica y su utilización han ido acompañadas de una serie de problemas (véase más adelante), su creación supuso el lanzamiento de la inferencia estadística como método de generalización de los resultados desde una experiencia concreta y como aval en la construcción de teorías científicas. Hasta ese momento, descontando algunos precedentes históricos incompletos (Gutiérrez Cabría, 1994) y los trabajos de la escuela de Karl Pearson (Vallecillos, 1966), no se ponía en práctica un procedimiento estandarizado para tomar decisiones probabilísticas sobre la veracidad de las hipótesis. La inferencia bayesiana se utilizaba con muy baja frecuencia, debido principalmente al problema de la estimación de las probabilidades subjetivas.

El procedimiento de Fisher es acorde con la perspectiva popperiana de la construcción del conocimiento (Borges, 1997): no es posible comprobar la veracidad de las declaraciones, pero sí su falsedad. De esta forma, la ciencia debe construir teorías

que sean falseables. Así, en la lógica fisheriana, puede mostrarse que H_0 es falsa (muy poco creíble a la luz de los resultados empíricos), pero nunca que sea verdadera.

Inconvenientes asociados a PSHN

El listado de publicaciones relacionadas con críticas a PSHN es muy amplio (Valera y Sánchez Meca, 1997). Sólo tres años después de la exposición de la lógica del proceso por Fisher, Berkson (1938) realizó algunas críticas sobre el procedimiento, que llega a ser calificado posteriormente de *bestia*, *cadáver sin cabeza* (Oakes, 1986) o *dragón* (Harris, 1991). No obstante, una visión histórica de los trabajos críticos con respecto a PSHN muestra cierto auge de la actitud negativa en torno a los años 60 y 70, además de un resurgimiento actual de esta vieja polémica. Dentro de las ciencias sociales y del comportamiento, quizá sean Rozeboom (1960), Bakan (1966), Cohen (1969) y Morrison y Henkel (1970) los trabajos críticos con mayor trascendencia en este tópico. En la actualidad, se vive otra época de intensa discusión en torno a PSHN, donde uno de los espolones es Cohen (1994) y toda una serie de publicaciones a modo de *efecto* que incluyen serias defensas del procedimiento (véase, por ejemplo, el excelente trabajo de Frick, 1996).

Sea como fuere, son dos las vías por las que la PSHN ha sido negativamente valorada:

1. El procedimiento en sí es defectuoso.
2. El uso que se hace en la práctica es frecuente y persistentemente (Vacha y Nilsson, 1998) incorrecto, resultado de creencias erróneas[1] sobre probabilidades, muestreo o la PSHN en sí.

---

[1] El término *falacia* es utilizado con frecuencia con el significado de una creencia errónea basada en un argumento de igual cualidad. No obstante, este uso en crecimiento es incorrecto en español. *Falacia* significa, según la Real Academia Española de la Lengua (página 946 del *Diccionario de la Lengua Española*, Tomo I, de la edición vigésima primera de 1992), «Engaño, fraude o mentira con que se intenta dañar a otro / Hábito de emplear falsedades en daño ajeno» y Seco (1997:195) indica que «es impropiedad, debida a anglicismo, dar a este nombre el sentido de 'error, sofisma o argumento falso'». De hecho, el significado español de falacia implica una falsedad y un daño intencionado que la acompaña, aspecto fundamental ausente en la utilización con que hemos encabezado esta nota. El término original inglés para el error, la creencia o el argumento falso es *falacy*. El uso incorrecto, resultado de una traducción no muy acertada, queda también potenciado por las equivalencias en los diccionarios. Así, en la edición quinta de 1997 del Collins de Grijalbo, el término *falacia* viene acompañado de la sugerencia en inglés: *falacy* o *error*. Nuestra impresión es que debería evitarse este término y utilizar en su lugar la expresión «creencia errónea» u otra similar.

Dentro de la primera afirmación se encuentran, básicamente, las siguientes críticas:

1. PSHN permite calcular $p(D/H_0)$ si se cuenta con una distribución muestral adecuada. Pero lo que interesa al investigador es realmente $p(H_0/D)$, sobre lo que PSHN no tiene competencias. De hecho, se encuentran ejemplos en los que puede generarse un bajo $p(D/H_0)$ (lo que lleva a rechazar H_0) pero con un alto $p(H_0/D)$ (Cohen, 1994).
2. PSHN es artificialmente dicotómica. El grado de significación suministra una información en un continuo que va desde la probabilidad 0 hasta la probabilidad 1. Sin embargo, en el procedimiento generado por Fisher se toma una decisión basada en la dicotomía «$p(D/H_0)$ es menor o no que α». Además, la decisión sobre α es subjetiva (Chow, 1988).
3. En PSHN se prueba una única de entre la infinidad de hipótesis que podrían explicar más o menos satisfactoriamente los hechos observados y que podrían ser compatibles con los postulados teóricos. Ésta es quizá la crítica principal que justifica la respuesta de Neyman y Pearson (1933), donde se favorece una toma de decisión entre dos hipótesis (nula y alternativa) a la luz de las probabilidades asociadas.

Por otro lado, hemos apuntado que el segundo tipo de críticas negativas que recibe la PSHN se ciñe a la percepción de su uso incorrecto, cuyo nivel de hábito ha sido catalogado de desastre (Hunter, 1997). Al respecto, Manzano (1997)[2] expone un conjunto de quince incorrecciones de concepción y utilización que pueden esquematizarse en:

1. Concluir que H_0 es cierta si $p(D/H_0)<\alpha$
2. Confundir $p(D/H_0)$ con $p(H_0/D)$
3. Indicar que α es el riesgo de equivocarse cuando $p(D/H_0)<\alpha$ y se rechaza H_0
4. Utilizar por defecto la distribución normal para el cálculo de $p(D/H_0)$
5. Intercambiar el radio de estimación (e_p) con otros conceptos
6. Considerar que existe una distribución de valores verdaderos alrededor de los hechos
7. No considerar la variedad de tablas a la hora de traducir distancias a probabilidades o viceversa
8. Confundir α, $p(D/H_0)$ y p
9. Decidir α tras conocer $p(D/H_0)$
10. Considerar que sólo existen dos valores utilizables para α: 0,05 y para situaciones más exigentes, 0,01

---

[2] Manzano utiliza $p(O/H_0)$ para representar el grado de significación. No obstante, en este extracto hemos seguido el casi idéntico $p(D/H_0)$ de Cohen (1994) anterior y más extendido.

11. Considerar que α debe ser suficientemente pequeño ante cualquier problema de investigación
12. Contemplar únicamente la significación estadística
13. Si se rechaza H_0, utilizar el valor encontrado en la muestra como una buena representación del valor en la población
14. Considerar que una prueba de una cola es más restrictiva que una de dos colas en cuanto al rechazo de H_0
15. Considerar que la prueba de significación elimina la subjetividad

No obstante, para todas las críticas recibidas por PSHN existe una respuesta que se centra en el modo en que este procedimiento debe utilizarse y lo habitual comienza a ser la publicación de textos donde los autores reconocen las limitaciones de PSHN y señalan los espacios en los que su uso es recomendable (véase, por ejemplo, Frick, 1996 o Cortina y Dunlap, 1997).

Aún así, existe una seria limitación de PSHN, combinación entre la lógica del procedimiento y su utilización habitual, que nos preocupa de forma especial al tener una trascendencia indiscutible en el campo de las investigaciones mediante encuestas en donde se realicen pruebas de hipótesis: la credibilidad de la *hipótesis cero*.

La hipótesis cero y sus efectos en la significación de las encuestas

La forma más habitual para definir la hipótesis nula es mantener que el tamaño del efecto es cero, es decir, que la correlación entre dos variables es cero, que la diferencia entre dos medias es cero, que la razón entre dos varianzas es uno (ambas varianzas coinciden exactamente), etc. Este comportamiento tan ampliamente extendido genera serias deficiencias en PSHN. En el tratamiento de este tópico, nos preocupa particularmente las incoherencias conceptuales y prácticas que subyacen a la hipótesis *cero* y sus efectos en las investigaciones mediante encuestas. Los siguientes subapartados se ocupan de estas cuestiones.

Deficiencias conceptuales en la hipótesis cero

Esta forma de construir la hipótesis (*nil hypothesis* para Cohen, 1994) resulta cómoda y poco comprometida puesto que el investigador se ve liberado de la obligación de reflexionar sobre un tamaño de efecto distinto de cero. Por otro lado, se trata de un comportamiento tan extendido y habitual que no representa inconveniente alguno para la publicación de trabajos en las revistas especializadas. Además, constituye la opción por defecto para las PSHN en los programas comerciales de análisis de datos implementados en ordenador.

No obstante, es difícil admitir que los valores de estricto cero existan realmente en las poblaciones. Así, por ejemplo, calculado el coeficiente de correlación producto-momento ρ entre dos variables (X e Y) sobre una población real ¿Cabe esperar que ρ_{xy}

suministre un valor de cero *exacto* (no 0,0003284 ó 0,00001, sino *exactamente* cero)? En función de la precisión del procedimiento de medida, el valor poblacional comenzará a ser distinto de cero a partir de alguna posición decimal (Tukey, 1991).

Si los efectos *cero* no existen en sentido estricto en la población, entonces las hipótesis *cero* son necesariamente falsas. El investigador, pues, puede prescindir de llevar a cabo investigaciones empíricas y encabezar las conclusiones con algo parecido a «*No se han recogido datos pero, con una cantidad suficiente de éstos, rechazaríamos la hipótesis nula. Luego, en consecuencia, ...*»

En el trasfondo conceptual de este proceder se encuentra cierta incompatibilidad entre los razonamientos cualitativos y cuantitativos. A saber:

Póngase por caso que el investigador está interesado en medir la relación entre dos variables, *X* e *Y*, y que escoge para su medida el coeficiente de correlación lineal simple ρ_{xy}. Este índice suministra valores en un continuo definido entre los extremos −1 y +1. De los infinitos resultados que pueden obtenerse al calcular un r_{xy}, 0 es sólo uno de ellos. En términos de probabilidad, pues, el suceso $\rho_{xy} = 0$ es imposible. Es cierto que ρ_{xy} es un valor concreto y no sigue una distribución de probabilidad (no es un estadístico), pero el razonamiento es conceptualmente aplicable considerando no el continuo (−1,+1) sino el intervalo de variación de ρ_{xy} que permitiría al investigador interpretar que *prácticamente* $\rho_{xy} = 0$.

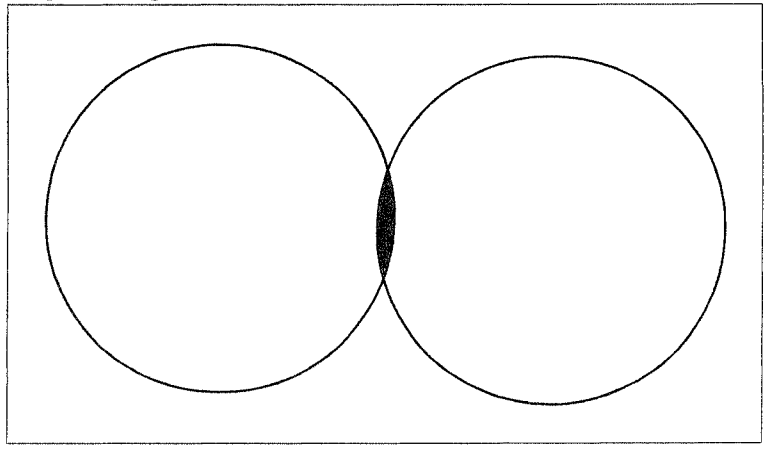
En su intención y objeto, el investigador no está interesado directamente en ρ_{xy} sino en la relación entre *X* e *Y*. Luego, y dado que va a traducir un ente cualitativo (el concepto de relación) en una cantidad numérica concreta (el valor de ρ_{xy}), debería reflexionar acerca de qué resultados de medida serán indicadores cuantitativos suficientes para la existencia de relación y cuáles no. En términos prácticos, este razonamiento obliga a definir un punto de corte en el continuo (−1,+1) que marque la presencia o ausencia de relación.

En el caso de ρ_{xy} se cuenta con un recurso útil para interpretar el ente cualitativo, puesto que el cuadrado de la correlación (ρ^2_{xy}) representa la proporción de variación lineal compartida (Kenny, 1979) entre *X* e *Y*. Bajo el supuesto de que la relación (sea la que fuere) entre *X* e *Y* sea lineal ¿Qué porcentaje mínimo de variación común es exigible?. Observemos que, por ejemplo, un valor $\rho_{xy} = 0{,}1$ implica un porcentaje de variación lineal compartida de $0{,}1^2 = 1\%$. La figura 1 intenta representar a dos variables cuya variación lineal compartida es del 1%. La zona común se ha indicado con el solapamiento de ambos círculos y ha sido sombreada. Debería concluirse con facilidad que ambas variables son prácticamente independientes (no obstante estas afirmaciones dependen siempre del contexto de investigación), es decir, que no están relacionadas. Sin embargo, su valor de relación es formal y estrictamente no nulo ($0{,}1 \neq 0$).

Pero la mayor incoherencia relacionada con el comportamiento del investigador se sitúa en su proceder ante el trabajo con muestras o con poblaciones. Cuando se trabaja con muestras, el hábito establece calcular r_{xy} (por ejemplo $r_{xy}=k$) , obtener el

valor de significación $p(D/H_0)$ asociado con la hipótesis $\rho_{xy}=0$, donde $D=|r_{xy}| \geq |k|$, y decidir si las variables X e Y están o no relacionadas entre sí en función de que $p(D/H_0)<\alpha$ ó $p(D/H_0) \geq \alpha$, respectivamente. Así pues, el investigador suele acudir únicamente al valor de la significación estadística para tomar decisiones cuando trabaja con muestras. No obstante, si opera directamente sobre una población, PSHN carece de sentido puesto que no procede la aplicación de la inferencia estadística (no existe una distribución muestral subyacente y su población no constituye una muestra aleatoria de nada), por lo que el investigador se ve obligado a reflexionar sobre el valor de ρ_{xy} que actuará como punto de corte, es decir, el valor mínimo para $|\rho_{xy}|$ que define la relación entre X e Y. Así, se llega a la incoherencia siguiente: es posible definir un umbral de $\rho_{xy}>\epsilon$ (donde $0 \leq \epsilon \leq 1$) para la asunción de relación entre variables en la población y, sin embargo, rechazar la hipótesis nula, asumiendo relación poblacional entre las variables, habiendo obtenido una correlación en la muestra con un valor $r_{xy}<\epsilon$.

Figura 1: *Representación de una variación lineal compartida del 1%*



Un ejercicio de simulación

En una población de tamaño infinito en la que se contemplan dos variables estrictamente independientes, se observará que $\rho_{xy}=0$. No obstante, las muestras de tamaño finito n suministrarán valores r_{xy} que pueden interpretarse según una distribución aleatoria alrededor de $\rho_{xy}=0$. Luego:

$$\lim_{n \to \infty} r_{xy} = 0$$

Pero al aumentar n también se incrementa la significación estadística de cuantías cada vez menores para r_{xy}. El valor límite para t será (donde se ha resuelto con $r_{xy}=n^{-1/2}$):

$$\lim_{\substack{n\to\infty \\ r_{xy}\to 0}} \sqrt{\frac{r_{xy}^2(n-2)}{1-r_{xy}^2}} = \lim_{n\to\infty} \sqrt{\frac{1-2/n}{1-1/n}} = 1$$

Luego, conforme la muestra es más grande, aumenta la probabilidad de mantener la hipótesis nula si el valor poblacional para la correlación es exactamente $\rho_{xy}=0$.

Sin embargo, las fluctuaciones aleatorias provocarán que, incluso en el contexto descrito, quepa esperar valores significativos para las correlaciones en la muestra. Dado que las poblaciones teóricas son inabordables, se recurre a la simulación para observar el comportamiento de los índices. Por ello, habrá que sumar a las fluctuaciones propias del azar, los sesgos que se deban a las máquinas imperfectas de generación de variables. Por otro lado, si bien conforme $n\to\infty$ entonces $r_{xy}\to 0$ (bajo el supuesto $\rho_{xy}=0$), las velocidades no son equiparables, de tal forma que si el tamaño de la muestra crece más de lo que decrece la cuantía de la correlación, se deberá observar que $p(D/H_0)$ tiende a ser inferior a α.

Para mostrar este punto, hemos simulado algunos experimentos aleatorios encaminados a calcular r_{xy} con respecto a la tirada de dos dados. En la vida real, la correlación también sería estrictamente no nula, puesto que los dados no son perfectos y la mano que los tira puede generar también algún tipo de tendencia reforzada con la práctica, si es que fuera posible un número infinito de ensayos. En la simulación cabe esperar otro tanto, puesto que tampoco en este mundo son manejables los infinitos.

Cuadro 1: *Algoritmo de generación de un vector de correlaciones para las tiradas de dos dados* [ent(x) es la parte entera de x]

```
 1:n=1, i=0, coc=934390642, mul=742938285, mod=2147483647
 2:  n=3n, i=i+1, sumax=0, sumay=0, suma2x=0, suma2y=0,
     sprod=0
 3:  num=1
 4:  vale=coc*mul, coc=vale-ent(vale/mod)*mod,
     valorx=ent(coc/mod*6)+1
 5:    vale=coc*mul, coc=vale-ent(vale/mod)*mod,
       valory=ent(coc/mod*6)+1
 6:    sumax=sumax+valorx, suma2x=suma2x+valorx²
 7:    sumay=sumay+valory, suma2y=suma2y+valory²
 8:    sprod=sprod+valorx*valory
 9:  si num<n entonces num=num+1 e ir a 4
10:  desvx=((suma2x-sumax²/n)/n)^(1/2)
11:  desvy=((suma2y-sumay²/n)/n)^(1/2)
12:  corr(i)=(sprod-sumax*sumay/n)/n/desvx/desvy
13:si i<13 entonces ir a 2
```

Para la simulación hemos utilizado un algoritmo (ver cuadro 1) de congruencia lineal (Barry, 1996) donde la semilla aleatoria fue suministrada por el reloj del sistema (934390642), el módulo fue el habitual para una precisión de 32 bits con signo (2147483647) y el multiplicador ha sido obtenido de entre los más recomendados (Fishman y Moore, 1986; y Sánchez-Bruno y San Luís-Costas, 1995) para los casos 2^{31}-1 (742938285). Se han generado 13 muestras con tamaños $n=3^i$ y en cada una de ellas se han realizado n tiradas de dos dados. En cada caso se ha calculado el índice de correlación producto-momento y su grado de significación asociado ($p[D/H_0]$). Los resultados se encuentran en la tabla 1.

Tabla 1: *Simulación de $n=3^i$ tiradas de dos dados, y valores de correlación y significación asociados*

i	n	r_{xy}	sign. bil.
1	3	0,8386	0,1834
2	9	0,0940	0,4054
3	27	0,2202	0,1349
4	81	0,1470	0,0952
5	243	0,0200	0,3806
6	729	0,0440	0,1177
7	2187	0,0240	0,1270
8	6561	0,0100	0,2664
9	19683	0,0110	0,0704
10	59049	0,0100	0,0153
11	177147	0,0110	0,0000
12	531441	0,0100	0,0000
13	1594323	0,0100	0,0000

Se observa que en el caso concreto de nuestra simulación, el valor de correlación no ha diminuido con la misma intensidad con que ha aumentado el valor del tamaño de la muestra, por lo que el valor de significación ha llegado a ser tal que el rechazo de la hipótesis nula está garantizado ante cualquier valor α a partir de la generación 11 incluida. Consideremos que la aproximación $r_{xy} = 1/\sqrt{n}$ genera un límite para t de valor 1, mientras que ya t=1,645 está asociado con un grado de significación de 0,05 con $n=\infty$. El margen no es amplio ni la aproximación exacta, por lo que pequeñas fluctuaciones provocarán correlaciones significativas aún bajo el supuesto en el que estamos trabajando: *independencia* entre las dos variables.

Así pues, aún provocando sucesos de tal forma que quede garantizada su independencia teórica, podemos esperar muestras grandes donde se concluya a favor de la dependencia entre las variables.

El caso particular de las encuestas

La encuesta, como procedimiento de medida, es utilizada para una infinidad de objetivos y situaciones (Rojas y otros, 1998), lo que la hace enormemente atractiva para la realización de investigaciones desde perspectivas muy dispares. Así, y sin ánimo de ser exhaustivos, se llevan a cabo encuestas en sociología, psicología, ciencias políticas, márketing o ciencias de la salud, por ejemplo. No obstante, aún compartiendo el mismo procedimiento, las tradiciones metodológicas no coinciden necesariamente. Así, en psicología, la mayoría de las revistas que publican trabajos empíricos exigen el esquema clásico de especificar el método seguido mediante los omnipresentes subapartados de "sujetos", "variables", "instrumentos", "procedimiento" y "resultados". En estos últimos, es casi perceptivo el uso de PSHN para sustentar las conclusiones mediante una estrategia formal de inferencia (Estes, 1997; Vacha y Ness, 1999). Sin embargo, la tradición sociológica permite mayor libertad en la exposición de los trabajos empíricos, abundando los estudios puramente descriptivos. No obstante, algunas voces (como Huxley, 1988) critican el que los estudios empíricos en sociología no contengan más pruebas estadísticas para sus declaraciones o aserciones.

Luego, la incidencia de la PSHN en las encuestas está en parte relacionada con la tradición disciplinar de los investigadores. No obstante, la proliferación de ordenadores y programas de análisis de datos orientados al manejo fácil y accesible aún para inexpertos, facilita e incita a la realización de análisis (Chatfield, 1985), incluso de los más complejos como los que se localizan en la inferencia estadística (quintaesencia de la abstracción, según Watts, 1991).

Sea cual fuere la perspectiva del investigador, la combinación entre PSHN y encuesta no da buenos frutos en la práctica y resulta poco recomendable su utilización en los términos en que es habitual, dado que a la extendida aplicación de la hipótesis cero se une los grandes tamaños de muestra frecuentes en las investigaciones mediante encuestas. Así, se encuentran consejos prácticos que señalan, como tamaño habitual en ocasiones o mínimo en otras, las cantidades concretas de 600 unidades (Pérez Cebrián, 1987), de 1000 a 5000 (Hedges, 1980), 1000 (Harvatopoulos, Livan y Sarnin, 1992), 2000 (Noelle, 1970) o al menos se asume que las muestras deben ser invariablemente grandes (Hyman, 1970).

¿Qué trascendencia tiene el tamaño de la muestra para la obtención de resultados significativos?. Ya hemos entrado en los argumentos que justifican el hecho de que el tamaño de la muestra aumenta la probabilidad de encontrar resultados significativos, aspecto que se ha venido encuadrando bajo el epígrafe de *potencia de la prueba*. Una prueba es tanto más potente cuanto es más sensible a la falsedad de la hipótesis nula (Cohen, 1992). Y el tamaño de la muestra incrementa la potencia. Sin embargo, si bien estas relaciones son intuitivas, una inspección y breve estudio de las tablas 2 y 3

nos permiten obtener un mayor grado de conocimiento práctico sobre las consecuencias que el tamaño de la muestra tiene sobre la significación de los resultados.

La tabla 2 contiene los tamaños de muestra que son necesarios para rechazar la hipótesis nula de no relación estricta (H_0: $\rho_{xy}=0$) con $\alpha=0{,}01$ (trazo delgado) y $\alpha=0{,}05$ (trazo grueso) para valores de correlación comprendidos entre 0,01 y 0,99, obtenidos en la muestra. Obsérvese que, por ejemplo, el valor $r_{xy}=0{,}10$ que hemos juzgado más atrás como excesivamente pequeño (las variables *X* e *Y* son ¡independientes al 99%!) permite concluir que las variables *X* e *Y* están relacionadas a partir de $n=272$ ($\alpha=0{,}05$).

Tabla 2: *Tamaño mínimo de muestra para considerar que ρ_{xy} es distinta de cero. En filas se encuentra el valor decimal de la correlación en la muestra; en las columnas, las centésimas.* α = **0,05** *y* 0,01

	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0		**27057** 54119	**6766** 13530	**3008** 6014	**1693** 3383	**1084** 2165	**753** 1504	**554** 1105	**424** 846	**336** 668
0,1	**272** 541	**225** 448	**190** 376	**162** 321	**140** 276	**122** 241	**107** 212	**95** 188	**85** 167	**77** 150
0,2	**69** 136	**63** 123	**58** 112	**53** 103	**49** 94	**45** 87	**42** 80	**39** 74	**36** 69	**34** 65
0,3	**32** 60	**30** 57	**28** 53	**26** 50	**25** 47	**24** 44	**22** 42	**21** 40	**20** 38	**19** 36
0,4	**19** 34	**18** 32	**17** 31	**16** 29	**16** 28	**15** 27	**14** 26	**14** 25	**13** 24	**13** 23
0,5	**12** 22	**12** 21	**12** 20	**11** 19	**11** 19	**10** 18	**10** 17	**10** 17	**10** 16	**9** 16
0,6	**9** 15	**9** 15	**9** 14	**8** 14	**8** 13	**8** 13	**8** 12	**7** 12	**7** 12	**7** 11
0,7	**7** 11	**7** 11	**7** 10	**6** 10	**6** 10	**6** 9	**6** 9	**6** 9	**6** 9	**6** 8
0,8	**6** 8	**5** 8	**5** 8	**5** 8	**5** 7	**5** 7	**5** 7	**5** 7	**5** 7	**5** 6
0,9	**4** 6	**4** 6	**4** 6	**4** 6	**4** 5	**4** 5	**4** 5	**4** 5	**4** 5	**3** 4

Con cierto interés práctico, hemos construido la tabla 3. Contiene tamaños de muestra que se han utilizado en investigaciones mediante encuesta con cierta trascendencia. En concreto:

- Barómetro español de Junio de 1999 (CIS, 1999), con una muestra a nivel nacional de n=2496. (*CIS-I*)
- La investigación realizada en 1995 para evaluar la incidencia de la ludopatía en Andalucía y su relación con otras variables (Irurita, 1996) con un tamaño de muestra n=4997. (*LUDOP*)
- Encuesta preelectoral para las elecciones a las cortes españolas de 1996 (CIS, 1996) con un tamaño de muestra de n=6642. (*PREEL*)
- El estudio internacional con marco en Europa, realizado entre 1986 y 1990 para identificar diversos comportamientos relacionados con la salud en niños escolarizados (Mendoza, Sagrera y Batista, 1994) con n=31928. (*ECCER*)
- El Estudio General de Medios (AIMC, 1999) que utiliza una muestra total de n=40000. (*AIMC*)
- La Encuesta de Población Activa (INE, 1998) realizada sobre una muestra efectiva de 64 mil familias que supone un tamaño aproximado de n=200000. (*EPA*)

Tabla 3: *valor mínimo de r_{xy} para rechazar la hipótesis de no relación según* n *y* α

Estudio	n	$\alpha = 0{,}001$	$\alpha = 0{,}01$	$\alpha = 0{,}05$
CIS-I	2496	0,0619	0,0466	0,0330
LUDOP	4997	0,0438	0,0330	0,0233
PREEL	6642	0,0380	0,0286	0,0202
ECCER	31948	0,0173	0,0131	0,0093
AIMC	40000	0,0155	0,0117	0,0083
EPA	200000	0,0070	0,0053	0,0037

Así, por ejemplo, se observa en la tabla, que un investigador interesado en identificar relaciones entre variables en el Estudio General de Medios con la muestra total, concluiría con una respuesta afirmativa de relación con sólo obtener $r_{xy} \geq 0{,}0083$ ($¡r_{xy}^2 = 0{,}00007!$) con α=0,05.

¿Es PSHN inútil en las investigaciones mediante encuestas?

El objetivo principal de las encuestas es la estimación de parámetros en problemas unitarios, preferentemente la estimación de proporciones (por ejemplo, Rojas y otros, 1998). Este criterio justifica los valores tan amplios para *n* en la búsqueda de estimaciones precisas y con un bajo nivel de error. No obstante, multitud de estudios (especialmente desde la psicología y el marketing) están más preocupados por las relaciones entre variables o las diferencias entre colectivos. Tales son los casos explícitos de algunas de las investigaciones mencionadas con motivo de la tabla 3. En el estudio sobre la incidencia de la ludopatía en la Comunidad Autónoma Andaluza, preocupaba especialmente identificar variables que aumentaran el nivel de comprensión del fenómeno mediante estudios de relación. En la investigación sobre las conductas de los escolares relacionadas con la salud se tenía como objetivo, entre otros, "poner a prueba hipótesis de relación entre variables" (pág. 28). Por otro lado, los datos generados en las investigaciones mencionadas del CIS, del INE o de la AIMC, se encuentran disponibles en forma de bancos de datos en soporte magnético, para su explotación por terceros. Cabe esperar, por tanto, que se realicen diversos estudios donde los investigadores busquen respuestas a cuestiones de relación entre variables. Así, por ejemplo, los abundantes bancos de datos del CIS sugieren multitud de análisis de relación que den luz sobre comportamientos o perfiles de votantes, estilos de opinión, etc. Por otro lado, programas informáticos orientados al tratamiento de datos provenientes de encuestas, como DYANE (Santesmases, 1997) incluyen entre sus posibilidades, las PSHN.

Y ahora ¿Es un inconveniente el tamaño habitualmente grande de las muestras en las encuestas para utilizar PSHN? Si bien a la luz de los anteriores argumentos es fácil responder que sí (Johnstone y Lindley, 1995), nuestra respuesta no es coincidente: las PSHN no son *necesariamente* incompatibles con las muestras provenientes de encuestas, siempre que no se recurra a la hipótesis *cero*.

La PSHN ha sido defendida, aunque quizá no con la misma contundencia utilizada por sus detractores. Así, Frick (1996) sostiene que resulta idónea para el trabajo con declaraciones de tipo ordinal, donde el tamaño del efecto no es sustancial pero sí su dirección (A > B, por ejemplo); y ocurre que las leyes en ciencias humanas son predominantemente ordinales. En el mismo trabajo (pág 379) se indica con respecto a PSHN: "una vía de pensamiento que ha sobrevivido décadas de feroces ataques debe tener probablemente algún valor". Y su uso sigue recomendándose para la corroboración de teorías (Chow, 1998).

Una solución inmediata es acudir a valores no estrictamente nulos para H_0 (Kirk, 1996; Howlin, 1997). No obstante, esta medida cuenta también con dos inconvenientes inmediatos:

1. Como ya hemos mencionado, se fuerza al investigador a reflexionar acerca de qué punto de corte representa la existencia *conceptual* de relación (por correlación, o diferencia de medias o cualquier otro indicador estadístico)
2. Los programas de ordenador, pensados más para el éxito comercial que para la consecución adecuada de los objetivos, se confeccionan considerando la hipótesis cero y condicionan, con ello, el uso que los usuarios hacen de las PSHN. Utilizar otro valor desproteje al investigador frente a estos artilugios informáticos.

Ya se ha indicado que la inferencia estadística no elimina la subjetividad, sino que la arrincona en momentos determinados (Manzano, 1997). Eludir la responsabilidad de tomar decisiones juiciosas en lugar de "dejarse arrastrar" por las opciones automáticas, no es un comportamiento confesable por parte del investigador. Así, el investigador no puede manifestarse explícitamente incapaz de identificar un punto de corte (arbitrario, pero conceptualmente justificado).

El segundo inconveniente se puede solucionar estableciendo modificaciones en los valores de partida o utilizando aproximaciones aceptables a partir de los resultados. Por ejemplo, si se considera justificado definir un mínimo de variación compartida del 10% (Cortina y Dunlap, 1997, llegan a manejar el 1%), se está exigiendo una $|r_{xy}| \geq 0{,}316$. Todo resultado muestral que no cumpla con este requisito no permite continuar con la prueba (Chia, 1997). Por otro lado, en el ejemplo, las correlaciones críticas aproximadas se obtendrán añadiendo la cuantía 0,316 a los valores contenidos en la tabla 3, bajo H_0 definida como $\rho_{xy}=0$. Por ejemplo, si alguien concluye que dos variables están relacionadas en el estudio AIMC, es porque ha obtenido $r_{xy} \geq 0{,}3243$.

Un índice de significación relativa

En 1962, Jacob Cohen publicó un trabajo especial en el *Journal of Abnormal and Social Psychology* (Cohen, 1962). En sustancia, analizó las condiciones en las que los artículos de la revista realizaban su pruebas de hipótesis, encontrando por término general un bajo nivel de potencia. Siete años después, publicaba un libro sobre el análisis de la potencia (Cohen, 1969) del que cabría esperar un alto grado de influencia en las investigaciones empíricas. No obstante, Sedlmeier y Gigerenzer (1989) mostraban que casi tres décadas después de la publicación del artículo o dos décadas después del libro, la potencia de los estudios no había mejorado. Cohen (1992), preocupado, sigue preguntándose cómo un procedimiento tan importante para la comprensión de los resultados, la potencia de las pruebas estadísticas, no se ha extendido de tal forma que sea de esperado y obligado cumplimiento en cualesquiera publicaciones que se refieran a trabajos empíricos y partan de PSHN o utilicen directamente pruebas de hipótesis.

La potencia de una prueba estadística ($1-\beta$) se refiere a la sensibilidad que posee la prueba para captar la falsedad de la hipótesis nula. Si H_0 es falsa, β es la probabili-

dad de mantenerla (¡error!, riesgo que también es denominado error tipo II) y 1-β es la probabilidad de rechazarla. En el esquema de Fisher, no aparece β. Lo único de interés es α o error tipo I: si H_0 es verdadera, α es la probabilidad de rechazarla (¡error!) y 1-α es la probabilidad de mantenerla.

Para los no iniciados, el párrafo anterior es difícilmente inteligible. A la complejidad de la lógica subyacente y al manejo inevitable de las probabilidades condicionadas asociadas al proceso, hay que añadir la dificultad de los conceptos implicados. Por otro lado, los cálculos para la potencia no son fáciles y requieren, en muchos casos, del auxilio de un programa de ordenador. La mayoría de los estudios no están realizados por investigadores que gocen de una base metodológica (lógica, estadística, matemática, informática...) generosa. Los investigadores aplicados, por lo general, son especialistas en áreas de contenido sustantivo y les preocupa cómo utilizar las técnicas para tomar decisiones con respecto a sus estudios, pero no las técnicas en sí. Luego, las complicaciones no son bien recibidas y, consecuentemente, tampoco seguidas. Por tanto, no es de extrañar que las llamadas de atención de Cohen no hayan tenido la repercusión que esperaba.

Son muchos los autores (por ejemplo, Alberdi, Lorente y Moreno, 1969; Boulanguer, 1971; Manheim, 1982; Rossi, Wright y Anderson, 1983; Solanas y Sierra, 1992; Judd, McClelland y Culhane, 1995) de los que puede extraerse una imagen de los investigadores como usuarios que recurren a las técnicas estadísticas no con placer y pleno conocimiento, sino más bien arrastrados por las exigencias del entorno y agradecidos por las facilidades de los programas informáticos. El usuario de estas técnicas sigue las *hipótesis de comodidad* (Lebart, Morineau y Fénelon, 1985); es decir, toma decisiones metodológicas no porque sean las más adecuadas, sino porque se muestran como las más cómodas.

En este contexto, es comprensible que los esfuerzos por fomentar el uso adecuado de la metodología de investigación, cuando ésta implica mayor dedicación por parte del investigador, no tengan efectos ni sensibles ni inmediatos. ¿Qué hacer pues con la recomendación de que se debe huir de la *hipótesis cero* y reflexionar acerca de qué valores del parámetro deben considerarse para concluir que las variables están relacionadas? No hay cabida para grandes esperanzas, puesto que la dedicación en este tópico es aún mayor que la que se ha exigido para contemplar la potencia de las pruebas.

Así pues, quizá sea necesario elaborar estrategias que, si bien no gocen del mismo nivel de adecuación a los principios básicos de la investigación, sí resulten más fáciles o más atractivas o menos comprometidas para los investigadores. De esta forma, los problemas metodológicos no se resolverán con rapidez, pero se permite al menos dar un paso en esta dirección.

Con esta justificación, en este punto del trabajo, nos preocupa identificar una estrategia que permita seguir utilizando la PSHN con una *hipótesis cero*, pero redu-

ciendo el problema de la falsedad a priori de H_0 y generando un procedimiento que pueda ser automatizado y regulado mediante convenio (al igual que α=0,05 ó β=0,2)

En las PSHN con *hipótesis cero*, la significación estadística es una medida insuficiente de la fuerza de la relación de las variables implicadas. En este trabajo hemos observado cómo el tamaño de la muestra disminuye muy sensiblemente el valor de $p(D/H_0)$. Un reflexión inmediata puede ser «las muestras grandes hacen que los estudios sean artificialmente más significativos». Y una solución acorde y también inmediata: «se debería acotar el tamaño de la muestra, con un límite superior que permita no rechazar con comodidad hipótesis claramente falsas, con cuantías de relación minúsculas». No obstante, esta solución cuenta con algunos inconvenientes importantes:

1. Es una estrategia indirecta que resuelve artificialmente un problema, pero deja abiertos otros, restando potencia a la investigación.
2. Se encuentra en clara contradicción con el espíritu de la ley de los grandes números.
3. No genera una salida práctica. Por un lado se aconseja recoger muestras lo más grandes que permitan el control y los medios. Por otro, se aconsejaría que las muestras fueran lo más pequeñas posibles, de tal forma que los valores bajos de $p(D/H_0)$ tendrían un alto nivel de credibilidad.

Tal y como hemos mencionado, el comportamiento habitual se centra en observar el grado de significación que suministran los programas de ordenador y comparar éste con el nivel α. En muchas ocasiones, el investigador tampoco debe realizar este esfuerzo, puesto que las utilidades informáticas suelen facilitar la tarea señalando los efectos significativos con un determinado nivel α establecido por defecto. Mientras α permanece inmutable, al aumentar el tamaño de la muestra, la distribución muestral disminuye su varianza y los valores de los estimadores cercan con mayor precisión al parámetro que, como también hemos razonado, difiere del que defiende la hipótesis cero. Además del tamaño de la muestra, el tamaño del efecto o el tipo de prueba pueden modificar la facilidad o dificultad con que se acierta al rechazar la hipótesis nula. 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. ¿Qué significa $p(D/H_0)$=0,001? ¿Un resultado *muy significativo*? Es posible que se trate de un valor de relación sensiblemente bajo, pero con una prueba excepcionalmente potente. Obsérvese en la tabla 3, por ejemplo, que un valor r_{xy}=0,007 (enormemente difícil de encontrar en la práctica) es suficiente para concluir que existe relación, con α=0,001 y n=200000. Luego, $p(D/H_0)$=0,001 puede no estar diciendo nada.

Si la interpretación de $p(D/H_0)$ depende de la potencia de la prueba (1-β), una posible salida es expresar la primera en términos de la segunda. Este índice permitiría interpretar más fácilmente la confianza en una relación, a pesar de utilizar una hipóte-

sis cero. Así, nuestra confianza en una relación será tanto mayor cuanto mayor sea también el valor de $p(D/H_0)$ y menor el de $1-\beta$. Por ejemplo, un resultado $p(D/H_0)=0,01$ es más importante si $1-\beta=0,4$ que si $1-\beta=0,8$. Digamos que, *a pesar de contar con una baja potencia (es decir, a pesar de carecer de facilidades para rechazar H_0) se ha rechazado H_0*. Esta estrategia permite contrarrestar el efecto de la disminución de $p(D/H_0)$ con el aumento del tamaño de la muestra y retorna el interés por interpretar un índice indirecto para la *fuerza* de la relación, basándose en $p(D/H_0)$.

Una concreción del índice puede ser:

$$1-\frac{p(D/H_0)}{\beta} \tag{1}$$

Siempre que $p(D/H_0)<\beta$ (aspecto que es coherente con los convenios sobre los valores de α y β), la expresión (1) suministrará valores comprendidos entre 0 y 1. Si se utilizan los referentes habituales ($\alpha=0,05$ y $\beta=0,2$, por ejemplo en Cohen, 1992), la expresión (1) suministra el valor 0,75. Para facilitar la interpretación del índice es deseable que los valores-criterio se encuentren en puntos especiales del continuo. Para ello, realizaremos una transformación encaminada a forzar que los valores $p(D/H_0) = 0,05$ y $\beta = 0,2$ generen un resultado de valor 0,5:

$$\left(1-\frac{p(D/H_0)}{\beta}\right)^k = 0,5 \qquad k = \frac{\ln 0,5}{\ln 0,75} = 2,40942984 \tag{2}$$

En la expresión (2) el valor k 2,41 permite obtener 0,5 como resultado del cálculo para $p(D/H_0)=0,05$ y $\beta=0,2$. No obstante, los referentes numéricos pueden mejorarse si se fuerza otro punto de fácil manejo para las situaciones consideradas por convenio como *más exigentes* y que recurren a un $p(D/H_0)$ mínimo (α) de valor 0,01. Para ello:

$$v = \left[\left(1-\frac{p(D/H_0)}{\beta}\right)^k - \frac{1}{2}\right]^h + \frac{1}{2} = 0,75 \tag{3}$$

$$h = \frac{\ln 0,25}{\ln (0,95^k - 0,5)} = 1,44740683$$

La expresión (3), con k=2,40942984 y h=1,44740683 se ajusta al comportamiento que se pretende: el índice suministra valores que son tanto mayores conforme disminuye $p(D/H_0)$ (es decir, conforme aumenta la significación estadística) y aumente β, y

facilita la interpretación localizando puntos críticos en 0,5 y 0,75. Obsérvese que los valores altos del índice ν (nu) implican una alta significación a pesar de que existe un riesgo también alto de mantener la hipótesis nula. En otros términos: a pesar de los inconvenientes, existe significación suficiente. El tamaño de la muestra disminuirá tanto $p(D/H_0)$ como β, por lo que se habrá neutralizado su efecto en la interpretación indirecta de la fuerza de la relación. Una aproximación aceptable se puede establecer para k=2,41 y h=1,45. No obstante, ν pierde utilidad si se recurre a su cálculo manual, puesto que ello implica aún mayor esfuerzo por parte del investigador que reflexionar acerca de los tópicos mencionados hasta el momento.

La potencia de valores negativos genera algunos problemas durante la ejecución de los programas. Luego, para la computación de ν es recomendable utilizar el algoritmo expuesto en el cuadro 2.

Por otro lado, debido a la prioridad por facilitar la interpretación de los valores-criterio, ν pierde el cero como cota inferior, puesto que al coincidir los valores de $p(D/H_0)$ y β,

$$\nu = -0{,}5^{h} + 0{,}5 \approx 0{,}13.$$

Cuadro2: *computación de ν*

1. Si $\beta < p(D/H_0)$, no calcular ν
2. $k = 2{,}40942084$; $h = 1{,}44740683$
3. $f = (1-p(D/H_0)/\beta)^{k}-{,}5$
4. si $f<0$ entonces $s=-1$ y $f=-f$; en otro caso, $s=1$
5. $\nu = s \cdot f^{h}+{,}5$

Cumpliendo, pues, con el objetivo de elaborar estrategias que permitan adecuar los comportamientos habituales de los investigadores, implicándoles un esfuerzo mínimo y con base en los valores de α y β establecidos por convenio, ν puede ser interpretado, mediante porcentajes (ν' = 100ν, véase la tabla 4 para algunos ejemplos concretos), con acuerdo a los siguientes puntos (traducción directa del hábito en el uso de la significación estadística):

- ν' ≤ 50: Resultado poco significativo. O bien $p(D/H_0)$ suministra un valor bajo, o la prueba es tan potente que el grado de significación es un resultado necesario, derivado no de la fuerza de la relación sino de la falsedad a priori de la hipótesis cero.
- 50 < ν' ≤ 75 Resultado significativo. Es posible que $p(D/H_0) \geq \alpha$, pero con una potencia tan baja, que el grado de significación obtenido es altamente meritorio.
- ν' > 75 Resultado muy significativo. $p(D/H_0)$ es relativamente bajo, a pesar de un valor de potencia que puede ser elevado.

Tabla 4: *Significación relativa v en función de algunos valores de p(D/H$_0$)* y β

v		Grado de significación $p(D/H_0)$						
		,1	,075	,05	,025	,01	,005	,001
Error tipo II (β)	,05	---	---	13,33	31,49	52,78	65,50	81,73
	,075	---	13,33	20,61	45,15	60,33	71,60	83,34
	,1	13,33	17,03	31,49	50,00	65,50	75,00	84,16
	,2	31,49	41,79	50,00	61,54	75,00	80,56	85,40
	,3	45,15	50,00	56,08	68,43	78,64	82,53	85,82
	,4	50,00	53,90	61,54	72,43	80,56	83,54	86,03
	,5	52,78	58,09	65,50	75,00	81,73	84,16	86,16

El esquema anterior puede no resultar suficiente. Las situaciones en las que v'>50 y se interprete una significación satisfactoria, pueden ir acompañadas de $p(D/H_0)>\alpha$. Una solución a esta incomodidad conceptual es aplicar el esquema anterior sólo en las ocasiones en las que el investigador fuera a rechazar la hipótesis nula, al obtener $p(D/H_0)<\alpha$. El procedimiento general no es coherente en todo momento con la información que suministra v, pero el resultado es interpretable como una estrategia conservadora. En otros términos: sólo cuando el resultado se juzga como significativo procede aplicar este procedimiento y evaluar en qué medida es fuerte o suficiente la significación, insertándola en un contexto más amplio. Con ello, el proceso global es:

1. Calcular $p(D/H_0)$ y β.
2. Comparar $p(D/H_0)$ con α.
3. Si $p(D/H_0)\geq\alpha$, se mantiene H_0 y se especifica β o 1-β. Fin del proceso.
4. Calcular v.
5. Utilizar el esquema anterior para interpretar los resultados en función de v.

Para que este procedimiento resulte finalmente útil, es necesario que sea implementado en programas de análisis de datos, de tal forma que el algoritmo anterior sea llevado a cabo directamente por el sistema informatizado. Es fácil observar que su automatización es inmediata e implica únicamente contemplar en el programa más opciones por defecto: riesgo β=0,2 y tamaño medio para el efecto estandarizado (Cohen, 1988). De esta forma, si no podemos evitar que el investigador-usuario interprete sus resultados únicamente por un valor de probabilidad (grado de significación) o por el número de asteriscos con que el programa marca los resultados, ni que

incluso este comportamiento sea exigido por un gran número de revistas que aceptan trabajos empíricos, sí habremos conseguido mitigar en parte la inadecuación del proceso.

Referencias

AIMC (1999) *Estudio General de Medios*. Información obtenida de la dirección electrónica «http://www.aimc.es/aimc/html/egm/caracteristicas.html»

Alberdi, J.; Lorente, S.; y Moreno, E. (1969) *Metodología de investigación por muestreo*. Madrid: Euramérica.

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

Berkson, J. (1938) Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33, 526-542.

Barry, T.M. (1996). Recommendations on the testing and use of pseudo-random number generators used in Monte Carlo analysis for risk assessment. *Risk Analysis*. 16 (1), 93-105.

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)

Boulanguer, G. (1971) *La investigación en ciencias humanas*. Madrid: Marova.

CIS (1996) *Encuesta preelectoral a las elecciones generales de 1996*. Información obtenida de la dirección electrónica «http://www.cis.es/cgi-bin/contenidobin ?nest=2207»

CIS (1999) *Barómetro de Junio de 1999*. Información obtenida de la dirección electrónica «http://www.cis.es/baros/frame.html»

Cohen, J. (1962) The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology*, 65, 145-153.

Cohen, J. (1969) *Statistical power analysis for the behavioral sciences*. Hillsdale (New Jersey): Erlbaum.

Cohen, J. (1992) A power primer. *Psychological Bulletin*, 112, 155-159.

Cohen, J. (1994) The earth is round ($p<.05$). *American Psychologist*, 49, 997-1003.

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

Chatfield, C. (1985) The Initial Examination of Data. *Journal of the Royal Statistical Society*, Series A, 148, 214-253.

Chia, K.S. (1997) "Significant-itis": an obsession with the p-value. *Scandinavian Journal of Work, Environment and Health*, 23, 152-154.

Chow, S.L. (1988) Significance test or effect size? *Psychological Bulletin*, 103, 105-100.

Chow, S.L. (1998) Precis of statistical significance: rationale, validity, and utility. *Behavioral and Brain Sciences*, 21, 169-239.

Estes, W.K. (1997) Significance testing in psychological research: some persisting issues. *Psychological Science*, 8, 18-20.

Fisher, R.A. (1956) *Statistical methods and scientific inference*. Edimburgo: Oliver and Boyd.

Fisher, R.A. (1966) *The design of experiments*. (2ª Edición) Edimburgo: Oliver and Boyd.

Fishman, G.S y Moore, L.R. (1982). A Statistical Evaluation of Multiplicative Congruential Random Number Generators with Modulus 2^{31}-1. *Journal of the American Statistical Association*, 77 (377), 129-136.

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

Gutiérrez Cabría, S. (1994) *Filosofía de la estadística*. Valencia: Servei de Publicacions de la Universitat de València.

Harris, M.J. (1991) Significance test are not enough: the role of effect-size estimation in theory corroboration. *Theory & Psychology*, 1, 375-382.

Harvatopoulos, Y.; Luvan, Y.F. y Sarnin, Ph. (1992) *El arte de la encuestas. Principios básicos para no especialistas*. Bilbao: Ediciones Deusto.

Hedges, B. (1980) Sampling. En G. Hoinville y R. Jowell. *Survey research practice*. London: Heinemann Educational Books, pp. 55-89.

Howlin, D. (1997) When is a significant change not significant? *Journal of Autism and Development disorders*, 27, 347-348.

Hunter, J.E. (1997) Needed: a bon on the significance test. *Psychological Science*, 8, 3–7.

Huxley, P. (1988) 'Quantitative-Descriptive' articles in the British Journal of Social Work 1-14. *British Journal of Social Work*, 18, 189-199.

Hyman, H. (1970) *Diseño y análisis de las encuestas sociales*. Buenos Aires: Amorortu Editores.

INE (1998) *Estudio de población activa*. Información obtenida de la dirección electrónica «http://www.ine.es/htdocs/daco/daco.htm»

Irurita, I.M. (Compilador) (1996) *Estudio sobre la prevalencia de los jugadores de azar en Andalucía*. Sevilla: Comisionado para la droga (Junta de Andalucía).

Johnstone, D.J. y Lindley, D.V. (1995) Bayesian inference given data "significant at a" tests of point hypothesis. *Theory and Decision*, 38, 51-60.

Judd, Ch.M.; McClelland, G.H. y Culhane, S.E. (1995) DATA ANALYSIS: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, 46, 433-465.

Kenny, D.A. (1979) *Correlation and causality*. New York: Wiley.

Kirk, R.E. (1996) Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.

Lebart, L.; Morineau, A. y Fénelon, J.P. (1985) *Tratamiento estadístico de datos. Métodos y programas*. Barcelona: Marcombo.

Manheim, H. (1982) *Investigación sociológica. Filosofía y métodos*. Barcelona: CEAC.

Manzano, V (1997) Usos y abusos del error tipo I. *Psicológica*, 18, 153-169.

Mendoza, R; Sagrera, MR y Batista, JM (1994) *Conductas de los escolares españoles relacionadas con la salud (1986-1990)*. Madrid: Centro Superior de Investigaciones Científicas.

Morrison, D.E. y Henkel, R.E. (Editores) (1970) *The significance test controversy.* Chicago: Aldine.

Neyman, J. y Pearson E.S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Transactions of the Royal Society*, Series A, 231, 289-337.

Noelle, E. (1970) *Encuestas en la sociedad de masas. Introducción a los métodos de la demoscopia.* Madrid: Alianza Editorial.

Oakes, M. (1986) *Statistical inference: a commentary for the social and behavioral sciences.* New York: Wiley.

Pérez Cebrián, F. (1987) *La planificación de la encuesta social.* Zaragoza: Secretariado de Prensas Universitarias de la Universidad de Zaragoza.

Rojas, A.J.; Fernández Prados, J.S. y Pérez Meléndez, C. (1998) *Investigar mediante encuestas. Fundamentos teóricos y aspectos prácticos.* Madrid: Síntesis.

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

Rossi, P.H.; Wright, J.P. y Anderson, A.B. (1983) Sample survey: history, current practice, and future prospect. En P.H.Rossi, J.P.Wright y A.B.Anderson, *Handbook of Survey Research.* Orlando (Florida): Academic Press Inc. pp 1-20.

Rozeboom, W.W. (1960) The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 86, 638-641.

Sánchez Bruno, A. y San Luís-Costas, C. (1995). A statistical analysis of seven multipliers for linear congruential random numbers generators with modulus 2^{31}-1. *Quality & Quantity*, 29, 331-337.

Santesmases, M. (1997) *DYANE: Diseño y análisis de encuestas en investigación social y de mercados.* Madrid: Pirámide.

Seco, M. (1997) *Diccionario de dudas y dificultades de la lengua española.* Madrid: Espasa.

Sedleimer, P. y Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.

Solanas, A. y Sierra, V. (1992) Bootstrap: fundamentos e introducción a sus aplicaciones. *Anuario de Psicología*, 55, 143-154.

Tukey, J.W. (1991) The philosophy of multiple comparisons. *Statistical Science*, 6, 100–116.

Vacha, T. y Ness, C.M. (1999) Statistical significance testing as it relates to practice: use within Professional Psychology: Research and Practice. *Professional Psychology: Research and Practice*, 30, 104-105.

Vacha, T. y Nilsson, J.E. (1998) Statistical significance reporting: current trends and use in MECD. *Measurement and Evaluation in Counseling and Development*, 31, 46-47.

Valera, A. y Sánchez Meca, J. (1997) Pruebas de significación y magnitud del efecto: reflexiones y propuestas. *Anales de Psicología*, 13, 85-90.

Vallecillos, A. (1996). *Inferencia estadística y enseñanza: un análisis didáctico del contraste de hipótesis estadísticas.* Granada: Comares.

Wainer, H. (1999) One cheer for null hypothesis significance testing. *Psychological Methods*, 4, 212-213.

Watts, D.G. (1991) Why is Introductory Statistics Difficult to Learn? And What Can We Do to Make It Easier? *The American Statistician*, 45, 290-291