Más de 50 años de Controversia del Contraste de Hipótesis: ¿una luz al final del túnel?

Dr. Gustavo Peña Torbay

Resumen

En este artículo se presenta una revisión conceptual de la llamada *controversia del contraste de hipótesis*, es decir, de los problemas que confronta el *contraste estadístico de hipótesis*, tanto desde el punto de vista de sus detractores como de sus defensores; a partir de lo cual se puede afirmar que el dilema entorno a los problemas inherentes al contraste de hipótesis no está solucionado. Luego, teniendo en cuenta que no parece existir una solución mágica a esta diatriba, se indican algunas de las propuestas recientes que pretenden mejorar la técnica del *contraste de hipótesis*, para, finalmente, presentar algunas recomendaciones en cuanto a qué hacer, a nivel operativo y en lo que toca a la docencia del *contraste de hipótesis*, con la idea de evitar perder lo bueno del procedimiento y agregarle elementos en pro de mejorar las conclusiones a las cuales se podría llegar mediante su empleo.

Palabras Clave: Contraste de Hipótesis, Enseñanza, Psicología.

Introducción

> El contraste de hipótesis es un método controversial de extraer información a partir de datos experimentales y de orientar la elaboración de conclusiones científicas. (Nickerson, 2000, p. 289)

> Sin embargo, es bastante evidente la casi abrumadora omnipresencia de las pruebas de significación estadística de hipótesis. (Mateo, 2001, p. 199)

La Escuela de Psicología de la Universidad Católica Andrés Bello (UCAB) comenzó sus actividades en el año de 1957, el 1º de Octubre, siendo la segunda escuela de su tipo en Venezuela, el año anterior había iniciado sus labores la escuela de la Universidad Central de Venezuela (UCV).

Cuando inicio sus actividades, el enfoque de la escuela era de corte psiquiátrico, una postura que luego pasó a llamarse *clínica*, con un programa de estudios que incluía pocas asignaturas referidas a la *metodología para la investigación*. Pero, al poco tiempo fue cambiando el panorama y ya para la segunda mitad de la década de los 60', para ser exactos en 1966, el pensum implicaba una carrera de 5 años y una fuerte carga de materias relacionadas con la epistemología, desde basamentos en *filosofía de la ciencia* hasta contenidos *estadísticos* y de *diseño de investigación* propios de cualquier especialidad de las ciencias empíricas (Gómez, 2003).

Este vuelco se dio a la par de cambios similares en muchas de las escuelas de psicología de América Latina (p.ej. en México, Brasil), respondiendo a una presión general sobre la disciplina que la llevaría a derroteros en lo metodológico que prontamente hicieron que algunos autores hablaran de la *cuantofrenia* en psicología (Koch, 1981); pero, como en casi todo itinerario disciplinar, esencialmente pendular, en los 90' el ambiente general de la psicología se movía en la dirección opuesta y bien se podía hablar para la época más bien de *cuantofobia* (Peña, 1995).

De hecho, en los 90' eclosionaron los textos en contra del *contraste de hipótesis*, uno de los pilares clásicos de la metodología cuantitativa; salvo que, los reproches no devenían tan solo de parte de los emergentes autores cualitativos, paradójicamente, los mismos partidarios de la cuantificación dieron los argumentos más dolorosos en desfavor de la estimación de la significación estadística y su uso en la investigación psicológica.

La crítica se debía, esencialmente, al embrollo de la lógica del contraste de hipótesis y al mal uso de la técnica por parte de los investigadores. Un tema que, por otra parte, no era nuevo para ese momento, se puede decir que desde el momento mismo de su aparición el método estadístico de inferencia ha sido criticado (Greenwald, 1993).

Esta situación llevó a que en el congreso de la American Psychological Association (APA), realizado Agosto de 1997, en la ciudad de Chicago, el tema de si se debía o no abandonar el uso del contraste de hipótesis fuera uno de los dilemas salientes (Cohen, 1997; Thompson, 1997). Aun más, poco después, como lo dicen Wilkinson et als (1999), "el Board of Scientific Affairs (BSA) de la American Psychological Association (APA) acordó la creación de un comité denominado el Task Force on Statistical Inference (TFSI) cuya misión era 'dilucidar algunos de los temas controversiales en relación a la aplicación de la estadística incluyendo el contraste de hipótesis y sus alternativas'" (p. 594. Comillas en el original).

Luego de todo esto, ya en el curso del nuevo siglo, el XXI, la diatriba ha seguido, no sólo en el contexto de la psicología, sino que además se ha extendido a otras disciplinas, como, por ejemplo, la educación, la medicina y la ecología (Anderson, Burnham y Thompson, 2000; Snijders, 2001; Gliner, Leech y Morgan, 2002; Singh, 2006).

Teniendo en cuenta que la situación dilemática de uno de los aspectos que usualmente se asume como determinante en la construcción del conocimiento científico y, de suyo, del saber de la psicología, sigue vigente y, con ello, amenaza a la disciplina, bien vale la pena dedicarle un tiempo al tema y, sobre todo, adelantar si hay alguna *luz al final del túnel.*

Con esta idea, en el presente trabajo, primero, se revisará el sentido esencial de la disputa en torno al contraste de hipótesis; segundo, se propondrán algunas formas de, al menos, paliar sus efectos sobre el hacer científico de disciplinas empíricas como la psicología; y, tercero, se rematará con algunas recomendaciones.

La controversia

Es un decir común que el objetivo de la ciencia es la formulación de relaciones de dependencia entre las variables; las cuales permitirían, en principio, una adecuada predicción de los eventos de estudio, y luego una explicación de los determinantes, es decir, la identificación de las causas (Kerlinger y Lee, 2001; Tabachnick y Fidell, 2007; Thompson, 2008).

El problema radica en cómo hacerlo, porque, de un lado, los dilemas filosóficos, y del otro los aspectos operativos, no hacen de esta una tarea fácil (Rivadulla, 1991); a lo cual hay que agregarle el hecho de que el desarrollo de

todo proyecto requiere de masa crítica y esta es usualmente opuesta a la complejidad, así que habitualmente se recomienda el uso de un proceder sencillo (Brenner-Golomb, 1993).

La búsqueda de tal procedimiento se puede remontar a 1700, de hecho, la primera prueba estadística que se publicó se debe a John Arbuthnot (1667-1735) en 1710, quien estaba fascinado con el hecho de que las proporciones de nacimientos humanos por género fueran tan similares (Kostenko y Hyndman, 2008); pero, la forma actual del modo de trabajar, el *contraste de hipótesis*, se desprende, inicialmente, del trabajo de K. Pearson (1857-1936) quien en 1900 propuso la muy conocida prueba de chi-cuadrado; y luego se desarrollaría, entre 1920 y 1940, como producto del trabajo, esencialmente, de tres destacados estadísticos: R. Fisher (1890-1962), E. Pearson (1895-1980) y J. Neyman (1894-1981) (Stigler, 1986).

Luego, al correr del tiempo y de un modo apresurado, entre 1940 y 1960 (Kline, 2004), con el trabajo de otros autores, como, por ejemplo, J. Guilford (1897-1987) y G. Snedecor (1881-1974), el esfuerzo redundó en el desarrollo y consagración del *método de contraste de la hipótesis nula* o, como se lo suele llamar de modo más corto, *el contraste de hipótesis*, una versión de la inferencia estadística que ha pasado a ser la forma más aceptada y usada de inferencia en la ciencia (Stigler, 1999; Peña, 1995). De hecho, para Fidler (2006, p. 3) "el contraste de hipótesis estaba ya completamente institucionalizado en las revistas, los libros de textos y los programas de enseñanza de psicología a mediados de los '50".

Para dar una idea de lo instaurado que está este procedimiento, "se puede decir que es el método más usado para el análisis de datos recabados en experimentos en psicología y que ha sido así desde hace unos 70 años" (Nickerson, 2000, p. 241); de un modo más específico, esta forma de argumentar "se emplea para justificar las conclusiones a partir de los datos en un 90% de los artículos de las revistas más importantes en psicología" (Loftus, 1991, p. 102).

Más recientemente Cumming et al (2007, p. 231) indican específicamente,

> Nosotros examinamos 10 publicaciones periódicas líderes en psicología que editan principalmente investigaciones empíricas... en tres periodos [se tomaron]: los 40 primeros artículos de 1998; para 2003-2004 los 40 artículos más recientes... y para 2005-2006 los 40 artículos más recientes alrededor de Abril de 2006.

El contraste de hipótesis se empleó en casi todos los artículos (97.8%, 97.7%, 96.6%, en los tres lapsos, respectivamente).

Ahora bien, aun cuando el uso del *contraste de hipótesis* es la forma más popular de inferencia en las investigaciones, no deja de tener críticas o, mejor dicho, no deja de tener muchos críticos; de hecho, "la literatura en esta área es muy grande" (Kline, 2004, p. 61). En 1997 David Parkhurst compiló 42 autores puestos en contra del procedimiento; por su parte, David Anderson en el mismo año recopiló 97 referencias adversas; Bill Thompson, en el 2001, recabó 402 referencias opuestas al *contraste de hipótesis*; y, finalmente, el que más, Fidler (2006, p.3) acota, "en mi registro de notas bibliográficas cuentan 700 referencias [adversas] en psicología".

Entre tanto material en disfavor vale destacar algunas citas, sólo como ejemplo de los argumentos que se esgrimen usualmente en contra del *contraste de hipótesis*:

> El énfasis que se le ha dado al contraste formal de hipótesis (R.A. Fisher)... ha causado que los investigadores le presten una atención indebida a los resultados de las pruebas de significación que realizan con los datos, particularmente con los que derivan de los experimentos, y muy poca a la estimación de la magnitud de los efectos que investigan. [...] El énfasis en las pruebas de significación y la consideración de los resultados de cada experimento de modo aislado, han tenido consecuencias muy infortunadas en el trabajo de los investigadores, quienes usualmente plantean el cálculo del contraste de hipótesis como el objetivo último de los experimentos. (Yates, 1951, p. 24)

> ¿Cuál es la probabilidad de obtener una persona muerta (D) dado que la persona fue ahorcada (H); en forma simbólica, cuál es la p(D|H)? Obviamente su valor es muy alto, quizás 0.97 o más. Ahora, revirtamos la pregunta: ¿Cuál es la probabilidad de que una persona haya sido ahorcada (H) dado que está muerta (D), es decir, cuál es la p(H|D)? En este caso la probabilidad sería, indudablemente, muy baja, quizás 0.01 o menor. Nadie debería estar dispuesto a sustituir la primera estimación (0.97) por la segunda (0.01); es decir, aceptar 0.97 como la probabilidad de que una persona fue ahorcada ya que está muerta. Aun cuando se podría pensar que este es un error muy infrecuente, este es exactamente el tipo de equivocación que se comete en la interpretación de las pruebas de significación estadística cuando, analógicamente, los valores estima-

> dos de la p(H|D) se interpretan como cálculos de la p(D|H). (Carver, 1978, p. 385)
>
> Argumentaremos en contra [de la significación de] la prueba, llegando a la conclusión de que su uso se abandone en toda la ciencia substantiva y no sólo en la investigación educativa y otras ciencias sociales que han comenzado a levantar voces contra la tiranía virtual de esta rama de la inferencia en el mundo académico. (Guttman, 1985, 4)
>
> Una pequeña reflexión revela un hecho extensamente entendido entre los estadísticos: la hipótesis nula, tomada literalmente (y ésa es la única manera en que usted puede tomarla en la prueba formal de hipótesis), siempre es falsa en el mundo real.... Si es falsa, incluso a un grado minúsculo, debería darse que una muestra lo suficientemente grande producirá un resultado significativo y conducirá a su rechazo. Por tanto, ¿si la hipótesis nula siempre es falsa, por qué el afán por rechazarla? (Cohen, 1990, p. 1306)

Y, como estas, muchas más, cuyos argumentos en contra del contraste de hipótesis se pueden reunir en dos grandes grupos. Uno relacionado con la dimensión estadística del contraste; y el otro referido a los aspectos de corte metodológico. Es interesante recordar, la visión de R. Fisher al respecto de la investigación era precisamente como una entidad dual: "el procedimiento estadístico y el diseño experimental son sólo dos aspectos diferentes de un mismo todo, el cual comprende todos los requerimientos lógicos del proceso integral para acrecentar el conocimiento empírico mediante la experimentación" (Fisher, 1935/1971, p. 3).

Concretamente, en lo que toca a la dimensión estadística, el problema se puede centrar, en dos sub-dimensiones. Por una parte, como señala Kirk (1996)

> En la inferencia científica, lo que deseamos saber es la probabilidad de que la hipótesis nula (H_0) sea cierta dado que se ha obtenido un cierto conjunto de datos (D); esto es, $p(H_0/D)$. Lo que el contraste de la hipótesis nula indica es la probabilidad de obtener esos datos o unos más extremos si la hipótesis nula es cierta $p(D/H_0)$. (p. 747)

Y, por la otra parte, como lo indica Gigerenzer (1993),

> Lo que se institucionalizó como *inferencia estadística* en psicología no es estadística Fisheriana. Es un amasijo incoherente de algunas de las ideas de Fisher por una parte, y algunas de las ideas de Neyman y de E. S. Pearson por la otra. Yo me refiero a esta mezcla como la "lógica híbrida" de la inferencia estadística. Fisher, Neyman, y Pearson la habrían rechazado, aunque por razones diferentes. (p. 314. Comillas y cursivas en el original)

Según Snijders (2001, p. 7122),

> Ejemplo de este carácter híbrido es que, de acuerdo con la postura Neyman/Pearson, la teoría del contraste de hipótesis se explica haciendo referencia a las hipótesis nula y alternativa, y a los errores tipo I y II, mientras que, en el espíritu del enfoque Fisheriano, los test estadísticos se reservan como procedimientos para obtener evidencia sobre una hipótesis particular que se intenta probar y no como reglas de comportamiento que aplicadas a largo plazo tienen una cierta tasa de error cuando se trabaja con un gran número de hipótesis y de conjuntos de datos.

En lo que se refiere a las críticas a nivel metodológico, según Granaas (2002. p. 2)

> Ya que las hipótesis de los investigadores en psicología usualmente asumen la existencia de un efecto de cierta magnitud, si bien esta última es desconocida, la hipótesis nula de "no efecto", una hipótesis nihilista, es un prefabricado opuesto a la hipótesis de investigación [ya que la limita a cero]. Así, para la mayoría de los investigadores en psicología la hipótesis nihilista es *la* hipótesis nula [es decir, asume que no hay efecto alguno del tratamiento o que la correlación es cero].
>
> Este enfoque acarrea tres problemas:
>
> - La hipótesis nula se propone como una hipótesis falsa a priori. Es falsa, en consecuencia, su rechazo ciertamente no se pone en duda. La única pregunta es si el investigador ha recogido suficientes datos como para demostrar este resultado predeterminado. El rechazo de tal hipótesis nula no provee de evidencia adicional y, por tanto, no permite progresar.
> - Rechazar la hipótesis nula sólo "prueba" la hipótesis de investigación en un sentido débil.

La práctica de rechazar el "no efecto" [H_0] a favor de "cierto efecto" [H_1] no incentiva al investigador a determinar el tamaño y naturaleza del efecto.

Krantz (1999, p. 1374), quien es estadístico, dice que la lectura de los argumentos en contra de la hipótesis nula en el ámbito de la psicología le llevan a decir que para los psicólogos:

> (1) el contraste de hipótesis es un ritual, (2) el contraste de hipótesis substituye a la inspección de los datos, (3) se confunde la $p(D/H_0)$ con la $p(H_0/D)$, (4) hay sobre-optimismo en relación a la replicabilidad, y (5) se confunde el rechazo de la hipótesis con la confirmación de una teoría. Además, otro error, vinculado con el (4) y el (5), que merece mencionarse: (6) se interpreta de modo erróneo la imposibilidad de rechazar la H_0 como una falla en replicar un estudio anterior.

Dadas estas limitaciones que se le atribuyen al contraste de hipótesis, no es de extrañar que se diga que este "no solo ha fallado en soportar y promover a la psicología como ciencia sino que la ha impedido gravemente" (Cohen, 1994, p. 997); es decir, "el contraste estadístico de hipótesis retarda el crecimiento del conocimiento científico; nunca ha hecho contribuciones positivas" (Schmidt y Hunter, 1997, p. 37); y representa "el hueso más pesado alguna vez institucionalizado en el plan de entrenamiento de los estudiantes" (Rozeboom, 1997, p. 335).

Ahora bien, no se debe descuidar que este es un tema que bien podría considerarse *mediático*, en otras palabras, que vive a la luz pública, siempre expuesto, de modo que, en buena medida, se modela a partir de las publicaciones que al respecto hacen los estudiosos del área. Ya se han dado pistas del modo y extensión en que la polémica aparece en las publicaciones periódicas especializadas. Pero, cómo se aborda esta diatriba en los textos de enseñanza, el otro medio de comunicación donde podría relucir la confrontación; aun más, cuál ha sido el efecto de esta diatriba en el modo de enseñar el contraste de hipótesis en los cursos de estadística.

Para comenzar, vale la referencia de Nickerson (2000)

> El debate acerca del *contraste de hipótesis* tiene sus raíces en desacuerdos irresolubles entre las mayores contribuciones al desarrollo de las teorías de la inferencia estadística en las cuales se basan los enfoques actuales. Gigerenzer et al (1989) revisaron de modo detallado la controversia entre R. A. Fisher por una parte y Jerzy Neyman y Egon Pearson por la otra,

> así como el desacuerdo entre estos dos puntos de vista y los seguidores de Tomas Bayes. Ellos se percataron del notorio hecho de que muy poco de la historia y de la controversia en sí misma se encuentra reseñada en los textos para la enseñanza de la estadística. La falta, por un lado, de una perspectiva histórica bien afinada y, por el otro, de una adecuada comprensión de la compleja y, en ocasiones, controversial justificación filosófica de varios de los enfoques de la estadística inferencial, serían razones suficientes para la aparente facilidad con que los textos de estadística son mal utilizados y mal entendidos. (p. 290)

Siguiendo con la idea, Gliner, Leech y Morgan (2002) seleccionaron "seis libros de texto usados en pregrado y otros seis empleados a nivel de maestría y de doctorado" (p. 87), para evaluar en qué medida se está difundiendo y dando a entender el tema de la controversia respecto del contraste de hipótesis. Luego de su análisis los autores concluyeron:

> Lo más desilusionante fue encontrar que la mayoría de estos textos no reconocen la existencia de la controversia en torno al contraste de hipótesis [...] ¿por qué esta discrepancia entre los muchos artículos que señalan los problemas con el contraste de hipótesis y los textos de enseñanza? Nosotros sugerimos tres explicaciones a este aparente desacuerdo [...] La primera tiene que ver con que la mayoría de los textos son revisiones, de tercera a sexta edición, de obras publicadas originalmente antes de los años 90' cuando el contraste de hipótesis era un tema menos controversial [...] La segunda se refiere al nivel de profundidad, dificultad de los conceptos, y los conocimientos previos de los alumnos [...] Nuestra tercera posible explicación se relaciona con la mejor práctica. Aunque se conoce que los tópicos que nosotros indagamos en los textos de metodología de la investigación y de estadística se deberían abordar, no existe acuerdo en cómo se debería hacer. (p. 90-91)

Dado el punto anterior, bien vale la pena un comentario relacionado con el estatus de la docencia de la estadística en los cursos universitarios, sobre todo en relación a si se toma en cuenta el tema de la disputa relativa al contraste de hipótesis.

En cuanto al contenido y extensión de los cursos de formación en estadística, Aiken, West y Millsap (2008),

> ... en un estudio tipo encuesta de todos los programas de PhD en psicología de Estados Unidos y Canadá... al cual respondieron 201 programas

> (el 86%)... [encontraron que] la mayoría de los programas apoyan el entrenamiento para estudios de laboratorio y no para la investigación de campo. La mediana de la duración del entrenamiento en estadística y teoría de la medición fue de 1.6 años y la mayor parte de este tiempo (1 año) se dedica a cursos introductorios en estadística, con lo cual se deja poco sitio para estudios avanzados.

Y, además, en lo que respecta a la enseñanza de la estadística, en los cursos universitarios de psicología, el entrenamiento para el *contraste de hipótesis* se da, sin mayores variaciones, en la *forma tradicional* desde mediados de los '50 al presente; de modo que, se podría decir que las diversas recomendaciones que se han hecho desde los '90, respecto de modificar el modo de adestramiento, no han impactado mucho en el ámbito académico de la enseñanza de la estadística en psicología (del Puerto, Seminara y Minnaard, 2007; Díaz y Batanero, 2007; Fidler, 2006; Batanero y Díaz, 2006).

Ahora bien, además de las críticas ya citadas hay otras más, entre ellas (Mateo, 2001, p. 193-195)

- Las pruebas de significación estadística de hipótesis se utilizan, de manera simplista e indebida, como procedimientos de decisión automatizados.
- Prácticamente nunca se tienen en cuenta los supuestos asociados.
- Las pruebas de significación estadística de hipótesis no proporcionan la información que se requeriría para dar cuenta de los fenómenos que interesan a la psicología.
- El valor fijado para el nivel de significación (en la práctica, muy a menudo $\alpha = 0.05$) es arbitrario.
- El tamaño muestral es arbitrario.
- Los valores de p y de α son indebidamente interpretados como medidas de significación.
- Las pruebas de significación de hipótesis presentan habitualmente una potencia muy pequeña (que, por otra parte, no se tiene en cuenta para nada).
- La significación estadística no informa la probabilidad de que la hipótesis sea cierta. Ni tampoco del verdadero valor del parámetro.
- Los errores Tipo I y Tipo II están inversamente relacionados. Los investigadores ignoran el error Tipo II al prestarle una atención indebida al error Tipo I.

Finalmente, los defectos achacados al *contraste de hipótesis* han promovido, en el tiempo, dos tipos de ataques esenciales hacia él (Sawilowsky, 2003, p. 467):

> El primero es la proposición de que el uso del contraste de hipótesis se deseche a favor de los intervalos de confianza [...] Un aspecto secundario de este ataque es la propuesta de que el contraste de hipótesis sólo es aceptable si es (a) reforzado con el reporte del tamaño del efecto, (b) acompañado de un despliegue gráfico, o (c) Bayesiano.
>
> El segundo estilo de ataque es que el contraste de hipótesis debe descartase por razones de tipo filosófico. Un ejemplo de esto está en la pregunta si la ciencia se ha beneficiado en algo con el contraste de hipótesis.

Sin embargo, si bien es cierto que hay muchas objeciones, no es menos cierto que la afirmación, de uno de sus detractores, "después de cuatro décadas de severas críticas, el ritual del contraste de la hipótesis nula –una bisección mecánica en torno al criterio del .05- aún persiste" (Cohen, 1994, p. 998), se puede considerar vigente; como se ha dicho previamente, es "evidente la casi abrumadora omnipresencia de las pruebas de significación estadística de hipótesis" (Mateo, 2001, p. 199), en los cálculos de Cumming et al (2007), no menos del 97% de los artículos lo emplean.

Ante esta situación cabe preguntarse ¿por qué ha sobrevivido el *contraste de hipótesis* a tal cúmulo de ataques? Claro está, esta es una pregunta para la cual no existe, al momento, una respuesta clara y tajante; pero, como es usual, se pueden esbozar varias soluciones parciales.

En primer lugar, hay que aceptar que algunos de los males, quizás no pocos, que se le achacan al *contraste de hipótesis* no son tanto deficiencias del método, como errores de los usuarios, "las concepciones erradas del contraste de hipótesis no son culpa del método" (Kline, 2004, p. 80); por ejemplo, entre las actitudes inconvenientes de los usuarios, quizás la más usual sea la que vieron hace 60 años Hotelling, Bartky, Deming, Friedman y Hoel (1948, p. 103): "desafortunadamente, a muchas personas les gusta hacer el trabajo estadístico del mismo modo que rezan sus oraciones, mediante una fórmula que está en un libro muy respetado el cual se escribió hace mucho tiempo".

Por su parte, Nickerson (2000) afirma que "algunos de los defensores del contraste de hipótesis reconocen que las concepciones erradas de los usuarios

y, en consecuencia, el uso inconveniente del procedimiento, lleva a muchos a etiquetar inadecuadamente a la táctica" (p. 277).

Un autor que desarrolla esta idea es Schmidt (1996), para quien

> Una parte importante de la explicación [del uso continuado del contraste de hipótesis] es que los investigadores tienen creencias falsas sobre el contraste de hipótesis, supersticiones que les dicen que las pruebas de significación ofrecen ventajas importantes que los hechos no brindan. Tres de estas creencias son particularmente importantes. La primera es la idea que el nivel de significación de un estudio indica la probabilidad de réplicas acertadas del estudio.... Una segunda creencia falsa muy frecuente entre los investigadores es que el nivel estadístico de significación proporciona un índice de la importancia o tamaño de una diferencia o de una relación.... La tercera creencia falsa es la más devastadora de todas para la empresa de investigación. Según esta suposición si una diferencia, o una relación, no es estadísticamente significativa entonces su valor es cero, o es tan pequeña que con seguridad se puede ver como cero. De acuerdo a esta creencia si la hipótesis nula no se rechaza entonces debe aceptarse. Ésta es la creencia en que una ventaja importante de las pruebas de significación es que nos dicen si una diferencia, o un efecto, es verdadero o "probablemente ocurren sólo por azar. (p. 476)

En segundo lugar, no todos los autores dan por cierta la diatriba, al menos no de modo extremo, por lo cual, algunos defienden el contraste de hipótesis, aun cuando cada quien lo haga por razones diferentes.

Entre los sostenedores está, por ejemplo, Abelson (1997a, 1997b, 1998), quien defiende el empleo del *contraste de hipótesis* si este se usa de modo juicioso, un fin para el cual recomienda ciertas reglas la cuales resume en el anagrama MAGIC (magnitud-articulación-generalidad-interés-credibilidad); como él lo dice: "he propuesto que la función propia de la estadística es enunciar buenos argumentos que expliquen diferencias comparativas, con la esperanza de que se haga de un modo interesante" (Abelson, 1998, p. 39).

Otro de los adalides del *contraste de hipótesis* es Chow (1996, 1998, 2000, 2002), quien asume que el procedimiento es útil a los efectos de discernir, de modo ordinal, cuál enunciado tiene un mejor comportamiento bajo ciertas circunstancias experimentales. Además, él distingue entre el contraste de hipótesis y la evaluación de una teoría; a los efectos de lo primero, el contraste estadístico

sería suficientemente eficiente, para lo segundo el procedimiento requiere de los resultados de muchos contrastes deducidos coherentemente de la teoría en cuestión. Como dice el mismo autor

> Un resultado significativo indica un efecto discernible en el sentido de que los efectos del azar están excluidos con referencia a un criterio bien definido. Igualmente, un efecto significativo es un efecto genuino en el sentido que no es producto de factores aleatorios. El contraste de hipótesis es una herramienta indispensable para estimar si se puede o no descartar la influencia del azar en la explicación de los datos. (Chow 1998, p. 328)

En este mismo sentido, otros defensores del contraste de hipótesis, Mulaik et al (1997, p. 81), puntualizan que

> ... no podemos librarnos del contaste de hipótesis porque él nos da un criterio mediante el cual distinguir provisionalmente los resultados que se deben al azar de las variaciones que representan efectos sistemáticos detectables en los datos. En el momento en que concebimos que las variaciones de las observaciones se pudieran deber al azar y que esa idea es igualmente aplicable a nuestra experiencia, entonces aparece la necesidad de disponer, de una u otra forma, de una prueba de significación. (p. 81)

Como resumen de los argumentos de los defensores, Mateo (2001, p. 196-197) apunta,

- *En el enfoque de Neyman y Pearson no hay interés por la verosimilitud de hipótesis alguna, el objetivo es tratar de decidir cuál de dos hipótesis propuestas es más apropiada para dar cuenta de unos datos que se han recolectado.*
- Las hipótesis siempre se definen a nivel poblacional, por lo cual la muestra tiene lo necesario para su contraste por vía de alfa (α).
- Dentro del marco decisorio de Neyman y Pearson, el concepto de nivel de significación representa una traducción cuantitativa de la expresión 'más allá de la duda razonable'.
- El tamaño de la muestra no debería ser nunca casual y se debería estimar sobre la base de estudios previos, de modo que optimice las probabilidades de que se manifiesten los efectos que son relevantes.

Como se puede ver, la diatriba en relación al *contraste de hipótesis* es de larga data, tiene más de 70 años, y no ha cesado; aunque en el tiempo ha mostrado variaciones, altos y bajos, continúa.

¿Qué hacer con el *contraste de hipótesis*?

Luego de lo dicho, que es sólo una pequeña parte de lo mucho que se ha escrito del tema, queda preguntarse ¿qué hacer con el contraste de hipótesis? Kline (2004, 85) especifica que

> ... podemos escoger uno de tres cursos de acción:
>
> 1. No hacer nada; esto es, continuar usando las pruebas estadísticas tal cual como lo hemos hecho durante los últimos 50 años.
> 2. Detener completamente el uso de las pruebas estadísticas. Dejar de enseñarlas en los cursos universitarios.
> 3. Trazar un curso de acción entre los dos extremos ya indicados, uno que implique a una variedad de grados de uso de las pruebas estadísticas – desde ninguna hasta una más bien nuclear, dependiendo del contexto de investigación, pero con estrictos requisitos para su uso.

Bien, sea cual sea la mejor opción, si es que hay tal cosa, acá se asume una postura cercana a la tercera, con la cual se pretende, uno, evitar perder lo bueno del procedimiento y, dos, agregarle elementos al método en pro de mejorar las conclusiones a las cuales se podría llegar mediante su empleo; concretamente, el punto es conservar lo mejor del contraste de hipótesis y no terminar *botando al niño junto con el agua de la batea* (Granaas, 2002).

Además, es muy importante destacar, por una parte, que "no se debe buscar una alternativa mágica a la formulación del contraste de hipótesis, ella no existe" (Cohen, 1994, p. 1001); y, por la otra, según Wagenmakers (2007, p. 795),

> ... el procedimiento seleccionado debería
>
> 1. depender sólo de los datos observados. Los datos que deberían haberse observado pero no que no se pudo lograr son irrelevantes para la situación presente.
> 2. dar una medida de la evidencia que toma en cuenta tanto a la hipótesis nula como a la alternativa.
> 3. ser fácil de implementar.

4. ser "objetivo", en el sentido que distintos investigadores, que disponen del mismo modelo y enfrentan los mismos datos, elaboren las mismas conclusiones.

Y, además, los resultados del procedimiento estadístico

5. no deberían depender de intenciones desconocidas del investigador.

En cuanto a las sugerencias específicas de cómo mejorar el procedimiento y/o las conclusiones, en la literatura se encuentran, ciertamente, muchas alternativas. Acá se indicarán sólo algunas de ellas y, para ordenarlas un poco, estas se dividirán en dos tipos, que ha falta de mejores nombres serán llamadas, unas, didácticas y, otras, metodológicas.

En relación a las primeras, lo esencial de estas sugerencias didácticas es que se orientan, por una parte, a mejorar la enseñaza del contraste de hipótesis y, por la otra, a enriquecer el discurso relativo al contraste.

En la investigación, desde un punto de vista epistemológico, nada ocurre por azar, por lo contrario, los sucesos se dan por razón de las causas y las circunstancias que los empelen y circundan en un momento determinado; pero, mayormente se desconocen los factores y sus parámetros, además el modelo lineal no es el prototipo del comportamiento de los fenómenos. Ante esta situación se opta por una formulación condicional y una técnica estadística, donde el azar es un recurso didáctico, una analogía, y no una explicación, lo cual, en principio pareciera contradictorio, ¿cómo estudiar fenómenos deterministas mediante asunciones probabilísticas?, pero, como dice Brenner-Golomb (1993),

> ... antes que ser contradictorio, las nociones de probabilidad y determinismo están intrínsecamente relacionadas: la primera se refiere a la frecuencia de eventos independientes, y la segunda es la consecuencia de estos agregados. De acuerdo a Fisher, la posibilidad de predecir depende de esta relación intrínseca. (p. 290)

Teniendo esto en cuenta, uno de los problemas principales surge del modo como se manejan estos elementos epistémicos al momento de explicar el *contraste de hipótesis*; en las palabras de Pollard (1993, p. 454)

> Cuando se presenta el contraste de hipótesis, el nivel de significación se identifica correctamente como condicionado a que la H_0 sea cierta, y el error Tipo I es casi correctamente definido como el rechazo de la H_0 siendo ella cierta. Sin embargo, las exposiciones posteriores tienden

> a igualar el nivel de significación con la probabilidad de un error Tipo I, al mismo tiempo que se desecha la naturaleza condicional de esta probabilidad.

En otras palabras, se propende a dar por seguro un lado del cuadro, que se refiere a la H_0, y se descarta el otro, que alude a la H_1; con esto, una situación que es por naturaleza condicional y un criterio que por definición es arbitrario, se venden como si fuesen una condición rígida y un arbitrio objetivo, respectivamente. Así, se tiende a mostrar el resultado del *contraste de hipótesis* como seguro e inobjetable, cuando es, ciertamente, condicional y discutible.

De este modo, sobre la enseñanza, el problema esencial es que el procedimiento se presenta como si fuese un método, es decir, un conjunto de reglas que si se siguen de modo correcto llevan inevitablemente a una solución correcta; se instruye a los alumnos a emplearlo de forma casi irracional, así literalmente la enseñanza se torna más bien en un adoctrinamiento rígido que anquilosa al alumno (Cohen, 1994).

Ante esto la sugerencia es sencilla de entrever, si bien quizás no es tan simple ponerla en práctica; una posibilidad radica en enfatizar el carácter determinista del fenómeno en estudio y la condición probabilística del contraste; es decir, la naturaleza es segura en algún sentido, mientras que la aproximación nosológica es contingente. Pero, cómo hacer esto sin generar desesperanza en los alumnos, la cual les podría llevar a la deserción o a una forma de solipsismo; sin duda, no hay una receta para esto, aun cuando vale iniciar los cursos de estadística inferencial distinguiendo la duda epistemológica del vacilar neurótico, destacando lo provisorio de todo saber, resaltado el valor de la recursividad constructiva, y, además, haciendo énfasis en cualquier otro elemento que sea útil a un enfoque el cual, apegándose a lo fundamental de la propuesta estocástica del procedimiento, no termine cosificando las metáforas del trabajo didáctico.

Como lo explica Batanero (2000, p. 96)

> ... frecuentemente encontramos que la estadística se enseña aisladamente, sin conectarla con un marco más general de metodología de investigación y diseño experimental. Desde nuestro punto de vista, es necesario discutir el papel de la estadística en la investigación experimental con los estudiantes y hacerlos conscientes de las posibilidades y limitaciones de la estadística en el trabajo experimental. Aún más, coincidimos con la sugerencia de Wood (1998) de enfocar el curso introductorio de estadística alrededor del razonamiento estadísticos, es decir el ciclo de

> aprendizaje Planificación-Conjetura-Comprobación-Acción. El análisis estadístico de datos no es un proceso mecánico y, por tanto, no debería ser enseñado o aplicado de esta forma. Puesto que la estadística no es una forma de hacer sino una forma de pensar que nos puede ayudar a resolver problemas en las ciencias y la vida cotidiana, la enseñanza de la estadística debería empezar con problemas reales mediante los cuales los estudiantes puedan desarrollar sus ideas, trabajando las diferentes etapas en la resolución de un problema real (planificar la solución, recoger y analizar los datos, comprobar las hipótesis iniciales y tomar una decisión en consecuencia).

Granaas (2002, p. 3) aporta otras sugerencias en este sentido, él señala que sería conveniente:

- Destacar que la hipótesis nula representa un valor teorético relevante (Fisher, 1955).
- Tratar el rechazo de la hipótesis nula como lo que es, evidencia en contra de la nula y no un elemento a favor de una hipótesis alternativa específica.
- Comenzar el entrenamiento de los investigadores motivándoles a que especifiquen o estimen los parámetros de sus modelos teóricos.

Por su parte, Steiger (2004, p. 69) indica otros elementos para tener en cuenta,

> Meehl (1978) produjo una crítica demoledora de la manera en que muchos psicólogos blandos emplean el contraste de hipótesis en la validación de las teorías. De dicho documento emergen algunas claves:
>
> 1. La hipótesis nihilista (hipótesis de absoluta no diferencia entre las medias o de correlación cero) siempre es falsa en la psicología blanda, por lo cual el porcentaje de rechazo es mayormente una función del poder estadístico.
> 2. Cuando contrastamos una teoría, usualmente probamos algo más –la teoría más varios elementos ancilares. Falsear la conjunción de la teoría y sus elementos ancilares no es necesariamente fatal para la teoría- podría indicar simplemente que los aspectos ancilares deberían ajustarse.
> 3. Hallar evidencia estadística acorde con una teoría no prueba que la teoría es cierta –esta es la falacia de *afirmar el consecuente.*

4. Por tanto, cuando se evalúa una teoría mediante la prueba de hipótesis, debemos dar mucho más peso a los resultados negativos que a los positivos.
5. Realizar un grupo de contrastes y "contar hitos" (p. ej. colocando acuerdos y desacuerdos en muchos contrastes de hipótesis para validar una teoría) ignora los principios anteriores y debería descontinuarse.
6. A pesar del punto (3), Meehl siente que si al contrastar una teoría se generan muchos datos concordantes con la predicción, no es plausible que la teoría esté errada.

En lo que respecta al segundo elemento, enriquecer el discurso relativo al contraste, en los últimos años se han propuesto una diversidad de alternativas, las cuales, sin dejar de lado lo esencial del contraste de hipótesis, aportan mayor cantidad de elementos de juicio, o de información como dicen otros, para la toma de decisiones (Wilkinson y Task Force on Statistical Inference, 1999).

Estas recomendaciones, indicadas por una variedad de autores (Thompson, 2008; Díaz, 2007; Díaz y Batanero, 2007; Tabachnick y Fidell, 2007; Young, Valero-Mora y Friendly, 2006; Hunter y Schmidt, 2004; Wilcox, 2004, 1999; Cohen, 2001; Good, 2001; Nickerson, 2000; Rosenthal, Rosnow y Rubin, 2000; Peña, 1999, 1998), se pueden resumir en:

- Restringir su uso a los estudios experimentales
- Apelar al análisis exploratorio de datos
- Utilizar métodos de análisis visual de datos
- Emplear límites de confianza
- Calcular y tomar en cuenta el poder de la prueba
- Estimar la magnitud del efecto
- Recurrir al meta-análisis
- Utilizar métodos robustos de estimación y contraste de hipótesis
- Acudir a técnicas de remuestreo
- Valerse de métodos bayesianos

Pasando ahora a las consideraciones desde un punto de vista epistemológico, quizás sería mejor decir, desde un enfoque metodológico, hay dos elaboraciones, más bien recientes, que valdría la pena comentar: el enfoque de modelos y el cálculo de la probabilidad de replicación.

Estadística inferencial basada en modelos

Como lo indica Dahl (1999, p. 3)

> Para lograr que los estudiantes evadan muchas de las pifias de la práctica usual de la estadística, yo creo que se debe conminar a que toda conclusión estadística sea relativa a un modelo. Típicamente, nunca se puede estar completamente seguro de que un modelo es correcto, pero si es posible verificar algunas de las consecuencias del modelo y de esa forma tener una impresión de la situación.

Otro autor que aboga por el uso del *contraste de hipótesis* basado en modelos es Granass (2002),

> Yo sugiero que se ponga a tono el contraste de hipótesis a favor de emplear la aproximación de la bondad-del-ajuste de la prueba de modelos y parámetros del modelo. Esto es, por su puesto, un giro de mano ya que la diferencia entre el contraste de hipótesis, como se usa habitualmente, y la bondad-del-ajuste es que este último emplea modelos teóricamente relevantes para justificar la nulidad. Por tanto, el cambio consiste no en prohibir el *contraste de hipótesis*, sino en reencuadrarlo en un marco más apropiado. Un esfuerzo en pro de la enseñanza y práctica para el desarrollo y prueba de modelos podría subsanar la debilidad inherente del *contraste de hipótesis* en su aplicación usual en psicología y educación, además de poner en práctica las recomendaciones del Task Force's. (p. 3)

Igualmente, para Wagenmakers (2007, p. 799)

> La solución al problema de la inferencia estadística es pasar de la metodología de los valores de *p* a una metodología de selección de modelos. Los métodos de selección de modelos evalúan varios modelos, y nada quita que H_0 y H_1 ofrezcan la mejor explicación de los datos.

Ahora, pasando de la propuesta de la idea a las formas de llevarlo a cabo, una recomendación concreta de cómo hacerlo la dan Jones y Tukey (2000, p. 412 y 414) cuando señalan que

> ... debemos evaluar los datos de la muestra y seleccionar una de tres conclusiones:
>
> (1) actuar como si $\mu_A - \mu_B > 0$;
>
> (2) actuar como si $\mu_A - \mu_B < 0$; o

(3) actuar como si el signo de $\mu_A - \mu_B$ es indefinido, p.ej., si (aun) no está determinado.

Esta especificación es similar a la "decisión de tres alternativas" propuesta por Tukey (1960, p. 245). Con esta formulación, una conclusión es errada sólo cuando es "reversa", es decir, cuando afirma una dirección mientras que la verdadera (desconocida) es en la otra dirección. Asumir que la dirección aun no está establecida podría ser una pérdida de oportunidad, pero no es un error.

En resumen,

- Prefiera los intervalos de confianza cuando están disponibles.
- Reconozca que las hipótesis puntuales, aun cuando matemáticamente convenientes, nunca se satisfacen en la práctica.
- Cuando se realiza un contraste simple, busque alguno de los tres resultados mencionados.

Por su parte, Granass (2007, p. 6-7) propone

En vez de hablar de la hipótesis nula como una "prueba de no diferencia", esta se presenta como una prueba de la necesidad, o de la precisión, de un parámetro del modelo.

Para esto se necesita un poco de notación adicional y algunos cambios conceptuales en relación a la interpretación de los resultados estadísticos. Para el caso de una t de Student la notación adicional involucra presentar un modelo para el parámetro hipotetizado (el modelo nulo) y un modelo para la estimación del parámetro (el modelo alternativo):

Modelo Nulo: $\hat{Y}_i = \mu$ (el cual se corresponde con H_0: $\mu = k$)

Modelo Alternativo: $\hat{Y}_i = \bar{Y}$ (el cual se corresponde con H_1: $\mu \neq k$)

Esta notación adicional clarifica que estamos contrastando un valor del parámetro basado en una teoría o en un resultado anterior (el modelo nulo) contra un parámetro estimado que se deriva de los datos (el modelo alternativo).

Cálculo de la probabilidad de replicación (p_{rep})

Un aserto usual en la metodología empírica, que se remonta a los tiempos de Robert Boyle (1627-1691), es que la replicación es un elemento fundamental

en la construcción del conocimiento científico (Ramos, Catena y Trujillo, 2004). En éste sentido, Cumming (2006, p. 1) señala

> Hay dos razones principales por las cuales la replicación es valiosa. Primera, ella reduce la posibilidad que un efecto observado sea un artefacto, causado por la variabilidad de la muestra. Segunda, e importante, cualquier replicación es en algo diferente al experimento original: ocurre en un momento diferente, y puede implicar a otros investigadores en un lugar distinto, quizás con ligeras variaciones en el procedimiento y el contexto. Si un efecto repite, estas diferencias ofrecen cierta evidencia de su robustez y generalidad.

El mismo Cumming (2006) considera que hay tres formas principales de abordar la replicación en los estudios experimentales: una, el empleo de los intervalos de confianza; dos, a través de la consideración del significado de la p de experimentos subsecuentes; y, tres, mediante el cálculo de la probabilidad de la replicación (p_{rep}). De seguido desarrollaremos los elementos esenciales de esta última alternativa.

Concretamente, esta propuesta, debida a Killeen (2005, 2006, 2007), asume que se debería estimar la probabilidad de que se pueda replicar un experimento. Killeen (2007, p. 107) señala,

> ¿Cuál es la probabilidad de que, usando las mismas operaciones experimentales y una misma población de sujetos, otro investigador pueda replicar los resultados originales? Esta probabilidad se puede estimar una vez que acordemos el significado de *replicar*. Consideren esta definición:
>
> *Replicar* significa repetir las mismas operaciones empíricas y registrar los datos que apoyan la deposición original
>
> - si la deposición es tan modesta como "las operaciones funcionan, generan un efecto positivo," entonces cualquier tentativa de la réplica que encuentre un efecto positivo se podría asumir como una réplica acertada. Aunque esto puede parecer un umbral demasiado modesto para la réplica, pongámosla en el contexto de qué significación usualmente verifica la contrastación: en una prueba de una cola, 1-p nos da la probabilidad de que nuestro estadístico *d* tiene el mismo signo que el parámetro de la población (Jones y Tukey, 2000).

- si la deposición señala "estas operaciones promueven una magnitud del efecto de por lo menos dL", entonces sólo las réplicas que logran d ≥dL cuentan como replicaciones exitosas.
- si la deposición es que "las operaciones generan un efecto significativo", entonces sólo los intentos de replica que redundan en una p < .05 se consideran replicaciones exitosas.
- si la deposición es "estas operaciones son esencialmente ineficientes, generan una magnitud del efecto menor que dU", entonces cualquier intento de réplica que genere una d<dU podría entenderse como una replicación exitosa. Uno podría tener que decidir de antemano si una d menor que, por ejemplo, -0.5 es consistente con la deposición o constituye una evidencia de una deposición alternativa más fuerte, tal como "esta operación podría defraudar".

Bajo esta definición, la probabilidad de replicación (p_{rep}), asumiendo la normalidad de la distribución de probabilidad, es una función del valor de la p de Fisher: $p_{rep} = \Phi[\Phi^{-1}(1\text{-}p)/\sqrt{2}]$.

Como lo indica textualmente Killeen (2006, p. 551), la probabilidad de replicar un experimento

> Tiene una correspondencia única uno-a-uno con la p de Fisher; particularmente, $p_{rep} = N\,[2^{-1/2}\,Z\,(1-p)]$, donde N es la distribución normal estandarizada y Z es su inversa. La distribución de la magnitud del efecto converge rápidamente en la curva normal. Para p = .05, .025, y .01, la probabilidad de replicación, es decir, de lograr un efecto del mismo signo en un nuevo experimento es: $p_{rep} \approx$ .88, .92, y .95 respectivamente.

En la Figura 1 (Killeen, 2007, p. 108, fig. 7.1) se dan las distribuciones de p y p_{rep} develando la relación entre ellas dos.

El área oscura a la derecha del tamaño del efecto observado (d_1) indica la probabilidad de observar datos más extremos que d_1, dada la H_0.

El área gris a la derecha de 0 indica la probabilidad de lograr un efecto positivo en la replicación.

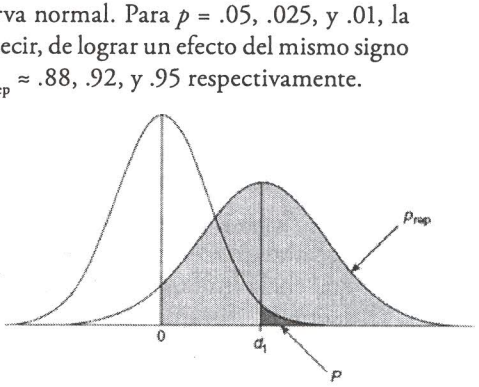

Figura 1: Distribución de p y de p_{rep}

Como lo explica (Killeen, 2007, p. 109)

La *p* es la probabilidad de rechazar la hipótesis nula dado que el dato, d_1, es muestreado de un mundo en el cual la hipótesis nula es cierta. Ella se muestra como el área a la derecha de d_1 en la figura 7.1, bajo una distribución normal centrada en 0 y con una varianza de δ_d^2 estimada a partir de S_d^2. El valor de p_{rep} de los mismos datos –la probabilidad de encontrar un efecto positivo en la replicación- es el área sombreada a la derecha de 0 en la curva normal centrada en d_1 y que tiene como varianza a δ_{dR}^2 estimada a partir de $S_{dR}^2 = S_d^2$.

La p_{rep} no da la probabilidad que la hipótesis experimental sea cierta, ni tampoco que la hipótesis nula es cierta, o de que alguna de las dos o las dos son falsas. Lo que indica es la probabilidad a largo plazo de que una replicación *exacta* de la investigación apoye la deposición original. Parafraseando, la idea sería, si un investigador detecta un efecto positivo resultado de cierta manipulación y alguien la repite el trabajo, hay una probabilidad igual a p_{rep} que este otro investigador logre un efecto igualmente positivo que el primigenio.

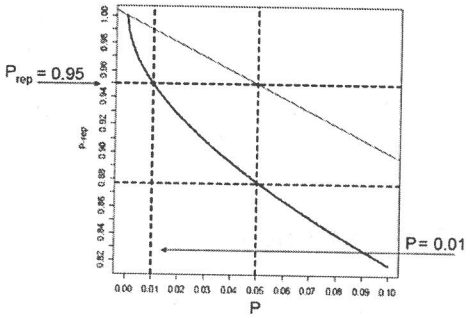

Figura 2: P_{rep} para valores de P< 0.1

Como se puede apreciar en la Figura 2 (Verkuilen y Davis-Stober, 2007, p. 22), a medida que la *p* es menor la p_{rep} es mayor; es decir, a menor probabilidad de cometer el error Tipo I, siendo la H_0 cierta, es mayor la probabilidad de que una replica del experimento original logre un resultado positivo. Además, en la medida que mayor número de réplicas logren resultados positivos, se incrementa tanto la seguridad en la certeza de la relación (validez interna), como su generalidad (validez externa).

Ahora bien, ningún procedimiento, sobre todo si es novedoso, sale ileso, es decir, está falto de críticos. En este caso, por una parte, Doros y Geier (2005,

p. 1006) argumentan "... cualquier medida que no es más que una mera transformación del valor clásico de *p*... heredará los defectos del valor de p".

Por otra parte, Iverson, Wagenmakers y Lee (2009) dicen

> ... aplaudimos el esfuerzo de Killeen en aras de que los investigadores en psicología computen una estimación Bayesiana para decidir si hay o no una relación causal entre las variables independientes y dependientes. Desafortunadamente, la opción de cálculo de la p_{rep} tiene problemas serios, uno de los cuales es que puede conducir al exceso de confianza y al optimismo indebido. Recomendamos que los investigadores no publiquen la p_{rep} sino que reporten una versión modelo-promedio de la p_{rep} o refieran las conclusiones de uno o más métodos alternativos de inferencia estadística. (p. 14)

Conclusiones

Resumiendo, como dicen Iverson, Wagenmakers y Lee (2009, p. 13), todo "esto plantea la pregunta de si hay un método para el contraste de hipótesis que no tenga el defecto de los demás. La respuesta, en principio, parece ser que no. El contraste de hipótesis es muy difícil". Y como dijese Nickerson (2000, p. 290), "el contraste de hipótesis tradicional tiene sus verrugas, pero la misma situación se aplica a todas las alternativas disponibles".

Ante lo cual hay dos conclusiones:

a) el dilema entorno a los problemas inherentes al contraste de hipótesis no está solucionado;

b) y no parece existir una solución mágica a esta diatriba, con lo cual se quiere decir que la meta de un *método para el contraste de hipótesis* no se podría lograr a corto plazo, si es que ello es posible en algún momento.

En consecuencia:

- primero, por ahora hay que reconocer que un procedimiento de *contraste de hipótesis* sin manchas es una *ficción teórica útil*, una especie de quimera que debería guiar los esfuerzos, enrumbarlos hacia un destino relevante, pero que, tal como la botija de oro al final del arcoíris, no es alcanzable.

- segundo, en este momento no se puede prescindir del *contraste de hipótesis* tal como lo conocemos, ya que no hay una solución mejor y el que existe al menos "nos da un criterio mediante el cual distinguir provisionalmente los resultados que se deben al azar de las variaciones que representan efectos sistemáticos detectables en los datos";
- tercero, es menester habituarse a la ambigüedad epistémica inherente a su empleo, al tiempo que se trata de subsanar sus ya muy cacareadas debilidades.

En cuanto a ¿cómo hacer esto último?, pragmáticamente, por una parte, se puede recurrir al uso de procedimientos diversos y, por la otra, apoyarse en la coincidencia de los resultados de las diferentes técnicas como un argumento a favor de las conclusiones.

En concreto,

a) en términos conceptuales, una aproximación que luce fundamental es la de los modelos teóricos; y aun cuando esta se ha desarrollado en años recientes mucho más en el contexto de los diseños observacionales, ahora se muestra indispensable en el ámbito de los experimentos; y

b) en términos prácticos, se debería:

- descartar el uso de la hipótesis nula como sinónimo de una relación nula de toda nulidad (la hipótesis nihilista); en todos los casos el investigador debe estimar a priori, pero con fundamento, cuál es el efecto esperable por azar y, claro está, cuál por acción de la(s) variable(s) independiente(s);
- efectuar el análisis exploratorio que mejor muestre las características de las variables en cuestión, de modo que se minimice el número de supuestos necesarios y se verifiquen todos aquellos que resultan imprescindibles;
- realizar el contraste de hipótesis tradicional, es decir, recurrir a los procedimientos usuales (como t, F, etc.) que convengan al tipo de variable y nivel de conocimiento del cual se disponga, cuidando los términos de representatividad y tamaño de la muestra, y teniendo en cuenta lo que ciertamente significa rechazar o aceptar las hipótesis en juego (nula y alternativas);
- calcular el poder de la prueba y la magnitud del efecto; así como la probabilidad de replicación. Y, además, hay que reconocer que todo

éxito es un paso en el camino, no la meta, y que todo fracaso es una advertencia, no el final de la ruta.

Claro está, del mismo modo, no se debe olvidar que también se debe cuidar el tema de la didáctica de la enseñanza del *contraste de hipótesis*; en este sentido, es recomendable:

- inducir en el estudiante la conciencia del error; en tanto que no existe forma de evitar la equivocación, pero si la hay de enmendar las desviaciones, se debe reconocer que la evidencia va dando a algunos asertos cada vez una mayor solidez;
- familiarizar a los alumnos con la historia de los procedimientos, indicando la cronología de los eventos y la controversia que ellos suscitan, con la idea que el estudiante se sensibilice en relación a la discusión de fondo y esté al tanto de las alternativas;
- vincular sistemáticamente los elementos teóricos (el saber acumulado), los metodológicos (el cómo obtener el conocimiento), los estadísticos (un modo de afrontar las situaciones de incertidumbre) y el contraste de hipótesis (una forma de inferencia) de modo que todos se acoplen de manera armónica y funcional;
- entrenar a los estudiantes inicialmente en la teoría matemática y estadística básica, recurriendo al cálculo manual al tiempo que se hacen evidentes las relaciones entre las operaciones abstractas y las empíricas, de manera que el alumno vivencia cómo estas se afectan mutuamente; logrado esto,
- adiestrar a los cursantes en el uso de herramientas computacionales para el manejo de datos (literales y numéricos), con énfasis en los paquetes estadísticos;
- incluir a los alumnos en indagaciones fácticas, de modo que mediante el ciclo de aprendizaje Planificación-Conjetura-Comprobación-Acción y escalando progresivamente los diversos niveles de complejidad, el estudiante logre desarrollar y poner a punto las competencias para la planificación y realización de investigaciones empíricas.

Finalmente, en relación a la metáfora en el título de este artículo, cabe preguntarse si este modo de actuar ofrece una alternativa que se puede considerar, ciertamente, una luz al final del túnel; la verdad es que sería exagerado verlo de esa manera, pero si bien no es tanto como eso, si representa, al menos, un conjunto de linternas que, bien usadas, alivian la tarea de la inferencia en la investigación empírica.

Asimismo, a lo anterior se le puede agregar una reinterpretación del sentido de la polémica, sobre todo en lo que corresponde a su larga extensión temporal, asumiéndola como una prolongación que da a entender que el tema es relevante y que hay muchos autores interesados en él, de lo cual se puede inferir que se dispone de un buen tropel de pensadores tratando de darle solución a un dilema relevante, muestra de lo cual son las muchas alternativas propuestas hasta el día de hoy, así como la existencia de varios artículos en prensa que contienen nuevos intentos, entre los cuales, quien quita, podría estar una mejor opción.

Teniendo todo esto en cuenta, aun cuando el panorama luce actualmente, en algún sentido, desbastador, no hay que perder la esperanza, ya que con tanto esfuerzo, en algún momento el tema debería resolverse; en fin: *ante el éxito, humildad; y ante el fracaso, optimismo.*

Referencias

Abelson, R. (1997a) On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12-15.

Abelson, R. (1997b) A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). En L. Harlow, S. Mulaik y J. Steiger (Eds.) *What if there were no significance tests?* Erlbaum: Hillsdale.

Abelson, R. (1998) *La estadística razonada: reglas y principios.* Paidos: Madrid.

Aiken L., West, S. y Millsap, R. (2008) Doctoral Training in Statistics, Measurement, and Methodology in Psychology: Replication and Extension of Aiken, West, Sechrest, and Reno's (1990) Survey of PhD Programs in North America. *American Psychologist,* 63, 1, 32–50.

Anderson, D. (1997) *A Few Quotes Regarding Hypothesis Testing.* En la dirección electrónica: http://welcome.warnercnr.colostate.edu/~anderson/nester.html

Anderson, D., Burnham, K. y Thompson, W. (2000) Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management,* 64, 4, 912-923.

Batanero, C. (2000) Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning,* 2, 1-2, 75-98.

Batanero, C. y Díaz, C. (2006) Methodological and didactical controversies around statistical inference. Proceedings of 38th Conference of the French Statistical Association. Paris: SFDE.

Brenner-Golomb, N. (1993) R.A. Fisher's philosophical approach to inductive inference. En G. Keren y C. Lewis (Eds) *A handbook for Data Analysis in Behavioral Sciences: Methodological Issues.* Lawrence Erlbaum: Hillsdale.

Carver, R. (1978) The case against statistical testing. *Harvard Educational Review,* 48, 378-399.

Chow, S. (1996) *Statistical significance: Rationale, validity, and utility.* Sage: Beverly Hills.

Chow, S. (1998) What statistical significance means. *Theory and Psychology, 8,* 323-330.

Chow, S. (2000) The Popperian framework, statistical significance, and Rejection of Chance. *Behavioral and Brain Sciences, 23,* 294-298.

Chow, S. (2002) Issues in Statistical Inference. *History and Philosophy of Psychology Bulletin, 14,* 1, 30-41.

Cohen, B. (2001) *Explaining psychological statistics* (2º Edc.). Wiley & Sons: New York.

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,* 997-1003.

Cohen, J. (1997, Agosto) *Much ado about nothing.* Ponencia presentada en la asamblea anual de la American Psychological Association, Chicago.

Cumming, G. (2006) Understanding Replication: confidence intervals, p values, and what's likely to happen next time. En la dirección:

http://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/7D3_CUMM.pdf

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N. y Wilson, S. (2007) Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230-232.

Dahl, H. (1999) Teaching Hypothesis Testing. Can it Still be Useful? *International Statistical Institute,* 52° Sesión.

Del Puerto, S., Seminara, S. y Minnaard , C (2007) Identificación y análisis de los errores cometidos por los alumnos en Estadística Descriptiva. *Revista Iberoamericana de Educación,* 43, 3, 1-8.

Díaz, C. (2007) *Suitability of teaching bayesian inference in data analysis courses directed to psychologists.* En la dirección: http://www.stat.auckland.ac.nz/~iase/publications/dissertations/07.diaz.pdf

Díaz, C. y Batanero, C. (2007) ¿Cómo puede el Método Bayesiano contribuir a la Investigación en Psicología y Educación? *Paradigma, 27,* 2, 35-54.

Doros, G. y Geier, A. (2005) Probability of replication revisited: Comment on "an alternative to null-hypothesis significance tests". *Psychological Science,* 16, 1005-1006.

Fidler, F. (2006) Should psychology abandon *p* values and teach CIs instead? Evidence-based reforms in statisics education. ICOTS-7.

Fisher, R. (1935/1971) *The design of experiments* (9º Edición). Hafner Press: New York.

Hotelling, H. Bartky, W. Deming, W. E. Friedman, M. y Hoel, P. (1948) The teaching of statistics. *Annals of Mathematical Statistics,* 19, 95–115.

Gigerenzer, G. (1993) The superego, the ego, and the id in statistical reasoning. En G. Keren y C. Lewis (Eds) *A handbook for Data Analysis in Behavioral Sciences: Methodological Issues.* Lawrence Erlbaum: Hillsdale.

Gliner, J., Leech, N. y Morgan, G. (2002) Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say? *The Journal of Experimental Education, 71*(1), 83–92.

Gómez, M. (2003) Breve Historia de la Escuela de Psicología de la Universidad Católica Andrés Bello (1957-1981). *Analogías del Comportamiento: Boletín de la Escuela de Psicología,* 6, 115-122.

Good, P. (2001) *Resampling methods: a practical guide to data analysis.* Birkhäuser: Boston.

Granaas, M. (2002) Hypothesis testing in psychology: throwing the baby out with the bathwater? En B. Phillips (Ed.) *Proceedings of the Sixth International Conference on Teaching of Statistics.* Ciudad del Cabo.

Granaas, M. (2007) Testing Models and an Alternative to Hypothesis Testing in Undergraduate Statistics. En la dirección: http://www.usd.edu/gpctop/Submissions2007/Granaas2007.pdf

Greenwald, A. (1993) Consequences of prejudice against the Null Hypothesis. En G. Keren y C. Lewis (Eds) *A handbook for Data Analysis in Behavioral Sciences: Methodological Issues.* Lawrence Erlbaum: Hillsdale.

Guttman, L. 1985. The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis,* 1, 3-10.

Hunter, J. y Schmidt, F. (2004) *Methods of meta-analysis: correcting error and bias in research findings* (2º Edc). Sage Publications: Thousand Oaks.

Iverson, G., Wagenmakers, E-J y Lee, M. (2009) *A Model Averaging Approach to Replication: The Case of* p_{rep}. *En la dirección: http://www.socsci.uci.edu/~mdlee/prepNullHypothesisWeb.pdf*

Jones, L. y Tukey, J. (2000) A sensible formulation of the significance testing. *Psychological Methods,* 5, 4, 411-414.

Kerlinger, F. y Lee, H. (2002) *Investigación del comportamiento: métodos de investigación en las ciencias sociales.* McGraw-Hill: México.

Killeen, P. (2005) An alternative to null hypothesis significance tests. *Psychological Science.*16, 345–353.

Killeen, P. (2006) Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review.* 13, 549–569.

Killeen, P. (2007) Replication statistics. En *J. Osborne (Editor) Best practices in quantitative methods.* Sage: Beverly Hills.

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

Kline, R. (2004) *Beyond significance testing: reforming data analysis methods in behavioural science.* American Psychological Association: Washington.

Kostenko, A. y Hyndman, R. (2008) *Forecasting without signicance tests?* En la dirección electrónica http://robjhyndman.com/papers/sst2.pdf

Krantz, D. (1999) The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association,* 94, 448, 1372-1381.

Koch, S. (1981) The nature and limits of psychological knowledge. *American Psychologist,* 36, 3, 257-269.

Loftus, G. (1991) On the Tyranny of Hypothesis Testing in the Social Sciences. *Contemporary Psychology,* 36, 2, 102-105.

Mateo, M. (2001) Reflexiones en torno a aspectos metodológicos de la psicología. En J. Fernández (Cord.) *La Psicología una ciencia diversificada.* Ediciones Pirámide: Madrid.

Mulaik, S., Raju, N. y Harshman, R. (1997) There is a time and place for significance testing. In L. Harlow, S. Mulaik, y J. Steiger (Eds.) *What if there were no significance tests?* Erlbaum: Hillsdale.

Nickerson, R. (2000) Null Hypothesis Significance Testing: a review of an old and continuing controversy. *Psychological Methods,* 5, 2, 241-301.

Parkhurst, D. (Comp.) (1997) *Commentaries on Significance Testing.* En la dirección electrónica http://www.indiana.edu/~stigtsts/

Peña, G. (1995) Dos factores explicativos de la cuantofobia: la revolución de la inferencia estadística y el factor humano en la concepción de los métodos estadísticos. *Revista de la Asociación Venezolana de Psicología Social,* XVIII, 2, 65-83.

Peña, G, (1998) Relevancia del efecto: el índice Λ. *Espacio: Temas de Ciencias Sociales,* Año 5, 5, 59-78.

Peña, G, (1999) *f* Eugene: una técnica para la evaluación gráfica de la relevancia del efecto. *Espacio: Temas de Ciencias Sociales,* Año 6, 6, 83-101.

Rivadulla, A. (1991) *Probabilidad e inferencia científica.* Antropos: Barcelona.

Ramos, M., Catena, A. y Trujillo, H. (2004) *Manual de métodos y técnicas de investigación en ciencias del comportamiento.* Biblioteca Nueva: Madrid.

Rosenthal, R., Rosnow, R. y Rubin, D. (2000) *Contrast and effect sizes in behavioural research: a correlational approach.* Cambridge University Press: Cambridge.

Rozeboom, W. W. (1997) Good science is abductive, not hypothetico-deductive. En L. Harlow, S. Mulaik y J. Steiger (Eds.) *What if there were no significance tests?* Erlbaum: Hillsdale.

Sawilowsky, S. (2003) Deconstructing Arguments From The Case Against Hypothesis Testing. *Journal of Modern Applied Statistical Methods*, 2, 2, 467-474.

Schmidt, F. (1996) Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods*, 1, 2, 115-129.

Singh, G. (2006) A shift from significance test to hypothesis test through power analysis in medical research. *Journal Postgrad Medical*, 52, 2, 148-150.

Schmidt, F., Hunter, J. y Urry, V. (1976) Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Snijders, T. (2001) Hypothesis Testing: Methodology and Limitations (pp. 7121-7127). *International Encyclopedia of the Social & Behavioral Sciences.* Elsevier Science Ltd.

Steiger, J. (2004) Paul Meehl and the evolution of statistical methods in psychology. *Applied & Preventive Psychology*, 11, 69–72.

Stigler, S. (1986) *The history o statistics: the measurement of uncertainty before 1900.* Harvard University Press: Cambridge.

Stigler, S. (1999) Statistics on the Table: the history of statistical concepts and methods. Harvard University Press: Cambridge.

Tabachnick, B. y Fidell, L. (2007) *Experimental designs using ANOVA.* Thomson: Belmont.

Thompson, B. (1997, Agosto) *If statistical significance test are broken/misused, what practices should supplement or replace then?* Ponencia presentada en la asamblea anual de la American Psychological Association, Chicago.

Thompson, B. (Comp.) (2001) *402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies.* En la dirección electrónica: http://welcome.warnercnr.colostate.edu/People/Bthompson.html

Thompson, B. (2008) *Foundations of behavioral statistics.* The Guilford Press: New York.

Verkuilen, J. y Davis-Stober, C. (2007) *Prep and You.* En la dirección:

Wagenmakers, E-J. (2007) A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14, 5, 779-804*

Wilcox, R. (1999) *Introduction to robust estimation and hypothesis testing.* Academic Press: San Diego.

Wilcox, R. (2004) *Statistics for the social sciences.* Academic Press: San Diego.

Wilkinson, L. y el Task Force on Statistical Inference (1999) Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist,* 54, 8, 594–604.

Yates, F. (1951) The influence of Statistical Methods for Research Workers on the development of the science of statistics. *Journal of the American Statistical Association,* 46, 19-34.

Young, F., Valero-Mora, P. y Friendly, M. (2006) *Visual statistics: seeing data with dynamic interactive graphics.* Wiley & Sons: Hoboken.

Yu, C. (2008) *Don't believe in the Null Hypothesis?* En la dirección electrónica: http://www.creative-wisdom.com/computer/sas/sas.html