Estadísticas sin sentido

Gerd Gigerenzer[1]

Instituto Max Planck para el Desarrollo Humano, Lentzeallee 94, 14195 Berlín, Alemania

The Journal of Socio-Economics 33 (2004) 587–606

doi:10.1016/j.socec.2004.09.033

Resumen

Los rituales estadísticos eliminan en gran medida el pensamiento estadístico en las ciencias sociales. Los rituales son indispensables para la identificación con los grupos sociales, pero deberían ser el tema y no el procedimiento de la ciencia. Lo que yo llamo el «ritual nulo» consta de tres pasos: (1) establezca una hipótesis nula estadística, pero no especifique su propia hipótesis ni ninguna hipótesis alternativa, (2) use el nivel de significación del 5% para rechazar la nula y aceptar su hipótesis, y (3) siempre realice este procedimiento. Reporto evidencia de la confusión colectiva resultante y los temores sobre las sanciones por parte de estudiantes y profesores, investigadores y editores, así como también escritores de libros de texto.

Palabras clave: Rituales; Ilusiones colectivas; Significación estadística; Editores de revistas científicas; Libros de texto

> ... ningún trabajador científico tiene un nivel fijo de significación en el que de año en año, y en todas las circunstancias, rechaza hipótesis; más bien dedica su mente a cada caso particular a la luz de sus pruebas y sus ideas.
>
> Sir Ronald A. Fisher (1956)

Una vez visité a un distinguido autor de libros de texto de estadística, cuyo libro pasó por muchas ediciones y cuyo nombre no importa. Su libro de texto representa lo relativamente mejor de las ciencias sociales. No era un estadístico; de lo contrario,

[1] Tel.: +49 30 82406 460; fax: +49 30 82406 394. Dirección email: gigerenzer@mpib-berlin.mpg.de.

su texto probablemente no se habría utilizado en una clase de psicología. En una edición anterior, había incluido un capítulo sobre estadística bayesiana y también mencionó (aunque en una sola proposición) que hubo un desarrollo en la teoría estadística de R. A. Fisher a Jerzy Neyman y Egon S. Pearson. Mencionar la existencia de métodos alternativos y los nombres asociados a ellos es prácticamente inaudito en psicología. Le pregunté al autor por qué eliminó el capítulo sobre Bayes y la proposición inocente de todas las ediciones posteriores. «¿Qué le hizo presentar las estadísticas como si tuviera un solo martillo, en lugar de una caja de herramientas? ¿Por qué mezcló las teorías de Fisher y Neyman-Pearson en un híbrido inconsistente que todo estadístico decente rechazaría?»

Para su crédito, debo decir que el autor no intentó negar que había producido la ilusión de que solo hay una herramienta. Pero me hizo saber quién era el culpable de esto. Había tres culpables: sus compañeros investigadores, la administración de la universidad y su editor. La mayoría de los investigadores, argumentó, no están realmente interesados en el pensamiento estadístico, sino solo en cómo publicar sus artículos. La administración de su universidad promovió a los investigadores según el número de sus publicaciones, lo que reforzó la actitud de los investigadores. Y pasó la responsabilidad a su editor, quien exigió un libro de cocina de receta única. Sin controversias, por favor. Su editor le había obligado a sacar el capítulo sobre Bayes, así como la proposición que nombraba teorías alternativas, explicó. Al final de nuestra conversación, le pregunté en qué tipo de teoría estadística creía él mismo. «En el fondo de mi corazón –confesó– soy bayesiano». Si el autor me estaba diciendo la verdad, había vendido su corazón por múltiples ediciones de un libro famoso en cuyo mensaje no creía. Había sacrificado su integridad intelectual por el éxito. Diez mil estudiantes han leído su texto, creyendo que revela el método de la ciencia. Docenas de escritores de libros de texto menos informados copiaron de su texto, produciendo una avalancha de libros de texto descendientes y sin darse cuenta del desorden.

1. El ritual nulo

Los libros de texto y los planes de estudio de psicología casi nunca enseñan la caja de herramientas estadísticas, que contiene herramientas como la estadística descriptiva, los métodos exploratorios de Tukey, la estadística bayesiana, la teoría de decisiones de Neyman-Pearson y el análisis secuencial de Wald. Conocer el contenido de una caja de herramientas, por supuesto, requiere un pensamiento estadístico, es decir, el arte de elegir una herramienta adecuada para un problema dado. En cambio, un solo procedimiento que llamo el «ritual nulo» tiende a aparecer en los textos y a ser practicado por los investigadores. Su esencia se puede resumir en unas pocas líneas:

El ritual nulo:

1. Establezca una hipótesis nula estadística de «sin diferencia de medias» o «correlación cero». No especifique las predicciones de su hipótesis de investigación o de ninguna hipótesis sustantiva alternativa.

2. Utilice el 5% como convención para rechazar la hipótesis nula. Si es significativo, acepte su hipótesis de investigación. Informe el resultado como p <0,05, p <0,01 o p <0,001 (lo que venga junto al valor p obtenido).

3. Realice siempre este procedimiento.

El ritual nulo tiene aspectos sofisticados que no cubriré aquí, como el ajuste alfa y los procedimientos ANOVA. Pero estos no cambian su esencia. A menudo, los libros de texto también enseñan conceptos ajenos al ritual, como el poder estadístico y el tamaño del efecto, pero estas adiciones tienden a desaparecer cuando se dan ejemplos. Simplemente no encajan. Más recientemente, el ritual ha sido etiquetado como prueba de significación de hipótesis nula, para abreviar, NHST[2] o, a veces, NHSTP (con P para «procedimiento»). Se institucionalizó en los planes de estudio, editoriales y asociaciones profesionales de psicología a mediados de la década de 1950 (Gigerenzer, 1987, 1993). La decimosexta edición de un libro de texto muy influyente, *Psicología y vida* de Gerrig y Zimbardo (2002)[3], describe el ritual nulo como una estadística per se y lo llama la «columna vertebral de la investigación psicológica» (p. 46). Su naturaleza mecánica se presenta a veces como las reglas de la gramática. Por ejemplo, el *Manual de Publicaciones de la Asociación Estadounidense de Psicología* de 1974 les dijo a los autores qué escribir en mayúscula, cuándo usar un punto y coma y cómo abreviar Estados y territorios. También les dijo a los autores cómo interpretar los valores p: "Precaución: No infiera tendencias a partir de datos que fallan por un pequeño margen en alcanzar los niveles habituales de significación. Tales resultados se interpretan mejor como causados por el azar y se informan mejor como tales. Trate la sección de resultados como un la devolución del impuestos sobre la renta. Tome lo que le viene, pero no más» (p. 19; este pasaje fue eliminado en la 3ª ed., 1983). No se invita al juicio. Esto me recuerda una máxima con respecto al índice crítico, el predecesor del nivel de significación: «Un índice crítico de tres, o no habrá doctorado».

El anonimato es fundamental. El ritual prácticamente siempre se presenta sin nombres, como estadísticas per se. Si en los libros de texto de psicología se mencionan nombres como Fisher o Pearson, generalmente se hace en conexión con un detalle menor, como agradecer a E. S. Pearson por el permiso para reimprimir una tabla. Las ideas principales se presentan de forma anónima, como si ellas proporcionaran verdades. ¿Qué texto escrito para psicólogos señala que la prueba de hipótesis nula fue idea de Fisher? ¿Y que Neyman y Pearson argumentaron en contra de las pruebas de hipótesis nulas? Si aparecen los nombres de los estadísticos, normalmente se le dice al lector que todos son de una misma opinión.

[2] Siglas de *null hypothesis significance testing*. (N. trad.)
[3] 17ª ed. Castellana: Prentice Hall México (N. trad.)

Por ejemplo, en respuesta a un artículo mío (Gigerenzer, 1993), el autor de un libro de texto estadístico, S. L. Chow (1998), reconoció que de hecho existen diferentes métodos de inferencia estadística. Pero unas líneas más tarde volvió a caer en la fábula de «todo-es-lo-mismo»: «Para K. Pearson, R. Fisher, J. Neyman y E. S. Pearson, NHSTP era de lo que trataba la investigación empírica» (Chow, 1998, p. xi). Lector, cuidado. Cada uno de estos eminentes estadísticos habría rechazado el ritual nulo como una mala estadística.

A Fisher se le culpa principalmente por el ritual nulo. Pero hacia el final de su vida, Fisher (1955, 1956) rechazó cada uno de sus tres pasos. Primero, «nulo» no se refiere a una diferencia de medias nula o correlación cero, sino a cualquier hipótesis que deba ser «anulada». Una correlación de 0,5, o una reducción de cinco cigarrillos fumados por día, por ejemplo, puede ser una hipótesis nula. En segundo lugar, como ilustra la cita inicial, en 1956, Fisher pensó que el uso de un nivel rutinario de significación del 5% indicaba una falta de sofisticación estadística. Ningún investigador respetable usaría un nivel constante. Sus posibilidades de encontrar esta cita en un texto estadístico de psicología son prácticamente nulas. En tercer lugar, para Fisher, la prueba de hipótesis nula era el tipo más primitivo de análisis estadístico y debería utilizarse solo para problemas sobre los que tenemos poco o ningún conocimiento (Gigerenzer et al., 1989, capítulo 3). Propuso métodos más apropiados para otros casos. Neyman y Pearson también habrían rechazado el ritual nulo, pero por diferentes razones. Rechazaron las pruebas de hipótesis nulas y favorecieron las pruebas competitivas entre dos o más hipótesis estadísticas. En su teoría, «hipótesis» está en plural, lo que permite a los investigadores determinar el error de Tipo II (que no es parte del ritual nulo y, en consecuencia, no es pertinente a NHSTP, como afirma Chow). La confusión entre el ritual nulo y la teoría de Fisher, y a veces incluso la teoría de Neyman-Pearson, es la regla más que la excepción entre los psicólogos.

La psicología parece ser una de las primeras disciplinas donde el ritual nulo se institucionalizó como estadística per se, durante la década de 1950 (Rucci y Tweney, 1980; Gigerenzer y Murray, 1987, capítulo 1). Posteriormente, se extendió a muchas ciencias sociales, médicas y biológicas, incluida la economía (McCloskey y Ziliak, 1996), la sociología (Morrison y Henkel, 1970) y la ecología (Anderson et al., 2000).

Si los psicólogos son tan inteligentes, ¿por qué están tan confundidos? ¿Por qué se realizan estadísticas como el lavado compulsivo de manos? Mi respuesta es que el ritual requiere confusión. Reconocer que hay una caja de herramientas estadísticas en lugar de un martillo significaría su fin, al igual que darse cuenta de que el ritual nulo no se practica ni en las ciencias naturales ni en la estadística propiamente dicha. Su origen está en la mente de los escritores de libros de texto estadísticos en psicología, educación y otras ciencias sociales. Fue creado como un híbrido inconsistente de dos teorías en competencia: la prueba de hipótesis nula de Fisher y la teoría de la decisión de Neyman y Pearson.

2. Lo que realmente propusieron Fisher y Neyman-Pearson

En las discusiones sobre los pros y los contras de las pruebas de significación en las ciencias sociales, comúnmente se pasa por alto (por ambas partes) que el ritual ni siquiera es parte de la estadística propiamente dicha. Veamos, pues, qué propusieron realmente Fisher y Neyman-Pearson. La lógica de la prueba de hipótesis nula de Fisher (1955, 1956) se puede resumir en tres pasos:

Prueba de hipótesis nula de Fisher:

1. Establezca una hipótesis nula estadística. La nula no tiene por qué ser una hipótesis nula (es decir, diferencia cero).

2. Informe el nivel exacto de significación (por ejemplo, $p = 0.051$ o $p = 0.049$). No utilice un nivel convencional del 5% y no hable de aceptar o rechazar hipótesis.

3. Utilice este procedimiento solo si sabe muy poco sobre el problema en cuestión.

La prueba de hipótesis nula de Fisher es, en cada paso, diferente del ritual nulo, pero también diferente de la teoría de decisiones de Neyman-Pearson. Carece de una hipótesis alternativa estadística específica. Como consecuencia, los conceptos de poder estadístico, tasas de error de tipo II[4] y tamaños de efectos teóricos no tienen cabida en el marco de Fisher: se necesita una alternativa específica para estos conceptos.

El matemático polaco Jerzy Neyman trabajó con Egon S. Pearson (el hijo de Karl Pearson) en el University College de Londres y más tarde, cuando las tensiones entre Fisher y él se intensificaron demasiado, se mudó a Berkeley, California. Neyman y Pearson criticaron la prueba de hipótesis nula de Fisher por varias razones, entre ellas que ninguna hipótesis alternativa se especifica (Gigerenzer et al., 1989, capítulo 3). En su versión más simple, la teoría de Neyman-Pearson tiene dos hipótesis y un criterio de decisión binario (Neyman, 1950, 1957).

[4] Los errores tipo I y tipo II se denominan también error de tipo α y error de tipo β, respectivamente. Más adelante se vuelve sobre esto. (N. trad.)

Teoría de la decisión de Neyman-Pearson:

1. Establezca dos hipótesis estadísticas, H_1 y H_2, y decida sobre α, β y el tamaño de la muestra antes del experimento, basándose en consideraciones subjetivas de costo-beneficio. Estos definen una región de rechazo para cada hipótesis.

2. Si los datos caen en la región de rechazo de H_1, acepte H_2; de lo contrario, acepte H_1. Tenga en cuenta que aceptar una hipótesis no significa que usted crea en ella, sino que actúa como si fuera cierta.

3. La utilidad del procedimiento se limita, entre otras cosas, a situaciones en las que existe una disyunción de hipótesis (p. ej., $\mu_1 = 8$ o $\mu_2 = 10$ es cierto) y en las que se pueden realizar compensaciones significativas de costo-beneficio para elegir alfa y beta.

Una aplicación típica de las pruebas de Neyman-Pearson es el control de calidad. Imagine un fabricante de placas de metal que se utilizan en instrumentos médicos. Considera que un diámetro medio de 8 mm (H_1) es óptimo y 10 mm (H_2) es peligroso para los pacientes y, por tanto, inaceptable. Por experiencias pasadas, sabe que las fluctuaciones aleatorias de los diámetros se distribuyen aproximadamente de manera normal y que las desviaciones típicas no dependen de la media. Esto le permite determinar las distribuciones muestrales de la media para ambas hipótesis. Considera que las falsas alarmas, es decir, aceptar H_2 mientras que H_1 es verdadero, es el error menos grave, y los errores de mal funcionamiento, es decir, aceptar H_1 mientras que H_2 es verdadero, es más grave. Los errores pueden causar daño a los pacientes y a la reputación de la empresa. Por lo tanto, establece la primera tasa de error pequeña y la segunda más grande, digamos $\alpha = 0{,}1\%$ y $\beta = 10\%$, respectivamente.

Ahora calcula el tamaño de muestra requerido *n* de placas que deben muestrearse todos los días para probar la calidad de la producción (véase Cohen, 1988). Cuando acepta H_2, actúa como si hubiera un mal funcionamiento y detiene la producción, pero esto no significa que crea que H_2 es cierto. Sabe que debe esperar una falsa alarma en 1 de cada 10 días en los que no hay ningún mal funcionamiento (Gigerenzer et al., 1989, capítulo 3).

Ahora está claro que el ritual nulo es un híbrido de las dos teorías. El primer paso del ritual, establecer una sola hipótesis estadística (la nula), se deriva de la teoría de Fisher, excepto que la nula siempre significa «azar», como una diferencia cero. Este primer paso es incompatible con la teoría de Neyman-Pearson; no especifica una hipótesis estadística alternativa, α, β, o el tamaño de la muestra. El segundo paso, tomar una decisión de sí o no, es consistente con la teoría de Neyman-Pearson, excepto que el nivel no debe fijarse por convención sino pensando en α, β y el tamaño de la muestra. Fisher (1955) y muchos estadísticos después de él (véase Perlman y Wu, 1999), por el contrario, argumentaron que, a diferencia del control de calidad, las decisiones de sí-no tienen poco papel en la ciencia; más bien, los

científicos deberían comunicar el nivel exacto de importancia. El tercer paso del ritual nulo es único en la teoría estadística. Si Fisher y Neyman-Pearson estuvieron de acuerdo en algo, fue en que las estadísticas nunca deben usarse mecánicamente.

Fisher es el más conocido de los inadvertidos «padres» del ritual nulo. Su influencia ha dividido profundamente a los psicólogos y, curiosamente, la brecha se extiende entre las grandes personalidades de la psicología, por un lado, y una masa de investigadores anónimos, por el otro. No he pillado a Jean Piaget calculando una prueba *t*. Las contribuciones fundamentales de Frederick Bartlett, Wolfgang Köhler y el premio Nobel I. P. Pavlov no se basaron en los valores *p*. Stanley S. Stevens, fundador de la psicofísica moderna, junto con Edwin Boring, conocido como el «decano» de la historia de la psicología, culparon a Fisher de una «prueba sin sentido de cálculos pedantes» (Stevens, 1960, p. 276). El psicólogo clínico Paul Meehl (1978, p. 817) calificó la prueba de hipótesis nula rutinaria como «una de las peores cosas que han sucedido en la historia de la psicología», y el conductista B. F. Skinner culpó a Fisher y sus seguidores por haber «enseñado estadística en lugar de del método científico» (Skinner, 1972, p. 319). El psicólogo matemático R. Duncan Luce (1988, p. 582) llamó a la prueba de hipótesis nula una «visión equivocada acerca de lo que constituye el progreso científico» y el premio Nobel Herbert A. Simon (1992, p. 159) simplemente afirmó que para su investigación, las «conocidas pruebas de significación estadística son inapropiadas».

Es revelador que pocos investigadores sean conscientes de que sus propios héroes rechazaron lo que practican habitualmente. El conocimiento de los orígenes del ritual y de su rechazo podría provocar una disonancia cognitiva virulenta, además de la disonancia con los editores, revisores y queridos colegas. La supresión de conflictos y la información contradictoria es la naturaleza misma de este ritual social.

3. Sentimientos de culpa

Permítanme presentarles al Dr. Pública-o-Perece[5]. Es el investigador medio, un consumidor devoto de paquetes estadísticos. Su superego le dice que debe establecer el nivel de significación antes de realizar un experimento. Un nivel del 1% sería impresionante, ¿no? Sí, pero... Teme que el valor *p* calculado a partir de los datos pueda resultar ligeramente superior. ¿Y si fuera 1,1%? Entonces tendría que informar de un resultado no significativo. No quiere correr ese riesgo. ¿Qué tal establecer el nivel en un 5% menos impresionante? Pero, ¿y si el valor *p* resultara ser menor al 1% o incluso al 0,1%? Entonces lamentaría profundamente su decisión, porque tendría que informar de este resultado como $p < 0{,}05$. A él tampoco le gusta eso. Entonces concluye que la única opción que queda es hacer un poco de trampa y

[5] Dr. Publish-Perish en el original. (N. trad.)

desobedecer a su superego. Espera hasta que ha visto los datos, redondea el valor *p* al siguiente nivel convencional e informa que el resultado es significativo a $p < 0{,}001$, 0,01 o 0,05, o lo que sea a continuación. Eso huele a engaño y su superego le deja con sentimientos de culpa. Pero, ¿qué debe hacer cuando la honestidad no compensa y casi todos los demás juegan a este pequeño juego de trampas? El Dr. Publica-o-Perece no sabe que su dilema moral es causado por una mera confusión, introducida por escritores de libros de texto que no lograron distinguir las tres interpretaciones principales del nivel de significación.

3.1. Nivel de significación = mera convención

Fisher escribió tres libros sobre estadística. Para las ciencias sociales, el más influyente de ellos fue el segundo, el *Diseño de experimentos*[6], publicado por primera vez en 1935. La definición de Fisher de un nivel de significación difiere aquí de sus escritos posteriores. En el *Diseño*, Fisher sugirió que pensemos en el nivel de significación como una convención: «Es habitual y conveniente que los experimentadores tomen el 5% como nivel estándar de significación, en el sentido de que están preparados para ignorar todos los resultados que fallan para alcanzar este estándar» (1935/1951, p. 13). La afirmación de Fisher de que el 5% (en algunos casos, el 1%) es una convención que deben adoptar todos los experimentadores y en todos los experimentos, mientras que los resultados no significativos deben ignorarse, se convirtió en parte del ritual nulo.

3.2. Nivel de significación = alfa

En la teoría de Neyman-Pearson, el significado de un nivel de significación como 2% es el siguiente: si H_1 es correcto y el experimento se repite muchas veces, el experimentador rechazará erróneamente H_1 en el 2% de los casos. Rechazar H_1 si es correcto se denomina error de tipo I y su probabilidad se denomina alfa (α). Se debe especificar el nivel de significación antes del experimento para poder interpretarlo como α. Lo mismo ocurre con beta (β), que es la tasa de rechazo de la hipótesis alternativa H_2 si es correcta (error de tipo II). Aquí obtenemos la segunda interpretación clásica del nivel de significación: la tasa de error α, que se determina antes del experimento, aunque no por mera convención, sino mediante cálculos de costo-beneficio que logran un equilibrio entre α, β y el tamaño de muestra *n*. Por ejemplo, si $\alpha = \beta = 0{,}10$, entonces no importa si el nivel exacto de significación es 0,06 o 0,001. El nivel de significación no influye en α.

[6] Trad. cast. *El planeo de experimentos*. Instituto Interamericano de Estadística, 1953, (N. trad.)

3.3. Nivel de significación = nivel exacto de significación

Fisher tuvo dudas sobre su propuesta de un nivel convencional y las expresó con mayor claridad en la década de 1950. En su último libro, *Métodos estadísticos e inferencia científica* (1956, p. 42), Fisher rechazó el uso de un nivel convencional de significación y ridiculizó esta práctica, junto con los conceptos de errores de Tipo I y Tipo II, por considerarlos «absurdamente académicos». y con origen en «la fantasía de círculos bastante alejados de la investigación científica» (1956, p. 100). Se refería a los matemáticos, específicamente a Neyman. En ciencia, argumentó Fisher, no se repite el mismo experimento una y otra vez, como se supone en la interpretación de Neyman y Pearson del nivel de significación como una tasa de error a largo plazo. Lo que los investigadores deberían hacer en su lugar, de acuerdo con los segundos pensamientos de Fisher, es publicar el *nivel exacto de significación*, digamos, $p = 0{,}02$ (no $p < 0{,}05$). Usted comunica información; no toma decisiones de sí o no.

Las diferencias básicas son las siguientes: para Fisher, el nivel exacto de significación es una propiedad de los datos, es decir, una relación entre un cuerpo de datos y una teoría. Para Neyman y Pearson, α es una propiedad de la prueba, no de los datos. En *Diseño* de Fisher, si el resultado es significativo, rechaza la hipótesis nula; de lo contrario, no sacará ninguna conclusión. La decisión es asimétrica. En la teoría de Neyman-Pearson, la decisión es simétrica. El nivel de significación y α no son lo mismo. Para Fisher, estas diferencias no eran bagatelas. Calificó la posición de Neyman como «infantil» y «horrorosa [para] la libertad intelectual de Occidente». De hecho, comparó a Neyman con

> los rusos [que] están familiarizados con el ideal de que la investigación en ciencia pura puede y debe estar orientada al desempeño tecnológico, en el esfuerzo organizado integral de un plan de cinco años para la nación ... [Aunque] en los EE. UU. también la gran importancia de la tecnología organizada creo que ha facilitado confundir el proceso apropiado para sacar conclusiones correctas con los que apuntan más bien a, digamos, acelerar la producción o ahorrar dinero. (Fisher, 1955, p. 70)

Probablemente no sea un accidente que Neyman había nacido en Rusia y, en el momento del comentario de Fisher, se había mudado a los Estados Unidos.

Volvamos al Dr. Publicar-o-Perecer y su conflicto moral. Su superego exige que especifique el nivel de significación antes del experimento. Ahora entendemos que la doctrina de su superego es parte de la teoría de Neyman-Pearson. Su ego personifica la teoría de Fisher de calcular el nivel exacto de significación a partir de los datos, combinada con la idea anterior de Fisher de tomar una decisión de sí o no basada en un nivel convencional de significación. El conflicto entre su superego y su ego es la fuente de sus sentimientos de culpa, pero él no lo sabe. Simplemente tiene un vago sentimiento de vergüenza por hacer algo mal. El Dr. Publica-o-Perece no sigue ninguna de las tres interpretaciones. Sin saberlo, trata de satisfacerlas todos, y termina presentando un nivel exacto de significación como si fuera un nivel alfa, redondeándolo a uno de los niveles convencionales de significación, $p < 0{,}05$, $p < 0{,}01$,

o p <0,001. El resultado no es α, ni un nivel exacto de significación. Es producto de un conflicto inconsciente. El conflicto está institucionalizado en los Manuales de Publicaciones de la Asociación Americana de Psicología. La quinta edición del Manual (2001, p. 162) finalmente agregó niveles exactos de significación a una tabla ANOVA (análisis de varianza), pero al mismo tiempo mantuvo los «asteriscos» p <0.05 y p <0.01 de la hipótesis nula ritual. El manual no ofrece ninguna explicación de por qué ambos son necesarios y qué significan (Fidler, 2002). El Dr. Publica-o-Perece tampoco puede encontrar en él información sobre las interpretaciones contradictorias del «nivel de significación»" y los orígenes de sus sentimientos de culpa.

4. Ilusiones colectivas

Los rituales exigen ilusiones cognitivas. Su función es hacer que el producto final, un resultado significativo, parezca altamente informativo y, por lo tanto, justifique el ritual. Intente responder a la siguiente pregunta (Oakes, 1986; Haller y Krauss, 2002):

> Suponga que tiene un tratamiento que sospecha que puede alterar el desempeño en una determinada tarea. Compara las medias de sus grupos de control y experimentales (digamos 20 sujetos en cada muestra). Además, suponga que usa una prueba t de medias independientes simple y su resultado es significativo (t = 2,7, g.l. = 18, p = 0,01). Marque cada una de las siguientes afirmaciones como «verdadera» o «falsa». «Falso» significa que la declaración no se sigue lógicamente de las premisas anteriores. También tenga en cuenta que varias o ninguna de las afirmaciones pueden ser correctas
>
> 1. Ha refutado absolutamente la hipótesis nula (es decir, no hay diferencia entre las medias de la población).
>
> [ ] verdadero / falso [ ]
>
> 2. Ha encontrado la probabilidad de que la hipótesis nula sea cierta.
>
> [ ] verdadero / falso [ ]
>
> 3. Ha probado absolutamente su hipótesis experimental (que hay una diferencia entre las medias poblacionales).
>
> [ ] verdadero / falso [ ]
>
> 4. Puede deducir la probabilidad de que la hipótesis experimental sea cierta.
>
> [ ] verdadero / falso [ ]

5. Sabe, si decide rechazar la hipótesis nula, la probabilidad de que esté tomando la decisión equivocada.

[ ] verdadero / falso [ ]

6. Tiene un hallazgo experimental fiable en el sentido de que si, hipotéticamente, el experimento se repitiera un gran número de veces, obtendría un resultado significativo en el 99% de las ocasiones.

[ ] verdadero / falso [ ]

¿Qué afirmaciones son verdaderas? Recuerde que un valor p es la probabilidad de los datos observados (o de puntos de datos más extremos), dado que la hipótesis nula H_0 es verdadera, definida en símbolos como $p(D|H_0)$[7]. Esta definición puede ser reformulada en una forma más técnica mediante la introducción del modelo estadístico subyacente al análisis (Gigerenzer et al., 1989, capítulo 3).

Las afirmaciones 1 y 3 se detectan fácilmente como falsas, porque una prueba de significación nunca puede refutar la hipótesis nula o la hipótesis experimental (no definida). Son ejemplos de la *ilusión de certeza* (Gigerenzer, 2002).

Las afirmaciones 2 y 4 también son falsas. La probabilidad $p(D|H_0)$ no es la misma que $p(H0|D)$ y, de manera más general, una prueba de significación no proporciona una probabilidad para una hipótesis. La caja de herramientas estadísticas, por supuesto, contiene herramientas que permitirían estimar probabilidades de hipótesis, como la estadística bayesiana. La afirmación 5 también se refiere a la probabilidad de una hipótesis. Esto se debe a que si uno rechaza la hipótesis nula, la única posibilidad de tomar una decisión incorrecta es si la hipótesis nula es verdadera. Por lo tanto, hace esencialmente la misma afirmación que la afirmación 2, y ambas son incorrectas. La afirmación 6 equivale a la falacia de la replicación (Gigerenzer, 1993, 2000). Aquí, p = 1% se toma para implicar que estos datos significativos reaparecerían en el 99% de las repeticiones. El enunciado 6 solo se puede hacer si se sabe que la hipótesis nula es verdadera. En términos formales, $p(D|H_0)$ se confunde con 1 - $p(D)$.

En resumen, las seis afirmaciones son incorrectas. Tenga en cuenta que los seis yerran en la misma dirección de las ilusiones: hacen que un valor p parezca más informativo de lo que es.

Haller y Krauss (2002) plantearon la pregunta anterior a 30 profesores de estadística, incluidos profesores de psicología, ayudantes y asistentes de enseñanza, a 39 profesores y ayudantes de psicología (que no impartían estadística) y a 44 estudiantes de psicología. Los profesores y estudiantes eran de los departamentos de psicología de seis universidades alemanas. Cada profesor de estadística había

[7] Probabilidad de D, sabiendo que H_0 es verdadera. (N. trad.)

enseñado pruebas de hipótesis nulas y cada alumno había aprobado con éxito uno o más cursos de estadística en los que se enseñó. La figura 1 muestra los resultados.

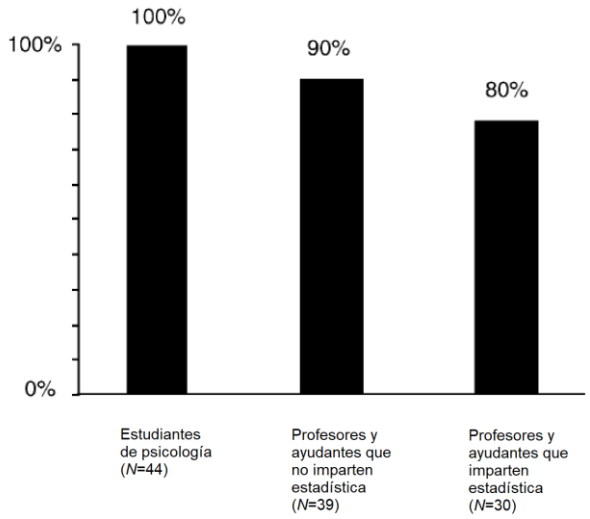

Fig. 1. La cantidad de delirios sobre el significado de «*p* = 0,01». Los porcentajes se refieren a los participantes de cada grupo que respaldaron una o más de las seis declaraciones falsas (véase Gigerenzer et al., 2004; Haller y Krauss, 2002).

Ninguno de los estudiantes notó que todas las afirmaciones estaban equivocadas; todos los estudiantes respaldaron una o más de las ilusiones sobre el significado de un valor *p*. ¿Quizás estos estudiantes carecían de los genes adecuados para el pensamiento estadístico? ¿O no prestaron atención a sus profesores y simplemente tuvieron suerte al aprobar los exámenes? Los resultados, sin embargo, indican una explicación diferente.

Los estudiantes heredaron las ilusiones de sus profesores. El noventa por ciento de los profesores y ayudantes creían que una o más de las seis afirmaciones eran correctas. Lo más sorprendente es que el 80% de los profesores de estadística compartían ilusiones con sus alumnos. Tenga en cuenta que no es necesario ser un matemático brillante para responder la pregunta «¿Qué significa un resultado significativo?» Uno solo necesita entender que un valor *p* es la probabilidad de los datos (o datos más extremos), dado que el H_0 es verdadero. La ilusión más frecuente fue la afirmación 5, respaldada por aproximadamente el 70% de los tres grupos. En un estudio anterior con psicólogos académicos en el Reino Unido (Oakes, 1986) hasta el 86% pensaba que esta afirmación era cierta. La falacia de la replicación (afirmación 6) fue la segunda ilusión más frecuente, considerada cierta por aproximadamente la mitad de los profesores y el 37% de los que enseñaban estadística. La cifra correspondiente para los psicólogos del Reino Unido fue del 60%. Aproximadamente el 60% de los estudiantes y un tercio de cada grupo de profesores creían que se puede deducir la probabilidad de que la hipótesis experimental sea cierta a partir del valor *p* (afirmación 4). En el estudio de Oakes,

dos tercios de los psicólogos académicos británicos creían esto. En promedio, los estudiantes respaldaron 2,5 ilusiones, sus profesores y ayudantes 2,0 ilusiones, y aquellos que enseñaron pruebas de significación respaldaron 1,9 ilusiones (Gigerenzer et al., 2004; Haller y Krauss, 2002). Con todo, a los profesores y ayudantes alemanes les fue algo mejor que a los psicólogos académicos británicos estudiados anteriormente por Oakes (1986), sin embargo, el número de ilusiones que tenían sigue siendo impresionante. Falk y Greenbaum (1995) agregaron la alternativa correcta («ninguna de las declaraciones es correcta») y también hicieron que los estudiantes israelíes leyeran el artículo clásico de Bakan (1966), que advierte sobre estas ilusiones. Sin embargo, el 87% de los estudiantes optaron por una o varias ilusiones. Una fantasía global parece viajar por transmisión cultural de maestro a alumno. Si los estudiantes «heredaron» las ilusiones de sus profesores, ¿dónde las adquirieron los profesores? La respuesta está ahí mismo en los primeros libros de texto que introdujeron a los psicólogos a las pruebas de hipótesis nulas hace más de 50 años. *Estadística fundamental en psicología y educación* de Guilford[8], publicado por primera vez en 1942, fue probablemente el libro de texto más leído en las décadas de 1940 y 1950. Guilford sugirió que la prueba de hipótesis revelaría la probabilidad de que la hipótesis nula sea verdadera. «Si el resultado sale de una manera, la hipótesis probablemente sea correcta, si sale de otra manera, la hipótesis probablemente sea incorrecta» (p. 156). La lógica de Guilford vaciló entre enunciados correctos e incorrectos y ambiguos que pueden leerse como manchas de tinta de Rorschach. Usó frases como «obtuvimos directamente las probabilidades de que la hipótesis nula fuera plausible» y «la probabilidad de desviaciones extremas del azar» de manera intercambiable para el nivel de significado. Guilford no es una excepción. Marcó el comienzo de un género de textos estadísticos que vacilan entre el deseo de los investigadores por las probabilidades de hipótesis y lo que realmente pueden proporcionar las pruebas de significado. Por ejemplo, en tres páginas de texto, Nunally (1975, págs. 194-196; cursiva en el original) usó todas las siguientes declaraciones para explicar lo que realmente significa un resultado significativo como el 5%:

• «la probabilidad de que una diferencia observada sea real»

• «la *improbabilidad* de que los resultados observados se deba a un error»

• «la *confianza estadística* ... con probabilidades de 95 sobre 100 de que la diferencia observada se mantenga en las investigaciones»

• «el peligro de aceptar un resultado estadístico como real cuando en realidad se debe únicamente a un error»

• el grado en que los resultados experimentales se toman «en serio»

• el grado de «fe [que] puede depositarse en la realidad del hallazgo»

[8] Trad. cast. México; Bogotá: McGraw-Hill, 1984.

• «el investigador puede tener un 95% de confianza en que la media de la muestra en realidad difiere de la media de la población»

• «si la probabilidad es baja, la hipótesis nula es improbable»

• «todas estas son formas diferentes de decir lo mismo»

¡Los pobres estudiantes que leyeron estas explicaciones! Es probable que atribuyan erróneamente la confusión del autor a su propia falta de inteligencia estadística. Este estado de desconcierto durará mientras el ritual continúe existiendo. Los estudiantes de hoy en día todavía encuentran declaraciones oraculares en los textos más leídos: «Las estadísticas inferenciales indican la probabilidad de que la muestra particular de puntajes obtenidos esté realmente relacionada con lo que sea que esté tratando de medir o si podrían haber ocurrido por casualidad» (Gerrig y Zimbardo , 2002, p.44).

Los primeros autores que promueven el error de que el nivel de significación especifica la probabilidad de hipótesis incluyen a Anastasi (1958, p. 11), Ferguson (1959, p. 133) y Lindquist (1940, p. 14). Pero la creencia ha persistido durante décadas: por ejemplo, en Miller y Buckhout (1973; apéndice estadístico de Brown, p. 523), y en los ejemplos recopilados por Bakan (1966), Pollard y Richardson (1987), Gigerenzer (1993) , Mulaik y col. (1997) y Nickerson (2000). A veces escucho que si se eliminaran las ilusiones asociadas, el ritual nulo emergería como un método significativo. Como mencioné antes, en cambio, creo que es necesario cierto grado de ilusión para mantener vivo el ritual nulo, y la evidencia empírica apoya esta conjetura (por ejemplo, Lecoutre et al., 2003; Tversky y Kahneman, 1971). Sin ilusiones, el ritual se reconocería fácilmente por lo que es.

5. Un editor con agallas

Todos parecen tener una respuesta a esta pregunta: ¿Quién tiene la culpa del ritual nulo? Siempre alguien más. Un estudiante de posgrado inteligente me dijo que no quería problemas con su asesor de tesis. Cuando finalmente obtuvo su doctorado y un postdoctorado, su preocupación era conseguir un trabajo real. Pronto fue profesor asistente en una universidad respetada, pero todavía sentía que no podía permitirse el pensamiento estadístico porque necesitaba publicar rápidamente para conseguir la titularidad. Los editores exigían el ritual, se disculpó, pero después de la titularidad, todo sería diferente y él sería un hombre libre. Años más tarde, se encontró titular, pero aún en el mismo entorno. Y le habían pedido que impartiera un curso de estadística, con el ritual nulo. Él lo hizo. Mientras los editores de las principales revistas castiguen el pensamiento estadístico, concluyó, nada cambiará.

Culpar a los editores no es del todo infundado. Por ejemplo, el ex editor *del Journal of Experimental Psychology*, Melton (1962), insistió en el ritual nulo en su editorial y

también dejó en claro que quiere ver p <0,01, no solo p <0,05. En su editorial, produjo las ilusiones habituales, afirmando que cuanto menor es el valor p, mayor es la confianza en que la hipótesis alternativa es verdadera y mayor es la probabilidad de que una réplica encuentre un resultado significativo. No se mencionó nada más allá de los valores p; hipótesis precisas, buenas estadísticas descriptivas, intervalos de confianza, tamaños de efecto y poder no aparecían en la definición del editor de buena investigación. Un pequeño valor p era el sello distintivo de una experimentación excelente, un criterio conveniente para aceptar o no un artículo en un momento en que el número de revistas, artículos y psicólogos se había disparado.

Hubo resistencia. Los skinnerianos fundaron una nueva revista, *Journal of the Experimental Analysis of Behavior*, para poder publicar su tipo de experimentos (Skinner, 1984, p. 138). De manera similar, una de las razones para lanzar el *Journal of Mathematical Psychology* fue escapar de la presión de los editores para realizar de forma rutinaria pruebas de hipótesis nulas. Uno de sus fundadores, R. D. Luce (1988), llamó a esta práctica una «prueba de hipótesis sin sentido en lugar de hacer una buena investigación: medir efectos, construir teorías sustantivas de cierta profundidad y desarrollar modelos de probabilidad y procedimientos estadísticos adecuados para estas teorías» (p. 582).

¿Deberíamos culpar a los editores? Sin embargo, la historia de Geoffrey Loftus, editor de *Memory and Cognition*, sugiere que la verdad no es tan simple como eso. En 1991, Loftus revisó *El imperio del azar* (Gigerenzer et al., 1989), en el que presentamos uno de los primeros análisis de cómo los psicólogos mezclaron las ideas de Fisher y Neyman-Pearson en una lógica híbrida. Cuando Loftus se convirtió en editor electo de *Memory and Cognition*, dejó en claro en su editorial que no quería que los autores enviaran artículos en los que los valores p, t o F hubieran sido calculados e informados sin pensar (Loftus, 1993). Más bien, su directriz fue: «Por defecto, los datos deben transmitirse como una figura que describa las medias de la muestra *con errores estándar asociados y/o, cuando corresponda, desviaciones típicas*» (p. 3; subrayado en el original). Su política alentó a los investigadores a usar estadísticas descriptivas adecuadas y los liberó de la presión de probar hipótesis nulas y tomar decisiones de sí o no cuya relevancia es oscura. Admiro aLoftus por el valor de dar ese paso.

Cuando conocí a Loftus durante su mandato como editor, le pregunté cómo iba su cruzada. Loftus se quejó amargamente de los muchos investigadores que rechazaron obstinadamente la oportunidad e insistieron en sus valores p y decisiones de sí-no. ¿Cuánto éxito tuvo a lo largo de los años? Loftus fue precedido como editor por Margaret Jean Intons-Petersen, que comenzó en 1990. En su editorial de ingreso mencionó el uso de estadísticas descriptivas incluyendo estimaciones de variabilidad, pero enfatizó las pruebas de significación habituales. Durante su mandato, el 53% de los artículos se basó exclusivamente en el ritual nulo (Finch et al., 2004). Bajo Loftus, quien se desempeñó como editor de 1994 a 1997, esta proporción disminuyó al 32%. Durante el mandato de la sucesora de Loftus, Morton Ann Gernsbacher (1998), quien no hizo comentarios sobre los

procedimientos estadísticos o las recomendaciones de Loftus en su editorial, la proporción volvió a subir a aproximadamente la mitad, alcanzando un nuevo máximo del 55% en 2000. La gran mayoría de los artículos restantes también se basaron en el ritual nulo, pero proporcionaron información adicional, como cifras con medias, errores estándar, desviaciones típicas o intervalos de confianza. La recomendación de Loftus de proporcionar esta información sin realizar el ritual nulo fue seguida en solo el 6% de los artículos durante su dirección como editor, y solo en un (!) caso en los años anterior y posterior (Finch et al., 2004). Antes de Loftus, solo el 8% de los artículos proporcionaban cifras con barras de error y/o informaban de intervalos de confianza, y en este caso no quedaba claro qué representaban las barras: ¿errores estándar, desviaciones típicas, intervalos de confianza? Loftus elevó esta proporción al 45% y redujo la de las barras de error poco claras (Finch et al., 2004). Pero bajo su sucesor, la proporción volvió a bajar al 27% y la de las barras poco claras aumentó.

Loftus informó que muchos investigadores mostraron una profunda ansiedad ante la perspectiva de abandonar sus valores *p*, confundieron los errores estándar con las desviaciones típicas y no tenían idea de cómo calcular un intervalo de confianza basado en sus paquetes ANOVA. Mirando hacia atrás, calculó que solicitó aproximadamente 300 intervalos de confianza, y probablemente calculó alrededor de 100 por sí mismo (Finch et al., 2004). ¿El experimento de Loftus tuvo el impacto deseado? Durante su mandato como director, logró reducir la dependencia del ritual nulo; luego, el efecto declinó. Si su ejemplo tiene un impacto a largo plazo es una cuestión abierta. Loftus se adelantó a su tiempo y solo puedo esperar que su admirable experimento finalmente inspire a otros editores.

Lo que está en juego aquí es la importancia de una buena estadística descriptiva y exploratoria en lugar de la prueba de hipótesis mecánicas con respuestas sí-no. Una buena estadística descriptiva (a diferencia de las cifras sin barras de error o barras de error poco claras, y el agregado de rutina en lugar del análisis individual, por ejemplo) es necesaria y, en general, suficiente. Tenga en cuenta que en los problemas científicos, la relevancia de los procedimientos de optimización como la teoría de decisiones de Neyman-Pearson es notoriamente poco clara. Por ejemplo, a diferencia del control de calidad, los sujetos experimentales rara vez se muestrean al azar de una población específica. Por lo tanto, no está claro para qué población debe hacerse la inferencia de una muestra, y las decisiones «óptimas» de sí-no son de poca relevancia. El intento de dar una respuesta «óptima» a la pregunta incorrecta se ha denominado «error de tipo III». El estadístico John Tukey (por ejemplo, 1969) abogó por un cambio en la perspectiva: una respuesta adecuada al problema correcto es mejor que una respuesta óptima al problema incorrecto (Perlman y Wu, 1999). Ni la prueba de hipótesis nula de Fisher ni la teoría de decisiones de Neyman-Pearson pueden responder a la mayoría de los problemas científicos. La cuestión de la optimización frente a la satisfacción es igualmente relevante para la investigación sobre la racionalidad limitada y la heurística rápida y frugal (Gigerenzer et al., 1999; Todd y Gigerenzer, 2000).

6. El superego, el ego y el ello

¿Por qué la gente inteligente se involucra en rituales estadísticos en lugar de en pensamiento estadístico? Toda persona de inteligencia promedio puede entender que $p(D|\mathrm{H})$ no es lo mismo que $p(\mathrm{H}|D)$. El hecho de que esta idea se desvanezca cuando se trata de probar hipótesis sugiere que la causa no es intelectual, sino social y emocional. He aquí una hipótesis (Acree, 1978; Gigerenzer, 1993): El conflicto entre estadísticos, tanto reprimido como inherente a los libros de texto, se ha internalizado en la mente de los investigadores. El ritual estadístico es una forma de resolución de conflictos, como lavarse las manos compulsivamente, lo que lo hace resistente a las discusiones. Para ilustrar esta tesis, utilizo los conflictos inconscientes freudianos como analogía (Fig. 2).

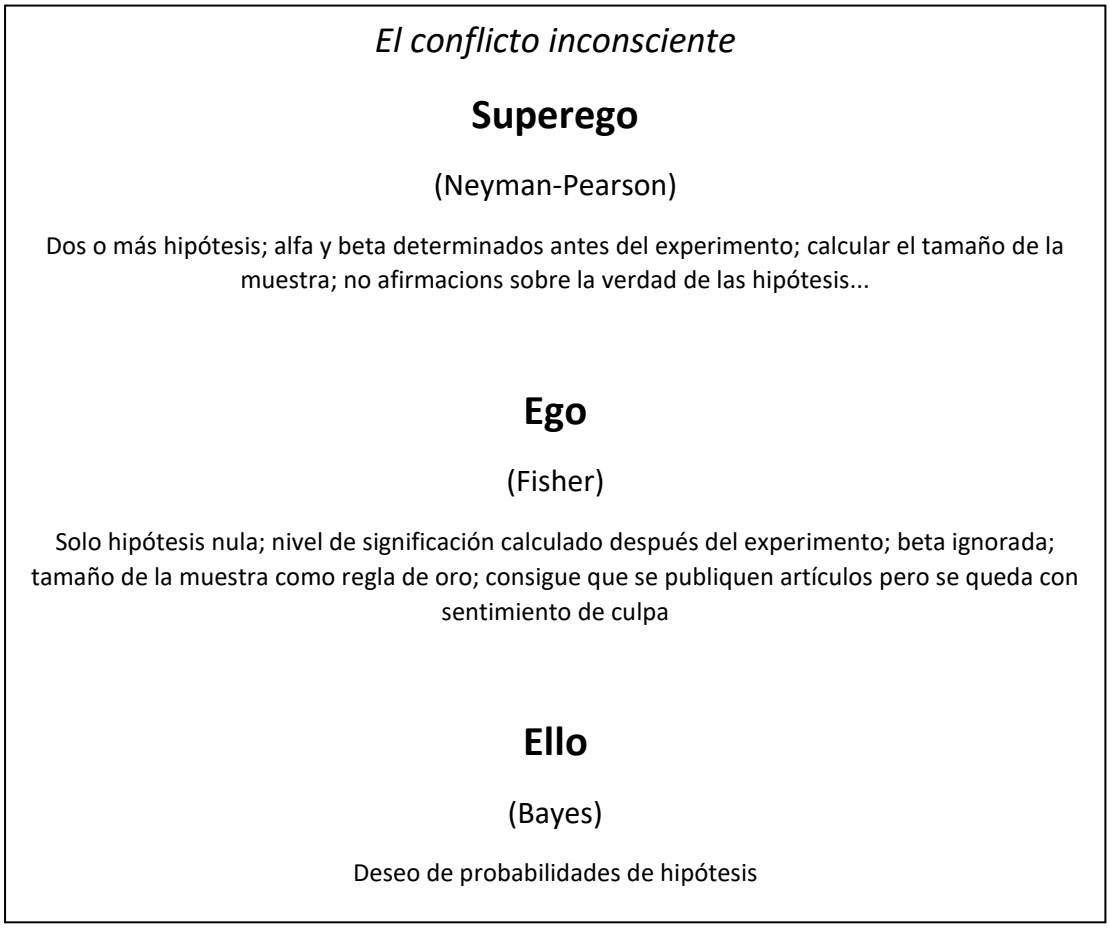

Fig. 2. Una analogía freudiana del conflicto inconsciente entre ideas estadísticas en la mente de los investigadores.

La teoría de Neyman-Pearson sirve como superego del Dr. Publica-o-Perece. Exige de antemano que las hipótesis alternativas, el alfa y el poder para calcular el tamaño de la muestra necesario se especifiquen con precisión, siguiendo la doctrina frecuentista del muestreo aleatorio repetido (Neyman, 1957). El superego prohíbe la interpretación de niveles de significación como el grado de confianza en que una hipótesis particular es verdadera o falsa. La prueba de hipótesis trata de qué hacer, es decir, uno actúa como si una hipótesis fuera verdadera o falsa, sin creer necesariamente que sea verdadera o falsa.

La teoría de Fisher de la prueba de hipótesis nulas funciona como el ego. El ego hace las cosas en el laboratorio y publica los artículos. Los niveles de significación se calculan después del experimento, se ignora la potencia de la prueba y el tamaño de la muestra se determina mediante una regla general. El ego no establece su hipótesis de investigación de una manera precisa, sino en el mejor de los casos en forma de predicción direccional, pero no duda en reclamar apoyo para ella rechazando una hipótesis nula. El ego hace abundantes declaraciones epistémicas sobre su confianza en hipótesis particulares. Pero se queda con sentimientos de culpa y vergüenza por haber violado las reglas.

El punto de vista bayesiano constituye el ello. Su objetivo es una declaración sobre las probabilidades de hipótesis, que es censurada tanto por el superego purista como por el ego pragmático. Sin embargo, estas probabilidades son exactamente lo que quiere el ello, después de todo. Se sale con la suya al impedir que el intelecto comprenda que $p(D|\mathrm{H})$ no es lo mismo que $p(\mathrm{H}|D)$. Esto permite hacer ilusiones. La analogía freudiana pone en primer plano la ansiedad y los sentimientos de culpa. Parece como si los furiosos conflictos personales e intelectuales entre Fisher y Neyman-Pearson, y entre estos frecuentistas y los bayesianos se proyectaran en una conflicto «intrapsíquico» en la mente de los investigadores. En la teoría freudiana, el ritual es una forma de resolver conflictos inconscientes, pero a costos considerables.

7. La conjetura de Meehl

Paul Meehl, un brillante psicólogo clínico con un amplio interés en la filosofía de la ciencia, fue uno de los que culpó a Fisher por el declive del pensamiento estadístico en psicología. «Sir Ronald nos ha confundido, hipnotizado y conducido por un camino de rosas. Creo que la confianza casi universal en simplemente refutar la hipótesis nula ... es un error terrible, es básicamente una estrategia científica errónea, deficiente y una de las peores cosas que han sucedido en la historia de la psicología» (Meehl, 1978, p. 817). Meehl es un poco severo al culpar a Fisher más bien del ritual nulo; recordemos que Fisher también propuso otras herramientas estadísticas y, en la década de 1950, pensó que las pruebas de hipótesis nulas eran adecuadas solo para situaciones en las que sabemos poco o nada. Meehl (1978) hizo una predicción desafiante con respecto a las pruebas de hipótesis nulas en entornos

no experimentales, donde la asignación aleatoria al grupo de tratamiento y control no es posible debido a limitaciones éticas o prácticas. Se puede resumir de la siguiente manera:

Conjetura de Meehl:

En entornos no experimentales con tamaños de muestra grandes, la probabilidad de rechazar la hipótesis nula de diferencias de grupo nulas a favor de una alternativa direccional es de aproximadamente 0,50.

¿No son buenas noticias? Suponemos que X es más grande que Y, y acertamos la mitad de las veces. Por ejemplo, si inventamos la historia de que los protestantes tienen una mayor capacidad de memoria que los católicos, tiempos de reacción más lentos, un tamaño de zapato más pequeño y niveles más altos de testosterona, cada una de estas hipótesis tiene aproximadamente un 50% de probabilidad de ser aceptada por una prueba de hipótesis nula. . Si no nos comprometemos con la dirección y simplemente adivinamos que X e Y son diferentes, lo hacemos bien prácticamente el 100% de las veces. Meehl razonó que en el mundo real, a diferencia de los escenarios experimentales, la hipótesis nula («nula», según la define el ritual nulo, no Fisher) siempre es incorrecta. Existe alguna diferencia entre los grupos naturales. Por lo tanto, con suficiente poder estadístico, casi siempre se encontrará un resultado significativo. Si uno adivina al azar la dirección de la diferencia, se deduce que será correcto en aproximadamente el 50% de los casos (con una hipótesis alternativa unidireccional, uno estará en lo correcto en aproximadamente el 100% de ellos).

Niels Waller (2004) se propuso probar empíricamente la conjetura de Meehl. Tuvo acceso a los datos de más de 81,000 personas que habían completado los 567 ítems del Inventario Multifásico de Personalidad de Minnesota-Revisado (MMPI-2). El MMPI-2 pregunta a las personas sobre una amplia gama de contenidos, que incluyen salud, hábitos personales, actitudes hacia el sexo y manifestaciones extremas de psicopatología. Imagínese un teórico de género que ha elaborado una nueva teoría que predice diferencias direccionales de género, es decir, las mujeres obtendrán una puntuación más alta en algún elemento que los hombres, o viceversa. ¿Podemos predecir la probabilidad de rechazar la hipótesis nula a favor de la nueva teoría? Según la conjetura de Meehl, es alrededor del 50%. En la simulación de Waller, la computadora seleccionó el primero de los 511 ítems del MMPI-2 (excluyendo 56 por su capacidad conocida para discriminar entre los sexos), determinó al azar la dirección de la hipótesis alternativa y calculó si la diferencia era significativa en la dirección prevista. Este procedimiento se repitió con los 511 elementos. El resultado: el 46% de las predicciones se confirmaron, a menudo con valores *p* muy

impresionantes. ¡Muchas de las diferencias medias de los ítems eran 50-100 veces mayores que los errores estándar asociados! Estos resultados empíricos apoyan la conjetura de Meehl, consistente con hallazgos anteriores de Bakan (1966) y el propio Meehl. Un poco de pensamiento estadístico puede hacer que la lógica de la conjetura sea transparente para un estudiante. Sin embargo, se pueden encontrar investigadores experimentados que informan con orgullo que han estudiado varios cientos o incluso miles de sujetos y han encontrado una diferencia media muy significativa en la dirección prevista, digamos p <0,0001. Sin embargo, la magnitud de este efecto no se informa en algunos de estos artículos. La combinación de un tamaño de muestra grande y valores p bajos tiene poco valor en sí misma.

El problema general que aborda Meehl es la falta de atención a los tamaños de los efectos en el ritual nulo. Cohen (1988) y Rosenthal y Rubin (1982) han analizado los tamaños del efecto. El Grupo de Trabajo sobre Inferencia Estadística (TFSI) de la *Asociación Americana de Psicología* (Wilkinson y TFSI, 1999) recomendó informar los tamaños del efecto (teóricos como en la teoría de Neyman-Pearson, o empíricos) como esencial. La quinta edición del *Manual de publicaciones de la Asociación Americana de Psicología* (2001) siguió esta recomendación, aunque a medias. En los ejemplos dados, los tamaños del efecto no se incluyen o no se explican e interpretan (Fidler, 2002). Sin un tamaño del efecto teórico, no se puede calcular la potencia estadística de una prueba. En 1962, Jacob Cohen informó que los experimentos publicados en una importante revista de psicología tenían, en promedio, solo un cincuenta por ciento de posibilidades de detectar un efecto de tamaño mediano, si lo hubiera. Es decir, el poder estadístico fue tan bajo como el 50%. Este resultado fue ampliamente citado, pero ¿cambió la práctica de los investigadores? Sedlmeier y Gigerenzer (1989) revisaron los estudios en la misma revista, 24 años después, un período de tiempo que debería permitir cambios. Sin embargo, solo 2 de 64 investigadores mencionaron el poder y nunca se estimó. Inadvertidamente, la potencia promedio había disminuido (los investigadores ahora usaban el ajuste alfa, que reduce la potencia). Por lo tanto, si hubiera habido un efecto de tamaño mediano, los investigadores habrían tenido más posibilidades de encontrarlo arrojando una moneda en lugar de realizar sus experimentos costosos, elaborados y que consumen mucho tiempo. En los años 2000–2002, entre unos 220 artículos empíricos, finalmente hubo 9 investigadores que calcularon el poder de sus pruebas (Gigerenzer et al., 2004). Cuarenta años después de Cohen, hay una primera señal de cambio. La cuarta edición del *Manual de publicaciones de la Asociación Americana de Psicología* (1994) fue la primera en recomendar que los investigadores se tomaran el poder en serio, y la quinta edición (2001) repitió este consejo. Sin embargo, a pesar de la abundancia de ejemplos sobre cómo informar valores p, el manual todavía no incluye ningún ejemplo de poder de informe (Fidler, 2002).

8. Conjetura de Feynman

La dependencia rutinaria del ritual nulo desalienta no solo el pensamiento estadístico sino también el pensamiento teórico. No es necesario especificar la hipótesis de uno, ni ninguna hipótesis alternativa desafiante. No se premian las hipótesis «audaces», en el sentido de Karl Popper o la comparación del modelo bayesiano (MacKay, 1995). En muchos artículos experimentales de psicología social y cognitiva, no existe una teoría sobre la «distancia del tiro», sino solo sustitutos como la redescripción de los resultados (Gigerenzer, 2000, capítulo 14). El único requisito es rechazar una hipótesis nula que se identifica con el «azar». Las teorías estadísticas como la teoría de Neyman-Pearson y la teoría de Wald, por el contrario, comienzan con dos o más hipótesis estadísticas.

En ausencia de teoría, la tentación es mirar primero los datos y luego ver qué es significativo. El físico Richard Feynman (1998, págs. 80-81) se ha dado cuenta de este mal uso de la prueba de hipótesis. Resumo su argumento.

Conjetura de Feynman:

Informar un resultado significativo y rechazar el nulo a favor de una hipótesis alternativa no tiene sentido a menos que la hipótesis alternativa se haya establecido antes de obtener los datos.

Cuando era un estudiante de posgrado en Princeton, Feynman tuvo una discusión con un investigador del departamento de psicología. El investigador había diseñado un experimento en el que las ratas corrían en un laberinto en T. Las ratas no se comportaron como se predijo. Sin embargo, el investigador notó algo más, que las ratas parecen alternarse, primero a la derecha, luego a la izquierda, luego a la derecha de nuevo, y así sucesivamente. Le pidió a Feynman que calcule la probabilidad bajo la hipótesis nula (posibilidad) de que se obtenga este patrón. En esta ocasión, Feynman (1998) planteó el nivel del 5%:

> Y es un principio general de los psicólogos que en estas pruebas se arreglan de manera que las probabilidades de que las cosas que ocurren sucedan por casualidad sean pequeñas, de hecho, menos de una en veinte. ... Y luego corrió hacia mí y me dijo: «Calcula la probabilidad para mí de que se alternen, de modo que pueda ver si es menos de uno en veinte». Dije: «Probablemente sea menos de uno en veinte, pero no cuenta». Él dijo: «¿Por qué?» Dije: «Porque no tiene ningún sentido calcular después del evento. Verá, usted encontró la peculiaridad y, por lo tanto, seleccionó el caso peculiar. ... Si quiere probar esta hipótesis, una de cada veinte, no puede hacerlo a partir de los mismos

> datos que le dieron la pista. Debe hacer otro experimento de nuevo y luego ver si se alternan. Lo hizo, y no funcionó.» (págs. 80-81)

La conjetura de Feynman es violada una y otra vez por las rutinarias pruebas de significación, donde uno mira los datos para ver qué es significativo. Los paquetes estadísticos permiten probar todas las diferencias, interacciones o correlaciones con el azar. Entregan automáticamente calificaciones de «significación» en términos de estrellas, estrellas dobles y estrellas triples, lo que fomenta el mal hábito posterior a los hechos. El problema general que abordó Feynman se conoce como sobreajuste. Ajustar un modelo a los datos que ya se han obtenido no es una prueba de hipótesis sólida, incluso si la varianza explicada resultante, o R^2, es impresionante. La razón es que no se sabe cuánto ruido se ha instalado, y cuantos más parámetros ajustables tenga, más ruido se puede ajustar. Los psicólogos habitualmente ajustan en lugar de predecir, y rara vez prueban un modelo con datos nuevos, como mediante una validación cruzada (Roberts y Pashler, 2000). La adaptación per se tiene los mismos problemas que la narración de historias después de los hechos, lo que conduce a un «sesgo retrospectivo» (Hoffrage et al., 2000). La verdadera prueba de un modelo es fijar sus parámetros en una muestra y probarla en una nueva muestra. Entonces resulta que las predicciones basadas en heurísticas simples pueden ser más precisas que las regresiones múltiples rutinarias (Czerlinski et al., 1999). Menos puede ser más. El uso rutinario de la regresión múltiple lineal ejemplifica otro uso insensato de la estadística.

9. Los albores del pensamiento estadístico

Los rituales parecen ser indispensables para la autodefinición de los grupos sociales y para las transiciones en la vida, y no tienen nada de malo. Sin embargo, deberían ser el tema y no el procedimiento de las ciencias sociales. Los elementos de los rituales sociales incluyen (i) la repetición de la misma acción, (ii) un enfoque en números o colores especiales, (iii) temores sobre sanciones graves por violaciones de las reglas, y (iv) deseos o ilusiones que virtualmente eliminan el pensamiento crítico. (Dulaney y Fiske, 1994). El ritual nulo tiene cada una de estas cuatro características: el mismo procedimiento se repite una y otra vez; el número mágico del 5%; miedo a las sanciones de los editores o supervisores, e ilusiones sobre el resultado, el valor *p*, que bloquea la inteligencia de los investigadores.

Sabemos, pero a menudo olvidamos, que el problema de la inferencia inductiva no tiene una solución única. No existe una prueba uniformemente más poderosa, es decir, ningún método que sea mejor para cada problema. La teoría estadística nos ha proporcionado una caja de herramientas con instrumentos efectivos, que requieren un juicio sobre cuándo es correcto usarlos. Cuando los libros de texto y

los planes de estudio comiencen a enseñar la caja de herramientas, los estudiantes aprenderán automáticamente a emitir juicios. Y se darán cuenta de que en muchas aplicaciones, un análisis de datos descriptivo hábil y transparente es suficiente y preferible a la aplicación de rutinas estadísticas elegidas por su complejidad y opacidad. El juicio es parte del arte de la estadística.

Para detener el ritual, también necesitamos más agallas y nervios. Necesitamos algunos kilos de valor para dejar de seguir el juego en este vergonzosa práctica. Esto puede causar fricciones con los editores y colegas, pero al final les ayudará a entrar en los albores del pensamiento estadístico.

Bibliografía

Acree, M. C., 1978. Theories of Statistical Inference in Psychological Research: A Historicocritical Study. Dissertation. University Microfilms International H790 H7000, Ann Arbor, MI.

American Psychological Association, 1974. Publication Manual, 2nd ed., 3rd ed., 1983; 4th ed., 1994; 5th ed., 2001. Garamond/Pridemark Press, Baltimore, MD.

Anastasi, A., 1958. Differential psychology, 3rd ed. Macmillan, New York.

Anderson, D. R., Burnham, K. P., Thompson, W. L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64, 912–923.

Bakan, D., 1966. The test of significance in psychological research. Psychological Bulletin 66, 423–437.

Chow, S. L., 1998. Pr´ecis of “Statistical significance: rationale, validity, and utility”. Behavioral and Brain Sciences 21, 169–239.

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

Cohen, J., 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Czerlinski, J., Gigerenzer, G., Goldstein, D. G., 1999. How good are simple heuristics? In: Gigerenzer, G., Todd, P. M., the ABC Reading Group, Simple Heuristics That Make Us Smart. Oxford University Press, New York, pp. 97–118.

Dulaney, S., Fiske, A. P., 1994. Cultural rituals and obsessive-compulsive disorder: is there a common psychological mechanism? Ethos 22, 243–283.

Falk, R., Greenbaum, C. W., 1995. Significance tests die hard. Theory and Psychology 5, 75–98.

Ferguson, L., 1959. Statistical Analysis in Psychology and Education. McGraw-Hill, New York.

Feynman, R., 1998. The Meaning of it All: Thoughts of a Citizen-Scientist. Perseus Books, Reading, MA, pp. 80–81.

Fidler, F., 2002. The fifth edition of the APA Publication Manual: why its statistics recommendations are so controversial. Educational and Psychological Measurement 62, 749–770.

Finch Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. 2004. Reform of statistical inference in psychology: the case of memory and cognition. Behavior Research Methods, Instruments and Computers 36, 312–324.

Fisher, R. A., 1935. The design of experiments, 5th ed., 1951; 7th ed., 1960; 8th ed., 1966. Oliver & Boyd, Edinburgh.

Fisher, R. A., 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society (B) 17, 69–77.

Fisher, R. A., 1956. Statistical Methods and Scientific Inference. Oliver & Boyd, Edinburgh.

Gernsbacher, M. A., 1998. Editorial comment. Memory and Cognition 26, 1.

Gerrig, R. J., Zimbardo, P. G., 2002. Psychology and Life, 16th ed. Allyn and Bacon, Boston.

Gigerenzer, G., 1987. Probabilistic thinking and the fight against subjectivity. In: Kr¨uger, L., Gigerenzer, G., Morgan, M. (Eds.), The Probabilistic Revolution. Vol. II: ideas in the Sciences. MIT Press, Cambridge, MA, pp. 11–33.

Gigerenzer, G., 1993. The superego, the ego, and the id in statistical reasoning. In: Keren, G., Lewis, C. (Eds.), A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues. Erlbaum, Hillsdale, NJ, pp. 311–339.

Gigerenzer, G., 2000. Adaptive Thinking: Rationality in the Real World. Oxford University Press, New York.

Gigerenzer, G., 2002. Calculated Risks: How to Know When Numbers Deceive You. Simon & Schuster, New York (UK edition: Reckoning with Risk: Learning to Live with Uncertainty. Penguin, London).

Gigerenzer, G., Krauss, S., Vitouch, O., 2004. The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In: Kaplan, D. (Ed.), Handbook on Quantitative Methods in the Social Sciences. Sage, Thousand Oaks, CA, pp. 389–406.

Gigerenzer, G., Murray, D.J., 1987. Cognition as Intuitive Statistics. Erlbaum, Hillsdale, NJ.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance. How Probability Changed Science and Every Day Life. Cambridge University Press, Cambridge, UK.

Gigerenzer, G., Todd, P. M., The ABC Research Group, 1999. Simple Heuristics that Make Us Smart. Oxford University Press, New York.

Guilford, J. P., 1942. Fundamental Statistics in Psychology and Education, 3rd ed., 1956; 6th ed., 1978 (with Fruchter, B). McGraw-Hill, New York.

Haller, H., Krauss, S., 2002. Misinterpretations of significance: a problem students share with their teachers? Methods of Psychological Research—Online [On-line serial], 7, 1–20. Retrieved June 10, 2003, from http://www.mpr-online.de.

Hoffrage, U., Hertwig, R., Gigerenzer, G., 2000. Hindsight bias: a by-product of knowledge updating? Journal of Experimental Psychology: Learning, Memory, and Cognition 26, 566–581.

Intons-Peterson, M. J., 1990. Editorial. Memory and Cognition 18, 1–2.

Lecoutre, M. P., Poitevineau, J., Lecoutre, B., 2003. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. International Journal of Psychology 38, 37–45.

Lindquist, E. F., 1940. Statistical Analysis in Educational Research. Houghton Mifflin, Boston.

Loftus, G. R., 1991. On the tyranny of hypothesis testing in the social sciences. Contemporary Psychology 36, 102–105.

Loftus, G. R., 1993. Editorial comment. Memory and Cognition 21, 1–3.

Luce, R. D., 1988. The tools-to-theory hypothesis. ReviewofG. Gigerenzer and D.J. Murray, “Cognition as intuitive statistics”. Contemporary Psychology 33, 582–583.

MacKay, D. J., 1995. Probable networks and plausible predictions: a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6, 469–505.

McCloskey, D. N., Ziliak, S., 1996. The standard error of regression. Journal of Economic Literature 34, 97–114.

Meehl, P. E., 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology 46, 806–834.

Melton, A. W., 1962. Editorial. Journal of Experimental Psychology 64, 553–557.

Miller, G. A., Buckhout, R., 1973. Psychology: The Science of Mental Life. Harper & Row, New York.

Morrison, D. E., Henkel, R.E., 1970. The Significance Test Controversy. Aldine, Chicago.

Mulaik, S. A., Raju, N.S., Harshman, R.A., 1997. There is a time and a place for significance testing. In: Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), What if there were no significance tests? Erlbaum, Mahwah, NJ, pp. 65–115.

Neyman, J., 1950. First Course in Probability and Statistics. Holt, New York.

Neyman, J., 1957. Inductive behavior as a basic concept of philosophy of science. International Statistical Review 25, 7–22.

Nickerson, R .S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods 5, 241–301.

Nunally, J. C., 1975. Introduction to Statistics for Psychology and Education. McGraw-Hill, New York.

Oakes, M., 1986. Statistical Inference: A Commentary for the Social and Behavioral Sciences. Wiley, NY.

Perlman, M. D., Wu, L., 1999. The emperor's new tests. Statistical Science 14, 355–381.

Pollard, P., Richardson, J.T.E., 1987. On the probability of making Type I errors. Psychological Bulletin 102, 159–163.

Roberts, S., Pashler, H., 2000. How persuasive is a good fit? A comment on theory testing. Psychological Review 107, 358–367.

Rosenthal, R., Rubin, D.R., 1982. Comparing effect sizes of independent studies. Psychological Bulletin 92, 500–504.

Rucci, A. J., Tweney, R. D., 1980. Analysis of variance and the "second discipline" of scientific psychology: a historical account. Psychological Bulletin 87, 166–184.

Sedlmeier, P., Gigerenzer, G., 1989. Do studies of statistical power have an effect on the power of studies? Psychological Bulletin 105, 309–316.

Simon, H. A., 1992. What is an "explanation" of behavior? Psychological Science 3, 150–161.

Skinner, B. F., 1972. Cumulative record. Appleton-Century-Crofts, New York.

Skinner, B. F., 1984. A Matter of Consequences. New York University Press, New York.

Stevens, S. S., 1960. The predicament in design and significance. Contemporary Psychology 9, 273–276.

Todd, P. M., Gigerenzer, G., 2000. Pr´ecis of simple heuristics that make us smart. Behavioral and Brain Sciences 23, 727–780.

Tukey, J. W., 1969. Analyzing data: sanctification or detective work? American Psychologist 24, 83–91.

Tversky, A., Kahneman, D., 1971. Belief in the law of small numbers. Psychological Bulletin 76, 105–110.

Waller, N. G., 2004. The fallacy of the null hypothesis in soft psychology. Applied and Preventive Psychology 11, 83–86.

Wilkinson, L., The Task Force on Statistical Inference, 1999. Statistical methods in psychology journals: guidelines and explanations. American Psychologist, 54, 594–604.

Traducción: Francesc J. Hernàndez

(Universitat de València)