anales de psicología, 1992, 8(1-2), 3-17

Cosas que he aprendido (hasta ahora)

Jacob Cohen [1]

New York University

Resumen. Esta es una relación de lo que he aprendido (hasta ahora) acerca de la aplicación de la estadística a la psicología y a otras ciencias sociobiomédicas. Incluye los principios "menos es más" (a menos variables, cuestiones más específicas y más sofisticadas), "lo simple es mejor" (representaciones gráficas, pesos unidad para los compuestos lineales) y "algunas cosas que usted ha aprendido no son así". He aprendido a evitar los numerosos equívocos que rodean al contraste Fisheriano de la hipótesis nula. También he comprendido la importancia que tienen el análisis de la potencia y la determinación del cuán grandes son los efectos que estudiamos (en lugar del cuán estadísticamente significativos son). Por último, he aprendido que no existe un único camino hacia la inducción estadística, que el buen juicio del investigador es el elemento crucial en la interpretación de los datos y que las cosas llevan tiempo.

Abstract. This is an account of what I have learned (so far) about the application of statistics to psychology and the other sociobiomedical sciences. It includes the principles "less is more" (fewer variables, more highly targeted issues, sharp rounding off), "simple is better" (graphic representation, unit weighting for linear composites), and "some things you learn aren't so". I have learned to avoid the many misconceptions that surround Fisherian null hypothesis testing. I have also learned the importance of power analysis and the determination of just how big (rather than how statistically significant) are the effects that we study. Finally, I have learned that there is no royal road to statistical induction, that the informed judgement of the investigator is the crucial element in the interpretation of data, and that things take time.

Lo que he aprendido (hasta ahora) procede de mi trabajo con estudiantes y colegas, de mi experiencia (a veces amarga) con los editores de revistas y comités de redacción, y de los escritos, entre otros, de Paul Meehl, David Bakan, William Rozeboom, Robyn Dawes, Howard Wainer, Robert Rosenthal, y más recientemente, Gerd Gigerenzer, Michael Oakes y Leland Wilkinson. Aunque no siempre citados explícitamente, muchos de ustedes serán capaces de detectar sus huellas en lo que sigue.

Algunas de las cosas que Vd. ha aprendido no son así

Una de las cosas que pronto aprendí fue que algunas de las cosas que se aprenden no son así. En la Facultad, justo después de la IIª Guerra Mundial, aprendí que para las Tesis Doctorales, y para otros muchos propósitos, cuando se comparan grupos, el tamaño muestral más adecuado es de 30 casos por grupo. El número *30* parece haber surgido de la convención de que, con menos de 30 casos estaríamos tratando con muestras "pequeñas" que requieren el manejo especializado de "estadísticos para muestras pequeñas", en lugar del procedimiento de la razón crítica que nos han enseñado. Algunos de nosotros conoció algo acerca de estos exóticos estadísticos para muestras pequeñas -de hecho, uno de mis compañeros de doctorado realizó una Tesis cuya característica peculiar era el tamaño muestral de sólo 20 casos por grupo, de modo que pudiera demostrar su destreza con los

[1] Esta conferencia invitada fue presentada a la Division of Evaluation, Measurement, and Statistics (Division 5) at the 98th Annual Convention of the American Psychological Association en Boston, 13 de agosto, 1990. Agradezco los comentarios sobre el borrador hechos por Patricia Cohen, Judith Rabkin, Raymond Katzell y Donald F. Klein.

La correspondencia referente a este artículo puede dirigirse a Jacob Cohen, Department of Psychology, New York University, 6 Washington Pl., 5th Floor, New York, NY 10003.

Este artículo apareció originalmente en inglés [Jacob Cohen (1990): Things I have learned (so far). *American Psychologist, 45(12),* 1304-1312].

Este artículo ha sido traducido con permiso del editor y del autor por Julio Sánchez Meca, bajo la supervisión del Servicio de Traducciones de la Universidad de Murcia.

 ISSN: 0212-09728.

estadísticos para muestras pequeñas. No fue hasta algunos años después cuando descubrí (adviértase que no digo "inventé") el análisis de potencia, uno de cuyos frutos fue la revelación de que, para una comparación entre dos medias de grupos independientes con n = 30 por grupo al santificado nivel bilateral del .05, la probabilidad de que un tamaño del efecto medio pudiera ser etiquetado como significativo por los métodos más modernos (una prueba t) era tan sólo de .47. Así, obtener un resultado significativo sería aproximadamente como lanzar una moneda al aire, incluso aunque, en realidad, el tamaño del efecto fuera importante. La potencia de mi amigo con n = 20 fue bastante peor (.33), aunque obviamente él no podía saberlo, y alcanzó resultados no significativos con los que procedió a demoler nada menos que una importante parte de la teoría psicoanalítica.

Menos es más

Algo que he tardado muchos años en aprender es la validez del principio general *menos es más*, excepto para el tamaño muestral, por supuesto (Cohen y Cohen, 1983, pp. 169-171). Me he encontrado demasiados estudios con cantidades prodigiosas de variables dependientes, o con lo que me parecieron demasiadas variables independientes, o (el cielo nos ayude) ambas cosas a la vez.

En cualquier investigación que no sea explícitamente exploratoria, deberíamos estudiar pocas variables independientes e incluso menos variables dependientes, por diversas razones.

Si se han de relacionar todas las variables dependientes con todas las variables independientes mediante simples análisis bivariados o por regresión múltiple, el número de contrastes de hipótesis que por fuerza se ejecutarán es al menos el producto de los dos grupos de variables. Utilizar el nivel del .05 para muchos contrastes incrementa la tasa de error Tipo I del experimento –o, dicho de otro modo, incrementa enormemente las posibilidades de descubrir cosas que no son. Si, por ejemplo, usted estudia 6 variables dependientes y 10 independientes y encuentra que su cosecha produce 6 asteriscos, usted sabe muy bien que, aunque no hubieran asociaciones reales en ninguno de los 60 contrastes, la probabilidad de obtener uno o más resultados "significativos" es bastante alta (tanto como $1 - .95^{60}$, que es igual, precisamente, a .95), y que en promedio debe esperar tres resultados espuriamente significativos. A continuación, tendrá que plantearse algunas preguntas embarazosas, tales como, bien ¿cuáles son las tres reales?, o incluso, ¿son las seis significativas *significativamente* más que las tres esperadas por azar? (y resultará que no lo son.)

Y por supuesto, como ya habrá descubierto, no es probable que usted resuelva el problema de los contrastes múltiples con la maniobra de Bonferroni. Al dividir .05 por 60, el criterio de significación por contraste pasa a ser .05/60 = .00083, y en consecuencia, un valor crítico bilateral de t en torno a 3'5. Los efectos con los que está trabajando pueden no ser lo suficientemente grandes como para producir valores t tan altos, a menos que tenga suerte.

Ni tampoco eludirá el problema haciendo seis regresiones múltiples por pasos sobre las 10 variables independientes. La cantidad de capitalización del azar que esto conlleva es más de lo que yo sé calcular, pero desde luego es superior a una simple remesa de asteriscos para los 60 coeficientes de regresión (Wilkinson, 1990, p. 481).

En resumen, los resultados de este enorme estudio son confusos. No existe una solución a su problema. Por supuesto, no presentaría el estudio para su publicación como si las improductivas tres cuartas partes de sus variables no existieran ....

Lo irónico del caso es que la gente que hace estudios como éste a menudo parte de alguna idea central útil que, si fuera perseguida con moderación mediante unas pocas variables e hipótesis puntuales, probablemente producirían resultados significativos. Estos podrían resistir con éxito el reto de Bonferroni o de cualquier otro procedimiento de ajuste del nivel alfa, si el buen criterio o nuestra formación así lo aconsejan.

En el análisis de regresión-correlación con numerosas variables independientes surge un caso especial del problema del exceso de variables. Conforme el número de variables independientes se incrementa, también lo hará su redundancia respecto de la relevancia sobre el criterio. Dado

que la redundancia infla los errores típicos de los coeficientes de regresión y de correlación parcial, reduciendo de este modo su significación estadística, es probable que los resultados sean inservibles.

He hecho tanto hincapié en la conveniencia de trabajar con pocas variables y tamaños muestrales grandes que algunos de mis estudiantes han difundido el rumor de que mi idea del estudio perfecto es el de 10.000 casos y ninguna variable. Exageran.

Una aplicación menos profunda del principio "menos es más" se refiere a los hábitos de presentación de los resultados numéricos. Existen programas de computador que por defecto dan los resultados numéricos con cuatro, cinco e incluso más decimales. Sus autores tienen que ser disculpados ya que, que sepa el programador, pueden ser utilizados hasta por científicos atómicos. Pero nosostros, los científicos sociales, deberíamos hacer algo más que presentar nuestros resultados con tantos decimales. ¿Qué significa, pues, un r = .12345? ¿o, para una distribución del CI, una media de 105'6345? Con N = 100, el error típico de r está en torno a .1 y el error típico de la media del CI es aproximadamente 1'5. Así, la parte *345* de r = .12345 es sólo el 3% de su error típico, y la parte *345* de una media en CI de 105'6345 es tan sólo el 2% de su error típico. Estos decimales superfluos no son mejores que los números aleatorios. Realmente son peor que inútiles debido a que la confusión que crean, especialmente en las tablas, sirve para distraer el ojo y la mente de las comparaciones necesarias entre los otros dígitos relevantes. En efecto, aquí menos es más.

Lo simple es mejor

También he aprendido que lo simple es mejor, lo cual es una especie de generalización del "menos es más". La idea "lo simple es mejor" es ampliamente aplicable a la representación, análisis e informe de datos.

Si, como el viejo tópico afirma, una imagen vale más que mil palabras, generalmente a la hora de describir una distribución valdrá más un polígono de frecuencias o, mejor aún, un diagrama en tallo y hojas de Tukey (1977, pp. 1–26) que los cuatro primeros momentos, es decir, la media, la desviación típica, la asimetría y la curtosis. Yo no pongo en duda que los momentos no resuman eficientemente una distribución ni que sean útiles en algunos contextos analíticos. Los paquetes estadísticos nos los dan alegremente y nosotros los publicamos obedientemente, pero por regla general no permiten ver la distribución a la mayoría de nosotros ni a la de los consumidores de nuestros productos. No nos dicen, por ejemplo, que no hay casos entre las puntuaciones 72 y 90, ni que la puntuación 24 está en algún lugar del lado izquierdo, ni que hay una acumulación de valores 9. Estos son los tipos de características de nuestros datos que seguramente queremos saber, y éstos se hacen inmediatamente evidentes con una simple representación gráfica.

Las representaciones gráficas son incluso más importantes en el caso de datos bivariados. Subyaciendo a cada coeficiente de correlación producto-momento en una gran cantidad de tales coeficientes descansa un simple diagrama de dispersión que r presume resumir, y bien pudiera hacerlo. Es decir, lo hace si la distribución conjunta es más o menos normal bivariada -lo que significa, entre otras cosas, que la relación debe ser lineal y que no existen puntos muy extremos. Sabemos que las medidas mínimo-cuadráticas, tales como la media y la desviación típica, son sensibles a los 'outliers'. Pues bien, las correlaciones de Pearson lo son más aún. Hace unos 15 años, Wainer y Thissen (1976) publicaron un conjunto de datos extraídos de la altura en pulgadas y el peso en libras de 25 sujetos, para los cuales r alcanzó un valor perfectamente razonable de .83. Pero si se cometía un error de transcripción, de modo que la altura y el peso de uno de los 25 sujetos se cambiaba, el valor de r se convertía en −.26, ¡un tremendo y costoso error!

Difícilmente exista alguna excusa al hecho de que nos pasen desapercibidos huecos, 'outliers', curvilinealidad u otra patología existente en nuestros datos. Los mismos paquetes estadísticos de computador con los que podemos hacer análisis tan complicados como una estimación no lineal cuasi-Newton o un escalamiento multidimensional con el coeficiente de alienación de Guttman también pueden proporcionarnos simples gráficos de dispersión y diagramas en tallo y hojas con los que

podemos ver nuestros datos. El análisis de regresión/correlación múltiple más apropiado no comienza con una matriz de coeficientes de correlación, medias y desviaciones típicas, sino con un conjunto de diagramas en tallo y hojas y gráficos de dispersión. A veces aprendemos más de lo que vemos que de lo que computamos; a veces lo que aprendemos de lo que vemos es que no deberíamos computar, al menos no sobre los datos tal y como se nos presentan.

Los computadores son una bendición, pero otra de las cosas que he aprendido es que no son una panacea. Hace cuarenta años, Antes de los Computadores (esto es, A.C.), hice para mi Tesis Doctoral tres análisis factoriales sobre los 11 subtests del Wechsler-Bellevue, con muestras de 100 pacientes psiconeuróticos, esquizofrénicos y lesionados cerebrales. Trabajando con un bloc y un lápiz, papel milimetrado, una tabla de productos de números con dos dígitos y una calculadora electromecánica de escritorio marca Friden que hacía raíces cuadradas "automáticamente", el proceso completo me llevó buena parte de un año. Hoy día, con un ordenador personal, el trabajo está hecho virtualmente en microsegundos (o al menos rápidamente). Pero otra diferencia importante entre entonces y ahora es que la absoluta laboriosidad de la tarea aseguraba que a todo lo largo del proceso yo estaba en estrecho contacto con los datos y su análisis. No había posibilidad alguna de que salieran cosas raras en mis datos ni resultados intermedios que no conociera, cosas que podrían viciar mis conclusiones.

Reconozco que esto puede sonarles a retrógado, pero no me malinterpreten –me entusiasman los computadores y me deleito con la facilidad con que se logra el análisis de datos con un buen paquete estadístico interactivo como SYSTAT y SYGRAPH (Wilkinson, 1990). No obstante, me aterra el hecho de que algunos editores de paquetes estadísticos vendan con éxito sus productos con el reclamo de que no se necesita saber estadística para utilizarlos. Lo que sí es cierto es que el mismo paquete que hace posible que un ignorante haga un análisis factorial con un menú descendente y el *click* de un ratón, también puede facilitar considerablemente la ejecución de análisis simples e informativos con una velocidad y eficiencia pasmosas.

Un excelente ejemplo del principio "lo simple es mejor" se encuentra en la composición de valores. Hemos aprendido y enseñado a nuestros estudiantes que para predecir un criterio a partir de un conjunto de variables predictoras, asumiendo por simplicidad (y como los matemáticos dicen, "sin pérdida de generalidad") que todas las variables están tipificadas, logramos la máxima predicción lineal haciendo un análisis de regresión múltiple y formando un compuesto mediante la ponderación de las puntuaciones z predictoras por sus pesos beta. Puede demostrarse matemáticamente que con estas betas como pesos, el compuesto resultante genera en la muestra una correlación con el criterio superior a la de cualquier otro compuesto lineal formado con otros pesos.

Sin embargo en la práctica, la mayoría de las veces es mejor utilizar los pesos unidad: $+1$ para los predictores relacionados positivamente, -1 para los relacionados negativamente y 0, es decir, para desechar los predictores pobremente relacionados (Dawes, 1979; Wainer, 1976). La trampa está en que las betas tienen garantías de ser mejores que los pesos unidad sólo para la muestra sobre la cual fueron determinadas. (Es como una TV cuyo funcionamiento solo estuviera garantizado en la tienda.) Pero el investigador no está interesado en hacer predicciones sobre esa muestra: él ya *conoce* los valores criterio para esos casos. Se trata de combinar los predictores de cara a una máxima predicción con muestras *futuras*. La explicación de la dudosa idoneidad de las betas con muestras futuras está en que posiblemente tienen errores típicos grandes. Para la situación típica de 100 ó 200 casos y 5 ó 10 predictores correlacionados, los pesos unidad funcionarán igual o mejor.

Permítaseme ofrecer un ejemplo concreto para clarificar este punto. Un ejemplo corriente en nuestro texto de regresión (Cohen y Cohen, 1983) presenta los salarios de una muestra del profesorado universitario estimados a partir de cuatro variables independientes: Años desde la lectura de la Tesis Doctoral, sexo (codificado a la nueva moda: 1 para mujer y 0 para hombre), número de publicaciones y número de citas. La correlación múltiple muestral alcanza el valor .70. Lo que pretendemos estimar es la correlación que obtendríamos si utilizáramos los pesos beta muestrales en la población, la correlación múltiple por validación cruzada, la cual desafortunadamente se reduce a un valor más

pequeño que la correlación múltiple restringida. Con N = 100 casos, utilizando la fórmula de Rozeboom (1978), se queda en .67. No está mal. Pero los pesos unidad lo hacen mejor: .69. Con 300 ó 400 casos, el incremento de la estabilidad muestral hace subir la correlación por validación cruzada, pero permanece ligeramente inferior al valor .69 de los pesos unidad. Aumentando el tamaño muestral a 500 ó 600, la correlación por validación cruzada alcanza en este ejemplo un punto en el que es mayor que el valor .69 de los pesos unidad, pero a duras penas, ¡por unos pocos puntos en el *tercer* decimal! Cuando el tamaño muestral es sólo de 50, la correlación múltiple por validación cruzada es sólo de .63, mientras que la correlación con los pesos unidad se mantiene en .69. El tamaño muestral no afecta a la correlación con los pesos unidad debido a que no estimamos coeficientes de regresión inestables. Eso sí, está sujeta a error de muestreo, pero también lo está la correlación múltiple por validación cruzada.

Ahora bien, los pesos unidad no siempre son tan buenos o mejores que los pesos beta. Con algunos patrones de correlación relativamente raros (la supresión es uno de ellos), cuando las betas varían considerablemente respecto de su media, o cuando la ratio del tamaño muestral con el número de predictores alcanza 30 a 1 y la correlación múltiple llega a .75, los pesos beta pueden ser mejores, aunque incluso en estas raras circunstancias, no serán mucho mejores.

Más aún, los pesos unidad funcionan bien fuera del contexto de la regresión múltiple donde tenemos datos en el criterio –es decir, una situación en la que pretendemos medir algún concepto combinando indicadores, o algún factor abstracto generado por un análisis factorial. Es probable que los pesos unidad sobre las puntuaciones tipificadas sean mejores para nuestros propósitos que las puntuaciones factoriales generadas por el programa de computador, las cuales, después de todo, son el fruto de un análisis de regresión para esa muestra de las variables sobre el factor como criterio.

Tenga en cuenta que cuando vamos a predecir el nivel medio de los estudiantes universitarios de primer año a partir de un test de 30 items, no hacemos análisis de regresión para obtener los pesos "óptimos" con los que combinar las puntuaciones de los items: simplemente las sumamos, como hizo Galton. Lo simple *es* mejor.

Sin embargo, no estamos aplicando el principio "lo simple es mejor" cuando "simplificamos" una variable graduada con múltiples valores (como el CI, o el número de hijos, o la gravedad de un síntoma) cortándola por cualquier punto y convirtiéndola en una dicotomía. Esto se hace a veces en prueba de modestia acerca de la calidad o precisión de la variable, o para "simplificar" el análisis. Esto no es una aplicación, sino más bien una perversión de "lo simple es mejor" debido a que esta práctica elimina información deliberadamente. Se ha demostrado que cuando se mutila así una variable, generalmente se reduce su correlación cuadrática con otras variables en torno a un 36% (Cohen, 1983). No lo haga. Este tipo de simplificación es similar a la práctica de "simplificar" el ANOVA de un diseño factorial mediante la reducción de todos los tamaños de celdilla al tamaño más pequeño eliminando casos. Son dos modos de tirar lo más valioso que tenemos: La información.

Desde un punto de vista más general, creo que he comenzado a aprender cómo utilizar la estadística en las ciencias sociales.

El ambiente que caracteriza a la estadística tal y como se aplica en las ciencias sociales y biomédicas es la de una religión secular (Salsburg, 1985), aparentemente de origen Judeo-Cristiano, ya que emplea como imagen más poderosa una cruz de seis puntas, a menudo presentada repetidamente para reforzar su autoridad. Confieso que soy un agnóstico.

El legado Fisheriano

Cuando comencé a estudiar inferencia estadística, me llevé una sorpresa compartida por muchos neófitos. Encontré que si, por ejemplo, quería saber si los niños pobres estiman el tamaño de las monedas mayor de lo que lo hacen los niños ricos, una vez extraídos los datos no podía probar esta hipótesis de investigación, sino más bien la hipótesis nula de que no había diferencia en la percepción del tamaño que los niños pobres y ricos tenían de un mismo conjunto de monedas. Esto me pareció

un tanto extraño y contraproducente, pero rápidamente me integré (o, si lo prefiere, me convertí, o quizá me lavaron el cerebro) en la creencia Fisheriana de que la ciencia avanza sólo a través de inferencia inductiva y que la inferencia inductiva se logra principalmente rechazando hipótesis nulas, usualmente al nivel del .05. (No fue hasta mucho después cuando aprendí que el filósofo de la ciencia, Karl Popper, 1959, defendía la formulación de hipótesis de *investigación* falsables y diseños de investigación que *las* puedan falsar.)

El que las ideas de Fisher se convirtieran rápidamente en *la* base de la inferencia estadística en las ciencias del comportamiento no es sorprendente: eran muy atractivas. Ofrecían un esquema determinista, mecánico y objetivo, independiente del contenido y dirigido a claras decisiones si–no. Durante años, educado por los libros de texto de estadística psicológica de los años 1940 y 1950, jamás imaginé que hubieran sido fuente de agudas controversias (Gigerenzer y Murray, 1987).

Tomemos, por ejemplo, la típica decisión si–no. Resultaba bastante apropiada en Agronomía, que era de donde Fisher procedía. El resultado de un experimento puede perfectamente ser la decisión de utilizar ésta en lugar de aquélla cantidad de abono, o plantar ésta o aquélla variedad de trigo. Pero nosotros no tratamos con abonos, al menos conscientemente. De igual modo, en otras tecnologías -por ejemplo, el control de calidad en ingeniería o la educación- la investigación está diseñada frecuentemente para producir decisiones. Sin embargo, las cosas no están tan claramente orientadas hacia la decisión en el desarrollo de las teorías científicas.

Consideremos a continuación el santificado (y santificante) nivel mágico del .05. Esta base para la decisión ha jugado un notable papel en las ciencias sociales y en las vidas de los científicos sociales. Al regir las decisiones acerca del estatus de las hipótesis nulas, vino a determinar decisiones sobre la aceptación de Tesis Doctorales y la concesión de becas de los fondos de investigación, sobre la publicación y la promoción, y sobre si tener o no un niño. Su arbitraria e irrazonable tiranía ha conducido a amañar datos con diversos grados de ingenio, desde alterarlos groseramente, hasta eliminar casos donde "deben haber habido" errores.

La hipótesis nula nos contrasta

No podemos cargar a R.A. Fisher con todos los pecados del último medio siglo que se han cometido en su nombre (o más a menudo anónimamente, aunque como parte de su legado), pero merecen ser catalogados (Gigerenzer y Murray, 1987; Oakes, 1986). Con el paso de los años, he aprendido a no cometer los siguientes errores:

Cuando una hipótesis nula Fisheriana es rechazada con una probabilidad asociada de, por ejemplo, .026, ello *no* significa que la probabilidad de que la hipótesis nula sea verdadera es .026 (o menor que .05, o que cualquier otro valor que podamos especificar). Dado nuestro enfoque de la probabilidad como el límite de la frecuencia relativa –por mucho que queramos que sea de otra forma–, este resultado no nos dice nada acerca de la veracidad de la hipótesis nula una vez dados los datos. (Para esto tenemos que ir a la estadística Bayesiana o por verosimilitud, en las que la probabilidad no es la frecuencia relativa, sino el grado de creencia.) Lo que nos da es la probabilidad de los datos, supuesto que la hipótesis nula es verdadera -lo cual no es lo mismo, aunque pueda sonar igual.

Si el valor p con el que rechazamos la hipótesis nula Fisheriana no nos dice nada acerca de la probabilidad de que la hipótesis nula sea verdadera, ciertamente tampoco nos puede decir nada acerca de la probabilidad de que la hipótesis alternativa o de la investigación lo sea. De hecho, no existe hipótesis alternativa en el esquema de Fisher: En efecto, él se opuso violentamente a su inclusión por Neyman y Pearson.

A pesar de los extendidos equívocos, el rechazo de una determinada hipótesis nula no nos aporta ninguna base para estimar la probabilidad de que una réplica de la investigación de nuevo dé lugar a un rechazo de esa hipótesis nula.

Por supuesto, todos sabemos que no poder rechazar la hipótesis nula Fisheriana no garantiza la conclusión de que ésta sea verdadera. Ciertamente Fisher lo sabía y lo enfatizó, y nuestros libros

de texto así nos lo enseñaron. Sin embargo, ¿cuán a menudo leemos en la discusión y en las conclusiones de los artículos que aparecen actualmente en nuestras revistas más prestigiosas que "no hay diferencias" o que "no hay relación"? (Y esto es 40 años después de que mi amigo con $N = 20$ utilizara un resultado no significativo para demoler la teoría psicoanalítica.)

La otra cara de esta moneda es la interpretación que acompaña a los resultados que superan la barrera del .05 y alcanzan el estado de gracia de la "significación estadística". "Todos" sabemos que lo único que esto significa es que el efecto no es nulo, y nada más. Sin embargo, cuán a menudo nos encontramos con que tal resultado es tomado para significar, al menos implícitamente, que el efecto es *significativo*, es decir, *importante*, *grande*. Si un resultado es *altamente* significativo, digamos $p < .001$, la tentación de cometer este error de interpretación se hace poco menos que irresistible.

Echemos un atento vistazo a esta hipótesis nula –el centro del esquema Fisheriano– que tan decididamente pretendemos negar. Una hipótesis nula es una afirmación precisa acerca de algún estado de la cuestión en una población, generalmente el valor de un parámetro y frecuentemente con valor cero. Se denomina hipótesis "nula" porque la estrategia es invalidarla o porque significa que "nada ha ocurrido". Así, "La diferencia entre las puntuaciones medias de los hombres y mujeres de U.S.A. sobre una escala de Actitud Hacia las U.N. es cero" es una hipótesis nula. "La correlación r producto-momento entre la altura y el CI en los estudiantes de escuela superior es cero" es otra. "La proporción de hombres en una población de adultos disléxicos es .50" es otra más. Cada una de ellas es una afirmación precisa –por ejemplo, si el r de la población entre la altura y el CI es de hecho .03, la hipótesis nula de que es cero es falsa. ¡También es falsa si r es .01, .001 ó .000001!

Un poco de reflexión revela un hecho ampliamente reconocido por los estadísticos: La hipótesis nula, en sentido literal (y éste es el único modo en que puede tomarse en el contraste de hipótesis formal), *siempre* es falsa en el mundo real. Sólo puede ser verdadera en las entrañas del procesador de un computador que realice un estudio Monte Carlo (y hasta ahí un electrón extraviado puede hacerla falsa). Si es falsa, incluso en grado minúsculo, es posible que una muestra suficientemente grande produzca un resultado significativo y conduzca a su rechazo. Así pues, si la hipótesis nula siempre es falsa, ¿qué conseguimos rechazándola?

Otro problema que me preocupaba ha sido el de la asimetría del esquema Fisheriano: Si su contraste excedía un valor crítico, se podía concluir, con riesgo alfa, que la hipótesis nula era falsa, pero si quedaba por debajo de ese valor crítico, *no podíamos* concluir que la hipótesis nula era verdadera. De hecho, lo único que podíamos concluir es que no podíamos concluir que la hipótesis nula era falsa. Con otras palabras, difícilmente podíamos concluir algo.

Otro problema fue que si la hipótesis nula era falsa, tendría que serlo en algún grado. Habría que distinguir si la diferencia entre las medias de población era de 5 ó de 50, o si la correlación de la población era .10 ó .30, y esto no ha sido tenido en cuenta en el método predominante. Había tropezado con algo que algún tiempo después supe que había sido una de las bases de la crítica de Neyman-Pearson al sistema de Fisher de la inducción estadística.

En 1928 (cuando yo estaba en parvulario), Jerzy Neyman y el hijo de Karl Pearson, Egon, comenzaron a publicar trabajos que ofrecían una perspectiva bastante diferente de la inferencia estadística (Neyman y Pearson, 1928a, 1928b). Entre otras cosas, argüían que en lugar de tener una sola hipótesis que era rechazada o no, las cosas podrían organizarse de modo que pudiera elegirse entre dos hipótesis, una de las cuales podría ser la hipótesis nula y la otra una hipótesis alternativa. Se podría asignar a la hipótesis nula precisamente definida un riesgo alfa, y a la hipótesis alternativa definida con la misma precisión un riesgo beta. El rechazo de la hipótesis nula cuando era verdadera sería un error de primera especie, controlado por el criterio alfa, pero el no poder rechazarla cuando la hipótesis alternativa fuera verdadera también era un error, un error de segunda especie, el cual sería controlado de forma que ocurriera a una tasa beta. Así, dada la magnitud de la diferencia entre las hipótesis nula y alternativa (es decir, dado el tamaño del efecto de la población hipotética), y fijando valores para alfa y beta, se podría determinar el tamaño muestral necesario para que se

cumplan estas condiciones. O, una vez fijados el tamaño del efecto, alfa y el tamaño muestral, se podría determinar beta, o su complemento, la probabilidad de rechazar la hipótesis nula, la potencia del contraste.

Ahora bien, R.A. Fisher fue sin duda el mayor estadístico de este siglo, merecidamente apodado "el padre de la moderna estadística", pero tenía un punto débil. Fue un testarudo y frecuentemente un temible oponente intelectual. Una disputa con Karl Pearson había dejado los trabajos de Fisher fuera de *Biometrika*, de la que Karl Pearson era su editor. Después de que el viejo Pearson se retirara, los esfuerzos de Egon Pearson y de Neyman por evitar la lucha con Fisher fueron inútiles. Fisher escribió que eran como los rusos, que pensaban que la "ciencia pura" debería estar "integrada en el funcionamiento tecnológico" como "en un plan de cinco años". Una vez más comenzó la discusión al comentar sobre un trabajo de Neyman en la *Royal Statistical Society* que Neyman debería haber elegido un tópico "sobre el que pudiera hablar con autoridad" (Gigerenzer y Murray, 1987, p. 17). Fisher condenó ferozmente la herejía de Neyman–Pearson.

Yo, por supuesto, no era consciente de nada de esto. Los textos de estadística con los que me formé y sus ediciones posteriores a las que repetidamente volví en los años 1950 y 1960 presentaban el contraste de la hipótesis nula 'à la Fisher' como un hecho dado, como *el modo* de hacer inferencia estadística. Las ideas de Neyman y Pearson apenas fueron mencionadas, si es que lo fueron, o desechadas por ser demasiado complicadas.

Cuando finalmente tropecé con el análisis de la potencia, y me las arreglé para superar el hándicap de una base en matemáticas no más allá del álgebra de la escuela superior (y no hablemos de estadística matemática), fue como si hubiera muerto e ido al paraíso. Después de aprender lo que eran las distribuciones no centrales y comprender que era importante descomponer los parámetros de no centralidad en sus constituyentes del tamaño del efecto y del tamaño muestral, me di cuenta de que disponía de un marco para el contraste de hipótesis con cuatro parámetros: El criterio de significación alfa, el tamaño muestral, el tamaño del efecto de la población y la potencia del contraste. En cualquier prueba estadística, cualquiera de éstos estaba en función de los otros tres. Esto significa, por ejemplo, que para una prueba de significación de una correlación producto-momento, utilizando un criterio alfa bilateral de .05 y un tamaño muestral de 50 casos, si la correlación de la población es .30, mi probabilidad a largo plazo de rechazar la hipótesis nula y encontrar una correlación muestral significativa era de .57, el lanzamiento de una moneda al aire. Otro ejemplo, para el mismo $\alpha = .05$ y valor de población $r = .30$, si quisiera tener una potencia de .80, podría determinar que el tamaño muestral necesario sería de 85.

Jugando con este nuevo juguete (y con una pequeña beca del *National Institute of Mental Health*) hice lo que vino a llamarse un meta-análisis de los artículos del volumen de 1960 del *Journal of Abnormal and Social Psychology* (Cohen, 1962). Encontré, entre otras cosas, que utilizando el criterio del .05 bilateral, la potencia mediana para detectar un efecto medio era de .46 -un pésimo resultado. Por supuesto, los investigadores no podían conocer la escasa potencia de su investigación, ya que su formación no les había preparado para saber nada acerca de la potencia, y menos cómo utilizarla en la planificiación de una investigación. Uno podía pensar que después de 1969, fecha en que publiqué mi primer tratado sobre la potencia que lo hacía tan fácil como buscar un logaritmo, los conceptos y los métodos del análisis de la potencia habrían conquistado los corazones de los contrastadores de hipótesis nulas. Pues sí, uno podía pensarlo.

Una de las ventajas menos claras del análisis de la potencia fue que hizo posible "probar" hipótesis nulas. Por supuesto, como ya he comentado, todos sabemos que no se pueden probar hipótesis nulas. Pero cuando un investigador quiere probar una hipótesis nula, el asunto no está en demostrar que el tamaño del efecto de la población es, digamos, cero con un millón de decimales o más, sino más bien en probar que es menos que insignificante, trivial (Cohen, 1988, pp. 16-17). Así, desde un análisis de la potencia con, digamos, $\alpha = .05$, con potencia fijada a, digamos, .95, de modo que también $\beta = .05$, puede determinarse el tamaño muestral necesario para detectar este efecto

insignificante con probabilidad .95. Ahora bien, si se lleva a cabo la investigación utilizando ese tamaño muestral y el resultado es *no* significativo, como había una probabilidad .95 de detectar este efecto insignificante y el efecto *no* fue detectado, quedaría justificada la conclusión de que existe un efecto trivial, al nivel $\beta = .05$. Esto, de hecho, prueba probabilísticamente la hipótesis nula deseada de un efecto insignificante. El razonamiento es impecable, pero cuando uno va a aplicarlo, descubre que requiere tamaños muestrales enormes. Por ejemplo, si adoptamos los parámetros de arriba para un contraste de significación de un coeficiente de correlación e interpretamos $r = .10$ como un tamaño del efecto insignificante, vemos que se requiere una muestra de casi 1.300 casos. Demandas de potencia más modestas, pero todavía razonables, requieren tamaños muestrales más pequeños, pero no lo suficientemente bajos como para interesar a la mayoría de los investigadores: incluso una potencia de .80 para detectar una correlación en la población de .10 requiere casi 800 casos. Así pues, generalmente se necesitan tamaños muestrales excesivamente grandes para probar hipótesis nulas tal y como las he definido; no obstante, el procedimiento deja claro qué se puede decir cuando no se consigue rechazar la hipótesis nula de que el efecto es trivial.

Un resultado beneficioso del análisis de la potencia es que nos arrastra forzosamente a considerar la magnitud de los efectos. En psicología, y especialmente en la psicología blanda, bajo el dominio del esquema Fisheriano, ha habido poca conciencia de cuán grandes son las cosas. Los tan populares diseños de ANOVA producen razones F, y son sus tamaños lo que interesa. Lo primero es la cuestión de si alcanzan el santificante corte del .05 y son así significativos, y a continuación cuán lejos se sitúan detrás de ese punto: ¿Tal vez fueron *muy significativos* (p menor que .01) o *muy altamente significativos* (menor que .001)? Dado que la ciencia trata inevitablemente con magnitudes no es sorprendente la frecuencia con que los valores p son tratados como sustitutos de los tamaños del efecto.

Una de las cosas que muy pronto me atrajeron hacia el análisis de correlación fue que producía un valor r, una medida del tamaño del efecto que luego era transformada a una t o a una F para valorar su significación, mientras que el análisis de varianza o de covarianza sólo producían una razón F y no me decían nada acerca del tamaño del efecto. Como muchas de las variables con las que *trabájabamos* estaban expresadas en unidades arbitrarias (puntos sobre una escala, número de ensayos para aprender un laberinto), y el esquema Fisheriano parecía tan completo, no exigiéndonos que pensáramos en los tamaños del efecto, nosotros simplemente no teníamos un lenguaje con el que tratarlo.

Mirando hacia atrás, me resulta bastante comprensible y al mismo tiempo ridículo intentar desarrollar teorías del comportamiento humano con valores p del contraste de hipótesis Fisheriano y con un sentido primitivo del tamaño del efecto. Y me gustaría poder estar hablando de algo ya muy lejano en el tiempo. En 1986, apareció en el *New York Times* un informe del UPI bajo el título "La Altura de los Niños Asociada a las Puntuaciones de los Tests". El artículo describía un estudio que implicó a casi 14.000 niños de 6 a 17 años de edad que presentaba una asociación *definitiva* entre la altura (ajustada por edad y sexo) y las puntuaciones en los tests de inteligencia y de rendimiento. La relación fue descrita como significativa y persistente, incluso después de controlar otros factores tales como el estatus socio-económico, el orden de nacimiento, el tamaño de la familia y la madurez física. Los autores señalaron que el efecto fue pequeño, pero *significativo*, y que no justificaba dar a los niños hormonas del desarrollo para hacerlos más altos y, por lo tanto más inteligentes. Ellos especularon que el efecto podía deberse al hecho de tratar a los niños más bajos como menos maduros, aunque existían explicaciones biológicas alternativas.

Ahora bien, ésto era un relato de un periódico, el fruto de la mente siempre curiosa de un reportero de ciencia, no un artículo de revista; quizás así sea comprensible que no hubiera ningún esfuerzo por ocuparse del tamaño real de este pequeño efecto. Pero busquemos cuán pequeña podría ser esta relación significativa. Bien, si tomamos como significativo un valor $p < .001$ (en beneficio de una mentalidad científica tenaz), resulta que una correlación de .0278 es significativa para 14.000 casos. Pero entiendo que cuando tratamos con variables expresadas en unidades cuya magnitud conocemos,

el tamaño del efecto en relaciones lineales queda mejor comprendido con coeficientes de regresión que con los de correlación. De este modo, aceptando el modelo causal implícito de los autores, obtenemos que para elevar el CI de un niño de 100 a 130 sería necesario dar al niño suficientes hormonas del desarrollo como para que incrementara su altura unos 4 metros (más o menos). Si la causalidad es en el otro sentido y pretendemos crear jugadores de baloncesto, un incremento en la altura de 10 centímetros requeriría elevar el CI en torno a los 900 puntos. Pues bien, ellos dicen que el efecto fue pequeño. (Cuando posteriormente chequeé el artículo de la revista que describía esta investigación, comprobé que la correlación fue mucho mayor que .0278. Fue realmente de .11, de modo que para un incremento de 30 puntos en el CI sería necesario tomar sólo las hormonas del desarrollo suficientes para producir un incremento de un metro de altura, o invirtiendo la causalidad, un incremento de 10 centímetros en la altura requeriría un aumento de sólo 233 puntos de CI.)

Me satisface decir que la persistente desatención hacia el tamaño del efecto parece estar llegando a su fin. El torpe y sobre todo inválido método del recuento en las revisiones de literatura, basado en los valores *p* está siendo reemplazado por el meta-análisis basado en los tamaños del efecto, formulado por Gene Glass (1977). La medida del tamaño del efecto más utilizada es la diferencia media tipificada, *d*, del análisis de la potencia. Ya se han publicado varios tratamientos profundos del meta-análisis[2] y están apareciendo numerosas aplicaciones en diversos campos de la psicología en el *Psychological Bulletin* y en otras prestigiosas publicaciones. En un meta-análisis típico, se revisa la literatura de investigación sobre algún tópico y se obtienen los tamaños del efecto encontrados en los estudios relevantes. Téngase en cuenta que la unidad observacional es el estudio. Estos datos no sólo proporcionan una estimación del nivel y variablidad del tamaño del efecto en un dominio basado en múltiples estudios y, por tanto, en muchas observaciones, sino que al relacionar el tamaño del efecto con diversas características sustantivas y metodológicas de los estudios, puede aprenderse mucho acerca del tópico bajo investigación y cómo investigarlo mejor. Espero que esta revolución persuada a los investigadores de que expliciten los tamaños del efecto y reduzcan así el peso de los meta-analistas y de otros que tienen que hacer suposiciones para extraerlos de los resultados de investigaciones presentados deficientemente. En un campo tan disperso (por no decir anárquico) como el nuestro, el meta-análisis constituye una grata herramienta en pro del avance del conocimiento. El meta-análisis me satisface mucho.

A pesar de mi ininterrumpida identificación con la inferencia estadística, creo, junto con eruditos tales como Meehl (1978), Tukey (1977) y Gigerenzer (Gigerenzer y Murray, 1987), que el contraste de hipótesis ha sido enfatizado en exceso en la psicología y en otras disciplinas que lo utilizan. Ha desviado nuestra atención de las cuestiones cruciales. Hipnotizados por un ritual simple, de propósitos generales, mecanizado y 'objetivo', en el que convertimos los números en otros números y alcanzamos una respuesta si-no, hemos descuidado el examen detenido de la procedencia de los números. Recuerde que en su deliciosa parábola acerca de promediar los números de las camisetas de los futbolistas, Lord (1953) señalaba que "los números no saben de dónde vienen". Pero *nosotros* sí que debemos saber de dónde vienen y deberíamos estar más interesados en el por qué, el qué y lo bien que estamos midiendo, manipulando condiciones y seleccionando nuestras muestras.

También hemos perdido de vista el hecho de que la varianza de error en nuestras observaciones debería retarnos a reducirla en lugar de arrinconarla sin más en el denominador de una prueba *F* ó *t*.

Cómo utilizar la estadística

Así pues, ¿cómo debería yo utilizar la estadística en la investigación psicológica? En primer lugar, descriptivamente. El *Exploratory Data Analysis* de John Tukey (1977) es un inspirado modelo de

[2]El lector puede consultar recientes publicaciones sobre meta-análisis (Cooper, 1989; Eddy, Hasselblad y Shachter, 1992; Hunter y Schmidt, 1990; Rosenthal, 1991; Wachter y Straf, 1990). En castellano pueden consultarse Gómez Benito (1987) y Sánchez Meca y Ato (1989). *[N. del T.]*

cómo realizar análisis gráficos y numéricos de los datos con objeto de comprenderlos[3]. Estas técnicas, aunque complicadas en su concepción, son simples de aplicar y no requieren más que papel y lápiz (Tukey dice que si usted tiene una calculadora de escritorio, mejor). Aunque reconoce la importancia de lo que él denomina confirmación (inferencia estadística), se las ingenia para llenar 700 páginas con técnicas de "mera" descripción, comentando en el prefacio que el énfasis sobre la inferencia en la moderna estadística ha provocado una pérdida de flexibilidad en el análisis de datos.

A continuación, entiendo que la planificación de la investigación se refiere a *cómo proyectarla*. Esto implica hacer juicios tentativos, entre otras muchas cosas, sobre el tamaño del efecto o efectos de la población que usted persigue, el nivel de riesgo alfa que quiere adoptar (convenientemente, aunque no necesariamente del .05) y la potencia que usted desea (por regla general, un valor relativamente elevado semejante a .80). Una vez especificado todo esto, determinar el tamaño muestral necesario es un problema sencillo. A continuación, será una buena idea repensar estas especificaciones. Si, como a menudo suele ocurrir, el tamaño muestral está más allá de sus recursos, considere la posibilidad de reducir sus demandas de potencia, o quizá el tamaño del efecto, o tal vez (el cielo nos asista) incrementar su nivel alfa. O bien, la muestra requerida puede ser más pequeña que lo que usted puede manejar fácilmente, lo cual también debería llevarle a repensar y posiblemente revisar sus especificaciones iniciales. Este proceso finaliza cuando usted alcanza un conjunto de especificaciones viables, o cuando descubre que no es posible ningún conjunto practicable debiendo ser abandonada la investigación tal y como fue concebida originalmente. Aunque difícilmente lo encontrará leyendo la literatura actual, el no someter su plan de investigación al análisis de la potencia es algo sencillamente irracional.

También he aprendido y enseñado que el producto primario de una investigación es una o más medidas del tamaño del efecto y no valores *p* (Cohen, 1965)[4]. Las medidas del tamaño del efecto incluyen las diferencias medias (en bruto o estandarizadas), las correlaciones y la correlación cuadrática de todo tipo, los "odds ratio", las kappas –y cualquier otra que transmita la magnitud del fenómeno de interés apropiada al contexto de la investigación. Si, por ejemplo, usted está comparando grupos sobre una variable medida en unidades bien conocidas por sus lectores (unidades de CI, dólares, número de niños, meses de supervivencia), las diferencias medias son medidas excelentes del tamaño del efecto. Cuando éste no sea el caso, lo que es más corriente, los resultados pueden ser transformados en diferencias medias tipificadas (valores *d*) o en alguna medida de correlación o asociación (Cohen, 1988). (Pero no tal y como solemos interpretar un determinado nivel de correlación [Oakes, 1986, pp. 88-92]. Está demostrado que los psicólogos generalmente sobreestiman la magnitud de la relación que representa una correlación, considerando una correlación de .50 no como su proporción de varianza explicada, como su cuadrado, .25, representa, sino más bien como su raíz cúbica de .80, ¡la cual representa sólo un espejismo! Pero eso es otra historia.)

Así pues, una vez encontrado el tamaño del efecto muestral, usted puede asignarle un valor *p*, pero resulta más informativo obtener un intervalo de confianza. Como sabe, un intervalo de confianza proporciona el rango de valores del índice del tamaño del efecto que incluye el valor de la población con una determinada probabilidad. Incidentalmente le dice si el efecto es significativo, pero le dice mucho más: le proporciona una estimación del rango de valores que pudiera tener, un dato ciertamente útil para el conocimiento en una ciencia que presume de ser cuantitativa. (Por cierto, no creo que debamos utilizar rutinariamente intervalos del 95%: A menudo nuestros intereses son mejor servidos por intervalos más tolerantes del 80%.)

Recuerde que a todo lo largo del proceso por el que usted concibe, proyecta, ejecuta y escribe una

[3]Más asequibles que el texto de Tukey (1977) para iniciarse en el Análisis Exploratorio de Datos son los de Hartwig y Dearing (1979), Velleman y Hoaglin (1981) y Marsh (1988). En castellano puede consultarse la excelente obra recientemente publicada por Freixa, Salafranca, Guardia, Ferrer y Turbany (1992). *[N. del T.]*

[4]Esta opinión se está imponiendo no sólo en las ciencias del comportamiento (Rosenthal, 1990; Yeaton, 1988), sino también en otros campos de la investigación biosocial (cf. por ejemplo, Slakter, Wu y Suzuki-Slakter, 1991; Thomas, Salazar y Landers, 1991). *[N. del T.]*

investigación, es en su experto juicio como científico en lo que debe basarse, y esto es válido tanto para los aspectos estadísticos del trabajo como para los demás aspectos. Esto significa que su experto juicio gobierna el conjunto de parámetros implicados en la planificación (alfa, beta, el tamaño del efecto de la población, el tamaño muestral, el intervalo confidencial) y que su experto juicio también gobierna las conclusiones que usted extraiga.

En su brillante análisis de lo que él denominó "la revolución de la inferencia" en psicología, Gerd Gigerenzer demostró cómo y por qué no es posible un único camino para extraer conclusiones de los datos, y concretamente ninguno que no dependa en gran medida de las cuestiones sustantivas de interés -es decir, de cualquier cosa que entre en una investigación además del engranaje de los números. Un ingrediente esencial en el proceso de investigación es el buen criterio del científico. Este debe decidir en qué medida una proposición teórica ha sido anticipada por los datos, del mismo modo que decide qué estudiar, qué datos extraer y cómo extraerlos. Creo que la inferencia estadística aplicada con buen criterio es un instrumento útil en este proceso, pero no es el instrumento más importante: No es tan importante como cualquier otra cosa que venga antes que él. Algunos científicos, por ejemplo los físicos, trabajan sin la estadística, aunque por supuesto, no prescinden del buen juicio. De hecho, psicólogos muy buenos han trabajado sin la inferencia estadística: Me vienen a la mente Wundt, Kohler, Piaget, Lewin, Bartlett, Stevens y, si se me permite, Freud, entre otros. En efecto, Skinner (1957) pensó en dedicar su texto *Verbal Behavior* (y cito textualmente) "... a los estadísticos y metodólogos científicos con cuya ayuda este libro nunca hubiera sido terminado" (p. 111). Yo pienso que la correcta aplicación de la estadística por metodólogos estadísticos sensatos (Tukey, por ejemplo) no hubiera perjudicado el trabajo de Skinner. Incluso pudiera haberle venido bien.

Las implicaciones de las cosas que he aprendido (hasta ahora) no están en consonancia con gran parte de lo que veo a mi alrededor en la práctica estadística estándar. La usual decisión si–no al mágico nivel del .05 en una única investigación tiene poco que ver con el buen criterio. Simplemente, la ciencia no funciona de esa forma. Una investigación provechosa no soluciona un problema, tan sólo hace más probable en alguna medida una proposición teórica. Sólo la replicación futura exitosa en el mismo y en otros ambientes (como podría hacerse a través del meta-análisis) proporciona un enfoque para disipar el problema. El grado en que esta investigación única haga más probable la proposición depende de muchas cosas, pero no de si el valor p es mayor o igual que .05; .05 no es un precipicio, sino un punto de referencia conveniente a lo largo del continuo posibilidad-probabilidad. No existe una base ontológica para la toma de decisiones dicotómicas en la investigación psicológica. Este punto ha sido claramente planteado por Rosnow y Rosenthal (1989) el año pasado en el *American Psychologist*. Ellos escribieron que "... con toda seguridad, Dios quiere al .06 casi tanto como al .05" (p. 1277). A lo que yo digo ¨Amén!

Por último, he aprendido, aunque no fácilmente, que las cosas llevan tiempo. Como ya he mencionado, hace casi tres décadas publiqué un estudio de la potencia de los artículos del volumen de 1960 del *Journal of Abnormal and Social Psychology* (Cohen, 1962) en el que encontré que la potencia mediana para detectar un tamaño del efecto medio bajo condiciones típicas era tan sólo de .46. La primera edición de mi texto sobre la potencia fue en 1969. Desde entonces, se han publicado más de dos docenas de estudios de potencia y del tamaño del efecto en psicología y en campos relacionados (Cohen, 1988, pp. xi-xii). También ha habido una gran cantidad de artículos sobre la metodología del análisis de potencia[5]. Los textos de estadística, incluso algunos para alumnos, dedican algún espacio al análisis de la potencia, y ya hay disponibles varios programas de computador (e.g., Borenstein y Cohen, 1988)[6] Me dicen que algunas entidades financieras requieren que sus becas

[5]Dos monografías sobre el análisis de la potencia estadística son las de Lipsey (1990) y Kraemer y Thiemann (1987). *[N. del T.]*

[6]Otro programa de computador actualmente en el mercado ha sido desarrollado por Bavry (1991). El módulo complementario *DESIGN* del paquete estadístico *SYSTAT* 5.01 (Wilkinson, 1990) también incluye la posibilidad de realizar cálculos de potencia y determinación del tamaño muestral. *[N. del T.]*

subvencionadas contengan análisis de potencia y que en una de esas agencias puede encontrarse mi libro de la potencia en cualquier oficina.

El problema radica en que la investigación actual, tal y como es realizada, apenas refleja alguna atención a la potencia. ¿Cuán a menudo ha visto usted en las secciones de método de los artículos alguna mención de la potencia en las revistas que lee, por no hablar de un auténtico análisis de la potencia? El año pasado en el *Psychological Bulletin*, Sedlmeier y Gigerenzer (1989) publicaron un artículo titulado "Los Estudios de la Potencia Estadística ¿tienen algún Efecto sobre la Potencia de los Estudios?". La respuesta fue no. Utilizando los mismos métodos que utilicé yo con los artículos de 1960 del *Journal of Abnormal and Social Psychology* (Cohen, 1962), ellos realizaron un análisis de la potencia del año 1984 del *Journal of Abnormal Psychology*, y encontraron que la potencia mediana bajo las mismas condiciones fue de .44, algo peor que el valor .46 que yo encontré 24 años antes. Fue peor aún (.37) cuando tomaron en cuenta el uso ocasional del criterio alfa ajustado al experimento. Pero lo que es peor todavía, aproximadamente en el 11% de los estudios las hipótesis de investigación fueron enmarcadas como hipótesis nulas y su no significación fue interpretada como confirmación. ¡La potencia mediana de estos estudios para detectar un efecto medio al nivel bilateral de .05 fue de .25! Estos no son resultados aislados: Rossi, Rossi y Cottrill (en prensa), con los mismos métodos, hicieron un estudio de la potencia de los 142 artículos de los volúmenes del año 1982 del *Journal of Personality and Social Psychology* y del *Journal of Abnormal Psychology*, y encontraron esencialmente los mismos resultados[7].

Un ejemplo menos notorio de la inercia del avance metodológico es la correlación de conjuntos, que es una aplicación altamente flexible del modelo lineal general multivariante. Yo lo publiqué en un artículo en 1982, y lo incluímos como apéndice en la edición de 1983 de nuestro texto de regresión (Cohen, 1982; Cohen y Cohen, 1983). La correlación de conjuntos puede ser considerada como una generalización de la correlación múltiple al caso multivariante, con la que usted puede estudiar la relación entre cualquier cosa y cualquier otra, controlando lo que quiera bien en una, bien en otra, bien en ambas. Creo que es un gran método; al menos mis colegas, habitualmente críticos, no se han quejado. Sin embargo, por lo que yo sé, apenas ha sido utilizado fuera de la familia. (La publicación de un programa como módulo suplementario del *SYSTAT* [Cohen, 1989] puede cambiar las cosas.)

Pero no pierdo la esperanza. Recuerdo que W.S. Gosset, el compañero que trabajó en una fábrica de cerveza y que apareció modestamente publicado como "Student", publicó la prueba *t* una década antes de que comenzara la I^a Guerra Mundial, y la prueba no llegó a los textos de estadística psicológica hasta después de la II^a.

Estas cosas llevan tiempo. Así pues, si usted publica algo que cree que es realmente bueno, y pasa un año o una década o dos y apenas nadie parece haberle hecho caso, recuerde la prueba *t*, y anímese[8].

Referencias de las notas del traductor

Bartko, J.J. (1991). Proving the null hypothesis. *American Psychologist, 46(10),* 1089-1089.

Bavry, J.L. (1991). *Statistical Design Analysis System.* (2nd Edition). Chicago, IL: Scientific Software, Inc.

Cooper, H.M. (1989). *Integrating Research: A Guide for Literature Reviews* (2nd ed.). Newbury Park, CA: Sage (1st ed.: 1984).

Chow, S.L. (1991). Some reservations about power analysis. *American Psychologist, 46(10),* 1088-1089.

Eddy, D.M.; Hasselblad, V. y Shachter, R. (1992). *Meta-analysis by the Confidence Profile Method.* Boston, MA: Academic Press.

[7] J.S. Rossi (1990) ha publicado recientemente este mismo estudio, incluyendo el año 1982 de la revista *Journal of Consulting & Clinical Psychology*; y obteniendo resultados similares. *[N. del T.]*

[8] La reflexión hecha en este artículo por J. Cohen ya ha suscitado alguna polémica que el lector puede consultar en el número 10 del volumen 46 correspondiente al año 1991 de la revista *American Psychologist* (concretamente, Bartko, 1991; Chow, 1991; Gorsuch, 1991; O'Neil, 1991). *[N. del T.]*

Freixa, M.; Salafranca, L.; Guardia, J.; Ferrer, R. y Turbany, J. (1992). *Análisis Exploratorio de Datos: Nuevas Técnicas Estadísticas.* Barcelona: PPU.

Gómez Benito, J. (1987). *Meta-análisis.* Barcelona: PPU.

Gorsuch, R.L. (1991). Things learned from another perspective (so far). *American Psychologist, 46(10),* 1089-1090.

Hartwig, F. y Dearing, B.E. (1979). *Exploratory Data Analysis.* Beverly Hills, CA: Sage.

Hunter, J.E. y Schmidt, F.L. (1990). *Methods of Meta-analysis.* Newbury Park, CA: Sage.

Kraemer, H.C. y Thiemann, S. (1987). *How many Subjects? Statistical Power Analysis in Research.* Newbury Park, CA: Sage.

Lipsey, M.W. (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage.

Marsh, C. (1988). *Exploring Data: An Introduction to Data Analysis for Social Scientists.* Cambridge, MA: Polity Press.

O'Neil, P.M. (1991). Disturbing threat to academic obfuscation. *American Psychologist, 46(10),* 1090-1090.

Rosenthal, R. (1990). Replication in behavioral research. En J.W. Neuliep (Ed.), Handbook of Replication Research in the Behavioral and Social Sciences. [Special Issue.] *Journal of Social Behavior and Personality, 5(4),* 1-30.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research* (revised ed.). Newbury Park, CA: Sage (ed. original: 1984).

Rossi, J.S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting & Clinical Psychology, 58(5),* 646-656.

Sánchez Meca, J. y Ato, M. (1989). Meta-análisis: Una alternativa metodológica a las revisiones tradicionales de la investigación. En J. Arnau y H. Carpintero (Eds.), *Tratado de Psicología General,* Vol. I (pp. 617-669). Madrid: Alhambra.

Slakter, M.J.; Wu, Y.B. y Suzuki-Slakter, N.S. (1991). *, **, and ***; statistical nonsense at the .00000 level. *Nursing Research, 40(4),* 248-249.

Thomas, J.R.; Salazar, W. y Landers, D.M. (1991). What is missing in $p < .05$? Effect size. *Research Quarterly for Exercise and Sport, 62(3),* 344-348.

Velleman, P.F. y Hoaglin, D.C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston, MA: Duxbury Press.

Wachter, K.W. y Straf, M.L. (Eds.)(1990). *The Future of Meta-analysis.* New York: Sage.

Yeaton, W.H. (1988). Treatment effect norms. En J.C. Witt, S.N. Elliott & F.M. Gresham (Eds.), *Handbook of Behavior Therapy in Education* (pp. 171-187). New York: Plenum Press.

Referencias

Borenstein, M. y Cohen, J. (1988). *Statistical power analysis: A computer program.* Hillsdale, NJ: Erlbaum.

Children's height linked to test scores (October 7, 1986). *New York Times, p. C4.*

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

Cohen, J. (1965). Some statistical issues in psychological research. En B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1982). Set correlation as a general multivariate data-analytic method. *Multivariate Behavioral Research, 17,* 301-341.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1989). *SETCOR: Set correlation analysis, a supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT, Inc.Cohen, J. y Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Gigerenzer, G. y Murray, D.J. (1987). Cognition as intuitive statistics. Hillsdale, NJ: Erlbaum.

Glass, G.V. (1977). Integrating findings: The meta-analysis of research. En L. Shulman (Ed.), *Review of research in education* (Vol. 5, pp. 351-379). Itasca, IL: Peacock.

Lord, F.M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750-751.

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.

Neyman, J. y Pearson, E. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A,* 175-240.

Neyman, J. y Pearson, E. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A,* 263-294.

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

Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.

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

Rossi, J.S.; Rossi, S.R. y Cottrill, S.D. (en prensa). Statistical power in research in social and abnormal psychology. *Journal of Consulting and Clinical Psychology.*

Rozeboom, W.W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin, 85,* 1348-1351.

Salsburg, D.S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician, 39,* 220-223.

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

Skinner, B.F. (1957). *Verbal Behavior.* New York: Appleton- Century-Crofts.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

Wainer, H. y Thissen, D. (1976). When jackknifing fails (or does it?). *Psychometrika, 41,* 9-34.

Wilkinson, L. (1990). *SYSTAT: The system for statistics.* Evanston, IL: SYSTAT, Inc.

Traducción autorizada: 16-7-1992