Antropología y estadísticas:

Batallas en torno de la Hipótesis Nula

Antropología y estadísticas:
Batallas en torno de la Hipótesis Nula

> Nada sucede por azar. "Azar" [*chance*] no es un término técnico en estadística. Algunos practicantes usan "azar" para referirse a sucesos con iguales probabilidades, otros pueden tener en mente una independencia estadística entre variables, y otros más pueden usarla meramente para indicar que todavía no se conoce que esté sujeta a leyes. En discusiones técnicas es mejor evitar la palabra. Del mismo modo, "nada sucede al azar" aunque el muestreo al azar sea posible.
>
> Louis Guttman (1977: 94-95)

1. Introducción[1]

Hacia fines del siglo XX el estudio de las estructuras en red de la World Wide Web y la Internet demostró de manera dramática que las diversas distribuciones estadísticas que les eran propias no respondían al modelo de las distribuciones normales sino que se ajustaban a leyes de potencia (Barabási 2002; Reynoso 2011). Estas leyes de potencia (en lo sucesivo LP), similares a las viejas leyes de Pareto y de Zipf, difieren de las distribuciones gaussianas y afines tanto como es posible que dos cosas difieran, cualitativa y cuantitativamente. Al ser la LP una distribución en la que no es ni útil ni posible definir promedios o ejemplares representativos, en la que no existe un valor que pueda reputarse "normal" (intermedio más o menos exacto entre los valores extremos) y a la que no puede llegarse mediante operaciones aleatorias de muestreo, gran parte de las estadísticas convencionales no le son aplicables.

El problema que deseo tratar aquí tiene que ver con el hecho de que si bien la vigencia de la LP en la mayor parte de los fenómenos y procesos sociales y culturales ha sido plenamente reconocida, las ciencias humanas en general y el análisis de redes sociales en particular siguen adelante su negocio sin tomar cabalmente en cuenta el nuevo estado de cosas. No se han cambiado correspondientemente ni los diseños investigativos ni las herramientas de software, que siguen aplicándose como si la distribución normal siguiera siendo, como se creyó que lo era hasta hace poco, la madre de todas las leyes. En ningún lugar esto es más evidente que en el uso continuado y acrítico del manual por

[1] La investigación que fundamenta este ensayo se realizó en el contexto del proyecto trianual UBACYT F-155 (2008-2010) de la Universidad de Buenos Aires ("Modelos de Casos en Antropología y Complejidad").

antonomasia del análisis de redes sociales (ARS), el tratado de Wasserman y Faust (1994). Respecto de él he escrito recientemente:

> Cuatro o cinco años después de editado ese manual considerado pináculo en su género se descubrió que las redes de la vida real no exhiben las propiedades estadísticas que Wasserman y Faust dan por sentadas. No son pocos los cálculos que propone este tratado que deberían plantearse ahora de otra manera; lo mismo se aplica a diversos supuestos metodológicos (distribuciones de Bernoulli, muestreo, monotonía) y a las correspondientes estrategias de modelado y visualización. Aquí y allá el texto de Wasserman-Faust habla […] de modelado estadístico y pruebas de significancia sin reconocer que estas técnicas de *statistical testing* (englobadas en la sigla NHST) hace mucho se saben problemáticas (véase p. ej. Bakan 1960; Berkson 1938; Rozeboom 1960; Meehl 1967; Morrison y Henkel 1970; Carver 1978; Carver 1993; Gigerenzer 1993; Cohen 1994; Falk y Greenbaum 1995; Harlow, Mulaik y Steiger 1997; Hunter 1997; Shrout 1997; Daniel 1998; Feinstein 1998; Krueger 2001; Haller y Krauss 2002; Gigerenzer 2004; Armstrong 2007a; 2007b; McCloskey y Zilliak 2008).
>
> Conceptos que se han vuelto fundamentales (la fuerza de los lazos débiles, los mundos pequeños, las transiciones de fase, la coloración de grafos y sus generalizaciones, la teoría de Ramsey, las cajas de Dirichlet, el principio de los *pigeonholes*, los grafos de intersección, de intervalo y de tolerancia, los grafos pesados, los árboles abarcadores mínimos, la tratabilidad, la percolación, la escala, la no-linealidad, las alternativas a la ley del semicírculo, la teoría extremal de grafos, la optimización combinatoria, el análisis espectral, las matrices laplacianas, la noción misma de vectores o de valores propios) no se tratan en absoluto o se despachan a la ligera. El texto, de apariencia extrañamente setentista, permanece anclado en una concepción estructural-estática de las redes que contrasta con la visión procesual-dinámica que hoy se cultiva en los principales centros de investigación. Lo más grave, consecuentemente, es que el libro consolidó una visión analítica de las redes sociales, sin interrogar a través de un modelado genuino los mecanismos que hacen a su accionar o la posibilidad de intervenir en ellas (Reynoso 2011a: 23-24, n. 5).

Aunque hay razones técnicas de peso para introducir más de un cambio drástico, las perspectivas estocastológicas (como las llama Paul Meehl [1978]), indiferenciables de la Realidad Lineal General (Abbott 1988) y del paradigma de Mediocristán (como lo bautiza el polémico Nassim Taleb [2007]) siguen siendo dominantes. Aunque este conservadurismo engendra muchos factores que invitan a una discusión de gran interés, en este ensayo sólo me ocuparé de poner en tela de juicio uno solo de ellos, que es el que concierne a la prueba estadística de la hipótesis nula [NHST], una práctica que se ha impuesto como un requisito imperioso en el ARS y en la investigación en general pese a haber sido cuestionada severamente desde mucho antes y con independencia del redescubrimiento de la preponderancia de la LP en una proporción significativa de las cosas humanas. Todo un capítulo del mencionado manual canónico del ARS (Wasserman y Faust 1994: 603-674), por e-

jemplo, continúa patrocinando métodos de prueba de hipótesis que hace rato se saben engañosos sin decir palabra sobre el conflicto que ha estallado en torno suyo.

Ante este escenario, en el estudio que aquí comienza me propongo no sólo arrojar alguna luz sobre el estado de la polémica a propósito de la NHST, sino demostrar que los factores que hacen a la fragilidad del método tienen menos que ver con la interpretación de los resultados de oscuras operaciones aritméticas que con las premisas epistémicas de linealidad, sumatividad, homogeneidad y distribución normal alentadas por las corrientes principales de la estadística en su conjunto, comenzando por la lógica que rige las operaciones de muestreo. Aunque en un lado se promueva la cuantificación a ultranza y en el otro se cultive una escritura intensamente esteticista, cuando se las mira bien se descubre además que aquellas premisas se fundan en las mismas concepciones de encapsulamiento del todo en la parte, de monotonía y de representatividad que gobiernan la hermenéutica del conocimiento local en la antropología interpretativa, en sus derivaciones posmodernas y en todo el campo de los estudios culturales (Geertz 1983: 16; 2000: 133-140; compárese con Kruskal 1979a; 1979b; 1979c; 1980).

Cabe presumir que estas concordancias se deben a que dichos postulados de linealidad, convexidad y proporcionalidad son (en un sentido que podría llamarse foucaultiano) de alcance epistémico, en la medida en que impregnan, desbordan y sirven de fundamento a las estrategias y tradiciones teóricas más diversas. Como habrá de comprobarse, esos supuestos de normalidad (como se los llamará de ahora en más) proporcionan el marco no sólo a los modelos que pondremos en tela de juicio sino también a las formulaciones que los han impugnado, por mucho que éstas se piensen a sí mismas como las posturas críticas más revulsivas, innovadoras y contrastantes que cabe imaginar.

En consonancia con ese estado de cosas, en este estudio se ahondará en el diagnóstico de los factores que han llevado al presente *impasse* en los estudios estadísticos, factores que tienen que ver con supuestos seculares respecto de las distribuciones características en el campo de la sociedad, el territorio y la cultura. Estos supuestos han generado problemas que dudosamente se solucionen mediante coeficientes de aproximación, exclusión de ejemplares fuera de norma, sistemas de traducción de distribuciones, recálculos de ajuste y otras tácticas coyunturales que han convertido los procedimientos estadísticos en un bochornoso conglomerado de prácticas de inducción (en el sentido coercitivo de la palabra) que los unos promueven con suficiencia y los otros aceptan con mansedumbre (cf. Pridmore 1974; Spedding y Rawlings 1994; Kruskal y Stigler 1997: 110; Guthery 2008). Probaré aquí que lo que se requiere es más bien reformular radicalmente los principios epistemológicos, la comprensión de los procedimientos y los métodos de modelado, incorporando lo que se fue aprendiendo en lo que va del siglo en el desarrollo de las teorías y las técnicas de la complejidad organizada.

2. Prueba de la Hipótesis Nula – Teoría e historia

> Él usa estadísticas igual que un ebrio usa el poste de una luz: para apoyarse, más que para obtener iluminación.
>
> Andrew Lang, 1949

La prehistoria y la historia de la NHST preceden largamente a lo que se acostumbra consignar en su crónica oficial. En el siglo XVIII, particularmente en los trabajos de John Arbuthnott (1710), Daniel Bernoulli (1734), el reverendo John Michell (1767) y Pierre Simon de Laplace (1773) ya se encuentran aplicaciones aisladas de un conjunto de ideas que poseen un fuerte aire de familia con las prácticas ulteriores.[2] En el siglo XIX métodos de prueba todavía más parecidos a los actuales aparecen en la obra de autores tales como Jules Gavarret (1840), Wilhelm Hector Lexis (1875; 1877) y sobre todo Francis Ysidro Edgeworth (1885a: 187).

Es Edgeworth quien menciona por primera vez la palabra "significante" en el contexto de la elaboración de un método descriptivo o comparativo que utiliza el examen de leves desviaciones a partir de una "ley del error" (o sea, de la distribución normal) para "poner a prueba" teorías de correlación (p. 200). Conviene subrayar que Edgeworth usa el calificativo "significante" y no el sustantivo "significancia", que es lo que por lo común se le atribuye. Lo hace en el acto de preguntarse si "las diferencias observadas entre las estaturas medias de 2315 criminales y la estatura media de 8585 adultos británicos de sexo masculino es significante". Aunque en lo personal dudo que Edgeworth haya pretendido acuñar una expresión técnica perdurable, la idea de significancia quedó tan embebida en la narrativa que sustenta lo que luego sería la prueba de hipótesis que la 'S' de su abreviatura más común en inglés (NHST) ha vuelto a denotar *Significance* antes que *Statistical* (Nickerson 2000: 241).

La era moderna de las pruebas de significancia comienza con el trabajo de Karl Pearson (1900) sobre el χ^2, expresivamente intitulado "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", y con los estudios de "Student", o sea William Sealy Gosset (1908), sobre la prueba *t*. El mero título del trabajo de Pearson constituye una descripción que describe con insuperable economía de palabras la esencia del método que se desarrollará después.

---

[2] He incluido referencias a todas estas obras y a las que siguen (con sus vínculos en línea) en el hipertexto bibliográfico al final de este libro.

La prueba estadística tal cual hoy se la conoce es una versión híbrida (de autor anónimo) de las pruebas propuestas por Sir Ronald Fisher (1922; 1925) por un lado y por Jerzy Neyman y Egon Pearson (1933a) por el otro.[3] A decir verdad no se trató de una creación armónica surgida de un trabajo cooperativo sino que fue fruto de un enfrentamiento dialógico, plasmado en un intertexto en el que nadie ha sabido encontrar a escala de cada una de las ideas la menor base de acuerdo. Fisher y Neyman, en particular, tenían en muy baja estima la capacidad intelectual del otro y polemizaron hasta el insulto. En una ponencia expresivamente titulada "Bodas de Plata de mi disputa con Fisher", Neyman (1961) alegaba que Fisher era especialmente inhábil para operar con conceptos. Éste respondió en su correspondencia quejándose de la abundancia de "neymanianos chiflados de California", gustosos de intimidar a la gente y en especial a refugiados extranjeros "ansiosos de obtener puestos en universidades norteamericanas" (Louçã 2008: 6). A la larga el enfrentamiento devino, en opinión de algunos, "una batalla que tuvo un efecto enormemente destructivo sobre la profesión estadística" (Zabell 1992: 382; Lehmann 1993: 1242).

Al lado de eso, que parecería puramente anecdótico, Fisher y Pearson poseían visiones muy distintas del papel de la probabilidad en el análisis estadístico. Las diferencias que mediaban entre ellos han alimentado escrituras erizadas y lecturas inconciliables; pero de un modo u otro la rutina logró consolidarse más allá de las incongruencias presentes en todas y cada una de las variantes de hibridación que se han propuesto. Con el correr de los años los usuarios de los métodos estadísticos dejaron de leer los textos originales, sustituyéndolos por manuales sin aparato erudito, sin contexto y sin temporalidad y adoptando hábitos académicos de enculturación profesional más inclinados a la mecanización de la práctica que al refinamiento reflexivo de la teoría. Las discordancias entre las posturas estadísticas originales no impidieron que se gestara una fusión entre las ideas del valor de *p* según Fisher y la tasa de error del tipo I (α) de Neyman-Pearson y así todo a lo largo del método hasta englobar ambas pruebas en un razonamiento que tanto los pocos fisherianos que restan como los muchos neymanianos que existen presumen inconsútil, como se examinará luego con mayor detalle (cf. pág. 54).

Esto ha redundado en una pérdida general de la calidad argumentativa, pues las ideas de Fisher alrededor de la prueba de significancia y la inferencia inductiva y las de Neyman-Pearson sobre prueba de hipótesis y conducta inductiva no sólo difieren en su enunciación sino que son visceralmente antitéticas en sus fundamentos, su terminología, su lógica y sus propósitos (Goodman 1993; Hubbard

[3] Jerzy Spława-Neyman [1894-1981] fue un matemático polaco que ya consagrado tuvo que trasladarse a Berkeley porque nadie menos que Fisher le hacía la vida imposible en el University College de Londres; aunque ha sido una figura mayor, Egon Pearson [1895-1980] no es en modo alguno el Pearson propiamente dicho sino el único hijo de Karl Pearson [1857-1936], figura incuestionablemente más célebre y fundador egregio de las estadísticas matemáticas.

y Bayarri 2003). El nombre mismo de la prueba híbrida, cuya paternidad no ha quedado establecida y que no se encuentra nunca en los textos canónicos, se cristalizó como si se hubiese buscado maximizar la conflictividad de la idea. Hasta la frase que la denota es un oxímoron: ni Fisher aceptaba la posibilidad de una "prueba estadística", ni Neyman o Pearson usaron jamás la expresión "hipótesis nula".

Neyman asignaba un significado conductual a un resultado significante, el cual era emergente de una decisión del investigador; Fisher excomulgaba esta concepción por creerla "pueril […], producto de una matemática sin contacto personal con las ciencias naturales" (1955: 75, 69). Los resultados de Fisher no estaban supeditados a una decisión interpretativa del estudioso sino que –decía él– se hallaban latentes en los datos. El contraste entre una visión y otra se asemeja a la disyuntiva que alguna vez se manifestó en el análisis lingüístico entre las estrategias que buscaban descubrir una "verdad de Dios" [*God's truth*] objetivamente presente en el objeto y las que instrumentaban una opción de "abracadabra" [*hocus pocus*], abstrayendo y organizando las propiedades del objeto hasta que aparecía lo que se estaba buscando (Householder 1952: 159). Ésta es una instancia que refleja la misma tensión que media entre empirismo y racionalismo, o más (pos)modernamente entre objetivismo y subjetivismo, como se llegaría a poner de manifiesto de manera explícita en el programa de las estadísticas bayesianas (Iversen 1984; Ellison 1996).

Si bien unos cuantos autores han minimizado las discrepancias entre ambos modelos estadísticos y los manuales nada dicen sobre ellas, a la hora del desenvolvimiento del método el estallido de las incoherencias es imposible de disimular. En psicoanálisis –sugieren Sedlmeier y Gigerenzer (1989: 313)– nadie osaría presentar una mixtura de las teorías de Freud y de Adler como representativa del psicoanálisis a secas; en estadísticas, sin embargo, la versión mixta del test se promueve con toda naturalidad como si fuese la prueba estadística por antonomasia. Más todavía, "la presentación anónima de una 'estadística inferencial' monolítica facilitó la supresión de los elementos de juicio controversiales" (Gigerenzer 1987: 20-21) algunos de los cuales ya hemos entrevisto.

Mientras más se escribe sobre la autoría y el significado de las hibridaciones más enrevesado se vuelve todo, pues las diversas lecturas, sean divergentes o conciliadoras, terminan adquiriendo tanto peso como (o acaban confundiéndose con) el contrapunto de lo que se encuentra objetivamente escrito en los originales. El campo abunda en razonamientos subjuntivos y condicionales que giran en círculo, preguntándose cosas tales como "¿qué pensaría Fisher (o Neyman, o Pearson) de la formulación mixta de la NHST?" o "¿rechazaría Pearson la filosofía estadística de Neyman-Pearson?". Mientras tanto, un número crecido de autores de primera línea ha sabido identificar importantes fallas del método imputables a las formas azarosas en que las ideas se han contaminado, yuxtapuesto o sustituido; sus textos ponen en relieve inconsistencias que la rutina ha invisibilizado pero que ni

es probable que se arreglen solas ni está de más sacar a la luz (Tukey 1960a; Zabell 1992; Lehmann 1993; 1995; Chow 1998; Mayo y Spano 2006; Goodman 2008).

Fisher elaboró sus procedimientos primordialmente en dos publicaciones, *Statistical methods for research workers* (1925) y *The design of experiments* (1971 [1935]). La idea de la hipótesis nula no es sino el componente central de una concepción más amplia, atinente a la inferencia inductiva. Fisher estaba convencido de que "es posible argumentar […] desde las observaciones a las hipótesis" (1971: 3). Para lograr este objetivo el investigador define primero una hipótesis nula; esta se ve "des-probada" si la muestra estimada se desvía de la media de la distribución de muestreo por una cantidad mayor a la de un criterio especificado, llamado el nivel de significancia o valor crítico de *p*, el cual se sugiere que se fije en el orden del 5% (Fisher 1971: 13). Se dice por ello que la prueba fisheriana de significancia se centra en el rechazo de la hipótesis nula al nivel de $p \leq 0{,}05$. En otros textos algo posteriores Fisher acepta que el nivel se configure en un 2% o un 1% si se estima necesario hacerlo.

El desarrollo del método está lleno de desmentidas, bifurcaciones y cambios de rumbo. En su último libro, *Statistical methods and scientific inference*, Fisher ridiculizó la idea de Neyman-Pearson de que el nivel de significancia debía fijarse a priori en un 5%, diciendo que era "absurdamente académica, pues de hecho ningún trabajador científico se atiene a un nivel de significancia fijo según el cual año tras año y en todas las circunstancias se rechazan las hipótesis; más bien abre su mente a cada caso particular a la luz de la evidencia y de sus ideas" (Fisher 1956: 42). Fisher también repudió la lógica de los múltiples experimentos con muestreo aleatorio propuesta por sus rivales y cuestionó su propia sugerencia de fijar un nivel convencional de significancia, recomendando que se hiciera público el nivel exacto encontrado con posterioridad al experimento (como por ejemplo $p \leq 0{,}03$) por las razones que en seguida se verán.[4]

Él consideraba, en efecto, que el valor de *p* constituía evidencia inductiva contra la HN: cuanto más pequeño el valor, más fuerte la evidencia, aunque no sea eso lo que *p* mide primordialmente. En una época en que tomaba vigor la creencia en que la inducción constituía el escándalo de la filosofía, Fisher pensaba que la estadística podía jugar un papel fundamental en la consolidación de la inferencia inductiva, o sea, en el desarrollo de la inferencia "a partir de los resultados de la experi-

[4] En una fascinante relectura de los textos de Fisher, el celebrado Leonard Savage (estadístico de la Universidad de Yale) ha identificado peculiaridades de su estilo de polemicidad que hacen que su interpretación sea particularmente complicada. No es inusual, por ejemplo, que Fisher escriba algo así como "Algunos han propuesto que…" y que en seguida arremeta fieramente contra la propuesta, despedazándola, sin aclarar que no es otro que él mismo el autor al que hace objeto de tamaña hecatombe. Algo así sucedió con el 5%. Savage cree no ser el único en sospechar que algunas de las concepciones más importantes de Fisher fueron elaboradas simplemente para evitar estar de acuerdo con sus oponentes (cf. Savage 1976).

mentación", sosteniendo que "es posible argumentar desde las consecuencias a las causas, de las observaciones a las hipótesis; como lo diría un estadístico, desde una muestra a la población de la cual la muestra fue tomada, o, como lo diría un lógico, de lo particular a lo general" (Fisher 1971: 3). De esta manera el valor de *p* asumía un rol epistemológico central (Hubbard y otros 2003: 172).

Se ha dicho que la inferencia inductiva implicada en la prueba estadística pertenece a la única clase de procesos lógicos mediante la cual se genera nuevo conocimiento. Fisher, que era biólogo, estadístico y genetista evolucionario, estaba consciente sin embargo de que muchos estudiosos no compartían su punto de vista, y "en especial los matemáticos [como Neyman] que han sido entrenados, como la mayoría de los matemáticos lo ha sido, casi exclusivamente en la técnica de razonamiento deductivo [y que por ende …] niegan desde el vamos que las inferencias rigurosas de lo particular a lo general incluso puedan ser posibles" (Fisher 1935: 39).

En la prueba de hipótesis de Fisher no se especifica ninguna hipótesis adicional fuera de H_0 y el valor de *p* (que resulta del modelo y los datos) se evalúa como la fuerza de la evidencia a favor de la hipótesis de investigación aunque no sea ésa su definición formal. No existe ninguna noción de potencia de la prueba ni se dice nada sobre aceptar (o no) una hipótesis alternativa. A la inversa, la prueba de Neyman-Pearson identifica hipótesis complementarias θ_A y θ_B tal que el rechazo de una involucra la aceptación de la otra, rechazo que se basa en un nivel α determinado a priori. Aunque los comentaristas que se han ocupado del tema discrepan, Neyman y Pearson (1933b) sí utilizan a veces la palabra "aceptar", lo cual es ostensiblemente impropio (Gill s/f: 4; Greenwald 1975: 2; Iacobucci 2005). Las recomendaciones de importantes publicaciones periódicas contemporáneas, empero, encarecen "nunca utilizar la infortunada expresión 'aceptar la hipótesis nula'" (Wilkinson 1999: 599): en ninguna esfera de la actividad científica, del pensamiento o de la práctica –argumentan– *no poder negar algo equivale a afirmarlo*. Que no se haya podido encontrar significancia en una prueba estadística particular –se sigue de ello– no implica que la HN sea verdad (Sprent y Smeeton 2001: 1.3.1).

La diferencia cardinal entre la postura de Fisher y la llamada de Neyman-Pearson (aunque se sabe que Pearson no participó en las polémicas con Fisher y luego tomó distancia de las ideas de Neyman) se establece en el momento en que Neyman y Pearson introducen nada menos que la hipótesis alternativa.

> [C]uando se selecciona un criterio para la prueba de una hipótesis particular *H* ¿debemos considerar sólo la hipótesis *H*, o algo más? Es sabido que algunos estadísticos son de la opinión de que se pueden desarrollar buenas pruebas no tomando en consideración nada más que la hipótesis [nula]. Pero mi opinión es que esto es imposible y que si de hecho se han desarrollado pruebas satisfactorias sin consideración explícita de nada más que la hipótesis pro-

> bada, ello se debe a que los respectivos autores toman subconscientemente en consideración ciertas circunstancias relevantes, a saber, la hipótesis alternativa que podría ser verdad si la hipótesis probada es errónea (Neyman 1952: 44).

Como dijimos, Fisher nunca había hablado de una hipótesis alternativa. Cuando ambas ideas al principio incompatibles se mezclan, la H_A comienza a jugar el papel de complemento de la H_0, aunque la especificación en ese sentido sea más que ambigua. Mirándolo bien las hipótesis en sí no se contrastan nunca, ya que el análisis se practica sobre conjuntos de datos cuyos mecanismos hipotéticos de origen no son (ni podrían ser) objeto de interrogación en términos estadísticos.

Pero H_A no es la única contribución de Neyman-Pearson; su marco de referencia introduce además las probabilidades de cometer dos clases de errores basados en consideraciones relativas al criterio de decisión, al tamaño de la muestra y al tamaño del efecto [*effect size*]. Esos errores son el falso rechazo (error de Tipo I) y la falsa aceptación (error de Tipo II) de la hipótesis nula. La probabilidad del primero se denomina α, la del segundo β. Más adelante describiré estos elementos uno por uno. Insólitamente, es en el intercambio entre Fisher y Neyman que comienza la trayectoria agonística y conflictiva de la NHST (cf. Tabla 1). Rechazando la necesidad de contar con una hipótesis alternativa, por ejemplo, Fisher escribía:

> La noción de un error de la así llamada "segunda clase" debido a la aceptación de la hipótesis nula cuando ella es falsa […] carece de sentido con respecto a una simple prueba de significancia, en la cual las únicas expectativas disponibles son aquellas que fluyen del hecho de que la hipótesis nula sea verdad (Fisher 1971 [1935]: 17).

Prueba de significancia	Prueba de hipótesis
La información es una extracción de una población hipotética infinita.	La prueba se concibe como una muestra obtenida por un muestreo repetido.
Las estadísticas son una propiedad de la muestra.	El tamaño y la potencia son propiedades del test.
1. Especificar H_0	1. Especificar H_0 y H_1
2. Especificar estadística de prueba (T) y distribución de referencia	2. Especificar estadística de prueba (T) y distribución de referencia
3. Recolectar datos y calcular el valor de T	3. Especificar valor de α y determinar región de rechazo (R)
4. Determinar valor de *p*	4. Recolectar datos y calcular el valor de T
5. Rechazar H_0 si el valor de *p* es pequeño; si no es así retener H_0	5. Rechazar H_0 en favor de H_1 si el valor de T está en la región de rechazo; si no es así retener H_0

Tabla 1 – Prueba de significancia de Fisher y prueba estadística de Neyman-Pearson.
Basado en Huberty (1993: 318) y Louçã (2008: 10).

En opinión de William Kruskal, la negativa de Fisher de siquiera tomar en cuenta una hipótesis opuesta a la hipótesis nula es desconcertante:

> Una de las cosas que intrigan cuando se lee a Fisher –escribió Kruskal– es su silencio sobre la relevancia de hipótesis alternativas a la que se encuentra bajo prueba, esto es, a la hipótesis bajo la cual se computa el nivel de significancia de la prueba o en la que se basa la prueba formal. Sin una especificación de hipótesis alternativas, aunque fuese rudimentaria, es difícil ver de qué manera se puede elegir una estadística de prueba o definir una familia de regiones críticas (Kruskal 1980: 1022).

Es posible conjeturar que Fisher omitió toda referencia a una hipótesis alternativa *a*) porque quería dejar que los hechos hablasen por sí mismos; *b*) porque intuía que un problema inverso de inducción (merced al principio de equifinalidad) posee muchas soluciones posibles, *c*) porque (conforme lo estableció más tarde Nelson Goodman [1972]) existe un número indefinido, posiblemente infinito de enunciaciones "opuestas" a cualquier argumentación específica, o *d*) por impulso de su personalidad peculiar. A mi juicio la primera y la última alternativa son las más plausibles; las otras dos hablarían de un pensador más riguroso y son epistemológica y metodológicamente más sustanciosas, pero no se podrían demostrar sin incurrir en una colección de anacronismos.

Algunos autores, como dije, procuran conciliar las visiones contrapuestas de Fisher y Neyman-Pearson; el problema que subsiste es que no sólo hay desacuerdos profundos en la interpretación y la metodología, sino en los resultados que pueden obtenerse según el modelo al que uno se atenga. Supongamos que tenemos unos datos tales como $X_1, \ldots X_n$ pertenecientes a una distribución $\mathcal{N}(\theta, \sigma^2)$, con un valor de σ^2 conocido y n=10, y que se desea testear $H_0: \theta=0$ contra $H_1: \theta\neq 0$. Si z=2,9 ó $z = \sqrt{n}\,\bar{x}/\sigma = 2{,}3$ es seguro que Fisher reportaría valores de p=0,021 ó p=0,0037. Si Neyman hubiera prescripto la probabilidad de error de tipo I α=0,05, reportaría en cambio α=0,05 en ambos casos (Berger 2003: 1). Induce a engaño entonces que los manuales de estadística enseñen primero la prueba frecuentista de Neyman-Pearson y sin solución de continuidad hablen de valores de *p*, sin poner suficiente énfasis en el hecho de que ambas ideas son tributarias de metodologías distintas (y sin aclarar, por ejemplo, que un valor de *p* de 0,05 corresponde a menudo a un error frecuentista de probabilidad de 0,5).

Si hoy hay que embarcarse en una especie de hermenéutica antropológica de la estadística para recabar elementos de juicio tan fundamentales es porque en su momento nadie indagaba semejantes cosas. Lo concreto es que el análisis de las contingencias de gestación y de los fermentos intelectuales en los que se desenvolvió la NHST distaba de ser un tema prioritario en los años treinta. Tanto la idea de una reflexividad epistemológica como la historiografía estadística adquirieron entidad varias décadas después, lo que en algunos respectos es como decir demasiado tarde. Los pioneros de la prueba estadística ya habían muerto para entonces y hasta sus epígonos inmediatos estaban al borde del retiro.

La historia temprana de la prueba recién fue reseñada con medio siglo de atraso por Carl Huberty (1993) de la Universidad de Georgia, quien examinó solamente materiales de libros sin incluir referencias a publicaciones periódicas. La semblanza de Huberty es sólida pero dista de ser definitiva, pues el autor no tuvo oportunidad de investigar el impacto de la bibliografía consultada en la investigación concreta. En la actualidad hay un fuerte movimiento revisionista en el que se promueve el examen en profundidad de los archivos y las fuentes para deslindar mejor la génesis y transformación de los conceptos y sus consecuencias filosóficas. Hasta la fecha, las contribuciones más valiosas de este revisionismo probablemente sean las de Deborah Mayo; ellas han logrado aclarar algunos puntos oscuros, aunque el sesgo de su propia postura interfiera bastante en la selección de los acontecimientos y mucho más aun en la comunicación de las ideas (1980; 1992; 1996; Berger 2003; Mayo y Spano 2006). El caso es que la estadística del siglo XIX se conoce hoy mejor que esta otra, más relevante y más tardía, pero de la que ya no sobreviven pioneros que nos puedan revelar las claves ocultas: ni a la época que nos ocupa le ha llegado su Stephen Stigler (1978), ni los ánimos se han tranquilizado lo suficiente, ni sus documentos de dominio público se han expuesto todavía en la Web.

Como sea, con lo que se ha visto hasta el momento ya alcanza para hacerse una idea de los propósitos de la técnica y de su polemicidad inherente, la cual alcanza extremos tan álgidos que todos y cada uno de los términos cardinales del vocabulario implicado (“aceptación de la hipótesis”, “hipótesis nula”, “hipótesis alternativa”, “errores de tipo I y II”, “conducta inductiva”, “inferencia inductiva”, “prueba estadística”, “prueba de significancia”, “decisión”, “muestreo repetido”, “valor de *p*”) se han visto alternativamente proscriptos o puestos en ridículo no tanto por una ociosa crítica epigonal sino, significativamente, por alguno de los tres estudiosos que les dieron origen.

3. El discurso del método

> Pronunciando cada palabra con enorme deliberación, el senador Resent preguntó: "¿Es usted hoy, o ha sido alguna vez, miembro de la Asociación Americana de Estadística?"
>
> Mirando al senador Resent directo a los ojos, Minnie replicó desafiante: "Me rehuso a contestar, dado que puede incriminarme".
>
> Frank Proschan, 1954

Este es el punto en el que corresponde describir el desarrollo de una prueba en términos de la versión híbrida que ha devenido estándar. Siguiendo a grandes rasgos la descripción de Geoffrey Loftus (2010) propongamos un típico experimento de prueba, en el que se trata de medir el efecto de una sola variable independiente (p. ej. la cantidad de información mantenida en la memoria) sobre una sola variable dependiente (p. ej. el tiempo requerido para barrer esta información). Típicamente la variable dependiente principal toma la forma de una media. En el ejemplo sería tiempo medio de reacción. Por lo general también se supone que las conclusiones de los experimentos se aplican a medias de la población. Notacionalmente, las medias de la muestra se refieren como M_j mientras que la media de la población se denota μ_j.

La NHST entraña establecer y evaluar dos hipótesis mutuamente excluyentes sobre la relación entre la variable independiente y la dependiente. La HN o H_0 habitualmente afirma que la primera variable no ejerce ningún efecto sobre la segunda. La hipótesis alternativa (H_1, a veces H_A) asevera, por el contrario, que sí ejerce algún efecto. Existe por ende una visible asimetría entre ambas hipótesis. La HN es una hipótesis exacta; la otra no. Hay una sola forma en que la HN puede ser correcta (esto es, que todas las μ_j sean iguales) mientras que hay muchas formas en que H_1 lo sea. Sirviéndonos de otra terminología, diríamos que en las ciencias empíricas siempre estamos en presencia de problemas inversos, los cuales, según anticipé, admiten una cifra posiblemente indefinida y probablemente muy grande de soluciones suficientemente correctas. El número de principios algorítmicos y de valores de parámetro que puede generar una distribución observable (un número más intuido que efectivamente computado) es entonces lo que se ha definido como inexactitud; la HN es, por decirlo de algún modo, una hipótesis exacta no perteneciente a ese conjunto plural de H_1 posibles.[5]

[5] Decir "plural" es, en todo caso, una expresión minimalista: "virtualmente infinito" sería un calificativo más apropiado. Ni aun en sus formulaciones matemáticamente más rigurosas (Lehmann y Romano 2005) se ha podido establecer un nexo formal acotado entre una hipótesis y otra que la niega, tal que la falsedad de una implique la verdad de la otra. Para cualquier afirmación *A*, su opuesto, su negación, su contrario (o como quiera que se designe a no-*A*) comprende el conjunto infinito de lo posible, exceptuando específicamente a

El ingrediente principal de la NHST es la comparación de los valores de las respectivas medias de la muestra M_j. En la medida en que éstas sean próximas, la evidencia sugiere la posible igualdad de las μ_j y por ende la validez de la HN. A la inversa, conforme al grado en que las M_j difieran entre sí, la evidencia se inclina hacia las diferencias asociadas entre las μ_j y, por ende, hacia la validez de la hipótesis alternativa.

La asimetría entre la HN (que es exacta) y la hipótesis alternativa (que no lo es) implica una asimetría asociada en las conclusiones sobre la respectiva validez de cada cual. Si las M_j difieren significativamente, se dice que uno "rechaza la hipótesis nula" a favor de aceptar la hipótesis alternativa; pero si las M_j no difieren sustancialmente, es incorrecto decir que uno "acepta la hipótesis nula"; lo que corresponde expresar es más bien que uno "falla en rechazar la hipótesis nula". La razón de este fraseo sinuoso pero lógicamente necesario es que dado que la hipótesis alternativa es inexacta, no es posible distinguir entre una HN genuinamente verdadera y una hipótesis alternativa que entrañe diferencias muy pequeñas entre las μ_j. La única conclusión fuerte posible en el contexto de la NHST es, por lo tanto, el rechazo de la HN (Loftus 2010: 6). Muy poca cosa, por cierto, si pensamos que la HN es casi siempre demasiado evidentemente falsa excepto para los casos más triviales, para situaciones en las que se impone una actitud escéptica (cf. pág. 42 y ss.) o para una multiplicidad de escenarios enunciativos que recién ahora se están comenzando a pensar.[6]

La prueba de significancia puede de hecho realizarse en función de observaciones y sin ninguna otra hipótesis fuera de la HN. Aunque habría preferido echar mano de un caso antropológico y no recurrir nunca a loterías y ruletas para ilustrar la idea, sugiero que pensemos con detenimiento en el ejemplo que sigue, dado que es un clásico imposible de superar en materia de pedagogía estadística:

Si queremos determinar si el revoleo sucesivo de una moneda no está sesgado y se aproxima al 50% de caras o cruces en una secuencia de 20 intentos, tomaríamos el valor de (por ejemplo) la cantidad de caras obtenidas. Supongamos que ese valor es 14. El valor de *p* sería entonces la probabilidad de que se obtengan por lo menos 14 caras en 20 intentos. Esta probabilidad puede calcularse de diversas formas; una de ellas sería mediante una cuenta muy simple basada en los coeficientes binomiales. En la notación de Andreas von Ettingshausen (derivada del triángulo de Yang Hui o de Pascal) este cálculo se desenvolvería de este modo:

A. Como decía el antropólogo Gregory Bateson (1980), una negación y una afirmación pertenecen a niveles distintos de tipificación lógica (ver también Goodman 1972). No es de extrañar entonces que los campos de estudio de problemas inversos y de prueba estadística de hipótesis hayan discurrido por carriles separados (cf. Ramm 2005; Kaipio y Somersalo 2006).

[6] Véase por ejemplo http://www.jasnh.com/. Visitado en julio de 2011.

$$\frac{1}{2^{20}}\left[\binom{20}{14}+\binom{20}{15}+\binom{20}{16}+\binom{20}{17}+\binom{20}{18}+\binom{20}{19}+\binom{20}{20}\right]=\frac{60.460}{1.048.576}\approx 0{,}05765914$$

Ecuación 1 – Probabilidad del valor de *p* para cada lado de la moneda

El número obtenido es el valor de *p* y mide la posibilidad de que la secuencia de 20 revoleos pueda dar un valor igual o superior al valor observado. En este punto, deberíamos rechazar la HN si el valor de *p* es menor o igual que el nivel de significancia, a menudo representado por la letra griega α y que fijaremos aquí clásicamente en 0,05. Esto implica que cualquier valor menor que 0,05 implicará el rechazo de la HN (al 5% del nivel de significancia). Las 14 caras que hemos obtenido se desvían de la paridad por un número de 4 en ambas direcciones. En el ejemplo de las 14 caras y 6 cruces, debemos calcular entonces la probabilidad de obtener un resultado que se desvíe de la paridad por lo menos en esa magnitud. Como la distribución binomial es simétrica para una moneda de dos caras, el valor de *p* para un test de doble cola [*two-tailed test*] es simplemente el doble del valor de *p* obtenido en la Ecuación 1, o sea 0,0576…x 2 = 0,1152.

Para hacerla simple, tenemos entonces que la HN (que afirma que no hay desvío) se establece en 0,5, que la observación O es de 14 caras sobre 20 lanzadas y que el valor de *p* para un test de doble cola es 0,1152. Como este valor de *p* excede a 0,05, la observación es consistente con la HN, esto es, con la afirmación de que el resultado observado puede deberse solamente al azar. Aunque la moneda no cayó en forma pareja, no nos es posible rechazar la HN al nivel del 5%. Si lo hiciéramos, incurriríamos en lo que en una prueba de hipótesis sería un error de Tipo I, al cual definiré más acabadamente un par de capítulos más adelante (pág. 32).

En teoría, el proceso de la NHST presupone el uso de varias técnicas específicas, tales como un muestreo aleatorio, el cálculo de los coeficientes binomiales, la prueba *t* (desarrollada por “Student” [1908]) o la prueba *F* (de Fisher). Muchos de estos cálculos y pruebas proporcionan un valor de *p*, el cual existe al menos desde las *Tables for the Staticians and Biometricians* de Karl Pearson (1914), anterior en unos diez años a la prueba fisheriana de significancia. Idealmente el estudioso debería tener robustas nociones de regresión, regresión múltiple y diversas variaciones del análisis de varianza. En la vida práctica, sin embargo, las reglas del juego son otras y varían conforme al perfil profesional de los investigadores, por lo que sería imposible que intente describir en este ensayo el desarrollo de esas pruebas en particular sin resultar ya sea tedioso para el conocedor o ininteligible para el no iniciado.

Sea que el usuario de los programas estadísticos esté consciente de ello o no, entre los datos que ingresan por un lado y las respuestas que salen por el otro todas las fases del procedimiento se encuentran, convenientemente, embutidos en una caja negra. Ella permite ejecutar las operaciones de inferencia que sea menester en el más alto nivel, en la más feliz ignorancia de los supuestos que

hacen que las muestras que se tienen entre manos hayan llegado a ser como son y de los tecnicismos que rigen el sentido de los cálculos de los que las muestras serán partícipes. Hoy se dispone de gran número de instrumentos para obtener un valor de *p* en contados milisegundos. La Calculadora Estadística de Daniel Soper, por ejemplo, suministra varios métodos de cálculo, incluyendo todas las pruebas que se nombran en este párrafo.[7] Con Microsoft® Excel™, por otro lado, se pueden calcular fácilmente valores de *p* a partir de las pruebas de Student, chi cuadrado, *F* y Durbin-Watson, entre otras.

Si de llevar adelante los cálculos se trata, no es preciso recapitular la tortuosa ontogénesis de la ciencia estadística para hacerse de una idea y obtener resultados que lucen significativos. Como en todo campo saturado de alternativas hay incluso un puñado de libros bastante operativos diseñados con el propósito de ser inteligibles y guiar la ejecución del método. Si tuviera que definir cuál es el mejor texto para aprender paso a paso todas las operaciones involucradas en diferentes escenarios de prueba de hipótesis sin supuestos previos de conocimientos de estadísticas yo diría que es (con las debidas disculpas por el agravio implicado) la versión electrónica de *Statistics for Dummies* de Deborah Rumsey (2003); en la misma tesitura se encuentra *Even you can learn statistics* de David Levine y David Stephan (2010).

El cielo es el límite, tal parece: si la crónica del método acabara en este punto, los investigadores podrían ser llevados a creer que existe una herramienta de mecanización de la persuasión y el progreso científico que por poco que se la alimente con un puñado de mediciones es capaz de poner la prueba automática de hipótesis al alcance de los dedos.

[7] Véase http://www.danielsoper.com/statcalc/default.aspx#c14. Visitado en junio de 2011. El Equipe Raisonnement Induction Statistique ha desarrollado un paquete, Le PAC (Programa para el Análisis de Comparaciones), que incluye tanto los procedimientos frecuentistas como los bayesianos (cf. la versión 2.0.2 en http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/ErisA.html).

4. El lado oscuro de la inferencia inductiva

La facilidad con que realizan estas operaciones corre pareja al grado de ofuscación conceptual que sobreviene apenas se pretende establecer qué significa cada término enclavado en la metodología. De todas las pruebas estadísticas básicas, la obtención del valor de *p* es la única cuya descripción viene acompañada, invariablemente, de una especificación de los múltiples equívocos que acostumbran circundarla.[8] Por más que los defensores de la NHST acostumbren echar la culpa al mensajero, alegando que éste ignora tal o cual elemento de juicio, o asegurando que hay miles de investigadores que han utilizado felizmente el método, o que los críticos no han sabido proponer técnicas mejores, la proclividad de la prueba a los malentendidos encubre acaso una debilidad lógica fundamental. Jacob Cohen (1994), autor de un texto clásico sobre análisis de potencia estadística que es reputado como poco menos que la Biblia para determinar el tamaño de la muestra (Cohen 1988), ha clarificado las razones que tornan inválida la lógica de la NHST. Consideremos, propone Cohen, este razonamiento:

> Si la hipótesis nula fuera correcta, entonces los datos (D) no podrían ocurrir.
> Sin embargo, han ocurrido.
> Por lo tanto la hipótesis nula es falsa.

Si este fuera el razonamiento de la prueba de H_0, continúa Cohen, sería formalmente correcto. Se trataría de lo que Aristóteles denominaba *modus tollens*, consistente en negar el antecedente negando el consecuente. Pero ésta no es la forma en que la NHST se desenvuelve por cuanto su razonamiento se ha tornado probabilístico:

> Si la hipótesis nula fuera correcta, entonces estos datos son altamente improbables.
> Los datos han ocurrido.
> Por lo tanto, la hipótesis nula es altamente improbable.

La incorrección lógica de la NHST tal vez no se perciba a primera vista, pero Pollard y Richardson (1987) lo demostraron a través de un ejemplo:

> Si una persona es norteamericana, entonces probablemente no es miembro del congreso.
> La persona es miembro del congreso.
> Por lo tanto, probablemente no es norteamericano.

Esto es formalmente idéntico a:

---

[8] Véase por ejemplo http://en.wikipedia.org/wiki/P-value#Misunderstandings. Visitado en junio de 2011.

> Si H_0 es verdad entonces este resultado (la significancia estadística) probablemente no ocurriría.
> El resultado ha ocurrido.
> Entonces H_0 es probablemente no verdadera y por ende formalmente inválida.

Formulaciones basadas en este modelo, atravesado por lo que los detractores llaman "la ilusión de alcanzar la improbabilidad" se repiten en texto tras texto en todas las disciplinas humanas y en algunas ciencias llamadas exactas, como si se creyera que la aplicación cuenta con el respaldo de una lógica que de ser como se la imagina no se sostendría siquiera en el ámbito abstracto (Cohen 1994: 998; ver no obstante Cortina y Dunlap 1997: 166). En la práctica de la NHST el ejemplo más acabado de esta falacia sale a relucir cuando se afirma (por ejemplo) que dado un rechazo de la HN al 1%, la probabilidad de que la hipótesis alternativa sea verdad es del 99%.

Si la especificación de Cohen sobre las formas lógicas luce forzada o inverosímil no hay más que echar una mirada a la bibliografía del género. Más recientemente Krämer y Gigerenzer volvieron a tratar el asunto en su estudio sobre el uso y abuso de las probabilidades condicionales, descubriendo que un gran número de manuales de estadística cometían exactamente el mismo error, bien conocido desde los tiempos de *How to lie with statistics* de Darrell Huff [1913-2001], el libro de estadísticas más vendido de la segunda mitad del siglo XX (Huff 1954: 75 *& passim*; Krämer y Gigerenzer 2005: 225).[9] Los críticos también documentaron que las mismas pautas de razonamiento aparecen en ámbitos que no necesariamente se presentan como pruebas de hipótesis:

> En 1221 asesinatos de mujeres entre 1984 y 1988, 44% fueron asesinadas por sus esposos o amantes, 18% por otros parientes y otro 18% por amigos o conocidos. Sólo 14% fueron asesinadas por extraños. ¿Prueba esto que la probabilidad de que un asesinato por el encuentro con un esposo sea mayor a la probabilidad de un asesinato debido al encuentro con un extraño, esto es, que el matrimonio favorezca el asesinato? (Krämer y Gigerenzer 2005: 226).

Hay otro estilo de razonamiento lógico falaz que suele encontrarse en la literatura de la NHST. Con frecuencia los promotores y usuarios del método presuponen que el rechazo de la HN involucra la aceptación automática de cualquier teoría que implique su falsedad. Pero la línea de inferencia que va desde "la HN es falsa" a "la teoría sustentada es por ende verdadera" involucra de plano la falacia lógica de afirmación del consecuente: $P \rightarrow Q$; Q, $\therefore P$. En la práctica de la prueba estadística la hipótesis alternativa recién podría probarse verdadera si se lograra articular un diseño de investigación que excluya toda otra explicación y que garantice la singularidad de H_A en tanto única

[9] Más adelante (pág. 56), podrá encontrarse una lista ampliada de los manuales que en su tratamiento de la NHST malinterpretan la lógica de la probabilidad condicional, incluyendo unos cuantos textos de antropología y arqueología.

forma de negar la posibilidad de obtener resultados nulos a través del azar en el caso en cuestión. Quien elaboró esta situación con más claridad de lo que yo he podido hacerlo es sin duda el psicólogo comunicacional Andrew F. Hayes:

> Que el mecanismo de "no-azar" que produce un resultado de investigación (es decir, uno que produce un efecto distinto de cero) sea el que propone el investigador sólo se puede determinar por medio de un buen diseño investigativo, eliminando explicaciones rivales a través del control apropiado de fuentes potenciales de confusión y mediante una traducción convincente de la cuestión sustantiva en la hipótesis empírica (Hayes 1998: 203).

En los 32 años que corren entre David Lykken (1968: 152) y Raymond Nickerson (2000: 254) numerosos autores han encontrado que la afirmación del consecuente plaga la inmensa mayoría de los reportes experimentales en psicología y que en ocasiones es utilizada con bastante éxito, aunque no por ello deja de ser una falacia (Cohen 1994: 999; Meehl 1990a; 1990b). Paul Meehl lo expone sin componendas, aportando una distinción conceptual que acaso agrava la crudeza del error:

> Es importante tener en cuenta la distinción fundamental entre una teoría sustantiva *T* y una hipótesis estadística *H*. Los manuales y los instructores de estadística no subrayan la distinción y algunos ni siquiera la mencionan en una sola frase de advertencia. Esta grave omisión pedagógica resulta en la tendencia de los estudiantes a fusionar [*conflate*] la refutación de H_0 con la prueba de la contranula $-H_0$, que de inmediato se confunde en sus mentes con "probar *T*". Esta tentadora línea de pensamiento combina de este modo un error en el razonamiento estrictamente estadístico con un error adicional en el razonamiento lógico, afirmando el consecuente en la inferencia empírica. […] El punto puramente lógico aquí es, como dije antes, que la inferencia empírica del hecho a la teoría [la inferencia inductiva, en suma] es una figura inválida del silogismo implicativo (1990a: 116).

Es verdad que los elementos físicos o conceptuales de una configuración experimental no son entidades sintácticas que mapean miembro a miembro sobre los términos de (por ejemplo) el cálculo de predicados: algunas falacias en un sistema lógico pueden interpretarse como implicaciones aceptables en un sistema que se funde en otras premisas. Pero que una falacia tan universalmente reconocida como tal a través de los diversos sistemas aflore en un papel tan preeminente en la arquitectura de la prueba estadística merecería alguna explicación si es que ésta pretende calificar como modelo de inferencia de corte clásico, tal como es efectivamente el caso.

Situaciones paradójicas parecidas a éstas han sido debatidas por György Pólya (1954a; 1954b) en dos volúmenes antológicos sobre la inducción y la analogía en matemáticas; sin embargo Pólya, por desdicha, nunca se interesó ni en la estadística en general ni en la NHST en particular. Como sea, delego al lector el ejercicio de buscar manifestaciones de esta falacia en las pruebas de hipótesis de su disciplina favorita, prediciendo que no volverá con las manos vacías. El problema se agrava aca-

so en los textos en lengua castellana: en idioma inglés se mantiene la diferencia entre un conjunto de operaciones estadísticas [*test*] y una prueba lógica o matemática de carácter formal [*proof*]; en nuestra lengua sólo se habla de "prueba", no siendo infrecuente que se trate el despliegue de una mera rutina de cálculo en una investigación periférica con la solemnidad que se imagina propia de las demostraciones de teoremas.

Por más que el deslinde del aparato lógico subyacente articulado por Cohen, Lykken, Nickerson, Krämer y Gigerenzer no describa con entera justicia las formas de inferencia vigentes en la totalidad de la literatura profesional, en el ámbito académico se percibe una cierta tendencia a aceptar sin crítica que la elaboración fisheriana, en particular, constituye una mecanización o automatización aceptable de la inferencia inductiva. Por más que las tablas de todos los manuales los den por sentados y los programas estadísticos de computadora incluyan hoy los mecanismos de cálculo correspondientes los fundamentos lógicos de la prueba estadística están hace décadas en tela de juicio, como en breve seguiremos comprobando. Los problemas lógicos, empero, no acaban allí.

La idea ya presente en *The Design of Experiments* (1971: 4) que aseveraba que los métodos de la inducción podían sustentarse de manera "perfectamente rigurosa e inequívoca" fue llevada a su extremo en obras posteriores, en las que Fisher afirmaba que sus métodos constituían una especificación relativamente exhaustiva del proceso inductivo, "tan satisfactoria y completa, por lo menos [sic], como la que se dio tradicionalmente a los procesos deductivos" (Fisher 1955: 74).[10] Cortando el nudo gordiano del problema de la inducción, la estadística, recién llegada a la escena científica, se preciaba de haber logrado lo que una lógica milenaria nunca había sido capaz de alcanzar. Más todavía, no era preciso abocarse a una demostración engorrosa ni comprender exhaustivamente el problema pues éste, cualquiera fuese, acababa resolviéndose por sí mismo aun cuando (como nos señalara Leonard Savage [1978: 448]) en los razonamientos de Fisher no se pueda encontrar ni una sola *prueba* de esa alegación en el sentido lógico o matemático de la palabra.

Vale la pena citar en extenso la crítica de David Bakan:

---

[10] Sobre la automatización de los procesos *deductivos* a través del cálculo de predicados de primer orden, véase mi vieja tesis de doctorado *Antropología y programación lógica* (Reynoso 1991). Contrariamente a lo que se cree (y aunque existe un repertorio inmenso de problemas de tratabilidad), los sistemas formales basados en la lógica clásica no se encuentran afectados por el célebre "problema de Gödel", el cual sólo atañe a los sistemas formales matemáticos basados en funciones recursivas que incluyen una porción importante de ciertas aritméticas, la de Peano entre ellas. El propio Kurt Gödel (1930) probó la consistencia y completitud del cálculo de predicados. El problema no afecta a toda la matemática; ni siquiera a toda la aritmética, a decir verdad. La aritmética de Mojżesz Presburger, por ejemplo, que sólo difiere de la de Peano por carecer de la operación de producto, es decidible, consistente y completa (Presburger 1929; Cooper 1972; Feferman 2006: 435).

> Los psicólogos interpretaron por cierto de este modo los procedimientos asociados con la prueba *t*, la prueba *F* y demás. En vez de involucrarse en la inferencia ellos mismos, no tuvieron sino que "correr el test" para realizar inferencias, dado que, según parecía, las pruebas estadísticas eran análogos analíticos de la inferencia inductiva. […] ¡[S]e podía, aparentemente, "operacionalizar" el proceso inferencial simplemente reportando los detalles del análisis estadístico! […] La contingencia de que el experimentador tomaba decisiones sobre el nivel de significancia se manejó de dos maneras. La primera, sumándose a una especie de consenso social que decía que 5% era bueno y 1% era mejor. La segunda, […] no tomando ninguna decisión sobre el nivel de significancia, sino sólo reportando el valor *p* como una "resultante" y una "medición" del nivel de confianza presumiblemente objetivas (Bakan 1966: 430).

Tradicionalmente se piensa que el de Fisher es un modelo de inferencia inductiva mientras que el de Neyman-Pearson encarna más bien una heurística operacional basada en indicadores, pero la cosa no es tan simple. Los desarrollos respectivos no han sido ni puros, ni transparentes ni inambiguos y han ido cambiando con el tiempo (cf. Mayo 1992; Lehmann y Romano 2005; Liese y Miescke 2008). Es evidente, sin embargo, que el modelo de Neyman-Pearson no busca razonar inductivamente ni proporcionar una medida de la evidencia, que es lo que la idea de "prueba de hipótesis" puede llevar a creer. Lo que estos autores tratan de hacer, con un alcance mucho más modesto, es articular alguna táctica para limitar el número de errores a través de una cantidad de experimentos. La imposibilidad de establecer el valor de verdad de una hipótesis mediante la inferencia probabilística simplemente se da por sentada:

> [N]ingún test basado en una teoría de la probabilidad puede por sí mismo proporcionar una evidencia valiosa sobre la verdad o falsedad de una hipótesis. Pero podemos contemplar el propósito de las pruebas desde otro punto de vista. Sin esperar saber si cada hipótesis por separado es verdadera o falsa, podemos buscar reglas que gobiernen nuestra conducta con respecto a ella, de tal modo que ateniéndonos a ellas podríamos estar seguros que, en el largo plazo, no estaremos equivocados muy a menudo. […] Tal regla no nos dice nada sobre si en un caso particular H es verdad cuando $x \leq x_0$ o si es falso cuando $x > x_0$. Pero a menudo se puede probar que si nos comportamos de acuerdo con esta regla, en el largo plazo rechazaremos H cuando es verdad no más de, digamos, una de cada cien veces, y que por añadidura tendremos evidencia de que rechazaremos H suficientemente a menudo cuando ella sea falsa (Neyman y Pearson 1933: 291).

Independientemente de la utilidad de esta regla de conducta, una vez que los propios autores ratifican que ninguna prueba basada en teoría de la probabilidad puede expedirse sobre el valor de verdad de una teoría, se torna dificultoso entender por qué se sigue llamando *prueba de hipótesis* al test de Neyman-Pearson. De hecho, en cualquiera de sus versiones el contenido de las hipótesis no interviene en el cálculo, dado que lo único que puede hacerse es establecer la probabilidad de determinados valores muestrales (con referencia a ciertas estimaciones teóricas) y proceder de un modo

o de otro según el valor que esa probabilidad alcance en un sistema de cálculo atestado de supuestos.

Sólo en tiempos recientes se ha caído en la cuenta que tanto la prueba de significancia como (en mayor grado) la prueba de hipótesis pertenecen más bien a la teoría de la decisión sobre cursos de acción (con sus rutinas de aceptación y rechazo) y ya no a la lógica propiamente dicha (Rivadulla 1991; Erwin 1998). La teoría de la decisión posee también objetivos que difieren formalmente de los de la inferencia estadística, existiendo conflictos matemáticos complicados pero bien conocidos en los principios relevantes de cada una (Tukey 1960a; Wilkinson 1977: 119); aquélla puede proporcionar orientaciones muy elaboradas pero no al punto de constituir una axiomática. Por más que la literatura formal sobre la decisión estadística promueva uno de los estilos simbólicos de especificación más asertivos e inmodestos que se conoce (v. gr. Liese y Miescke 2008), el prestigioso Henry Kyburg, Jr [1928-2007], uno de los mayores especialistas en lógica de la inferencia estadística, nos invitaba a poner los pies sobre la tierra cuando decía:

> Hablar acerca de aceptar o rechazar hipótesis […] es *prima facie* hablar epistemológicamente; y sin embargo en la literatura estadística aceptar la hipótesis de que el parámetro μ es menor que μ^* es meramente una forma fantasiosa e indirecta de decir que Mr Doe no debería aceptar más que $ 36,52 por una bolsa de tornillos (Kyburg 1971: 82-83).

El propio Fisher había elaborado una crítica parecida de la interpretación conductual cuando declaró que no había que confundir "los procesos apropiados para sacar determinadas conclusiones con aquellos que buscan más bien, por así decirlo, acelerar la producción o ahorrar dinero" (Fisher 1955: 70). Nadie prestó la debida atención a esas observaciones coloquiales, pronunciadas en el vértigo convulso de las escaramuzas entre Fisher y Neyman y sus partidarios respectivos. Por eso es turbador que una teoría que presume expedirse sobre lógica o metamatemática haya llegado tan tarde, tan tortuosamente y tan sobrecargada de imprecisiones interpretativas a dicha percepción.

Lo importante de todo esto (como ha sido ampliamente reconocido por Fisher [1956] en torno de la prueba de significancia y por una cohorte de especialistas en el tema a propósito de las enmiendas introducidas al respecto) es que el logro fundamental de la inferencia inductiva, esto es, *el salto desde las propiedades cuantitativas de una muestra a los valores de verdad de hipótesis referidas a una población*, es un dilema científico mayor que no ha podido hasta hoy implementarse mecánicamente ni validarse con el rigor requerido.

La discusión estadística del asunto (que no he de tratar aquí) es de una complejidad inenarrable y ha discurrido por carriles ajenos a las conceptualizaciones de la lógica, a la filosofía de la ciencia o incluso a la pelea alrededor de la NHST; pero sus conclusiones lapidarias no dejan lugar a dudas. Aunque no es posible resumir todos los razonamientos involucrados en una fórmula simple, lo e-

sencial del problema finca en el hecho de que la probabilidad inferencial derivada de datos observacionales es inherentemente no-coherente, en el sentido de que sus implicaciones no pueden representarse mediante una sola distribución de probabilidad en el espacio de parámetros (Buehler y Federsen 1963; Dempster 1963a; 1963b; Wilkinson 1977). En sus últimos años Fisher intentó compensar este fracaso desarrollando la idea de una inferencia fiducial, pero falló de nuevo:

> Inicialmente incapaz de justificar su intuición sobre el pasaje desde una aserción de probabilidad sobre una estadística (condicional a un parámetro) hacia una aserción de probabilidad sobre un parámetro (condicional a una estadística) Fisher pensó en 1956 que él había descubierto finalmente la salida de este enigma con su concepto de *subconjunto reconocible*. Pero a pesar de la autoridad con la que Fisher afirmó su nueva posición en su último libro, *Statistical Methods and Scientific Inference* [1956], el argumento crucial para la relevancia de este concepto se fundaba en otra intuición, una que, ahora claramente especificada, Buehler y Feddersen [1963] demostraron más tarde que era falsa (Zabell 1992: 382).

No se ha experimentado lo suficiente en la vinculación entre los argumentos tardíos sobre la inferencia inductiva y los desarrollos tempranos de la prueba de significancia; por eso es que el fracaso de Fisher no ha significado ninguna sorpresa para sus críticos y biógrafos: a la fecha todavía no existe una teoría integrada de este tipo. Sandy L. Zabell, matemático de la Northwestern University, lo expresa categóricamente:

> [E]l intento de Fisher por timonear un rumbo entre el Escila de los métodos incondicionales conductistas [de Neyman y Pearson] que desalientan todo intento de "inferencia" y el Caribdis del subjetivismo [bayesiano] en ciencia, estaba fundado en preocupaciones importantes; esta falla personal en el hallazgo de una solución satisfactoria sólo significa que el problema permanece sin resolver y no que ese problema no exista (Zabell *loc. cit.*)

No por nada se atribuía al filósofo de la ciencia Morris Cohen haber dicho que "[T]odos los textos lógicos se dividen en dos partes. En la primera, sobre la lógica deductiva, se explican las falacias; en la segunda, sobre la lógica inductiva, se las comete" (según Meehl 1990a: 110).

5. El surgimiento de la crítica

> Probablemente desde hace treinta o más años muchos han sentido que las pruebas de significancia se han sobre-enfatizado y utilizado mal con suma frecuencia, y que se debería poner más énfasis en la estimación y la predicción. Mientras que ese cambio de énfasis parece estar ocurriendo, por ejemplo en estadísticas médicas, el uso continuado y muy extensivo de las pruebas de significancia es por un lado alarmante, pero por el otro testimonia una respuesta, aunque sea imperfecta, a una necesidad ampliamente sentida.
>
> Sir David Roxbee Cox, 1986

En materia de críticas puramente estadísticas se cree que el primer cuestionamiento de la prueba de significancia fisheriana ha sido el del educador Ralph Winfred Tyler (1931), especialista en evaluación de currícula; dado que se refiere más bien a la necesidad de distinguir entre significancia estadística y significación empírica revisaremos ese documento más adelante, en un apartado específico (cf. pág. 38).

Poco después pero todavía antes que la NHST se afincara, el físico y estadístico Joseph Berkson (1938) registró algunas dificultades de interpretación que se encuentran en una aplicación de la prueba de χ^2 que repercuten con un efecto deletéreo sobre la prueba de significancia. Los trabajos de Berkson sobre pormenores estadísticos fueron bien conocidos, en especial su modelo de errores para el cálculo de regresión que contradecía al modelo clásico y su propuesta a favor de usar la más versátil distribución logística en lugar de la distribución normal. Promovida por Berkson, esta distribución simétrica pero de cola pesada (que no suele requerir aproximación numérica y que se puede resolver analíticamente) ha conocido una multiplicidad de aplicaciones.[11] Pero no todas han sido flores en la trayectoria de este autor; él mismo ganó celebridad al implementar sus metodolo-

[11] Se la utiliza rutinariamente en biología para describir el crecimiento de especies en competencia, en mercadotecnia para predecir la difusión de nuevos productos, en epidemiología para modelar la difusión de enfermedades, en tecnología para planificar la sucesión de innovaciones, en psicología cognitiva para describir el aprendizaje y en ecología para comprender la producción agrícola y la sustitución de nuevas fuentes energéticas (Balakrishnan 1992; Evans, Hastings y Peacock 1993: 98-101; Johnson, Kotz y Balakrishnan 1994: 113-154). Yo mismo recurro con frecuencia a la ecuación fundamental que genera una versión discreta de la ley cada vez que enseño los rudimentos de la dinámica no lineal (Reynoso 2003). Aunque a esta altura del siglo se encuentran por debajo de lo que es el caso con la LP (cf. más adelante, pág. 82) las credenciales empíricas de la ley logística son impresionantes, órdenes de magnitud por encima de lo que es el caso de la ley normal; pero no he sabido que se la haya implementado consistentemente en prueba estadística de hipótesis ni siquiera en los campos en que su adecuación es manifiesta.

gías personalizadas de prueba de hipótesis para probar que "la evidencia, tomada como un todo, no establece sobre una base científica razonable que fumar cigarrillos provoque cáncer de pulmón".[12]

Tras un interregno de un par de décadas, en los 50s el zoólogo Lancelot Hogben (1957) arremetió contra la NHST a lo largo de todo un libro luego que Lyle V. Jones (1950) recomendara sustituirla por estimaciones de punto e intervalos de confianza, dos de las alternativas más populares desde entonces. En los 70s dos sociólogos, Denton Morrison y Ramon Henkel (1970), publicaron una compilación sobre *The significance test controversy*, con colaboradores de la talla de William Rozeboom (1960), Paul Meehl (1967), David Bakan (1966) y el estudioso de la conducta antisocial David Lykken (1968). Veinte años más tarde Paul Meehl (1990a) lamentaría que un libro de semejante nivel de excelencia [*epoch-making, path-breaking*] no fuese más conocido; a más de cuarenta de su publicación (es decir, hoy) la mayor parte de los ensayos circula intensamente en la Web, como puede comprobarse en el hipertexto de nuestra bibliografía.

Más de una generación después de *Controversy…* Lisa Harlow, Stanley Mulaik y James Steiger (1997) convocaron a especialistas para que imaginaran como sería la vida científica si las pruebas de significancia de golpe dejaran de existir. Entretanto las críticas se habían multiplicado más allá de toda medida (Rozeboom 1960; Bakan 1966; Coats 1970; Morrison y Henkel 1970; Spielman 1973; Guttman 1977; 1985; Carver 1978; Shaver 1985a; 1985b; 1993; Cox 1986; Rosnell y Rosenthal 1989; Slakter y Suzuki-Slakter 1991; Hubbard y Armstrong 1992; Cohen 1994; Dar, Serlin y Omer 1994; Falk y Greenbaum 1995; Schmidt 1996). Tras la segunda gran polémica el movimiento crítico ha seguido en efervescencia hasta nuestros días (Schmidt y Hunter 1997; Abelson 1997; Hubbard 1997; 2005; Gill 1999; Anderson 2000; Krueger 2001; Nickerson 2000; Kline 2004; Armstrong 2007a; 2007b; Guthery 2008; McCloskey y Ziliak 2008; Marewski y Olssohn 2009; Sestini y Rossi 2009; Bedeian, Taylor y Miller 2010; Loftus 2010; Monterde, Frías-Navarro y Pascual-Llovell 2010).

Al lado de la literatura impresa hay por lo menos seis portales de la Web con punteros a bibliografía masiva sobre la discusión. Estos son los de Mark Nester (1997), Bill Thompson (2001), el Southwest Fisheries Science Center (2010), el del Equipe de Raisonnement Induction Statistique (ERIS, bayesiano fundamentalista), el del Management Junk Science (2011) y el que alberga el libro que se está leyendo (Reynoso 2011b). Ante la comprobación reiterada de sesgos editoriales que promueven la censura de todos los diseños de investigación que no estén alineados con la metodología fa-

[12] Véase http://tobaccodocuments.org/profiles/people/berkson_joseph.html. Visitado en julio de 2011.

vorecida por defecto, hace poco se ha sumado un *Journal* digital específicamente orientado a respaldar la verificación de hipótesis nulas.[13]

En ciencias de la educación varios autores han llamado la atención sobre las precauciones que hay que tomar para realizar inferencias a partir de la prueba de significancia estadística pero con escasa repercusión en la comunidad de educadores que desarrollan diseños de investigación (Morrison & Henkel 1970; Shaver, 1985a; 1985b; Slakter, Yu, & Suzuki-Slakter 1991; Stevens 1968; Thompson 1999a; 199b; Tyler 1931). La situación en psicología es aproximadamente la misma (Dar, Serlin y Omer 1994; Gigerenzer 1987), igual que lo es en las disciplinas más variadas, antropología y arqueología inclusive, aunque las advertencias han sido más tibias (Cowgill 1977; Hole 1980; Chibnik 1985). En psicología, neurociencia y educación se consagraron al problema números especiales de *Psychological Science* (8[1]), *Journal of Experimental Education* (61[4]), *Behavioral and Brain Science* (21[2]) y *Research in the Schools* (5[2]). En antropología sociocultural, mientras tanto, el primero y hasta ahora único trabajo íntegramente dedicado al tema es, tal parece, el que se está leyendo.[14]

Las polémicas en torno de la NHST han venido en oleadas o borbotones de fuerza creciente más o menos cada veinte años de manera consistente con los principios de corrección generacional (Levin 1998: 43).[15] La literatura ha alcanzado tal grado de sedimentación que ha sido necesario publicar *surveys* que se ocupan nada más que de reseñarlas (Nickerson 2000; Kline 2004); se prevé cercano el día en que los *surveys* mismos sean objeto de catalogación sistemática en una literatura prospectiva de tercer orden. A lo largo de décadas no pocos especialistas han propuesto el abandono total de las pruebas de significancia (Rozeboom 1960; Bakan 1966; Carver 1978; Cohen 1994; Schmidt y Hunter 1997; Krueger 2001; Nickerson 2000; Armstrong 2007; Guthery 2008).

El hecho es que al margen de las que ya señalamos se reconocen varias clases de problemas en la aplicación de la prueba. Algunas de ellas tienen que ver con una alarmante serie de inconsistencias lógicas mientras que otras son más bien de naturaleza interpretativa. Los autores indecisos o equidistantes son pocos y las opiniones extremadamente críticas han sido tan crispadas como cuantiosas.

---

[13] Véase http://www.jasnh.com/. Visitado en julio de 2011.

[14] Los ideólogos de la distribución normal sostienen que la NHST (junto con su cohorte de supuestos de normalidad) todavía se encuentra en pie en física y en otras ciencias duras. Deirdre McCloskey y Stephen Ziliak (2007) demostraron acabadamente que la frecuencia de su uso no es, valga la paradoja, estadísticamente significativa. En el inmenso campo transdisciplinario de la dinámica no lineal que se estableció en las últimas décadas puede afirmarse taxativamente que no hay sombras de distribuciones normales ni, por supuesto, de prueba estadística de hipótesis.

[15] "Se siente casi como si cada generación librara la 'guerra de las estadísticas' cada vez de nuevo, con reformas de las políticas de publicación y fuerzas de tareas consagradas a impedir los usos automatizados de las pruebas de significancia, similares a recetas, que han sido reprobados hace tanto tiempo" (Mayo 2006: 325).

La NHST ha sido calificada como “una ordalía sin sentido de computaciones pedantes” (Stevens 1960: 276); “una variedad de falta total de sentido en el desarrollo de la investigación” (Bakan 1966: 436); “un método basado en un malentendido fundamental sobre la naturaleza de la inferencia racional, rara vez apropiado a los fines de la indagación científica” (Rozeboom 1960: 417); “una prueba inadecuada en sus propios términos” (Spielman 1973: 202); “un ejercicio de futilidad académica” (Chew 1977); “una forma corrupta de método científico que es, en el mejor de los casos, de importancia científica trivial” (Carver 1978: 378); “un error terrible, una estrategia pobre y básicamente infundada, y una de las peores cosas que jamás han sucedido en la historia de la disciplina” (Meehl 1978: 817); un procedimiento “o bien profundamente fallido o mal utilizado por los investigadores” (Serley y Lapsley 1993); una prueba que “oscurece verdaderamente las regularidades y procesos subyacentes en los estudios individuales y en la literatura de investigación, llevando a conclusiones sistemáticamente erróneas” (Schmidt 1992); una técnica que ha “retardado sistemáticamente el desarrollo del conocimiento acumulativo” (Schmidt 1996) y que constituye “una metodología incoherente de inferencia estadística que es perjudicial para el progreso de la psicología como ciencia” (Marewski y Olssohn 2009: 49); “una estrategia seudocientífica que proporciona una falsa sensación de objetividad y rigor” (Anderson 2000: 921); “una conducta ritualística que obstaculiza la acumulación del conocimiento” (Guthery 2008: 1872); “una idea excepcionalmente mala, […] una parte de la estadística matemática terriblemente equivocada” (Ziliak y McCloskey 2009: 2302); “una mitología que torna la inmensa mayoría de los trabajos de investigación publicados en un cuerpo de hallazgos infundados, no-científicos, característicamente incorrectos” (Kmetz 2011); “un desastre […], una prueba que lisa y llanamente no funciona” (Hunter 1997), “profundamente fallida y ampliamente mal usada” (Gill 1999), “exhaustivamente desacreditada […] lógica y conceptualmente” (Schmidt y Hunter 1997), e “innecesaria aún cuando se la ejecute y se la interprete correctamente” (Armstrong 2007b). La prueba estadística –dice por último David Rindskopf (1998) ilustrando el tono irritado de la polémica– “no es completamente estúpida, pero la estadística bayesiana es mejor”.

Por más que se pueda señalar un alto porcentaje de errores en los dichos de la vertiente opositora, que algunos críticos de excelencia cometieran desatinos colosales en otros rubros científicos y que gran número de principiantes arremetieran contra ella sin dar la talla, no todos los críticos de la NHST han sido operarios empíricos de segundo orden, ni han sido siempre bayesianos que promueven doctrinas incompatibles, ni personajes que desconocen la filosofía científica o las técnicas estadísticas, ni gente interesada en mantener viva una polémica superflua e inconcluyente como se ha llegado a insinuar tanto entre los partidarios (Giere 1972; Cortina y Dunlap 1997; Mulaik, Raju y Harshman 1997; Chow 1998: 170; Hoover y Siegler 2008) como entre los enemigos de la prueba

(Meehl 1990b: 222-223; Nickerson 2000: 241). Tras haber leído con aprecio las críticas de Meehl (1967) y de Lykken (1968), nadie menos que Imre Lakatos llegó a preguntarse

> [s]i la función de las técnicas estadísticas en las ciencias sociales no es primariamente proporcionar una maquinaria para producir corroboraciones espurias [*phony*] y por tanto una apariencia de "progreso científico" allí donde, de hecho, no hay nada más que un incremento en la basura seudo-intelectual. […] Me parece que buena parte de la teorización condenada por Meehl y Lykken puede ser *ad hoc*$_3$. Por lo tanto, la metodología de programas de investigación puede ayudarnos a desarrollar leyes para contener esta polución intelectual capaz de destruir nuestro ambiente cultural aun antes de que la polución de la industria y el tráfico destruyan nuestro ambiente físico (Lakatos 1978: 88-89, nota §4).[16]

Tampoco la postura de William Kruskal deja lugar a dudas, aunque desborda el blanco de la NHST como tal:

> Existen algunos otros roles que las actividades llamadas "estadísticas" pueden, desafortunadamente, jugar. Dos de esos roles mal concebidos son (1) santificar u otorgar un sello de aprobación (se escucha hablar, por ejemplo, de consejeros de tesis o editores de revistas que insisten en ciertos procedimientos estadísticos formales, sean ellos o no apropiados); (2) impresionar, ofuscar o mixtificar (por ejemplo, algunos *papers* de investigación en ciencias sociales contienen masas de fórmulas no digeridas [o pruebas de significancia] que no sirven a ningún propósito excepto el de indicar lo brillante que es el autor) (Kruskal 1968b: 209).

Se ha ido generando cierto consenso en torno a la idea de que la NHST ya no se encuentra a la altura de los tiempos y que configura un "*folkways* estadístico de un pasado más primitivo" (Rozeboom 1960), un "procedimiento estadístico pasado de moda" (Slakter y Suzuki-Slakter 1991) cuyo uso "todavía muy extendido […] es alarmante" (Cox 1986). Medio siglo atrás William Rozeboom evaluaba el síndrome de anacronismo inherente a los métodos de prueba de hipótesis de esta manera:

> [D]ado que las conductas que alguna vez se ejercen tienden a cristalizar en hábitos y a veces en tradiciones, no debería sorprender que se encuentre que los rituales tribales para el procesamiento de datos que se ven a lo largo de cursos de grado sobre métodos experimentales contengan elementos que se justifican más por la costumbre que por la razón (Rozeboom 1960: 416).

---

[16] Importantes filósofos de la ciencia se ocuparon de la confrontación entre las posturas de Fisher y Neyman-Pearson o de los factores filosóficos implicados. Deborah Mayo y Aris Spano (2006) registran los nombres de John Earman, James Fetzer, Ronald Giere, Donald A. Gillies, Clark Glymour, Ian Hacking, Paul Horwich, Colin Howson, Henry Kyburg Jr, Isaac Levi, C. S. Peirce, Roger Rosenkrantz, Wesley Salmon, Teddy Seidenfeld, Stephen Spielman y P. Urbach. Insólitamente no hay casi discusiones filosóficas de los elementos más básicos, tales como la problemática del muestreo, la incidencia de las distribuciones empíricamente dadas sobre el método de prueba, las políticas de descarte de los datos que se salen de norma [*outliers*] o las premisas probabilísticas a veces extravagantes de la distribución normal.

Los bayesianos James Berger y Donald Berry, treinta años más tarde, lo expresaron así:

> El análisis estadístico juega un papel central en la indagación científica. La adopción de los métodos estadísticos actuales ha conducido a mejoras enormes en la comprensión de la evidencia científica. Pero el uso común de la estadística parece haberse fosilizado, principalmente debido a la perspectiva de que la estadística estándar es la forma objetiva de analizar los datos (Berger y Berry 1988: 165).

Otros muchos autores han hecho palanca en la dimensión del tiempo en sus evaluaciones críticas. Cherry Ann Clark (1963: 466) alegaba que "se ha juzgado que [l]a hipótesis nula de no diferencia ya no es más una base coherente o fructífera de la investigación estadística"; Jacob Cohen (1994) creía que había sido "gravemente mal utilizada durante demasiado tiempo"; L. S. Shulman (1970: 389) proclamaba que "ha llegado el tiempo para que los investigadores en educación se saquen de encima el yugo de la prueba estadística de hipótesis"; el criminólogo Michael Maltz (1994) exploró "la significancia declinante de la significancia" y Lee Cronbach (1975: 124) terminaba certificando que "ha llegado la hora de exorcizar la hipótesis nula". Robert McGrath (1998), finalmente, argumentaba que "en razón de todo lo que se ha ganado a través de su uso, pienso que sería muy apropiado elogiar el brillo de la NHST; pero una vez dicho eso, quizá ya sea tiempo de darle sepultura".

Si tuviera que singularizar la crítica que mejor certifica la calidad del movimiento opositor a la NHST y a las ideas de significancia en la teoría y en la práctica, yo diría que ella se encuentra en los *surveys* escritos para la primera edición de la *Enciclopedia Internacional de Ciencias Sociales* por William Kruskal (1968a, 1968b), profesor de la Universidad de Chicago, ex-presidente de la Asociación Americana de Estadísticas, amigo personal de Fisher y celebridad de la algorítmica reticular.[17] Las ideas de Kruskal sobre la prueba de significancia no están unificadas en una sola obra pero su efecto acumulativo es demoledor. A tono con las blanduras diplomáticas del Pensamiento Débil posmoderno y con las políticas académicas de no mencionar literatura epigonal de más de

[17] William ("Bill") Kruskal [1919-2005] es coautor del célebre test no-paramétrico de Kruskal-Wallis, el cual se aplica cuando las variables de medición no satisfacen los supuestos de normalidad y homocedasticidad (homogeneidad de varianza) requeridos por el análisis de varianza de una vía. Este último puede producir valores absurdos de *p* cuando la población está incluso levemente alejada de la normalidad. A diferencia de lo que aseveran no pocos libros introductorios, la prueba no verifica la HN de que las poblaciones que se están evaluando tengan medias o medianas idénticas (McDonald 2009: 165-172). No se debe confundir a este Kruskal con sus hermanos menores Martin David Kruskal [1925-2006], descubridor de los solitones, y Joseph Kruskal [1928-2010], creador del algoritmo de grafos que lleva su nombre. Los logros estadísticos de Bill han sido colosales; él ha sido capaz de iluminar la naturaleza subyacente de ciertos problemas estadísticos con una claridad que no se encuentra cuando se presuponen determinados marcos de referencia, como lo es en nuestro caso la presunción de normalidad. Los trabajos de Kruskal sobre las retóricas del muestreo "representativo" corroboran en un portentoso *tour de force* (como demostraré más tarde) el parentesco entre las semánticas de la inferencia inductiva fisheriana y los supuestos dominantes en la hermenéutica antropológica del último tercio del siglo XX (cf. Kruskal y Mosteller 1979a; 1979b; 1979c; 1980).

(digamos) 5 años de añejamiento, los artículos de Kruskal fueron excluidos de las ediciones ulteriores de la *Enciclopedia*. Como sea, encuentro significativo que un pensador con tan amplias miras haya tomado partido al respecto en el sentido en que lo hizo, motivado por diagnosticar y exponer sin sensacionalismos la confusión interpretativa imperante en la estadística en general y en la NHST en particular.

David Rindskopf (1997: 319) ha dicho que "[d]ados los muchos ataques que ha recibido, la prueba de la hipótesis nula ya debería estar muerta". Como veremos más adelante es evidente que no lo está. Aunque en lo que va de la segunda década del siglo XXI su curva de crecimiento se ha moderado un poco, sigue siendo, para bien o para mal, la forma primaria mediante la cual los números que comprenden los datos de un experimento se traducen en conclusiones sobre las preguntas que el experimento se había propuesto abordar.

6. Errores de tipo I y II

Pese a que en el modelo original de la prueba de significancia nunca se trató semejante cosa, ninguna discusión sobre la NHST en su formato híbrido luciría completa sin una semblanza de los errores de Tipo I y II. Primero que nada hay que decir que los nombres que Neyman asignó a los errores han sabido causar impacto; aunque el tratamiento original de la idea no es particularmente estimulante y la tipología es escueta en extremo, la nomenclatura sugiere una sistematización de un orden exhaustivo, como si las problematicidades involucradas se comprendieran en su totalidad y estuvieran bajo control aunque el meollo de la cuestión se refiriese a errores.

En este campo, como en muchos otros, "[r]ara vez la terminología es superficial; puede ser clave en la adopción de nuevas ideas, o por lo menos puede señalar su llegada" (Kruskal y Stigler 1997: 87). Pero las dudas sobre la taxonomía de los errores surgieron a poco de empezar. El primer pensador de importancia que se opuso tajantemente a la tipología y a los errores de la Segunda Clase en particular no fue otro que Sir Ronald Fisher: "La frase 'errores de la Segunda Clase' –llegó a escribir– aunque en apariencia es sólo una pieza inofensiva de jerga técnica, es útil como indicadora del tipo de confusión mental con el que fue acuñada" (Fisher 1955: 73).

A despecho de las protestas de Fisher la tipología de Neyman-Pearson inspiró en un puñado de antropólogos de orientación interpretativa poco inclinados u hostiles a las matemáticas un conjunto de reflexiones de honda matriz estética. La instancia más radiante de esta práctica quizá haya sido este cuádruple encadenamiento de antítesis que Clifford Geertz concertó con exquisita prolijidad en *Conocimiento local*:

> En las formas de ciencia más estándar el truco consiste en manejarse entre lo que los estadísticos llaman errores del tipo uno y errores del tipo dos: aceptar hipótesis que sería más sensato rechazar y rechazar otras que sería más inteligente aceptar; aquí se trata de arreglárselas entre la sobreinterpretación y la subinterpretación, entre leer más en las cosas de lo que la razón permite y menos de lo que ella demanda (Geertz 1983: 16).

La caracterización geertziana es estilísticamente preciosa pero técnicamente disparatada, ya que la aceptación o el rechazo tal como los define el autor nada tienen que ver con las especificaciones respectivas de los tipos de error en la literatura estadística o con las formas lógicas implicadas. Por otro lado, en dichas especificaciones (que en la versión geertziana han quedado asombrosamente al revés) no se habla de "hipótesis" sin cualificar (de esas que pocas palabras más tarde equivalen a "cosas" a ser leídas) sino concretamente de la hipótesis nula (o lo que haga las veces de tal en el marco de Neyman-Pearson), una entidad que Geertz nunca menciona pero que es la única sobre la cual a la prueba de hipótesis le cuadra expedirse.

Tampoco parece entender Geertz que ni el investigador está condenado a equivocarse ni le está permitido afincarse premeditadamente en posiciones equidistantes o intermedias *entre* ambas clases de errores: simplemente incurre en uno o bien en el otro, si es que se da el caso (que se espera sea excepcional) de que perpetre alguno. Contradiciendo una vez más lo que Geertz pretende enseñarnos de estadística, en la disciplina de origen el *truco* no consiste en manejarse entre una y otra clase de errores sino en evitar cometer cualquiera de las dos.

En el procedimiento de prueba estadística a la manera fisheriana, por añadidura, no se procura "aceptar" ninguna hipótesis de manera directa, pues lo más afortunado que puede sucederle a uno es tener éxito en el rechazo de la H_0 en cuestión. Como quiera que lo considerara Neyman, Fisher tenía razón cuando alegaba que "es una falacia, tan bien conocida como para ser un ejemplo *estándar*, concluir a partir de una prueba de significancia que la hipótesis nula queda en ella establecida; a lo sumo podría decirse que se la confirma o fortalece" (Fisher 1955: 73).

El matiz implicado en el alegato de Fisher, en fin, estropea el efecto de simetría que la retórica geertziana requiere para mantener su musicalidad y el equilibrio de sus oposiciones, de cuyas asignaciones equivocadas ninguno de los epígonos que la reprodujeron o de los *reviewers* que la comentaron pareció darse cuenta (cf. Petersen 1983; Lieberson 1984; Catt 1984; Shankman 1985; Foster 1985; Curry 1985; Wegener 1985; Keesing 1985; Piot y Scult 1985; Ostrow 1990; Rinehart 1998: 14). Cuesta un poco encontrar explicación a un yerro discursivo que hace que en una frase tan palpablemente trabajada las definiciones acaben siendo más infieles a la realidad de lo que lo serían si estuviesen meramente invertidas; excluidas la mala fe y la simple ignorancia, mi conjetura más benigna sobre las razones del equívoco remite a la estética de la enunciación: si el orden de atribución de la tipología de errores hubiera sido el correcto y si se hubiera sido fiel a las asimetrías de origen, la cadencia prosódica de la frase quizá no sería tan perfecta y algo de lo que la ha hecho memorable se habría perdido.

Llama asimismo la atención que en el mismo libro en el que publica una visión antropológica del derecho y la lógica legal atiborrada de distinciones que presumen lucidez, Geertz confunda el hecho de no poder rechazar la HN con la acción de aceptarla. Lo más llamativo del caso es que desde los años tempranos de la discusión sobre la prueba de hipótesis se ha vuelto común ilustrar esa diferencia, precisamente, mediante una analogía jurídica. Un ejemplo a cuento es éste que sigue, publicado en *American Anti*quity un año antes que se editara *Conocimiento Local*:

> La distinción es análoga a la que existe entre los conceptos de "inocente" y "no culpable" en la jurisprudencia estadounidense. Una declaración de "no culpable" sancionada por el jurado está muy lejos de declarar al imputado "inocente". Simplemente dice que la evidencia ha sido insuficiente para convencer al jurado más allá de algún criterio arbitrario de duda razonable

> de condenar al acusado. Parecidamente, en estadística "no rechazar" está muy lejos de "aceptar" o "sustentar". Simplemente dice que la "evidencia" (en este caso, los datos) han sido insuficientemente convincentes más allá de algún estándar arbitrario de duda razonable (en este caso, el nivel de significancia elegido para la prueba) de "condenar" (es decir, rechazar) al "acusado" (H_0) (Scheps 1982: 839).

He dejado esta cita completa, incidentalmente, no sólo por su exactitud y su valor pedagógico en cuanto a lo que en ella se puntualiza, sino para contrastar la definición geertziana, de grano grueso, con lo que debería ser el canon de refinamiento del debate que él mismo estipulara como el único objetivo susceptible de alcanzarse en una ciencia humana (Geertz 1973: 29 [1987: 39]).

Con malentendidos o sin ellos, la inferencia estadística fisheriana y la inferencia clínica interpretativa coinciden en un mismo modelo inductivo. En lo que concierne a las relaciones entre una muestra parcial y una población total y el escenario global por el otro, nuestro antropólogo planteó la pregunta en *La Interpretación de las Culturas* (1973: 23 [1987: 35]) sin suministrar entonces una respuesta aparte de decir que "los pequeños hechos hablan de las grandes cuestiones […] porque están hechos para hacerlo así". Pero es palpable y explícito que cuando la globalización se le vino encima Geertz (1983; 2000: 137) no tuvo reparos en generalizar sus propias observaciones microscópicas de trabajo de campo a totalidades de "dos millones de personas [Bali], o quince millones [Marruecos], o sesenta y cinco millones [Java]", pasando de la inducción a la teoría, "de las observaciones a las hipótesis, […] de lo particular a lo general" exactamente igual que los fisherianos presumieron haber hecho (Fisher 1971: 3). En su lógica peculiar Geertz acabó dando el mismo salto de fe que se atrevieron a dar los estadísticos aunque sin poner sus hipótesis de generalización en claro ni reflexionar sobre los mecanismos de inferencia que se requerirían para consumar esos portentos abductivos. Y sin aplicar tampoco lo que en un marco cualitativo podría hacer las veces de una prueba hermenéutica de significancia, capaz, llegado el caso, de poner a la luz los alcances y las tribulaciones de la demostración.

Llegados a este punto puede que al lector le sorprenda la cantidad y magnitud de las chapuzas que Geertz es capaz de embutir en un solo enunciado, comenzando por el anonimato de sus fuentes y por el silenciamiento del carácter espinoso de la cuestión. Dado que él ha supeditado siempre el rigor de la argumentación al refinamiento del estilo (encontrando en ello su marca de fábrica) debo decir que a mí no me sorprende demasiado. Cada vez que Geertz se ocupa de menesteres teóricos que se encuentran por encima de un modesto umbral de dificultad un régimen de inconsistencias comparables a las que aquí tuve oportunidad de documentar impregna con intensidad parecida una parte significativa de su argumentación (cf. Reynoso 2008: caps. 1 y 2).

No hay nada peculiar en la escritura geertziana que motive esta tasa de inexactitudes, sin embargo; la imprecisión que afecta a su discurso (y que pone de relieve limitaciones impensadas de la descripción densa y de la interpretación de significados ante un simple trabajo de glosa a libro abierto) no es sino lo que cabe esperar que ocurra cuando un autor pontifica sobre tópicos técnicos de alta complejidad que le son conceptualmente ajenos.[18] Cabría tal vez aplicarle a Geertz, literalmente, lo que alguna vez él espetó con saña a Mary Douglas a propósito de fallas de mucho menor calibre: "Los comentarios –como escribía Gertrude Stein– no son literatura" (Geertz 1987: 37).

Ahora bien, de ningún modo Geertz se lleva la palma de la interpretación más desatinada de los tipos de errores existentes, pues las han habido todavía más torpes: más recientemente el antropólogo Kevin Avruch (2003), citando a Geertz pero no a su lectura tipológica, asegura que uno perpetra un error de tipo I si se muestra culturalmente insensitivo mientras que incurre en un error de tipo II si sobreestima el impacto de lo cultural. Otra vez las definiciones quedan al revés que en el modelo clásico; y otra vez se confunde una decisión dicotómica con un continuum de posibilidades de elección, como si todas las diferencias fueran iguales.

Más allá de estas disquisiciones antropológicas de alto empaque en las que las pretensiones formales de las referencias tipológicas conviven con (y son desmentidas por) una hermenéutica que cada quien modula a su antojo, la definición canónica de los tipos de errores según Neyman-Pearson es la que sigue:

- Un error de Tipo I, conocido también como falso positivo, falso rechazo o error de la primera clase, ocurre cuando en una prueba estadística se rechaza una hipótesis nula que resulta ser verdadera. La tasa de este error se suele representar mediante la letra griega α y por lo general es igual al nivel de significancia de la prueba. Se considera que un error de tipo I es un error que se debe a una "excesiva credulidad".

- Un error de tipo II, llamado también falso negativo, falsa aceptación o error de la segunda clase, se manifiesta cuando en una prueba estadística se falla en rechazar una hipótesis nula que resulta ser falsa. La tasa de este tipo de error se designa mediante la letra griega β y se relaciona con la potencia de la prueba estadística, que es en rigor $1-\beta$, lo cual se conoce también como su sensitividad. Esta potencia mide la probabilidad de que la prueba rechace la hipótesis nula cuando ella es falsa; también puede decirse que la potencia estadística es la capacidad de una prueba para detectar un efecto

[18] Este es tal vez el principal dilema endémico a las formas tradicionales de multi-, inter- y transdisciplinariedad cuando los grados de separación entre las disciplinas están por encima de cierto grado de proximidad. He tratado sistemáticamente la cuestión a propósito de Edgar Morin en Reynoso (2009).

cuando ese efecto existe. Habitualmente se procura que este valor se sitúe en torno a 0,8 (Park 2008). Se considera concomitantemente que un error de tipo II es una falla debida a un excesivo escepticismo. En rigor, un error de este tipo no implica tanto "elegir lo falso", sino más bien "quedarse con lo falso a falta de una mejor alternativa". En cuanto a la potencia estadística, se reconocen que hay tres factores que inciden en ella: el tamaño o magnitud del efecto de la población, el nivel de significancia y el número de observaciones. El tamaño del efecto expresa la discrepancia entre H_0 y H_1; otra forma de expresar lo mismo es definiéndolo como la fuerza de la relación entre la variable dependiente y la independiente (Gliner, Leech y Morgan 2002: 85). Permaneciendo todo lo demás constante, a mayor tamaño del efecto y número de observaciones mayor potencia, y menor potencia cuanto menor nivel de significancia.

En general se admite que los procesos interpretativos prevalecientes se han concentrado en evitar los errores de Tipo I prestando poca atención a los errores de Tipo II. No estoy seguro, sin embargo, que se haya realizado un relevamiento sistemático para probar esta aserción a través de las disciplinas y a lo largo del tiempo. Sin duda resta investigar este punto en profundidad, deslindando el papel que las políticas editoriales pudieran tener en las tendencias resultantes. Éste es el llamado *file drawer problem*, constituido por un número crecido y creciente de estudios que no llegan a la publicación por no haber podido rechazar la hipótesis nula (Rosenthal 1979; Kmetz 2011: 29).

Este apartado no estaría completo si no mencionara el hecho de que varios autores propusieron otros tipos de errores desde fechas tempranas. Tal vez el más relevante a nuestros fines sea el error de Tipo III propuesto por Robert Schlaifer (1959: 654), consistente en la aplicación indebida de procedimientos estadísticos. Antes y después de la enmienda de Schlaifer otros estudiosos propusieron sus propias definiciones más o menos irónicas de los errores de la tercera y (a veces) de la cuarta clase: (1) "rechazar correctamente la HN por la razón errónea" (Mosteller 1948: 61); (2) "dar la respuesta correcta al problema equivocado" (Kimball 1957: 134); (3) "resolver el problema correcto demasiado tarde" (Raiffa 1968: 264); (4) "interpretar incorrectamente una hipótesis correctamente rechazada" (Marascuilo y Levin 1970: 398); (5) "desarrollar buenas respuestas a las preguntas equivocadas" (Mitroff y Silvers 2009).

Al lado de las dificultades de interpretación el procedimiento mediante el cual se calcula *p* ha sido cuestionado por los estadísticos bayesianos. El matemático Sir Harold Jeffreys [1891-1989], por ejemplo, sentía que la lógica de basar los valores *p* en la región de cola de la distribución (en lugar de hacerlo en los datos mismos) era una tontería. El razonamiento de Jeffreys es sutil y difícil de seguir y él mismo se encuentra desacreditado por haberse opuesto a la teoría de la deriva continen-

tal; pero dado que en este caso él trata las definiciones de los elementos estrictamente tal cual se han establecido, el efecto de su inferencia es devastador y merece ser considerado con detenimiento:

> Si P es pequeño eso significa que han habido desviaciones inesperadamente amplias de la predicción [bajo la hipótesis nula]. Pero ¿por qué deberían éstas ser formuladas en términos de P? Este último identifica la probabilidad de desviaciones medidas de una manera particular, igual *o mayor que* el conjunto observado, y la contribución del valor concreto [de la estadística del test] es casi siempre despreciable. *Pero lo que el uso de P implica, por lo tanto, es que una hipótesis que puede ser verdad puede ser rechazada porque no ha predicho resultados observables que no han ocurrido*. Esto parece ser un procedimiento curioso. En función de él, el hecho de que tales resultados no hayan ocurrido debería ser tomado más razonablemente como evidencia a favor de la ley [o la hipótesis nula] y no en contra de ella (Jeffreys 1961: 385; énfasis en el original).

La clave para comprender la objeción de Jeffreys radica en que la propia definición *probabilística* del valor de *p* presume la certidumbre de que haya desviaciones *mayores* (potencialmente observables) a las que se presentan en el conjunto observado. Jeffreys también advertía que el procedimiento lógico había quedado inadvertidamente dado vuelta. Tal como lo expresan McCloskey y Ziliak (2009: 46), el estudioso de la facción Cinco Porcientadora anhela encontrar un cuerpo de datos "significante y consistente con" alguna hipótesis. La búsqueda en sí es intelectualmente comprensible; pero se está buscando el elemento equivocado de la manera errónea:

> Les guste o no a los estadísticos, sus resultados se usan para decidir entre hipótesis; y es elemental que si *p* entraña *q*, *q* no necesariamente entraña *p*. No podemos pasar de 'los datos son improbables dada la hipótesis' a 'la hipótesis es improbable dados los datos' sin que medie alguna regla de pensamiento adicional (Jeffreys 1963: 409)

A todo esto, ni la estadística ni las ciencias empíricas que han implementado la prueba de significancia han explorado suficientemente el curso de acción cuando lo que se requiere es afirmar tajantemente la hipótesis nula, como cuando el científico escéptico o el consultor en estadística pretende probar que ciertos aparentes "misterios" de percepción extrasensorial, telepatía o desaparición de barcos y aviones en un triángulo oceánico sólo son fenómenos que bien podrían ocurrir con mediana probabilidad y por mero efecto del azar. Si bien la estadística afinó sus propias técnicas de muestreo y prueba de significancia en el campo ocultista (cf. Edgeworth 1885b; 1887; Fisher 1929; Hacking 1988; Utts 1991) la refutación estadística de lo paranormal (o del diseño inteligente, o del creacionismo científico) dista mucho de haber sido sido, como veremos más adelante (págs. 42 y 76), un tranquilo paseo por el campo.

7. Significancia y significado

Mientras que en el apartado anterior pudimos observar la forma en que los hermeneutas supieron retorcer las conceptualizaciones de la estadística, en lo que sigue podrán comprobarse, simétricamente, los atropellos perpetrados por los estadísticos en materia de semántica e interpretación. La historia oficial sostiene que la NHST ha sido cuestionada en este rubro desde que el educador Ralph Winfred Tyler (1931) alertara sobre el uso acrítico de la prueba estadística de significancia y la confusión imperante en torno de esta palabra:

> Las interpretaciones que comúnmente se han elaborado en estudios recientes indican con claridad que somos propensos a concebir la significancia estadística como equivalente a la significancia social. Estos dos términos son esencialmente distintos y no deberían confundirse […] Las diferencias que son estadísticamente significantes no siempre son socialmente importantes. El corolario es también verdad: diferencias de las que se puede mostrar que no son estadísticamente importantes pueden ser sin embargo socialmente significantes (Tyler 1931: 115-117).

Ahondando en los repositorios bibliográficos he encontrado que la crítica más temprana de las connotaciones de la significancia se formuló mucho antes que eso en al menos dos artículos de un psicólogo de intuición asombrosa pero insólitamente apellidado Boring, Edwin Boring (1919; 1926). En el segundo de esos *papers* olvidados Boring no sólo advierte de la diferencia entre dos significancias distintas sino que revela que lo esencial del método de la NHST ya estaba plenamente articulado (presunción de normalidad y contraste con opción nula incluidas) antes que Fisher comenzara a escribir sobre la cuestión y le pusiera nombres a las cosas:

> Hace algunos años me atreví a expresar mi opinión sobre la adecuación relativa del método estadístico en lo que concierne a los problemas de descripción y generalización. Discutía entonces la cuestión de la significancia de las diferencias, y sugería que uno podría necesitar distinguir entre "significancia matemática" y "significancia científica". Supongo que el problema elemental más frecuente de la estadística es establecer una diferencia en la tendencia entre dos series [*arrays*] de particulares. En tal caso uno analiza cada serie en una tendencia central y una medida de precisión, anota la diferencia entre las dos tendencias centrales y luego determina la "significancia" de esta diferencia considerándola en su relación con su medida de precisión. La medida más simple de "significancia" es la razón entre la diferencia y su error probable, pero si se presume la universal normalidad de la distribución, esta razón se puede convertir en una "probabilidad de que la diferencia no se deba al azar" (Boring 1926: 303).

Ambos artículos de Boring prodigan observaciones sobre el muestreo aleatorio, la normalidad y la representatividad de las muestras que desde la perspectiva actual suenan de estilo arcaico pero de honda inspiración metodológica. Igualmente anticipatoria es la convicción de que es la dimensión

científica antes que las abtracciones matemáticas lo que debe tener precedencia: “Este caso es uno de los muchos en los que la habilidad estadística, divorciada de una intimidad científica con las observaciones fundamentales, no lleva a ninguna parte” (Boring 1919: 338). En las ciencias humanas recientes frases como éstas aparecen todos los días y hasta tienen un tinte de obligatoriedad; pero no es común encontrar juicios de este tenor en la escritura de científicos educados para el cálculo.

En sociología la discusión sobre las diferencias que conviene establecer entre la significancia estadística y la significación sustantiva se continuó en un artículo del sociólogo Hanan Selvin (1957: 523-524), quien concluyó que las pruebas estadísticas eran inaplicables a contextos de investigación que no fueran plenamente experimentales, que no garantizaran la estricta independencia de las variables y que no permitieran, por ejemplo, aleatorizar los datos conforme a los requisitos del método.[19]

Atenuada la polémica sobre la independencia de los datos y sobre la improbabilidad de que una muestra sea genuinamente aleatoria, la disputa sobre el significado empírico de la significancia conoció un clímax a fines de los 60s pero luego se estabilizó, quedando a la postre en un segundo plano (Kish 1959; Bolles 1962; Gold 1969; Reynolds 1969; Morrison y Henkel 1969; Winch y Campbell 1969; Taylor y Frideres 1972). Poco a poco los argumentos empezaron a girar en torno a variaciones de unos pocos temas inconcluyentes; casi las mismas observaciones que hemos visto planteadas por Tyler, por ejemplo, resurgen en los ensayos del investigador de la conducta Frederick Nicholas Kerlinger (1979):

> La significancia estadística dice poco o nada sobre la magnitud de una diferencia o de una relación. Con un gran número de ejemplares […] las pruebas de significancia muestran significancia estadística incluso si una diferencia entre medias es muy pequeña, y hasta trivial, o si un coeficiente de correlación es muy pequeño y trivial. […] Para usar las estadísticas adecuadamente, se deben comprender los principios involucrados y ser capaz de juzgar si los resultados obtenidos son estadísticamente significantes *y* si son significativos en el contexto de la investigación particular (pp. 318-319).

Todavía a fines del siglo pasado Bruce Thompson (1996) y Pedhazur y Schmelin (1991: 202) explicaban la vigencia anacrónica de esta ambigüedad de percepción atribuyéndola a la tendencia de los investigadores a utilizar (y de las publicaciones periódicas a publicar) manuscritos que con-

[19] La práctica de la aleatorización recién se impuso tras mucha discusión en la década de 1930. Aunque se la da por sentada en la corriente principal, su utilidad sigue siendo ampliamente resistida en estadística y no sólo entre los bayesianos. A fin de no ampliar demasiado el frente de la presente discusión no trataré el tema en este ensayo. La bibliografía sobre el carácter polémico de la aleatorización, la retórica de la “representatividad” y los supuestos ocultos en los métodos de muestreo en general es de todos modos masiva. Véase Gosset (1942: *paper* §5 [1911], §11 [1923] y §13 [1926]); Harville (1975); Kruskal y Mosteller (1979a; 1979b; 1979c; 1980); Hacking (1988); Krishnaiah y Rao (1988: 1-14).

tienen la expresión "significante" en vez de "estadísticamente significante". A caballo de este conveniente efecto de elipsis, devino práctica común eliminar toda referencia a la palabra "estadística" y hablar en cambio de "diferencias significantes", "correlaciones significantes" y demás, como si de hermenéutica se tratara.

Muchos han denunciado estos ardides de semantización de segundo orden que se creerían imposibles de consumar en un modelado matemático. M. G. Kendall y A. Stuart (1951: 163) reconocieron estos y otros excesos retóricos, recomendando la frase "tamaño de la prueba" en vez de "nivel de significancia". Del mismo modo, los sociómetras Denton Morrison y Ramon Henkel (1970: 198) sugirieron que la "prueba de significancia" sea sustituida por la expresión fea pero menos pomposa de "procedimiento de decisión de error de la muestra". Eso no evitó que el hábito siguiera asomando aquí y allá a lo largo de las décadas, cada vez como si fuera la primera (Berkson 1942; Gold 1969; Winch y Campbell 1969; Chow 1988; Shaver 1993). En un espléndido artículo que examina la retórica de los estudios estadísticos de economía siguiendo la tradición inaugurada por Bill Kruskal (1978), D. McCloskey (1985: 204) llama a una variante de ese recurso la "falacia de equivocación".

El consenso contemporáneo está generalmente de acuerdo con la forma en que David Roxbee Cox plantea la cuestión cuando escribe que "el punto central es que la significancia estadística es muy diferente de la significancia científica y por tanto la estimación […] de la magnitud de los efectos es por lo general esencial, más allá que se logre o no una diferencia estadísticamente significante respecto de la hipótesis nula" (Cox 1978). El dilema no es tanto teórico como práctico y pertenece menos al orden de la semántica que al de la pragmática; en este contexto escribía W. Edwards Deming [1900-1993], creador del método Total Quality: "La significancia estadística de *B* sobre *A* no proporciona conocimiento ni bases para la acción" (Deming 1975: 149). Por algo ha sido que Deirdre McCloskey y Stephen Ziliak (2007) consideraron que al llamar "significancia" a su concepto, Edgeworth homologó un término desastrosamente equívoco.[20]

A la hora del balance, el contraste entre la significación empírica y la significancia estadística se correlaciona con la diferencia entre las teorías sustantivas y las hipótesis estadísticas, antítesis que a su vez tiene algo más que un aire de familia con la clásica contraposición entre las teorías sustantivistas y las formalistas en antropología económica (Carrier 2005: 18-20, 502-504). También en-

[20] Llama la atención encontrar a lo largo y a lo ancho de la estadística nombres sobre los que hay escaso acuerdo o de los que se admite que son inapropiados. Uno de ellos, por ejemplo, es el error probable (la desviación relativa a la medida central dentro de la que se espera que la mitad de los casos caiga por azar, equivalente a 0,67456 o =2/3 de la desviación estándar): es ostensiblemente engañoso llamar "error" a la variabilidad. Como fuere, "tres veces el error estándar", o más o menos dos veces la desviación estándar, fue la cifra escogida por Fisher para delimitar la significancia (1925: 47-48; Cowles y Davis 1982).

cuentro que todos estos contrastes son correlativos a las diferencias que señalara Berkson en una de las críticas más tempranas que se formularon al método, atinentes a las clases contrastantes de procedimientos mediante los cuales se obtiene un valor de *p*:

> [E]xiste una importante distinción entre la connotación física de una prueba para, digamos, la significancia de una diferencia entre medias o varianzas y una diferencia de chi cuadrados. Podemos concebir una verdadera diferencia de medias, o una verdadera diferencia de varianzas, las cuales corresponden a las distribuciones verdaderas. Pueden ser definidas operacionalmente. Las pruebas son, por así decirlo, comentarios sobre nuestras estimaciones de estas diferencias verdaderas. Pero no hay nada que corresponda a una verdadera diferencia de chi cuadrado entre las distribuciones verdaderas. El chi cuadrado no se corresponde con ningún carácter definible específico de la distribución verdadera. No es un parámetro descriptivo, como sí lo es la desviación estándar (Berkson 1938: 527, n. 2).

Algunos procedimientos estadísticos consagrados, en síntesis, obvian o aniquilan todo isomorfismo o analogía imaginable entre los valores estadísticos y las propiedades estructurales de las poblaciones "verdaderas". Mientras que esta observación de Berkson casi no se tuvo en cuenta en los años subsiguientes, nadie desarrolló la problemática de los contrastes entre lo físico y lo estadístico mejor de Paul Meehl a propósito de la diferencia en el uso de la NHST en psicología y en agronomía, la disciplina de origen del propio Ronald Fisher. Decía Meehl que existe

> una distancia lógica entre las hipótesis estadísticas y la teoría sustantiva que, cuando se combina con el factor de impureza,[21] introduce una diferencia entre la prueba correlacional de teorías en la psicología blanda y la manipulación experimental en agronomía que involucra una diferencia de clase y no de grado. […] Es precisamente la distancia lógica entre la hipótesis estadística y la teoría sustantiva cuando se articula con la ubicuidad de las correlaciones que no dan cero, lo que hace que la estrategia vigente sea radicalmente defectuosa y probablemente imposible de mejorar [*probably not improvable*], incluso si los demás ofuscadores pudieran ser eliminados o se redujera su dimensión e influencia (Meehl 1990b: 226)

Las complicaciones se originan, sostiene el autor, en el hecho de que los libros introductorios y los profesores de estadística elemental utilizan la palabra "hipótesis" de una manera un tanto indiscriminada, sin marcar ninguna diferencia entre una teoría sustantiva (causal, estructural o composicional) y una hipótesis estadística sobre los valores numéricos de los observables. En un artículo sintomáticamente titulado "Valores de *p* e intervalos de confianza: Dos lados de una misma moneda insatisfactoria", el epidemiólogo Alvan R. Feinstein contempla el mismo problema desde un ángulo levemente distinto:

---

[21] Sobre este factor véase más adelante, pág. 50.

> Si un cálculo resulta en un valor de p que se encuentra por debajo de α, o si un intervalo de confianza $1–\alpha$ excluye el resultado nulo de "no diferencia" se proclama "significancia estadística". Tanto el valor de p como los métodos de intervalo de confianza son esencialmente recíprocos, dado que utilizan los mismos principios de cálculo probabilístico; y ambos pueden dar resultados distorsivos o equívocos si los datos no se conforman a los requisitos matemáticos subyacentes. La mayor desventaja científica de ambos métodos es que su "significancia" es meramente una inferencia derivada de principios de la probabilidad matemática, y no una evaluación de la importancia sustantiva de la magnitud "grande" o "pequeña" de la distinción observada. […] Después de un siglo de "significancia" inferida exclusivamente a partir de probabilidades, un desafío científico básico es desarrollar métodos para decidir qué es lo que es ya sea sustantivamente importantísimo o qué es lo que es más bien trivial (Feinstein 1998: 355).

Una de las soluciones que se han propuesto para poder compensar la falta de relevancia de la significación estadística en materia de significación científica consiste en alentar a que en todo estudio se reporte la potencia estadística, o sea $1–\beta$, siendo β la probabilidad de cometer un error de tipo II. El problema que ello acarrea es que esa decisión involucra que el tamaño de la muestra intervenga en el cálculo con los siguientes efectos: si el tamaño es pequeño la potencia será baja y la HN será falsamente aceptada; pero si el tamaño de la muestra es grande la potencia será tan extremadamente alta que será imposible rechazar una HN aun cuando conceptualmente sea insostenible. Una forma de salir del paso es reducir sustancialmente el nivel de α que decide el rechazo de la HN cuando el tamaño de la muestra y la potencia estadística son elevados; pero no existen lineamientos claros para la reducción de la magnitud de α en relación sistemática con los tamaños de las muestras.

Fuera del campo de la literatura crítica reciente, pocos estudiosos han tomado conciencia del hecho de que la contundencia de los valores de significancia de la prueba estadística no puede penetrar ni en el significado de las argumentaciones en juego, ni en su consistencia lógica, ni en su relevancia pragmática. A este respecto propongo examinar dos testimonios que ilustran los riesgos y limitaciones de la NHST cuando ella interviene en la práctica.

El primero de los testimonios documenta una situación de extrema gravedad. Cuando en los laboratorios Merck se hicieron las pruebas para evaluar los efectos colaterales adversos en el uso de la droga anti-inflamatoria Rofecoxib (distribuido con la marca Vioxx) se encontró que los individuos del grupo al que se había suministrado el producto experimentaban sucesos cardíacos diversos, incluyendo (según los documentos) entre 5 y 8 infartos fatales que no se manifestaron en el grupo de control al cual se administró naproxen. El laboratorio estimó, sin embargo, que estos efectos adversos existían pero no eran estadísticamente significativos y la Corte acompañó su postura. Hubo advertencias por parte de los especialistas pero fueron inicialmente desestimadas y la droga se lanzó a la venta. Con el tiempo, sin embargo, y ante nuevas víctimas fatales, miles de millones de dólares

perdidos y un vendaval jurídico, la droga debió ser retirada del mercado (Brief of Amici Curiae 1993; Lisse y otros 2003; FDA Consumer 2004: 38; Ziliak y McCloskey 2008; McCloskey y Ziliak 2010; Wall Street Journal 2011). Recientemente la Corte Suprema de los Estados Unidos también se expidió (lo mismo que la FDA a propósito de los implantes mamarios) respecto de la insuficiencia material de la prueba estadística en otros casos de litigio (Supreme Court of the United States 2010).

El segundo testimonio del desacople entre la significancia estadística y la significación concreta se manifiesta en la evaluación que uno de los más grandes estudiosos de la teoría de la probabilidad, Francis Ysidro Edgeworth [1845-1926], elaborara en 1885 sobre los datos inexplicables recabados por el futuro Premio Nóbel Charles Richet [1850-1935] en diversos experimentos de telepatía y percepción extrasensorial que este último llevó a cabo cuando era joven. La conclusión de Edgeworth fue que la probabilidad de que los fenómenos reportados por Richet resultaran del mero azar era muy baja, de alrededor de 0,00004, de modo que la confianza de que no se debieran al azar "puede considerarse como certidumbre física" y "la conclusión se podría considerar segura". Y luego dictaminó, serenamente:

> Tal es la evidencia que el cálculo de probabilidades concede a la existencia de una agencia distinta a la del mero azar. El cálculo guarda silencio en lo que atañe a la naturaleza de esa agencia. Esta es una cuestión a ser decidida no mediante fórmulas y figuras, sino a través de la filosofía general y el sentido común (Edgeworth 1885b: 199).

En una de sus indagaciones más punzantes, el filósofo canadiense Ian Hacking (1988: 441) encuentra que esta advertencia es argumentativamente idéntica a una celebrada aserción de Fisher sobre la lógica de la prueba de significancia. En esta prueba se opera una "disyunción lógica" mediante la cual la estadística deja de ser un mero cálculo para devenir inferencia: o bien algo muy poco común ha ocurrido por azar, o una hipótesis de "no efecto" debe ser rechazada. La naturaleza de la agencia que hace que ocurra lo que ha ocurrido (el espacio vacío de lo que luego será la hipótesis alternativa) es también (en el modelo de Fisher) objeto de silencio. En términos de lógica, significado y retórica, en fin, el isomorfismo entre el argumento ocultista y la formulación estadística es incontrovertible.

Lo curioso del caso (como puede verse siguiendo los hipertextos al final de la versión digital de la bibliografía) es que, cerrando el círculo, Fisher introduce explícitamente la disyunción tomando como modelo "los estudios conocidos como para-psicología" en los que las técnicas de muestreo y aleatorización hicieron sus primeras armas (Fisher 1956: 43; cf. también Fisher 1929). Después de todo, la primera prueba de hipótesis que realizó un contraste con una opción de azar fue la de John Arbuthnott (1710), quien probó de este modo la existencia de Dios; la misma lógica prevalece

también en la idea de "diseño inteligente" en el creacionismo contemporáneo y específicamente en el "filtro explicativo" de William Demski (Danziger 1990; Gigerenzer y Murray 1987: 4-5; Gigerenzer 1998a; http://www.designinference.com/). Vale la pena citar el argumento de Arbuthnott, quien pretendía explicar el hecho de una mayor proporción de nacimientos de varones que de mujeres en Londres durante 82 años consecutivos:

> De lo que ha sido dicho resulta evidente que la Proporción de *A*s para cada Año es menos que ½; (pero para que el Argumento resulte más fuerte) dejemos que esta Proporción sea igual a ½ por cada año. Si se trata de hacer lo mismo 82 veces consecutivas, su Proporción será $½^{82}$, lo que se puede encontrar fácilmente mediante la Tabla de Logaritmos que equivale a 1/4836000000000000000000000. Pero que no sólo el Número de Varones exceda al de Mujeres cada Año, sino que este Exceso ocurra en Proporción constante, y que la Diferencia permanezca dentro de límites fijos; y esto no sólo por 82 Años, sino por Eras y Eras, y no sólo en *Londres*, sino a lo largo de todo el Mundo, a mucho menos que cualquier fracción asignable, de ello se sigue que es el Arte, y no el Azar, lo que gobierna (Arbuthnott 1710: 188-189).

Si se compara este razonamiento con la fundamentación de la prueba de hipótesis de Fisher se verá que los mecanismos de inferencia subyacentes son idénticos.[22]

De ningún modo creo en los argumentos sustantivos de John Arbuthnott o de William Demski, desde ya. Hay innumerables refutaciones de ambos en el plano de la biología, la filosofía de la ciencia e incluso la teoría de la información que son acertadas y bien conocidas. La refutación de una hipótesis de azar, por más que sea aplastante, tampoco implica la aceptación de ninguna hipótesis alternativa específica; si hay alguna prueba formal de la existencia de Dios decididamente no puede venir por esta vía. Pero no conozco una sola refutación probabilística de argumentos como los de Arbuthnott de la cual no se infiera también, categóricamente, la invalidez de la NHST. Y a la inversa, no es posible pensar en una sola prueba de la validez de la NHST que no involucre también reconocer la validez de la inferencia de Arbuthnott.

No estoy implicando que la literatura de fundamentación de la NHST en su conjunto tenga el mismo valor que la que sustenta el fervor religioso de Arbuthnott, las investigaciones psíquicas del tardío siglo XIX o el creacionismo científico; lo que sí afirmo, con el aval de todos los autores implicados y a la luz de una evaluación de significancia impecablemente ejecutada tanto por Arbuthnott como por Edgeworth (y en ocasiones también por Demski), es que la disyunción lógica en que reposa el método ni posee frente a los recursos de la superchería una fuerza crítica que pueda reputarse aplastante, ni puede decirnos nada sobre la índole y significación de las causas intervinientes.

---

[22] Véase más adelante, pág. 71.

No es casual entonces que de todos los conceptos vinculados al asunto, la significancia estadística haya sido uno de los que se han definido de manera más volátil e impropia, propiciando por ello errores de antología. El psicólogo Donald Hebb, por ejemplo, ha escrito:

> Cuando se encuentra una afirmación de que una diferencia es significante, eso significa, por convención, de que la probabilidad es por lo menos de 19 contra 1 de que ella se deba a operaciones de azar al obtener nuestra muestra (Hebb 1966: 173).

Esta curiosa fantasía del 19 a 1 se origina en el hecho de que 0,95 y 0,05 divididos ambos por 0,5 resultan 19 y 1 respectivamente. Se interpreta entonces que un resultado estadísticamente significante al nivel de 0,05 significa que sólo 5 veces de cada 100 este resultado se deberá al azar, o al muestreo, o a errores de muestreo, o a otras entidades del mismo género. Lee Cronbach y Richard Snow ponen las cosas en su lugar:

> Un valor de *p* alcanzado por métodos clásicos no es un resumen de los datos. Tampoco el valor de *p* agregado a un resultado nos dice cuan fuerte o confiable resulta ser ese resultado. Escritores y lectores tienden a leer 0,05 como p(H|E) “la probabilidad de que la Hipótesis sea verdad, dada la Evidencia”; como los libros de texto de estadística reiteran casi en vano, *p* es más bien p(E|H), la probabilidad de que surja la Evidencia si la Hipótesis (nula) es verdad (Cronbach y Snow 1977: 52).

Tal como lo afirma Ronald Carver (1978) y como hemos tenido oportunidad de comprobarlo, este es quizá el principio más fundamental y menos comprendido de la NHST.

8. El elusivo significado de la hipótesis nula

Hasta donde he podido investigar la expresión "hipótesis nula" fue acuñada en un libro seminal por Ronald A. Fisher (1971 [1935]: 19) por lo menos diez años después que comenzara a elaborar su método de prueba de significancia. Es sugestivo que el autor afirme que la HN "nunca se prueba ni se establece, sino que es posiblemente des-probada en el curso de la experimentación. Puede decirse que cada experimento sólo existe con el propósito de dar a los hechos la oportunidad de des-probar [*disproving*] la hipótesis nula" (p. 16). Lo notable del caso es que Fisher elabora estas ideas unos pocos años antes que Karl Popper, por ejemplo, hiciera conocer sus ideas sobre falsabilidad.

En una crítica que raya entre las más tempranas, Joseph Berkson (1942) expone el significado y la práctica de la HN en el contexto de la estadística en general primero y de la prueba estadística en particular después:

> Dudosamente sea una exageración decir que las estadísticas, tal como se enseñan actualmente en la escuela dominante, consisten casi enteramente en pruebas de significancia, aunque no siempre presentadas como tales, algunas comparativamente simples y directas, otras elaboradas y abstrusas. Por detrás de esto hay una doctrina de análisis que consiste en configurar lo que se llama una "hipótesis nula" y en ponerla a prueba. En esta concepción, ciertamente, el procedimiento no sólo caracteriza el método de la estadística, sino que se lo considera la esencia misma de toda la ciencia experimental. (Berkson 2003: 687).

Joseph Berkson (que es el mismo autor que hemos visto promoviendo el uso de la distribución logística y liberando de culpa a la comercialización del tabaco en la etiología del cáncer) se oponía a la idea de que el planteo sobre la HN requiriera siempre una respuesta negativa:

> En el esquema de la hipótesis nula estamos tratando de nulificar algo. "La hipótesis nula nunca es probada o establecida sino posiblemente des-probada en el curso de la experimentación". Pero la evidencia ordinaria no toma esta forma. Con el *corpus delicti* delante nuestro no decimos "Hay evidencia contra la hipótesis de que nadie está muerto". Decimos, más bien, que "Evidentemente alguien ha sido asesinado" (loc. cit.).

En la misma línea de razonamientos, Jum Nunnally, un especialista en estadísticas para la psicología y la educación cuyos textos han leerse con especiales precauciones, ha afirmado que

> los modelos de hipótesis nula comparten una debilidad que las vulnera: en la vida real la hipótesis nula casi nunca es verdad, por lo que carece de sentido desarrollar un experimento con el solo objetivo de rechazarla. […] Si el rechazo de la HN fuese la intención real de los experimentos psicológicos, no habría ninguna necesidad de recopilar datos (Nunnally 1960: 210).

En rigor, la idea de que "la hipótesis nula de no-diferencia se sabe usualmente que es falsa antes de recolectar los datos" había sido formulada unos años antes por Leonard Savage (1957: 332-333), el

genial “consultor estadístico” de John von Neumann. Ahora bien, a menudo la HN puede ser trivial o artificiosa, pero (dada la variedad de modelos de prueba imaginables) no es imperativo que sea falsa en todos los casos aunque buena parte de la crítica lo presuma así. Esta presunción, incidentalmente, habla a las claras de la ambigüedad de todo cuanto concierne a la prueba, crítica incluida. En un cuestionamiento de tono más pragmático que el de Nunnally el psicólogo cognitivo Geoffrey Loftus asegura:

> Nadie jamás pareció saber exactamente qué nos puede decir exactamente la prueba de hipótesis que sea en alguna medida interesante o importante. Muchos colegas psicólogos, quizá en la desesperación de su expresión de deseos, han llegado a creer en una variedad de implicaciones generalmente oscuras e incorrectas del proceso (por ejemplo, que la magnitud del valor de *p* nos dice algo sobre la magnitud de un efecto, o la replicabilidad de un efecto, o la probabilidad de que la hipótesis nula o la alternativa sea verdadera o falsa). En algún punto a lo largo de esta línea, sin embargo, todos hemos internalizado una lección que es enteramente correcta: cuanto más tú rechaces la hipótesis nula, más probable es que consigas trabajo académico (Loftus 1991: 103).

Años más tarde, apenas unos pocos meses atrás, el mismo Loftus (2010) sintetizaría los principales inconvenientes de la pruebas de la HN en los siguientes tres puntos:

- Una HN nunca puede ser literalmente verdadera. En casi todas las ciencias es casi una verdad evidente que es improbable en extremo que una variable independiente no ejerza algún efecto en las variables dependientes (considerando una resolución con un número infinito de posiciones decimales). El rechazo de la HN (la única conclusión fuerte posible) no dice entonces al investigador nada que éste no hubiera podido intuir de antemano. Paradójicamente, este rechazo es lo único que puede dar crédito a una hipótesis de investigación que ni se especifica ni se comprueba mediante el procedimiento estadístico (Frias, Pascual y García 2002: 182).

- La naturaleza humana hace que la aceptación de la HN sea casi irresistible. Aunque es difícil que la HN sea siquiera verosímil, no se hace fácil aceptar una conclusión tan débil como que “uno ha fallado en rechazar la HN”. Cuando eso sucede, Loftus observa que en las conclusiones de los trabajos empujados a ese extremo esta expresión correcta muta silenciosamente en algo que puede interpretarse como “la HN es verdadera”.

- La NHST enfatiza conclusiones descarnadamente dicotómicas. La prueba en sí nada puede decirnos sobre la medida en que los M_j observados son buenas estimaciones de los μ_j inobservables. Debe desarrollarse entonces extremo cuidado al hacer inferencias

> sobre los μ_j a partir de los M_j, a riesgo de incurrir en las más variadas falacias si así no se hiciere.

A lo largo y a lo ancho de la literatura se ha hecho costumbre enseñar a los bisoños que la hipótesis nula tiene algo que ver con un "efecto nulo" (de la aplicación de un medicamento o de la implementación de una política, por ejemplo). Una vez más, alcanza con reflexionar un poco para comprobar que esa interpretación no es exacta en todos los escenarios, dependiendo en gran medida de la argumentación que rige el diseño experimental. En un sentido estricto, el calificativo "nulo" designa a la hipótesis a "nulificar"; esto no implica entonces que las políticas o procedimientos que pudiesen mencionarse en el *wording* de la hipótesis no tengan efecto alguno en la práctica (Berkson 1942; Bakan 1966: 423).

Tras mucho trajinar por la bibliografía he encontrado que pocas observaciones sobre la naturaleza de la HN son tan esclarecedoras como esta nota al pie de William Kruskal:

> El término "hipótesis nula" ha sido sumamente distorsionado. En su sentido original se refería a alguna afirmación sobre la distribución del punto de muestra [*sample point*] que, de ser verdad, rara vez sería rechazado. Aparte de este requerimiento, [la HN] podría no ser evidencia de nulidad; por ejemplo, θ=17 o $G=F^3$ (donde "θ" denota un parámetro y "G" y "F" distribuciones acumulativas) podrían ser afirmaciones abreviadas de hipótesis nulas. Muchas o la mayoría de las hipótesis nulas podrían ser trivialmente re-expresadas como igualdades con cero de un lado, C–17=0, o $G–F^3$=0, pero esto no es de gran interés. Lo que podría llamarse una hipótesis nula fuerte, o una hipótesis nula nula, es una afirmación que estipula que alguna transformación del punto de muestra deja inalterada una distribución; por ejemplo, una hipótesis nula nula en un problema de muestras múltiples podría ser que las muestras son realmente de una distribución común, de modo que permutando sus nombres no se altera la distribución conjunta (Kruskal 1980: 1021, n. 3).

La evidencia de nulidad es, entonces, un concepto delicado. Como se ve claramente ante temas tan álgidos como el cáncer de pulmón, la exposición a la violencia televisiva o el calentamiento global hay que ser muy cuidadosos en cuanto a los matices de la lógica y de la formulación experimental: "que no haya evidencia de efecto no quiere decir que haya evidencia de que no hay efecto", o más estrictamente: "ausencia de evidencia no es evidencia de ausencia" (Altman y Bland 1995).

Louis Guttman iba más lejos y proponía que una HN no debería hipotetizar nulidad; dada su improbabilidad sustancial la nulidad debería ser, por lo general, una hipótesis alternativa razonable (Guttman 1977: 95). Este punto merece, a mi juicio, la más minuciosa reflexión, pues es probable que toda la lógica involucrada en la prueba de hipótesis se encuentre exactamente al revés de lo que debería estar: lejos de constituir el grado cero del acontecimiento significativo y la opción más probable bajo un régimen de azar, la ley normal perfecta y el ruido blanco correlativos a la HN configuran

frente a cualquier hipótesis sustantiva una instancia de antítesis cuya improbabilidad ronda lo absoluto. Habría que plantear muy estúpidamente la prueba para no obtener en el cotejo contra la nulidad los resultados que se necesitan. En todo caso, que se siga discutiendo si aquello contra lo cual contrastamos es lo más probable o por el contrario lo más improbable debería ser, pienso, materia de preocupación.

Un aspecto francamente inaceptable del uso de la HN en la literatura de investigación empírica radica en la frecuencia con que se han postulado HNs triviales o bien lisa y llanamente imposibles, figuras de paja construidas solamente al efecto de establecer un contraste con cualesquiera hipótesis alternativas. Ignoro cuál pueda ser cuantitativamente hablando la situación en nuestras disciplinas, pero Anderson, Burnham y Thompson (2000) investigaron la literatura biológica, encontrando que sólo 5 entre 95 artículos en el *Journal of Wildlife Management* incluían HNs que podrían juzgarse opciones plausibles; el resto proponía argumentos que carecían de plausibilidad biológica aun antes que se emprendiera cualquier estudio tendiente a su refutación. Una vez más, Louis Guttman ha ponderado la seriedad del problema:

> [M]uchos –si no la mayoría– de los practicantes no desarrollan el pensamiento científico que debe preceder a la inferencia estadística. No llevan a cabo una elección de la hipótesis nula *versus* la hipótesis alternativa que esté adecuadamente a la medida de su problema sustantivo específico. Se comportan como si estuvieran bajo la ilusión de que la disyuntiva no está en sus manos, de que la hipótesis nula está pre-determinada ya sea por los matemáticos que crearon la inferencia estadística moderna o por principios de parsimonia inmutables y carentes de contenido (Guttman 1977: 82).

La mención de los matemáticos que crearon la inferencia estadística moderna no es ociosa: a partir de este juicio y aunque Guttman ha decidido no dar un solo nombre, el estado de cosas puede entenderse como un efecto colateral de los procedimientos mecánicos de inducción instaurados por Fisher. Guttman juzga asimismo que el adjetivo “nula” es desafortunado; debería hablarse, sostiene, de hipótesis incumbentes *versus* hipótesis desafiantes, tal que aquéllas resulten desalojadas si y sólo si se llegase a acumular en su contra una evidencia abrumadora (Guttman 1977: 87).

En su clásico y corrosivo artículo “The earth is round: $p \leq 05$” Jacob Cohen (1994: 1000-1001) se refiere a las HNs de riesgo cero y escaso valor metodológico (p. ej. “el precio de los automóviles no está vinculado con las cifras de venta”) como *nil hypotheses*, algo así como “hipótesis de la nada” o “hipótesis de nihilismo”. El psicólogo David Bakan [1921-2004], un intelectual conocido por sus comparaciones magistrales entre el psicoanálisis y la mística judía, había aportado mucho antes estos corrosivos elementos de juicio:

> El meollo del asunto es que realmente *no hay una buena razón para esperar que la hipótesis nula sea verdad en alguna población*. ¿Por qué debería la media, digamos, de todos los puntajes al este del Mississippi ser *idéntica* a todos los puntajes al oeste del Mississippi? ¿Por qué cualquier coeficiente de correlación sería *exactamente* de 0,00 en una población? ¿Por qué esperaríamos que la relación entre varones y mujeres sea *exactamente* 50:50 en una población? ¿O por qué distintas drogas deberían tener exactamente el mismo efecto sobre cualquier parámetro de la población? *Una mirada a cualquier conjunto de estadísticas sobre poblaciones totales confirma rápidamente la rareza de la hipótesis nula en la naturaleza* (Bakan 1966: 426).

Hay razones para creer, concluye Bakan (p. 434), que la existencia de HNs perspicaces es improbable, lo cual no ha sido óbice para que un número de respetados estadísticos haya recomendado medio siglo atrás que se postulen HNs de cierta plausibilidad, más potentes y aceptables para dejar a salvo la dignidad del método y su imagen pública (Good 1958; Lindley 1958; Sterling 1960; Savage y otros 1962; Clark 1963: 467; Greenwald 1975). Con posterioridad incluso este empeño modesto parece haber perdido prioridad. De los muchos aportes de Bakan a la ciencia y a la política académica subsiste su convicción de que en la prueba de hipótesis es importante considerar en qué dirección se comete un error, pues bien podría suceder que un error en determinada dirección carezca de consecuencias, pero en el sentido contrario sus consecuencias sean catastróficas. Algunos científicos, entre paréntesis, piensan que a pesar de sus premisas alarmantes este examen sólo puede derivar en un ejercicio vacío (p. ej. Shaver 1993: 29).

El régimen de razonamientos con el que entronca la idea de las hipótesis de la nada (o la propuesta de Guttman en el sentido de que la nulidad debería ser una hipótesis alternativa) es similar al que propuso Paul Meehl (1990b: 204-210) bajo el nombre del factor de impureza [*crud factor*]. Este factor expresa que en las ciencias de lo viviente, desde la biología a la sociología, casi todas las variables que podemos llegar a medir están correlacionadas en alguna medida. Con muestras suficientemente amplias y representativas, la armonía Leibniziana –alega Meehl– desaparece por completo. Casi lo mismo sostenía David Lykken (1968) treinta años antes cuando hablaba del "ruido correlacional ambiente" propio de los muestreos de gran calado en ciencias sociales. Cuanto más se refina la investigación o se incrementa el repositorio de los datos, asegura Lykken, si se apuesta en contra de la pureza platónica de los números perfectos y los patrones distribucionales simétricos no hay forma de salir perdiendo.[23]

[23] Como habría dicho Bateson (1980: 12), "esto me recuerda una historia". En uno de sus artículos más clásicos, Claude Lévi-Strauss alega que el psicoanálisis cree confirmar lo que argumenta cuando encuentra el síntoma que está buscando; cuando no lo encuentra, afirma que el síntoma existe de todos modos pero que se lo ha reprimido. A lo que voy es a que (como concluía genialmente el maestro) una dialéctica que gana a todo trance siempre encuentra el modo de llegar a la significación (Lévi-Strauss 1995 [1955]: 130).

Lo que más me preocupa, a todo esto, es que la HN no detenta el monopolio de la trivialidad. Mientras todo el mundo está distraído en espera de que la HN sea tan estúpida como en ocasiones lo es, muchas veces es la hipótesis alternativa la que aporta la excusa para llevar adelante experimentos estadísticos superfluos hasta el extremo de lo vergonzante. El ecólogo y naturalista Fred Guthery aporta la crónica siguiente:

> Las experiencias aparentemente sin sentido de la NHST se siguen manifestando con sorprendente frecuencia. Recientemente he visto pruebas de significancia para determinar si los tratamientos de irrigación afectan la germinación de las semillas (Cornaglia y otros 2005), si el líquido corporal afecta la masa de un carnívoro (Hwang y otros 2005), si el podado afecta la altura de la vegetación herbácea (Washburn y Seamans 2007) y si la tala afecta la densidad de árboles en una selva (Morris y Maret 2007). Un *paper* (inédito) reporta que el tiempo de descanso de una especie de pájaros se correlaciona ($p<0{,}0091$) con la hora de la puesta del sol. Asistí a un seminario en el que un orador reportó que la enfermedad reduce el número de días que la gente trabaja ($p<0{,}05$). […] No hay más remedio que adherir al ritual de hacer pruebas estadísticas de lo inevitable. Aplicando esas pruebas, la ciencia de la vida silvestre despliega una parodia de sí misma (Guthery 2008: 1872).

También es Guthery quien aporta este ejemplo reciente de lo que él propone llamar seudo-diferencias. Éstas se pueden reproducir en contados minutos, con fulminante efecto pedagógico, en una planilla de Excel:

> A medida que crece el tamaño de la muestra, cualesquiera diferencias entre ≥ 2 medias, sin que importe cuan pequeñas sean, devendrán estadísticamente significantes. […] Para decirlo de un modo distinto, todas las diferencias distintas de cero entre 2 medias son estadísticamente significantes en el límite. En un registro parecido, recientemente hice un análisis de regresión (inédito) de las propiedades de puntos usados *versus* puntos al azar. La estimación de un coeficiente generada por un software estadístico fue 0,000 ($p<0{,}001$, n=751). Esto podría llamarse, un poco paradójicamente, un efecto nulo estadísticamente significante (Guthery 2008: 1873).

Revisando la bibliografía se encuentra que ya Meehl (1967: 109) había demostrado algo de lo mismo hace ya mucho tiempo. Trabajando con datos de 55.000 estudiantes de Minnesota encontró "relaciones estadísticamente significantes en el 91% de las asociaciones entre una congerie de pares de 45 variables misceláneas tales como sexo, orden de nacimiento, preferencia religiosa, número de herman@s, elección vocacional, pertenencia a clubes, elección de colegios, educación de la madre, danza, interés en la talla en madera, gusto por ir a la escuela, etcétera". Esto no es más que un efecto matemático trivial que se deriva del valor de los parámetros usados en el cálculo. Después de todo, el error estándar –por ejemplo– se calcula como $(s^2/N)^{1/2}$, lo cual hace que si la muestra es suficientemente amplia cualquier cosa diferirá de cualquier otra: la inversa de la raíz cuadrada de un núme-

ro muy grande es a fin de cuentas un valor muy pequeño (McCloskey 1985: 202). Esto quiere decir que cualquier científico social o arqueólogo con muestras de tamaño suficiente tiene todas las probabilidades a su favor para demostrar lo que se le ocurra, sea ello (glosando a Guthery) una opinión trivial relativa a la importancia del agua en la navegación o (reivindicando a Arbuthnott) la hipótesis que afirma la existencia de Dios.

Para acabar de rizar el rizo alrededor de la idea de nulidad y como si los embrollos revisados hasta aquí hubiesen sido pocos, al lado de la definición que viene desde Fisher se habla también de HN para indicar que un parámetro es cero, o que las diferencias entre las medias de la población es cero, o que las diferencias de las proporciones en la población es cero (Lindquist 1940: 15). Aunque estas acepciones divergentes testimonian un desborde semántico que se diría fuera de lugar en un modelo cuantitativo y aunque mejor sería acaso disponer de una definición estable para cada concepto, al final del día la situación no parece implicar un problema tan tremendo. Hasta donde sé, en todas las ciencias sucede aproximadamente lo mismo; el carácter mutable y resbaloso de la significación de los conceptos no es, en las disciplinas formales, algo que desvele a muchos espíritus; tampoco debería ser fuente de preocupación en estadística, sobre todo cuando en torno de la NHST todavía hay latentes y sin asomos de resolución dilemas que, como se verá en seguida, resultan ser órdenes de magnitud más fastidiosos que los meros significados confusos de las palabras.

9. Los valores de *p*

La discusión en torno a los valores de *p* y a su significación es una de las más vivas y recurrentes en el campo de los enfrentamientos referidos a la NHST. No es inusual que se encuentren textos con nombres tales como “Criteria for selecting a significance level: On the sacredness of .05” (Labovitz 1968), “The sacredness of .05: A note concerning the uses of statistical levels of significance in social science” (Skipper, Guenther y Nash 1970), “Confidence intervals rather than *p* values: estimation rather than hypothesis testing” (Gardner y Altman 1986), “The irreconcilability of *P* values and evidence” (Berger y Sellke 1987), “The end of the *p* value?” (Evans, Mills y Dawson 1988), “A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age” (Loftus 1993), “The earth is round ($p< .05$)” (Cohen 1994), “Toward evidence-based medical statistics: 1. The *p* value fallacy” (Goodman 1999), “What your statistician never told you about *P*-values” (Blume y Peipert 2003), “Incongruence between test statistics and *p* values in medical papers” (García-Berthou y Alcaraz 2004), “A farewell to p-values?” (Moran y Solomon 2004), “A dirty dozen: Twelve *P*-value misconceptions” (Goodman 2008), “Exposing the P value fallacy to young residents” (Sestini y Rossi 2009) y “Much ado about the p value” (van der Pas 2010).

En el marco de la prueba de hipótesis frecuentista, el valor de *p* se define como la probabilidad de observar eventos tanto o más extremos que los que se manifiestan en los datos observados en caso que la hipótesis nula (H_0) sea verdad. Si el valor de *p* es suficientemente pequeño (característicamente $p \leq 0{,}05$) puede decirse que los datos proporcionan evidencia contra la HN, la que debe ser rechazada. Técnicamente *p* se define entonces como el más pequeño nivel de significancia que lleva al rechazo de la HN con probabilidad 1 (Lehmann y Romano 2005: 58, 64).

Sin embargo, el valor de *p* no es –como se ha echado a rodar– la probabilidad de que la HN sea falsa; es más bien una medida indirecta para evaluar si la HN es verdadera o no (Berger y Sellke 1987: 114; Lee 2011: 1). En notación de probabilidades condicionales esto sería $p(\mathrm{D}|H_0)$, o sea la probabilidad de obtener los mismos datos si H_0 se sostiene; de ningún modo puede ser $p(H_0|\mathrm{D})$ (o sea la probabilidad de que la HN sea verdad, dados los datos), que es como a menudo se lo interpreta (Nickerson 2000: 247).

El hábito de vincular *p* con 0,05 se consolidó en las antiguas tablas publicadas por Fisher (1925) en su libro de métodos estadísticos para el investigador. Esas tablas, más compactas y sencillas que otras de género parecido, se convirtieron en estándares para los usos más diversos. El lector puede encontrar tabulaciones parecidas en libros de metodología antropológica como el clásico manual de H. Russell Bernard (1995: 529-534). En los años 60 se hizo costumbre reportar con un asterisco (*) los valores de $p \leq 0{,}05$, dos asteriscos (**) para $p \leq 0{,}01$ y tres asteriscos (***) para $p \leq 0{,}001$, lo cual,

por la familiaridad que tiene con las calificaciones de filmes en rangos como 👍, 👍👍 o 👍👍👍, ayudó a que se popularizara la notación de estrellas [*star notation*] sobre todo en los *papers* de sociología (Leahy 2005). En algún momento se agudizó un enfrentamiento gulliveriano entre los partidarios de los asteriscos y los adeptos de la notación numérica que algunos asimilaron irónicamente a una "guerra de las estrellas" [*star wars*] (Gregoire 2001: 2).

A despecho de esos gestos (que no sé si juzgar de presunción o de condescendencia), en las disciplinas más diversas muchos autores encuentran que la idea del valor *p* es inaceptable o ridícula desde el vamos, dado que se trata de una entidad teórica que es explícitamente condicional al hecho de que la HN sea verdadera, lo que en muchísimos contextos es imposible. Como fuere, la crítica no impidió que se desarrollara la preceptiva, aunque ésta no es en modo alguno unánime. El valor de *p* se obtiene mediante una prueba estadística (habitualmente F, t, z o χ^2) en la cual, naturalmente, no se pueden tomar en cuenta los valores de variables que no hayan sido objeto de observación o que se hayan eliminado por ser *outliers* (Anderson, Burnham y Thompson 2000: 914).[24]

A esta altura ya se va poniendo en evidencia que (por más que esté connotando la improbabilidad de obtener datos como los considerados si la HN fuese verdad) el valor de *p* tampoco es una medida exacta de la fuerza de la hipótesis alternativa; más bien es un indicador contrafáctico del grado de consistencia o inconsistencia de los datos respecto de la HN. Decir contrafáctico es empero decir lo menos; tal como salió a relucir cuando se debió explicar a la Suprema Corte de los Estados Unidos cómo era el caso, la lógica de la prueba estadística está decididamente alborotada: se basa en presuponer primero que no existe ninguna vinculación y en determinar luego la posibilidad de que solamente el azar haya producido los resultados que se toman como punto de partida (Wall Street Journal 2010).

La discusión sobre los valores aceptables o inaceptables de *p* ha llevado a considerar métodos de prueba alternativos a los que se les han asignado méritos variables en tanto herramientas para deslindar evidencia científica:

> Hay muchos tipos de pruebas de significancia. Una clase importante consiste en pruebas (usualmente pruebas *t*) de hipótesis nulas en las que no se cree y a las que uno espera rechazar en beneficio de afirmaciones sustantivas. Esta es la práctica estadística del "colesterol malo" que se cree peligrosa para la salud del campo. Se dice que el uso de valores de *p* a partir de pruebas de la hipótesis nula confunde a la gente por cuanto acarrea significados no deseados; peor todavía, el valor de *p* es altamente sensible a diferencias arbitrarias en el tamaño de la muestra y por lo tanto no puede ser considerado una propiedad intrínseca del fenómeno bajo

[24] Los *outliers* se definen y discuten más adelante, pp. 94 y ss.

> estudio. Sin embargo hay también un "colesterol bueno", una prueba de bondad de adecuación [*goodness of fit*] de los modelos a los datos, llevada a cabo sobre residuos, usualmente utilizando pruebas de *chi* cuadrado con un gran número de grados de libertad. Los modelos log-lineales de datos categóricos […] y los modelos de estructura de covarianza son dos de los tipos de análisis de esta clase (Abelson 1997: 12).

Una vez más, los cuestionamientos a la imposición de valores de *p* que se saben arbitrarios no han sido óbice para que muchas revistas especializadas estipulen que $p \leq 0{,}05$ es el valor que decide la aceptación o el rechazo ya no de la HN sino de los artículos académicos (p. ej. Melton 1962; Altman 1998; García-Berthou y Alcaraz 2004; Altmann 2007: 6). Incluso los estudios que en algún momento cuestionan las políticas editoriales a este respecto consignan obedientemente los valores de *p* y los márgenes de error de sus propios *surveys* argumentativos (Goodman, Altman y George 1998; Strasak y otros 2007).

El valor numérico con el que se contrasta *p* ha venido descendiendo con el correr de los años. El problema con esta propensión a la exactitud micrométrica es que mientras los valores declinantes trasuntan una cierta imagen de refinamiento progresivo, concomitantemente se sigue escamoteando toda información sobre la potencia estadística y los tamaños del efecto de la muestra.[25] Esta dualidad encubre el hecho de que la potencia de las pruebas estadísticas en la literatura tienda a la baja, llegando a menos de 0,50 para un efecto mediano, lo cual no es mucho más eficiente que revolear una moneda. Que autores de diversa orientación lo señalaran (Cohen 1994; Gigerenzer 1993) no ha logrado cambiar la situación. Un cuarto de siglo después que Cohen (1962) publicara su descubrimiento, la potencia de las pruebas de HN no ha hecho más que debilitarse (Sedlmeier y Gigerenzer 1989). Tampoco se ha reflexionado gran cosa sobre su sentido. "Más bien, las hipótesis nulas se arman y ponen a prueba de una manera en extremo mecánica, reminiscente del lavado compulsivo de manos" (Gigerenzer 1993).

Fóbica o no, la digresión fantasiosa alrededor de los valores de *p* ha llegado a extremos irrisorios. En un libro de texto bien conocido, *Introduction to statistics for psychology and education*, Jum Nunnally asevera que mantener *p* en el nivel de 0,05 ya es una exigencia superada por el progreso científico.[26] "Hasta hace 20 años no era poco común ver importantes reportes de investigación en

---

[25] La potencia estadística se ha descripto en la pág. 35; el tamaño del efecto se define en la pág. 101.

[26] Es verdad, sin embargo, que para establecer los ajustes necesarios para garantizar la independencia mutua de los datos el valor de *p* debería ajustarse en torno a 0,005 u otras cifras extremas. Pero esos requerimientos, ligados a la problemática planteada en el famoso "problema de Galton" no solamente afectan al valor de *p*; tampoco se especificaron hasta casi veinte años después de publicado el manual de Nunnally (cf. Dow [1993] y más adelante, pág. 91). Como sea, es cierto también que en la era de las computadoras se pueden alcanzar rutinariamente niveles mucho más altos de precisión; sin embargo (como lo documenta Huberty

los cuales las diferencias eran significantes sólo a nivel de 0,05. Hoy en día esos resultados no se toman en serio, y es costumbre ver que se publican resultados sólo si alcanzan niveles de 0,01 o aún más bajos" (Nunnally 1975: 195). No es de sorprender que este autor haya sido el mismo que sostuvo y popularizó la idea de que el nivel de significancia especifica la probabilidad de que un resultado de investigación sea replicado.

Orientado por estos indicios, Gerd Gigerenzer encontró en unas pocas páginas del mismo libro de Nunnally un caudal de observaciones oscuras e incorrectas sobre los múltiples significados del nivel de significancia que vale la pena citar en extenso:

> Nunnally explicó que el "nivel de significancia" significa todo lo siguiente: (a) "Si la probabilidad es baja, la hipótesis nula es improbable" (p. 194); (b) "la improbabilidad de que los resultados observados se deban a un error" (p. 195); (c) "la probabilidad de que una diferencia observada sea real" (p. 195); (d) "la confianza estadística […] con marcas de 95 sobre 100 de que las diferencias observadas se mantengan en otras investigaciones" (p. 195); (e) el grado en que los resultados experimentales se tomen "seriamente" (p. 195); (f) "el peligro de aceptar un resultado estadístico como real cuando en realidad sólo se debe a error" (p. 195); (g) el grado de "fe (que) se puede tener en la realidad del hallazgo (p. 196); (h) "la hipótesis nula se rechaza al nivel de 0,05"; y (i) "el investigador puede tener 95 por ciento de confianza en que la media de la muestra difiere en realidad de la media de la población". Después de las últimas dos versiones el autor asegura a sus lectores: "Todas estas son formas diferentes de decir lo mismo" (p. 196).

Sólo en las interpretaciones pos-estructuralistas o morinianas de la ciencia de la complejidad se puede encontrar un caudal de desatinos semejante (cf. Reynoso 2009). El problema con todo esto es que Nunnally está mucho más cerca de ser un autor representativo que un caso marginal.

A lo que voy es a que, en el límite, el hecho que ideas equivocadas sean representativas acaba materializando el principio de Goebbels, en el sentido de que lo que se repite mil veces puede terminar consagrado por aclamación, cristalizando en paradigma y hasta resultando verdad: en el precario estado actual de la reflexión estadística, por cada definición descaminada es posible hallar unas cuantas autoridades que la respaldan. Con toda la deferencia que ellos puedan merecerme en otros órdenes, diversos autores profusamente publicados y re-editados replican una y otra vez ésas y otras formas de la falacia, como podrá comprobar dolorosamente quien siga el rastro de las referencias que aquí acompaño (cf. Lindquist 1940: 14; Guilford 1942: 156-166; Underwood y otros 1954: 107; Ferguson 1959: 13, 133; Anastasi 1958: 9; Wilson 1961: 230; Bolles 1962; Melton 1962: 553; Miller y Buckhout 1973: 523 [Apéndice de F. L. Brown]; Pelto y Pelto 1975: 162-164; Thomas 1976:

1993: 328), y aunque las tablas impresas en papel ya casi no se usan, se sigue utilizando por ejemplo $p < 0{,}05$ en lugar de (digamos) $p = 0{,}029$.

459-468; Schuchard-Fischer y otros 1982: 83; Wyss 1991: 547; Ellison 1996; Currell y Dowman 2009: 250). Hay referencias adicionales a estudios equivocados en Falk y Greenbaum (1995), Gigerenzer (2000: cap. 13) y Nickerson (2000: 242).

Estos problemas en particular no se solucionan aprendiendo estadísticas. Contrariamente a lo que sugeriría el sentido común, se ha encontrado que no hay grandes contrastes entre los novicios, los practicantes de nivel intermedio y los estadísticos experimentados en cuanto a interpretar adecuadamente las pruebas de significación. Lecoutre, Poitevineau y Lecoutre (2003), tras investigar este tópico en profundidad y con amplio sustento experimental, sostienen que a diferencia de lo que alegan Schmidt (1995) y Goodman (1999) no es una tarea fácil, incluso para estadísticos profesionales, interpretar los valores de *p* "de una manera razonable". No siempre es posible, prosiguen los autores, explicar los errores de exégesis debido a la falta de maestría técnica, ya que los expertos, documentadamente, también la yerran por amplio margen. Aunque Lecoutre y sus coautores son partidarios de mantener las pruebas en vigencia, lo más probable –concluyen– es que los resultados negativos revelen la inadecuación constitutiva de la NHST respecto a las necesidades conceptuales de la investigación en la vida real.

Tomando alguna distancia y echando sobre el problema del significado de los valores de *p* una mirada distante antropológica, por así decirlo, percibo tanto en la práctica de la NHST como en su crítica una complicación a la que no se ha prestado la atención debida. Siendo que cualquier afirmación sobre el sentido de esos valores (incluyendo las que se reputan correctas) puede impugnarse ya sea en función de lo que afirmara por un lado Fisher o por el otro Neyman y Pearson, considerando que la NHST "híbrida" se fue armando a través de manuales que reflejan pensamientos construidos colectivamente antes que las palabras textuales de esos padres fundadores, y teniendo en cuenta que la replicación de definiciones derivativas es aluvional, lo que cabe preguntarse primero que nada es si en realidad existe una definición adecuada y monolítica del concepto o si más bien hay (como los llaman los fenomenólogos) múltiples universos finitos de sentido igualmente legítimos o (como se dice en informática) diversos estándares de facto en circulación.

10. El arte de la interpretación sistemáticamente indebida

Al lado de las ambigüedades intrínsecas a lo que podríamos llamar el desciframiento cuantitativo de la significancia, he encontrado una serie de equívocos que son claramente de naturaleza hermenéutica pero que se salen del cuadro de la confusión entre significado y significancia. Larry Daniel (1998) de la Universidad del Norte de Texas, por ejemplo, ha definido cinco percepciones erróneas de este tipo:

- La percepción errónea de que la significancia estadística informa al investigador sobre la probabilidad de que un resultado determinado sea replicable ("la fantasía de la replicabilidad" de Carver [1978]).
- La percepción errónea de que la significancia estadística informe al investigador sobre la posibilidad de que los resultados se deban al azar (o, como lo llama Carver [1978: 383] "*the odds-against-chance phantasy*").
- La percepción errónea de que un resultado estadísticamente significante indique la probabilidad de que la muestra sea representativa de la población.
- La percepción errónea de que la significancia estadística sea la mejor forma de evaluar los resultados estadísticos.
- La percepción errónea de que los coeficientes estadísticamente significantes de confiabilidad y validez basados en cifras en una prueba que se administren a una muestra dada impliquen que la misma prueba obtendrá puntajes válidos o confiables trabajando con una muestra diferente.

Estas hermenéuticas fallidas se han documentado fehacientemente mediante un número crecido de experimentos. Oakes (1986), Haller y Krauss (2002) y Gigerenzer y otros (2004) describen un caso dramático referido a una prueba en la que se pide a un grupo de estudiantes y profesores que respondan a un cuestionario como el que sigue:

> Supongamos que usted tiene un tratamiento que sospecha que puede alterar la performance de una cierta tarea. Usted compara las medias de su grupo de control y del grupo experimental (digamos, unos 20 sujetos en cada muestra). Más todavía, supongamos que usted usa una prueba simple de *t*-test de las medias y que su resultado es significante (t=2,7, df=18, p=0,01). Por favor marque cada una de las afirmaciones de más abajo como "Verdadera" o "Falsa". "Falsa" significa que la afirmación no se sigue de las premisas. Nótese que varias de las afirmaciones (o ninguna de ellas) pueden ser correctas.
>
> (1) Usted ha des-probado absolutamente la hipótesis nula o sea, no hay diferencias entre las medias. □ Verdadero - Falso □
>
> (2) Usted ha encontrado la probabilidad de que la hipótesis nula sea verdad. □ Verdadero - Falso □

(3) Usted ha probado absolutamente su hipótesis experimental (que existe una diferencia entre las medias de las poblaciones). □ Verdadero - Falso □

(4) Usted puede deducir la probabilidad de que la hipótesis experimental sea verdad. □ Verdadero - Falso □

(5) Usted conoce, si decide rechazar la NH, la probabilidad de que tome la decisión equivocada. □ Verdadero - Falso □

(6) Usted tiene un hallazgo experimental confiable tal que si, hipotéticamente, el experimento se repitiera muchas veces, usted obtendría un resultado significativo 99% de las veces. □ Verdadero - Falso □

Haller y Krauss plantearon el test a 44 estudiantes avanzados de psicología, 39 *lecturers* y profesores de psicología y 30 profesores de estadísticas. Todos los estudiantes habían tomado cursos de estadística en los que se practicaba NHST; todos los profesores enseñaban rudimentos del tema en sus materias. Los participantes en la prueba provenían de seis destacadas universidades alemanas.

Afirmaciones (abreviadas)	Alemania 2000			UK 1986
	Estudiantes de psicología	Profesores e instructores que no enseñan estadísticas	Profesores e instructores que enseñan estadísticas	Profesores e instructores
1. H0 desprobada absolutamente	34	15	10	1
2. Se encuentra probabilidad de H0	32	26	17	36
3. H1 absolutamente probada	20	13	10	6
4. Se encuentra probabilidad de H1	59	33	33	66
5. Probabilidad de la decisión errónea	68	67	73	86
6. Probabilidad de replicación	41	49	37	60

Tabla 2 – Porcentajes de respuestas incorrectas – Datos de Alemania y UK (Oakes 1986)

Para hacerla breve, diré que los resultados de la prueba fueron mucho más que decepcionantes. Aunque los guarismos de la Tabla 2 puedan llamar a engaño, ni uno solo de los estudiantes advirtió que todas las aserciones eran falsas: cada uno de ellos homologó una o más de las ilusiones en vigencia. El 90% de los instructores y profesores también compartía esas ilusiones; en el caso de los profesores de estadísticas la cifra descendió apenas al 80%. En promedio, los estudiantes sustentaron 2,5 ilusiones por cabeza, sus profesores no estadísticos 2,0 y los estadísticos 1,9 sobre 6 posibles. Esto implica que en cada uno de los estamentos el porcentaje de error fue del 41%, 33,3% y 31,6% respectivamente. Aunque la extrapolación puede que no sea del todo legítima (y aunque los valores de *p* no sean indicadores de la probabilidad de error) esto excede por amplio margen el 5%

de significancia que los estadísticos conceden como valor máximo aceptable para sus propias pruebas estadísticas.

En otro experimento, Falk y Greenbaum (1995) encontraron medidas parecidas entre estudiantes de Israel, no obstante agregar la opción de "Ninguna de las afirmaciones es correcta" y haber establecido como bibliografía previa el artículo clásico de David Bakan (1966) sobre las perversiones más comunes del razonamiento estadístico. El problema empero es que una alta proporción de los textos usuales de estadística para las ciencias sociales incurren en errores semejantes. En una ponencia titulada "Fantasía de la prueba de significancia estadística en manuales introductorios de psicología" J. C. McMan (1995) identificó errores sustanciales de interpretación de la NHST en 24 libros publicados entre 1964 y 1995. Una especificación representativa es la siguiente:

> "Aceptar el nivel 0,05 de significancia" al rechazar la hipótesis nula significa que 95 veces de cada 100 estaremos en lo correcto con nuestra decisión, pero, 5 veces de cada 100, corremos el riesgo de rechazar la hipótesis nula cuando de hecho es verdad ([Louis] Cohen y Holliday 1982: 124).

La culminación del método de la interpretación sistemáticamente fallida (aunque con alguna redundancia y con referencias a algún material que ya hemos visto) se documenta en el artículo del bayesiano Steven Goodman (2008) sobre las doce equivocaciones hermenéuticas en torno a (una vez más) los valores de *p*. Según Goodman las concepciones erróneas más conspicuas en torno suyo y las respuestas que corresponde darles (con comentarios míos agregados) serían:

- Si $p = 0{,}05$ la HN posee sólo un 5% de probabilidades de ser correcta. → Esta pasa por ser la falla más pervasiva y perniciosa de todas, por cuanto perpetúa la falsa idea de que los datos por sí solos, sin que medie teoría alguna, nos pueden decir en qué medida estamos en lo cierto o no en nuestras conclusiones. Otra forma parecida de decir lo mismo sería aseverar que cuanto menor es el valor de *p*, más fuerte es la prueba contra el modelo nulo. En cualquiera de sus redacciones esta idea es falsa porque el propio valor de *p* se calcula bajo el supuesto de que la HN es verdadera, por lo que no puede en modo alguno ser indicador de la medida de su falsedad. Por más que resulte evidente que esta concepción es equivocada, me he sorprendido de encontrarla en al menos un texto de Sir Ronald Fisher (1970: 80).
- Una diferencia no significativa (por ejemplo $p > 0{,}05$) significa que no hay diferencias entre ambos grupos. → Una diferencia no significativa significa sólo que un efecto nulo es estadísticamente consistente con los resultados observados. Esto no ocasiona que el efecto nulo sea más probable. El efecto mejor soportado por los datos en cualquier experimento siempre son los datos observados, cualquiera sea su significancia.

- Un hallazgo estadísticamente significante es clínica (o antropológicamente) importante. → El valor de *p* no contiene información sobre la magnitud de un efecto. Aunque cueste creerlo, la significancia de una prueba nada dice sobre relaciones de causalidad ni (por abrumadora que parezca ser) sobre la magnitud de un efecto. Aunque he podido localizar algunos cientos de documentos que afirman lo contrario (las más de las veces reinterpretando la NHST) ninguna estadística puede expedirse sobre causalidad; para referirse a ello se requieren bases analíticas, lógicas y sustantivas a los que los modelos estadísticos no se atienen por definición y que sólo los modelos mecánicos pueden articular (Cohen y Cohen 1983: 210; Shaver 1993: 16-18; Reynoso 2006; 2011: 23-30). Presuponer que la correlación es prueba de causación es y será por siempre una falacia bien conocida bajo cualquier modelo lógico: *cum hoc ergo propter hoc*. La prueba estadística en sí sólo se refiere a los datos y es por completo ciega e indiferente respecto de los enunciados de cualesquiera hipótesis, impliquen ellas o no algún régimen de causalidad.
- Estudios con valores de *p* en lados opuestos de 0,05 son recíprocamente conflictivos. → Dependiendo de diversos factores, los estudios pueden dar medidas diferentes de significancia aun si su significancia efectiva fuese idéntica.
- Estudios con el mismo valor de *p* proporcionan la misma evidencia contra la hipótesis nula. → Efectos observables dramáticamente distintos pueden tener el mismo valor de *p*.
- 0,05 significa que hemos observado datos que ocurrirían sólo 5% de las veces bajo la hipótesis nula. → El problema es que el valor de *p* concierne a valores iguales a los observados o a otros más extremos que son no-observados por definición y que introducen dificultades operacionales extremas.
- $p=0{,}05$ y $p\leq 0{,}05$ significan lo mismo. → Esta proposición ilustra en qué medida es diabólicamente difícil ya sea explicar o entender los valores de *p*.
- Los valores de *p* se deben escribir como desigualdades, por ejemplo $p\leq 0{,}02$ cuando $p=0{,}015$. → Si lo que se busca es comunicar la fuerza de la evidencia, quizá sea mejor reportar el valor exacto. Algunos estadísticos, sin embargo, se han manifestado contrarios a lo que consideran precisión espuria, sugiriendo limitarse al primer decimal significativo, convenientemente aproximado (p. ej. Vickers 2010: 100).

- 0,05 significa que si usted rechaza la HN la probabilidad de un error de Tipo I es sólo del 5%. → Aquí estamos metiéndonos en arena movediza; esta falacia es lógicamente similar a la §1, aunque eso sea difícil de visualizar.
- Con un umbral de significancia de $p = 0{,}05$ la probabilidad de un error de Tipo I es del 5%. → Si sabemos que la HN es falsa, no hay posibilidad de un error de este tipo.
- Se debería usar un valor de *p* de una sola cola cuando a uno no le interesan los resultados en una dirección, o cuando una diferencia en esa dirección es imposible. → Este es un asunto muy complejo sobre el que se ha discutido mucho y sobre el que no hay acuerdo entre los especialistas.
- Una conclusión científica o una política de tratamiento debería basarse en determinar si el valor de *p* es o no significante. → Esta falacia abarca a todas las demás y equivale a decir que la magnitud de un error no es relevante.

A medida que se recorre la literatura se comprueba que los malentendidos surgen a cada paso. En este terreno toda desmentida parece, por así decirlo, de efecto más nulo que el de las hipótesis de la nada. Fuera del caso especial de la medicina los *surveys* que documentan estas fallas no han sido considerados en su conjunto hasta que se escribiera el ensayo que se está leyendo, como si ni siquiera los críticos quisiesen constatar con la asiduidad necesaria y en términos comparativos la incorregibilidad de las prácticas a las que todos estamos expuestos.

11. Pragmática e imagen de la NHST

> El Mal nunca triunfa, porque cuando triunfa se lo llama Bien.
>
> Atribuido a Perich (s/f)

En este apartado corresponde interrogarse, reflexivamente, sobre la utilidad y relevancia de estudios como el presente. Como se infiere del epígrafe, no creo que en la discusión científica tal como se ha venido dando resulte siempre técnicamente fácil o ideológicamente ecuánime tomar partido a favor de la verdad, dado que ni las cosas son tan simples ni hay solamente dos facciones internas en pugna. En primer lugar, en estadísticas rara vez es posible dirimir definitivamente una cuestión por el expediente de la prueba matemática, a fuerza de teoremas, lemmas, contrapruebas y generalizaciones; de allí que las contiendas disciplinarias acostumbren ser tanto o más feroces de lo que es el caso (por ejemplo) en la antropología. Por otro lado, y como ha sucedido otras veces (y el colapso de la antropología matemática lo ilustra con elocuencia), bien podría suceder que los efectos colaterales de un posicionamiento inflexible en contra de la NHST acaben siendo funcionales a intereses de talante anticientífico hace rato aposentados en las disciplinas empíricas, los cuales ni siquiera admitirían la necesidad de contar con herramientas de modelado cualesquiera fuesen los formalismos que están en juego (cf. D'Andrade 2000). La pregunta que cabe hacerse, de todos modos, es, kuhnianamente, si la NHST se verá en efecto reformulada en razón de la abrumadora acumulación de anomalías documentadas en su contra, o si en el futuro prevalecerá la inercia y todo seguirá siendo más o menos como lo ha sido hasta hoy.

A esta pregunta muy pocos responden de maneras matizadas. En el campo de la educación matemática, Rama Menon (1993) de la Universidad Tecnológica de Nanyang, ha propuesto que la prueba de significancia estadística sea discontinuada de los programas académicos. A tal efecto pone en duda lo que él llama cinco mitos en torno de la prueba, y que son (1) que se trata de un método parecido a una receta y libre de controversias que permite una limpia toma de decisiones, (2) que proporciona respuesta a la pregunta sobre si una baja probabilidad en los resultados de una investigación se debe al azar, (3) que su lógica equivale a la de la prueba matemática por contradicción, (4) que tiene algo que ver con cuestiones de confiabilidad y replicabilidad, y (5) que es una condición necesaria pero no suficiente que atañe a la credibilidad de los resultados.

Sin embargo, y tal como lo reportan Gigerenzer, Krauss y Vitouch (2004: 391), los editores de las revistas científicas más prestigiosas en ciencias humanas han convertido a la NHST en la condición necesaria para la aceptación de *papers* y han promovido los valores pequeños de *p* como el signo de la experimentación de excelencia. La prueba se ha convertido en una especie de ritual, cuyos pasos

esenciales ya hemos visto descriptos y sujetos a cuestionamiento. Lo paradójico del caso es que la NHST no forma parte de las estadísticas que se encuentran matemática o lógicamente bien fundadas:

> El ritual nulo no se origina ni en Fisher ni en ningún otro estadístico renombrado y no existe en la estadística propiamente dicha. En vez de eso fue fabricado en la imaginación de los escritores de libros de texto de estadísticas en psicología y educación (loc. cit.).

Un número de autores más elevado del que consentiría encapsularse en una hipótesis nula también hablan del procedimiento de la NHST como de un ritual y de la práctica de la prueba de significancia como un culto (Rozeboom 1960: 416; Thomas 1978: 233; Salsburg 1985; Warren 1986; Cohen 1990; Shaver 1993: 23; Huysamen 2005; Lecoutre, Lecoutre y Poitevineau 2001: 400; Gigerenzer 2004: 588-589; Marx 2006; Guthery 2008: *passim*; McCloskey y Ziliak 2008; Marewski y Ohlsson 2009).

Los efectos del crecimiento acrítico en el uso de las técnicas vinculadas con la NHST con o sin complementos se reflejan en esta cita del crítico Gerd Gigerenzer:

> Hace muchos años, pasé un día y una noche en una biblioteca leyendo números del *Journal of Experimental Psychology* de la década de 1920 y 1930. Profesionalmente fue una experiencia de lo más deprimente, pero no porque los artículos fueran metodológicamente mediocres. Por lo contrario, muchos de ellos harían que la investigación actual luzca pálida en comparación con su diversidad de métodos y estadísticas (Gigerenzer 1998a: 201).

La misma sensación suscita, a mi juicio, el cotejo de la antropología de aquel entonces con lo que la disciplina, a contrapelo de las herramientas disponibles, ha llegado a ser el día de hoy. Como quiera que sea, en la década de 1990 la Asociación Americana de Psicología debatió seriamente la posibilidad de sancionar una prohibición de la presentación de resultados basados en la NHST en las publicaciones periódicas de la APA; la propuesta fue rechazada no porque careciera de méritos, sino porque lucía como un acto de censura políticamente incorrecto (Meehl 1997; Johnson 1999: 763). La prohibición más taxativa hasta la fecha proviene de la epidemiología, donde Kenneth Rothman, antiguo director del *American Journal of Public Health* que en los años 80 había predicho la muerte de la epidemiología, llegó a rubricar sin pelos en la lengua esta amenaza insólita:

> En *Epidemiology* pueden mejorar sus perspectivas [de publicación] si omiten referirse a la prueba de hipótesis. […] Toda referencia a la prueba estadística de hipótesis y a la significancia estadística deberá ser removida del *paper*. Les pido entonces que borren los valores de *p* así como todo comentario sobre significancia estadística. Si no están de acuerdo con mi estándar […] se pueden sentir libres de debatir el punto. Como editor, sin embargo, dudosamente puede esperarse que acepte *papers* que se aparten del principio científico que acabo de exponer (Rothman 1986: 9, 559; 1998: 334).

Dicho sea de paso, las referencias a la NHST cayeron del 63% en 1982 (dos años antes que asumiera Rothman) a 6% en 1986 (dos años después que dejara el cargo). Curiosamente, Rothman no se oponía a la publicación ocasional de valores de *p* siempre que se los expresara como igualdades estrictas.

Cualquiera haya sido la virulencia de la crítica y el efecto inmediato de estas decisiones autoritarias al filo de la extravagancia, la NHST se encuentra más consolidada que nunca. No hay evidencia de que la masividad de los ataques que recibió haya tenido efecto sobre su popularidad. El examen de un ejemplar al azar por año del *Journal of Applied Psychology* reveló que el uso de la técnica ha crecido desde un modesto 17% entre 1917 y 1929 hasta un abrumador 94% durante los tempranos 90s (Hubbard 1997; Nickerson 2000: 245). Las cifras registradas para *papers* publicados en el área de mercadeo son casi tan impresionantes, reportando 37,4% para 1960-1969, 65,5% para 1970-1979, 76,6% para 1980-1989, 80,4% para 1990-1999 y 85,3% para 2000-2002 (Hubbard 2005: 9). Entre 1995 y 2000 el 91% de los artículos sociológicos incluyeron análisis de significancia, con el 86% de los autores escogiendo el nivel α del 0,05, 67% el nivel 0,01 y 52% el 0,001 (Leahy 2005: 3). Aislado en una disciplina periférica, hoy se ve que Rothman no fue capaz de lograr un consenso que le permitiera aniquilar el género fuera de su territorio de influencia.

Mientras todo sube algunas cosas caen. Debido sin duda a la influencia de las políticas editoriales, las cifras de publicación de resultados nulos (esto es, la falla en rechazar la HN) pueden caer a un ritmo de hasta un 50% en cada década que pasa. No faltan autores que especulan que aunque la NHST en general sigue creciendo sin mesura, la práctica de publicar resultados nulos en particular puede considerarse una especie en vías de extinción (Hubbard y Armstrong 1992).

Pero tampoco es verdad que todos los medios de publicación muestren un comportamiento homogéneo. Huberty (1993: 329) ha señalado un desfasaje entre los métodos favorecidos por los manuales de estadística por un lado y los *papers* de congresos y revistas especializadas por el otro. En los libros el rol de los valores de *p* a la manera fisheriana ha pasado de un fuerte énfasis en las décadas de 1940 y anteriores a una presencia muy limitada entre la de 1950 y 1970. En los 80s volvió a crecer un poco pero en los 90s se lo veía declinando otra vez. El crecimiento de este cálculo en las publicaciones periódicas (las que no suelen estar ligadas a la estadística en abstracto sino a disciplinas empíricas) sigue desde los orígenes un curso ascendente bajo la guisa del modelo híbrido, como si el método no fuera en absoluto problemático.

Sea cual fuere el bando al cual uno termine concediendo razón, el caso es que en la batalla en torno a la HN no se discierne un vencedor inequívoco. Haller y Krauss (2002) concluyen, con razón, que ambos bandos parecen haber triunfado: por un lado los críticos, porque gran parte de la crítica es sustancial y no ha sido satisfactoriamente contestada; por el otro sus defensores, porque la NHST

todavía se enseña a los estudiantes en las universidades como el método por antonomasia para la evaluación de las hipótesis científicas.

A lo largo de setenta años, los defensores de la NHST han producido un rico acervo de argumentos defensivos que sería injusto no mencionar. Se ha dicho por ejemplo que la culpa de tantas fallas no es imputable al método sino a sus malos usos, y que "el hecho que haya muchos incendiarios en el mundo no hace que el fuego sea una cosa mala"; se ha argumentado también que no existen métodos alternativos que sean la mar de mejores; que en todos los campos del saber se pueden perpetrar usos perversos; que muchas de las críticas son equivocadas o que se deben a una mala comprensión de las estadísticas; que siempre se puede complementar la prueba de hipótesis con algún otro instrumento conceptual o hacerla subsidiaria a otros mecanismos de evaluación de datos; que si bien los razonamientos basados en probabilidades condicionales suelen ser resbaladizos y la afirmación del consecuente es una falacia hay toda suerte de verdades útiles o ideas defendibles dentro y fuera de la estadística que no tolerarían un examen serio de sus mecanismos de inferencia; que "la validez formal tiene poco o nada que ver con la argumentación científica razonable"; que la cosa se arreglaría si pensásemos hipótesis alternativas más precisas e hipótesis nulas no vacías, y que todos deberíamos ser más humildes y reconocer el carácter provisional de las hipótesis que formulamos (Carver 1978; 1993; Chow 1988; 1998; Hagen 1997; 1998; Mulaik, Raju y Harshman 1997: 74; Wainer 1999; Nickerson 2000; Robinson y Wainer 2002).

Como puede verse, la estilística y la retórica dominantes entre los partidarios de la prueba sugieren que, arrinconados por las evidencias, está tomando cuerpo entre ellos un leve espíritu de reforma, una postura que los más recalcitrantes procurarán quizá que degenere en una iniciativa de retorno a las bases, en una alianza estratégica con los bayesianos o en una táctica más conservadora todavía. A ninguno de los adeptos se le cruza por la cabeza la necesidad de impulsar un cambio en profundidad. Ni uno solo entre los argumentos de la defensa, mientras tanto, examina la posibilidad de someter a examen reflexivo los supuestos de muestreo aleatorio, linealidad, distribución normal y representatividad que subyacen a la prueba de hipótesis desde su origen, que atraviesan e impregnan todas sus operaciones, que hasta sus más acérrimos adversarios comparten, que prevalecen cualitativamente idénticas en las estrategias descriptivas de la sociedad, el territorio y la cultura y que ahora pasamos a interrogar.

12. Los abismos de la normalidad y las distribuciones sin media

> [S]e está volviendo cada vez más obvio que la base de la estadística aplicada no es la distribución normal, sino la distribución multinomial y de Poisson. Dejemos la primera a esos estadísticos que trabajan en la asíntote.
>
> Lindsey (1995)

> Sólo cabe reconocer la ocurrencia de la curva normal –la curva Laplaciana de errores– como un fenómeno muy anormal. Es aproximada de manera muy ruda en ciertas distribuciones; por esta razón, y teniendo en cuenta su hermosa simplicidad, podemos, quizá, usarla como una primera aproximación, particularmente en investigaciones teóricas.
>
> K. Pearson (1901: 111)

Los procedimientos que conforman la NHST no pueden prescindir de la comparación de los valores de las respectivas medias de la muestra M_j. En un número crecido y creciente de escenarios, sin embargo, carece de sentido tomar como cota de referencia la media de ciertas clases de distribuciones alejadas de la normalidad. No se trata solamente del hecho de que exista una dicotomía entre las distribuciones normales y el azar dócil por un lado y las leyes de potencia y el azar salvaje por el otro, suficientemente bien caracterizada por Benoît Mandelbrot (Mandelbrot y Hudson 2006); no se trata tampoco de que las distribuciones normales que constituyen implícitamente la carne del método de prueba estadística sean un artefacto construido por las operaciones de muestreo, como bien se sabe desde el surgimiento de los teoremas del límite central (Le Cam 1986; Fischer 2011: 115, 123); y no se trata por último del hecho de que una media no proporcione ninguna indicación sobre las propiedades de los individuos que componen ya sea las poblaciones o las muestras o (aunque más no fuere) sobre las propiedades estructurales del conjunto.[27] Lo que acabo de decir involucra por cierto consecuencias muy graves; pero hace ya tiempo que los científicos cognitivos habían denunciado las falacias que se anidan en las ideas de probabilidad, azar y muestreo, y que son mucho más constitutivas:

> La gente sostiene intuiciones erróneas sobre las leyes del azar. En particular, considera que una muestra tomada al azar de una población es altamente representativa, esto es, similar a la

[27] Más todavía, pretender inferir características de los individuos a partir de rasgos de las poblaciones (o de las muestras) configura la bien conocida falacia ecológica, en la que que invita a incurrir una proporción importante de los descriptores de econometría (ingreso *per capita*), geografía humana (densidad de población), antropología psicológica (personalidad modal), psicometría (uso del coeficiente de inteligencia a nivel de poblaciones como predictor del desempeño de individuos), etcétera (cf. Robinson 1950; Reynoso 2010: 111-158).

> población en todas sus características esenciales. La prevalencia de esta creencia y sus consecuencias infortunadas para la investigación […] han quedado ilustradas en las respuestas de los profesionales […] a un cuestionario concerniente a decisiones de investigación (Tversky y Kahneman 1971).

En los hechos, la idea de un muestro al azar (o del azar sin más) y la de una distribución normal vienen siempre inextricablemente aparejadas. Bajo distintos nombres la idea de la distribución normal se origina hace bastante tiempo, pero se fue asentando hacia la época de Carl Friedrich Gauss [1777-1855], que es como decir no hace mucho. Fue de hecho Francis Galton (quien se reconocía antropólogo, pero que fue más bien antropómetra) quien canonizó de este modo embelesado el irresistible encanto de la distribución normal:

> Apenas si conozco otra cosa que impresione la imaginación tanto como la forma maravillosa del orden cósmico expresada por la "Ley de Frecuencia del Error". La ley habría sido personificada y endiosada por los griegos si la hubieran conocido. Reina con serenidad y con modestia absoluta en medio de la más salvaje confusión […] Cuando quiera que tomemos con la mano una muestra grande de elementos caóticos y los pongamos según el orden de su magnitud, una insospechada y bellísima forma de regularidad prueba haber estado latente en la totalidad (Galton 1889: 66).

El lector reconocerá en el acto de tomar con la mano la operación que hoy se designa como muestreo y en la regularidad de la muestra el efecto de lo que pronto sería el ya mencionado teorema del límite central.[28] Apenas unos pocos años después que lo hiciera Charles Sanders Peirce (1873) (quién si no) y después de intentar apelativos tales como ley de frecuencia del error, ley exponencial (1875), ley de desviación de un promedio, ley de la constancia estadística, montaña expónica y ley de los errores de observación (1869), Galton bautizó la ley normal con el nombre con que la conocemos. Comenzó entonces (como lo llaman Kruskal y Stigler 1997: 93) el "culto de la ley normal", practicado por muchos y resistido por unos pocos (p. ej. Pridmore 1974).

En el análisis de este culto, y anticipando las exploraciones en la retórica científica de los antropólogos posmodernos o de los estudios culturales de la ciencia, Kruskal llegó a prefigurar hasta en los más mínimos detalles lo que algunos años más tarde habríamos saludado sin duda como una poética o una política de la prueba estadística:

> […] Galton y sus contemporáneos, cuando escribían sobre la distribución normal, no sólo utilizaban términos como "la distribución exponencial" sino que, para evitar la monotonía, recurrían a expresiones más vagas: "la distribución usual", "la distribución comúnmente encontrada" y, por supuesto, "la distribución normal". […] Pensamos que de alguna manera "nor-

[28] Sobre este teorema véase más adelante, pág. 94.

> mal" triunfó entre los sinónimos. ¿Por qué debió ganar? Especulo que "normal" es una palabra con connotaciones poderosas y positivas debido a la ambigüedad que existe entre sus dos significados: (1) algo deseable, y (2) algo que se encuentra comúnmente. Un tema mayor en nuestra cultura, después de todo, es la deseabilidad de lo que se encuentra comúnmente, de modo que los dos sentidos se refuerzan entre sí (Kruskal 1978: 227)

Todavía en la década de 1950 el estadístico, químico y tipógrafo W. J. Youden ensalzaba la ley normal en sus tarjetas profesionales con el efecto gráfico que se muestra en la Figura 1.

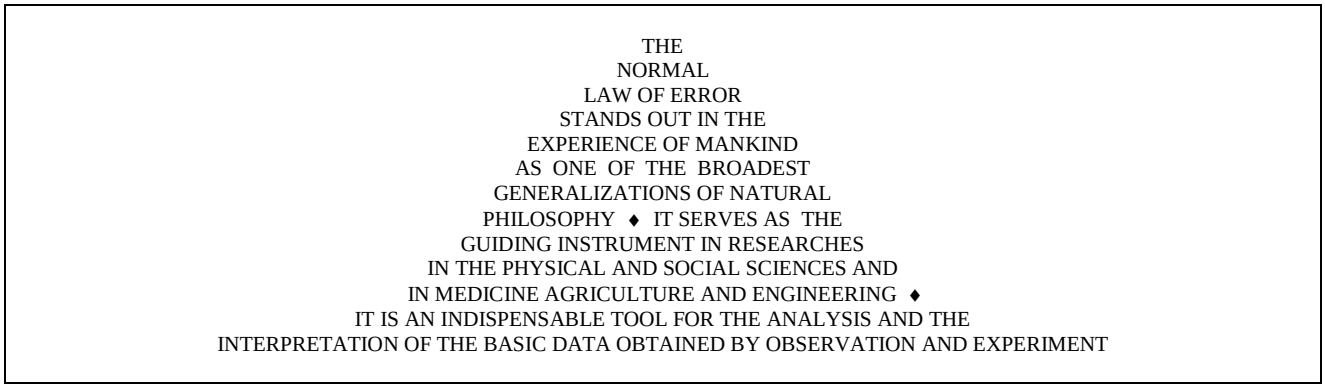

Figura 1 – Elogio de la ley normal por W. J. Youden [1900-1971]
Según Wallis y Roberts (1956: 359)

Pero lo singular del caso en lo que a nuestras disciplinas concierne ha sido que la seducción del análisis estadístico bajo el supuesto de la universalidad de la distribución normal fue el fundamento mediante el cual Émile Durkheim [1858-1917] erigió los rudimentos de las facetas cuantitativas de su sociología, no siempre expresadas bajo la forma de tablas y números. La sociología durkheimiana estaba concebida como una ciencia que permitía distinguir las formas normales de una sociedad de los estados patológicos que daban lugar, por ejemplo, a la anomia (un estado de "falta de normalidad"), el crimen (una "desviación") o el suicidio (el cual "varía en función [linealmente] inversa al grado de integración" a la pauta normal). Inspirándose en el concepto de *l'homme moyen* de Adolphe Quételet [1796-1874], especificado en su *physique sociale* (1835) por referencia al valor de la media de las variables que se atienen a una distribución normal, Durkheim también definía el concepto de normalidad estadísticamente:

> Llamaremos normales a los hechos que presentan las formas más generales y daremos a los otros el nombre de mórbidos o patológicos. Si se conviene en nombrar tipo medio al ser esquemático que se constituiría uniendo en un mismo todo los caracteres más frecuentes con sus formas más frecuentes, se podrá decir que el tipo normal se confunde con el tipo medio y que toda desviación con relación a esta marca de salud es un fenómeno mórbido. […] La mayor frecuencia de lo normal es también la prueba de su superioridad (Durkheim 1895: 41-42, 44).

Los conceptos durkheimianos de patología social y la estadística erigida en base a la distribución normal han sido, en fin, frutos de una episteme monolítica que todavía impera en las ciencias huma-

nas, signada por criterios de equilibrio, simetría, homogeneidad y correspondencia entre los observables y las totalidades. La terminología durkheimiana, cuyo origen oscuro y cuyo raro estilo siguen desconcertando a los historiadores, procede en una proporción apreciable (como acaba de verse) del vocabulario técnico relativo a la ley estadística primordial (Durkheim 1893: libro III; 1895: cap. §3; 1897: libro III, cap. §2). Esta imaginería, sin embargo, no es privativa de las sociologías de cuño conservador. El propio Karl Marx en *El Capital* (Libro I, cap. XIII) elaboraría sus razonamientos sobre el valor del trabajo exactamente en función de los mismos cánones, referencia a Quételet incluida:

> Toda magnitud promedio, […] es meramente el promedio de un número de magnitudes separadas todas de la misma clase, pero difiriendo en calidad. […] Estas diferencias individuales, o "errores" como se las llama en matemáticas, se compensan las unas a las otras toda vez que cierto número mínimo de trabajadores trabajan juntos. […] Estoy muy seguro, a partir de mis mejores observaciones, que cualesquiera cinco hombres aportarán, en total, una proporción de trabajo igual a la de cualesquiera otros cinco; […] esto es, que entre tales cinco hombres habrá uno que posea todos los atributos de un buen trabajador, uno malo y los otros tres mediando y aproximando al primero y al último. Compárese con Quételet sobre el individuo promedio (Marx 1909: 355, nota 1)

La implicancia de que cualesquiera cinco hombres aportarán la misma distribución (normal) que cualesquiera otros cinco, sorprendentemente, reproduce (o más bien *anticipa*) el principio que articula nada menos que el teorema del límite central (cf. más abajo, pág. 94). Siempre he considerado una desdicha que Marx no llegara a tomar contacto con distribuciones más congruentes con la realidad, y en particular con la distribución (continua) de Vilfredo Pareto [1848-1923]. Ciertamente cuando Marx escribió el primer libro de *El Capital* (1867), el único de los volúmenes que pudo publicar en vida, faltaban aun décadas para que se escribieran las obras estadísticas fundamentales en economía y ciencias sociales (cf. Pareto 1896; Kleiber y Kotz 2003: 59-106).

Más de cien años después de Durkheim y Marx, la preceptiva estadística de las ciencias sociales sigue presuponiendo contrariamente a las demostraciones de Pareto, Zipf, Mandelbrot y otros, y a contrapelo de la nueva evidencia, que la distribución normal es dominante en todos los órdenes, que todos estos órdenes son más o menos idénticos en todas partes simplemente porque sí y que el fragmento de cultura a la que el antropólogo tiene acceso (lo mismo que el educto del muestreo que el estadístico tiene entre manos) es *representativo* de la población, del orden global o de la totalidad que circunstancialmente definamos como objeto (Kruskal y Mosteller 1979a; 1979b; 1979c; 1980). El efecto deletéreo de esta creencia no es para nada minúsculo; el uso del formuleo usual para tratar muestras o poblaciones que poseen distribuciones diferentes a la distribución normal involucra una distorsión de extrema magnitud, comenzando por las mismas operaciones de muestreo y por el sen-

tido semántico de la HN, por más que el marco terminológico del que estemos haciendo uso se mantenga en un plano cualitativo y no haga referencia a semejantes parámetros.

Los equivalentes no muestrales, no lineales y no paramétricos para llevar adelante una NHST o son de una dificultad prohibitiva, o no han sido elaborados todavía, o no se sabe muy bien cuáles podrían llegar a ser. En las ciencias complejas contemporáneas ciertas operaciones estadísticas comunes bajo el supuesto de la distribución normal, como el cálculo del coeficiente de correlación de Pearson, han debido ser abandonadas debido a dificultades insalvables cuando la red es de gran tamaño y la distribución es una LP (Serrano y otros 2006; Ferrer i Cancho y otros 2007: 67).

Lo llamativo del caso es que cuando se consulta la literatura original se percibe que Fisher, el padre de la NHST, tenía clara noción de la existencia de distintas clases de distribuciones y de su impacto en la estadística. En particular él abordó en detalle la distribución normal, las series de Poisson y las distribuciones binomiales, como se puede ver en nuestro hipertexto (Fisher 1925: 44 y ss.). Pero aunque el número de distribuciones que él interrogó establece un principio de diversidad, la primera en ser tratada es la que resultó ser definitoria, literalmente. De hecho la definición misma de significancia surge de la inesperada y abrupta caída de la probabilidad de ocurrencia de un determinado valor a medida que uno se va apartando de la media en una distribución normal:

> La rapidez con que la probabilidad cae a medida que aumenta la desviación [respecto de la media] se muestra claramente en estas tablas. Una desviación que exceda la desviación estándar ocurre más o menos una vez cada tres intentos. El doble de la desviación estándar se excede sólo una vez en 22 intentos, tres veces la desviación estándar sólo una vez en 370 intentos, mientras que la tabla […] muestra que para exceder la desviación seis veces se necesitarían cerca de mil millones de intentos. El valor para el cual $P \leq 0{,}05$ o 1 en 20, es 1,96 o cerca de 2; es conveniente tomar este punto como un límite para juzgar si una desviación ha de ser considerada significante o no. Las desviaciones que excedan dos veces la desviación estándar son por ende formalmente consideradas significantes (1925: 46-47).

Lo que Fisher no explora con el detenimiento debido es el hecho de que precisamente por esa caída exponencial de la probabilidad que experimenta todo lo que se aparte de la norma, la distribución normal resulta ser una mala heurística para juzgar la probabilidad de un evento o de un fenómeno. El escritor contemporáneo que ha echado más luz sobre este escenario es quizá el heterodoxo Nassim Taleb, un personaje con un sarcasmo demasiado a flor de piel pero que de a ratos es brillante como pocos. En una página que es un paralelo exacto de la cita de Fisher que acabo de consignar (seguramente desconocida para él), Taleb propone examinar la probabilidad de encontrar una persona de determinada estatura a medida que uno se aparta una "unidad de desviación" arbitraria (10 cm) de la media de 1,67 metros:

10 cm más alta que el promedio (1,77 m): 1 en 1,63

20 cm más alta que el promedio (1,87 m): 1 en 44

30 cm más alta que el promedio (1,97 m): 1 en 740

40 cm más alta que el promedio (2,07 m): 1 en 32.000

50 cm más alta que el promedio (2,17 m): 1 en 3.500.000

60 cm más alta que el promedio (2,27 m): 1 en 1.000.000.000

70 cm más alta que el promedio (2,37 m): 1 en 780.000.000.000

80 cm más alta que el promedio (2,47 m): 1 en 1.600.000.000.000.000

90 cm más alta que el promedio (2,57 m): 1 en 8.900.000.000.000.000.000

100 cm más alta que el promedio (2,67 m): 1 en 130.000.000.000.000.000.000.000

110 cm más alta que el promedio (2,87 m): 1 en 36,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

Es obvio que esta aceleración de la improbabilidad no describe adecuadamente ni siquiera aquello que la distribución normal estaba llamada a modelar con mayor contundencia pedagógica, a saber: la estatura de las personas. Con una población mundial de unos 6.854.196.000 habitantes, el número de personas fehacientemente registradas que miden más de 2,37 cm supera la treintena; conforme a la tabla anterior no debería haber ninguna, ni la habría tampoco si la población fuese diez veces mayor de lo que actualmente es.

En fin, *en todos los ámbitos de aplicación imaginables* la ley normal predice probabilidades que están billones o trillones de veces por debajo de las magnitudes que se dan en la realidad apenas uno se aparta un poco de la media: no precisamente un leve efecto de alas de mariposa.[29] Si alguien sostiene que a pesar de todo es mejor disponer de alguna herramienta estadística que no tener ningún instrumento de aproximación, si ésta es la configuración numerológica que nos servirá de fundamento me permito entonces discrepar, pues cualquier ponderación a ojo de buen cubero, cualquier impresión de la vida cotidiana resulta funcionar mejor. Y si alguien sugiere (como de hecho me lo han sugerido) que el fracaso de la ley normal para dar cuenta de los casos que se salen de norma es culpa de estos casos precisamente por desbordar la pauta de la diversidad permitida, entonces me temo que no es la antropología como yo la concibo la disciplina que debería avalar esa preceptiva.

Buena parte del problema de la estadística de la prueba de hipótesis radica en su dependencia de la distribución normal, cuyo emblema es la curva en forma de campana [*bell curve*] a la que los

[29] Véase un inventario actualizado de las estaturas en http://www.taringa.net/posts/imagenes/9788460/la-persona-mas-alta-del-mundo.html. Los datos de esa página han sido recabados informalmente y carecen de aparato erudito pero son fidedignos; mucho más, en todo caso, que las predicciones de la distribución normal en tanto gráfico de la función de probabilidad.

medios de habla castellana han rebautizado con el aberrante apelativo de la "curva de Bell".[30] Taleb (ajeno a este género de traiciones filológicas) señala que es sorprendente que el billete de 10 marcos alemanes muestre un retrato de Gauss al lado de su famosa curva:

> La chocante ironía es que el último objeto posible que podría vincularse a la moneda alemana es precisamente esa curva: el Reichsmark (como se lo llamaba antes) pasó de cuatro por dólar a *cuatro trillones* por dólar en el espacio de unos pocos años durante la década de 1920, un comportamiento que nos dice que la curva en forma de campana carece de sentido como descripción de la fluctuación aleatoria de la cotización de la moneda. Todo lo que se necesita para rechazarla es que un movimiento así ocurra una vez, y sólo una vez (Taleb 2007: 240).

Curiosamente –concluye nuestro autor– la curva en forma de campana es la herramienta de evaluación de riesgo que manejan de preferencia banqueros y financistas, lo cual es (digo yo, según se mire) o bien suicida o bien criminal: una pequeñísima diferencia de medición del sigma, como suele llamarse, conduce inexorablemente a una masiva subestimación de la probabilidad (*op. cit.*: 232). Chirriante y rústica como pueda parecer, la conclusión de Taleb estaba latente (casi *verbatim*) en las severas elaboraciones de Fisher (1925: 47), quien prudentemente no se animó a mostrar el salto que existe entre la probabilidad de 1 en 1.000.000.000 para mediciones que se apartan de la media en 6 desviaciones estándar y la de 1 en 780.000.000.000 que nos espera una sola desviación más allá. A las breves aunque alarmantes disquisiciones probabilísticas de este último se las llevó sin embargo el tiempo; nadie quiso escuchar la voz del buen sentido y las distribuciones estadísticas acabaron entronizando la normalidad y aplanando las diferencias a través de la idea de media o de otros criterios del mismo linaje.

Pero aún en el orden de las magnitudes relativas (y ya no de las probabilidades) las distribuciones normales y las leyes de potencia responden a diferentes clases de escala; no sólo se diferencian en la dispersión de las cantidades implicadas, sino en el hecho de que las primeras son lineales y las segundas logarítmicas. Un conjunto de datos que exhibe (o que es muestreado a partir de) una distribución normal no puede compararse con otro que se distribuye según una LP: como puede verse en ambas puntas de la curva (Figura §1, izq.), en una distribución normal siempre hay muy pocos individuos altísimos y muy pocos también de bajísima estatura, o poquísimos genios y gente de poca inteligencia (o como los llamen los psicómetras). La diferencia cuantitativa entre los ejempla-

[30] Por supuesto, no ha habido ningún estadístico apellidado Bell que prohijara la curva epónima, por más que una búsqueda de la expresión encomillada "curva de Bell" en Google™ devuelva hoy (diciembre de 2011) la monstruosa cifra de 92.400 resultados, implicando que una proporción estadísticamente significativa de quienes imparten pedagogía en la materia promueven una patafísica insostenible. Si el Bell que todos estos enseñadores de falacias tienen en mente es, como imagino, Alexander Graham Bell, las únicas campanas que se podrían asociar razonablemente con este inventor son –conjeturo– la que su apellido invoca y la del timbre del teléfono.

res extremos y la cresta de la campana sería de menguada magnitud: cuatro o cinco órdenes lineales como mucho, jamás del orden de los miles o los millones como sucede en una LP característica como la que se encuentra en la medida del volumen de comercio exterior de los países, en la cantidad de discos que venden los artistas *pop*, en el número y capital de las corporaciones o en la dispersión de intensidades de los terremotos (Figura §2, der.; cf. Sornette 2006: 94).

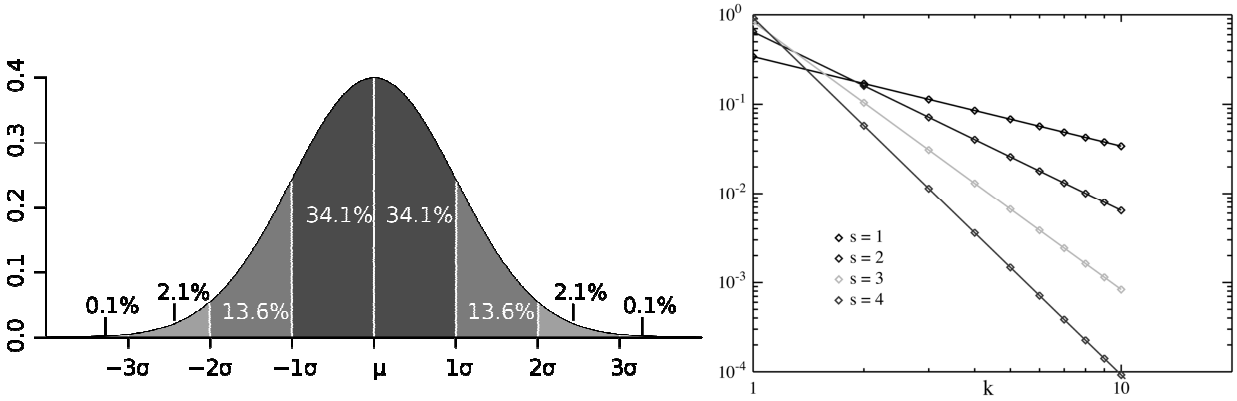

Figura 2 – Distribución normal (con desviación estándar) y LP

Incluso en el extremo de aleatoriedad absoluta de una ley gaussiana, las desviaciones de la media mayores a unas pocas desviaciones estándar son muy raras, como si hubiera límites precisos a los grados de libertad del mismo azar. Desviaciones mayores a 5, por ejemplo, rara vez o nunca se ven en la práctica. Ni hablar de categorías tales como media/promedio, mediana, moda, variancia, desviación estándar o los valores *t*, *z* o χ^2 que se manejan en los cálculos implicados en la NHST. El adulto más alto (tal vez Robert Pershing Wadlow, de 2,71 m) no puede ser nunca más que entre 5 y 6 veces de mayor estatura que el más bajo (hasta ayer Gul Mohammed con 0,57 m). La estatura promedio (o "normal") está seguramente en las cercanías de la suma de ambos extremos, dividida por dos: 1,64 m en este caso, apenas 3 cm por debajo de la media de Taleb.

Si volvemos a observar con detenimiento la Figura 1, veremos que alrededor del 68% de los valores de una distribución normal (≈ 0,6827) se encuentran a una desviación estándar σ apartados de la media, un 95% (≈ 0,9545) yace a dos desviaciones estándar (de donde viene $p \leq 0{,}05$) y un 99,7% (≈ 0,9973) a tres. Este hecho se conoce popularmente como la regla 68-95-99,7, regla empírica o regla de 3-sigma y es peculiar de esta distribución. Redondeando un poco, el cómputo de la probabilidad de que un valor *x* se encuentre a 2 desviaciones estándar de la media es:

$$\Pr(\mu - 2\sigma < x < \mu + 2\sigma) = \Phi(2) - \Phi(-2) \approx 0{,}9772 - (1 - 0{,}9772) \approx 0{,}9545$$

Esto se puede expresar en términos de intervalos de confianza: diríamos entonces que el $\mu \pm 2\sigma$ propiciado por Fisher es un intervalo de confianza de aproximadamente un 95%. Cifras parecidas a éstas se mantienen para muchas otras distribuciones simétricas.

Dada su aparente simplicidad la distribución normal se ha utilizado como la ley de referencia para lograr aproximaciones a distribuciones que se desvían (más o menos ligeramente) de la normalidad; en el registro de estas aproximaciones consagradas por el uso aparecen la distribución binomial, la de Poisson, la binomial negativa, la hipergeométrica, la beta, las de von Mises y Birnbaum-Saunders, la de Neyman tipo A, la distribución de chi cuadrado y gama, la *t* de Student, la *F* de Fisher, las formas cuadráticas y alguna otra que se me escapa (Patel y Read 1982: 168-223; Johnson, Kotz y Balakrishnan 1994: 111-122). No son aproximables normalmente, desde ya, las distribuciones de Cauchy y las demás leyes de la clase LP, que son las que proliferan en la abrumadora mayoría de las investigaciones en ciencias sociales (ARS incluido) que se han realizado en el siglo que corre (Reynoso 2010; 2011a: 207-236).

Más allá de la simplicidad platónica y de las resonancias pitagóricas de la ley normal, es absolutamente obvio que una entidad caracterizada por los constreñimientos que se han señalado refleja muy pocas características de la vida social, de los fenómenos territoriales o de la realidad económica. La fortuna promedio de las personas que viven en el planeta no se obtiene sumando los 50 mil millones de dólares que posee Carlos Slim o Bill Gates al centavo de dólar que atesora un digambara mendicante, dividiendo luego ese guarismo por la mitad. Desafortunadamente (valga la expresión) tampoco el 95% de las personas del mundo se apiñan en simetría y a una leve distancia ($\pm 2\sigma$) de ese orden promedio de magnitud.

Como se sabe por lo menos desde Vilfredo Pareto, la probabilidad de incluir a alguien del nivel económico de Slim o Gates en una muestra de población es, asimismo, ridículamente baja por más refinamiento que insuflemos a nuestras técnicas de muestreo (y las hay en extremo refinadas) (cf. Krishnaiah y Rao 1988). Tal como lo prescribe el TLC, excepto en el caso de poblaciones con distribuciones estrictamente normales y salvo que ocurra un milagro astronómicamente improbable, ningún muestreo aleatorio podría reproducir en la muestra *exactamente* la misma clase de ley que impera en la población. No es sino Mandelbrot, el padre de la geometría fractal, quien acaba con el mito de la distribución gaussiana como imagen de la normalidad de los aconteceres:

> [L]a campana de Gauss se ajusta muy poco a la realidad. Desde 1916 hasta 2003, los movimientos diarios del índice Dow Jones no se distribuyen sobre el papel como una campana de Gauss simple. Las colas se elevan demasiado, pues hay más cambios grandes de lo esperado. La teoría [gaussiana] sugiere que, a lo largo de todo ese tiempo, debería haber habido 58 días en que el Dow Jones variara más del 3,4 por ciento; en realidad hubo 1001. La teoría predice seis días de variaciones por encima del 4,5 por ciento; hubo 366. Y las oscilaciones del índice por encima del 7 por ciento deberían darse una vez cada 300.000 años, mientras que el siglo XX contempló 48 de tales días. Una era ciertamente calamitosa que insiste en burlarse de

> todas las predicciones. O tal vez son nuestros supuestos los que están equivocados (Mandelbrot y Hudson 2006: 36).

Un examen de las observaciones de Benoît Mandelbrot a propósito de las cotizaciones de bolsa, sumadas a las de Fisher y Taleb sobre las estaturas de las personas nos permite explicar, incidentalmente, el estupor de Francis Edgeworth, de Ian Hacking y del propio Fisher sobre el aval que la teoría de probabilidad basada en la distribución normal estaría prestando a los fenómenos paranormales con significancias de hasta 0,00004 (cf. más arriba, pág. 42; Burdick y Kelly 1977; Gilmore 1989: 338; 1990; Utts 1991). El asunto admite una simple explicación: dada la abrupta caída de la probabilidad por poco que el caso se aleje de la media, la prueba basada en la ley normal obliga a rechazar la HN aun cuando el mero azar constituya una explicación satisfactoria de los hechos observables y aun cuando la hipótesis alternativa postule la verdad de la transmisión extrasensorial. Por más que no se le pueda acompañar en sus explicaciones esotéricas, tiene razón entonces J. B. Gilmore cuando concluye que "[l]os datos anómalos no se encontrarían tan a menudo si la estadística clásica ofreciera un modelo válido de la realidad" (1990: 54).

Mientras que las distribuciones normales reinaron soberanas hasta hace pocos años, ahora la mayor parte de los textos avanzados de estadísticas convencionales y de estimación robusta documentan los problemas infinitos e insidiosos que acarrea la presunción de normalidad, bien conocidos desde la época de Henri Poincaré [1854-1912], quien certificadamente escribió: "Todos creen [en la ley normal de los errores]: los experimentadores porque piensan que es un teorema matemático, los matemáticos porque piensan que es un hecho experimental" (Poincaré 1912: 171; cf. Cramer 1946).[31]

Poniendo en valor este género de indicios históricos, hoy se piensa más bien que "ninguna distribución es normal" (Wilcox 2005), que la no-normalidad es lo verdaderamente universal (Gosset [Student] 1908), "que la normalidad es un mito, [que] nunca ha habido, ni nunca habrá, una distribución normal" (Geary 1947: 210, 241), y que dicha ley ha sido desde siempre una fuente inagotable de distorsiones y problemas prácticos y teóricos (Fisher 1922; Pridmore 1974; Bradley 1977; 1980; Hill y Dixon 1982; Tan 1982; Abbott 1988; Micceri 1989; Hacking 1991; Maltz 1994; Spedding y Rawlings 1994; Kruskal y Stigler 1997; Hald 1998: 649). "Dios ama a la curva normal", escribían Hopkins y Glass (1978: 95); yo decididamente no. Las tendencias recientes y otras que no lo son tanto refrendan, incluso con Dios en contra mía, el punto de vista que se sustenta en este ensayo:

> Los supuestos de normalidad han jugado un papel crucial en el análisis estadístico a través de los años, pero desde la década de 1960 se ha prestado más atención al cuestionamiento de

[31] "Tout le monde y croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théorème de mathématiques, et les mathématiciens que c'est un fait expérimental."

> estos supuestos, requiriéndose estimadores que sean robustos cuando se violan esos supuestos y desarrollando más pruebas iluminadoras sobre su validez (Patel y Read 1982: 11).

Ya Louis Guttman había escrito unos años antes:

> La distribución normal no es un fenómeno empírico normal. Nunca, si es que alguna vez, se la ha observado en la naturaleza. Ha sido generada en gran medida por los estadísticos cuando desarrollaron las matemáticas de la teoría del muestreo. Este hecho se ha venido enseñando desde hace mucho, pero parece que es necesario repetírselo constantemente a los estudiantes después que han sido expuestos a cursos de inferencia estadística (Guttman 1977: 92).

Y W. A. Pridmore antes aun:

> No voy tan lejos como Geary […] quien demandaba que todos los manuales debían tener un texto en tipografía de gran porte que dijera "La normalidad es un mito […]". Pero por cierto no seré segundo de nadie en mi celo por encontrar, seleccionar y transformar todos los datos para que sean efectivamente Normales. Log, log-log, *probit*, *rankit*, arco-seno, raíz cuadrada… Incluso me rebajaré al recurso de tratar todos los elementos que rehusen ser desmenuzados a la normalidad como si fueran intrusos provenientes de una segunda distribución de "disidentes" o "salvajes" que aparecen 1 vez en *n* veces. Pero una distribución nunca, repito, nunca, comienza siendo verdaderamente normal (Pridmore 1974: 623).

Más cerca de nuestros tiempos escribe el criminólogo Michael Maltz:

> Se nos ha acondicionado para presuponer que un solo conjunto de datos producirá un solo patrón que puede caracterizarse por sus valores de media. Este supuesto implícito de unimodalidad –un solo modo de conducta para toda la población bajo estudio– también se encuentra en nuestros supuestos sobre la influencia unitaria de varios procesos sobre todos los individuos. […] Deben tomarse medidas para poner a prueba estos supuestos. Los conjuntos de datos deben analizarse para ver si hay más de un tipo de individuo contenido en los datos, si hay más de un tipo de conducta que parece manifestarse y si más de un escenario resultará de un tratamiento. […] Tuvimos que conformarnos con los supuestos y limitaciones de las técnicas estadísticas estándar porque hasta hace muy poco no teníamos alternativas realistas para analizar grandes conjuntos de datos. Para manipularlos debíamos emplear modelos inexactos cuya virtud primaria era que resultaban tratables. Pero ya no tenemos más que "modelar los datos". La creciente disponibilidad de computadoras de alta velocidad y capacidad y de excelentes programas de análisis y graficación significa que ahora podemos dejar que los datos hablen por sí mismos (Maltz 1994: 457).

Los manuales contemporáneos de estadística robusta ya acostumbran presentar los hechos distribucionales de maneras así de taxativas:

> [P]or lo general se entiende que los modelos formales son simplificaciones de la realidad y que su validez es en el mejor de los casos aproximada. El modelo de formalización más ampliamente utilizado radica en el supuesto de que los datos observados poseen una distribución

> normal (Gaussiana). Este supuesto ha estado presente en las estadísticas por dos siglos, y ha sido el marco de referencia para todos los métodos clásicos de regresión, análisis de varianza y análisis multivariado. Han habido intentos por justificar el supuesto de normalidad por medio de argumentos teoréticos, tales como el teorema del límite central. Estos intentos, sin embargo, fácilmente se demuestran erróneos. La principal justificación para presuponer una distribución normal es que proporciona una representación aproximada de muchos conjuntos de datos, y al mismo tiempo es teoréticamente muy conveniente porque permite derivar fórmulas explícitas para métodos estadísticos óptimos […]. Nos referimos a esos métodos como los métodos estadísticos *clásicos*, y tomamos nota de que ellos reposan en el supuesto de que la normalidad rige *exactamente*. Las estadísticas clásicas son muy fáciles de computar para los estándares modernos de computación. Por desdicha, la conveniencia teorética y computacional no siempre proporciona una herramienta adecuada para la práctica de la estadística y el análisis de datos (Maronna, Martin y Yohai 2006: xv).

En uno de los mejores estudios sobre los significados cambiantes de la normalidad escriben por último Kruskal y Stigler:

> En 1994 el portavoz de la Casa de Representantes de los Estados Unidos admitía (en el *New York Times*, 10 de noviembre de 1994) que se le había dicho que "el uso de la palabra normal es políticamente incorrecto". […] Mareas cambiantes de usanzas de esta clase han afectado también al uso estadístico, con una generación rebelándose contra el término porque éste inspira presupuestos de normalidad demasiado fáciles, mientras que la siguiente urde transformaciones ingeniosas para asegurarse que la normalidad se mantiene, como una profecía matemática que se autocumple. Pero la fuerza de lo "normal" (con la deliciosa ambigüedad que trae tanto a la discusión científica como a la pública por encarnar tanto lo usual como lo ideal) parece asegurar que cuando retrocede no ha de ser por mucho tiempo. El uso obligatorio de "normal" ha continuado por más de dos siglos y es probable que lo siga haciendo por los próximos dos. Que esto sea así parece, bueno…, sólo normal (Kruskal y Stigler 1997: 105).

De todas maneras el hecho es que desde comienzos del siglo XXI en particular la distribución normal se encuentra en retirada. Característica de esta actitud defensiva es la forma en que los aleatoristas intentan racionalizar el hecho de que en disciplinas enteras no existen formas no distorsivas de forzar los datos para que se acomoden a la curva en forma de campana. Tras mostrar un gráfico no normal correspondiente a niveles de antígenos específicos de próstata (PSA) en pacientes que han sufrido operaciones de cáncer y otro que muestra rangos de intensidad de dolores de cabeza en diversos pacientes escribe Andrew Vickers:

> Ambos gráficos lucen bastante similares entre sí y bastante distintos a una distribución normal. La simple explicación de lo que ocurre aquí es que la investigación médica típicamente involucra estudiar pacientes con alguna clase de enfermedad. Por definición, estas poblaciones no son normales. Quizá es esto lo que está detrás del comentario de mi profesor respecto de la rareza de distribuciones normales en medicina; rara vez se ven distribuciones normales

> en medicina porque rara vez se estudian poblaciones “normales” como un todo, sino subconjuntos inusuales. Una forma más matemática de decir esto es que mientras los procesos normales usualmente involucran sumas, […] los procesos de enfermedad usualmente involucran multiplicación (Vickers 2010: 34)

Aplicando una transformación logarítmica, esto es, suplantando las multiplicaciones por sumas y desfigurando la escala según lo impongan las metáforas, los datos pueden ser movidos de grado o por fuerza hacia algo que no es la normalidad en plenitud pero que se le parece; eso es lo que Vickers hace finalmente. El problema que subsiste es que toda la estadística que nos ocupa requiere normalidad estricta, comenzando por la aleatoriedad de los mecanismos que generan las distribuciones observables; de otro modo todo se torna indecidible: “No hay reglas claras para determinar si un conjunto de datos está lo suficientemente cerca de una distribución normal para utilizar procedimientos estadísticos que presuponen normalidad” (Op. cit.: 39).

Desde ya que existen técnicas de muestreo y cálculo diseñadas para la estadística no lineal y el azar salvaje, tales como el método de linealización (o expansión de Taylor), la minimización de la suma de los cuadrados de los residuos y las técnicas de reutilización de muestras como el jackknife, la replicación repetida balanceada (BRR), y el *bootstrap* (Krishnaiah y Rao 1988: 436-444; Wasserman 2006). Pero ni uno solo de los textos que recomiendan el uso de la NHST se ocupa de semejantes menesteres, particularmente oscuros, mal conocidos, tratados a las apuradas en una bibliografía periférica y plagados de desajustes y supuestos de ley normal y uniformidad de escala que, expulsados por la puerta, siempre vuelven, apenas se presenta la ocasión, a colarse en la teoría por ventanas que todavía no sabemos cómo cerrar.

Esto dicho, es evidente que (por razones mucho más fuertes que la hibridez entre las ideas de Fisher y las de Neyman-Pearson, la arbitrariedad de 0,05, el problema de Galton, la [re]apropiación del método por creacionistas y parapsicólogos o el significado ambiguo del valor de *p*) la NHST es, cualesquiera sean sus suplementos redentores, de cabo a rabo incongruente con las clases de distribución que conciernen a nuestra disciplina, a las prácticas de gestión territorial y a la mayor parte de las ciencias empíricas.

Aun cuando mi concepción de lo aleatorio difiere enormemente de la suya, vale la pena citar una vez más al heterodoxo Taleb:

> Las desviaciones estándar no existen fuera del mundo gaussiano, o si existen no importan nada y tampoco explican mucho. Pero la cosa es peor. La familia gaussiana (que incluye varios amigos y parientes, tales como la ley de Poisson) es la única clase de distribución para las cuales la desviación estándar es descripción suficiente. No se necesita más nada. La curva en forma de campana satisface el reduccionismo de lo engañoso.

> Hay otras nociones que poseen poca o ninguna significación fuera de lo gaussiano: correlación, o peor todavía, regresión. Y sin embargo están profundamente engranadas en nuestros métodos. [...]
>
> Para ver cuán carente de sentido es la correlación fuera de Mediocristán, tome usted una serie histórica que involucre dos variables que visiblemente son de Extremistán, tales como los bonos y el mercado de acciones, o dos precios de pólizas de seguros, o dos variables como, digamos, los cambios en los precios de libros infantiles en los Estados Unidos y la producción de fertilizantes en China; o los precios de los bienes inmobiliarios en Nueva York y los retornos del mercado de acciones de Mongolia. Mida la correlación entre los pares de variables en diferentes sub-períodos, por ejemplo, para 1994, 1995, 1996, etcétera. La medida de correlación exhibirá probablemente una severa inestabilidad; dependerá del período para el cual fue computada. Y sin embargo la gente habla de correlación como si fuera algo real, haciéndolo tangible, invistiéndolo de una propiedad física, reificándolo.
>
> La misma ilusión de concretidad afecta a lo que llamamos desviaciones "estándar". Tome usted una serie de precios o valores históricos. Sepárela en varios sub-segmentos y mida su desviación "estándar". ¿Sorprendido? Cada muestra exhibe una desviación "estándar" diferente. ¿Por qué la gente habla entonces de desviaciones *estándar*? Pues vaya uno a saber… (Taleb 2007: 230-240).

Pero no se trata sólo de una disyunción excluyente –como conviene a Taleb– entre Mediocristán y Extremistán. Aunque todo ordenamiento taxonómico es más o menos arbitrario y nunca es correcto preguntarse unívocamente "cuántas clases de *x* hay", el número de distribuciones estadísticas que han sido honradas con un nombre ronda la cincuentena. Tarde o temprano se hará necesario escribir un buen manual de distribuciones características en la vida sociocultural, bien razonado y hasta condescendientemente pedagógico, sin dar nada por sentado, sin alardes de incomprensibilidad, análogo al manual matemático de distribuciones de Evans, Hastings y Peacock (1993) o al precioso compendio de Kalimuthu Krishnamoorty (2006). Ya hay antecedentes de esta iniciativa en una ciencia semiblanda: en economía y ciencias actuariales existen al menos dos volúmenes en esa tesitura, el de Christian Kleiber y Samuel Kotz (2003) y el de Svetlozar Rachev (2003). También *Financial modeling under non-gaussian distributions* de Eric Jondeau, Ser-Huang Poon y Michael Rockinger (2007) apunta en esa dirección.

Si es que la antropología aspira a estar en la misma liga de pensamiento esclarecido que la contabilidad y el mercadeo, una parte esencial del diagnóstico pasa entonces por establecer un cruzamiento entre la estructura del objeto y alguna de las distribuciones conocidas. En términos de la nomenclatura estadística las distribuciones posibles son numerosas: entre las que se me ocurren ahora (con alguna que otra homonimia o nombre colectivo) están la de Benford, Benini, Benktander, Bernoulli, beta, binomial, binomial negativa, de Bose-Einstein, Bradford, Bull, Burr, Cantor, Cauchy (o Breit-

Wigner, o Lorentz), Champernowne, Chernoff, chi cuadrado, de Davis, Dirichlet, doble gamma, doble Weibull, de Erlang, exponencial, geométrica, de Gauss, Gibrat, Gompertz, gamma, Heaps, hiperexponencial, hipergeométrica, de Horton, Kleiber, Kumaraswamy, Laplace, Lévy, logarítmica, logística, lognormal, Lotka, de Moyal, multinormal, de Nakagami, Pareto, Poisson, Pólya, Rademacher, Rayleigh, Rice, secante hiperbólica, de Wigner o semicircular, Skellam, de Student, triangular, uniforme, de von Misses, Wald, Wallenius, Yule-Simon, zeta, los tres tipos de valor extremo (Gumbel, Fréchet, Weibull) y por supuesto la distribución de Zipf, Zipf/Mandelbrot o LP (Kagan, Linnik y Rao 1973; Patel y Read 1982; Evans, Hastings y Peacock 1993; Johnson, Kotz y Balakrishnan 1994; Kotz y Nadarajah 2000; Walck 2000; Balakrishnan y Nevzorov 2003; Zelterman 2004; Johnson, Kemp y Kotz 2005; Consul y Famoye 2006; Newman 2006; Forbes, Evans, Hastings y Peacock 2011).[32] De ningún modo en todas las instancias pero sí en unas cuantas de ellas puede que las proporciones, los atractores, los sesgos, las formas alcancen a trasuntar información; e información *es*, como decía Gregory Bateson (1980: 62), una diferencia que hace una diferencia.

Es que en las distribuciones no hay sólo medidas sino fundamentalmente pautas. Cada una de ellas tiene su historia, diagnosis, idiosincracia, significado, campo de aplicación y etiología. Una distribución es, además, un artefacto narrativo con una fuerte connotación espacial y visual, uno de esos *habitus* estructurantes de los cuales nos habla (empleando otras palabras) la nueva ciencia de la cognición matemática, desde Marcus Giaquinto (2009) en más. Encontrar cuál es la distribución que aplica con mayor probabilidad a un caso concreto involucra no sólo fundar un conocimiento del objeto que se nutre de un amplio campo transdisciplinario, sino habilitar las posibilidades de su comparación sistemática y de una intervención coherente en las prácticas complejas que lo conforman.

Saber cuál entre todas las distribuciones converge mejor con los datos es tarea necesaria pero inextricable, pues tampoco lleva cada una su nombre marcado en la frente; a veces obtenemos respuestas antagónicas atinentes a su identidad formulando preguntas apenas dispares. Igual que sucede con los venenos en el peritaje forense, el análisis sólo proporciona respuestas a las preguntas que efectivamente se hagan. Cada variante de distribución debe perseguirse con una probabilidad de aproximación incierta mediante pruebas estadísticas en extremo disímiles y (en lo que a la LP atañe) todavía mal conocidas que (a diferencia de una NHST engañosamente fácil) muy pocos antropólogos o arqueólogos educados a la usanza tradicional estarán en condiciones de parafrasear. Ya no se

[32] La familia LP es una de las más amplias, comprendiendo distribuciones como la de Benford, Bradford, Cauchy, Gibrat (crecimiento proporcional), Gutenberg-Richter (terremotos y tsunamis), Heaps, Horton (cuencas hídricas), Kleiber (metabolismo/tamaño y otras leyes alométricas), Lévy, Lotka, Pareto (regla 80/20), Stefan-Boltzmann, Steven (psicofísica), Yule-Simon, Zipf-Mandelbrot. Quienes piensen que con Poisson o Bernoulli tienen suficiente diversidad deberían dedicarle algún tiempo a la exploración de estos mundos distintos.

aplican tanto las pruebas de Shapiro-Wilk, Jarque-Bera, Cramér-von Mises, χ^2 o Kolmogorov-Smirnoff, sino que más bien cuadra pensar en variaciones de los tests de Kuiper, Lilliefors o Anderson-Darling (Jondeau, Poon y Rockinger 2007: 16-21; Saichev, Malevergne y Sornette 2010: 4). De más está decir que ninguno de los textos de antropología, análisis de redes o estudios territoriales que abrazan la NHST menciona siquiera estas pruebas esenciales.

Los campos en que se manifiestan leyes de potencia son innumerables y extraordinariamente diversos; se las encuentra en el número y población de las ciudades, el número y longitud de las calles en una ciudad, las longitudes de los trayectos en cualquier viaje, las erupciones volcánicas, la intensidad de los terremotos y tsunamis, las proporciones constructales de la naturaleza y la cultura, los diámetros de los cráteres lunares (o de los cráteres en general), las erupciones solares, la cantidad y volumen de las cuencas petrolíferas, el tamaño de los archivos de computadora, los catálogos de música por compositor, las víctimas en guerras y contiendas, la frecuencia del uso de palabras en cualquier lengua, texto o género literario, los hablantes de cada una de las lenguas, los nombres de persona y los apellidos en la mayor parte de las culturas, el número de *papers* que escriben los científicos, las citas bibliográficas por autor, los *hits* recibidos por las páginas de la Web, los vuelos que llegan y salen de un aeropuerto, los libros o discos vendidos por autor o intérprete, el número de especies en las categorías biológicas o de personas que son seguidores de un equipo de fútbol, los individuos sobre los que los influenciadores ejercen influencia, el número y caudal de los ríos, la estructura de ramificación de los árboles y de los sistemas circulatorios, el número de fieles de cada confesión religiosa o el de seguidores en una red social, los ingresos anuales de las personas o el volumen del comercio exterior (o del presuesto militar) de los países (Kleiber y Kotz 2003; Newman 2006).

Cuando Fisher elaboró su método muy poco de esto se conocía; como lo expresa Savage (1976: 449), crípticamente, Fisher “le dio la espalda a las funciones de potencia”, prefiriendo concentrarse en una distribución normal que, al margen de innumerables *data sets* que resultan de la aplicación de métodos de muestreo, sólo se ha probado muy deficientemente apropiada para el análisis de las estaturas de las personas, sus coeficientes intelectuales, sus pesos, sus perímetros torácicos, el tamaño de sus pies o el largo de sus penes, siempre y cuando el rango de dispersión de valores no sea mayor a 2 desviaciones estándar respecto de la media (Abbott 1988; Reynoso 2011a: 207-236).

Seguramente hay otras distribuciones relevantes aparte de la LP, muchas de las cuales se confunden con ella. La identificación e interpretación de las distribuciones tienen no poco de arte o de hermenéutica: a veces ellas difieren entre sí en un grado muy pequeño, y el grano grueso del *assessment* en ciencias humanas (o la impropiedad del diseño algorítmico, o el sesgo del muestreo) casi siempre arroja dudas de monta sobre la distribución que se tiene entre manos. Y como se ha probado hasta el

hartazgo en tiempos recientes, son muy pocos los científicos que dominan los mecanismos del aparato probatorio y menos todavía los que llegan a comprender qué es con exactitud (valga la expresión) lo que las pruebas de significancia logran probar (Falk y Greenbaum 1995; Haller y Krauss 2002).

Desde ya que en una antropología sensible a la complejidad hay mucho más por hacer que abrir la mirada hacia la diversidad de las distribuciones estadísticas. Lo primero en que deberíamos empeñarnos sería en poner bajo sospecha el supuesto de que las innumerables cifras de tendencia que nos entregan los programas de análisis son elementos diagnósticos significativos en cualquier escenario.[33] También habrá que decidirse a abandonar los encuadres de referencia irreales que nos ofrece la idea de media y recordar que (como una vez más decía Bateson 1980: 47-48) una cantidad no involucra una pauta; y habrá que resignarse al hecho de que todo lo que hagamos debe ser congruente con la dinámica no lineal que atraviesa al objeto tal cual hoy es posible percibirlo, antes que con la estática encarnada en aquéllas y en otras numerologías simplificadoras. Todo esto excede la mera estadística, por cierto; pero ganar conciencia reflexiva de la existencia y relevancia de otras estructuras y procesos alejados de la normalidad ya sería sin embargo un buen comienzo.

[33] Me refiero a variables tales como el grado promedio de vértices, el valor medio de ajuste para el modelo de bloques, la densidad o el *span* de un actor, la media o la varianza de *indegree* o *outdegree*, etc., todas ellas muy comunes en analítica de redes. En un modelo en que la población es pequeña imagino que tales guarismos quizá proporcionen algún indicador más o menos útil; pero es indudable que esa pauta escalará muy mal cuando las distribuciones sean del género de la LP, cuando se quieran comparar redes con distintas (o múltiples) distribuciones subyacentes o cuando haya que pasar del escenario local al contexto global.

13. NHST en antropología, arqueología y estudios territoriales

El primer y más extenso tratamiento sobre las vicisitudes de la prueba estadística en arqueología fue desarrollado por George Cowgill (1977) en un artículo de *American Antiquity* que dista de la perfección pero debería ser mejor conocido. El objetivo de la publicación aparece definido con claridad en estos párrafos:

> El problema finca en que cuando uno se fija en la vasta mayoría de los textos introductorios a la estadística, encuentra protestas en el sentido de que "este no es otro de esos libros de recetas de cocina" y un capítulo o dos sobre los fundamentos lógicos de la inferencia estadística, antes que el autor se mueva hacia diez o veinte capítulos sobre cómo llevar a cabo una variedad de procedimientos. Por desdicha, la sección sobre la lógica de la inferencia estadística a menudo deja mucho que desear y con frecuencia alienta unas cuantas concepciones equivocadas.
>
> Hay, entonces, un vacío que rara vez se cubre adecuadamente ya sea por los libros de filosofía o los de estadística, y este artículo se propone llenar ese vacío. Más en particular, pretendo corregir varios errores serios en la interpretación de los resultados estadísticos que a menudo cometen los arqueólogos (¡así como otros científicos sociales!). Sin embargo no se pueden explicar claramente estos errores o las razones que hay para pensar en mejores alternativas excepto en el contexto de una discusión más amplia de algunos aspectos de los métodos de inducción; esto es, procedimientos para justificar aseveraciones que alegan referirse a algo más que (o a algo distinto de) lo que efectivamente se ha observado o ha sido deducido a partir de premisas lógicas abstractas (Cowgill 1977: 351).

Dado que todos los métodos que circundan a la NHST presuponen un muestreo aleatorio, Cowgill puntualiza correctamente que muchas situaciones arqueológicas no se concilian con el muestreo al azar sino con alguna otra variante probabilística tal como el *cluster sampling*. También es acertada la observación respecto de que un 5%, un 1% o el porcentaje de significancia que fuere implica algo muy distinto en el caso de disponer de una muestra pequeña o de una muestra grande. Metodológicamente se requiere también dar cuenta del tamaño de la muestra y de la potencia estadística –añade– así como estar en guardia frente al hecho de que "la información ambigua no es manejada muy bien por la estrategia de Neyman-Pearson" (p. 359). El problema con la NHST –concluye, con fina percepción pero exactitud parcial– es que ella es primariamente una evaluación de la evidencia relevante para H_0 y sólo de manera indirecta y retorcida una evaluación de la evidencia relevante de cara a cualquier alternativa específica (p. 363).

Si bien hacía falta que alguna autoridad en la disciplina aclarara de una buena vez que la prueba estadística no se ocupa en absoluto de la probabilidad de una hipótesis alternativa, el problema que subsiste es que tampoco es verdad que el valor de *p* exprese de manera directa una evaluación de la

evidencia respecto de H_0, dado que (como es necesario repetirlo) lo que en realidad mide (siempre que se satisfagan los supuestos del caso) es la probabilidad de obtener una muestra igual o más extrema que aquella que se tiene entre manos en caso que H_0 sea verdad.

Asentadas sus consideraciones, Cowgill desarrolla su elocución sin poner en tela de juicio la robustez de los parámetros usuales ni los supuestos de aleatoriedad, normalidad y homocedasticidad invariablemente requeridos por la operatoria (y presupuestas en la lógica) de cualquiera de las formas puras o templadas de prueba estadística. Cowgill adopta una lectura personal de la versión híbrida de la NHST atribuyéndola a Neyman-Pearson, sin recurrir a la literatura original, reposando en interpretaciones de figuras epigonales (primordialmente Wesley Salmon, sospecho) y sin advertir la naturaleza contradictoria de los postulados en juego. Utiliza a cada momento la expresión fisheriana "hipótesis nula", por ejemplo (pp. 353, 359, 363-365, 367), pese a que Neyman detestaba ese nombre y no lo homologó en ninguno de sus trabajos. Tampoco cita Cowgill ningún texto concreto de Neyman y Pearson, lo cual es curioso si se piensa que su trabajo se refiere específicamente a la clarificación de la metodología propuesta por dichos autores.[34] Conocedor de las críticas emanadas de la colección de Morrison y Henkel (1970) pero sin ahondar en ellas, y admitiendo la arbitrariedad de un "número mágico" de significancia del 5% (p. 359), Cowgill termina asegurando que si bien la NHST es propensa a multitud de errores de procedimiento e interpretación, es posible ejecutarla de manera correcta y provechosa, si se tiene además en cuenta que existen otras alternativas de igual valor.

Dado que el autor no ahonda en la ejemplificación de las formas incorrectas e improductivas de ejecutar la prueba de hipótesis que se encuentran en la literatura arqueológica, no queda finalmente en claro cuáles son las modificaciones del método a implementar de aquí en más, a excepción de trabajar con poblaciones enteras antes que con muestreos (pp. 366-367), de examinar la relación entre intervalos de confianza y pruebas de significancia (p. 364), de adoptar una complicada estrategia de "estimación" (pp. 360, 362-366) y de prestar atención a la potencia estadística (p. 358), ideas que tampoco son llevadas a su acabamiento respecto de escenarios alejados de la normalidad. La pregunta que queda en el aire es, por supuesto, cuál es el objeto de obtener estos estimadores (y sobre todo qué significan en este contexto los valores de p, α, β o lo que fuere) si se supone que se

[34] Si bien el desconocimiento de la literatura estadística clásica de la estadística hace cuarenta años era excusable, hoy está muy claro que ya no lo es. Tanto el texto capital de Fisher como el de Neyman-Pearson están hoy disponibles en la Web, en http://psychclassics.yorku.ca/Fisher/Methods/index.htm y por la vía académica a través de JSTOR, respectivamente. *The design of experiments* se puede obtener en http://www.4shared.com/file/0QcsfzFm/design_of_experiment_fisher.htm?aff=7637829. No obstante la amplia disponibilidad de materiales los autores se siguen resistiendo a la lectura de las obras canónicas, aun cuando los temas que ponen en discusión involucren a las teorías plasmadas en ellas. Como caso extremo véase p. ej. van der Pas (2010).

ha trabajado no con una muestra sino con la totalidad de la población. Cowgill parece no haber reflexionado suficientemente sobre el hecho de que la prueba estadística sólo tiene sentido como estimación relativa a la lógica específica del muestreo aleatorio; la falta de consulta de la literatura original ha tenido sin duda su impacto en ello. Estimo que ha sido D. McCloskey quien más claramente discurrió sobre este punto:

> La prueba usual no discute sobre estándares. Los deja de lado a favor de una perorata irrelevante sobre la probabilidad de un error de tipo I de cara a la lógica del muestreo. Muchos economistas [y muchos arqueólogos, se diría] parecen haber olvidado cuán estrecha es la pregunta a la que responde la prueba estadística de significancia. Ésta dice al investigador intrépido cuán probable es que, *debido al pequeño tamaño de la muestra que él* [sic] *tiene*, cometerá un error de excesivo escepticismo si rechaza una hipótesis verdadera (en este caso, β=1,0). Aunque no da para burlarse de esto, honestamente no es mucho. Sólo nos permite estar alerta frente a una clase muy estrecha de estupidez (McCloskey 1985: 202).

Pocos meses más tarde de publicado el artículo de Cowgill el arqueólogo y antropólogo David Hurst Thomas (1978) enumeró algunos de los errores estadísticos más ubicuos en arqueología, incluyendo (1) adherir de manera esclavizante al nivel de 0,05 de significancia, (2) inferir relaciones causales a partir de la significancia estadística, (3) confundir la significancia estadística con la fuerza de la asociación, (4) modificar hipótesis a priori de niveles de significancia para dar cuenta de datos muestreados específicos, (5) manipular tablas de contingencia para obtener resultados estadísticamente significativos, (6) testear hipótesis mediante estrategias de "expedición de pesca", (7) malinterpretar *p* como medida de significancia, y (8) utilizar de manera defectuosa o ignorar los supuestos de los modelos estadísticos (Thomas 1978: 233). Por desdicha, ni el tema se elaboró más allá de esa enumeración, ni los documentos responsables de esos errores se identificaron al punto de poder seguirles la pista. Si sólo se leyera la semblanza de Thomas quedaría además la impresión de que la estadística no se refiere a otra cosa que a la prueba de significancia.

Para mayor abundamiento, muchas aseveraciones de Thomas sobre éste y otros temas tratados en su exitoso clásico *Figuring Anthropology* (1976; 1986) se saben desde hace mucho equivocadas. Los errores de calibre más grueso se encuentran en sus apreciaciones de la media aritmética como un parámetro estadístico confiable, importante, estable y sobre todo robusto:

> La media es la medida más eficiente de tendencia central porque cada variable tiene impacto sobre la computación final. No se desperdicia ningún dato. La media es también la medida más estable de tendencia central. Cuando se tomen muestras a partir de una población, se encontrará que la media sólo varía mínimamente entre muestras sucesivas (Thomas 1976: 69)

Por más que la disciplina no reaccionó como debió haberlo hecho, saludó al libro con una salva de elogios (Ammerman 1977: 456; Hodson 1977; Scheps 1982) y hasta generó demanda para una edi-

ción suplementaria, no tengo que ser yo quien refute esas apreciaciones contrarias a la estadística elemental. En un intenso artículo publicado en los *Annual Reviews of Anthropology* que parece ser su única contribución científica, Bonnie Laird Hole las confrontó sin contemplaciones:

> Los últimos 15 años presenciaron una cantidad enorme de literatura en lugares tan prominentes como el *Journal of the American Statistical Association*, *The Annals of Statistics*, *The American Statistician* y *Technometrics* preguntándose cuáles estimaciones de locación son eficientes y robustas y buscando alternativas a la media, de la que se sabe ampliamente que es inestable y no robusta. Hay cursos exhaustivos en departamentos de estadística sobre esa cuestión; se han escrito libros enteros sobre el asunto […] y muchos estadísticos bien conocidos se asocian específicamente con el campo de la estimación robusta. […] La corriente principal en la investigación de la estimación robusta difiere sustancialmente de Thomas al respecto. […] Un estudio en Princeton examinó 68 estimadores de locación incluyendo la media muestral y la mediana. En respuesta a la pregunta sobre cuál era el peor entre los 68 estimadores bajo estudio, los estadísticos escribieron que "si hay un candidato claro para esa afirmación global, se trata sin duda de la media aritmética" (Hole 1980: 228-229.

El caso es que la media no es lo que se llama una medida robusta de tendencia central porque un solo *outlier* manifiesto o unos pocos ejemplares que se aparten levemente de los supuestos del modelo normal ocasionan mediciones desmesuradamente impropias. La media posee un punto de quiebra [*breakdown point*] de 0%, de modo que un pequeño desvío posee efectos distorsivos fuera de toda proporción; la desviación estándar (que con frecuencia se utiliza como estimador de escala) tampoco es muy confiable porque los cuadrados de las desviaciones de la media también intervienen en el cálculo, de modo que los efectos de su desvío se exacerban. Pensándolo bien, la falta de robustez de los estimadores usuales, casi medio siglo después de las observaciones pioneras de John Tukey (1960b), fue el motivo esencial para el surgimiento de las estadísticas paramétricas robustas y eventualmente de las estadísticas no paramétricas y no paramétricas robustas (Sprent y Smeeton 2001; Maronna, Martin y Yohai 2006; Wasserman 2006; Huber y Ronchetti 2009).

La afirmación de Thomas es apenas la punta del iceberg. En cuanto se navega por la bibliografía se encuentra que a lo largo de las ciencias sociales se ha fortalecido desde hace mucho un régimen conservador, cristalizado en los libros introductorios de estadística, desde el cual no sólo se niegan sin mayor fundamento los problemas de pérdida de robustez sino que se invita alegremente a dejar de lado toda preocupación cuando el modelo estadístico no responde a condiciones estrictas de normalidad, linealidad, homocedasticidad y simetría. El tenor de su discurso es casi el mismo que el que prevalece en las opiniones vertidas por Thomas sobre la robustez de la media:

> En el pasado se asignaba considerable importancia a los supuestos de normalidad y homogeneidad de varianza y a técnicas para determinar si esos supuestos se habían satisfecho. En años recientes, sin embargo, se ha prestado mucha menos atención a la necesidad de satisfacer

> esos supuestos. C. A. Boneau, Young y Veldman y muchos otros han mostrado que incluso una desviación muy grande de la normalidad o de la homogeneidad de varianza poseen relativamente poca influencia sobre estas pruebas [de significancia]. Nuestra posición es, por ende, que mientras que es mejor que se satisfagan los supuestos de la prueba, las violaciones a los supuestos de normalidad y homogeneidad de varianza tendrán probablemente poco efecto sobre las conclusiones que se hayan de sacar (Young y Veldman 1965: 270).

Hace ya más de tres décadas que James Bradley (1978) demostró sobre bases sistemáticas que una vez que se define cuantitativamente un criterio de robustez y se modulan los tamaños de la muestra y la magnitud de la potencia estadística, definiendo para determinados valores de la tasa de error del tipo I (α) el rango de valores de p a los cuales se aplica la prueba de robustez, los resultados que se obtienen no respaldan el optimismo de Young, Veldman y demás partidarios de la estadística paramétrica convencional. Las estadísticas robustas de Bradley (que de ello se trata) demostraron fehacientemente la invalidez de las pruebas de hipótesis incluso en condiciones de muy leve violación de los supuestos.

Desde entonces se ha podido determinar también que muchas de las estadísticas cableadas en programas tales como SPSS o SAS eran y siguen siendo incorrectas en materia de robustez, que es como decir que son por completo inútiles a los efectos de las pruebas de hipótesis y significancia excepto en escenarios altamente idealizados (cf. Wilcox 2005; Erceg-Hurn y Mirosevic 2008). De más está decir que la comunidad de los estadísticos convencionales hizo caso omiso de estos hallazgos o se resistió a ellos activamente mucho más en el plano del poder académico que en el de la polémica científica. Todavía hoy los "métodos modernos" alternativos siguen siendo poco frecuentados aun cuando hay varios sitios en los que se encuentran disponibles. En nuestras disciplinas las estadísticas robustas todavía no se toman seriamente en consideración.

Mientras la arqueología se desentendió mayormente de la posibilidad de incurrir en errores de Tipo I debido al incumplimiento de los supuestos de normalidad, en la antropología de la corriente principal el asunto se desenvolvió en un registro parecido pero con intensidad mucho más leve. En un artículo característico de los *Annual Reviews* de mediados de los 80, el antropólogo Michael Chibnik (1985: 140), de la Universidad de Iowa, señaló que una falla común que se encuentra en los análisis estadísticos de la antropología y otras ciencias sociales es un énfasis excesivo en evitar los errores de Tipo I en detrimento de la evitación de los errores de Tipo II. Familiarizado en apariencia con la amplia literatura sobre el uso y el abuso de las pruebas de significación y conocedor de la debilidad antropológica en materia de muestreo, Chibnik describió el segundo tipo alegando que consiste en "rechazar incorrectamente la hipótesis alternativa". Dicha descripción no forma parte de la que rige en el dogma estadístico desde Neyman y Pearson (1933a), quienes definen los tipos de error no en base a la(s) hipótesis alternativa(s) sino en función de la hipótesis "sometida a prueba"

[*tested*] como ellos la llaman, y que no es otra que H_0. Como sea, Chibnik no tomó partido en la querella a favor o en contra de la NHST ni analizó el impacto de sus deficiencias en la investigación disciplinar.

Las posturas de Cowgill y de Chibnik tuvieron algún impacto en la elaboración del capítulo sobre prueba de hipótesis en *Statistics for Archaeologist: A commonsense approach* de Robert Drennan (1996: cap. 11), uno de los títulos más influyentes de finales del siglo XX. El texto es representativo de la aceptación levemente crítica de la NHST como parte del paquete metodológico de todo buen profesional. Más que servirse de la técnica para producir una proposición afirmativa o negativa desde el vamos, Drennan argumenta que debemos contemplar los niveles de significancia como "un esfuerzo por ponderar [*assess*] la probabilidad de que nuestros resultados no reflejen sólo los antojos del muestreo" (p. 163). De este modo, un nivel de significancia de 0,10, digamos, sugeriría que el error de muestreo "no es muy probable" como explicación del patrón de los datos, mientras que un nivel de 0,80 indicaría que el error de muestreo es "ampliamente probable". Aun cuando el autor promueva los "buenos usos" de la prueba de significancia, ni la lógica inherente a la prueba, ni la robustez de los parámetros, ni los supuestos de normalidad, ni las operaciones de normalización forzada (pp. 20-21) son puestos aunque más no fuere parcialmente en duda.

Las observaciones de Cowgill, Thomas y Chibnik acerca de las alternativas favorecidas por nuestros profesionales tampoco se han visto refrendadas en la práctica cotidiana de la investigación arqueológica, donde los números que se observan no son los que la teoría predice. Si bien no he podido articular un inventario exhaustivo (y en la ignorancia respecto de cuáles pudieron haber sido las políticas editoriales a propósito de la prueba estadística), de la revisión sumaria de la bibliografía no me es posible inferir si los arqueólogos de veras favorecen masivamente la evitación de los errores de Tipo I invirtiendo escaso esfuerzo en eludir el Tipo II. En el estudio de Jean Arnold y Annabel Ford (1980) sobre los patrones de asentamiento en Tikal, Guatemala, por ejemplo, las autoras tampoco realizan el ritual de rechazo de la HN que según los rumores en boga los arqueólogos tienden a privilegiar:

> Nuestros hallazgos arrojan dudas sobre el supuesto de que los centros del período Maya Clásico manifestaban un patrón de zonación concéntrica en términos del status de los residentes. El cálculo computacional de *tau*-B produjo el resultado de *tau*-B=0,03. Este coeficiente de correlación es extremadamente bajo, aproximándose a cero. La conclusión obvia es que los dos órdenes de rango tal como aquí se calcularon se ordenan casi al azar con respecto al otro y que toda correlación entre ambos es tan baja como para ser insignificante. Dados estos resultados no se puede rechazar la hipótesis nula. Por ende, se puede afirmar con cierta certidumbre que no hay una correlación positiva entre inversión de trabajo (o status, tal

como se utilizó aquí) y distancia a partir del punto central definido en Tikal (Arnold y Ford 1980: 722).

A lo largo de la era interpretativa, del período posmoderno y del giro ulterior hacia los estudios culturales la NHST se ha retraído en antropología tanto como lo hizo el análisis estadístico en su conjunto. Éste conoció su apogeo hace ya unos cuarenta años en el seno de la escuela murdockiana del análisis transcultural, la cual se mantiene viva en unos pocos enclaves olvidados de la mano de Dios sin haber llegado a ser nunca una estrategia cuyo prestigio estuviera a la altura del trabajo que en ella se invirtió. Tanto en la versión transcultural como en las modalidades independientes de escuela, la literatura disciplinaria meramente registra el uso de la NHST sin reportar nada específico o incurriendo en las equivocaciones de costumbre (cf. McEwen 1963; Watson y Graves 1966; Chaney y Ruiz Revilla 1969; Naroll 1970; Naroll y Cohen 1973; Henry 1976; y'Edinak y otros 1976; Thomas 1976; Hirschfeld, Howe y Lewin 1978; Pelto y Pelto 1978).

Lo llamativo del caso es que incluso la literatura antropológica que se pretende crítica de los métodos estadísticos vigentes dista mucho de ser genuinamente radical, por cuanto se agota en el cuestionamiento de los usos ocasionales pero no de los principios básicos, difiriendo de la corriente principal en la generalidad de las disciplinas sólo en la medida en que alguno que otro antropólogo se manifiesta incapaz de rechazar alguna que otra HN con mayor asiduidad de lo común (p. ej. Wilson 1957).

En antropología biológica, mientras tanto, la NHST se utiliza para dirimir polémicas allí donde los argumentos discursivos o numéricos inconciliables se muestran en paridad de fuerzas; sin que importe mucho la naturaleza de la querella original (que en rigor versa sobre los orígenes africanos de la Eva primordial) este razonamiento es típico de la especie:

> La segunda línea de evidencia ofrecida a favor de los orígenes africanos es que los africanos poseen niveles más elevados de diversidad de mtDNA. Yo señalé […] que nunca se presentó ninguna prueba de hipótesis nula de igual diversidad, y hasta que esta hipótesis nula sea rechazada no hay bases para afirmar la mayor diversidad de los africanos. [Mark] Stoneking sigue sin presentar esas pruebas, y utilizando las medidas de diversidad de Vigilante y otros […] que él cita, no puedo detectar una diferencia significativa entre el valor africano de 0,0208 y el valor asiático de 0,0175 (Templeton 1994: 142).

Este patrón de razonamiento, consistente en diferir la ejecución de la prueba estadística para mejor oportunidad, se muestra también y primordialmente en los estudios territoriales, incluso en los casos en que las únicas elaboraciones alternativas son de carácter verbal:

> Las pruebas adecuadas del modelo de defensibilidad económica utilizarían casos en los cuales varían las medidas cuantitativas de densidad de recursos y predicitibilidad, sea dentro de un

> grupo (a través del tiempo, como con los Ojibwa, o para diferentes clases de recursos, como con los Karimojong) o a través de grupos que comparten tecnologías y organizaciones sociales parecidas (como los grupos indios de la Cuenca-Meseta). Esta es la estrategia que hemos tratado de adoptar más arriba, pero debido a que los datos son inadecuados nos hemos visto forzados a adoptar un modo de argumentación cualitativo. Por añadidura, el modelo también podría ponerse a prueba examinando la evidencia para la hipótesis nula. En particular, si se pudiera demostrar que el cambio desde los sistemas no territoriales de organización espacial a los sistemas territoriales bien definidos ocurre con alguna frecuencia sin incrementos correlativos en las medidas de densidad de recursos y/o predictibilidad, el modelo tal como lo hemos presentado tendría que rechazarse (Dyson-Hudson y Smith 1978: 38).

Años más tarde las referencias a la prueba estadística aparecen de manera rutinaria en el análisis de redes sociales mixturada con el uso de estadísticas *lato sensu*, siendo indistinguible la forma que ha asumido en nuestra disciplina de lo que es el caso en otras ciencias humanas. En la literatura reticular se percibe aquí y allá una leve resistencia ante la NHST; se trata de una rebeldía inorgánica que no necesariamente se funda en las razones de mayor peso, que no ha sabido poner el dedo en la llaga de la diversidad de distribuciones existentes y que no ha tomado tampoco en consideración la polémica instalada desde hace tanto tiempo. A veces, sin embargo, allí se pone en relieve (aunque con mesura y bajo perfil) otra serie de problemas no menos insidiosos:

> [M]uchas de las herramientas de estadística inferencial estándar que hemos aprendido del estudio de distribuciones de atributos no se aplican directamente a datos de redes. La mayoría de las fórmulas estándar para calcular error estándar estimado, computar pruebas estadísticas y establecer la probabilidad de hipótesis nulas que hemos aprendido en estadística básica no funcionan con datos de red (y si se los usa nos pueden dar más respuestas de "falsos positivos" más a menudo que "falsos negativos"). Esto se debe a que las "observaciones" de valores en datos de red no son muestras "independientes" a partir de poblaciones (Hanneman 2005: cap. 18).

La falta de independencia de los elementos de la muestra señalada por Hanneman es exactamente la misma condición en la que se origina el famoso "problema de [Sir Francis] Galton", un impedimento empírico que el célebre estadístico, eugenista, geógrafo y psicómetra reveló a los antropólogos en los albores de la era evolucionista y que nunca terminó de resolverse a satisfacción de todos.

El problema de Galton (como lo llamó Raoul Naroll [1961; 1965]) no es sino lo que en geografía se conoce como autocorrelación reticular y espacial, uno de los mayores obstáculos para los cálculos estadísticos en los estudios que involucran GIS y territorialidad (Dow y otros 1984). El problema surge porque los elementos muestreados no son estadísticamente independientes, dado que se relacionan por influencia, contacto, derivación genética, contigüidad, incidencia, comunicación, migración, difusión, fisión. La presencia inevitable de este fenómeno en los estudios de antropología

transcultural y en la investigación de los fenómenos de globalización, sumado al uso ingenuo y generalizado de pruebas de independencia de chi cuadrado, ocasiona un alto número de rechazos incorrectos de la HN (errores de Tipo I), característicos de toda la investigación anterior a la publicación de los métodos correctivos de Dow (1993) y otros autores. Si se pretende preservar la prueba de hipótesis, casi todos los estudios de ARS y antropología comparativa existentes deberían ajustarse conforme a las nuevas pautas. Éstas implican cambios metodológicos en profundidad y una drástica reducción del nivel de significación de $p \leq 0{,}05$ a $p \leq 0{,}005$ o similar, un orden de magnitud por debajo de lo que Fisher se atrevió a conceder (Korotayev y de Munck 2003; Jahn 2006). A pesar de su tremendo impacto el problema ni siquiera aparece referido en el manual de ARS de Wasserman y Faust (1994).

La minimización del impacto acarreado por la falta de independencia de los datos en el análisis de redes, de hecho, es proverbial. Como si se olvidara de los reparos que él mismo trajo a colación Hanneman sigue luego desarrollando la prueba de hipótesis reticular a la manera acostumbrada, indicando además proactivamente la forma en que dicha operatoria debería llevarse a cabo mediante el software de ARS utilizado, el cual resulta ser, sintomáticamente, UCINET: Network > Compare densities > Against theoretical parameters… y así el resto. La información emergente de este procedimiento, derivada de un mecanismo conceptual cuyos fundamentos se mantienen prudentemente en una caja negra sin ser objeto de compulsa, es no obstante el indicador que definirá sin más trámite el rechazo o la aceptación de la HN. De más está decir que debido a que la relación entre los elementos es constitutiva de la estructura de toda red o grafo, en tanto se imponga un requisito de muestreo aleatorio e independencia de datos no se puede aplicar a las redes sociales o a las redes en general (sostengo aquí) ninguna clase imaginable de prueba estadística de hipótesis o significancia. Tampoco le son aplicables, ciertamente, gran parte de los procedimientos de las estadísticas paramétricas descriptivas.

A quince años de la epifanía de la ley de potencia y de la fusión entre las teorías de redes y la ciencia de la complejidad y el caos, es lástima que una facción del ARS, de tan alto potencial en la investigación compleja contemporánea, siga apegada a los lineamientos aleatoristas de la escuela de Harvard[35] y continúe concediendo crédito a metodologías que ya se sabían fallidas y que resultaron imposibles de defender incluso en la arqueología estadística de hace un cuarto de siglo (cf. Thomas

[35] Con el apoyo activo de importantes científicos y matemáticos (como Anatole Rapoport, George Yule y Herbert Simon), la escuela tuvo un rol activo en el rechazo ancestral de las distribuciones de Zipf incluso en los dominios lingüísticos y territoriales donde ellas prevalecen sin la menor sombra de duda. Miembros de esa escuela conservadora han sido Harrison White, David Krackhardt y por supuesto Wasserman & Faust (1994).

1978; Hole 1980; Ammerman 1992). Cuando la antropología tuvo oportunidad de expedirse sobre la controversia en torno de la NHST, sorprendentemente, tomó partido a favor de la conservación del método. Raoul Naroll, uno de los apóstoles de la antropología transcultural, escribe a propósito de *The significance test controversy*:

> En las dos últimas décadas, muchos críticos escépticos han sostenido que esta pregunta carece por completo de interés para la mayoría de los científicos de la conducta. Estos escépticos creen que las pruebas de significancia estadística son usualmente una pérdida de tiempo para los científicos de la conducta. Morrison y Henkel nos han hecho un servicio al recolectar las mejores de estas críticas. Los editores no tienen pretensiones de objetividad; están francamente en contra del uso general de pruebas de significancia; su selección es unilateral; y sus conclusiones (pp. 305-311) no son hallazgos de jueces imparciales, sino los argumentos culminantes de un par de abogados del diablo (Naroll 1971: 1437).

Sabiendo que la NHST ha sido mal utilizada y malinterpretada demasiadas veces, Naroll busca resolver su doble vínculo mediante una cabriola retórica, admitiendo que (conforme a las ideas de Popper) las pruebas estadísticas no pueden probar nada pero podrían ser admirables instrumentos de refutación. En una ingeniosa vuelta de tuerca, propone que la NHST se redefina como "prueba de *in*significancia" estadística. De lo que se trata es entonces de poner todo el empeño en refutar la HN y no intentar probar ninguna otra cosa más que el hecho de que el mero azar no parece sustentar nuestros hallazgos. "Si el azar por sí solo –termina pregutándose Naroll– no puede explicar nuestros hallazgos ¿Qué es lo que podría hacerlo?" (op. cit.: 1439). Un giro sagaz, lo admito; pero demasiado transparentemente interesado en dejar que las cosas queden como están.

Casi lo mismo cabe decir de otro empeño panglossiano, el que elaboraron Andrey Korotayev y Victor de Munck (2003) con el propósito de convertir el problema de Galton en el "recurso de Galton" [*Galton's asset*] haciendo de la necesidad virtud y sirviéndose a tal efecto de la idea de autocorrelación de redes propuesta por Dow, Burton, White y Reitz (1984). Este modelo propone no pensar tanto en unidades culturales aisladas y discretas, sino en redes de comunicación y en procesos históricos de difusión de rasgos, introduciendo severos ajustes en el cálculo de la correlación antes de abordar la prueba estadística propiamente dicha. La solución de Korotayev-Munck no resuelve en modo alguno el problema de la auto-correlación sino que apenas difiere su tratamiento, ocupándose de la prueba de hipótesis sólo de manera marginal y asistemática: la consideración de las unidades culturales –aseguran– debe ser juzgada caso por caso y dependiendo de la clase de hipótesis que se quiera probar. En cuanto a la prueba de significancia hay una letra chica que pocos entenderán por completo y que dice, expresamente:

> Una estrategia alternativa para testear el $\dot{p}$ generado ya sea por IGLS o por IRR para la significancia sería aceptar el $\dot{p}$ que se produce como una buena aproximación al $\dot{p}$ de ML e

> insertar el valor IGLS o IRR en un test de relación de verosimilitud [*likelihood*]. Sin embargo, los resultados de simulación aquí reportados sugieren que esto ofrecería sólo una prueba de significancia muy aproximativa, de modo que no examinamos esta posibilidad en este momento (Dow y otros 1984: 764).

Aunque he revisado gran parte de la bibliografía estadística de las publicaciones periódicas en antropología y arqueología, todavía está pendiente una evaluación sistemática del estado de la discusión y del uso de la NHST en ambas disciplinas. Si es menester sintetizar en pocas palabras qué es lo que este uso tiene de singular, habrá que reconocer que el papel que ha jugado nuestra academia en el campo en que se ha desenvuelto la polémica ha estado signado por un embarazoso conformismo epistemológico.

Ninguna de las dos especialidades, por lo pronto, ha tomado acción en lo que respecta a una práctica científica aberrante como lo es la eliminación de los *outliers*, es decir, la exclusión de los ejemplares cuyos valores se salen del rango admitido por las premisas de normalidad que estipulan los métodos de prueba (ver p. ej. Brewer 1986; Barnett y Lewis 1994). En ocasiones los *outliers* pueden ser sintomáticos del hecho de que la población posee una distribución no normal de "cola pesada", que casi siempre resulta ser logística, de Cauchy o de LP. Alcanza en rigor con un solo *outlier* para que las cantidades que rigen la inferencia en la estadística clásica (intervalos de confianza, estadísticas *t*, valores de R^2 y por supuesto valores de *p*) se salgan de madre (Maronna, Martin y Yohai 2006: 1). En las disciplinas humanísticas casi siempre se ha optado por excluir esos elementos aduciendo motivos contrapuestos:

- Los estudiosos de orientación más discursiva suelen invocar el prestigioso "criterio de Peirce". Si bien fue el ignoto Benjamin Peirce quien creara ese criterio en una de las pruebas de significancia más tempranas que se han publicado, nadie menos que Charles Sanders Peirce (el Peirce propiamente dicho, dirán algunos) homologó esa política de expulsión de los intrusos entre 1872 y 1878 (B. Peirce 1852; Ch. S. Peirce 1986).
- Los autores de perfil más técnico, en contraste, han tendido a sobreinterpretar los poderes de los respetados teoremas del límite central (TLC), restringiendo la libertad de los datos a apartarse de las pautas que rigen la distribución gaussiana.

Los teoremas del límite central (parientes próximos de la ley de los grandes números) establecen que la media de un número suficientemente grande de variables muestreadas al azar a partir de una población "de buen comportamiento" [*well behaved*] exhibirá una distribución normal (Fischer 2011). Por supuesto que se sabe que si la muestra es realmente *muy* grande la ley fallará en caso que la población posea una distribución sin media, asimétrica, de alta dimensión fractal o de cola pesada; pero si el muestreo se mantiene en los márgenes minimalistas que prescriben las técnicas

consagradas, los TLC garantizan que se producirán muestras con una distribución normal a partir de poblaciones moderada o fuertemente alejadas de la normalidad. Los libros de texto de estadística aplicada desaconsejan confiar en los TLC si la muestra es menor de 30 ejemplares; pero ninguno garantiza o siquiera mide la exactitud de los TLC con respecto al grado de la muestra y el grado de no-normalidad (Chatfield 1976; Spedding y Rawlings 1994).

Con todo, insisto en que el efecto de los TLC dista de ser tan universal como se pretende; ante distribuciones de Cauchy realmente severas casi cualquier técnica de muestreo (aleatorio simple, estratificado, *clustered*, por cuotas, por propósito, de bola de nieve, de conveniencia) genera otras distribuciones de Cauchy (Hole 1980: 229; Le Cam 1986: 81). Aplicando métodos de teoría de la información se ha demostrado recientemente que dependiendo de condiciones precisas el TLC ocasiona convergencia hacia otros tipos de atractores estables, tales como leyes de Cauchy, Bernoulli, Poisson y ley gaussiana inversa (o sea, distribución de Lévy) (cf. Johnson 2004: cap. 5).

Aunque dignos de ser tenidos en consideración por el papel histórico que han jugado en la consolidación de la ley normal como la distribución de referencia, los TLC no pueden considerarse robustos en presencia de *outliers*. Ahora bien, el mero hecho de llamar los *outliers* de ese modo denota que la estadística se encuentra en la misma situación descripta por Benoît Mandelbrot (2003: 87-124) cuando los geómetras pensaban que las curvas de extraña dimensionalidad que aparecían aquí y allá (el polvo de Cantor, las curvas de Koch, Gosper y Peano, la escalera del diablo) eran “curvas monstruosas”, anomalías indignas de tratamiento matemático. La política de supresión de datos estadísticos y de negación de la heterogeneidad o la discontinuidad cualitativa no es sólo una curiosidad matemática. Los métodos computacionales programados para identificar y suprimir *outliers* son responsables de haber retrasado la investigación sobre el agujero de ozono durante años (desde 1976 a 1985, por lo menos) por considerar que las desviaciones del 10% por debajo de la normalidad (180 unidades de Dobson) detectadas por los instrumentos TOMS del satélite Nimbus 7 se debían a errores en la toma y filtrado de datos. Revisados los programas del satélite y eliminados los filtros, se comprobó que el agujero venía siendo detectado por los sensores satelitales desde mucho antes sin que nadie hiciera nada al respecto (Sharman, Gardiner y Shanklin 1985). Tenemos aquí, sin duda alguna, por la propia especificación de un método de prueba que se autodestruye a medida que se despliega, uno de los más obscenos errores del Tipo II de la historia de la ciencia.

Dejando de lado las escuelas comparativas murdockianas de mediados del siglo pasado los antropólogos no han practicado demasiada estadística, ni qué decir tiene; pero, meteorólogos aparte, en pocas disciplinas se percibe una propensión tan vehemente y expeditiva a la eliminación de *outliers* (demasiado prestamente definidos como errores, ruidos o contaminaciones) como la que se encuentra en estadística arqueológica, en estudios de datación y en arqueometría en general. Al lado de

estas políticas que suscitaron polémicas de extrema tibieza se percibe también un amplio repertorio de prácticas de transformación logarítmica, normalización de rangos, estabilización de varianza y estandarización (Kamminga y Wright 1988; Frink 1992; Wolpoff 1993: 382-383; Neupert 1994: 715, 719; Baxter 1995: *passim*; Drennan 1996: 20-21; Baxter y Beardah 1997; Heidelberg 2001; Die 2004; Bubenzer, Hilgers y Riemer 2007; Ramsey 2009; Weber s/f). Esto no puede sino recordarnos la admirable frase de Steven Stigler: "Mientras más torture usted sus datos, más probable será que ellos confiesen; pero las confesiones obtenidas bajo dureza pueden no ser admisibles en la corte de la opinión científica" (Stigler 1987: 148).

A pesar que la disciplina se sueña siempre en un rol protagónico en la línea de fuego de las dudas metódicas, de la diferencia, de las pruebas ácidas y del cuestionamiento transgresor de la normalidad, la experiencia antropológica frente a la estadística ha sido hasta hoy más conservadora y menos inclinada a la rebelión creativa que la de la medicina, la psicología, la economía, la criminalística, la ecología o la veterinaria, acaso en ese orden decreciente.

La NHST fue expresamente rechazada en psicología por figuras de la talla de R. Duncan Luce, Herbert Simon (1992: 159) y B. F. Skinner (1972: 319) y en matemática estadística por John Tuckey y William Kruskal (1968a); nunca pudo instalarse tampoco en las ciencias formales más que a título precario (Gigerenzer y Murray 1987; Gigerenzer 1998b: 199; McCloskey y Ziliak 2007). En antropología, mientras tanto, aparte de las estrategias hermenéuticas o posmodernas que reniegan de los números a priori pero que dejan en pie todos y cada uno de los supuestos esenciales de la vieja episteme, jamás hubo un gesto de caución como no sea el que se propone en este ensayo. Fuera de incurrir en extravíos metodológicos que en nombre de la prueba de hipótesis niegan una diversidad que debería ser connatural a nuestras tácticas, nada hemos hecho en nuestra disciplina, en suma, que valga la pena destacar.

14. Conclusiones

> ¿Cuál es la probabilidad de que alguien tenga dos veces tu estatura? ¡Esencialmente cero! La altura, el peso y muchas otras variables están distribuidas en funciones de probabilidad "dóciles" con un valor típico bien definido y relativamente poca variación en torno suyo. La ley gaussiana es el arquetipo de las distribuciones "dóciles".
>
> ¿Cuál es la probabilidad de que alguien tenga el doble de tu fortuna? La respuesta depende por supuesto del monto de ella, pero en general hay una fracción no despreciable de la población que será dos, diez o incluso cien veces más adinerada que tú. Esto fue descubierto a finales del siglo [ante]pasado por Pareto, por quien se ha llamado así la ley que describe la [distribución de] ley de potencia de las fortunas, el ejemplo tipico de una distribución "salvaje".
>
> Didier Sornette (2006: 104)

Alguna vez habrá que hacerse cargo de que unos cuantos rasgos del estilo de las estadísticas vigentes en las ciencias humanas se derivan del hecho de que las estadísticas mismas no se desarrollaron ni en el corazón de las matemáticas ni en las ciencias llamadas duras o formales. En el campo estadístico cada tanto se publican manifiestos que reclaman un retorno a las matemáticas, o un encuentro entre éstas y la estadística (Yates y Healy 1964; Bailey 1998; Sprent 1998). Ya hemos visto que el propio Fisher (1935: 39), al defender su prueba de las acometidas de Neyman y Pearson, se quejaba de la falta de comprensión de "los matemáticos". En la gestación de los tests que nos preocupan "los matemáticos" del campo neymaniano se expresaron mediante pruebas [*proofs*] en el sentido cabal, con sus teoremas, generalizaciones, lemmas y simbolismos; Fisher no, en absoluto. No fue ni reclamó ser un matemático, aunque desde una ciencia humana que desconoce los estilos y las convenciones académicas pueda parecer hasta que lo fue con creces (cf. Savage 1967). Es sus últimos años Fisher se quejaba amargamente de la matematización de la estadística, aduciendo también que "es probable que un estadístico educado sólo en matemáticas espere que un problema tenga una sola solución" (Kruskal 1980: 1027; Savage 1976: 444). Aunque con el tiempo se articularon elaboraciones formales monolíticas de la teoría de la decisión (Wald 1950; Liese y Miescke 2008) y de la prueba estadística (Lehmann y Romano 2005) de muy escaso uso (y sin una sola implementación de referencia) en las ciencias humanas empíricas, la formulación de base de la NHST ha sido rara, incongruente, reflejando en cada acento sus condiciones agonísticas de gestación y sus contradicciones constitutivas; la ciencia estadística monolítica, apacible, consensuada y en estado puro que promueven los estereotipos –nos damos cuenta ahora– no es más que una desacreditada leyenda urbana.

En lo que a los estereotipos respecta, en las ciencias sociales hay quienes creen que el modelado matemático es invariablemente cuantitativo, que la estadística y las matemáticas son la misma cosa, que aquélla es coextensiva a la NHST y que los métodos cualitativos que valen la pena sólo pueden ser del orden de la discursividad y que ni las matemáticas, ni la lógica, ni la algorítmica compleja califican como tales (Denzin y Lincoln 2000; Given 2008). Nada de esto es exacto, desde ya, ni tampoco es cierto que las estrategias interpretativas o de orden literario (sea por no cuantificar sus parámetros o por rehusarse a interrogar sus métodos) estén exentas de los mismos constreñimientos de linealidad, proporcionalidad, simetría, monotonía y representatividad que hemos visto minan la operatoria de la prueba de hipótesis y que constituyen la raíz del problema (cf. Kruskal y Mosteller 1979a; 1979b; 1979c; 1980; Abbott 1988).

Ahora bien, corregir o armonizar esas visiones dudosas en las condiciones en las que la experiencia de la NHST nos ha dejado es cualquier cosa excepto fácil. En materia de estadísticas por ningún lado hay un corpus o una escuela a la que pueda singularizarse como el ejemplo a seguir. La literatura especializada que aquí hemos interpelado no ayuda mucho a clarificar el terreno: no sólo todavía no ha ajustado sus mecanismos para dar cuenta de una diversidad de distribuciones que se conoce desde Pareto, sino que aun no respondió satisfactoriamente al planteo del problema de Galton, el primer dilema metodológico con el que se enfrentó la antropología.

Lo que es más grave, un puñado entre los científicos de mayor prestigio que hemos visto oponiéndose a las fallas lógicas de la NHST (Hans Eysenck, David Lykken, Paul Meehl) ha tomado partido a favor de los argumentos de esa apoteosis de la distribución normal, de las inferencias de base unimodal y de las correlaciones lineales que es el infame *The Bell Curve* (Herrnstein y Murray 1994), alegando, más específicamente, que incluso la población negra de clase más pudiente se encuentra unos cuantos puntos por debajo de los blancos más pobres de los Estados Unidos en materia de coeficiente intelectual (The JBHE Foundation 1996: 19).

A lo largo de los cuarenta años que ya lleva la siesta interpretativa y posmoderna, también las estrategias del modelado matemático en las ciencias humanas y en la antropología en particular se fueron relajando. La revista señera *Mathematical Anthropology* dejó de circular (sustituida por *Mathematical Anthropology and Cultural Theory*),[36] los métodos cuantitativos desaparecieron de la currícula, las técnicas comparativas se contrajeron hacia el interior de un par de instituciones residuales, el análisis de redes sociales se desechó como “un caballo muerto” y hasta la analítica del parentesco se disolvió en el aire (cf. Reynoso 2011: 373-408). Así como la teoría de las redes sociales tomó distancia de sus fundamentos en la teoría de grafos, las estadísticas en general y la NHST en

[36] Véase http://mathematicalanthropology.org/.

particular, al amparo de un cuantitativismo que cada día parece más bizarro y autista, encontraron la forma de desentenderse casi por completo de los rigores formales sin dejar por ello de presentarse como la encarnación de las matemáticas *lato sensu* de cara a nuestras disciplinas. En definitiva, hacia los años 60 la práctica estadística cristalizó como una especie de consultoría consagrada a procedimientos que presuponen una sola clase de función de probabilidad, privativa de una familia de distribuciones que hace poco más de un siglo se encontraba en todas partes y ahora parece no encontrarse en ninguna.

Es el mismo Edgeworth que acuñara la idea de significancia a fines del siglo XIX y que aprobara las pruebas de hipótesis de los experimentos psíquicos de Richet quien nos da indicios de las razones de este despropósito:

> Los métodos de la Estadística son tan diversos como las definiciones de la ciencia. Los límites de este estudio se han fijado con referencia a las definiciones de Estadística mejor acreditadas. Hay tres definiciones que parecen merecer atención como respectivamente la más popular, la más filosófica y la que constituye el mejor compromiso entre los requisitos en conflicto de una buena definición. De acuerdo con la primera de estas definiciones, la Estadística es la porción aritmética de la Ciencia Social (que trata no sólo con cifras que fluctúan en torno de una Media tales como las tasas de mortandad, sino con lo que se puede llamar retornos solitarios, tales como el número de personas que han sido muertos en una batalla, o el número de ganado que ha muerto en una plaga). De acuerdo con la segunda definición, la Estadística es la ciencia de las Medias en general (incluyendo las observaciones físicas); de acuerdo con la tercera definición, de esas Medias que están presentadas por los fenómenos sociales (Edgeworth 1885a: 181-182).

Es imposible no advertir que junto con el apego conservador a la idea de *media* (que carece de sentido –o posee un sentido inusual– en las distribuciones sujetas a una LP) ya se percibía tempranamente una distinción radical entre las cifras de población que efectivamente fluctúan en torno a una media y las que obedecen a distribuciones que podríamos llamar de [Lewis Fry] Richardson, tales como el número de contiendas en las que se presenta un cierto número de víctimas fatales, el número de ciudades que tienen cierta cantidad de habitantes, o las magnitudes y frecuencias con que ocurren los cambios en los procesos de criticalidad auto-organizada (cf. más arriba, pág. 82). En estos escenarios es más orientador pensar en la idea de invariancia de escala (una característica cualitativa que luego se revelará fractal) que tratar de determinar bizantinamente medidas de promedio (Richardson 1948; cf. Reynoso 2011a: 212-214). Como quiera que sea, a renglón seguido el propio Edgeworth no acierta a poner en foco el carácter irreductible a la media que es típico de esta clase de distribuciones y sigue adelante como si la media fuera un parámetro robusto y ley normal la pauta a la cual hay que atenerse.

Si bien existen abundantes métodos no paramétricos de estimación y descripción que no presuponen la vigencia de determinadas clases de leyes distributivas y que en algunos casos son bastante parcos en materia de presuposiciones,[37] en la literatura técnica de las disciplinas empíricas no existe siquiera el esbozo de cómo podría ser una NHST adaptada a distribuciones no normales sin medias o varianzas a la vista, o con medias caóticas infinitamente susceptibles a las más mínimas variaciones en las condiciones iniciales: una prueba de hipótesis respetuosa del hecho de que los datos quizá no provengan de muestreos al azar en una población simétrica, sino de alguna clase de abstracción congruente con las dinámicas implicadas en la estructura peculiar del objeto.

La normalidad quizá haya sido el escenario estadístico de más fácil gestión, pero no por aferrarse a ella la práctica en torno de la NHST resultó exitosa. La revisión de la literatura crítica, aunque somera, alcanza para poner la idea misma de prueba estadística bajo la más seria vigilancia aun cuando las distribuciones normales y las escalas lineales sean la norma, lo cual se sabe bien que no es el caso. En último análisis está muy claro, sean cuales hayan sido los gestos teatrales, las impericias y los sesgos de la crítica, que hoy en día la NHST es más un obstáculo que una herramienta productiva. “No es ninguna sorpresa –escribía hace ya mucho el legendario W. Edwards Deming– que los estudiantes experimenten problemas [con la prueba estadística]: Están tratando de pensar” (Deming 1975: 150).

Desde ya que existen otros campos de la estadística que son objeto constante de malentendidos y estereotipos (Gore, Jones y Rytter 1976; Scheps 1982; Spirer, Spirer y Jaffe 1998; Strasak y otros 2007). James Brewer (1986: 129), por ejemplo, menciona equívocos en torno de los TLCs y de las distribuciones de muestreo que son característicos, tales como la creencia en que las medias de cualquier variable que sea muestreada repetidamente y al azar mostrarán tendencia a adoptar una distribución normal, tanto más cuanto mayor sea la muestra y sin que deba cumplirse ningún otro requisito. Bonnie Hole (1980: 229) ya había anticipado una crítica parecida a un prejuicio similar. Tal parece que la documentación de las ilusiones estadísticas, por ponerle un nombre, es un género literario establecido. Pero las equivocaciones que involucran a la NHST son por amplio margen las más flagrantes, las más tediosas y las más populares del mercado.

---

[37] Entre los más comunes cabe mencionar (por orden alfabético) la prueba de Anderson-Darling, los métodos de *bootstrap*, la prueba *Q* de Cochran, el coeficiente kappa de Cohen, el análisis de varianza de dos vías de Friedman, el estimador de Kaplan-Meier, el coeficiente de correlación de rango τ de Kendall, el coeficiente de concordancia *W* de Kendall, la prueba de Kolmogorov-Smirnov, el análisis de varianza de una vía de Kruskal-Wallis, la prueba de Kuiper, la prueba de *logrank* o de Mantel-Cox, la prueba *U* de Mann-Whitney, la prueba de Siegel-Tukey, el coeficiente de correlación de Spearman, los métodos de ondículas [*wavelets*], la prueba de Wald-Wolfowitz y la prueba de Wilcoxon. En el ámbito académico ninguna de estas pruebas ha sido jamás competencia para la de Fisher-Neyman-Pearson y sus prerrequisitos paramétricos (Sprent y Smeeton 2001; Wasserman 2006).

En la batalla por dirimir o re-establecer la estrategia dominante en el campo metodológico, la NHST, representativa de las ideas frecuentistas, parece estar cediendo terreno a metodologías bayesianas decididamente subjetivistas que no todos los estadísticos, matemáticos y lógicos encuentran adecuadas. Previsiblemente, dentro y fuera de los campos frecuentista y bayesiano se han imaginado procedimientos alternativos (información de Kullback-Leibler, intervalos de confianza, intervalos fiduciales, potencia estadística, análisis del tamaño del efecto, estimación de parámetros, meta-análisis, *model-fitting*, procedimientos PPE, métodos gráficos, pruebas tri-valuadas, replicación experimental) aunque tampoco hay acuerdo en lo que a su eficacia concierne (Denis 2003; Fidler y otros 2006). Pero ni aun la suma de todas las alternativas llega a reunir la masa crítica que se requiere para sustituir la prueba de hipótesis sin que subsista la impresión de que algo importante se ha perdido.

La mayor parte de las alternativas examinadas, por otro lado, siguen sosteniendo tenaces supuestos de linealidad; el tamaño del efecto, por ejemplo, sobre el que se cifran esperanzas que se desenvuelven a lo largo de libros enteros (v. gr. Ellis 2010), se define como la fuerza de la relación entre la variable independiente y la dependiente. Es fácil ver que el parámetro sólo se aplica a sistemas monótonos y rígidamente lineales, lo cual deja fuera de consideración cualquier escenario medianamente complejo en el que los exponentes de los argumentos de las ecuaciones que lo rigen difieran de la unidad o en el que se manifiestan conductas emergentes y no-convexidad. Diversos autores, incidentalmente, han sugerido que la expresión “tamaño del efecto” sea sustituida por “tamaño del resultado” para evitar connotaciones de causalidad; pero la cosmética de los nombres no llega a corregir la precaria estructuración de estas ideas en particular (cf. Shaver 1993: 19).

En otro orden de cosas, los métodos clásicos de la estadística, primordialmente paramétricos, han cedido unas pocas posiciones a otras concepciones metodológicas que parecen mejor preparadas para afrontar datos de la vida real. Hacia fines del siglo pasado, sin embargo, ha llegado a constituirse una imagen idealizada de lo robusto, lo independiente-de-distribución y lo no-paramétrico; algunos reputan estos reacomodamientos un canto de sirenas (Johnson 1998: 2000), otros los asimilan a un *bandwagon* (Huber y Ronchetti 2009: xv); yo, como incurable antropólogo, los encuentro más bien afines a un *cargo cult* (Sprent y Smeeton 2001; Wasserman 2006; Guthery 2008). Sin ánimo de minimizar la promesa de estas estadísticas adaptativas que sólo parecen efectivas ante poquísimas desviaciones muy pequeñas y que no corrigen las fallas lógicas constitutivas de la inferencia probabilística, sostengo que el próximo cambio debería ser algo diferente a una mera reforma; cien años de irreflexividad metodológica resultan suficientes para que debamos resignarnos a más de lo mismo.

Por eso es que (al lado de las exploraciones de William Kruskal en los confines de la no-normalidad) una de las pocas visiones que comparto a propósito de las alternativas opuestas o complementarias al "ritual nulo" es la de Julian Marewski y Henrik Olsson (2009) del Instituto Max Planck, quienes piensan que ya ha pasado el tiempo de contrastar las teorías contra el azar y que ya es hora de comenzar a contrastar las teorías entre sí mediante técnicas de modelado no necesariamente cuantitativas. Ya lo había sintetizado magistralmente Gerd Gigerenzer (1998b: 200), también del Max Planck: "En una ciencia que pugna por precisos modelos de procesos, se necesitan métodos que contrasten las predicciones de un modelo con las de modelos alternativos, y no un ritual que pruebe una hipótesis no especificada contra el azar".

Aunque a veces parecería que la polémica sobre la NHST no ha servido para otra cosa que para robustecer el status quo en un extremo y para alimentar una tribu desproporcionadamente numerosa de refutadores de leyendas en el otro, la neutralidad no es una opción. Dado que en muchas disciplinas empíricas las prácticas que se han consolidado alrededor de la técnica no han probado ser particularmente sustentables y hasta poseen sobrado potencial para causar un daño muy serio a los objetos de su estudio (la gente, la salud, la nutrición, las especies, el territorio, el ambiente) la apertura hacia el cambio, aunque fuese tentativa y exploratoria, luce como un curso de acción más razonable que la insistencia en el uso excluyente de técnicas que ya han tenido su oportunidad histórica, que han producido más pérdidas que ganancias y que han superado el término de su vida útil (cf. Fidler y otros 2006: 1543; McCloskey y Ziliak 2008).

La postura expresada en este ensayo no promueve la alianza con epistemologías sospechosas de subjetivismo incapaces de (o renuentes a) intervenir en la transformación de la realidad, ni propone enriquecer el repertorio de técnicas estadísticas alternativas que al cabo comparten los mismos o parecidos supuestos de aleatoriedad, linealidad, unimodalidad y monotonía. Existe un número formidable de métodos estadísticos que no son precisamente de prueba de hipótesis, y un número mucho más grande todavía de estilos de análisis, diagnosis, prueba, simulación y proyección que no son necesariamente estadísticos.

Sin que me resulte posible describir aquí dichas heurísticas, cuya implementación he descripto detenidamente en otros textos (Reynoso 2006; 2010; 2011a), la perspectiva que aquí se sustenta finca más bien en la convicción de que la existencia de distribuciones estadísticas sin sombra de normalidad, multimodales, no convexas, ruidosas e independientes de escala, características de la complejidad organizada, reclama ya no rutinas de prueba imposibles de cumplimentar y disparadoras de interpretaciones divergentes, sino una epistemología del modelado complejo que se encuentre a la altura de lo que hoy es posible pensar y llevar a la práctica.

Referencias bibliográficas

Abbott, Andrew. 1988. "Transcending General Linear Reality". *Sociological Theory*, 6(2): 169-186. http://www.jstor.org/pss/202114. Visitado en julio de 2011.

Abelson, Robert. 1997. "On the surprising longevity of flogged horses: Why there is a case for the significance test". *Psychological Science*, 8(1): 12-15.

Abelson, Robert. 1997. "A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented". En: L. Harlow, S. Mulaik y J. Steiger (editores), *Op. cit*., pp. 117-144. http://homepage.psy.utexas.edu/homepage/class/Psy391P/Abelson.pdf. Visitado en julio de 2011.

Altman, Douglas G. 1998. "Statistical reviewing for medical journals". *Statistics in Medicine*, 17: 2661-2674.

Altman, Douglas G. y J. Martin Bland. 1995. "Absence of evidence is not evidence of absence". *BMJ*, 311: 485311.

Altmann, Gerry. 2007. "Editorial. Journal policies and procedures". *Cognition*, 102: 1-6.

Ammerman, Albert. 1977. Revisión de *Figuring anthropology* de David Hurst Thomas. *American Anthropologist*, 79: 456.

Ammerman, Albert. 1992. "Taking stock of quantitative archaeology". *Annual Review of Anthropology*, 21: 231-255.

Anastasi, Anne. 1958. *Differential psychology*. 3ª edición, Nueva York, Macmillan.

Anderson, David, Kenneth Burnham y William Thompson. 2000. "Null hypothesis testing: Problems, prevalence, and an alternative". *Journal of Wildlife Management*, 64(4): 912-923. http://warnercnr.colostate.edu/~anderson/PDF_files/TESTING.pdf. Visitado en julio de 2011.

Arbuthnott, John. 1710. "An argument for divine providence taken from the constant regularity observ'd in the births of both sexes". *Philosophical Transactions of the Royal Society*, 27: 186-190. http://www.stats.org.uk/statistical-inference/Arbuthnott1710.pdf. Visitado en julio de 2011.

Armstrong, Scott. 2007a. "Significance tests harm progress in forecasting". *International Journal of Forecasting*, 23: 321-327. http://marketing.wharton.upenn.edu/ideas/pdf/Armstrong/StatSigIJF26.pdf. Visitado en julio de 2011.

Armstrong, Scott. 2007b. "Statistical significance tests are unnecessary even when properly done". *International Journal of Forecasting*, 23: 335-336. http://works.bepress.com/j_scott_armstrong/50/. Visitado en julio de 2011.

Arnold, Jeanne y Annabel Ford. 1980. "A statistical examination of settlement patterns at Tikal, Guatemala". *American Antiquity*, 45(4): 713-726.

Bailey, Rosemary A. 1998. "Statistics and mathematics: The appropriate use of mathematics within statistics". *Journal of the Royal Statistical Society*, 47(2): 261-271.

Bakan, David. 1960. "The test of significance in psychological research". *Psychological Bulletin*, 66: 423-437. http://stats.org.uk/statistical-inference/Bakan1966.pdf. Visitado en julio de 2011.

Balakrishnan, Narayanaswami. 1992. *Handbook of the logistic distribution*. Nueva York, Marcel Dekker.

Balakrishnan, Narayanaswami y Valery B. Nevzorov. 2003. *A primer on statistical distributions*. Hoboken, Wiley.

Barabási, Albert-László. 2002. *Linked. The new science of networks*. Cambridge, Perseus Publishing.

Barnett, Vic y Toby Lewis. 1994. *Outliers in Statistical Data*. 3ª edición, Nueva York, John Wiley & Sons.

Batanero, Carmen. 2000. "Controversies around the role of statistical tests in experimental research". *Mathematical Thinking and Learning*, 2(1-2): 75-97. http://www.ugr.es/~batanero/ARTICULOS/mtl.PDF. Visitado en julio de 2011.

Batanero, Carmen y Carmen Díaz. 2006. "Methodological and didactical controversies around statistical inference". http://www.ugr.es/~batanero/ARTICULOS/Paradigma.pdf.

Bateson, Gregory. 1980. *Espíritu y naturaleza.* Buenos Aires, Amorrortu.

Baxter, M. J. 1995. "Standardization and transformation in principal component analysis, with applications to archaeometry". *Journal of the Royal Statistical Society*, series C, 44(4): 513-527.

Baxter, M. J. y C. C. Beardah. 1997. "Some archaeological applications of kernel density estimates". *Journal of Archaeological Science*, 24: 347–354. http://faculty.ksu.edu.sa/archaeology/Publications/arch.%20Survey/Archaeology%20and%20geostatistics_files/Archaeological%20Applications%20of%20Kernel%20Density%20Estimates.pdf. Visitado en agosto de 2011.

Bedeian, Arthur G., Shannon Taylor y Alan N. Miller. 2010. "Management science on the credibility bubble: Cardinal sins and various misdemeanors". *Academy of Management Learning & Education*. 9(4): 715-725.

Berger, James O. 2003. "Could Fisher, Jeffreys and Neyman could agreed on testing?". *Statistical Science*, 18(1): 1-32. http://www.stat.duke.edu/~berger/papers/02-01.pdf. Visitado en julio de 2011.

Berger, James O. y Donald A. Berry. 1988. "Statistical analysis and the illusion of objectivity". *American Scientist*, 76: 159-165. http://drsmorey.org/research/rdmorey/bibtex/upload/Berger:Berry:1988.pdf. Visitado en julio de 2011.

Berger, James O. y Thomas Sellke. 1987. "Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion)". *Journal of the American Statistical Association*, 82(397): 112-122. http://www.stat.duke.edu/courses/Spring07/sta215/lec/BvP/BergSell1987.pdf. Visitado en julio de 2011.

Berkson, Joseph. 1938. "Some difficulties of interpretation encountered in the applications of the chi-square test". *Journal of the Americal Statistical Association*, 33(203): 526-542. http://www.stats.org.uk/statistical-inference/Berkson1938.pdf. Visitado en julio de 2011.

Berkson, Joseph. 1942. "Tests of significance considered as evidence". *Journal of the American Statistical Association*, 37: 325-335. Reimpreso en *International Journal of Epidemiology*, 32 (2003): 687-691. http://www.botany.wisc.edu/courses/botany_940/06EvidEvol/papers/1Berkson.pdf. Visitado en julio de 2011.

Bernard, H. Russell. 1995. *Research methods in anthropology. Qualitative and quantitative approaches*. 2ª edición, Nueva York, AltaMira Press.

Bernoulli, Daniel. 1734. "Quelle est la cause physique de l'inclinaison des plans des orbites des planétes par rapport au plan de l'équateur de la revolution du soleil autour de son axe; Et d'óu vient que les inclinaisons de ces orbites sont differentes entre elles". *Recueil des Pièces qui ont Remporté le Prix de l'Académie Royale des Sciences*, 3: 93-122. Traducción al inglés en http://www.cs.xu.edu/math/Sources/DanBernoulli/1734_planets%20and%20comets.pdf. Visitado en agosto de 2011.

Blume, Jeffrey y Jeffrey Peipert. 2003. "What your statistician never told you about *P*-values". *The Journal of the American Association of Gynecologic Laparoscopists*, 10(4): 439-444. http://www.cceb.med.upenn.edu/pages/courses/EPI520/2006/Blume_pvalues_2003.pdf. Visitado en agosto de 2011.

Bolles, Robert C. 1962. "The difference between statistical hypotheses and scientific hypotheses". *Psychological Reports*, 11: 639-645.

Boneau, C. Alan. 1960. "The effects of violations of assumptions underlying the *t* test". *Psychological Bulletin*, 57: 49-64.

Boring, Edwin G. 1919. "Mathematical vs scientific significance". *Psychological Bulletin*, 16: 335-338.

Boring, Edwin G. 1926. "Scientific induction and statistics". *The American Journal of Psychology*, 37(2): 303-307.

Brief of *Amici Curiae* of Professors Kenneth Rothman, Noel Weiss, James Robins, Raymond Neutra and Steven Stellman, in Supp. of *Pet'rs, Daubert v. Merrell Dow Pharms, Inc*. 1993. US 579, N° 92-102, WL 12006438, at *5.

Bradley, James V. 1977. "A common situation conducive to bizarre distribution shapes". *The American Statistician*, 31: 147-150. http://www.jstor.org/pss/2683535. Visitado en agosto de 2011.

Bradley, James V. 1978. "Robustness?". *British Journal of Mathematical and Statistical Psychology*, 31: 144-152. http://www.psych.umn.edu/faculty/waller/classes/mult11/readings/Bradley1978.pdf. Visitado en julio de 2011.

Bradley, James V. 1980. "Nonrobustness in *z*, *t*, and *F* tests at large sample sizes". *Bulletin of the Psychonomic Society*, 16: 333-336.

Brewer, James. 1986. "Behavioral statistics textbooks: Source of myths and misconceptions?". *ICOTS*, 2: 127-131. http://www.jstor.org/pss/1164796. Visitado en julio de 2011.

Bubenzer, Olaf, Alexandra Hilgers y Heiko Riemer. 2007. "Luminescence dating and archaeology of Holocene fluvio-lacustrine sediments of Abu Tartur, Eastern Sahara". *Quaternary Geochronology*, 2: 314-321.

Buehler, Robert J. y A. P. Federsen. 1963. "Note on a conditional property of Student's t". *Annals of Mathematical Statistic*s, 34(3): 1098-1100.

Burdick, D. S. y E. F. Kelly. 1977. "Statistical methods in parapsychological research". En: B. B. Wolman (editor), *Handbook of Parapsychology*, Nueva York, Van Nostrand Reinhold, pp. 81-130.

Carrier, James (editor). 2005. *A handbook of economic anthropology*. Cheltenham, Edward Elgar.

Carver, Ronald P. 1978. “The case against statistical hypothesis testing”. *Harvard Educational Review*, 48: 378-399. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.780&rep=rep1&type=pdf. Visitado en julio de 2011.

Carver, Ronald P. 1993. “The case against statistical significance testing, revisited”. *Journal of Experimental Education*, 61: 287-292. http://www.jstor.org/pss/20152382. Visitado en julio de 2011.

Catt, Issac E. 1984. Revisión de *Local knowledge*, de C. Geertz. *Literature in Performance*, 5(1): 59-60.

Chaney, Richard y Rogelio Ruiz Revilla. 1969. “Sampling methods and interpretation of correlation: A comparative analysis of seven cross-cultural samples”. *American Anthropologist*, 71(4): 597-633.

Chatfield, Christopher. 1976. *Statistics for Technology: A course in applied statistics*. Londres, Chapman & Hall.

Chew, V. 1977. “Statistical hypothesis testing: an academic exercise in futility”. *Proceedings of the Florida State Horticultural Society*, 90 : 214-215. http://www.fshs.org/Proceedings/Password%20Protected/1977%20Vol.%2090/214-215%20%28CHEW%29.pdf. Visitado en julio de 2011.

Chibnik, Michael. 1985. “The use of statistics in sociocultural anthropology”. *Annual Review of Anthropology*, 14: 135-157.

Chow, Siu L. 1988. “Significance test or effect size?”. *Psychological Bulletin*, 70: 426-443.

Chow, Siu L. 1998. “Précis of Statistical significance: Rationale, validity, and utility”. *Behavioral and Brain Sciences*, 21: 169-239.

Clark, Cherry Ann. 1963. “Hypothesis testing in relation to statistical methodology”. *Review of Educational Research*, 33: 455-473. http://www.stats.org.uk/statistical-inference/Clark1963.pdf. Visitado en julio de 2011.

Coats, W. 1970. “A case against the normal use of inferential statistical models in educational research”. *Educational Researcher*, 6-7.

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

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, Lawrence Erlbaum.

Cohen, Jacob. 1990. “Things I have learned (so far)”. *American Psychologist*, 95(12): 1304-1312. http://www.uvm.edu/~bbeckage/Teaching/DataAnalysis/AssignedPapers/Cohen_1990.pdf. Visitado en julio de 2011.

Cohen, Jacob. 1994. “The earth is round (p< .05)”. *American Psychologist*, 49: 997-1003. http://www.psych.yorku.ca/sp/Amer%20Psychologist%201994%20Cohen%20Earth%20is%20Round.pdf. Visitado en julio de 2011.

Cohen, Jacob y P. Cohen. 1983. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, Lawrence Erlbaum.

Cohen, Louis y Michael Holliday. 1982. *Statistics for social scientists. An introductory text with computer programs in Basic*. Londres, Harper & Row.

Consul, Prem y Felix Famoye. 2006. *Lagrangian probability distributions*. Boston-Basilea-Berlín, Birkhäuser.

Cooper, D. C. 1972. “Theorem Proving in Arithmetic without Multiplication”. En: B. Meltzer y D. Michie (compiladores), *Machine Intelligence*. Edinburgo, Edinburgh University Press, pp. 91–100.

Cornaglia, P. S., G. E. Schrauf, M. Nardi y V. A. Deregibus. 2005. “Emergence of dallisgrass as affected by soil water availability”. *Rangeland Ecology & Management*, 58: 35-40.

Cortina, José M. y William P. Dunlap, 1997. “On the logic and purpose of significance testing. Psychological Methods”. *Psychological methods*, 2(2): 161-172. http://www2.psych.ubc.ca/~schaller/349and449Readings/CortinaDunlap1997.pdf. Visitado en julio de 2011.

Cowgill, George. 1977. “The trouble with significance tests and what can we can do about it”. *American Antiquity*, 42(3): 350-368.

Cowles, Michael y Caroline Davis. 1982. “On the origins of the .05 level of statistical significance”. *American Psychologist*, 37(5): 553-558.

Cox, David Roxbee. 1986. “Some general aspects of the theory of statistics”. *International Statistical Review*, 54: 117-126.

Cramér, Harald. 1946. *Mathematical methods of statistics*. Princeton, Princeton University Press.

Cronbach, Lee J. 1975. “Beyond the two disciplines of scientific psychology”. *American Psychologist*, 30: 116-127.

Cronbach, Lee J. y Richard E. Snow. 1977. *Aptitudes and instructional methods: A handbook for research in interactions*. Nueva York, Irvington.

Currell, Graham y Antoni Dowman. 2009. *Essential mathematics and statistics for science*. 2ª edición, Chichester, John Wiley & Sons. http://www.blackwellpublishing.com/currellmaths2/. Visitado en julio de 2011.

Curry, Michael G. 1985. Revisión de *Local knowledge*, de C. Geertz. *Annals of the Association of American Geographers*, 75(2): 291-293.

Daniel, Larry. 1993. “Statistical Significance Testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals”. *Research in the schools*, 5(2): 23-32. http://www.personal.psu.edu/users/d/m/dmr/sigtest/3mspdf.pdf. Visitado en julio de 2011.

Dar, Reuven. 1987. “Another look at Meehl, Lakatos, and the scientific practices of psychologists”. *American Psychologist*, 42: 145-151.

Dar, Reuven, Donald Serlin y Haim Omer. 1994. “Misuse of Statistical Tests in Three Decades of Psychotherapy Research”. *Journal of Consulting and Clinical Psychology*, 62(1): 75-82. http://www.uv.es/~friasnav/DarEtAl1994.pdf. Visitado en julio de 2011.

Deming, W. Edwards. 1975. “On probability as a basis for action”. *American Statistician*, 29: 146-152. http://www.stat.osu.edu/~jas/stat600601/articles/article1.pdf. Visitado en agosto de 2011.

Dempster, A. P. 1963a. “Further examples of inconsistencies in the fiducial argument”. *Annals of Mathematical Statistics*, 34(3): 884-89.

Dempster, A. P. 1963b. “On a paradox concerning inference about a covariance matrix”. Annals of Mathematical Statistics, 34(3): 884-89.

Denis, Daniel J. 2003. “Alternatives to null hypothesis significance testing”. *Theory and Science*, 14(1), http://theoryandscience.icaap.org/content/vol4.1/02_denis.html.

Denzin, N. K. y Y. S. Lincoln (editores). 2000. *Handbook of qualitative research*. 2ª edición. Thousand Oaks, Sage.

Dow, Malcolm. 1993. "Saving the theory: Chi-squared tests with cross-cultural survey data". *Cross-cultural research*, 27: 247. http://dx.doi.org/10.1177%2F106939719302700305. Visitado en junio de 2011.

Dow, Malcolm, Michael Burton, Douglas R. White y Carl Reitz. 1984. "Galton's problem as network auto-correlation". *American Ethnologist*, 11(4): 754-770

Drennan, Robert. 1996. *Statistics for archaeologist: A commonsense approach*. Nueva York, Plenum Press.

Durkheim, Émile. 1893. *De la division du travail social*. http://www.uqac.ca/zone30/Classiques_des_sciences_sociales/classiques/Durkheim_emile/division_du_travail/division_travail.html. Visitado en julio de 2011.

Durkheim, Émile. 1895. *Les règles de la méthode sociologique*. http://www.uqac.ca/zone30/Classiques_des_sciences_sociales/classiques/Durkheim_emile/regles_methode/regles_methode.html. Visitado en julio de 2011.

Durkheim, Émile. 1897. *Le suicide. Étude de sociologie*. http://www.uqac.ca/zone30/Classiques_des_sciences_sociales/classiques/Durkheim_emile/suicide/suicide.html. Visitado en julio de 2011.

Dye, Thomas. 2004. "Bayesian statistics for archaeologists". IGERT Program, University of Arizona, http://www.tsdye.com/research/ua/ua-bayesian-lecture.pdf.

Dyson-Hudson, Rada y Eric Alden Smith. 1978. "Human territoriality: An ecological reassessment". *American Anthropologist*, 80(1): 21-41.

Edgeworth, Francis Ysidro. 1885a. "Methods of statistics". *Jubilee Volume of the Statistical Society*, Royal Statistical Society of Britain, 22–24 June, pp. 181-217. http://www.jstor.org/pss/25163974. Visitado en julio de 2011.

Edgeworth, Francis Ysidro. 1885b. "The Calculus of Probabilities Applied to Psychical Research". *Proceedings of the Society of Psychical Research*, 3: 190-199.

Edgeworth, Francis Ysidro. 1887. "The Calculus of Probabilities Applied to Psychical Research - II". *Proceedings of the Society of Psychical Research*, 4: 189-208.

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results*. Cambridge, Cambridge University Press.

Ellison, Aaron M. 1996. "An introduction to Bayesian inference for ecological research and environmental decision-making". *Ecological Applications*, 6: 1036-1046.

Erceg-Hurn, David y Vikki Mirosevic. 2008. "Modern robust statistical methods: An easy way to maximize the accuracy and power of your research". *American Psychologist*, 63(7): 591-601. http://www.unt.edu/rss/class/mike/5700/articles/robustAmerPsyc.pdf. Visitado en julio de 2011.

Erwin, Edward. 1998. "The logic of null hypothesis testing". *Behavioral and Brain Sciences*, 21(2): 197-198.

Evans, Merran, Nicholas Hastings y Brian Peacock. 1993. *Mathematical distributions.* 2ª edición, Nueva York, John Wiley & Sons.

Evans, Stephen, Peter Mills y Jane Dawson. 1988. "The end of the p value?". *British Heart Journal*, 60: 177-180. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1216550/pdf/brheartj00081-0001.pdf. Visitado en julio de 2011.

Eysenck, Hans J. 1960. “The concept of statistical significance and the controversy about one-tailed tests”. *Psychological Review*, 67: 269-271.

Falk, Ruma y Charles W. Greenbaum. 1995. “Significance tests die hard: The amazing persistence of a probabilistic misconception”. *Theory and Psychology*, 5: 75-98. http://tap.sagepub.com/content/5/1/75.short. Visitado en julio de 2011.

Farman, Joe C., Brian G. Gardiner y Jonathan D. Shanklin. 1985. “Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction”. *Nature*, 315: 207-10. http://www.ciesin.org/docs/011-430/011-430.html. Visitado en julio de 2011.

FDA Consumer. 2004. “Merck withdraws Vioxx”. *FDA Issues Public Health Advisory*, noviembre-diciembre, p. 38. http://www.highbeam.com/doc/1G1-124698350.html. Visitado en agosto de 2011.

Feferman, Solomon. 2006. “The impact of the incompletness theorems in mathematics”. *Notices of the AMS*, 53(4): 434-439.

Feinstein, Alvan R. 1998. “*P*-values and confidence intervals: Two sides of the same unsatisfactory coin”. *Journal of Clinical Epidemiology*, 51: 355-360.

Ferguson, George A. 1959. *Statistical analysis in psychology and education*. Nueva York, McGraw- Hill

Ferrer-i-Cancho, Ramón, Alexander Mehler, Olga Pustylnikov y Albert Díaz-Guilera. 2007. “Correlations in the organization of large-scale syntactic dependency networks”. En: *TextGraphs-2: Graph-based algorithms for natural language processing*. Rochester, Association for Computational Linguistics, pp. 65-72.

Fidler, Fiona, Mark Burgman, Geoff Coming, Robert Buttrose y Neil Thomason. 2006. “Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology”. *Conservation Biology*, 20(5): 1539-1544. http://www.faculty.biol.ttu.edu/Strauss/Stats/Readings/FidlerBurgmanCummingButtroseThomason2006.pdf. Visitado en julio de 2001.

Fischer, Hans. 2011. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Nueva York, Springer.

Fisher, Ronald Aylmer. 1922. “On the mathematical foundations of theoretical statistics”. *Philosophical Transactions of the Royal Astronomical Society of London*, A, 222: 309-368.

Fisher, Ronald Aylmer. 1925. *Statistical methods for the research workers*. Edimburgo, Oliver & Boyd. http://psychclassics.yorku.ca/Fisher/Methods/index.htm. Visitado en junio de 2011.

Fisher, Ronald Aylmer. 1929. “The statistical method in psychical research”. *Proceedings of the Society for Psychical Research*, 39: 189-192. http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15204/1/79.pdf. Visitado en julio de 2011.

Fisher, Ronald Aylmer. 1935. “The logic of inductive inference”. *Journal of the Royal Statistical Society*, 98: 39-54.

Fisher, Ronald Aylmer. 1955. “Statistical methods and scientific induction”. *Journal of the Royal Statistical Society*, B, 17: 69-78.

Fisher, Ronald Aylmer. 1956. *Statistical methods and scientific inference*. Edimburgo, Oliver & Boyd.

Fisher, Ronald Aylmer. 1970 [1925]. *Statistical methods for the research workers*. 14ª edición revisada, Nueva York, Hafner Publishing Company.

Fisher, Ronald Aylmer. 1971 [1935]. *The design of experiments*. 9ª edición, Nueva York, Hafner Press. http://www.4shared.com/file/0QcsfzFm/design_of_experiment_fisher.htm?aff=7637829 – Visitado en junio de 2011.

Forbes, Catherine, Merran Evans, Nicholas Hastings y Brian Peacock. 2011. *Statistical distributions*. 4ª edición, Hoboken, Wiley.

Foster, Stephen William. 1985. Revisión de *Local knowledge*, de C. Geertz. *American Anthropologist*, 87(1): 164-165.

Frías, M. D., J. Pascual y J. F. García. 2002. "La hipótesis nula y la significación práctica". *Metodología de las ciencias del comportamiento*, 4: 181-185. http://www.uv.es/garpe/C_/A_/C_A_0020.pdf. Visitado en julio de 2011.

Frink, Douglas. 1992. "The chemical variability of carbonized organic matter through time". *Archaeology of Eastern North America*, 20: 67-79

Gabor, George. 2004. "Classical statistics: Smoke and mirrors". http://www.stats.org.uk/statistical-inference/Gabor2004.pdf.

Galton, Francis. 1869. *Hereditary Genius. An inquiry into its laws and consequences*. Londres, Macmillan. http://www.galton.org/books/hereditary-genius/text/pdf/galton-1869-genius-v3.pdf. Visitado en julio de 2011.

Galton, Francis. 1875. "Statistics by intercomparison, with remarks on the law of frequency of error". *Philosophical Magazine*, 49: 33-46. http://galton.org/essays/1870-1879/galton-1875-intercomparison.pdf. Visitado en julio de 2011.

Galton, Francis. 1889. *Natural inheritance*. Londres, Macmillan. http://galton.org/books/natural-inheritance/pdf/galton-nat-inh-1up-clean.pdf. Visitado en julio de 2011.

García-Berthou, Emili y Carles Alcaraz. 2004. "Incongruence between test statistics and P values in medical papers". *BMC Medical Research Methodology*, 4:13, http://www.biomedcentral.com/content/pdf/1471-2288-4-13.pdf.

Gardner, Martin y Douglas Altman. "Confidence intervals rather than P values: estimation rather than hypothesis testing". *British Medical Journal*, 22: 746-750.

Gavarret, Jules. 1840. *Principes Généraux de Statistique Médicale, ou Dévelopement des règles qui doivent présider a son emploi*. Paris, Bechet Jeune et Labé. http://carlosreynoso.com.ar/archivos/varios/Gavarret-Statistique-medicale.pdf. Visitado en agosto de 2011.

Geary, R. C. 1947. "Testing for normality". *Biometrika*, 34: 209-242.

Geertz, Clifford. 1973. *The interpretation of cultures.* Nueva York, Basic Books [Traducción castellana editada por Carlos Reynoso, *La interpretación de las culturas.* Barcelona, Gedisa, 1987].

Geertz, Clifford. 1983. *Local knowledge: Further essays in interpretive anthropology*. Nueva York, Basic books.

Geertz, Clifford. 1987. "The anthropologist at large". Reseña de Mary Douglas, *How Institutions Think. The New Republic*, 25 de mayo, pp. 34 y 36-37.

Geertz, Clifford. 2000. *Available light: Anthropological reflections on philosophical topics*. Princeton, Princeton University Press.

Giaquinto, Marcus. 2007. *Visual thinking in mathematics: An epistemological study*. Oxford, Oxford University Press.

Giere, Ronald. 1972. “Review: The significance test controversy”. *The British Journal for the Philosophy of Science*, 23(2): 170-181. http://www.jstor.org/pss/686441. Visitado en julio de 2011.

Gigerenzer, Gerd. 1987. “Probabilistic thinking and the fight against subjectivity”. En: L. Krüger, G. Gigerenzer y M. S. Morgan (editores), *The probabilistic revolution. Vol. 2. Ideas in the sciences*. Cambridge (USA), MIT Press, pp. 11-33.

Gigerenzer, Gerd. 1993. “The superego, the ego, and the id in statistical reasoning”. En: G. Keren y C. Lewis (editores), *A handbook of data analysis in the behavioral sciences: Methodological issues*. Hillsdale, Erlbaum, pp. 311-339.

Gigerenzer, Gerd. 1998a. “Surrogate for theories”. *Theory & Psychology*, 8(2): 195-204.

Gigerenzer, Gerd. 1998b. “We need statistical thinking, not statistical rituals”. *Behavioral and brain sciences*, 21(2): 199-200. http://library.mpib-berlin.mpg.de/ft/gg/GG_We%20need_1998.pdf. Visitado en julio de 2011.

Gigerenzer, Gerd. 2000. *Adaptive thinking – Rationality in the real world*. Nueva York, Oxford University Press.

Gigerenzer, Gerd. 2004. “Mindless statistics”. *The Journal of Socio-Economics*, 33: 587-606. http://www-unix.oit.umass.edu/~bioep740/yr2009/topics/Gigerenzer-jSoc-Econ-1994.pdf. Visitado en julio de 2011.

Gigerenzer, Gerd, Stefan Krauss y Oliver Vitouch. 2004. “The null ritual: What you always wanted to know about significance testing but were afraid to ask”. En: D. Kaplan (editor), *The SAGE handbook of quantitative methodology for the social sciences*, Londres, SAGE, pp. 391-408. http://library.mpib-berlin.mpg.de/ft/gg/GG_Null_2004.pdf. Visitado en julio de 2011.

Gigerenzer, Gerd y D. J. Murray. 1987. *Cognition as intuitive statistics*. Hillsdale, Erlbaum.

Gill, Jeff. 1999. “The insignificance of null hypothesis significance testing”. *Political Research Quarterly*, 52(3): 647-674, http://polmeth.wustl.edu/media/Paper/gill99.pdf.

Gill, Jeff. S/f. “The Current State of the Null Hypothesis Significance Test”. Texto disponible sin referencia de origen.

Gill, Jeff. S/f. “How do we do hypothesis testing?”. http://artsci.wustl.edu/~jgill/papers/hypos.pdf.

Gilmore, J. B. 1989. “Randomness and the search for psi”. *Journal of Parapsychology*, 53: 309-340.

Gilmore, J. B. 1990. “Anomalous significance in pararandom and psi-free domains”. *Journal of Parapsychology*, 54: 53-58.

Given, Lisa M. (editora). 2008. *The SAGE Encyclopedia of Qualitative Research Methods*. Los Angeles, Sage.

Glaser, Dale N. 1999. “The controversy of significance testing: Misconceptions and alternatives”. *American Journal of Critical care*, 8(5): 291-296. http://www.glaserconsult.com/docs/Peer_Reviewed_Articles/controversy_of_signicance_testing.pdf. Visitado en julio de 2011.

Gliner, Jeffrey, Nancy Leech y George Morgan. 2002. “Problems with null hypothesis significance testing (NHST): What do the textbooks say”. *The Journal of Experimental Education*, 71(1): 83-92. http://www.andrews.edu/~rbailey/Chapter%20two/7217331.pdf. Visitado en julio de 2011.

Gödel, Kurt. 1930. “Die Vollständigkeit der Axiome des logischen Funktionen-kalküls", *Monatshefte für Mathematik und Physik* 37: 349-360. [Trad. Esp.: “La suficiencia de los

axiomas del cálculo lógico de primer orden". En: *Obras completas*. Madrid, Alianza, 1981: 20-34).

Gold, David. 1969. "Statistical tests and substantive significance". *American Sociologist*, 4: 42-46. http://www.jstor.org/pss/27701454. Visitado en julio de 2011.

Good, I. J.. 1958. "Significance tests in parallel and in series". *Journal of the American Statistical Association*, 53: 799-813.

Goodman, Steven. 1999. "Toward evidence-based medical statistics: 1. The *P* value fallacy". *Annals of Internal Medicine*, 130(12): 995-1004.

Goodman, Steven. 2008. "A dirty dozen: Twelve *P*-values misconceptions". *Seminars in Hematology*, 45: 135-140.

Goodman, Steven, Douglas G. Altman y Steven George. 1998. "Statistical reviewing policies of medical journals: Caveat lector?". *Journal of General Internal Medicine*, 13(11): 753-756.

Gore, S. M., I. G. Jones y E. C. Rytter. 1976. "Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976". *BMJ*, 1: 85-7.

Gosset, William Sealy ["Student"]. 1908. "The probable error of a mean". *Biometrika*, 6: 1-25. http://www.york.ac.uk/depts/maths/histstat/student.pdf. Visitado en julio de 2011.

Gosset, William Sealy ["Student"]. 1942. *'Student's' collected papers*. En: E. Pearson y J. Wishart. Cambridge, Cambridge University Press.

Greenwald, Anthony. 1975. "Consequences of prejudice against the null hypothesis". *Psychological Bulletin*, 82(1): 1-20. http://faculty.washington.edu/agg/pdf/Gwald_PsychBull_1975.OCR.pdf. Visitado en julio de 2011.

Gregoire, Timothy. 2001. "Biometry in the 21st century: Whither statistical inference?". Conference on Forest Biometry and Information Science, 25 al 29 de julio, http://cms1.gre.ac.uk/conferences/iufro/proceedings/. Visitado en julio de 2011.

Guilford, Joy Paul. 1942. *Fundamental Statistics in Psychology and Education*, Nueva York, McGraw-Hill.

Guthery, Fred. 2008. "Statistical ritual versus knowledge accrual in wildlife science". *Journal of Wildlife management*, 72(8): 1872-1875. http://www.auburn.edu/~tds0009/Articles/Guthery%202008.pdf. Visitado en julio de 2011.

Guthery, Fred, Jeffrey Lusk y Markus Peterson, 2001. "The fall of the null hypothesis: Liabilities and opportunities". *The journal of wildlife management*, 65:3: 379-384.

Guttman, Louis. 1977. "What is not what in statistics". *The Statistician*, 26: 81-107.

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

Hacking, Ian. 1988. "Telepathy: Origins of Randomization in Experimental Design". *Isis*, 79(3): 427-451. http://www.jstor.org/pss/234674. Visitado en julio de 2011.

Hacking, Ian. 1991. "How Shall We Do the History of Statistics?". En: Graham Burchell, Colin Gordon y Peter Miller (editores), *The Foucault Effect: Studies in Governmentality*. Chicago, University of Chicago Press, pp. 181-195.

Hagen, Richard L. 1997. "In praise of the null hypothesis statistical test". *American Psychologist*, 52: 15-24.

Hagen, Richard L. 1998. "A further look at wrong reasons to abandon statistical testing". *American Psychologist*, 53: 801-803.

http://kochanski.org/gpk/misc/papers_that_shouldnt_be_lost/1998/Tryon_1998_The_Inscrutable_Null_Hypothesis.pdf. Visitado en julio de 2011.

Hager, Willi. 2000. "About some misconceptions and the discontent with statistical tests in psychology". *Methods of Psychological Research Online*, 5(1), http://www.dgps.de/fachgruppen/methoden/mpr-online/issue9/art1/hager.pdf. .

Hald, Alders. 1998. *A History of Mathematical Statistics from 1750 to 1930*. Nueva York, Wiley

Haller, Heiko y Stephan Krauss. 2002. "Misinterpretations of significance: a problem students share with their teachers?". *Methods of Psychological Research—Online* [On-line serial], 7, 1–20. http://www.dgps.de/fachgruppen/methoden/mpr-online/issue16/art1/haller.pdf. Visitado en julio de 2010.

Hanneman, Robert. 2005. *Introduction to social network methods*. Publicación en línea, University of California at Riverside. http://faculty.ucr.edu/~hanneman/. Visitado en julio de 2010.

Harlow, Lisa, Stanley Mulaik y James Steiger (compiladores). 1997. *What if there were no significance tests?*. Mahwah, Erlbaum.

Harville, David A. 1975. "Experimental randomization: Who needs it?". *American statistician*, 29: 27-31.

Hayes, Andrew F. 1998. "Reconnecting research design and data analysis: Who needs a confidence interval?". *Behavioral and Brain Sciences*, 21: 203-204.

Hebb, Donald Olding. 1966. *A handbook of Psychology*. Filadelfia, Saunders.

Heidelberg, Kurt R. 2001. "Feasibility Study: Predictive Model for the Management and Interpretation of Cultural Resources, Yuma Proving Ground, Arizona". *Technical Report*, 01-38, Statistical Research, Tucson.

Henry, Edward. 1976. "The Variety of Music in a North Indian Village: Reassessing Cantometrics". *Ethnomusicology*, 20(1): 49-66.

Herrnstein, Richard y Charles Murray. 1994. *The bell curve: Intelligence and class structure in American life*. Nueva York, Free Press.

Hill, M., y W. J. Dixon. 1982. "Robustness in real life: A study of clinical laboratory data". *Biometrics*, 38: 377-396.

Hirschfeld, Lawrence A., James Howe y Bruce Levin. 1978. "Warfare, Infanticide, and Statistical Inference: A Comment on Divale and Harris". *American Anthropologist*, 80(1): 110-115.

Hodson, Frank Roy. 1977. Revisión de *Figuring anthropology* de David Hurst Thomas. *American Antiquity*, 42(2): 299-300.

Hogben, Lancelot. 1957. *Statistical theory*. Londres, Allen & Unwin.

Hole, Bonnie Laird. 1980. "Sampling in archaeology: A critique". *Annual Review of Anthropology*, 9: 217-234.

Hoover, Kevin y Mark Siegler. 2008. "Sound and Fury: McCloskey and Significance Testing in Economics". *Journal of Economic Methodology*. http://econ.duke.edu/~kdh9/Source%20Materials/McCloskey/Sound%20and%20Fury%206%20March%202007.pdf. Visitado en julio de 2011.

Hopkins, Kenneth D. y Gene V. Glass. 1978. *Basic statistics for the behavioral sciences*. Englewood Cliffs, Prentice-Hall.

Householder, Fred. 1952. Revisión crítica de *Methods in structural linguistics* de Zelig Harris. *International Journal of American Linguistics*, 18: 260-268.

Hubbard, Raymond. 2005. “Why we don’t really know what “statistical significance” means: A major educational failure”. http://escholarshare.drake.edu/bitstream/handle/2092/413/WhyWeDon't.pdf?sequence=3. Visitado en junio de 2011.

Hubbard, Raymond y J. Scott Armstrong. 1992. “Are null results becoming an endangered species in marketing?”. http://fourps.wharton.upenn.edu/ideas/pdf/nullresults.pdf. Visitado en junio de 2011.

Hubbard, Raymond y M. J. Bayarri. 2003. “P values are not error probabilities”. http://www.uv.es/sestio/TechRep/tr14-03.pdf - Visitado en junio de 2011.

Hubbard, Raymond y M. J. Bayarri. 2003. “Confusion over measures of Evidence (*p*’s) versus Error (α) in classical statistical testing”. *The American Statician*, 57(3): 171-182. http://www.pucrs.br/famat/viali/cursos/especializacao/ceea/testes/Textos/Valor-pXAlfa.pdf. Visitado en julio de 2011.

Huber, Peter y Elvezio Ronchetti. 2009. *Robust statistics*. 2ª edición. Hoboken, John Wiley and Sons.

Huberty, Carl J.. 1993. “Historical Origins of Statistical Testing Practices: the Treatment of Fisher Versus Neyman-Pearson views in textbooks”. *Journal of Experimental Education*, 61(4), 317-333.

Huff, Darrell. 1954. *How to lie with statistics*. Nueva York, W. W. Norton & Co.

Hunter, John. 1997. “Needed: A ban on the significance test”. *Psychological science*, 8: 3-7.

Huysamen, G. K. 2005. “Null hypothesis significance testing: Ramifications, ruminations and recommendations”. *South African Journal of Psychology*, 35(1): 1-20.

Hwang, Y. T., S. Larivière y F. Messier. 2005. “Evaluating body condition of striped skunks using non-invasive morphometric indices and bioelectrical impedance”. *Wildlife Society Bulletin*, 33: 195-203.

Iacobucci, Dawn. 2005. “On *p*-values”. *Journal of Consumer Research*, 32(1): 6-11.

Iversen, Gudmund R. 1984. *Bayesian statistical inference*. Beverly Hills, Sage.

Jahn, Detlef. 2006. “Globalization as Galton’s problem: The missing link in the analysis of missing patterns in welfare state development”. *International Organization*, 60: 401-431. http://intersci.ss.uci.edu/wiki/pw/DetlefJahn2006.pdf. Visitado en julio de 2011.

Jeffreys, Sir Harold. 1961 [1939]. *Theory of probability*. 3ª edición, Oxford, Oxford University Press.

Jeffreys, Sir Harold. 1963. Revisión de L. J. Savage y otros, *The Foundations of Statistical Inference*. *Technometrics*, 5: 407-410.

Johansson, Tobias. 2011. “Hail the impossible: *p*-values, evidence, and likelihood”. *Scandinavian Journal of Psychology*, 52: 113-125

Johnson, Douglas H. 1995. “Statistical sirens: The allure of non- parametrics”. *Ecology*, 76(6): 1998-2000.

Johnson, Douglas H. 1999. “The insignificance of statistical significance testing”. Journal of Wildlife Management, 63(3): 763-772. http://faculty.washington.edu/skalski/classes/QERM597/papers/Johnson.pdf. Visitado en julio de 2011.

Johnson, Douglas H. 2004. “What hypothesis tests are not: a response to Colegrave and Ruxton”. *Behavioral Ecology*, 16(1): 323-324. http://digitalcommons.unl.edu/usgsnpwrc/35/. Visitado en julio de 2011.

Johnson, Norman, Adrienne Kemp y Samuel Kotz. 2005. *Univariate discrete distributions*. Hoboken, John Wiley & Sons.

Johnson, Norman, Samuel Kotz y N. Balakrishnan. 1994. *Continuous univariate distributions. Vol 1*. 2ª edición, Hoboken, John Wiley & Sons.

Johnson, Oliver. 2004. *Information theory and the central limit theorem*. Singapur, World Scientific.

Jondeau, Eric, Ser-Huang Poon y Michael Rockinger. 2007. *Finantial modeling under non-Gaussian distributions*. Londres, Springer London.

Jones, Lyle V. 1950. “Statistics and research design”. *Annual Review of Psychology*, 6: 405-430.

Kagan, Abram, Yurii Vladimirovich Linnik y C. Radhakrishna Rao. 1973. *Characterization problems in mathematical statistics*. Nueva York, Wiley.

Kaipio, Jari y Erkki Somersalo. 2006. *Statistical and computational inverse problems*. Nueva York, Springer.

Kamminga, J. y R. V. S. Wright. 1988. “The upper cave at Zhoukoudian and the origins of the mongoloids”. *Journal of Human Evolution*, 17: 739-767.

Keesing, Roger M. 1985. Revisión de *Local knowledge*, de C. Geertz. *American Ethnologist*, 12(3): 554-555.

Kendall, M. G. y A. Stuart. 1951. *Advanced Theory of Statistics*. 3ª edición, Londres, Griffin.

Kerlinger, Frederick Nicholas. 1979. *Behavioral research: A conceptual approach*. Nueva York, Holt, Rinehart and Winston.

Kilgarriff, Adam. 2005. “Language is never, ever, ever, random”. *Corpus Linguistics and Linguistic Theory*, 1-2: 263-275. http://www.kilgarriff.co.uk/Publications/2005-K-lineer.pdf. Visitado en julio de 2011.

Kimball, A. W., 1957. “Errors of the Third Kind in statistical consulting”. *Journal of the American Statistical Association*, 52(278): 133–142.

Kish, Leslie. 1959. “Some statistical problems in research design”. *American Sociological Review*, 24: 328-338. http://www.epidemiology.ch/history/PDF%20bg/Kish%20L%201959%20some%20statistical%20problems.pdf. Visitado en julio de 2011.

Kleiber, Christian y Samuel Kotz. 2003. *Statistical size distribution in economics and actuarial sciences*. Hoboken, Wiley Interscience.

Kline, Rex. 2004. *Beyond Significance Testing: reforming data analysis methods in behavioral research*. Washington, American Psychological Association.

Kmetz, John. 2010. *Management Junk Science*. http://sites.udel.edu/mjs/statistical-significance-references/. Visitado en agosto de 2011.

Kmetz, John. 2011. “Fifty lost years: Why international business scholars must not emulate the Us Social-Science research model”. 14th Annual Business Research Conference, 28 de abril, Dubai. http://www.wbiconpro.com/422%20KMETZ.pdf. Visitado en agosto de 2011.

Korotayev, Andrey y Victor de Munck. 2003. “Galton’s asset and Flower’s problem: Cultural networks and cultural units in cross-cultural research”. *American Anthropologist*, 105: 353-358.

Kotz, Samuel y Saralees Nadarajah. 2000. *Extreme value distributions: Theory and applications*. Londres, Imperial College Press.

Krämer, Walter y Gerd Gigerenzer. 2005. "How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities". *Statistical Science*, 20(3): 223-230. http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdfview_1&handle=euclid.ss/1124891288. Visitado en julio de 2011.

Krantz, David. 1999. "The null hypothesis testing controversy in psychology". *Journal of the American Statistical Association*, 44(448): 1372-1381. http://www.unt.edu/rss/class/mike/5030/articles/krantznhst.pdf. Visitado en julio de 2011.

Krishnaiah, Paruchuri R. y Charumpuri R. Rao (editores). 1988. *Handbook of statistics. Vol 6: Sampling*. Amsterdam, North Holland.

Krishnamoorty, Kalimuthu. 2006. *Handbook of statistical distributions with applications*. Boca Raton, Chapman & Hall / CRC.

Krueger, Joachim. 2001. "Null hypothesis significance testing: On the survival of a flawed method". *American Psychologist*, 56(1): 16-26. http://www.psych.uncc.edu/pagoolka/krueger2001.pdf. Visitado en julio de 2011.

Kruskal, William. 1968a. "Tests of statistical significance". En: David Sills (editor), *International Encyclopedia of the Social Sciences*, vol. 14. Nueva York, MacMillan, pp. 238-250.

Kruskal, William. 1968b. "Statistics: The field". En: David Sills (editor), *International Encyclopedia of the Social Sciences*, vol. 14. Nueva York, MacMillan, pp. 206-224.

Kruskal, William. 1978. "Formulas, numbers, words: statistics in prose". *The American Scholar*, 47(2): 223-229.

Kruskal, William. 1980. "The significance of Fisher: A review of *R. A. Fisher: The life of a scientist*", de Joan Fisher Box. *Journal of the American Statistical Association*, 75(372): 1019-1030.

Kruskal, William y Frederick Mosteller. 1979a. "Representative sampling. I. Non-scientific literature". *International Statistical Review*, 47: 13–24.

Kruskal, William y Frederick Mosteller. 1979b. "Representative sampling. II. Scientific literature, excluding statistics". *International Statistical Review*, 47: 111–127.

Kruskal, William y Frederick Mosteller. 1979c. "Representative sampling. III. The current statistical literature". *International Statistical Review*, 47: 245–265.

Kruskal, William y Frederick Mosteller. 1980. "Representative sampling. IV. The history of the concept in statistics, 1895–1939". *International Statistical Review*, 48: 169–195.

Kruskal, William y S. M. Stigler. 1997. "Normative terminology: 'Normal' in statistics and elsewhere". En: Bruce Spencer (editor), *Statistics and Public Policy*. Oxford, Oxford University Press, pp. 85-111.

Kyburg, Henry, Jr. 1971. *The logical foundations of statistical inference*. Dordrecht, Reidel.

Labovitz, Sanford. 1968. "Criteria for selecting a significance level: On the sacredness of .05". *The American Sociologist*, 3(3): 220-222. http://www.jstor.org/pss/27701367. Visitado en julio de 2011.

Labovitz, Sanford. 1970. "The Nonutility of Significance Tests: The Significance of Tests of Significance Reconsidered". *The Pacific Sociological Review*, 13(3): 141-148. http://www.jstor.org/pss/1388411. Visitado en julio de 2011.

Lakatos, Imre. 1978. “Falsification and the methodology of scientific research programmes”. En J. Worral y G. Curie (editores), *The methodology of scientific research programs: lmre Lakatos' philosophical papers (Vol. 1).* Cambridge, Cambridge University Press.

Laplace, Pierre Simon de. 1773 [pub. 1776]. “Mémoire sur l’inclinaison moyenne des orbites des comètes, sur la figure de la Terre et sur les fonctions”. *Memoire de l’Academie Royale des Sciences*, París, 7: 503-524. http://gallica.bnf.fr/ark:/12148/bpt6k77596b/f284. Visitado en agosto de 2011.

Leahy, Erin. 2005. “Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology”. *Social Forces*, 84(1): 1-24. http://www.soc.washington.edu/users/bpettit/soc504/leahey_significance.pdf. Visitado en julio de 2011.

Le Cam, Lucien. 1986. “The central limit theorem around 1935”. *Statistical Science*, 1(1): 78-91.

Lecoutre, Bruno. 1999. “Beyond the significance test controversy: Prime time for Bayes?”. *International Statistical Institute, 52nd Session*. http://www.stat.auckland.ac.nz/~iase/publications/5/leco0735.pdf.

Lecoutre, Marie-Paul, Bruno Lecoutre y Jacques Poitevineau. 2001. “Uses, Abuses and Misuses of Significance Tests in the Scientific Community: Won't the Bayesian Choice Be Unavoidable?”. *International Statistical Review / Revue Internationale de Statistique*, 69(3): 399-417. http://www.tc.umn.edu/~alonso/Lecoutre_ISR_2001.pdf. Visitado en julio de 2011.

Lecoutre, Marie-Paul, Jacques Poitevineau y Bruno Lecoutre. 2003. “Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests”. *International Journal of Psychology*, 38: 37-45. http://www.unt.edu/rss/class/mike/5030/articles/evenstatsguys.pdf. Visitado en julio de 2011.

Lee, J. Jack. 2011. “Demystify statistical significance – Time to move on from the *P* value to Bayesian analysis”. *Journal of the National Cancer Institute*, 103: 1. http://jnci.oxfordjournals.org/content/103/1/2.full.pdf. Visitado en julio de 2011.

Lehmann, Erich Leo. 1993. “The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?”. *Journal of the American Statistical Association*, 88(424): 1242-1249. http://www.phil.vt.edu/dmayo/PhilStatistics/Other/Lehmann%201993%20%20Fisher%20and%20NP%20theories%20of%20testing%20%20hypotheses%20one%20theory%20or%20two.pdf. Visitado en julio de 2011.

Lehmann, Erich Leo. 1995. “Neyman’s statistical philosophy”. *Probability and Mathematical Statistics*, 15: 29-36. http://www.math.uni.wroc.pl/~pms/files/15/Article/15.4.pdf. Visitado en julio de 2011.

Lehmann, Erich Leo y Joseph P. Romano. 2005. *Testing statistical hypotheses*. 3ª edición, Nueva York, Springer.

Levin, Joel R., 1998. “What if there were no more bickering about statistical significance tests”. *Research in the Schools*, 5(2): 43-53. http://www.personal.psu.edu/users/d/m/dmr/sigtest/6mspdf.pdf. Visitado en julio de 2011.

Levine, David y David Stephan. 2010. *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. 2ª edición, Upper Saddle River, Pearson Education. http://www.ftpress.com/store/product.aspx?isbn=9780137010592. Visitado en julio de 2011.

Lévi-Strauss, Claude. 1995 [1955]. “La estructura de los mitos”. En: *Antropología estructural*. 2ª reimpresión, Barcelona, Paidós.

Lexis, Wilhelm Hector Richard Albrecht. 1875. *Einleitung in die Theorie der Bevölkerungsstatistik*, Strassburg. http://dz-srv1.sub.uni-goettingen.de/sub/digbib/loader?did=D307894. Visitado en agosto de 2011.

Lexis, Wilhelm Hector Richard Albrecht. 1877. *Zur Theorie der Massenerscheinungen in der Menschlichen Gesellschaft.* Freiburg, Fr. Wagner'sche Buchhandlung. http://dspace.utlib.ee/dspace/bitstream/handle/10062/3542/lexistheorieocr.pdf;jsessionid=21A0C856C0AA13A3E8C879784E8BDC96?sequence=7. Visitado en agosto de 2011.

Lieberson, Jonathan. 1984. "Interpreting the interpreter". *New York Review of books*, 31: 39-46.

Liese, Friedrich y Klaus-J. Miescke. 2008. *Statistical decision theory: Estimation, testing, and selection*. Nueva York, Springer.

Lindley, D. V. 1958. "Professor Hogben's 'Crisis': A survey of the foundations of statistics". *Applied Statistics*, 7: 186-198.

Lindquist, Everett Franklin. 1940. *Statistical analysis in educational research*. Boston, Houghton Mifflin.

Lindsey, James K. 1995. *Modelling frequency and count data*. Oxford, Clarendon Press.

Lisse, Jeffrey R. y otros. 2003. "Gastrointestinal Tolerability and Effectiveness of Rofecoxib versus Naproxen in the Treatment of Osteoarthritis". *139 Annals Internal Med*., 539: 543-544.

Loftus, Geoffrey. 1991. "On the Tyranny of Hypothesis Testing in the Social Sciences". *Contemporary Pschology*, 36(2): 102-105. http://faculty.washington.edu/gloftus/Downloads/CPChance.pdf. Visitado en julio de 2011.

Loftus, Geoffrey. 1993. "A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age". *Behavior Research Methods, Instruments & Computers*, 25(2): 250-256. http://faculty.washington.edu/gloftus/Downloads/ThousandpValues.pdf. Visitado en julio de 2011.

Loftus, Geoffrey. 1996. "Psychology will be a much better science when we change the way we analyze data". *Current Directions in Psychological Science*, 1996: 161-171. http://faculty.washington.edu/gloftus/Downloads/CurrentDirections.pdf. Visitado en julio de 2011.

Loftus, Geoffrey. 2010. "Null hypothesis". http://faculty.washington.edu/gloftus/Downloads/Loftus.NullHypothesis.2010.pdf.

Louçã, Francisco. 2008. "The widest cleft in statistics – How and why Fisher opposed Neyman and Pearson". http://www.iseg.utl.pt/departamentos/economia/wp/wp022008deuece.pdf. Visitado en julio de 2011.

Lykken, David T. 1968. "Statistical significance in psychological research". *Psychological Bulletin, 70*, 151-159. http://www.psych.umn.edu/courses/spring05/mcguem/psy8935/readings/lykken1968.pdf. Visitado en julio de 2011.

Maltz, Michael. 1994. "Deviating from the mean: The declining significance of significance". *Journal of Research in Crime and Delinquency*, 31(4): 434-463. http://tigger.uic.edu/~mikem/Deviating.PDF. Visitado en julio de 2011.

Mandelbrot, Benoît y Richard L. Hudson. 2006. *Fractales y finanzas: Una aproximación matemática a los mercados*. Barcelona, Tusquets.

Marascuilo, L. A. y J. R. Levin. 1970. "Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type-IV errors". *American Educational Research Journal*, 7(3): 397–421.

Marewski, Julian y Henrik Olsson. 2009. “Beyond the null ritual: Formal modeling of psychological processes”. *Journal of Psychology*, 217(1): 49-60. http://library.mpib-berlin.mpg.de/ft/jm/JM_Beyond_2009.pdf. Visitado en julio de 2011.

Maronna, Ricardo, Douglas Martin y Victor Yohai. 2006. *Robust statistics: Theory and methods*. Chichester, Wiley.

Marx, Karl. 1909. *Capital. A critique of political economy. Vol. I: The process of capitalist production*. Traducción de Samuel Moore y Edward Aveling. Revisado y ampliado a partir de la 4ª edición alemana por Ernest Untermann. Chicago, Charles H. Kerr & Company.

Marx, Wolfgang. 2006. “Das Null-Ritual und einer seiner Bewunderer”. *Psychologische Rundschau*, 57(4): 256-258. http://www.fachportal-paedagogik.de/fis_bildung/suche/fis_set.html?FId=774688. Visitado en julio de 2011.

Mayo, Deborah. 1980. “The philosophical relevance of statistics”: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol. 1: 97-109.

Mayo, Deborah. 1992. “Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?”. *Synthese*, 90 (2): 233-262. http://www.phil.vt.edu/dmayo/personal_website/%281992%29%20DID%20PEARSON%20REJECT%20THE%20NEYMAN-PEARSON%20Philosophy%20of%20Statistics.pdf. Visitado en julio de 2011.

Mayo, Deborah. 1996. *Error and the growth of experimental knowledge*. Chicago, University of Chicago Press.

Mayo, Deborah y Aris Spano. 2006. “Severe testing as a basic concept in a Neyman-Pearson philosophy of induction”. *British Journal for the Philosophy of Science*, 57: 323-357. http://www.phil.vt.edu/dmayo/conference_2010/Mayo%20Spanos%20Severe%20Testing%20as%20a%20Basic%20Concept%20in%20NP%20Theory%20of%20Induction.pdf. Visitado en julio de 2011.

McCloskey, Deirdre N. 1998. “Other things equal: Quarreling with Ken”. *Eastern Economic Journal*, 24(1): 111-115.

McCloskey, Deirdre N. y Stephen T. Ziliak. 2007. “Signifying nothing: Reply to Hoover and Siegler”. *Journal of Economic Methodology*. http://www.deirdremccloskey.org/docs/hoover.pdf. Visitado en julio de 2011.

McCloskey, Deirdre N. y Stephen T. Ziliak. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (Economics, Cognition, and Society)*. Ann Arbor, The University of Michigan Press. http://press.umich.edu/titleDetailPraise.do?id=186351. Visitado en julio de 2011.

McCloskey, Deirdre N. y Stephen T. Ziliak. 2009. “The Unreasonable Ineffectiveness of Fisherian ‘Tests’ in Biology, and Especially in Medicine”. *Biological Theory*, 4(1): 44-53. http://www.deirdremccloskey.org/docs/fisherian.pdf. Visitado en julio de 2011.

McCloskey, Deirdre N. y Stephen T. Ziliak. 2010. Brief of amici curiae statistics experts professors Deirdre N. McCloskey and Stephen T. Ziliak in support of respondents. http://www.americanbar.org/content/dam/aba/publishing/preview/publiced_preview_briefs_pdfs_09_10_09_1156_RespondentAmCu2Profs.authcheckdam.pdf. Visitado en julio de 2011.

McCloskey, Donald [Deirdre]. S/f. “Rhetoric within the citadel: Statistics”. http://www.deirdremccloskey.com/docs/pdf/Article_181.pdf. Visitado en julio de 2011.

McCloskey, Donald [Deirdre]. 1985. “The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests”. *The American Economic Review*, 75(2), Papers and Proceedings of the Ninety-Seventh Annual Meeting of the American Economic Association, pp. 201-205.

McDonald, J. H. 2009. *Handbook of Biological Statistics*. 2ª edición, Baltimore, Sparky House Publishing.

McEwen, William. 1963. “Forms and problems of validation in social anthropology”. *Current Anthropology*, 4(2): 155-183.

McGrath, Robert. 1998. "Significance testing: Is there something better?". *American Psychologist*, julio. http://kochanski.org/gpk/misc/papers_that_shouldnt_be_lost/1998/Tryon_1998_The_Inscrutable_Null_Hypothesis.pdf. Visitado en julio de 2011.

McMan, J. C. 1995. “Statistical significance testing fantasies in introductory psychology textbooks”. Ponencia presentada en la 103rd Annual Convention of the American Psychological Association, Nueva York.

Meehl, Paul E. 1967. “Theory testing in psychology and physics: A methodological paradox”. *Philosophy of science*, 34: 103-115.

Meehl, Paul 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. http://www.haverford.edu/psych/ddavis/psych212h/meehl.1978.html. Visitado en julio de 2011.

Meehl, Paul. 1990a. “Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it”. *Psychological Inquiry*, 1: 108-141. http://www.tc.umn.edu/~pemeehl/147AppraisingAmending.pdf. Visitado en julio de 2011.

Meehl, P. E. 1990b. “Why summaries of research on psychological theories are often uninterpretable”. *Psychological Reports*, 66: 195-244. http://www.tc.umn.edu/~pemeehl/144WhySummaries.pdf. Visitado en julio de 2011.

Meehl, Paul E. 1999. “The problem is epistemology, not statistics: Replace statistical tests by confidence intervals and quantify accuracy of risky numerical predictions”. En: L. L. Harlow, S. A. Mulaik y J. H. Steiger (editores), *What if there were no significance tests?*, pp. 393-425.

Melton, Arthur W. 1962. “Editorial”. *Journal of Experimental Psychology*, 64: 553-557.

Menon, Rama. 1993. “Statistical significance testing should be discontinued in mathematics education research”. *Mathematics education research journal*, 5(1): 4-18. http://www.merga.net.au/documents/MERJ_5_1_Menon_1.pdf. Visitado en julio de 2011.

Micceri, Theodore. 1989. “The unicorn, the normal curve, and other improbable creatures”. *Psychological Bulletin*, 105: 156-166. http://www.unt.edu/rss/class/mike/5030/articles/micceri89.pdf. Visitado en julio de 2011.

Michell, John. 1767. "An Inquiry into the Probable Parallax and Magnitude of the Fixed Stars, Stars, from the Quantity of Light Which They Afford us, and the Particular Circumstances of Their Situation". *Philosophical Transactions*, 57: 234-264. http://www.philoscience.unibe.ch/documents/TexteHS09/Mitchell1767.pdf. Visitado en agosto de 2011.

Miller, George Armitage y Robert Buckhout. 1973. *Psychology: The Science of Mental Life*. 2ª edición, Nueva York, Harper and Row.

Mitroff, Iain y Abraham Silvers. 2009. *dirty rotten strategies: How we trick ourselves and others into solving the wrong problems precisely*. Stanford, Stanford Business Press.

Monterde-i-Bort, Héctor, Dolores Frías-Navarro y Juan Pascual-Llovell. 2006. "Errores de interpretación de los métodos estadísticos: importancia y recomendaciones". *Psicothema*, 18(4): 848-856. http://redalyc.uaemex.mx/pdf/727/72718426.pdf. Visitado en julio de 2011.

Monterde-i-Bort, Héctor, Dolores Frías-Navarro y Juan Pascual-Llovell. 2010. "Uses and abuses of statistical significance tests and other statistical resources: a comparative study". *European Journal of Education and Psychology*, 25: 429-447.

Moran, John L. y Patricia J. Solomon. 2004. "A farewell to P-values?". *Critical Care and Resuscitation*, 6: 130-137

Morris, K. M. y T. J. Maret. 2007. "Effects of timber management on pond-breeding salamanders". *Journal of Wildlife Management*, 71: 1034–1041.

Morrison, Denton F. y Ramon E. Henkel. 1969. "Significance test reconsidered". *American Sociologist*, 4: 131-140. http://www.jstor.org/pss/27701482. Visitado en julio de 2011.

Morrison, Denton F. y Ramon E. Henkel. 1970. *The significance test controversy: A reader*. Chicago, Aldine.

Mosteller, F. 1948. "A *k*-Sample slippage test for an extreme population". *The Annals of Mathematical Statistics*, 19(1): 58–65.

Mulaik, Stanley, Nambury Raju y Richard Harshman. 1997. "There is a time and a place for significance testing". En: L. L. Harlow, S. A. Mulaik y J. H. Steiger (editores), *Op. Cit*., pp. 65-115.

Naroll, Raoul. 1961. "Two solutions to Galton's problem". *Philosophy of Science*, 28: 15-29, http://dx.doi.org/10.1086%2F287778. Visitado en juniuo de 2011.

Naroll, Raoul. 1965. "Galton's problem: The logic of cross-cultural research". *Social Research*, 28: 428-451.

Naroll, Raoul. 1970. "What have we learned from cross-cultural surveys?". *American Anthropologist*, 72(6): 1227-1288.

Naroll, Raoul. 1971. Revisión crítica de Denton Morrison y Ramon Henkel (editores), *The significance test controversy*. *American Anthropologist*, 73(6): 1437-1439.

Naroll, Raoul y Ronald Cohen (editores). 1973. *A handbook of method in cultural anthropology*. Nueva York, Academic Press.

Nester, Mark. 1997. "A few quotes regarding hypothesis testing". http://warnercnr.colostate.edu/~anderson/nester.html.

Neupert, Mark. 1994. "Strength testing archaeological ceramics: A new perspective". *American Antiquity*, 59(4): 709-723.

Newman, Mark E. J. 2006. "Power laws, Pareto distributions and Zipf's law". arXiv:cond-mat/0412004v3, http://arxiv.org/abs/cond-mat/0412004v3. Visitado en abril de 2011.

Neyman, Jerzy. 1952. *Lectures and Conferences on Mathematical Statistics and Probability*. 2ª edición. Washington, DC, Graduate School, U.S. Department of Agriculture.

Neyman, Jerzy. 1961. "Silver Jubilee of my dispute with Fisher". *Journal of the Operational Research Society of Japa*n, 3: 145-154.

Neyman, Jerzy y Egon S. Pearson, 1933a. "On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, Vol. 231, pp. 289-337.

Neyman, Jerzy y Egon S. Pearson. 1933b. “The testing of statistical hypotheses in relation to probabilities a priori”. *Proceedings of the Cambridge Philosophical Society*, 24: 492-510.

Nicholls, Neville. 2000. “Commentary and analysis: The insignificance of significance testing”. *Bulletin of the American Meteorological Society*, 81: 981-986.

Nickerson, Raymond S. 2000. “Null hypothesis significance testing: A review of an old and continuing controversy”. *Psychological Methods*, 5(2): 241-301.

Nix, Thomas y J. Jackson Barnette, 1998. “A review of hypothesis testing revisited: Rejoinder to Thompson, Knapp, and Levin”. *Research in the Schools*, 5(2): 55-57.

Nix, Thomas y J. Jackson Barnette. 1998. “The data analysis dilemma: Ban or Abandon. A review of null hypothesis significance testing”. *Research in the Schools*, 5(2): 3-14. http://www.uqtr.ca/metho-lcs/html/Pdfstatis/Data.pdf. Visitado en julio de 2011.

Nunnally, Jum. 1960. “The place of statistics in psychology”. *Educational and Psychological Measuremen*t, XX(4) : 641-650.

Nunnally, Jum. 1975. *Introduction to statistics for psychology and education*. Nueva York, McGraw-Hill.

Oakes, Michael. 1986. *Statistical inference: A commentary for the social and behavioral sciences*. Chichester, UK: Wiley

Ostrow, James. 1990. “The availability of difference: Clifford Geertz on problems of ethnographic research and interpretation”. *International Journal of Qualitative Studies in Education*, 3(1): 61-69.

Pareto, Vilfredo. 1896. “La courbe de la répartition de la richesse”. Reimpreso en 1965 en G. Busoni (editor), *OEeuvres complètes de Vilfredo Pareto, vol. 3: Écrits sur la courbe de la répartition de la richesse*. Ginebra, Librairie Droz. Traducción al inglés en Rivista di Politica Economica, 87 (1997): 647-700.

Park, Hun Myoung. 2008. *Hypothesis Testing and Statistical Power of a Test*. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.” http://www.indiana.edu/~statmath/stat/all/power/index.html. Visitado en julio de 2011.

Patel, Jagdish y Campbell Read. 1982. *Handbook of the normal distribution*. Nueva York, Marcel Dekker, Inc.

Pearson, Karl. 1900. “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”. *Philosophical Magazine*, Series 5(50): 157-175. [Reimpreso en *Karl Pearson’s Early Statistical Papers*, Cambridge University Press, 1956]. http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf. Visitado en agosto de 2011.

Pearson, Karl. 1901. “On some applications of the theory of chance to racial differentiation. From the work of W. R. Macdonell, M. A., LL. D., and Cicely D. Fawcett, B. Sc.” *Philosophical Magazine*, 6ª serie, 1: 110-124.

Pearson, Karl. 1914. *Tables for staticians and biometricians*. Cambridge, Cambridge University Press. http://ia700508.us.archive.org/21/items/p2tablesforstati00pearuoft/p2tablesforstati00pearuoft.pdf. Visitado en julio de 2011.

Peirce, Benjamin. 1852. “Criterion for the rejection of doubtful observations”. *The Astronomical Journal*, 2(21) (n° 45): 161-163. http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1852AJ.....2..161P;data_type=PDF_HIGH. Visitado en julio de 2011.

Peirce, Charles Sanders. 1986 [1873]. “On the theory of errors of observation”. Apéndice 21. En Christian J. Kloesel y otros (editores), *Writings of Charles S. Peirce: A Chronological Edition*. Volumen 3, 1872-1878. Bloomington, Indiana University Press, pp. 140–160.

Pelto, Pertti y Gretel Pelto. 1978. *Anthropological research*. 2ª edición, Cambridge, Cambridge University Press.

Petersen, Glenn.1983. Revisión de *Local Knowledge* de C. Geertz. *Library Journal*, 108(140): 1497.

Piot, Charles y Allen Scult. 1985. Revisión de *Local knowledge*, de C. Geertz. *Quarterly Journal of Speech*, 71(3): 390-392.

Poincaré, Henri. 1912. *Calcul des probabilités*. París, Gauthier-Villars. http://www.archive.org/details/calculdeprobabil00poinrich. Visitado en julio de 2011.

Pollard, P. y J. T. E. Richardson. 1987. “On the probability of making Type I errors”. *Psychological bulletin*, 102: 159-163.

Pólya, György. 1954a. *Mathematics and plausible reasoning: vol. 1, Induction and analogy in mathematics*. Princeton, Princeton University Press.

Pólya, György. 1954b. *Mathematics and plausible reasoning: vol. 2, Patterns of plausible inference*. Princeton, Princeton University Press.

Presburger, Mojżesz. 1929. “Über die Vollständigkeit eines gewissen Systems der Arithmetik ganzer Zahlen, in welchem die Addition als einzige Operation hervortritt”. En: *Comptes Rendus du I congrès de Mathématiciens des Pays Slaves*, Varsovia, pp. 92–101. [Traducción inglesa: “About the completeness of a certain system of integer arithmetic in which addition is the only operation”, http://cs.fit.edu/~ryan/papers/presburger.pdf]. Visitado en setiembre de 2011.

Pridmore, W. A., 1974. Revisión de *Statistics in small doses*, de Winifred Castle. *Journal of the Royal Statistical Society (A)*, 137: 623-624.

Quételet, Adolpe. 1835. *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. París, Bachelier. Vol. 1: http://gallica.bnf.fr/ark:/12148/bpt6k81570d.pdf; Vol. 2: http://gallica.bnf.fr/ark:/12148/bpt6k817719.pdf. Visitado en julio de 2011.

Rachev, Svetlozar T. 2003. *Handbook of heavy tailed distributions in finance*. Amsterdam, Elsevier/North Holland.

Raftery, Adrian. 2001. “Statistic in sociology, 1950-2000: A selective review”. *Sociological methodology*, 31: 1-45.

Raiffa, H. 1968. *Decision Analysis: Introductory lectures on choices under uncertainty*. Reading, Addison-Wesley.

Ramm, Alexander. 2005. *Inverse problems: Mathematical and analytical techniques with applications to engineering*. Boston, Springer.

Ramsey, Christopher Bronk. 2009. “Dealing with outliers and offsets in radiocarbon dating”. http://c14.arch.ox.ac.uk. Visitado en junio de 2011.

Reynolds, R. 1969. “Replication and substantive import: a critique on the use of statistical inference in social research”. *Sociology and Social Research*, 53:,299-310.

Reynoso, Carlos. 1991. *Antropología y programación lógica: Una propuesta sistemática*. http://carlosreynoso.com.ar/antropologia-y-programacion-logica-1991/. Visitado en julio de 2011.

Reynoso, Carlos. 2003. [Planilla de cálculo de caos y dinámica no lineal]. http://carlosreynoso.com.ar/caos.xls. Visitado en setiembre de 2011.

Reynoso, Carlos. 2006. Complejidad y Caos: Una perspectiva antropológica. Buenos Aires, Editorial Sb.

Reynoso, Carlos. 2008. *Corrientes teóricas en antropología: Perspectivas desde el siglo XXI.* Buenos Aires, Editorial Sb.

Reynoso, Carlos. 2009. *Modelos o metáforas: Crítica del paradigma de la complejidad de Edgar Morin*. Buenos Aires, Editorial Sb.

Reynoso, Carlos. 2010. *Análisis y diseño de la ciudad compleja: Perspectivas desde la antropología urbana*. Buenos Aires, Editorial Sb.

Reynoso, Carlos. 2011a. *Redes sociales y complejidad: Modelos interdisciplinarios en la gestión sostenible de la sociedad y la cultura*. Buenos Aires, Editorial Sb.

Reynoso, Carlos. 2011b. *Atolladeros del pensamiento aleatorio: Batallas en torno de la Prueba Estadística de la Hipótesis Nula*. http://carlosreynoso.com.ar/atolladeros-del-pensamiento-aleatorio-batallas-en-torno-de-la-prueba-estadistica/. Visitado en julio de 2011.

Richardson, Lewis Fry. 1948. "Variation of the frequency of fatal quarrels with magnitude". *Journal of the American Statistical Association*, 43: 523-546.

Rigby, Alan S. 1999. "Getting past the statistical referee: moving away from P-values and towards interval estimation". *Health Education Research*, 14(6): 713-715.

Rindskopf, David M. 1997. "Testing 'small,' not null, hypotheses: Classical and Bayesian approaches". En: L. L. Harlow, S. A. Mulaik y J. H. Steiger (editores.), *Op. cit*., pp. 319-332.

Rindskopf, David M. 1998. "Null-Hypothesis Tests Are Not Completely Stupid, but Bayesian Statistics Are Better". *Behavioral and Brain Sciences*, 21(2): 215-216.

Rinehart, Robert. 1998. *Players all. Performance in contemporary art*. Bloomington, Indiana University Press.

Rivadulla, Andrés. "Mathematical Statistics and Metastatistical Analysis". *Erkenntnis*, 34 (2): 211-236.

Robinson, Daniel y Howard Wainer. 2002. "On the past and future of Null Hypothesis Significance Testing". *Journal of Wildlife Management*, 66(2): 263-271.

Robinson, Scott. 2008. "Hypothesis and hypothesis testing".

Robinson, William S. 1950. "Ecological correlations and the behavior of individuals". *American Sociological Review*, 15(3): 351-357.

Rosenthal, R. 1979. "The "file drawer problem" and tolerance for null results". *Psychological Bulletin*, 86(3): 838-641.

Rosnell, R. L. y R. Rosenthal. 1989. "Statistical procedures and the justification of knowledge and psychological science". *American Psychologist*, 44**:** 1276-1284

Rothman, Kenneth J. 1986. *Modern epidemiology*. Boston, Little, Brown.

Rothman, Kenneth J. 1998. "Writing for *Epidemiology*". *Epidemiology*, 9(3): 333-337.

Rozeboom, William W. 1960. "The fallacy of the null hypothesis significance test". *Psychological bulletin*, 57: 416-428. http://stats.org.uk/statistical-inference/Rozeboom1960.pdf. Visitado en julio de 2011.

Rumsey, Deborah. 2003. *Statistics for dummies*. Nueva York, John Wiley & Sons.

Saichev, Alexander, Yannick Malevergne y Didier Sornette. 2010. *Theory of Zipf's law and beyon*d. Berlín y Heidelberg, Springer.

Salsburg, David S. 1985. "The religion of statistics as practiced in medical journals". *The American Statistician*, 39: 220-223. http://www.jstor.org/pss/2683942. Visitado en julio de 2011.

Savage, Leonard. 1957. "Nonparametric significance". *Journal of the American Statistical Association*, 52: 331-344.

Savage, Leonard. 1967. "On rereading R. A. Fisher". *The Annals of Statistics*, 4(3): 441-500. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176343456. Visitado en julio de 2011.

Savage, Leonard y otros. 1962. *The foundations of statistical inference*. Nueva York, John Wiley & Sons.

Scheps, Sheldon. 1982. "Statistical blight". *American Antiquity*, 47(4): 836-851.

Schervish, Mark J. 1996. "P Values: What they are and what they are not". *The American Statistician*, 50(3): 203-206. http://www.cs.ubc.ca/~murphyk/Teaching/CS532c_Fall04/Papers/schervish-pvalues.pdf. Visitado en julio de 2011.

Schlaifer, Robert. 1959. *Probability and statistics for business decisions*. Nueva York, McGraw-Hill.

Schmidt, Frank. 1992. "What do data really mean?: Research findings, meta-analysis, and cumulative knowledge in Psychology". *American Psychologist*, 47: 1173-1181.

Schmidt, Frank. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers". *Psychological Methods*, 1: 115-29.

Schmidt, Frank y John Hunter. 1997. "Eight common but false objections to the discontinuation of significance testing in the analysis of research Data". En: L. Harlow, S. Mulaik y J. Steiger (editores), *Op. Cit*., pp. 37-64. http://homepage.psy.utexas.edu/homepage/class/Psy391P/Schmidt&Hunter.pdf. Visitado en julio de 2011.

Schuchard-flscher, C., K. Backhaus, H. Hummel, W. Lohrberg, W. Plinke y W. Schreiner. 1982. *Multivariate Analysemethoden - Eine anwendungsorientierte Einführung*, 2a edición, Berlín, Springer.

Sedlmeier, Peter y Gerd Gigerenzer. 1989. "Do studies of statistical power have an effect on the power of studies?". *Psychological Bulletin*, 105: 309-316. http://library.mpib-berlin.mpg.de/ft/gg/GG_Do%20Studies_1989.pdf. Visitado en julio de 2011.

Selvin, Hanan. 1957. "A critique of tests of significance in survey research". *American Sociological Review*, 22: 519-527. http://www.jstor.org/pss/2089475. Visitado en julio de 2011.

Serlin, Ronald C. y Daniel K. Lapsley. 1993. "Rational appraisal of Psychological Research and the Good-enough Principle". En: G. Keren y C. Lewis (editores), *A Handbook of Data Analysis the Behavioral Sciences: Methodological Issues*. Hillsdale, Lawrence Erlbaum Associates.

Serrano, María Angeles, Marian Boguñá, Romualdo Pastor-Satorras y Alejandro Vespignani. 2006. "Correlations in complex networks". En: G. Caldarelli y A. Vespignani (compiladores), *Structure and dynamics of complex networks: From information technology to finance and natural science*. Singapur, World Scientific.

Sestini, Piersante y Stefania Rossi. 2009. "Exposing the P value fallacy to young residents". 5th International Conference of Evidence-Based Health-Care Teachers and Developers. Taormina, 29 de octubre. (Ya no más disponible en la Web)

Shankman, Paul. 1985. “Gourmet anthropology: the interpretive menu”. *Reviews in anthropology*, 12: 241-248.

Shaver, James P. 1985a. “Chance and nonsense: a conversation about interpreting tests of statistical significance, Part 1”. *Phi Delta Kappan*, 67: 57-60.

Shaver, James P. 1985b. “Chance and nonsense: a conversation about interpreting tests of statistical significance, Part 2”. *Phi Delta Kappan*, 67: 138-141. Erratum, 1986, 67:624. http://www.jstor.org/pss/20387558. Visitado en julio de 2011.

Shaver, James P. 1993. “What statistical significance testing is, and what it is not”. *Journal of Experimental Education*, 61: 293-316. http://www.eric.ed.gov/PDFS/ED344905.pdf. Visitado en julio de 2011.

Shrout, Patrick E. 1997. “Should significance tests be banned? Introduction to a special section exploring the pros and cons”. *Psychological Science*, 8(1): 1-2.

Shulman, Lee S. 1970. “Reconstruction of educational research”. *Review of Educational Research*, 40: 371-393.

Simon, Herbert A. 1992. “What is an ‘explanation’ of behavior?”. *Psychological Science*, 3: 150-161.

Skinner, Burrhus F. 1972. *Cumulative record*. Nueva York, Appleton-Century-Crofts.

Skipper, James, Anthony Guenther y Gilbert Nash. 1970. “The sacredness of. 05: A note concerning the uses of statistical levels of significance in social science”. *The American Sociologist*, 2: 16-18. http://www.jstor.org/pss/27701229. Visitado en julio de 2011.

Slakter, Malcolm J., You-Wu Wu y Nancy S. Suzuki-Slakter. 1991. “*, **, ***; statistical nonsense at the .00000 level”. *Nursing Research*, 40: 248-249.

Solomonoff, Ray. 1960. *A preliminary report on a general theory of inductive inference*. http://world.std.com/~rjs/z138.pdf. Visitado en julio de 2011.

Sornette, Didier. 2006. *Critical phenomena in the natural sciences: Chaos, fractals, selforganization and disorder. Concepts and tools*. 2ª edición, Berlín-Heidelberg, Springer.

Southwest Fisheries Science Center. 2010. “Papers discussing significance testing”. http://swfsc.noaa.gov/textblock.aspx?Division=PRD&ParentMenuId=228&id=17036.

Spedding, T. A. y T. L. Rawlings. 1994. “Non-normality in Statistical Process Control Measurements”. *International Journal of Quality & Reliability Management*, 11(6): 27-37.

Spielman, Stephen. 1973. “A refutation of the Neyman-Pearson theory of testing”. *The British Journal for the Philosophy of Science*, 24(3): 201-222.

Spirer, Herbert, Louise Spirer y Abram Jaffe. 1998. *Misused statistics*. 2ª edición. Boca Raton, CRC Press.

Sprent, Peter. 1998. “Statistics and mathematics: Trouble at the interface?”. *Journal of the Royal Statistical Society*, 47(2): 239-244. http://www.jstor.org/pss/2988664. Visitado en julio de 2011.

Sprent, Peter y Nigel C. Smeeton. 2001. *Applied nonparametric statistical methods*. Boca Raton, Chapman & Hall/CRC.

Sterling, Theodore. 1960. “What is so peculiar about accepting the null hypothesis?”. *Psychological Reports*, 7: 363-364.

Sterling, Theodor D., W. L. Rosenbaum y J. J. Weinkam. 1995. “Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa”. *The American Statitian*, 49(1): 108-112.

Sterne, Jonathan. 2003. “Commentary: Null points—has interpretation of significance tests improved?”. *International Journal of Epidemiology*, 32: 693-694.

Sterne, Jonathan y George Davey Smith. 2000. “Sifting the evidence—what’s wrong with significance tests?”. *Physical Therapy*, 81(8): 1464-1469. http://ptjournal.apta.org/content/81/8/1464.full. Visitado en julio de 2011.

Stevens, Stanley S. 1968. “Measurement, statistics, and the schemapiric view”. *Science*, 161: 849-856.

Stevens, Stanley S. 1960. “The predicament in design and significance”. *Contemporary Psychology*, 9: 273-276.

Stigler, Stephen. 1978. “Mathematical statistics in the early States”. *The Annals of Statistics*, 6(2): 239-265.

Stigler, Stephen. 1987. “Testing hypotheses or fitting models? Another look at mass extinctions”. En: M. H. Nitecki y A. Hoffman (editores), *Neutral models in biology*. Nueva York, Oxford University Press, pp. 147-159.

Strasak, Alexander, Qamruz Saman, Karl Pfeiffer, Georg Göbel y Hanno Ulmer. 2007. “Statistical errors in medical research – A review of common pitfalls”. *Swiss Medical Weekly*, 137: 44-49. http://www.smw.ch/docs/pdf200x/2007/03/smw-11587.PDF. Visitado en julio de 2011.

Supreme Court of the United States. 2010. *Syllabus*. N° 09-1156. Matrixx Initiatives, inc., et al. V. Siracusano et al. Certiorari to the United States court of appeals for the ninth circuit. Decidido el 22 de marzo. http://www.deirdremccloskey.org/docs/decision.pdf. Visitado en julio de 2011.

Suter, Glenn W. II. 1996. “Abuse of hypothesis testing statistics in ecological risk assessment”. *Human and Ecological Risk Assessment*, 2(2): 341-347.

Taleb, Nassim Nicholas. 2007. *The black swann: The impact of the highly improbable*. Nueva York, Random House.

Tan, W. Y. 1982. “Sampling distributions and robustness of *t*, *F* and variance-ratio in two samples and ANOVA models with respect to departure from normality”. *Communications in Statistics*, A 11: 2485-2511.

Taylor, K. W. y James Frideres. 1972. “Issues versus controversies: Substantive and statistical significance”. *American Sociological Review*, 37: 464-472. http://www.jstor.org/pss/2093185. Visitado en julio de 2011.

Templeton, Alan. 1994. “‘Eve’s’ hypothesis compatibility versus hypothesis testing”. *American Anthropologist*, 96(1): 141-147.

The JBHE Foundation. 1995-1996. “Who Are the Academic Proponents of the Theory of Inferior IQs of Black People?”. *The Journal of Blacks in Higher Education*, 10: 18-19.

Thomas, David Hurst. 1976. *Figuring anthropology: First principles of probability and statistics*. Nueva York, Holt, Rinehart & Winston.

Thomas, David Hurst. 1978. “The awful truth about statistics in archaeology”. *American Antiquity*, 43(2): 231-244.

Thomas, David Hurst. 1986. Ref*iguring anthropology: First principles of probability and statistics*. Waveland Press.

Thompson, Bill. 2001. “402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies”. http://warnercnr.colostate.edu/~anderson/thompson1.html.

Thompson, Bruce. 1999a. “If statistical significance tests are broken/misused, what practices should supplement or replace them?”. *Theory and Psychology*, 9: 167-183. http://www.eric.ed.gov/PDFS/ED413342.pdf. Visitado en julio de 2011.

Thompson, Bruce. 1999b. “Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity”. *Theory and Psychology*, 9:191-196

Tukey, John W. 1960a. “Conclusions vs decisions”. *Technometrics*, 2(4): 423-433.

Tukey, John W. 1960b. “A survey of sampling from contaminated distributions”. En: I. Olkin (editor), *Contributions to Probability and Statistics*. Stanford, Stanford University Press.

Tversky, Amos y Daniel Kahneman. 1971. “Belief in the law of small numbers”. *Psychological Bulletin*, 76, 105-110. http://pirate.shu.edu/~hovancjo/exp_read/tversky.htm. Visitado en julio de 2011.

Tyler, Ralph Winfred. 1931. “What is statistical significance?”. *Educational Research Bulletin*, 10: 115-118,142.

Underwood, Benton J., Carl P. Duncan, Janet A. Taylor y John W. Cotton. 1954. *Elementary statistics*. Nueva York, Appleton-Century-Crofts.

Utts, Jessica. 1991. “Replication and meta-analysis in parapsychology”. *Statistical Science*, 6(4): 363-403.

van der Pas, S. L. 2010. *Much ado about the p-value. Fisherian hypothesis testing versus an alternative test, with an application to highly-cited clinical research*. Disertación de licenciatura, Mathematisch Instituut, Universiteit Leiden.

Vicente, Kim J. y Gerard L. Torenvliet, 2000. “The Earth is spherical ($p < 0.05$): alternative methods of statistical inference”. *Theoretical Issues in Ergonomics Science*, 1(3): 248-271. https://agora.cs.illinois.edu/download/attachments/28936216/Vicente-stats.pdf?version=1&modificationDate=1295122326000. Visitado en julio de 2011.

Vickers, Andrew. 2010. *What is a* p*-value anyway? 34 stories to help you to actually understand statistics*. Boston, Addison-Wesley.

Wainer, Howard. 1999. “One cheer for null hypothesis significance testing”. *Psychological Methods*, 4: 212-213.

Wald, Abraham. 1950. *Statistical decision functions*. Nueva York, John Wiley.

Walk, Christian. 2000. *Handbook of statistical distributions for experimentalists*. Internal Report, SUF-PFY/96-01, Fysikum, Particle Physics Group, Universidad de Estocolmo.

Wallis, W. A. y H. V. Roberts. 1956. *Statistics: A new approach*. Glencoe, The Free Press.

Wall Street Journal. 2011. “A statistical test gets its closeup”. *Wall Street Journal*, 1 de abril, http://www.deirdremccloskey.org/docs/wsj1.pdf. Visitado en julio de 2011.

Warren, W. G. 1986. “On the presentation of statistical analysis: reason or ritual”. *Canadian Journal of Forest Research*, 16: 1185-1191.

Washburn, B. E. y T. W. Seamans. 2007. “Wildlife responses to vegetation height management in cool-season grasslands”. *Rangeland Ecology & Management*, 60: 319-323.

Wasserman, Larry. 2006. *All of non-parametric statistics*. Nueva York, Springer.

Wasserman, Stanley y Katherine Faust. 1994. *Social networks analysis: Methods and applications*. Nueva York, Cambridge University Press.

Watson, Michael y Theodore Graves 1966. “Quantitative research in proxemic behavior”. *American Anthropologist*, 68(4): 971-985.

Weber, Andrzej. s/f. “Evaluation of radiocarbon dates from the Middle Holocene hunter-gatherer cemetery Khuzhir-Nuge XIV, Lake Baikal, Siberia”. http://baikal.arts.ualberta.ca/_NEW/private/dissemination/papers/JAS_KN14_C14_Draft_02_2-spaced_ALL.pdf

Wegener, Charles. 1985. Revisión de *Local knowledge*, de C. Geertz. *American Journal of Sociology*, 91(1): 164-166.

Wilcox, Randy. 2005. *Introduction to robust estimation and hypothesis testing*. 2ª edición, Amsterdam, Elsevier.

Wilkinson, G. N. 1977. “On resolving the controversy on statistical inference”. *Journal of the Royal Statistical Society*, 39(2): 119-171.

Wilkinson, Leland and the Task Force on Statistical Inference. 1999. “Statistical methods in psychology journals”. *American Psychologist*, 54(8): 594-604. http://www.loyola.edu/library/ref/articles/wilkinson.pdf. Visitado en julio de 2011.

Williams, Ken. 1972. “A criticism of the Neyman-Pearson theory of testing”. *Journal of the Royal Statistical Society, series D (The Statician)*, 21(2): 128-131. http://www.jstor.org/pss/2987324. Visitado en julio de 2011.

Wilson, Kellogg. V. 1961. “Subjectivist statistics for the current crisis”. *Contemporary Psychology*, 6: 229-231.

Wilson, Thurlow. 1957. “The Statistical Analysis of Whiting and Child’s Child Training and Personality”. *American Anthropologist*, 59(2): 338-342.

Winch, Robert y Donald Campbell. 1969. “Proof? No. Evidence? Yes. The significance of tests of significance”. *American Sociologist*, 4: 140-143.

Wolpoff, Milford. 1993. “Reply to Dr Foote”. *American Journal of Physical Anthropology*, 90: 381-384.

Yates, Frank y M. J. R. Healy. 1964. “How should we reform the teaching of statistics?”. *Journal of the Statistical Society*, A, 127: 199-210. http://www.jstor.org/pss/2344003. Visitado en julio de 2011.

y’Edynak, Gloria Jean, Brad Bartel, Carl B. Compton, Robert W. Ehrich, David A.Fredrickson, Alexander Gallus, Leo S. Klejn, Matthias Laubscher, Tadeusz Malinowski,Raymond R. Newell, Ari N. Poulianos, Milan Stloukal, Susan C. Vehik y Robert A. Benfer. 1976. “A Test of a Migration Hypothesis: Slavic Movements into the Karst Region of Yugoslavia” [con Comentarios y Respuesta]. *Current Anthropology*, 17(3): 413-428.

Yoccoz, Nigel G. 1991. “Use, overuse, and misuse of significance tests in evolutionary biology and ecology”. *Bulletin of the Ecological Society of America*, 72: 106-111. http://www.abdn.ac.uk/biologicalsci/uploads/files/Yoccoz%20Use%20mis%20use%20abuse%20significance%20testing%20bullesa91.pdf. Visitado en julio de 2011.

Young, R. K. y D. J. Veldman. 1965. *Introductory statistics for the behavioral sciences*. Nueva York, Holt, Rinehart and Winston.

Zabell, S. L. 1992. “R. A. Fisher and the fiducial argument”. *Statistical Science*, 7: 369-387.

Zelterman, Daniel. 2004. *Discrete distributions. Applications in the health sciences*. Chichester, John Wiley & Sons.

Ziliak, Stephen y Deirdre McCloskey. 2009. “The cult of statistical significance”. http://www.deirdremccloskey.org/docs/jsm.pdf. Visitado en julio de 2011.