LUIS CARLOS SILVA AYÇAGUER

Cultura estadística e investigación científica en el campo de la salud:

una mirada crítica

DIAZ DE SANTOS

Cultura estadística e investigación científica en el campo de la salud: una mirada crítica

LUIS CARLOS SILVA AYÇAGUER
Investigador Titular
Instituto Superior de Ciencias Médicas de La Habana

Cultura estadística e investigación científica en el campo de la salud: una mirada crítica

DIAZ DE SANTOS

Ediciones Díaz de Santos, S.A.
Juan Bravo, 3A. 28006 MADRID
España

ISBN : 84-7978-320-6
Depósito legal: M. 34.134-1997

Diseño de Cubierta: Angel Calvete
Fotocomposición: Angel Gallardo, S.L.
Impresión: Lavel, S.A.
Encuadernación: Rústica-Hilo, S.L.

A mis amigos de Uruguay y Cuba, Checoslovaquia y España, los países en que he ido renaciendo con los años.

Para darme un rinconcito en sus altares, me vienen a convidar a indefinirme. Caminando fui lo que fui. Yo me muero como viví.

SILVIO RODRíGUEZ

Contenido

Presentación

Las sociedades, a decir verdad, no han estimado jamás alpensador Lo han considerado, y con razón, como un hereje. No le perdonan sobre todo su originalidad, porque la originalidad es una de las formas de la indisciplina. Frente a un pensador que surge, la sociedad ha seguido dos caminos: o atraerlo para domesticarlo, o perseguirlo para concluir con él.

ANÍBAL PONCE

Este libro no es un conjunto de ideas pretendidamente ingeniosas, hilvanadas en forma de divulgación científica, sobre el uso de la estadística. Está dirigido a un público relativamente especializado: profesionales e investigadores relacionados con la estadística, particularmente a los que laboran en el circuito de la salud. La obra procura abordar con seriedad -que no uso como sinónimo de solemnidad, sino de rigor- algunos problemas técnicos no necesariamente simples. Su propósito central es hacer una contribución a la ***cultura de investigación*** del lector en dos sentidos: uno estrecho, ya que pone el énfasis fundamental en una de las herramientas de la ciencia, la estadística; y otro más amplio, en la medida que intenta colocar ese análisis en el complejo contexto general de la producción de conocimientos.

Su mayor aspiración es ayudar tanto al desmontaje de una dictadura metodológica de origen oscuro y efecto paralizante, como a la desmitificación de algunas prácticas estadísticas, por estar viciadas, una y otras, ya sea de dogmatismo o de error. El mito mayor, quizás, consiste en la convicción de que, al usar la estadística, el investigador está a salvo de la subjetividad, como si ello fuera posible, y exonerado de marcar las conclusiones con su impronta personal, como si ello fuera deseable.

Este propósito no debe confundirse con un afán de socavar los cánones actuales de la producción científica (aunque no estén eximidos *a priori* de examen crítico), ni de minimizar a través de una estrategia de francotirador el inestimable valor de la estadística como herramienta del método científico. De lo que se trata es de «poner

en su lugar» algunos procedimientos, desenmascarar ciertas falacias y abrir líneas de debate.

«El hombre que no crea», escribía Ortega y Gasset, «tenderá a no sentir auténticas necesidades, ya que se encuentra con un repertorio de soluciones antes de haber sentido las necesidades que provocaron aquellas». De ahí mi personal inclinación a favorecer la creación de áreas que toleren e incluso promuevan la controversia y la contradicción. Es en esos espacios de gestación, a menudo asfixiados por pautas oficiales o por la inercia, donde suele germinar la semilla de la creatividad.

He intentado entonces elaborar un discurso enclavado en el dominio de la investigación biomédica y epidemiológica, que va dando cuenta de algunos errores frecuentes, tanto en lo que concierne al manejo de datos como a su interpretación. Estas pifias a veces encarnan una vocación formalista, divorciada del necesario realismo que exige la práctica; otras veces operan en la dirección inversa: dimanan de la aplicación pragmática -ya sea mimética o libertinamente- de técnicas no cabalmente entendidas por quienes las aplican.

No se está presentando, por tanto, un texto para ***enseñar*** estadística. Sin embargo, abrigo la esperanza de que el lector halle algunos enfoques originales y pasajes de valor didáctico, parte de los cuales están destinados a rescatar propuestas poco conocidas.

El tono escogido para la exposición es deliberadamente informal y, en ocasiones, quizás, irreverente. En ese punto me adhiero totalmente a Paul Feyerabend cuando se refiere al lenguaje habitual en la polémica científica:

> ***Éste se ha tomado tan insípido y discreto como el traje de negocios que ahora usan todos, ya sean académicos, hombres de negocios o asesinos profesionales. Acostumbrado a un estilo árido e impersonal, el lector se siente molesto por cualquier digresión con respecto a la monótona norma y ve en ella un signo inequívoco de arrogancia y de agresión; el respeto casi religioso que siente por la autoridad hace que se ponga frenético cuando alguien le tira de la barba a su profeta favorito.***

A lo largo del texto el lector encontrará una colección de recursos de índole diversa, desde anécdotas y ejemplos hasta desarrollos técnicos, pasando por transcripciones literales de textos debidos a diversos pensadores, preguntas abiertas, sugerencias de lecturas y, sobre todo, reflexiones personales que he querido compartir para promover el espíritu crítico que muchos colegas parecen tener adormecido o domesticado.

Se hallará asimismo un nutrido número de citas bibliográficas como sustrato de ese discurso. Un repaso a la bibliografía permitirá comprobar que una parte de la obra citada es de vieja data. La naturaleza de este libro lo impone, ya que también se quiere llamar la atención sobre problemas planteados y alertas lanzadas hace muchos años, aunque con frecuencia olvidados en virtud de la pereza intelectual que conduce a eludir lo conflictivo.

Si mis opiniones provocan desasosiego en el lector o promueven discrepancias de su parte (aunque no es mi objetivo), no lo consideraría desastroso; sería saludable indicio de que su pensamiento no se halla en reposo. Pero la tónica del libro es exactamente opuesta: dar una perspectiva más lógica y flexible a su práctica investigadora, ayudarlo a comprender que los dioses tecnológicos están hechos a imagen y semejanza de sus creadores y heredan, por tanto, muchas de sus debilidades; y creo que cuando se comprende mejor la irracionalidad inherente a todo endiosamiento, se está más habilitado para ejercer la libertad, usufructuar con mayor soltura los espacios que genera y asumir los compromisos que de ella se deriven.

Este libro es el resultado de muchos años de estudio, discusiones, polémicas y dudas. Recoge buena parte del aprendizaje conseguido como subproducto de tales experiencias. Me ha tocado escribirlo, pero tiene muchos autores. Entre ellos, quiero reconocer el aporte de Maggie Mateo, filóloga y amiga entrañable de quien disiento más veces de las que concuerdo; de mi padre, Celiar Silva Rehermann, por su probidad intelectual y su ejemplo de lealtad a lo que cree, y del doctor Francisco Rojas Ochoa, por su notable ductilidad para aceptar lo nuevo y su capacidad para sortear las trampas de lo consagrado.

Otros quizás ni siquiera sospechan cuán importantes han sido en mi formación científica, puesto que ninguno, salvo un par de excepciones, es estadístico, en tanto que ninguno me ha tenido como alumno (algunos sí, como profesor). Su enseñanza, no ha sido ni formal ni sostenida; en algunos casos se trata de una sola idea, aparentemente nimia, pero que resultó ser una llave con la que luego pudieron abrirse muchas puertas. Agradezco por ello a mis compañeros y amigos Arsenio Carmona (Hospital «Julito Díaz» de La Habana), Juan Cabello (Hospital General de Alicante), Eduardo Casado (Facultad de Física, Universidad de La Habana), Alfredo Dueñas (Instituto de Cardiología y Cirugía Cardiovascular de La Habana), Rodrigo Arquiaga (Instituto Nacional de Salud, Valladolid), Manuel Alvarez (Ministerio de Salud Pública de Cuba), Antonio Pareja (Instituto Nacional de Salud, Palma de Mallorca), Jorge Bacallao (Instituto Superior de Ciencias Médicas de La Habana), José Francisco García (Instituto Nacional de Salud, Salamanca), Armando Aguirre (Escuela Nacional de Sanidad de Madrid), José Tapia (Organización Panamericana de la Salud, Washington), Pedro Ordúñez (Hospital «Gustavo Aldereguía» de Cienfuegos) y Humberto Fariñas (Hospital «Hermanos Ameijeiras» de La Habana).

La Habana, agosto de 1997

Prólogo

La tarea que se ha propuesto Luis Carlos Silva en este libro es trazar una panorámica general de la utilización de la estadística en la investigación actual en ciencias de la salud. La obra exige un cierto nivel de conocimientos previos, pues el autor da por supuesto que el lector conoce los conceptos más básicos de la estadística. En ese sentido, el libro no es uno de esos textos introductorios para profanos, ni se limita a describir métodos y procedimientos estadísticos, sino que presenta los usos frecuentes y no tan frecuentes de esas técnicas y expone los abusos que se cometen con ellas.

Normalmente, las obras sobre metodología de la investigación o técnicas estadísticas aburren al más interesado. Suelen adquirirse cuando se hace algún curso y se usan como textos, de forma que tras la lectura de un capítulo o, si hay suerte, de unos pocos, el libro pasa a dormir el sueño de los justos en un estante. A mi juicio, este libro es de otro tipo y he de confesar que al leer el manuscrito solté alguna carcajada. El autor sabe presentar las cosas de forma amena y la obra puede leerse perfectamente de corrido, de la primera a la última página. Claro está que en el libro -que no es una novela- hay partes complejas que exigen cierto esfuerzo. De todas formas, el autor ha sabido explicar de forma sencilla muchas cosas complicadas y son pocos los temas intrincados que se resisten a que el lector les hinque el diente simplemente con una lectura atenta, quizás en algún caso con lápiz y papel. Los capítulos admiten lecturas aisladas, pero es muy recomendable la lectura íntegra de la obra. Quien esté resuelto a profundizar paso a paso en la comprensión de los procedimientos estadísticos en la investigación sanitaria - y científica en general- hallará aquí una lectura provechosa -al menos, para mí lo fue. Cuando acabe el libro, el lector no sabrá hacer regresión multifacial con el SAS o el SPSS -si no sabía antes- pero habrá recorrido con el autor docenas de ejemplos teóricos y prácticos que de seguro le habrán hecho avanzar en su capacidad de razonamiento estadístico. El estupendo librito de Durrel Huff *Cómo mentir con estadísticas* [1], del que en el mundo anglohablante se han vendido más de medio millón de copias

[1] *How to lie with statistics,* Nueva York, W. W. Norton, 1993 (1.ª ed.: 1954).

-que yo sepa, nunca se tradujo al castellano- llevaba como epígrafe una cita de H. G. Wells: «Llegará el día en que, para ser un buen ciudadano, el razonamiento estadístico será tan necesario como saber leer y escribir». Aún no llegó el día previsto por ese visionario -pese a las apariencias vivimos en un mundo que fomenta las tareas para las que se requiere poco razonamiento, sea estadístico o de otro tipo- pero es evidente que quienes quieren hoy hacer investigación no pueden conformarse con saber las cuatro reglas aritméticas. De lo contrario se autoengañarán o serán *estadísticamente* engañados.

Nunca estuve en Cuba y la única vez que he visto a Luis Carlos Silva fue en Guadalajara (México), donde coincidimos en el VIII Congreso Mundial de Medicina Social. Desde entonces ni siquiera hemos hablado por teléfono. Sin embargo, un intercambio inicial sobre cuestiones de terminología epidemiológica se complicó con otros asuntos y las cartas comenzaron a ir y venir con comentarios, discrepancias y matizaciones sobre autores y temas tan diversos como Otto René Castillo, Paul Feyerabend, la cuantificación del desarrollo humano o la definición de la representatividad de una muestra. Finalmente, un día recibí una nota en la que Luis Carlos me proponía escribir un prólogo para este libro. Yo ya había leído entonces varios capítulos y, por supuesto, me hizo una gran ilusión aceptar su propuesta.

Luis Carlos Silva es matemático, pero se nota a la legua que hace años dejó los axiomas y los teoremas archivados y se dedicó a cosas mucho más terrenas, en concreto, a aplicar sus conocimientos matemáticos y de computación a la investigación biomédica y sanitaria. Él no me lo ha contado, pero me imagino que ha debido participar como asesor metodológico o estadístico de docenas o quizá centenares -ya tiene algunos años- de investigaciones clínicas y epidemiológicas. Quizá es esa la razón por la que a menudo de su pluma salen dardos afilados dirigidos contra los que él llama *metodólogos* -en *cursiva o* entre comillas-, que sin haber hecho jamás una investigación se permiten el lujo de dar instrucciones sobre cómo hacerla. Sin duda, en los razonamientos de Luis Carlos Silva siempre está muy presente el criterio de la práctica.

Para acabar este prólogo me gustaría citar un texto del afamado Gregorio Marañón, que viene muy a propósito del tema de este libro, cuyos lectores serán médicos en proporción considerable.

> ***Nuestra generación*** -decía Marañón- ***advino a la profesión y a la ciencia bajo el signo de las estadísticas. Primero los cirujanos y después también los internistas, consideraban como la expresión de su sabidunía médica el reunir un número grande de pacientes de cada enfermedad. Una gran escuela médica, la norteamericana, hizo de la estadística su fundamento principal. Y aun nosotros, europeos modestos, escribíamos con orgullo el número crecido de nuestros ulcerosos de estómago o de nuestros tabéticos.***
>
> ***Gran error Aquello, tan manoseado, de que los árboles impiden ver el bosque, a ninguna actividad humana puede aplicarse con tanta razón como a la clínica. Los muchos, los demasiados enfermos, dan, sin duda, seguridad y prestancia para ir y venir entre ellos***

> ***y agudizar el golpe de vista del patólogo. Pero el verdadero conocimiento de la enfermedad lo da el estudio profundo de «cada caso», en el que se resume no sólo el esquema del proceso sino todas sus posibles variedades*** [2].

Quienes sigan aferrados a puntos de vista como éstos no podrán hacer buena investigación médica o sanitaria. Efectivamente, los árboles pueden ocultar el bosque y cada paciente es único como persona y como «caso», pero «los muchos enfermos» nunca son demasiados si de lo que se trata es de extraer inferencias generales sobre el curso de la enfermedad, su clínica o su tratamiento. Por otra parte, quienes piensen que la efectividad de los tratamientos se demuestra en la práctica médica diaria -esto creo que por desgracia es bastante corriente- sin controles ni métodos de aleatorización, podrán ser buenos terapeutas y dar una atención de calidad y máxima empatía a sus pacientes - y eso es tan importante como la bioquímica de los fármacos o la estadística de los ensayos clínicos- pero no podrán hacer buena investigación. Para ello hacen falta ideas mucho más claras que éstas de Marañón sobre lo que la estadística aporta al desarrollo de las ciencias de la salud, sea en el ámbito clínico o epidemiológico. Este libro de Luis Carlos Silva ilumina muy bien distintos aspectos de esta aportación.

JOSÉ A. TAPIA GRANADOS.
Nueva York, junio de 1997.

[2] «Medicina estadística» en ***La medicina y los médicos,*** recopilación de escritos de Gregorio Marañón a cargo de A. Juderías (Madrid, Espasa-Calpe, 1962, p. 39), donde se da como fuente original el prólogo al libro *Veinticinco años **de labor: historia y bibliografia del profesor G. Marañón, por sus alumnos*** (Madrid, 1935).

La estadística en su contexto

Lejos de ser autónoma, la ciencia florece o se marchita junto con la sociedad. Lo mismo pasa con la tecnología, las humanidades y las artes.

MARIO BUNGE

1.1. De la marginalidad a la cima

Cuatro décadas atrás, muchos investigadores médicos carecían de entrenamiento profesional y de asesoría pertinente para la aplicación adecuada de métodos bioestadísticos [1]. Según Hill (1965) verdadero gestor de la estadística aplicada a la medicina (véase Doll, 1992) para explicar la ausencia de esta herramienta en sus trabajos, en aquella época los investigadores esgrimían con frecuencia el argumento de que su uso era injustificado en virtud de la dudosa calidad de los datos a los cuales habría de aplicarse. Tal argumento era claramente endeble, como el propio Bradford Hill se ocupó de subrayar: si la naturaleza del problema demanda que los datos sean analizados estadísticamente, es inaceptable pasar por alto esta exigencia; si no son confiables, lo más razonable sería desecharlos.

A lo largo de las décadas siguientes esa realidad se ha ido invirtiendo paulatinamente hasta arribar a una situación de signo claramente opuesto: muchos investigadores biomédicos -sobre todo si sus trabajos no han recorrido un camino metodológicamente riguroso y maduro- se muestran ansiosos por «aderezar» sus análisis con técnicas estadísticas. Algunos de ellos están persuadidos de que lo ideal sería recurrir a las más intrincadas. Es obvio que tal aberración tiene un origen básicamente exógeno: el impetuoso desarrollo -paralelo a la creciente institucionalización de la producción científica- de un metodologismo [2] trepidante, que halla en la matematización uno de sus disfraces más elegantes.

---

[1] En algunos puntos a lo largo del texto se usará el término «bioestadística» para aludir a la estadística como disciplina cuando su aplicación se ubica en el ambiente biológico y, por extensión, en el campo de la salud; en este último caso, naturalmente, se hace uso de cierta licencia, ya que el componente biológico constituye sólo una de las dimensiones que integran el sistema cognoscitivo de la salud.

[2] Nótese desde ahora la distinción entre «metodologismo» y «metodología». *Método* es una palabra de raíz grie-

En un sencillo pero esclarecedor material concebido para profesionales sanitarios, basándose en el estudio de 168 artículos publicados en 6 revistas médicas, Castle (1979) llegó a la conclusión de que muchos médicos parecen usar la estadística como los borrachos los faroles: para apoyarse, no para iluminarse.

Ese panorama, lejos de desdibujarse con el tiempo, se ha consolidado en un nuevo escenario donde hay más publicidad orientada a persuadir a los investigadores de que pueden ser estadística y computacionalmente autosuficientes, que voces alertando sobre los riesgos inherentes al ejercicio de tal independencia.

En cualquier caso, lo cierto es que las técnicas estadísticas, al menos en el mundo sanitario, parecen haber completado el tránsito que las han llevado de la marginalidad a ocupar importantes espacios de poder.

Hay un poder legítimo: el que dichas técnicas confieren al investigador que las domina; pero hay otro aberrante: el que se genera cuando las técnicas parecen adquirir vida propia y ejercen su influjo sobre los propios usuarios para producir una imagen lastimera: un profesional atrapado por demandas que no entiende; un trabajador intelectual quien, en lugar de sentirse amparado por recursos metodológicos razonables, es presa del temor asociado a la eventual descalificación de sus acciones técnicas, y que se apresura sumisamente a procurarse un aval que a menudo ni siquiera necesita.

Para Ingenieros (1957):

> ***Los investigadores ennoblecerán su propia ética cuando se desprendan de los dogmas convencionales que perturbaron la lógica de sus predecesores. Sin la firme resolución de cumplir los deberes de la crítica, examinando el valor lógico de las creencias, el hombre hace mal uso de la función de pensar; convirtiéndose en vasallo de las pasiones propias o de los sofismas ajenos.***

1.2. Los metodólogos usan el poder

1.2.1. Una suma de presiones: el último sumando

Einstein, adelantándose de algún modo al pensamiento popperiano señalaba en una carta inédita [3] que la naturaleza nunca dice sí, sino sólo no o, en el mejor de los casos, ***quizá.*** ¿Son las leyes de la naturaleza la única barrera que limita a los investigadores en el ejercicio de su libertad creativa? Lamentablemente no; Bunge (1985) consigna que «los científicos, por destacados que sean, no logran sustraerse

---

ga que significa «camino para alcanzar un fin» ***(meta,*** fin; ***odo,*** camino). El metodologismo es la perversión de la metodología, una práctica que hace de la aplicación de los métodos una finalidad en sí misma.

[3] Así lo consigna Jesús Mosterín, catedrático de Lógica, Historia y Filosofía de la Ciencia en la Universidad de Barcelona (Mosterín, 1995).

a todas las presiones del medio, por lo cual a veces se comportan de manera no científica».

El físico Richard Feynman, ganador del premio Nobel y una de las más atractivas personalidades científicas de nuestro siglo, escribe con vehemencia sobre la integridad científica como un principio del pensamiento científico. En su fascinante libro titulado *Seguramente usted está bromeando Mr. Feynman,* llega a decir (Feynman, 1985):

> ***Lo que más desearía es que usted tenga la fortuna de radicar en un lugar donde nunca se sienta forzado a perder su integridad para mantener su posición o conseguir apoyo financiero.***

Las presiones pueden ser - y son, de hecho- de naturaleza diversa: abarcan el área de lo político y lo económico, de lo social y lo académico; nuestro interés se centra ahora, sin embargo, en las que se despliegan al amparo de un presunto rigor y en nombre del sagrado interés por la verdad: las presiones metodológicas ejercidas por ciertos especialistas de la crítica, profesionales de la metodología y fiscales de toda laya, uno de cuyos sospechosos rasgos suele ser la mediocridad propia de quienes se especializan en enseñar a hacer lo que nunca han hecho, o en criticar lo que no hubieran sido capaces de realizar por sí mismos.

Alexander Zinoviev (citado por Fernández, 1991) señala con dureza que:

> ***De ciencia que, en lo fundamental, proporcionaba ciertos consejos sencillos, pero positivos, la metodología se convirtió en una colección de obras críticas que lo que proporciona son refutaciones complicadas, fundamentalmente negativas, a las soluciones positivas de los problemas. Y cuando los especialistas en metodología dan consejos positivos no se puede evitar el compararlos con los consejos de los alquimistas. Al igual que éstos vendían gustosamente sus recetas para la obtención de oro sin ponerlas ellos en práctica, también los especialistas en metodologia enseñan gustosamente a todos la forma de hacer descubrimientos científicos, pero se las ingenian para no hacerlos ellos ni siquiera en su propio ámbito.***

1.2.2. Códigos de actuación para producir conocimientos

A mi juicio los puntos de vista examinados en la sección precedente son reacciones contra determinadas formas que han asumido la práctica y la enseñanza de la metodología de la investigación antes que negaciones en bruto de esta disciplina como tal.

En su famoso ***Discurso del método,*** Descartes enunció un conjunto de sencillas reglas acerca de cómo discurrir lógicamente en el acto de investigación. El filósofo y geómetra francés sugería acciones tales como «dividir cada dificultad en cuantas

porciones sea preciso para mejor atacarlas». Sin embargo, ello no lo salva de juicios como el de Goethe: «Descartes escribió y reescribió su libro muchas veces; y, aun así, sigue siendo inútil» [4] o como el que hace Ramón y Caja1 (1945) cuando, refiriéndose a dichas reglas, escribe:

> ***Tengo para mí que el poco provecho obtenido de la lectura de tales obras, y en general de todos los trabajos concernientes a los métodos filosóficos de indagación, depende de la vaguedad y generalidad de las reglas que contienen, las cuales, cuando no son fórmulas vacías, vienen a ser la expresión formal del mecanismo del entendimiento en función de investigar. Este mecanismo actúa inconscientemente en toda cabeza regularmente organizada y cultivada, y cuando, por un acto de reflexión, formula el filósofo sus leyes psicológicas, ni el autor ni el lector pueden mejorar sus capacidades respectivas para la investigación científica. Los tratadistas de métodos lógicos me causan la misma impresión que me produciría un orador que pretendiera acrecentar su elocuencia mediante el estudio de los centros del lenguaje, del mecanismo de la voz y de la inervación de la laringe. Como si el conocer estos artificios anátomo-fisiológicos pudiera crear una organización que nos falta 0 perfeccionar la que tenemos.***

Sin embargo, el propio Caja1 se pregunta si debe dejarse desorientado al principiante, entregado a sus propias fuerzas y marchando sin consejo por una senda llena de dificultades y peligros. Su respuesta es la siguiente:

> ***De ninguna manera. Pensamos, por lo contrario, que si, abandonando la vaga región de los principios filosóficos y de los métodos abstractos, descendemos al dominio de las ciencias particulares y al terreno de la técnica moral e instrumental indispensable al proceso inquisitivo, será fácil hallar algunas normas positivamente útiles al novel investigador: Algunos consejos relativos a lo que debe saber, a la educación técnica que necesita recibir; a las pasiones elevadas que deben alentarle, a los apocamientos y precauciones que será forzoso descartar; opinamos que podrán serle harto más provechosos que todos los preceptos y cautelas de la lógica teórica.***

Esta realidad ha sido objeto de análisis por los más interesantes pensadores contemporáneos. Por ejemplo, Wright (1961), en un libro profundamente crítico de las tendencias normalizadoras en el mundo de la sociología sajona, tres décadas atrás nos brindó un juicio que merece, en mi opinión, la máxima atención:

> ***Cuando hacemos una pausa en nuestros estudios para reflexionar sobre la teoría y el método, el mayor beneficio es una reformulación de nuestros problemas. Quizás es por eso por lo que, en la práctica real, todo investigador social activo debe ser su propio metodólogo y su propio teórico, lo cual sólo quiere decir que debe ser un artesano intelectual. Todo artesano puede, naturalmente, aprender algo de los intentos generales para***

[4] En ***Maximen und reflexionen,*** según se cita en Pera (1994).

> ***codificar los métodos, pero con frecuencia no mucho más que un conocimiento de tipo muy general. Por eso no es probable que los «programas ruidosos» en metodología contribuyan al desarrollo de la ciencia social (...) es mucho mejor la información de un estudioso activo acerca de cómo proceder en su trabajo que una docena de «codificaciones de procedimiento» hechas por especialistas que quizás no han realizado ningún trabajo de importancia. Únicamente mediante conversaciones en que pensadores experimentados intercambien información acerca de su manera real de trabajar puede comunicarse al estudiante novel un concepto útil del método y de la teoría.***

Nótese el adjetivo «ruidoso» que elige Wright Mills: queda claro que no se opone a la formulación de orientaciones generales, ni al debate epistemológico, ni al intercambio de experiencias fecundas; reacciona contra el dogma y la banalidad de las recetas.

En otro lugar (Silva, 1989) ya opiné que era menester reconocer que la enseñanza formal de la metodología no ha sido medular en el desarrollo de la investigación en el mundo. Para Medawar (1984):

> ***...la mayoría de los hombres de ciencia no ha recibido ninguna instrucción formal en el método científico y quienes parecen haberla recibido no muestran una superioridad sobre quienes no la recibieron. Los asuntos cotidianos de la ciencia requieren el ejercicio del sentido común apoyado por un poderoso entendimiento. El método científico es una potenciación del sentido común.***

Sin embargo, creo que la instrucción en esta materia puede ser fecunda, siempre que se trate de una docencia viva, flexible, que tenga en cuenta, como advierte Kapitsa (1985) que al desarrollar las facultades creadoras del alumno, se precisa un enfoque individual. Desde mi punto de vista, la enseñanza de la metodología de la investigación debe centrarse en la discusión de problemas concretos, procurando educar y no adocenar, informar y no uniformar, animar la vocación crítica mediante el intercambio vivo de impresiones en torno a proyectos reales. Una de las expresiones más negativas que han mediatizado la utilidad de esta disciplina ha sido la de poner más énfasis en transferir códigos de procedimiento para resolver problemas que en la formulación adecuada de éstos.

Muy pocos cursos sobre el tema se ocupan de examinar los problemas fundamentales que aquejan a las preguntas que se formulan los investigadores noveles. En un artículo publicado hace algunos años (Silva, 1991) llamé la atención sobre lo que en mi experiencia parece ser la más común de las dificultades y la que con más claridad revela la falta de elaboración del problema: que su enunciado incluya parte del método para resolverlo. Planteado un problema científico, las vías para resolverlo pueden ser diversas. Diferentes enfoques y recursos pueden usarse, y unos serán más fecundos o ingeniosos que otros; puede ocurrir incluso que algunos sean totalmente inaceptables (por razones prácticas, materiales, o aun científicas), pero el problema sigue siendo exactamente el mismo. Ello subraya la veracidad de que *la*

formulación de un problema bien planteado debe prescindir de toda alusión al método o métodos que habrán de usarse para resolverlo.

Sin embargo, no es nada infrecuente tropezar con objetivos como los que, a modo de ilustración, se enumeran y comentan para concluir esta sección [5].

1. ***Determinar el nivel medio de colesterol de la población de Pinar del Río a través de una muestra representativa de la provincia***

Es improcedente consignar en el propio enunciado a través de qué procedimiento se hará la determinación deseada: el muestreo es una técnica, una forma de trabajo que se implementa ***en función*** del objetivo trazado y, en esa medida, no es parte de él.

2. Comparar el desempeño de los cirujanos antes y después de recibir el curso de entrenamiento

Comparar es una acción claramente metodológica. Nunca el acto de comparar constituye una finalidad ***per sé.*** Quizás la verdadera finalidad del investigador en este ejemplo es evaluar la eficiencia de cierto entrenamiento para los cirujanos; en el enunciado se está mezclando esa finalidad con el esbozo del diseño por conducto del cual habría de consumarse la evaluación.

3. ***Determinar si el consumo de alcohol es mayor entre los casos de demencia senil que entre los controles tomados de la consulta oftalmológica***

Nuevamente, lo que aparentemente se desea es determinar si el hábito alcohólico influye o no en el desarrollo posterior de la demencia senil: pero es absurdo y metodológicamente incorrecto comprometerse de antemano a realizar tal determinación mediante esta o aquella vía (estudio de casos y controles en este caso). Precisamente, este compromiso apriorístico reduce el espacio para meditar acerca de la forma óptima de encarar el asunto.

4. Realizar una encuesta sobre accidentes entre jóvenes universitario de La Habana

Este es el caso extremo, en que el acto metodológico (realizar la encuesta) ¡se ha convertido completamente en una finalidad! Si tal encuesta quiere realizarse, seguramente es con el fin de acopiar información que consienta dar respuesta a alguna pregunta sustancial sobre accidentalidad. Naturalmente, el problema consiste precisamente en obtener esa respuesta, nunca en el camino para conseguirlo.

[5] Todos son tomados de tesis de especialización realizadas por médicos cubanos.

1.2.3. **El protocolo: puente y escollo**

Hay múltiples «guías» para la confección de los proyectos de investigación. El hecho de que unas sean más exigentes que otras responde en general a las inclinaciones personales de sus autores y a requerimientos administrativos variables.

Existe un conjunto de elementos básicos que no pueden estar ausentes en un proyecto si se quiere que cumpla su cometido. Pero, al mismo tiempo, opino que, si tales elementos aparecen de manera secuencial y comprensible para cualquier lector, el documento cumpliría con su razón de ser. Así, al margen de subdivisiones estructurales cuya definición puede quedar sujeta al estilo de los investigadores, el problema científico debe quedar nítidamente expresado en términos de preguntas o hipótesis, fundamentada la necesidad de encararlo y expuestos tanto el marco teórico en que se inscribe como los antecedentes en que reposa. Por otra parte, suele resultar en general conveniente concretar separadamente en forma de preguntas o de objetivos explícitos los propósitos del estudio. Deben finalmente consignarse el método general y las técnicas particulares que en principio se proyecta utilizar para alcanzarlos. Otros aspectos como los recursos necesarios o el cronograma del proceso, se exigirán o no en dependencia del contexto institucional correspondiente.

Una guía será fecunda, en síntesis, en la medida que conduzca con eficiencia a la transmisión clara y concisa de las acciones que se proyectan, a la vez que se aleje de todo encasillamiento. Cabe aclarar que no me estoy refiriendo al esquematismo formal (el que establece reglas sobre la estructura del texto y el modo de ordenarlo); éste es tan ridículo que no merece ser dignificado por la réplica. En cambio, el esquematismo conceptual (el que defiende determinados pasos aunque no ayuden a llegar al destino deseado), es mucho más preocupante. Esquematismo al fin, es paralizador y atenaza la creatividad, confunde los medios con los fines y, de paso, confunde también al investigador.

Una regla formalmente esquemática establecería: ***«El proyecto debe tener su capítulo de definiciones escrito en cuartilla independiente y colocarse a renglón seguido de los objetivos».*** El esquematismo conceptual exigiría en cambio: ***«Todo proyecto de investigación tiene que contener una sección destinada a definiciones».*** En mi opinión basta con exigir que el documento sea claro e inteligible; al autor compete decidir, teniendo en cuenta los conceptos incluidos en el texto y, sobre todo, las características de sus potenciales lectores, si para lograrlo debe o no hacer delimitaciones semánticas explícitas.

Estimo que los mitos que se han creado en torno al proceso de investigación y su planificación son múltiples. Muchos de ellos se han generado a partir de una buena intención didáctica, o han nacido del sano afán de estructurar el pensamiento científico de los principiantes, pero, por desgracia, en ocasiones han derivado hacia expresiones de sumisión a los cánones, colocándose así en paradójica contradicción con la flexibilidad y capacidad autocrítica inherente a la ciencia.

El primer aspecto que merece desmitificación es la propia utilidad del protoco-

lo. Su confección es un paso de importancia cardinal, pero no por las razones que suelen invocarse, en especial la de ser un documento para la consulta permanente del investigador sobre los pasos que habrá de ir dando. Lo cierto es que, una vez confeccionado, quienes conducen la investigación raramente acuden a él durante el curso de la indagación y, cuando lo hacen, usualmente no es con la finalidad de consultarlo sino para modificarlo. Esto es lógico, ya que lo normal es que los investigadores no necesiten consultar lo que ellos mismos concibieron, maduraron y finalmente plasmaron en el documento; en cambio, la situación se invierte cuando se trata de enfrentar las dificultades no planificadas.

En mi opinión, el protocolo es útil por dos razones: por un lado, como elemento para la aprobación administrativa y el control técnico por parte del aparato institucional que financia o ampara al estudio; por otro, porque el hecho mismo de que el investigador se vea obligado a plasmar explícitamente sus motivaciones, ideas y planes constituye una práctica esclarecedora, un ejercicio intelectual que contribuye a la organización conceptual y a la autodisciplina.

En cualquier caso, el proyecto tiene que ponerse en función del investigador y no viceversa. Muchas veces se ve a profesionales envueltos en la tarea estéril de dar la «forma debida» a sus proyectos, de suerte que concuerden con expectativas arbitrarias de eventuales censores.

Quizás ningún ejemplo sea más universal (al menos en algunos entornos académicos) que el de construir «objetivos generales» y «objetivos específicos». Muy pocos libros buenos dedicados a la metodología de la investigación establecen la necesidad u obligación de fijar tales objetivos generales, ni sugieren hacerlo; muchos ni siquiera mencionan la posibilidad: lo que todos consignan es que hay que establecer claramente los propósitos del estudio, pero lo típico es que enfaticen la conveniencia de formularlos mediante preguntas. Según mi opinión, no se comete un error si en una guía o en una clase se menciona ***la posibilidad*** de definir un objetivo general del que se derivan los propósitos del estudio, pero bajo ningún concepto puede establecerse sectariamente como un paso cuya ausencia descalifique el proyecto. Hacerlo sería desconocer la práctica y la literatura científica universales.

Pero esto no es lo más grave: lo realmente lastimoso es ver a un joven científico que, ***después*** de haber establecido diáfanamente sus preguntas, proceda entonces a la construcción de un llamado «objetivo general», recurriendo a complejas acrobacias verbales para lograr una síntesis que nada agrega ni quita (salvo el tiempo y la energía destinados a ello) y que más bien se parece a un ejercicio de retórica.

1.2.4. Racionalidad y anarquismo gnoseológico

El anarquismo epistemológico, postura sostenida por polemistas de indudable brillantez (véase, por ejemplo, el inquietante ***Tratado contra el método*** de ***Feyerabend, 1981), se*** coloca en el extremo más alejado de este totalitarismo metodológico.

Feyerabend (1982) diagnostica con notable agudeza muchos de los males del metodologismo y sus áreas conexas. Por ejemplo, reflexionando en torno a las coordenadas sociales en que ha de desempeñarse la acción intelectual, escribe:

> ***Nuestras mentes y nuestros cuerpos*** están ***sometidos*** a ***limitaciones muy diversas. Nuestra educación no nos ayuda a reducir tales limitaciones. Desde la misma infancia estamos sometidos a un proceso de socialización y aculturación (por describir con palabras nada gratas un proceder que tampoco es nada grato), comparado con el cual el adiestramiento de los animales domésticos, de las fieras del circo o de los perros-policía son un simple juego de niños.***

El escepticismo -herramienta cardinal de la ciencia y punto de partida de muchas de sus iniciativas [6]– tampoco puede constituir una finalidad en sí misma. Si no se admitieran pautas valorativas del contenido de verdad de una teoría científica, y de las reglas para usarlas, se estarían abriendo las puertas a la pseudociencia. Ahí radica el aspecto más pernicioso del anarquismo gnoseológico: por su conducto se pone en pie de igualdad (¡así lo hace literalmente Feyerabend!) la magia con la tecnología, la acupuntura o la medicina psiónica con la medicina científica, las leyes de la astronomía con el zodiaco.

Materazzi (1991), destacado psiquiatra comunitario argentino, denuncia como habitual que las estructuras autoritarias generen instituciones especiales, tendentes a reemplazar el pensamiento libre y los criterios personales por posiciones congeladas y mesiánicas acerca de lo que la gente tiene que decir o pensar.

El proceso de producción científica no escapa a ese mal. Sería por lo menos una ingenuidad suponer que un sistema tan rentable como el que rige la actividad científica contemporánea esté inmunizado contra el afán natural que toda estructura social tiene por alcanzar cuotas de poder que van más allá de las que corresponden a sus logros reales. Presumo que Materazzi no se refería exclusivamente a las estructuras que se concretan en forma de instituciones formales; las hay que pueden -tal es el caso del metodologismo- consolidarse como tradición a partir de una praxis utilizada interesadamente por algunos de los que viven de ella y admitida acríticamente (o simplemente padecida) por la mayoría.

Guttman (1977) plantea el problema en los siguientes términos:

> ***La experiencia muestra que la intolerancia suele venir de los firmes creyentes en prácticas sin fundamento. Tales devotos actúan frecuentemente como árbitros y jueces científicos, y no se arredran a la hora de adelantar críticas irrelevantes y decisiones negativas sobre aquellos nuevos desarrollos que no comulguen con sus errores conceptuales preferidos.***

[6] Cuando se le preguntó a Carlos Marx cuál era su precepto favorito, contestó «De omnibus dubitandum», o sea, dudar de todo, con lo cual reivindicaba un elemento cardinal del racionalismo materialista.

Parecería que el péndulo metodológico que se mueve de un extremo a otro -del autoritarismo castrador a la liberalidad inconducente- necesita de regulaciones periódicas. Este libro intenta, precisamente, hacer una modesta contribución reguladora tanto en una dirección del proceso como en la opuesta.

1.3. Los pilares de la producción científica

No ha de parecernos blasfemia infundamentada el anuncio de que en el mundo metodológico (y por ende, en el de la estadística) existen múltiples vicios de que precavernos, incluyendo no sólo el dogmatismo y la inercia intelectual sino, incluso, la deshonestidad y el simple fraude. Es relativamente natural que así sea por al menos dos razones. 1º) Porque los «metodólogos» -admitan o no desempeñar tal profesión- son seres humanos y, como tales, no están a salvo de las mismas tentaciones que alcanzan a un taxista, a un cartero o a una corista, de actuar según intereses personales, generalmente con un trasfondo económico[7]. 2º) Porque todo el sistema de producción científica se da en un entorno que está lejos de haber erradicado esos mismos males. Las reflexiones que siguen procuran, precisamente, fundamentarlo.

La ciencia contemporánea es uno de los pocos sistemas sociales autorregulados, en el sentido de que carece de mecanismos *externos* de control; sin embargo, casi toda norma de actuación socializada (el sistema escolar, el código del tránsito o la práctica del sacerdocio, por mencionar tres ejemplos) o bien está sujeta a la auditoría externa o bien se practica, por lo menos, a través de estructuras jerárquicas verticales (ejercida por inspectores, policías u obispos). Resulta en extremo peligroso para la propia eficacia de un sistema que el control recaiga exclusivamente sobre sus protagonistas (que los maestros se evaluaran unos a otros, que los conductores se sancionaran mutuamente, o que los clérigos fueran los responsables de reprender a sus colegas).

Cabe detenerse, siquiera brevemente, en el examen del mencionado proceso autorregulador en el caso de la producción científica. ¿Cuáles son sus mecanismos concretos? ¿cuán sólidos resultan? ¿Qué consecuencias se derivan de sus debilidades?

Los pilares en que se sustenta el control de la investigación científica contemporánea son, en esencia, los dos siguientes: la replicación y el llamado arbitraje científico, o *peer review* de los anglosajones[8]. A continuación se bosqueja el *modus operandi* de cada uno de estos procedimientos.

[7] «Hablen de lo que hablen, estarán hablando de dinero», afirma una de las *Leyes de Murphy*.

[8] Usaré en lo sucesivo la conocida expresión inglesa. Si me viera obligado a castellanizarla sin pérdida de matices, escribiría «revisión crítica de los resultados de un científico a cargo de sus pares». Las razones de mi elección son obvias.

1.3.1. La replicación

Puesto que los resultados científicos, para serlo, tienen que estar sujetos a contrastación o verificación, su propia lógica de producción debe dar la oportunidad para comprobar el contenido de verdad que encierran. Dicho de otro modo, a diferencia de lo que ocurre en tantas otras áreas de la vida, supuestamente no existe autoridad alguna, por mucho poder, carisma o prestigio que tenga, que pueda hacer valer sus puntos de vista por encima del dictamen surgido de la aplicación del método científico.

En efecto, el teorema de Pitágoras y la estructura heliocéntrica del sistema solar son ciertos, aunque el Papa o el presidente de los Estados Unidos opinasen lo contrario; del mismo modo que la parapsicología y la teoría de los biorritmos son supercherías, aunque hayan conseguido popularidad, ya que un error no deja de serlo por el hecho de que se repita con mucha frecuencia [9].

Todo resultado científico es, en principio, de público dominio y puede ser evaluado en procura de corroboración. El conocimiento científico difiere del que no lo es, precisamente, en que aquel ha de ser verificable. La replicación es el recurso corroborativo por antonomasia.

Puesto que un resultado dado -en especial el surgido de la experimentación- pudiera haberse establecido erróneamente, ya sea por mero azar, por manipulación espuria de los datos o por equivocación de sus descubridores, sus puntos débiles habrán de salir a la luz cuando el proceso que lo produjo es reproducido en una o, preferiblemente, en varias ocasiones y en diversas circunstancias. En las nuevas experiencias, un ocasional efecto insidioso del azar no volvería a aparecer, los posibles errores -de haber existido- no se cometerían nuevamente, y saldrían a la luz indicios de una posible manipulación.

La lógica de este planteamiento es impecable. Su implantación práctica es harina de otro costal.

El escollo fundamental radica en que ***los investigadores no tienen motivaciones de peso para replicar el trabajo ajeno.*** El sistema de recompensas de la ciencia es especialmente desestimulante de la replicación, ya que sólo premia al que ofrece una novedad, pero no da virtualmente nada al que la convalida sin modificación alguna. Es incluso muy difícil conseguir la publicación de un artículo que se limite a reproducir literalmente el trabajo de otro. Asimismo, las bases de los concursos y premios, las normas para la aceptación de tesis doctorales y las reglas de financiamiento de los proyectos reclaman, antes que nada, originalidad.

Broad y Wade (1982) señalan otras dos razones complementarias para el desestímulo: a saber, que la descripción de que son objeto los métodos empleados

[9] Recuérdese la historia del hombre del que hablaba Wittgenstein, quien, al dudar de la veracidad de algo que ha leído en el periódico, compra 100 ejemplares del diario para constatar su veracidad.

es frecuentemente incompleta y que la repetición de una experiencia suele ser tan cara como el esfuerzo original.

1.3.2. El arbitraje científico

Típicamente, los gobiernos y las agencias deciden los recursos correspondientes a cada área de investigación, pero el destino específico lo determinan los comités científico-asesores. Ellos conforman uno de los ejes del llamado *peer review*. Supuestamente, este sistema canaliza los fondos hacia los investigadores con las ideas más promisorias y la mayor habilidad para concretarlas.

El *peer review* abarca también la actividad de árbitros y editores orientada a determinar si un trabajo será o no publicado [10]. A ellos toca, entre otras cosas, evaluar si las preguntas de investigación son pertinentes, si la metodología es correcta y los resultados constituyen un aporte a lo que ya se conoce, si se han hecho las referencias debidas y si la obra responde a las normas éticas consabidas.

El sistema no está exento, sin embargo, de aristas conflictivas; así lo demuestran algunos tropiezos dramáticos como el que tuvo en 1937 el trabajo de Hans Krebs sobre el ciclo del ácido cítrico, originalmente rechazado por *Nature, y* que más tarde sería la pieza clave para que el autor fuese galardonado con el premio Nobel.

1.4. El fraude científico: más crimen que castigo

En el sustrato de todo el sistema se halla la convicción generalizada de que los que se dedican a esta actividad poseen una moral diferente, más elevada, una ética mucho más acrisolada que la del resto de los profesionales. Se supone que el hecho de que la misión del científico sea, precisamente, la búsqueda de la verdad, da lugar a que jamás traicione el propósito fundamental de su quehacer *(Editorial,* 1995).

Entre los principios éticos que todo hombre de ciencia aprende desde muy temprano se hallan mandamientos tales como no omitir los créditos debidos a sus antecesores, comunicar con transparencia los procedimientos empleados, no escamotear resultados relevantes ni «recortar esquinas» para dar una falsa impresión, no publicar dos veces el mismo material sin autorización de la primera fuente, no interpretar resultados de forma tendenciosa para favorecer los intereses de quienes han proporcionado recursos financieros al investigador [11] y, desde luego, no inventar ni plagiar resultados. La violación de estos preceptos ha promovido creciente atención

[10] Para marcar su especificidad, éste es conocido como «editorial peer review».

[11] Este es un problema sumamente polémico, recientemente tratado por el conocido epidemiólogo Rothman (1991), quien reacciona airadamente contra la posibilidad de considerar que el trabajo científico financiado por la industria pueda no ser valorado como confiable.

en los últimos años; sin embargo, escapa al marco de este libro extendernos teóricamente sobre el tema; el lector interesado en fuentes recientes de análisis sobre ello puede acudir a Danforth y Schoenhoff (1992), Siegel (1991), Simmons ***et al*** (1991) o Chop y Silva (1991) y, sobre todo, a Lock y Wells (1993), Erwin, Gendin y Kleiman (1994) y Evered y Lazar (1995).

La estadística se presta con más facilidad como vehículo para consumar cierto tipo de fraudes que para otros, tal y como discuten Bross (1990) y Bailar (1976); esta última llegó a afirmar -quizás cargando un poco las tintas- que «puede correrse mayor peligro para el bienestar público por concepto de la deshonestidad estadística que producto de cualquier otra forma de deshonestidad». De modo que vale la pena hacer un breve comentario especialmente pertinente.

Obviamente, la invención de datos y el plagio directo son fechorías conscientes, propias de un delincuente. Un escandaloso ejemplo fue protagonizado muy recientemente por Roger Poisson de un hospital de Montreal en el estudio multicéntrico sobre cáncer de mama conducido por la Universidad de Pittsburgh, parte de cuyo sinuoso decursar se describe en el artículo de Altman (1994). Y otro, aun más sonado porque involucró a un premio Nobel (el biólogo molecular David Baltimore), tuvo lugar cuando la investigadora Thereza Imanishi-Kari del Instituto Tecnológico de Massachusetts fue declarada culpable en 1994 por el ***Departamento de Investigación de la Integridad*** de EE UU de 19 cargos de falsificación de datos y pruebas, incluidas en un artículo científico publicado en 1986 por la prestigiosa revista *Cell* (Weaver ***et al, 1986).*** Baltimore, que había participado como coautor del artículo en cuestión, se negó a apoyar la investigación sobre el posible fraude y desarrolló una intensa campaña de defensa de la investigadora, sancionada a la postre a 10 años de exclusión de investigaciones financiadas por agencias federales *(El* ***País,*** 1994).

Mucho más sutiles y comunes son otras conductas, como la de introducir modificaciones aparentemente irrelevantes en los resultados, o seleccionar aquellos más aceptables. Ajuicio de Broad y Wade (1982) sin embargo, la diferencia entre este «cocinado» y la invención de datos es sólo una cuestión de grado. Como «botón de muestra», que según mi percepción viene a dar la razón a estos autores, vale que nos detengamos en un caso elocuente.

Robert A. Millikan fue un médico norteamericano con enorme prestigio entre sus contemporáneos durante la primera mitad del siglo xx. Ganó 16 premios y 20 grados honorarios a lo largo de su exitosa vida. Llegó a ser nada menos que presidente de la ilustre ***Asociación Americana para el Avance de la Ciencia.*** Alrededor de 1910 sostuvo una sonada polémica con Félix Ehrenhaft, profesor de la Universidad de Viena, en torno a la existencia o no de partículas subelectrónicas. Mientras él afirmaba haber corroborado experimentalmente la inexistencia de tales subelementos, el grupo vienés testimoniaba que sus experiencias no convalidaban dicho resultado. El debate abarcó a científicos tan encumbrados como los físicos Einstein, Planck y Börn. Para refutar a Ehrenhaft, Millikan publicó una nueva serie de resultados en 1913 y los acompañó explícitamente del siguiente texto: «éste no es un gru-

po seleccionado sino que representa la totalidad de los hallazgos durante 60 días consecutivos». Holton (1978) dos décadas después de la muerte de Millikan, hurgando en sus notas experimentales originales, descubrió que las 58 observaciones presentadas en 1913 habían sido cuidadosamente elegidas de entre las 140 realizadas. La veracidad de las teorías de los vieneses fue recientemente corroborada por físicos de la Universidad de Stanford. Millikan, por su parte, había obtenido el premio Nobel en 1923, en buena medida gracias al estudio mencionado.

El fraude es un delito tan grave que cabría esperar, por lo menos, la inmediata expulsión de los autores de sus respectivos enclaves académicos. Siendo la actividad científica contemporánea altamente privilegiada por la sociedad, que costea salarios y subvenciones enormes, las conductas fraudulentas podrían y quizás deberían tener implicaciones penales. Lejos de ello, muchas veces los que denuncian fenómenos de fraude, lo hacen con temor a las consecuencias legales que pudiera traerles tal iniciativa. Eisenberg (1994) relata cómo Steven Koonin, presidente de la sección de física nuclear de la ***Sociedad Norteamericana de Física*** fue avizorado por sus abogados de no usar la palabra «fraude» en ocasión del desenmascaramiento de una de las más connotadas farsas científicas de los últimos años, protagonizada por los físicos Martin Fleischmann y Stanley Pons, quienes en 1989 habían hecho un anuncio propio de una película de ciencia ficción: la obtención de energía a partir de la fusión fría de núcleos atómicos en un tubo de ensayo. Konin hubo de limitarse a mencionar la «incompetencia y, quizás, engaño» de los impostores.

La historia de la ciencia fraudulenta muestra fehacientemente que el fenómeno del plagio -por mencionar una sola de sus expresiones- no es insólito, ni mucho menos; con harta frecuencia toma mucho tiempo para salir a la luz (Garfield, 1980) y, en no pocos casos, los delincuentes, aun habiendo quedado demostrada su culpa, pueden continuar sin demasiados sobresaltos sus carreras. Además de las incontables referencias al respecto, personalmente he podido apreciar tal impunidad, no sin estupor, en mi propio ambiente académico.

Aparte de la posibilidad de que se cometa fraude en el manejo estadístico de los datos, como en cualquier otro tratamiento del material empírico, el autoritarismo metodológico que atenaza algunas esferas de la estadística induce a formas muy sutiles de deshonestidad, tema del que nos ocuparemos más adelante.

1.5. Del poder a su verdadero sitio

Tras este panorama del ambiente en que está llamada a operar la estadística, y de la posición alcanzada en la actualidad, cabe preguntarse cuál es el papel que realmente desempeña. En la Sección 2.6 ampliaremos este tema; de momento, se resaltan algunas ideas fundamentales.

Cuando se encara un problema científico en el mundo biomédico, existe la idea de que la estadística ***tiene*** que estar presente. «Cree el aldeano presuntuoso que su

aldea es el mundo», escribió José Martí. La estadística no es «la tecnología del método científico», como proclamaran los afamados estadísticos norteamericanos Mood y Graybill(1969).

En cierta ocasión, hace más de tres lustros (Silva, 1977), suscribí alegremente esa peregrina afirmación, tan breve como egocéntrica. Tras muchos años de reflexión y experiencia práctica, he arribado a un criterio que considero muchísimo más justo, aunque menos aparatoso. El procesamiento estadístico de datos puede ser sin duda un importante componente tecnológico del método científico. Para el examen de muchos problemas de la clínica y la salud pública, su presencia puede no sólo ser iluminadora, sino imprescindible; en el campo de la epidemiología, no hay dudas de que alcanza creciente protagonismo; en las ciencias básicas, a menudo surgen preguntas que demandan su aplicación. Pero bajo ningún concepto puede creerse en su ubicuidad. Las técnicas de consenso, el análisis de sistemas, la teoría de estructuras algebraicas, la semiótica, la simulación, el reconocimiento de patrones, las ecuaciones diferenciales, las técnicas de matemáticas finitas, la lógica formal, la programación lineal, el álgebra de Boole y la teoría de conjuntos borrosos son algunos ejemplos de las alternativas o complementos que pueden comparecer en el acto metodológico para procesar e interpretar información biomédica y sanitaria.

Es cierto que muchos problemas biomédicos están colocados, por su propia naturaleza, en un marco no determinístico y, por ende, reclaman de un tratamiento probabilístico (Soriguer, 1993). Pero también lo es que una gran cantidad de resultados de enorme impacto se han conseguido sin su concurso. Para no ir más lejos, la investigación biomédica de mayor trascendencia en el siglo xx (según la opinión de Lawrence Bragg, también laureado con el premio Nobel) fue la que llevaron adelante Watson y Prick sobre la estructura de la molécula de ADN. La lectura de *La doble hélice,* el descarnado testimonio que dio Watson (1981) sobre el proceso según el cual discurrió ese estudio, revela que la estadística no representó papel alguno.

Muchas reflexiones que el lector hallará en lo sucesivo procuran ofrecer una visión más ajustada del lugar que realmente ocupa esta importante herramienta en el proceso productivo de conocimientos.

Bibliografía

Altman LK (1994). *Probe into flawed cancer studyprompts federal reforms.* The New York Times, Medical Section, 26 de abril: C2-C3.

Bailar JC (1976). *Bailar's law of data analysis.* Clinical Pharmacology Theory 20: 113-120.

Broad W, Wade N (1982). *Betrayers of the truth. Fraud and deceit in the halls of science.* Simon and Schuster, New York.

Bross ID (1990). *How to eradicate faudulent statistical methods: statisticians must do science.* Biometrics 46: 1213-1225.

Bunge M (1985). ***Seudociencia e ideología.*** Alianza, Madrid.

Castle WM (1979). ***Statistics in operution.*** Churchill Livingstone, Edinburgh.

Chop RM, Silva MC (1991). ***Scientific fraud: definitions, policies, and implications for nursing research.*** Journal of Professional Nursing 7: 166-171.

Danforth WH, Schoenhoff DM (1992). ***Fostering integrity in scientifc research.*** Academy of Medicine 67: 351-356.

Do11 R (1992). ***People of consequence. Sir Austin Bradford and theprogress of medical science.*** British Medical Journal 305: 1521-1526.

Editorial (1995). ***Shall we nim a horse?*** Lancet 345: 1585-1586.

Eisenberg A (1994). ***El arte del insulto científico.*** Investigación y Ciencia 215: 80.

El País (1994). ***Fraude científico,*** Sección Futuro, 30 de noviembre.

Erwin E, Gendin S, Kleiman L (1994). ***Ethical issues in scientifc research.*** Garland, New York.

Evered D, Lazar P (1995). ***Misconduct in medical research.*** Lancet 345: 1161-1162.

Fernández F (1991). ***La ilusión del método.*** Crítica, Barcelona.

Feyerabend P (1982). ***La ciencia en una sociedad libre, 2.ª*** ed, Siglo XXI, México DE

Feyerabend P (1981). ***Tratado contra el método.*** Tecnos, México DE

Feynman R (1985). *Surely* ***you're joking, Mr Feynman.*** Bantam Books, New York.

Garfield E (1980). ***From citation amnesia to bibliographi plagiarism.*** Currents Contents 23: 5-9.

Guttman L (1977). ***What is not what in stutistics?*** The Statistician 26: 81-107.

Hill AB (1965). ***Principios de estadística médica.*** Instituto Cubano del Libro, La Habana.

Holton G (1978). ***Subelectrons, presuppositions, and the Millikan-Ehrenhaft dispute.*** Historical Studies in the Physical Sciences 9: 166-224.

Ingenieros J (1957). ***Las fuerzas morales.*** Latino Americana, México DE

Kapitsa P (1985). ***Experimento, teoría, práctica.*** Mir, Moscú.

Lock S, Wells F (1993). ***Fraud and misconduct in scientific research.*** BMJ Publishing Group, London.

Materazzi MA (1991). ***Propuesta de prevención permanente.*** Paidós, Buenos Aires.

Medawar P (1984). Consejos ***a un joven científico.*** Fondo de Cultura Económica, México DF.

Mood AM, Graybill FA (1969). ***Zntroduction to the theory of statistics.*** McGraw Hill, New York.

Mosterín J (1995). ***Popper como filósofo de la ciencia.*** El País, 25 de enero: 34.

Pera M (1994). ***The discourses of science.*** University of Chicago Press, Illinois (traducción del castellano por C. Boltsford).

Ramón y Caja1 S (1945). ***Los tónicos de Za voluntad.*** Espasa-Calpe, Madrid.

Rothman KJ (1991). ***The ethics in research sponsorship.*** Journal of Clinical Epidemiology 44: 25 S-28 S.

Siegel HS (1991). ***Ethics in research.*** Poultry Science 70: 271-276.

Silva LC (1977). ***El razonamiento en las aplicaciones de la estadística; nociones con-***

ceptuales y terminológicas. Revista Cubana de Administración de Salud 3: 275-279.

Silva LC (1989). ***Apuntes sobre elproyecto de investigación del estudiante deposgrado.*** Educación Médica Superior 3: 29-40

Silva LC (1991). ***La formulación de problemas de investigación en salud.*** Revista Cubana de Cardiología y Cirugía Cardiovascular 5: 64-71.

Simmons RL, Polk HC, Williams B, Mavroudis C (1991). ***Misconduct and fraud in research: social and legislative issues.*** Surgery 110: l-7.

Soriguer FJC (1993). ***¿Es la clínica una ciencia?*** Díaz de Santos, Madrid.

Watson JD (1981). ***La doble hélice.*** Consejo Nacional de Ciencia y Tecnología, México DE

Weaver D, Reis MH, Albanese C, Constantini F, Baltimore D, Imanishi-Kari T (1986). ***Altered repertoire of E.endogenous immunoglobulin gen expression in transgenic mice containing a rearranged mu heavy chain gene.*** Ce11 45: 247-257.

Wright C (1961). ***La imaginación sociológica.*** Revolucionaria, La Habana.

¿La estadística al alcance de todos?

Se puede querer pensar de otro modo que como, en efecto, se piensa y trabajar lealmente por cambiar de opinión e inclusive conseguirlo. Pero lo que no se puede es confundir nuestro querer pensar de otro modo con la ficción de que ya pensamos como queremos.

JOSÉ ORTEGA Y GASSET

2.1. La s del Sr. Standard es mejor que el porcentaje

Corría el año 1977 y me hallaba en Praga cursando estudios de doctorado. Una tarde se reunió un nutrido grupo de estudiantes latinoamericanos de posgrado para realizar un paseo en barco sobre el hermoso río Moldava. En medio de una animada conversación sobre cuestiones generales, un médico que se hallaba entre los presentes comentó que, a su juicio, los porcentajes deberían enseñarse con mayor insistencia y rigor durante la enseñanza media, ya que había advertido que muchos de sus colegas tenían dificultad con tan elemental recurso [1]. Entonces, un compatriota que se hallaba en la ciudad haciendo una pasantía posgradual de otorrinolaringología procedió a ilustrarnos con la antológica afirmación siguiente: «No, no: ya el porcentaje casi no se usa; ahora lo que se usa es la desviación estándar, que es muchísimo mejor».

Pocos años después, en el vestíbulo de cierto ministerio donde tenía que tramitar un documento y mientras hacía la espera de rigor, me puse a hojear algunas revistas de índole diversa allí disponibles para hacer menos pesada la espera de los ciudadanos.

Cayó así en mis manos un pequeño folleto técnico en el que se daba cuenta de

[1] Opinión que, por cierto, compartía entonces y sigo compartiendo aún.

un estudio realizado por investigadores de dicho ministerio. Teniendo en cuenta que mi espera se dilataba, comencé a leerlo con mediano interés. Tras la introducción y el enunciado de los propósitos, los autores se internaban en la explicación del método seguido; una vez expuestos algunos procedimientos para la captación de datos, los autores pasaban a comunicar las técnicas de análisis utilizadas. Con estupor pude leer que en cierto punto se decía lo siguiente: «para el análisis estadístico, se aplicó la t de Student y la s de Standard» (sic). ¿Podrán alguna vez estos autores recuperar la coherencia y llegar a distinguir entre un símbolo y un matemático imaginario, entre un estadígrafo y una prueba de hipótesis? El mal, ¿será irreparable? No son preguntas retóricas sino legítimas dudas personales.

Ambas anécdotas, materia prima para el título de la presente sección, revelan un rasgo muy curioso de la visión que algunos tienen de la estadística como disciplina: a diferencia de lo que ocurre con las ecuaciones diferenciales, el diseño gráfico o la bibliotecología -materias que cuentan con especialistas a los que un profesional típico ha de recurrir cuando le surgen necesidades en las áreas respectivas- se ha ido cincelando la convicción de que cualquiera puede y debe dominar la estadística, como si se tratara del inglés -hoy por hoy, verdadero esperanto de la ciencia-, de la gramática básica o de los procesadores de texto.

Acaso esta convicción sea la responsable de que con extrema frecuencia tropecemos con apreciaciones estadísticas que revelan que muchos profesionales han aprendido una terminología, un lenguaje, pero no los conceptos subyacentes, fenómeno que se expresa no solamente en ambientes de tercera línea o en la prensa, sino también en medios que, aunque no especializados, presumen de poseer alto nivel académico.

2.2. El lenguaje de la ciencia estadística y la solución del diablo

Las palabras son, en ocasiones, soportes insidiosos de las ideas, herramientas en cierto sentido rudimentarias para abarcar la complejidad y riqueza del pensamiento que procuran comunicar. E inversamente, pueden evocar mucho más de lo que realmente pretendían transmitir, especialmente cuando no están insertadas en un diálogo vivo. Sin embargo, lo cierto es que toda publicación de un texto científico convierte al autor en rehén de lo que ha escrito.

Ingenieros (1957), con su habitual incisividad, expresaba:

> ***El estilo que anhela expresar la verdad se estima por su valor lógico: su claridad es transparente, sus términos precisos, su estructura crítica. Es el lenguaje de las ciencias.***

La perversión sintáctica y semántica en el caso de la estadística puede tener, como en cualquier otro campo, su origen en la deshonestidad, ya que algunos acuden a un lenguaje deliberadamente turbio para maquillar u ocultar su propio des-

concierto; pero una parte no despreciable es atribuible a cierto mimetismo en el uso ignorante de los términos.

El siguiente texto, citado por Kruskal (1978) y debido a Schneider y Mass (1975) salió nada menos que en la revista *Science.*

> *Es difícil contrastar nuestros resultados contra observaciones porque no existe un registro global de temperaturas que se remonte a 1600 y que sea estadísticamente significativo.*

No es fácil deducir qué entenderán estos autores por un «registro estadísticamente significativo».

El uso indiscriminado de los vocablos técnicos puede hallarse incluso en textos académicos que incursionan profusamente en la estadística. Por ejemplo, en Rey (1989) se lee:

> *En estadística, el estudio de un fenómeno hay que reducirlo, por su gasto y lo engorroso que supondrán un estudio de la población general, a una muestra que sea estadísticamente significativa y extensible a la población general.*

No existe nada, que yo sepa, a lo que se pueda llamar «muestra estadísticamente significativa». Pudiera pensarse que el autor quiso decir que la muestra debe ser «representativa» de la población; pero en tal caso, ¿qué quiere decir la otra condición que se exige: que la muestra sea «extensible a la población general»?

En otro punto de esta misma obra se escribe que «el riesgo es la probabilidad de que ocurra un fenómeno epidemiológico». El riesgo de que se produzca cierto suceso adverso en una población dada se suele cuantificar a través de una probabilidad: la de que un individuo tomado al azar de dicha población lo padezca. Pero la oración es totalmente equívoca, puesto que «fenómeno epidemiológico» es un concepto genérico y, a estos efectos, sumamente vago: el brote de una epidemia, el abandono generalizado de la lactancia materna y el impacto de un programa contra el consumo de grasas saturadas sobre la mortalidad son, sin duda, fenómenos epidemiológicos para los que la afirmación en cuestión carece de todo sentido.

Armijo (1994) escribe textualmente:

> *Existen numerosas definiciones de epidemiología. Entre tantas, uno puede permitirse el lujo de acuñar una más: el uso del cráneo* **para** *resolverproblemas de salud.*

Personalmente, sólo conozco una manera de usar el cráneo como tal, y es para cabecear pelotas de fútbol; por lo demás, esta «definición» parece no ameritar más comentarios.

Pueden darse muchos ejemplos de oraciones que son meras divagaciones sintácticas, de contenido críptico en el mejor de los casos, vacío en el peor. Tanto es así

que la situación suele traer a mi recuerdo el siguiente pasaje del *Fausto,* donde Goethe pone al diablo instruyendo al Estudiante:

> Mefistófeles .... ***ateneos a las palabras; entonces entráis en el templo de la certeza por la puerta segura.***
> Estudiante ***Pero la palabra debe entrañar un concepto.***
> Mefistófeles ***¡Desde luego! Pero no hay que preocuparse mucho por eso, pues precisamente allí donde faltan los conceptos*** **se** *presenta* ***una palabra adecuada, en sazón. Con palabras se puede discutir a las mil maravillas: con palabras es posible erigir un sistema.***

2.3. Cuatro sugerencias para neófitos

Se puede estipular que un investigador reciba cursos de estadística; lo que resulta imposible es decretar que aprenda, como las anécdotas e ilustraciones precedentes demuestran. No parece razonable esperar que una disciplina compleja y esquiva como la estadística sea algo que pueda estar al alcance de todos, como si se tratara de una calculadora de bolsillo. Mucho menos si no ha mediado una perseverante y reflexiva disposición de estudio unida a una práctica sostenida. A creer tal desatino contribuye el hecho de que, con el advenimiento de las computadoras personales, se nos intenta vender esa convicción junto con el ***software*** que presuntamente la respalda [2].

Significa todo lo anterior que sostengo la opinión de que un especialista en cardiología, un bioquímico o un planificador de salud deben mantenerse al margen de esta disciplina, sumergidos en la más absoluta ignorancia al respecto y dejando el asunto en manos de los estadísticos profesionales?

Desde luego, estoy muy lejos de opinar de ese modo. Creo incluso que el gran mérito de Bradford Hill, de quien según Do11 (1992) «no es exageración decir que (...) tuvo más influencia en los últimos 50 años de ciencia médica que muchos ganadores del premio Nobel en medicina», consistió en difundir la estadística dentro del personal sanitario.

El propio Hill, que no alcanzó título ni grado alguno, ni en estadística ni en medicina[3], consideraba que, sólo conociendo algo de estadística, los médicos podrán considerar a los estadísticos como genuinos interlocutores y colegas en materia de investigación.

En este terreno comparto las opiniones de una autoridad como Alvan R. Feins-

[2] Véase Capítulo 14.

[3] Valga, de paso, este dato como elemento de reflexión para quienes gustan de etiquetar a los profesionales y reducirlos a lo que estudiaron en la universidad, quizás 15 o 20 años antes, en lugar de clasificarlos según lo que saben y lo que saben hacer.

tein. En su libro ***Epidemiología clínica,*** Feinstein (198.5) se interna en el análisis de las dificultades inherentes al intercambio transdisciplinario entre la medicina y la estadística; allí sostiene que:

> ***Las*** *principales* ***técnicas*** *descriptivas pueden* ***ser comprendidas por cualquier persona inteligente que haya hecho el esfuerzo de aprender acerca de medidas de tendencia central, variabilidad y concordancia de una o dos variables. Cuando*** *tales índices* ***se extienden al ambiente multivariante, los detalles se toman complicados y frecuentemente parecen inescrutables. Los principios básicos de las ideas, sin embargo, no son demasiado difíciles de captar: Otro tanto ocurre con los principios de la inferencia estadística,*** *supuesto que* ***la persona ha hecho el*** **esfuerzo** *por* ***descubrir las*** *estrategias* ***para hacer contrastes estocásticos con métodos paramétricos y no paramétricos. Sin embargo, puesto que el personal médico está tan desacostumbrado al pensamiento numérico*** *y,* ***en virtud de que su instrucción estadística en la escuela de medicina a menudo fue poco fructífera o nula,*** *este* ***personal usualmente confronta una dificultad mucho mayor que los estadísticos para atravesar la frontera interdisciplinaria.***

A partir de las dificultades enumeradas, resumiría mis opiniones al respecto por medio de las siguientes cuatro sugerencias, dirigidas a ese profesional de la salud que carece hasta ahora de formación exitosa en la materia y se dispone a conseguirla o, simplemente, desea alcanzarla.

1. No avanzar en el estudio de un tema estadístico hasta que no domine cabalmente los precedentes

Esta regla parece tomada de las obras completas de Perogrullo, pero mi experiencia -especialmente como profesor- me inclina a reiterarla. Obviamente, el otorrinolaringólogo de la primera anécdota no sólo necesita que se le ilustre sobre la desviación estándar sino que, también y sobre todo, demanda de varias horas de explicación y entrenamiento en materia de porcentajes, a su juicio sustuibles por aquella. Los autores del trabajo publicado en forma de folleto, por su parte, harían bien en olvidarse por ahora de las pruebas de significación y ponerse a estudiar con ahínco y humildad algo de estadística descriptiva.

Actualmente no es nada raro encontrarse con un epidemiólogo, pongamos por caso, abrumado por el intento de comprender los entresijos del ***cluster analysis*** a pesar de que sólo es capaz de balbucear algunas incoherencias cuando tiene que explicar la diferencia entre el riesgo relativo y el ***odds ratio*** [4].

[4] Cuando dice las incoherencias sin balbucear, el caso es mucho más grave.

II. Esforzarse por dominar el lenguaje, el propósito, la lógica y las condiciones de aplicación de los métodos estadísticos antes que su aritmética interna

Con el surgimiento de las computadoras personales parte de esta exigencia se ha tornado virtualmente innecesaria, ya que los paquetes estadísticos se ocupan -a la manera de una caja negra- de las manipulaciones aritméticas. Sin embargo, las propias computadoras han venido a conspirar contra la demanda de que el usuario de un método lo comprenda cabalmente, ya que, si bien lo exime de manipular aritméticamente los datos, contribuye a adocenarlo (véase el Capítulo 14).

III. No dejarse amedrentar por la siempre creciente cantidad de técnicas existentes que no domina: en lo esencial, no le hacen falta

Una de las razones en que reposa este consejo es que, contra lo que muchos suponen, una altísima porción de la estadística que se usa en la «ciencia exitosa» corresponde a los métodos más simples (véase Sección 2.5).

IX Aplicar autónomamente sólo aquello que entiende totalmente y acudir a un especialista en todos los demás casos

Es cierto que no siempre se tiene un estadístico a mano a quien consultar, pero ello no hace legítimos los disparates que se cometan por cuenta propia. Obviamente, es mejor abstenerse. Una cosa puede conocerse o desconocerse; pero hay una situación peor que el desconocimiento: estar convencido de algo que es erróneo. La lógica bivalente establece que toda proposición es o bien falsa, o bien verdadera; sin embargo, no puede olvidarse el vastísimo dominio de lo que no está ni en uno ni en otro caso porque, simplemente, carece de sentido. Si a usted se le pregunta cuál es la capital de Francia y responde «Moscú», habrá cometido un error; pero si a esa pregunta responde «raíz cuadrada de 5», no es una clase de geografía lo que necesita.

2.4. No pienso, pero existo y... recibo cursos

Nada tengo contra los cursos de posgrado, ni de estadística ni de ninguna otra materia. He adquirido unos cuantos conocimientos por esa vía, aunque confieso que he aprendido mucho más impartiéndolos.

Pero también he sentido que perdía el tiempo siempre que ha concurrido al menos una de las siguientes circunstancias: cuando no me ha resultado divertido, cuando me han confundido con un receptor acrítico de información, o cuando lo que se abordaba en el curso no figuraba entre las necesidades sentidas para mi formación.

Aprovecharé esta sección para compartir algunas reflexiones sobre el ejercicio de la docencia en estadística, que posee algunas especificidades en el mundo complejo y harto cambiante de la pedagogía.

2.4.1. La enseñanza estadística prostituida

En la actualidad, como apunta Von Glasersfeld (1987), muy pocos dudan de la valía del llamado *constructivismo,* recurso pedagógico según el cual el alumno «construye su propio conocimiento» mediante una interacción activa frente al planteamiento de problemas; tal enfoque se plantea como una alternativa a la práctica de embutir al alumno con los conocimientos que ha de asimilar.

La comunicación mecánica de información técnica poco tiene que ver, en efecto, con una verdadera y fecunda gestión del proceso docente. Pero ese mecanicismo es mucho más inquietante aun cuando se produce en ancas de lo que yo llamaría *transferencia comercializada de conocimientos.*

La enseñanza por correspondencia de dibujo técnico, corte y costura o taquigrafía es un negocio de vieja data. Ignoro cuán eficiente puede resultar para el alumno semejante procedimiento, pero intuyo fuertemente que, en caso de que funcione, a ello contribuye que el carácter de lo que se debe aprender se reduce a cierto repertorio de habilidades formales y actos de destreza.

Pero la estadística aplicada es una herramienta con escaso sentido si no se inscribe dentro de un universo conceptual flexible, que contemple, además de los algoritmos, el componente heurístico imprescindible para que dicha herramienta funcione fructíferamente. No descarto que se pueda instruir por vía postal a un alumno acerca de la fórmula para aplicar la *prueba de Wilcoxon, o* sobre qué teclas del *SPSS* ha de apretar, pero abrigo serias dudas acerca de que se pueda insuflar a vuelta de correo (una vez recibido el cheque correspondiente) el espíritu creativo y dúctil que demanda la aplicación fecunda de esta disciplina.

En torno al tema, vale la pena recordar dos reflexiones debidas a sendos investigadores y pedagogos eminentes, ambos ganadores del premio Nobel: Piotr Kapitza y Bernardo Houssay.

El físico ruso comentaba (Kapitza, 1985):

> ***Al asistir a los exámenes dados por los estudiantes de postgrado, con frecuencia observé que el profesorado de la enseñanza superior valora más al estudiante que sabe que al que*** entiende y ***la ciencia necesita, antes que nada, gente que*** entienda.

En tanto que el fisiólogo argentino (Houssay, 1955) observaba que:

> ***La educación pasiva y con vista a calificaciones o exámenes, acostumbra a la sumisión intelectual y al deseo de congraciarse con*** *los* ***profesores, perdiendo toda autonomía y sacrificando el sagrado afán de la veracidad.***

No desdeño los recursos ágiles que faciliten procesos docentes obstaculizados por la distancia y la falta de tiempo, pero sospecho del espíritu empresarial como protagonista de la enseñanza estadística. Sé de casos bochornosos en que, como condición para que un candidato pueda inscribirse en un cursillo a distancia de estadística, se le exige que curse un entrenamiento de cinco días destinados a aprender un sistema operativo y el MS-Word. Si el sujeto ya domina este procesador de textos o prefiere estudiarlo pausadamente por su cuenta[5], puede ser exonerado... ¡siempre que lo pague!: un verdadero atraco intelectual y económico.

Aunque entiendo todo proceso docente como algo mutuamente enriquecedor, y sigo creyendo que el educador genuino no ha tenido ni tendrá nunca la cacería obscena del dinero como incentivo fundamental, mi razón fundamental para renegar de la transferencia tarifada de conocimientos no es, sin embargo, de principios. Se fundamenta en algo totalmente ajeno a consideraciones éticas: una experiencia de muchos años con estudiantes de posgrado me ha permitido constatar, simplemente, que la mayoría de los alumnos que pasaron por el proceso de aprendizaje postal no saben y, que la mayoría de los pocos que saben, no entienden.

Ésta es una constatación empírica de algo teóricamente esperable, ya que tal procedimiento educacional transgrede las más importantes premisas teóricas de la enseñanza de la estadística consignadas por Garfield (1995) en un artículo altamente especializado en el tema.

Para que el lector conforme su propia composición de lugar, a continuación enumero y comento las cuatro primeras.

1. Los estudiantes aprenden construyendo el conocimiento

Este principio se basa en los estudios de Resnick (1987) y Von Glasserfeld (1987), quienes demuestran que ignorar, subvalorar o desaprobar las ideas propias de los alumnos solo consigue dejarlas esencialmente intactas. Creo que si, en lugar de decretar la suplantación de unas nociones por otras, se procede según una búsqueda creativa por parte de cada alumno, las nuevas nociones adquieren carta de ciudadanía en el educando.

2. **Los estudiantes aprenden mediante actividades que los involucren activamente**

Tal propósito se consigue fundamentalmente con trabajo en grupos para la solución colectiva de problemas en un ambiente de debate y cooperación,

[5] Personalmente creo que recibir un curso formal para el aprendizaje de un procesador de texto es innecesario, ya que mucho más eficiente es encararlo de forma autodidacta (véase Sección 14).

tal y como recomienda el National Research Council (1989) y subraya Silver (1990).

3. **Los estudiantes aprenden a hacer bien sólo aquello que practican**

Como señala la American Association for the Advancement of Science (1989), es imposible que un alumno aprenda a pensar críticamente y a comunicarse productivamente si el proceso no incluye ciclos de obtención de resultados en un entorno práctico efectivo.

4. **Los profesores no deben subestimar la dificultad de los alumnos para comprender los conceptos básicos de probabilidades y estadística**

Como se ha demostrado (Shaughnessy, 1992) los alumnos suelen enfrentarse a esas nociones con convicciones e intuiciones propias, usualmente muy arraigadas y con frecuencia erróneas. La simple comunicación de resultados correctos, si no se ventilan en un ambiente de contradicción, no consigue modificar los prejuicios iniciales del alumno.

Con frecuencia un alumno contesta correctamente a una pregunta porque sabe qué es lo que se quiere que responda, pero mantiene los conceptos equivocados. En un estudio desarrollado al efecto, Konold (1989) testimonia, por ejemplo, cómo algunos alumnos identifican como igualmente probables varias secuencias diferentes de caras y cruces pero, cuando se les pregunta cuál de ellas es la que con menos probabilidad puede ocurrir, seleccionan una u otra sin pestañear.

Los famosos «cursos de estadística a distancia», expresión de educación pasiva si la hay, me recuerdan, quiéralo o no, a la relación sexual prostituida como sucedáneo del vínculo amoroso: el acto mecánico es esencialmente el mismo; el entorno emotivo, el ejercicio de la sensibilidad y, sobre todo, la comunicación, sólo pueden ser integrales en el segundo caso.

2.4.2. Los peligros de la media aritmética y de la inercia

Ahora bien, independientemente de que la comunicación entre profesor y alumno se ejerza o no directamente, el riesgo de mecanicismo conceptual está siempre presente. Garfield (1995), por ejemplo, advierte claramente:

Los profesores deben experimentar con diferentes enfoques y actividades, así como monitorizar los resultados, no solo mediante exámenes convencionales sino, también y

sobre todo, oyendo a los estudiantes y evaluando la información que refleja diferentes aspectos de su aprendizaje.

Consideremos un ejemplo sumamente elemental: el aprendizaje de las llamadas ***medidas de tendencia central.***

Para describir un fenómeno, provenga de la investigación científica o de la realidad cotidiana, es común que los datos asociados a él se condensen mediante un número que sintetice y refleje su esencia. La media aritmética -resultado de dividir la suma de todos las observaciones entre el número de sumandos- es el índice que más frecuentemente se usa con esa finalidad. Así, por ejemplo, la nota promedio de las diferentes asignaturas cursadas por un estudiante durante un año académico puede usarse para resumir y caracterizar su actuación.

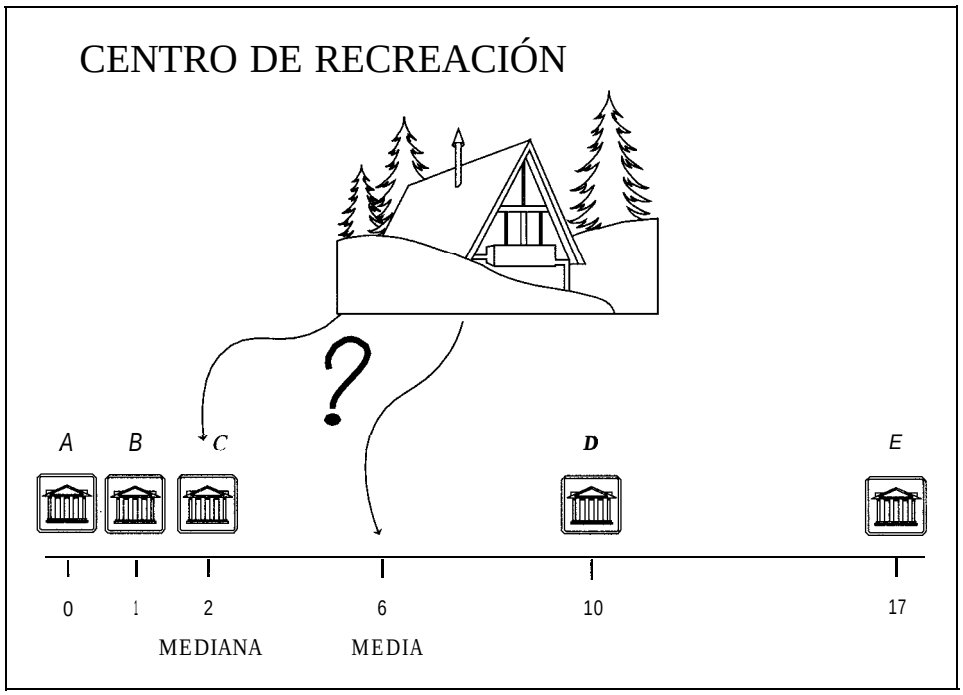
Si bien la capacidad descriptiva de este indicador resulta virtualmente universal, su aplicación puede ser inconveniente y ocasionalmente absurda. Es bien conocido, sin embargo, que ésta no es la única «medida de tendencia central»; existen otras, a saber: la mediana, la moda, la media geométrica y la media armónica. Así se comunica en cualquier texto básico y así lo informarán todos los profesores de estadística descriptiva. Lo que es menos frecuente es que se intente hacer entender con claridad qué significado real tienen esas otras medidas.

Los problemas siguientes, que bien pudieran servir para construir ejercicios llamados a ser colectivamente discutidos y resueltos por los alumnos, resumen diversas situaciones en las cuales es alguna de esas otras medidas, y no la media aritmética, la que se ajusta al problema [6]. Deliberadamente he seleccionado uno de los temas estadísticos más básicos para ilustrar un camino didáctico mucho más rico que la simple y tradicional enumeración de fórmulas y denominaciones seguida de ejercitación formal; es decir, un recurso para conseguir que el alumno ***entienda y*** no solo que ***sepa.***

Un problema de distancias

Supongamos que a lo largo de una carretera rectilínea se hallan cinco escuelas y se quiere construir un centro de recreación para uso común del alumnado. El problema radica en determinar el mejor punto de ubicación de dicho centro, de manera que la suma de las distancias a las cinco escuelas sea mínima. Para ello se sitúa una escala y un origen que, por mera conveniencia, supondremos que se halla en el lugar ocupado por la primera escuela (véase Figura 2.1).

[6] En cada caso, la pregunta podría ser: ¿por qué la media aritmética no da solución al problema planteado?



Escuela	A	B	C	D	E	Total
Distancia entre la escuela y el origen	0	1	2	10	17	30

Figura 2.1 Ubicación del centro de recreación en un punto de la carretera.

Un primer impulso podría llevarnos a calcular la distancia media de las escuelas al origen, igual a $\frac{30}{5} = 6$, y colocar allí el centro deportivo. Nótese que la posición relativa de dicho punto respecto a las escuelas es independiente del punto seleccionado como origen.

La solución no está dada, sin embargo, por ese valor promedio sino por la ***mediana*** de la serie, definida como ***cualquier número que, miembro o no del conjunto de números, ni exceda ni sea sobrepasado por más de la mitad de las observaciones.*** A partir de esta definición no es difícil advertir que si el total de las observaciones es impar -como en nuestro ejemplo- entonces la mediana es única y coincide con el punto ubicado en el medio de la serie (supuesto que ésta ha sido ordenada de menor a mayor o viceversa): En este caso la mediana es 2.

En la Tabla 2.1 se resumen las distancias de las cinco escuelas del ejemplo, tanto al valor promedio como a la mediana.

Tabla 2.1. ***Distancias desde las escuelas hasta la media y hasta la mediana***

	Escuelas					
	A	B	C	D	E	Total
Distancia entre las escuelas y el valor promedio (6)	6	5	4	4	11	30
Distancia entre las escuelas y el valor mediano (2)	2	1	0	8	15	26

Puede demostrarse que cualquier otra ubicación del centro de recreación haría que la suma de las cinco distancias a dicho punto excediese de ***26.***

El valor de moda

Al hacer un cálculo numérico más o menos laborioso todos hemos repetido el proceso varias veces para corroborar la corrección del cómputo. Supongamos que, tras haber reiterado 7 veces el proceso de sumar ***20*** números de 3 cifras, se obtienen los siguientes resultados:

6038 6035 6035 6041 6035 6042 6040

¿Cuál de ellos sería más razonable sindicar como correcto? El sentido común no nos permitiría optar, ciertamente, ni por la media aritmética ni por la mediana de los 7 resultados (que en este caso coinciden ambas con 6038) sino por ***el número de la serie que más se repite.*** Tal resultado, 6035 en este caso, es lo que se conoce como ***moda*** del conjunto. Personalmente, no conozco ninguna otra aplicación práctica de esta medida de tendencia central.

El crecimiento geométrico

Una comunidad ha aumentado su población a lo largo de 50 años según los ritmos reflejados en la Tabla 2.2, en los que el número de individuos registrados cada decenio se expresa como porcentaje de la población censada el decenio precedente.

Esto quiere decir que la población de 1950 supera la de 1940 en un 5% y así sucesivamente, hasta el año 1990 en que la de 1980 se incrementó en un 95%.

Se quiere caracterizar el aumento anual durante el quinquenio; o sea, se trata de hallar un porcentaje constante de incremento anual que, aplicado sucesivamente, coincida con la población alcanzada al concluir el periodo.

Tabla 2.2. Incrementosporcentuales de población a lo largo de cinco décadas

Primer año del decenio	1950	1960	1970	1980	1990
Porcentaje respecto al decenio anterior	105	105	110	110	195

La media aritmética de los cinco porcentajes (12.5) no da, tampoco en este caso, la respuesta adecuada. En efecto, si suponemos que había 10.000 individuos en 1940, con un aumento anual constante del 25%, la población hubiera evolucionado de 10.000 a 12.500, de 12.500 a 15.625 y así sucesivamente, hasta alcanzar aproximadamente 30.518 unidades en 1990.

Sin embargo, al aplicar los incrementos de la Tabla 2.2, se observa que la secuencia de los cinco volúmenes poblacionales tuvo que ser:

$$10.500 \Rightarrow 11.025 \Rightarrow 12.128 \Rightarrow 13.340 \Rightarrow 26.014$$

de modo que el monto poblacional en 1990 ascendería a 26.014 individuos y no a 30.518.

La respuesta adecuada se obtiene sólo mediante la ***media geométrica*** de los porcentajes anuales, ya que ésta no es otra cosa que ***aquel valor que,*** *multiplicado* ***por sí mismo tantas veces como datos haya, resulte igual al producto de todos ellos (o*** sea, la media geométrica es la raíz n-ésima del producto de los ***n*** datos).

En este caso se corrobora que la media geométrica de los cinco porcentajes es 121,07 y puede constatarse que, ciertamente, un aumento fijo del 21,07%, año tras año, hace que los 10.000 sujetos de 1990 se conviertan en los 26.014 correspondientes a 1990.

La media armónica y las velocidades

El conductor de una moto se desplaza de A a B, puntos que distan entre sí 30 km, a razón de 30km/h. ¿A qué velocidad debe recorrer el camino de regreso para que la velocidad media de todo el trayecto -es decir, la distancia total recorrida entre el tiempo invertido para ello- sea igual a 60 km/h?

Pudiera pensarse que la velocidad necesaria para el segundo tramo habría de ser 90 km/h, ya que 60 es el promedio de 30 y 90. Esto es, sin embargo, erróneo.

En efecto: a estas velocidades, los primeros 30 km se recorren en 1 hora exacta, en tanto que los otros 30 km insumen sólo 20 minutos (un tercio de hora). La velocidad media para el recorrido entero no sería 60 km/h, sino que, según la definición, equivaldría ala razón entre la distancia total (60 km) y el tiempo total invertido(4/3 de hora):

$$\frac{60 \text{ km}}{4/3 \text{ h}} = 45 \text{ km/h}$$

Para hacer el trayecto de ida y vuelta a razón de 60 km/h se necesita 1 hora, algo imposible, ya que hubo de invertirse exactamente 1 hora sólo para el viaje de ida. Es decir, no existe velocidad alguna para este segundo tramo, por alta que sea, que permita elevar la velocidad media a 60 km/h.

No es difícil convencerse de que la velocidad media de todo el recorrido está dada por la ***media armónica*** de las velocidades de ida y de vuelta, definida c ***el inverso de la media aritmética de sus inversos.***

Si se conduce a razón de 30 y 90 km/h, según la regla de la media armónica, la velocidad media general será:

$$\frac{1}{\frac{1}{2}\left(\frac{1}{30}+\frac{1}{90}\right)} = 45$$

lo cual corrobora el cálculo previo. Análogamente, es fácil convencerse de que la ecuación:

$$\frac{1}{\frac{1}{2}\left(\frac{1}{30}+\frac{1}{X}\right)} = 60$$

no tiene solución para X, tal y como ya habíamos analizado.

2.5. Fábula estadística con dos moralejas

De las secciones precedentes podría deducirse que los planteamientos disparatados son patrimonio exclusivo de los «intrusos». No es cierto. A modo de ilustración de cuán escarpadas pueden ser las laderas de la estadística, incluso para los «alpinistas» profesionales, recreo a continuación un relato tomado de Hunter (1981).

Cuando un estadístico, que acababa de obtener su doctorado, se presentó a su puesto de trabajo en una planta química, fue de inmediato sometido al rito habitual de iniciación. Se le proveyó de una enorme colección de datos obtenidos a lo largo de los años, procedentes de diferentes áreas de la planta y se le pidió que examinara tal información y sacara algo de ella.

Actuando en la cuerda de cierta tradición estadística, examinó la producción mediante gigantescos análisis de regresión, primero de un modo, luego de otro, y de otro. Consiguió elaborar un voluminoso informe. Claramente escrito como estaba, el informe cubrió las expectativas de su supervisor. Su segmento protagónico era una lista de variables ordenadas según su importancia, tal y como habían quedado tras su laborioso tratamiento estadístico. El ordenamiento había sido realizado de acuerdo con la magnitud de la significación estadística asociada a cada coeficiente de regresión.

A partir de la favorable impresión general causada por el informe, se le pidió que ofreciera un seminario al equipo técnico en pleno de la planta para explicar sus hallazgos. La pieza clave de su exposición, en calidad de resumen del trabajo, era la lista de las variables ordenadas por orden de importancia.

Inició su conferencia y todo iba muy bien hasta que llegó a la última línea de su transparencia final. Entonces dijo: «Y, como se aprecia, la variable menos importante es la cantidad de agua presente». En ese punto observó estupefacto cómo la audiencia estallaba unánimemente en carcajadas.

Lo que todo el mundo, salvo él, sabía, era que la cantidad de agua presente distaba de ser la variable menos importante. Era, de hecho y sin duda alguna, la más importante, puesto que cualquier cantidad de agua por encima de un umbral muy reducido produciría una gloriosa y devastadora explosión del complejo.

En efecto, este peligro configuraba el centro de la preocupación de muchos de los ingenieros que allí laboraban. Un refinado sistema de monitores controlaba día y noche el contenido de agua en muchos puntos de la planta. Numerosas alarmas estaban previstas para anunciar un ocasional nivel de agua próximo a aquel umbral de tolerancia. Por razones de seguridad, entonces, el nivel de agua era mantenido muy cercano a cero, hecho que se traducía en que el recorrido de esa particular variable regresora estaba fuertemente contraído, independientemente de los niveles de producción que se hubieran alcanzado. Consecuentemente, el análisis de regresión no podía descubrir signos de su importancia.

Hunter extrae dos moralejas de esta historia:

1. Se debe aprender tanto como razonablemente se pueda sobre la materia general y el entorno específico del cual proceden los datos inherentes al problema que se encara.
2. La correlación medida en un estudio observacional no implica causalidad, y los estadísticos deben ser muy cuidadosos con el lenguaje cuando se refieren a esos temas.

Lo cierto es que los grandes desarrollos y los paquetes informaticos sofisticados no sólo no garantizan la corrección del análisis, como se verá en el Capítulo 14, sino que pueden oscurecerlo.

2.6. Paseo estadístico por el mundo de la ciencia exitosa

2.6.1. Práctica y tecnología estadística

A diferencia de los resultados tecnológicos, los conocimientos científicos no se patentan ni constituyen una mercancía: se publican. Consecuentemente, es posible conocer qué métodos estadísticos se utilizan ***realmente*** en lo que llamaremo ***investigación biomédica exitosa.***

Pronunciarse sobre si un trabajo merece realmente el calificativo de *exitoso* puede ser en principio polémico. Sin embargo, no hay dudas de que aquellos que consiguen ser publicados en revistas de alto prestigio internacional admiten, aunque solo fuera por ese hecho, tal calificación. Se ha dicho (Spinak, 1996), y es cierto, que no todos los trabajos acogidos en este tipo de revistas son necesariamente buenos, ni son siempre mediocres los que terminan en una de segundo o tercer orden (o, incluso, en un archivo); pero creo que aquellos, en el peor de los casos, siempre alimentarán la circulación de ideas, así como que nadie podrá calificar a los últimos como «exitosos».

La actividad investigativa necesitada de métodos cuantitativos para el análisis de sus observaciones, en todas las ciencias de la salud, crece incesantemente. Ello haría pensar que la preparación del profesional de la salud en disciplinas matemáticas debe ser mucho más versátil y abarcadora que la necesitada en épocas precedentes. Sin embargo, lo cierto es que el repertorio de métodos realmente necesarios para llevar adelante la mayoría de las investigaciones de excelencia es más bien reducido.

La primera investigación orientada a valorar el uso de la estadística en las revistas biomédicas fue publicada a mediados de los años 70 por Feinstein (1974); casi 10 años después, otro trabajo, basado en artículos contenidos en New ***England Joumal of Medicine y*** debido a Emerson y Colditz (1983), aportó un método de análisis altamente influyente para el esfuerzo posterior en esta dirección. A partir de este momento se revitalizó el interés por el tema. Entre 1985 y 1992 se realizaron diversos estudios sobre los métodos estadísticos usados en áreas específicas; entre ellos pueden citarse los de Avram ***et al.*** (1985), Hokanson, Luttman y Weiss (1986), Hokanson ***et al.*** (1987a), Hokanson ***et al.*** (1987b), Rosenfeld y Rockette (1991) y Juzych ***et al.*** (1992); o en revistas especializadas, como reflejan los trabajos de Fromm y Snyder (1986) y de Cruess (1989).

Casi todos estos trabajos procuran responder a la misma pregunta: qué técnicas estadísticas necesitaría dominar un lector para comprender los artículos publicados en las áreas 0 revistas respectivas.

Dado que lo que se conocía sobre este tema parecía limitarse al registro de las técnicas más usadas (información, por otra parte, no actualizada), Silva, Pérez y Cuéllar (1995) se propusieron responder algunas preguntas que, expresadas en síntesis, fueron: ¿Qué métodos son objetivamente los usados en la investigación biomédica exitosa actual? ¿Qué grado de complejidad tienen los métodos que demanda esta producción [7]?

En el fondo del debate se trata de evaluar si lo ideal es que un investigador típico tenga un conocimiento extenso, aunque quizás no muy acabado, de la tecnología estadística; o si lo mejor sería que dominara recursos sencillos, pero de modo profundo y completo.

Tras un examen detenido, se decidió elegir como fuentes básicas para el análisis

[7] Otra área abordada en este estudio se comenta en la Sección 2.6.2.

dos revistas punteras en las áreas de clínica y epidemiología: ***The New England Journal of Medicine (NEJM) y American Joumal of Epidemiology (AJE),*** publicaciones de primera línea mundial y de notable impacto en sus esferas respectivas.

The New England Joumal of Medicine, que se publica semanalmente y aborda un amplio espectro de especialidades clínicas y quirúrgicas ocupó, según el Journal Citation Report (1990) el décimo lugar dentro de las 25 revistas más citadas de todas las ramas de la ciencia en 1990, con 78.767 citas en la literatura médica de ese año. Paralelamente, NEJM estuvo ese mismo año entre las revistas de más alto factor de impacto: octavo lugar en general y primero dentro del área de medicina general e interna.

American Joumal of Epidemiology, cuya frecuencia de publicación era mensual en esa etapa, contiene artículos en su casi totalidad de epidemiología. Su factor de impacto la coloca en el segundo lugar dentro de su área, sólo superada en el año 1990 por Epidemiologic Reviews.

Se examinaron todos los artículos publicados en estas dos revistas entre los años 1986 y 1990, ambos incluidos: 1.341 trabajos correspondientes a NEJM y 1.045 a AJE. Cada artículo se clasificó de acuerdo con el grado de complejidad de las técnicas estadísticas utilizadas según lo que el propio trabajo comunica.

Cada técnica o método estadístico se clasificó dentro de uno de los siguientes niveles de complejidad estadística:

Nivel 1: Procedimientos de estadística descriptiva.
Nivel 2: Técnicas convencionales univariadas o de muestreo.
Nivel 3: Postestratificación, regresión logística, regresión múltiple o análisis de supervivencia.
Nivel 4: Otras técnicas multivariadas, recursos inferenciales avanzados o técnicas de alta especificidad.

Se definió que un artículo tiene cierto nivel si hace uso de al menos un método clasificado en ese nivel, pero no de métodos correspondientes a un nivel superior. Por ejemplo, si cierto artículo usa tres métodos: uno de Nivel 1 y dos de Nivel 4, se dice que ese artículo es de Nivel 4. Si no hace uso de técnica estadística alguna, se clasificó como de Nivel 0. En un anexo del trabajo citado se enumeran detalladamente los métodos que quedan comprendidos en cada nivel. La Tabla 2.3 ofrece la distribución de los artículos según nivel para ambas revistas.

En primer lugar se aprecia que los artículos que prescinden totalmente de la estadística constituyen sólo la sexta parte de los publicados en NEJM, y apenas un 4% para AJE. Parecen no quedar dudas de que este segmento de la ciencia exitosa tiene en la tecnología estadística de análisis una fuente metodológica muy relevante.

Resulta llamativo, sin embargo, que para NEJM, la revista biomédica más famosa del mundo, el mayor porcentaje de los artículos (dos de cada tres), o bien no hacen uso de estadística, o bien acuden sólo a técnicas estadísticas elementales.

Tabla 2.3. ***Distribución porcentual y acumulada de artículos según nivel en NEJM y AJE, 19861990***

Nivel	N.º de artículos		Porcentaje		Porc. acumulado	
	NEJM	AJE	NEJM	AJE	NEJM	AJE
0	226	41	16,9	3,9	16,9	3,9
1	80	71	6,0	6,8	22,9	10,7
2	552	275	41,1	26,3	64,0	37,0
3	252	530	18,8	50,8	82,8	87,8
4	231	128	17,2	12,2	100,0	100,0
Total	1.341	1.045	100,0	100,0		

Solamente uno de cada seis trabajos exigieron recursos de Nivel 4. Resulta claro que la complejidad de los métodos predominantes es reducida.

En el estudio realizado con artículos publicados por NEJM 15 años atrás, Emerson y Colditz (1983) habían determinado que el 58% de los trabajos revisados no hacía uso de la estadística, o se circunscribía a su vertiente descriptiva; sólo el 4% usaba en aquella época técnicas pertenecientes al Nivel 4. La apreciable diferencia entre este panorama y el que se observa actualmente (los artículos de Nivel 0 pasaron de ser el 58% al 17%, y los del Nivel 4 se elevan de 4% a 23%) da indicios de una transición hacia la complejidad.

Sin embargo, otras revistas menos connotadas exhiben, para el mismo periodo de nuestro estudio, un patrón similar al de NEJM en 1977. En efecto, en 1986, el análisis que realizan Fromm y Snyder (1986) de los procederes estadísticos, usados en ***Journal of Family Practice,*** arrojó que el 46% de los artículos no utiliza método estadístico alguno, el 13% se auxilia sólo de técnicas descriptivas, y el 25% de tablas de contingencia, en tanto que sólo un pequeño porcentaje de los trabajos hace uso de recursos más avanzados.

De los 1.045 artículos examinados en AJE, la mitad demanda técnicas multivariadas elementales (postestratificación y regresión) o para el análisis de supervivencia (Nivel 3). Casi el 90%, sin embargo, puede prescindir de técnicas multivariadas avanzadas o relativamente complejas tales como ***MANOVA,*** análisis de clusters o de componentes principales (de Nivel 4).

A todo esto debe añadirse que en el trabajo original el estudio se extendió a otras dos revistas, representantes ya no de la «ciencia exitosa» sino de lo que yo llamaría «ciencia decorosa». Se trata de revistas serias y rigurosas, aunque de segundo

nivel en cuanto a trascendencia y factor de impacto, en parte debido, en un caso, a que se publica en castellano y, en el otro, a que geográficamente pertenece a la periferia de las influencias. Se trata, respectivamente, del ***Boletín de la Oficina Sanitaria Panamericana*** (BOSP) y ***del Indian Joumal of Experimental Biology*** (IJEB).

Los resultados en ambos casos apuntan en la misma dirección que los ya mencionados para NEJM y AJE, pero de manera mucho más marcada. Por ejemplo, los artículos de Nivel 4 constituyen el 0,6% entre los 111 publicados en ese quinquenio por el BOSP y el 2,0% de los 390 aparecidos en IJEB.

En general, una vez constatada la limitada presencia de procedimientos de nivel elevado, el rasgo más llamativo es la ausencia ***total*** y ***absoluta*** durante todo el quinquenio y en las cuatro revistas, de una amplia gama de recursos estadísticos que en muchos ambientes se reputan como de gran trascendencia. Tal es el caso, entre muchos otros, de los siguientes procedimientos:

- Análisis de componentes principales.
- Análisis espectral.
- Correlación canónica.
- Técnicas de autocorrelación.
- Regresión múltiple no lineal.
- Análisis factorial.
- Análisis «Box-Jenkins».
- Prueba de Wald-Wolfowitz.
- Prueba de Moses.
- Diagnóstico de la regresión.

Con respecto a tal constatación, e independientemente de cuál sea la razón, una primera y obvia consideración salta a la vista: existe una gran cantidad de recursos estadísticos que los autores de una parte relevante de la ciencia exitosa no usan. Si bien no es posible sacar conclusiones definitivas acerca de las causas [8], procede considerar algunas posibles explicaciones.

1.” Los objetivos de los trabajos pueden ser alcanzados sin necesidad de aplicar dichas técnicas.

Tratándose de trabajos publicados en revistas especialmente celosas de que las contribuciones tengan preguntas transparentemente formuladas y de que ellas hayan sido respondidas de manera debida, es muy verosímil que ésta sea una de las explicaciones.

[8] Ello exigiría acciones tales como entrevistas a los autores y quizás un examen más profundo de los propósitos de los trabajos.

2.º **Los autores desconocen las técnicas en cuestión.**

Esa posibilidad viene abonada por el hecho de que las que no se usan son precisamente las más complejas y recientes. Pero, en tal caso, también cabe pensar que los autores no sintieron la necesidad de usarlas. Es decir, si los problemas abordados no hubieran podido ser resueltos con las técnicas que conocían, más tarde o más temprano los autores habrían sentido la necesidad de aplicar otras, y hubieran tomado contacto entonces con ellas.

3.º **El perfil de las revistas elegidas es tal que los problemas allí tratados no demandan de las técnicas aludidas.**

Si bien ello es posible, debe tenerse en cuenta que NEJM es una revista de perfil sumamente ancho, no tributario de una preferencia metodológica específica. Por lo demás, muchas de las técnicas que no fueron usadas (tales como el análisis discriminante o la regresión múltiple no lineal) distan de ser recursos de alta especificidad disciplinaria, como pudiera ocurrir con el análisis basado en el ***método probit*** (usado, por cierto, en algunos trabajos de NEJM), típico - y casi exclusivo- de los estudios farmacológicos.

Aunque las tres explicaciones pudieran ser plausibles, la primera parece ser la más razonable.

2.6.2. Uso y desuso de la literatura especializada

La publicación en estadística, como en cualquier otra disciplina, salvo excepciones, puede dividirse en dos grandes áreas: el de las revistas especializadas, que responde a los resultados más vivaces y en cierto sentido más provisionales y discutibles, y el área de los libros de texto, con resultados organizados ya de modo sistemático, que han pasado en lo esencial por el tamiz de la polémica y la práctica.

Otro propósito del trabajo ya mencionado de Silva, Pérez y Cuéllar (1995) fue el de sondear en qué medida la «ciencia exitosa» hacía uso de las fuentes metodológicas ***originales*** en estadística.

Para ello se trabajó, al igual que antes, con la totalidad de los artículos publicados en ***The New England Journal of Medicine y American Joumal of Epidemiology*** entre 1986 y 1990, pero concentrando ahora el interés en las referencias bibliográficas a los trabajos procedentes de revistas especializadas en estadística.

De los 1.341 artículos publicados por NEJM, el 82% *no* ***cita a revista estadística alguna,*** y lo mismo ocurre con el 59% de los 1.045 artículos de AJE. Como promedio hay aproximadamente una cita de este tipo por cada cuatro artículos de NEJM

y una por cada dos de AJE. Sin embargo, lo más llamativo es que la edad media de las citas a publicaciones especializadas en estadística en las dos revistas examinadas es sumamente elevada: 19,3 años para NEJM y 16,3 años para AJE, cifras cuatro veces mayores que la edad promedio del resto de las citas. Este resultado es sumamente elocuente habida cuenta de la actual dinámica de acceso a la literatura publicada, que es casi instantánea. De hecho, lo que ocurre es que hay un manojo de artículos clásicos que se citan recurrentemente y que muy probablemente no siempre han sido realmente leídos por quienes ahora los citan.

En síntesis, la bibliografía estadística utilizada por los autores para realizar sus análisis se caracteriza por prescindir, salvo contadas excepciones, de la literatura reciente. Dado que el número de artículos que hacen algún uso de las revistas especializadas ya es de por sí escaso, en general el uso de los métodos recientemente publicados es casi nulo.

Para examinar más incisivamente estos datos, cabe tener en cuenta que las revistas especializadas en estadística (que pasan de 40) no han dejado de publicar novedades en su área en los últimos años. Por ejemplo, el número de artículos publicados entre 1983 y 1990 por dos de las revistas estadísticas más prestigiosas y de corte más aplicativo, *Statistics in Medicine y Biometrics,* ascendió según el MEDLINE, a 705 y 541 respectivamente. Un cómputo rápido e informal permite estimar que una revista de estadística puede haber publicado anualmente unos 75 artículos de fondo como promedio. Si consideramos sólo 20 revistas de esta índole (para circunscribirnos a las de naturaleza más aplicativa), en los últimos tres lustros podrían haberse publicado unos 22.000 trabajos. NEJM solo ha hecho 68 citas correspondientes a ese lapso. Si se repara en que esas 68 citas corresponden a muchos menos trabajos, ya que algunos se citan reiteradamente, la influencia casi nula de esa producción en la ciencia exitosa se evidencia de manera meridiana.

El hecho de que la inmensa mayoría de las investigaciones exitosas actuales prescinda de las publicaciones especializadas en metodología estadística para consolidar sus análisis, unido al de que la poca literatura estadística que se cita es francamente anticuada, subraya que los avances en el conocimiento clínico y epidemiológico no han dependido medularmente de una gran actualización bibliográfica en esa materia.

Esto no puede, desde luego, considerarse una tendencia positiva; nuestros propios resultados mostraron que, en la medida que la investigación biomédica contemporánea hace uso de técnicas estadísticas más complejas, utiliza con mayor frecuencia la literatura estadística. Pero también revela que tal falta de actualización no es óbice para que se consiga una producción de muy alto nivel.

2.7. Impresiones después del viaje

¿Qué se infiere de estos resultados para la práctica investigativa actual? Los

resultados descritos, precisamente porque atañen a la investigación de excelencia, parecen convalidar la presunción inicial: no es necesario ni estar al tanto de los últimos avances ni dominar recursos estadísticos altamente especializados para hacer buena ciencia.

En relación con este tema, no se ha hecho un estudio profundo y completo. No me ha sido fácil, al menos, hallar rastros de otros esfuerzos de peso orientados en esa dirección.

Resulta muy interesante, sin embargo, reparar en el modo en que dos figuras de prestigiosas escuelas de salud pública (de la de Harvard uno, de la de Columbia el otro) sostienen una curiosa sesión de esgrima al respecto. El ***American Journal of Public Health*** acogió una polémica entre Walker (1986) y Fleiss (1986) en torno al tratamiento estadístico de ciertos temas epidemiológicos.

Walker hizo allí esta tajante afirmación:

> ***Nunca se ha hecho una observación epidemiológica importante que no haya*** *podido* ***serplanteada claramente en*** *unos pocos* ***cuadros o tablas de datos numéricos en bruto y algunas estadísticas descriptivas sencillas*** [9]***.***

Fleiss responde airadamente diciendo que eso no es cierto y a renglón seguido pone un ejemplo (uno sólo) en que -según él- tal simplificación es imposible. A mi juicio la afirmación de Walker es desproporcionada y, como toda aseveración absoluta hecha por un ser que no es omnisciente, puede ser, en uno u otro momento, refutada. Pero lo realmente significativo es la tibieza de la respuesta. Es como si presenciáramos el siguiente diálogo:

- Todos los ingleses son altaneros.
- Eso es completamente falso: conozco a uno que no lo es.

Uno se quedaría con la clara sensación de que, estadísticamente hablando, todos los ingleses son altaneros.

Se puede considerar, en síntesis, que la investigación biomédica exitosa (al menos una parte relevante de ella) se limita con elevada frecuencia al uso de recursos estadísticos de muy antigua data, que no sobrepasan las técnicas multivariadas elementales (postestratificación y regresión múltiple).

De ello no se deduce ni mucho menos que las técnicas más avanzadas deban ser desechadas, pero es obvio que sería más fructífero que muchos investigadores, que no son estadísticos profesionales, se abstuvieran de profundizar en ellas hasta que no consiguiesen una comprensión acabada de las más elementales. Con ellas bastaría no sólo para asimilar el grueso de los resultados de alto nivel que se publican

[9] Llama la atención el tono cortante y categórico de la afirmación, tan poco frecuente en el ambiente epidemiológico sajón y que parece más propio de la grandilocuencia latina.

sino, incluso, para producirlos.

Bibliografía

American Association for the Advancement of Science (1989). ***Science for all americans,*** Washington DC.

Armijo R (1994). ***Epidemiología básica en atenciónprimaria de la salud.*** Díaz de Santos, Madrid.

Avram MJ, Shanks CA, Dikes MHM, Ronai AK, Stiers WM (1985). ***Statistical methods in anesthesia articles: an evaluation of two American joumals during two six-moth periods.*** Anesthesiology and Analgesics 64:607- 611.

Cruess DF (1989). ***Review of use of statistics in the American Journal of Tropical Medicine and Hygiene for January-December 1988.*** American Journal of Tropical Medicine Hygiene 41: 619-626.

Do11 R (1992). ***People of consequence: Sir Austin Bradford and theprogress of medical science.*** British Medical Journal 305: 1521-1526.

Emerson JD, Colditz GA (1983). ***Use of statistical analysis in the New England Journal of Medicine.*** The New England Journal of Medicine 309: 709-713.

Feinstein AR (1974). ***Clinical biostatistics: a survey of the statistical procedures in general medical joumals.*** Clinical Pharmacology Therapeutic 15: 97-107.

Feinstein AR (1985). Clinical Epidemiology. ***The Architecture of Clinical Research.*** W.B. Saunders Company, Philadelphia.

Fleiss JL (1986). ***Significance tests have a role in epidemiologic research: reactions to A.M. Walker*** (Different Views) American Journal of Public Health 76: 559-560. Tradución al castellano publicada en el ***Boletin de la Oficina Sanitaria Panamericana 1993; 115: 155-159.***

Fromm BS, Snyder VL (1986). ***Research design and statisticalprocedures used in the Joumal of Family Practice.*** Journal of Family Practice 23: 564-566.

Garfield J (1995). ***How students learn statistics.*** International Statistical Review 63: 25-34.

Hokanson JA, Stiernberg CM, McCraken MS, Quinn FB (1987a). ***The reporting of statistical techniques in otolaryngology joumals.*** Archives of Otolaryngology, Head and Neck Surgery 113: 45-50.

Hokanson JA, Ladoulis CT, Quinn FB, Bienkowski AC (1987b). ***Statistical techniques reported in pathology joumals during 1983-1985.*** Archives of Pathology 111: 202-207.

Hokanson JA, Luttman DJ, Weiss GB (1986). ***Frequency and diversity of use of statistical techniques in oncology joumals.*** Cancer Treatment Reports 70: 589-594.

Houssay BA (1955). ***La investigación científica.*** Columba, Buenos Aires.

Hunter WG (1981). ***Six statistical tules.*** The Statistician 30: 107- 117.

Journal Citation Report (1990). ***A bibliometric analysis of science joumals in the ISI database.*** Institute for Scientific Information.

Juzych MS, Shin DH, Seyedsadr M, Siegner SW, Juzych LA (1992). ***Statistical tech-***

niques in ophthalmic journals. Archives of Ophthalmology 110: 1225-1229.

Ingenieros J (1957). ***Las fuerzas morales.*** Latino Americana, México DE

Kapitsa P (1985). ***Experimento, teoría, práctica.*** Mir, Moscú.

Konold C (1989). ***An outbreak of belief in independence?*** En: Maher C, Goldin G, Davis B, editores. ***Proceedings of the llh Annual Meeting of the North American Chaper of the Intemational Group for the Psychology of Mathematics Education*** Rutgers University Press, Rutgers, Volume 2: 203-209.

Kruskal W (1978). ***Formulas, numbers, words:statistics in prose.*** En: Fiske D (editor) New ***Directions for Methodology of Social and Behavioral Science, 1981*** Jossey-Bass, San Francisco.

National Research Council (1989). ***Everybody counts: A report to the nation on the future of mathematics education.*** Academy Press, Washington.

Resnick L (1987). ***Education and learning to think,*** National Research Council, Washington DC.

Rey J (1989). ***Método epidemiológico y salud de la comunidad.*** McGraw Hill Interamericana de España, Madrid.

Rosenfeld RM, Rockette HE (1991). ***Biostatistics in otolaryngology journals.*** Archives of Otolaryngology, Head and Neck Surgery 117: 1172-1176.

Schneider SH, Mass C (1975). ***Volcanic dust, sunspots, and temperature trends.*** Science 190: 741-746.

Shaughnessy JM (1992). ***Research in probability and satistics: Rejlections and directions.*** En: Grouws DA (editor) ***Handbook of Research on Mathematics Teaching and Leaming*** Macmillan, New York: 465-494.

Silva LC, Pérez C, Cuéllar 1 (1995). ***Uso de métodos estadísticos en la investigación publicada en dos revistas médicas con alto factor de impacto.*** Gaceta Sanitaria 9:189-195.

Silver E (1990). ***Contributions to Research to Practice: Applying Findings, Methods and Perspectives.*** En: Cooney T (editor) ***Teaching and Leaming Mathematics in the 1990s*** Reston: NCTM; l-ll.

Spinak E (1996). ***Los análisis cuantitativos de la literatura y su validez para juzgar la producción latinoamericana.*** Boletín de la Oficina Sanitaria Panamericana 120:139-145.

Von Glasersfeld E (1987). ***Learning as a constructive activity.*** En: Janvier C (editor) ***Problems of representation in the teaching and leaming of mathematics*** Lawrence Erlbaum Associates, Hillsdale: 3-17.

Walker AM (1986). ***Reporting the results of epidemiologic studies.*** American Journal of Public Health 76: 556-558. Citado según la traducción al castellano «Como publicar los resultados de los estudios epidemiológicos» publicada en el Boletín de la Oficina Panamericana 1993; 115: 148-154.

3

Escalas e indicadores

Para quien sabe leer las cifras, ellas difunden una luz astral sobre la esfera del conocimiento. Para usar de ellas importa disponer de muchos conocimientos laterales y poseer un cierto olfato moral, parecido a la segunda potencia del sentido común.

BALDOMERO SANÍN CANO

El investigador involucrado en problemas sociomédicos, clínicos o epidemiológicos, opera con abstracciones tales como colesterolemia, calidad de vida, ansiedad, edad, tabaquismo, capacidad diagnóstica, depresión, mortalidad materna o eficiencia de la gestión hospitalaria. Los conceptos pueden corresponder a un individuo, a una comunidad o a una institución, entidades que -ya en el plano operativo- se denominan ***unidades de análisis.***

Las preguntas de investigación siempre están formuladas a nivel conceptual y necesitan de tales nociones para ser expresadas de manera inteligible. Las respuestas que se obtengan, parciales o no, necesariamente habrán de resumirse, a la postre, en los mismos términos, usando las mismas categorías y a la luz del mismo marco teórico de la pregunta formulada.

Muchos problemas (no todos, véase Sección 8.4) exigirán que la fase empírica del estudio -aquella en que se obtiene información de la población o universo al que conciernen las preguntas- se verifique a través de mediciones formales de esos conceptos. A tales mediciones se las denomina «variables» del problema [1]. El capítulo que nos ocupa se destina a discutir algunos de los problemas asociados a las escalas que se usan para conformar las variables y a reflexionar sobre una herramienta clave para el trabajo de salubristas y epidemiólogos: los indicadores.

[1] No pocas veces el acto metodológico demanda incorporar otras variables que no se derivan directamente de la formulación original, como son aquellas que se incluyen con la exclusiva finalidad de ser controladas.

3.1. Operacionalización de variables

El proceso que permite hacer el tránsito que parte del concepto y desemboca en el recurso cuantitativo (o cualitativo) con que se mide (o clasifica) dicho concepto se denomina ***operacionalización de variables.*** El término proviene de que se trata, precisamente, de llevar la noción desde el plano teórico al operativo, y concierne al acto de medición del grado (o la forma) en que el concepto se expresa en una unidad de análisis específica.

Consideremos un ejemplo simple y familiar: la edad. Formalmente, la edad de un sujeto habría de definirse como el número de vueltas que ha dado la tierra en torno al sol, más la fracción del recorrido realizado desde la última vuelta completa hasta el instante en que se hace la indagación. Sin embargo, en la mayoría de los estudios [2], la edad se operacionaliza tomando simplemente el número de años cumplidos (vueltas completas). Éste es un ejemplo en que la operacionalización es muy sencilla y directa.

Pero si el estudio, por ejemplo, se preguntara en qué medida la edad de un médico de atención primaria se relaciona con el grado de conocimientos farmacológicos que posee, la noción ***conocimientos farmacológicos*** demandaría de una operacionalización mucho más elaborada; para concretarla, probablemente haya que elaborar un cuestionario y, luego, una variable sintética compuesta por unas cuantas variables intermedias. Una propuesta de cómo llevar adelante tal operacionalización mediante una variable sintética se hallará en la Sección 4.5.2, dentro del capítulo dedicado específicamente a ese tema.

Stevens (1946) realizó un inestimable aporte taxonómico en relación con las escalas de medición de las variables. Su enfoque, que hoy parece universalmente aceptado, caracteriza las diversas escalas posibles en las que pueden medirse las variables y examina sus propiedades.

No se hará aquí una reproducción detallada de la teoría de Stevens; nos detendremos, más bien, en el examen de algunas incomprensiones o errores que se han ido radicando a ese respecto en la práctica investigativa.

3.2. Variables nominales en el contexto cuantitativo

Cuando se ha definido un conjunto de categorías mutuamente excluyentes que no guardan entre sí relación de orden, y la «medición» consiste en ubicar la unidad de análisis en una de dichas clases, entonces se dice que la noción subyacente se ha operacionalizado a nivel ***nominal.***

[2] Ocasionalmente, como ocurre en algunos estudios auxológicos con niños, o de neonatología, se trabaja con un grado mayor de precisión: la llamada edad decimal.

Muchas técnicas estadísticas están concebidas específicamente para el manejo de este tipo de escalas (por ejemplo, la que se asocia al llamado ***modelo logarítmico lineal).*** Otros procedimientos, como los que hacen uso de los modelos de regresión (lineal, no lineal o logístico), no contemplan en principio la incorporación de variables nominales. Puesto que el entorno en que se inscribe dicho problema es cuantitativo, resulta inaceptable incluir de manera directa a las variables cualitativas entre las independientes.

Sin embargo, es indiscutible que rasgos tales como la raza, la religión que se profesa o el grupo sanguíneo, variables que se expresan a través de escalas nominales, podrían ser modificadores de la variable dependiente. Por tanto, con frecuencia su incorporación en el análisis de regresión resulta conveniente y hasta insoslayable.

Un primer impulso podría conducir a la asignación de números a las diversas categorías de una variable nominal. Por ejemplo, para la religión, se podría valorar una asignación como la siguiente:

Religión	Valor de la variable
Católico	1
Protestante	2
Otra	3
Ateo	4

Pero tal «solución» es de todo punto inadmisible, ya que la regresión no es más que un algoritmo y, por tanto, actuaría como si se tratara de una variable cuantitativa (como lo haría, por ejemplo, con la variable ***número de hijos),*** y haría interpretaciones absurdas como que ser ateo es cuatro veces mayor que ser católico. A este problema se le han dado distintas soluciones. Hosmer y Lemeshow (1989) manejan varias posibilidades, en especial la que parece más simple y natural, objeto de la siguiente explicación.

Supongamos que la variable nominal opera con C categorías. Deben crearse entonces C – *1* variables dicotómicas (cada una de las cuales puede tomar el valor 1 o el valor 0); se trata de las llamadas ***variables dummy*** [3], a las que denotaremos Z_1, Z_2, ... Z_{c-1}. El vector integrado por estas C – *1* variables contendrá la misma cantidad de información que la variable nominal y habrá de suplirla en el contexto de la regresión. A cada categoría o clase de la variable nominal le corresponde un vector (Z_1, Z_2, ..., Z_{c-1}) que la identifica, cuyas coordenadas se definen de la manera siguiente:

[3] En algunos textos se ha traducido esta expresión como «variables de diseño». Aquí se asumirá el término «dummy», acuñado así en las publicaciones estadísticas en inglés.

Si el sujeto pertenece a la primera categoría, a la que suele llamársele ***categoría de referencia,*** entonces las ***C-l*** variables dummy valen 0 : $Z_1 = Z_2 = \ldots = Z_{c-1} = 0$. Si el sujeto se halla en la segunda categoría, entonces para ese individuo $Z_1 = 1$, y las restantes C – 2 variables valen 0; para aquellos individuos que están en la tercera categoría, Z_2 vale *1* y las otras variables toman el valor 0; y así sucesivamente, hasta llegar a la última categoría, para la cual Z_{c-1} ***es*** la única que vale *1* y el resto 0.

En el ejemplo de las religiones se tendría la siguiente correspondencia entre las categorías y los valores de las variables dummy:

Variable nominal (Religión)	Z_1	Z_2	Z_3
Católico	0	0	0
Protestante	1	0	0
Otra	0	1	0
Ateo	0	0	1

Una explicación detallada de la interpretación que debe hacerse de los coeficientes de regresión correspondientes a estas C – *1* variables puede hallarse en Silva (1995) para el caso de la regresión logística, y en Draper y Smith (1981) para el caso de la regresión lineal múltiple. Pero lo que realmente interesa subrayar es que, cuando se hace esta maniobra, ***existe*** una interpretación; cuando se tratan las categorías como si fueran números, no hay manera de recuperar conceptualmente los resultados de la regresión, de manera que en tal caso la interpretación que se haga de los resultados será casi con seguridad disparatada.

3.3. Las variables ordinales: crónica de una prohibición arbitraria

Una situación que merece comentario más detenido es la de las ***variables ordinales,*** que se definen del mismo modo que las nominales, pero con la singularidad de que las categorías guardan un orden de precedencia.

Tal es el caso, por ejemplo, cuando se reclama la opinión de un ciudadano sobre cierta medida y se le pide que se ubique a sí mismo en una de las cinco categorías siguientes: TOTALMENTE DE ACUERDO, DE ACUERDO, ME RESULTA INDIFERENTE, EN DESACUERDO, EN TOTAL DESACUERDO.

Un viejo error de concepto relacionado con las escalas ordinales de medición amenaza con tomar carta de ciudadanía en el mundo de las aplicaciones. Se trata de una excelente ilustración de cómo una prohibición, establecida de manera esen-

cialmente arbitraria e infundamentada, surge como propuesta, luego evoluciona a la condición de pauta metodológica que se va heredando sucesivamente, hasta convertirse en regla cuyos efectos inhibidores consiguen neutralizar el razonamiento.

Todo empezó cuando, en relación con las escales ordinales, el propio Stevens (1951) escribió [4]:

> ***Las medidas estadísticas usuales, tales como*** la media o la desviación estándar no deberían, en rigor, usarse con escalas ordinales ***ya que ellas suponen el conocimiento de algo más que el mero orden jerárquico relativo de los datos. Por otra parte, este uso «ilegal> de la estadística está avalado por una suerte de ley pragmática: resulta muchas veces objetivamente útil y fecundo. Pero, si bien la negación a ultranza de estos recursos no sería aconsejable, cabe señalar que cuando los intervalos sucesivos de la escala son de diferentes tamaños, ellos*** *pueden* ***ciertamente conducir a conclusiones falaces. (El subrayado es mío, LCS.)***

De ese texto se deducen tres elementos centrales: a) calcular medias o varianzas para este tipo de variables sería erróneo, b) hacerlo puede, no obstante, producir dividendos de interés, y c) ese error no es preocupante si los intervalos son de igual tamaño.

La primera objeción que surge al examinar estas consideraciones concierne a su carácter contradictorio, ya que no es fácil admitir (salvo que asumamos un pensamiento dogmático) que un proceder «útil y fecundo» pueda ser erróneo; según mi comprensión, un procedimiento solo puede ser calificado como «erróneo» si es inútil, o cuando conduce a la obtención de respuestas equivocadas.

Por otra parte, el concepto de *tamaño del intervalo* resulta vago en este contexto, ya que la escala ordinal, por definición, no reposa en una métrica que permita la medición de un *tamaño* como tal. El sentido común y el valor semántico que se le atribuya a las denominaciones de las clases parecen ser, por tanto, los únicos árbitros disponibles para conocer en qué medida las clases «equidistan» entre sí o no. Naturalmente, si se parte de una escala ilógica, es verosímil que se saquen conclusiones falaces, pero no como resultado de una manipulación inadecuada de los datos sino como consecuencia de una ruptura con el sentido común, cuya incidencia en el proceso cognoscitivo hubo de ser *anterior* a cualquier tratamiento estadístico.

Podrían, por ejemplo, atribuirse los valores 1, 2, 3, 4 y 5 a las categorías respectivas del ejemplo arriba mencionado en que se solicitaba del interrogado que expresara su grado de adherencia a una disposición, y tratar en lo sucesivo a esta variable como una dimensión cuantitativa más. Esto parte del supuesto de que la «distancia» entre cualquier pareja de categorías contiguas (por ejemplo, entre DE ACUERDO

[4] Declaraciones similares aparecen en otros artículos suyos de esos años.

y ME ES INDIFERENTE) es la misma. Se trataría de una decisión operativa cuya validez, como he dicho, reposaría en el sentido común del investigador. Consecuentemente, la recomendación de Stevens podría reducirse a lo siguiente:

No ASIGNAR VALORES A TONTAS Y A LOCAS A LAS DIFERENTES CLASES DE UNA ESCALA ORDINAL; PERO, SI SE HA SOPESADO ADECUADAMENTE QUÉ ASIGNACIÓN NUMÉRICA DARLES, ENTONCES MANEJAR LOS NÚMEROS DEL MODO QUE PAREZCA MÁS FRUCTÍFERO.

Con ella es imposible no concordar, pero el modo elegido por él para decirlo ha dado lugar al dogma conceptual al que me referiré más adelante. Mi opinión es que expresiones de ***audacia metodológica*** como esta no deben ser desestimuladas (salvo que haya objeciones fundadas) con el fin de no crear, o consolidar, un clima de parálisis, acaso más pernicioso que las consecuencias derivadas de admitir supuestos informales como el que subyace en la propuesta.

Por otra parte, en algunos contextos se ha corroborado, incluso, que los resultados finales suelen ser básicamente los mismos, independientemente de cuáles sean los valores que se atribuyan a las diferentes categorías. Una discusión especialmente transparente y persuasiva en esta dirección puede hallarse en el trabajo de Moses, Emerson y Hosseini (1984). Otro análisis detallado se desarrolla en el libro de Streiner y Norman (1989) sobre construcción de escalas para la medición en salud; los autores concluyen sus reflexiones sobre el tema afirmando que, desde un punto de vista pragmático, no cabe esperar sesgos de importancia como consecuencia del tratamiento cuantitativo de datos ordinales. Incidentalmente, cabe apuntar en este contexto que la literatura recoge intentos de establecer el valor numérico que en la práctica diaria se otorga subjetivamente a determinadas expresiones. Por ejemplo, Beyth-Maron (1982) realizó una investigación en la que se indica a los participantes que hagan su traslación cuantitativa para diversas expresiones inglesas de probabilidad: ***not likely, very low chance, poor chance, doubtful, possible, close to certain,*** etc. Se detectó una notable variabilidad en cuanto al «valor» que los individuos atribuyen a las mismas expresiones. Un trabajo similar y con análogos resultados habían realizado 15 años antes Lichtenstein y Newman (1967).

Ahora bien, volviendo a la sugerencia de Stevens, varios autores de textos parecen haber encontrado una oportunidad para hacer prohibiciones taxativas en su nombre. Tal es el caso, por ejemplo del muy influyente libro de Siegel (1956) y, mucho más recientemente, de los trabajos de Twaite y Monroe (1979) y Townsend y Ashby (1984).

Siegel, concretamente, escribió en su texto original lo siguiente:

> *Aunque* ***parezca excesivamente insistente, quisiera el que esto escribe recalcar que las pruebas estadísticas paramétricas, que usan las medias y las desviaciones estándares (en las que hay que efectuar operaciones aritméticas sobre*** *los* ***puntajes originales),*** no deben

> usarse ***con datos de una escala ordinal. Cuando se emplean técnicas paramétricas de inferencia estadística con tales datos, todas las decisiones acerca de las hipótesis son*** dudosas. ***En vista de que la mayoría de las mediciones hechas por los científicos de la conducta culminan en escalas ordinales, este punto merece*** especial énfasis. ***(Los subrayados son míos, LCS.)***

Una cosa es la recomendación de que, al hacer la traslación de las escalas ordinales al terreno numérico, no se violente el sentido común; otra muy diferente que ***seprohiba*** usar un método porque la escala *original* no cumpla tal o cual condición. Adviértase, de paso, que Siegel nos comunica con ***especial énfasis*** que el tratamiento numérico ***no debe*** usarse; pero aparentemente considera que no necesitamos saber por qué, como puede corroborarse tras un escrutinio de su libro.

Como señaló Lord (1953) en un breve pero esclarecedor trabajo crítico: «Los números no conocen de dónde vienen». Quiere decir: tome usted toda las precauciones del caso y asuma su responsabilidad al hacer la cuantificación pero, una vez realizada, está libre de dar a los datos el tratamiento estadístico que le parezca mejor. En la misma línea razona Boneau (1961) cuando opina que «los números asignados mediante el proceso de medición constituyen un problema de medición, no un problema estadístico».

En sucesivos trabajos realizados a lo largo de i26 años!, John Gaito, estadístico de la York University, ha venido insistiendo en la necesidad de sacudirse esta absurda coyunda, como se aprecia en la secuencia de sus trabajos: Gaito (1960) Gaito (1980) Gaito (1986), Gaito y Yokubynas (1986).

La discusión precedente refleja un tipo de úcases metodológicos cuyo origen no siempre es fácil rastrear. Por ejemplo, en más de una ocasión he oído a sociólogos y psicólogos afirmar que «el número de categorías en una escala ordinal debe ser impar a fin de que haya un punto de equilibrio que sirva de referente al que responde». Según esa ley, por ejemplo, un juego de 4 alternativas (digamos EXCELENTE, BUENO, MALO, PÉSIMO) dejaría sin «punto de equilibrio» al encuestado. No creo sinceramente que éste se vaya a derrumbar por la demanda de que elija una de esas 4 posibilidades. A ese respecto Streiner y Norman (1989) escriben: «Un número par de categorías fuerza a los interrogados a colocarse de uno o de otro lado. Pero no hay una regla absoluta; dependiendo de las necesidades de cada investigación particular puede o no ser deseable que se permita una posición neutral».

3.4. Índice de posición para escalas ordinales

A continuación se expone un ejemplo que cumple varias funciones: permite apreciar la utilidad del manejo cuantitativo de datos ordinales, introduce un índice que en ciertos contextos puede resultar útil ***Indice de Posición)***, *e* ilustra una manera de analizar los resultados producidos por un tipo muy particular de preguntas.

Supongamos que se tiene una muestra de n individuos que se han evaluado a través de una escala ordinal compuesta por las clases $A_1, A_2, \ldots, A_k$ de manera que en la clase A_i se ubican n_i de ellos; de modo que $\sum_{i=1}^{k} n_i = n$.

Atribuyamos arbitrariamente los valores o puntajes 1, 2, ... **k** a las clases A_1, $A_2, \ldots, A_k$ respectivamente. Se tiene entonces que el acumulado de «puntos» por los n sujetos es:

$$1\,n_1 + 2\,n_2 + \ldots + k\,n_k = \sum_{i=1}^{k} i\,n_i$$

El promedio de puntos por sujeto es:

$$M = \frac{1}{n} \sum_{i=1}^{k} i\,n_i$$

Es fácil ver que M se mueve necesariamente entre 1 y k. El valor mínimo se alcanza sólo cuando todos los sujetos caen en la clase A_1 **y** el máximo cuando todos están en **A,.** Es decir, $M = 1$ si y sólo si:

$$n_1 = n \quad ; \quad n_2 = n_3 = \ldots = n_k = 0$$

mientras que (M = ***k) se*** obtiene sólo si se cumple que:

$$n_1 = n_2 = \ldots = n_{k-1} = 0 \quad ; \quad n_k = n$$

En el afán de disponer de un índice, llamémosle I, que cuantifique la posición global de la muestra respecto de la escala ordinal ***sin necesidad de tener en cuenta el número de clases que la componen,*** resulta natural plantearse una transformación adecuada de M. Consideremos $I = a + bM$ (transformación lineal de M) de modo que $0 \leq I \leq 1$ con $I = 0$ cuando toda la muestra esté ubicada en el extremo inferior, A_1, ***e I = 1*** cuando todos los sujetos estén en **A,.**

Esto se reduce a encontrar ***a y b*** que satisfagan el siguiente sistema de ecuaciones

$$a + b\,k = l$$
$$a + b\,l = O$$

Su solución conduce a lo que llamaremos ***Índice de Posición:***

$$I = \frac{M - 1}{k - 1} \qquad [3.1]$$

Así definido, el índice tiene algunas propiedades que lo convierten en un instrumento valioso para el análisis [5].

Por ejemplo, es lógico desear que cuando la distribución de los sujetos en la escala sea simétrica respecto del centro (si k es impar) o de las clases centrales (si k es par), entonces el índice tome el valor $I = 0{,}5$. Un ejemplo en que el índice vale 0,5 por ese concepto se tiene con la siguiente configuración:
para $\boldsymbol{n} = 110$ sujetos en $k = 6$ categorías.

A_1	A_2	A_3	A_4	A_5	A_6
10	25	30	30	25	10

Es fácil demostrar que la condición es, en efecto, suficiente para obtener $I = 0{,}5$, pero no necesaria. Puede ocurrir que cierta combinación de frecuencias a uno y otro lado de la clase central produzca el «equilibrio» y haga que I valga 0,5, no obstante la presencia de asimetría. Ello ocurre en el ejemplo siguiente ($k = 5$ y $n = 135$):

A_1	A_2	A_3	A_4	A_5
10	30	50	40	5

Consideremos ahora una aplicación del Índice de Posición. Supongamos que se quieren evaluar cuatro alternativas para la organización de un servicio según el criterio de 100 sujetos que integran una muestra de profesionales a quienes se consulta. A cada individuo se le plantea lo siguiente:

> A continuación se mencionan cuatro modelos organizativos. Colóquelos en orden según su preferencia. Asigne la calificación 1 al mejor, 2 al siguiente, etc.
>
> _____ Modelo A _____ Modelo B _____ Modelo C _____ Modelo D

Realizada la encuesta, los resultados se resumen en un cuadro que contiene las cuatro distribuciones de frecuencias:

[5] Nótese que cuando sólo hay dos categorías (variable dicotómica), I no es otra cosa que la proporción de sujetos que caen en la segunda categoría.

	Lugar				
	1.º	2.º	3.º	4.º	Total
Modelo A	7	57	31	5	100
Modelo B	64	5	17	14	100
Modelo C	2	16	2	80	100
Modelo D	27	22	50	1	100

Esto quiere decir que, por ejemplo, 80 de los 100 profesionales interrogados colocaron el Modelo C como el peor, y sólo 2 como el mejor. ¿Cuál es el Modelo más favorecido por la opinión de los individuos consultados? A partir de la tabla, no es fácil en principio pronunciarse; podría pensarse que el preferido es el *B,* por haber sido clasificado como el mejor por el mayor número de personas; pero el *D* tiene el mérito de ser el que menos veces fue considerado como peor. En fin, lo natural es procurarse alguna medida de resumen que ayude a la adopción de un criterio objetivo; si se usa el índice de posición, se tienen los siguientes resultados:

$$I_B = 0.270 \qquad I_D = 0.416 \qquad I_A = 0.446 \qquad I_c = 0.867$$

El índice permite apreciar que el Modelo *B es* el mejor evaluado (con *I* más cercano al extremo izquierdo de la escala, que en este caso es el que refleja «lo mejor») y que aventaja al *D* mucho más de lo que el propio *D* aventaja al *A;* estos dos últimos están próximos entre sí pero con considerable ventaja sobre el C, que quedó en último lugar.

Debe señalarse que, en este caso, el mismo análisis puede llevarse adelante sin necesidad de usar *I* sino trabajando directamente con la media *M,* número que se mueve entre 1 y 4. La utilidad específica del Índice de Posición es más clara si se trabaja con preguntas para las que rigen diferentes valores de *k,* ya que permite comparar la posición de la muestra en lo que concierne a una pregunta respecto de la que tiene en lo que concierne a otra.

Por ejemplo, imaginemos que dos investigadores independientes han formulado en respectivos estudios a cada encuestado lo siguiente: ¿En qué medida está Ud. satisfecho de la atención que le brinda el sistema público de salud?

Imaginemos que el *Investigador A* encuesta a 312 sujetos y que, al ofrecer 5 posibles respuestas, obtuvo los siguientes resultados:

Muy satisfecho	Satisfecho	Dudoso	Insatisfecho	Muy insatisfecho
101	73	30	81	29

El ***Investigador B,*** por su parte ofreció 4 posibles respuestas a sus 511 interrogados y los resultados fueron:

Satisfecho	Esencialmente satisfecho	Con algunas razones para estar insatisfecho	Insatisfecho
125	207	108	71

Mientras M-que asciende a ***2,58 y 2,24*** para los casos *A* y ***B*** respectivamente- no basta para pronunciarse por tener diferentes cotas superiores, los valores de *I* ***(***0,395 y 0,413 respectivamente) permiten apreciar que la población investigada por ***B*** estaba más insatisfecha que la encuestada por A.

Un ejemplo en el que usan recursos similares al descrito puede hallarse en el trabajo de Zdravomishov e Iadov (1975).

3.5. **Índices e indicadores**

En la sección precedente hemos visto un índice específicamente basado en una variable medida en escala ordinal. Se trata de un ejemplo de lo que también se conoce como un ***indicador.***

Curiosamente, no es nada fácil hallar definiciones de este concepto de cotidiana aplicación. El conocido diccionario de epidemiología de Last (1982) no incluye ***indicator*** como tal entre sus entradas; sí incorpora sin embargo ***health indicator, o*** sea ***indicador de salud.***

Genéricamente, un ***indicador*** es una construcción teórica concebida para ser aplicada a un colectivo (población o muestra) y producir un número por conducto del cual se procura cuantificar algún concepto o noción asociada a ese colectivo.

Cabe aclarar que, si bien el término «indicador» se remite usualmente a una agrupación (tal es el caso de muchas de las tasas usadas en demografía y salud pública), el término «índice» suele usarse en ambos sentidos: para aludir a alguna medida resumen grupa1 (como el ***Índice de Desarrollo Humano*** [6]***)*** y también para aludir a una magnitud construida a partir de otras, a menudo, una simple razón de dos varia-

[6] Véase Sección 4.5.3.

bles, la segunda de las cuales es tomada como base respecto de la cual se observa el comportamiento de la primera. En este segundo caso el índice se aplica a una unidad de análisis; tal es caso del ***Índice Cintura Cadera,*** o del ***Índice de Masa Muscular,*** utilizados en antropometría.

Los indicadores constituyen uno de los recursos metodológicos básicos de la salud pública y la epidemiología. En esta categoría caen, como es bien sabido, construcciones tales como las tasas específicas de mortalidad, el producto interno bruto per cápita o la tasa de abortos inducidos por cada 1.000 mujeres en edad reproductiva.

El uso de indicadores resulta especialmente propicio para generar confusión entre forma y contenido: aveces se olvida que su cómputo tampoco es una finalidad en sí, y que ellos son simplemente ***intermediarios operativos*** para alcanzar objetivos previamente especificados.

Una expresión de esta confusión se hace tangible en algunos trabajos -especialmente en tesis de licenciatura, maestría o especialización médica- que colocan ***resultados*** en el sitio destinado a las ***conclusiones.*** No es raro hallar «conclusiones» del tipo siguiente: «El 45% de los encuestados expresó satisfacción con la atención recibida». El porcentaje de satisfechos, un indicador quizás bautizable como ***Indice de Satisfacción,*** debe servir para ***valorar*** o ***enjuiciar*** la situación de la satisfacción y hacerlo de modo objetivo. Es decir, a partir de que sólo el 45% de los individuos estaban satisfechos, el investigador podría concluir algo así como que «se observan claros signos de insatisfacción» o «si bien la situación no es favorable, el grado de satisfacción es mucho mayor del que cabría esperar, habida cuenta de...»; pero el dato en sí mismo no puede suplir a una verdadera conclusión; ésta habrá de constituir el resultado de algún proceso interpretativo, mediante el cual se haya conseguido dar un salto cualitativo respecto de los números propiamente dichos.

3.6 Indicadores positivos y negativos

Recientemente, Abelin (1986) y Hartman (1993) han hecho convocatorias a la construcción y propuesta de indicadores ***positivos*** de salud y bienestar. Debe repararse en que, ateniéndonos a una conceptualización amplia del término ***tecnología*** definido por Bunge (1985) como el «diseño y planeación que utiliza conocimientos científicos con el fin de controlar cosas o procesos naturales, de diseñar artefactos o procesos, o de concebir operaciones de manera racional», los indicadores constituyen una de sus expresiones.

En un mundo crecientemente cautivado por la tecnología se corre el riesgo de suplir las interpretaciones más fecundas e incisivas de la realidad por meras aplicaciones tecnológicas. El concepto de ***indicador positivo de salud (o bienestar)*** resulta equívoco. Su introducción en el contexto de la salud pública parece responder al afán de llamar la atención sobre aquellos puntos nodales del desarrollo, que revelan

modificaciones en aspectos favorables de la salud a nivel comunitario; sería, supuestamente, la herramienta acorde con el tránsito que va de una visión medicalizada a una percepción social de la salud pública, y a la vez coherente con la tendencia a hacer de la promoción de salud un componente protagónico de la gestión de salud, tal y como reclama la Carta de Ottawa para la Promoción de la Salud (1986).

Sin embargo, esta etiqueta ***para un instrumento de medición*** -ya que un indicador no es más que un instrumento para medir o reflejar una situación, o aquilatar un cambio- es, por lo menos, discutible. En principio haría pensar que tendría sentido adjetivar de ese modo, por ejemplo, a un esfigmomanómetro; resulta chocante, sin embargo, que este equipo pudiera ser calificado como un aparato «positivo» o «negativo». Él sirve para medir un parámetro fisiológico cuya magnitud hará pensar que estamos ante un sujeto enfermo, 0 ante uno que, en lo que concierne a la tensión arterial, no lo está. Es el ***resultado*** de haber aplicado tal instrumento el que, en todo caso, dará lugar a una ***valoración*** positiva o negativa del estado del paciente.

Según aquellas invocaciones, indicadores comunitarios tales como el ***Porcentaje de Viviendas Adecuadas, o*** la ***Tasas de Alfabetismo*** serían preferidos antes que otros como ***Porcentaje de Viviendas Inadecuadas*** o la ***Tasa de Analfabetismo.*** Es evidente que todo se reduce a jugar con las palabras, ya que la información contenida en los primeros y los segundos es exactamente la misma y, consecuentemente, toda comparación o valoración realizada sobre la base de unos llevaría al mismo resultado que si se hiciera usando los otros.

3.7. Nuevos indicadores

La discusión más relevante es si hace falta o no crear indicadores novedosos; si basta con los existentes, o es menester generar nuevas herramientas de esta naturaleza para realizar una evaluación fecunda de la realidad y, por ende, para conseguir una gestión más eficiente, o una comprensión más acabada de lo que se examina.

La respuesta es claramente afirmativa; la propia historia de los indicadores revela que los nuevos enfoques, las nuevas realidades socioeconómicas y las nuevas conceptualizaciones, reclaman nuevos instrumentos. No fue hasta que el problema de la inequidad económica al interior de las comunidades saltó al primer plano de las preocupaciones de los salubristas y economistas que no se crearon indicadores que, como el ***Indice de Gini, se*** orientan a cuantificar esa desigualdad (véase Cortés y Rubalcava, 1984).

Hace unos años propuse un indicador bautizado como ***Índice de Lactancia Acumulada*** [7], con el cual se procura cuantificar qué parte de toda la experiencia de lactancia materna potencialmente acumulable por los niños de una comunidad

[7] Véanse los detalles en Silva (1995).

durante los primeros cuatro meses de vida se produce efectivamente en el seno de esa comunidad. Posteriormente, y dentro de la misma área nutricional, introdujimos el ***Índice de Deserción de la Lactancia Materna*** (véase Silva, Amador y Valdés, 1994). Ambas operacionalizaciones hubieran carecido de mayor sentido en épocas anteriores a la comprensión generalizada de la importancia cardinal de esa práctica alimenticia, así como de su naturaleza profundamente social. En los dos casos, además, las potencialidades técnicas hoy disponibles a partir del acceso a las computadoras personales desempeñaron un papel clave, ya que sin ellas estos indicadores específicos serían virtualmente incalculables, y la teoría subyacente, punto menos que diletantismo estadístico.

El ejemplo más elocuente de un nuevo indicador sanitario nacido (e impetuosamente desarrollado) al calor de necesidades expresivas no cubiertas es el de los ***años de vida ajustados en función de la discapacidad (AVAD),*** descrito en detalle por Murray (1993). Con él se procura medir la llamada ***carga de la enfermedad,*** que intenta cuantificar no sólo los años de vida potencialmente perdidos por un colectivo (pérdida vinculada con la mortalidad) sino también el efecto de la enfermedad como agente que resta años de vida saludable a la comunidad. Una reciente aplicación de sumo interés puede hallarse en Murray, López y Jamison (1994).

La drogadicción, la contaminación ambiental, las nuevas áreas de la ergonomía, los desafíos de la ecología y la capacidad para comunicarse, son ejemplos de la larga lista de problemas relativamente nuevos para la humanidad cuya medición demanda nuevas herramientas formales.

3.8. Una nota contra los fetiches

La sumisión fetichista a los indicadores no es rara. Ya en la Sección 2.1 se expuso un ejemplo elocuente. Es común hallarla en el contexto de planes o programas cuando se usan para la fijación de metas numéricas. Por ejemplo, puede plantearse que, como resultado de cierto programa, para el año 2005, el porcentaje de niños que extienden la lactancia materna exclusiva hasta los 4 meses de edad deberá de elevarse, del actual 60%, al 85%.

En principio, la fijación de tales metas tiene el mérito de colocar el programa en un marco medible, y se plantea que sólo en ese caso podrán evaluarse oportuna y objetivamente sus resultados. Lo malo es que, por lo general, el modo en que esas metas se han fijado constituye un verdadero misterio. De modo que nuevamente estamos ante el fenómeno de la seudoobjetividad.

En efecto, la tasa de lactancia materna exclusiva que se alcance en el 2005 podrá medirse, pero si la meta del 85% se ha establecido de modo arbitrario, no se habrá avanzado nada desde la perspectiva metodológica a efectos de evaluar objetivamente la eficacia del programa.

Una estrategia que ponga el énfasis en dichas metas en lugar de colocarla en las

acciones necesarias para modificar la realidad que esos números procuran reflejar, es esencialmente perniciosa. Puede dar lugar a sentimientos de frustración por no haber alcanzado cierto valor del indicador en circunstancias en las que se hizo todo lo objetivamente posible por lograrlo y en las que el resultado conseguido es, incluso, meritorio. Puede, asimismo, producir el efecto opuesto: una actitud triunfalista porque se superó el número ansiado, aunque no se haya agotado el margen de acción existente para modificar aquella realidad.

Lo natural es establecer con claridad todas las acciones que, con el fin de mejorar la situación, pueden realizarse dentro de las restricciones presupuestarias, organizativas, logísticas, políticas y socioculturales en las que se ha de desenvolver el programa. Sólo cuando sea factible vaticinar el efecto que tales acciones tendrían, cabe poner estas metas como pautas valorativas del programa.

Es cierto que no es nada fácil hallar procedimientos que modelen procesos complejos de esta naturaleza y consientan vaticinar el valor de ciertos indicadores como función de determinadas acciones supuestamente modificadoras de la situación prevaleciente. Sin embargo, ello no legitima el establecimiento voluntarista de números que suplan esa ignorancia. Por otra parte, aproximaciones tales como la de los riesgos ***atribuibles múltiples*** calculados a partir de la regresión logística (véanse Bruzzi ***et. al.,*** 1985 y un ejemplo detallado en Silva, 1995) constituyen un enfoque mucho menos arbitrario que el compromiso de conseguir una disminución (o aumento) de cierta tasa sobre la base ingenua y metodológicamente irrelevante de que ello sería loable o conveniente.

Bibliografía

Abelin T (1986). ***Positive indicators in healthpromotion andprotection.*** World Health Statistics 39(4), WHO, Geneva.

Beyth-Maron R (1982). ***How probable is probable? A numerical translation of verbal probability expressions.*** Journal of Forecasting 1: 257-269.

Boneau CA (1961). ***A note on measurement scales and statistical tests.*** American Psychologist 16: 260-261.

Bruzzi PS, Green SB, Byar DP, Brinton LA, Schairer C (1985). ***Estimating the population attributable risk for multiple risk factors using case-control data.*** American Journal of Epidemiology 122: 904-919.

Bunge M (1985). ***Seudociencia e ideología.*** Alianza, Madrid.

Carta de Ottawa para la Promoción de la Salud (1986). ***Conferencia Internacional sobre la Promoción de la Salud: hacia un nuevo concepto de la saludpública.*** Ottawa, Canada.

Cortés F, Rubalcava RM (1984). ***Técnicas estadísticas para el estudio de la desigualdad social.*** Flacso, México DE

Draper N, Smith H (1981). ***Applied regression analysis. 2.ª*** ed, Wiley, New York.

Gaito J (1960). ***Scale classification and statistics.*** Psychological Review 67: 277-278.

Gaito J, Yokubynas R (1986). ***An empirical basis for the statement that measurement scaleproperties (and meaning) are irrelevant in statistical analysis.*** Bulletin of the Psychonomics Society 24: 449-450.

Gaito J (1986). ***Some issues in the measurement statistics controversy.*** Canadian Psychology 27: 63-68.

Gaito J (1980). ***Measurement scales and statistics: Resurgence of an old misconception.*** Psychological Bulletin 87: 564-567.

Hartman SB (1993). ***Indicadores «positivos» de salud y su relación con ciudades saludables.*** Trabajo presentado en la Primera reunión de registros de salud y estadística médica, México DE

Hosmer DW, Lemeshow S (1989). ***Applied logistic regression.*** Wiley, New York.

Last JM (1982). ***A dictionary of epidemiology.*** Oxford University Press, Oxford.

Lichtenstein S, Newman JR (1967). ***Empirical scaling of common verbalphrases associated with numericalprobabilities.*** Psychonomics. Science 9: 563-564.

Lord FM (1953). ***On the statistical treatment of football numbers.*** American Psychologist 8: 750-751.

Moses LE, Emerson JD, Hosseini H (1984). ***Analyzing data from ordered categories. New*** England Journal of Medicine 311: 442-448.

Murray CJL (1994). Q***uantifying the burden of disease: the technical basis for disability adjusted life years.*** Bulletin of the World Health Organization 72: 429-445.

Murray CJL, López AD, Jamison DT (1995). ***La carga global de enfermedad en 1990: resumen de los resultados, análisis de la sensibilidad y orientaciones futuras.*** Boletín de la Oficina Sanitaria Panamericana 118: 510-528.

Siegel S (1956). ***Nonparamettic statistics for the behavioral sciences.*** McGraw Hill, New York.

Silva LC (1995). ***Excursion a la regresion logística en ciencias de la salud.*** Díaz de Santos, Madrid.

Silva LC, Amador M, Valdés F (1995). ***Discontinuity indices: a tool for epidemological studies on breastfeeding.*** International Journal of Epydemiology 24:965-969.

Stevens SS (1946). *On* ***the theory of scales of measurements.*** Science 103: 667-680.

Stevens SS (1951). ***Mathematics, measurement and psychophysics*** en ***Handbook of Experimental Psychology*** Wiley, New York.

Streiner DL, Norman GR (1989). ***Health measurement scales. A practica1 guide to their development and use*** Oxford University Press, Oxford.

Townsend JT, Ashby FE (1984). ***Measurements scales and statistics: the misconception misconceived.*** Psychological Bulletin 96: 394-401.

Twaite JA, Monroe JA (1979). ***Introductory statistics.*** Glenview, 111.: Scott, Foresman.

Zdravomishov AG, Iadov P (1975). ***Efecto de las diferencias vocacionales en la actitud hacia el trabajo en la URSS.*** En: ***Industria y trabajo en la URSS*** Ciencias Sociales. La Habana: 139-179.

4

Cuantificación de nociones abstractas

Las facultades humanas de percepción, juicio, capacidad diferenciadora, actividad mental, e incluso preferencia moral se ejercen solamente cuando se hace una elección. Las potencias mental y moral, al igual que la muscular; solo se mejoran si se usan. Las facultades no se ejecutan haciendo una cosa meramente porque otros la hagan, ni tampoco creyendo algo solo porque otros lo crean.

JOHN STUART MILL

Uno de los desafíos metodológicos a los que se ve abocado el investigador biomédico, especialmente el interesado en los dominios más próximos a la sociología (salud pública y epidemiología) es la elaboración de instrumentos teóricos que permitan la operacionalización cuantitativa de nociones abstractas. Tales instrumentos son el objeto de análisis de este capítulo.

La medición de magnitudes tales como la concentración de hemoglobina, la edad gestacional, el volumen pulmonar o la colesterolemia puede ofrecer dificultades prácticas, pero es un problema tecnológica y conceptualmente resuelto; la de ciertas categorías abstractas como la ansiedad, la armonía familiar o la calidad de vida presenta, sin embargo, dificultades de uno y otro tipo.

Feinstein (1971) ha planteado que el famoso ***dictum*** de Lord Kelvin, según el cual ***si uno logra medir lo que está diciendo y lo puede expresar en números, es que sabe lo que dice; pero si no lo puede expresar con números es que el conocimiento que tiene de ello es escaso e insatisfactorio,*** constituye una especie de maldición que pesa perniciosamente sobre el ánimo de los investigadores biomédicos. En concordancia con esa doctrina, según este autor, algunos se empecinan en dar una interpretación métrica a todas las categorías con que trabajan; confunden la ***cuantificación*** con la ***metrización.*** Feinstein hace un reclamo vigoroso de que el investigador no abandone miméticamente la precisión en el uso de las palabras -la cual conduciría a valio-

sas cuantificaciones en forma de frecuencias- en favor de forzadas escalas numéricas para variables no medibles físicamente. Es difícil no compartir esta invocación, pero es obvio que no siempre que se intenta «metrizar» una categoría abstracta es porque se esté sometido a la «maldición de Kelvin». En muchos casos, la construcción de una operacionalización adecuada resulta vital para dar objetividad al proceso de investigación.

Es evidente que existen nociones cuya concreción cuantitativa resultará altamente polémica; pero ello no puede de antemano llevarnos de forma fatalista a renunciar a conseguirla. Cualquier construcción operativa, si se pretende que sea medianamente fecunda, exige una definición clara; de lo que se trata es de intentar aprehenderla cuantitativamente. El resultado difícilmente satisfará totalmente a todos, pero si el consenso evoluciona sin dogmatismos, tanto la propia definición de la categoría como su operacionalización habrán a su vez de modificarse en procura de conseguir aplicaciones más útiles y de aceptación más amplia.

Este capítulo se destina a debatir algunos aspectos de tan conflictivo asunto, cuyo estrecho parentesco con la estadística es, por lo demás, evidente.

4.1. Variables sintéticas

Denominamos *variable sintética (VS)* a una función de un conjunto de variables intermedias, cada una de las cuales contribuye a cuantificar algún rasgo del concepto cuya magnitud quiere sintetizarse. Nótese la generalidad del concepto de variable sintética: si las variables intermedias que la conforman fuesen indicadores — magnitudes susceptibles de ser obtenidas para agrupaciones- entonces dicha VS es también un indicador [1].

La materia prima de tal variable integrada suele ser el conjunto de respuestas a un cuestionario, en cuyo caso la *VS* se construye mediante alguna regla integradora de esas respuestas. La situación típica es similar a la que se produce con las famosas y controvertidas pruebas de inteligencia: tras intentar dar solución a una serie de problemas que se puntúan separadamente, al sujeto se le atribuye un puntaje global, con el que se calcula el polémico cociente de inteligencia conocido como *IQ*. Otro ejemplo clásico, en este caso de la clínica, es la propuesta de Apgar (1953) para cuantificar la vitalidad de un recién nacido en función del pulso cardiaco, el esfuerzo respiratorio, el tono muscular, el color de la piel y la respuesta al estímulo producido por la colocación de un catéter en las fosas nasales.

En esa línea se han desarrollado numerosos procedimientos con los que se miden nociones tales como la capacidad de liderazgo, el dolor, la gravedad de un

[1] Por otra parte, nada contradice la posibilidad de que la *VS* conste tan sólo de una variable «intermedia», aunque lo típico sea que la integren varias.

proceso asmático, la discapacidad funcional del anciano o la calidad de vida del trasplantado renal. La creación de una *VS* para la medición de la salud personal (física y psíquica) por medio del escrutinio múltiple de los sujetos, posteriormente conjugados en puntajes integrados, fue metodológicamente impulsada en Estados Unidos con motivo de la segunda guerra mundial y en virtud de la necesidad de valorar grandes cantidades de reclutas (Dowell y Newell, 1987). Un uso muy extendido de este tipo de variables se produce en el campo de la psicología, disciplina que quizás haya acopiado la mayor experiencia al respecto, tal y como testimonian los múltiples esfuerzos realizados desde la década del 40 bajo el auspicio de la ***American Psycho-Zogical Association,*** a los cuales se refieren profusamente artículos clásicos de la época como el de Cronbach y Meehl(1955).

Este proceso es llamado en ocasiones «construcción de una escala», expresión que, indudablemente, se ha acuñado con bastante firmeza, aunque a mi juicio no es muy adecuada, ya que, a diferencia de las VS, las escalas (nominal, ordinal, etc.) no se construyen: se usan en el acto de operacionalización. Ocasionalmente -aunque afortunadamente es muy poco frecuente- también para este tipo de construcción se ha usado el término «indicador», no solo en el caso legítimo en el que los ingredientes son propiamente indicadores, sino también cuando la *VS* está llamada a ser aplicada a una unidad de análisis aislada (por ejemplo, un paciente). Coherentemente con lo que se ha dicho sobre las escalas y con la definición propuesta para el concepto de indicador (véase Sección 3.5) aquí se mantendrá la expresión distintiva de ***variable sintética*** para este último caso.

Puesto que en materia de construcción y valoración de variables sintéticas parece reinar un considerable estado de confusión y cierto clima polémico, a continuación se fija un marco de partida que permita el examen ulterior de algunas expresiones de ese debate.

Junto con el requisito natural de que sea sencilla, las propiedades técnicas fundamentales que se suelen demandar para una variable sintética son que posea: fiabilidad ***(reliability)*** y validez ***(validy).***

Las alternativas metodológicas (en especial, estadísticas) para encarar estos dos requisitos han sido objeto de reflexión desde épocas «estadísticamente remotas». No muy avanzado el siglo se trató el tema en libros como el de Kelly (1924) o el de Thurstone (1931). Habida cuenta de la voluminosa bibliografía producida desde entonces, aquí solo se hará mención a los aspectos estadístico-computacionales que resulten imprescindibles para profundizar en los de índole conceptual.

4.2. Fiabilidad

Se dice que una ***VS*** tiene ***fiabilidad*** cuando mide de modo reproducible lo que se quiere; el sentido fundamental que está detrás de esta idea es el de la ***estabilidad*** en mediciones sucesivas. Más formalmente, se considera que una ***VS es*** fiable si se ob-

tiene esencialmente el mismo resultado cuando la medición se aplica reiteradamente. De lo que se trata en definitiva es de examinar si la variabilidad en mediciones sucesivas se mantiene dentro de cierto margen razonable, o si es tan grande que deba declararse la variable como no fiable.

Es evidente que las razones posibles para que los resultados de diversas aplicaciones de una medición no coincidan entre sí (fuentes de variabilidad) pueden ser múltiples. En principio, entre ellas se incluyen no sólo las variaciones atribuibles al instrumento de medición propiamente, sino también a los observadores, que no siempre operan del mismo modo, y al objeto medido, que puede experimentar variaciones entre una y otra medición.

En este punto bien podría cuestionarse la pertinencia de interpretar las posibles variaciones del sujeto que se mide como parte de la inestabilidad del instrumento con que se hace la medición. Por ejemplo, supongamos que se está cuantificando la concentración de colesterol en sangre de un individuo y que al hacerlo en tres ocasiones sucesivas se obtienen valores diferentes; ello pudiera deberse a que los valores de ese parámetro sanguíneo hubiesen efectivamente variado, hecho del cual el instrumento «no tiene la culpa». Sin embargo, en ocasiones, especialmente en el caso de ciertas variables sintéticas, puesto que en las respuestas registradas interviene el testimonio o la opinión del sujeto que responde (como pudieran ser ciertas declaraciones sobre el grado de satisfacción con la atención recibida durante una hospitalización) la falta de estabilidad pudiera ser inducida por el modo en que el instrumento de medición (el cuestionario) está diseñado, o por el modo en que se aplica.

En cualquier caso, lo importante es que cualquier expresión de inestabilidad es inconveniente, independientemente de su origen. Con el fin de poder discutir algunos conceptos generales y sin ningún afán de exhaustividad en un tema altamente especializado [2], a continuación se exponen algunas apreciaciones sobre el modo de evaluar estas condiciones.

Comencemos considerando el caso en que la *VS* da como resultado un número; dicho de otro modo, cuando la *VS* sea una variable continua o pueda manejarse como tal. Imaginemos que a cada sujeto de una muestra se le ha aplicado más de una vez el mismo cuestionario del cual se desprende un puntaje global del tipo que nos ocupa.

El ***coeficiente de fiabilidad (C_f) se*** define como la razón de dos medidas de variabilidad:

$$C_f = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

[2] Para tomar contacto de manera acabada con el examen estadístico de la fiabilidad se sugiere acudir a libros altamente especializados como el de Dunn (1989) o el más reciente de Ansell y Philips (1994). Una revisión amplia, con cientos de citas sobre el tema puede hallarse en Feinstein (1985).

donde σ_s^2 mide la variabilidad atribuible a que las unidades observadas difieren entre sí y σ_e^2 la que corresponde a las diferencias entre mediciones sucesivas en la misma unidad.

Obviamente, el coeficiente varía entre 0 y 1. $C_f = 0$ equivale a que $\sigma_s^2 = 0$; es decir, a que toda la variabilidad es atribuible al error inherente al instrumento, en cuyo caso se considera que su fiabilidad es nula. $C_f = 1$, el otro caso extremo, se produce cuando $\sigma_e^2 = 0$; es decir, cuando no hay variabilidad debido a la falta de fiabilidad; en tal caso, la estabilidad es máxima.

Para computar el C_f en el caso más frecuente (cuando se hacen sólo dos mediciones en cada una de las *n* unidades de análisis) [3], el procedimiento es muy simple, tal y como se expone a continuación. Supongamos que $x_1, x_2, \ldots . x_n$ es el resultado de haber medido la VS a *n* sujetos y que $y_1, y_2, \ldots . y_n$ *es* el de haberla medido en una oportunidad posterior a los mismos *n* individuos. Con esos datos se realizan los siguientes cálculos:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$M = \frac{\bar{x} + \bar{y}}{2} \qquad m_i = \frac{x_i + y_i}{2} \qquad (i : 1, \ldots . n)$$

y, de inmediato, los siguientes cómputos parciales:

$$S = 2 \sum_{i=1}^{n} (m_i - M)^2 \qquad R = n[(\bar{x} - M)^2 + (\bar{y} - M)^2]$$

$$T = \sum_{i=1}^{n} [(x_i - M)^2 + (y_i - M)^2] \qquad E = T - R - S$$

Finalmente, el coeficiente de fiabilidad se calcula mediante la fórmula siguiente:

$$C_f = \frac{S - E}{S + E} \qquad [4.1]$$

Por ejemplo, imaginemos que se ha hecho una encuesta a 10 individuos en dos oportunidades y que los resultados son los que se recogen en la Tabla 4.1.

[3] La solución para el caso más general, con *k* repeticiones *(k>2) se* deriva del análisis de la varianza.

Tabla 4.1. ***Resultados de medir cierta variable a 10 sujetos en dos ocasiones.***

Primera encuesta *(x)*	Segunda encuesta *(y)*
10	25
20	45
23	51
8	21
11	27
30	65
2	9
18	41
40	85
7	19

Es fácil corroborar que $M = 27{,}85$, $\bar{X} = 16{,}9$ y $\bar{y} = 38{,}8$. Aplicando las fórmulas anteriores se puede constatar que:

$$S = 5557{,}05 \qquad T = 8572{,}55 \qquad R = 2398{,}05 \qquad E = 617{,}45$$

De modo que el coeficiente de fiabilidad asciende a $C_f = 0{,}80$.

Cabe ahora intercalar un comentario relacionado con la incomprensión, tanto del significado del coeficiente de correlación lineal de Pearson (al que denotaremos *r)* como del papel de las pruebas de hipótesis. Es bien conocido que lo que dicho coeficiente permite medir es el grado de proximidad con el que un conjunto de puntos se ubica en torno a una línea recta. ¿Es realmente útil tal estadígrafo para medir, además, estabilidad entre dos mediciones sucesivas de una variable? Aunque la respuesta es negativa, eso es lo que en diversos sitios se ha sugerido para el caso de que la *VS* sea de índole cuantitativa. Ante un señalamiento crítico en ese sentido Mora-Maciá y Ocón (1992), se amparan en que así lo recomiendan importantes instituciones norteamericanas. En un libro dedicado a exponer una larga lista de VS para la medición de la salud individual, Bowling (1994) da cuenta de varias decenas de trabajos en que se incurre en dicha práctica. La recomendación, finalmente, también se puede hallar en textos sobre metodología epidemiológica. Por ejemplo, en el libro de Almeida (1992) epidemiólogo brasileño, destacado por su estimulante vocación crítica y polemizadora, se afirma algo que, lamentablemente, es probablemente cierto:

La medida de fiabilidad más empleada cuando se trata de variables continuas es el coeficiente de correlación de Pearson, basado en un modelo simple de regresión lineal de dos variables.

Almeida da su aval al procedimiento al ilustrar cómo ha de usarse. Plantea así, por ejemplo, que al aplicar una prueba en dos oportunidades sucesivas a 26 sujetos (con intervalo de 1 mes entre uno y otro) obtuvo el valor $r = 0{,}815$, «medida estadísticamente significativa al nivel del l,0%».

La propuesta ilustra el uso inercia1 de métodos formales sin la debida reflexión independiente. De hecho, aquí se va más allá: no sólo se concede importancia al valor del coeficiente de correlación lineal entre x y y, sino que se plantea que un valor resulte significativamente mayor que cero reflejaría una buena estabilidad (fiabilidad) del indicador. Cualquiera de las dos razones siguientes basta para comprender que ésta es una afirmación equivocada.

En primer lugar, el coeficiente de correlación lineal puede ser muy pequeño y, no obstante, resultar significativamente diferente de cero; basta con que haya cierta relación (aunque sea mínima) y que el tamaño muestra1 sea suficientemente grande. Por ejemplo, un valor de $r = 0{,}25$, siempre que se haya obtenido con no menos de 85 pares, ya se declarará significativamente distinto de cero al nivel 0,05 (IC:O,O3 – 0,45, con confianza del 95%).

Por otra parte, un valor de r relativamente alto -digamos, por ejemplo, $r = 0{,}7$-, casi con seguridad será «significativo»; es decir, muy probablemente revelará que la correlación lineal existe más allá del azar (basta que este $r = 0{,}7$, por ejemplo, se haya obtenido con una muestra de tamaño igual a 9). Sin embargo, esta «significación» no quiere decir que haya estabilidad. Puede ocurrir que el valor de r sea muy alto, incluso igual a 1, y que la estabilidad sea pésima.

Consideremos una situación quizás poco realista pero elocuente. Supongamos que se trabaja con el número de hijos en una muestra de mujeres embarazadas, pero que la pregunta esté formulada de manera tal que resulte equívoco si hay o no que incluir al niño en gestación. Si en la primera encuesta se consideró que éste no debía contemplarse, pero durante la aplicación de la segunda se interpretó que sí debía contarse, se tendrá en todos los casos que $y = x + 1$. La correlación lineal (que es lo que mide r) será perfecta ($r = 1$); sin embargo, la fiabilidad de la variable, obviamente, dista de ser óptima.

Reconsideremos los datos de la Tabla 4.1 que producían el valor $C_f = 0{,}8$. El coeficiente de correlación alcanza para esos datos el máximo valor posible. No hace falta realizar el cómputo para corroborarlo; basta advertir en la Tabla 4.1 que, para cada uno de los 10 individuos, se cumple la relación $y = 2x + 5$. De modo que los puntos están sobre una línea recta, lo cual equivale a que $r = 1$.

Nuestro sentido común nos dice que, independientemente de lo que se estuviera procurando medir, unos resultados como los reproducidos en esa Tabla están lejos de reflejar que la KS sea altamente fiable. Si se empleara el coeficiente de

correlación como criterio valorativo de la fiabilidad en este caso, tendríamos que decir que la fiabilidad es insuperable. El coeficiente de fiabilidad, en cambio, registra que la estabilidad deja bastante que desear [4].

Debe señalarse que, si bien se ha utilizado el C_f como recurso para poner en evidencia el error que entraña usar el coeficiente de regresión lineal, este indicador clásico también adolece de notables limitaciones. Para apreciarlas, supongamos que se cumpliera la relación $y = \alpha + \beta x$; en tal caso, puede demostrarse que se acerca a su máximo en la medida que β esté más cerca de 1. Sin embargo, contradictoriamente con lo que cabría esperar de un recurso que mide fiabilidad, C_f es independiente de α. Esto quiere decir, por ejemplo, que si en lugar de tener $y = 2x + 5$ en la Tabla 4.1, se tuviera $y = 2x + 30$, el valor de C_f seguiría siendo igual a 0,8. Y en el ejemplo de las mujeres embarazadas, donde se cumplía $y = x + 1$ ($\alpha = \beta = 1$), se alcanzaría el máximo $C_f = 1$ a pesar de que, como ya se discutió, la pregunta solo estaría redactada de modo enteramente fiable si se tuviera $y = x$ (es decir, $\alpha = 0$; $\beta = 1$).

Lo que ocurre es que la fórmula [4.1] corresponde a una versión «optimista» del coeficiente: la que se basa en el supuesto de que no hay variabilidad entre observadores. Si se aplica la que se deriva del supuesto (más realista) de que tal variabilidad sí existe, el valor del coeficiente de fiabilidad habría de calcularse mediante [4.2]:

$$C'_f = \frac{n\ (S\text{-}E)}{n\ (T\text{-}E) + (n\ -2)\ (T\text{-}S)} \qquad [4.2]$$

Para los datos de la Tabla 4.1, esta corrección arroja el valor $C'_f = 0{,}48$ y, si se cumpliera la relación $y = 2x + 30$, el valor de C'_f descendería a 0,19. De hecho, C'_f sí mide el grado en que la nube de puntos se ubica en una vecindad de la primera bisectriz (caso $\alpha = 0$, $\beta = 1$, equivalente a la situación $y = x$ para todos los sujetos).

Existen alternativas para valorar la fiabilidad. En este sentido remito al lector a los trabajos de Bland y Altman (1986) y de Brennan y Silman (1992). Véase también Candela (1992).

Para el caso en que la VS no es continua sino politómica, el recurso más recomendable es el coeficiente de concordancia *(agreement coefficient),* para el cual existen varias modalidades, todas relacionadas con el *coeficiente kappa* introducido por Cohen (1960).

Lo que hemos manejado hasta aquí es la medición de la llamada fiabilidad *externa,* una demanda cuyo interés para otorgar confianza a la VS es altamente intuitivo. Sin embargo, se han desarrollado algunos indicadores para medir otra forma de

[4] Weiner y Stewart (1984) sugieren que el valor de C_f debe ser, como mínimo, igual a 0,85 para considerar fiable a la *VS.*

fiabilidad de una VS: ***la consistencia interna.*** Ésta se produce cuando hay una alta concordancia entre los ítems que integran la *VS*[5].

Su medición exige realizar maniobras tales como dividir al azar en dos subgrupos a los ítems que integran la VS, luego utilizar separadamente una y otra mitad para evaluar a los sujetos, y finalmente corroborar la existencia de una alta asociación entre ambas construcciones. La conveniencia de asegurarse de esta forma de fiabilidad es muy discutible, ya que una alta asociación tras la maniobra bosquejada exigirá algún grado de redundancia en la información que se registra; sin embargo, es lógico aspirar a que los componentes de la *VS* recorran dimensiones mutuamente incorrelacionadas. El indicador más connotado para medir esta forma de fiabilidad es el llamado ***coeficiente alfa,*** propuesto por Cronbach (1951); cuando todos los ítems o variables intermedias son dicotómicos, este coeficiente equivale al conocido KR-20, ***coeficiente de Kuder-Richardson.***

4.3. Validación

Imaginemos que se está valorando la eficiencia de un dispositivo cuya función es la de contar los kilómetros recorridos por un automóvil. Si se transita desde la ciudad A hasta la B en varias ocasiones y las distancias registradas, salvo pequeñas fluctuaciones, coinciden entre sí, se dirá que el dispositivo es fiable. Pero eso no basta para considerarlo útil; es necesario constatar que la distancia registrada (el promedio si no fueran idénticas) coincide con la que realmente separa a ambas ciudades. En tal caso, se dirá que el instrumento es, además de fiable, válido.

La ***validez*** de un instrumento de medición –y, por tanto, también de una *VS*– es, en esencia, la capacidad que tiene de medir realmente el concepto que pretende medir. En el caso que nos ocupa, existen varias alternativas teóricamente aplicables para valorar esta condición; a saber: validez de aspecto ***(face validity),*** validez de criterio ***(criterio validity),*** validez de contenido ***(content validity),*** validez predictiva ***(predictive validity) y*** validez por construcción ***(construct validity).***

Antes de examinar con cierto detalle cada una de estas formas de validación, cabe recalcar que la teoría a que se alude aquí es la que concierne a construcciones relativamente complejas para medir ciertas nociones abstractas, para las que existe una definición y, sobre todo, una sustentación teórica bien desarrollada, sobre cuya base, precisamente, se construye la *VS.*

A diferencia de lo que ocurre con la fiabilidad, la discusión de la validez no tiene mayor sentido cuando se trata de la medición de rasgos físicos tales como hemoglobina en sangre o circunferencia cefálica del recién nacido, o incluso de índices

[5] Nótese que, a diferencia de la fiabilidad externa, esta variante sólo se puede aplicar a una reducida parte de las VS: aquellas conformadas a partir de un gran número de variables intermedias.

compuestos como el índice cintura-cadera que usan los auxólogos. Sin embargo, sí suele reclamarse validez para un indicador como el ***producto interno brutoper cápita*** en su calidad de presunta medida del nivel adquisitivo de los ciudadanos; ejemplos de este tipo resultan ser casos particulares de lo que hemos llamado ***variables sintéticas,*** de manera que quedan abarcados en el análisis que sigue.

4.3.1. Validez de aspecto

Esta primera forma de validación concierne al hecho de si el instrumento «parece» medir lo que se quiere. Por su naturaleza, es evidente que, para constatarla, hay solamente un recurso: el juicio de expertos. Las tareas no son otras que las de, por una parte, examinar la congruencia teórica entre la construcción teórica y el marco conceptual en que se inscribe el concepto y, por otra, valorar el grado en que éste ha quedado aprehendido de manera tal que el resultado sea coherente con el sentido común prevaleciente en el entorno científico y social en el que está llamado a operar. El consenso de los expertos es, precisamente, la forma en que puede concretarse ese sentido común.

4.3.2. Concordancia con una regla externa

La ***validez de criterio*** (también llamada ***validez por concurrencia)*** exige la existencia -como el nombre indica- de un criterio ***externo*** contra cuyos resultados contrastar los que produce la *VS* que se estudia. Este criterio actuaría como «árbitro» de dicha *VS:* si se observa un grado alto de concordancia, la nueva *VS* es válida; en caso contrario, no lo es. Una situación en que procede utilizar esta modalidad de validación se produce cuando lo que se busca es crear una VS tan eficiente como la que existe pero cuya aplicación sea más económica, 0 más sencilla, 0 menos peligrosa. Esto es lo que ocurre a veces con la evaluación de algunos medios diagnósticos. Un caso especial es aquel en que el ***gold standard*** no puede aplicarse en la práctica regular, como ocurre cuando el resultado se obtiene a partir de una autopsia; este caso, sin embargo, poco tiene que ver con el tipo de VS que nos ocupa.

La idea central consiste en evaluar alguna forma de concordancia entre los resultados que se obtienen al aplicar la nueva construcción y los de la externa, que se da por válida y a la que suele llamarse ***patrón de referencia*** (en inglés ***gold standard).*** Si tal criterio externo no existiera, este modo de validar, simplemente, no puede aplicarse. La alternativa en tal caso es hacer primero una construcción ***ad hoc*** del patrón de referencia; el modo de conseguirlo sería, nuevamente, a través de expertos que cuantifiquen de manera independiente la categoría con que se trabaja para una muestra (real o ficticia) de unidades. Tomando el concenso de tales resultados como el criterio correcto, cotejar entonces los datos obtenidos con los de la

VS De hecho, se estaría validando la variable a través de los expertos como en el caso de la validez de aspecto, sólo que ahora se les pediría un pronunciamiento cuantitativo y no cualitativo como antes.

Pero lo cierto es que muchas veces se está trabajando en la construcción de una nueva VS, precisamente porque la que existe no se considera satisfactoria, de modo que ***carece de sentido valorar una alta concordancia como indicio de validez,*** ya que esa que existe, no puede, lógicamente, admitirse como paradigma. Sobre este punto volveremos en la Sección 4.4.

4.3.3. Contenido correcto

La ***validez de contenido*** concierne al grado en que los componentes de la *VS* recorren todo el espectro del concepto involucrado. Normalmente, la medida en que ello se consigue se remite al grado de adhesión y rigor con que el instrumento contempla el entorno teórico en el que se inscribe la categoría sintetizada. Con mucha frecuencia, la noción que se mide es una dimensión no bien estructurada pero cuyo marco teórico permite identificar áreas o componentes [6]. De lo que se trata es de corroborar que todos ellos aparezcan representados entre los elementos que integran la medida final.

Por ejemplo, si para medir la inteligencia de un individuo se parte de que su capacidad para resolver problemas nuevos se asocia con cinco áreas concretas (a saber: pensamiento lógico, rapidez del razonamiento, creatividad, capacidad de abstracción y capacidad de asociación), entonces una prueba de inteligencia con validez de contenido habrá de contener preguntas y problemas que cubran las cinco áreas mencionadas[7]. De manera que esta forma de validación concierne al dominio teórico-lógico; es decir, no se conecta usualmente con esfuerzo empírico alguno, y la consulta a expertos, como bien advierten Dowell y Newell (1987), es el recurso pertinente.

4.3.4. Capacidad predictiva

En ocasiones (lamentablemente no siempre) la naturaleza de la *VS* es tal que si ella mide efectivamente aquello que se supone que mide, entonces es posible deducir un desarrollo o desenlace para la unidad medida en función del valor alcanzado.

Imaginemos, por ejemplo, que se ha creado una *VS* que supuestamente cuantifica la potencialidad intelectual de un alumno que se inicia en la universidad. Si se

[6] Obviamente, esta forma de validación carece de sentido si para la *VS* no existieran dichos «componentes» (de hecho, esto equivale a que la *VS* realmente no sintetiza nada).

[7] La mención a estas 5 áreas constituye un recurso para canalizar la explicación y no un punto de vista teórico que yo comparta. De hecho, el concepto de inteligencia pertenece a un dominio altamente especializado y es fuente de enconadas discusiones teóricas (véase Sección 5.6).

corrobora que aquellos sujetos que producen altos valores de la variable son los que consiguen el mejor desempeño académico, se habrá logrado *validar por predicción* dicha variable. Algunos autores ubican esta forma de validación dentro de la categoría más amplia que es objeto de la siguiente sección.

4.3.5. Validez por construcción

Éste es el procedimiento más interesante y creativo pero, a la vez, el más borroso desde una perspectiva teórica. Usualmente se trata de identificar ciertas condiciones, entornos o estratos, para los cuales la *VS* «debe» exhibir determinado tipo de valores, y luego corroborar que sus resultados son coherentes con dicha previsión. Es decir, teniendo en cuenta tanto la naturaleza de lo que se mide, como las características de dichos entornos y el encuadre teórico del problema, se anticipa un desempeño de la variable; se considera que la corroboración empírica de ese vaticinio valida la variable. De manera más general la validación por construcción se establece por el grado en que la *VS* se relaciona con otras mediciones que sean consistentes con hipótesis teóricamente derivadas del concepto que se quiere medir (Carmines y Zeller, 1979). Consideremos algunos ejemplos.

Ejemplo 1. Una ilustración ofrecida por Streiner y Norman (1989) nos sitúa en 1920, cuando no se disponía de procedimientos de laboratorio para medir la glucosa en sangre, y nos hace suponer que se está evaluando una prueba bioquímica concebida con ese fin[8]. Teóricamente, cabe esperar que se obtengan valores mayores para diabéticos que para los que no lo son, que produzca mediciones mayores en perros a los que se les ha suprimido el páncreas que en los que lo conservan, y que se registren resultados menores en diabéticos tratados con insulina que en enfermos no tratados. Si se confirmaran tales expectativas, se habría conseguido *validar por construcción* a la prueba bioquímica.

Ejemplo 2. *Se* ha creado una VS para medir «armonía en la convivencia familiar». Se espera que los valores de la *VS* aumenten sostenidamente en la medida que se incremente el nivel de ingresos económicos de la familia, siempre que este dato no se haya incluido como componente de la variable construida.

Ejemplo 3. *Se* quiere cuantificar mediante un cuestionario el nivel de ansiedad de un individuo. Se considera la posibilidad de medir la cantidad de sudor en las palmas de las manos en los minutos previos a un examen como recurso de validación. Si hay concordancia entre el grado de sudoración y los puntajes que produce el cues-

[8] El ejemplo no se relaciona con una VS típica, nacida de una elaboración teórica; por lo demás el hecho de que se haya seleccionado una ilustración tan rebuscada despierta automáticamente mi suspicacia sobre la aplicabilidad de esta forma de validación en la práctica regular.

tionario, se considera que la *VS* que mide la ansiedad ha sido validada por construcción.

Este enfoque produce al menos dos fuentes de inquietud. Primero, que al final todo reposa en supuestos (que hay más azúcar en la sangre de un diabético tratado con insulina que en uno que no está en ese caso, que el monto de ingresos económicos es favorecido por la armonía familiar, o que las manos de un sujeto sudan más en una situación extrema cuanto más ansioso sea), los cuales pudieran ser tan discutibles (o tan procedentes) como los presupuestos teóricos que avalan directamente a la *VS* Nótese que no se dice que esos supuestos sean discutibles sino que pueden ser *tan* discutibles como aquellos en los que reposa la operacionalización inicial. Desde luego, hay situaciones en que los supuestos son suficientemente razonables como para dejar poco espacio a la suspicacia. Por ejemplo, si se intenta medir la práctica regular de ejercicios físicos mediante un cuestionario, se puede suponer que aquel sujeto que realiza una ejercitación sostenida habrá de estar en mejor forma física; esta última se puede medir de manera directa a través de pruebas funcionales y por esa vía validar (o invalidar) el cuestionario.

En segundo lugar, por ese conducto quizás se podría hacer una validación cualitativa general, pero no se obtendría información sobre el grado de validez ***de la magnitud*** que se registra; es decir, no permite evaluar cuán sensible resulta la VS a los cambios de gradiente. La finalidad de construirla, sin embargo, no es la de tener una idea vaga sino de proveer al investigador de un recurso para ***cuantificar,*** de tal manera que si la *VS* arroja un valor 4 veces mayor para una unidad que para otra, podamos aceptar que la magnitud que se mide es, en efecto, 4 veces mayor en el primer caso que en el segundo. Por lo general, la validación por construcción no ayuda absolutamente nada en ese sentido.

Otra alternativa (Harman, 1976) que se considera dentro de las posibilidades para validar por construcción y que ha conseguido amplia popularidad consiste en realizar un ***análisis factorial*** con los diversos ítems o componentes intermedios de la *VS* como variables; si el primer factor resultante explica un porcentaje muy alto de la variabilidad total (por ejemplo, más del 90%), se considera que ese conjunto de ítems abarca una única dimensión; es decir, que existe una variable subyacente no susceptible de medirse directamente y que puede razonablemente suplida por la *VS* que se está considerando. Alternativamente, se ha sugerido (véase una aplicación, por ejemplo, en Schuling ***et al,*** 1993) la realización de un ***análisis de componentes principales*** a partir del mismo razonamiento.

4.4. El fantasma de la validación

Es muy común oír que alguien dude acerca de la calidad de una *VS* que se ha creado ***«porque aún no se ha validado» o*** que un fiscal metodológico cuestione o directamente condene algún esfuerzo operacionalizador de cierto concepto por esta

razón. Lamentablemente, en la mayor parte de estos casos se está invocando a un fantasma que, como todos los fantasmas, por una parte, no existe y, por otra, sólo sirve para amedrentar a su víctima.

La confusión mayor, fuente del grueso de las objeciones, se produce en relación con la validación por concurrencia. Resumamos la situación. Cuando se quiere medir determinada noción abstracta, se abren en principio cuatro posibilidades:

a) No existe ningún intento anterior de medir dicha noción o, por alguna razón, se considera inaceptable todo posible antecedente; consecuentemente, se decide crear una VS totalmente nueva.
b) Existe una operacionalización anterior que se considera de interés aunque se valora como insuficiente o inaplicable tal y como se ha propuesto, de modo que se adopta como fuente de inspiración para realizar una nueva construcción; se toman de allí algunas preguntas tal y como fueron originalmente concebidas, se modifican o suprimen otras, se adicionan nuevos ítems, y tal vez se redefine el modo de computar la *VS* como tal.
c) Existe una *VS* que se considera correcta pero se aspira a construir un instrumento equivalente solo que más simple 0 más económico. Con ese fin se realizan acciones como las descritas en el punto anterior.
d) Se asimila sin más cierta VS existente; en caso de que el idioma original no fuese el mismo al que se utiliza en el sitio donde habrá de aplicarse ahora, será necesario, naturalmente, traducir el instrumento original.

Examinemos la posibilidad de validar por concurrencia en cada una de estas cuatro situaciones.

En el primer caso, cuando cualquier posible estándar se considera improcedente, o simplemente no existe manera confiable alguna de medir lo que se quiere, queda cancelada toda posibilidad de validar la nueva *VS* por medio de un criterio externo [9].

Si existiera alguna operacionalización previa, pero la nueva *VS* se estuviera construyendo para enmendarla, estaríamos en el segundo caso; por simples razones lógicas, tampoco procede esta forma de validación. En efecto: ¿cuál sería el resultado deseable para el cotejo entre el criterio novedoso y el estándar que se usaría como referencia?

Imaginemos que, como resultado de la aplicación de esta *VS,* se consigue clasificar a cada unidad en una *k* categorías excluyentes (por ejemplo: ENFERMO – SANO, para $k = 2$, o PSICÓTICO – NEURÓTICO – SANO, para $k = 3$). Una concordancia alta (por ejemplo, un valor del *coeficiente kappa* ascendente a 0,85, que

[9] Nótese que, para cualquier concepto o noción que se considere, en caso de existir un estándar, éste quizás no pudo ser en un su momento validado por esta vía, ya que se carecía entonces de antecedentes. Es decir, siempre tiene que haber existido un primer intento operacionalizador que, por serlo, no pudo ser validado por concurrencia.

Fleiss (1981) calificaría como excelente) sería decepcionante; vendría a decir que el nuevo indicador es inútil, pues ya existe uno anterior que resuelve lo mismo; lo procedente en tal caso sería utilizar directamente la *VS* original. Una concordancia débil, por otra parte, sería indicio de que el nuevo instrumento mide de manera ***diferente*** el concepto. Pero es obvio que en tal caso no se ha hecho validación alguna, pues nada asegura que nuestra *VS* no esté haciendo peor las cosas. La convicción de que es más eficiente, si existe, es anterior al cotejo; éste, en el mejor de los casos, confirmaría que los procedimientos no miden lo mismo.

La situación es la misma en caso de que la *VS* registre un «puntaje»; es decir, cuando sea una variable continua, como podría ocurrir con un supuesto ***coeficiente de autonomía funcional del anciano*** que varíe entre 0 y 100. Solo si hubiera una manera muy buena aunque inconveniente (por su costo o sus riesgos asociados) de medir la magnitud de esa autonomía, podría validarse otra medida que no tuviese tales inconvenientes mediante alguna forma de concordancia.

En el tercer caso no sólo es posible sino que resulta ineludible validar la *VS* que se propone contrastándola con aquella a la que supuestamente habrá de suplir.

Para concluir, consideremos el cuarto y último caso: se ha decidido aplicar cierta *VS* tal y como fue originalmente concebida. Valoremos primero la situación en que si bien el sitio de procedencia de la *VS* no es el mismo al de su próxima aplicación, el idioma que se maneja en ambos lugares sí coincide y por tanto no hace falta hacer traducción alguna. Es evidente que carece de sentido valorar siquiera la posibilidad de validar por concurrencia la variable, ya que tal validación ***exige*** la aplicación de dos instrumentos diferentes a la misma muestra, algo imposible en este caso por razones obvias. Finalmente, si el material de partida fuera un cuestionario creado en otro idioma, entonces se sugiere corroborar la calidad de la traducción mediante una «retrotraducción» (traducir nuevamente al idioma original el matrial que resultó de hacer la primera traducción). La expectativa es que la versión original no difiera apreciablemente del resultado de la retrotraducción; en tal caso podrá confiarse en que el documento con el que se trabajará ha respetado el espíritu del original y, sobre todo, que ha reproducido adecuadamente su contenido. Pero esto no debe confundirnos; ***tal maniobra no integra el acto de validación*** en sí mismo sino que es parte del proceso de asimilación de una tecnología y se desarrolla con el afán exclusivo de evitar diferencias artificialmente introducidas por conducto de la traducción. Una vez hecha la traducción y habiéndose confirmado su eficiencia, estamos en el mismo punto que cuando no era menester traducir: no hay forma alguna de constatar empíricamente que el cuestionario «funciona» en un sitio del mismo que lo hace en otro.

Un ejemplo, que ilustra la incomprensión de estas realidades lo ofrece García-García (1995) cuando impugna la aplicación que hacen Gutiérrez ***et al*** (1994) de un conocido índice de calidad de vida creado por Spitzer ***et al*** (1981) (ICVS) debido a que, según García-García, siendo «evidente que el índice necesita ser adaptado al nuevo entorno cultural, se requiere de un laborioso proceso de traducciones y retrotraducciones... y valorar si el "nuevo" instrumento se comporta como el original».

El caso que nos ocupa está aparentemente ubicado en la segunda de las cuatro alternativas (los propios objetores hablan de «nueva versión» y de la posibilidad de patentarla), pero aun si se tratara de que la única modificación del cuestionario en que se basa el índice hubiera sido el idioma utilizado, el reclamo de validar por concurrencia («valorar si se comporta como el original») solamente puede provenir de la tendencia a repetir frases hechas sin otro asidero que la inercia metodológica. García-García comunica en el artículo citado que en su momento había llevado adelante una validación de aspecto y de constructo para el *ICVS* y anuncia que próximamente intentará corroborar que el propio *ICVS* (traducido y avalado por una retrotraducción) se comporta como el original usando una muestra procedente de la atención primaria española. Estoy francamente intrigado por conocer lo que realmente haría este investigador si es que se propone valorar si hay concordancia entre los resultados del *ICVS* en el país de origen y los que arroje en España. También me intriga qué es lo que se entiende por que unos y otros sean concordantes. Si dos objetos diferentes se miden con respectivas reglas (incluso en el caso de que una sea réplica de la otra) me temo que nada se podrá sacar en claro sobre la validez de dichos instrumentos a partir de la comparación de las magnitudes obtenidas por uno y otro.

Sintetizando, en la mayoría de las situaciones prácticas es ***imposible*** demostrar por conducto de la concurrencia que una *VS* mide lo que se desea de modo tanto o más eficiente que el patrón de referencia. Por otra parte, en relación con las restantes formas de validación, también es posible que seamos víctimas de demandas más inerciales que fundamentables, ya que lo usual es que la situación no se preste para aplicar criterios de validación predictiva, en tanto que la validez por construcción puede ser tan discutible que resulte perfectamente legítimo que alguien decida prescindir de ella. En última instancia, con frecuencia uno siente fuertemente que este recurso se usa menos por necesidad o convicción que como hoja de parra: no nos viste realmente, pero nos defiende del embate potencial de los puritanos.

Resumiendo, con extrema frecuencia, a las *VS* no se les ha confirmado la validez -aparte del criterio de expertos y de la reflexión crítica de índole teórica-porque no ***existe un método para hacerlo,*** e incluso ***porque carece de sentido plantearse esa tarea,*** por mucho que la exijan los adalides de la metodología infuncional. Si se dan las condiciones, lo ideal es llevar adelante acciones orientadas a confirmar las cinco formas de validación; pero si no se cuenta con la posibilidad de aplicar algunas de ellas, lo ideal no es precisamente declarar que el esfuerzo operacionalizador no sirve para nada.

4.5. Variables sintéticas para la reflexión y el debate

A continuación se exponen con cierto detalle varias ilustraciones que muestran el modo como puede concretarse la creatividad en materia de construcción de varia-

bles sintéticas; además de ser útiles para repasar algunos de los puntos de vista desarrollados pueden, en algunos casos, incentivar valoraciones críticas y ofrecer sugerencias susceptibles de ser adaptadas a situaciones similares. En el libro de Bowling (1994) el lector podrá encontrar un amplio y detallado inventario de media centena de variables sintéticas para medir aspectos tales como capacidad funcional, bienestar psicológico, estado de ánimo y autoestima.

4.5.1. Riesgo cardiovascular

Desde que las enfermedades cardiovasculares alcanzaron el grado de prominencia que hoy tienen en los países desarrollados, han surgido diversas variables sintéticas que procuran medir el riesgo de padecerlas. Entre muchas otras, cabe reparar en las que han propuesto Shaper *et al* (1987), Chambless *et al* (1990) y Orduñez *et al* (1993).

La primera, por ejemplo, procura estimar el riesgo de desarrollar isquemia coronaria e ilustra el caso en que la construcción se apoya en un modelo formal, ya que se usan los coeficientes surgidos de la regresión logística como fuente para la definición de ponderaciones de siete variables. El índice en cuestión es el siguiente:

$$I = 51X_1 + 5X_2 + 3X_3 + 100X_4 + 170X_5 + 50X_6 + 95X_7$$

donde:

X_1 = colesterol sérico (mmol/l).
X_2 = años acumulados como fumador.
X_3 = tensión arterial sistólica (mm de Hg).
X_4 = 1 si tiene dolores anginosos; 0 si no los tiene.
X_5 = 1 si el sujeto recuerda un diagnóstico de enfermedad isquémica y 0 en otro caso.
X_6 = 1 en caso de que alguno de los padres haya muerto por una dolencia relacionada con el corazón; 0 en caso opuesto.
X_7 = 1 si es diabético; 0 si no lo es.

La propuesta de Ordúñez *et al. se* basa, en cambio, en un algoritmo bastante simple, como se muestra a continuación. Se propone un *Sistema Lógico de Riesgo Cardiovascular* (SILORCA), que consiste en darle cierta puntuación a cada uno de seis «marcadores de riesgo» (tensión arterial, colesterolemia, tabaquismo, obesidad, ingestión de alcohol y sedentarismo) para instrumentar una clasificación útil en materia de prevención. SILORCA es un ejemplo de construcción elemental de una VS a partir de la simple suma de 6 puntajes intermedios para cuantificar el riesgo de una enfermedad para la cual se han identificado más de 300 marcadores de riesgo. Tres marcadores de riesgo fueron considerados «mayores»; el puntaje para ellos

osciló entre 0 y 2; los otros tres se manejaron a nivel dicotómico: o no aportan nada a la VS, 0 aportan 1 punto. La suma oscila, entonces, entre 0 y 9.

La definición de la *VS* se resume en la siguiente Tabla.

Sistema Lógico de Riesgo Cardiovascular (SILORCA)

Marcadores de riesgo	Puntuación		
	0	1	2
Tensión arterial diastólica (mmHG)	menos de 90	90-104	más de 104
Colesterol (mmol/L)	menos de 5,2	5,2-6,l	más de 6,l
Tabaquismo (cigarrillos)*	menos de 1	l-10	ll y más
Índice de masa corporal (Peso/talla en Kg/m^2)	26 o menos	27 o más	
Sedentarismo*	No	Si	
Consumo de alcohol*	No	Si	—

* El tabaquismo se refiere a consumo diario y se hizo una equivalencia de 5 cigarrillos por cada habano y 2 cigarrillos por cada pipa. Se consideran no sedentarios los sujetos que realizan ejercicios físicos por lo menos trs veces por semana durante 30 minutos o más en cada ocasión. Son consumidores de bebidas alcohólicas los que las ingieren esporádicamente pero de forma excesiva y los que las consumen diariamente, cualquiera que sea la cuantía.

Como todo sistema de puntuación, el SILORCA está «contaminado» por la subjetividad de sus creadores; la conceptualización y la concreción operativa de las variables incorporadas se basó en la apreciación teórica que los autores conformaron a partir de los diversos datos disponibles en la literatura. Como se ve, la definición de algunas variables intermedias es algo laxa; una formalización exquisita en la medición, por ejemplo, del alcoholismo, pudiera ser atractiva desde el punto de vista métrico, pero en esa misma medida deja de serlo desde la perspectiva operativa. A veces el refinamiento operacional es contraproducente, pues el afán de precisión puede, insidiosamente, trocarse en resultados menos precisos debido a las dificultades de aplicación. Esta VS (al igual que la anterior) puede ser validada por predicción; la comparación de tasas de incidencia de cardiopatía para diferentes valores de SILORCA, por ejemplo, podría servir para ese fin.

4.5.2. Medición de conocimientos farmacoterapéuticos

La prescripción de medicamentos según Mercer (1978) constituye la expresión terapéutica más frecuente y de mayor repercusión económica global en la mayoría de los países. La OMS (1985) ha reconocido que la práctica de prescripción de medicamentos es, con frecuencia, ilógica e irracional y, consiguientemente, peligro-

sa. Una parte de la responsabilidad inherente a esta irracionalidad puede atribuirse al profesional recetador, debido a la insuficiencia o desactualización de sus conocimientos.

González *et al* (1991) realizaron un estudio empírico sobre las prácticas prescriptoras en la Habana que confirmaba esta realidad para los médicos de familia en esa ciudad. Estas circunstancias promovieron el interés de Silva y Nogueiras (1991) por evaluar el grado de conocimientos que posee el médico de la atención primaria en Cuba sobre el manejo de fármacos y conceptos afines. Para medir ese nivel de conocimientos se consideró necesaria la creación de un cuestionario y, a partir de él, de una *VS*. El análisis teórico del problema condujo a la convicción de que dicha variable debía abarcar el siguiente conjunto de áreas o esferas:

I. Pertinencia de la prescripción según síntomas y signos.
II. Empleo de sucedáneos.
III. Manejo de antibióticos.
IV. Manejo de sicofármacos.
V. Reacciones adversas a los medicamentos.
VI. Contraindicaciones.

De lo que se trataba era de confeccionar preguntas que «recorrieran» estas 6 esferas, proceder típico en este tipo de situaciones. La encuesta completa, que constó de 14 apartados, se reproduce en el ANEXO 1. Para ilustrar la lógica de la VS asociada, detengámonos en tres de estos apartados. Por ejemplo, la primera pregunta de la encuesta reza así:

¿En cuáles de estos casos utilizaría usted el CLORAMFENICOL como droga de elección?

— Absceso Dental: SÍ _____ NO _____ NO TENGO OPINIÓN _____
— Fiebre tifoidea: SÍ _____ NO _____ NO TENGO OPINIÓN _____
— Amigdalitis: SÍ _____ NO _____ NO TENGO OPINIÓN _____
— Faringitis Estreptocócica «B» SÍ _ _ _ N O _____ NO TENGO OPINIÓN _____
— Infección urinaria: SÍ _____ NO _____ NO TENGO OPINIÓN _____

En ella se indaga acerca de la utilización de una droga peligrosa y con indicación ***muy precisa y restringida,*** que ha de ser aplicada sólo para la fiebre tifoidea. Cada uno de los 5 ítems [10] que lo componen sirve para valorar dos áreas (la 1 y la III).

El segundo apartado fue diseñado del modo siguiente:

[10] Cabe aclarar que se consideró como un ítem a cada una de las posibilidades que se plantean tras el enunciado de cada apartado. Por ejemplo, el primer apartado abarca 5 ítems y el segundo 4, pero el cuarto contiene un único ítem (véase Anexo 1).

¿En cuál de los siguientes casos cabría considerar el uso conjunto de AMPICILINA ***y*** CLORAMFENICOL?

Peritonitis por apendicitis perforada:	SÍ _____	NO_____	NO SÉ _____
Sépsis meningocócica:	SÍ _____	NO_____	NO SÉ _____
En determinadas infecc. hospitalarias:	SÍ _____	NO_____	NO SÉ _____
Neumonía por ***Haemophilus***	SÍ _____	NO_____	NO SÉ _____

Se ha sugerido la posibilidad de utilizar una combinación que no se debe admitir en ningún caso. Con ello se valora el conocimiento sobre un aspecto relacionado con el manejo de antibióticos, y se agregan otros 4 ítems que también hacen aportes tanto al área 1 como a la III.

Como tercer ejemplo, consideremos el sexto apartado de la encuesta:

Señale cuáles de las siguientes entidades suponen contraindicaciones para la administración de BUTACIFONA:

– Diabetes mellitus:	SÍ _____	NO _____	NO SÉ _____
– Antecedentes de úlcera péptica:	SÍ _____	NO_____	NO SÉ _____
– Antecedentes de discrasia sanguínea:	SÍ _____	NO_____	NO SÉ _____
– Hipotensión arterial	SÍ _____	NO_____	NO SÉ _____

Aquí se sondea el conocimiento acerca de las contraindicaciones para el uso de una droga de por sí potente y peligrosa que, usada en pacientes con antecedentes, historia de úlcera péptica o de discrasias sanguíneas, puede ser causante de alteraciones graves. De aquí se derivan 4 ítems vinculados a la esfera VI.

De este modo se va concretando una estrategia de recorrido de las diversas áreas dentro del instrumento según la cual un mismo ítem se relaciona con una o más esferas de interés. La encuesta contiene formalmente 14 apartados pero, en realidad, consta de 51 ítems. El aspecto clave que debe considerarse es, naturalmente, la evaluación global a través de la cual se sintetiza la noción de interés. Se valoró que quizás no todos los ítems tenían porqué poseer la misma «importancia» de cara a la evaluación. Sin embargo, se descartó la posibilidad de atribuir ponderaciones a los diferentes ítems porque los autores no pudieron identificar pautas racionales para establecer jerarquías y optaron por eludir la gran carga de subjetividad a la hora de determinar la importancia de cada ítem [11].

Se sugirió entonces el ***índice de conocimientos farmacoterapéuticos (ICF)*** definido del modo siguiente:

[11] Debe repararse sin embargo en que al «no fijar ponderaciones» se elude la atribución de pesos sólo en apariencia, ya que, de hecho, se está decidiendo otorgar la misma «importancia» a todos los ítems.

$$ICF = \frac{a}{n} 100$$

donde ***n es*** el número total de ítems evaluados (51 en este caso) y el número de ítems contestados correctamente.

Por otra parte, se consideró que no todas las respuestas incorrectas son de igual naturaleza: contestar «No sé» (o «No tengo opinión»), además de que no se puede admitir como una respuesta correcta, también refleja un estado de conciencia de que el asunto se desconoce; contestar mal a pesar de existir la opción «No sé» (o «No tengo opinión»), también revela ignorancia pero acompañada de una convicción errónea, matiz'de gran interés a los efectos de eventuales programas de formación pre o posgradual. Por ello se creó una *VS* que midiera de qué parte del desconocimiento es consciente el individuo: se definió así el ***indice de desconocimiento consciente (IDC)*** del modo siguiente:

$$IDC = \frac{b}{n - a} 100$$

donde ***b es*** el número de ítems para los que explícitamente el encuestado reconoce ignorancia.

El numerador del ***IDC*** refleja la parte del desconocimiento atribuible a una ignorancia consciente del encuestado; por su parte, el denominador refleja el monto total del desconocimiento, tanto consciente como inconsciente. De modo que el ***IDC*** permite obtener, como se pretendía, una medida de la cantidad de desconocimiento que el médico reconoce como tal.

Cabe advertir que el cuestionario sugerido es, en esencia, una especie de examen. De hecho la «nota» que clásicamente se otorga a los estudiantes en ocasión de un examen no es más que una variable sintética para medir sus conocimientos y cuya «validación» siempre se ha reducido (en el mejor de los casos) al criterio razonado de expertos (jefes departamentales o colegas del profesor). En el caso que nos ocupa se corroboró que el ***ICF*** crecía sostenidamente con los años de actividad profesional acumulados por el encuestado. Este hecho viene a constituir una especie de validación por construcción, supuesto que la mayor experiencia profesional se asocia con mayor nivel de información y cultura farmacoterapéuticas.

4.5.3. IDH: una manipulación ingeniosa

En enero de 1966 la ***Organización de Naciones Unidas*** creó el ***Programa de Naciones Unidaspara el Desarrollo*** (PNUD), cuyo propósito declarado es ***impulsar el desarrollo humano*** y ***propiciar acciones encaminadas a fomentarlo*** (PNUD, 1992). Un nuevo concepto de desarrollo humano fue introducido en 1991, vertebrado alrede-

dor de la noción de bienestar y, más concretamente, en torno a cuatro dimensiones: la diversidad de opciones de las personas, su capacidad tanto en materia de salud como de educación, y su acceso a los recursos.

A partir de este marco conceptual surgió la necesidad de crear una *VS* orientada a cuantificar este concepto a nivel de una población: el ***Índice de Desarrollo Humano (IDH),*** concebido para ser calculado globalmente para un país. Se trata de un número que se ubica entre 0 y 1, cotas indicativas del mínimo y máximo desarrollo posibles. El índice supuestamente permite, por tanto, ordenar a los países según su nivel de desarrollo humano. Por ejemplo, en 1993 los primeros 5 países según el IDH fueron los siguientes: Japón (0,984), Canadá (0,982) Noruega (0,978), Suiza (0,978) y Suecia (0,976) en tanto que los 5 últimos, ocupantes de los lugares del 169 al 173, fueron: Níger (0,080), Burkina Faso (0,074), Afganistán (0,066), Sierra Leona (0,065) y Guinea (0,045).

En la literatura publicada por el PNUD, en especial en el informe de 1993, aparecen notas técnicas que, además de esclarecer varios aspectos metodológicos, intentan comunicar el modo de calcular esta *VS*. La explicación que ofrece el propio PNUD (1993) al respecto resulta confusa e incompleta. Algunos trabajos posteriores, como el de López (1994) y el de Rosenberg (1994) exponen con cierto detalle adicional los aspectos computacionales. No obstante, por su interés metodológico, tanto en lo que tiene de positivo como en sus aspectos criticables, a continuación se ofrece una explicación detallada.

El índice comprende tres componentes esenciales (longevidad, conocimientos e ingresos) expresados respectivamente en términos de: ***esperanza de vida al nacer (EVN), logro educativo (LE) y producto interno bruto per cápita ajustado (PIBA).*** Para adentrarnos en el procedimiento de cómputo es necesario explicar cómo se obtiene, para un país dado, cada uno de sus tres componentes. Para ilustrar la explicación se ha elegido un país concreto, Austria, cuyo ***IDH*** para 1993 ascendió a 0,952.

El primero de estos componentes no demanda de explicación: se trata simplemente del conocido parámetro poblacional del mismo nombre. A Austria se le atribuía en 1993, según el PNUD, una ***EVN*** de *74,8* años.

En el cálculo del ***PIBA*** intervienen dos parámetros: ***el producto interno bruto per cápita (PIB) y*** el ***umbral de pobreza,*** a los que llamaremos Y y *U* respectivamente. El PNUD define este umbral como ***el nivel del producto interno bruto per cápita*** (PIB) ***por debajo del cual no es posible garantizar; desde el punto de vista económico, tanto una dieta adecuada como requerimientos no alimenticios esenciales.*** Un aspecto que no aparece de forma clara en los informes del PNUD (1990-1993) es cómo se ha procedido para definirlo; lo que queda inequívocamente establecido es que en 1990 fue fijado en 4.861 dólares y que, a partir de 1991, se redujo a 4.829 dólares [12].

El ***PIBA*** de cada país se obtiene a partir de la siguiente fórmula (que hemos

[12] Es muy probable que en versiones posteriores a 1993 los procedimientos y parámetros usados por PNUD para computar el ***IDH*** hayan sido modificados, ya que así lo ha venido haciendo con regularidad este organismo.

deducido a partir de la explicación que se da en PNUD (1992) y corroborada con las tablas que allí aparecen):

$$\boldsymbol{PIBA} = (k+1)\sqrt[k+1]{Y - kU} + \sum_{i=1}^{k} i\sqrt[i]{U} \qquad [4.3]$$

donde *K* es la parte entera del cociente entre Y y *U.* Se deduce de [4.3] que, para los países cuyo ***PIB*** está por debajo del umbral (*Y*<*U*), se tiene ***k = 0 y,*** consecuentemente, ***PIBA*** = *Y; **es*** decir, el producto interno bruto per cápita no es en este caso objeto de ajuste alguno.

Esta compleja función del ***PIB,*** que se comenta más adelante, tiene como finalidad atenuar drásticamente el efecto de los ingresos sobre el *IDH* para aquellas naciones cuyo ***PIB*** per cápita se halla por encima del llamado umbral de pobreza. Tal maniobra responde al criterio del PNUD, según el cual los dólares que exceden el nivel de pobreza deben «pesar» mucho menos, puesto que ese dinero ya no es tan necesario para el desarrollo personal; para países cuyo ***PIB se*** coloca por encima de otros múltiplos de *U,* ese «peso» debe ser aún menor porque serían recursos crecientemente superfluos para tener una vida digna y adecuada.

Para Austria, Y = 16.504 y, como *U* = 4.829, se obtiene $\frac{Y}{U} = 3{,}42$, cuya parte entera es *K* = 3. Entonces, según [4.3], se tiene que:

$$\boldsymbol{PIBA = 4.829 + 2\sqrt{4.829} + 3\sqrt[3]{4.829} + 4\sqrt{416.504 - 3\,(4.829)} = 5.045}$$

La contribución de la esfera educacional se basa en ***el porcentaje de alfabetismo*** ($\boldsymbol{X_a}$) ***y*** la ***mediana en años de escolarización*** ($\boldsymbol{X_b}$)***;*** estos valores conforman la materia prima para calcular el llamado ***logro educativo (LE)*** de un país. Se empieza por computar X_a^* y X_b^* como se muestra a continuación:

$$X_a^* = \frac{\max_j X_{aj} - X_a}{\max_j X_{aj} - \min_j X_{aj}}$$

$$X_b^* = \frac{\max_j X_{bj} - X_b}{\max_j X_{bj} - \min_j X_{bi}}$$

donde j se mueve por todas las unidades (países) involucradas en el cómputo (173 en el informe de 1993). La fórmula según la cual se calcula el ***logro educativo*** es, por último, la siguiente:

$$\text{LE} = 2(1 - X_a^*) + (1 - X_a^*)$$

Como se ve, para esta medición integrada del logro educativo se ha otorgado un peso dos veces mayor al alfabetismo que a la mediana de escolaridad. Algunos recla-

man fundamentaciones formales ante tal tipo de decisiones y reaccionan airadamente si éstas no se dan detalladamente. Personalmente me parece una exigencia injusta. La asignación de pesos ***siempre*** será en alguna medida «arbitraria»; pero es muy diferente que sea arbitraria a que sea caprichosa. Procurar que se explique por qué se usaron unos pesos dados y no otros me parece correcto, pero considero «abusivo» con quien tiene la iniciativa de crear algo exigirlo en el mismo tono con que se puede exigir la demostración de un teorema; recomiendo evitar la crítica facilista, que se puede hacer siempre, y que tiene una vocación poco constructiva. Salvo que se trate de alguien que sostenga doctrinariamente que las ponderaciones de este tipo deben ser definitivamente suprimidas de la metodología de construcción de índices, me temo que el propio crítico se vería en una situación difícil si se le pidiera que explicara cómo hay que definir los pesos. Por otra parte, los que exigen que se les expliquen las ponderaciones y que se les convenza de su pertinencia, parecen olvidar que cuando no se hace uso de ponderaciones, lo que se está decidiendo es que todos los sumandos pesen igual, presupuesto que pudiera demandar igualmente de una explicación persuasiva. En este caso, por ejemplo, ¿por qué dar la misma importancia al porcentaje de alfabetismo que a la mediana de escolarización?

En definitiva, estamos de nuevo en el tema del consenso. Si al decidir las ponderaciones, y tras un examen informal (no por ello frívolo), los pesos que se otorgan no resultan ilógicos, entonces deben a mi juicio admitirse. Si resultaran «chocantes», lo que eso quiere decir es que no comulgan con nuestro «sentido lógico»; finalmente, si ello ocurre a muchos, entonces es claro que los pesos no condicen con el «sentido común», ya que éste último no es otra cosa que la síntesis del sentido lógico de la mayoría. Lo que sí me parecería respetable y atendible sería un argumento concreto que señalara la improcedencia de favorecer del modo en que se haya hecho a una variable intermedia más que a otra; y más respetable y atendible aún si el objetor hiciera una propuesta alternativa en lugar de circunscribirse a pedir explicaciones como si estuviera investido de la condición de fiscal, muchas veces adquirida para compensar su incapacidad para no producir nada fuera de las críticas.

Volviendo al ejemplo, y a modo de ilustración, el ***Recuadro 1*** presenta los datos de Austria para el cálculo del logro educativo.

Recuadro 1. ***Datos necesarios para el cómputo del logro educativo de Austria para 1993***

Tasa de alfabetismo de adultos en Austria (X_a = 99,0.
Tasa máxima de alfabetismo en adultos (Suecia) = 99,0.
Tasa mínima de alfabetismo en adultos (Burkina Faso) = 18,2.

Mediana en años de escolarización para Austria (X,) = 11,l.
Mediana en años de escolarización máxima (EEUU) = 12,3.
Mediana en años de escolarización mínima (Níger) = 0,l.

Estos datos producen lo siguiente:

$$X_a^* = \frac{99{,}0 - 99{,}0}{99{,}0 - 18{,}2} = 0{,}000$$

$$X_b^* = \frac{12{,}3 - 11{,}1}{12{,}3 - 0{,}1} = 0{,}098$$

LE = 2,90

Una vez que se tienen los valores de ***EVN LE y PIBA*** para todos los países incluidos en el análisis, si llamamos I_1, I_2 e I_3 respectivamente a estos tres valores para un país dado, se computa:

$$I_t^* = \frac{\max_j (I_{jt}) - I_t}{\max_j (I_{jt}) - \min_j (I_{jt})}$$

para $t = 1$, ***2, 3*** y ***j:*** 1, . . . , 173. El *IDH,* finalmente, se calcula mediante la fórmula siguiente:

$$\boldsymbol{IDH} = 1 - \frac{1}{3}\sum_{t=1}^{3} I_t^* \qquad [4.4]$$

El ***Recuadro 2*** ofrece los datos necesarios para computar la fórmula [4.4] en el caso de Austria.

Recuadro 2. ***Datos necesarios para el cómputo del IDH de Austria en 1993***

Esperanza de vida de Austria (I_1) = 74,8.
Esperanza de vida máxima (Japón) = 78,6.
Esperanza de vida mínima (Sierra Leona) = 42,0.

Logro educativo de Austria (1,) = 2,90.
Logro educativo máximo (Estados Unidos) = 3.
Logro educativo mínimo (Sierra Leona) = 0,13.

PIBA de Austria (1,) = 5.045.
PIBA máximo (Estados Unidos) = 5.075.
PIBA mínimo (Zaire) = 367.

A partir de aquí se computan los I_t^*:

$$I_1^* = \frac{78,6 - 74,8}{78,6 - 42,0} = 0,104$$

$$I_2^* = \frac{3,00 - 2,90}{3,00 - 0,13} = 0,033$$

$$I_3^* = \frac{5.075 - 5.045}{5.075 - 367} = 0,006$$

El paso final consiste en aplicar [4.4]: computar el complemento respecto de 1 del promedio simple de los tres indicadores antes calculados:

$$IDH = 1 - \frac{0,104 + 0,033 + 0,006}{3} = 1 - 0,048 = 0,952$$

Este indicador es un claro ejemplo de *VS* para la que no es posible realizar una validación ni por concurrencia ni por predicción, tal y como se discute en las secciones precedentes. Es posible como en cualquier otro caso, evaluar la validez de contenido. No conozco, sin embargo, ningún esfuerzo riguroso por parte del PNUD orientado a sondear el consenso existente sobre si las variables integrantes han de ser las que esa agencia ha elegido. Por ejemplo, algunos consideran absolutamente improcedente el uso del *PIB* per cápita, entre otros motivos por la consabida razón de que detrás de este promedio se ocultan las enormes desigualdades que suele haber en la distribución de la riqueza dentro de un país (Tapia, 1995). Diversas valoraciones críticas del IDH aparecieron en la revista especializada *World Development* casi desde el momento de su creación (Bhanoji Rao, 1991; McGillivray, 1991; Hopkins, 1991).

Por otra parte, incluso admitidos los componentes que se han seleccionado para sintetizar el desarrollo humano, el *IDH* carece de validez de aspecto.

Por ejemplo, a mi juicio, el truco de usar el *PIBA* en lugar del *PIB es* simplemente inaceptable. Cualquiera sospecha que alguna razón existirá para que casi nadie -ni una persona, ni un país- esté dispuesto a renunciar a esos dólares que, según PNUD, son superfluos para llevar una vida digna. Examinemos detenidamente la función a través de la cual se consuma la transformación. La Tabla 4.2 refleja numéricamente el papel que ella desempeña en la transformación del *PIB.*

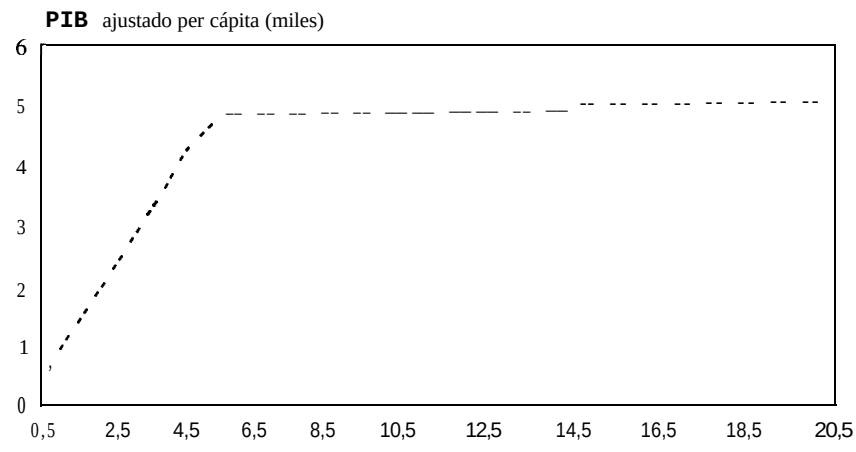
Se aprecia, por ejemplo, que el *PIB* per cápita de Uruguay es el 27,6% del de Estados Unidos; tras el ajuste, resulta que el valor del indicador que se ha tomado como representante del poder adquisitivo del ciudadano uruguayo es el 96,4% del de un estadounidense, dato que seguramente llenaría de estupor a los uruguayos.

Tabla 4.2. ***Valores del PIBA para países escogidos con diferentes niveles de PIB per cápita para*** 1993

País	PIB	PIBA
Estados Unidos	21.449	5.075
Suiza	20.874	5.074
Alemania	18.213	5.050
Austria	16.504	5.045
España	11.723	5.006
Portugal	8.770	4.955
República Checa	7.300	4.928
Venezuela	6.169	4.902
Uruguay	5.916	4.895
Panamá	3.317	3.317
Bangladesh	872	872
Burundi	625	625
Burkina Faso	618	618
Etiopía	369	369

Una idea quizás más expresiva del papel de la función [4.3], se obtiene a través de la Figura 4.1. Obsérvese, tanto en la Tabla como en la Figura que, en esencia, la decisión de ***ajustar*** el ***PIB*** a través del ***PIBA y*** de operar en lo sucesivo con este último, es equivalente a decidir que, cuando el ***PIB*** sea alto -superior a cierta cantidad, en este caso al «umbral de pobreza»- se decide sin más... que no es alto. Tal decisión se concreta sustituyendo el ***PIB*** por un número que cambia de manera casi imperceptible respecto de dicho umbral (4.829 dólares).

El efecto de este manejo es el de reducir artificialmente la estimación de la verdadera distancia que separa a los países pobres de los ricos. Si no se hiciera esa pudorosa (quizás sea más justo decir impúdica) «atenuación», entonces la diferencia entre el IDH de los ricos y el de los pobres sería mucho más marcada. Por una parte, ese sería el procedimiento legítimo, ya que es el que resulta de aplicar el verdadero producto interno (supuesto que pasamos por alto otras críticas que pudieran caber al ***PIB) y*** no un amanado engendro tecnicista como el que se esconde detrás de la fórmula [4.3]. Por otra parte, una diferencia mucho más acusada reflejaría la verdadera y bochornosa separación que hay, por ejemplo, entre



NOTA: El PIBA se computó con la fórmula [4.3] usando el «umbral de pobreza» igual a 4.829

Figura 4.1. *Producto interno bruto per cápita ajustado en función del producto interno bruto per cápita.*

Etiopía y Austria. La lectura que puede hacerse de la justificación del PNUD es más 0 menos ésta:

> LA DIFERENCIA POR LA QUE HAY QUE PREOCUPARSE ES LA QUE CONCIERNE A TENER O NO UN VASO DE LECHE, A DISPONER O NO DE AGUA POTABLE. PERO LA DIFERENCIA EXPRESADA EN YATES, APARATOS ELECTRÓNICOS, VACACIONES EN MARBELLA, COCHES LUJOSOS Y MANJARES NO TIENE NINGUNA IMPORTANCIA POR QUE NADA DE ESO HACE FALTA REALMENTE. DE MANERA QUE HAREMOS UNA MANIPULACIÓN ARITMÉTICA QUE LA HAGA DESAPARECER.

Así, como por arte de magia, Austria no tiene realmente un opulento PIB per cápita de 16.504 dólares sino que posee un austero PIBA de 5.045. ¡Austria está a 216 dólares del umbral de pobreza! Esta limitación del índice es a mi juicio de la máxima importancia, porque una de las pocas cosas para las que el IDH puede ser útil es para, justamente, evaluar la distancia norte-sur.

4.5.4. **Índice de desarrollo en salud comunitaria**

El ejemplo de la sección precedente sirvió de inspiración para que Silva y Cuéllar (1995) desarrollaran un algoritmo de construcción de lo que denominaron como ***índice de desarrollo en salud comunitaria (IDSC).*** La línea de razonamiento central y el algoritmo de construcción obtenido se bosquejan a continuación.

Se toman como punto de partida dos bases conceptuales. Por una parte, la conocida definición de ***salud*** acuñada por la ***OMS*** a mediados de siglo como un ***estado de completo bienestar físico, social y mental y no sólo la ausencia de enfermedad.*** Mirada desde una perspectiva comunitaria, esta definición subraya la improcedencia de estudiar a un colectivo humano como un conjunto de entes anátomo-fisiológicos, aislados de su ambiente. Su propio contenido obliga a ubicar el fenómeno de la salud dentro de las coordenadas históricas y culturales que correspondan. Los seres humanos son seres eminentemente sociales; consecuentemente, los riesgos y tensiones a los que se enfrentan como resultado de esa socialización adquieren inevitablemente una importancia creciente en temas relacionados con su salud. Por otra parte, se tomó en cuenta la consideración de Dowell y Newell (1987) en el sentido de que una medición integrada de salud habrá de vertebrarse a partir de la combinación de cierto número de indicadores de salud, cada uno de los cuales represente adecuadamente alguna dimensión del concepto global.

Se trataba entonces de identificar indicadores que representaran las esferas biológica, psicológica y social, con atención a la noción de bienestar social y emocional, que luego serían objeto de un proceso de síntesis. En busca de la racionalidad de este indicador y de que su aplicación fuese ágil y factible, se procuró reducir tanto como fuese posible ese grupo inicial de indicadores. El desafío era notable dada la complejísima dimensión del concepto que se quería sintetizar; pero era imprescindible asumirlo a fin de resolver del mejor modo la contradicción inherente a todo modelo: la aspiración de que sea manuable y simple por una parte, y la necesidad de que sea suficientemente complejo como para conferirle realismo y expresividad por otra.

En el trabajo de Silva y Cuéllar (1995) se hace una aplicación del ***IDSC*** a las 15 provincias de Cuba; allí se explica en detalle la base racional para seleccionar los siguientes 13 indicadores de partida, que se consideraron pertinentes y factibles en ese caso [13]:

1. Tasa de mortalidad de menores de cinco años.
2. Índice de bajo peso al nacer.

[13] Allí se fundamenta, por ejemplo, la inconveniencia en el caso de Cuba de admitir el *porcentaje de alfabetismo,* que estaba originalmente contemplado en un conjunto de 14 indicadores, porque es virtualmente idéntico para todas las provincias.

3. Tasa bruta de mortalidad por enfermedades diarreicas.
4. Tasa de incidencia de sífilis.
5. Consultas estomatológicas por cada 1000 habitantes.
6. Número de abortos por cada 1000 mujeres entre 12 y 49 anos.
7. Tasa bruta de divorcialidad.
8. Tasa bruta de mortalidad por suicidio.
9. Número de enfermeras por 1000 habitantes.
10. Tasa bruta de mortalidad por enfermedades del corazón.
11. Consumo per cápita de electricidad.
12. Porcentaje de menores de 2 años adecuadamente vacunados.
13. Porcentaje de población servida con agua adecuada.

La estrategia general para la construcción de tal indicador de desarrollo, susceptible de ser aplicado a un conjunto dado (supondremos que son N en total) de unidades político-administrativas, espaciales o comunitarias, fue la de hacer una elaboración gradual. Partiendo de ese conjunto de indicadores, decididos teóricamente, y cuyos valores son conocidos para todas y cada una de las n unidades involucradas, se empieza por elegir aquel que se considere el más representativo a los efectos del concepto cuya medición se quiere sintetizar. A partir de ahí se van adicionando indicadores de manera que, con cada uno, se construye una ***expresión parcial del IDSC. Se*** evalúa en cada paso si el indicador adicionado a dicha expresión parcial hace una contribución informativa apreciable respecto de lo que hasta ese momento se había conseguido. De ser así, se prueba con la adición de otro nuevo indicador; en caso contrario, se da por concluido el proceso. El procedimiento detallado se expone a continuación.

Supondremos que el número de indicadores iniciales es r (en el ejemplo anterior $r = 13$ y $N = 15$). Para exponer el método hay que realizar tres definiciones previas:

1.º Privación relativa de salud

Para cada indicador se computa ***la privación relativa de salud*** que le corresponde a cada unidad. Esta es una medida de cuánto dista el valor del indicador para esa unidad, en términos relativos, del mejor valor alcanzado entre las restantes unidades.

Dos posibilidades han de considerarse al efectuar el cálculo de la privación relativa: existen indicadores para los cuales lo deseable es que tomen valores bajos (por ejemplo, mortalidad en el menor de cinco años o índice de bajo peso al nacer), y otros para los que, por el contrario, lo ideal es, en principio, que alcancen valores altos (tales como camas hospitalarias por 1.000 habitantes o consultas estomatológicas por 1.000 habitantes).

La privación se denotará genéricamente mediante P_{ij} (privación correspondiente al indicador ***i*** para la unidad j) y se define, formalmente, del modo que sigue.

Para el primer caso, la fórmula utilizada es:

$$P_{ij} = \frac{X_{ij} - X_{im}}{X_{iM} - X_{im}}$$

en tanto que para el segundo es:

$$P_{ij} = \frac{X_{iM} - X_{ij}}{X_{iM} - X_{im}}$$

donde X_{ii} es el valor del indicador ***i*** para la unidad j, X_{iM} es el máximo X_{ij} entre las *n* unidades y X_{im} es el mínimo.

2.º Índice parcial de desarrollo en salud

Si consideramos ***k*** de los *r* indicadores iniciales, se define lo que llamaremos ***IDSC parcial de orden k*** para la unidad j del modo siguiente:

$$I_{kj} = 1 - \frac{1}{k} \sum_{i=1}^{k} P_{ij} \qquad [4.5]$$

Dado que, por definición, se tiene que $0 \leq P_{ij} \leq 1$, es evidente que I_{kj} también cumple tal condición; y puesto que, cuanto mayor sea P_{ij}, peor es la posición de la unidad j en cuanto al indicador *i*, con I_{kj} se produce la situación opuesta: cuanto mayor sea, mejor es la situación de salud de la unidad j en lo que concierne a los ***k*** indicadores tomados en conjunto.

3.º Distancia promedio entre dos vectores

Si se tienen dos vectores:

$$X = (X_1, X_2, \ldots, X_N)$$
$$Y = (Y_1, Y_2, \ldots, Y_N)$$

se define la ***distancia promedio*** entre X y Y del modo siguiente:

$$D(X, Y) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (X_j - Y_j)^2}$$

A partir de estas tres definiciones preliminares, los pasos que se siguen para construir el ***IDSC*** definitivo son los siguientes:

1. Elegir aquel indicador aislado que teóricamente se considere el que mejor resuma el desarrollo en materia de salud; calcular entonces

$$I_{1j} = 1 - P_{1j}$$

para las n unidades ($j = 1, \ldots . N$).

2. Computar $r - 1$ valores de I_{2j} usando, en cada caso, el indicador elegido en el paso (1) y cada uno de los otros $r - 1$ indicadores; es decir, aplicar $r - 1$ veces la fórmula [4.5] para $k = 2$ indicadores: el primero elegido y cada uno de los restantes.

   Seguidamente se computa la distancia entre el vector de los I_1 para las n unidades y los vectores correspondientes a cada una de los $r - 1$ ***IDSC*** parciales de orden 2 arriba considerados; es decir, computar $r - 1$ veces:

$$D(I_1, I_2) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (I_{1j} - I_{2j})^2}$$

   Se elige como ***IDSC*** parcial de orden 2 aquel I_2 que más diste de I_1. Así quedan incorporados los dos primeros indicadores definitivos del ***IDSC.***

3. Computar I_{3j} en $r - 2$ oportunidades: en cada una de ellas se involucran los 2 indicadores ya elegidos y, sucesivamente, cada uno de los $r - 2$ no incorporados aún. Se calculan nuevamente las distancias, ahora entre el vector I_2 y los $r - 2$ vectores correspondientes a I_3. Se identifica el vector más alejado de I_2.

El proceso se repite hasta que se identifique un conjunto de k indicadores cuyo I_k correspondiente sea tal que, cualquiera que sea un nuevo indicador que se adicione (de los $r - k + 1$ aún no incorporados), el I_{k+1} resultante diste de I_k en una magnitud despreciable. Se consideró razonable adoptar la regla de detener el proceso cuando tal distancia tomase un valor menor que 0,05 para los $r - k + 1$ valores de I_{k+1}. Nótese que la máxima distancia que puede haber entre los ***IDSC*** parciales es 1, de modo que el umbral 0,05 es un 5% de ese máximo.

Puesto que I_{k+1} contiene siempre a los k indicadores incorporados en I_k es razonable esperar [14] que, a medida que se incorporan indicadores al ***IDSC*** parcial, la distancia entre índices parciales sucesivos vaya disminuyendo; o, dicho de otro modo,

[14] Así ocurrió en el ejemplo que se desarrolló para las provincias cubanas que se reproduce en Silva y Cuéllar (1995).

que el método sea convergente. La demostración formal al respecto, sin embargo, está aún pendiente. Una vez completado el proceso, se habrá computado I_k para las ***n*** unidades consideradas, de manera que por su conducto ellas pueden ordenarse de menor a mayor, además de que se tendrá una cuantificación del nivel integrado de salud que le corresponda a cada una. Esta información pudiera constituir, por ejemplo, un recurso para la definición de prioridades en el contexto de un programa, un elemento para el examen evolutivo de estas unidades y otros procesos valorativos análogos.

Bibliografía

Almeida N (1992). ***Epidemiología sin números.*** Serie Paltex n.º 28, OPS/OMS, Washington.

Ansell JI, Philips MJ (1994). ***Practica1 methods for reliability data analysis.*** Claredon, Oxford.

Apgar V (1953). ***Proposal for method of evaluation of newbom infant.*** Anesthesiology and Analgesics 32: 260-267.

Bhanoji Rao VV (1991). ***Human development report 1990: review and assessment.*** World Development 19: 1451-1460.

Bowling A (1994). ***La medida de la salud: revisión de las escalas de medida de la calidad de vida.*** Masson, Barcelona.

Brennan P, Silman A (1992). ***Statistical methods for assessing observer variability.*** British Medical Journal 304: 1491-1494.

Bland JM, Altman DG (1986). ***Statistical methods for assessing agreement between two methods of clinical measurement.*** Lancet a: 307-310.

Candela AM (1992). ***Validación de aparatos y métodos de medida: concordancia sí, correlación no.*** Carta al director. Medicina Clínica 99:314.

Carmines EG, Zeller RA (1979). ***Reliability and validity assessment.*** Sage Publications, Beverly Hills.

Cohen J (1960). ***A coefficient of agreement for nominal scales.*** Educational and Psychological Measurement 20: 37-46.

Cronbach L (1951). ***Internal consistency of tests: analyses old and new.*** Psychometrika 16: 297-334.

Cronbach L, Meehl P (1955). ***Construct validity in psychological test.*** Psychological Bulletin 52: 281-302.

Chambless LE, Dobson A , Patterson CC, Raines B (1990). ***On the use of a logistic risk score in predicting risk of coronary heart disease.*** Statistics in Medicine 9: 385-396.

Dowell 1, Newell C (1987). ***Measuring Health,*** Oxford University Press, New York.

Dunn G (1989). ***Design and analysis of reliability studies.*** Oxford University Press, New York.

Feinstein R (1971). ***On exorcising the ghost of Gauss and the curse of Kelvin.*** Clinical Pharmacology and Therapeutics 12: 1003-1016.

Feinstein R (1985). ***A bibliography of publications on observer variability*** Journal of Chronic Diseases 38: 619-631.

Fleiss JL (1981). ***Statistical methods for rates and proportions,*** 2.ª ed, Wiley, New York.

García-García JA (1995). ***El índice de calidad de vida de Spitzer: ¿validado?*** Carta al editor. Medicina Clínica *105*:319.

González M, Laborde R, Pino A, Silva LC (1991). ***Hábitos de prescripción en los médicos del programa de atención primaria al adulto.*** Revista Cubana de Salud Pública 17: 68-73.

Gutiérrez T, Latour J, López V, Bonastre J, Giner JS, Rodríguez M ***et al.*** (1994). ***Efecto de los factores sociales sobre la calidad de vida de los supervivientes de un infarto de miocardio.*** Medicina Clínica *103*:766-769.

Harman HH (1994). ***Modern factor analysis.*** University of Chicago Press, Chicago.

Hopkins M (1991). ***Human development revisited: a new UNDP report.*** World Development 19: 1469-1473.

Kelly TL (1924). ***Statistical method.*** McMillan, New York.

López C (1994). ***Indice de desarrollo humano: el caso Cuba.*** Boletín del Ateneo «Juan César García» 2: 3-37. Organización Panamericana de la Salud, La Habana.

McGillivray M (1991). ***The human development index: yet another redundant development indicator?*** World Development 19: 1461-1468.

Mercer H (1978). ***La prescripción de medicamentos.*** Revista Salud Problema UAM-X 2: 8-13.

Mora-Macía J, Ocón J (1992). ***Carta al director*** Medicina Clínica *99*:314.

OMS (1985). Uso ***racional de los medicamentos.*** Informe de la Conferencia de Expertos, Nairobi.

Orduñez PO, Espinosa AD, Alvarez OM, Apolinaire JJ, Silva LC (1993). ***Marcadores múltiples de riesgo para enfermedades crónicas no transmisibles.*** Informe técnico ISCM/H, La Habana.

PNUD (1992). ***Catalizadores de la cooperación. Desarrollo humano y erradicación de la pobreza.*** Naciones Unidas, Nueva York.

Rosenberg H (1994). ***El índice de desarrollo humano.*** Boletín de la Oficina Sanitaria Panamericana 117: 175-181.

Shaper AG, Pocock SJ, Phillips AN, Walker M (1987). ***A scoring system to identify, men at high risk of heart attack.*** Health Trends 19: 37-39.

Schuling J, de Haaen R, Limburg M, Groenier KH (1993). ***The frenchay activities index; assesment of functional status in stroke patients.*** Stroke 24: 1173-l 177.

Silva LC, Cuéllar I(l995). ***Una medición integrada de salud comunitaria.*** Boletín del Ateneo «Juan César García» Volumen 3 (en prensa) Organización Panamericana de la Salud, La Habana.

Silva LC, Nogueiras (1991). ***Conocimiento farmacoterapéutico del médico de la familia.*** Tesis de Maestría en Salud Pública, Facultad de Salud Pública, La Habana.

Spitzer W, Dobson A, Hall J, Chesterman E, Levi J, Shepherd R ***et al.*** (1978). ***Measuring the quality of*** life ***of cancer patients. A concise QL-index for use by physicians.*** Journal of Chronic Diseases *34:*585-597.

Streiner DL, Norman GR (1989). ***Health measurement scales: a practical guide to their development and use.*** Oxford University Press, New York.

Tapia JA (1995). ***Algunas ideas críticas sobre el índice de desarrollo humano.*** Boletin de la Oficina Sanitaria Panamericana 119: 74-87.

Thurstone LL (1931). ***The reliability and validity of tests.*** Edward Bross, Ann Arbor.

Weiner EA, Stewart BJ (1984). ***Assessing individuals,*** Little Brown, Boston.

5

Estadísticamente posible; conceptualmente estéril

El aborto se está haciendo tan popular en algunos países que la espera para conseguir uno se está alargando rápidamente. Los *expertos* ***predicen que, a ese ritmo, pronto habrá un año de espera para lograr un aborto.***

JOHN ALLEN PAULOS

Lo usual no es que los reclamos técnicos de los estadísticos sean resultado de innecesarios melindres formales, atribuibles -por lo demás- a la mediocridad de sus cultores. No debe confundirse la defensa de la audacia y de la creatividad, ni el ataque a los dogmas, con el apoyo al liberalismo metodológico.

La aplicación acrítica o mecánica de las técnicas estadísticas suele conducir tanto a valoraciones equivocadas como al callejón sin salida de lo ininterpretable. Este capítulo tiene como propósito alertar sobre ese fenómeno e ilustrarlo con algunos ejemplos que recorren diversas esferas conceptuales y varias áreas de aplicación.

5.1. La necesidad de un espacio epidemiológico

En cierta ocasión fui consultado por una funcionaria ministerial, responsable de salud de un área urbana que abarcaba a unas 3.000 familias. Entre otros aspectos de su interés, me expresó su inquietud por el crecimiento que había experimentado la tasa de mortalidad infantil en el área que estaba bajo su responsabilidad.

Durante el año anterior al momento de la entrevista habían nacido en el área 125 niños, y había muerto, sólo un niño antes de cumplir el año; durante el año en curso, sin embargo, habían nacido hasta ese momento 114 niños, y había muerto un total de 3 menores de un año. Formalmente, la tasa de mortalidad infantil *(TMI)* había ascendido entonces de 8,0 por 1.000 nacidos vivos (la *TMI* de Noruega) a 31,9 (la *TMI* de Surinam). Mi interlocutora consideraba muy «injusta» esta evolución ya

que, si bien uno de los niños había fallecido por una enteritis a los tres días de nacido, los otros dos eran hermanos gemelos de diez meses, que habían fallecido recientemente en un mismo accidente debido a un trágico descuido de los padres.

En el razonamiento de mi consultante se estaba produciendo una curiosa inversión: la tasa de mortalidad infantil, indicador indirecto e integrado que supuestamente refeja (véase Freedman, 1991) el nivel alcanzado por una sociedad o comunidad en materia de atención materno-infantil y de desarrollo económico-social, está siendo considerado como un ***gestor*** del desarrollo. Ante tales apreciaciones, da la impresión de que se piensa que el nivel de atención y de desarrollo pudiera modificarse ***como consecuencia*** de un cambio en el valor obtenido para el indicador, independientemente de cuál fuese la razón o la vía por la cual ese cambio se produjo [1]. Tal parecía que mi colega estuviera convencida, en fin, de que si el accidente se hubiera evitado, entonces el nivel de desarrollo se hubiera mantenido alto.

El solo hecho de que sea posible contar la anécdota inherente a los casos que conforman el numerador de la tasa (un niño falleció por una enteritis en tanto que otros dos murieron en un mismo accidente evitable) subraya la improcedencia de calcularla.

En general puede decirse que allí donde hay margen para la anécdota, no lo hay para la estadística, y viceversa. Si se está examinando la mortalidad infantil de un país donde, por ejemplo, nacen 150 mil niños al año y donde mueren alrededor de 3.000, ningún hecho anecdótico cambiará la tasa, que se mantendrá en un entorno de 20 muertos por cada 1.000 nacidos vivos. Esa es la tasa ***que le corresponde*** a ese país de acuerdo con su nivel de atención materno-infantil y a su desarrollo económico social. Que se produzca o no un accidente como el mencionado no habrá de modificarla; del mismo modo que un hecho fortuito adicionó 2 muertos, algún otro hecho fortuito operaría en el sentido inverso (por ejemplo, 2 o 3 niños que nacieron con 600 gramos de peso se salvan contra todo pronóstico). Y aun en caso de que no se produjera tal «compensación», tampoco la tasa experimentaría una modificación apreciable.

La tasa de mortalidad es la expresión de una realidad; de hecho, reflejará una ley socio-sanitaria siempre que se dé a otra ley, la de los grandes números, ***espacio*** para expresarse.

Esto trae a colación un concepto de aliento más general: el de ***espacio epidemiológico.*** Para poder evaluar un problema desde una perspectiva epidemiológica es preciso tener una masa crítica de información, un «espacio» mínimo: sea físico, cuantitativo o temporal. En el ejemplo precedente, por ejemplo, no había suficiente «espacio cuantitativo» como para esperar un comportamiento estable del indicador.

Cabe subrayar que, para algunas prácticas derivadas de viejas presunciones teó-

[1] Nótese que el tránsito del nivel de Noruega al de Surinam se produjo en un solo día; de hecho, en un solo minuto.

ricas, sólo ahora es cuando se empieza a contar con el *espacio epidemiológico* -en este caso, temporal- necesario para una evaluación ecuánime y cabal.

Por ejemplo, para aquilatar los dividendos reales de la mamografía como práctica preventiva masiva, que se vertebró alrededor de un entramado teóricamente impecable, hubo que esperar un buen número de años: los necesarios para poder computar tasas confiables de mortalidad. La generación de un espacio adecuado (una masa crítica de información) ha permitido a Greenberg y Stevens (1986) llamar la atención sobre el hecho de que las tasas de mastectomías, incrementadas extraordinariamente en los Estados Unidos hasta niveles muy superiores a los del Reino Unido, no se haya traducido en tasas de mortalidad esencialmente diferentes entre ambos países. Naturalmente, esta constatación no permite, *per se,* sacar conclusiones definitivas [2], pero aporta sin duda una buena razón para mirar con suspicacia el radical tratamiento mencionado.

5.2. Los algoritmos no piensan

Para ilustrar a dónde puede conducir la aplicación mecánica de un procedimiento estadístico, nos internaremos con cierto detalle en un problema práctico tomado de Silva (1995).

Imaginemos que se estudia el efecto de la edad de la madre sobre el peso del recién nacido. La hipótesis en juego es que las madres en edades extremas -adolescentes o añosas- tienden a producir, con más frecuencia que las que se hallan en edades reproductivas intermedias, niños con peso por debajo de lo conveniente.

Tratándose de dos variables continuas, se podría pensar, en principio, en la realización de un análisis de correlación entre la edad materna y el peso del recién nacido. Sin embargo, al menos dos razones conducen a desechar la idea:

a) Se considera aceptable un peso al nacer de 2.500 gramos; un peso mayor que ese umbral no es necesariamente indicio de que el niño posee mejor estado de salud. Ello aconsejaría manejar esta variable a nivel dicotómico: bajo peso y normopeso [3].
b) El manejo de la edad de manera directa podría producir que la posible relación existente se diluyera, debido a que en el tramo de edades fisiológicamente adecuadas para el parto -de 18 a 34 años- no cabe esperar teóricamente que la edad tenga capacidad explicativa para la variación en el peso del recién nacido. La variabilidad en el peso de los niños cuyas madres tienen edades en tal recorrido podría deberse, además de a las natu-

[2] Ver la Sección 8.4 en la que se aborda la ***falacia ecológica***

[3] Nótese, de paso, que estamos ante un ejemplo en que resulta recomendable perder información para conseguir un análisis más claro.

rales diferencias biológicas, a variables tales como peso, talla y edad gestacional de la madre, su alimentación durante el embarazo, o el uso de medicamentos teratogénicos, que serían objeto de consideración en una segunda fase del análisis, pero difícilmente a la edad como tal.

Supongamos que se tiene en cuenta sólo la primera de estas dos consideraciones y que, consecuentemente, se maneja la variable de respuesta a nivel dicotómico. La idea de utilizar la regresión logística, de manera que se ajuste la probabilidad de ser un «bajo peso» en función de la edad materna (y quizás de algunas otras, si se encara el problema multifactorialmente), no tardará en comparecer [4].

Si se decide hacer un ajuste univariado de la regresión logística y no se contempla la segunda objeción, se procederá a estimar α y β en el modelo logístico:

$$P(BP) = [1 + \exp(-\alpha - \beta X)]^{-1}$$

donde X representa la edad materna medida en años cumplidos y *P(BP)* es la probabilidad de que el niño nazca con menos de 2.500 gramos.

Como los algoritmos no piensan, ni un programa computacional tiene la capacidad de protestar sino que «funciona» siempre que sea aplicado según las demandas formales de su arquitectura, la regresión puede llevarse adelante siempre que se tengan parejas de datos (BP_i, X_i) donde BP_i ***es 1 o 0,*** en dependencia de que el niño de la i-ésima madre tenga o no menos de 2.500 gramos respectivamente, y X_i es la edad en años de esa madre.

Para que la regresión logística tenga un sentido claro, las variables explicativas deben guardar una relación monótona con la probabilidad del evento que se estudia. Esto quiere decir que la dependencia entre la variable explicativa y *P(BP)* tiene que ser, o bien directa, o bien indirecta. Dicho de otro modo: debe evitarse incluir una variable X tal que *P(BP)* aumente para cierto recorrido de sus valores, y disminuya para otro segmento de valores posibles. Este fenómeno indeseable se produce, obviamente, con la edad de la madre y la probabilidad de que el hijo sea «de bajo peso». En esa situación cabe esperar que *P(BP)* disminuya en la medida X que se acerca desde la izquierda al intervalo (18-34) y crezca para edades superiores a 35 (madres añosas).

En una situación como ésta, lo que realmente interesa evaluar no es el efecto de ***la edad de la madre*** como tal sino, más bien, el que tiene la distancia entre la edad real y la «edad óptima» para el parto. Si se admitiera, por ejemplo, que 25 años es esa «mejor edad», en lugar de *X*, podría usarse una «edad corregida»: $X_c = |X-25|$.

Si se considera que cualquier punto del intervalo (18,34) es igualmente aceptable, podría hacerse una definición algo más elaborada; por ejemplo la siguiente:

[4] El disparate de realizar un ajuste lineal múltiple (o simple) cuando la variable de respuesta es binaria no será valorado por nadie que posea un conocimiento mínimo de estos recursos.

$$X_c = \begin{cases} 18\text{-}X \text{ si } X<18 \\ 0 \text{ si } 18\leq X\leq 34 \\ X\text{-}34 \text{ si } X>34 \end{cases} \qquad [5.1]$$

Así definida, X_c aumenta en la medida que X se aleja del intervalo admitido como aceptable para tener hijos, sea por la izquierda o por la derecha.

Otra alternativa sería ajustar la curva logística agregando un término cuadrático:

$$\boldsymbol{P}\,(\boldsymbol{Y} = 1) = [1 + exp\ (-\alpha - \beta X - \gamma X^2)]^{-1}$$

Sin embargo, puede demostrarse que en el ejemplo que sigue el ajuste es mejor cuando se emplea la edad corregida definida en [5.1].

En Silva (1995) se considera una cohorte de 8.012 embarazadas cuyas edades se ubican entre los 14 y los 41 años; se calcula la distribución de las madres según edades simples y las tasas de incidencia de bajo peso observadas en cada una de esas edades. El resultado se resume en la Tabla 5.1.

Tabla 5.1. ***Distribución de una cohorte de 8.012 embarazadas según edades simples y tasas de prevalencia de bajo peso***

Edad	N.º de embarazadas	N.º de niños con bajo peso	Tasa de bajo peso	Edad	N.º de embarazadas	N.º de niños con bajo peso	Tasa de bajo peso
14	46	19	41,3	28	490	42	8,6
15	83	22	26,5	29	372	39	10,4
16	125	31	24,8	30	311	35	11,3
17	264	38	14,4	31	287	37	12,9
18	312	37	11,8	32	200	29	14,5
19	394	40	10,1	33	210	36	17,1
20	420	39	9,3	34	192	27	14,1
21	514	46	8,9	35	108	16	14,8
22	487	44	9,0	36	97	16	16,5
23	501	39	7,8	37	101	20	19,8
24	512	48	9,4	38	82	21	25,6
25	623	55	8,8	39	76	23	30,3
26	618	63	10,2	40	50	24	48,0
27	511	60	11,7	41	26	17	65,3

Al ubicar las tasas de bajo peso por edades simples en un gráfico, se aprecia que el patrón de modificación de la probabilidad empírica de ser un bajo peso como fun-

ción de la edad dista de ser monótono. Los resultados tienen el aspecto que refleja la Figura 5.1.

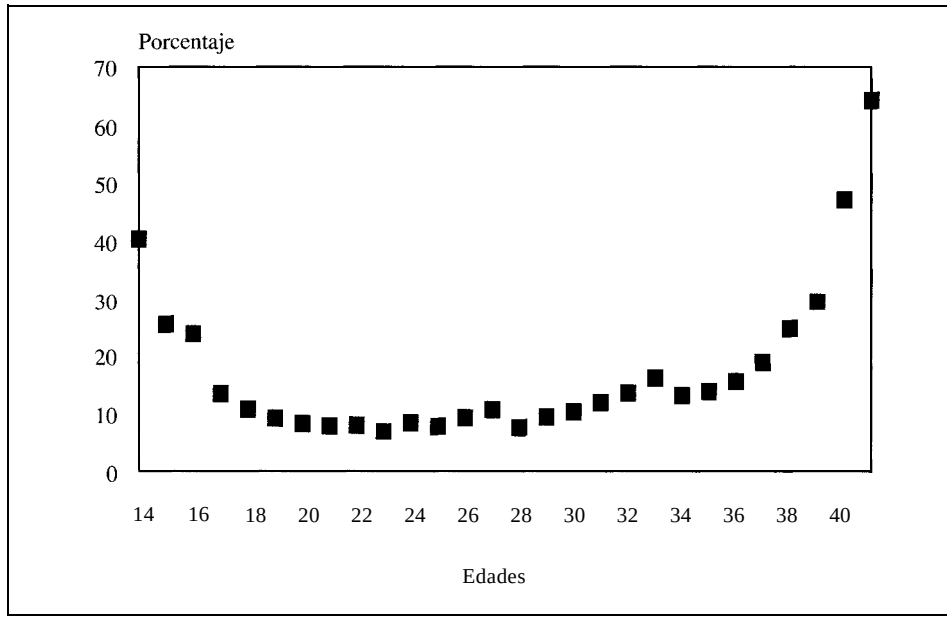


Figura 5.1. Tasas de niños con bajo** peso **según edad de la madre.

Si se aplica la regresión logística univariada a estos 8.012 pares de valores, la técnica fluye bien ***desde el punto de vista formal o aritmético;*** concretamente, se obtienen las siguientes estimaciones para α y β, respectivamente:

$$a = -2{,}83 \qquad b = 0{,}03$$

Al dibujar la curva se aprecia (ver Figura 5.2) que el ajuste a los datos empíricos es pésimo.

Siguiendo el enfoque clásico, al evaluar la hipótesis $\beta = 0$ mediante el ***test de Wald (se(b) =*** 0,06; ***Z = 5,0;*** $P<0{,}001$), se rechaza enfáticamente la hipótesis de nulidad para este parámetro; por tanto, teniendo en cuenta el signo de ***b se*** diría que ***la edad materna influye positivamente en la probabilidad de ocurrencia de un bajo peso:*** a más edad, más riesgo de que se produzca esta anomalía.

Esto es falso, tanto teórica como empíricamente, para un tramo de la vida, precisamente para el que más embarazos aporta: el que va de 14 a 34 años. Si se hubiera hecho un gráfico como el de la Figura 5.2, o se hubiese aplicado una prueba de bondad de ajuste, quizás no se hubiera evaluado siquiera la hipótesis sobre el valor de β, y se hubiera evitado así el error. Pero si ello condujese a la conclusión de que «no hay relación entre la edad materna y la condición anómala en el peso al nacer», se estaría cometiendo otro error.

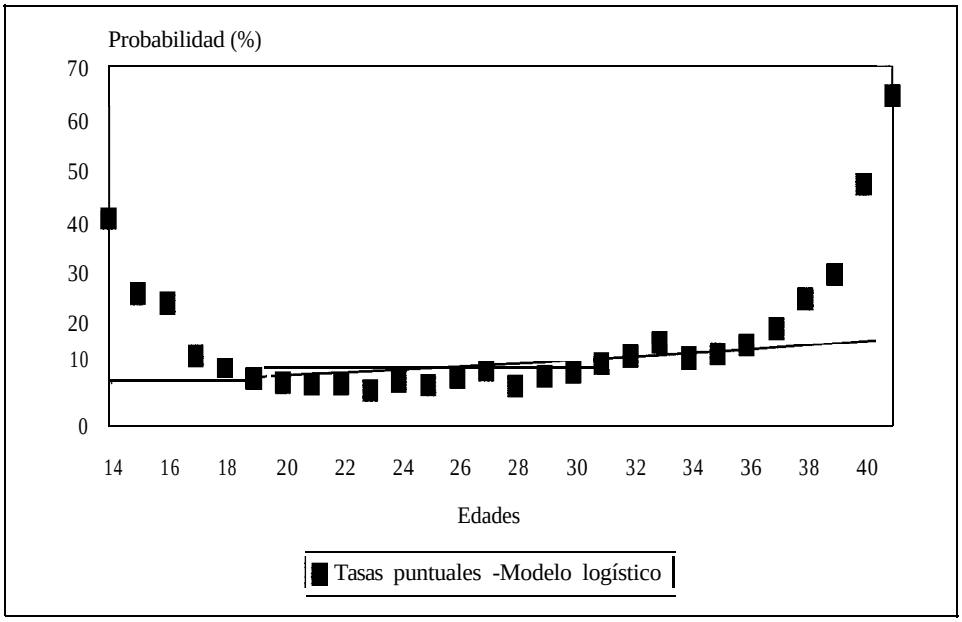


Figura 5.2. ***Probabilidad empírica y ajustada de bajo peso según edad materna.***

Si se aplicara la segunda variante de modificación para el manejo de la edad materna según la fórmula [5.1], la relación entre X y X_c sería como refleja la Tabla 5.2.

Tabla 5.2. ***Relación entre la edad corregida y la edad materna***

X	de 18 a 24	17035	16036	15037	14038	39	40	41
X_c	0	1	2	3	4	5	6	7

Usando ahora los datos de la Tabla 5.1, se deducen las tasas correspondientes a los 8 valores de X_c, tal y como recoge la Tabla 5.3.

Al ajustar el modelo logístico a las 8.012 parejas (BP_i, X_{Ci}), la estimación de α y β viene dada por los valores siguientes:

$$a = -2{,}16 \qquad b = 0{,}34$$

La Figura 5.3 refleja las tasas empíricas dentro de los nuevos grupos (definidos por los valores posibles de X_c) y el ajuste realizado. La calidad del ajuste, como se aprecia claramente, es satisfactoria y -en cualquier caso- abismalmente mejor que la conseguida con la edad sin modificar.

Tabla 5.3. ***Distribución de una cohorte de embarazadas según edad con-egida y tasas de prevalencia de bajo peso***

Edad corregida	N.º de embarazadas	N.º de niños con bajo peso	Tasa de niños con bajo peso
0	6.954	716	10,2
1	372	54	14,5
2	222	47	21,2
3	184	42	22,8
4	128	40	31,2
5	76	23	30,3
6	50	24	48,0
7	***26***	17	63,3

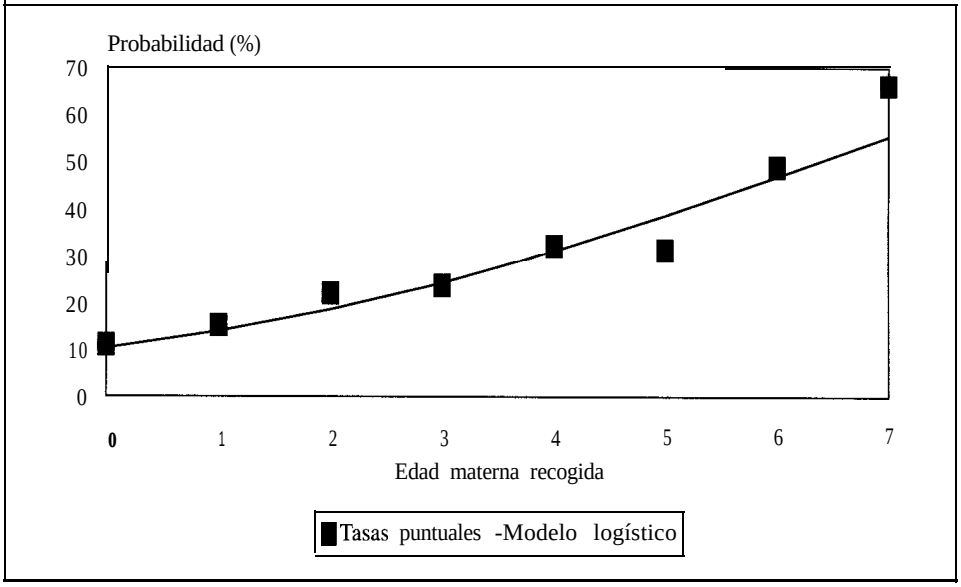


Figura 5.3. ***Probabilidad empírica y ajustada de ser un bajo peso según edad materna corregida.***

La ilustración refleja que ***a quien toca pensar no es al algoritmo sino a su usuario:*** en este caso, la aplicación mecánica de un procedimiento poderoso como la regresión logística produciría una respuesta errónea, y malograría la posibilidad de hallar

un resultado de alto interés: el que surge del uso de esta misma técnica pero tras un proceso de reflexión que contemple el encuadre teórico pertinente.

5.3. El laberinto de la normalidad

5.3.1. El recorrido normal

Un problema de gran trascendencia para la práctica clínica -y, en buena medida, para la epidemiológica- es el llamado ***recorrido de normalidad*** de los parámetros fisiológicos. Por su naturaleza, todos los enfoques para resolver este problema pasan, de un modo LI otro, por el uso de la estadística.

Uno de los procedimientos más utilizados es en extremo simple. Comienza con la selección de una muestra de individuos que supuestamente no tienen patológicamente alterado el parámetro en cuestión. Hacer entonces para cada sujeto la determinación de laboratorio (llamémosle *X),* y computar tanto el valor medio $\bar{x}$ como la desviación estándar *s.* Finalmente, se resta de, y se suma a $\bar{x}$, la magnitud 2*s*. De este modo, el intervalo determinado por los extremos $\bar{x} - 2s$ y $\bar{x} + 2s$ es considerado un ***recorrido de normalidad*** para la dimensión fisiológica de interés.

Por ejemplo, para la creatinina, se plantea que los ***valores normales*** que corresponden a un hombre adulto están dentro del intervalo que va de 7 a 18 mmo1/24h. A partir de ello, un sujeto que tenga un valor por debajo de 7 mmo1/24h o por encima de 18 mmo1/24h sería considerado «anormal».

A continuación se enumeran y comentan varios problemas que cabe considerar ante tan simple y, a la vez, socorrido recurso.

La primera cuestión que debe examinarse es por qué se procede del modo expuesto. El propósito es fijar dos límites *L1 y L2* que tengan los siguientes rasgos:

(A) Por debajo de *L1* queda el 2,5% de los valores.
(B) Entre *L1* y *L2* ha de hallarse el 95% de los datos.
(C) Por encima de L2 se ubica el restante 23% de las observaciones.

Si la distribución de la variable X es ***gausiana o normal,*** entonces los límites $\boldsymbol{Ll = \bar{x} - 2s}$ ***y*** $\boldsymbol{L2 = \bar{x} + 2s}$ cumplen aproximadamente con las tres condiciones enumeradas.

Tal práctica nace de una propuesta de Ronald Fisher, formulada con el afán de identificar los «valores más comunes» (los que se ubican en el intervalo *[Ll, L2]) y* los «poco comunes»: el 5% que se reparte entre los extremos.

La búsqueda de *L1 y L2* que cumplan (A), (B) y (C) es una decisión totalmente arbitraria, surgida de una propiedad de la distribución normal y sin asiento en consideraciones relacionadas con el problema mirado desde la perspectiva clínica o fisiológica.

Con todo derecho uno puede preguntarse: ¿por qué se eligió el 5% para esta-

blecer quiénes están «fuera de la normalidad»? ¿Por qué no 4% o 6,2%? Personalmente pienso que eso está motivado por una razón bastante alejada de consideraciones pertenecientes al ámbito del laboratorio clínico: el número 5, simplemente, tiene un atractivo casi místico, derivado de que ese es el número de dedos que los humanos poseemos en cada mano, del mismo modo que se creó el sistema métrico decimal debido a que poseemos 10 dedos entre las 2 manos. Sospecho que si, en lugar de 5, tuviéramos 6 dedos, el sistema métrico sería duodecimal y Fisher hubiera propuesto usar 92% y no 95% para fijar los extremos del intervalo, y hubiera dejado el 4% de los individuos con valores «poco comunes» a cada lado [5].

Y, a propósito, ¿por qué esa fijación con la simetría? Si se decidió que 9.5 iba a ser el porcentaje de «normales», ¿cuál es la razón para repartir el 5% restante en partes alícuotas? ¿No puede haber situaciones en que convendría buscar ***Ll y L2*** de manera que se cumpla la condición (B) pero para las cuales, en lugar de (A), se plantee que por debajo de ***Ll*** haya el 0,5% de los individuos y, en lugar de (B), que por encima de L2 quede el 4,5% de ellos?

En cualquier caso los límites $\bar{x} - 2s$ y $\bar{x} + 2s$ cumplen las tres condiciones sólo si la distribución de X es normal, rasgo que no tiene por qué producirse, como quedó demostrado hace más de medio siglo por Rietz (1927), en ocasión de su medular trabajo sobre estadística matemática. Es bien conocido que variables tales como el calcio sérico o las proteínas totales siguen distribuciones marcadamente asimétricas. De hecho, autores como Elveback, Guillier y Keating (1970) afirman que ***la mayoría*** de las variables fisiológicas distan de estar en ese caso.

Por otra parte, puede ocurrir que $2s$ sea mayor que $\bar{x}$; en ese caso ***Ll*** sería inferior a cero, situación que, salvo excepciones, carecería de todo sentido, ya que usualmente los parámetros no pueden alcanzar valores negativos.

Una solución para estas dos últimas dificultades sería usar estimaciones no paramétricas de ***Ll y L2;*** concretamente, trabajar con percentiles. Si se ordenan los valores de la muestra de menor a mayor, ***Ll se*** define como aquel número por debajo del cual queda el 2,5% de los sujetos; análogamente, el número por debajo del cual queda el 97,5% se puede definir como L2. Obviamente, por una parte se cumplirán las tres condiciones, sea la distribución gausiana o no, y por otra, la última dificultad (límite inferior negativo) no se presentará nunca.

5.3.2. Cuánto más anormal, ¿más normal?

Sin embargo, el establecimiento de ***Ll y L2*** como límites para establecer la normalidad, cualquiera sea la vía utilizada, entraña algunas contradicciones difíciles de pasar por alto.

[5] Notar que el «equivalente» a $\frac{90}{100}$ en un sistema en base 12 sería $\frac{132}{144} = 0{,}92$.

Se exige partir de una muestra de cierta población que sea considerada, en principio, sana. Pero el método es tal que el 5% de la muestra que se elija para establecer los límites *necesariamente* terminará siendo «anormal», hecho que nos coloca en una contradicción, pues si los integrantes de la población son sanos respecto de ese parámetro, también lo serán los de la muestra, y por tanto ella no puede a la vez contener sujetos con valores patológicos.

Una alternativa pudiera ser la siguiente: tomar la muestra poniendo el máximo escrúpulo en que esté enteramente formada por personas sanas; usar entonces el valor mínimo y el valor máximo que aparezcan en dicha muestra como límites de normalidad. Obviamente, la dificultad inherente a esta variante es que nunca se estará seguro de que no se haya introducido en la muestra una persona clínicamente anormal (con valores patológicos para el parámetro), aunque no sea evidente que está enferma; o bien una sana pero cuyo valor personal sea aberrantemente pequeño o grande. Un solo individuo que se halle en uno u otro caso ya podría modificar radicalmente los límites. En última instancia, si el procedimiento se desarrolla precisamente para poder determinar quién es normal y quién no, es imposible, por lógica, establecer si alguien está en el primero o en el segundo caso antes de dar por terminada la aplicación del procedimiento.

A todo esto se añade el problema que se desprende de la siguiente ilustración. Imaginemos que se quieren conocer los «valores normales» para las funciones pulmonares (capacidad vital, capacidad vital forzada en un segundo, etc). ¿Qué debemos entender por *una población sana* de la cual tomar la muestra?

Algunos opinan que debe tomarse una muestra representativa de la población general, excluyendo sólo a aquellos de cuya condición patológica (en lo que concierne a dolencias que comprometen la función respiratoria, tales como silicosis o cáncer pulmonar) se tenga constancia. Otros consideran que deben eliminarse *a priori*, por ejemplo, a los fumadores, ya que ellos han modificado negativamente su función pulmonar a través del tabaquismo.

Pero en esa línea, otros podrían exigir que tampoco se admitieran en la muestra a los sedentarios, pues el buen funcionamiento pulmonar demanda la práctica de ejercicios. Y aun podría haber quienes exigieran que la muestra estuviese exclusivamente integrada por montañeses que no sólo no fuesen fumadores ni sedentarios sino que no hubiesen respirado nunca monóxido de carbono, de modo que el sistema respiratorio conserve el estado más próximo posible a su constitución fisiológica original.

Ahora, imaginemos que según estos límites (obtenidos a través de una muestra de impolutos con acuerdo al último y más restrictivo de los criterios), se valora a un conjunto de trabajadores de una fábrica en el contexto de un tamizaje organizado por autoridades de salud ocupacional. Con ese rasero, ¡quizás todos resulten ser anormales! La muestra elegida para fijar el criterio diagnóstico era tan normal que lo más común y corriente resulta ser anormal.

Resulta imposible escapar de este laberinto si no se establece *para qué* se es-

tá procurando encontrar esos límites. Si lo que se quiere, por ejemplo, es fijarlos con la finalidad de establecer una regulación de carácter jurídico, según la cual un trabajador que presente valores anómalos resulte beneficiado con un periodo de descanso o algún tipo de compensación, entonces la muestra apropiada sería la de la población general, fumadores incluidos. Si lo que se quiere es estudiar el funcionamiento del sistema respiratorio como tal -quizás para aquilatar la desviación atribuible a los contaminantes ambientales cuando se comparen aquellos valores de referencia con los valores de una población que sí está bajo sus efectos- entonces la muestra más adecuada sería la de los montañeses no fumadores.

Otro problema que cabe recordar en esta materia es el hecho de que un sujeto dado puede ser «normalmente» lábil a los efectos que nos ocupan. Es decir, sin que se hayan producido cambios de índole sistémico en su fisiología, los resultados correspondientes a un mismo individuo pueden variar apreciablemente de un momento a otro.

5.3.3. El hilo de Ariadna

Todo lo anterior parece conducir a la convicción de que el problema de la identificación de valores anormales no tiene solución, hecho contraproducente y descorazonador, habida cuenta de la necesidad de referentes, que sin duda tiene el clínico para su gestión diagnóstica.

Pero ocurre que ni se está negando la conveniencia de contar con ese marco de referencia, ni dando por cierta la imposibilidad de obtenerlo; de lo que se reniega es del esquematismo facilista para su construcción y su aplicación. El análisis esquemático tiene la enorme ventaja de ser simple y directo; pero tiene el defecto de que puede no servir para nada.

El primer cargo de esquematismo que cabe al tema que nos ocupa dimana de su absurdo e irrealista afán dicotomizador que lleva las cosas al plano polar: «normal-anormal».

En un magnífico artículo sobre este tema, Murphy (1973) toma el ejemplo de las dolencias mentales y escribe:

> ***Los defectos mentales se discuten a menudo, al menos por el hombre común, como si hubieran dos grupos, el de aquellos con un desarrollo mental normal y el de los que no lo tienen: dos clases distinguibles sin ambigüedad, del modo en que pueden distinguirse las plantas de los insectos.***

Y tras un análisis de las consecuencias y riesgos implícitos en tal convicción, señala que la línea divisoria entre la normalidad y la anormalidad es muy frecuentemente arbitraria (aunque no caprichosa, que es algo bien diferente) y que, cuando se establece, ello se debe solamente a un imperativo operacional, ya que sin tal

demarcación no se podrían adoptar decisiones, por ejemplo, en el ámbito jurídico. En relación con esto, se pregunta:

> ***¿En qué punto de su desarrollo la sociedad decidió que el tonto del pueblo ya no debía ocuparse de barrer las calles o llevar las vacas a pastar y fue puesto bajo el amparo de una institución? Es difícil evitar la conclusión de que la respuesta simplemente es: «Cuando la sociedad estuvo en condiciones de asumirlo».***

Este carácter relativo de las acciones, que se verifican más en función de las posibilidades reales que a partir de determinaciones intrínsecas de «normalidad», se aprecia en hechos como que el desarrollo de la sociedad contemporánea consiente, al menos en los países más avanzados, que los oftalmólogos no dejen de recetar gafas graduadas al 95% de la población adulta general, aunque su capacidad visual está «dentro del recorrido normal» y se les procure la adquisición de dicho recurso paliativo. Illich (1975) señalaba que «toda dolencia es una realidad que posee una configuración social; tanto su significado como la reacción a que da lugar tienen una historia».

Es paradójico el hecho de que solemos decir y reiterar que el hombre es un ser ***bio-psico-social,*** pero de inmediato aspiramos a poseer reglas dicotómicas y unidimensionales que nos digan para cada parámetro *fisiológico* si un valor específico es o no normal. El carácter «normal» de ciertas funciones no puede ser descontextualizado, ni de su relación con otros parámetros fisiológicos conexos, ni del entorno ecológico, social y psicológico en que se mueve el individuo.

En principio, no hay ningún derecho a considerar que los límites que definen al intervalo en que se ubican los valores más habituales o próximos a la media -por ejemplo, los correspondientes al 95% de los individuos- sirvan para marcar la «normalidad» en el sentido clínico del término. Tal advertencia ha sido hecha por varios autores. Por ejemplo, Riegelman y Hirsch (1992) señalan adecuadamente que «el intervalo de lo normal es descriptivo y no diagnóstico».

De hecho, parece haberse enraizado un enorme malentendido al trasladar el término «normal», proveniente del descubrimiento del físico-matemático Karl F. Gauss, relacionado con las mediciones sucesivas de un mismo objeto, y el alcance semántico que tiene en la clínica para separar a los sujetos que no están enfermos de los que sí lo están. Según Kruskal (1978), no está muy claro cómo surgió el adjetivo «normal» para identificar a la distribución descubierta por Gauss: aparentemente fue introducido en 1877 por Francis Galton, debido a la positiva connotación del término, que navega ambiguamente entre lo que es «deseable» y lo que es «común». Esa «positiva ambigüedad» -en absoluto ajena a la teoría de la eugenesia, creada por el propio Galton- podría estar causando, un siglo más tarde, iatrogenias nacidas del malentendido subsiguiente.

De hecho son relativamente pocos los estados que pueden considerarse ***incondicionalmente patológicos.*** Según Murphy, los que más se acercan a dicha condición

son aquellos que representan desórdenes capaces de acortar o imposibilitar la vida, de los cuales algunas expresiones cancerosas pudieran ser el ejemplo más claro.

En general, un clínico eficiente examinará un conjunto de parámetros fisiológicos integralmente; lo hará sin desdeñar la historia del individuo concreto, y su juicio dependerá del enclave histórico, económico y cultural en que se halle. Es muy conveniente que disponga de datos que le informen cuáles son los ***valores más comunes*** para cada parámetro, cuál el ***recorrido usual*** de éstos en la población de la que procede el paciente, pero sin etiquetar ***a priori*** el dato como «normal» o «anormal», algo que él decidirá después del complejo análisis cuyos ejes básicos se han bosquejado arriba. Tal matización cabe también para el examen epidemiológico al nivel poblacional.

En síntesis, lo que parecería más atinado y fructífero sería combatir el mito de que los estadísticos han conseguido delimitar cuándo un valor señala una condición patológica.

5.4. Inferencia estadística para interpretar la historia y viceversa

Un curioso ejemplo del uso festinado de la estadística inferencia1 lo hallamos en una esfera insospechada: la investigación histórica.

Aparentemente entusiasmados por la facilidad con que algunos investigadores de las ciencias sociales y biomédicas estaban (o afirmaban estar) demostrando sus teorías con el auxilio de esta herramienta, a finales de la década de los 50 surgió la llamada «nueva historia económica».

Al decir de Fontana (1982), este fenómeno se produce en medio de la inquisitorial vigilancia ideológica de los años de la ***guerra fría.*** Quizás se explique así la carga de voluntarismo de que estuvo signada. Esta línea de pensamiento se estructuró sobre la base de que, si bien la historia no podía estudiarse a la manera determinística de algunas zonas de la física o la química, sí podía intentarse una explicación de los hechos históricos mediante un enfoque estocástico, haciendo uso de ecuaciones, variables, componentes aleatorios (en los que se englobarían acontecimientos debidos a causas fortuitas), control de variables confusoras, etc.

En cuestión de pocos años, los trabajos de ***historia econométrica*** se multiplicaron en algunas revistas norteamericanas, especialmente en lo que Fontana llama el «órgano oficioso de la secta»: el ***Joumal of Ecomomic History.***

A modo de ilustración del carácter aberrante de esta tendencia, recreo a continuación la reseña que se hace del estudio hecho por Fogel y Engerman (1974) según el cual se demuestra con refinados recursos matemáticos la insólita conclusión de que la ***ley de la oferta y la demanda*** fue capaz de producir beneficios en materia de limpieza moral de la sociedad.

El objeto del análisis era el fenómeno de la prostitución en Nashville a finales del siglo XIX. A partir de datos censales, los autores detectan que no había pros-

titutas esclavas en esa ciudad de Tennessee. Sin reparar en que el censo sólo recogía información sobre las profesiones ejercidas por los habitantes libres (es decir, que nunca hubiera registrado esclavas dedicadas a esa labor aunque las hubiera en abundancia), proceden a descubrir las causas. En palabras de Fontana:

> ***Semejante error resulta disculpable al lado del disparate cometido al pretender explicar la supuesta ausencia de esclavas-prostitutas con un precioso gráfico de curvas de oferta y demanda que parte de la suposición de que las prostitutas, a consecuencia de lo mucho que disfrutan en su actividad laboral, aceptarán trabajar en ese oficio por unos ingresos que estén por debajo del salario mínimo de las obreras sin calificar Siendo así; se explicaría que sus dueños no las dejasen dedicarse a un oficio tan mal pagado, puesto que lo que les interesaba era que maximizasen sus ganancias monetarias, se divirtiesen o no.***

En general, el propósito de esta corriente era el de conseguir expresar las relaciones históricas en términos de ecuaciones y funciones numéricas cuya interpretación habría de esclarecer el porqué de los acontecimientos.

Se trata en fin sólo de una ilustración de la trivialización que puede experimentar la estadística, y de la manipulación de que puede ser objeto la sociedad por ese conducto.

Paralelamente, así como algunos historiadores sublimaron el papel de la estadística, los estadísticos y epidemiólogos han desdeñado con frecuencia el papel de la historia. El examen histórico de los acontecimientos que estudian puede ser, en efecto, cardinal para el proceso de identificación, no sólo de su diagnóstico -ya que toda realidad tiene una historicidad en la que el epidemiólogo debe reparar- sino incluso de sus determinantes. Resulta oportuno intercalar un ejemplo notable en que sí se saca partido a esta potencialidad.

A principios de 1992, en la provincia extremo occidental de Cuba, Pinar del Río, se empezaron a registrar casos de pacientes que presentaban trastornos visuales y desórdenes neurológicos periféricos.

Con el paso de los meses se diagnosticaron decenas de miles de casos en todo el país: se había consolidado una epidemia de neuropatía. Este hecho era en extremo llamativo, pues no existían antecedentes de una epidemia de tal naturaleza en un país con el desarrollo sanitario de Cuba, sin dudas la nación subdesarrollada con más logros en materia de salud.

Poco después, en una primera aproximación, se consideró oficialmente que se trataba de una ***Ambliopía Tabáquico-Alcohólica Nutricional.*** Esta abarcadora y complicada denominación esclarecía bastante poco sobre las causas. La epidemia comenzó a ceder a finales de 1993, pero el análisis causal aún esperaba por resultados categóricos. A lo largo de su desarrollo, se produjo una agitada polémica científica en torno a las causas de la enfermedad. Tres hipótesis se disputaban la explicación del fenómeno. En esencia, eran las siguientes:

a) La denominada como *tóxico-metabólica,* que colocaba la existencia de alguna toxina ingerida por las víctimas como causa primaria. Pero el componente tóxico no se dejaba ver. Una excepción era el tabaco, cuyo consumo sí aparecía asociado a la enfermedad; sin embargo, el ejercicio del tabaquismo no se había modificado recientemente, de modo que esa práctica difícilmente podría explicar por sí sola la irrupción abrupta y cuantiosa de la enfermedad. Por otra parte, de haber estado actuando un tóxico alimentario, la incidencia por grupos de edades tendría que haber sido esencialmente la misma, con excepción quizás de los niños. Las tasas eran, sin embargo, muy altas en las edades laboralmente activas y muy bajas entre los sujetos de mayor edad. La distribución espacial de la epidemia, finalmente, tampoco favorecía esta hipótesis, ya que se diseminó por todo el país y, por otra parte, con los focos de mayor intensidad ubicados en puntos muy distantes entre sí.

b) La hipótesis *viral* fue enfáticamente defendida por algunos investigadores; pero los indicios a su favor eran bastante débiles, y era inconsistente con un dato clínico: no se encontró rastro alguno de contagio. En particular, era notable la ausencia de casos entre los niños, y muy baja la tasa entre adolescentes y ancianos.

C) La llamada hipótesis *nutricional,* sustentada por el hecho de que se había producido una marcada disminución en la calidad general de la dieta, muy especialmente a partir de 1990 y entre los adultos de la isla. Ésta se había tornado monótona, cargada de energéticos y deficiente en proteínas y grasas. Uno de los argumentos en contra de esta tercera conjetura era el carácter «explosivo» de la epidemia, ajeno a un efecto carencia1 que, de estar actuando, debería expresarse más bien de manera gradual y relativamente aislada.

Esta última hipótesis, sin embargo, iba ganando adeptos: en Cuba se había producido un súbito trauma nutricional del que participó virtualmente toda la población dentro de un plazo reducido a muy pocos meses. Era una consecuencia del deterioro económico debido a la desaparición de sus mercados centroeuropeos a comienzos de la década del 90, coincidente con un recrudecimiento del bloqueo comercial, económico y financiero ejercido por los Estados Unidos sobre la isla. La doble protección, estatal y familiar -típica esta última de la familia latina- que en ese contexto se dispensó a la población ubicada en las edades extremas, explicaba los diferenciales de la epidemia según edades. Los estudios realizados para la identificación etiológica, basados en técnicas epidemiológicas y estadísticas formales, avalaban casi unánimemente la verosimilitud de esta tercera conjetura; algunos de ellos, como el de Gay *et al.* (1994), aportaron indicios muy persuasivos en esa dirección.

Millones de personas -no cientos, ni miles- sostenidamente enfrentadas a una disvitaminosis del complejo B y a un déficit de tioaminoácidos (con la conse-

cuente inhibición de los mecanismos de desintoxicación que habitualmente mitigan el efecto del cianuro aportado por el consumo de tabaco) habrían terminado por dar lugar -transcurrido un lapso de cerca de dos años- a manifestaciones clínicas características de las neuromielopticopatías en forma de epidemia, concentradas, como es de esperar, en los sujetos genética y constitucionalmente predispuestos; o sea, a cargo de las 50 mil personas que enfermaron.

Sin embargo, la duda no había muerto: las hipótesis alternativas aún defendían su espacio. En ese punto, los investigadores cubanos Francisco Rojas, Pedro Ordúííez y Alfredo Espinosa hacen un significativo descubrimiento: un artículo anónimo, publicado en ***Crónica Médico Quirúrgica de La Habana*** el año en que se inició el presente siglo (Anónimo, 1900), cuyo autor sería -según estos investigadores- el doctor Santos Fernández, eminente oftalmólogo cubano de la época.

En ese testimonio se reseña una epidemia de ***ambliopía por desnutrición*** en la Cuba de finales de siglo; se mencionan los trastornos originados por una alimentación escasa y defectuosa como responsables de una especie de neuritis periférica, prácticamente idéntica a la de 1991, y se destaca que, en principio, había sido confundida con los trastornos de análoga apariencia pero de etiología tóxica.

Siguiendo el hilo de la madeja, se descubre un trabajo que había sido publicado dos años antes por Madan (1898) donde se había escrito:

> ***La Isla de Cuba, país fértil y rico, experimentó los efectos de la devastación que origina siempre la guerra a la que completó la medida estratégica adoptada por el gobierno español de reconcentrar en los poblados a los campesinos, privando así de auxilios al enemigo, pero también de brazos a la agricultura que quedó paralizada... aumentando la miseria por la falta de auxilio del exterior en un país asolado.***

Todo esto dio lugar a lo que, según palabras de Madan, llegó a conocerse familiarmente en círculos sanitarios como «ambliopía del bloqueo, que se atribuye a la escasez y la mala calidad de los alimentos».

Las dudas relacionadas con el origen de la epidemia quedaron disipadas: su carácter carencia1 fue universalmente aceptado, y la responsabilidad que en buena medida correspondía al bloqueo financiero, comercial y económico impuesto por Estados Unidos a Cuba quedó bien establecida (véase por ejemplo el contundente artículo al respecto publicado por Román (1995) en ***Annals of Internal Medicine).***

Ordúñez, Nieto, Espinosa y Caballero (1995) publicarían más tarde una completa reseña histórica en que se da cuenta de este iluminador antecedente la cual concluye con una afirmación que me releva de más comentarios:

> ***Todos los que hemos estado envueltos en el estudio de esta epidemio tenemos una importante lección que aprender La comprensión de las causas de la epidemia que azotó a Cuba no dependerá de las respuestas que puedan ofrecer los mejores estudios epidemiológicos; las evidencias que dimanan de la historia pueden desempeñar un papel cardinal a los efectos de prevenir la reaparición de esta enfermedad en Cuba y en cualquier otro sitio.***

5.5. La interpretación de los riesgos

El concepto de riesgo en su sentido epidemiológico ha dado lugar a una vastísima producción teórica y práctica, pero también a no pocas confusiones, controversias y falacias. La presente sección se destina a comentar e ilustrar algunos de los puntos polémicos más frecuentemente olvidados.

5.5.1. Riesgo y conceptos conexos

Para comenzar, cabe señalar que la situación en materia de definiciones es algo confusa. A continuación se realiza un breve repaso del manejo conceptual dado por la literatura al tema.

En el diccionario de Last (1988), en la entrada correspondiente a *riesgo,* figura lo siguiente:

> ***Probabilidad de que un acontecimiento definido ocurra, por ejemplo la aparición de enfermedad, o de que el individuo enfermo muera dentro de un determinado periodo o edad. Como término no técnico comprende una variedad de medidas con respecto a la probabilidad de que ocurra un acontecimiento, generalmente adverso.***

El concepto de más interés, sin embargo, es el de *factor de riesgo.* Según Colimón (1978), éste es un miembro del «conjunto de fenómenos de los cuales depende el riesgo», definición laxa o ambigua, en buena medida debido al carácter equívoco del verbo «depender».

Por su parte, Kannel (1988) señala que los ***factores de riesgo «se*** basan, exclusivamente, en asociaciones demostradas en los estudios epidemiológicos; pueden ser directamente causales, manifestaciones secundarias de anormalidades o síntomas precoces de enfermedad». Tampoco resulta inequívoco qué entiende este autor por «asociación demostrada».

Piédrola ***et al*** (1990) se pronuncian de manera mucho más clara cuando lo definen como:

> ***Factor endógeno o exógeno, que puede ser controlado, precede al comienzo de la enfermedad, está asociado a un incremento de la probabilidad de incidencia de una enfermedad y tiene responsabilidad en su producción.***

Y lo distinguen, por una parte, de la noción de ***marcador de riesgo,*** que sería un: «concepto reservado a las variables de persona y, por tanto, endógenos, que no son controlables y definen a los individuos particularmente vulnerables. Señalan un aumento del riesgo de padecer la enfermedad, aunque no tienen influencia directa en su producción». Por otra parte, definen también el llamado ***indicador de riesgo***

como una «característica significativamente unida a la enfermedad en su estadio preclínico, sin influencias en su producción».

Finalmente, Last (1988) considera que *factor de riesgo* es un término que ha sido usado por lo menos con tres diferentes significados. Sin embargo, la explicación que da de ellos es, a mi juicio, confusa hasta el punto de que no resulta fácil distinguir una de otra.

Como se ve, la discrepancia central consiste en que, algunas definiciones exigen un papel causal, directo o «demostrado», en tanto que otras sólo demandan una asociación con una vaga insinuación sobre su posible implicación causal. En el Capítulo 7, donde me extiendo algo más sobre estos conceptos, el factor de riesgo se define como cualquier factor asociado a la enfermedad o daño que, sin ser causa propiamente (en el sentido de que su supresión no elimina necesariamente la aparición del problema), su presencia puede favorecer que el agente causal actúe. En cualquier caso, un factor de riesgo constituye un indicio, un punto de partida para indagaciones más profundas en procura de confirmación causal.

Lo que se quiere resaltar ahora, sin embargo, es que una cosa son los conceptos de ***riesgo*** y ***factor de riesgo,*** y otra el modo de medirlos, problema que exige contemplar el diseño del estudio.

Las herramientas básicas (y emblemáticas) para medir el riesgo son ***las tasas***, ya sea de prevalencia o de incidencia. Típicamente se construyen mediante la razón entre el número de individuos afectados por un daño dado (enfermos, muertos, accidentados, etc.) como numerador, y alguna medida del grado de exposición (número de sujetos susceptibles de ser afectados, un intervalo temporal, número de años acumulados por los expuestos, número total de kilómetros recorridos, etc.) como denominador. Por otra parte, el grado en que cierto factor entraña un riesgo se puede medir de diversas maneras (diferencia de tasas, riesgo relativo, odds ratio, tasas proporcionales de mortalidad, coeficientes estandarizados de regresión, entre otras).

Sin embargo, no es nuestro propósito detenernos ahora en este vastísimo campo; sólo procede hacer algunas puntualizaciones, especialmente sobre aquellos aspectos conflictivamente vinculados a la estadística.

Algunos conceptos estadísticos se han identificado erróneamente con métodos epidemiológicos específicos, hasta generar una perversión de la que resulta un laberinto conceptual dentro del que en ocasiones se extravía el propio concepto.

Especial confusión parecen generar los famosos «odds ratio», a partir de su conexión con el riesgo relativo y con el tipo de estudio epidemiológico utilizado [6].

Por ejemplo, en un texto publicado por la ***Organización Panamericana de la Salud*** para ejecutores de programas de salud, se aprecia claramente tal confusión; allí Almeida (1992) escribe:

[6] La traducción al castellano de la expresión «odds ratio» es bastante conflictiva. Tanto es así que ha dado lugar, incluso, a varios artículos destinados a ese único asunto: véase Rigau (1990), Martín (1990) y Tapia y Nieto (1993). Personalmente, prefiero asimilar sin más, y en consonancia con Porta (1990), la expresión inglesa.

Otra medida de asociación es el llamado odds ratio, o estimación del riesgo relativo, específico para el análisis de un diseño de investigación muy especial, el estudio de casos y controles.

El odds ratio es un concepto ***per se*** y no un modo de estimar el riesgo relativo; tampoco es específico de ningún tipo de estudios. Al igual que ocurre con la edad de una persona a quien se acaba de conocer, el odds ratio es un número que existe independientemente de la vía por la que se intente determinar su magnitud. Esta confusión está tan extendida que merece una consideración pausada.

Los *odds* asociados a cierto suceso se definen como la razón que resulta de dividir la probabilidad de que dicho suceso ocurra entre la probabilidad de que no ocurra; es decir, es un número que expresa cuántas veces más probable es que se produzca el hecho en cuestión frente a que no se produzca.

Así, si llamamos *E* a dicho suceso, P(E) a la probabilidad de que ocurra, y *O(E)* a los *odds* que le corresponden, entonces se tiene:

$$O(E) = \frac{P(E)}{1 - P(E)}$$

Por ejemplo, si se estima que el 75% de los pacientes que ingresan en un servicio hospitalario de cuidados intensivos sobreviven, entonces se dice que «los ***odds*** de que un paciente genérico sobreviva son 3», ya que 0,75/0,25=3.

Resulta interesante que éste sea el modo en que a menudo se resume esta realidad probabilística en algunas zonas de la cultura sajona, incluso en ambientes no académicos [7]. La información equivalente -la probabilidad de que un sujeto que ingresa sobreviva es del 75%- es la usada regularmente en el mundo latino.

Por otra parte, conocidos los ***odds*** de un suceso, se puede deducir su probabilidad. En efecto, si los ***odds*** de un suceso *E* ascienden a *O(E),* entonces la probabilidad de que éste se produzca es $P(E) = \frac{O(E)}{O(E) + 1}$. Por ejemplo, si se sabe que los ***odds*** de sobrevivencia que tiene un paciente de cáncer pulmonar un año después de ser operado ascienden a $O(E) = 0{,}4$, esto es lo mismo que saber que la probabilidad de que ese hecho ocurra es $P(E) = \frac{0{,}4}{1{,}4} = 0{,}285$.

'De modo que ambas informaciones son equivalentes y expresan la misma noción: cuantifican cuán verosímil es que algo ocurra y, en particular, cuál es el riesgo de un acontecimiento, si ese «algo» representa un daño para la salud.

Obviamente, entre la probabilidad del suceso y los ***odds*** correspondientes hay

[7] Por ejemplo, se suele hablar de «los odds que tiene un equipo de baloncesto» antes de un juego; con ello se alude a cuántas veces más probable es que gane frente a que pierda.

una clara relación directa: si aquella aumenta, éstos también lo hacen. Si P(E) = 0, entonces *O(E)* también es nulo; pero, en la medida que P(E) tiende a la unidad, *O(E)* tiende a infinito.

La Figura 5.4 refleja gráficamente la relación existente entre ambas magnitudes en el intervalo [O,l – 0,9].

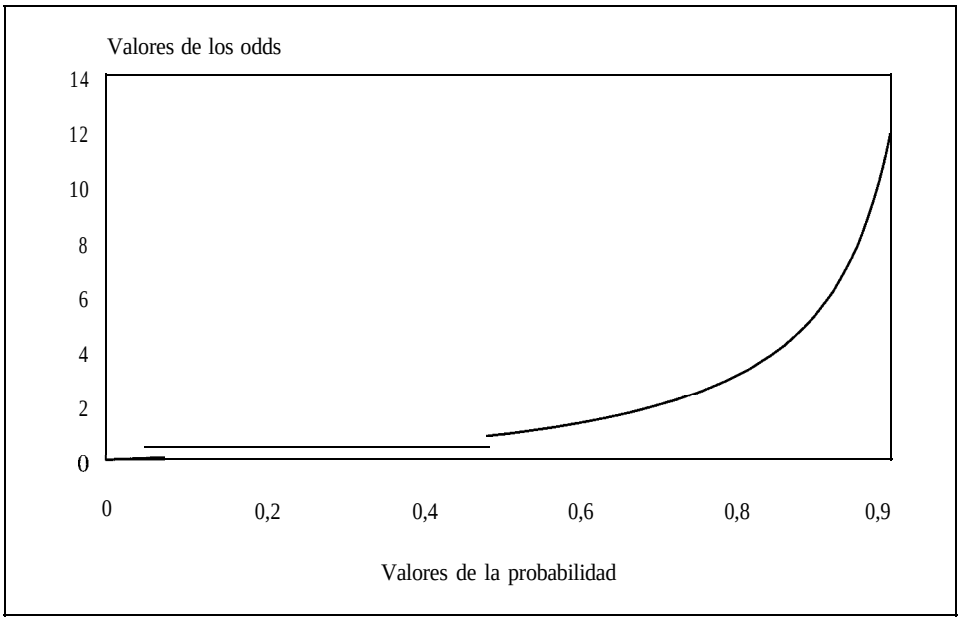


Figura 5.4. ***Transformación del espacio probabilístico al de odds.***

En este punto es menester recordar ese cardinal concepto de la epidemiología actual conocido como ***riesgo relativo.*** Supongamos que $P_A(E)$ denota el riesgo [8] de que se produzca la enfermedad [9] cuando está presente la condición ***A, y*** que ***B es*** otra condición, de manera que $P_B(E)$ denota el riesgo que se corre cuando es ella la que rige. Entonces, la razón:

$$RR = \frac{P_A(E)}{P_B(E)}$$

[8] Usualmente, esta noción, como ya se dijo, se mide mediante una tasa de prevalencia o de incidencia pero, en cualquier caso, se trata de algo que puede entenderse como una probabilidad. La condición A puede representar un factor aislado o un complejo de factores.

[9] Puede tratarse de una enfermedad, que es el caso más común; pero también de otro acontecimiento, tal como no haber sido vacunado contra la tuberculosis, tener un accidente o ser portador de un virus. Por mera comodidad, con frecuencia hablaremos de «enfermedad», aunque el concepto es obviamente más amplio.

expresa el riesgo *relativo* de padecer la enfermedad *E* cuando se está en la condición *A* respecto de padecerla cuando se está en la condición B. Dicho de otro modo, sintetiza cuántas veces más probable es desarrollar la enfermedad si se está en el primer caso que si se está en el segundo. El hecho de que RR sea mayor que la unidad hace pensar, en principio, que la condición *A es* más peligrosa que la *B* a los efectos de desarrollar *E*.

Un caso particular de este concepto que resulta de especial interés para los epidemiólogos es aquel en que, al analizar la etiología de cierta enfermedad *E*, la condición *A es* haber estado expuesto a cierto factor *F, y* la condición *B* es la complementaria, no haber estado expuesto a él. En tal caso, se dice que RR es el *riesgo relativo inherente al factor F, y* queda sobreentendido que la probabilidad del numerador se circunscribe a los sujetos expuestos a *F* en tanto que la del denominador corresponde a los que no lo están. En esta situación, el RR no es más que una medida de asociación entre *E y F*.

Nótese que en la definición de RR no hay alusión alguna ni al diseño del estudio realizado ni, mucho menos, al modo en que tal estimación habría de producirse. Sin embargo, como ya se ha dicho, una confusión frecuente consiste en no distinguir entre el parámetro -número cuya existencia se postula al margen de que vaya o no a ser estimado- y el modo de computar la estimación, que sí puede depender del tipo de estudio que vaya a llevarse adelante o de los recursos estadísticos disponibles.

Ahora bien, del mismo modo que los ***odds*** son una manera equivalente (aunque diferente) de expresar la probabilidad de un acontecimiento, la razón de dos odds es una manera alternativa a la de manejar la razón entre dos probabilidades.

Así, se define el llamado ***odds ratio*** *(OR)* la razón de los ***odds*** correspondientes a un suceso bajo cierta condición entre los que le corresponden bajo otra. Por esa vía se procura cuantificar la misma noción, en esencia, que con el RR.

Con el *OR y* el RR ocurre algo muy parecido a lo que pasa con dos buenos exámenes para medir la destreza de un sujeto para la geometría, construidos ambos sobre los mismos presupuestos conceptuales y aplicados al mismo individuo: a pesar de que cuantifican la misma noción, sus valores (aun en el supuesto de que las pruebas usaran la misma escala) sólo en casos excepcionales darán lugar al mismo número; esto es así -aparte de la variabilidad experimental- por la simple razón de que son instrumentos *diferentes* intentando aprehender cuantitativamente la misma dimensión conceptual.

De este modo, en lugar de trabajar con el RR de cierta dolencia *E* correspondiente a cierto factor *F* para cuantificar el grado de asociación entre éste y aquella, se puede manejar el ***odds ratio*** asociado al factor:

$$OR = \frac{\dfrac{P_F(E)}{1 - P_F(E)}}{\dfrac{P_{\bar{F}}(E)}{1 - P_{\bar{F}}(E)}}$$

donde $\bar{F}$ denota que nos estamos circunscribiendo a los sujetos que no están expuestos a *F*.

Este parámetro tiene especial atractivo en determinados entornos aplicativos. Tal es el caso, por ejemplo, cuando se ha hecho uso de la regresión logística [10], o cuando se realiza un estudio de casos y controles.

En esta última situación, lo que ocurre es que, por razones técnicas que no procede detallar ahora (véanse textos como Schlesselman, 1982 o Silva, 1987), no es posible estimar el , pero sí el . De modo que en ese tipo de estudio, ésta es la manera usual de medir la magnitud relativa del riesgo bajo una condición respecto de otra.

Por otra parte, puede demostrarse que si la prevalencia de *E* es muy baja (suele bastar con que sea inferior al l0%), entonces los valores correspondientes a ambos conceptos -riesgo relativo y ***odds ratio***- se asemejan extraordinariamente.

5.5.2. Los peligros de la letra C

El más burdo de los errores que se cometen en nombre de la «teoría de riesgos» consiste en considerar el hecho de que algún factor se presente con alta frecuencia entre los que sufren una dolencia como un indicio (y, en casos extremos, como una prueba) de que dicho factor entraña un riesgo de padecerla. Este error se viene cometiendo desde hace décadas y, aunque la frecuencia de su aparición es decreciente, no parece por ahora dispuesto a hacer mutis.

El ejemplo que elijo para ilustrarlo constituye un modesto homenaje personal a José Augusto Coll, profesor de la cátedra de bioestadística de la Universidad del Litoral (Argentina) quien, a mediados del presente siglo, ya había dado valiosas muestras de sus perspicacia, espíritu crítico y lucidez. El ejemplo en cuestión puede hallarse en Col1 (1950) y es en esencia el siguiente:

En un estudio sobre gigantismo fetal Coatz (1945) comunica que de un total de 51.000 partos ocurridos en una clínica, se detectaron 125 casos para los cuales el peso del recién nacido superó los 5.000 gramos. Los 125 fetos gigantes estudiados se distribuyen según paridad de la madre en la forma que recoge la Tabla 5.3.

Al analizar estos resultados, el autor escribe:

> En ***resumen: solo 9 casos, es decir el 7%, son primíparas; el resto, 116, es decir 93% eran multíparas; y de estas últimas, 24 eran secundíparas, lo que representa el 19% del total. Éste es el grupo más numeroso, lo que nos autorizaría a aceptar que la secundiparidad es el momento más favorable para la producción de fetos gigantes.***

[10] Esto es debido a que, en tal caso, el exponencial de un parámetro de la regresión logística es el ***OR*** correspondiente a la variable que se le asocia (véanse detalles en Silva, 1995).

Tabla 5.3. ***Distribución de los 125 casos de gigantismo fetal según paridad de la madre***

Paridad de la madre	N.º de casos	Porcentaje
1	9	7%
2	24	19%
3	17	14%
4	16	13%
5	9	7%
6y7	19	15%
8y9	9	7%
10 y más	22	18%
Total	125	100%

El comentario es obviamente muy aventurado, ya que no se ha tomado en cuenta la distribución según paridad de la ***población*** de ***51.000*** partos. Puesto que tal distribución no se comunica en el libro, para hacer el análisis correcto hay que procurarse dicha información en alguna fuente capaz de ofrecer un patrón que, verosímilmente, no sea muy diferente del que regía en esta población de embarazadas. Podremos, corroborar entonces que, además de aventurado, el análisis es erróneo.

Examinemos qué sucede si estos 51.000 partos se hubieran distribuido según paridad en la misma proporción que los 117.644 nacimientos acaecidos en el Estado de Nueva York (excluida la propia ciudad de Nueva York), de acuerdo con datos recogidos en el sexagésimo-cuarto informe anual del Departamento de Salud de ese estado para 1943, año del estudio de Coatz. La distribución porcentual, según esa fuente, fue como se recoge en la Tabla 5.4.

Aplicando esos porcentajes a los 51.000 partos se obtienen los denominadores necesarios para construir las tasas de gigantismo fetal que se resumen en la Tabla 5.5 (segunda columna).

La Figura 5.5 permite apreciar gráficamente la notable diferencia entre la secuencia de porcentajes y la de tasas de macrofetos según paridad.

Es evidente que el riesgo de que se produzca un feto gigante crece sostenidamente al aumentar la paridad; consecuentemente, afirmar que el mayor riesgo corresponde a las secundíparas es, simplemente, disparatado.

Aunque el presente comentario se escribe exactamente medio siglo después de

Tabla 5.4. ***Distribución porcentual según paridad de los ll 7.644 nacimientos registrados en el Estado de nueva York en 1943***

Paridad de la madre	Porcentaje del total de partos
1	37,9
2	30,4
3	14,8
4	7,2
5	3,7
6 y 7	3,5
8 y 9	1,4
10 y más	1,1
Total	100,0

Tabla 5.5. ***Tasa de gigantismo fetal según paridad si la distribución de nacimientos en el estudio hubiera sido iguala la de los nacimientos producidos en Nueva York en 1943***

Paridad	N.º de nacimientos	N.º de fetos gigantes	Tasas por 1.000
1	19.329	9	0,5
2	15.504	24	1,5
3	7.548	17	2,3
4	3.672	16	4,3
5	1.887	9	4,8
6 y 7	1.785	19	10,6
8 y 9	714	9	12,6
10 0 más	561	22	39,2
Total	51.000	125	2,5

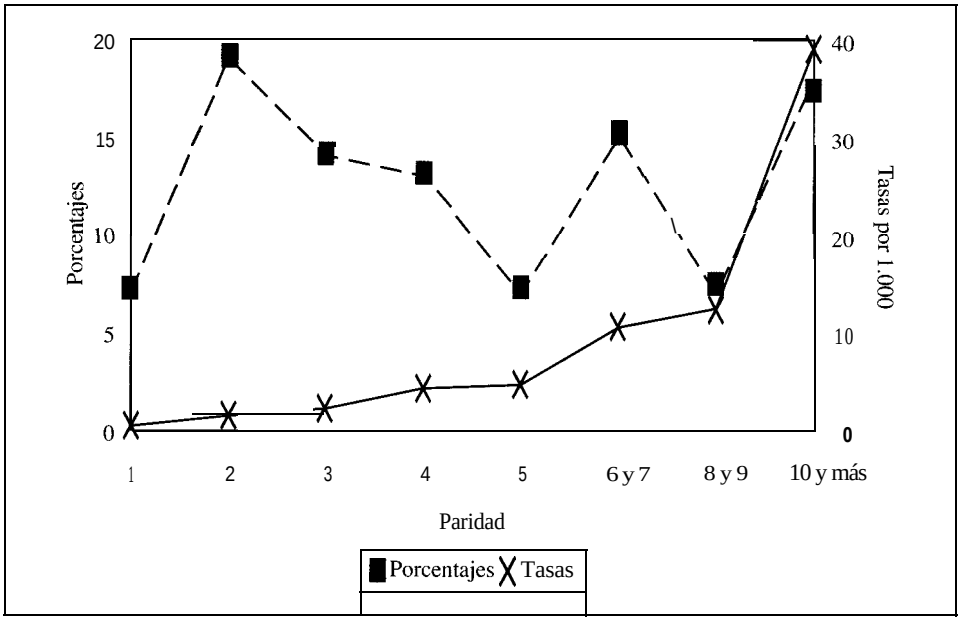


Figura 5.5. Porcentajes y tasas de macrofetos según paridad.

publicado el trabajo de Coatz, no debe pensarse, como ya se apuntó, que se trata de un error ya superado.

En un reciente programa de televisión, dedicado a la educación sanitaria de la población cubana en materia de accidentes, se alertaba a las madres de niños muy pequeños sobre *los peligros de la letra C.* La explicación que allí se dio a esta enigmática advertencia fue la siguiente: la mayoría de los accidentes en niños de tal edad se producen en enclaves cuyos nombres comienzan con dicha letra: cuna, calle, cochecito, corral, cocina, cama...

A mi juicio esta advertencia, aun cuando el «recurso de la letra C» tenga un propósito básicamente nemotécnico, epidemiológicamente, carece de mayor sentido, ya que un bebé no suele ir a sitios que empiecen con *a* (azoteas, avenidas, aviones) o con *b* (bares, barcos, burdeles): su vida discurre en lo esencial en los seis sitios inicialmente señalados; ¿dónde, sino allí, podrían producirse entonces los accidentes? Para identificar entornos de riesgo hay que calcular *tasas* que permitan estimar datos como los siguientes: cuántos accidentes se producen por cada cierto número prefijado de horas de permanencia en un entorno específico (cuna, cochecito, corral, ...).

Éste es un caso extremo, pero la falacia es un poco más insidiosa en el caso siguiente: se ha constatado que el 80% de los accidentes de tránsito se producen en un radio de 10 km en torno al domicilio del conductor: ¿significa esto que es más peligroso conducir cerca de la vivienda que lejos de ella o, simplemente, que el 80%

(o más) del tiempo de circulación de los automóviles se verifica en el entorno mencionado? Para hablar de que algún factor entraña riesgo es imprescindible contar con algún referente que situar en el denominador, de modo que se pueda construir, como ya se dijo, una tasa.

Desde luego, errores de este tipo serán difíciles de hallar actualmente en un artículo científico publicado en una revista de epidemiología; sin embargo, no es nada insólito que aparezca en ambientes diversos del pensamiento científico actual, la prensa incluida.

Por ejemplo, en el contexto de un discurso contra las prácticas preventivistas contemporáneas, Skrabanek (1994) suscribe la opinión de McCarthy (1992) según quien la exigencia de que los conductores de motos usen cascos es una medida inútil, salvo para conferir a las víctimas la responsabilidad por las heridas, ya que: «se producen más heridas en la cabeza en automovilistas y peatones que entre motoristas y, si hubiera un verdadero interés por defender a los ciudadanos, se extendería la prohibición a todos ellos».

Tengo la sospecha de que la disposición no se ha basado en que sus promotores tengan un obsesivo y oculto odio hacia los motociclistas y se empeñen en hostigarlos con una exigencia estéril. Pienso, más bien, que ella se asienta en que ***las tasas*** de mortalidad por trauma craneal son mucho mayores entre los que se accidentan mientras transitan en una moto que entre los peatones y automovilistas que se accidentan mientras se desempeñan como tales.

5.5.3. Una caricatura de los factores de riesgo y el riesgo de las caricaturas

El texto siguiente procede del mismo libro, altamente crítico contra las tendencias de promoción contemporáneas, escrito por Skrabanek (1994):

> ***Hablando técnicamente, los factores de riesgo no tienen nada que ver con las causas de las enfermedades; su introducción ha sido un ejemplo de las triquiñuelas estadísticas que proveen de una «explicación» para mecanismos causales que, de hecho, se desconocen. Por ejemplo, la homosexualidad es un factor de riesgo para contraer SIDA. Sin embargo, es claro que la homosexualidad no*** causa ***la enfermedad e, incluso, que si todos los homosexuales fueran exterminados, ello no la erradicaría. La posesión de una licencia de conducir es un factor de riesgo para los accidentes automovilísticos. La habilidad para nadar lo es para ahogarse. Ser japonés era un factor de riesgo para morir*** *por* ***concepto de un harakiri. En general, el estudio de los factores de riesgo y su detección en los individuos no nos acerca a la comprensión de los mecanismos causales. Los factores de riesgo ayudan la*** *mayoría* ***de las veces a oscurecer antes que a iluminar el trayecto hacia una adecuada comprensión de las causas. Kiihn (1993) señalaba que la prevención basada en la epidemiología de factores de riesgo está gobernada por el tipo de lógica según la cual la temperatura de una habitación es reducida mediante la ubicación del termómetro dentro de un cubo con hielo.***

Creo que todo es caricaturizable. La caricatura sin embargo no es a mi juicio una vía adecuada para ejercer la crítica científica. Tiene la ventaja de ser divertida pero, por su propia naturaleza, orientada a exagerar los rasgos distintivos del objeto caricaturizado, padece de tres defectos: promueve la desmesura, deforma las proporciones entre unos y otros componentes del todo, y suprime rasgos de interés, todo lo cual compromete el presupuesto básico del pensamiento científico: la objetividad.

Un examen serio de la teoría de factores de riesgo es una cosa; una estrategia de francotirador, es otra bien diferente. Lo primero que cabe señalar al texto de Skrabanek es que, obviamente, el homosexualismo no es una causa del SIDA. Lo que me pregunto es si puede citarse un solo trabajo serio en que tal afirmación haya sido hecha o siquiera insinuada.

Si alguien interpreta erróneamente a los factores que se asocian a un daño como explicaciones causales, no se debe por lo general a que usa «triquiñuelas estadísticas», sino a que está siendo mecánico. Pero de ahí a afirmar que la identificación de un factor de riesgo no nos sirve para nada, salvo para oscurecer el camino que conduce al conocimiento de las causas, va un trecho apreciable.

En lo que concierne a este aspecto, y para no extendernos teóricamente en un tema que se aborda en otro punto (véase Capítulo 7), baste reparar en el hecho de que el homosexualismo, como marcador de riesgo, constituyó un importantísimo indicio que contribuyó a identificar el carácter vira1 de la dolencia.

En 1981 los epidemiólogos repararon en la aparición en tres hospitales de Los Angeles de neumonía *porpneumocystis carinii,* una dolencia sólo común entre inmunodeprimidos, en cinco varones jóvenes y sanos. A partir de ese desconcertante hallazgo inicial y del examen de los nuevos casos que se iban presentando, pudieron identificarse cuatro subpoblaciones definidas por los que, en su momento, fueron considerados respectivos «grupos de riesgo»: homosexuales, hemofílicos, receptores de sangre y consumidores de drogas por vía parenteral.

Tales rasgos, que en ningún momento fueron considerados causas de la nueva enfermedad, sirvieron para evaluar hipótesis sobre las formas de transmisión de la infección y condujeron, como subraya Glass (1986), a una mutua y fructuosa colaboración entre clínicos, cientistas básicos y epidemiólogos que dio lugar, a la postre, a la identificación del retrovirus responsable directo de la dolencia.

Por otra parte, al decir que la posesión de una licencia de conducción es un factor de riesgo, se incurre claramente en una falacia: la misma que examinamos en la sección precedente. Poseer licencia es una condición para conducir legalmente; consecuentemente, el porcentaje de conductores (y, por tanto, también de accidentados) que la poseen es altísimo; pero ello no le confiere la condición de factor de riesgo, del mismo modo que no ser ciego, un rasgo que posee el 100% de los conductores (y una abrumadora mayoría de los accidentados) no adiciona riesgo al conductor. Lo cierto es que lo que diría un epidemiólogo no es lo que Skrabanek le atribuye sino lo contrario: que no tener licencia es un marcador de riesgo, ya que *la tasa*

de accidentes es mucho mayor entre los pocos que conducen sin poseerla que entre los que la poseen. Ignoro si, al hacer un viaje largo por carretera, a Skrabanek le hubiera sido indiferente que el conductor tuviera o no licencia. Pero tengo la sensación de que la mayoría de mis lectores, como yo, preferiría que sí la tuviera. Quizás tal preferencia no se deba a que hemos pasado por alto el hecho de que la mayoría de los accidentados tenga licencia, sino a que sabemos o intuimos que ***es más probable que se produzca el accidente si el conductor no tiene licencia que si la tiene.***

Análogamente, casi con seguridad la mayoría de los que se han ahogado sabían nadar; de hecho, la mayoría de la población y, en especial, la inmensa mayoría de los que se echan al agua, saben hacerlo. Pero si yo tuviera que ser rescatado por alguien, preferiría que se tratara de un nadador, precisamente porque saber nadar ***no es*** un factor de riesgo, como afirma Skrabanek, sino solamente ***un factor frecuente*** entre los que se ahogan, aunque ciertamente, más frecuente aun entre los que ***no se*** ahogan.

5.5.4. Para evitar el SIDA

En mi opinión, la educación para la salud ha de conducirse como cualquier otra forma de educación moderna; es decir, debe compartir información científicamente corroborada, pero sin anexarle juicios de valor; su cometido debe ser el generar y difundir información orientada a desarrollar en el ciudadano su personal capacidad de elección sin amedrentarlo.

Nada mejor que el polémico ejemplo del SIDA sirve para ilustrar cómo se transgrede esta norma. Quizás poca gente tenga una idea clara de la infectividad -es decir la probabilidad de contagio- de esta enfermedad. A pesar de las serias dificultades metodológicas que supone el diseño y la puesta en práctica de estudios capaces de dar respuesta a esta interrogante (Brookmeyer y Gail, 1994), hace ya bastante tiempo que se dispone de información aceptable al respecto. Con ese fin se han desarrollado varios «estudios de pareja» ***(partner studies),*** basados en la recolección de datos sobre el estado de sujetos susceptibles de infestarse por ser compañeros sexuales de individuos portadores del VIH.

Uno de los más importantes estudios de este tipo ***(The California Partner Study)*** examinó a mujeres sexualmente relacionadas con hombres infestados. Con datos precedentes de dicho estudio, Jewell y Shiboski (1990) hallan que la probabilidad de contagio en un solo contacto heterosexual ***con una persona infestada es*** de 1 en 1.000 (véase también Wiley, Hershkorn y Padian, 1989). Si llamamos γ a dicha probabilidad, según esa fuente, una fórmula elemental para calcular el riesgo de infestarse tras ***k*** contactos sexuales con una persona portadora del virus es la siguiente:

$$p(k) = 1 - (1 - \gamma)^k \qquad [5.2],$$

de manera que si tal acción se realizara diariamente durante un año, la probabilidad de convertirse en portador en ese caso es sólo del 30%. Si se utiliza condón, esta última circunstancia (relaciones heterosexuales diarias con un portador durante un año) hace que el riesgo baje a niveles mínimos (Paulos, 1980). Si finalmente, el contacto se realiza una vez y con una persona de la que se sabe que no pertenece a los grupos de alto riesgo, la probabilidad se reduce extraordinariamente, hasta ser inferior a la de un accidente fatal durante el viaje hacia el motel donde tendrá lugar la cita.

Me temo que existe una fuerte tendencia a evitar que la gente domine datos como éstos [11]. Si los argumentos estadísticos se usan sólo cuando favorecen una política establecida de antemano y se silencian cuando la ponen en entredicho, estamos ante una manipulación tan espuria como la que padecemos cuando los propios datos se tergiversan.

No es cuestión de promover una distensión en la lucha contra el terrible mal, ni de dejar de advertir que el uso de condones ayuda a prevenirlo, que es muy peligroso compartir jeringuillas u omitir el bien documentado y altísimo riesgo de la relación homosexual sin protección por vía anal. Pero es éticamente inaceptable que no se informe objetivamente todo lo que se conoce sobre la dolencia. Nadie tiene derecho a escamotear información ni mucho menos a suplir las células grises de los demás.

En este sentido resultan preocupantes grandes titulares de prensa como el que, con el texto «La alarma es necesaria para combatir el SIDA», atribuye el periódico ***El País*** al director del ***Centro Internacional del SIDA*** de la Universidad de Harvard (Mann, 1992). Quizás sin quererlo, se podría estar licitando la emisión de mensajes como el del cardenal Narcis Jubany para quien «desde el punto de vista preventivo el único remedio contra el SIDA es la fidelidad conyugal» (Jubany, 1992). Afirmaciones como ésta se ubican en el extremo más pernicioso de los mensajes supuestamente preventivos. Hablar de «remedios únicos» no parece coherente con las complejidades de un problema como éste, cuya etiología, por lo demás, no se circunscribe a la esfera sexual. Por otra parte, cualquier mensaje que procure ser efectivo tiene que empezar por ser factible. Como demuestra Engels (1965) en su notable obra sobre el origen de la familia y la propiedad privada, la «infidelidad» tiene profundas raíces culturales y económicas y, aunque fuese deseable eliminarla, no serán los consejos intolerantes y admonitorios quienes lo consigan.

5.6 Gauss y la curva racista

El físico William Shocley, ganador del premio Nobel, conmovió a la comunidad científica internacional en 1966 al proponer la esterilización de los ciudadanos con

[11] Sáez (1994) consigna que el 50% de los ciudadanos de la Comunidad Económica Europea cree que el SIDA puede adquirirse como consecuencia de compartir un vaso con un portador del virus.

un bajo coeficiente de inteligencia y la creación de un banco de semen para la construcción de superdotados. Se reavivó así la polémica sobre las normas éticas que han de regir la manipulación genética. Sin embargo, lo sorprendente para muchos fue el resurgimiento, un vicenio después de la derrota del fascismo en Europa, de un pensamiento «culto», orientado en la línea eugenésica de la ideología nazi.

En 1839, un médico de Philadelphia, Samuel Morton, había inventado la *craneometría;* mediante la medición del cráneo, esta «disciplina» permitiría determinar la inteligencia de las diferentes razas. Su estudio de blancos, indios y negros colocó en ese orden la inteligencia de los humanos a partir de los mencionados criterios métricos. Morton alcanzó una enorme reputación y el agradecimiento expreso de las clases dominantes por haber dado, finalmente, aval científico a lo que hasta ese momento era sólo una convicción basada en el prejuicio. Pero Morton había caído en la trampa del dogmatismo: aunque su tesis central se presentó como el resultado de la inducción que va de los datos objetivos a las conclusiones, la realidad fue que, partiendo de conclusiones predefinidas, manejó los datos de suerte que pudiera arribar a ellas.

Un siglo y medio más tarde Stephen Gould, antropólogo de la Universidad de Harvard, reconsideró los datos originales de Morton y corroboró la falsedad de sus conclusiones (Gould, 1978). Sin embargo, ya mucho antes, Bertrand Russell, refiriéndose a esta línea de pensamiento había aconsejado (Russell, 1949) a «cualquiera que desee pasar una hora divertida» que atendiera «a las tergiversaciones de eminentes craneólogos, en sus intentos de probar por medidas cerebrales que las mujeres son más estúpidas que los hombres».

Como señala Galeano (1994):

> ***Se puso de moda otra manera de medir la inteligencia: la capacidad intelectual dependía del peso del cerebro, método que tenía el inconveniente de que sólo permitía admirar o despreciara los muertos. Los científicos andaban a la caza de cráneos famosos, y no se desalentaban a pesar de los resultados desconcertantes de sus operaciones. El cerebro de Anatole France, por ejemplo, pesó la mitad que el de Iván Turguenev, aunque sus méritos literarios se consideraban parejos. Gabriel René Moreno, la gran figura intelectual de Bolivia del siglo pasado, había descubierto que el cerebro indígena pesaba menos que el cerebro blanco. Como ocurre con la policía en los allanamientos, el racismo encuentra lo que pone.***

Tal fue el ***modus operandi*** de otro destacado científico: el francés Paul Broca quien, 20 años después de Morton, afirmaba que el cerebro masculino era mayor que el femenino, que pesaba más para hombres eminentes que para los mediocres, y que este indicador era útil para determinar las razas superiores. Murió sin saber que el suyo tenía un peso bastante inferior a la media.

El último eslabón de esta oprobiosa cadena ha contado con el concurso de la estadística como ***vedette.***

The Bel1 Curve (La curva acampanada) es una obra de más de 800 páginas debida a dos investigadores norteamericanos: el sociólogo Charles Murray y el psicólogo Richard Herrnstein (Herrnstein y Murray, 1994). Allí se pretende demostrar que los pobres no sólo se caracterizan por su escasez de recursos sino, también y sobre todo, por carecer de inteligencia.

El voluminoso estudio tiene como piedra angular las pruebas de inteligencia, a partir de cuya aplicación se obtienen datos concluyentes en la dirección mencionada, que les llevan a proclamar la esterilidad de procurar ayuda a los niños con mayor retraso, en especial los de raza negra.

La población negra norteamericana, a diferencia de las etnias asiáticas, se mantiene por debajo de los umbrales de pobreza, hecho que el libro atribuye a que la media en su coeficiente intelectual (el famoso *IQ)* se ubica 15 puntos por debajo de lo normal. Operando con la distribución de Gauss para el IQ, se identifica que la curva correspondiente a los negros se ubica a la izquierda de la de los blancos de modo tal que uno de cada cuatro negros norteamericanos tendría coeficiente mental 25 puntos por debajo de la media de los blancos, cifra propia del retraso mental, y muy inferior a la de la sociedad completa. McCord y Freeman (1990) revelaron hace unos años que la esperanza de vida de un hombre negro de Harlem (en el opulento Nueva York) era menor que la de uno de Bangla Desh, el país más pobre del mundo. ¿Será realmente debido a la inferior inteligencia de esta raza?

En un país en que una elevada cantidad de madres de niños negros no están integradas en programas de atención materno infantil, cuyos hijos nacen bajos de peso y crecen en medio de agudas carencias alimentarias y espirituales, donde la contaminación del aire urbano, mucho más alta en los barrios negros, produce una tasa de mortalidad por asma tres veces mayor entre jóvenes negros que entre jóvenes blancos, donde los negros constituyen el 12% de la población pero aportan el 34% de los niños que mueren antes de cumplir el primer año, y donde el riesgo perinatal es para negros mayor que para blancos para todo el espectro de pesos al nacer (Wilcox y Russell, 1986), se nos regala una tesis antiquísima envuelta en un nuevo envase: el de la estadística. En un excelente artículo publicado en *American Journal of Epidemiology*, Mutaner, Nieto y O'Campo (1996) citan dos decenas de trabajos que documentan mecanismos de discriminación étnica con relevancia potencial a los efectos de la salud y que abarcan esferas de la sociedad norteamericana tales como segregación residencial, acceso a bienes y servicios, y oportunidades en el mercado laboral.

La polémica en torno a las pruebas de inteligencia es tan vieja como su propia creación; no es una frase hecha sino rigurosa verdad histórica. Aparte de las suspicacias que siempre puede despertar la construcción de una variable sintética (véase Capítulo 4), en este caso particular cabe recordar la singular circunstancia de su aparición. El propio Alfred Binet, creador a comienzos del siglo de la primera prueba formal de este tipo de que se tenga noticia, no demoró en poner alarmados reparos al uso indiscriminado que podía dársele a su instrumento, y dejó claro que sus resul-

tados no indicaban ningún rasgo intrínseco ni permanente de los sujetos sino, meramente, el grado de necesidad de ayuda docente que tenían los niños a quienes se aplicaba. Pero ello no fue óbice para que, pocos años después, el psicólogo Lewis Terman, de la Universidad de Stanford, asimilara el modelo con el fin expreso de identificar perversiones genéticas que permitieran eliminar a sus portadores de la sociedad.

Como una premonición, Broad y Wade (1983), 12 años antes de la publicación de ***The Bell Curve,*** escribieron:

> ***Una y otra vez, los datos reflejan fuerte correlación entre los resultados de las pruebas de inteligencia y las variables ambientales. Una y otra vez, aquellos que crean o aplican las pruebas inventan tortuosas explicaciones ad*** *hoc* ***para respaldar sus prejuicios sobre el carácter hereditario de la inteligencia.***

En efecto, múltiples hallazgos científicos permiten descartar categóricamente que el total de la inteligencia sea heredada, y confirman que, en buena medida, es un producto del ambiente socio-económico en el que se desarrolla el individuo. Entre los más persuasivos e interesantes estudios de este tipo es el debido a Orley Ashenfelter y Alan Krueger, economistas de Princeton, quienes examinaron las diferencias de salario entre gemelos idénticos: aquel de los hermanos que tenía más educación ganaba como promedio un 16% más por cada año adicional de formación académica.

Como contrapartida a la polémica concepción de inteligencia medida a través del ZQ el psicólogo Peter Salovey de la Universidad de Yale propuso el concepto de «inteligencia emocional» (Gibbs, 1995) relacionado con la capacidad de controlar las emociones de manera que se optimice la forma de llevar adelante la vida. En ese contexto se ha propuesto el llamado ***cociente emocional,*** que resulta ingeniosamente ilustrado por la siguiente experiencia, descrita por Gibbs.

Se trabaja con una muestra de niños de cuatro años; cada niño es introducido en una habitación y el investigador le indica que puede disponer de una golosina en el momento; pero que si espera hasta que él regrese luego de un trámite que ha de realizar, entonces recibirá dos golosinas en lugar de una. La gama de reacciones es amplia: algunos sucumben al instante, otros consiguen esperar unos minutos y luego se abalanzan sobre la golosina; otros más resisten estoicamente: cantan, intentan distraerse y hasta dormir, pero aguardan en procura del premio. Un estudio reveló que diez años más tarde, aquellos que supieron contenerse eran adolescentes socialmente exitosos y seguros de sí mismo; los que no pudieron resistir la tentación eran con más frecuencia chicos inestables y frustrados. La media de rendimiento académico de los primeros superaba abismalmente a la de estos últimos. El *IQ,* sin embargo, era básicamente el mismo. Goleman (1996), redactor científico del ***New York Times,*** ha publicado un libro divulgativo con gran éxito de ventas sobre la inteligencia emocional.

El voluminoso tratado gestado al amparo de la curva en forma de campana no pasa de ser un ejemplo más de manipulación que tiene cabida gracias al anume-

rismo de la sociedad. En efecto, aunque la inteligencia fuese algo inherente al individuo, y aun cuando en su conformación existiera un fuerte componente genético-hereditario, los resultados presentados entrañan una falacia, como pone en evidencia Gould (1994).

Supongamos que se miden las tallas de hombre adultos en una comunidad sometida a una fuerte privación nutricional donde la talla media asciende a 168 centímetros. La heredabilidad dentro de la comunidad es alta: padres altos -quizás de 174 centímetros- tienden a producir descendientes altos, y padres bajos -164 centímetros en promedio- tienden a tener hijos pequeños. Pero la alta heredabilidad no descarta que una mejor nutrición no fuese capaz de elevar la talla media por ejemplo a 182 centímetros en cuestión de 2 o 3 generaciones. Exactamente del mismo modo, la aparentemente bien documentada diferencia de 1.5 puntos en el *IQ* de negros y blancos, con un alto peso hereditario dentro de cada grupo, no permite concluir que un sistema de oportunidades iguales vaya a ser incapaz de elevar el IQ de los negros hasta alcanzar el nivel de los blancos, e incluso superarlo.

Según Gould, el libro adolece de varios problemas adicionales en el uso e interpretación de la regresión múltiple y de otros recursos estadísticos usados en el libro. Esta expresión de danvinismo social resulta ser, en fin, algo muy alejado de un verdadero tratado científico para ser simplemente, un manifiesto de ideología conservadora en la línea de la tenebrosa teoría eugenésica creada por Francis Galton en 1898.

Aquel año en el estado de Michigan se aprobó la ley que disponía la esterilización eugenésica mediante el recurso de emasculación a débiles mentales y epilépticos; en los años 30, más de un tercio de los estados de la Unión ya habían adoptado regulaciones similares. Aunque exista tan sombría tradición, causa estupor el resurgimiento recurrente de la mitología racista en Estados Unidos, aunque ahora aparezca disfrazada de ciencia. Afortunadamente vivimos tiempos en que no es tan fácil embaucar a la sociedad con el manejo falaz de los conceptos y los números; cientos de profesores universitarios y científicos norteamericanos se han pronunciado contra este despropósito maquillado con afeites estadísticos, y muchos de ellos lo han enfrentado con todo el rigor metodológico necesario para desmontar el engendro (véanse una veintena de ensayos sobre el tema en Fraser, 1995).

Bibliografía

Almeida N (1992). ***Epidemiología sin números.*** Serie Paltex N.º 28,OPS/OMS, Washington.

Anónimo (1900). ***Ambliopía por neuritis periférica debida a auto-intoxicación de origen intestinal por alimentación defectuosa.*** Crónicas Quirúrgicas de La Habana 27: 330-334.

Broad W, Wade N (1983). ***Betrayers of the truth. Fraud and deceit in the halls of science.*** Simon and Schuster, Inc., New York.

Brookmeyer R, Gail MH (1994). ***AZDS Epidemiology. A Quantitative Approach.*** Oxford University Press, New York.

Coatz AS (1945). ***Gigantismo fetal.*** El ateneo, Buenos Aires.

Colimón KM (1978). ***Fundamentos de epidemiología.*** Colimón, Medellín.

Col1 JA (1950). ***Defectos en el análisis en la interpretación de trabajos médicos.*** Anales de Medicina Pública ll: 12-23.

Elveback LR, Guillier CL, Keating FR (1970). ***Health, normality, and the ghost of Gauss.*** Journal of the American Medical Association 211: 69-75.

Engels F (1965). ***El origen de la familia, la propiedad privada y el estado.*** Editorial R, La Habana.

Fogel RW, Engerman SL (1974). ***Time on the cross,*** Little Brown, Boston.

Fontana J (1982). ***Historia: análisis del pasado y*** *proyecto* ***social.*** Grijalbo, Barcelona.

Fraser S (1995). ***The bell curve wars: Race, intelligence and the future*** *of* ***America.*** Basic Books, New York.

Freedman MA (1991). ***Health status indicators for the year 2000.*** Healthy People 2000, Statistical Notes Vo1 1, N.º1.

Galeano E (1994). ***La peste.*** Semanario Brecha, 10 (471): 32.

Gay J, Porrata C, Hernández M, Clúa AM, Argüellez JM, Cabrera A, Silva LC (1994). ***Factores dietéticos de la neuritis epidémica en la Isla de la Juventud, Cuba.*** Boletín de la Oficina Panamericana de la Salud 117: 389-399.

Gibbs N (1995). ***Las emociones y no el coeficiente intelectual pueden ser la base de la inteligencia humana.*** El País, Sección Sociedad, Madrid, pág. 27,8 de octubre de 1995.

Glass RI (1986). ***Newprospects for epidemiologic investigations.*** Science 234: 951-955.

Goleman D (1995). ***Inteligencia emocional.*** Kairos, Barcelona.

Gould SJ (1978). ***Morton's ranking of races by cranial capacity.*** Science 200: 503-509.

Gould SJ (1994). ***Curveball.*** The New York, 28 de noviembre: 139-150.

Greenberg ER, Stevens M (1986). ***Recent trends in breast surgery in the United States and the United Kingdom.*** British Medical Journa1292: 1487-1491.

Herrnstein RJ, Murray C (1994). ***The bel1 curve: intelligence and class structure in American life.*** The Free Press, Nueva York.

Illich 1 (1975). ***Medical nemesis.*** Calder and Boyars, Londres.

Jewell NP, Shiboski SC (1990). ***Statistical analysis of HIV infectivity based on partner studies*** Biometrics 46: 1113-1150.

Jubany N (1992). ***La plaga del siglo XX,*** citado por Miguel Sen, Periódico «El Mundo», Madrid, 3 de junio de 1992.

Kannel WB (1988). ***Una perspectiva sobre los factores de riesgo de las enfermedades cardiovasculares.*** En: Buck, C. ***et al.*** El desafío de la epidemiología. Problemas y lecturas seleccionadas. 1988 OPS/Washington. Publicación Científica n.º 505. p. 758-780.

Kruskal W (1978). ***Formulas, numbers, words: statistics inprose.*** En: Fiske D (Editor) New Directions for Methodology of Social and Behavioral Science, 1981 Jossey-Bass, San Francisco.

Kiihn H (1993). ***Healthismus. Eine Analyse der Praventionspolitik und Gesundheits-forderung in den USA.*** Sigma, Berlin.

Last JM (1988). A ***Dictionary of Epidemiology,*** 2nd Edition, Oxford University Press, Oxford.

Madan D (1898). ***Notas sobre una forma sensitiva de neuritis periférica, ambliopía por neuritis óptica retrobulbal:*** Crónica Médico Quirúrgica de La Habana 24: 81-86.

Mann J (1992). Periódico «El País», Sección de Sanidad, Madrid, 22 de junio.

Martín JM (1990). ***Oportunidad relativa: reflexiones en tomo a la traducción del término «odds ratio».*** Gaceta Sanitaria 16: 37.

McCarthy M (1992). ***Do cycle helmetsprevent serious injuly?*** British Medical Journal 305: 881-882.

McCord C, Freeman HP (1990). ***Excess mortality in Harlem.*** New England Journal of Medicine 322: 173-177.

Murphy EA (1973). ***The normal.*** American Journal of Epidemiology 98: 403-411.

Mutaner C, Nieto FJ, O'Campo P (1996). ***The bell curve: On race, social class, and epidemiologic research.*** American Journal of Epidemiology 144: 531-536.

Ordúñez P, Nieto FJ, Espinosa A, Caballero B (1996). ***Cuban epidemic neuropathy 1991 to 1994: History repeat itself a century after the «Amblyopia of the Blockade».*** American Journal of Public Health 80: 738-743.

Paulos JA (1990). ***El hombre anumérico.*** Alfaguara, Madrid.

Piédrola G et al (1990). ***Medicina preventiva y salud pública.*** Salvat, Barcelona.

Porta M (1990). ***Traducir or no traducir: ¿es esa la cuestión?*** Gaceta Sanitaria 16: 38-39.

Riegelman RK, Hirsch RP (1992). ***Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica.*** Publicación Científica n.º 531, Organización Panamericana de la Salud, Washington.

Rietz HL (1927). ***Mathematical statistics.*** Carus Mathematical Monograph N.º 3, Mathematical Association of America, Open Court Publishing, Chicago.

Rigau JG (1990). ***Traducción del término «odds ratio».*** Gaceta Sanitaria 16: 35.

Román GC (1995). ***Onpolitics and health: An epidemic of neurologic disease in Cuba.*** Annals of Interna1 Medicine 122: 530-533.

Russell B (1949). ***The scientifc outlook.*** George Allen and Unwin, Londres.

Sáez F (1994). ***El hombre y la técnica.*** América Ibérica, Madrid.

Schlesselman JJ (1982). ***Case-control studies.*** Oxford University Press, New York.

Silva LC (1987). ***Métodos estadísticos aplicados a la investigación epidemiológica,*** Cuaderno 14, Instituto Vasco de Estadística, Bilbao.

Silva LC (1995). ***Excursión a la regresión logística en ciencias de la salud.*** Díaz de Santos, Madrid.

Skrabanek P (1994). ***The death of humane medicine and the rise of coercive healthism*** The Social Affairs Unit, Publication n.º 59, London.

Tapia JA, Nieto FJ (1993). ***Razón de posibilidades: una propuesta de traducción de la expresión odds ratio,*** Salud Pública de México 35: 419-424.

Wilcox AJ, Russell IT (1986). ***Birthweight and perinatal mortality: III. Toward a new method of analysis.*** International Journal of Epidemiology 15: 188-196.

Wiley JA, Hershkorn SJ, Padian NS (1989). ***Heterogeneity in the probability of HIV transmission per sexual contact: The case of male-to-female transmission in penile-vaginal intercourse.*** Statistics in Medicine 8: 93-102.

6

¿Qué significan las pruebas de significación?

La verdad de las cosas reside en sus matices.

PAUL VALÉRY

El uso formal de *pruebas estadísticas de significación,* también conocidas como *pruebas o dócimas de hipótesis,* data de la década de los 30 del presente siglo. Desde entonces, en pocas áreas de la estadística más que en ésta se han entronizado las recetas y el adocenamiento.

Tanto las decenas y decenas de libros sobre estadística inferencia1 que -con pocas diferencias entre sí- no han dejado de producirse a lo largo de estos 60 años, como muchos cursos de estadística desembocan rápidamente en una secuencia de rutinas para la evaluación de diversas hipótesis y promueven, de hecho, su aplicación mecánica como árbitros que dicen la última palabra. Una vez que se pronuncian, ya sea responsabilizando por los hechos observados a una hipótesis, ya sea incriminando al azar como posible explicación, se nos enseña que solo resta aderezar las conclusiones con unas cuantas frases convencionales. Las cosas, sin embargo, no son tan simples. Para cada problema hay un contexto que atender, ciertas cautelas que adoptar y hay, en fin, todo un mundo conflictivo y en evolución al que prestar atención y en torno al cual posicionarse.

6.1. Una polémica escamoteada

Prácticamente ningún texto de estadística, salvo que se especialice en el tema, informa a los lectores sobre el vivo debate que durante las últimas décadas ha venido desarrollándose en torno al uso de esta herramienta. Se conforma así otro notable ejemplo de conducta contradictoria con la ética científica; en este caso, con el precepto que demanda comunicar transparentemente todos los elementos relevantes sobre el tema que se aborda. ¿Por qué la literatura general tiende a omitir casi

toda referencia a ese debate? ¿Por qué los programas docentes de estadística silencian la mayoría de las contradicciones que le dan vida?

Aparte de la responsabilidad que corresponde a la inercia, ya que muchos cursos y libros se diseñan de cierta forma por la simple razón de que otros... se han diseñado así, la respuesta básica es que muchos especialistas no tienen mayor interés en que la gente piense con autonomía. Algunos «metodólogos» prefieren conducirse como brujos de tribu: saben cómo hacer las cosas; saben incluso que algunas no son como dicen que son, pero no comunican sus interioridades a quienes solo toca aplicar las rutinas que ellos enseñan y sobre las cuales, a menudo, erigen sus pedestales.

Curiosamente, según mi experiencia, muchos «usuarios» de la estadística se sienten perfectamente cómodos con las recetas, y no sólo las prefieren sin matizaciones, sino que reaccionan con irritación cuando éstas aparecen en escena. Suelen ser refractarios a una formulación del tipo: ***Esto puede considerarse así, aunque hay otra corriente de opinión que señala que...*** Prefieren la fórmula que simplemente reza: ***Esto es así.***

El connotado economista Galbraith (1991), reflexionando sobre esta forma de la conducta, escribe:

> ***Forma parte de la vanidad humana el que el esfuerzo mental proporcione una satisfacción intrínseca. Esto es cierto para algunos, sin duda; para la mayoría, el esfuerzo mental es algo que resulta excepcionalmente agradable eludir: De esto se deriva el carácter de toda gran organización. Los que las sirven tienen fuerte compromiso de lealtad a la creencia oficial y*** *por* ***tanto a la actuación oficial.***

Con mucha frecuencia he apreciado que algunos investigadores se conducen exactamente así, procurando enterarse de las «posiciones oficiales» de la estadística para someterse a ellas con disciplinada lealtad y ahorrarse tanto esfuerzo mental como sea posible.

Actúan, en fin, como quien corta los hilos telefónicos para no enterarse de las malas noticias.

6.2. La lógica interna de las pruebas

Aunque parto del supuesto de que el lector tiene adecuada información sobre la lógica fundamental según la cual operan las pruebas de significación, tal vez no sea ocioso comenzar recordándola a grandes rasgos.

Expresado del modo más general, se declara que hay ***significación estadística*** en relación con una hipótesis, a la que se llama ***hipótesis nula y*** que se denota como H_o, cuando, usando el supuesto de que sea cierta, se obtiene una probabilidad sorprendentemente pequeña de algún acontecimiento. En tal caso, se da por cierta la llamada ***hipótesis alternativa H,.*** En general, se trata de pronunciarse sobre la veracidad de H_o ***y*** en principio se podrían cometer dos errores al hacerlo: declararla fal-

sa siendo cierta (*error de tipo I), y* declararla cierta a pesar de que sea falsa (*error de tipo II)* [1].

Detengámonos en un ejemplo muy simple que permita «repasar» la filosofía y la mecánica en que se fundamentan las pruebas de significación. Supongamos que H_o afirma que una novedosa técnica quirúrgica para el tratamiento de la otoesclerosis *no es* más eficaz que otra ya existente. Usualmente el investigador aspira a demostrar que H_o *no es* cierta; en nuestro ejemplo, obviamente, lo que se desea es *rechazar* la hipótesis planteada en favor de la hipótesis alternativa H_1, que afirma que la técnica novedosa es más eficaz que la convencional.

Imaginemos que se cuenta con 10 pacientes que padecen de este mal en ambos oídos, todos con el mismo nivel de deterioro audiométrico. Se realiza una experiencia consistente en aplicar la técnica novedosa en uno de los dos oídos y la técnica convencional en el otro, decidiendo al azar para cada paciente cuál se aplica en el derecho y cuál en el izquierdo. La decisión sobre la validez de H_o *se* adoptará en función de los resultados y de la estructura probabilística de la experiencia.

Hechas las 20 intervenciones, se practican las mediciones correspondientes y se cuenta el número *m* de pacientes para los que el nuevo tratamiento consigue más recuperación auditiva que el otro. Se calcula entonces la probabilidad de tal resultado «bajo H_o». Por ejemplo, ¿cuál es la probabilidad de que la técnica novedosa produzca mejores resultados para *todos* los pacientes, supuesto que es válida la hipótesis de que ambas técnicas sean igualmente eficientes? Se trata de la misma probabilidad de que la experiencia de lanzar 10 veces una moneda no trucada dé lugar a 10 caras y a ningún escudo *(m = 10).* De la teoría elemental de probabilidades sabemos que este número es igual ap = $(0,5)^{10}$ = 0,00098.

Ha aparecido la famosa ***p:*** la probabilidad de haber obtenido los resultados que se obtuvieron, supuesto que H_o *es* cierta. La regla que se aplica ahora depende de la magnitud de ese número: si es muy pequeño, entonces se concluye que H_o ha de ser falsa. ¿Cuán pequeño debe ser para considerar que los resultados son suficientemente incompatibles con H_o como para descartar su validez? La cota más «popular» es 0,05; otros más exigentes optan por un valor menor, 0,0l 0 incluso 0,001; los menos conservadores consideran que basta con que dicha probabilidad no supere a 0,l. Conozco un solo autor (Winer, 1971) que reivindica la posibilidad de trabajar, cuando se dan ciertas condiciones, con niveles mayores (tan grandes como 0,2 o 0,3).

De modo que si la experiencia concreta con los 10 enfermos desembocara en que para todos ellos la nueva técnica es más efectiva, puesto que es menor que 0,001, con cualquier rasero se rechazaría la hipótesis nula en favor de la hipótesis alternativa H_1. Esa es la formalización del razonamiento que nos llevaría a concluir

[1] Coincido con Paulos (1980) en que, al bautizar los dos errores posibles, los estadísticos no hicieron, precisamente, un alarde de imaginación.

en el ejemplo de la moneda que la obtención de 10 caras consecutivas es incompatible con el supuesto de que aquella no está trucada.

Si la superioridad de la técnica novedosa se registrara en sólo 9 de los 10 casos *(m* = 9), la hipótesis de nulidad estaría soportando un fuerte embate experimental aunque, sin duda, algo menos categórico que en el caso anterior. ¿Cuál es en este caso el grado de incompatibilidad entre el valor de ***m y*** la veracidad de H_o? La probabilidad de obtener 9 o más resultados favorables a la nueva técnica, supuesto que H_o es cierta, es igual ap = $\frac{11}{1.024}$ = 0,01067. Según la tradición ya comentada, la mayoría de los analistas consideraría tal probabilidad suficientemente pequeña como para poder afirmar que hay evidencia muestra1 de que H_o ***es*** falsa.

La probabilidad de que la experiencia dé lugar a por lo menos ***m = 8*** «éxitos» para la nueva técnica asciende ap = 0,055: estando por encima de 0,05, algunos intérpretes encadenados a este último número dirían que ya no puede descartarse el azar como explicación; y si ***m*** = 7, entonces *p* sería igual a 0,172, un número que ya nadie consideraría suficientemente pequeño como para declarar que los resultados se apartan significativamente de lo que H_o permite esperar. En tal caso, el experimentador se abstendrá de sacar conclusión alguna sobre H_o.

El lector habrá observado que al computar *p* se ha supuesto que el valor ***m*** obtenido representa una gama más amplia de posibilidades: la de que el número de éxitos sea «mayor o igual» que ***m.*** Es decir, lo que se calcula es la probabilidad de obtener ***m o más*** éxitos. Se trata de una singularidad muy discutible de esta teoría, pues es razonable que una hipótesis sea rechazada por ser incompatible con ciertos hechos observados, pero de esta práctica se colige que una hipótesis que pudiera ser verdadera pueda resultar que se vea rechazada porque no cubre o no explica resultados ¡que no han ocurrido! (Jeffreys, 1961).

El ejemplo ilustra la lógica de las pruebas de significación; en la práctica casi todas las situaciones exigen, desde luego, cómputos probabilísticos más complejos, pero el razonamiento básico es el mismo. ¿Cómo se consolidó, hasta universalizarse, esta manera de valorar hipótesis científicas?

Si un investigador consiguiera demostrar que determinado procedimiento terapéutico ha sido capaz, más allá de toda duda, de revertir la situación de un par de pacientes con cáncer pulmonar en fase terminal, entonces seguramente no necesitará de aval estadístico alguno para que ese hallazgo sea considerado con la máxima seriedad[2]. Nadie dudaría -en principio- de que se trata de un aporte «significativo», donde el término está siendo usado en su alcance semántico habitual (no estadístico), como algo fuera de lo regular, de importancia sustantiva, no trivial en el sentido clínico. Tal fue el caso, por ejemplo, de la cura de la endocarditis bacteriana con penicilina.

[2] Esto no equivale a incurrir en la ingenuidad, típica de los cultores de la pseudociencia, de considerar automáticamente *demostrada* la eficacia del procedimiento.

Sin embargo, pocos hallazgos de la investigación médica consiguen ser tan elocuentes y persuasivos como éstos. Los que abundan son los trabajos confusos, cuyos resultados no dan lugar a una interpretación inequívoca. La afluencia de artículos de dudosa relevancia, o que ofrecían conclusiones endebles, condujo a que los editores de revistas médicas empezaran a reclamar constataciones estadísticas que trascendieran lo que podría no ser más que simple anécdota clínica debida a la casualidad.

Así, junto con demandas como las de usar grupos de control o asignaciones aleatorias en los ensayos clínicos, se empieza a reclamar que las conclusiones novedosas vinieran acompañadas del apoyo que otorgan las ***pruebas de significación estadística;*** de ese modo se instalan en la literatura los famosos valores ***p*** como recursos para poder afirmar que no es por azar por lo que los datos dan indicios de un nuevo conocimiento, sino porque lo que se pensaba al respecto era falso.

6.3. Abusos y cautelas; errores y enmiendas

6.3.1. Diferentes significaciones

La más universal de las interpretaciones abusivas en que se incurre al usar las pruebas de significación consiste en no distinguir entre ***significación estadística y significación clínica*** [3]. A pesar de que no es frecuente hallar textos que aborden el tema sin hacer la advertencia de que estos conceptos no son equivalentes, en la práctica se sigue actuando como si la primera garantizara la segunda, cuando en realidad cualquier diferencia o asociación, sea o no clínicamente relevante, puede ser estadísticamente significativa.

Refiriéndose a este pernicioso mimetismo, Feinstein (1985) afirmaba:

> ***Si la demanda crítica hubiera sido que la investigación produjese ambos tipos de significación (la que concierne al área estocástica y la que se vincula con los atributos cualitativos) entonces la alienación intelectual de hoy no hubiera ocurrido. Desafortunadamente, sin embargo, la palabra «significación» fue reservada sólo en su connotación estocástica, y la palabra «estadística» le fue adjuntada para crear la «significación estadística» como paradigma de calidad e importancia en la investigación médica... Usando *, **, y *** como símbolos para representar que p<0,05, p<0,01 y p<0,001 respectivamente, el investigador puede presentar tablas celestiales en las cuales los datos han sido reemplazados por estrellas.***

Llamaba así la atención sobre el efecto negativo de una confusión vigente en el proceso investigativo y en su reflejo editorial, a pesar de que en su momento había

[3] Se acostumbra a usar este adjetivo a pesar de que también se podría hablar de significación ***biológica, química o social,*** ya que el concepto a que se alude no necesariamente procede del ambiente clínico.

sido advertido por el propio Fisher (1959), y más tarde condenado por otros líderes prominentes del mundo estadístico, tales como Yates (1968), o los enciclopédicos Kendall y Stuart (1983). Con cierta resignación, Feinstein concluía:

> ***Puesto que la historia de la investigación médica también muestra una larga tradición de mantenerse por mucho tiempo fiel a doctrinas establecidas después de que esas doctrinas han sido desacreditadas, o de haberse demostrado su escaso valor,* no *podemos esperar un súbito cambio en esta política por el mero hecho de que ha sido denunciada por connotados conocedores de la estadística.***

Debe señalarse, sin embargo, que si bien la homologación mecánica entre ambos tipos de significación mantiene *de facto* su vigencia, lo cierto es que en la actualidad la suplantación de datos por asteriscos va siendo crecientemente desterrada de la literatura; algo similar está ocurriendo incluso con el uso de los emblemáticos 0,l, 0,05, 0,0l y 0,001.

6.3.2. Rechazar o no rechazar: ¿es esa la cuestión?

Una de las más notables limitaciones que lastran a las pruebas de significación en su planteamiento original, tal y como lo formuló Fisher, consiste en que no permiten realmente evaluar de manera cabal la validez de una hipótesis. Al aplicar una prueba para cierta hipótesis nula H_o, sólo hay dos posibilidades legítimas: o bien rechazar la validez de H_o, ***o*** bien no sacar conclusión alguna; es decir, su lógica interna no permite aceptar la hipótesis nula a partir de los resultados muestrales. En palabras del propio Fisher (1935): «la hipótesis nula jamás puede ser demostrada o establecida; es posible, sin embargo, refutarla mediante un experimento».

En efecto, la aritmética de la prueba reposa en el supuesto de que H_o ***es*** cierta y lo que se computa es la probabilidad asociada a la observación obtenida bajo dicho supuesto. Si ésta fuera muy pequeña, puesto que no hay duda alguna de que la observación fue esa, la responsabilidad por tal incompatibilidad sólo puede recaer en la incorrección del supuesto empleado para calcularla; es decir, se infiere que H_o debe ser falsa. Pero si la probabilidad es alta, ello no constituye una corroboración de H_o, ya que esa misma probabilidad, pero calculada bajo el supuesto de que sea cierta alguna otra hipótesis, pudiera ser tanto o más alta que *p*, sólo que ninguna de esas otras posibles hipótesis es valorada.

El modo como se han construido las pruebas lleva a decir: «Sólo cuando hay resultados en contra de la hipótesis nula daremos el paso de descartarla como válida; en caso contrario, todo queda como estaba». Esta restricción parecería no ser excesivamente descorazonadora; puede considerarse incluso conveniente. Téngase en cuenta que la hipótesis H_o es la expresión conservadora, en tanto que H_1 sería «lo nuevo», lo que desafía el conocimiento actual; de modo que dicha restricción está en línea con la saludable cautela que se recomienda en el proceso de sacar conclu-

siones. Esto trae consigo, sin embargo, un grave problema: por conducto de las pruebas de hipótesis nunca se podrá, en propiedad, considerar demostrada la falsedad de una hipótesis alternativa. Por ejemplo, nunca podría refutarse estadísticamente la afirmación, pongamos por caso, de que el uso sostenido de calcetines amarillos se relaciona con la aparición de tuberculosis. Si tal hipótesis fuera puesta a prueba y no se obtuvieran resultados contradictorios con H_0 (la hipótesis que declara que no existe asociación alguna entre una y otra condición), habría que decir que no se pueden sacar conclusiones al respecto. Algunos van más lejos y sugieren que ***siempre*** se deje claro que, aunque no se puede concluir nada, es posible que en el futuro sí se pueda rechazar la hipótesis nula. Sheehan (1980), por ejemplo, alerta en los términos siguientes sobre el error que se comete cuando se identifica el hecho de ***no rechazar*** la hipótesis H_o con el de ***aceptarla.***

> ***Planteada la hipótesis nula de que «no existe asociación entre fumar cigarrillos y padecer cáncer de pulmón», el investigador intenta diseñar el estudio de manera que pueda rechazarla. Si tras aplicar la prueba estadística adecuada no consigue rechazar la hipótesis nula, intuitivamente podría deducir que ésta es entonces verdadera, y concluir que no existe relación alguna entre fumar cigarrillos y padecer cáncer de pulmón. Pero la intuición es a veces engañosa, y en este caso puede conducir a un error muy frecuente al que tal vez no presten atención los lectores menos experimentados. Una cosa es obtener una conclusión negativa (no hay efecto o no hay asociación) y otra es no llegara conclusión alguna. La conclusión de que «no hay un efecto» dice que el tratamiento no funciona, o que fumar cigarrillos no produce daño. Esto no equivale a decir que en este estudio no se pudo alcanzar conclusión alguna.***

Hasta aquí, la explicación de Sheehan deja ver «la cara buena» de este rasgo de las pruebas de hipótesis; pero más adelante la complementa de una manera que permitirá apreciar claramente la otra cara de la moneda:

> De manera que, si los datos no exhiben una relación estadísticamente significativa entre fumar cigarrillos y contraer cáncer pulmonar, es incorrecto concluir que tal asociación no existe. ***Lo correcto es decir que los datos de esta investigación no mostraron una relación, pero que un estudio de mayor sensibilidadpodn'a mostrarla.*** (El subrayado es mío, LCS.)

Lo que se está diciendo es: «Atención, nunca concluya que H_o es cierta pues pudiera estar cometiendo el error de segundo tipo». En caso de que sí se halle significación, jamás se advierte que dicho hallazgo puede deberse a que se cometió el error de primer tipo porque dicha contingencia se considera una limitación intrínseca de la prueba: no sólo se sabe que tal error puede cometerse sino que el propio investigador ha fijado el riesgo que acepta de cometerlo (el «nivel de significación»). El error de tipo II, en cambio, no se controla de antemano; para no cometerlo nunca, simplemente no se acepta nunca H_o.

Ahora volvamos al ejemplo de la tuberculosis; si somos consecuentes con la sugerencia de Sheehan, habría que decir: «por ahora no hay indicios de que el color de los calcetines tenga responsabilidad etiológica en el desarrollo de la tuberculosis, pero otro estudio basado en una muestra mayor pudiera producirlos». Este ejemplo que resulta burdo, precisamente porque se sabe de antemano que no existe tal asociación, permite apreciar el limitado valor demarcatorio de la prueba; pero también pone en evidencia el carácter tendencioso del ejemplo elegido por Sheehan, pues concierne a una situación en que no hace ningún daño que no se saquen conclusiones a partir de la prueba porque... también la conclusión se conocía de antemano.

La literatura está claramente signada por esta práctica; la diferencia con los ejemplos arriba mencionados radica en que en la mayoría de las aplicaciones regulares sí se estaría, supuestamente, usando este procedimiento estadístico para pronunciarse sobre un asunto en disputa. Abundan las situaciones en que, al valorar por ejemplo el peso etiológico de cierto factor en el desarrollo de una enfermedad, los investigadores obtienen un riesgo relativo quizás bastante elevado pero no significativamente diferente de 1. ¿Qué se concluye regularmente en esos casos? Lo usual es que tal resultado no apee al investigador de su convicción inicial, ya que cuenta con el recurso de anunciar que la susodicha significación no se halló en virtud del tamaño muestra1 insuficiente. Lo malo radica en que esto es enteramente cierto (en la Sección 6.4.1 se discute este punto en detalle); precisamente por serlo, el valor instrumental de la prueba de hipótesis se tambalea.

6.3.3. Contra su propia lógica

Esta limitación es tan contraproducente que no es infrecuente que las pruebas se usen olímpicamente aun en aquellos casos en que, justamente, lo que se espera o desea es no rechazar la hipótesis nula. Ocasionalmente, en efecto, lo que se procura es probar su veracidad; es decir, demostrar que no hay asociación o diferencias. En tal caso, si se quiere ser teóricamente coherente, las pruebas de hipótesis no deberían emplearse en absoluto. Pero ya es una práctica establecida que las pruebas se tomen como recurso cómodo que suple la reflexión crítica, de manera que con frecuencia no se repara en los «detalles» y también en estas situaciones se aplican sin más trámite. Consideremos algunos ejemplos.

Ilustración 1: Aval estadístico para una no-asociación

Imaginemos que el señor Warren Sánchez afirma tener la capacidad de mitigar los dolores lumbares mediante una oración milagrosa que susurra al oído del paciente. Para probar que el único efecto de semejante procedimiento -si es que tiene alguno- es de tipo placebo, un investigador lleva adelante un experimento en el que se pone a prueba este recurso contrastándolo contra el resultado que produ-

ce el empleo de una «falsa oración». Lo que se espera es que los porcentajes de recuperación correspondientes a uno y otro procedimientos no difieran; lo que se procura, en fin, es dejar sentado por esa vía que la oración no tiene nada de milagrosa. Si no se hallaran diferencias entre los efectos de una y otra oración, se declara que Warren Sánchez es un farsante y no que se deben seguir haciendo pruebas ***ad infinitum*** hasta que se pueda extraer la única conclusión posible. A despecho de lo señalado arriba, ante situaciones como ésta [4], profusamente ilustradas en el Capítulo 13, la prueba de hipótesis es el recurso inferencia1 rutinariamente empleado.

Ilustración 2: Pruebas de bondad de ajuste

Se está en un caso similar, aunque muchísimo más frecuente, cuando se aplican pruebas de bondad de ajuste. La hipótesis nula establece que la distribución de la muestra es igual a cierta distribución teórica; normalmente tales pruebas se usan con la esperanza de poder ***confirmar*** la validez de H_o *y,* cuando no se halla significación, nadie duda de que se ha corroborado estadísticamente la hipótesis nula. Nótese que el efecto de trabajar con un nivel de significación muy pequeño es en este caso exactamente opuesto al deseado: lejos de preservarnos contra la tendencia a sacar conclusiones alegremente, ayuda a que así se haga. Para apreciar este fenómeno más claramente, consideremos un ejemplo típico.

Imaginemos que se ha hecho un estudio descriptivo de las opiniones de los adultos de una ciudad sobre el sistema sanitario basado en una muestra de ***n*** = 400 individuos. Supongamos que se conoce teóricamente que las opiniones pueden diferir en dependencia de la edad, de manera que los investigadores quisieran confirmar la representatividad de la muestra a esos efectos; la vía natural es corroborar que la distribución muestra1 por edades es similar a la distribución poblacional. Al hacer los cómputos correspondientes, se obtienen los datos que se recogen en la Tabla 6.1.

La prueba de hipótesis clásica se basa en el estadígrafo siguiente:

$$\chi^2_{obs} = \sum_{i=1}^{3} \frac{(o_i - e_i)^2}{e_i} \qquad [6.1]$$

que se distribuye χ^2 con tantos grados de libertad como categorías haya menos 1, (en este caso, 2) donde O_i es la frecuencia observada para la i-ésima clase, p_i es la probabilidad teórica de pertenecer a ella y e_i es la frecuencia esperada correspondiente, que se computa mediante la fórmula: $e_i = n\ p_i$.

[4] Joyce y Welldon (1965) dan cuenta de un estudio «a doble ciego» en que las oraciones se realizan en ausencia de los pacientes; ni los que podrían ser «beneficiados» por ellas ni los del grupo control, ni tampoco los médicos que han de evaluar los resultados, saben por cuáles pacientes se oró y por cuáles no.

Tabla 6.1. ***Frecuencias observadas y esperadas para una muestra de 400 sujetos distribuidos según tres grupos de edad***

	16-39 años	40-64 años	65 años y más
Distribución porcentual en la población $100\,p_i$	45%	40%	15%
Frecuencia muestra1 esperada e_i	180	160	60
Frecuencia muestra1 observada o_i	175	166	59

Aplicando [6.1] en este caso, se obtiene:

$$\chi^2_{obs} = \frac{(175 - 180)^2}{180} + \frac{(166 - 160)^2}{160} + \frac{(59 - 60)^2}{60} = 0{,}38$$

Como este estadígrafo se distribuye ji cuadrado con 2 grados de libertad, la probabilidad asociada con 0,38 es igual ap = 0,827. Siendo ***p*** un número enorme, la hipótesis H_o no se puede rechazar; pero lo que usualmente se hace en estas situaciones no es abstenerse de sacar conclusión alguna sino concluir que H_o ***es*** válida; es decir, que la distribución muestra1 es esencialmente igual a la poblacional.

Si en lugar de la distribución muestra1 de la Tabla 6.1 (dada por las frecuencias 175, 166 y 59 respectivamente), se hubiera obtenido una mucho más alejada de la teórica, el valor de χ^2_{obs} sería mayor y, consecuentemente, menor el dep. Por ejemplo, para la distribución 170-182-48, los resultados serían $\chi^2_{obs} = 5{,}98$ y ***p*** =0,0503. Según las reglas ortodoxas, seguiríamos afirmando que H_o es correcta (ya que p>0,05). O sea, para no admitir que la muestra representa adecuadamente a esta población habría que tener una discrepancia muy considerable.

Por otra parte, si la distribución muestra1 fuese mucho más discrepante con la teórica (por ejemplo: 165-190-45) pero se operara con un nivel de significación más exigente (digamos, $\alpha = 0{,}01$), entonces tampoco se rechazaría H_o (en ese caso p = ***0,038).*** Este hecho resalta el siguiente contrasentido: cuanto más intransigente que sea ante la posibilidad de cometer el único de los errores que se controlan de antemano (el de tipo 1), con más facilidad se arribará a la conclusión deseada.

No conozco que tal propuesta se haya formulado nunca [5]. Conozco en cambio (y

[5] En cualquier caso, tampoco la menciono en calidad de alternativa real, ya que debe recordarse que ella no resuelve la objeción central consistente en que se da por demostrada la hipótesis nula.

además comparto, como señalé en Silva, 1995), el punto de vista de Khan y Sempos (1989) según el cual puede ser mejor no hacer ninguna prueba de hipótesis para evaluar la bondad del ajuste sino realizar una inspección informal de las frecuencias esperadas y observadas, y evaluar el grado de concordancia poniendo a funcionar el sentido común.

Esta alternativa, que a muchos pudiera parecer blasfema, está en línea con un problema sobre el que volveremos más adelante: si el tamaño de muestra es suficientemente grande, cualquier distribución muestra1 (salvo que sea exactamente igual a la poblacional) llevará al rechazo de H_o. Por ejemplo, si en lugar de $n = 400$ el tamaño muestra1 fuera 30 veces mayor ($n = 12.000$) y si la distribución de frecuencias relativas en la muestra fuera la misma que se reflejó en la Tabla 6.1, entonces los resultados serían los que muestra la Tabla 6.2.

Tabla 6.2. ***Frecuencias observadas y esperadas para una muestra de 1.200 sujetos distribuidos según tres grupos de edad***

	16-39 años	40-64 años	65 años y más
Distribución porcentual en la población $100\,p_i$	45%	40%	15%
Frecuencia muestra1 esperada e_i	5.400	4.800	1.800
Frecuencia muestra1 observada o_i	5.250	4.980	1.770

Al aplicar el estadígrafo en este caso, se obtiene un número 30 veces mayor:

$$\chi^2_{obs} = \frac{(5.250 - 5.400)^2}{5.400} + \frac{(4.980 - 4.800)^2}{4.800} + \frac{(1.770 - 1.800)^2}{1.800} = 11{,}42$$

La probabilidad asociada ahora a χ^2_{obs} es igual a $\boldsymbol{p} = 0{,}003$, de manera que la hipótesis H_o debe ser rechazada. El sentido común dice que en esto hay algo profundamente absurdo, pues si con determinado tamaño muestra1 el panorama observado nos llevó a sacar cierta conclusión, exactamente ese mismo patrón, pero observado a partir de una cantidad de información mucho mayor, debería llevarnos a ratificar la conclusión original y no, como realmente ocurre, conducirnos a la conclusión opuesta.

Ilustración 3: Algoritmos con pruebas de hipótesis sucesivas

El desarrollo de algunos algoritmos estadísticos exige respuestas parciales para decidir el curso de las acciones subsiguientes; en muchos casos, para obtener dichas

respuestas se usan pruebas de significación. Un ejemplo típico es el de la ***regresión paso a paso;*** en cada uno de esos pasos se hacen pruebas, y la acción que se desarrolla en el paso siguiente depende de que se haya rechazado o ***aceptado*** una hipótesis. Algo análogo ocurre cuando se utiliza la ***prueba de ji cuadrado de Bartlett*** para valorar si dos o más varianzas son iguales, requisito teórico para tener «derecho» a aplicar el ***análisis de la varianza; es*** decir, suele emplearse la prueba de Bartlett para, en dependencia del resultado, proceder o no a aplicar aquella prueba: si se ***acepta*** la hipótesis de homocedasticidad, se aplica el análisis de la varianza; de lo contrario, se realiza una prueba no paramétrica (usualmente la ***prueba Q de Cochran).*** Algunos paquetes estadísticos -por ejemplo EPIINFO- aplican esa regla de manera rutinaria, con lo que tácitamente da por sentado que es tan legítimo rechazar como aceptar H_o.

6.3.4. Una herramienta para cazar

La lógica de las pruebas comentada en la sección precedente, según la cual sólo se puede «ganar» (ya que si se pierde se hace como si nunca se hubiera jugado), ha contribuido a cincelar una costumbre de «mal perdedor». Asociado íntimamente a ella aparece la llamada «cacería de diferencias significativas».

Se trata de la práctica de computar largas series de medidas de asociación entre unas y otras variables, y evaluarlas a través de pruebas estadísticas de hipótesis para identificar aquellas que puedan considerarse significativas. De tal suerte, sin contar con hipótesis teóricamente fundamentadas de antemano, poseedoras de algún grado de plausibilidad, se atrapan «interesantes» asociaciones que, en su momento, serán embalsamadas con la teoría necesaria.

Tal estrategia ha sido reiteradamente objetada con el argumento de que, debido a la propia mecánica de las pruebas, casi siempre aparecerán asociaciones «cazables». En efecto, si se usa un nivel de significación, digamos del 5%, y se manejan por ejemplo 35 variables ***no correlacionadas entre sí, se*** pueden examinar casi 600 distribuciones bivariadas, resultantes de tomar respectivas parejas de variables de entre las 35 originales; cabe esperar que ***sólo*** *por* ***azar,*** unas 30 (5% de 600) de esas parejas exhiban asociaciones que se separen significativamente de la nulidad.

Visto el mismo problema desde otro ángulo, si se eligen 20 muestras independientes de cada una de dos poblaciones ***que no difieren entre sí,*** la probabilidad de hallar al menos una diferencia significativa entre las respectivas medias cuando se realizan las 20 pruebas admitiendo un error de primer tipo del 5%, ***no es*** igual a 0,05 como pudiera pensarse: la probabilidad es ¡nada menos que 0,64! En efecto, ésta es igual al complemento de la probabilidad de no hallar diferencia significativa en ninguno de los 20 intentos:

$$1 - (0{,}95)^{20} = 1 - 0{,}36 = 0{,}64$$

Un recurso que se sugiere para encarar la tarea de evaluar muchas diferencias a la vez es el de reducir el nivel de significación inicial para cada prueba; es decir, en lugar de hacer cada una de las comparaciones usando el nivel 0,05, trabajar con uno mucho menor, suficientemente pequeño como para compensar la facilidad con que se podría capturar la presa. La realidad práctica, por una parte, es que tales técnicas han sido sugeridas y fundamentadas con denuedo, pero su aplicación real es virtualmente nula. Por otra parte, se trata de una seudosolución al problema, ya que el verdadero error no estriba en el nivel de significación con que el investigador esté armado cuando sale a cazar sino en el acto mismo de salir de cacería [6].

6.3.5. Pruebas de significación para datos poblacionales

Es bien conocido que la teoría formal ubica a las pruebas de significación en el ámbito inferencial: ***a partir de los datos muestrales se infieren conclusiones sobre aquella población supuestamente representada por la muestra.*** Consecuentemente, desde un punto de vista teórico no tiene sentido aplicar tales pruebas cuando los datos, en lugar de conformar una muestra, coinciden con todo el universo.

Esa regla se viola con extrema frecuencia; en mi opinión ello no ocurre tanto porque los investigadores no la conozcan como por carecer de recursos alternativos; y el hecho revela no tanto que las pruebas se usen incorrectamente como que son insuficientes para resolver los problemas que se presentan.

El primer caso que cabe considerar es en extremo frecuente. Imaginemos que en un hospital se estudia a ***todos*** los pacientes intervenidos quirúrgicamente durante 1995 y que se calculan las tasas de infección postoperatoria para diferentes grupos de edad. Si los investigadores quieren simplemente describir el comportamiento de la infección postquirúrgica de ese año y para ese hospital, entonces no tiene sentido hacer prueba de hipótesis alguna para comparar las tasas específicas entre grupos de edad, pues la información disponible se maneja en su totalidad. Si se realiza una prueba de hipótesis para evaluar la diferencia entre estas tasas, es necesariamente porque se está suponiendo que los sujetos estudiados conforman una muestra. En tal caso, cabe preguntarse, una muestra ¿de qué población?

Para considerar una respuesta razonable, procede intercalar el concepto de ***superuniverso*** introducido por Stouffer (1934) quien lo definió como «un universo infinito de posibilidades, formado por todos los universos finitos que pudieran haberse producido en el momento de la observación y del cual nuestra población finita puede considerarse una muestra aleatoria».

[6] Hay situaciones (especialmente para discutir el problema de las ***comparaciones múltiples en*** el contexto del análisis de varianza) en que el uso de este recurso -justo es consignarlo- se inscribe como parte de un proceso inferencial secuenciado y no en un acto de búsqueda a toda costa de significaciones.

Si la prueba se lleva a cabo es porque la verdadera pregunta formulada, aun en el caso de que los investigadores no lo declaren así, no se remite a ese hospital ni a ningún hospital específico, sino genéricamente a un superuniverso. La pregunta concierne genéricamente a «la enfermedad nosocomial en sujetos operados»; es decir, atañe a un aspecto de la fisiología humana (la propensión a infectarse) en función de la edad. De modo que la población sobre la que se infiere abarca a todos los sujetos operados, pasados y futuros. Es en ese sentido que la población estudiada puede entenderse como una muestra. La dificultad teórica de mayor entidad, sin embargo, radica no tanto en el hecho de que haya que «inventarse» un superuniverso sino en conseguir «ver» esta «muestra» como el resultado de haber aplicado un procedimiento *aleatorio* de selección en el superuniverso.

De hecho la «solución» del superuniverso es muy singular pues viene a decir, poco más o menos, que como la aplicación de la prueba de significación exige cierta condición que no se cumple, y puesto que se va a aplicar de todos modos, ha de inventarse una elaboración teórica que le dé sentido. Pero el asunto no es tan simple.

Supongamos que la tasa de infección en ancianos asciende a 19,2%, mientras que entre adultos jóvenes es sólo de 8,6%, así como que, al realizar una prueba de hipótesis, esta importante diferencia resulta estadísticamente significativa [7]. Como se ha dicho, la conclusión de que la tasa es mayor para ancianos que para quienes no lo son obviamente no recae sobre los pacientes de ese año en dicho hospital; para establecer que 19,2 es un número mayor que 8,6 no hace falta una prueba estadística: bastan los conocimientos de un escolar. ¿Puede decirse entonces que hay mayor propensión a infectarse entre ancianos que entre jóvenes, independientemente del país en que se realiza el estudio, del tipo de hospital y de las características de los pacientes? Obviamente, no. Pudiera ocurrir que este hospital estuviera enclavado en un área de bajo nivel económico y que si el estudio se repitiera con pacientes en mejor situación (y por ello mejor alimentados, con más instrucción o hábitos de vida más saludables), las dos tasas fueran esencialmente iguales, quizás porque en tal caso la desventaja fisiológica inherente a la vejez no se expresaría a través de la infección.

De modo que en un caso como el que nos ocupa, en el que el enfoque de identificar un *superuniverso* parece ineludible, la identificación del que resulte adecuado no es una tarea susceptible de ser resuelta mecánicamente sino que, una vez más, reclama el concurso del sentido común, del conocimiento profundo de la materia y, al fin, de una dosis inevitable de subjetividad.

Por otra parte, en estrecha relación con ese modo de pensar está la necesidad de superar la convicción, presente en algunos equipos investigadores, de que su traba-

[7] Si ello ocurre o no con esas dos magnitudes depende de los tamaños de los grupos de edad.

jo es *el* que dará *la* respuesta a la pregunta planteada, cuando en realidad seguramente se tratará de *un* trabajo más, llamado a sumarse al debate que la comunidad científica desarrolla sobre el problema. Ese trabajo debe arriesgar una propuesta de superuniverso; en su momento se producirán las acotaciones o enmiendas que procedan. Sólo el cotejo con otros esfuerzos y el examen integrado de resultados -acaso ayudado por enfoques como el del metaanálisis y en medio de un permanente proceso de ajustes- será capaz de generar respuestas científicamente fructíferas.

Hace casi medio siglo que el célebre estadístico Frank Yates advirtió (Yates, 1951):

> ***El énfasis en las pruebas de hipótesis y la consideración de los resultados de cada experimento separadamente han tenido la desafortunada consecuencia de que los trabajadores científicos han considerado con frecuencia la ejecución de una prueba de significación o un experimento como el objetivo último; actúan sobre la base de que los resultados o son o no son significativos y de que éste es el fin de todo.***

Los propios creadores de las pruebas de hipótesis, Egon Pearson y Jerzy Neyman, advirtieron con toda transparencia que éstas no fuesen utilizadas a la manera en que, en definitiva, se hace a diario: como instrumentos para evaluar la validez o la falsedad de una ley científica. En sus propias palabras (Neyman y Pearson, 1933), advertían:

> ***Ninguna prueba basada en la teoría de probabilidad puede por sí misma generar índices válidos sobre la verdad o la falsedad de una hipótesis. Las pruebas de hipótesis deben ser miradas desde otra perspectiva. Siguiendo la regla de aceptar o rechazar una hipótesis no estamos diciendo nada definitivo sobre si la hipótesis es o no verdadera... Lo que se puede demostrar es que si somos consecuentes con esa regla, a la larga, la rechazaremos cuando sea cierta no más, digamos, que una de cada 100 veces; adicionalmente, la rechazaremos con alta frecuencia cuando sea falsa.***

Consideremos ahora una segunda situación. El decano de la facultad de medicina observa que en 1994 el 84% de los alumnos terminó exitosamente el primer año, en tanto que en 1995 este porcentaje «cayó» al 77%. Ante tales datos, acudir a un estadístico para que los analice y le informe si tal reducción es o no «significativa» carece *de todo sentido.* El problema que en ese punto el decano está considerando *no es de naturaleza estadística;* esta disciplina terminó su función una vez que los datos fueron resumidos. Ahora corresponde a un especialista pedagógico, a un político o (si se quiere demorar el asunto) a una comisión, pronunciarse sobre la significación *cualitativa* de los datos. Si esa diferencia de 7% es o no alarmante, si merece o no una indagación más profunda, si cabe adoptar o no medidas especiales, ya son asuntos que no pertenecen a la órbita estadística. Lo que ocurre es que en este caso no hay ninguna población razonable, ni hipotética ni real, ni finita ni infinita, a la cual extrapolar los resultados.

6.4. El juez en el banquillo de los acusados

Hasta aquí se han comentado diferentes interpretaciones y usos de esta herramienta; el debate se ha circunscrito en lo fundamental al ámbito de sus limitaciones operativas y teóricas. Sin embargo, con el tiempo se ha gestado un considerable cúmulo de «cargos» contra la técnica propiamente dicha. Según Morrison y Henkel (1970), la literatura crítica sobre el uso de las pruebas de hipótesis data de la década del 50, aunque la controversia en torno a ellas tiene una historia más larga y profunda.

Cox y Hinkley (1974) alertan sobre la insuficiencia de las pruebas de hipótesis en la medida en que no se complementen con otros recursos. A juicio de ellos, «atribuir interés a una prueba de significación en estado puro es esencialmente una expresión de economía intelectual». Tal declaración, como señala Poole (1987), es la elegantísima y gentil manera elegida por estos autores para repudiar la renuncia al pensamiento crítico. Refiriéndose a quienes usan o enseñan las pruebas de significación con ese fin, este autor es tajante: «no puede haber una manera más cínica y pesimista de ver la ciencia que la sugerencia a los científicos de que economicen su intelecto»; y señala que la discusión crítica en ciencias requiere reflexión y es, en consecuencia, difícil, en tanto que la aplicación mecánica de una regla, no exige pensar y es por tanto muy fácil. La esperanza ingenua de que hallemos una manera simple de llevar adelante una discusión crítica en el campo de la ciencia es una propuesta a que se abandone nuestra habilidad y nuestro deber de pensar. Veamos someramente qué ha ocurrido cuando algunos investigadores han ubicado su pupila crítica sobre las pruebas de hipótesis sin desentenderse de ese deber.

6.4.1 El tamaño muestral decide

Una de las primeras expresiones de «rebeldía» ante la norma de utilizarlas fue protagonizada por Lipset, Trow y Coleman (1956). En un apéndice de ese libro, los autores explican que la decisión de no emplearlas se debe a que las consideran irrelevantes en los estudios no experimentales. Uno de sus más importantes argumentos consiste en que una asociación muy débil puede ser estadísticamente significativa si la muestra es suficientemente grande, en tanto que una muy intensa puede no resultar significativa debido a la reducida magnitud de la muestra. Nótese que no se trata de reclamar cautela debido a esos riesgos sino de rechazar la técnica misma porque ese defecto se considera intrínsecamente descalificador. Esta es, en mi opinión, una crítica fundamental, pues subraya que a la postre la decisión se remite a un componente externo al contenido de los datos: el tamaño muestral.

Savage (1957) llega a decir:

> ***Con mucha frecuencia se sabe de antemano que las hipótesis de nulidad son falsas antes incluso de recoger los datos; en ese caso el rechazo o la aceptación, simplemente***

> ***es un reflejo del tamaño de la muestra y no hace, por tanto, contribución alguna a la ciencia.***

Tal inquietud pudiera parecer de importancia marginal, pues se limita a las situaciones en las que ya se sabe con anticipación que H_o es falsa. Sin embargo, la realidad es que el planteamiento entraña una durísima crítica, ya que, como señalaban Edwards, Lindman y Savage (1963), «en las aplicaciones típicas se sabe desde el principio que H_o es falsa». Si tal afirmación le resulta chocante a alguien es porque en su subconsciente está arraigada la idea de que la falsedad de H_o equivale a la existencia de una relevante relación biológica o epidemiológica, cuando lo único que está diciendo es que la hipótesis no se cumple, aunque sea por escaso margen. Bakan (1966) subraya la idea de manera más explícita y persuasiva:

> ***Es un hecho objetivo que casi nunca hay buenas razones para esperar que la hipótesis nula sea verdadera. ¿Por qué razón la media de los resultados de cierta prueba habría de ser exactamente igual al este que al oeste del río Mississipi? ¿Por qué deberíamos esperar que un coeficiente de correlación poblacional sea igual a 0.00? ¿Por qué esperar que la razón mujeres/hombres sea exactamente 50:50 en una comunidad dada? o ¿por qué dos drogas habrán de producir exactamente el mismo efecto? Una mirada a cualquier conjunto de estadísticas que incluyan totales poblacionales confirmará de inmediato que la nulidad se presenta muy raramente en la naturaleza.***

En síntesis, la decisión sobre la falsedad de la hipótesis nula está casi siempre en manos del tamaño muestral y, por tanto, del investigador que a su vez decide esta magnitud. Para conjurar esta incómoda circunstancia, además de advertir que a la hora de seleccionar el tamaño muestra1 debemos procurar que éste sea suficientemente grande como para poder detectar cierta diferencia mínima, también se ha señalado que debe evitarse que sea tan grande que permita declarar como significativa a una diferencia conceptualmente irrelevante. Abramson (1990), además de suscribir esta opinión, cita la siguiente frase de Sackett (1979): «Las muestras demasiado pequeñas pueden servir para no probar nada; las muestras demasiado grandes pueden servir para no probar nada». Tal opinión equivale a poner en manos de una prueba de significación la decisión acerca de si una asociación es o no clínicamente relevante, algo que -como ya se señaló- es simplemente absurdo.

Pero independientemente de cuál sea la motivación esgrimida, este argumento es descabellado: es obvio que carece de lógica plantear que resulta mejor no observar bien la realidad; es como eludir el uso del microscopio al examinar una muestra serológica, no vaya a ser que se identifique la presencia de una bacteria que no puede ser observada a simple vista. Por otra parte, actuar de ese modo con las pruebas es como si al medir un objeto con una regla se le cambiara la escala en cada ocasión para que dicho objeto tenga la dimensión que se desea de antemano.

6.4.2. Resolviendo la tarea equivocada

Uno de los más interesantes argumentos cuestionadores del valor de las pruebas de hipótesis fue desarrollado por Rozeboom (1960) en un lúcido y agudo artículo cuya conclusión central el autor resume así:

> ***A pesar de la preeminencia que ha alcanzado este método en las revistas experimentales y los textos de estadística aplicada, su empleo se basa en una incomprensión clave de la naturaleza de la inferencia racional; por ello las pruebas de significación casi nunca constituyen un medio apropiado para la investigación científica, si es que lo son en algún caso.***

La línea argumenta1 de Rozeboom se basa en que las pruebas de significación tratan la aceptación o el rechazo de hipótesis como si fueran decisiones que uno hace sobre la base de los datos experimentales, cuando ***el propósito central de un experimento no es precipitar la toma de decisiones sino propiciar un reajuste en el grado en que uno acepta (o cree en) la veracidad de la hipótesis que se valora.*** La creencia en una proposición no es un asunto de todo o nada; la tarea del científico no es prescribir acciones sino establecer convicciones razonables sobre las cuales aquellas habrían de basarse. «Una hipótesis», argumenta Rozeboom, «no es algo como un pedazo de tarta que uno puede aceptar o rechazar a través de una decisión voluntaria». Y prosigue:

> ***Si hay algo que señala la irrelevancia práctica de las pruebas de hipótesis convencionales, es su incapacidad para proveer de genuinos recursos al investigador para su conducta inferencial. ¿Quién alguna vez ha renunciado a una hipótesis sólo*** *porque* ***un experimento produjo un resultado no significativo? El lector muy bien puede permanecer impasible ante los cargos que se imputan a las pruebas precisamente porque, quizás sin comprenderlo, nunca ha tomado el método muy en serio.***

Casi cuatro décadas más tarde, Lilford y Braunholtz (1996) reproducen esta crítica al afirmar que «las pruebas estadísticas convencionales son inadecuadas (...) porque dicotomizan los resultados y tienden a producir una respuesta si/no en los que adoptan decisiones», pero agregan otra al consignar que «ellas no incorporan de manera explícita las evidencias adicionales, hayan sido generadas fuera o dentro del estudio, de manera explícita».

Como oportunamente señaló Poole (1987), la defensa de esta práctica, al menos en epidemiología, es más bien rara y, cuando se produce, parece provenir de autores no muy convencidos. Una excepción ha sido la airada carta de Fleiss (1986) al editor de ***American Journal of Public Health*** en la cual las pruebas son reivindicadas como un recurso razonable para facilitar la toma de decisiones, que provee a los científicos de criterios especificados con antelación para pronunciarse. La mayor

debilidad de este argumento consiste en que confunde la toma de decisiones con el afán por comprender y explicar la realidad.

Lanes y Poole (1984), en un trabajo en que destacan que «aunque existe un precedente histórico de larga data en el empleo de pruebas de significación como recurso básico para interpretar datos epidemiológicos, en realidad no existe ninguna razón sólida para ello», ilustran la separación entre el conocimiento y las decisiones con el siguiente argumento:

> ***El pronóstico matutino del tiempo puede sernos útil para decidir si llevaremos o no el paraguas. El metereólogo puede ofrecer una complicada exposición y una declaración probabilística para vaticinar la inclemencia del tiempo. La decisión final sobre acarrear o no el paraguas, sin embargo, queda en manos del receptor de mensaje; ella puede depender del lapso durante el cual él se hallará al aire libre, de su proclividad a llevar un paraguas en el metro o de que su sobretodo esté disponible. El meteorólogo no nos aconseja acerca de cómo actuar sino que solo nos comunica aquel subconjunto de información sobre el tema que él domina.***

Por otra parte, resulta muy dudoso que sea conveniente el establecimiento de pautas que uniformen mecánicamente las interpretaciones; todo hace pensar que la verdad científica se verá más favorecida por la controversia que por la aplicación de reglas que la desestimulen.

6.4.3. Elementos adicionales para la reflexión

Entre otras consideraciones críticas que se han señalado se hallan las siguientes:

a) Casi todas las pruebas que se han desarrollado están concebidas para el caso en que los datos proceden de una muestra simple aleatoria; este procedimiento muestral sin embargo, casi nunca es el que se utiliza.

b) Puede ocurrir que un tratamiento haya producido un incremento significativamente no nulo tanto para una subpoblación como para su complemento (por ejemplo, tanto para mujeres como para hombres), pero que no llegue a serlo si la prueba se realiza trabajando con la muestra completa. Que ocurra al revés no sería nada sorprendente por la simple razón de que la disminución del tamaño muestra1 deprime la potencia de las pruebas, pero que pueda ocurrir en el sentido descrito es obviamente contraproducente, pues revela una inconsistencia lógica del método.

c) La elección de trabajar con una o con dos colas [8] es un acto altamente conflictivo y dependiente de la visión subjetiva del investigador; los resultados

[8] Aspecto técnico conectado con lo que realmente se está considerando como hipótesis alternativa.

> pueden modificarse sustancialmente en dependencia de que la decisión sea una u otra. Lo inquietante no es que la subjetividad desempeñe un papel —cosa siempre inevitable- sino que, en lugar de reducir la influencia del componente subjetivo, las pruebas de hipótesis puedan servir en realidad solamente para ocultar su participación.

Otros motivos para el examen crítico pueden hallarse en la obra de muchos objetores actuales del uso de las pruebas de hipótesis; entre ellos cabe mencionar a Rothman (1978), Salsburg (1985), Walker (1986), Gardner y Altman (1986), Thompson (1987) y Goodman y Roya11 (1988).

Algún lector puede reaccionar airadamente por el hecho de que estos puntos de vista lo dejarían huérfano de recursos, ya que sólo aportan un enfoque destructivo, sin ofrecer alternativas eficientes. Suponiendo que así fuera, tal reacción sería similar a la de un náufrago que descalifica la información de que beber agua de mar favorece el proceso de deshidratación, no porque los argumentos basados en las concentraciones sódicas a nivel celular le parezcan endebles, sino porque no se le comunica dónde obtener agua potable. Pero lo cierto es que la orfandad sería, en todo caso, de recursos mecánicos. Como señala Lykken (1968):

> ***Una prueba de hipótesis*** **nunca es** ***condición suficiente para concluir que una teoría se ha corroborado, que un hecho empírico útil se ha establecido con confianza razonable, o que el informe sobre un experimento debe ser publicado. El valor de cualquier investigación puede ser determinado, no a partir de resultados estadísticos, sino solo mediante una evaluación incisiva, aunque necesariamente subjetiva, de la coherencia y la racionalidad de una teoría, del grado de control empleado, de la calidad de las técnicas de medición y de la importancia práctica o científica del fenómeno estudiado*** (el subrayado es de Lykken).

Por otra parte, en vista de las limitaciones de las pruebas de hipótesis, se han manejado algunas alternativas; a modo de ilustración general, a ellas se destina la próxima sección.

6.5. Las alternativas

6.5.1. Intervalos de confianza

Desde hace algún tiempo se ha venido imponiendo la práctica de suplir las pruebas de hipótesis por intervalos de confianza, alternativa defendida con vehemencia por diversos autores entre los que se destacan los connotados estadísticos británicos Martin Gardner y Douglas Altman (véanse Gardner y Altman, 1986; Gardner y Altman, 1987) y secundada actualmente por muchos editores de revistas médicas.

Estos autores sugieren que los intervalos sean empleados como recurso expresivo *básico (*no se indica que los valores *p* sean necesariamente eliminados pero, en cualquier caso, sí que ocupen un lugar secundario) «siempre que se haga una inferencia de los resultados a un ámbito más abarcador y que concierna a medidas de resumen (no a características individuales) tales como tasas, diferencias de medianas, coeficientes de regresión, etc». Se advierte que si bien «el grado de confianza más usado es 95%, cualquier intento de estandarizarlo resulta indeseable».

El argumento central en que se sustenta esta corriente de opinión proclama que los intervalos son mucho más informativos que el mero valor de ***p***, ya que éste no entraña información alguna sobre la magnitud de la diferencia o de la asociación que se valora, en tanto que el intervalo nos provee de un recorrido de valores posibles para el valor poblacional en lugar de una dicotomía arbitraria. Se añade, finalmente, que los intervalos incluyen toda la información necesaria para aplicar la prueba de significación si se deseara realizarla. Cuando se trata de valorar la diferencia entre dos parámetros, lo que debe hacerse es un intervalo de confianza para la diferencia. Es decir, se alerta (Altman, 1980) que en tal caso no se.construyan dos intervalos (uno para cada parámetro); en caso de que los datos estuvieran pareados, tal práctica no sería solamente inconveniente sino directamente errónea.

En mi opinión, el uso de intervalos de confianza es menos inadecuado que el de las pruebas de hipótesis, en especial porque proveen más información y aportan un enfoque más flexible.

De hecho, si se procede a estimar, pongamos por caso, una diferencia entre medias muestrales y se corrobora que el error de esa estimación es suficientemente pequeño como para considerar que dicha estimación es eficiente (por ejemplo que el error es inferior al 5% de la magnitud de la propia estimación), lo cual equivale a que el intervalo de confianza sea suficientemente estrecho, ya se cuenta con toda la información necesaria para hacerse un juicio acerca del problema abordado, ya que lo que corresponde es simplemente pronunciarse sobre la sustantividad o significación clínica de la diferencia en cuestión.

Pero lo cierto es que los intervalos de confianza están sujetos a buena parte de las críticas antes señaladas para las pruebas de hipótesis si sólo se emplean como sus sucedáneos. Ésta es la base en que se apoya otra corriente de opinión que se bosqueja en la sección que sigue.

6.5.2. Verosimilitud de las hipótesis

Un enfoque alternativo, cuya base lógica es sin dudas interesante, aunque no ha tenido éxito alguno en la práctica (no conozco de ninguna tendencia a sugerirlo como recurso por los editores de revista alguna, ni sé de trabajos trascendentes que lo hayan empleado) ha sido el de usar el criterio *del peso probatorio de las hipótesis*, que según sus defensores refleja la verdadera base sobre la que se realizan las inferencias.

Goodman (1992) ridiculiza el principio que justifica el uso de los valores -es decir: que si la observación de ciertos datos es un fenómeno raro bajo el supuesto de que se cumple cierta hipótesis, entonces ellos tienen un valor indicativo, probatorio, contra dicha hipótesis- con un ejemplo en que se supone que en la ruleta de un casino sale la secuencia de números 3 - 14 - 6 - 27. Bajo el supuesto de que el dispositivo otorga igual probabilidad a cada uno de los 37 números posibles, la probabilidad de tal suceso es $(1/37)^4 = 0{,}0000005$. Goodman señala que nadie interpretaría esta bajísima probabilidad como indicio de que es falsa la hipótesis que se ha usado para calcularla: que la ruleta no está trucada. De modo que «no solemos interpretar los hallazgos que resulten raros bajo una hipótesis dada como indicios probatorios en su contra».

De hecho equipara esa lógica con la que se usa en las pruebas de significación al preguntar: «¿Qué ocurre si un estudio encuentra una asociación imprevista con una $P = 0{,}01$? ¿Correremos a publicar? Lamentablemente, la respuesta es sí en este caso». Esta línea de pensamiento se basa en que los datos en sí mismos no otorgan «peso probatorio» a favor (o en contra) de una hipótesis salvo que exista una hipótesis rival bajo la cual estos datos sean más (o menos, según el caso) probables.

La argumentación de Goodman es, sin embargo, algo tendenciosa, pues cualesquiera 4 números que salgan en la ruleta tendrán la misma probabilidad que corresponde a la secuencia señalada; sin embargo, es obvio que al aplicar, por ejemplo, la *prueba t de Student* para comparar dos medias muestrales, los posibles valores que puede tomar el estadígrafo son infinitos, y también lo son las respectivas probabilidades asociadas. Esa es la primera diferencia entre las dos situaciones. Por otra parte, en las aplicaciones reales, como la de comparar dos medias, el modo en que desarrollan las pruebas de hipótesis es tal que H_o tiene una expresión paramétrica clara, cosa que no ocurre con su ejemplo de la ruleta, tal y como lo maneja. Por ejemplo, si se quiere expresar paramétricamente la hipótesis de que la ruleta no está trucada ella podría, postularse:

$$H_o : P_1 = P_2 = ... = P_{37} = \frac{1}{37}$$

donde P_i denota la probabilidad de que sea el i-ésimo de los 37 resultados posibles el que salga ($i = 0, 1, 2, \ldots . 36$), y la experiencia para evaluarla podría consistir en echar a andar la ruleta, por ejemplo 10.000 veces, para luego realizar una prueba de bondad de ajuste a través de [6.1], donde $e_i = \frac{10.000}{37}$ y O_i es la frecuencia con que aparece el número *i*.

Otra variante más específica sería considerar:

$$H_o : P\ (3,\ 14, 6, 7) = \left(\frac{1}{37}\right)^4$$

donde ***P (n„*** n_2, n_3, n_4**)** denota la probabilidad de que en 4 usos de la ruleta salgan los números n_1, n_2, n_3 ***y*** n_4 en ese orden. En tal caso, la experiencia podría ser la siguiente: se rueda la ruleta 10 millones de veces; si la fracción en que se produce la secuencia 3-14-6-27, por ejemplo, fuera 0,25 en lugar de 5 diezmillonésimas (o un número próximo a éste último, que es «el que le toca» bajo la hipótesis de nulidad) se diría que hay indicios muestrales en contra de H_o. Pero si la experiencia se hace una sola vez, el resultado es ininformativo, pues no hay oportunidad de computar ninguna frecuencia empírica. La probabilidad $(1/37)^4$ para la secuencia mencionada es teórica y se conoce de antemano: no es un cálculo que se haga usando la hipótesis nula sino que en eso consiste precisamente la hipótesis misma.

Sin embargo, aunque el argumento inicial resulte objetable, es muy cierto que el valor de ***p*** en cualquier prueba depende de una sola hipótesis (la nula) y no informa nada sobre hipótesis alternativa alguna. El carácter probatorio de unos datos sobre cierta hipótesis sólo existe en la medida en que se exprese en términos relativos; es decir, solamente tienen tal virtualidad cuando los datos son más compatibles con una que con otra hipótesis.

El enfoque de usar razones de verosimilitud propone operar con la probabilidad condicional de los datos en función de diferentes hipótesis. Por ejemplo, supongamos que, al observar que los 4 números mencionados ocupan espacios contiguos en la ruleta, un investigador sospecha que el dispositivo está trucado y formula la hipótesis alternativa siguiente:

$$H_1 : \mathrm{P}\ (3,\ 14, 6, 7) = \left(\frac{1}{37}\right)^3$$

Entonces, tras hacer la experiencia arriba descrita (10 millones de experimentos con la ruleta), se puede computar la ***razón de verosimilitud:***

$$RV = \frac{P\ (datos \mid H_o)}{\boldsymbol{P\ (datos \mid H,)}}$$

donde ***P (datos*** $\mid H_o$**)** es la probabilidad de que, valiendo H_o, se hayan obtenido los datos que se observan y ***P (datos*** $\mid H_1$**)** es lo mismo pero bajo el supuesto de que rige ***H,;*** luego se preferiría H_1 a H_o con tanta mayor intensidad cuanto menor que 1 sea ***RV.***

Este enfoque podría tener un espacio real en un futuro no lejano; en cualquier caso, cabe esperar la aparición de alternativas a una práctica que perdura más por la ausencia de tales alternativas que por sus méritos.

6.5.3. El advenimiento de la era bayesiana

La propuesta esbozada en el apartado anterior se inscribe dentro de la llamada ***estadística bayesiana.*** El elemento distintivo entre la escuela bayesiana y la estadística frecuentista parte de que ésta entiende la probabilidad como la frecuencia relativa con que ocurre un acontecimiento dentro de un marco experimental bien definido, en tanto que aquella la aprecia como un grado de convicción personal que, si bien puede modificarse a la luz de nuevos datos, no depende en principio de resultados experimentales u observaciones formales. Para un «frecuentista» estricto una oración cotidiana tal como «es muy probable que Pedro haya sido el ladrón» carece de todo sentido; para un bayesiano esto sería perfectamente sensato, (como lo es, por cierto, para todo ciudadano común).

Esta ***probabilidad a priori*** de transforma entonces en una ***probabilidad a posteriori*** para cuya construcción se emplean los datos de que se disponga en cada momento específico.

Estas convicciones ***a priori*** son necesariamente subjetivas, pueden variar de un momento a otro y de un analista a otro, pero ello no constituye una acusación demasiado grave, pues tal es la manera más natural de reflexionar en un marco de incertidumbre. Las probabilidades ***a posteriori,*** por su parte, configuran el grado de convicción prevaleciente luego de haberse incorporado la nueva información al conocimiento precedente.

Para fijar las ideas acerca de la diferencia de enfoques ante un problema concreto, imaginemos que se discute si una técnica quirúrgica novedosa es más eficaz que otra convencional. Supongamos que se aplican ambas técnicas a respectivos grupos de 40 pacientes: 27 se recuperan cuando son intervenidos mediante el procedimiento novedoso en tanto que se alcanza un resultado exitoso en solo 21 de los tratados con la vieja tecnología.

El enfoque frecuentista clásico procedería a realizar una prueba de hipótesis de comparación de porcentajes (equivalente a una prueba de Ji-cuadrado) y obtendría el valor $p = 0{,}17$. Verosímilmente, el investigador, que había quedado contento con la apreciable diferencia observada, tendrá que admitir (probablemente a regañadientes y sin creérselo en el fondo) que «la diferencia no es significativa», que no hay evidencia muestra1 suficiente como para rechazar la hipótesis que afirma que ambas técnicas son igualmente eficaces.

El enfoque bayesiano seguirá otro trillo: en lugar de abstenerse de sacar conclusión alguna, conseguirá «actualizar» el grado de convicción que prevalecía hasta ese momento sobre las técnicas rivales; los datos resultantes del ensayo clínico constituirán la materia prima para dicha «puesta al día». Es obvio que esta manera de apreciar la información novedosa se acerca mucho más al razonamiento natural de todo investigador, ya que es absurdo desdeñar una información por el solo hecho de que no permita hacer juicios categóricos. No en balde el cultor de las técnicas clásicas acepta, si es que lo hace, de mala gana determinados «no rechazos de la hipótesis»; y

la realidad es que con frecuencia, después del fracaso en el empeño de ***hallar significación,*** suele reaccionar con el típico: «si bien en nuestro estudio no se ha podido rechazar la hipótesis de nulidad, es posible que con una muestra más grande...».

Las técnicas cuantitativo-computacionales con que opera el enfoque bayesiano para conseguir expresar la probabilidad que sintetiza el nuevo estado de opinión desborda el nivel (y, sobre todo, el objetivo) de este libro. Además de su substrato natural -el teorema de Bayes- tales técnicas exigen cierto dominio de la teoría probabilística de distribuciones y, a los efectos de computar la probabilidad ***a posteriori,*** de los recursos de simulación tipo montecarlo. Esto conduce casi inevitablemente al empleo de programas computacionales no incluidos en los paquetes estadísticos convencionales.

Una ilustración relativamente simple de cómo aplicar esta técnica, concerniente al presunto efecto pernicioso de contraceptivos de tercera generación, puede consultarse en Lilford y Braunholtz (1996). Estos autores, informan además acerca de cómo obtener un ***software*** (el programa BUGS) apropiado para llevar adelante el procedimiento bayesiano.

Ahora bien, cabe recalcar que las técnicas bayesianas son considerablemente más complejas que las que se han venido aplicando en el campo de la salud durante los últimos decenios. Precisamente, el predominio del paradigma frecuentista puede explicarse parcialmente por esa razón; a ello se suma, como señala Freedman (1996) la dificultad de establecer la probabilidad *a priori,* empresa muy alejada de la pereza intelectual que suele asociarse al empleo de las pruebas de hipótesis.

Bacallao (1996) consigna que:

> ***Otra razón podría tener que ver con la hipótesis Kuhniana de que los paradigmas sobreviven mientras su fracaso no es demasiado frecuente y ostensible, y realmente, el enfoque clásico de la estadística inferencia1 continúa brindando respuestas válidas a un buen número de problemas y una engañosa apariencia de solución a otro buen número.***

El paradigma frecuentista, en especial con relación a las pruebas de significación, empieza, sin embargo, a mostrar síntomas claros de agotamiento; por tanto, no me quedan dudas de que el acercamiento gradual a las técnicas bayesianas constituye actualmente un saludable posicionamiento para todo profesional interesado en la investigación que quiera consolidar la cultura estadística.

Bibliografía

Abramson JH (1990). ***Survey methods in community medicine: Epidemiological studies, program evaluation, clinical trials.*** Churchill Livingstone, 4.ª ed, Edinburgh.

Altman DG (1980). ***Statistics and ethics in medical research: VI-presentation of results.*** British Medical Journa1281: 1542-1544.

Bacallao J. (1996). ***La perspectiva exploratorio-confirmatoria en las aplicaciones biomédicas de la estadística: dos diálogos (1) Bayesianismo frente a frecuencialismo: sus respectivas implicaciones prácticas en relación con el análisis de datos.*** Medicina Clínica 107: 467-471.

Bacallao J. (1996). ***La perspectiva exploratorio-confirmatoria en las aplicaciones biomédicas de la estadística: dos diálogos (y II) Consideraciones cn'ticas acerca de las pruebas de significación.*** Medicina Clínica 107: 539-543.

Bakan D (1966). ***The test of significance in psychological research.*** Psychological Bulletin 66: 423-437.

Cox DR, Hinkley DV (1974). ***Theoretical statistics.*** Chapman and Hall, Londres.

Edwards W, Lindman H, Savage LJ (1963). ***Bayesian statistical inference for psychological research.*** Psychological Review 70: 193-242.

Fisher R (1935). ***The design of experiments.*** Oliver and Boyd, Londres.

Feinstein AR (1985). ***Clinical epidemiology: The architecture of clinical research.*** W.B. Saunders Company, Philadelphia.

Fisher R (1959). ***Statistical methods and scientifical inference.*** Oliver and Boyd, 2.ª ed, Edinburgh.

Fleiss JL (1986). ***Significance tests have a role in epidemiologic research: reactions to A.M. Walker:*** (Different Views) American Journal of Public Health 76: 559-560.

Freedman L (1996). ***Bayesian statistical methods. A natural way to assess evidence.*** (Editorial) Bristish Medical Journal 313: 569-570.

Galbraith JK (1991). ***La cultura de la satisfacción.*** Ariel Sociedad Económica, 2.ª ed, Madrid.

Gardner MJ, Altman DG (1986). ***Confidence intervals rather than P values: estimation rather than hypothesis testing.*** British Medical Journal 292: 746-750.

Gardner MJ, Altman DG (1987). ***Using confidence intervals.*** Lancet i: 746.

Goodman SN, Roya11 R (1988). ***Evidence and scientific research.*** America Journal of Public Health 78: 1568-1574.

Goodman SN (1992). *p* ***values, hypothesis test, and likelihood: implications for epidemiology of a neglected historical debate.*** American Journal of Epidemiology 137: 485-495.

Jeffreys H (1961). ***Theory of probability.*** Oxford University Press, 3.ª ed, Oxford.

Joyce C, Welldon R (1965). ***The objective efficacy of prayer: a double blind clinical trial.*** Journal of Chronic Diseases 18: 367-372.

Khan HA, Sempos CT (1989). ***Statistical methods in epidemiology.*** Oxford University Press, New York.

Kendall MG, Stuart A (1983). ***The advanced theory of statistics.*** 4.ª Ed, Hafner Press, New York.

Lanes SF, Poole C (1984). ***The unwrapping of epidemiologic research.*** Journal of Occupational Medicine 26: 571-574.

Lilford RJ, Braunholtz D (1996). ***The statistical basis of public policy: aparadigm shift is overdue.*** Bristish Medical Journal 313: 603-607.

Lipset SM, Trow MA, Coleman JS (1956). ***Statistical problems.*** Apéndice 1-B de ***Union Democracy, Glencoe,*** Free Press (427-432) Illinois.

Lykken DT (1968). ***Statistical significance in psychological research.*** Psychological Bulletin 70: 151-159.

Morrison RE, Henkel DE. (1970). ***The significance test controversy*** Aldine Publishing Company, Chicago.

Neyman J, Pearson E (1933). *On* ***the problem of the most efficient tests of statistical hypotheses.*** Philosophical Trans of the Roya1 Society of London A 231: 289-337.

Paulos JA (1990). ***El hombre anumérico.*** Alfaguara, Madrid.

Poole C (1987). ***Beyond the confidence interval.*** American Journal of Public Health 77: 195-199.

Rothman KJ (1978). *A* ***show*** **of** ***confidence. New*** England Journal of Medicine 299: 1362-1363.

Rozeboom WW (1960). ***The fallacy of the null hypothesis significance test.*** Psychological Bulletin 56: 26-47.

Sackett DL (1979). ***Bias in analytic research.*** Journal of Chronic Diseases 32: 51-57.

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

Savage IR (1957). ***Nonparametric statistics.*** Journal of the American Statistical Association 52: 332-333.

Sheehan TJ (1980). ***The medical literature: Let the reader beware.*** Archives of Internal Medicine 140: 472-474.

Silva LC (1995). ***Excursión a la regresión logística en ciencias de la salud.*** Díaz de Santos, Madrid.

Stouffer SA (1934). ***Sociology and sampling*** en Bernard LL (Ed) ***Fields and methods of sociology,*** Long and Smith, New York.

Thompson WD (1987). ***Statistical criteria in the interpretation of epidemiologic data.*** American Journal of Public Health 77: 191-194.

Walker AM (1986). ***Reporting the results of epidemiologic studies.*** American Journal of Public Health (Different Views) 76: 556-558.

Winer BJ (1971). ***Statistical principles in experimental designs.*** McGraw Hill, 2.ª ed, New York.

Yates F (1951). ***The influence of Statistical Methods for Research Workers on the development of the science of statistics.*** Journal of the American Statistical Association 46: 32-33.

Yates F (1968). ***Theory and practice in statistics.*** Journal of the Roya1 Statistical Society (Series A) 131: 463-475.

7

El acertijo de la causalidad

El realismo es definitivo y el idealismo es prematuro. Ni uno ni otro poseen esa actualidad que reclama el pensamiento científico.

GASTON BACHELARD

El presente capítulo, por encima de todo, se propone dar una semblanza del debate que promueve el tema de la causalidad, tanto desde una perspectiva histórica como en lo que concierne a sus actuales circunstancias. Siendo la misión central de la ciencia la de explicar la realidad (vale decir: identificar los mecanismos causales que la gobiernan) y, siendo la estadística una de las herramientas más utilizadas para examinar sus expresiones empíricas, en el campo de la salud la problemática de la causalidad está presente, tácita o explícitamente, en muchas de las secciones del libro. Este capítulo se destina íntegramente a la reflexión genérica sobre el tema, aunque poniendo el énfasis, naturalmente, en el entorno investigativo de la sanidad.

A mi juicio, la determinación de relaciones causales en materia de salud, además de poseer una complejidad intrínseca, está sumida en una tupida madeja de malentendidos. Por una parte se hallan los que se derivan del sustrato filosófico sobre el que se erige el problema, que es de por sí controvertido; por otra, están los malentendidos que se originan, una vez más, a raíz de la confusión prevaleciente entre los conceptos y su manejo operativo.

Para iniciar la discusión, examinaremos brevemente el enclave histórico del problema y las principales líneas del debate teórico; luego, el problema conceptual de la causalidad en sí y, finalmente, algunos aspectos prácticos.

7.1. El paradigma inductivo-verificacionista

Desde el punto de vista metodológico, la estrategia teórica predominante para la identificación de cadenas causales ha descansado en la combinación inducción-verificación, proceso racional para el hallazgo de explicaciones generales que se inicia con la observación de lo particular, procede a la generación de hipótesis, y cierra su ***modus operandi*** verificándolas empíricamente.

La esencia de esta línea de pensamiento, sintetizada en palabras de Russell (1949), es la siguiente:

> ***El silogismo inductivo opera del modo siguiente: si cierta hipótesis fuera verdadera, entonces determinados hechos habrían de ser observados; ahora bien, estos hechos son observados; consiguientemente, la hipótesis es verosímilmente verdadera.***

La endeblez de tal silogismo, conocido como ***la afirmación del consecuente,*** estriba en que la conclusión no resulta de una deducción lógica que parta de las premisas; tal limitación fue señalada hace más de dos siglos por el filósofo inglés David Hume (1711-1776). Desde entonces, una discusión, que en ocasiones parece bizantina, ha venido produciéndose.

El debate se verifica actualmente en ese mismo marco de corte filosófico y con especial referencia a la epidemiología. Me refiero a la discusión desarrollada en torno al pensamiento del connotado epistemólogo austriaco Karl Popper (1902-1994). Vale la pena internarse, siquiera someramente, en los entresijos de esta apasionante polémica.

El célebre manifiesto que los inquisidores prepararon para que Galileo Galilei (1564-1642) leyera en el Convento de Minerva el 22 de junio de 1633 refleja con nitidez la naturaleza dogmática del pensamiento dominante durante la oscura época que le tocó vivir. En uno de sus segmentos, la famosa retractación decía:

> ***Yo, Galileo Galilei, hijo del difunto Vincenzo Galilei, de Florencia, de setenta años de edad, siendo citado personalmente a juicio y arrodillado ante vosotros, los eminentes y reverendos cardenales, inquisidores generales de la República universal cristiana contra la depravación herética, teniendo ante mí los Sagrados Evangelios, que toco con mis propias manos, juro que siempre he creido y con la ayuda de Dios, creeré en*** *lo futuro,* ***todos los artículos que la Sagrada Iglesia católica y apostólica de Roma sostiene, enseña y predica. Por haber recibido orden de este Santo Oficio de abandonarpara siempre la opinión falsa que sostiene que el Sol es el centro e inmóvil, siendo prohibido el mantener, defender o enseñar de ningún modo dicha falsa doctrina; y puesto que después de habérseme indicado que dicha doctrina es repugnante a la Sagrada Escritura, he escrito y publicado un libro en el que trato de la misma condenada doctrina y aduzco razones con gran fuerza en apoyo de la misma, sin dar ninguna solución; por eso he sido juzgado como sospechoso de herejía, esto es, que yo sostengo y creo que el Sol es el centro del mundo e inmóvil, y que la Tierra no es el centro y es móvil, deseo apartar de las mentes de vuestras eminencias y de todo católico cristiano esta vehemente sospecha, justamente abrigada contra mi; por eso, con un corazón sincero y fe verdadera, yo abjuro, maldigo y detesto los errores y herejías mencionados, y en general, todo error y sectarismo contrario a la Sagrada Iglesia; y juro que nunca más en el porvenir diré o afirmaré nada, verbalmente o por escrito, que pueda dar lugar a una sospecha similar contra mí*** (citado por Russell, 1949).

El planteamiento verificacionista en su versión moderna [1] nació como reacción al dogmatismo medieval encarnado en este extraordinario documento. En lo que constituyó un trascendente acto de emancipación intelectual, el nuevo discurso se vertebró en torno a la declaración de que el mundo real, susceptible de ser observado cuidadosamente, constituía la única autoridad reconocida por la ciencia. La actividad científica se sacudió así un pesadísimo lastre. Pero no fue hasta el siglo XIX cuando el inductivismo alcanzó el prestigio que aún hoy posee, después de que el naturalista británico Charles Darwin (1809-1882) se apoyara en él para asestar un golpe demoledor a la biología creacionista.

El enfoque inductivo establece, por una parte, que una teoría puede (y debe) emerger de la observación, y por otra, que dicha teoría ha de ser verificada a través de la contrastación con la práctica de sus derivaciones. En su versión más pura, tal y como lo defendía Francis Bacon (1561-1626), el inductivismo reclama que la observación se desarrolle a partir de un estado de virginidad teórica, la misma convicción que, tres siglos después, mantendría un investigador de la talla del premio Nobel Konrad Lorenz al sostener (Lorenz, 1950):

> ***Toda idea preconcebida es sumamente peligrosa para la investigación inductiva. La ciencia natural inductiva ha de comenzar irrenunciablemente con la observación, y hacerlo de manera independiente a toda teoría o hipótesis previa***.

Si hay varias alternativas de explicación a cierto fenómeno, el inductivismo optará por aquella que haya sido más veces verificada; y cuantas más confirmaciones empíricas reciba, más proclives estará a considerarla cierta. Si, además, esa teoría es capaz de hacer predicciones, tendrá más valor científico como tal; y si, finalmente, tales predicciones se cumplen, se concederá a la teoría un sólido valor explicativo.

El planteamiento alternativo que se expone a continuación reniega, punto por punto, de este enfoque.

7.2. Popper entre el absurdo y la genialidad

Avanzado el siglo XX emerge un enfoque alternativo, a primera vista absurdo: la clave para avanzar no se halla en la verificación sino en su antítesis, la refutación. Para Popper, impulsor protagónico de una exitosa corriente filosófica, el método inductivo es lógicamente inválido y, por tanto, esencialmente estéril.

Según Popper (1972), la ciencia nunca establece sus resultados partiendo de la

[1] No conozco casi ninguna reflexión filosófica que no tenga algún antecedente en la antigüedad, específicamente en el mundo griego. El inductivismo no es una excepción: ya Aristóteles, por ejemplo, había exaltado el papel de la observación objetiva como medio fundamental para la obtención de nuevos conocimientos.

observación de casos singulares para llegar por esa vía a la formulación de leyes generales: el único procedimiento aceptable es el hipotético-deductivo. Éste exige formular conjeturas, que parten de la intuición y no de la observación y cuya plausibilidad tiene valor marginal. En rigor, Popper concede incluso más interés a la hipótesis que más se aleje de lo que el conocimiento vigente permite esperar, ya que, si no pudiera refutarse, tal resultado sería mucho más informativo. Estas conjeturas serán blanco inmediato de una crítica despiadada, orientada a demostrar su incorrección. El procedimiento para probar la falsedad de tales hipótesis se basaría en la ***deducción*** de sus posibles consecuencias, las cuales se intentaría refutar por vía preferiblemente experimental. Las teorías nunca pueden ser corroboradas en sí mismas, sino sólo en relación con teorías alternativas que resulten más claramente refutadas. Cuando una teoría resiste los intentos de derribarla, la preferencia por esa explicación se incrementa.

La lógica de ese enfoque reposa en un hecho formalmente cierto: por muchas confirmaciones que se consigan para un resultado, éste no queda por ello demostrado; un solo caso negativo, sin embargo, permite descartarlo como cierto.

Para los interesados en este tema hago una breve digresión deteniéndome en la desconcertante «paradoja de Hempel» que, en su versión clásica y de manera concisa, puede formularse como sigue. La afirmación: «Todos los cuervos son negros» es lógicamente equivalente a su contrarrecíproco: «Todo aquello que no sea negro, no puede ser un cuervo». Si admitimos -como afirma el inductivismo- que toda corroboración empírica de una teoría aumenta la verosimilitud de su veracidad (por ejemplo, que al observar muchos cuervos y corroborar que son negros, ello reafirma la convicción de que ***todos*** los cuervos tienen ese color), habrá que admitir que cualquier corroboración de su proposición equivalente, también aumenta su credibilidad. Así las cosas, la observación, por ejemplo, de una rana verde o de una nube blanca sería una confirmación empírica de que lo que no es negro no es un cuervo y, por ende, de la afirmación inicial. Sintetizando, la observación de un objeto que no sea negro ni sea un cuervo, constituye, de acuerdo con el pensamiento inductivo, un aporte adicional a la creencia de que todos los cuervos son negros [2].

Tal conclusión es muy poco intuitiva pero no llega a ser contradictoria. Sin embargo, unos segundos de reflexión persuadirán al lector de que el mismo razonamiento conduce a que la observación de una rana verde apoyaría la hipótesis de que todos los cuervos son amarillos, lo cual sí entraña una contradicción, ya que la misma observación conduciría a conclusiones incompatibles entre sí. Sintetizando, la combinación del pensamiento inductivo con un razonamiento lógico impecable conduce a una sinrazón; el refutacionismo, en cambio, está a salvo de esa dificultad pues no concede valor a las confirmaciones sino solamente a los contraejemplos: un solo cuervo que no sea negro bastaría para descartar esta última hipótesis, y el fra-

[2] Entre muchísimas fuentes de consulta sobre esta paradoja, sugiero el libro debido a Gardner (1975), que es, a la vez, divertido y claro.

caso en el intento de hallar cuervos no negros prestigia la hipótesis de que todos lo sean sin dar por concluido el asunto.

Podría pensarse que el enfoque popperiano, al plantear que el progreso científico discurre a través de conjeturas y refutaciones y no mediante comprobaciones, no constituye realmente una visión opuesta al inductivismo, pues ¿no son acaso *verificación y refutación* categorías complementarias?, ¿no es en definitiva equivalente *refutar* la validez de una hipótesis a *aceptar* la afirmación que la contradice? La respuesta es negativa, y en ese detalle radica parte del mérito de Popper. La negación de una hipótesis explicativa bien definida entraña una multitud de explicaciones alternativas. La confirmación de que una asociación es real (en el sentido de que no puede explicarse solamente por el azar) es, en efecto, equivalente a refutar la afirmación de que tal asociación es debida a la casualidad; pero no identifica a ninguna de las muchas otras explicaciones que pueda tener el hecho de haberla observado, tales como el efecto de uno o más sesgos, los errores de medición, o la interferencia de factores de confusión. Popper y sus seguidores señalan por consiguiente que siempre pueden existir explicaciones alternativas, y reclaman una permanente actitud de escepticismo que nos resguarde de la tentación de dar por cierto algo por el mero hecho de que se ha observado muchas veces.

Los inductivistas desdeñan el apego estricto a esta doctrina, considerado por ellos paralizante, a la vez que reivindican importantes éxitos en materia de causalidad. La historia de la ciencia biomédica les concede, en principio, una importante cuota de razón. En el ambiente clínico, por ejemplo, es emblemática la historia de Ignacio Semmelweis quien, a mediados del siglo pasado, muchos años antes del descubrimiento de las bacterias, sostenía que la mayoría de los episodios de fiebre puerperal y, consecuentemente, la mayoría de las muertes maternas, podían evitarse mediante el simple expediente de que los médicos se lavaran sus manos con una solución antiséptica antes de manipular a las parturientas. Todo hace pensar que el pensamiento del galeno húngaro siguió un exitoso trayecto inductivo que más tarde consiguió contundentes avales empíricos: él mismo los obtuvo al reducir (y virtualmente a eliminar) la mortalidad materna en su servicio, mientras que los restantes hospitales europeos seguían exhibiendo pavorosas tasas de mortalidad materna, que oscilaban entre 10% y 20%. En esta fase, aunque no se tratara de un ensayo clínico controlado en su sentido moderno, la hipótesis en cuestión ya pasaba por el riguroso dictamen de la confirmación experimental.

En el ambiente epidemiológico, el ejemplo del SIDA refleja con transparencia el éxito relativo del pensamiento inductivo: por ejemplo, desde la aparición de la epidemia se observó una similitud tal entre sus grupos de riesgo y los de la hepatitis que la hipótesis de que actuaba un agente infeccioso no tardó en aparecer; este y muchos otros aspectos de los estudios concretos sobre SIDA se articulan coherentemente por Ng (1991) en defensa del camino inductivo, mostrando que, aun padeciendo, en efecto, de cierta inconsistencia lógica, tiene una expresión operacional comprobada y ha dejado dividendos tangibles.

En ese punto del debate cabe mencionar un elemento que coloca a este dilema epistemológico en un segundo plano. Incluso aunque no se suscriban afirmaciones como la de Schlesinger (1988), quien considera innecesaria la reflexión filosófica sobre aspectos gnoseológicos generales de los investigadores [3], es imposible librarse del inaplazable imperativo práctico de adoptar decisiones. Si sólo fuera posible refutar hipótesis y no confirmarlas, ¿sobre qué bases, entonces, diseñar programas, elaborar normas, difundir mensajes sanitarios, fijar políticas y emitir regulaciones o leyes?

Puestas las cosas en ese marco, los inductivistas suelen acudir no sin cierta razón a la caricatura. Por ejemplo, considerando el tan llevado y traído ejemplo del tabaquismo como causa del cáncer pulmonar, Greenland (1988) parodia el punto de vista popperiano, al expresar que, en buena lid, éste exigiría que, en lugar de eliminar la publicidad al tabaco, se orientara a los publicistas a que difundieran el siguiente mensaje:

> **LA HIPÓTESIS DE QUE EL HÁBITO DE FUMAR ES DAÑINO PARA SU SALUD ES DE MOMENTO LA MEJOR EXPLICACIÓN PARA LO QUE SE HA OBSERVADO PERO, TAL Y COMO OCURRE CON TODAS LAS TEORIAS, ESTA NO DEBE CONSIDERARSE COMO INEQUÍVOCAMENTE CIERTA.**

Llevando la idea a su extremo, yo agregaría que a nuestros alumnos de medicina no habríamos de insistirle en la obligación de desinfectarse las manos antes de realizar el trabajo de parto; en su lugar, les diríamos:

> **LA CONCLUSION A QUE ARRIBÓ SEMMELWEIS CARECE DE VALOR DEFINITIVO; LA PRÁCTICA DEL ÚLTIMO SIGLO Y MEDIO ES UNA CORROBORACION EPISTEMOLÓGICAMENTE INVÁLIDA, PUES NADA NOS ASEGURA QUE TODO NO HAYA SIDO OBRA DE CIERTA RAZÓN AÚN IGNORADA, AJENA A LAS BACTERIAS.**

Aquí cabe intercalar, sin embargo, que la promulgación de leyes, políticas o mensajes, y en general la toma de decisiones, se produce en medio de un entramado de restricciones. Uno de ellos es, sin duda, el conocimiento prevaleciente en cada momento histórico, independientemente del modo como se haya obtenido. Pero el peso que tenga ese conocimiento será mayor o menor según las circunstancias, pues los mensajes y las regulaciones, querámoslo o no, siempre serán, en el mejor de los casos, una versión política de las convicciones técnicas. Incluso, en el caso improbable de que las decisiones solamente dependieran del conocimiento científico acumulado, siempre será necesaria, al menos, una adecuación al lenguaje del desti-

[3] **Para Schlesinger la reflexión filosófica no es de ayuda alguna al investigador, al igual que el artista de la cuerda floja no ha de conocer las leyes de la física para realizar su espectáculo.**

natario. Por otra parte, de los irónicos puntos de vista críticos al refutacionismo parece desprenderse que lo ideal sería que los mensajes sanitarios fueran siempre categóricos; sin embargo, en mi opinión, tampoco estaría nada mal que esos mensajes fuesen matizados con acuerdo con el grado de certeza científica que se posea. Ese es el ***leit motiv*** del discurso de Skrabanek (1994) en su pormenorizada y dura crítica contra los mensajes rígidos y no claramente justificados que impone la que él llama ***medicina coercitiva contemporánea*** en relación con estilos y hábitos de vida.

En cualquier caso, los investigadores adscritos a la corriente refutacionista, sean del ámbito clínico (McIntyre, 1988) o del epidemiológico (Lanes, 1988), salen parcialmente del escollo que supone la ineludible toma de decisiones diciendo que ésta ha de guiarse por la hipótesis que mejor resista los embates de las pruebas más exigentes.

Lo interesante es que, para la investigación biomédica y epidemiológica, tanto uno como otro enfoque reivindican la intervención de la estadística, a la vez que ninguno de los dos «bandos» rechaza procedimiento estadístico alguno sobre la base de sus respectivos presupuestos teóricos. Por ejemplo, las pruebas de hipótesis (polémicas por otros motivos, véase Capítulo 6) son vistas por unos como un instrumento corroborativo y por otros como un medio para la refutación, La epidemiología inductivista, desde luego, las aprecia desde la perspectiva ortodoxa: a partir de la muestra, se sacan conclusiones sobre la población. Los popperianos las entienden al revés; por ejemplo, Maclure (1985), firme defensor de esta corriente, escribe:

> ***Los epidemiólogos emplean la inducción cuando infieren informalmente que una asociación observada en una muestra pudiera ser válida a nivel poblacional. Debe notarse, sin embargo, que la inferencia estadística formal es deductiva en su forma. Una prueba de la hipótesis nula es una inferencia hipotética de la población a la muestra, ya que el que la aplica, razona así: «Si no hay asociación en la población general, un resultado como el que observamos en la muestra tendría probabilidad P».***

Consiguientemente, si bien el posicionamiento epistemológico de los investigadores respecto de la dicotomía inducción-refutación puede signar el talante de las conclusiones, la estadística como instrumento no parece conmoverse por ello y sirve por igual a unos y a otros.

7.3. Concepto de causa: divergencias para comenzar

La definición del concepto de ***causa*** en el ambiente biomédico, frecuentemente relacionado con la etiología, es motivo de una visión tan poliédrica que no puede sino generar bastante desconcierto.

La obra portentosa de Newton, fuente de explicaciones perfectas -como la que da, por ejemplo, a las órbitas que describen los cuerpos celestes- y de fascinantes

revelaciones sobre el mundo natural, dejó una honda herencia determinística que aún predomina en una parte del pensamiento epidemiológico.

Lilienfeld y Lilienfeld (1980) dan cuenta del punto de vista sostenido por algunos que, inscritos en esa tradición, consideran que para que a un factor pueda llamársele causa de cierto efecto, dicho factor ha de ser *suficiente y necesario.* Resulta curioso el modo como estos dos autores interpretan lo de «necesario y suficiente». Ellos escriben textualmente:

> ***Tal como podemos sospechar intuitivamente, el término*** necesario ***se refiere al hecho de que el factor debe estar presente para que la enfermedad pueda ocurrir; mientras que*** suficiente ***denota que, si el factor está presente, la enfermedad puede ocurrir (pero que su presencia*** no siempre ***resulta en la aparición de aquella)*** [4].

Sin embargo, que la enfermedad ***pueda*** ocurrir en presencia del factor significa simplemente que éste no es incompatible con aquella; y que ***no siempre*** la genere significa justamente que ese factor no es suficiente. La interpretación de la suficiencia que hacen Lilienfeld y Lilienfeld es insólita: hace pensar que ellos creen, por ejemplo, que conservar las amígdalas es una causa suficiente para la leucemia; en tanto que habérselas extirpado... también lo es, ya que ninguno de esos dos rasgos es incompatible con la dolencia, a la vez que ninguno de ellos la genera inexorablemente.

En el marco etiológico, lo que realmente quiere decir que un factor sea necesario y suficiente para producir una enfermedad es que si está presente la enfermedad es porque actuó necesariamente ese factor; y viceversa: basta con que él comparezca para que se produzca aquella. Este doble condicionamiento, por otra parte, hace de la definición un instrumento estéril, ya que, según él, la enfermedad y la causa son la misma cosa.

Stehbeens (1985), refiriéndose a la etiología de las enfermedades, ha llamado la atención sobre el «uso totalmente ambiguo del término «causa», que abarca a factores contribuyentes, modificadores, predisponentes y condicionales», y, aunque a mi juicio este autor no comunica con claridad qué entiende por cada uno de tales factores, sí da cuenta parcial de la diversidad de opiniones que se manejan al respecto.

Menciona así que en la *Enciclopedia de las Ciencias Sociales* Simon (1968) define como *causa* «aquello que ocasiona un resultado, o bien constituye un antecedente uniforme de cierto fenómeno». En mi opinión la primera parte de esta definición es simple y correcta pero improductiva, pues la indefinición se transfiere al concepto de *ocasionar;* la segunda parte, en cambio, es directamente inaceptable. Es difícil no apreciar la primavera como un «antecedente uniforme» del verano, pero más difícil es aceptar que una estación sea causa del advenimiento de la otra.

[4] Los subrayados se hacen en el original por los autores.

Por su parte, Claude Bernard escribía (Bernard, 1961):

> ***Los médicos reconocen una multitud de causas para una misma dolencia; pero no todas las circunstancias enumeradas pueden considerarse causas: a lo sumo son medios o procesos por cuyo conducto la enfermedad puede producirse.***

Bernard aludía de hecho a tres nociones, aunque lo hace de manera más bien confusa: explícitamente se refería al agente causal y a los intermediarios imprescindibles para su actuación, y tácitamente a otros factores que no llegan siquiera a ser estos medios. Procuraremos más adelante esclarecer este asunto.

En mi opinión, cualquiera que sea la definición formal de esta categoría, ***la definición funcional*** -es decir, útil a los efectos prácticos- de la causa de un efecto dado puede expresarse del modo siguiente: ***cualquier factor; condición o característica, cuya supresión elimina la posibilidad de que se produzca el efecto, es una causa del mismo.***

Bollet (1964), en esa misma línea, es categórico al afirmar que «la única prueba de que un fenómeno desempeña un papel parcialmente causal en relación con otro es comprobar que, una vez suprimido éste, desaparece aquel». El uso del adverbio «parcialmente» parece subrayar que este pragmático enfoque atañe a fenómenos que no tienen que ser suficientes.

Abundando en la definición propuesta, cabe comentar que el manido ejemplo del bacilo de Koch como causa de la tuberculosis obviamente se inscribe dentro de sus límites. Pero también se ajustan a ella otros factores cuya condición de causa pudiera parecer algo chocante. Consideremos un caso no relacionado con la patología y que es familiar a todos: ¿qué podemos entender por causa de un embarazo? [5] Tal efecto es producido por un espermatozoide sano. Es obvio que, si se elimina la posibilidad de que este agente arribe al óvulo, el embarazo no puede producirse; y el sostenimiento de relaciones sexuales, según este enfoque, también sería una causa de embarazo, pues su supresión elimina el posible efecto. Sin embargo, la ausencia de métodos anticonceptivos durante la relación sexual no puede considerarse una causa de embarazo, ya que el uso de tales medios no suprime necesariamente la posibilidad de que se produzca.

Sintetizando, esta definición no distingue entre el agente directo y los rasgos posibilitadores para que éste se exprese sensiblemente, factores sin cuyo concurso no puede producirse el efecto que se estudia; tanto uno como los otros son causas del fenómeno.

Es importante subrayar que el conocimiento de algunos de estos factores en un proceso morboso no exige que se sepa la identidad del agente que actúa directamente. De hecho, éste puede ser ignorado e identificarse mucho después que tales

[5] Me refiero, desde luego, a un embarazo natural, no inducido por medios de inseminación artificial.

factores intermediarios, como ilustra entre muchos otros el caso del control epidemiológico de la difteria, muy anterior al dominio de los mecanismos intracelulares que producen la dolencia.

Un viejo chiste habla de un hombre que se sintió mareado cada vez que bebió agua carbonatada; la primera vez la acompañó de ron, la segunda vez tomó agua y vino tinto y en la tercera ocasión combinó el agua con coñac. Es obvio que sólo siguiendo métodos acordes con nuestra definición podrá demostrarse la «inocencia» del agua carbonatada y, en su momento, también la responsabilidad causal del alcohol.

Ahora bien, hay una tercera categoría de interés práctico: los elementos propiciatorios o ***factores de riesgo,*** que son los que generan más confusión porque, como es lógico, suelen venir acompañados de las causas y «se parecen» a ellas, del mismo modo que las medidas paliativas de una dolencia (por ejemplo, analgésicas) se parecen a la curación. Se trata de ***factores asociados al efecto que, sin ser causas propiamente, pueden favorecer que el agente causal actúe.*** En esta categoría caen la práctica de no usar anticonceptivos, circunstancia que favorece el embarazo, el consumo de grasas saturadas a los efectos de las cardiopatías, o la obesidad como factor que favorece el surgimiento de dolencias vasculares.

Hegel decía que «un ladrillo no mata a un hombre porque sea un ladrillo, sino que produce este resultado solamente en virtud de la velocidad que adquirió». El mensaje central está en que ni la práctica preventiva ni la curativa están necesaria y obsesivamente interesadas en descubrir el agente causal directo (el ladrillo) para eliminarlo; de hecho puede bastar con cortar la cadena causal e impedir así su actuación. Si se eliminan todos los ladrillos, ellos no podrán caer y, ciertamente, no se producirán muertes por ese concepto; pero si se evita que caigan, el efecto es el mismo. Por otra parte, la ausencia de casco protector en un trabajador de la construcción al que mata un ladrillo ni es el agente que causa la muerte ni es un condicionante necesario para que se produzca (o sea, no es una causa); sin embargo, si se cumple la norma de usar cascos protectores, se reducirá drásticamente la probabilidad de que ocurra la muerte, de manera que la ausencia de casco es un factor de riesgo [6].

Una dificultad operativa radica en que no siempre puede distinguirse entre un verdadero factor propiciatorio o de riesgo (aquel cuya presencia modifica la probabilidad de ocurrencia del problema) y uno que aparece parasitariamente asociado al efecto sin llegar, no ya a ser una causa, sino siquiera a ser un factor de riesgo. Un ejemplo sumamente elocuente en que un simple factor concomitante -un mero acompañante frecuente del efecto- es apreciado como una posible causa o, al

[6] Todo el texto se ha escrito con referencia a los factores de riesgo. En muchos casos, la probabilidad de ocurrencia de la enfermedad *disminuye* como resultado de que el factor esté presente; de ser así, se dice que éste es un *factor de protección.* Tal es el caso de la presencia del casco o del uso de métodos anticonceptivos. Como en toda situación polar, la inversión del punto de vista que se use, demanda del correspondiente ajuste léxico.

menos, como un factor de riesgo para contraer el SIDA, es el que presentan Vanderbroucke y Parodel (1989), desarrollado con cierto detalle en la Sección 8.8.3.

Puede decirse, en síntesis, que hay tres categorías afines a las relaciones causales que revisten interés teórico y práctico: el agente causal directo, los intermediarios imprescindibles para que éste actúe y los rasgos que favorecen la acción del agente causal (factores de riesgo).

Ortega y Gasset (1958) escribió:

> ***La razón pura tiene que ser sustituida por una razón vital, donde aquella se localice y adquiera movilidad y fuerza de transformación.***

En el manejo con fines preventivos de las causas y los factores que posibilitan o favorecen su acción es imprescindible asumir esa flexibilidad. Quiere esto decir que, aunque los factores de riesgo no son causas, pueden ser considerados virtualmente como tales con el propósito de prevenir (o propiciar) determinado efecto.

7.4. Pautas de causalidad

Las bases teóricas para el examen de causalidad en medicina y epidemiología responden a un paradigma que, aunque según Morabia (1991) tiene sus raíces en Hume, es relativamente reciente y fue expresado de modo acabado por Bradford Hill en relación con la etiología de origen medioambiental en un célebre y, por cierto, muy disfrutable [7] artículo (Hill, 1965). No es ocioso recordar y comentar críticamente los nueve ***rasgos*** fundamentales que este autor enumera como elementos que incrementan el valor de una asociación ***en tanto indicio de causalidad.*** Se parte de que se ha observado una asociación claramente definida entre dos fenómenos, y que se ha descartado la posibilidad de que ella sea exclusivamente debida al azar. Los rasgos en cuestión son los siguientes:

1. Fuerza de la asociación

La existencia de una asociación (que puede medirse a través de estadígrafos tan diversos como, por ejemplo, un riesgo relativo, un ***odds ratio,*** un coeficiente de correlación o una diferencia de medias aritméticas) es condición ***sine qua non*** de la causalidad. Ahora bien, cuanto más intensa sea, más verosimilitud adquiere la hipótesis subyacente. Ello se debe a que una de las incertidumbres estriba en que la correlación observada pudiera ser sólo un reflejo del efecto que ejercen otros condicionantes; si la correlación es intensa, la posibilidad de que

[7] Hasta que no lo leí de primera mano, siempre creí que sería un texto frío y estirado, y no el material fresco y de espíritu zumbón que realmente es.

pueda explicarse enteramente por el efecto de tales factores (los llamados ***factores confusores) es*** mucho menor. Suele considerarse que un riesgo relativo mayor que 2 es suficientemente grande como para considerarlo «fuerte» (Beaglehole, Bonita y Kjellström, 1994). Importantes figuras de la epidemiología contemporánea demandan valores mayores para dignarse a prestar atención a un odds ratio; Marcia Angell, por ejemplo, editora de ***New England Journal of Medicine*** reclama que éste ascienda por lo menos a 3. Huelga extenderse, sin embargo, sobre los peligros que entrañan tales ***thumb rules;*** en la Sección 8.8.3 se describe una «autopsia» cuyos resultados ilustran crudamente cuán endebles pueden ser los diagnósticos basados en esas «fortalezas».

2. **Consistencia**

Cuando la relación de causalidad apunta a la identificación de una ley natural, ella no puede presentar inconsistencias empíricas: debe verificarse, en principio, en cualquier entorno socioeconómico y cultural. El VIH es el agente causal directo del SIDA en San Francisco y en Burundi, del mismo modo que en ambas latitudes actúa por igual la ley de conservación de la energía. El criterio que nos ocupa demanda la máxima consistencia posible. Esta pauta, sin embargo, tiene un valor relativo cuando no se trata de principios de las ciencias básicas. Por ejemplo, al estudiar las causas de aumento de los suicidios en una ciudad de Uruguay, la exigencia de que éstas sean consistentes con las causas identificadas en Sarajevo para el mismo problema sería un evidente despropósito.

3. **Especificidad**

Según esta regla, es más razonable esperar que el factor tenga carácter causal si está asociado de manera ***específica*** con el fenómeno que se estudia; que si influye causalmente, además, en otros fenómenos. Este requisito es uniformemente olvidado por una parte importante de la investigación analítica contemporánea. Así, no es extraño ver cómo algunos investigadores «descubren» que la nutrición inadecuada es una de las causas de tal o cual dolencia crónica. ¿Es que existe alguna que no se vea favorecida por esta carencia? [8]

Por otra parte, el hecho de que este rasgo se mencione una y otra vez, pero que casi nadie le haga el menor caso, tiene una explicación. Según él, es más probable que X sea una causa de Y si no se asocia casualmente con ningún otro factor Z. La polución ambiental podría ser una causa del aumento de la incidencia de crisis asmáticas, pero, puesto que es un factor causal de la extinción de las cigüeñas, la convicción de que tenga el efecto anterior se vería disminuida. El

[8] Esto no quiere decir, desde luego, que los problemas nutricionales no puedan ser la causa fundamental de determinadas ¿desviaciones de la salud. En la Sección 5.4 se expone con cierto detalle el caso en que un déficit de vitaminas y otros nutrientes causa una epidemia de neuritis.

planteamiento es suficientemente inconsistente y relativo como para que Hill desdeñe su importancia en el propio trabajo que nos ocupa.

4. **Secuencia temporal correcta**

Toda hipótesis de causalidad involucra un efecto y una presunta causa; lo que se subraya es que ésta ha de ser previa a aquel. A pesar de su aparente obviedad, la violentación de este principio -que en este caso es una condición verdaderamente imprescindible y no solamente un rasgo que aumente nuestra convicción de que la hipótesis sea cierta- es tan frecuente que se le destina íntegramente una sección del presente capítulo (Sección 7.7).

5. **Existencia de un gradiente biológico**

La observación de una sostenida relación dosis-respuesta (a mayor dosis del factor, mayor el efecto registrado) aumentará el grado de confianza que se deposite en la validez del silogismo causal. Beaglehole, Bonita y Kjellström (1994), citando información de la OMS, llaman la atención sobre el impresionante gradiente que vincula la prevalencia de sordera tanto con el nivel de ruido soportado como con los lapsos de exposición a él, tal y como se aprecia en la Tabla 7.1.

Tabla 7.1. ***Tasas de prevalencia (porcentaje) de sordera según nivel medio de ruido soportado y años de exposición***

Nivel medio de ruido durante una jornada laboral de 8 horas (decibelios)	Período de exposición (años)		
	5	10	40
<80	0	0	0
85	1	3	10
90	4	10	21
95	7	17	29
100	12	29	41
105	18	42	54
110	26	55	62
115	36	71	64

6. **Plausibilidad biológica**

Es conveniente que la hipótesis sea verosímil. No tiene mayor sentido perder el tiempo profundizando en el estudio de una hipótesis sin respaldo teórico como las que, por ejemplo, suelen proponer algunos farsantes (en algunos casos, se trata de sujetos «iluminados» que actúan de buena fe) como los que toda ins-

titución académica del mundo ha tenido que «atender». La esperanza de que la plausibilidad biológica no sea exigida es la levadura de muchas expresiones de la pseudociencia.

7. **Coherencia con lo que ya se conoce**

Según este precepto, las hipótesis de causalidad no han de contradecir hechos ya constatados científicamente. Se trata de una regla muy discutible. La investigación es por naturaleza un acto de rebeldía intelectual contra lo que se cree hasta ese momento; un apego irrestricto a dicha regla cancelaría una porción sustancial de la investigación, quizás la más interesante. En última instancia, su aplicación se remite a lo que se entienda por «hechos científicamente constatados». En realidad, todo resultado científico es provisional, y los nuevos paradigmas que mencionara Kuhn se forjan sobre la base de la irreverencia ante la obra, generalmente encomiable, de los antepasados. En este sentido, la posición de Popper es mucho más razonable y transparente: no hay ningún compromiso con las convicciones anteriores; cualquier hipótesis será tratada con idéntica saña.

8. Indicios experimentales

El poderoso componente experimental del método científico aparece en el mundo en la segunda mitad del siglo XVI con Galileo, y en menor grado, con su contemporáneo Kepler. Dicho en su esencia y de manera informal, se trata de una indagación en la que el investigador ***modifica*** la realidad según un plan y registra lo que ocurre como resultado de la maniobra para evaluar el efecto de la intervención.

Desde su aparición el experimento ha constituido un recurso en permanente evolución, con hitos muy importantes en figuras como el químico francés Antoine Lovoisier (1743-1794), cuya obra experimental revolucionó la química del mismo modo que había ocurrido con Galileo y la física un siglo atrás. Aquellas disciplinas que lo han usado con más rigor son sin duda las que han conseguido resultados más concluyentes. Puesto que la investigación epidemiológica sufre serias barreras éticas y prácticas para incorporar la experimentación en su desempeño regular, esta premisa y los recursos para mitigar los efectos de su incumplimiento revisten la máxima importancia y actualidad [9]. El debate en torno a la utilidad de los estudios no experimentales para demostrar relaciones de causalidad ha cobrado especial intensidad en los últimos años.

Aunque nadie llega a afirmar categóricamente que los estudios no experimentales puedan ser suficientes para probar causalidad, es obvio que a muchos les cuesta trabajo desprenderse de esta ilusión. Quizás por ello no faltan voces

[9] Cabe recordar que estas barreras alcanzan a muchas disciplinas, como la meteorología, la demografía o la investigación histórica, por mencionar sólo tres áreas potencialmente vinculadas a la salud.

muy vehementes (véase, por ejemplo, McCormick, 1988) desvalorizando la potencialidad probatoria de los estudios observacionales. Sus puntos de vista se reafirman a través del señalamiento de las debilidades que aquejan a muchos de los trabajos desarrollados para desentrañar la etiología de las dolencias que más preocupan a los epidemiólogos y salubristas de hoy. Un análisis técnico estadístico en que se fundamenta esta insuficiencia puede hallarse en Greenland (1990).

El pensamiento epidemiológico predominante defiende el punto de vista de que los esfuerzos de investigación que no hacen uso de la experimentación son, a pesar de ello, de extrema utilidad, básicamente porque contribuyen a fundamentar empíricamente las hipótesis. Personalmente, comparto esta apreciación, pero se advierte una desenfrenada producción de trabajos observacionales -estudios de casos y controles en particular- destinados a la identificación de factores de riesgo. Y, además de que tales esfuerzos nunca darán respuestas definitivas, da la impresión de que esa práctica pudiera estar generando más confusión que esclarecimiento. Dando palos de ciego -es decir, sin hipótesis de causalidad bien fundamentadas y con oportuna corroboración experimental- sólo se logra en el mejor de los casos desarrollar un acto de especulación sobre la magnitud de asociaciones aisladas, carente de un marco teórico integrador, e incapaz de producir, por tanto, avances claros.

Algunas revistas de epidemiología se llenan últimamente de trabajos de este tipo. Considérense los siguientes tres artículos tomados del mismo volumen de ***American Journal of Epidemiology,*** que abordan la relación entre: defectos congénitos y uso de mantas eléctricas (Duglosz ***et al.,*** 1992); ser zurdo y la necesidad de resucitación entre recién nacidos (Williams, Kimberly y Eskenazi, 1982); y uso de aire acondicionado y mortalidad en verano (Rogot, Sorlie y Backlund, 1992). Henri Poincaré, en su momento, advirtió:

> ***La ciencia se construye con hechos, del mismo modo que un edificio se hace con ladrillos; pero hacer pasar una colección de hechos como ciencia es como atribuir a un montón de ladrillos la condición de edificio.***

A mi juicio es obvio que la experimentación constituye un paso de importancia cardinal y demarcatorio entre los estudios convincentes y los que no lo son en materia de causalidad. No en balde, uno de los recursos típicos de la pseudociencia es eludirla (véase el Capítulo 13).

9. **Analogía**

Si otros factores siguen caminos causales análogos a los de la hipótesis que ahora se valora, la convicción de que ella sea válida se incrementa. Éste es un viejo planteamiento que se manejó intuitivamente desde siempre. Tal y como lo formula el propio Hill, «puede ser útil en algunas circunstancias», pero dista de constituir una demanda, a diferencia de lo que ocurre con la experimentación.

En resumen, los nueve preceptos de Hill pueden constituir pautas de utilidad, pero no han de considerarse en su totalidad mandamientos ineludibles de valor universal, ni pueden concebirse -obviamente- como garantías para la causalidad.

7.5. Investigación descriptiva, estudios ecológicos y causalidad

La presente sección se destina a esclarecer dos aspectos vinculados entre sí, aunque de naturaleza diferente: los estudios descriptivos y los estudios ecológicos en particular. Tienen en común la «mala fama» de no servir como recurso para el estudio de la causalidad, e incluso hasta de no servir para nada. Nuestro propósito es destacar no solo la legitimidad que ambos comparten sino la necesidad imperiosa de revitalizar su presencia en la investigación actual.

7.5.1. La primera tarea del epidemiólogo

Con frecuencia, los alumnos de posgrado me preguntan si una indagación descriptiva puede considerarse realmente una investigación o si, para merecer tal nombre, ha de trascender la descripción de una realidad y proceder a explicarla o a descubrir por qué es como es. La pregunta equivale, palabras más o menos, a si un trabajo desarrollado con el fin de obtener nuevos conocimientos sólo constituye una investigación cuando procura desentrañar causas de cierto fenómeno.

Los esfuerzos orientados en esta última dirección corresponden a la llamada *investigación analítica,* ya sea experimental u observacional (en este caso, destacadamente, los estudios de cohorte y los de casos y controles). Éstas son las que, supuestamente, tienen la capacidad de ayudar a identificar causas. Ya vimos que, en rigor, tal rasgo solamente lo poseen las investigaciones experimentales, aunque la comprensión del proceso de investigación como un todo complejo e integrado, en movimiento, me hace preferir no parcelar las cosas de ese modo. En un sentido amplio, la investigación siempre tiene finalidad explicativa. De modo que la descripción no sólo es una forma legítima de investigación biomédica sino que constituye un pilar básico de todas las expresiones restantes. Tanto es así que, como señala Greenland (1990) para el caso de la epidemiología, «la primera tarea del epidemiólogo es descriptiva».

Según esta concepción, no podrían desarrollarse investigaciones analíticas realmente fecundas si no se contara con el conocimiento del que se nutren las hipótesis que ellas evalúan; usualmente tal aval empírico, o bien procede de la descripción, o bien se consolida con ella.

La investigación descriptiva, en suma, además de cumplir una función valorativa de máxima trascendencia, satisface otra importante función, que es antecedente natural de cualquier intento por aproximarse al esclarecimiento causal: la generación o consolidación de hipótesis. Desde luego, esta generación nunca podrá ser

tarea de una computadora porque para ello es imprescindible el concurso de la sensibilidad y de la creatividad humanas; Hempel(1973) decía:

> ***No hay reglas de aplicabilidad general para inferir hipótesis o teorías a partir de las observaciones. Esta transición demanda imaginación; las hipótesis y teorías no se derivan de los hechos observados sino que se inventan para explicarlos.***

Por lo demás esta tarea muy raramente será resuelta por la casualidad porque, como indicaba Pasteur, en materia de observación, el azar sólo favorece a los espíritus preparados.

7.5.2. Reivindicación de los estudios ecológicos

Otro aspecto lastrado por ese tipo de malentendidos que una vez establecidos se arraigan y perduran sostenidamente se relaciona con los llamados ***estudios ecológicos.*** Se conoce como tales a aquellos estudios (observacionales casi siempre) en que las mediciones, tanto de factores condicionantes como de daños, se verifican a nivel de grupos o agregados poblacionales y no al de los sujetos que portan dichos factores o sufren los daños.

Con el tiempo, los estudios ecológicos han ido perdiendo su peso específico, y su presencia ha disminuido marcadamente en la investigación epidemiológica contemporánea. En parte debido al temor que despierta la «falacia ecológica» (véase Sección 8.4) y en parte debido a otros prejuicios (Schwartz, 1994), éstos han llegado a conceptualizarse sólo como sucedáneos (como un mal menor) de los estudios en que las unidades de análisis son los individuos, a pesar de que, como demuestra **Susser (1994)**, constituyen una herramienta de la salud pública con su propia legitimidad y de haber producido conocimiento que permanece enteramente vigente después de muchos años.

El más arraigado y pernicioso de tales prejuicios es la convicción de que las variables medidas a nivel de grupos no representan agentes causales de enfermedad. Quizás la más emblemática declaración que muestra hasta donde ha llegado el reduccionismo de la epidemiología al nivel individual de análisis sea la que hizo Rothman (1986) cuando escribió que «la clase social no se relaciona causalmente con ninguna o casi ninguna enfermedad». Para dar sólo un elemento persuasivo en dirección contraria, basta detenerse en los trabajos de Hertzman (1986) y Davey et al (1990) quienes muestran no sólo el marcado gradiente en esperanza de vida y otros indicadores de salud según grupos socioeconómicos (operacionalizados mediante una combinación de indicadores que involucran educación, ingresos y situación laboral) sino su notable persistencia a lo largo de decenas de años.

Se ha perdido de vista incluso que cuando se mide un rasgo individual muchas veces se está midiendo algo distinto que cuando se trata de ese mismo rasgo pero mirado a nivel colectivo.

Por ejemplo, el concepto de pobreza referido a un sujeto o a una familia no es el mismo que el que corresponde a la pobreza de la comunidad en la que vive el sujeto o la familia, la cual inexorablemente afecta a todos los que residan allí, cualquiera sea su nivel socioeconómico. La categoría *ocupación,* pongamos por caso, tiene una dimensión ecológica que los estudios al nivel del individuo suelen pasar por alto. Un maestro en un ambiente urbano marginal no está sometido a las mismas restricciones o tensiones que uno que se desempeña en un entorno urbano saludable o que uno que lo hace en un medio rural empobrecido, y a la vez, esas restricciones o tensiones pueden ser muy similares a las de una secretaria o un taxista que comparten un mismo espacio. Hasta la mismísima variable que mide si un sujeto es o no desocupado habría que considerarla con cautela, ya que «tener trabajo» no es una condición uniforme. Galbraigth (1991) señalaba que «No hay mayor espejismo en la actualidad, mayor fraude incluso, que el uso del mismo término *trabajo* para designar lo que para algunos es monótono, doloroso y socialmente degradante y para otros placentero, socialmente prestigioso y económicamente provechoso».

Si en un estudio, por ejemplo, se demostrara que el nacimiento de niños con bajo peso al nacer es más frecuente en las comunidades con altos niveles de desempleo que en las que no tienen ese problema, entonces la afirmación no puede quizás trasladarse al nivel de los sujetos. Pero tal vez no haya ningún interés en hacer ese traslado, quizás el interés esté directamente centrado en el efecto de ese indicador sobre todos los residentes de cada comunidad, trabajen o no.

De modo que los rasgos generales pueden tener impactos globales sobre todos los sujetos abarcados por ellos, pero no sólo en el sentido en el que lo hace, por ejemplo, la contaminación ambiental (que no distingue entre unos y otros individuos al margen de las decisiones que adopten), sino que influye en las conductas individuales; por ejemplo, como hacen notar Evans y Stoddart (1990), fumar o no es una *acción* individual, pero es posible que sea algo tan fuertemente condicionado por circunstancias sociales que no constituya una *elección* individual. Del mismo modo, las muertes por accidentes automovilísticos en un sitio dado pudieran estar tan condicionadas por la apreciación social que allí se haga acerca del modo de conducir como por el hecho de que se emplee o no cinturón de seguridad; análogamente, un sistema de valores moralmente opresivo o altamente competitivo puede tener enorme relevancia en materia de suicidios o del ejercicio de la violencia, tal y como pueden tenerla los problemas individuales de personalidad de los agresores o los sujetos con tendencias suicidas.

Las «variables ecológicas» son vitales para el examen de los efectos estructurales sobre la enfermedad, cualquiera sea su naturaleza. Por otra parte, puesto que la búsqueda de factores etiológicos ubicados al nivel del individuo ha resultado ser tan poco fructuosa (véase Sección 7.8) y dado que la finalidad última de todo el esfuerzo investigativo es mejorar el estado de salud de la población, podría ser muy conveniente que se retornara la práctica de examinar los problemas en su dimensión socio-epidemiológica, ya que no quedan dudas de que existen factores colecti-

vos cuyo valor etiológico puede ser crucial, y mucho menos, de que ellos podrían modificarse con tanto o más éxito que el alcanzado por los programas que se proponen modificar conductas individuales. Y lo que es más importante: tal vez con más dividendos que éstos, pues, como apunta Syme (1989) «Muy a menudo pueden conseguirse efectos muchísimo mayores alterando el ambiente que intentando alterar los comportamientos individuales».

En síntesis, hay que distinguir los factores que pudieran explicar la presencia de la enfermedad *en los individuos* de los que explicarían la distribución *entre la población, y* deben rescatarse estos últimos como fuente explicativa de los procesos que deterioran la salud.

El cúmulo de indicios sobre los efectos estresantes de ciertos ambientes laborales sobre la salud -niveles de catecolamina, enfermedades coronarias, alcoholismo, neurosis, etc.- es ciertamente copioso y convincente. Y, a la misma vez, la mayor parte de los factores estresantes que hoy pudieran estar influyendo en muchas de las dolencias no transmisibles no son de naturaleza individual sino de índole sociocomunitaria.

Es usual que el ciudadano -si es que no se halla bajo la degradante circunstancia de ser un desocupado- reparta la mayor parte de sus días de la siguiente manera: durante un tercio de su tiempo subsiste en un recinto que le procura la tarea monótona, dolorosa y socialmente degradante de que habla Galbraight. Ahí se desempeña muchas veces en un clima de competencia enervante. Además invierte varias horas en viajar por redes viales atestadas o en transportes más o menos hostiles. En el seno familiar le espera la banalización que encierra la televisión y una prensa de un nivel generalmente bochornoso. Todo esto, en el contexto de un patrón social de consumo que invoca al desarrollo externo en detrimento del desarrollo interno. Es una tarea pendiente de la epidemiología evaluar la responsabilidad de esta «realidad ecológica» en la evasión representada por el alcohol, las drogas y el consumo de tranquilizantes y, más indirectamente, en las diversas enfermedades que aquejan a la sociedad.

Consideraciones similares pueden hacerse sobre los efectos de «variables ecológicas» tales como polución ambiental, regulaciones jurídicas, formas de organización laboral, valores religiosos, grado de desigualdad, etc., pero cabe detenerse en el ámbito de la salud mental (asunto abordado extensamente en Silva, 1993) por ser uno de los más expresivos del reduccionismo a que nos ha constreñido la teoría de factores de riesgo centrada en los riesgos relativos, que típicamente desdeña los condicionamientos sociales.

El sujeto que, más allá de cierto límite, defiende su identidad, su independencia intelectual frente a las exigencias normativas de su sociedad, de un modo u otro entra en colisión con ella. Son bien conocidos los casos de sujetos que son tratados como enfermos mentales sólo porque se escapan de una norma laboral, cultural o política. Esta contradicción dialéctica individuo-sociedad se asienta en que, como agudamente apunta Materazzi (1991) «Lo que cuenta no es lo que uno transmite, muestra o representa, sino lo que la comunidad asimila, interpreta o recrea».

Ha de tenerse en cuenta que las tensiones que derivan en trastornos psíquicos no configuran un problema marginal; no se trata de un problema puntual reducido a unos cuantos: se estima que la tasa de sujetos con depresión o ansiedad en la población adulta varía entre 12 y 15%. Algunos estudios mencionan cifras más elevadas; por ejemplo 20% en Holanda (Giel, 1992) y 22% en Italia (Sicliani, 1992).

La teoría de factores individuales de riesgo ha alcanzado un desarrollo impetuoso; los artículos que concentran su atención en ellos pueblan las revistas médicas como hongos después de la lluvia. Muchos de tales artículos, relacionados con los problemas epidemiológicos, se basan en la identificación y exaltación de formas de vida orientadas a superar el origen de ciertos males y precaven contra algunas prácticas o condiciones (colesterol, tabaquismo, obesidad, sedentarismo, etc.), que pertenecen a la esfera de la conducta; pero resulta por lo menos desproporcionado cargar las tintas en ellos a la vez que se mantienen, por ejemplo, ambientes insalubres y espacios contaminados, restricciones objetivas para realizar ejercicios, presiones publicitarias en materia alimenticia y muchos otros condicionamientos sociales dañinos para el desarrollo equilibrado del ciudadano.

Heredada de la reducción impuesta por la supuesta supremacía metodológica de los estudios sobre riesgos individuales, existe una tendencia subyacente a considerar que los problemas de salud mental tienen su etiología radicada fundamentalmente en el cerebro o la mente, o en la personalidad del sujeto que lo padece.

Me parece obvio que cuando un sujeto aislado tiene un comportamiento socialmente incivilizado y, por ejemplo, agrede brutal e inmisirecordemente a un conciudadano, probablemente se esté ante una grave desviación mental radicada al nivel del sujeto; si tal conducta se produce a manos de una horda, como ocurre hoy con los «cabezas rapadas» y grupos similares, se está ante una desviación de etiología social y no ubicada en la psique de cada uno de sus integrantes.

Caplan (1966) ha elaborado una teoría que permite ilustrar adecuadamente la estrechez en boga a que nos estamos refiriendo. Dicha teoría se estructura a partir de las crisis que afectan a los sujetos y en los por él denominados «aportes» físicos, socioculturales y psicosociales, y constituye un enfoque atractivo para encarar los problemas que pueden aquejar al individuo. Su tesis central establece que ***para no sufrir un trastorno mental, una persona necesita continuos aportes, adecuados a las diversas etapas de crecimiento y desarrollo,***

Este enfoque ignora los más dañinos factores de riesgo. Lo óptimo no es idear la incorporación de aportes que compensen al sujeto; ello pudiera ser una buena alternativa ante imponderables inevitables como la muerte de un ser querido; pero bajo ningún concepto cuando se trata de problemas cuyos más importantes factores etiológicos pueden ser eliminados.

Por ejemplo, puede ser útil extender el repertorio de habilidades de un ciudadano o ciudadana homosexual para enfrentar las agresiones sociales de que puede ser objeto, pero siempre serán meros recursos paliativos si el homosexual está socialmente mal conceptuado, inhabilitado para pertenecer a una organización polí-

tica, discriminado a los efectos laborales, ofendido por chistes difundidos en los medios masivos de comunicación, etc. El hecho de que las tasas de prevalencia de neurosis sean mayores entre los homosexuales que entre los que no los son denuncia con mucha mayor fuerza la intolerancia arbitraria de la sociedad que su incapacidad para proveerlos de aportes compensatorios.

A pesar de todo lo anterior, en resumen, el papel del ambiente ha quedado en un segundo plano (o tercero, o simplemente nulo) en la investigación epidemiológica contemporánea; tal fenómeno se produce en virtud de una marcada sumisión acrítica o inercia1 al paradigma que establece que las causas han de buscarse al mismo nivel en que se produce el problema. Los estudios ecológicos deben recuperar su espacio original y complementar los restantes recursos de que disponemos, así como asumir el debido protagonismo cuando la naturaleza de la investigación lo aconseje.

7.6. Enfoque clínico y epidemiológico

Un atolladero típico para el investigador principiante se deriva de la contaminación que el peso inercia1 del pensamiento clínico ejerce sobre su proceder epidemiológico. Una expresión nada infrecuente que ilustra claramente este fenómeno es que el cuestionario de una encuesta, diseñada para la identificación de factores etiológicos, resulte atiborrado de detalles propios de una historia clínica.

Supongamos que se desarrolla una investigación con el fin de identificar factores de riesgo de la parálisis cerebral infantil. En un tramo del instrumento creado para la recogida de datos primarios puede aparecer una secuencia de preguntas a la madre como la siguiente [10]:

> ¿TUVO USTED UN ACCIDENTE DURANTE EL EMBARAZO? SI LA RESPUESTA ES SI, SEÑALE SI FUE DE NATURALEZA TRAUMÁTICA O DE OTRA íNDOLE. SI FUE UN TRAUMA, ¿ACUDIO AL MÉDICO?, ETC.

Cuando a este investigador se le pregunta por qué genera tal laberinto informativo, responde atónito que, ¿cómo, si no, va a poder detectar el origen de la parálisis? Ésta sería la reacción legítima del clínico, que está pensando en un paciente, con una historia personal singular e irrepetible. Pero ahora su problema no es identificar las causas de la dolencia para ese paciente concreto; el problema que encara concierne a la biología humana en genérico, no a la de ningún humano específico. El tiene que prepararse para el manejo estadístico (seguramente con apoyo informático) de la montaña de cuestionarios rellenados que le espera.

[10] Naturalmente, en un cuestionario real las preguntas seguirán cierta disposición formal que ahora no interesa.

Sólo un pensamiento globalizador y una disposición a sacrificar detalles sinuosos permitirán apreciar la información desde un ángulo que consienta, con buena suerte, identificar las leyes generales que laten (y que a veces se insinúan) detrás de los datos, o evaluar las hipótesis que se han planteado.

Esta antinomia, que suele observarse en el terreno operativo, es reflejo de una confusión más general y atañe directamente a la manera de entender las relaciones de causalidad. Cuando se examina a un paciente concreto, en búsqueda de una explicación para que haya arribado al estado en que se encuentra, se procura identificar qué factores específicos concurrieron en *su* caso para producir tal efecto. Pero el pensamiento que procura desentrañar leyes naturales se orienta en el sentido opuesto: desdeña el derrotero anecdótico de cada caso para que emerjan los patrones generales según los cuales funciona el proceso estudiado.

7.7. Avanzando hacia atrás

Desde muy temprano un niño identifica que el sonido del timbre hogareño y la presencia de un visitante son acontecimientos asociados; y no demorará en comprender que no es el sonido el que produce la visita sino viceversa, ya que una y otra vez irá corroborando que el orden de los acontecimientos es invariablemente el mismo: primero llega el visitante, luego suena el timbre. Es decir, todos sabemos que *la* ***presunta causa de cierto efecto, necesariamente, ha de precederlo en el tiempo.***

Como he subrayado con énfasis especial en otro sitio (Silva 1995), ningún análisis de causalidad tiene un sentido claro cuando el diseño del estudio no ha tenido en cuenta una regla tan básica como la enunciada.

Tal inadvertencia, asombrosamente frecuente, es una trampa abierta, especialmente insidiosa en los estudios transversales y retrospectivos, en los que se indaga sobre hechos ocurridos con anterioridad al momento del estudio. La clave del problema radica en que es ***imposible*** en estos casos establecer ***mediante observación*** cuál fue el orden en que ocurrieron los hechos que se registran.

Cabe intercalar, sin embargo, que el carácter prospectivo de un estudio no garantiza que no se presente en cierto sentido el problema. Supongamos, por ejemplo, que se estudian 2.000 individuos sanos aunque poseedores de factores sindicados como de riesgo para el infarto de miocardio (personalidad tipo A, estrés, edad superior a 50 años, etc.), pero que se distinguen internamente por el hecho de que 1000 de ellos son sedentarios mientras que los otros 1000 practican ejercicios regularmente. Al año de observación se registra que entre los sedentarios se produjeron 12 infartos mientras que hubo sólo 3 entre los activos. El riesgo relativo asociado al sedentarismo asciende a 4,0, pero no es descartable que tal resultado se deba a que parte de los sedentarios hubieran adoptado tal forma de vida precisamente porque ya venían teniendo síntomas precursores del infarto, aunque éste no se hubiera producido aún.

Volviendo a los estudios no prospectivos, frecuentemente, los investigadores estudian a un individuo que ***en el momento de la encuesta*** padece cierta dolencia -digamos, por ejemplo, hipertensión- y registran datos de su pasado, tales como si su padre era o no hipertenso, o si el individuo ha consumido o no alcohol con intensidad en los últimos años. En este ejemplo, la hipertensión del padre, de haber existido, es un hecho muy probablemente ***anterior*** a la situación que presente este individuo ahora; quizás fue incluso diagnosticada antes del nacimiento del hijo. Pero, si se quiere evaluar el posible efecto causal del alcohol en el desarrollo de la dolencia, lo que realmente interesa registrar *no* es si el sujeto consumió esta sustancia, sino si la consumía o no ***antes*** de que apareciera la enfermedad.

La inadvertencia de este «detalle» desmantela la estructura lógica del estudio y arruina cualquier interpretación de sus resultados. El hecho puede ser devastador, aunque muchos investigadores no reparen en él, o lo consideren, ingenuamente, como un mal menor.

El problema es típico de las situaciones en que aparecen involucradas las enfermedades crónicas: el conocimiento de lo que ocurrió antes de su comienzo puede ser muy difícil (o imposible), simplemente debido a la dificultad (o imposibilidad) identificar el momento en que comenzó el trastorno. Pero no se expresa solamente en este caso; se presenta muy frecuentemente en situaciones socioepidemiológicas [11] en las que los fenómenos se «retroalimentan» mutuamente como causa y efecto.

Por ejemplo, el planteamiento de una pregunta tal como si hay asociación entre la salud de una familia [12] y su situación económica tiene un sentido borroso. Obviamente, las dificultades económicas de una familia dada pueden contribuir al deterioro de su salud (por ejemplo, puede producir limitaciones para costear servicios médicos o medicamentos, o producir, directamente, su desestructuración como tal); pero también ocurre que los problemas de salud pueden ser responsables parciales de las dificultades económicas (por ejemplo, ocasionan pérdidas de salario y gastos adicionales). Cada factor puede ser causa contribuyente del otro, de modo que una pregunta neutra sobre la existencia de asociación no conduce a ninguna parte. El acto de investigación puede servir para examinar, por un lado, el efecto del primero sobre el segundo, y por otro, el del segundo sobre el primero; pero ello exige un diseño que contemple la observación de los hechos de manera que quede debidamente registrada el orden temporal en que ocurren.

En una ocasión me hallaba en cierta celebración que generaba una gran aglomeración de público en las cercanías de las puertas de acceso, aún cerradas, del

[11] No en los ensayos clínicos que, como todos los experimentos, son prospectivos y que gracias a la asignación aleatoria no pueden padecer del defecto ilustrado a través del ejemplo del infarto que se bosquejó arriba.

[12] La *salud familiar* es un concepto relativamente novedoso que concierne a esta entidad como célula económica y social de la comunidad. No es necesario, sin embargo, definirlo ahora a los efectos del ejemplo.

recinto en que habría de producirse el acto. Entonces oí a uno de los organizadores, megáfono en mano, dirigirse a la multitud que pugnaba por ganar los primeros puestos; su exhortación consistía en la paradójica invocación siguiente: «Por favor, vamos a avanzar hacia atrás». En los estudios que no son prospectivos (en estos se avanza, naturalmente, hacia los desenlaces), tampoco se puede «avanzar hacia atrás»; sólo se puede «mirar hacia atrás».

Resulta interesante constatar que en una gran cantidad de ejemplos de conclusiones sesgadas, el problema se reduce, en última instancia, a una violación de este elemental precepto. Sugiero al lector que relea «la fábula estadística» que se incluyó en la Sección 2.3 y la aprecie ahora desde esta perspectiva. Por la entidad del asunto y por la extrema frecuencia con que se produce, parece conveniente profundizar en él. Susser (1973) y Gray-Donald y Kramer (1988) aportan algunas reflexiones complementarias sobre el tema, pero, para concluir ahora, se exponen cuatro ejemplos de diferente naturaleza en los que se pagan las consecuencias de «avanzar hacia atrás».

7.7.1. Tabaquismo: una práctica preventiva para el enfisema pulmonar

Imaginemos que, al comenzar el año 1995, se estudian dos grupos: un conjunto de 200 enfermos afectados por enfisema pulmonar durante el año precedente (es decir, a lo largo de 1994), y un conjunto de igual número de «controles», sujetos que no han padecido tal enfermedad en el período. Supongamos que a unos y a otros se les pregunta si fuman o no.

Al examinar los resultados, bien puede ocurrir que se halle una correlación inversa a la esperada entre el padecimiento de la dolencia y el tabaquismo; en este caso, que la ***ausencia de enfisema*** se asocie ***positivamente*** con el ***hábito de fumar.*** Los datos ficticios que se resumen en la Tabla 7.2 constituyen un resultado perfectamente verosímil.

Tabla 7.2. ***Distribución de sujetos con y sin enfisema pulmonar según sean o no fumadores***

	Casos (enfermos)	Controles (sanos)
Fumadores	20	90
No fumadores	180	110
Total	200	200

Es decir, el 45% de los controles son fumadores mientras que tal hábito prevalece sólo para el 10% de los casos. La estimación de la *odds ratio* (OR) de enfermar

dado que se es fumador se estima mediante la ***razón de productos cruzados,*** que arroja un valor muy por debajo de 1:

$$\frac{(20)\ (110)}{(90)\ \ (180)} = 0{,}14$$

El intervalo de confianza al 95% para la ***OR*** es [0.80 - 0.24]; puesto que su extremo superior está ubicado muy a la izquierda de la unidad, el enfoque de las pruebas de significación permite descartar el azar como explicación del aparente efecto protector que exhibe el tabaco. Como es bien conocido, siendo la tasa de incidencia de enfisema muy reducida, la ***OR*** es esencialmente igual al riesgo relativo, de manera que este número se traduce en que la probabilidad de enfermar para un *no* fumador es 7,4 veces mayor que la de un fumador.

El detalle que explica este sorprendente resultado consiste en lo siguiente: el fenómeno observado puede ser debido exclusivamente al hecho de que muchos sujetos que han padecido la enfermedad dejaron, precisamente por ello, de fumar; paralelamente, otros muchos continúan fumando porque no han padecido aún la dolencia y no han sentido la necesidad de abandonar esa práctica.

La clave del asunto radica en que la pregunta que se ha hecho sobre el hábito de fumar es irrelevante (a la postre, peor que irrelevante: perniciosa); lo que interesa *no* es si los sujetos fuman o no ***en el momento de la encuesta.*** Teniendo en cuenta que lo que se registró para distinguir casos de controles fue si se produjo o no la enfermedad ***durante 1994,*** quizás la pregunta adecuada sea: ***¿Fumaba usted a lo largo del trienio 1991-1993?*** Solamente de ese modo se podría estar seguro de que se están registrando los datos acaecidos ***en el orden compatible con la hipótesis que se valora.*** Sería una «seguridad» relativa, desde luego, pues estará sujeta a la calidad de un dato basado en el testimonio del sujeto. Estructuralmente, sin embargo, esa sería la formulación correcta.

7.7.2. Educación sanitaria inducida por las enfermedades venéreas

La falacia presente en el problema de la transgresión de la precedencia temporal puede ser muy insidiosa, como ilustra el siguiente ejemplo. Imaginemos que entre 1990 y 1994 se ha desarrollado una campaña educativa orientada a la prevención de enfermedades venéreas y que se hace un estudio para evaluar sus efectos, particularmente su capacidad potencial para inducir cambios en las conductas sexuales. Se lleva adelante una encuesta de la población objeto de la campaña según la cual se puede clasificar a cada sujeto en una de dos categorías según sea poseedor de una cultura sanitaria adecuada o no.

La idea es computar las tasas específicas de individuos que mantienen prácticas «negativas» desde el punto de vista profiláctico (por ejemplo, no usar preservativos) en estos dos grupos, definidos por la posesión de convicciones diferentes. Este enfo-

que metodológico, además de que no individualizaría a la campaña como responsable del posible cambio (ya que no se estudia un grupo de referencia; es decir, un grupo no abarcado por la campaña) olvida que bien puede ocurrir que la convicción favorable se haya consolidado en algunos sujetos ***como consecuencia*** de la misma práctica negativa que se está midiendo en calidad de presunta respuesta. Es decir, no se estaría controlando la posibilidad de que para una parte de los individuos pudieran valer silogismos tales como: «Adquirí una gonorrea por no usar preservativo: ella me hizo comprender la importancia de utilizarlos».

Si no se prevé y elimina el papel de inversiones perfectamente verosímiles como ésta, difícilmente se podrá obtener una medición «limpia» del posible efecto que tienen las convicciones sobre las prácticas.

7.7.3. Emigrando hacia la bronquitis

Si se observa que una zona geográfica exhibe una tasa de prevalencia de bronquitis crónica sustancialmente más alta que otra, se está sin duda constatando la existencia de una asociación (en este caso, entre zona de residencia y padecimiento de bronquitis).

Pero la inferencia tácita o explícita de que tal diferencia es *debida* a las características de la zona (por ejemplo, de que es el clima de la zona en cuestión el que causa el daño) puede ser totalmente espuria.

Podría ocurrir todo lo contrario: que el clima de esa zona fuese tan beneficioso que muchos individuos que padecen la dolencia se trasladan hacia allí. 0 también que se trate de una zona cuyos rasgos climáticos beneficiosos eleven mucho el costo de la vida, de modo que predominantemente emigran hacia ella individuos adinerados y de mayor edad; siendo la bronquitis crónica mucho más frecuente entre personas mayores que entre jóvenes, tan alta prevalencia puede realmente deberse a la diferencia en la estructura por edades de la población y no al efecto del clima per se.

De hecho estamos ante la misma inversión: la enfermedad ocurre primero, el presunto factor de riesgo (vivir en la zona «peligrosa») es consecuencia de ella. En cualquier caso, lo importante es que la observación plana y transversal de los datos no permite deslindarlo.

7.7.4. El peligro de no ser militar

La tasa de mortalidad durante cierto período fue muy inferior en el ejército que en la sociedad civil; aparentemente, desempeñarse en las fuerzas armadas es más seguro que hacerlo fuera de ellas. De hecho así se razona, por ejemplo, con la tasa de mortalidad infantil computada en diferentes países o para diferentes grupos sociales.

Pero en este caso es obvio que la institución castrense está casi exclusivamente integrada por personal joven y saludable; no están enlistados, por ejemplo, ni lactantes, ni ancianos, ni cardiópatas.

En el lenguaje coloquial de la investigación biomédica se dice que el problema radica en que ***los grupos no son comparables.*** Personalmente, esa expresión no me parece muy atinada pues los grupos (en este caso, militares y civiles) sí son susceptibles de ser comparados. De hecho, en este ejemplo han sido comparados y, además, el balance es correcto: la tasa de mortalidad es efectivamente menor dentro de la institución militar. Lo incorrecto es la inferencia causal que se ha hecho: no puede afirmarse que el hecho de pertenecer al ejército tiene un efecto causal de naturaleza protectora porque lo que se ha registrado como presunta causa no es anterior al supuesto efecto sino exactamente al revés: la propensión a morir, por decirlo de algún modo, contribuye a que no se pertenezca al ejército. En un artículo debido a Seltzer y Jablon (1974) se examina con detalle este sesgo.

7.8. Los síntomas de una crisis

De hecho, va resultando imposible desconocer los primeros, inquietantes síntomas de una crisis de la epidemiología analítica observacional. En un incisivo trabajo publicado por la revista ***Science,*** Taubes (1995) construye un discurso crítico vertebrado a partir de las propias declaraciones de figuras tan emblemáticas de la epidemiología actual como Greenland, Sackett, Rothman, Mac Mahon, Breslow, Feinstein y Peto, en que los propios protagonistas van dando indicios de la desazón que merodea por sus predios metodológicos. En síntesis, la crítica esencial es muy simple: los resultados alcanzados por esta disciplina en la ***explicación*** de mecanismos causales ha sido desproporcionadamente modesta con respecto al tiempo y los recursos empleados.

Pero, desde luego, este no es el primero ni el único reflejo de la insatisfacción que aflora en este terreno. Syme (1989) ha elaborado un penetrante discurso para reclamar lo que denominó como «una epidemiología más relevante». Según Syme:

> ***Hoy en día, casi ninguna investigación epidemiológica explica con éxito, ni siquiera de una manera satisfactoria, la aparición de las enfermedades que estudiamos. Los éxitos con que contamos responden, casi invariablemente, al modelo de causalidad antiguo.***

El centro de la inquietud radica en que cuatro décadas de intensos y costosos estudios orientados a esclarecer las más importantes enfermedades que aquejan al mundo desarrollado, los resultados son decepcionantes. En efecto, hasta los factores o conductas de riesgo más claramente corroborados como tales (hábito de fumar, consumo de grasas saturadas, sedentarismo, etc.) son marcadamente inespecíficos, ya que la mayoría de los enfermos no están aquejados por ellos y, sobre

todo, poseen muy escaso valor predictivo, pues la inmensa mayoría de los que tienen tales factores nunca desarrollan la enfermedad.

Los resultados del Pooling Project Research Group (1978), que consolidó los hallazgos de seis importantes proyectos sobre cardiopatías, confirman inequívocamente esta doble afirmación. Por una parte, cerca del 10% de los más de 7000 sujetos estudiados eran a la vez fumadores, hipercolesterolémicos e hipertensos; de estos sujetos «de alto riesgo», solo el 13% sufrieron un ataque cardíaco a lo largo del siguiente decenio. Y la inmensa mayoría de los infartados en ese lapso estaba libre de tales factores.

Mc Kinlay (1994) llega a caracterizar las situación del modo siguiente[13]:

> ***Lo que actualmente se considera epidemiología establecida se caracteriza por el reduccionismo biofisiológico, absorción por la biomedicina, ausencia de una verdadera teoría sobre las causas de la enfermedad, pensamiento dicotómico sobre la enfermedad (cada persona está enferma o está sana), un amasijo de factores de riesgo, confusión entre asociaciones observacionales y causalidad, dogmatismo acerca de cuáles diseños de estudios son aceptables y excesiva repetición de estudios.***

¿Cómo explicarse esta desalentadora realidad? Se ha argüido que las enfermedades, en especial las coronarias y los tumores malignos, son entidades muy complicadas y dependientes de demasiadas variables mutuamente correlacionadas. A partir de esta explicación, en los últimos años se ha multiplicado el empleo de complejos modelos estadísticos multivariados que supuestamente podrían esclarecer las cosas. Sin embargo, algunos de estos esfuerzos (que, ciertamente, no parecen haber producido un giro visible en la situación) me recuerdan a los constructores de máquinas de movimiento perpetuo, quienes ignoraban la ley de la conservación de la energía y, creyendo que sus fracasos se derivaban de que el diseño del aparato no era suficientemente eficaz, procedían a desgastarse con nuevos y más sofisticados ingenios. El ejemplo que se desarrolla en la Sección 8.8 ilustra cuán decepcionantes pueden resultar las manipulaciones estadísticas cuando no se enmarcan en una teoría lúcida y coherente.

No casualmente se ha llamado la atención (Charlton, 1996) sobre la sorprendente circunstancia de que, con excepción de unos pocos contraejemplos, la epidemiología más productiva ha estado a cargo de científicos clínicos interesados en problemas específicos de patología y no de especialistas en epidemiología cuyo interés se centra en la estadística y la metodología.

Sin embargo, el debate relevante gira en torno a cuáles son los propios límites epistemológicos de una disciplina que generalmente no puede acudir al recurso experimental y de la que podríamos quizás estar esperando más de lo que puede

[13] Citado por Pearce (1996).

aportar (Mc Cormick, 1996). Un cultor de la especialidad como Miettinen (1985) suscribe implícitamente tal punto de vista cuando apunta en su ***Epidemiología Teórica*** que los epidemiólogos se ocupan usualmente de las relaciones descriptivas, las cuales «son ajenas a la interpretación causal de las asociaciones».

Así las cosas, actualmente pueden leerse afirmaciones como la de Charlton (1996) cuando escribía en la páginas del ***Journal of Clinical Epidemiology*** que:

> ***La epidemiología cada día se considera más a sí misma como una disciplina autónoma con sus propios patrones intelectuales para encontrar demostraciones (...) a*** *pesar* ***de que está sistemáticamente incapacitada para resolver debates concernientes a los mecanismos causales.***

La realidad, afortunadamente, dista de ser esa. A la vez que, ciertamente, se hacen más obvias y reconocidas las limitaciones intrínsecas de la epidemiología en su cauce actual, se van creando alternativas, proponiendo y debatiendo rectificaciones, y abriendo avenidas novedosas. Una línea se orienta hacia la integración con otras disciplinas, en especial con las ciencias básicas (Cabello, 1996). Vineis y Porta (1996), por ejemplo, ilustran ese proceso secuencial de integración de resultados procedentes de los niveles básico, clínico y epidemiológico en el campo de los biomarcadores.

La otra vertiente (véase Susser, 1996), por cierto en absoluto incompatible con el afán integrador arriba mencionado, se orienta hacia el establecimiento de un nuevo paradigma y una nueva manera de ver las cosas.

Pearce (1996) lo explica así:

> ***La epidemiología está inevitablemente imbricada en la sociedad, y no es ni factible ni deseable estudiar las causas de la enfermedad en abstracto. Para entender esas causas en una población es esencial entender su contexto social e histórico, así como reconocer la importancia de los conocimientos locales en lugar de buscar solamente relaciones universales. Esto requiere un mayor involucramiento de las ciencias sociales y un enfoque multidisciplinario. La epidemiología aporta uno de los enfoques para identificar los determinantes de la salud en una población y debe complementarse con otros recursos cuantitativos de las ciencias sociales, así como con estudios cualitativos e históricos. El énfasis debe ponerse en usar una metodología adecuada y no en procurar que el problema se ajuste a las metodologías.***

En esa misma cuerda se ubica el artículo ya citado de Syme (1989) sobre la irrelevancia de la epidemiología actual, la cual invierte cada mez más recursos en esclarecer asuntos más fútiles. El autor reclama e ilustra el desarrollo de modelos conceptuales más abarcadores coherentemente con la siguiente posición:

> ***Cuando comprendemos*** *poco,* ***todo parece caótico, desorganizado y complicado (...) Ciertamente, en la etiología de una enfermedad intervienen muchos factores de riesgo,***

pero si dispusiéramos de un modelo teórico coherente, se podría esperar que todos estos factores jugaran un papel mucho más significativo.

Estas nuevas líneas de pensamiento aún esperan por aportes más concretos y operativos, pero puedo citar un libro llamado a tener, en mi opinión, gran trascendencia estratégica en ese propósito de reestructurar nuestra distorsionada perspectiva actual; se trata del volumen titulado ***¿Porqué alguna gente enferma y otros no?*** de Evans, Morris y Marmor (1994) [14].

Bibliografía

Beaglehole R, Bonita R, Kjellström T (1994). ***Epidemiología básica.*** Organización Panamericana de la Salud, Publicación Científica Nº 551, Washington.

Bernard C (1961). ***An introduction to the study of experimental medicine.*** Collier Books, New York.

Bollet AJ (1964). ***On seeking the cause of disease.*** Clinical Research 12: 305-310.

Cabello J (1996). ***El futuro de la práctica clínica. La investigación necesaria.*** Libro de Ponencias. Seminario REUNI, Albacete: 17-58.

Caplan G (1996). ***Principios de psiquiatría preventiva.*** Paidós, Buenos Aires.

Charlton BG (1996). ***The scope and nature of epidemiology.*** Journal of Clinical Epidemiology 49: 623-626.

Davey G et al (1990). ***The black report on socioeconomic inequalities in health: 10 years on.*** British Medical Journal 301: 373-377.

Duglosz L, Vena J, Bayers T, Sever L, Bracken M, Marshall E (1992). ***Congenital defects and electric bed heating in New York State: a register based case-control study.*** American Journal of Epidemiology 135: 1000-1011.

Evans RG, Stoddart GL (1990). ***Producing health, consuming health care.*** Social Science and Medicine 31: 1347-1363.

Evans RG, Morris LB, Marmor TR (1994). ***Why are Some People Healthy and Others Not? The determinants of Health of Populations.*** Aldine de Gruyter, New York.

Galbraith JK (1991). ***La cultura de la satisfacción.*** 2ª ed., Ariel Sociedad Económica, Madrid.

Gardner M (1975). ***Aha! Gotcha. Paradoxes to puzzle and delight,*** Scientific American, New York.

Giel R (1993). ***Trastornos mentales en atención primaria.*** En ***Nuevos Sistemas de Atención Primaria: evaluación e investigación.*** Pág ***245-254.*** Consejería de Sanidad y Servicios Sociales, Principado de Asturias.

Gray-Donald A, Kramer G (1988). ***Causality inference in observational vs. experimental studies.*** American Journal of Epidemiology 127: 885-892.

[14] Este libro ha sido traducido al castellano y publicado en 1996 por la editorial Díaz de Santos.

Greenland S (1988). ***Probability:*** *an* ***elaboration of the insufficiency of current*** *popperian approaches for epidemiological analysis.* En: Rothman KJ (editor) ***Causal Znference*** Chestnut Hill: Epidemiology Resources, Boston.

Greenland S (1990). ***Randomization, statistics and causal inference.*** Epidemiology 1: 421-429.

Hempel CG (1973). *Filosofía* ***de la ciencia natural.*** Alianza, Madrid.

Hertzman C (1986). ***The Health Context of Worklife Choice.*** Canadian Mental Association, Ottawa.

Hill AB (1965). ***The environment and disease: association of causation?*** Proceedings of the Roya1 Society of Medicine 58: 295-300.

Lanes SF (1988). ***The*** *logic* ***of causal inference.*** En: Rothman KJ (editor) ***Causal*** *Inference* Chestnut Hill: Epidemiology Resources, Boston.

Lilienfeld AM, Lilienfeld DE (1980). ***Foundations of epidemiology,*** 2.ª ed, Oxford University Press, New York.

Lorenz K (1950). ***The*** *comparative* ***method of studying innate behavior patterns.*** Cambridge University Press, Cambridge.

Maclure M (1985). *Popperian refutation in* ***epidemiology.*** American Journal of Epidemiology 121: 343-350.

Materazzi MA (1991). ***Propuesta de Prevención Permanente.*** Paidós, Buenos Aires.

McCormick J (1988). ***The multifactorial*** *aetiology* ***of*** *coronary* ***heart disease: a*** *dangerous* ***delusion.*** Perspectives in Biology and Medicine 32: 103-108.

McCormick J (1996). ***Medical hubris and the*** *public* ***health: the ethical dimension.*** Journal of Clinical Epidemiology 49: 619-621.

McIntyre N (1988). ***The*** *truth,* ***the whole*** *truth,* ***and nothing but the*** *truth.* En: Rothman KJ (editor) ***Causal Znference*** Chestnut Hill: Epidemiology Resources, Boston.

Mc Kinlay JB (1994). ***Towards*** *appropriate levels* ***of analysis, research methods and health*** *public* ***policy.*** International Symposium on Quality of Life and Health: Mayo 25-27, Berlin.

Miettinen OS (1985). ***Theoretical Epidemiology.*** Wiley, New York.

Morabia A (1991). *On* ***the*** *origin* ***of*** *Hill's* ***causal*** *criteria.* Epidemiology 2: 367-379.

Ng SKC (1991). *Does* ***epidemiology need a new philosophy? A case study of logical inquiry*** *in* ***the acquired immunodeficiency*** *syndrome epidemic.* American Journal of Epidemiology 133:1073-1077.

Ortega y Gasset J (1958). ***El tema de nuestro tiempo.*** Revista de Occidente, 13ª ed, Madrid.

Pearce N (1996). ***Traditional epidemiology, modem epidemiology, and*** *Public* ***Health.*** American Journal of Public Health *86:*678-683.

Pooling Project Research Group (1978). *«Relationship* ***of blood*** *pressure, serum cholesterol,* ***smoking*** *habit,* ***relative weight and EKG abnormalities to the incidence of major*** *coronary* ***events: Final report of the Pooling*** *Project».* Journal of Chronic Diseases, 31 (Special Issue): 201-306.

Popper KR (1972). ***Objective knowledge: an evolutiona y approach.*** Clarendon Press, Oxford.

Rogot E, Sorlie PD, Backlund E (1992). ***Air-conditioning and mortality in hot weather.*** American Journal of Epidemiology 136: 106-116.

Rothman JK (1986). ***Modern epidemiology.*** Little, Brown and Col., Boston.

Russell B (1949). ***La perspectiva científica.*** Ariel, Barcelona, 1983 (traducción de G. Saus Huelin y Manuel Sacristán). Ed orig: ***The scientific outlook.*** George Allen and Unwin, Londres.

Schlesinger GN (1988). ***Scientists and philosophy.*** En: Rothman KJ (editor) ***Causal Inference*** Chestnut Hill: Epidemiology Resources, Boston.

Schwartz S (1994). ***The fallacy of the ecological fallacy: Thepotential misuse of a concept and the consequences.*** American Journal of Public Health 1994; 84: 819-824.

Seltzer CC, Jablon S (1974). ***Effects of selection on mortality*** American Journal of Epidemiology 100: 376-389.

Sicliani S et al (1988). ***A community survey on mental health in South-Verona.*** Research Report, Instituto di Psichiatria, Universitá di Verona.

Silva LC (1993). ***Prevención en salud mental desde una perspectiva comunitaria.*** Memorias de las Segundas Jornadas sobre Actividades Preventivas en el Area de Salud: 111-129, Burgos.

Silva LC (1995). ***Excursión a la regresión logística en ciencias de la salud.*** Díaz de Santos, Madrid.

Simon HA (1968). ***Causation. En International Encyclopedia of the Social Sciences*** Vo1 2: 350-356.

Skrabanek P (1994). ***The death of humane medicine and the rise of coercive healthism.*** The Social Affairs Unit, Publication No 59, London.

Stehbeens WE (1985). ***The concept of cause in disease.*** Journal of Chronic Diseases 38: 947-950.

Susser M (1973). ***Causal thinking in the health science.*** Oxford University Press, New York.

Susser M (1994). ***The logic in ecological: I. The logic of analysis.*** American Journal of Public Health 84: 825-829.

Susser M (1994). ***The logic in ecological: II. The logic of design.*** American Journal of Public Health 84: 830-833.

Susser M, Susser E (1996). ***Choosing a Future for Epidemiology: I. Eras and Paradigms.*** American Journal of Public Health 86: 668-673.

Susser M, Susser E (1996). ***Choosing a Future for Epidemiology: II. From black box to chinese boxes and eco-epidemiology.*** American Journal of Public Health 86: 674-677.

Syme L (1989). ***La investigación sobre la salud y la enfermedad en la sociedad actual: la necesidad de una epidemiología más relevante.*** Anthropos 1181199: 39-46.

Taubes G (1995). ***Epidemiology faces its limits.*** Science 269:164-169.

Vanderbroucke JP, Parodel UPAM (1989). ***An autopsy of epidemiologic methods: the***

case of <<poppers» in the early epidemic of the acquired immunodeficiency syndrome (AZDS). American Journal of Epidemiology 129: 455-457.

Vineis P, Porta M (1996). ***Causal thinking, biomarkers, and mechanisms of carcinogenesis.*** Journal of Clinical Epidemiology 49: 951-956.

Williams CS, Kimberly AB, Eskenazi B (1992). ***Infant resuscitation is associated with an increased risk of left-handedness.*** American Journal of Epidemiology 136: 277-286.

8

El sesgo que acecha

El error ignora la crítica; la mentira le teme; la verdad nace de ella.

JOSÉ INGENIEROS

En cierta ocasión leí que hacer una buena investigación científica en biomedicina no es más que seguir los dictados del sentido común, pero precaviéndose de los sesgos que acechan permanentemente al proceso de investigación. Ésta es una manera de decir que no siempre el ingenio creativo o la aparición de ese ingrediente de la ciencia conocido como ***serendipia*** [1] desempeñan un papel clave; y que sí es, en cambio, imprescindible estar alerta contra los sesgos.

El término «sesgo», como ocurre tan frecuentemente, tiene varias acepciones. En nuestro idioma, según el diccionario de la Real Academia Española, la acepción de sesgo que más se ajusta al sentido con que se usa en estadística es la siguiente: ***Oblicuidad o torcimiento de una cosa hacia un lado, o en el corte, o en la situación, o en el movimiento.***

Genéricamente, al hablar de sesgos, en este contexto nos referimos a esas insidiosas manifestaciones de la realidad que disfrazan la verdad de manera que nos parece estar observando lo que no existe, o de forma que pasemos por alto los indicios necesarios para no sacar conclusiones incorrectas.

En este capítulo nos ocuparemos de comentar algunos de ellos. Cabe subrayar, sin embargo, que hay dos tipos de sesgos: los objetivos, que amenazan a todo investigador, y los que toman forma gracias a algún tipo de prejuicio o huella mental asentada en su propia cabeza. Los dos parecerían estar esperando su oportunidad para hacerse presentes, y ambos pueden conseguirlo; la combinación puede ser devastadora.

[1] Me adhiero a la forma en que Pérez (1980) traduce el vocablo «serendipity», inventado en el siglo XVIII por el escritor inglés Horacio Walpole para aludir a la circunstancia de hacer descubrimientos inesperados cuando se procura realmente hallar otra cosa.

8.1. La lógica primero

El primer sesgo que debe resaltarse es el que se produce cuando hay una quiebra de la lógica. Este problema no es, evidentemente, de la esfera cuantitativa, ni puede resolverse por mucho que se dominen las técnicas que de allí provengan.

Peter Medawar, connotado investigador cuyos descubrimientos en materia de histocompatibilidad le valieron el Nobel en 1963, relata (Medawar, 1967) que solía proponer una ***prueba de inteligencia*** ante sus auditorios científicos en los siguientes términos:

> ***A muchas personas, las figuras de los cuadros de*** El Greco ***les parecen antinaturalmente altas y delgadas. Un oftalmólogo supuso que habían sido pintadas así porque*** El Greco ***padecía de un defecto de la vista que le hacía vera*** *las* ***personas de tal manera y, tal como las veía, necesariamente las pintaba. ¿Puede ser válida semejante interpretación?***

Medawar comenta que, al plantear esta pregunta, solía añadir que cualquiera que pudiera comprender ***instantáneamente*** que esta explicación es absurda, y que lo es más por razones filosóficas que estéticas, era indudablemente brillante; y que aquel que no entendiera la explicación, que es de índole epistemológica, no pictórica ni clínica, ni siquiera después de que se le explicara, haría bien en no dedicarse a la investigación.

La explicación, es la siguiente:

> ***Supongamos que el defecto de visión de un pintor fuera, lo que es perfectamente posible, la diplopia, que consiste en verlo todo doble. Si la explicación del oftalmólogo fuera correcta, entonces ese artista pintaría sus figuras duplicadas; pero de hacerlo así, entonces, al inspeccionar su obra, ¿no vería cuádruples todas las figuras, y por tanto sospecharía que algo andaba mal? Por otra parte, al hacer un solo trazo, él estaría viendo cómo se producen dos trazos, procedentes de dos manos y de dos pinceles. Pero nosotros, naturalmente, veríamos moverse una sola mano, un solo pincel y, desde luego, un solo trazo en el lienzo. Es decir; aun cuando el pintor tuviera ese trastorno, éste no quedanría reflejado en su obra. En síntesis, si el pintor padeciera de un defecto de visión, las figuras que le parezcan naturales a él, también habrán de parecerles naturales a los demás. Si algunas de las figuras de*** El Greco ***parecen antinaturalmente altas y delgadas, es porque esa fue la intención estética del artista.***

Vale decir: si el investigador se divorcia del marco lógico en que métodos y cómputos están llamados a desempeñarse, de nada sirve el intento de eludir sutiles acechanzas metodológicas, ¡mucho menos acudir a poderosos instrumentos estadísticos o computacionales! Recientemente un periódico madrileño reproducía un artículo (Itzhaki, 1995) aparecido en ***New Scientist*** donde se detallan los experimentos de Stuart Anstis, psicólogo de la Universidad de California, para demostrar ***por esa vía*** que el estilo de El Greco no podía atribuirse a un defecto oftálmico.

8.2. Muestras sesgadas

En el ámbito de la teoría de estimación, cuando el interés se concentra en conocer la magnitud de un parámetro poblacional (una media, un porcentaje, un coeficiente de correlación, etc.), usualmente se obtiene una estimación basada en una muestra. En el contexto de ese proceso, se denomina ***sesgo de la estimación*** a la distancia que cabe esperar que haya entre el valor del parámetro y la estimación realizada.

Esa magnitud puede no ser nula cuando, por ejemplo, el tipo de muestreo utilizado produce una distorsión sistémica, un ***torcimiento hacia un lado,*** tal como el que se produciría si, al estimar el consumo per cápita de electricidad en una comunidad, se usara una muestra procedente de las familias ubicadas en su zona más elegante («torcimiento» hacia la sobreestimación) o una muestra sólo de los sectores más deprimidos («torcimiento» hacia la subestimación).

Muchas veces se puede saber, o sospechar fundadamente, de la posible existencia de un sesgo aunque no se pueda prever en qué dirección él actúa. Consideremos el artículo titulado ***Consumo de drogas en una muestra de médicos rurales de la provincia de Valladolid*** de Carvajal ***et al.*** (1984). El propio título del trabajo anuncia algo anómalo: la magnitud o forma del consumo ***en una muestra*** no interesa a nadie. Lo que puede interesar es esta información para ***la población*** de médicos. El uso de una muestra es un aspecto metodológico ajeno por completo al propósito de un estudio, cualquiera que sea éste; si la muestra ***representa*** adecuadamente a la población, los resultados obtenidos de aquella podrán extenderse o extrapolarse a esta última y entonces se habrá alcanzado aquel propósito; pero lo que nunca interesa es un resultado muestral per se sino el que se desprende de ese proceso de extrapolación.

En el artículo se exponen los resultados obtenidos en una encuesta realizada por medio del correo a la que 73 médicos aportaron testimonios sobre sus hábitos de consumo de drogas (tanto de las llamadas «institucionales» -alcohol, cafeína y tabaco- como de otras, tales como anfetaminas, ácidos, opiáceos y alucinógenos).

La población de interés estaba conformada por 211 médicos. A todos ellos les fue originalmente remitido el cuestionario, pero sólo respondió el 34,5% de ellos. Dada la naturaleza obviamente comprometedora de las preguntas [2], resulta altamente dudosa la representatividad de una muestra autoconfigurada por sus integrantes.

Es muy difícil establecer a qué población esta muestra podría representar: ¿a los que no tienen nada que ocultar? ¿A los que no tienen aprensión en admitir sus hábitos?, ¿a los que aprovechan la ocasión para dar pistas falsas sobre sus prácticas? Considero imposible responder estas preguntas. Sin embargo, de lo que no quedan dudas es de que la muestra representa... a los médicos que acceden a contestar, subgrupo cuya diferencia con la de los que optan por no responder es algo más que

[2] **Los** propios autores adelantan en el trabajo un juicio moral negativo hacia un médico que consuma estas sustancias.

verosímil. Se trata a todas luces de una muestra llamada a arrojar estimaciones sesgadas.

Por otra parte, aun cuando la totalidad de los médicos hubiese respondido, habría motivos para sospechar la presencia de un sesgo: es bien conocido que cuando se formula una pregunta embarazosa, los resultados suelen arrojar subestimaciones (o sobrestimaciones, en dependencia de lo que se investigue) de la realidad.

El trabajo, por ejemplo, comunica que el porcentaje de médicos que consumen cocaína (ya fuese habitual o esporádicamente) es nulo; ni hombres ni mujeres de la muestra admiten tal consumo. ¿Podrá razonablemente confiarse en que, si la realidad fuese otra, ella hubiera quedado fielmente reflejada en las respuestas?

De hecho, se trata de un antiguo problema que ha sido objeto de atención cuidadosa. Un examen detallado del tema y de diversas soluciones posibles puede hallarse en el libro de Baruch y Cecil (1979).

8.3. Una solución aleatoria para la confidencialidad

En una situación en que las preguntas formuladas sean de naturaleza altamente comprometedora (consumo de drogas, prácticas sexuales, conductas ilegales, etc.) sería iluso esperar que las respuestas sean veraces y, por tanto, poco riguroso sacar conclusiones globales de tal información salvo que se hayan adoptado precauciones metodológicas especiales.

Como se comentó en la sección precedente, el interrogado puede sentirse incómodo ante la situación y por ello, aun en caso de que se le haya persuadido del carácter confidencial de la encuesta, es probable que se inhiba de comunicar la verdad e incluso de contestar.

A continuación se expone la descripción que se hace en Silva (1982) de un problema de este tipo y de una solución que se ha propuesto para encararlo.

En 1973 se realizó una encuesta de fecundidad (véase Krotki y Mc Daniel, 1975) en Alberta, Canadá, país donde el aborto provocado era en aquel momento ilegal salvo que mediaran razones terapéuticas. Se seleccionaron 3 muestras independientes, de 327,269 y 342 mujeres en edad fértil respectivamente.

Entre otras, se formulaban las siguientes preguntas:

1. HA TENIDO UN ABORTO PROVOCADO (TERAPEUTICO O ILEGAL) DURANTE 1972?
2. ¿SE HA CASADO EN ALGUNA OPORTUNIDAD?

Es fácil advertir el carácter altamente sensitivo de la primera pregunta en aquel medio, a la vez que la segunda carece en principio de todo efecto inhibitorio.

Se siguieron procedimientos diferentes con cada una de las muestras. Para comenzar, expliquemos cómo se procedió con dos de ellas. La primera se abordó a través del interrogatorio directo; a las integrantes de la segunda se les indicó enviar su respuesta por correo *sin consignar el remitente. Se* procuraba eliminar así en este segundo grupo las fuentes de distorsión que verosímilmente aquejarían al primero.

Los porcentajes de respuestas afirmativas que se obtuvieron para la pregunta comprometedora fueron 0,3% y 0,8% respectivamente: el segundo porcentaje asciende a casi el triple del primero, aparentemente debido al efecto desinhibitorio de aquel procedimiento. Para la otra pregunta, en cambio, estos porcentajes fueron 82,3% y 81,8%; la notable similitud de estos dos números refleja y confirma la condición no comprometedora de la pregunta. Cabe preguntarse, sin embargo, si es confiable la encuesta anónima. Para valorarlo, miremos el problema desde otro ángulo.

Se quería conocer el número total de abortos ilegales que se verificaron en la ciudad durante el año 1972. Al tener en cuenta el total de mujeres en edad fértil en la ciudad y el resultado de la primera muestra, se estimó que 1.148 mujeres se habían practicado aborto de algún tipo; cuando se utilizó el resultado de la encuesta anónima, el número estimado ascendió a 3.058, cifra casi tres veces mayor.

Todo parece lógico pero, por su carácter oficial, se conocía el número de abortos terapéuticos (y por ello, legales): ¡ascendía a 4.040 en el período! Ello revela que incluso la encuesta anónima padeció de un sorprendente subregistro, ya que el número de abortos terapéuticos no puede ser mayor que la totalidad de interrupciones [3] (terapéuticas e ilegales). Con esto se ilustra convincentemente la inoperancia de la encuesta anónima para evitar el temor de que respuesta e identidad puedan ser conectados.

Este problema venía planteando un desafío a los estadísticos: ¿cómo obtener conclusiones confiables logrando a la vez que cada interrogado sepa (no que crea ni que confíe, sino que sepa) que no es posible establecer su posición respecto de la condición embarazosa? La solución hallada fue la que se aplicó a la tercera muestra en el estudio canadiense.

A mediados de la década de los 60, Warner (1965) había sembrado la primera semilla teórica de lo que luego constituyó un recurso en creciente expansión: *la técnica de respuesta aleatorizada.*

La idea central consiste en que el interrogado haga un experimento regido por el azar (tal como lanzar un dado) y, *sin revelar el resultado obtenido,* dé una información que dependa, según cierta regla predeterminada, de tal resultado y de su verdadera situación ante la cuestión indagada. Así, el encuestador nunca conocerá la situación que realmente corresponde al individuo; el estadístico, sin embargo, usando los datos recogidos y las leyes probabilísticas que rigen el experimento, podrá obtener datos *globales* correctos sobre la población así investigada.

Consideremos el problema de los abortos a través de una expresión sencilla del método, tomado de entre los múltiples procedimientos ideados con el mismo

[3] Aquí se está equiparando el número de mujeres que abortaron en un año con el de abortos producidos en ese lapso, ya que el caso de una sola mujer con dos o más interrupciones en tan breve lapso, aunque posible, es muy poco probable.

principio. En un cartón se dibuja un círculo, que se divide en dos partes dentro de las cuales aparecen afirmaciones complementarias de la manera indicada en la Figura 8.1. La parte más pequeña abarca la cuarta parte del área total del círculo.

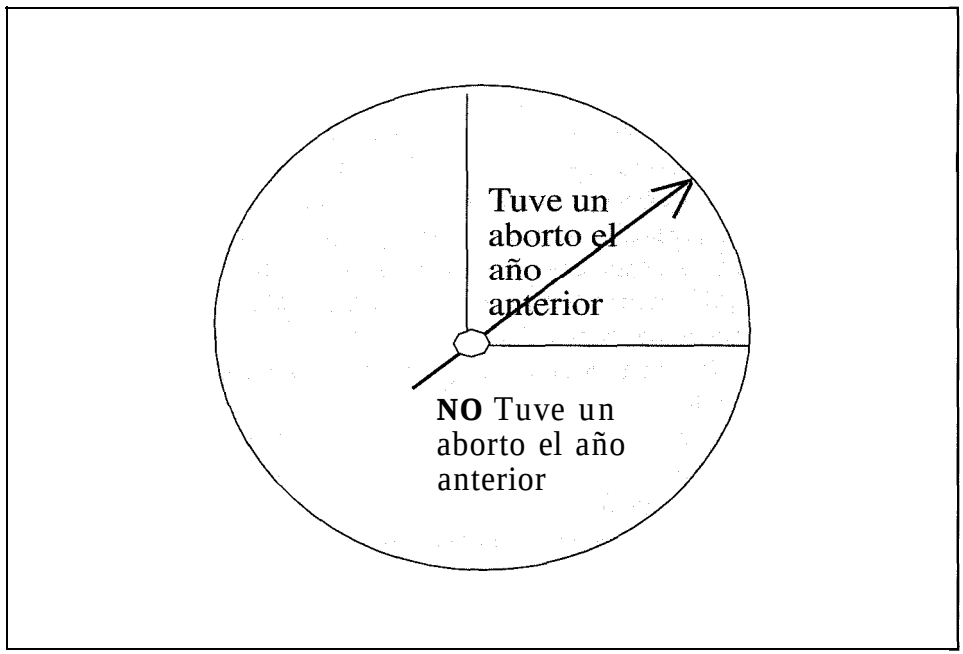


Figura 8.1. *Dispositivo aleatorio para llevar adelante la encuesta.*

Fija al centro del círculo hay una saeta metálica que, tras un impulso inicial del individuo interrogado, gira libremente. Cada mujer encara aquella de las dos afirmaciones que resulte señalada por la flecha al detenerse. En dependencia de su situación real, declara «verdadera» o «falsa» la afirmación seleccionada ***por el mecanismo;*** el encuestador simplemente anota esta respuesta, sin conocer, naturalmente, a qué pregunta corresponde.

Una vez interrogadas ***n*** mujeres, se conocerá el número de las que contestaron «verdadero». Puede suponerse que sólo a una cuarta parte de las mujeres les haya correspondido evaluar la afirmación contenida en la sección pequeña, y también que el resultado de cada experimento (la región en que cae la saeta) es independiente de que la mujer se haya o no practicado un aborto. A partir de estas suposiciones y usando la teoría elemental de probabilidades, es fácil deducir que el porcentaje desconocido ***P*** de mujeres que tuvieron un aborto se estima en este caso mediante la fórmula siguiente:

$$P = \frac{3n-4a}{2n}$$

Como se recordará, en la encuesta se habían tomado 3 muestras. A cada una de las 342 mujeres de la tercera muestra se le interrogó mediante un procedimiento similar al que se ha descrito. Cabe preguntarse ahora ¿es efectivamente eficaz el procedimiento? El resultado obtenido con los tres procedimientos se resume en la Tabla 8.1.

Tabla 8.1. ***Estimaciones de los porcentajes de diferentes tipos de abortos según método de encuesta***

	Encuesta directa	Anónima por correo	Respuesta aleatorizada
Tamaño muestra1	327	269	342
¿Ha tenido un aborto provocado (terapéutico o ilegal) durante I972?	0,3%	0,8%	3,2%
¿Se ha casado en alguna oportunidad?	82,3%	81,8%	84,4%

Por otra parte, las estimaciones para el número de abortos ilegales se resumen en la Tabla 8.2.

Tabla 8.2. ***Número estimado del total de abortos según método de encuesta***

Número estimado de abortos	Encuesta directa	Anónima por correo	Respuesta aleatorizada
Provocados (desconocido)	1.148	3.058	14.197
Terapéuticos (conocido)	4.040	4.040	4.040
Ilegales (diferencia)	-2.829	- 982	10.157

Las cifras son de una notable elocuencia; el porcentaje para la pregunta comprometedora se multiplica por 10, en tanto que el de la pregunta inocua no exhibe prácticamente variación alguna. Y mientras los procedimientos convencionales arrojaron conclusiones tan disparatadas como que el total de abortos ilegales era negativo, la técnica de respuesta aleatorizada consiguió ***arrancar*** una estimación razonable de dicho número: el azar asegura la privacidad de cada cual, pero la verdad se abre paso a través de las leyes que lo rigen.

Múltiples procedimientos similares al de Warner (en que se involucran cartas, dados o monedas) fueron creados para encarar problemas como éste. Por ejemplo, Dalenius y Vitale (1974) plantearon un ingenioso procedimiento para estimar la media μ de una variable discreta. Consideremos la variable: ***edad de la primeras relaciones sexuales (X). Se*** parte de que X puede tomar cualquiera de los 35 valores que van desde 15 a 49 años y que se trabaja con una muestra de ***n*** mujeres que admiten haber tenido tales relaciones en el momento de la encuesta.

En un cartón se dibuja un círculo que se divide en 35 secciones iguales y que se numeran del 15 al 49. Fija al centro del círculo hay, como antes, una saeta metálica que la interrogada hace girar; cuando el dispositivo se detiene, señala un número que el encuestador no conoce. Cada mujer se circunscribe a decir NO en caso de que sus primeras relaciones sexuales se hayan verificado estrictamente después de la edad señalada por la saeta y SÍ en caso opuesto. Puede probarse que $\hat{\mu} = 15 + 35\,\frac{a}{n}$ es un estimador insesgado de la media, donde es el número de mujeres que contestaron NO.

Si por ejemplo el porcentaje de respuestas negativas es 40%, entonces $\hat{\mu} = 15 + (35)\ (0{,}4) = 29$; o sea, se estima que la edad media del comienzo de relaciones sexuales es 29 años.

La técnica de respuesta aleatorizada alcanzó un considerable aval práctico. Múltiples experiencias se realizaron en esferas tales como fecundidad, conducta sexual, consumo de alcohol, actos ilegales y fraude académico. Massey, Ezatti y Folsom (1989) la sugirieron para estimar el porcentaje de personas que niegan falsamente mantener conductas de riesgo en relación con el SIDA. Zdep y Rhodes (1971) a través de una encuesta basada en respuesta aleatorizada encontraron, por ejemplo, que la estimación del porcentaje de individuos que golpean a sus hijos es cinco veces mayor que lo que arrojó el método de respuesta anónima por correo. En algunos países de alto desarrollo el procedimiento ha servido para mostrar que la prevalencia de drogadicción es mucho mayor de lo que los métodos tradicionales hacían suponer; por ejemplo, el estudio de Brown y Harding (1973) -en que se encuestaron miles de individuos- produjo estimaciones dos veces mayores para submuestras tratadas con respuesta aleatorizada que para submuestras interrogadas anónimamente.

Otros estudios informan resultados igualmente elocuentes; es obvio, sin embargo, que la mera diferencia entre las estimaciones no constituye una prueba irrefutable de la eficiencia del procedimiento.

Los trabajos de validación realizados agregaron en su momento algún aliento adicional a las expectativas creadas por el método. Para llevar adelante tal validación es menester comparar los verdaderos parámetros (suponiendo, claro, que estos sean conocidos) tanto con las estimaciones obtenidas por conducto de la técnica novedosa como con las que proceden de métodos tradicionales. Lamb y Stem (1978) y Tracy y Fox (1981) obtuvieron resultados bastante estimulantes en esta línea.

La experiencia acumulada hace pensar en general que el grado de confianza del

interrogado aumenta considerablemente entre los que acceden a participar; sin embargo, el grado de participación no se ve sensiblemente incrementado. En efecto, los métodos estadísticos aún generan suspicacia y asombro entre los interrogados y-según se informa en la literatura- su aplicación no ha producido la disminución esperada en las tasas de no respuesta.

La manera en que se explican y aplican los procedimientos, el mecanismo aleatorio utilizado y, particularmente, el nivel cultural de los encuestados influyen decisivamente en el éxito de procedimientos como éste.

En este terreno Silva (1984) apuntaba algunas sugerencias:

a) La técnica debe incorporarse en una parte de la muestra piloto y el método regular en la otra parte, a fin de evaluar su comprensión, grado de aceptación y funcionamiento general en la población.
b) En su fase de aplicación es preciso constatar que cada interrogado ha comprendido tanto aquello que él debe hacer como que el método efectivamente confiere absoluta privacidad.
c) Las preguntas tratadas por respuesta aleatorizada deben aparecer al final del cuestionario, después de las que se formulan por vías convencionales y previa explicación de que se trata de un procedimiento para cuya aplicación se solicita especial cooperación.

8.4. Falacia ecológica

Formalmente se conoce como *falacia ecológica* al error en que se incurre cuando, equivocadamente, se da por sentado que una relación observada a nivel de agrupaciones también se verifica al de los individuos. Debe advertirse que el señalamiento de la existencia de este problema no significa que deba renunicarse a los estudios que miden variables a nivel grupal, como parece haberse creido en amplios sectores de la investigación epidemiológica contemporánea. Por el contrario, este es un poderoso instrumento en buena medida desdeñado (véase Sección 7.6), aunque sea menester cuidarse de no incurrir en la falacia que ahora nos ocupa y que fuera estudiada por primera vez por Robinson (1950). Un ejemplo que ilustra ese espejismo se ofrece a continuación.

Supongamos que se está evaluando un programa de educación para la salud desarrollado en cuatro ciudades, una de cuyas acciones consistió en ofrecer cursillos de educación sanitaria a los ciudadanos que pudieron captarse con ese fin. A los efectos de este estudio, se considera que un individuo tiene ***hábitos de vida saludables (HVS)*** si practica ejercicios regularmente y no fuma, dos objetivos del programa. Al examinar las tasas de prevalencia de adultos que poseen *HVS* en las cuatro ciudades, se observa una notable variabilidad: las cifras se mueven en el recorrido que, a grandes rasgos, va de 30% a 90%.

Tales datos condujeron a un examen de diversos indicadores en dichas ciudades a fin de elaborar posibles explicaciones. Uno de los aspectos estudiados produjo resultados muy llamativos: los porcentajes de sujetos que tenían cursos *educativos completados (CEC)* también exhibían un recorrido amplio (de 5% a 44%), pero según un patrón inverso a lo esperado, tal y como permite apreciar la Tabla 8.3.

Tabla 8.3. *Porcentajes de sujetos con HVS y CEC en las 4 ciudades*

	Ciudad			
	A	B	c	D
Porcentaje de sujetos con ***CEC***	4,7	9,0	20,0	44,3
Porcentaje de sujetos con HVS	90,0	83,3	60,0	28,6

Al ubicar estas parejas de valores en un sistema de ejes, se obtiene el resultado que muestra la Figura 8.2.

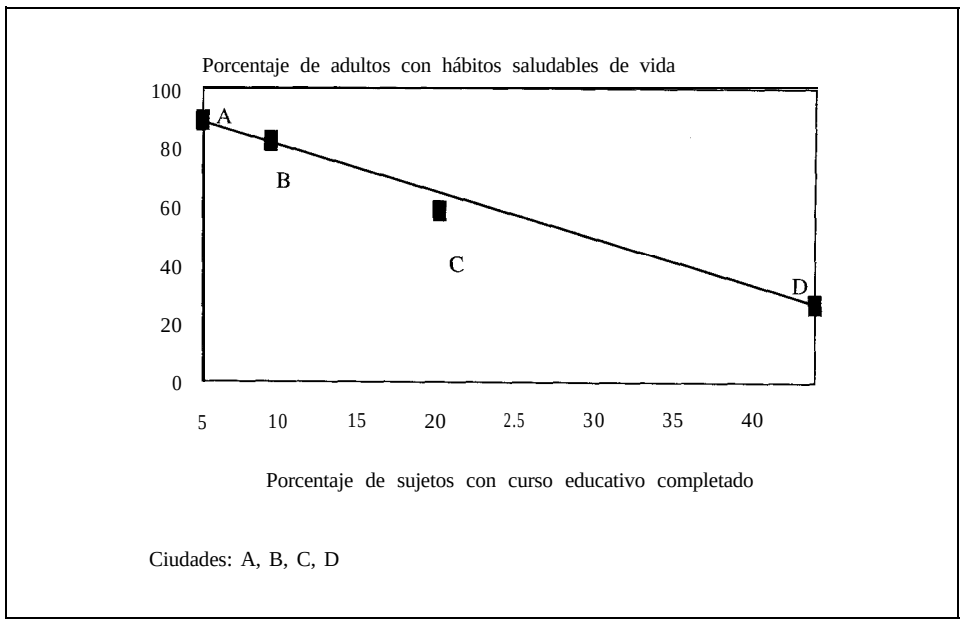


Figura 8.2. *Porcentajes* ***de adultos con curso educativo completado y hábitos saludables de vida en cuatro ciudades.***

En principio, pudiera pensarse que los cursos educativos desestimulan los hábitos saludables. En la búsqueda de una valoración más profunda de la paradójica conclusión que parece emerger de estos resultados, se decidió realizar sendos estudios transversales *(cross sectional studies)* en cada una de las 4 ciudades.

Una vez seleccionadas respectivas muestras de alrededor de 600 adultos en cada ciudad, y habiéndose encuestado a sus integrantes, se construyeron tablas de contingencia con las dos variables (*HVS* y *CEC*) del tipo siguiente:

Ciudad	CEC	No CEC	Total
HVS	a	b	a + b
No HVS	c	d	c + d
Total	a + c	b + d	n

Se computaron las tasas de: $HVS = \frac{a+b}{n}$ y $CEC = \frac{a+c}{n}$. Luego, en cada caso *se* computó el coeficiente de correlación tetracórica, definido por:

$$\varphi = s\sqrt{\frac{\chi^2}{n}}$$

donde χ^2 es el valor de ji cuadrado observado en cada tabla de 2 por 2 y *s* es el signo que tenga la expresión *ad-bc*.

El análisis (Tabla 8.4) permite apreciar de inmediato que estos estudios, a la vez que corroboran los resultados que aparecen en la Tabla 8.3 (ya que en cada caso se reproducen en esencia las tasas de prevalencia), revelan la existencia de una asociación, al *nivel de los individuos,* de naturaleza opuesta a la anterior: coeficientes de correlación positivos en todos los casos, y valores χ^2 de significativos al nivel cx = 0,05 para 3 de las ciudades.

De hecho, la conclusión que puede sacarse al nivel de los individuos - y es a ese nivel al que *debe* sacarse, ya que ¡son ellos y no las ciudades los que reciben cursos o modifican sus hábitos de vida!- es la contraria a la inicialmente considerada; la asociación entre el hecho de haber completado el curso educativo se asocia positivamente con el de poseer hábitos saludables [4].

[4] Nótese que afirmar que lo segundo se ***debe*** a lo primero (es decir, que hay una relación causa-efecto) sería aventurado: la naturaleza transversal de los estudios no lo permite (véase Capítulo 7).

Tabla 8.4. ***Resultado de cruzar los valores de CEC y HVS en las cuatro ciudades***

Ciudad A	CEC	No CEC	Total
HVS	28	494	522
No HVS	1	57	58
Total	29	551	580

Tasa de *HVS* = $\frac{522}{580}$ = *0,900*

Tasa de *CEC* = $\frac{29}{580}$ = *0,050*

χ^2 = 1,46 φ = +0,05

Ciudad B	CEC	No CEC	Total
HVS	55	493	548
NoHVS	4	108	112
Total	59	601	660

Tasa de *HVS* = $\frac{548}{660}$ = *0,830*

Tasa de *CEC* = $\frac{59}{660}$ = *0,094*

χ^2 = 4,77 φ = +0,09

Ciudad C	CEC	No CEC	Total
HVS	98	274	372
NoHVS	27	224	251
Total	125	498	623

Tasa de HVS = $\frac{372}{623}$ = 0,597

Tasa de *CEC* = $\frac{125}{623}$ = 0,201

χ^2 = 22,70 φ = +0,19

Ciudad D	CEC	No CEC	Total
HVS	114	164	278
No HVS	66	285	351
Total	180	449	629

Tasa de *HVS* = $\frac{278}{629}$ = 0,442

Tasa de *CEC* = $\frac{180}{629}$ = *0,286*

χ^2 = 37,44 φ = +0,24

8.5. La tortura de Procusto

En estrecha relación con la falacia ecológica se halla toda una serie de procedimientos y procesos intelectuales equívocos que tienen en común la interpretación errónea del coeficiente de correlación entre dos variables.

La asociación es una premisa indispensable para la causalidad; pero, aparte de esto, por alguna razón para mí no muy clara, a dicho coeficiente se le atribuyen virtudes que está lejos de poseer. A continuación se muestra una ilustración de cuán engañoso puede llegar a ser.

El hecho es que el coeficiente de correlación lineal tiende a aumentar en la medida en que las unidades de análisis se van «compactando» y dan lugar a unidades agregadas de mayor magnitud (en caso, naturalmente, de que tal compactación sea posible). Supongamos que, combinando diversas mediciones de contaminantes tales como dióxido de azufre, humos, óxidos de nitrógeno y monóxido de carbono, se ha construido un cierto *Índice de Contaminación Ambiental (ICA),* susceptible de ser medido diariamente en una ciudad. Imaginemos que esta medición se ha realizado durante 48 días consecutivos y que se ha registrado, asimismo, la *Incidencia de Asmáticos Agudos (IAA)* que acudieron cada día a los servicios de urgencia de la ciudad [5].

Por ejemplo, el primer día se registró un *ZCA* de 5,8 unidades y se produjeron 7 casos de sujetos con crisis asmática; para el segundo, el valor del *ZCA* fue de 22,0 unidades con 12 casos que acudieron a los servicios de urgencia. De ese modo se registraron 48 pares de datos, tal y como lo recoge la Tabla 8.5.

Para aquilatar la asociación entre ambas variables, se piensa entonces en aplicar técnicas de correlación, que son, según WHO (1983), «las técnicas más útiles y generalmente utilizadas en la epidemiología medioambiental».

Si se computa el coeficiente de correlación de Pearson entre estas dos variables, se obtiene un valor positivo y alejado de cero aunque, ciertamente, no muy alto: r = 0,628. Se trata, por otra parte, de un valor significativamente superior a cero (intervalo de confianza al 95%: 0,0301-0,823), aunque éste no es un dato relevante en el contexto de lo que se quiere ilustrar.

Imaginemos ahora que alguien plantea que, para que el posible efecto de la contaminación se exprese con más claridad, sería conveniente dar menos espacio a las fluctuaciones aleatorias, por lo cual sería conveniente «compactar» días contiguos. Es decir, podrían sumarse los valores del índice de días consecutivos, y hacer lo propio con los datos de asma [6].

De tal suerte, se tendrían 24 pares de datos. Para la primera pareja de días, por ejemplo, el *ICA* acumulado sería 27,8 -resultado de sumar 5,8 y 22,0-, en tanto que se habrán acumulado 19 casos de asmáticos críticos. El resultado de tal proceso se reproduce en la Tabla 8.6.

---

[5] En rigor, los estudios de este tipo suelen realizarse pareando el nivel del agente contaminante con la incidencia de enfermos correspondiente *a un lapso posterior* a la fecha en que se hizo la medición del contaminante (con un llamado «retardo»). A los efectos de la presente ilustración, sin embargo, es equivalente considerar que las mediciones son simultáneas.

[6] Recuérdese que si, en lugar de sumarse, los dos valores diarios fueran promediados, el resultado no cambiará, ya que el coeficiente de correlación entre los promedios sería el mismo que el que se obtiene con las sumas.

Tabla 8.5. ***Registros del Índice de Contaminación Ambiental y de la Incidencia de Asmáticos Agudos a lo largo de 48 días***

Día	ICA	IAA	Día	ICA	IAA
1	5,8	7	25	10,8	2
2	22,0	12	26	4,9	4
3	12,9	9	27	12,1	7
4	26,2	12	28	14,7	10
5	8,1	4	29	9,9	3
6	4,3	3	30	4,3	4
7	10,0	9	31	2,6	2
8	11,0	7	32	3,8	3
9	11,0	6	33	2,8	0
10	12,1	1	34	3,0	1
11	4,8	5	35	6,1	11
12	0,3	2	36	14,7	8
13	52,0	32	37	18,0	23
14	18,2	18	38	22,0	12
15	30,0	17	39	26,3	4
16	12,0	9	40	27,0	16
17	4,3	3	41	18,0	43
18	3,8	5	42	33,3	10
19	3,3	3	43	6,1	2
20	6,1	6	44	7,2	9
21	19,0	6	45	9,1	6
22	8,1	7	46	11,0	1
23	0,4	7	47	11,4	4
24	2,5	9	48	9,9	5

Ahora el coeficiente de correlación se eleva considerablemente y pasa a ser r = 0,846 (intervalo de confianza: 0,672 – 0,932). Si ese proceso se reproduce una vez más y se conforman segmentos de 4 días contiguos, se tendrán 12 pares de datos, que aparecen en la Tabla 8.7.

El coeficiente de correlación se incrementa ahora hasta un nivel francamente alto: r = 0,916 (intervalo de confianza: 0,721- 0,977). Finalmente, si ese proceso se reprodujera una vez más y se tomaran 6 acumulados de 8 días cada uno, se obtendrían los resultados de la Tabla 8.8.

El coeficiente de correlación de Pearson [7] se eleva a r = 0,925 (intervalo de confianza: 0,455 - 0,992).

[7] Por razones técnicas suele sugerirse que, cuando el tamaño muestra1 sea tan pequeño, se use el *coeficiente de Spearman* (versión del coeficiente de Pearson cuando se trabaja con rangos en lugar de con las cifras propiamente

Tabla 8.6. ***Registros del Indice de Contaminación Ambiental y de la Incidencia de Asmáticos Agudos (acumulados) a lo largo de 24pares de días***

Pares de días	ICA acumulado	IAA acumulado
l-2	27,8	19
3-4	39,l	21
5-6	12,4	7
7-8	32,0	21
9-10	23,l	7
ll-12	5,l	7
13-14	70,2	50
15-16	42,0	26
17-18	8,l	8
19-20	9,4	9
21-22	27,l	13
23-24	2,9	16
25-26	15,7	6
27-28	26,8	17
29-30	14,2	7
31-32	6,4	5
33-34	5,8	1
35-36	20,8	19
37-38	40,0	35
39-40	53,3	20
41-42	51,3	53
43-44	13,3	11
45-46	20,l	7
47-48	21,3	9

¿Qué conclusión se puede extraer de este proceso numérico? Mills (1990) comenta que en su oficina se ha vuelto un lugar común la apreciación de que, si los datos son torturados durante un tiempo suficientemente largo, terminarán por decir lo que queremos oír.

En cierto sentido hemos asistido a una sesión de tortura de los datos originales. La mitología griega da cuenta del ***modus operandi*** de Procusto, un bandido quien, luego de asaltar a los caminantes, forzaba sus cuerpos a ajustarse exactamente a una cama de hierro mediante el expedito recurso de mutilarlos o descoyuntarlos, según hiciera falta. Un tratamiento como el que se ha esbozado se inscribiría en lo que Mills llamaría «forma procusteana» de torturar los datos.

De hecho, puede demostrarse que, en la mayor parte de las situaciones, si se

dichas). A los efectos de lo que ahora se quiere subrayar, sin embargo, se computó el mismo coeficiente anterior (de Pearson) para esta muestra de 6 unidades.

Tabla 8.7. ***Registros del Índice de Contaminación Ambiental y de la Incidencia de Asmáticos Agudos (acumulados) a lo largo de 12 segmentos de 4 días***

Grupos de 4 días	ICA acumulado	IAA acumulado
1-4	55,9	35
5-8	44,4	28
9-12	28,2	14
13-16	112,2	76
17-20	17,5	17
21-24	30,0	29
25-28	42,5	23
29-32	20,6	12
33-36	26,6	20
37-40	93,3	55
41-44	64,6	64
45-48	41,4	16

Tabla 8.8. ***Registros del Índice de Contaminación Ambiental y de la Incidencia de Asmáticos Agudos (acumulados) a lo largo de 6 segmentos de 8 días***

Grupos de 8 días	ICA acumulado	IAA acumulado
1-8	100,3	63
9-16	140,4	90
17-24	47,5	46
25-32	63,1	35
33-40	119,9	75
41-48	106,0	80

cuenta con un conjunto inicial de datos suficientemente grande y procedentes de unidades de análisis modificables mediante agregación [8] es posible obtener el coeficiente de correlación que virtualmente se desee, siempre que hagamos la manipulación «debida».

Tal artificio podría conseguirse, por ejemplo, trabajando con un registro de tasas de mortalidad y porcentajes de analfabetismo correspondientes a varias decenas de municipios de un país: la correlación entre estas variables crecerá si se opera con las provincias, que no son más que compactación de municipios, y decrecerá si se consideran regiones menores en que se dividan por los municipios.

[8] Nótese que, dentro de un marco racional, esta manipulación no es posible en caso de que las unidades de análisis con que se trabaje sean, por ejemplo, mujeres embarazadas a las que se mide edad y hemoglobina; la suma (o promedio) de edades o concentraciones de hemoglobina procedentes de dos o más embarazadas carece de todo sentido físico.

8.6. Sesgo inducido por la baja calidad de los datos

Es bien conocida por los investigadores la importancia que se otorga a la calidad de los datos primarios. Siendo elementos de entrada *(input)* de un sistema, se cumple el adagio de los informáticos conocido familiarmente como *GIGO (Garbage In, Garbuge Out): es* imposible obtener otra cosa que basura si lo que se procesa es basura.

Sin embargo, es también obvio que los datos obtenidos empíricamente *siempre* tendrán algún grado de imprecisión o error. Ello no los convierte necesariamente en basura, como ocurre en el patético caso en que se mide algo diferente de lo que se quiere. Con los errores de medición, lo que se presupone con frecuencia es que, al ser de índole aleatoria (no sistemáticos, no inclinados o sesgados en cierta dirección), entonces los resultados que de ellos se desprendan tampoco estarán sesgados en un sentido específico. Por ejemplo, si se mide sucesivamente la longitud de un objeto con un instrumento válido, las observaciones no serán todas iguales debido al error inevitable que se comete; pero su promedio «tenderá» a coincidir con la verdadera dimensión del objeto siempre que aumente el número de mediciones.

Sin embargo, esto no es necesariamente válido para todo tipo de evaluación. El prolífico epidemiólogo Richard Peto, connotado impulsor del moderno metaanálisis, ha llamado la atención sobre el impacto que los errores presentes en el proceso de medición de cierta variable pueden tener sobre la evaluación *de sus efectos.*

Palca (1990) propone en la prestigiosa revista *Science* un modelo teórico muy simplificado con el que se ilustra transparentemente el razonamiento de Peto. Lo que sigue es una adaptación de dicho modelo.

Imaginemos que en cierta comunidad hay sólo tres tipos de personas: las que tienen colesterol *bajo,* las de colesterol *medio y* las que lo tienen *alto; y* que aproximadamente un tercio de las personas pertenece a cada grupo. Más específicamente, en esta comunidad los valores posibles para dicha variable son exactamente 220,230 y 240 mg/dl, que corresponden, respectivamente, a los integrantes de cada uno de los grupos mencionados.

Admitamos, además, que la tasa de incidencia de mortalidad por infarto de miocardio *(IM)* entre los 40 y los 50 años de edad es de 4, 6 y 8 (por 1000 habitantes) dentro de los respectivos grupos. Quiere esto decir que, por cada 10 mg/dl, el riesgo de muerte por *IM* crece en 2 por 1000.

Un investigador -quien, naturalmente, ignora la información precedente- desea estimar esa relación entre la concentración de colesterol en sangre y la mortalidad, y se propone hacerlo a través de un estudio longitudinal en ese grupo de edad. Concretamente, supongamos que este epidemiólogo obtiene una muestra representativa de 30 mil individuos de 40 años y le mide la concentración de colesterol a cada uno. Diez años después registra cuántos murieron por *IM.*

Partiendo de la *ley de los grandes números,* en la muestra habrá alrededor de 10 mil individuos procedentes de cada grupo, y se producirán aproximadamente 180 muertes (40 en el grupo de nivel *bajo,* 60 en el de nivel *medio y 80* en el de nivel *alto).*

Si no hubiera error en la medición del parámetro sanguíneo, la relación observada entre concentración de colesterol (C) y la tasa de mortalidad (*T*) sería entonces la que se ha descrito:

$$C_1 = 220 \qquad T_1 = 4$$
$$C_2 = 230 \qquad T_2 = 6$$
$$C_3 = 240 \qquad T_3 = 8$$

donde C_1, C_2 y C_3 son las tres concentraciones de colesterol existentes y T_1, T_2 **y** T_3 las tasas de mortalidad respectivas.

Ahora admitamos que la medición del colesterol está sujeta al siguiente error de tipo aleatorio: en cada determinación de este parámetro sanguíneo se comete ***siempre*** una adulteración consistente en que el valor observado es aumentado o disminuido, con probabilidad igual a 0,5 para cada posibilidad, en 10 unidades [9]. De modo que, por ejemplo, para la mitad de los 10 mil sujetos cuya concentración de colesterol es 240, la medición arrojará un valor de 230 en tanto que, para los otros 5.000, se observará un valor de 250.

El valor esperado del error cometido en la medición es nulo: es fácil convencerse de que si con esos datos se quisiera, por ejemplo, conocer el promedio de la concentración de colesterol en la comunidad, este parámetro poblacional sería estimado con toda exactitud; la variabilidad, sin embargo, resultaría obviamente sobreestimada. ¿Qué efecto tendrá esta sobrestimación sobre la apreciación que el investigador, basándose en sus datos, hará acerca del impacto del colesterol sobre la mortalidad? Lo que realmente se habrá de observar es lo que recoge la Tabla 8.9.

Tabla 8.9. ***Resumen de los resultados observados de concentración de colesterol y tasas de mortalidad en la cohorte de 30.000 sujetos***

Nivel observado de colesterol: C_i^*	Tamaño muestral	Sujetos que mueren	Tasa de mortalidad T_i^*
210	5.000	20	4
220	5.000	30	6
230	10.000	60	6
240	5.000	30	6
250	5.000	40	8

[9] Unos *gremlins* que habitan en el laboratorio lanzan una moneda y adulteran el resultado de cada medición verdadera, suman 10 si sale caray restan 10 en caso contrario.

Por ejemplo, el estudio arrojará, para 10 mil sujetos, la observación $C_3^* = 230$; de ellos, 60 morirán, para una tasa observada en ese grupo igual a T_3 = 6. Veamos: la muestra contiene 10 mil individuos con el verdadero colesterol igual a C = 220 y otros 10 mil para los que C = 240; para la mitad de los primeros se registrará C_3^* = 230 y ellos aportarán 20 muertos (4 por cada 1.000); para la mitad de los otros 10 mil se registrará ese mismo valor de colesterol pero, por ser realmente del tercer grupo, aportará 40 muertos (8 por cada 1.000). De modo que, en efecto, habrá en total 10 mil individuos con ese valor de C_3^*, y 60 muertos entre ellos. Análogamente se corroboran los otros cuatro pares de valores.

Si se representan gráficamente estos datos -verdaderos y observados- poniendo la tasa de mortalidad en función del nivel de colesterol, se ajustarían las rectas que se reflejan en la Figura 8.3.

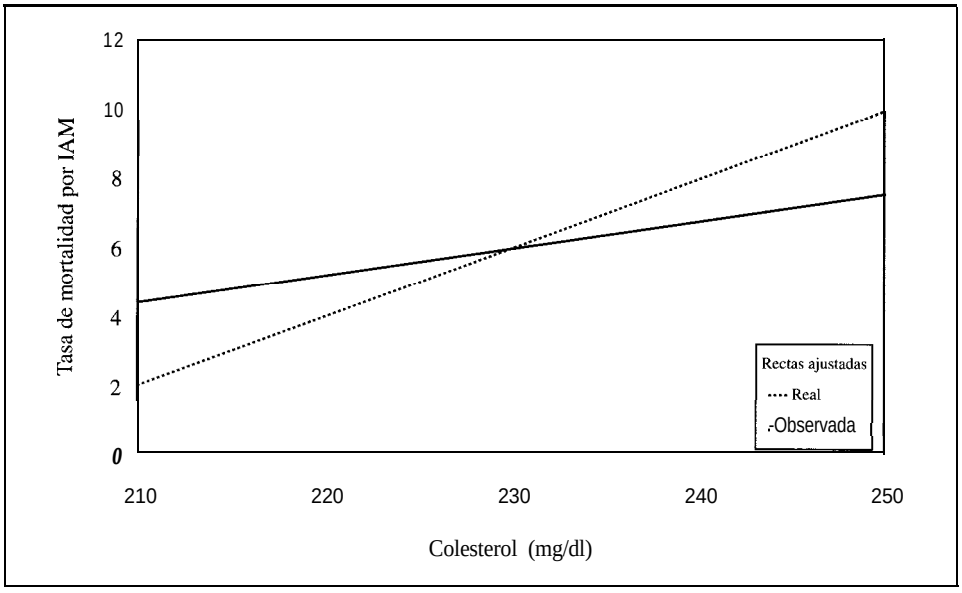


Figura 8.3. ***Rectas ajustadas de mortalidad según concentración de colesterol (real y observada).***

Se observa, en síntesis, un efecto sorprendente: como consecuencia exclusiva de un ***error no sistemático*** de medición, del tipo que regularmente se produce en cualquier determinación de laboratorio, el investigador observará una relación mucho más atenuada de la que realmente se verifica. Si se conociera la magnitud del error de las mediciones, se podrían hacer las correcciones correspondientes y evitar la falacia en que se incurre.

Lo cierto es que, al no contemplarse este efecto de los errores, la literatura ha venido recogiendo impactos menores que los verdaderamente debidos a los factores sindicados como causales de ciertos daños.

8.7. Sesgo de Berkson

A mediados de este siglo, Berkson (1946) llamó la atención por primera vez sobre un sesgo que puede presentarse en los estudios basados en información hospitalaria. Este sesgo aparece cuando la combinación de una enfermedad y un presunto factor de exposición resulta sobrerrepresentada como consecuencia de que las circunstancias que determinan la admisión de los pacientes al centro de atención distorsionan la verdadera relación que se da en la población.

Este fenómeno, conocido como sesgo de *Berkson* en honor a su descubridor, constituye un peligro de cierta importancia, especialmente para los estudios de casos y controles, ya que con mucha frecuencia éstos se basan en información procedente de una fuente institucional. A continuación se ofrece una ilustración del fenómeno.

Supongamos que se sabe que la prevalencia de hemorroides en una población es del 25%. Un proctólogo observa que la interconsulta procedente del servicio de dermatología en su hospital es más frecuente que la requerida por el de oncología; a partir de ahí realiza una indagación más profunda. Examina a los 732 cancerosos recluidos en el hospital e identifica que 192 de ellos padecen de hemorroides.'Esto es: la prevalencia de dicha dolencia entre los cancerosos asciende a 26%, cifra compatible con la prevalencia general en la población. Pero cuando repite este análisis entre los 231 pacientes hospitalizados que padecen problemas dérmicos, halla que 186 de ellos (el 80%) padecen de hemorroides. 0 sea, la tasa de hemorroides entre estos enfermos es mayor que el triple de la que se observa entre los cancerosos. Los resultados se resumen en la Tabla 8.10.

Tabla 8.10. ***Número de sujetos con y sin hemorroides entre cancerosos y pacientes con dolencias dermatológicas en el hospital***

	Dolencias dermatológicas	Cáncer
Con hemorroides	186	192
Sin hemorroides	45	540
Total	231	'732

¿Existirá, como sugieren estos datos, una asociación de índole causal entre los problemas de la piel y las hemorroides? Un examen más detenido del problema permite comprobar que los ***resultados hallados son mero efecto de que las tasas de admisión de pacientes son diferentes para las distintas dolencias.***

Imaginemos que el 90% de los sujetos que padecen de cáncer en la población

son remitidos hacia este hospital, cosa que sin embargo ocurre sólo con el 5% de los que adolecen de problemas de la piel. Por otra parte, supongamos que en ese hospital se trata aproximadamente al 60% de todos los que sufren de hemorroides [10].

Si en la población hay en total 800 sujetos con cáncer y 1.200 con trastornos de piel, es fácil ver que la composición hospitalaria será la de la Tabla 8.10, aunque las hemorroides no tengan asociación alguna ***al nivel de la población*** ni con una ni con otra enfermedad. En efecto, partiendo de este supuesto de independencia, se tiene lo siguiente.

De los 1.200 sujetos con problemas dermatológicos en la comunidad, 300 (la cuarta parte) tendrán hemorroides: de ellos, 186 irán al hospital (el 60% de los 300 — es decir, 180- por tener hemorroides, y el 5% de los 120 que no ingresan por ese concepto -otros 6— que lo hacen por razones dermatológicas). Hay 900 sujetos con dermatitis que no tienen hemorroides: el 5% de ellos (45) son ingresados. Es decir, de 231 individuos con problemas dérmicos, 186 tienen hemorroides, tal y como se recoge en la Tabla.

Ahora, en relación con los 800 enfermos de cáncer, la situación es la siguiente: 200 tendrán hemorroides y 600 no. El 90% de estos últimos (540) estarán en ese hospital por razón de su dolencia oncológica. De los otros 200, 180 ingresan por ser cancerosos y 12 de los otros 20 (el 60%) ingresan por concepto de sus hemorroides, para un total de 192 que tienen cáncer y hemorroides a la vez. En síntesis, de 732 cancerosos hospitalizados, 192 tienen hemorroides, lo que completa la corroboración de que las distribuciones dentro del hospital serán las de la Tabla 8.10.

Otros detalles matemáticos y varios ejemplos adicionales de esta falacia pueden hallarse en los trabajos de Brown (1976), Roberts et al (1978) y Conn et ***al.*** (1979).

8.8. Factores de confusión

8.8.1. La paradoja de Simpson y un problema para el lector

Hace unos años (Silva, 1987) expuse la llamada ***paradoja de Simpson*** en el contexto de unas conferencias sobre métodos estadísticos en epidemiología. En esa oportunidad construí un ejemplo original para ilustrarla. El texto publicado fue exactamente el siguiente:

[10] Tal situación es perfectamente verosímil, ya que depende de circunstancias tales como si existen o no programas de pesquisaje, del hecho de que ciertas enfermedades inclinan más que otras a los enfermos a acudir al hospital, o de la propia política hospitalaria.

En esencia, cuando la estimación del efecto de un factor (F) se distorsiona a raíz de la mezcla con el efecto de otro factor ***F', se*** dice que este último es un factor confusor.

Esta situación ha sido identificada (Rothman, 1976) con la «paradoja» sobre la que Simpson (1951) llamó la atención por primera vez y que pudiera ilustrarse con el siguiente ejemplo:

> Una persona quiere evaluar si existe alguna asociación entre la aparición de sus dolores reumáticos y el hecho de que llueva.
>
> En el mes de junio registra 10 días lluviosos y advierte dolores en 9 de ellos (90%); por otra parte, sintió dolores en sólo 15 de los 20 días no lluviosos (75%).
>
> Repite la experiencia en julio y observa que el dolor sobrevino en 4 de 20 días lluviosos (20%) y sólo en 1 de los ll no lluviosos (9%).
>
> En cada uno de los meses se produce entonces que el dolor aparece más frecuentemente en los días lluviosos que en días que no lo son. Al considerar el bimestre completo, sin embargo, del número total de días de lluvia -treinta- en sólo 13 (43%) se produjo el dolor, en tanto que éste apareció en 16 (52%) de los 31 días secos.

El texto original de Simpson (1951) es un trabajo teórico en que aparece un ejemplo relacionado con juguetes elegidos por un niño, otro basado en la baraja y, finalmente, un tercero sobre supervivencia en los dos sexos. Puesto que allí no se menciona nada sobre meteorología o dolores (ni reumáticos, ni de ningún otro tipo), me llenó de estupor encontrar publicado, unos años después, el párrafo de Rey (1989) que se reproduce textualmente a continuación:

> Simpson (1951) llamó la atención sobre la paradoja de asociar un factor de confusión ***F'*** a la relación factor a enfermedad ***(F*** a E), que ilustra mediante la relación entre los dolores reumáticos y el hecho de llover.

¿Qué deduce el lector al cotejar la publicación de 1987 con la de Rey Calero dos años después?

Pero, al margen del tema de la deshonestidad intelectual (abordado con detalle en la Sección 1.4 y que abarcaba la obligación tanto de no citar trabajos que no se conocen, salvo que se consigne la fuente de la que se tomó como la de citar aquellos de los que se haya tomado una idea), el asunto es delimitar cuál es el razonamiento correcto. Concretamente, en mi ejemplo de la lluvia y el dolor: ¿qué se puede sacar en limpio del estudio de la asociación entre ambas variables? El análisis de cada uno de los meses arroja una asociación positiva; el del bimestre, una negativa: ¿a cuál de las dos alternativas debemos dar crédito? ¿Puede inferirse algo sustancial después de examinar integralmente ambos enfoques?

La aproximación epidemiológica clásica dirá en esencia que la asociación surgi-

da del análisis bimestral es *espuria*, y que el análisis estratificado según meses revela la ***verdadera*** naturaleza de la asociación (en este caso, la que afirma que la frecuencia de dolores es mayor durante los días lluviosos).

Naturalmente, en el ámbito etiológico real, la situación suele ser enormemente compleja (véase Capítulo 7). Pero, en cualquier caso, cabe alertar sobre las trampas semánticas que pueden estar viciando este debate.

Cuando se habla de ***correlación verdadera y*** de ***correlación espuria,*** ¿qué se quiere decir? Puesto que una correlación que esté en este último caso ***existe objetivamente,*** el adjetivo ***espurio*** viene a equivaler a ***no causal;*** de lo contrario, ¿qué puede significar ***verdadera*** para calificar a la asociación que no es ***espuria?***

Cuando se alerta machaconamente -aunque ello es enteramente legítimo- sobre el hecho de que ***asociación no implica causación,*** se está invocando la necesidad de reexaminar esa asociación después de controlar posibles efectos confusores. Lo que no está nada claro en el presente estado de la investigación epidemiológica no experimental es cuándo deja de ser espuria una asociación por el hecho de que tal asociación subsista (se invierta su signo o no), ni cuándo deja de ser verdadera porque se desvanezca cuando se controlan ciertos factores que eran posibles confusores. La siguiente sección se destina a profundizar en este importante problema.

8.8.2. La Sinfónica de Londres no interpreta música salsa

Consideremos nuevamente el problema de los dolores reumáticos y la lluvia pero con un monto de información suficientemente amplio como para permitir un examen más detallado. Imaginemos ahora que otra persona quiere evaluar a comienzos de 1993 si en su caso existe tal asociación entre lluvia y dolor. Puesto que tenía un registro de los hechos relevantes para cada día del cuatrienio 1989 - 1992, realiza el siguiente análisis.

Primero contempla los 4 años completos (1.461 días) y constata que hubo un total de 719 días de lluvia y que sólo en 314 (44%) de ellos experimentó dolor. Éste se produjo en 404 (54%) de los restantes 742 días (no lluviosos). La situación se resume en la Tabla 8.11:

Tabla 8.11. ***Distribución de días del cuatrienio 1989-l 992 según lloviera o no y según aparición de dolores reumáticos***

	Dolor	No dolor	Total
Días lluviosos	314	405	719
Días secos	404	338	742

Descrita la situación en términos del riesgo relativo de experimentar dolor bajo el *factor lluvia,* tendríamos que RR = 0,80, resultado de dividir el porcentaje de días con dolor bajo condiciones de lluvia entre el que se obtiene bajo condiciones secas. El intervalo de confianza (95% de confiabilidad) es: [0,72 – 0,891. Los que gustan de las pruebas formales de significación pueden comprobar que la asociación que indica que a más lluvia menos dolor es claramente significativa ($\chi^2 = 16{,}97$, $p = 0{,}00004$). Se decide ahora examinar el problema por bienios y se encuentra la situación que se resume en la Tabla 8.12.

Tabla 8.12. ***Distribución de días de cada uno de los dos bienios del periodo 1989-1992 según lloviera o no y según aparición de dolores reumáticos***

1989-1990	Dolor	No dolor	Total
Días lluviosos	235	18	253
Días secos	382	95	477

1991-1992	Dolor	No dolor	Total
Días lluviosos	79	387	466
Días secos	22	243	265

Es fácil constatar que, para ambos bienios, las tasas de dolor son mayores dentro de los días de lluvia que en los días secos: respectivamente 93% (235 de 253) y 80% (382 de 477) para el primero de ellos, y 17% (79 de 466) contra sólo 11% (22 de 265) en el segundo bienio. Puesto que el patrón se repite en cada caso, de acuerdo con la práctica habitual, este señor daría por válido que hay una asociación positiva: la asociación negativa obtenida inicialmente, para el cuatrienio completo, era «espuria».

Si se examina el problema según los cánones clásicos de análisis, se obtiene que para el primer bienio el riesgo relativo inherente a la lluvia es:

$$RR, = 1{,}16\ (1{,}10 - 1{,}23 \quad \text{con} \quad \chi^2 = 20{,}71\ (p < 0{,}0001);$$

para el segundo bienio, este riesgo asciende a:

$$RR_2 = 2{,}04\ (1{,}30 - 3{,}20) \text{ con } \chi^2 = 10{,}62\ (p = 0{,}0001).$$

La estimación del riesgo relativo conjunto (ponderado) según el método de Mantel-Haenszel es: $RR_{MH} = 1{,}24$ con un intervalo de confianza igual a: [1,16 – 1,34];

el valor de ji cuadrado resumen es $\chi^2_{MH} = 29{,}8$ ($p < 0{,}0001$), todo lo cual según la ***teoría oficial*** «convalida estadísticamente» la conclusión arriba mencionada.

Vayamos ahora más lejos e imaginemos que en ese punto al investigador se le ocurre volver a hacer el análisis, pero trabajando con los cuatro años separadamente. De modo que desdobla cada una de los dos cruzamientos bienales y obtiene los resultados que recoge la Tabla 8.13.

Tabla 8.13. ***Distribución de días de cada uno de los cuatro años del lapso de 1989-1992 según lloviera o no y según aparición de dolores reumáticos***

1989	Dolor	No dolor	Total
Días lluviosos	229	13	242
Días secos	117	6	123

1990	Dolor	No dolor	Total
Días lluviosos	6	5	11
Días secos	265	89	354

1991	Dolor	No dolor	Total
Días lluviosos	74	289	363
Días secos	1	1	2

1992	Dolor	No dolor	Total
Días lluviosos	5	98	103
Días secos	21	242	263

Lejos de confirmarse la última conclusión, este nuevo examen -pensado para eliminar efectos confusores- viene a añadir considerable confusión: la nueva situación se resume en la Tabla 8.14.

Según estos datos, la situación ¡vuelve a invertirse! Como cuando se analiza el cuatrienio completo, los días lluviosos tienen menor tasa de dolor que los días secos para todos y cada uno de los años.

Tabla 8.14. ***Porcentajes de días con dolor según hubiese llovido o no y riesgos relativos para cada uno de los cuatro años***

	Días lluviosos	Días sin lluvia	Riesgo relativo	Intervalo de confianza
1989	94,6	95,1	0,99	0,95-1,05
1990	54,4	79,9	0,73	0,42-1,25
1991	20,4	50,0	0,41	0,10-1,65
1992	4,9	8,0	0,61	0,24-1,57

Sin embargo, en ninguno de los cuatro casos es posible descartar el azar como explicación de las diferencias entre estas tasas (los cuatro intervalos de confianza contienen al 1).

La estimación del riesgo relativo de Mantel-Haenszel es RR_{HM} = 0,94 con un intervalo de confianza de 0,87 – 1,02. La conclusión resumida sería que no hay indicio alguno de que exista relación entre lluvia y dolor.

Cabe preguntarse: ¿cómo explicar esta paradoja? Ocurre que no tiene «explicación». A mi juicio, ninguna paradoja la tiene: la realidad es como es; si nos resulta paradójica, es en virtud de un prejuicio o un error radicado ***en nuestra cabeza.*** Es el hecho de por qué convivimos con tal error el que ocasionalmente puede ser explicado.

En este caso opera una huella mental debida a un hecho cierto. Si todos los miembros de cierto conjunto tienen un rasgo y todos los de un segundo conjunto también lo tienen, entonces los integrantes del conjunto resultante de la unión de los anteriores necesariamente tendrán el rasgo en cuestión. Pero esto no es válido cuando el rasgo concierne a los conjuntos como tales.

Por ejemplo, todos los ejecutantes de la Orquesta Sinfónica de Londres son talentosos y también lo son todos los músicos de la orquesta de salsa «Tempestad Latina»; de manera que todos y cada uno de los integrantes de una supuesta orquesta combinada de ambas agrupaciones musicales tendrán talento. Por otra parte la orquesta sinfónica puede sonar excelentemente y la de salsa también; pero lo que se deje oír de la unión de ambas pudiera no ser agradable para quien las disfrutaba separadamente.

El ejemplo de los dolores reumáticos, en resumen, ilustra con crudeza el hecho de que el recurso de la postestratificación para evaluar el efecto de factores confusores puede en muchos casos no pasar de ser un mero artificio numerológico: un olmo cuyas bellas hojas están siendo contempladas en calidad de peras.

Desde luego, actualmente, con el advenimiento de la regresión logística y la factibilidad computacional adjunta, la postestratificación está en franco desuso. La ocasional esterilidad del control de factores resulta en este caso menos obvia, y

muchos usuarios quedan enceguecidos por las luces deslumbrantes del análisis multivariado.

Pero la situación es exactamente la misma. De hecho, el problema considerado puede examinarse con dicha técnica: tomando los días como unidades muestrales que son «dolorosos» o «no dolorosos» (variable de respuesta), podrá corroborarse que el coeficiente correspondiente a la variable independiente (lluvia) es negativo si no se controla ninguna otra variable, positivo si se controla el bienio y virtualmente nulo si se controlan los años.

Una de las moralejas que cabe extraer de este ejemplo es que el control de variables no debe practicarse si no es para evaluar hipótesis claramente expuestas y fundamentadas con anticipación. Sin embargo, aunque resulte descorazonador, ha de quedar claro que esta condición es necesaria pero no suficiente para que las conclusiones además de inteligibles, sean correctas.

8.8.3. La autopsia de una hipótesis

¿Es el ejemplo de la sección anterior el resultado de una hábil construcción numérica para conseguir una evidencia artificiosa, que raramente se producirá en la práctica, de los peligros inherentes al famoso control de variables con recursos estadísticos en la fase de análisis?

En cierto sentido sí lo es, pues el ejemplo es algo alambicado (y confieso que me costó cierto trabajo armarlo). Pero basta con que se ***pueda*** construir un ejemplo de este tipo para generar dudas sobre la validez del método en general, ya que en la práctica nunca estaremos seguros de que no estemos padeciendo el mismo problema. Ocurre como con la clínica y la anatomía patológica: se ha dicho que los clínicos ***adivinan y*** los patólogos ***saben.*** En muchas ocasiones sólo una autopsia permitirá confirmar o refutar un diagnóstico. Veamos qué ocurrió cuando se practicó la autopsia de una hipótesis de causalidad manejada a nivel «clínico» a través de la regresión logística.

En un elocuente trabajo, Vanderbroucke y Parodel (1989) examinan el curso que siguió la investigación sobre las causas del SIDA en la etapa inicial de la epidemia. Entre otras posibles causas, se especulaba sobre el posible efecto de los «poppers», nombre coloquial con el que se designaba a un nitrito usado como estimulante sexual dentro de la comunidad homosexual en Estados Unidos.

Los «poppers» aparecen en el horizonte etiológico a partir de la búsqueda exhaustiva de algún «nuevo factor» para una «nueva enfermedad»; se constató (véase McManus, Starret y Harris, 1982) que su consumo era intenso entre las víctimas del SIDA. Un estudio de casos y controles publicado por Mormor ***et al.*** (1982) en la prestigiosa revista ***Lancet*** arroja una odds ratio de 8,6 para los consumidores intensos de la sustancia respecto de los que no lo son.

Sin embargo, se sospechaba que la promiscuidad sexual pudiera estar desem-

peñando un papel confusor; el estudio mencionado permite calcular una odds ratio de 4,9 para aquellos que habían tenido más de 10 compañeros sexuales por mes en el año anterior a la aparición del síndrome. De modo que se examinó la asociación entre el uso de los «poppers» y el SIDA ***controlando el efecto de la promiscuidad.*** El análisis a través de la regresión logística arrojó que la odds ratio correspondiente al presunto agente causal se elevaba a 12,3 cuando se controlaba la promiscuidad, en tanto que la de esta última se reducía a 2,0 cuando se ajustaba el efecto confusor del estimulante sexual. A partir de esta «evidencia» se llegaron a publicar trabajos que ofrecían plausibilidad biológica a la hipótesis y bosquejaban un posible mecanismo bioquímico según el cual podría estar actuando la sustancia.

A principios de 1983 se produjo el célebre (y polémico por otras razones) aislamiento del virus de inmunodeficiencia humana, con lo cual la efímera e incorrecta, pero instructiva hipótesis acusatoria de los «poppers» hizo mutis definitivo de la literatura sobre el tema y pasó al olvido. Lamentablemente, los recursos de la epidemiología observacional, lejos de contribuir a esclarecer el problema, condujeron a los investigadores por trillos erróneos; sólo un hallazgo procedente de las ciencias básicas puso las cosas en su lugar.

Bibliografía

Berkson J (1946). ***Limitations of the application of fourfold table analysis to hospital data.*** Biometrics Bulletin 2: 47-53.

Baruch RF, Cecil JS (1979). ***Assuring the confidentiality of social research data.*** University of Pennsylvania Press, Pennsylvania.

Brown GW (1976). ***Berkson fallacy revisited.*** American Journal of Diseases of Children 130: 56-62.

Brown GH, Harding FD (1973). ***A comparison of methods of studying illicit drug usage.*** HUMRO Technical Report 73, Arlinton.

Carvajal A, García JL, Holgado E, Velasco A (1984). ***Consumo de drogas en una muestra de médicos rurales de Valladolid,*** Medicina Clínica 83: 444-446.

Conn HO et al (1979). ***The Berkson bias in action.*** Yale Journal of Biological Medicine 2:141-147.

Dalenius T, Vitale RA (1974). ***A new randomized response design for estimating the mean of a distribution.*** Report No. 78 of the Errors in Surveys Research Project University of Stockholm (mimeo).

Itzhaki J (1995). El ***estilo de El Greco.*** Periódico ***El Mundo, 28*** de septiembre, Sección Salud: 3, Madrid.

Krotki KJ, McDaniel SA (1975). ***Three estimates of illegal abortion in Alberta, Canada: survey mail-back questionnaire and randomized response technique.Trabajo*** presentado en la 40.ª Sesión del ***Intemational Statistical Institute,*** Varsovia.

Lamb CW, Stem EE (1978). ***An empirical validation of the randomized response technique.*** Journal of Marketing Research 15: 616-621.

Massey JT, Ezzati TM, Folsom R (1989). ***Survey methodology requirements to determine the feasibility of the national household seroprevalence survey.*** Quality Assessment Task Force Report, NCHS.

McManus TJ, Starret LA, Harris JR (1982). ***Amyl nitrito use by homosexuals.*** (letter), Lancet 1: 503.

Medawar P (1967). ***Consejos a un joven científico.*** Fondo de Cultura Económica, México DE

Mills JL (1990). ***Data torturing. New*** England Journal of Medicine 329: 1196-1199.

Mormor M ***et al,*** (1982). ***Risk factors for Kaposi's sarcoma in homosexual men.*** Lancet 1: 1083-1086.

Palca J (1990). ***Getting to the heart of the cholesterol debate.*** Science 247:1170-1171.

Pérez R (1980). ***Serendipia.*** Siglo XXI, México DE

Rey J (1989). ***Método epidemiológico y salud de la comunidad.*** Interamericana McGraw Hill de España, Madrid.

Roberts RS ***et al.*** (1978). ***An empirical demonstration of Berkson's bias.*** Journal of Chronic Diseases 31: 119-128.

Robinson WS (1950). ***Ecological correlations and the behavior of individuals.*** American Sociological Review 15: 531-537.

Silva LC (1982). ***La conjidencialidad: Un desafio aleatorio.*** Revista Juventud Técnica 176: 33-35.

Silva LC (1984). ***La técnica de respuesta aleatorizada: un método para la reducción de conductas evasivas en las encuestas depoblación.*** Revista Cubana de Administración de Salud 10: 53-59

Silva LC (1987). ***Métodos estadísticos para la investigación epidemiológica.*** Instituto Vasco de Estadística, Cuaderno 14, Bilbao.

Simpson EH (1951). ***The interpretation of interaction in contingency tables.*** Journal of the Roya1 Statistical Society, Series B 13: 238-241.

Tracy PE, Fox JA (1981). ***The validity of randomized response for sensitive measurements.*** American Sociological Review 46: 187-200.

Vanderbroucke JP, Parodel UPAM (1989). ***An autopsy of epidemiologic methods: the case of «poppers» in the early epidemic of the acquired immunodefciency syndrome (AIDS).*** American Journal of Epidemiology 129: 455-457.

Warner S (1965). ***Randomized response: a survey technique for eliminating evasive answer bias.*** Journal of the American Statistical Association 60: 63-69.

WHO (1983). ***Guidelines on studies in environmental epidemiology.*** Environmental Health Criteria 27, Geneva.

Zdep SM, Rhodes IN (1971). ***Making the randomized response technique work.*** Public Opinion Quarterly 41: 531-537.

9

El mundo de lo sensible y lo específico

Cuando la intensidad de la intervención biomédica traspasa un umbral crítico, la yatrogénesis clínica se convierte de error; accidente o culpa en una perversión incurable de la práctica médica.

IVAN ILLICH

Una de las tareas típicas de la medicina como práctica hondamente implicada en el marco cultural, social y económico de la sociedad, es la de asignar «etiquetas» que caractericen a los enfermos y le confieran formal y oficialmente la condición de tales. Esto explica en parte el interés que siempre ha despertado la clasificación binaria enfermo-sano.

Las pruebas diagnósticas que arrojan resultandos de este tipo tienen, ciertamente, una clara función para la acción terapéutica, aunque generalmente insertadas en procesos diagnósticos escalonados o como punto de partida de las decisiones, tal y como ocurre cuando se quiere conocer si un sujeto porta o no el virus del SIDA, o determinar si cierto agente bacteriano está o no presente en un proceso infeccioso.

Pero tal dicotomía (enfermo-sano) desempeña también un papel instrumental en el ámbito social y jurídico; Skrabanek y McCormick (1989) señalan:

> ***Un beneficio de la etiqueta es la legitimación del papel de enfermo. Fue un norteamericano, Talcott Parsons, quien llamó la*** *atención* ***por vezprimera sobre el hecho de que en nuestra sociedad sólo hay una manera de quedar exonerados de cumplir nuestras obligaciones: obtener la condición de enfermo. Sólo de ese modo podemos escapar del trabajo, de ir a la escuela, de lavar los platos, ir a una fiesta o sostener relaciones sexuales.***

Es también en este sentido en el que las pruebas diagnósticas de naturaleza binaria cobran protagonismo. En la Sección 5.3 se desarrollaron algunas ideas de índole conceptual sobre las fronteras para fijar esa dicotomía. En las secciones que

siguen se analizan algunos de los problemas estadísticos asociados con la evaluación y el uso de ese tipo de pruebas diagnósticas.

9.1. Las pruebas diagnósticas y sus parámetros

Mucho se ha escrito en torno a las pruebas diagnósticas y a su eficacia como instrumentos para la correcta clasificación nosológica de un paciente bajo análisis. Una revisión exhaustiva de la teoría desarrollada al respecto hasta mediados de los años 70 se encuentra en el libro de Galen y Gambino (1975). En Silva (1987) puede hallarse un panorama algo más actualizado del tema.

La calidad del proceso diagnóstico resulta importante tanto a nivel individual como poblacional. En el primer caso se trata, como se ha dicho, de identificar si un sujeto está afectado o no por cierta dolencia, de lo cual puede derivarse, al menos teóricamente, la determinación de una u otra conducta terapéutica. El problema se presenta a nivel poblacional cuando se quieren estimar parámetros de morbilidad, información clave para tareas tales como la planificación de recursos por parte del sistema de salud, para el conocimiento del estado de salud de la población y para el diseño eficiente tanto de programas como de estudios epidemiológicos. Tal es el caso, por ejemplo, cuando se intenta establecer la prevalencia de una enfermedad en un conglomerado humano.

Desde el punto de vista operacional, la distinción básica entre ambas situaciones estriba en que, si bien a nivel individual el interés se concentra en que el diagnóstico de cada sujeto sea correcto, a nivel poblacional lo que se desea es minimizar - y de ser posible, evitar- el error con que pudiera hacerse la estimación de la situación de salud o sus condicionantes.

La ***sensibilidad y especificidad*** de una prueba diagnóstica, que se definen más adelante, constituyen los dos indicadores clásicos para evaluar su capacidad demarcatoria. Aun siendo conceptos relativamente sencillos, suelen producirse confusiones, tanto con su interpretación y alcance, como con la relación que guardan con otros parámetros de interés. Tales problemas se observan a menudo entre aquellos profesionales supuestamente llamados a usarlas a diario en su trabajo. En efecto, Schecheter y Sheps (1985) señalan:

> ***Desafortunadamente algunos, a partir de su falta de comprensión, tienden a concluir que esos conceptos son esotéricos y clínicamente irrelevantes. Nada podría estar más alejado de la verdad.***

Es claro que algunas entidades nosológicas pueden ser padecidas de acuerdo a una gradación; por ejemplo, un hipertenso puede sufrir su trastorno con una u otra intensidad, dependiendo de cuán altas sean sus tensiones diastólica y sistólica y de la combinación en que éstas se presenten. Sin embargo, en muchos casos la tarea de

interés primario es, como se ha dicho, delimitar si el sujeto tiene o no una enfermedad especificada.

Para muchas entidades existen nítidas fronteras teóricas inequívocas que permiten separar esas dos categorías. Por ejemplo, el significado de que un sujeto tenga fractura del cráneo o sea portador del VIH -por su propia naturaleza- no se presta a equívocos, independientemente de que un medio concreto para establecerlo pueda dar lugar a errores. En otros casos, como ocurre con la esquizofrenia, tanto el proceso diagnóstico como la propia definición teórica son borrosos y, por tanto, polémicos.

Un elemento metodológicamente cardinal es, en cualquier caso, distinguir entre el criterio que define teóricamente la posesión del estado patológico y *los procedimientos* susceptibles de ser usados para evaluar si dicho criterio se cumple.

En lo sucesivo aceptaremos el supuesto de que, idealmente, tal condición discriminadora acerca de si un sujeto posee o no cierta condición patológica está perfectamente definida. Un individuo concreto, por tanto, cumple o no cumple con esa condición, circunstancias que se denotarán por E *y* $\bar{E}$ respectivamente. Simultáneamente, supondremos que existe una prueba que, aplicada sobre cierto sujeto, puede dar lugar a sólo uno de dos resultados posibles: $T+$ y $T-$, símbolos con que se representan respectivamente los resultados positivo y negativo de la prueba.

En principio, el resultado $T+$ obtenido para un sujeto dado constituye un *indicio* de que éste tiene la condición E *(es* decir, de que está enfermo) y el resultado $T-$ *induciría* a pensar que el sujeto en cuestión tiene la condición complementaria $\bar{E}$ (o sea, que no tiene la enfermedad). Pero el grado de eficiencia inherente a una prueba diagnóstica se resume en los dos parámetros conocidos como *sensibilidad y especificidad,* que parten de la condición morbosa y no del resultado de la prueba.

El primero mide la capacidad de la prueba para detectar a un sujeto enfermo; expresa cuán «sensible» es ella a la presencia de la enfermedad y viene definido por la probabilidad condicional siguiente:

$$\alpha = P(T + \mid E)$$

0 sea, la sensibilidad es la probabilidad de que la prueba identifique como enfermo a aquel que realmente lo es.

El otro parámetro mide la capacidad que tiene la prueba de diagnosticar como sanos a los que efectivamente lo son. Se define como la probabilidad condicional:

$$\beta = P(T + \mid \bar{E})$$

Es decir, mide cuán específica es la prueba diagnóstica en el sentido siguiente: cuanto mayor sea β, menor será su complemento $P(T + \mid \bar{E})$; *o* sea, menor es la probabilidad de que declare como enfermos a sujetos que no sufren esta enfermedad, ya sea porque están sanos o porque realmente padecen otra dolencia.

Los parámetros α y β, sugeridos originalmente por Yerushalmy (1947), sintetizan la calidad intrínseca de una prueba y constituyen la materia prima, por ejemplo, para comparar dos o más pruebas que compiten entre sí. Una temprana medida conjunta de la eficiencia de una prueba fue propuesta poco después. Se trata del Índice de Youden (1950):

$$I_y = \alpha + \beta - 1$$

En condiciones mínimamente razonables I_y se mueve entre 0 y 1.

El caso en que $\alpha + \beta = 2$ ($I_y = 1$) se alcanza sólo cuando la prueba diagnóstica es óptima ($\alpha = \beta = 1$). El caso en que $\alpha = \beta = 0{,}5$ (cuando la realización de la prueba equivale a pronunciarse a través de un procedimiento ajeno al conocimiento y la razón: el lanzamiento de una moneda) produce $I_y = 0$. Un test para el cual I_y sea negativo, ya correspondería a una construcción diagnóstica perversa.

9.2. Estimación de α y β: ¿un círculo vicioso?

Para estimar los valores α y β que le corresponden a una prueba es imprescindible contar con un ***criterio de verdad,*** también llamado ***prueba de oro*** [1]; ***es*** decir, un método susceptible de ser aplicado en la práctica (al menos en una experiencia ***ad hoc)*** tal que la corrección de sus resultados no ofrezca duda alguna.

En posesión de tal recurso, pueden seguirse dos diseños básicos. En primer lugar, el de tomar una muestra representativa de la población, clasificar a cada sujeto como ***E o*** como $\bar{E}$, y luego aplicar la prueba ***Ta*** todos ellos. La otra variante consiste en obtener dos muestras: una de verdaderos enfermos y otra de sujetos comprobadamente sanos; someter entonces a cada sujeto de ambas muestras a la prueba diagnóstica. En cualquiera de los dos casos se obtendría el resultado que se resume en la Tabla 9.1.

Tabla 9.1. ***Esquema que reume el resultado cruzado de aplicar una prueba de oro y una prueba diagnóstica en estudio***

	Prueba de oro		
Prueba en estudio	E	$\bar{E}$	Total
T+	a	b	a + b
T-	c	d	c + d
Total	a + c	b + d	n

[1] En inglés: ***gold standard.***

A partir de esos datos se pueden estimar los indicadores en cuestión, la ***sensibilidad*** y la ***especificidad*** respectivamente, mediante las fracciones siguientes:

$$A = \frac{a}{a + c} \qquad B = \frac{d}{b + d}$$

Las fórmulas para computar intervalos de confianza para estos parámetros pueden hallarse en Diamond (1989).

Una pregunta clave es: ¿cuándo tiene sentido valorar por esta vía [2] una nueva prueba diagnóstica? En principio podría decirse que *nunca,* ya que para valorarla es ineludible contar con una prueba mejor (más eficiente) que sirva como estándar o referencia y, si tal prueba existiera, ¿para qué validar una que por definición es peor?

Se trata sin embargo de un círculo vicioso sólo aparente, pues hay al menos cuatro posibles circunstancias que justifican tal empeño:

a) Cuando la prueba novedosa es más económica que la de oro.
b) Si la prueba de referencia es peligrosa para la salud del paciente y la nueva lo sea menos, 0 sea inocua.
c) En caso de que, no teniendo ninguno de los dos rasgos precedentes, sea impracticable en condiciones regulares (por ejemplo, si la prueba de oro procede de una autopsia).
d) Siempre que la corroboración a través de la prueba de oro sólo se pueda conocer a partir de la evolución del paciente, en cuyo caso la prueba en estudio se convierte de facto en un ejercicio de predicción más tarde sancionado por la experiencia.

La otra pregunta relevante en este punto es la siguiente: ¿existe siempre una genuina prueba de oro? Lo cierto es que para muchas situaciones y dolencias no existe nada que se pueda llamar así en propiedad; tal es el caso de una prueba para diagnosticar la esquizofrenia paranoide, y siempre se padecerá esa carencia cuando la condición corresponda a una construcción teórica novedosa como ocurriría si se quisiera «diagnosticar» si una familia es o no «funcional».

En muchas otras ocasiones, lo que se toma en calidad de criterio de verdad dista de ser un procedimiento de clasificación perfecto; simplemente, se usa una prueba imperfecta pero cuya eficiencia es suficientemente alta como para admitirla en calidad de criterio de referencia. Por ejemplo, el resultado de un examen histopatológico para diagnosticar si un tumor de mama es o no maligno puede producir cla-

[2] Nótese que la valoración de una prueba puede decursar por vías ajenas a este enfoque estadístico. Por ejemplo, se puede admitir, por razones meramente teóricas y sin necesidad de contrastación alguna, que la tomografía axial computarizada es más eficiente que la radiología convencional para diagnosticar ciertas dolencias.

sificaciones erróneas en uno y otro sentido, pero siempre será más eficiente que una mamografía; siendo así, aquel puede usarse como un estándar para evaluar la mamografía. El uso de estándares imperfectos para realizar la evaluación de pruebas novedosas genera sus propios conflictos y ha sido objeto de alguna atención (véase, por ejemplo, Begg y Metz, 1990).

La literatura médica recoge una enorme cantidad de pruebas diagnósticas usadas en la práctica clínica y epidemiológica. Para que se tenga una idea de la amplia gama de situaciones que se presenta en relación con la eficiencia de las pruebas, la Tabla 9.2 recoge una muestra de procedimientos diagnósticos con sus respectivas medidas de eficiencia.

Tabla 9.2. *Valores de sensibilidad para diversas pruebas diagnósticas identificadas en la literatura reciente*

Prueba diagnóstica y fuente bibliográfica	Rasgo o dolencia que diagnostica	α	β
Técnica serológica EITB. Kaddah *et al.* (1992)	Hidatidosis	100	100
Ensayo inmunoenzimático AgHB. Quiñonez et al. (1992)	Antígenos de superficie de la hepatitis B	100	95
CAGE Magruder, Stevens y Alling (1993)	Dependencia actual del alcohol	100	61
Ultrasonido transvaginal. Kurjal *et* al. (1992)	Neoplasias malignas	96	9.5
VAST. Magruder, Stevens y Alling (1993)	Dependencia genérica del alcohol	95	80
Marcadores tumorales (CEA). Javirsalo et al. (1982)	Cáncer de pulmón	95	17
Test de tolerancia a la glucosa con criterio UGDP Valleron *et al.* (1975)	Diabetes mellitus	91	94
MAST. Magruder, Stevens y Alling (1993)	Dependencia genérica del alcohol	90	82
Técnica serológica ELISA. Kaddah *et al.* (1992)	Hidatidosis	89	89

Tabla 9.2. *Continuación*

Prueba diagnóstica y fuente bibliográfica	Rasgo o dolencia que diagnostica	α	β
Test de detección de antígenos urinarios. McIntosh y Jeffery (1992)	Enfermedades por streptococcus tipo B.	88	98
Test simplificado para identificar desórdenes relacionados con el alcohol. Rost, Burnam y Smith (1993)	Enfermedades relacionadas con el consumo de alcohol.	87	90
Papanicolau. Nesbit y Brack (1956)	Carcinoma de cérvix.	86	91
Papanicolau. Weinberger y Harger (1993)	Infección asintomática por trichomonas vaginales.	86	83
Test simplificado para identificar enfermedades depresivas. Rost, Burnam y Smith (1993)	Enfermedades depresivas	83	91
Electrocardiograma de alta resolución. Makijarvi (1993)	Taquicardia ventricular después del infarto del miocardio.	80	80
Test de aglutinación látex. Quentin *et al.* (1993)	Treptococcus agalactyae en el cuello uterino.	78	98
Test de progesterona. Macía *et al.* (1993)	Actividad endometrial proliferativa.	76	100
Sonografía de vesícula biliar. Duchatel et al. (1993)	Fibrosis en el conducto cístico (prenatal).	75	100
Electrocardiograma bidimensional. Pasquini et al. (1993)	Arterias coronarias intramurales en transposición con las grandes arterias.	75	99
Mamografía. Sienko et al. (1993)	Cáncer de mama.	71	98
DIA. Scieux et al. (1992)	Clostridium dificciles en diarreas.	67	84
Ecocardiograma fetal. Shields et al. (1993)	Enfermedades fetales del corazón.	66	100

Tabla 9.2. *Continuación*

Prueba diagnóstica y fuente bibliográfica	Rasgo o dolencia que diagnostica	α	β
QBC. Mak, Normaznah, Chiang (1992)	Malaria	56	95
Cuestionario para detectar úlcera péptica. Dunn y Cobb (1962)	Úlcera péptica.	50	98
Cuestionario Rose. Cedorlöf, Johnsson y Lundman (1996)	Angina pectoral y bronquitis	44	93

9.3. Curvas ROC: un producto importado

Las pruebas reales que se usan con finalidad diagnóstica suelen arrojar medidas no dicotómicas, aunque tales medidas sean usadas con frecuencia para «dicotomizar». Tal es el caso de innúmeros parámetros que se mueven en un *continuum,* como la bilirrubina, la tensión arterial diastólica, la capacidad vital o el peso para la talla.

Y también es frecuente el uso de pruebas que contemplan un conjunto de alternativas estructuradas ordinalmente y por una de las cuales habrá de pronunciarse quien diagnostique; por ejemplo, pudiera tratarse de optar por una de las siguientes *5* categorías:

a. Inequívocamente normal.
b. Posiblemente normal.
c. Dudoso.
d. Posiblemente anormal.
e. Inequívocamente anormal.

Para la evaluación de la capacidad diagnóstica de pruebas como las mencionadas suelen ser utilizadas las llamadas *curvas ROC (Receiving Operating Characteristic),* generadas en el contexto de las telecomunicaciones a partir de los problemas de recepción de señales, e «importadas» por Lusted (1971) al ámbito médico a través de un breve pero trascendente artículo. A continuación se expone con cierto detalle el manejo de tales curvas.

En cualquiera de los dos casos -escala continua o escala ordinal- se trabaja con los llamadospuntos *de corte (cut off points):* un conjunto de valores que señalan ciertos «hitos» en el recorrido de los valores posibles de la escala.

Consideremos el caso en que se valora la glucosa en sangre de un individuo como factor discriminante entre diabéticos y no diabéticos. Tomando un ejemplo de

Lilienfeld y Lilienfeld (1983), supongamos que se ha realizado una prueba postprandial de dos horas para dosificación de la glicemia a 580 individuos: 70 diabéticos confirmados y 510 no diabéticos.

Al considerar positivo al sujeto cuyo nivel de glucosa en sangre excede los 110 mg/100 cc, y negativo al que exhiba un valor inferior, se obtienen los resultados de la Tabla 9.3.1. Si el punto de corte elegido es 140 mg/100 cc, entonces los resultados son los de la Tabla 9.3.11.

Tabla 9.3. *Clasificación en positivos y negativos de un grupo de 70 diabéticos y 510 no diabéticos a partir de una prueba prostpandreal de dos horas según se alcance o no el nivel 110 mg/100 cc en el primer caso y 140 mg/100 cc en el segundo*

I	Diabéticos	No diabéticos
Más de 110 mg/ 100 cc	65	263
110 mg/100 cc o menos	5	247
Total	70	510

II	Diabéticos	No diabéticos
Más de 140 mg/ 100 cc	58	45
140 mg/100 cc 0 menos	18	46.5
Total	70	510

Las estimaciones de sensibilidad y especificidad para estas situaciones son, respectivamente:

$$A_I = 0{,}93 \qquad B_I = 0{,}48$$
$$A_{II} = 0{,}74 \qquad B_{II} = 0{,}91$$

En el ejemplo se observa lo que típicamente ocurre en tales situaciones: una modificación en el punto de corte que aumente la especificidad, disminuye la sensibilidad y viceversa[3]. Por otra parte, es obvio que para cada punto que se use para discriminar, se tendría una «configuración 2 x 2» como las que se incluyen en la Tabla 9.3.

La Tabla 9.4 reproduce los datos de *A y B* que se ofrecen en Lilienfeld y Lilienfeld (1983) para 13 niveles (puntos de corte) escogidos.

[3] Ocasionalmente puede ocurrir que uno crezca y el otro se mantenga igual; lo que no puede ocurrir es que, de uno a otro punto, ambos parámetros varíen en el mismo sentido.

Tabla 9.4. *Sensibilidad y especificidad para varios niveles de glucosa en sangre aplicados a 70 diabéticos y 510 no diabéticos*

Nivel de glucosa (mg/100 cc)	A	B
80	1,00	0,01
90	0,99	0,07
100	0,97	0,25
110	0,93	0,48
120	0,89	0,68
130	0,81	0,82
140	0,74	0,91
150	0,64	0,96
160	0,56	0,99
170	0,53	1,00
180	0,50	1,00
190	0,44	1,00
200	0,37	1,00

De hecho, se está ante una «familia de pruebas» con la que se puede construir una curva ROC, que represente la capacidad discriminativa de esta prueba a lo largo de todos los «puntos de corte» posibles.

Si nos situamos en un sistema de ejes cartesianos, la curva se construye como sigue: se ubican los distintos valores estimados de la sensibilidad en el eje de ordenadas y en el de abscisas las correspondientes tasas estimadas de falsos positivos $(1 - B)$; los puntos contiguos se unen mediante segmentos de recta, generando así la curva ROC, cuya expresión empírica habrá de ser, naturalmente, un polígono. La Figura 9.1 muestra el aspecto idealizado de una curva ROC. Tal y como se había mencionado, se aprecia que el valor de α crece en la medida que crece $1 - \beta$ (es decir, en la medida que disminuye).

Una situación extrema se produce cuando la curva coincide con la diagonal; éste sería el caso en que para todos los puntos de corte se tuviera $\alpha + \beta = 1$: cuando la prueba nunca ofrece la más mínima información útil. El otro extremo es aquel en que la curva coincide con la paralela al eje de abscisas que pasa por el punto (0,l); es decir, cuando para todos los puntos de corte se tiene $\alpha = 1$ y $1 - \beta = 0$).

En la Figura 9.2 se refleja la curva ROC empírica correspondiente a los datos de la Tabla 9.4

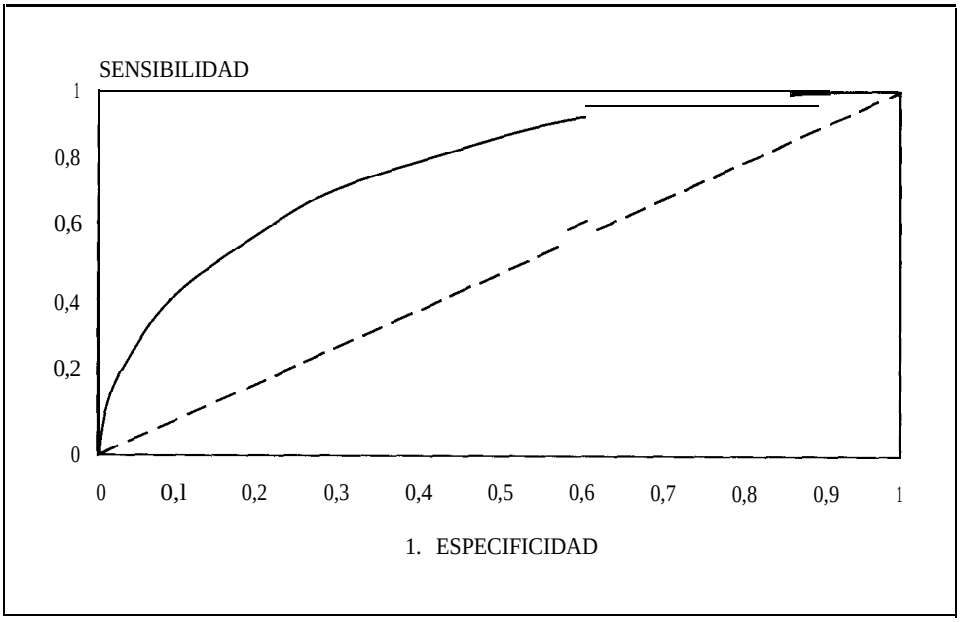


Figura 9.1. ***Curva ROC típica.***

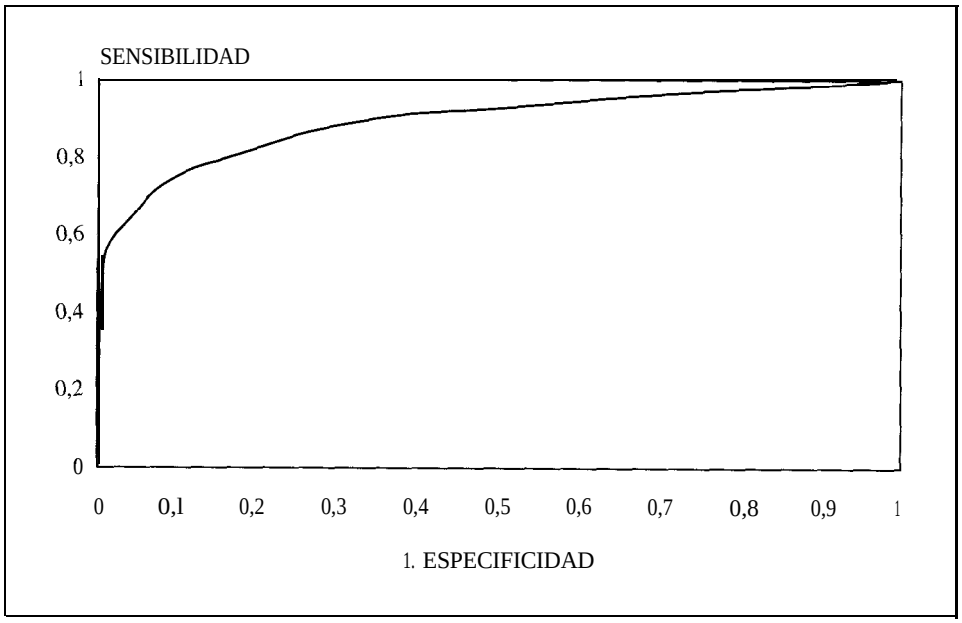


Figura 9.2. ***Curva ROC para la determinación de glucemia.***

El caso en que el clínico se pronuncia a través de una escala ordinal, exhibe algunas singularidades. Consideramos el ejemplo expuesto por Centor (1991) en el que se usa la escala ordinal mencionada al comienzo de esta sección.

Se aplica una técnica radiológica a 120 enfermos de cáncer de páncreas y a 77 sujetos libres de ese mal. Un observador se pronuncia ante las 197 imágenes y genera los resultados que se recogen en la Tabla 9.5.

Tabla 9.5. *Resultados obtenidos al aplicar una prueba diagnóstica en escala ordinal a 120 sanos y 77 portadores de cáncer de páncreas*

	Enfermos E	Sanos $\bar{E}$
1. Inequívocamente normal	2	63
2. Posiblemente normal.	5	38
3. Dudoso.	20	15
4. Posiblemente anormal.	30	3
5. Inequívocamente anormal	20	1
Total	77	120

Llamémosle A_i y B_i respectivamente a las estimaciones de α y β que corresponden a la i-ésima categoría. De acuerdo con lo explicado, la curva ROC se construye ubicando en el plano los puntos $(1 - B_i, A_i)$. En este caso, para hacer tales cómputos, hay que formar las consabidas tablas de 2×2.

Concretamente, han de construirse 4 tablas (número de categorías menos 1). Para comenzar, el criterio de positividad sería el de considerar $T+$ a aquellos que estén en una clasificación superior a la primera (cualquiera de las cuatro categorías entre «posiblemente normal» e «inequívocamente anormal») y T– a los que se consideren «inequívocamente normales». Como resultado puede construirse la Tabla 9.6.

Tabla 9.6. *Resultados de distribuir cancerosos (E) y sanos ($\bar{E}$) según se consideren o no inequívocamente normales (T-) o no (T+)*

	E	$\bar{E}$
T+	75	57
T–	2	63
Total	77	120

De modo que $A_1 = \frac{75}{77} = 0{,}97$ y $B_1 = \frac{63}{120} = 0{,}52$.

Análogamente, se construiría la siguiente tabla, para la cual se declararían positivos a todos aquellos que estén en la escala por encima de la segunda categoría y negativos a los demás. Procediendo análogamente se construyen las restantes dos tablas.

Agregando el caso $A_o = 1$, $B_o = O$ (correspondiente a la situación en que se consideran enfermos todos los individuos y la sensibilidad es, por tanto, máxima) y $A_6 = 0$, $B_6 = l$ (caso en que la especificidad es total como resultado de tipificar como enfermos a todos los individuos), se puede formar la Tabla 9.7.

Tabla 9.7. ***Parejas de valores que determinan los puntos de la curva ROC correspondiente a la Tabla 9.5***

Criterio de positividad	$1 - B_i$	A_i
Todos	1,00	1,00
Mayor que 1	0,48	0,97
Mayor que 2	0,16	0,91
Mayor que 3	0,03	0,65
Mayor que 4	0,01	0,26
Ninguno	0,00	0,00

La Figura 9.3 refleja la curva ROC correspondiente.

Es altamente intuitivo que, cuanto más alejada del eje de abscisas esté la curva que se genera uniendo estos puntos, más eficiente resulta la prueba diagnóstica. Pero el grado de «alejamiento» de la curva al eje de abscisas debe ser formalizado. Simpson y Fitter (1973) demostraron el carácter óptimo del ***área bajo la curva ROC*** como medida de la capacidad diagnóstica global de la variable. La interpretación probabilística de este número -que necesariamente no excede a 1, ya que el área bajo la curva se inscribe dentro de un cuadrado con esa longitud en cada lado- se explica a continuación.

Supongamos que se elige aleatoriamente un sujeto enfermo y que se hace otro tanto con uno sano. Los dos sujetos son presentados a un observador, quien habrá de identificar mediante la prueba diagnóstica cuál de ellos es el que está enfermo. ***El área bajo la curva ROC coincide con la probabilidad de que la identificación sea correcta.***

Obviamente, el área bajo la curva ROC es un recurso útil para comparar dife-

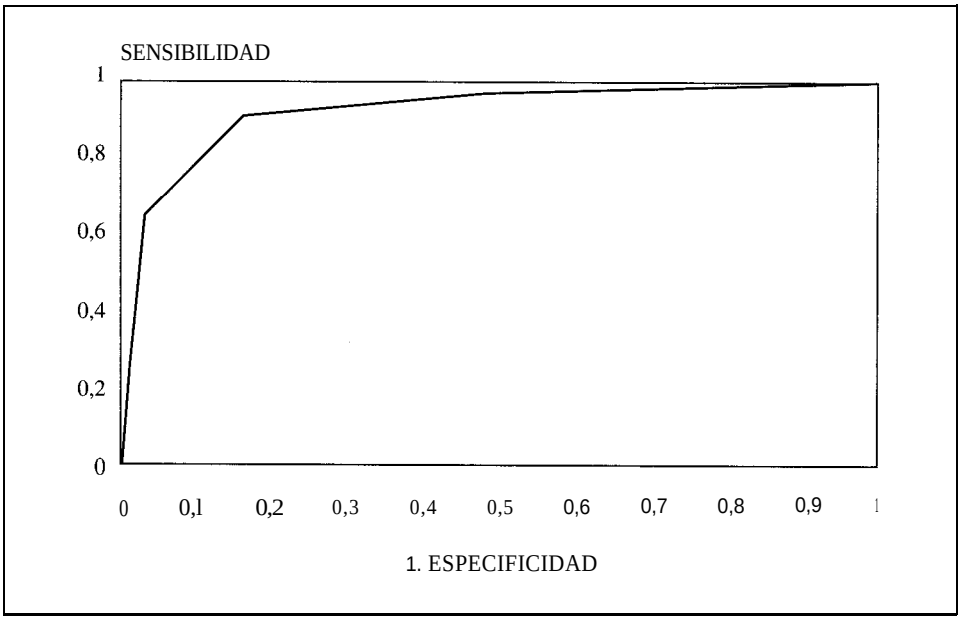


Figura 9.3. ***Curva ROC para la determinación de cáncer de páncreas.***

rentes criterios diagnósticos [4]. Si se valoran 2 variables y la curva ROC que corresponde a una de ellas está por encima de la correspondiente a la otra a lo largo de todo el intervalo (0,l), puede pensarse que, en principio, el primer modelo diagnóstico es más eficiente que el segundo. Si ambas curvas se intersectan, entonces hasta que no se computen las áreas no es posible pronunciarse

Puede demostrarse que si los puntos considerados son tales que $A_m = B_0 = 0$ y $A_0 = B_m = 1$ y los datos se ordenan de manera que la sensibilidad vaya aumentando, el área viene dada por la fórmula general siguiente:

$$A = \frac{1}{2} \sum_{i=0}^{m-1} (B_i - B_{i+1}) (A_i + A_{i+1}) \qquad [9.1]$$

La fórmula [9.1] expresa el área bajo la curva ROC empírica en términos de las estimaciones de sensibilidad y especificidad. Sin embargo, lo más natural es contar con la información directamente en términos de frecuencias de sujetos sanos ($\bar{E}$) y enfermos (E) para cada punto de corte, tal y como aparece en la Tabla 9.5. A conti-

[4] No tiene, por cierto, utilidad alguna a los efectos de diagnosticar a un paciente concreto.
[5] Un enfoque que abarca a las dos situaciones es el de evaluar si un área es ***significativamente*** mayor que la otra.

nuación se expone la solución para calcular tanto el área en cuestión como su error estándar cuando los datos se tienen de esta manera.

Supongamos que hay ***m*** categorías y que llamamos:

a_i: número de sujetos anormales en la categoría i-ésima.
n_i: número de sujetos normales en la categoría i-ésima.

Se definen además:

$a = \sum_{i=1}^{m} a_i$: número de sujetos anormales (con la condición E)

$n = \sum_{i=1}^{m} n_i$: número de sujetos sanos (con la condición $\bar{E}$)

$A_1 = a - \sum_{j=1} a_i$: acumulado de enfermos desde la categoría ***i + 1*** en adelante

$N_i = \sum_{j=1}^{i-1} n_i$: acumulado de sanos hasta la categoría ***i*** - 1.

Se puede entonces demostrar que el área bajo la curva ***ROC*** viene dada por la fórmula [9.2]:

$$\Delta = \frac{1}{n\,a} \sum_{i=1}^{m} \left(n_i A_i + \frac{n_i a_i}{2}\right) \qquad [9.2]$$

y que el error estándar de A puede estimarse mediante [9.3]:

$$\text{se (A)} = \sqrt{\frac{\text{A (1-A)} + (a-1)\,(U-\Delta^2) + (n-1)\,(V-\Delta^2)}{n\,a}} \qquad [9.3]$$

donde:

$$U = \frac{1}{na^2} \sum_{i=1}^{m} n_i \left(A_i^2 + A_i a_i + \frac{a_i^2}{3}\right)$$

$$V = \frac{1}{n^2 a} \sum_{i=1}^{m} a_i \left(N_i^2 + N_i n_i + \frac{n_i^2}{3}\right)$$

Si se aplican estos resultados a los datos de la Tabla 9.5 se tiene que: A = 0,926 y *se* (A) = 0,020.

Los resultados precedentes (fórmulas [9.2] y [9.3]) se han deducido del trabajo de Hanley y McNeil (1982) y pueden obtenerse haciendo uso del programa ***ROC Analizer,*** creado por Centor y Keightley (1989).

Un caso de particular interés, que coloca el problema en el marco de la predicción, se desarrolla en Silva y Alcarria (1994) y en Silva (1995): se trata del uso de este recurso para la comparación de modelos pronósticos.

Imaginemos que la condición *E* es que cierto proceso tenga determinado desenlace futuro y $\bar{E}$ es que dicho proceso tenga el desenlace opuesto. Por ejemplo, un paciente que ingresa en un servicio de quemados muere o sobrevive, un alumno que comienza en la escuela de medicina concluye exitosamente su carrera o no, etc.

Supongamos que se ha identificado un conjunto de variables que pueden, verosímilmente, contribuir a vaticinar el resultado, y que se conocen los valores de dichas variables para el sujeto *al inicio del proceso.* En el ejemplo del servicio de quemados, tal sería el caso de la edad, la condición de diabético o no y el porcentaje de la superficie corporal afectada por quemaduras de tipo hipodérmico en el momento en que ingresa el paciente.

El problema planteado se da en un marco de incertidumbre y su solución ha de ser, inevitablemente, de naturaleza probabilística. Una manera de encararlo es a través de la *regresión logística.* Es decir, construyendo una fórmula que ponga la probabilidad de que se produzca el suceso *E* en función de las variables elegidas como posibles predictoras.

Hecho esto, para cada sujeto concreto se tendrá un «perfil de entrada» (valores de las variables en cuestión para ese sujeto) y, consecuentemente, se podrá estimar la probabilidad de que evolucione hacia el desenlace *E.* Si se tiene una muestra de *n* individuos para los cuales se conoce tal perfil, una vez realizado el ajuste logístico, se podrán calcular las probabilidades $p_1, p_2, \ldots . p_n$ que corresponden a los *n* sujetos.

Ahora bien, en este caso, se contará con un genuino *gold standard:* el desenlace que efectivamente se produzca a la postre para cada sujeto. Supongamos que en este contexto se fija un valor *P* entre 0 y 1. Si la p_i obtenida para el i-ésimo sujeto es inferior a ese *P,* la probabilidad de que ocurra *E se* considera tan baja que se le pronostica el resultado $\bar{E}$; es decir: se declara . -. Inversamente, si $p_i \geq P$, *se* considera suficientemente alta como para declarar *T+*: es decir, se vaticina *E.* Lo que se ha hecho es crear un mecanismo de clasificación en función del resultado probabilístico y de un punto de corte *P.*

Como consecuencia de aplicar tal mecanismo, se producirán falsos *positivos y falsos negativos,* además de aciertos en uno u otro sentido. A partir de aquí, para cada *P* que se elija, se puede formar una configuración según la cual pueden clasificarse los datos de una muestra de acuerdo con los resultados reales y vaticinados; y construir por tanto una distribución como la de la Tabla 9.1 con la cual computar los indicadores clásicos: la *sensibilidad y* la *especificidad.* La curva ROC empíricas construye finalmente a partir de los valores que se obtienen usando los distintos puntos de corte que se definan en el espectro de 0 a 1.

Supongamos que se eligen *m* puntos de corte en ese intervalo. Llamémosles $P_1, \ldots . P_m$. Por ejemplo, pueden ser en este caso los $m = 9$ puntos siguientes:

$$P_1 = 0{,}1 \qquad P_2 = 0{,}2 \qquad \ldots \qquad P_9 = 0{,}9$$

Imaginemos que se valoran dos modelos predictivos alternativos -de hecho, dos conjuntos de variables predictivas que se usarán en respectivos ajustes logísticos- y que se quiere evaluar cuál de ellos es más eficiente. Supongamos que se aplican ambos modelos a un conjunto de 1.200 sujetos, 500 de los cuales tuvieron el desenlace *E,* en tanto que los restantes 700 evolucionaron hacia la condición $\bar{E}$. Se tomaron los 9 puntos de corte sugeridos anteriormente, que dan lugar a 10 intervalos (clases o categorías posibles en este caso). La Tabla 9.8 recoge los resultados de la distribución según esas categorías y según el resultado para cada uno de los ajustes.

Tabla 9.8 *Distribución de sanos y enfermos según los intervalos conrrespondientes a* 9 *puntos de corte para dos ajustes predictivos*

Intervalos	Ajuste 1		Ajuste 2	
	E	$\bar{E}$	E	$\bar{E}$
0,0 – 0,1	160	28	80	14
0,1 – 0,2	135	70	25	14
0,2 – 0,3	5	42	60	35
0,3 – 0,4	30	63	105	77
0,4 – 0,5	15	91	70	21
0,5 – 0,6	5	119	60	77
0,6 – 0,7	75	147	10	112
0,7 – 0,8	25	56	5	91
0,8 – 0,9	20	49	10	77
0,9 – 0,10	30	35	75	182
Total	500	700	500	700

La Figura 9.4 muestra las 2 curvas *ROC* resultantes. El cómputo de las áreas usando la fórmula [9.2] arroja, respectivamente, los valores

$$\Delta_1 = 0{,}696 \quad , \quad \Delta_2 = 0{,}736$$

de lo cual se deduce que el segundo ajuste tiene, aparentemente, mayor capacidad predictiva que el primero.

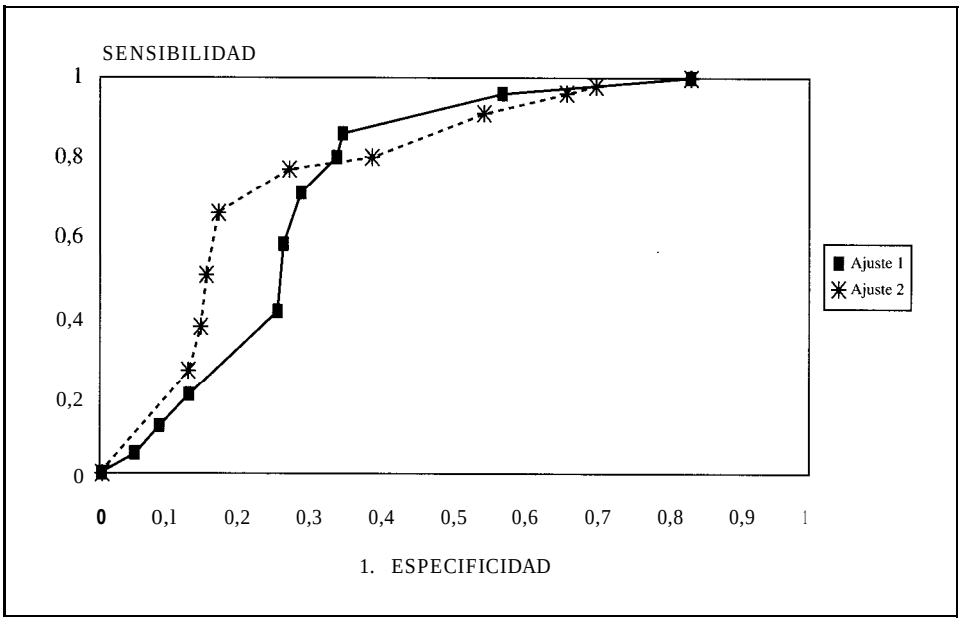


Figura 9.4. *Curvas ROC correspondientes a dos ajustes predictivos.*

El programa *ROC ANALYZER* computa los errores estándar siguientes:

$$\text{se}\,(A_1) = 0{,}0168 \quad , \quad se\,(\Delta_2) = 0{,}0154$$

El propio programa comunica que la diferencia es, por lo demás, estadísticamente significativa al nivel cx = 0,05 (p = 0,038).

9.4. Estimación de una tasa de prevalencia con una prueba imperfecta

Se define teóricamente como *prevalencia* de la enfermedad en cuestión a la probabilidad *P(E)* de que un sujeto cualquiera tenga la enfermedad en estudio; *P(E) es* la expresión probabilística de lo que, en sentido epidemiológico, se conoce como tasa *de prevalencia* (véase, por ejemplo, Hill, 1967) y se define como el resultado de dividir el número de enfermos existentes en un instante dado entre el número de individuos susceptibles de padecer dicha enfermedad en ese momento.

Con frecuencia, este importante parámetro se estima a través de datos obtenidos mediante una prueba como la descrita previamente, y como estimador de *P(E)* suele tomarse directamente la proporción de sujetos de una muestra para los que se obtiene un resultado positivo (T+); tal es el procedimiento usual, por ejemplo, para

las estimaciones hechas a partir de datos surgidos en el contexto de los programas de pesquisaje (*screening programs*) y de los exámenes masivos.

Ese estimador está afectado, naturalmente, por el error muestra1 que dimana del hecho de que no toda la población es objeto de examen (Silva, 1993). Sin embargo, de tanta o mayor importancia son las consecuencias de que raramente se cuenta con pruebas diagnósticas perfectas: no todos los enfermos son detectados como tales, ni todos los sanos resultan clasificados como negativos. Los efectos de tal circunstancia se examinan a continuación.

Supongamos que α y β son conocidos para cierta prueba *T y* cierta dolencia *E, y* que se desea estimar la prevalencia *P(E)* de esta última en una población dada. El método natural consiste en seleccionar una muestra de tamaño *n* y aplicarle la prueba diagnóstica en cuestión a todos sus integrantes. La estimación regular de *P(E)* sería, simplemente, la tasa de positivos $p = \frac{t}{n}$, donde *t es* el número de sujetos para los cuales la prueba produjo un resultado positivo *(T+).* Puesto que la prueba no es perfecta, es evidente quep es, a su vez, un estimador imperfecto de *P(E).*

Rogan y Gladen (1978), partiendo de que el valor esperado de *p* puede ponerse, según la teoría elemental de probabilidades, del modo que sigue:

$$\begin{aligned} P(T+) &= P(T+ \mid E)\ P(E) + P(T+ \mid \bar{E})\ P(\bar{E}) \\ &= \alpha\ P(E) + (1 - \beta)\ (1 - P(E)) \end{aligned}$$

y, despejando *P(E)* de esta relación, obtienen una corrección que elimina el sesgo atribuible a la falta de sensibilidad y especificidad:

$$p^* = \frac{(p + \beta - 1)}{(\alpha + \beta - 1)} \qquad [9.4]$$

Los efectos de esta corrección pueden ser considerables, aun en el caso de pruebas diagnósticas que tengan alta sensibilidad y especificidad. Por ejemplo, si $p = 0{,}20$ y $cx = \beta = 0{,}9$, entonces la estimación corregida de la prevalencia sería $p^* = 0{,}125$.

Esto quiere decir que el acto, común y corriente, de estimar una tasa de prevalencia mediante la mera fracción de sujetos para los que la prueba diagnóstica arroja un resultado positivo entre el total de examinados es casi siempre sesgado y, ocasionalmente, muy lejano de la realidad. Lo cierto es que, a pesar de que esta advertencia data ya de varios lustros, en la práctica sólo excepcionalmente es tenida en cuenta. Debe señalarse, sin embargo, que la solución de Rogan y Gladen no es una panacea, ya que la fórmula [9.4] puede dar lugar a estimaciones aberrantes. Por ejemplo, sip > α, entonces *p** será mayor que 1. Y no se sabe qué es peor: una estimación sin sentido, o una fuertemente sesgada.

9.5. Estimación del impacto relativo en un marco realista

En muchas situaciones de investigación la determinación de la condición E o $\bar{E}$ no es susceptible de error, ya que no siempre ella se hace mediante una prueba diagnóstica como tal. El ejemplo más obvio es aquel en que el desenlace es la muerte o la sobrevivencia del sujeto a raíz de cierto proceso, determinación que, obviamente, no está sujeta a error alguno. En muchos otros casos, sin embargo, esta dificultad sí suele presentarse. La presente sección procura discutir el efecto de la imperfección de *las pruebas diagnósticas* en tal caso.

Dos razones nos llevan a que la discusión que sigue se circunscriba al ámbito de los estudios prospectivos, que abarcan tanto a los observacionales de cohorte como a los experimentales. Por una parte, el tema no se aborda con una intención didáctica -que aconsejaría cierto grado de exahustividad- sino con la de alertar acerca de que las estimaciones de ciertos parámetros se pueden llevar adelante de una manera bastante ingenua, aspecto tratado con mucha menos insistencia de la que a nuestro juicio merece[6]. Por otra parte, esta inquietud no tiene, a nuestro juicio, especial entidad en los estudios retrospectivos, pues el establecimiento de la condición E o E (casos o controles respectivamente) es el punto de partida. Una vez creados tales grupos, se emprende un estudio comparativo que discurre del efecto a sus presuntas causas, de modo que tal definición es objeto de atención priorizada y, consiguientemente, es mucho menos susceptible de error. En los prospectivos, sin embargo, ese proceso diagnóstico no es parte del diseño sino que se desarrolla apelando a los procedimientos regulares de la práctica médica comentados en las secciones precedentes. Tal convicción es compartida por otros autores (véase, por ejemplo, Cousens *et al.,* 1988).

9.5.1. Riesgo relativo

Consideremos el caso harto frecuente en que se quiere estimar el *riesgo relativo* de padecer cierto daño E asociado a determinado factor de riesgo F. Veamos con un ejemplo simple cómo la falta de sensibilidad y especificidad de la prueba usada para evaluar si está presente ese daño puede conducir a una notable subestimación del parámetro.

Imaginemos que entre los individuos expuestos a cierto factor de riesgo, la verdadera tasa de incidencia del daño E asciende a 14%, en tanto que para los no expuestos es del 2%. Esto quiere decir que el riesgo relativo real asciende a 7. En términos formales:

[6] Una fuente de consulta es el trabajo de Copeland *et al.* (1977).

$$R^* = \frac{P(E \mid F)}{P(E \mid \bar{F})} = \frac{0{,}14}{0{,}02} = 7$$

Ahora bien, la presencia o no de *E se* establece a través de una prueba imperfecta; por ejemplo, supongamos que $\alpha = \beta = 0{,}9$. Es decir:

$$P(T+ \mid E) = 0{,}9 \qquad P(T+ \mid \bar{E}) = 0{,}1 \qquad [9.5]$$

Consideremos ahora que en cierta comunidad el 30% de los sujetos está expuesto al factor *F,* en tanto que el 70% restante está libre de él, y que un investigador desarrolla un estudio de cohorte con 1.000 sujetos de la población. ¿Cuál será el riesgo relativo estimado a partir de esta experiencia? Lo que realmente se observará habrá de ser:

$$R = \frac{P(P+ \mid F)}{P(T+ \mid \bar{F})} \qquad [9.6]$$

El esquema que recoge la Figura 9.5 permite «seguir» el proceso que conduce al cómputo de [9.6].

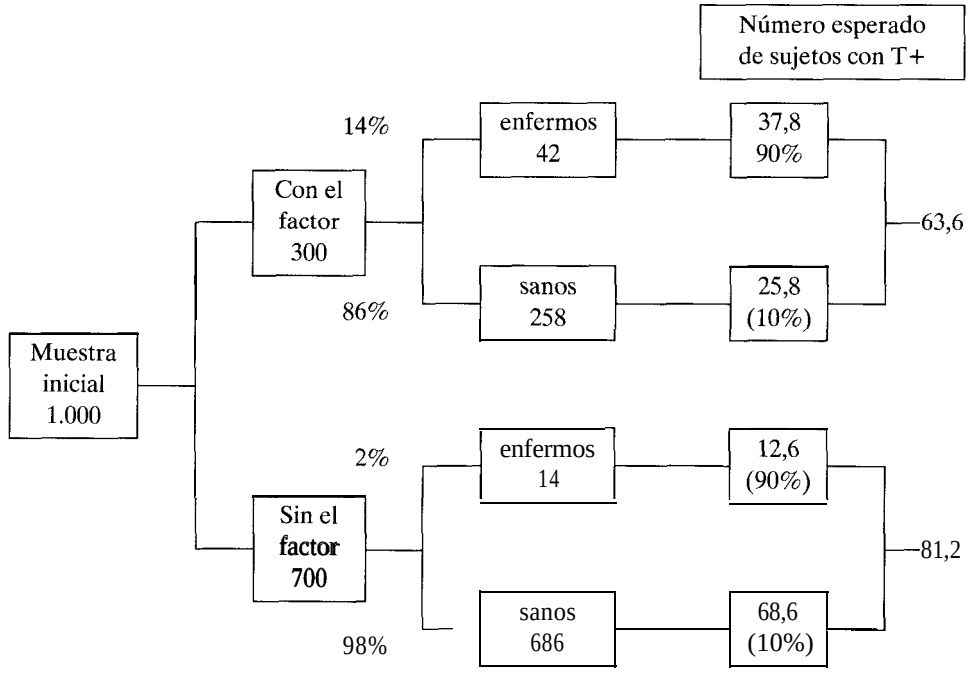


Figura 9.5. Esquema según el cual se desagrega la cohorte en enfermos y sanos, los cuales se clasifican luego como positivos 0 no.

Así. se obtendrán las estimaciones:

$$P(T+ \mid F) = \frac{63,6}{300} = 0,212 \qquad P(T+ \mid \bar{F}) = \frac{81,2}{700} = 0,116$$

De modo que la estimación del riesgo relativo será:

$$\boldsymbol{R} = \frac{0,212}{0,116} = 1,83$$

Como puede apreciarse, la subestimación de ***R*es*** enorme, a pesar de que la sensibilidad y la especificidad de la prueba eran francamente altas. Si estas fueran menores, como ocurre en muchas situaciones prácticas, el problema sería más acusado. Si se aplicara [9.4] a las tasas observadas (es decir, tanto al numerador como al denominador de R), se obtendrían estimaciones corregidas de las tasas de incidencia, tanto entre expuestos como entre no expuestos:

$$P^*\ (T+ \mid F) = \frac{0,212 + 0,9 - 1}{0,9 + 0,9 - 1} = 0,14$$

$$P^*\ (T+ \mid \bar{F}) = \frac{0,116 + 0,9 - 1}{0,9 + 0,9 - 1} = 0,02$$

La razón de estos dos números producirá, a su vez, una estimación corregida (y correcta) del riesgo relativo.

La Tabla 9.9 muestra, con los datos del mismo ejemplo, cómo se degradaría la estimación de R^* para diferentes valores de α y β.

Tabla 9.9. ***Riesgo relativo estimado de padecer cierto daño E para pruebas diagnósticas con grado variable de imperfección con los datos de la Figura 9.5***

		BETA					
		0,5	0,6	0,7	0,8	0,9	1,0
ALFA	0,5	1,00	1,03	1,08	1,17	1,44	7,00
	0,6	1,02	1,06	1,12	1,23	1,55	7,00
	0,7	1,05	1,09	1,16	1,29	1,64	7,00
	0,8	1,07	1,12	1,19	1,34	1,74	7,00
	0,9	1,09	1,15	1,23	1,39	1,83	7,00
	1,0	1,12	1,17	1,27	1,44	1,92	7,00

Ahora, mirado el problema desde un ángulo práctico, la pregunta de interés sería: ¿cómo corregir la estimación observada del riesgo relativo si la observación se hizo con una prueba para la que conocemos α y β?

Si llamamos p_1 y p_2 a las tasas de incidencia observadas y $R = \frac{p_1}{p_2}$; entonces la estimación corregida del parámetro viene dada por la fórmula [9.7];

$$R^* = \frac{R - \frac{1-\beta}{p_2}}{1 - \frac{1-\beta}{p_2}} \qquad [9.7]$$

Consideremos los datos realmente observados pero en la forma de una tabla clásica de 2 x 2:

	Diagnóstico de la enfermedad		
Factor de riesgo	$T+$	$T-$	Total
F	a	b	n_1
$\bar{F}$	c	d	n_2

Así las cosas, el riesgo relativo corregido se estimaría mediante la siguiente fórmula, equivalente a [9.7]:

$$R^* = \frac{(1-\beta) - a}{(1-\beta) - c}$$

El valor correcto de χ^2 asociado a esta tabla no sería el clásico:

$$\chi^2_{obs} = \frac{(\text{Ud} - bc)^2 n}{(a + \text{c})\ (b + \quad n_1 n_2}$$

donde $n = n_1 + n_2$, sino:

$$\chi^2_c = \frac{(cn_1 - an_2)^2 n}{[(1-\beta)\, n - (a + \text{c})]\ [(\text{a} + \text{c}) - \alpha n]\, n_1 n_2}$$

Cabe destacarse un hecho singular que se deduce de [9.7]: para hacer la corrección no es necesario conocer ni la tasa de prevalencia del factor ni la sensibilidad de

la prueba. Por otra parte, esta misma ecuación permite apreciar, tras una simple manipulación algebraica, que, lo que refleja la Tabla 9.9 para el ejemplo considerado no es casual: siempre que el riesgo relativo R sea mayor que 1, el valor de R que se observe necesariamente será una subestimación del verdadero.

Por otra parte, con un poco de álgebra es posible probar que el valor de χ^2 corregido siempre arrojará un valor mayor que el del estadígrafo obtenido sin corregir; es decir, siempre se cumple que $\chi^2_{obs} \leq \chi^2_c$. El efecto de este hecho es que el intervalo de confianza para el riesgo relativo es en realidad más ancho de lo que aparenta ser cuando no se hace corrección alguna; esto puede apreciarse a través de las fórmulas que se presentan a continuación, las más comúnmente usadas para obtener los límites inferior y superior:

$$LI = R^{(1-Z_{1-\alpha/2}/\sqrt{\chi^2})} \qquad LS = R^{(1+Z_{1-\alpha/2}/\sqrt{\chi^2})}$$

Finalmente, también se deduce de [9.7] que si β fuese igual a 1, entonces la estimación no será sesgada (se tendrá $R = R^*$), aunque la sensibilidad sea deficiente. Este último hecho puede tener una connotación práctica medular en determinadas circunstancias. A modo de ilustración, consideremos la siguiente situación.

Se quiere evaluar el efecto de un programa de educación sexual orientado a evitar embarazos indeseados en adolescentes. Para ello se eligen dos comunidades en las que se hace un censo de mujeres de 13 a 15 años. En una de ellas se desarrolla el programa ($\bar{F}$) y en la otra (F) no se lleva adelante. Dos años más tarde se entrevistan a todas las adolescentes censadas en ambas comunidades y se les pregunta si durante ese lapso tuvieron ($T+$) o no ($T-$) un embarazo no deseado. Es razonable esperar que este «método diagnóstico» habrá de tener una especificidad máxima (toda adolescente que no se embarazó habrá de testimoniarlo así) y una sensibilidad por debajo de uno (algunas, quizás muchas, de las que hayan tenido este percance, lo negarán).

A pesar de ello, de acuerdo con la fórmula [9.7], el riesgo relativo *observado* de embarazarse asociado a no pasar por el programa coincidirá con el riesgo relativo ***real*** del hecho.

Naturalmente, esto es cierto sólo en el supuesto de que el programa no modifique la actitud de la adolescente en cuanto a admitir o negar un embarazo tenido (es decir, suponiendo que se cumpla el presupuesto de las pruebas diagnósticas según el cual el valor de α es inherente a la prueba y, por ende, independiente de que esté o no presente el factor F).

9.5.2. Odds ratio

Si en las mismas circunstancias del ejemplo inicial, en lugar del riesgo relativo, se procurase estimar la odds ratio O*asociada a F, el verdadero valor sería:

$$O^* = \frac{\dfrac{P(E|F)}{P(\bar{E}|F)}}{\dfrac{P(E|\bar{F})}{P(\bar{E}|\bar{F})}} = \frac{\dfrac{0,14}{0,86}}{\dfrac{0,02}{0,98}} = 7,98$$

Si estuviéramos ante una tabla de 2 x 2 como la de la sección anterior, esta fórmula se reduce a la «razón de productos cruzados», cuyos datos proceden de la Figura 9.5:

$$O^* = \frac{a\,d}{b\,c} = \frac{42 \cdot 686}{258 \cdot 14} = 7,98$$

Por otra parte, al realizar la estimación usando los datos obtenidos con la prueba imperfecta, también se tendrá una marcada subestimación:

$$o = \frac{\dfrac{P(T+|F)}{P(T-|F)}}{\dfrac{P(T+|\bar{F})}{P(T-|\bar{F})}} = \frac{\dfrac{0,212}{0,788}}{\dfrac{0,116}{0,884}} = 2,05$$

Nuevamente, si de lo que se dispone es de las tasas observadas p, y p_2 (afectadas por los errores inherentes al carácter defectuoso de las pruebas), es posible computar una odds ratio corregida siempre que se conozcan α y β; puede probarse con un poco de álgebra elemental que para corregir la estimación β la fórmula adecuada es:

$$O^* = \frac{(\beta - q_1)}{(\beta - q_2)} \frac{(\alpha - p_2)}{(\alpha - p_1)} \qquad [9.8]$$

donde $q_1 = 1 - p_1$ y $q_2 = 1 - p_2$.

En nuestro ejemplo, poniendo p, =0,212, p_2 = 0,116, α = 0,9 y β = 0,9 en [9.8] se tiene, en efecto, la verdadera odds ratio: 0* = 7,99. Es fácil ver que, curiosamente, a diferencia de lo que ocurre con el riesgo relativo, en este caso no basta con que $\beta = 1$ para que no se produzca un sesgo por este concepto (es decir, no necesariamente se cumple en este caso que 0 = 0"). La Tabla 9.10 permite apreciar el efecto que tienen diferentes valores de α y β en la subestimación de *O** para los datos del ejemplo.

Tabla 9.10. ***Odds ratio estimada de padecer cierto daño E para pruebas diagnósticas con grado variable de imperfección con los datos de la Figura 9.5***

		BETA					
		0,5	0,6	0,7	0,8	0,9	1,0
ALFA	0,5	1,00	1,05	1,12	1,23	1,53	7,45
	0,6	1,05	1,10	1,18	1,31	1,66	7,55
	0,7	1,10	1,16	1,24	1,39	1,79	7,65
	0,8	1,16	1,22	1,31	1,47	1,92	7,76
	0,9	1,21	1,28	1,37	1,56	2,05	7,86
	1,0	1,27	1,34	1,44	1,65	2,18	7,99

9.6. Valores predictivos en el ambiente clínico

Es necesario tener claro que, desde el punto de vista operativo, los conceptos que realmente interesan en relación con las pruebas diagnósticas *no* son la sensibilidad y la especificidad. El clínico procede en la dirección opuesta: del resultado de la prueba intenta deducir la condición del paciente.

Por lo tanto, lo que él reclama de una prueba es que, si el resultado de la prueba es positivo, la probabilidad de que el sujeto esté efectivamente enfermo sea muy alta y, análogamente, que sea muy alta la de que el individuo esté sano, supuesto que la prueba arroje un resultado negativo.

En términos formales, lo ideal es que sean muy altos los valores

$$\text{y} \qquad P \quad | T\text{-})$$

que son probabilidades condicionales a las que ha dado en llamarse *valores predictivos de la prueba.*

¿Qué utilidad puede reportar conocer de antemano estos valores? Cualquier clínico avezado, ante un cuadro sintomático y un conjunto de datos básicos, es capaz de hacer un «diagnóstico presuntivo» y, llegado el caso, de atribuir una probabilidad al hecho de que el paciente portador de ese perfil padezca la dolencia en cuestión. Establecido tal diagnóstico inicial, la prueba que se practica tiene la finalidad de arrojar más luz sobre la situación: o bien consolida la validez de aquel (si el resultado es *T*+), o bien le resta verosimilitud (si la prueba da lugar al resultado *T*-).

Es decir, para evaluar el interés práctico del resultado de una prueba a nivel clínico hay que centrar la atención en el grado en que sus resultados modificarían realmente el conocimiento que se tenía sobre el estado del paciente antes de ser practi-

cada (Sox, 1986). Es ampliamente conocido que, si bien α y β son números inherentes a la prueba (en el sentido de que no dependen de cuál sea la población o el sujeto específico a la que se aplique), no ocurre lo mismo con sus valores predictivos.

Si llamamos ***P = P(E)*** a la *probabilidad a priori* de que el sujeto esté enfermo, y 0 = 1 -P a su complemento, aplicando el teorema de Bayes se obtienen de inmediato las siguientes relaciones, que expresan la forma concreta que alcanzan estos valores cuando se ponen en función de los tres parámetros [7]:

$$P(E \mid T+) = \frac{P\alpha}{P\alpha + O(1-\beta)} \qquad [9.9]$$

$$P(\bar{E} \mid T-) = \frac{O\beta}{O\beta + P(1-\alpha)} \qquad [9.10]$$

Supongamos que en una comunidad la prevalencia de padecer una enfermedad coronaria (EC) entre sujetos mayores de 50 años es 5%. Esto quiere decir que, de cada 100 sujetos que están en esa franja de edad, 5 tendrán la dolencia; o, dicho de otro modo, que la ***probabilidad a priori*** de que un sujeto elegido al azar se halle en ese caso es ***P*** = 0,05.

Si a cierto individuo se le practica una angiografía (una prueba invasiva y, por ende, peligrosa), hay dos resultados posibles: *T*+ y ***T-.*** Después de hacerlo, ¿cuál es la probabilidad de que el sujeto esté enfermo en cada caso? Teniendo en cuenta que la sensibilidad y especificidad de una angiografía para el diagnóstico de una ***EC*** son, según Austin (1982), respectivamente iguales a α = 0,87, β = 0,54 y, aplicando [9.9] y [9.10], se tiene que:

$$P(E \mid T+) = \frac{(0,05)\ (0,87)}{(0,05)\ (0,87) + (0,95)\ (0,46)} =$$

$$P(\bar{E} \mid T-) = \frac{(0,95)\ (0,54)}{(0,95)\ (0,54) + (0,05)\ (0,13)} = 0,99$$

[7] Como nota curiosa, es interesante reparar en un resultado teórico de Connell y Koespell (1985): después de consideraciones metodológicas y desarrollos algebraicos relativamente complicados, basándose en los valores predictivos, estos autores arriban a un indicador que mide lo que ellos llaman *la ganancia porcentual en certeza a que da lugar un resultado positivo y un resultado negativo.* Luego crean un indicador combinado de esas dos «ganancias». De ese elaborado proceso resulta un nuevo indicador... ¡que no es otra cosa que el ya mencionado índice de Youden! Aunque ello aporta otra interpretación cualitativa para I_y, sorprende la capacidad de algunos resultados para ser «redescubiertos» (los autores no mencionan que este indicador había sido propuesto tres décadas antes).

de modo que la probabilidad de que el individuo esté enfermo, si la prueba fue positiva, se eleva a 0,09 y si se obtuvo T-, tal probabilidad se reduce a 0,01. Los valores 0,09 y 0,01 son las llamadas probabilidades *a posteriori* de estar enfermo.

Ahora imaginemos que de este sujeto se sabe, además, que tiene más de 60 años y que es fumador; y que se conoce que para la población de los que cumplen estas dos restricciones la prevalencia es $P = 0{,}18$. Es fácil entonces corroborar que:

$$P(E+ \mid T+) = 0{,}29 \qquad P(E+ \mid T-) = 0{,}05$$

Si, finalmente, se trata de un sujeto perteneciente a un subgrupo de la población para el cual $P = 0{,}62$, entonces los valores predictivos serían:

$$P(E+ \mid T+) = 0{,}75 \qquad P(E+ \mid T-) = 0{,}28$$

Una síntesis de esta breve casuística se recoge en la Tabla 9. ll.

Tabla 9.11. *Valores predictivos para 3 prevalencias diferentes en un prueba para la que* $\alpha = 0{,}87$ y $\beta = 0{,}54$

Sujeto	Prevalencia (probabilidad *a priori)*	Probabilidad *a posteriori* si T es positivo	Probabilidad *a posteriori* si T es negativo
1	0,05	0,09	0,01
2	0,18	0,29	0,05
3	0,62	0,75	0,28

En síntesis, antes de hacer la prueba, se puede conocer cómo y cuánto se modifica tal probabilidad *a priori* para cada uno de sus resultados posibles. Puesto que los valores de $P(E+ \mid T+)$ y $P(E+ \mid T-)$ pueden calcularse *antes* de aplicarse la prueba, el clínico podría, reparando en ellos, abstenerse incluso de ordenar su aplicación. Ese sería el caso si se sabe que, cualquiera que sea el resultado, no se modificaría la conducta clínica posterior.

Así, un resultado T+ para el sujeto 1 de la Tabla 9.11 es tal que su probabilidad *a posteriori se* eleva demasiado poco (pasa de a 0,09) como para justificar la iatrogenia potencial de la angiografía: en cualquiera de los dos casos, considerará de hecho que el paciente no está enfermo. El sujeto 2 tiene una probabilidad inicial bastante alta de padecer la enfermedad coronaria (0,18). Pero si la angiografía se realiza y da negativa, entonces la probabilidad de que, no obstante, el individuo padezca de una *EC es* tan baja (0,05) que el clínico cambiará el curso de sus acciones (quizás actúe entonces como si el paciente no tuviese esa dolencia). En el caso 3, un resultado *T+* eleva muchísimo la convicción de que el sujeto tiene *EC,*

pero *T- no* cambiaría la presunción inicial: se mantendría, a pesar de ello, el diagnóstico original. Resumiendo: al examinar los datos de la Tabla 9.11, probablemente la conducta más racional del médico sería indicar una angiografía para el paciente 2, pero no para el 1 ni para el 3. De tal suerte se evitaría una práctica según la cual, al decir de Wong y Lincoln (1983), algunos clínicos «disparan antes de apuntar».

9.7. Cuando eficiencia no equivale a utilidad

En la Sección 9.1 se había señalado que los conceptos de sensibilidad y especificidad solían generar confusiones importantes. Una de ellas dimana de la dificultad para interiorizar el concepto de probabilidad condicional e interpretarlo adecuadamente en un entorno práctico.

$P(T+|E)$ y $P(T-|\bar{E})$ son medidas de la *eficiencia* de la prueba *T* que expresan la calidad de su comportamiento ante la presencia o ausencia de la enfermedad, pero no necesariamente de su *utilidad* para pronunciarse en enclaves prácticos reales sobre si un sujeto está o no enfermo. El grado en que la prueba cumple adecuadamente esta última función viene dado, como se ha explicado, por los valores predictivos $P(E|T+)$ *y* $P(\bar{E}|T-)$.

La literatura recoge frecuentes permutaciones entre estos conceptos, todo lo cual se torna más confuso cuando se amplía el marco semántico y se incorporan los conceptos de *tasas de falsos positivos, tasas de falsos negativos,* etc. Tales percances, sin embargo, pueden entenderse en general como erratas o *lapsus mentis* de autores y editores. Pero ese no es el caso de una gran cantidad de estudios cuyo contenido evidencia la incomprensión del hecho de que las dos medidas de eficiencia son indicadores intrínsecos, inherentes a la prueba; es decir, independientes de cuál sea el grupo poblacional al que se apliquen. De manera que *no tiene sentido alguno* buscar los valores de a y β para diferentes subgrupos poblacionales (por ejemplo, para diferentes grupos de edad separadamente).

En tal error incurren por ejemplo Scieux *et al.* (1992) quienes estudian una técnica basada en cierto inmunoensayo para la detección de *Chlamydia trachomatis* en orina, y se proponen desde el mismísimo título estimar los valores de α y β tanto para hombres como para mujeres. Cuando estos parámetros se estiman dentro de ciertos subgrupos, los resultados no son idénticos, como es natural, debido a que no lo son las condiciones experimentales, a la intervención del azar y a la variabilidad biológica. Pero *conceptualmente si lo son.* El error es equivalente al que se cometería si nos planteáramos estimar cuál es la distancia entre Roma y París los domingos y cuál es esa distancia los restantes días de la semana.

Situaciones muy similares se presentan con extrema frecuencia, como evidencian trabajos tales como los de Mills *et al.* (1992), Weinberger y Harger (1993), Quentin *et al.* (1993), Rost, Burnam y Smith(1993) y Ainslie y Murden (1993).

Ransohoff y Feinstein (1978) han señalado que, al evaluar el valor diagnóstico de una prueba es necesario que la muestra de enfermos recorra un amplio espectro de pacientes y que otro tanto ocurra con los sujetos no enfermos. Así, según estos autores, la muestra de enfermos debe abarcar, por ejemplo, casos con diversos grados de la dolencia, en diferentes localizaciones, con diferentes grados de cronicidad y alta diversidad en materia de síntomas; análogamente, el conjunto de sujetos que no padecen la dolencia debe incluir pacientes con enfermedades parecidas y signos clínicos que pudieran estar asociados con resultados que sean falsos positivos. ¿Por qué plantear esta advertencia si los valores de α y β no dependen de los rasgos que tengan los pacientes? Todo reside en que cuando se procura diagnosticar una enfermedad, la denominación usada puede abarcar una gama de variantes o de intensidades. Por ejemplo, para estimar los parámetros asociados a una prueba diagnóstica de cáncer gástrico, la muestra de enfermos debe incluir sujetos con diferentes grados de malignidad, ya que en cierto sentido se trataría de diferentes enfermedades englobadas en una. Del mismo modo, un pequeño infarto del miocardio puede entenderse como una enfermedad diferente a un gran infarto; una prueba puede ser muy sensible en este último caso pero no así en el primero. Si al usarla no se distingue en el tipo de infarto que se procura diagnosticar, será menester que entre los «casos» de la muestra existan ambas expresiones para conseguir una estimación de la sensibilidad.

9.8. El salto a la práctica preventiva

Ahora detengámonos a examinar el efecto que puede producirse ***a nivel epidemiológico*** como consecuencia de que los valores predictivos dependen de la prevalencia de la enfermedad. Situémonos para ello en el contexto de los programas de tamizaje ***(screening programs); es*** decir, en la situación en que se trata de aplicar pruebas diagnósticas para la detección masiva de enfermedades.

Supongamos que se trabaja con una grave enfermedad que exhibe una prevalencia que, teniendo en cuenta su seriedad, puede considerarse relativamente alta. Por ejemplo, supongamos que se trata de una enfermedad padecida por una de cada 150 personas de la población. Es decir, estamos en el caso en que: $\boldsymbol{P} = 0{,}0067$ [8]. Imaginemos que se cuenta con una prueba diagnóstica de alta eficiencia: concretamente, que tiene sensibilidad y especificidad del 90% (recuérdese que no muy frecuentemente se opera con una prueba con tal alto nivel de eficiencia, como evidencia la Tabla 9.2).

En una población de 9.000 personas habrá, por tanto, 60 enfermos. Si la pobla-

[8] Nótese que eventos tales como ser portador del virus del SIDA o de cáncer de la pared uterina son muchísimo menos prevalentes.

ción fuera íntegramente sometida a dicha prueba, ésta detectará al 90% de los enfermos (54 individuos), y declarará como sanos al 90% de los 8.940 sujetos que lo están (o sea, diagnostica como enfermos al 10% restante: a 894 sanos).

En total, habrá por tanto 948 individuos diagnosticados como enfermos, pero sólo el 5,7% de ellos (54 de los 948) realmente lo están.

Los resultados precedentes pueden deducirse directamente de la aplicación de la fórmula [9.9] ya que, para $\alpha = \beta = 0{,}9$ y $\boldsymbol{P} = 0{,}0067$, se tiene que:

$$P\,(E \mid T+) = \frac{(0{,}00667)\ (0{,}9)}{(0{,}00667)\ (0{,}09)\ +\ (0{,}99333)\ (0{,}1)} = 0{,}057$$

de donde se deriva que, en efecto, la tasa de falsos positivos es nada menos que del 94%. Para que se aprecie que no se trata de una mera especulación teórica, téngase en cuenta por ejemplo que la tasa de carcinoma de cérvix entre mujeres en edad fértil asciende en algunos enclaves a cifras del orden de 1 por 1.000 ($\boldsymbol{P} = 0{,}001$) y que, según Nesbit y Brack (1956), la sensibilidad y especificidad de la prueba citológica son $\alpha = 0{,}91$ y $\beta = 0{,}76$. La tasa de falsos positivos sería entonces muy similar a la del ejemplo. De hecho, el sistema británico de citología cervical, que invierte unas 280.000 libras esterlinas por cada vida salvada, realiza 40.000 pruebas y 200 biopsias por cada caso verdaderamente beneficiado por la detección precoz (Charny, Farrow y Roberts, 1989).

El interés básico de esta ilustración radica en que llama la atención sobre el hecho de que los dividendos que se desprenden de este tipo de tamizajes no son siempre mayores que los problemas que traen consigo: posiblemente muchos de esa inmensa mayoría de sanos entre los positivos serán objeto de pruebas adicionales con sus respectivos riesgos; como mínimo, muchas de las víctimas del error habrán de sumergirse en un estado de zozobra, acaso acompañado de un injustificado temor a la muerte. Muchas veces, los exámenes masivos se desarrollan sin reparar en esta nefasta consecuencia de la imperfección de las pruebas. Para apreciar una semblanza a la vez profunda y abarcadora de tales contingencias para las pruebas de cáncer en la pared uterina, sugiero la lectura del inquietante artículo de McCormick (1989). Este problema ha promovido considerable polémica. Recientemente, ésta se ha reanimado por la problemática del SIDA, tal y como puede apreciarse en Meyer y Paulker (1987) y Germanson (1989).

En términos generales se han planteado (Terris, 1967) dos tácticas preventivas fundamentales: el diagnóstico temprano y la llamada *supervisión preventiva*.

El enfoque clásico ejercía una espera pasiva a que el sujeto pasara a la condición de paciente [9] para -una vez en manos competentes- darle el tratamiento debido;

[9] Nótese la connotación del término «paciente», acuñado por la medicina asistencial: el sujeto enfermo es alguien que ha quedado en las redes de quienes habrán de decidir por él.

el enfoque complementario se propone ejercer acciones sobre la población aparentemente sana.

La inmunización contra la difteria, el tétanos y la poliomelitis, el control del sarampión y la tosferina, y la lucha contra muchas dolencias ocupacionales, son todas experiencias exitosas de la prevención, corroboradas más allá de la anécdota, como corresponde al pensamiento epidemiológico, radicalmente diferente del estilo que prevalece en el ambiente clínico.

Sin embargo, el hecho de que el tratamiento pase a ser sólo una segunda línea de defensa ha conducido a lo que dos décadas atrás Illich (1975) llamó *cacería preventiva de enfermedades* tendencia que nos pone en riesgo de otorgar proporciones epidémicas al diagnóstico. Es quizás insuficientemente conocida la siguiente experiencia realizada hace ya medio siglo, descrita en ese libro clásico.

Mil niños de once años, tomados al azar de escuelas públicas de Nueva York, fueron examinados por médicos pertenecientes a una institución que prestaba atención gratuita. Una primera inspección estableció que 610 habían sido objeto de una amigdalectomía. Los restantes 390 fueron sometidos a un examen practicado por varios especialistas que dictaminaron que 175 (el 45%) demandaban una extirpación de amígdalas. Aquellos 215 niños no elegidos fueron reexaminados por otros especialistas: 99 de estos niños (46%) fueron entonces declarados como necesitados de la operación. Finalmente, los 116 escolares remanentes fueron vistos por tercera vez: el 44% de ellos recibió la susodicha indicación de amigdalectomía. La notable regularidad de los porcentajes (45, 46 y 44) adiciona, por cierto, elocuencia a la experiencia.

El diagnóstico precoz supone que es posible detener o mitigar la enfermedad si se detecta en sus primeras etapas. Para la tuberculosis, por ejemplo, la radiografía de un individuo supuestamente sano permite identificar la dolencia antes de la aparición de síntoma alguno, lo cual incrementa la probabilidad real de detener el proceso morboso. El intento de diagnóstico precoz del cáncer ha sido el más practicado. La quimioterapia, la intervención quirúrgica, u otra línea de acción usada en las etapas iniciales, antes de que el proceso maligno se haya consolidado o expandido, ofrece -supuestamente- mayor esperanza de vida que cuando el oncólogo se enfrenta a un tumor ya consolidado o quizás ramificado en el organismo.

Pero, ijustifica esta presunción la práctica de las búsquedas masivas de morbilidad? Una respuesta, hoy considerada clásica, fue establecida por Wilson y Jungner (1968) al enunciar los diez rasgos de la enfermedad que los legitiman y que resumo a través de las cinco reglas siguientes:

a) Debe tener prevalencia considerablemente alta.
b) Debe ser suficientemente grave.
c) Ha de contarse con un tratamiento efectivo y accesible para encararla.
d) Habrá de disponerse de pruebas eficientes para su detección que sean, además, aceptadas por la población.
e) Ha de conocerse adecuadamente su historia natural.

Man y Fowler (1991), en consonancia con estas premisas, advierten que con frecuencia los entusiastas se lanzan a la detección sin adiestrarse suficientemente, sin un protocolo para definir un resultado positivo y el tratamiento de estos pacientes, y sin apreciar la enorme carga de trabajo que generan la necesidad de estudiar y seguir a los pacientes durante un largo período de tiempo. Estas observaciones provienen de la amarga realidad y no de la teoría: simplemente, el entusiasmo por hacer el bien, no basta.

Puesto que, como señalé en otro sitio (Silva, 1992), el propio concepto de prevención entraña una paradoja, ya que el discurso en que se basa puede resumirse en buena medida a través de la máxima: ***Prívate ahora de algo que te agrada con el fin de disfrutar de grandes ventajas en el futuro.*** Tal es el caso del mensaje contra el tabaco y el consumo de grasas saturadas, a favor del cinturón de seguridad en los automóviles, del condón o el Papanicolau. Todos entrañan una renuncia o una incomodidad, de modo que las condiciones para la polémica están a la vista.

Skrabanek y McCormick (1989) alertan que, puesto que la prevención tiene un precio, en cada caso cabe preguntarse si éste es o no exhorbitante. La respuesta a esa pregunta tiene que derivarse de una medición objetiva y mesurada, no de presunciones, por muy razonables que parezcan. De hecho, la necesidad de evaluar críticamente las tecnologías preventivas ha sido priorizada por instituciones responsables de generalizar su aplicación. En este sentido se han pronunciado importantes organizaciones como el ***Centro de Control de Enfermedades*** (véanse Teutsch, 1992 o el trabajo de US Preventive Services Task Force, 1989).

A modo de ilustración final, cabe recomendar la lectura de Russell(1994) donde se examinan desde esa perspectiva las prácticas de pesquisaje en materia de cáncer cervical, colesterolemia excesiva y cáncer prostático. En este último caso, por ejemplo, el cúmulo de razones que aconsejan erradicar el tacto rectal (incluso la imagen ultrasónica y la prueba del antígeno específico) de los chequeos de rutina en hombres adultos es tan notable que actualmente casi nadie defiende tal práctica.

Bibliografía

Ainslie NK, Murden RA (1993). ***Effect of education on the clock-drawing dementia screen in non-demented elderly persons.*** Journal of the American Geriatric Society 41: 249-252

Austin EH (1982). ***Prospective evaluation of radionuclide angiocardiography for the diagnosis of coronary heart disease.*** American Journal of Cardiology 50: 1212-1216.

Begg CB, Metz CE (1990). ***Consensus diagnosis and "gold standards».*** Medical Decision Making 10: 29-30.

Cedorlöf R, Johnsson E, Lundman T (1966). ***On the validity of mailed questionnaires***

in diagnosing «angina pectoris» and «bronchitis». Archives of Environmental Health 13: 738-742.

Centor RM (1991). ***Signal detectability: the use of ROC Analysis.*** Medical Decision Making ll: 102-106.

Centor RM, Keightley GE (1989). ***Receiver operating charactetistic (ROC) curve urea analysis using the ROC ANALYZER.*** SCAMC Proceedings 13: 222-226.

Connell FA, Koespell TD (1985). ***Measures of gain in certainty from a diagnostic test.*** American Journal of Epidemiology 121: 744-763.

Copeland KT, Checkoway H, McMichael AJ, Holbraook RH (1977). ***Bias due to misclassification in the estimation of relative risk***.American Journal of Epidemiology 105: 488-495.

Cousens SN, Feachem RG, Kirlwood B, Mertens TE, Smith PG (1988). ***Case-controls studies of childhood diarrhoea.*** CDD/EDP/88.2, World Health Organization.

Charny MC, Farrow SC, Roberts CJ (1987). ***The cost of saving a life through cervical cytology screening: implications for health policy.*** Health Policy 7: 345-359.

Diamond GA (1989). ***Limited assurances.*** American Journal of Cardiology 63: 99-100.

Duchatel F, Muller F, Oury JF, Mennesson B, Boue J, Boue A (1993). ***Prenatal diagnosis of cystic fibrosis: ultrasonography of the gallbladder at 17-19 weeks of gestation.*** Fetal Diagnosis and Therapy 8: 28-36.

Dunn JP, Cobb S (1962). ***Frequency ofpeptic ulcer among executives craftsmen, and foremen.*** Journal of Occupational Medicine 4: 343-348.

Galen RS, Gambino SR (1975). ***Beyond normality: thepredictive value and efficiency of medical diagnosis.*** Wiley, New York.

Germanson T (1989). ***Screening for HIV: can we afford the confusion of the false positive rate?.*** Journal of Clinical Epidemiology 42: 1235-1237.

Hanley JA, McNeil BJ (1982). ***The meaning and use of the urea under a receiver operating charactetistic (ROC) curve.*** Radiology 143: 29-36.

Hill AB (1967). ***Principios de estadística médica.*** Instituto Cubano del Libro, La Habana.

Illich 1 (1975). ***Medical nemesis.*** Calder and Boyars, Londres.

Jarvisalo J, Hakama M, Knekt P ***et al.*** (1982). ***Serum tumor markers CEA, CA 50, TATI, and NSE in lung cancer screening.*** Cancer 15: 71-78.

Kaddah MH, Maher KM, Hassanein HI, Farrag AI, Shaker ZA, Khalafallah AM (1992). ***Evaluation of different immunodiagnostic techniques for diagnosis of hydatidosis in Egypt*** Journal of the Egyptian Society of Parasitology 22: 653-665.

Kurjak A, Schulman H, Sosic A, Zalud l, Shalan H (1992). ***Transvaginal ultrasound, color flow, and Doppler waveform of thepostmenopausal adnexal mass.*** Obstetrics and Gynecology 80: 917-921.

Lilienfeld AM, Lilienfeld DE (1983). ***Fundamentos de epidemiología*** Fondo Educativo Interamericano, Mexico DE

Lusted B (1971). ***Signal detectability and medical diagnosis.*** Science 171: 1217-1219.

Macía M, Novo A, Ces J, González M, Quintana S, Codesido J (1993). ***Progesterone challenge test for the assessment of endometrial pathology in asymptomatic menopausal women.*** International Journal of Gynecology and Obstetrics 40: 145-149.

Magruder HK, Stevens HA, Alling WC (1993). ***Relative performance of the MAST, VAST, and CAGE versus DSM-III-R criteria for alcohol dependence.*** Journal of Clinical Epidemiology 46: 435-441

Mak JW, Normaznah Y, Chiang GL (1992). ***Comparison of the quantitative buffy coat technique with the conventional thick blood film technique for malaria case detection in the field.*** Singapore Medical Journal33: 452-454.

Makijarvi M (1993). ***Identification ofpatients with ventricular tachycardia after myocardial infarction by high-resolution magnetocardiography and electrocardiography.*** Journal of Electrocardiology 26: 117-124.

Man D, Fowler G (1991). ***Detección masiva: teoría y ética.*** British Medical Journal 6:62-66.

McCormick JS (1989). ***Cervical smears: a questionable practice?*** Lancet; i: 207-209.

McIntosh ED, Jeffery HE (1992). ***Clinical application of urine antigen detection in early onset group B streptococcal disease.*** Archive of Diseases of Children 67: 1198-1200.

Meyer KB, Paulker SG (1987). ***Screeningfor HIV: can we afford the falsepositive rate? New*** England Journal of Medicine 317: 238-241.

Mills SJ, Ford M, Gould FK, Burton S, Neal DE (1992). ***Screening for bacteriuria in urologicalpatients using reagent strips.*** British Journal of Urology 70: 314-317.

Nesbit REL, Brack CB (1956). ***Role of cytology in detection of carcinoma of cervix.*** Journal of the American Medical Association 161: 183-188.

Pasquini L, Parness IA, Colan SD, Wernovsky G, Mayer JE, Sanders SP (1993). ***Diagnosis of intramural coronary artery in transposition of the great arteries using two-dimensional echocardiography.*** Circulation 88: 1136-1141.

Quentin R, Dubarry 1, Gignier C, Saulnier M, Pierre F, Goudeau A (1993). ***Evaluation of a rapid latex test for direct detection of Streptococcus agalactiae in various obstetrical and gynecological disorders.*** European Journal of Clinical Microbiology and Infectious Diseases 12: 51-54.

Quiñonez ME, Martínez D, Taibo ME, Rota TR, Flores J (1992). ***Comparación entre diferentes reactivos comerciales para el diagnóstico de la infección por los virus de la inmunodeficiencia humana (VIH-1) y de la hepatitis B (AgHBs y anti-HBc).*** Revista del Instituto Nacional de Higiene «Rafael Rangel» 23: 32-40.

Ransohoff DF, Feinstein AR (1978). ***Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New*** England Journal of Medicine 299: 926-930.

Rogan WJ, Gladen B (1979). ***Estimating prevalence from the results of a screening test.*** American Journal of Epidemiology 107: 71-76.

Rost K, Burnam MA, Smith GR (1993). ***Development of screeners for depressive disorders and substance disorder history.*** Medical Care 31: 189-200.

Russell LB (1994). ***Educated guesses: making policy about medical screenings tests.*** University of California Press, Berkeley.

Scieux C, Bianchi A, Henry S ***et al.*** (1992). ***Evaluation of a chemiluminometric immunoassay for detection of Chlamydia trachomatis in the urine of male and female patients.*** European Clinical Microbiology and Infectious Diseases ll: 704-708.

Schecheter MT, Sheps SB (1985). ***Diagnostic testing revisited: pathways through uncertainty.*** The Canadian Medical Association Journal 132: 755-760.

Shields LE, Gan EA, Murphy HF, Sahn DJ, Moore TR (1993). ***Theprognostic value of hemoglobin Alc in predicting fetal heart disease in diabetic pregnancies.*** Obstetrics and Gynecology 81: 954-957.

Sienko DG, Hahn RA, Mills EM et al (1993). ***Mammography use and outcomes in a community. The Greater Lansing Area Mammography Study.*** Cancer 1; 71: 1801-1909.

Silva LC (1987). ***Métodos estadísticos para la investigación epidemiológica.*** Instituto Vasco de Estadística, Cuaderno 14, Bilbao.

Silva LC (1992). ***Valoración epidemiológica de las actividades preventivas.*** Memorias de las Primeras Jornadas sobre Actividades Preventivas en el Area de Salud: 28-42, Burgos, España.

Silva LC (1993). ***Muestreo para la investigación en ciencias de la Salud.*** Díaz de Santos, Madrid.

Silva LC, Alcarria A (1993). ***Predicción del rendimiento académico a partir del perfil de entrada de los estudiantes de enfermería de la Habana.*** Educación Médica Superior 7: 97-106.

Silva LC (1995). ***Excursión a la regresión logística en ciencias de la salud.*** Días de Santos, Madrid.

Simpson AJ, Fitter MJ (1973). ***What is the best index of detectability?*** Psychological Bulletin 80: 481-488.

Skrabanek P, McCormick J (1989). ***Follies and fallacies in medicine.*** The Tarragon Press, Glasgow.

Sox HC (1986). ***Probability theory in the use of diagnostic tests.*** Annals of Interna1 Medicine 104: 60-66.

Terris M (1967). ***Una política social para la salud.*** Discurso pronunciado por el presidente de la ***Asociación Americana de Salud Pública (AASP),*** Segunda Sesión General, 95.ª Sesión Anual,Miami.

Teutsch SM (1992). *A* ***fiamework for assessing the effectiveness of disease and injury prevention.*** MMWR 41 (n.º RR-3).

US Preventive Services Task Force (1989). ***Guide to clinical preventive services: an assessment of the effectiveness of 169 interventions.*** Williams and Wilkins, Baltimore.

Valleron AJ, Eschwege E, Papoz L ***et al.*** (1975). ***Agreement and discrepancy in the evaluation of normal and diabetic oral glucose tolerance test.*** Diabetes 24: 585-593.

Weinberger MW, Harger JH (1993). ***Accuracy of the Papanicolaou smear in the diagnosis of asymptomatic infection with*** trichomonas vaginalis. Obstetrics and Gynecology 82: 425-429.

Wilson JMG, Jungner G (1968). ***Principles and practice of screening for disease.*** Public Health Papers n.º 34, WHO, Geneva.

Wong ET, Lincoln TL (1983). ***Ready' Fire'... Aim': an inquiry into laboratory test ordering.*** Journal of the American Medical Association 250: 2510-2513.

Yerushalmy J (1947). ***Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques.*** Public Health Reports 62: 1432-1449.

Youden WJ (1950). ***Index for rating diagnostic tests.*** Cancer 3: 32-35.

10

La subjetividad se disfraza

Desgraciadamente casi nunca la realidad parece tener en cuenta nuestros deseos y esa melancólica tendencia es la causa de innumerables desconsuelos.

ERNESTO SÁBATO

Cuando no se han buscado - o no ha sido posible obtener- datos que permitan un conocimiento sustantivo de la realidad, no hay modo de remediar tal carencia con filigranas estadísticas, por refinadas o ingeniosas que sean. Esta advertencia es pertinente porque algunos investigadores, para dar respuesta a sus preguntas, actúan como si los procedimientos cuantitativos -en especial, los estadísticos- pudieran suplir las deficiencias de la muestra o el diseño y, en ocasiones, relevarles del uso de su cultura y de sus conocimientos sobre el tema que abordan.

Se supone que, al aplicar cierto procedimiento estadístico a un conjunto de datos, lo que se procura es que el análisis gane en objetividad; es decir, que la sensibilidad personal del analista, en particular sus prejuicios y deseos, no puedan modificar sustancialmente las conclusiones. De modo que cuando el instrumental estadístico es usado como máscara para, a la postre, avalar ideas preconcebidas o respaldar resultados obtenidos mecánicamente, lo que realmente se consigue no es objetivizar los procesos de análisis sino, en el mejor de los casos, disfrazarlos de objetividad [1].

La-verdad es, sin embargo, que ni los métodos estadísticos garantizan automáticamente la objetividad, ni las aproximaciones a la realidad protagonizadas por métodos no cuantitativos constituyen necesariamente recursos metodológicos de segunda clase (véase Sección 10.6), aunque estos últimos sean con frecuencia desdeñados por la relativa subjetividad que los caracteriza.

En muchas ocasiones la perversión no se produce por razones espurias (manipulaciones deliberadas) sino porque, en lugar de asumir con las debidas madurez y

[1] Un caso típico se produce con la determinación del tamaño muestral; a él, por su singularidad y relevancia, se destina íntegramente el próximo capítulo.

flexibilidad los recursos cuantitativos formales, algunos investigadores se sienten obligados a dar respaldo a sus conclusiones mediante tales tratamientos y desembocan en el uso rígido y acrítico de los métodos.

En aras de conseguir una productividad científica real, considero necesario tomar conciencia del verdadero alcance de los métodos estadísticos, no atribuirles más méritos de los que efectivamente tienen y aprender a prescindir de ellos cuando no representen otra cosa que un formalismo estéril. Inversamente, procede aquilatar en su cabal sentido el papel que ocupan esos procedimientos, e identificar las circunstancias en que su utilización dista de ser un adorno para convertirse en clave ineludible del examen objetivo de la realidad que se explora o estudia.

Este capítulo no se propone, naturalmente, ofrecer un inventario exhaustivo de los mecanicismos estadísticos en que se incurre para conferir apariencia de objetividad a los trabajos ni, mucho menos, de las omisiones metodológicas en materia estadística que abonan el subjetivismo; procura, sin embargo, hacer algunas reflexiones generales e ilustrar ambas caras de esa moneda. Se pretende dar así un alerta que cada lector podría contemplar provechosamente en sus circunstancias concretas.

10.1. Una fábrica de publicaciones

Las consecuencias de manipular la obtención de conocimientos mediante el manejo formal de procedimientos estadístico-computacionales pueden llegar a tener bastante entidad. El artículo de Vanderbroucke (1990), incisivo epidemiólogo holandés, titulado ***¿Cuán fiable es la investigación epidemiológica?, se*** cuestiona duramente el ***modus operandi*** de los epidemiólogos [2] ante los datos que emergen de los estudios prospectivos a gran escala, considerados como los más prestigiosos o confiables después de los experimentales.

En este tipo de trabajos se recolectan cientos de datos referentes a miles de personas, enormes series de observaciones que son entonces examinadas mediante técnicas automáticas de descripción. Unas pocas horas escudriñando las salidas informáticas iniciales producirán un bosquejo de análisis orientado a buscar conexión entre muy diversos puntos de partida (subpoblacionales de interés) y muchos puntos de llegada (resultados a partir de los cuales, típicamente, se estiman tasas para aquellas subpoblaciones). Una nueva sesión con la computadora conducirá a desechar un buen número, quizás la mitad, de los resultados, por carecer de interpretación razonable. El examen ya más detenido y formal del material remanente siempre dará lugar a unas cuantas asociaciones promisorias.

[2] No se refiere a situaciones excepcionales. «Quien no lo haya hecho alguna vez, que tire la primera piedra», escribe el autor.

En ese punto sólo resta aplicar el maquillaje final. Se elegirán puntos de corte adecuados, se dicotomizarán variables cuyo carácter continuo original no sea suficientemente expresivo, se aplicarán las pruebas de significación más productivas y se elaborarán las hipótesis que hagan falta, las cuales se verán luego «confirmadas» por los mismos datos de las que nacieron.

No se han inventado datos ni se han falseado análisis. Sólo se ha acicalado cuidadosamente el resultado de cribar la montaña inicial de datos hasta conferirle un aspecto publicable. Comentando el tema, Marshall (l990) señala que la trascendencia real de un hallazgo es inversamente proporcional a la intensidad con que se usó la maquinaria estadística que lo produjo.

«Esta estrategia», resume Vanderbroucke, «convierte tales estudios en fábricas semirrobotizadas de publicaciones». A mi juicio, el problema mayor radica en que el artículo que regularmente sintetizará este proceso, lejos de comunicar el orden verdadero de los hechos, presentará una hipótesis, luego el método seguido para evaluarla, a continuación los resultados que la corroboran y, finalmente, la conclusión que la consagra como verdad.

Ante tal realidad, la primera pregunta que surge naturalmente es: ¿constituye la formulación *previa* de una hipótesis una condición inexcusable para que las observaciones que la avalan deban ser tenidas en cuenta? La respuesta en mi opinión es negativa: sería absurdo recomendar que ante un resultado llamativo se actúe como un avestruz. De hecho, el llamado ***análisis exploratorio de datos,*** asentado sobre procedimientos gráficos y técnicas informales, es una lúcida propuesta (Tukey, 1977) para el examen razonado de los datos en procura de rasgos o patrones que pueden operar como guía para la acción confirmatoria (Cobos, 1995). Y, más en general, cabe preguntarse: ¿carecen de todo valor los análisis secundarios de este tipo? Mi respuesta es nuevamente negativa, siempre que se satisfagan dos exigencias: una, que a los destinatarios se les informe íntegramente del procedimiento seguido y no una parte de él[3]; otra, que todos los resultados (tanto «negativos» como «positivos») sean comunicados.

10.2. Ignorancia paso a paso

Como es bien conocido, el análisis de regresión lineal múltiple tiene como propósito general expresar una magnitud dada (variable dependiente) como función de otras, calificadas como ***explicativas, regresoras o predictoras,*** según el contexto, y también genéricamente llamadas ***variables independientes.***

Con frecuencia se procura incorporar en esta condición dentro del modelo a

[3] A este respecto cabe recordar, por más señas, la conocida afirmación de Fisher (1934) de que «es bien sabido en estadística que la interpretación de un cuerpo de datos requiere el conocimiento de cómo fue obtenido».

aquellas variables que, en principio, se consideren capaces de «explicar» el comportamiento de la variable dependiente.

Si para *n* sujetos, se observaron ***k*** variables independientes $X_1, X_2, \ldots, X_k$, además de la variable dependiente *Y*, la información puede ponerse en forma matricial del modo siguiente:

$$\begin{matrix} Y_1 X_{11} X_{12} \ldots X_{1k} \\ Y_2 X_{21} X_{22} \ldots X_{2k} \\ \ldots \\ Y_n X_{n1} X_{n2} \ldots X_{nk} \end{matrix}$$

El propósito central es encontrar valores $\beta_0, \beta_1, \ldots . \beta_k$ tales que pueda conformarse una función de la siguiente forma:

$$\mathbf{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Una vez realizado el ajuste, se obtienen las estimaciones $\beta_0^*, \beta_1^*, \ldots . \beta_k^*$ y para cada sujeto puede computarse el valor estimado de *Y*; es decir:

$$Y_i^* = \beta_0^* + \beta_1^* X_{i1} + \beta_2^* X_{i2} \cdot \cdot . + \beta_k^* X_{ik}$$

Suele entonces calcularse el llamado ***coeficiente de determinación R^2,*** definido como:

$$R^2 = \frac{S_R}{S_{YY}}$$

donde $S_R = \sum_{i=1}^{n} (Y_i^* - \bar{Y})^2$ y $S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, con $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Se puede demostrar formalmente que R^2 ***es*** un número ubicado necesariamente entre 0 y 1. Cuanto más próximo esté de la unidad, mayor es la calidad del ajuste; o sea, R^2 es una medida conjunta de la cercanía de los *n* vectores de la muestra al hiperplano ajustado. Si, por ejemplo, $R^2 = 0{,}78$, suele decirse que las variables X_1, $X_2, \ldots . X_k$ «explican» el 78% de la variabilidad de *Y*.

El uso de este verbo, sin embargo, ha dado lugar a una peligrosa y falsa creencia sobre las virtudes de la regresión. ***Explicar*** algo, según el diccionario de la Real Academia Española, es ***dar a conocer la causa de alguna cosa.*** Pero ocurre que las variables independientes no tienen la cualidad de dar a conocer causa alguna; ni mucho menos de desentrañar la urdimbre causal que podría vincular a la variable dependiente con las independientes, como la expresión puede hacer creer a muchos lectores (y, lamentablemente, también a muchos usuarios de la técnica). La crítica más persuasiva e interesante que conozco al respecto es la que desa-

rrolló King (1986) en el contexto de un atractivo artículo titulado ***Cómo no mentir con la estadística.***

Aunque tal advertencia ha sido reiterada por connotados estadísticos, este uso confuso y la subsiguiente interpretación errónea no dejan de aparecer en la literatura. La fábula estadística que se expone en la Sección 2.5 da cuentas con elocuencia del riesgo implícito en este abuso de lenguaje.

Por otra parte, el uso ingenuo de la regresión como recurso «explicativo» tiene un aliado muy activo en los algoritmos usados para llevar adelante la llamada ***selección de variables,*** proceso mediante el cual se eligen algunas de las variables independientes iniciales para conformar un modelo más simple.

El más conocido de ellos es el llamado ***método paso a paso (step- wise method).*** De hecho hay varias modalidades de aplicación; las dos más notables son:

a) Ir incorporando variables al modelo ***(step up).***
b) Ir eliminando variables de él ***(step down).***

La idea básica de ambos procedimientos consiste en construir de manera gradual el modelo con el que a la postre nos quedamos. En la variante (a), por ejemplo, se comienza el proceso tomando una de las variables y se procede a examinar mediante una prueba de significación si la incorporación al modelo de regresión de alguna de las otras variables consigue aumentar el valor de R^2 en una cuantía apreciable. De ser así, se intenta incorporar una tercera, y así sucesivamente, hasta que el «aporte explicativo» que haga una variable llegue a ser despreciable respecto a lo que «explican» las que ya se han incorporado. De tal suerte, al finalizar el proceso queda dentro del modelo el subconjunto de las variables originalmente consideradas que maximiza el valor de R^2 con la restricción de no hacerlo más complejo con una adición cuyo aporte sea muy escaso.

En la práctica no es inusual que se ajuste un modelo de regresión múltiple y de inmediato se aplique un procedimiento como éste para determinar cuáles variables han de «quedarse» en calidad de factores detectados como verdaderamente influyentes y cuáles habrán de despreciarse. Algunos paquetes informáticos contemplan incluso la posibilidad de proceder directamente al ajuste de un modelo por esa vía.

El problema está en la pretenciosa y a la vez ingenua interpretación que suele hacerse del resultado, la cual puede bosquejarse como sigue: las variables que se «quedan» dentro del modelo final son las ***responsables (o*** principales responsables) de las modificaciones que experimenta la variable dependiente; las que «salen», o bien no influyen causalmente en el proceso, o su influencia no es apreciable. Como señala Evans (1988), es muy grande la tentación de usar un nuevo método antes de constatar su pertinencia, por el solo hecho de que está disponible. Las víctimas de esa trampa, especialmente en la «era cibernética», no son pocas (aunque una visita al ***MEDLINE*** persuadirá al lector de que el empleo de esta técnica es mínima).

Se ha dicho (Guttman, 1977) que «el uso de la regresión paso a paso es en la actualidad una confesión de ignorancia teórica sobre la estructura de la matriz de correlaciones».

Comparto totalmente esta opinión. Cuando la regresión múltiple se usa para descubrir los patrones de causalidad según los cuales ciertas variables actúan sobre otra, la regresión paso a paso equivale a cubrir esa ignorancia con un algoritmo que piense por el investigador. No en balde el procedimiento de ***stepwise regression*** fue rebautizado irónicamente (Leamer, 1985) como ***unwise regression*** (juego de palabras intraducible que aprovecha que el vocablo ***wise*** denota en inglés la *manera* o el ***modo*** de hacer algo pero también significa ***sabio,*** de modo que ***unwise regression*** vendría a ser algo así como ***regresión torpe).***

Cabe advertir, sin embargo, que cuando la técnica no se usa con finalidad explicativa sino pronóstica, o con la de estimar los valores que tomará la variable dependiente, parece enteramente razonable buscar el modelo más sencillo, y no hay razones para objetar el uso de algoritmos para hallarlo. En ese caso es irrelevante que ocurran cosas como que una variable que pudiera ser «directamente causal» resulte eliminada o quede suplida por una o más variables que no tengan influencia real alguna pero que se vinculen con aquella de manera que la predicción sea en definitiva eficiente, o por lo menos aceptable.

Por otra parte, no hay ningún argumento que permita suponer que ir paso a paso «hacia abajo» sea mejor o peor que hacerlo «hacia arriba». Tal circunstancia nos permite identificar un claro e inapelable indicio de la improcedencia de depositar en un algoritmo como la regresión paso a paso la tarea de ***explicar*** la realidad, ya que es fácil corroborar que al usar uno u otro método, las variables que conforman el modelo final suelen no ser las mismas. Tal desempeño de la regresión paso a paso está bien documentado en la literatura; por ejemplo, McGee, Reed y Yano (1984) exponen detalladamente un ejemplo basado en la regresión logística, en el que cada uno de tres procedimientos de selección producen resultados finales que son drásticamente diferentes entre sí.

El ejemplo sirve para recordar que esta técnica no distingue entre las asociaciones de índole causal y las debidas a terceros factores involucrados en el proceso. Mucho menos aun sirve para distinguir asociaciones causales de las observadas como consecuencia de un sesgo en el estudio. Por otra parte, ***las variables que se retienen en el modelo dependen, como es obvio, de las que originalmente se hayan incluido en él,*** elección ésta que suele padecer de un alto grado de subjetividad.

Finalmente, este proceso de selección está firmemente asentado sobre las pruebas de significación; por lo tanto, su validez está sujeta a todas las suspicacias que ellas despiertan (véase Capítulo 6). En particular, cuando la muestra es grande, una variable puede quedar incluida por el método, aunque su sustantividad clínica o biológica sea nimia.

Ramón y Caja1 advertía que pensar sin observar es un error, pero que hacer lo contrario, observar sin pensar, es igualmente peligroso. En consonancia con tal

advertencia y a modo de resumen, sostengo la opinión de que los procedimientos algorítmicos para seleccionar variables relevantes carecen de interés (y, más aun: deben evitarse) salvo que el modelo de regresión múltiple se utilice con finalidad pronóstica. En este caso, lo que se desea es contar con un recurso que permita anticipar el valor de cierta variable conocidos los de otras que se le asocian; poco importa si una de ellas desempeña un papel sustantivo en el proceso causal o es simple reflejo de otra que sí lo desempeña siempre que su contribución a la obtención de un vaticinio adecuado sea eficaz.

10.3. Bajo el barniz multivariado

Algunos procedimientos multivariados reflejan con particular elocuencia el fenómeno de la subjetividad estadísticamente disfrazada. Consideremos con cierto detalle el caso del análisis factorial *(factor analysis),* una de las técnicas multivariadas más ingeniosas, atractivas y elegantes de cuantas se han creado.

En su sentido más general, el método sirve, supuestamente, para explorar un conjunto de variables y, en definitiva, entender mejor la realidad de la que proceden. Para examinar detenidamente los problemas que le son inherentes, es necesario repasar sus propósitos y características específicas.

El nacimiento de esta técnica, también conocida como ***análisis factorial de correlaciones, se*** remonta a la primera década de nuestro siglo y se inicia con un extenso artículo en el que Spearman (1904) encara el problema de «determinar y medir objetivamente la inteligencia». Como se ve, la pretensión de «objetividad» estaba, curiosamente, en la raíz del asunto. Aquella fue la época en que surgió el primer recurso formal para medir capacidades intelectuales, la prueba de Binet, que en su momento dio vida a una polémica aún vigente (véase la Sección 5.6).

Como en casi todos los procedimientos multivariados, se parte de un conjunto de *k* variables $X_1 X_2 \ldots X_k$, mediante las cuales se podría caracterizar a una unidad específica. Por ejemplo, las unidades podrían ser países y las variables reflejar rasgos de interés a los efectos de caracterizarlos sanitariamente (tasa de mortalidad infantil, porcentaje de la población con agua potable, número de médicos por 1.000 habitantes, etc.).

El propósito es redimensionar la información que dichas variables abarcan, de manera que pueda ser descrita a través de un número menor (digamos ***p,*** donde ***p < k)*** de ***factores.*** En ese sentido, el análisis factorial de correlaciones es similar al ***de componentes principales.*** Sin embargo, su finalidad última tiene honda connotación explicativa y no sólo descriptiva como la de aquel; la esencia del procedimiento se bosqueja a continuación.

Se trata de hallar ***p*** índices $F_1, F_2, \ldots F_p$ ***(factores comunes)*** mutuamente no correlacionados, con los cuales construir las funciones siguientes:

$$X_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1p}F_p$$
$$X_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2p}F_p$$
$$\ldots$$
$$X_k = a_{k1}F_1 + a_{k2}F_2 + \ldots + a_{kp}F_p$$

Conceptualmente, los F_i son ciertos factores no susceptibles de ser directamente medidos pero que admiten alguna interpretación cualitativa (abstracciones tales como «higiene comunitaria», «mortalidad evitable» o «recursos sanitarios»).

Idealmente hay dos alternativas: el análisis factorial *confirmatorio,* en que teóricamente ya se conocen cuántos factores existen, y el *exploratorio,* en el que el número ***p*** no se conoce y ha de fijarlo el analista (aunque tampoco han faltado procedimientos algorítmicos para determinarlo automáticamente). La realidad es que la variante exploratoria es la única que, con excepción de algún ejemplo de museo, se usa en la práctica.

Los valores $a_{i1}, a_{i2}, \ldots . a_{ip}$, que han de cumplir la condición $|a_{ij}| \leq 1$, son las llamadas ***cargas factoriales (factor loadings)*** para la i-ésima variable. Lo que se espera es conseguir que los valores $|a_{ij}|$ sean, o bien muy pequeños (próximos a cero) o bien muy grandes (próximos a 1), de manera que cada variable original dependa fuertemente de alguno o de pocos factores, y que los factores de los que dependa sean diferentes, en lo posible, para las distintas variables (véanse detalles, por ejemplo, en Manly, 1994).

El proceso de determinación de la forma definitiva de las ecuaciones supone obtener primero unos factores iniciales o provisionales; con ellos se construyen entonces nuevos factores (combinaciones lineales de los iniciales). Al proceso de determinación de los coeficientes que corresponden a tales combinaciones, que son las nuevas «cargas factoriales», se le denomina ***rotación factorial.***

Hay varios procedimientos para extraer los factores iniciales, y otros varios para generar la rotación. Por ejemplo, el paquete de programas estadístico BMDP (Dixon, 1990) permite elegir, por una parte, entre cuatro modos de seleccionar variables y, por otra, entre ocho variantes para hacer la rotación. Vale decir, para los mismos datos originales, que este programa ofrece 32 procedimientos. Al desarrollar una aplicación concreta, el usuario tiene que seleccionar cuál de ellas habrá de utilizar. El problema que tal variedad de posibilidades genera es que no se cuenta con un criterio razonable para hacer una elección. Ésta es una encrucijada inevitable, y ocurre que las opciones posibles ¡arrojan 32 resultados diferentes!

Por añadidura, una vez en posesión de los factores que dimanen de todo este andamiaje, corresponde construir una interpretación para cada uno de ellos: darle un nombre y un sentido cualitativo. Tanto el bautizo como la decodificación conceptual constituyen un acto de imaginación, a veces exuberante.

No me alarma, desde luego, que la elaboración creativa del investigador intervenga en el proceso. De un modo u otro, ello es siempre inevitable. Lo malo es que, aplicado el elegante método de análisis factorial, cada autor conferirá, para los

mismos resultados, su personal significado cualitativo a cada uno de los factores, como si se tratara de interpretar un pasaje bíblico.

Mulaik (1987) realizó una profunda reflexión sobre las fuentes filosóficas en que descansa la teoría del análisis factorial de correlaciones. Así, halla raíces diversas que van desde Aristóteles, quien preconizaba la inducción y la búsqueda de factores comunes en los fenómenos que quería explicar, hasta los grandes estadísticos empiricistas, especialmente Pearson y Yule, padrinos de la exploración descriptiva como punto de partida para la inducción. Es precisamente al amparo de esta concepción de ingenuo inductivismo como nace el análisis factorial de correlaciones. La convicción de que las inferencias inductivas pueden producir resultados únicos e inequívocos sin que sea necesario imponer supuestos previos es conocida como la «falacia inductivista» (véase Chomsky y Fodor, 1980); en ella incurren, conscientemente o no, casi todos los que aplican el análisis factorial de correlaciones.

En un libro destinado a explicar y difundir esta técnica (Yera, 1967), el propio autor, refiriéndose a los múltiples derroteros internos que pueden seguirse en su aplicación (32 según BMDP), escribe textualmente:

> ***Debo advertir que los resultados obtenidos por los distintos métodos son en el fondo equivalentes y varían tan solo en la interpretación que los diversos autores les dan, de acuerdo con su postura científica y con su filiación filosófica*** [4].

Se ha señalado que la regla de oro del investigador es la ***máxima audacia en la formulación de hipótesis y la extrema cautela al sacar conclusiones.*** El espíritu del análisis factorial parecería ser exactamente el opuesto: la cautela en la formulación de hipótesis es total puesto que, simplemente, no se formulan; las conclusiones se sacan con total desenfadado una vez que se está frente a los resultados.

Otras críticas de naturaleza similar han sido señaladas por varios autores. Chatfield y Colin (1980), por ejemplo, desarrollan seis objeciones de este tipo y sugieren, simplemente, no usarlo. Seber (1984) destaca su esterilidad apoyándose en la simulación. Guttman (1977) resume drásticamente la situación así:

> ***No quedan dudas de que tras 70 años de «exploración» y «confirmación», los libros de texto sobre análisis factorial aún no presentan un solo ejemplo, en ningún área de la ciencia, de una ley empírica bien establecida sobre la base del análisis factorial.***

[4] El subrayado es mío y señala que el autor viene a decir que las alternativas son equivalentes salvo en lo único que, en definitiva, interesa: la interpretación; dudo que pueda haber un texto más expresivo de cómo la subjetividad reina detrás de todo el andamiaje «objetivo» del procedimiento.

10.4. El azar y la paradoja muestra1

Es bien conocida la demanda de los metodólogos en el sentido de que las muestras sobre las que habrán de reposar las inferencias sean elegidas haciendo uso del azar. Dos preguntas emergen de inmediato: ¿se trata de algo legítimo, o estaremos frente a otro reclamo dogmático? Y, en caso de que haya una razón de peso, ¿cuál es realmente su fundamento?

La pertinencia de esta exigencia está fuera de toda discusión cuando se trata de estudios experimentales. La discusión es, sin embargo, válida para el caso de muestras en estudios descriptivos y, en general, observacionales.

Generalmente, cuando se procura caracterizar cierta realidad, el uso del azar es no sólo procedente sino que resulta cardinal e insustituible. Sin embargo, como suele ocurrir con toda afirmación categórica relacionada con aspectos metodológicos, la exigencia ciega de incorporarlo es blanco fácil de la crítica y eventual objeto de ridiculización. Aunque son ciertamente excepcionales, se dan situaciones en las que el empleo del azar durante el proceso muestra1 sería contraproducente.

Imaginemos, por ejemplo, que se ha decidido elegir un mes del año para estudiar un servicio de urgencia hospitalaria con el fin de realizar una descripción del tipo de morbilidad que éste atiende, y que se ha decidido hacerlo a partir del examen de todos los pacientes que llegan a dicho servicio a lo largo de ese mes. Desde luego, la restricción de circunscribirse a un solo mes es altamente cuestionable, pero no es imposible que haya sido impuesta por alguna coyuntura administrativa.

El hecho de que los meses exhiben un patrón de morbilidad variable y el de que los pacientes que acuden durante determinado mes son preponderantemente portadores de algún tipo específico de dolencia, nos persuadirá de que la elección de ese mes por vía del azar sería absurda. En lugar de arriesgarnos a que resulte seleccionado un mes de invierno (típico generador de influenzas) o uno de vacaciones (con reducida afluencia de pacientes y mayor número de accidentes) sería más recomendable elegir *racionalmente* el que mejor refleje o represente la gama de situaciones diferentes que generan la demanda de atención, que quizás es aquel mes en que la variabilidad de causas es mayor. Tal opción es altamente intuitiva, pero además se puede fundamentar teóricamente (véase Silva, 1993).

Suele creerse que la razón subyacente para exigir el uso del azar dimana de que, por ese conducto, se estaría *inyectando representatividad* a la muestra. Así dicha, la afirmación entraña un error conceptual y tiende a consolidar una convicción falsa. En efecto, es fácil convencerse de que la selección aleatoria de la muestra no garantiza en modo alguno que se consiga representatividad. No obstante, la introducción del azar en la selección muestra1 es un requisito de la buena práctica de investigación. A la fundamentación de ambos extremos se destinan las secciones que siguen.

10.4.1. Los cinco mejores alumnos se eligen al azar

Imaginemos la siguiente situación. Un inspector escolar tiene que evaluar el trabajo de cierto maestro y, entre otras acciones, ha de examinar a 5 niños de los 30 que integran su alumnado. Para decidir a cuáles hará el examen, pide al maestro que éste señale 5 alumnos concretos. Hecha la evaluación, obtiene magníficos resultados, que incluye en el informe que rinde a sus superiores. Cuando explica el método de selección utilizado, recibe la recriminación del caso: se le dice que, verosímilmente, el maestro ha elegido a los 5 mejores y que para que la valoración de la gestión docente sea correcta este método no debe ser utilizado en lo sucesivo. La próxima vez deberá escribir los nombres de los 30 alumnos en respectivas tarjetas de las cuales, una vez mezcladas concienzudamente, ha de elegir cinco. Así se determinará la muestra de alumnos que serán examinados.

Tres meses más tarde, nuestro inspector regresa al aula y hace lo que se le ha dicho. Pero he aquí que la muestra elegida al azar resulta estar formada por los mismos cinco alumnos que había sugerido el maestro el trimestre anterior [5].

El inspector presenta su informe. Su jefe, atento a la situación generada tres meses atrás por la torpeza del inspector, repara en que la muestra ha sido exactamente la misma. Sin embargo, no puede objetarla, ya que esta vez el inspector ha seguido estrictamente sus orientaciones. La paradoja se concreta en que, según esa lógica, ***la misma información*** puede o no ser válida según el modo en que se obtuvo. Es como si sólo pudiéramos pronunciarnos sobre la calidad de un poema en caso de haberlo leído en un libro, pero que ese mismo juicio no fuera válido en caso de haberlo leído en una revista.

Para discutir el asunto con más rigor, es preciso detenerse en el concepto de representatividad. En su momento volveremos con el inspector.

10.4.2. Representatividad: un concepto esquivo

La noción de representatividad es esencialmente intuitiva. No existe una definición formal, ni siquiera una definición operacional que permita -ante una muestra dada- determinar si ella es o no representativa de la población. Alrededor de esta noción se producen no pocas confusiones. El epistemólogo y salubrista argentino Juan Samaja plantea (Samaja, 1994):

> ***La pata de una silla es una parte de una silla, pero no una muestra de la silla. Las personas que están sentadas en la fila 5 de un cine son una parte de la concurrencia al cine, pero no una muestra. En cambio, una foto de la silla es una muestra de ella, del***

[5] Obviamente la probabilidad de que sea esa la muestra que resulte elegida es muy baja. Es igual, concretamente, al inverso del número combinatorio C_5^{30}, cifra que asciende (más bien, desciende) a 0,000007. Pero eso no ha de importarnos puesto que lo relevante ahora es que tal hecho es perfectamente posible.

> ***mismo modo que un grupo «construido» mediante la selección aleatoria, de una cantidad suficiente de concurrentes al cine para que el atributo que estudiamos tenga una probabilidad alta de quedar representado, es una muestra de la concurrencia al cine (...) Si alguien dijera que un «esclavo negro» es un «hombre de la raza negra», estaría cometiendo el mismo error que el que afirma que*** una ***muestra es un subconjunto de un conjunto mayor: No basta que haya un hombre de la raza negra: se necesita, además y de manera esencial, que se den las relaciones sociales que hacen que un hombre sea esclavo. Análogamente, para que algo*** sea ***una «muestra» no basta que «unos» elementos sean un subconjunto de un conjunto: deben darse ciertas relaciones de semejanza entre un todo y otro todo para que uno de ellos pueda ser considerado una muestra del otro.***

Varias acotaciones cabe hacer a este texto. En primer lugar, al decir que la pata de una silla es una parte de la silla pero no una muestra de ella, se está usando un símil algo tendencioso, ya que el concepto de «silla» es algo integral e indivisible. Sin embargo, la madera de que se compone esa pata puede ser una muestra de la que se usó para construir la silla. Por otra parte, los que se hallan en la fila quinta del cine sí constituyen una muestra de los asistentes. Esa muestra puede no ser útil, no ser «buena» o no ser «representativa» (¡aunque también pudiera serlo!), pero no deja por ello de ser una muestra [6]. Es metodológicamente inconveniente que el concepto de muestra lleve implícito rasgos deseables para ella, del mismo modo que no procede exigir que una ecuación, para serlo, deba tener solución.

En lo que concierne al ejemplo del esclavo, su improcedencia estriba en que se puede establecer con todo rigor cuáles han de ser las relaciones sociales que le confieren a ese hombre negro tal condición, pero ante una muestra concreta, como se ha dicho, no es posible determinar si es o no representativa de la población.

Es imposible, en primer lugar, por la ya señalada naturaleza intuitiva (no formalizable) del concepto de representatividad; en segundo lugar, porque ésta es una noción relativa: una muestra concreta puede representar a la población a unos efectos y a otros no; y, en tercer lugar porque, usualmente, no se conoce la población que se quiere modelar mediante la muestra y es por ende imposible el cotejo entre ambas [7]: es ese desconocimiento, precisamente, el que genera la necesidad de obtener la muestra.

Para la aplicación fecunda de la estadística en general, y para discutir este problema particular, es imprescindible ***no confundir el método usado para resolver un problema con el resultado de haberlo aplicado.*** Esta confusión entre método y resultado trasciende, ciertamente, el ámbito de las muestras. Por ello lo abordaremos con otro ejemplo.

---

[6] Por lo demás, como se enfatizó en la sección precedente, la selección aleatoria bien pudiera tener como resultado a los que ocupan la fila 5 del cine.

[7] Se sobrentiende que me refiero al cotejo ***respecto de los parámetros que, precisamente, se quieren conocer a través de la muestra;*** otras comparaciones quizás sí puedan realizarse con fines orientativos.

10.4.3. Un estadígrafo esotérico

Consideremos la situación en que se quiere estimar la media poblacional correspondiente a cierta variable. Por ejemplo, supongamos que se desea estimar la talla media de todos los niños de 6 años de Cuba, parámetro al que llamaremos μ. Supongamos que esa tarea se encara a través de una muestra [8] de n niños de esa edad y que sus tallas se representan por $x_1, x_2, \ldots . x_n$. Con esos datos hay que usar un ***estadígrafo*** [9] en calidad de estimador. Naturalmente, el primero que viene a la mente es la media aritmética de esos *n* números:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Pero éste no es necesariamente el único que sirve para el propósito señalado. La mediana, por ejemplo, pudiera ser equivalente a la media [10]. No es correcto decir, entonces, que la media muestra1 es «el» estimador de la media poblacional; puede y debe decirse que es «un» estimador de dicho parámetro.

Ahora, supongamos que en lugar del promedio de los *n* valores se sugiere aplicar el siguiente estadígrafo:

$$\psi = \sqrt{n^3 \sum_{i=1}^{n} ln(x_i + 1) + \frac{n}{\pi}}$$

¿Por qué no aceptar que ψ es también «un estimador» de la media poblacional? En principio no hay ninguna razón para no admitirlo. Nótese que, a pesar de su insólito aspecto, no es imposible que para alguna muestra particular se tenga incluso una estimación perfecta: $\psi = \mu$.

Sin embargo, cualquier persona con sentido común estará más dispuesto a admitir que el promedio muestra1 sea un «buen» estimador de μ a que lo sea ψ. ¿Por qué? Se han elaborado ciertos *criterios* (tales como el de insesgamiento, mínima varianza, consistencia, admisibilidad, suficiencia, etc.) que los estadísticos matemáticos consideran deseables y utilizan para evaluar la bondad de un estimador.

La media aritmética y la mediana satisfacen algunos de esos criterios (es difícil hallar un estimador que los cumpla todos), pero el que se acaba de inventar, posi-

---

[8] Podrían usarse otros métodos, tales como hacer un censo y medir a todos los niños, preguntarle su opinión a un pediatra o, incluso, a una cartomántica. Cada uno puede tener sus ventajas y desventajas; el primero, por ejemplo, puede ser prohibitivamente caro; el último tiene la ventaja de ser muy barato pero el inconveniente de no encajar en los cánones de indagación admitidos por la comunidad científica.

[9] Algunos le llaman «estadístico».

[10] En la mayor parte de las situaciones lo es; y hay algunas en que la mediana resulta ser una medida de tendencia central más expresiva. Véase la Sección 2.4.1.

blemente, no cumpla ninguno. Ello habilita a considerar que ψ no es un «buen estimador». Es decir, puede considerarse como un estimador, pero el sentido común de los estadísticos -consagrado en criterios que nadie razonable discute- niega en cambio el derecho a considerarlo «bueno». Del mismo modo que nadie puede discutir que el horóscopo es un método para pronosticar el futuro, aunque sobren motivos para conceptuarlo como ineficaz.

La talla media de la muestra, por su parte, quizás sea muy parecida (o, en algún caso excepcional, igual) a la del universo, pero, por la propia naturaleza del problema, no es posible confirmarlo. Puede ocurrir por otra parte que así sea, pero que el porcentaje de sujetos con sangre tipo AB en el subconjunto sea muy diferente al del universo: ¿es este subconjunto acaso una muestra el lunes -que es el día en que se examinan las tallas- y deja de serlo el martes, cuando le toque el turno al estudio hematológico? Según el texto de Samaja citado en la sección precedente habría que convenir en que ese subconjunto puede ser y a la vez no ser una muestra.

10.4.4. Distinguir la muestra del método muestra1

La clave del asunto consiste en que la calidad de una muestra, o la confianza que se pueda depositar en ella, dimana exclusivamente de la calidad *del método* usado para obtenerla, o de la confianza que él merezca. En la práctica, lo único que puede legítimamente hacerse es aplicar un método que, por el modo de hacer la selección y por el tamaño del subconjunto elegido, produzca un alto porcentaje de muestras «buenas», de manera que resulte muy probable que la que nos toque en suerte pertenezca al subgrupo específico de muestras posibles que tienen ese atributo. De ser así, diríamos informalmente que *el método es* confiable, aunque no conozcamos ni las virtudes ni los defectos de uno u otro de sus posibles resultados.

Más generalmente, la confianza que se tiene en un método (ya sea estadístico, ya sea un método para aprender a mecanografiar, o para vender enciclopedias a domicilio) no reposa en que se sepa o se crea que *siempre es* efectivo, sino en que lo es en una magnitud o una porción suficientemente alta de aplicaciones como para que lo más práctico sea usarlo, aun bajo el riesgo de que «nos toque» una de aquellas de sus manifestaciones en que funciona mal [11].

Inversamente, no se desecha un método dado porque sepamos que *siempre* es inválido o ineficaz; prescindimos de él cuando sea tan probable que no produzca el dividendo deseado, que resulte desmedido el riesgo de que nos toque un resultado indeseable en lugar de la posible excepción en que funciona.

Por ejemplo, para bajar desde un cuarto piso, regularmente usamos el ascensor

[11] Ese no es el caso, claro está, de métodos canónicos como podría ser el de Sarrus para calcular el determinante de una matriz. Nos referimos, obviamente, a cualquier método cuyo desempeño pueda depender en alguna medida del azar.

en lugar de lanzarnos por el balcón, recurso éste ciertamente más rápido. Esa elección no se debe a que sea imposible terminar felizmente la experiencia voladora (existen antecedentes que así lo confirman) sino porque la inmensa mayoría de las veces el resultado será trágico. Análogamente, el uso del ascensor puede malograr nuestro propósito de llegar rápidamente (como consecuencia, por ejemplo, de una rotura o de un corte de fluido eléctrico), pero no nos abstenemos de utilizarlo, porque lo regular es que el viaje se verifique sin dificultad [12].

La dinámica que rige el uso de muestras es exactamente igual a la descrita; para fijar ideas, consideremos el siguiente problema concreto. Supongamos que se desea estimar el porcentaje de seropositividad a cierto virus en una población de $N = 3.000$ individuos y que se valoran dos métodos muestrales alternativos:

a) Seleccionar una muestra simple aleatoria de tamaño $n = 500$.
b) Tomar una muestra, también al azar, de tamaño $n = 2$.

Supongamos por un momento que exactamente la mitad de los 3.000 individuos posee el anticuerpo en cuestión; es decir, que el porcentaje desconocido P asciende realmente al 50%. Una muestra de tamaño $n = 2$ puede contener exactamente un sujeto con anticuerpos y otro sin ellos, en cuyo caso la estimación será perfecta [13]. De hecho, así ocurrirá con más de la mitad de las muestras posibles a la vez que se producirá sólo excepcionalmente con la de $n = 500$. Sin embargo, dudo que nadie con sentido común elija, consideraciones económicas aparte, el método (b) antes que el (a). Elegiremos sin dudas este último, pero no porque sepamos que la muestra va a ser necesariamente representativa a los efectos de los anticuerpos, sino porque es muy baja la probabilidad de que resulte objetivamente seleccionada una de las poquísimas que conducen a una estimación descabellada del porcentaje de interés. Aunque al optar por (a) no se repara explícitamente en ello, un enorme porcentaje de las muestras posibles en caso de trabajar con el procedimiento (b) dará lugar a estimaciones extremadamente deficientes.

La diferencia básica entre esta situación y la de bajar desde un cuarto piso estriba en que en aquel caso, hecha la experiencia, es posible (e incluso inevitable) enterarnos de cuán eficientemente funcionó el método elegido, en tanto que a través de la propia muestra es imposible evaluar si el resultado es o no eficiente. Por ejemplo, no puede saberse si el porcentaje muestra1 de individuos con el anticuerpo es o no próximo a ese porcentaje en la población; pero ello no es óbice para preferir *el método* de elegir 500 individuos antes que el de seleccionar sólo dos.

[12] Esa misma manera de pensar es la que aconseja no jugarse sistemáticamente el salario a la ruleta. Aunque en este caso el problema se manifiesta de manera suficientemente sutil como para que muchas personas inteligentes caigan en la trampa, el asunto es el mismo.

[13] Y como nota lateral, es interesante reparar en que en una situación como la descrita (n=2, P=50), la probabilidad de que la estimación sea perfecta es estrictamente mayor que 0,5, cualquiera sea la magnitud de N.

Al decir que es imposible evaluar un método a partir del resultado que puntualmente pudo haber arrojado su utilización en cierta ocasión, podría pensarse que se está insistiendo en algo evidente. Sin embargo, no es tan inusual tropezar con este error en la literatura.

Por ejemplo, tiene interés didáctico detenerse a examinar el trabajo de López *et al.* (1992). Estos autores se plantearon dirimir, entre cuatro métodos de muestreo, cuál sería el mejor para estimar diversos aspectos relacionados con la morbilidad registrada en la atención primaria española. Las unidades de muestreo eran días laborables y las unidades de análisis, consultas médicas. Los procedimientos alternativos eran seleccionar:

a) Un día semanal, cambiando el día de una semana a otra.
b) Una semana aleatoria por trimestre.
c) Dos semanas tomadas al azar por trimestre.
d) Once días al azar por trimestre.

Una vez establecidos claramente los cuatro diseños de muestreo, el problema queda bien planteado y su solución constituye un ejemplo de evaluación de tecnologías, que en este caso los investigadores se proponen llevar adelante con datos de 1986. Pero los autores de este trabajo no comparan los métodos sino los resultados que resultan de aplicarlos ***en una*** oportunidad. En sus propias palabras: «se evaluaron los cuatro diseños mediante la comparación de las estimaciones obtenidas por cada muestra con los valores observados para los datos de todo el año», conocidos en este caso excepcional a partir de registros censales.

Constatan así, por ejemplo, que la estimación del porcentaje de consultas de las que se derivan peticiones de laboratorio (12,93%) resultó ser más cercana al verdadero porcentaje (13,12%) cuando se usa el método (d) que cuando se utiliza cualquiera de sus competidores (13,78%, 13,19%, 13,86% para (a), (b) y (c) respectivamente) y concluyen que, a esos efectos, el cuarto procedimiento es el más eficiente.

Esto es exactamente lo mismo que si, en el problema de los anticuerpos, conociendo que el verdadero porcentaje es 50%, para pronunciarnos por la calidad de uno u otro método, usáramos como árbitro los resultados de respectivas muestras de tamaño 2 y 500, y declaráramos mejor la estrategia de usar $n = 2$ si la estimación que esa experiencia produzca está más cerca de 50% que la que haya resultado al usar $n = 500$ [14].

Debe señalarse que los autores no son ajenos al problema; de hecho, escriben textualmente:

[14] Curiosamente, en este caso se puede incluso demostrar que la probabilidad de que salga «victorioso» el método peor (el que usa n=2) es mayor que la de que «gane» el que realmente es mejor. El programa URNA, que simula una situación de este tipo y que se halla en el disco que acompaña el libro de Silva (1993), permite apreciarlo empíricamente.

El trabajo aquí presentado supone una aproximación al estudio de la validez de cuatro diseños muestrales habitualmente utilizados en Atención Primaria, puesto que, en sentido estricto, la determinación de dicha validez requeriría la repetición de numerosas muestras (idealmente todas las posibles) para cada uno de los diseños, calculando en cada caso la media de los estimadores y el sesgo correspondiente.

El problema es que no se trata de una «aproximación» ineficiente sino de que se usa un recurso equivocado. El texto es equivalente a decir: «Aunque mi interés era ver la obra de teatro, como aproximación, obtuve una foto de uno de los actores».

En realidad, por otra parte, no es imprescindible, ni ideal, repetir la experiencia para todas las muestras posibles: lo que hay que hacer es comparar las varianzas de los estimadores, ya sea teóricamente o, al menos, a través de sus estimaciones. En el ejemplo de los anticuerpos, aunque el procedimiento (b) «ganaría» más veces que el (a), es mucho más ineficiente que éste, pues el promedio de las distancias (al cuadrado) entre todas las posibles estimaciones y $P = 50$ (promedio que no es otra cosa que la varianza del estimador) es muchísimo mayor en aquel caso. Dicha varianza, por otra parte, se puede estimar con una sola muestra, aunque cuando se conocen los datos poblacionales -como ocurre en este caso excepcional- puede calcularse exactamente. Esa comparación es, en esencia, lo que se hace en un estudio realizado por García *et al.* (1987) con el mismo propósito que se planteaba el que ahora nos ocupa.

Por otra parte, el trabajo de López *et al. (1990)* incurre en otra inadvertencia conceptual. Para cada uno de los cuatro métodos se contabiliza el número de veces en que los intervalos de confianza construidos para las diversas estimaciones contienen al parámetro verdadero. Se considera mejor aquel procedimiento que registre más aciertos de este tipo. Esto es absurdo, ya que por la propia definición del intervalo de confianza, éste contendrá al parámetro aproximadamente para el $(1 - \alpha)$ 100% de las muestras que se usen.

Si, por ejemplo, el porcentaje de sujetos con anticuerpos se estimara 800 veces con respectivas muestras de tamaño $n = 10$ y otras 800 veces con muestras de tamaño $n = 500$, se podrían construir 800 intervalos (por ejemplo, con un 95% de confianza) en cada caso. Tanto para el ineficiente procedimiento de tomar $n = 10$ como para el mucho más preciso de tomar $n = 500$, se tendrá que aproximadamente 760 intervalos (95% de 800) habrán de contener a *P.* Lo que ocurre es que los intervalos (los que aciertan y los que no) en el segundo caso serán mucho más estrechos (y, por ende, más informativos) que en el primero.

Sintetizando la discusión, hay que reparar en que, si bien el azar no inyecta representatividad, sí ***inyecta confianza.*** El método aleatorio puede, alguna vez, jugarnos una mala pasada (dar lugar a una muestra no representativa), pero al usarlo como regla, estamos asegurando la objetividad e imparcialidad del proceso valorativo y, a la larga, por su conducto habrá de emerger necesariamente la verdadera cara de la realidad que se investiga. Cuando el inspector de la Sección 10.4.1 selecciona su muestra al azar puede obtener un resultado extremo como el del ejemplo,

pero si todos los inspectores se ciñen a esa regla, tales resultados atípicos, serán excepcionales y prevalecerán en cambio las valoraciones que registran las regularidades del proceso docente que se quiere evaluar.

10.5. Marco de extrapolación

Entre los estudiantes de muestreo es bien conocido el adagio que establece que *las **inferencias que** se **produzcan a partir de una muestra han de circunscribirse a la población muestreada.*** Técnicamente, eso es impecablemente cierto. Pero un apego estricto a dicha regla tiene efectos paralizantes difíciles de justificar. En la práctica suele ocurrir que el método de selección es tal que, por una razón u otra, no todos los individuos de la población tienen oportunidad de integrar la muestra y, sin embargo, la inferencia realizada abarca a toda la población y no sólo a la porción de la que procede la muestra.

El grado en que una transgresión como esa sea «perdonable» no es un asunto de naturaleza estadística sino algo inherente al problema que se aborda; depende de la valoración de los investigadores, basada en su sentido común y su «cultura» sobre el problema, elementos a partir de los que se dirá la última palabra.

Supongamos que se quiere estimar la prevalencia de insuficiencia renal crónica (IRC) en una ciudad pero que, por razones prácticas, la muestra se elegirá de un listado incompleto, que contiene solamente al 90% de la población actual (por ejemplo, sólo a los que poseen teléfono, ya que, supongamos, el marco muestra1 será la guía telefónica).

La pregunta clave, que evidentemente no es de índole estadística, sería: ¿hay motivos para sospechar que la posesión o no de teléfono se relaciona de algún modo con el hecho de padecer IRC? Si la respuesta es positiva, no habrá «perdón» [15]. Pero si, por mucho que se especule teóricamente, no aparece ningún vínculo, ni directo ni indirecto, entre ambos rasgos, yo me inclino por la «absolución metodológica». Me baso en que, de todos modos, el conocimiento que se obtenga por vía muestra1 habrá de ser provisional, perfectible y sujeto a refinamiento. ¿Por qué entonces no ejercer la flexibilidad también en este sentido? En última instancia, se trata de ser flexible en el marco de un talante riguroso, que es mucho mejor que ser rígido e implacable sobre un sustrato conceptualmente borroso, como ocurre con tanta frecuencia y en tantos contextos.

En un libro clásico, aunque en buena medida olvidado, Hagood (1941) encara el tema desde una perspectiva más general, que desborda el marco del muestreo en poblaciones finitas con fines descriptivos, para abarcar también el de las pruebas de

[15] Tal es el caso del consumo de drogas entre médicos rurales que se comenta en la Sección 8.1: producto del carácter autoseleccionado de la muestra, allí la probabilidad de pertenecer a ella era diferente entre los consumidores y los no consumidores de drogas.

hipótesis. Lo que se discute es cuándo y sobre qué bases se pueden hacer inferencias a un *universo hipotético,* un superuniverso del cuál nuestro universo finito (es decir, la población observada) puede ser considerada una muestra aleatoria. Allí se barajan varias alternativas que justifican teóricamente la flexibilidad que en este terreno yo defiendo.

10.6. Técnicas cualitativas, técnicas cuantitativas

Como se comentaba al final de la Sección 1.5, la estadística no siempre tiene asegurado el protagonismo en los procesos de análisis. Por ejemplo, muchos problemas de índole social se abordaron durante muchos años sin hacer uso de las técnicas formales de encuesta y, con frecuencia, tomando información de manera tal que no existía la posibilidad de manejar los datos en un entorno cuantitativo. Los procedimientos de análisis eran de orden básicamente cualitativo, especialmente entre antropólogos y etnógrafos, pero también entre sociólogos, psicólogos e, incluso, clínicos.

Investigadores de la talla de Max Weber sostenían que la sociología hallaba sus resultados por medio de la *comprensión,* a diferencia de las ciencias naturales que procedían por vía de la *explicación* (Boudon, 1978).

Posteriormente fueron expandiéndose diversos desarrollos teóricos -en especial las técnicas de muestreo- que, al propiciar la cuantificación de los resultados, y permitir expresarlos, por tanto, de manera más formal, fueron desplazando de la práctica, y devaluando metodológicamente, a las técnicas cualitativas de análisis. En rigor, las prácticas cualitativas nunca desaparecieron y en cierta medida tuvieron incluso expresiones importantes de renovación y vitalidad; lo que sí se consolidó como realidad fue la separación casi absoluta [16] entre unas y otras.

El epidemiólogo típico de la nueva hornada positivista, cuya gestión, quiéralo o no, tiene profunda connotación sociológica, miraría con desdén al antropólogo que se pierde en lo que, a su juicio, no pasaba de ser diletantismo subjetivo y visión novelada de la realidad. El científico social clásico, por su parte, recelaba de las tecnologías encartonadas que le proponían los estadísticos y que lo obligarían a reducir la riqueza de sus observaciones a tablas y porcentajes. Gurvitch (1950), por ejemplo, escribía que «cuando las estadísticas no se aplican en un marco cuidadosamente acotado y verificado, no constituyen más que manipulaciones puramente matemáticas de grandes cifras» y, refiriéndose a las técnicas formales de encuestas, encarnadas en la archifamosa empresa que las popularizara en Estados Unidos, agregaba que «los procedimientos de Gallup son irrisorias búsquedas de promedios arbitrarios que no existen y operan en el vacío».

[16] Tal separación tuvo excepciones desde muy temprano. El estudio, por citar un ejemplo, de Warner (1947), en que se combinan las encuestas formales con la «observación etnográfica», da cuenta de ello.

Almeida (1992) atribuye el divorcio entre técnicas cuantitativas y cualitativas al profundo compromiso de la epidemiología con las primeras, lo cual impidió una integración más estrecha entre las estrategias de investigación de las ciencias sociales en general. Y agrega:

> ***Decir que uno se pierde en lo específico, o que el otro siempre ofrece una aproximación superficial de cuestiones complejas, perdidas en los grandes números, es una actitud por lo menos ingenua que algunas veces aparece entre investigadores de ambas disciplinas... La naturaleza desigual y multifacética del objeto epidemiológico y su determinante justificará el empleo de un sensato «pluralismo metodológico».***

Castellanos (1989) ya había subrayado una opinión muy próxima a ésta cuando escribió:

> ***Nuestra apreciación es que frente a cada problema correctamente definido y delimitado en función de su complejidad y del marco conceptual con el cual se abordará su estudio, el investigador debe optar por aquella combinación de técnicas, cuantitativas y no cuantitativas, que le permitan el mejor éxito en su empeño, dentro del marco de las posibilidades.***

En ese trabajo, Castellanos recupera la taxonomía de Mitrov y Killman (1978) que clasifica los problemas en los ***bien estructurados*** (aquellos para los que se conocen todas las variables que participan del problema, los cuales a su vez subdivide en tres tipos) y los ***cuasiestructurados*** (aquellos en que al menos algunas de las variables relevantes son desconocidas). Personalmente pienso que, a los efectos de las investigaciones sanitarias y, en especial, de las socio-epidemiológicas, la clasificación, además de resultar algo confusa, es más bien estéril, ya que es imposible conocer con certeza cuáles son ***todas*** las variables relevantes; de modo que ***la totalidad*** de estas investigaciones estarían asociadas a problemas «cuasiestructurados». Castellanos (1989) comenta al respecto:

> ***Existe, sin embargo, la posibilidad de reducir el universo de eventos o variables a un universo construido, práctico, operativo, dentro del cual sean razonablemente conocidas las probabilidades de ocurrencia, siempre que las no conocidas puedan ser asumidas como muy bajas o de escasa significación práctica por sus consecuencias. Pero este proceso de reducción y de sopesar el valor relativo de las variables tiene un elevado contenido subjetivo que no puede ser resuelto por las matemáticas, aun cuando pueden ser un auxiliar*** muy ***importante.***

Entre las más connotadas técnicas cualitativas, todas bosquejadas y referenciadas en el citado trabajo de Castellanos, se hallan la asamblea o forum comunitario, el famoso ***brainstorming y*** el menos conocido ***brainwriting,*** las entrevistas en profundidad, las técnicas grupales (nominal, de discusión y Delfos), la historia de vida, el uso de informantes claves, la observación estructurada y la observación participante.

Todas ellas suponen la existencia de lo subjetivo (tanto en la realidad como en quien la estudia) y lo explotan sin ruborizarse. De hecho no hay motivos para el rubor si se repara en que para cierto tipo de problemas resultan tanto o más recomendables que determinadas aplicaciones estadísticas cuya objetividad dista mucho de ser absoluta.

Las encuestas estructuradas, por ejemplo, por lo común restringen marcadamente el espacio de expresión de los interrogados. Ello viabiliza y simplifica notablemente el procesamiento estadístico, pero la subjetividad del investigador contamina el proceso tan pronto éste fija, tanto sintáctica como conceptualmente, las respuestas posibles. La diferencia básica entre los procedimientos cuantitativos y los cualitativos no estriba en que aquellos sean objetivos y éstos no, sino en el punto y el modo en que se introduce la subjetividad: los últimos ponen el énfasis en permitir que los actores sociales participen con su propia subjetividad en el proceso; en los primeros la subjetividad -poca o mucha- es monopolizada por los investigadores.

Las técnicas cualitativas tienen, sin embargo, un alcance muy limitado. Si bien pueden ser un magnífico instrumento alternativo o complementario de las encuestas para resolver tareas tales como sopesar un estado de opinión u ordenar jerárquicamente un paquete de necesidades según prioridades, resultan inoperantes en la mayoría de los problemas de investigación epidemiológica, en casi todos los de la clínica y en la totalidad de los de las ciencias básicas. La bioestadística, en cambio, puede ser útil en cualquiera de estos ambientes y en la mayoría de sus problemas. Pero lo que no debe perderse de vista es que el rasgo señalado no es privativo de las encuestas: todo problema encarado con ayuda de la estadística tendrá siempre un componente subjetivo, si no en varios puntos, al menos en las fases de análisis e interpretación.

Bibliografía

Almeida N (1992). ***Epidemiología sin números.*** Serie Paltex n.º 28, OPS/OMS, Washington.

Boudon R (1978). ***Los métodos en sociología.*** El Ateneo, Buenos Aires.

Castellanos PL (1989). ***Algunas técnicas para el estudio de lo subjetivo, los problemas*** *cuasiestructurados* ***y el estudio de la situación de salud.*** OPS/OMS. Presentado en la ***Reunión sobre abordajes y métodos para estudiar diferenciales de salud según condiciones de vida,*** Brasilia 7-11 de agosto.

Cobos A (1995). ***Diferencia entre confirmar y explorar*** JANO 49:1017-1018.

Chatfield C, Colin AJ (1980). ***Introduction to multivariate analysis.*** Chapman and Hall, London.

Chomsky N, Fodor J (1980). ***The inductivist fallacy.*** En Piattelli - Palmarini, M. (Ed.), ***Language and learning,*** Harvard University Press, Cambridge.

Dixon WJ (1990). *BMDP* ***Statistical software manual.*** University of California Press, Berkeley.

Evans SJW (1988). Uses ***and abuses of multivariate methods in epidemiology.*** Journal of Epidemiology and Community Health 42: 311-315.

Fisher R (1934). ***The effect of methods of ascertainment upon the estimation of frequencies.*** Annals of Eugenics 6: 13-2.5.

García LM, Pérez MM, Bassolo A, Abraira V, Gervás JJ (1986). ***Estudios de morbilidad: ¿qué muestra elegir?*** Atención Primaria 4:136-139.

Guttman L (1977). ***What is not what in statistics.*** The Statistician 26: 81-107.

Gurvitch G (1950). ***La voaction actualle de la sociologie.*** Presses Universitaires de France, Paris.

Hagood MJ (1941). ***Statistics for sociologists.*** Reynal and Hitchcock, New York, NY.

King G (1986). ***How not to lie with statistics: Avoiding common mistakes in quantitative political science.*** American Journal of Political Science 30: 666-687.

Leamer EE (1985). ***Sensitivity analysis wolud help.*** American Economic Review 75: 308-313.

López A, Esnaola S, Guinea J, Gómez MC (1992). ***Limitaciones del muestreo en estudios de atención primaria: comparación de cuatro diseños muestrales.*** Gaceta Sanitaria 6: 19-24.

Manly BFJ (1994). ***Multivariate Statistics: A Primer:*** Chapman and Hall, 2.ª ed, London.

Marshall W (1990). ***Data dredging and noteworthiness.*** Epidemiology 1: 5-7.

McGee DL, Reed D, Yano K (1984). ***The results of logistic analyses when the variables are highly correlated.*** American Journal of Epidemiology 37: 713-719.

Mitrov 1, Killman R (1978). ***Methodological approach to social sciences.*** Jossey-Buss, San Francisco, CA.

Mulaik SA (1987). ***A brief history of thephilosophicalfoundations of exploratory factor analysis.*** Multivariate Behavioral Research 22: 267-305.

Samaja J (1994). ***Vigilancia epidemiológica de los ambientes en que se desarrollan los procesos de la reproducción social.*** Ponencia presentada al 8.º Congreso Mundial de Medicina Social, Guadalajara, Jalisco.

Seber GAF (1984). ***Multivariate observations.*** Wiley and Sons, New York.

Silva LC (1993). ***Muestreo para la investigación en ciencias de la salud.*** Díaz de Santos, Madrid.

Spearman C (1904). ***General intelligence objectively determined and measured. American*** Journal of Psychology 15: 201-293.

Tukey JW (1977). ***Exploratory data analysis.*** Addison-Wesley, Massachussets.

Vanderbroucke JP (1990). ***How trustworthy is epidemiologic research?*** Epidemiology 1: 83-84.

Yera M (1967). ***La técnica del análisis factorial. Un método de investigación en psicología.*** Editorial R, La Habana.

Warner L (1947). ***The status system of a modem community.*** Yale University Press, New Haven.

11

El enigma del tamaño muestral

Creo en el suave poder de la razón.

GALILEO GALILEI

Uno de los más asombrosos ejemplos de autoengaño en materia estadística se da en relación con el famoso asunto de la determinación del tamaño de la muestra, «la determinación de la n», dicen algunos, como si dicho tamaño tuviera necesariamente que denotarse mediante la letra n. En este capítulo se discute con detalle este importante problema.

11.1. Un caso de hipnosis colectiva

En relación con el tema, parecería que se hubiese consumado una especie de autismo generalizado: los usuarios no consiguen salir de su estado hipnótico; los autores de texto y muchos profesores, o bien son víctimas de la parálisis, o bien se suman a una función hipnotizadora que yo prefiero no cohonestar.

Lo que sí está fuera de toda duda es que algo muy singular ocurre con este asunto. A pesar de haber sido profusamente tratado en libros y artículos, de que se explica en clases y conferencias y de que existen tablas *ad hoc* para hacer las determinaciones, este tema parecería renuente a dejarse dominar por los interesados: llegado el momento de calcular el tamaño muestral, la mayoría de los investigadores se sienten incapacitados para hacerlo por sí solos o, en el mejor de los casos, inseguros de la corrección de lo que han hecho.

¿Cuál será el enigma subyacente? ¿Cómo explicar tan recurrente y curiosa circunstancia? Por su enorme relevancia y su carácter polémico, considero pertinente desarrollar este punto sin apremios y profundizar en ciertas cuestiones técnicas cuando sea menester.

11.2. Problema de estimación y pruebas de hipótesis

Para comenzar, deben reconocerse dos situaciones bien diferenciadas:

a) Aquella en que se necesita determinar el tamaño muestra1 necesario para *realizar estimaciones.*

b) Cuando se está planificando un estudio analítico, sea de tipo observacional o experimental, y lo que se quiere es determinar los tamaños muestrales para los grupos involucrados en unaprueba *de hipótesis.*

La primera de ellas corresponde, en esencia, a los llamados *estudios descriptivos* (diagnósticos de salud, caracterizaciones epidemiológicas, estudios de prevalencia, etc.). En el segundo, el análisis transita en algún punto por el acto metodológico de la *comparación* [1].

Cuando estamos en el ambiente descriptivo, lo que se quiere es realizar estimaciones de parámetros (fundamentalmente proporciones o porcentajes, razones, medias, varianzas y totales; pero en ocasiones, también con afán descriptivo, coeficientes de correlación o de regresión).

En el entorno analítico, el examen estadístico de los datos suele conducir a maniobras tales como la comparación estadística de porcentajes o medias, o a la evaluación de la significación de coeficientes de concordancia o de correlación (nótese que, de hecho, en estos últimos dos casos se trata también de comparar; lo que se compara son las estimaciones correspondientes con el número cero).

Ambos procesos se desarrollan por medio de muestras y es natural que, trátese de lo uno o de lo otro, se desee operar con la menor cantidad posible de datos con el fin de economizar recursos. El problema es hallar ese número mínimo de unidades con el cual puedan resolverse eficientemente tales tareas.

11.3. La teoría oficial sobre tamaños de muestra

Repasemos primero de manera sucinta la *teoría oficial* sobre este tema. Curiosamente, a pesar de ser uno de los temas más borrosamente solucionados por la estadística, la uniformidad del tratamiento que le dan los textos es casi total; ello haría pensar que la interfase entre la teoría y la práctica correspondiente está completamente consolidada y carece de toda fisura, que una y otra guardan una armoniosa relación mutua.

---

[1] Dicho sea de paso, la comparación ***siempre es un*** recurso metodológico. «Comparar» nunca es una finalidad en sí misma, como muchos parecen creer cuando formulan sus objetivos haciendo uso de ese verbo. Para más detalles, véase Sección 1.2.2.

Desde hace relativamente poco tiempo contamos, incluso, con una especie de biblia, portavoz de ese oficialismo: un libro exclusivamente destinado al tema, debida a Lemeshow ***et al.*** (1990), ahora publicado por la prestigiosa editorial Wiley, pero heredero del manual de Lwanga y Lemeshow (1989), editado varias veces por la Organización Mundial de la Salud. La existencia de este libro, titulado ***Corrección del tamaño muestra1 en estudios de salud*** constituye, por su alta especificidad temática, una magnífica referencia para el examen que sigue.

11.3.1. Estudios descriptivos

En el caso descriptivo, lo más conocido es el tratamiento que debe darse al cómputo del tamaño muestra1 cuando se lleva adelante un ***muestreo simple aleatorio (MSA).*** Además de hallarse en el libro arriba citado, la solución a este problema puede consultarse en decenas de obras sobre muestreo, desde los textos clásicos como el de Cochran (1963) o el de Yamane (1970), hasta en mi propio libro sobre el tema (Silva, 1993).

Puesto que la naturaleza del debate que se desarrollará es, en esencia, independiente del parámetro específico que se quiera estimar, centrémonos en el caso más sencillo, la estimación de un porcentaje.

Sintetizando, se trata de la siguiente situación: se quiere estimar el porcentaje ***P*** de aquellos sujetos de una población de tamaño *N* que cumplen con cierta condición o poseen determinado rasgo. Para ello hace falta seleccionar primero y encuestar luego una muestra simple aleatoria cuyo tamaño, naturalmente, es necesario fijar de antemano. Llamemos *n* a ese número desconocido.

Lo que se quiere, en fin, es determinar el ***valor mínimo necesario*** que ha de tener *n* para conseguir una estimación ***adecuada*** de ***P.*** Si llamamos p ***al porcentaje muestral*** (estimación de ***P),*** lograr que esta estimación sea ***adecuada*** equivale a lograr que se cumpla, con una probabilidad alta (por ejemplo $1 - \alpha = 0{,}95$), que $|p - P|$ sea menor que cierta magnitud prestablecida.

Es claramente intuitivo (y además puede demostrarse formalmente) que, tomando *n* suficientemente grande, por pequeño que sea un número dado *E* y por reducido que sea el valor de α, se conseguirá que la probabilidad de que $|p - P| < E$ sea superior a $1 - \alpha$. Por lo tanto, una vez establecido el valor máximo que puede aceptarse para ***E*** (el error máximo admisible) y el grado de confianza $1 - \alpha$ que se desee, se puede hallar el valor mínimo de ***n*** que cumpla con la mencionada restricción.

Supongamos que, tras realizar un análisis detenido del problema, se llega a la convicción de que basta conocer ***P*** con error -a lo sumo- de ***E,.*** Por ejemplo, consideremos una población de $N = 2.500$ mujeres en edad fértil y supongamos que se quiere conocer el porcentaje de las que desean tener al menos un hijo en los próximos 5 años con error no mayor de 4% ***(E, = 4).*** Esto quiere decir que se está pensando en términos como los siguientes:

SI EL VERDADERO PORCENTAJE FUERA 40%, PERO LA MUESTRA ARROJASE UNA ESTIMACIÓN DE 37%, SE CONSIDERARÍA ADECUADAMENTE CONOCIDO EL DATO (EL ERROR COMETIDO SERÍA DEL 3%, INFERIOR A LA COTA MÁXIMA QUE SE DIJO ADMITIR); SIN EMBARGO, SI LA ESTIMACIÓN DE ESE PORCENTAJE ASCENDIERA A 45%, ENTONCES SE CONSIDERARÍA QUE LA ESTIMACIÓN ES INADMISIBLEMENTE DISTANTE DEL PARÁMETRO (LA DIFERENCIA CON ÉL SERÍA DE UN 5%, QUE SUPERA LO ACEPTABLE).

La fórmula que se desprende [2] de la mencionada demanda acerca del máximo error admisible, es la siguiente:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \qquad [11.1]$$

donde

$$n_0 = \frac{Z^2_{1-\alpha/2}\ P(100 - P)}{E_0^2} \qquad [11.2]$$

en la cual $Z_{1-\alpha/2}$ representa al percentil$(1 - \alpha/2)100$ de la distribución normal estandarizada (por ejemplo, si $\alpha = 0{,}05$, entonces $Z_{1-\alpha/2} = Z_{0.975} = 1{,}96$).

Supongamos que se quiere aplicar la fórmula [11.1] a la situación descrita. Entonces surge la primera dificultad de peso: se fijó E_0 ***y*** $Z_{1-\alpha/2}$, ***y se*** conoce *N*, pero para hacer uso de [11.2] es necesario conocer el valor de *P*.

Menudo círculo vicioso: ***todo este andamiaje tiene como finalidad conocer el valor de Ppero, para utilizarlo, es necesario conocer P.*** Nótese que no se trata de un detalle lateral: un enfoque riguroso aconsejaría detener el proceso en este punto por elementales razones de coherencia.

¿Cómo sobrevive el discurso oficial del muestreo a pesar de esta flagrante incongruencia? Se plantea que usualmente «se puede tener una idea aproximada» del valor de ***P*** a partir de un estudio piloto, o de una investigación similar realizada anteriormente[3]. Por el momento, admitamos que tenemos alguna idea del valor de *P*.

Imaginemos que consideraremos ***P*** = 75%, «idea aproximada» que se ha obtenido, por ejemplo, de un estudio anterior. Así las cosas, la aplicación de [11.1] y [11.2] (usando ***P*** = 75, $N = 2.500$, $\alpha = 0{,}05$ y E_0 = 4) se concluye que ***n*** = 381.

[2] La deducción puede hallarse en cualquiera de las citadas referencias especializadas.

[3] Otra variante es que se ponga directamente $P = 0{,}5$; de ella nos ocuparemos específicamente en la Sección 11.5.

Puesto que un estudio de esta índole no se circunscribe usualmente a estimar un único parámetro, y como el tamaño de muestra ha de ser el mismo para todo el estudio, aquí surge otra dificultad. En efecto, el desarrollo precedente produciría tamaños de muestras distintos para los diferentes parámetros considerados. La «solución» más frecuente que se da (por aquellos pocos que reconocen y abordan este problema) es la de seleccionar «el más importante» y hacer el análisis para éste.

Ahora bien, ésta es la solución que se da al caso en que se usa el muestreo simple aleatorio. Si el diseño muestra1 es otro, esta teoría no es válida. En particular, si se trata de un muestreo en etapas (que es el que se utiliza por lo menos 95 de cada 100 veces en la práctica), el error que regularmente se comete al estimar ***P es*** mayor que el obtenido para MSA, supuesto que se está usando el mismo tamaño de muestra. O, dicho de otro modo: con muestreo por conglomerados, para alcanzar el grado de precisión prefijado, es necesario seleccionar una muestra de tamaño mayor que el que demanda el MSA.

La sugerencia que se da entonces es, simplemente, multiplicar el tamaño surgido de [11.1] por un número, el ***efecto de diseño,*** usualmente denotado como ***deff,*** apócope de la expresión inglesa ***design effect.*** Es decir, el tamaño corregido n_c sería:

$$n_c = (deff)\ (n) \qquad [11.3]$$

Lemeshow y sus tres coautores, tras extenderse sobre el caso en que se usa MSA, dicen textualmente lo siguiente:

> ***Este nunca*** *sería* ***el diseño empleado en una encuesta de terreno verdadera. Como resultado de ello, el tamaño de muestra ha de elevarse en una magnitud igual al efecto de diseño. Por ejemplo, si se fuera a usar muestreo por conglomerados, el efecto de diseño pudiera estimarse en 2.***

Hasta aquí la síntesis de la «solución oficial» para estudios descriptivos.

11.3.2. Estudios analíticos

Ahora consideremos un ejemplo relacionado con estudios analíticos: de lo que se trata ahora es de evaluar la hipótesis que afirma que dos porcentajes P_1 ***y*** P_2 son iguales, con la hipótesis alternativa de que son diferentes. Imaginemos que se trata de un ensayo clínico en que P_1 ***es*** la tasa de recuperación de pacientes que reciben un tratamiento convencional y P_2 la de los que reciben uno experimental.

Se quiere hallar el tamaño de muestra mínimo n que debe tomarse en cada grupo (el mismo para ambos) de modo que la prueba sea capaz de detectar como sig-

nificativa (no atribuible al azar) una diferencia mínima prefijada entre P_1 y P_2. La fórmula correspondiente [4] es:

$$n = \frac{[Z_{1-\frac{\alpha}{2}}\sqrt{2P^*(1-P^*)} + Z_{1-\beta}\sqrt{P_1(1-P_1)} + P_2(1-P_2)]^2}{(P_1 - P_2)^2} \qquad [11.4]$$

donde α y β representan las probabilidades máximas admisibles de cometer, respectivamente, los errores de tipo 1 (la probabilidad de rechazar indebidamente la hipótesis nula) y de tipo II (el complemento de la llamada ***potencia de la prueba) y*** $P^* = \frac{P_1 + P_2}{2}$.

En esta situación hay que prefijar α (por ejemplo, puede elegirse el sacralizado 0,05) y β (se toma con frecuencia $\beta = 0{,}2$). Supongamos que ***P, es*** conocido en la práctica clínica y asciende a 60% ***(P,*** = 0,6) y que la diferencia se considerará clínicamente relevante si la tasa de recuperación se eleva por lo menos hasta 70% (P_2 = 0,7). En tal caso la aplicación de [ll.4] arrojaría $n = 400$.

11.4. ¿Qué oculta la teoría oficial?

El discurso oficial tiende a ejercer un tipo de censura que silencia algunos problemas reales y los suple con formulaciones académicas de escasa o nula aplicabilidad. Varios puntos son usualmente omitidos (algunos, a cal y canto) cuando se aborda el tema. A continuación se analizan cinco de ellos tomando como paradigma, nuevamente, el caso de la estimación de ***P*** en el contexto descriptivo [5].

A) SUBJETIVIDAD EN LA DETERMINACIÓN DE LOS DATOS QUE EXIGEN LAS FÓRMULAS

Se supone que hay un «error máximo» que se puede aceptar; sin embargo, no siempre resulta fácil la identificación ***a priori*** de ese error. Esta tarea exige del investigador que piense en unos términos para los que con frecuencia no está preparado.

Pero esto no es atribuible al método sino a sus usuarios. Sin embargo, un problema del que no es posible escapar consiste en que, ***de todos modos, se trata de una decisión esencialmente subjetiva.***

---

[4] En rigor, ésta es una de las múltiples situaciones posibles (quizás una de las más frecuentes); incluso para esta misma situación existen otros enfoques, pero todos son susceptibles de los mismos juicios que cabe hacer para la que hemos elegido con fines expositivos.

[5] Los apartados que siguen se han tomado, con escasas modificaciones, de mi libro sobre muestreo (Silva, 1993).

Volvamos al ejemplo en que se quiere estimar el porcentaje de mujeres que quieran otro hijo: ¿qué error considerar suficientemente pequeño como para que resulte admisible? No parece existir árbitro alguno que determine inequívocamente si un error de un 4% es admisible, o si, para que lo sea, éste no pueda exceder por ejemplo el 2%.

Algo muy similar ocurre con la confiabilidad $1 - \alpha$. ¿Tomar 95%, o 99%, como sugieren otros con no menos argumentos?

Como se recordará la «pre-estimación» de ***P es*** un proceso igualmente cargado de subjetividad. Puesto que vamos a trabajar con «una idea aproximada» del valor de *P,* estamos condenados a conocer sólo «una idea aproximada» del valor del *n* necesario. Si tal aproximación fuera muy errónea, así será el tamaño muestral, lo cual es obviamente muy inconveniente. Pero si estuviéramos persuadidos de que la preestimación fuera muy cercana al verdadero valor de ***P,*** entonces... la situación sería aún más inconveniente, porque en tal caso no haría falta hacer la estimación, ya que la finalidad de todo el proceso es precisamente obtener una buena aproximación de ***P;*** consiguientemente, mucho menos interés tendría el cómputo de *n.*

En el ejemplo, bien podría haber ocurrido que un estudio previo hubiese arrojado que un 75% de las mujeres estaban en ese caso, pero también es posible que una pequeña encuesta piloto hubiese producido una preestimación de, por ejemplo, 86%. ¿Cómo escoger entre ambas alternativas?

A toda esta subjetividad hay que añadir la que se desprende del hecho de que los diseños son, por lo general, totalmente diferentes al muestreo simple aleatorio. Los errores muestrales -en el caso de muestras complejas- no sólo dependen del tamaño total de la muestra sino también de las asignaciones muestrales a los estratos y de los tamaños de muestra en las etapas intermedias (es decir, del número de conglomerados en cada etapa y de sus tamaños) cuando se usa muestreo polietápico. Este hecho, cuando no es pasado por alto, se resuelve con la variante de usar el coeficiente expansor del tamaño muestra1 (el *deff* mencionado en la Sección 11.3.1) según la cual debe aumentarse el tamaño producido por la fórmula [ll.l]

Todo un acto de prestidigitación numérica: el libro destinado a dar soluciones técnicamente fundamentadas se limita a comunicar que el ***deff*** «pudiera estimarse en 2». Naturalmente, lo esperable es que el investigador que acuda al libro para buscar su receta, proceda siempre de ese modo: multiplicar por 2 el tamaño muestra1 salido del MSA.

Uno se pregunta, ¿para qué tantas fórmulas previas y tantas tablas, si a la postre se dice que se multiplique el número obtenido por un 2 cuya pertinencia es totalmente especulativa? Se podría argüir que el efecto de diseño «suele ser de esa magnitud». Pero es falso: basta echar una ojeada a algunos estadios que calculen los ***deff*** para corroborar que este número cambia radicalmente de una encuesta a otra, de un diseño muestra1 a otro y que, dentro de la misma encuesta y del mismo diseño, suele modificarse notablemente en dependencia del parámetro elegido.

El recurso de ayuda que aparece en el módulo STATCALC del EPIINFO, en concordancia con el talante más realista y flexible que caracteriza a este sistema,

sugiere utilizar un *deff* entre 1,5 y 2. Lo cierto es que con el *deff* se hace la contribución final a la de por sí abultada colección de entradas subjetivas en el sistema.

A modo de ilustración, computemos los tamaños de muestra que se obtendrían con los dos juegos de decisiones que hemos mencionado, ambos igualmente razonables (y, ciertamente, no demasiado dispares). Para la población de $N = 2.500$ mujeres en edad fértil, usando [ll.l], [ll.2] y [11.3], y se tendría:

	Alternativa 1	Alternativa 2
P	86%	75%
Máximo error E_0	4%	3%
Confiabilidad 1 - α	95%	99%
deff	1,5	2,0
Tamaño muestra1	388	1.780

Quiere decir que dos investigadores independientes, siguiendo la misma estrategia general (usando las mismas fórmulas para la estimación del mismo parámetro), llegarían a tamaños ***abismalmente*** diferentes, sólo por el hecho de que aprecian (legítimamente ambos) de manera distinta los valores que deben darse a los elementos que las fórmulas demandan.

0 sea, que ese inevitable margen de subjetividad, aun cuando dé lugar a pequeñas diferencias en las decisiones iniciales, puede producir diferencias muy notables en el tamaño muestra1 [6]. ¿Podrá mantenerse que, al usar las fórmulas, se está haciendo uso de un método objetivo? Cualquier investigador honesto reconocerá que, usualmente, ha llevado el proceso de determinación de tamaños muestrales al revés: ha ido escogiendo los valores de ***P,*** α, E_o ***y deff*** de manera tal que la fórmula [ll.3] arroje el valor de *n* ***que ha decidido de antemano.*** Es precisamente el marco de subjetividad que inescapablemente padece el proceso el que permite (e invita a) tal manipulación.

[6] Pudiera pensarse que la diferencia entre una confianza del 95% y una del 99% no es «pequeña»; pero debe recordarse que si se decidiera trabajar con confiabilidad mayor del 95%, según el ritual consagrado, se escogerá 99%, nunca 96,7% o 98,3%. De modo que si alguien siente que 95% no es un número cercano a 99%, habrá de admitir que 99% es «cualitativamente» el número más cercano a 95% por el que se optaría.

B) SIEMPRE SE QUIERE ESTIMAR UN NUTRIDO GRUPO DE PARÁMETROS

Una encuesta en la práctica nunca se proyecta para estimar uno o dos parámetros sino que normalmente exige realizar decenas o cientos de estimaciones. Debe notarse que, usualmente, una tabla demanda una estimación por cada celda y, en ocasiones, suelen construirse muchas tablas. Es evidente, entonces, que no resulta demasiado práctico aplicar fórmulas como la examinada para cada uno de estos parámetros.

Pero aunque ello fuese factible y aunque se contara con información disponible para hacerlo, nos encontraríamos con un panorama desconcertante: los tamaños que demandarían las diferentes estimaciones recorrerían, probablemente, *un amplísimo espectro de valores* como candidatos a ser el tamaño muestral.

Esta dificultad, como ya se dijo, suele olvidarse bajo la premisa de que se puede elegir el parámetro «más importante» (en el mejor de los casos, «los dos o tres más importantes»). Obviamente, incluso suponiendo que tenga sentido hacer tal discriminación [7], el margen de subjetividad para decidir cuáles son tales «parámetros más importantes» es enorme. Y aun así, esos «más importantes» pueden dar lugar a números muy dispares entre sí.

C) CARÁCTER REDUCTOR DE LOS TAMAÑOS MUESTRALES INDUCIDOS POR LA NECESIDAD DE REALIZAR ESTIMACIONES DENTRO DE SUBCLASES

Lo usual es que se seleccione una muestra de unidades y luego se hagan análisis circunscritos a subconjuntos de la población, basados -naturalmente- en los correspondientes subconjuntos muestrales, cuyos tamaños son menores (a veces mucho menores) que el de la muestra original.

Por ejemplo, en el estudio de las mujeres en edad fértil que se bosquejó antes, es muy probable que se quieran hacer estimaciones del porcentaje de aspirantes a tener un hijo en el próximo quinquenio, no sólo para toda la población sino también para el subgrupo de las solteras, o el de las que usan anticonceptivos, o dentro de según las diferentes categorías determinadas por el número de embarazos previos de la mujer.

De hecho, las sesudas formulaciones para determinar tamaños de muestra suelen desconocer totalmente que más tarde se harán estimaciones dentro de subconjuntos para los que las muestras se reducen, en algunos casos muy notablemente.

Veamos un ejemplo real que refleja cómo se produce esta reducción del tamaño muestra1 y también cómo puede (y suele) ser manejada. Provisionalmente, abandonaremos el ámbito de la salud.

[7] Es muy probable que no existan parámetros «más importantes»: los problemas suelen exigir enfoques integrados, que superen su desconcatenación metafísica en parcelas.

El periódico español *El País* (1993) reproduce el resultado de una encuesta realizada por la empresa DEMOSCOPIA en la que se sondeaba la opinión popular sobre el debate televisado entre dos candidatos presidenciales: José María Aznar y Felipe González, al día siguiente de producido.

La ficha técnica, incluida en el artículo de prensa, dice textualmente:

> Tamaño y distribución de la muestra: 800 ***entrevistas fijadas mediante muestreo estratificado por región y tamaño de hábitat proporcional a la distribución de la población y con ampliación de cuotas de sexo y edad*** [sic].

Confieso que el texto me resulta simplemente tríptico: no me resulta posible entender casi nada a partir de la palabra «entrevistas» [8].

Pero ahora lo que importa es que nos han informado que el tamaño muestra1 es $n = 800$. La primera pregunta que se formuló a los encuestados fue: ¿VIO ENTERO O EN PARTE EL DEBATE?

Nos comunican que sólo 431 entrevistados lo vieron durante un lapso suficientemente largo como para opinar, y casi todo el resto del análisis se remite a ese número de espectadores. Por ejemplo, a continuación se preguntó:

> CON INDEPENDENCIA DE SUS SIMPATÍAS POLÍTICAS, ¿QUIÉN LE HA RESULTADO MÁS CREÍBLE?

Y registran los siguientes resultados según las tres grandes agrupaciones políticas españolas [9]:

	Partido Socialista %	Partido Popular %	Izquierda Unida %
Felipe González	5.5	0	25
José M. Aznar	28	94	34
Los dos por igual	7	3	29
Ninguno de los dos	13	1	12
No sabe/no contesta	3	1	0

[8] Y sospecho fuertemente que a los lectores de «El País» les pasará lo mismo, aunque quizás muchos de ellos queden bastante anonadados con la «cientificidad» que, a juzgar por su apariencia, destila tal enunciado.

[9] La suma de porcentajes dentro del Partido Socialista asciende absurdamente a 106%. No es una errata de este libro: así aparece en el informe de Demoscopia.

Los autores no comunican cuántos entrevistados correspondieron a cada una de estas tres agrupaciones pero, teniendo en cuenta datos conocidos, cabe esperar que, por ejemplo, los entrevistados simpatizantes de Izquierda Unida hayan sido alrededor de 40, ya que esa fuerza agrupaba en aquel momento aproximadamente al 10% del electorado. De modo que, para estimar el porcentaje de individuos de Izquierda Unida a quienes resultó más creíble el Sr. González, se trabajó con una muestra de 40, ***un número*** 20 veces ***menor que el tamaño de muestra original.*** Esta situación es enteramente típica: aunque la ficha técnica informe que la muestra fue de 800 entrevistados, el tamaño efectivo varía en dependencia de los vericuetos computacionales que exige el estudio. Ahora detengámonos a examinar qué efecto tiene este hecho.

Si se computa el error de muestreo en que se ha incurrido (suponiendo que se usó muestreo simple aleatorio), hay que aplicar la fórmula siguiente [10]:

$$e = Z_{1-\alpha/2}\sqrt{\frac{p\,(100-p)}{n-1}} \qquad [11.5]$$

¿Qué dice en su ficha técnica el artículo que nos ocupa sobre este tema? Textualmente, lo que sigue:

> Error de muestreo: ***asumiendo los criterios de muestreo aleatorio simple, para un nivel de confianza de*** 95,57% ***(dos sigmas)*** y ***para la hipótesis más desfavorable (p = q = 50), el errorpara el total de la muestra sería de ±3,5%.***

La frase «para la hipótesis más desfavorable *(p = q* = 50)» aparece sistemáticamente y obsesivamente en todas las ***fichas técnicas*** de este tipo. ¿Qué se quiere decir exactamente con ella'? Se trata de que el valor *p* (100 -*p*) que está bajo el radical de [11.5] alcanza su máximo en el caso en que ***p*** = 50. Consecuentemente, el mayor valor que puede alcanzare corresponde a dicho caso. Por ejemplo, tomando ***n*** = 800 y $Z_{1-\frac{\alpha}{2}} = 2$, el valor del error será a lo sumo igual a 3,5, que es el que figura en la ficha técnica. Pero, ¿es ésta realmente «la hipótesis más desfavorable»? Ciertamente no. Consideramos otra pregunta del cuestionario; según el informe, el 6% de los 800 encuestados contestó que probablemente no vería el segundo debate, programado para una semana más tarde. El error asociado a tal estimación sería entonces:

$$e = 2\sqrt{\frac{(6)\,((94)}{799}} = 1.7$$

[10] Véase cualquier texto de muestreo; por ejemplo, Silva (1993).

Este número es, en efecto, menor que 3,5. Sin embargo, la calidad de la estimación es inferior pues 1,7 representa el 28% de p = 6 mientras que 3,5 es solamente el 7% de p = 50.

En general el error relativo, definido como $e_r = \frac{e}{p}$, se incrementa en la medida que p disminuye, de manera que es absurdo afirma que el caso en que la calidad de la estimación es más desfavorable corresponde a preguntas para las cuales la estimación es *p = 50*.

Sin embargo, lo verdaderamente grave es que no se trabaja, para casi ninguna de las estimaciones de la encuesta, con ese tamaño de muestra sino con números muchísimo menores.

En el ejemplo que nos ocupaba, *p* = 25% (dato registrado en la tabla para Izquierda Unida) $n = 40$ y $Z_{1-\alpha/2} = 2$, la estimación del error máximo en que se ha incurrido es $e = 13{,}9\%$.

Por otra parte, esta sería la estimación si se hubiera tratado de un muestreo simple aleatorio. Si fue un diseño complejo (como parece colegirse del texto incluido en la ficha técnica) entonces el error verdadero ha de ser mayor. Concretamente, si llamamos e_c al error correspondiente al verdadero diseño, se tendrá (Kish, 1965): $e_c = e\sqrt{deff}$.

Supongamos que al diseño muestra1 que nos ocupa le corresponde un *deff* igual a 2, tal y como sugieren Lemeshow y sus tres coautores. Es un supuesto conservador, ya que por la naturaleza del problema (cierta tendencia a la homogeneidad de opiniones y valores dentro de una misma localidad), cabría esperar un *deff* mayor. Pero considerémoslo así para poder prosperar en el análisis.

En tal caso el error estimado, una vez hecho el ajuste, resulta igual a:

$$e_c = 13{,}9\ \sqrt{2} = 19{,}7$$

Quiere esto decir que el error cometido al estimar ese porcentaje de 25% podría razonablemente ser de 20%. Se trata de un error enorme: casi igual a la magnitud de lo que se estima. Es como si me preguntaran mi edad y yo dijera, en lugar de mis 44 años actuales, que tengo *aproximadamente* 9 *años*.

En síntesis, lo que se ha ilustrado es que al computar un tamaño muestra1 con las fórmulas, normalmente se actúa pensando en estimaciones para la población completa; pero el error «máximo» que se está dispuesto a cometer y que se usa con el fin de realizar aquella determinación resulta luego mucho menor que el que realmente se comete en «el fragor» de la tabulación real.

D) EL COSTO DE LOS PROCESOS ES USUALMENTE MÁS DETERMINANTE QUE CUALQUIER CONSIDERACIÓN TEÓRICA

El análisis aislado de la eficiencia carece de sentido: si no hubiese limitaciones de recursos, no se plantearía siquiera el uso de muestras sino que se investigaría íntegramente la población. Las disponibilidades de tiempo, personal y presupuesto ocupan un lugar determinante en las decisiones, aunque muchas veces la influencia de esta limitante se ejerza de manera implícita o solapada.

Se puede argüir que ha habido esfuerzos teóricos para hacer intervenir los aspectos económicos en las fórmulas para la determinación del tamaño de muestra. Por ejemplo, a lo largo del libro de Hansen, Hurwitz y Madow (1953) el lector hallará numerosos resultados en que se reflejan valores óptimos del número de conglomerados, de las asignaciones de tamaños a estratos, etc., para diversos diseños en los que intervienen tales aspectos. Se trata, sin embargo, de fórmulas -en general extremadamente abigarradas- cuya aplicación exige especificaciones de costos y de variabilidades que sólo pueden ofrecerse a través del procedimiento de... inventarlas.

Consideremos el siguiente ejemplo en que se trata de una muestra de escolares. Se quiere realizar un muestreo trietápico para estimar una media poblacional. Primero se eligen m escuelas; en cada una de ellas se tomarán como promedio $\bar{n}$ grupos; finalmente, dentro de cada uno de estos últimos se elegirá un número medio de q alumnos. La determinación de los números $m, \bar{n}$ y q que maximizan la precisión para un presupuesto económico dado, según Hansen, Hurwitz y Madow (1953) se realizará como sigue:

$$q = \frac{W_w}{\sqrt{W_b^2 - W_w^2 / Q}} \sqrt{\frac{C_2}{C_4}}$$

$$\bar{n} = \frac{1}{q} \frac{W_w}{B} \sqrt{\frac{C_1}{C_4}}$$

$$m = \frac{C}{C_1 + C_2 \bar{n} + C_3 \bar{n} q}$$

donde Q es el número medio de alumnos elegidos por escuela (computado usando el total de escuelas de la población), C es el presupuesto total de que se dispone; C_1 es el costo de ir a una escuela, C_2 el de ir a un grupo, y C_3 el de encuestar a un alumno. W_w, W_b y B son *complejísimas* expresiones que representan medidas de variabilidad relativa entre y dentro de las unidades de muestreo.

Los valores de C, $C_1, C_2, C_3, W_w, W_b, B$ y Q han de conocerse antes de realizar el

estudio[11]. Creo que no es menester extendernos en un juicio crítico de toda esta parafernalia, habida cuenta del laberinto en que ya nos colocaba el sencillísimo problema del tamaño muestra1 para estimar un modesto porcentaje en el contexto del MSA.

E) SE DEBEN ESTIMAR PARÁMETROS DE DIVERSA NATURALEZA

Usualmente se computa un tamaño muestra1 para estimar una media o un porcentaje dados; pero luego, en el estudio propiamente dicho, suelen hacerse estimaciones de todo tipo de parámetros, tales como coeficientes de correlación, pendientes de regresión, coeficientes de concordancia, etc.

En el mejor de los casos se construyen intervalos de confianza para estos parámetros, saludable práctica que permite aquilatar el grado de conocimiento alcanzado sobre el parámetro en cuestión. Pero casi nunca se utilizan procedimientos formales para determinar el tamaño muestra1 necesario para estimarlos. El manual de Lemeshow y sus colaboradores, por ejemplo, ni siquiera los menciona.

Desde luego que tales fórmulas serían en extremo complicadas; pero el hecho de que lo sean no legitima -si es que los cultores de la «objetividad» quieren ser coherentes- que a los efectos del cómputo de tamaños se pueda actuar como si ellos no fueran luego a ser estimados. Tal conducta es análoga a la de buscar las llaves extraviadas al lado de un farol, no porque se hayan perdido ahí sino porque esa es la zona iluminada.

Aunque el pormenorizado recorrido que hemos hecho se ha referido al proceso de estimación, virtualmente todo lo que se ha dicho es válido para el caso en que lo que se procura es determinar el tamaño muestra1 para un estudio analítico, como se resume a través de las siguientes seis observaciones.

a. Las decisiones previas (α, β, P_1, P_2, etc.) son igualmente subjetivas.
b. A menudo no se hace una sola prueba de significación; quizás, se realizan 10 o 15.
c. Ocasionalmente se hacen comparaciones entre parámetros inherentes a subconjuntos de la población original (ese es el caso frecuente, por ejemplo, de la postestratificación).
d. Los problemas de costo son tan influyentes en este tipo de estudios como en los descriptivos.
e. Aunque se estime el tamaño de muestra necesario para probar la diferencia de dos porcentajes o evaluar la diferencia entre una ***odds ratio y*** 1, no es

[11] Nótese que en esta formulación no aparecen la confiabilidad ni el error porque el criterio usado no se basa en poner una cota al intervalo de confianza sino en optimizar la precisión dentro de una restricción presupuestaria.

infrecuente que en el mismo estudio se hagan pruebas para contrastar, además, otras hipótesis, por ejemplo relacionadas con coeficientes de correlación o de regresión.

f . Una fórmula como [ll.4] se deduce a partir del supuesto de que se ha realizado un MSA, circunstancia casi desconocida en la epidemiología y la investigación clínica actuales.

Algunas de las fuentes de subjetividad que plagan a este proceso son ocasionalmente reconocidas en tal calidad por la literatura, pero uno puede hallar «explicaciones» doctrinarias como la que aparece en un reciente artículo (Mejía, Fajardo, Gómez *et* ***al.,*** 1995) cuyos nueve autores escriben:

> ***... podría parecer que la suposición de estos valores es extremadamente arbitraría; sin embargo, es mejor intentar esta aproximación a llevar a cabo el estudio sin intentarlo...***

Lo que no dicen Mejía y sus ocho colaboradores es por qué es mejor hacer suposiciones extremadamente arbitrarias.

11.5. Pseudosoluciones

Esta sección tiene como finalidad examinar ciertas «soluciones» que se han ofrecido para resolver algunas de las contradicciones arriba señaladas.

1) FIJAR UN PORCENTAJE DE LA POBLACIÓN COMO TAMAÑO MUESTRAL

En ocasiones se han hecho recomendaciones como la siguiente [12]: «si no tiene elementos para decidir el tamaño muestra1 de manera rigurosa, tome el 10% de la población para formar la muestra». Por ejemplo, al redactar una norma para realizar una auditoría de la gestión hospitalaria, en la que se fijan los pasos a dar por el equipo auditor, puede aparecer una orientación del tipo siguiente:

> SELECCIONAR EN CADA HOSPITAL UNA MUESTRA ALEATORIA FORMADA POR EL 15% DE LAS HISTORIAS CLÍNICAS INICIADAS DURANTE EL ÚLTIMO TRIMESTRE Y, SI EL PORCENTAJE DE HISTORIAS QUE TIENEN TAL CARACTERÍSTICA EXCEDE EL 25%, ENTONCES...

[12] No conozco, ciertamente, de textos serios que hagan esta recomendación. Este comentario se incluye no para hacer una crítica a lo que llamé «teoría oficial» sino para precaver a los lectores sobre una regla que aparece ocasionalmente en ambientes técnico-administrativos.

Ésta es una indicación improcedente. Con ella, y contrariamente a lo buscado, la evaluación de los hospitales grandes se verificará con extraordinario rigor, mientras que la de los pequeños puede ser en extremo imprecisa, ya que dependerá mucho más del azar. Es bien sabido que la calidad de una estimación depende sobre todo del ***tamaño absoluto*** de la muestra, y solo mínimamente del poblacional.

Si se quieren o se necesitan indicaciones de tipo general, hay que dar números absolutos. Por ejemplo, en ***World Fertility Survey (1975) se*** recomienda que los estudios nacionales de fecundidad se realicen con tamaños de 2.000 a 8.000 mujeres en edad fértil. Nótese el enorme margen de elección.

II) PARTIR DE QUE EL PORCENTAJE DE SUJETOS CON CIERTO RASGO ASCIENDE AL 50% PARA OBTENER EL MAYOR TAMAÑO DE MUESTRA SIMPLE ALEATORIA POSIBLE

Este error parece estar bastante extendido y es el que mejor evidencia el estado de hipnosis colectiva al que se aludió al comienzo. Es harto frecuente hallarlo en los textos de muestreo (Azorín y Sánchez-Crespo, 1986), metodología de la investigación (Argimón y Jiménez, 1991), epidemiología (Jenicek y Cleroux, 1987) y estadística (Domenech, 1990). Desde luego, también se muestra en el recetario a cargo de Lemeshow y sus tres colaboradores. Estos últimos hacen textualmente la tajante afirmación siguiente:

> ***Cuando el investigador no tenga la menor idea acerca de cuál puede ser el valor de P, sustituya 50 en lugar de P y siempre obtendrá suficientes observaciones, cualquiera que sea el verdadero valor de P.***

La fundamentación de esta mágica receta puede esbozarse así:

> ***Puesto que 50 es el valor de P para el cual*** *el producto* ***P(100 – P) es máximo, se asegura así el mayor valor*** *posible* ***para*** n_0. ***Éste nunca podría ser menor que lo que resulta de tal manipulación ya que*** n_0 ***es directamente proporcional a P(100 – P). Finalmente, puesto que, según la fórmula*** [1 1.1], ***a mayor valor de*** n_0, ***mayor es n, nunca sería posible obtener una muestra menor que la requerida.***

Sin embargo, se trata de una falacia, de una regla carente de fundamentación real. Analicémosla cuidadosamente.

Antes de entrar en detalles, hago una rápida invocación a la intuición del lector: ¿no sospecha usted que la estimación en una ciudad de la tasa de prevalencia de portadores de VIH demanda mayor tamaño de muestra que la del porcentaje de individuos necesitados de atención dental? Es bastante intuitivo que en este segundo caso una muestra de, por ejemplo, 100 sujetos podría servir. Pero el más elemental sentido común nos permite comprender que una muestra de 100 personas

tomadas de la población general será absolutamente insuficiente para estimar el primer parámetro: lo más probable es que ella no contenga a enfermo alguno (en cuyo caso, sacaríamos la absurda conclusión de que no hay portadores del virus en la ciudad); pero si cayera al menos un portador en la muestra, se concluiría en principio que la tasa es por lo menos l%, dato tan absurdo como el anterior, pues se sabe que la tasa es muchísimo menor.

De modo que es obvio que la estimación del porcentaje de portadores del virus demanda una muestra mucho mayor que la del porcentaje de individuos que requieren atención dental. Sin embargo, el primer porcentaje está ***muchísimo*** más lejos de 50% que el segundo.

Formalmente es cierto que la expresión $n_0 = \dfrac{(1{,}96)^2\, P(100 - P)}{E_0^2}$, como función de P, alcanza su máximo valor para $P = 50$. En tal caso:

$$n_0 = \frac{(1{,}96)^2\ 2.500}{E_0^2} \approx \frac{10.000}{E_0^2}$$

Pero ello sólo es válido, ***siempre que se suponga que E_o está fijado*** de antemano. Sin embargo, el máximo error absoluto E_o que se puede admitir a la hora de estimar P no *puede* ***establecerse razonablemente hasta que no se tenga una idea de la magnitud de P.*** Esto es así, del mismo modo que no sabemos si resulta caro un objeto que se vende al precio de 100 dólares mientras no sepamos de qué objeto se trata.

Imaginemos que se tiene cierta dolencia específica cuya prevalencia ha sido estimada, y que se sabe que el error en que se ha incurrido al hacerlo no excede al 1%. ¿Es grande o pequeño ese error? ¿Se ha conseguido estimar razonablemente bien el valor de esa prevalencia ***P?***

Si el lector medita durante unos segundos y procura responder a estas dos preguntas, no demorará en comprender que es ***imposible*** darles respuesta hasta tanto no se le comunique cuál es el valor de ***P o,*** al menos, de qué dolencia se trata.

En efecto, si se trata, por ejemplo, de la prevalencia de cáncer pulmonar (un número próximo a 1 en 20 000), un error de 1% sería descomunal; pero, si se trata de la prevalencia de hipertensión, tal error sería totalmente razonable pues con seguridad es menor que la décima parte de *P.*

Dicho de otro modo: cuando se va a calcular n_o, hay que preestimar *P,* no sólo porque aparece explícitamente en la fórmula [ll.2] sino porque sin ese conocimiento es ***imposible*** decidir el valor del error absoluto que también aparece en ella.

Imaginemos que le pedimos al Sr. Lemeshow (o a alguno de sus tres colaboradores) que calcule el tamaño muestra1 para un estudio que realizaremos en una población de 2.000 habitantes. Para simplificar las cosas diremos que se hará un MSA y que sólo queremos estimar un parámetro: el porcentaje de sujetos que poseen un componente sanguíneo denominado ***farsemia. Se*** trata de un componente que se tiene o no se tiene; pero «no tenemos ni la menor idea» de cuál será el porcenta-

je de sujetos que poseen ese rasgo. Ellos conocen N, pueden fijar α en 0,05, y suponer que $P = 50$ pero, ¿qué valor pondrán en lugar de E_o dentro de [11.2]? Carece, simplemente de sentido fijar ese número de manera razonable hasta que no se tenga una idea de la verdadera prevalencia de *farsemia* entre los seres humanos.

Lo que sí podría fijarse de antemano es el valor del error *relativo;* por ejemplo, podría decidirse que éste no sobrepasara al 10% de P: $E_r = \frac{E_o}{P} = 0{,}1$. Si se divide por P^2 tanto el numerador como el denominador de [11.2], se tendrá:

$$n_o = \frac{(1{,}96)^2 \ \frac{100-P}{P}}{E_r^2} \qquad [11.6]$$

De modo que para $E_r = 0{,}1$, al aplicar [11.6] se tiene que $n_o = 384 \ \frac{100-P}{P}$. Ya no aparece el fastidioso E_o en la fórmula. Pero es fácil ver que, a diferencia de $P(100 - P)$, la expresión $\frac{100-P}{P}$ no está superiormente acotada sino que tiende a infinito en la medida que P *se* aproxima a 0.

Dicho de otro modo: cuando se fija el error en términos relativos, n_o crece en la medida que P disminuye, un resultado que -como ya vimos- es coherente con la intuición: si el rasgo cuya prevalencia se quiere estimar es muy poco frecuente, entonces el tamaño de muestra necesario ha de ser muy alto.

Por último, para ilustrar prácticamente estas ideas sobre el error relativo, imaginemos que dos muestristas pretenden estimar mediante MSA el mismo parámetro P en una población de $N = 2.000$ unidades. El primero -llamémosle A - tiene motivos para creer que el valor P puede estar cerca de 11% y el otro -investigador B- «no tiene la menor idea» al respecto, de modo que aplicará la regla que indica Lemeshow para cuando no se tiene la menor idea: tomará $P = 50$. Ambos deciden trabajar con un nivel de confianza del 95% y error relativo del 10% (recordar que el investigador B no tiene derecho alguno a considerar aceptable ningún error absoluto); esto significa tomar como valor de E_o el 10% del valor supuesto para P.

Los resultados para los respectivos cómputos de la fórmula [11.1] (usando $\alpha = 0{,}05$) serán entonces los siguientes:

	Muestrista A	Muestrista B
Proporción P	11,0%	50,0%
Máximo error admisible E_o	1,1%	5,0%
Tamaño muestra1	1.217	322

Como se ve, el tamaño de muestra obtenido por ***B*** (que, supuestamente, es «el mayor que se podría obtener») resulta casi cuatro veces ***menor*** que el que obtuvo A. Ello se debe a que usaron el mismo error relativo (y, por ende, valores muy diferentes de ***E,).***

En síntesis, la regla que se ha examinado es absurda porque olvida que el conocimiento previo del valor de la prevalencia es necesario no sólo para sustituir en lugar de ***P*** en la fórmula, sino también para poder fijar E_0, «detalle» que la mágica sustitución por 50 no resuelve.

11.6. La solución del enigma

En Silva (1993) ya expuse que la mayoría de los textos y de los profesores prescinden de estas realidades. Una excepción notable se produce en el libro de Rothman (1986) quien, aunque sin desarrollar estas ideas *in extensis*, reconoce sin ambages la inviabilidad de una solución teórica cuando escribe:

> ***En resumidas cuentas, el problema de determinar el*** *tamaño* ***de muestra más adecuado no es de naturaleza técnica, susceptible de ser resuelto por vía de los cómputos sino que ha de encararse mediante el juicio, la*** *experiencia* ***y la intuición.***

Pero la mayoría de los autores propician que se consolide en estudiantes e investigadores la convicción de que, para cada problema existe ***un*** número único, independiente del enfoque personal, la intuición y la experiencia del investigador; un número que puede determinarse técnicamente por aquellos «elegidos», capaces de desentrañar complejas formulaciones. Y eso, como se ha visto, es simplemente falso.

Muchos metodólogos profesionales pueden poner (y, de hecho, lo hacen) en un serio aprieto a modestos investigadores exigiéndoles que justifiquen el tamaño muestra1 que han elegido por analogía con lo que han visto en la literatura, o porque es el que permiten sus recursos. Sin embargo, me temo que en la inmensa mayoría de los casos, los propios inquisidores se verían en similar dificultad si, en lugar de dedicarse a pedir respuestas, tuvieran que producirlas [13]. En tal caso, quizás acudirían a la aplicación de fórmulas que, como se ha explicado e ilustrado, contienen una carga de subjetividad acaso tan grande como la de quien elige el tamaño muestra1 guiado por su propio y saludable sentido común.

Lo importante es comprender que ***cualquiera que sea el tamaño de muestra, tanto los errores de muestreo como la probabilidad de rechazar erróneamente una hipótesis de nulidad pueden ser calculados a*** *posteriori*. 0 sea, tanto las fórmulas para el

[13] ***No*** en balde el notable sociólogo norteamericano Wright (1961) lanzaba desde varias décadas atrás la exhortación: «*¡Metodólogos: a* ***trabajar!»***

cómputo de errores como las de los estadígrafos en que se basan las pruebas de hipótesis contemplan los tamaños muestrales empleados y en ambos casos la estructura de esas fórmulas es tal que el investigador desconfiará de la validez de estos recursos en caso de que la escasa magnitud de la muestra así lo aconseje. Los investigadores se sienten a menudo desconcertados e inseguros por la simple razón de que se les impone una teoría desconcertante e insegura, plagada de incoherencias. Los teóricos hacen sus elegantes propuestas y, si surgen cuestionamientos a los absurdos en que se basan, miran hacia otro lado. Uno no puede menos que recordar la frase de Churchill: «En ocasiones, el hombre tropieza con la verdad pero, casi siempre, evita caerse y sigue adelante».

Los investigadores reales, en cambio, no pueden usar ese cómodo recurso pues necesitan verdaderamente de un tamaño muestra1 concreto, no para hacer manuales basados en recetas mágicas, sino para llevar adelante estudios tangibles.

¿Cuál es finalmente la recomendación que han de seguir estos últimos ante tan acuciante exigencia práctica, que no puede esperar por soluciones que hoy se ignoran?

Como siempre, roto el hechizo, todo es simple. No me sonrojo al decir que, a partir de los recursos disponibles, una excelente alternativa es usar el sentido común y tener en cuenta los tamaños usados en trabajos similares (es decir, incorporar el sentido común de los demás).

No casualmente la inmensa mayoría de los trabajos serios y trascendentes (por ejemplo, los publicados en revistas de impacto real como ***Lancet*** o ***British Medical Journal),*** no se preocupan en buscar taparrabos técnicos para explicar sus tamaños muestrales: se circunscriben a comunicar los que fueron usados. Muchos de estos investigadores, sin embargo, tuvieron que hacerlo al presentar el proyecto en procura de financiación; en tal caso la presentación de las fórmulas es, con frecuencia, puntualmente demandada. Es natural que los financiadores reclamen argumentos para el tamaño muestra1 propuesto, ya que de él dependen vitalmente los recursos que habrían de asignarse. Lo que no es natural es que se consideren relevados de responsabilidad tan pronto se les ofrezca un artificio numerológico. Quizás los practicantes del autoengaño consideren que mi sugerencia es herética pero, mientras no se den argumentos racionales en contra de los míos, tal acusación no confiere ningún género de aval al fraudulento balbuceo pseudotecnológico que ellos defienden. Como Galileo, creo en el suave poder de la razón. Si alguien demostrara que las recetas examinadas son mejores que mi sugerencia, no seré yo quien reniegue de usarlas.

Bibliografía

Argimón JM, Jiménez J (1991). ***Métodos de investigación. Aplicados a la atención primaria de salud.*** Doyma, Barcelona.

Azorín F, Sánchez-Crespo JL (1986). ***Métodos y aplicaciones del muestreo.*** Alianza, Madrid.

Cochran WG (1963). ***Sampling techniques.*** Wiley, New York.

Domenech JM (1990). ***Métodos estadísticos en ciencias de la salud.*** Unidad Didáctica 5, Gráficas Signo, Barcelona.

El País (1993). ***Triunfo claro de Aznar en el primer debate. 26*** de mayo, página 15, Madrid.

Hansen MH, Hurwitz WN, Madow WG (1953). ***Sample survey methods and theory.*** Wiley, New York.

Jenicek M, Cleroux R (1987). ***Epidemiología: principios-técnicas-aplicaciones.*** Salvat, Barcelona.

Kish L (1965). ***Survey sampling.*** Wiley, New York.

Lemeshow S, Hosmer Jr DW, Klar J, Lwanga SK (1990). ***Adequacy of sample size in health studies.*** Wiley, New York.

Lwanga SK, Lemeshow S (1989). ***Sample size determination in health studies: a user's manual.*** World Health Organization, Geneva.

Mejía JM, Fajardo A, Gómez A, Cuevas ML, Hernández H, Garduño J ***et al.*** (1995). ***El tamaño de muestra: un enfoque práctico en la investigación clínica pediátrica.*** Boletín Médico del Hospital Infantil de México 52:381-391.

Rothman JK (1986). ***Modern epidemiology.*** Little, Brown and Col., Boston.

Silva LC (1993). ***Muestreo para la investigación en ciencias de la salud.*** Díaz de Santos, Madrid.

World Fertility Survey (1975). ***Manual on sample designs.*** The Hauge: International Statistical Institute.

Wright C (1961). ***La imaginación sociológica.*** Revolucionaria, La Habana.

Yamane T (1970). ***Elementary sampling theory.*** Editorial R, La Habana.

12

Comunicación científica; el peso de la estadística

Los débiles de espíritu han escrito miles de libros vanos y vacuos (...) es un hecho fatal que lo que se encuentra en los libros es aceptado instantáneamente como la verdad, especialmente si los libros son viejos... pero no todo lo que ves en los libros es prueba convincente, el mentiroso miente tan fácilmente con la pluma como con la lengua.

MOISÉS MAIMÓNIDES

Uno de los asuntos que más inquieta al investigador es el proceso de comunicar sus resultados. Es una circunstancia enteramente natural puesto que, como se ha dicho, tratándose de conocimientos, difundirlos es en primera instancia lo único que puede hacerse con ellos. Entre los componentes de ese proceso, la estadística juega con frecuencia un papel relevante no sólo por la participación que pudo haber tenido en los hallazgos que han de comunicarse, sino porque con su ayuda los mensajes pueden ganar en transparencia y poder persuasivo.

La profundidad y el modo en que se informa la participación de la estadística en el proceso investigativo, así como la manera en que se inserta dentro del cuerpo de información que ha de transmitirse, constituyen motivos para la reflexión. En este capítulo se comparten algunas opiniones sobre el tema de la comunicación científica en general y sobre la intervención de la estadística en particular.

12.1. La función múltiple del artículo científico

Las formas de comunicación científica son diversas, y con la expansión de la informática y de las telecomunicaciones se van abriendo nuevas alternativas, tales como las novedosas ***revistas electrónicas*** (Huth, 1995). Sin embargo, el mecanismo por antonomasia para la transmisión de nuevos conocimientos es, y aparentemente

seguirá siendo, el artículo científico [1]. Los libros, presentaciones en congresos, reportes técnicos, monografías y demás opciones comunicativas siguen constituyendo importantísimas vías para el flujo de las ideas científicas, pero el artículo ha consolidado su prestigio como vehículo no sólo para compartir conocimientos e ideas sino para, al menos, otras tres funciones que se enumeran y comentan a continuación.

A) DAR CABIDA A UN DEBATE CIENTÍFICO-TÉCNICO MÁS INTEGRAL QUE CUALQUIER OTRO MEDIO

Podría pensarse que las discusiones personales (por ejemplo, en seminarios y congresos), tienen más vivacidad porque dan la posibilidad de un diálogo directo y sin demoras. En cierto sentido es así, pero con frecuencia se trata de una dinámica bastante superficial. Mi experiencia es que las preguntas, réplicas o inquietudes más interesantes no se generan *in situ,* a la misma vez que se toma contacto con el estímulo que habrá de producirlas, sino que se gestan tras un proceso de reflexión crítica y de cotejo con otras fuentes. Ponerlas por escrito en un artículo científico exige ponderación y cautela propias del texto escrito, siempre sujeto al escrutinio potencial de futuros polemistas. Por otra parte, el artículo científico es un vehículo evidentemente mucho más ágil y propicio para el debate que el de los libros. El artículo se halla, en fin, entre la volatilidad de la palabra hablada y el talante pausado del libro, entre la vorágine de las jornadas científicas y la lentitud de los volúmenes académicos.

B) EXAMINAR CUANTITATIVA Y CUALITATIVAMENTE LAS TENDENCIAS PREVALECIENTES EN LA PRODUCCIÓN CIENTÍFICA

El análisis bibliométrico, hoy favorecido por la existencia de bases de datos computarizadas y eficientes programas de búsqueda y recuperación de información, está desempeñando un papel cardinal en la identificación de las líneas de investigación en boga, e incluso en su redireccionamiento, ya que tanto autores como editores se valen de él para elegir áreas de trabajo y de difusión. La intensidad de este entramado de influencias mutuas, estructurado en torno al artículo científico, es un fenómeno novedoso que ha potenciado toda la producción científica contemporánea.

[1] Cuando en ocasión del 125 Aniversario de la revista *Nature,* posiblemente la más prestigiosa del mundo, se le preguntó a su director, John Maddox, cómo se imaginaba la ilustre publicación dentro de otros 125 años, respondió «Seguirá publicándose semanalmente, impresa en papel», adelantándose así a la suspicacia de que pudiera «cibernetizarse».

La investigación bibliométrica responde a una especie de «epidemiología de la literatura científica», con sus particulares tasas de incidencia y prevalencia, sus series cronológicas y hasta sus medidas de asociación. En relativamente pocos años se han generado y difundido diversos indicadores bibliométricos tales como el *índice de aislamiento* (porcentaje de citaciones realizadas en las publicaciones producidas dentro de cierto espacio geográfico, por ejemplo un país, que corresponden a ese mismo espacio) o el llamado *factor de impacto* de un artículo o de una revista (número de citas recibidas por el artículo o la revista en un lapso, usualmente de 2 años, luego de su publicación). Aunque su empleo universal ha sido cuestionado (Spinak, 1996), lo cierto es que por su conducto se mide, como el nombre indica, el impacto real que dicha producción ha tenido en la comunidad científica.

Por esa vía se han podido conocer datos que dan una medida del reducido impacto global que tiene una buena parte de la enorme producción científica actual. Por ejemplo, los sociólogos norteamericanos Cole y Cole (1972) afirmaban en la revista *Science* que la cantidad de científicos que consiguen hacer un aporte trascendente es mínima, ya que la inmensa mayoría de lo publicado no es siquiera citado en un lapso razonable después de su difusión. La situación no ha cambiado sustancialmente en 20 años. Piñero y Terrada (1992) señalan que aproximadamente un tercio de la literatura publicada no recibe cita alguna, la mitad recibe solamente una, y sólo el 4% de los trabajos son citados 4 o más veces.

Esta función de los artículos subraya la importancia de que las citas que se hagan sean correctas y precisas, y ocupen un lugar legítimo en los trabajos que las emplean, requisito cuyo cumplimiento deja, por cierto, bastante que desear, como bien documentan Eichorn y Yankaner (1987) y King (1987).

C) EVALUAR A LOS PROPIOS INVESTIGADORES

Los artículos científicos han constituido desde hace mucho la vía más natural para evaluar a los investigadores. Arnold Relman, editor de New *England Journal of Medicine,* los caracterizó 20 años atrás como «el registro colectivo de los logros académicos individuales» (Relman, 1977). Sin embargo nunca como en la actualidad habían sido tan usados con ese fin. Es cada vez más frecuente que, en lugar de solicitar voluminosos y muchas veces confusos *curricula*[2], las convocatorias de concursos para cubrir plazas académicas o los comités que conceden financiamientos o becas, soliciten de los aspirantes la presentación, por ejemplo, de aquellos 3 (o 5)

[2] He tenido la impresión de que la frondosidad de un *curriculum* vitae (apodado como *ridiculum vitae* en algunos medios) depende mucho de cuán acucioso sea el titular y de cuánto él se haya ocupado de inflarlo. Lo que ha publicado y, en menor medida, la docencia que ha impartido, son datos indicativos de tipo *hard;* las distinciones, cursos recibidos, cargos, congresos, etc., ofrecen información de tipo *soft.* Muchas páginas rellenas con este último tipo de datos dicen actualmente muy poco sobre su calidad científica. El número de jornadas científicas o congresos a los que ha asistido, por ejemplo, se desprende, más que nada, de coyunturas extracientíficas.

artículos de su producción que el propio investigador considere más indicativos de sus logros y potencialidades.

Ésta es una sana reacción no sólo contra la tendencia a considerar que el prestigio de los científicos depende de actividades más sociales que académicas sino contra la convicción de que es muy revelador el *número* de artículos científicos publicados.

12.2. La carrera publicadora

Durante mucho tiempo podía conseguirse renombre de manera automática publicando una gran cantidad de trabajos. Uno de los efectos de ese mecanismo ha sido la apetencia por aparecer muchas veces como autor o coautor, hecho que explica parcialmente el sostenido incremento en el número promedio de firmantes por artículo, tal y como revela el trabajo de Silva (1990).

El apremio por publicar alcanza, como nunca antes, no sólo a los investigadores a tiempo completo (que son relativamente pocos) sino también a los docentes universitarios ya que, como apuntaba recientemente Zolla-Pazner (1994), cada vez se da mayor importancia a la investigación, anteponiéndola a la docencia. La universidad se ha ido transformando en un laboratorio subsidiado por la industria privada. De manera que, a través del sistema de becas, contratos y mecanismos indirectos de estímulo, la publicación científica se ve altamente incentivada y, por tanto, resulta muy apetecida por los profesionales de casi todas las disciplinas. Según Casino (1996), solo en el mundo biomédico se publican cuatro millones de artículos anuales. La cifra puede ser exagerada; fuentes quizás más comedidas (Siegel, Cummings y Woodsmall, 1990) la situaban un lustro antes en poco más de dos millones, pero no caben dudas de que la avalancha informativa en este terreno es abrumadora.

En esta materia parece haberse llegado a un extremo: del harto conocido ***dictum: Quien no publica, perece,*** en algunos círculos parece que se ha consolidado otro que reza: ***La calidad del investigador es directamente proporcional al número de publicaciones*** *por* ***año.*** Algunas expresiones de este afán casi patológico son asombrosas. Por ejemplo, el afamado doctor Robert Good de un laboratorio de Manhatan, el investigador más citado de la historia según Broad y Wade (1982), en sólo 5 años de la década del 70 publicó 700 artículos; es decir, 140 artículos científicos por año, casi 3 semanales. Sobran los comentarios, exceptuando quizás el hecho de que este prolífico autor resultó envuelto en un caso de fraude del que fue directamente responsable William Summerlin, uno de sus colaboradores (y coautores). Dicho caso resultó quizás más sonado debido a que durante su rutilante carrera, Summerlin recibió ***grants*** del orden de 100 mil dólares para sus trabajos.

En un trabajo antiquísimo, del que luego se han hecho diversas versiones informales, Graham (1957) ofreció una visión irónica de los artículos científicos. Ya desde entonces existían serias inquietudes motivadas por el uso de recursos estandarizados con los cuales se suple el rigor y se consigue incrementar la productividad.

Graham «traduce» el significado de ciertas frases típicas de los artículos al lenguaje de la cruda realidad. Transcribo seis ejemplos:

1.	***Dice:***	Se sabe desde hace mucho tiempo
	Léase :	No me he tomado el trabajo de buscar la referencia original
2.	***Dice:***	. . . de gran importancia tanto teórica como práctica
	Léase:	... a un par de personas más le interesa el asunto
3.	***Dice:***	Probablemente se cumpla también para lapsos más largos
	Léase:	No tuve paciencia suficiente para comprobarlo
4.	***Dice:***	Estos resultados serán publicados más adelante
	Léase:	A ver si tengo la constancia de trabajar algo más
5.	***Dice:***	Es evidente que se requiere una considerable labor adicional para que se llegue a entender completamente
	Léase:	Yo no lo entiendo
6.	***Dice:***	Agradecemos a Smith su ayuda en la fase experimental y a Jones sus interesantes comentarios
	Léase :	Smith hizo el trabajo y Jones nos explicó su significado

Naturalmente, pudiera pensarse que se trata de un mero juego de ingenio que hizo fortuna [3], pero lo cierto es que frases de uso frecuente tales como «los resultados deben apreciarse cautelosamente» son como códigos acuñados a los que se apela con regularidad y que carecen de un sentido claro, ya que no es fácil imaginarse resultados que se puedan legítimamente valorar a tontas y a locas.

Quien escribe con el deseo de que le entiendan, usualmente es capaz de hallar recursos para conseguirlo. Ehrenberg (1982) sugiere que, siempre que sea posible, el autor se procure un crítico amistoso que haga anotaciones en un margen del borrador cuando algo no esté claro. Y en tal caso, jamás defendernos explicando lo que quisimos decir. Seguramente eso que se quiso comunicar no fue lo que se comunicó. Cuando alguien no ha entendido (incluyendo editores y árbitros) la responsabilidad es en primer lugar nuestra, ya que ellos responden a lo que hemos escrito.

Es frecuente tropezar con conflictos similares en la esfera específicamente estadística. Por poner solamente un ejemplo, al no haber obtenido un valor de *p* por debajo de los sacralizados 0,01 o 0,05, algunos autores comentan que, ***si bien los resultados no arrojan significación, debe tenerse en cuenta que, quizás con un tamaño de muestra mayor...*** Si se utilizó determinada prueba, se supone que haya sido porque se confió en su valor demarcatorio entre el azar y otra explicación. No es legítimo matizar determinado resultado cuando no nos gusta, pero jamás cuando es compatible con nuestras expectativas [4].

[3] El periódico madrileño *El País* reproducía recientemente una versión anónima recogida del *ciberespacio.*
[4] Sobre este particular, consúltese la Sección 6.4.1.

12.3. Todo autor es un rehén voluntario

En un estimulante artículo publicado por ***British Medical Joumal,*** McIntyre y Popper (1983) proponen un nuevo código ético, relacionado sobre todo con los errores que se cometen en el ambiente clínico. Allí subrayan que es mucho más importante aprender de los errores que cometemos que obtener nueva información, punto de vista que se halla en clara consonancia con el espíritu de este libro. Pero en otro punto señalan que la crítica racional debe estar signada por el afán de aproximarse a la verdad, razón por la cual debe ser ***impersonal. No*** me queda completamente claro qué quieren decir con ello los autores, quizás porque no entiendo el silogismo que lleva de la primera afirmación (con la cual es imposible no concordar) a la segunda, que considero altamente discutible.

Mi opinión es que, por supuesto, carece de sentido constructivo dirigir la crítica hacia ***las personas*** que supuestamente han errado. Pero en el caso de las publicaciones, no veo razón alguna para no mencionar con precisión la referencia del trabajo en que se ha cometido el presunto error; más aun, considero esencial que así se haga si es que realmente queremos acrisolar el proceso de obtención de nuevos conocimientos en general y preservar la calidad de las publicaciones. Sin esa información resulta imposible hacer el cotejo real entre el contenido de la crítica y el material que es objeto de ella. En su momento me sorprendió negativamente que en un lúcido y oportuno artículo destinado a denunciar lo que él llama ***«el escándalo de la mala investigación científica»,*** Altman (1994) no incluya un solo ejemplo del mal contra el que tan vehementemente reacciona. Me aflige reparar en que usualmente el engaño produce menos escándalo que un acto de cruda sinceridad; posiblemente, si superáramos el temor a producir este tipo de «escándalos», contribuiríamos mucho más eficientemente a desestimular los que motivan la justa demanda de Altman.

Creo que cada cual es rehén de lo que soberanamente ha decidido (y conseguido) publicar, y parto de antemano del supuesto de que el autor de un trabajo al que se le señale una presunta deficiencia no incurrirá en la necedad de considerarlo como una agresión a su persona sino que se alegrará, incluso, de que su trabajo, en lugar de pasar inadvertido, haya contribuido a dinamizar el libre flujo de las ideas. En última instancia, el factor de impacto no distingue entre citas «favorables» y citas «negativas» que recibe un trabajo.

Estimo, en síntesis, que todo autor tiene el deber de enmendar su error (íntima o públicamente, según proceda), pero también el derecho de reafirmarse en sus opiniones en caso de que así se derive legítimamente del análisis suscitado por la diferencia de opiniones.

A este respecto, suscribo la lúcida apreciación de Stuart Mill (1806-1873) cuando, un siglo y medio atrás, escribió [5]:

[5] Citado en Feyerabend (1974).

***Lo que hay de especialmente negativo en silenciar una opinión es que se trata de un robo a la especie humana, tanto a la presente generación como a la posteridad. Un robo incluso mayor para aquellos que discrepan de ella que para quienes la suscriben. Si la opinión es correcta, se** les **priva de la oportunidad de trocar el error por la verdad; si fuera errónea, pierden lo que es casi un servicio igual de grande: una percepción más viva de la verdad corno consecuencia de su confrontación con el** error.*

12.4. La estadística en los artículos científicos

Acaso por la íntima convicción que poseen muchos investigadores de que poseer esta condición los obliga a dominar las técnicas estadísticas [6], unida a la realidad de que pocos las dominan cabalmente, el estudio de la «calidad estadística» de los artículos publicados ha revelado desde muchos años atrás muy serias deficiencias.

Por ejemplo, Ross (1951) identificó graves insuficiencias de diseño en los trabajos de la época. Para ello se basó en una muestra aleatoria de 100 artículos tomados de 5 prestigiosas revistas médicas tales como ***Joumal of the American Association y Annals of Interna1 Medicine.*** En aquella etapa, el uso incorrecto de la estadística era de importancia relativamente secundaria, pues los problemas eran metodológicamente «anteriores» a la participación de la estadística como tal; errores tan burdos como la ausencia de controles en situaciones que lo exigen aparecen hoy muy raramente en revistas de cierta calidad mínima.

Pero 15 años después de aquel estudio, Schor y Karten (1966) encontraron que en el 53% de casi 300 trabajos procedentes de las revistas más connotadas (New ***England Joumal of Medicine, Annals of Medicine y Joumal of Clinical Investigation*** entre ellas), se cometieron importantes errores estadísticos. Y aún una década más tarde, Gore, Jones y Rytter (1976) detectan que más de la mitad de todos los trabajos de ***British Medica2 Joumal*** contenían al menos un error de índole estadística. Sheehan (1980), basándose en el citado estudio de Schor y Karten (1966) y en otro similar de Freiman ***et al.*** (1978), alerta a los lectores sobre cuán lejos de la verdad pueden ser conducidos como resultado de los «desmanes estadísticos» presentes en la literatura a que acceden cotidianamente.

Como reacción ante tales realidades, algunas revistas han adoptado la medida de fortalecer el equipo de revisores con personal avezado en la materia. El ejemplo de ***Circulation Research,*** revista que decidió someter todos los artículos que tuvieran relación con estadística a una revisión por parte de profesionales de alto nivel en esta rama, ha sido citado como altamente positivo (Rosen y Hoffman, 1978).

Como parte de los esfuerzos orientados a modificar favorablemente este panorama, últimamente se han publicado varias listas de recomendaciones acerca de

[6] Para una discusión detallada al respecto, véase el Capítulo 2.

cómo exponer los datos estadísticos y de qué aspectos han de ser contemplados en los artículos. En este sentido pueden consultarse los trabajos de Bailar y Mosteller (1988), Gardner, Machin, Campbell (1986) y Altman ***et al.*** (1986).

Aparte de los errores estadísticos que puedan cometerse, se suele producir un problema a mi juicio insuficientemente enfatizado. Me refiero al hecho de que el artículo científico casi siempre contiene -como debe ser- una detallada exposición de los procedimientos usados en el terreno experimental o para la observación, explica las técnicas de laboratorio o de campo usadas, así como el diseño utilizado, pero no es extraño que al final de la sección de ***Material*** y ***Métodos,*** se enumeren en una secuencia rutinaria los procedimientos estadísticos, desconcatenados de las preguntas formuladas en el trabajo, como si se tratara de cumplimentar un ritual. En tales casos hay que hurgar en los resultados para detectar el uso específico que se dio a uno u otro procedimiento.

Las técnicas estadísticas siempre han de tener una funcionalidad; se aplican para ayudar a responder algo concreto. La expectativa del lector es hallar un texto estructurado del modo siguiente: ***«Para responder la pregunta tal, se compararon las medias mediante la prueba de Fulánez; el examen de la asociación entre las variables X e Y se realizó mediante el coeficiente de Zutanovich y no mediante la técnica de Menganovsky debido a . ..»,*** y no la simple y anodina lista de procedimientos empleados.

12.5. Los congresos-dinosaurios en extinción

Se ha afirmado que los dinosaurios desaparecieron como consecuencia de su enorme tamaño, incompatible con el hábitat de nuestro planeta. A los congresos médicos podría esperarles el mismo destino, ya que crecen indeteniblemente [7]; con su tamaño, crece la confusión y la influencia del azar en los movimientos de los participantes.

Siendo, supuestamente, un espacio para el debate y la comunicación libre, este crecimiento no puede menos que inquietar, pues cada vez se hace más difícil de conseguir tanto lo uno como lo otro. En algunos enclaves, a los problemas organizativos que supone la participación de miles de delegados, se suma el peso que tienen los laboratorios farmacéuticos con sus mecanismos de financiación, el cual ya trasciende la mera presencia publicitaria para intervenir en casi todos los aspectos del congreso. Por ejemplo, el Comité Organizador del XIII Congreso semFYC de La Coruña (1993) denunciaba la masificación y «la actual parafernalia», que provoca disminución de la participación real de los delegados, así como la enorme dependencia de la industria farmacéutica, que tiende a hipotecar el contenido y la libertad del colectivo profesional.

La actividad científica que se desarrolla en jornadas de este tipo tiene, como se

[7] El record mundial lo ostentaba en 1995 la Sociedad Americana del Corazón ***(American Heart Association)*** que consiguió reunir en Anaheim, California a más de 33.000 delegados en su 68º congreso.

ha dicho, una dinámica muy peculiar. Según mi opinión el mayor interés de estos cónclaves radica en propiciar el encuentro vivo entre profesionales, donde la exposición y debate de ponencias tiene, objetivamente, un papel secundario, ya que resulta muy difícil seguir ***realmente*** el hilo de un discurso técnico comunicado en los plazos apremiantes que son típicos de la mayor parte de estas actividades.

Las sesiones ***de posters*** han sido uno de los mecanismos instrumentados en los últimos años para compensar esa limitación. Mi experiencia, sin embargo, es que resulta casi imposible que éstas consigan sacudirse el formalismo que las mediatiza; contribuyen, eso sí, a promover el conocimiento mutuo de las personas, mérito indiscutible de esta modalidad comunicativa.

Sintetizando, cualquiera sea la forma organizativa que adopte la comunicación en este contexto, la aspiración máxima es que los participantes se enteren de lo que sus colegas quieren decir, pues difícilmente se conseguirá que, además, capten lo que verdaderamente dicen. Los aspectos estadísticos, por su carácter instrumental, están por tanto entre los que más «sufren». No obstante, considero oportuno hacer algunas breves observaciones acerca de cómo enfrentar «estadísticamente» tanto la contingencia de exponer un trabajo científico en pocos minutos como la de asimilarlo.

La falta de experiencia (y, en ocasiones, de sentido común) lleva a que algunos expositores presenten transparencias repletas de números, tablas atiborradas de valores ***p*** y demás expresiones cuantitativas destinadas al limbo intelectual de auditores aburridos. Para decirlo rápidamente: personalmente me resisto a incluir (y, desde luego, a intentar leer) más de 4 o 5 cifras en una sola exhibición; ellas bastan usualmente para lo que realmente importa en estas circunstancias: compartir a grandes trazos las ideas centrales de un argumento. En cuanto al lenguaje, lo mejor es evitar alambicamientos inútiles. Pocas cosas resultan más patéticas que el afán por convertir en una ***hard science*** lo que todo el mundo puede entender si se comunica con sencillez. En su diccionario sobre el uso del idioma inglés, Fowler (1965) escribía:

> ***La sociología es una ciencia nueva que no se relaciona con materias esotéricas fuera de la comprensión del hombre común, sino que trata de los asuntos ordinarios de la gente ordinaria. Esto parece engendrar en aquellos que escriben sobre esa materia la convicción de que la carencia de complicaciones de su materia demanda de un lenguaje inextricable.***

En cuanto a la actitud del oyente, suele producirse una tendencia a perder el tiempo; y, ya que uno está dentro de un salón, lo razonable es tratar de sacarle partido. Cierta vez, en un aula donde estaba impartiendo un curso, hallé una hoja anónima con un decálogo de normas que deben seguirse cuando se escucha una ponencia. Creo que son buenas y, con ligeras adecuaciones, las transcribo; espero que su autor sepa disculpar que no me sea posible darle los créditos debidos.

1. Encontrar áreas de interés; preguntarse ¿qué significa esto para mí?
2. Evaluar el contenido aunque el expositor no sea un gran actor.

3. Dominar los sentimientos y no prejuzgar.
4. Escuchar ideas y no centrarse en los datos.
5. Ser flexible, tomar notas sólo de lo esencial.
6. Escuchar activamente, esforzándose por captar los mensajes básicos.
7. Evitar distracciones en lugar de fingir atención.
8. Mantener la mente abierta a lo nuevo.
9. Ejercitar la mente; no solo destinar la atención a los temas divertidos sino encarar los más complejos como un desafío intelectual.
10. Aprovechar la diferencia entre la velocidad del pensamiento y la de la expresión oral para sacar más partido a lo que se oye.

12.6. Los libros crecen y se multiplican

En un atractivo artículo que señala una serie de perversiones que, al decir de los autores, son «moneda corriente en el negocio de la publicación», Benach y Tapia (1995) consignan que en la actualidad se publican aproximadamente un millón de nuevos libros por año; de ellos, según López y Díaz (1995), unos 17.000 corresponden al sector biomédico. Como en tantas otras esferas, la oferta es apabullante y confusa. El consumidor intenta orientarse en ese bosque informativo y para él lo grave no es que tenga que elegir, algo que en una u otra medida siempre ha ocurrido, sino que se torna cada vez más difícil hallar *criterios* para hacerlo.

La informática computarizada ha venido a agilizar los procesos que implican alguna forma de elección. Sin embargo, en materia de selección de libros, no conozco de recursos informáticos que la optimicen.

En el caso de los libros especializados de estadística, el panorama es ciertamente frondoso, como permite apreciar, por ejemplo, una publicación mensual del *Instituto Internacional de Estadística* destinada a ofrecer una breve revisión de los libros nuevos sobre la materia. En su número de agosto de 1995 el *Short Book Reviews* (véase *International Statistical Institute, 1995)* reseña unos 20 libros y registra el arribo de unos 30 más. Teniendo en cuenta que esta fuente documental registra sólo una parte de los libros nuevos (por lo pronto, los que allí aparecen son solamente los escritos en idioma inglés, salvo alguna excepción), una estimación conservadora hace pensar que surge un libro nuevo de estadística cada día. ¿Qué uso se da a esta literatura en la producción científica actual? Una primera aproximación al modo como se usan los libros de estadística[8] en un importante segmento de la investigación biomédica actual del más alto nivel la brindan los resultados de un estudio de Silva, Pérez y Cuéllar (1995) varias veces mencionado en este libro (véanse detalles en la Sección 2.6).

[8] Se usa la expresión «libros de estadística» haciendo cierto abuso de lenguaje, pues se incluyen algunos que pertenecen a esta área en el espíritu, pero que no están dedicados a dicha disciplina en estado puro.

Los artículos publicados por ***American Joumal of Epidemiology*** (AJE) y New ***England Journal of Medicine*** (NEJM) en el período que va de 1986 a 1990, exhiben la situación que se sintetiza a continuación.

En los casi 1400 artículos publicados por NEJM se mencionan 120 libros diferentes de estadística; de ellos, 74 se citan una sola vez y solamente nueve, más de 10 veces. Estos últimos se relacionan[9] en la Tabla 12.1.

Tabla 12.1. ***Libros citados más de 10 veces en NEJM en el lapso 1986-1980***

Libros	*N.º de artículos en que fue citado*
Snedecor y Cochran (1980)	***34***
Kleinbaum, Kupper y Morgenstern (1982)	***31***
Fleiss (1972)	***28***
Armitage (1971)	***25***
Kalbfleisch y Prentice (1980)	***23***
Schlesselman (1982)	***21***
Breslow y Day (1980)	***18***
Winer (1971)	***13***
Siegel(1956)	***12***

En los más de 1.000 artículos que aparecen en AJE durante el período, se citan 150 libros diferentes. 91 se mencionan una sola vez, y sólo diez de ellos más de 10 veces en el quinquenio. Éstos son los que se presentan en la Tabla 12.2.

Tabla 12.2. ***Libros citados más de 10 veces en AJE entre 1986 y 1990***

Libros	*N º de artículos en que fue citado*
Breslow y Day (1980)	112
Kleinbaum, Kupper y Morgenstern (1982)	78
Fleiss (1972)	59
Schlesselman (1982)	57
Snedecor y Cochran (1980)	39
Armitage (1971)	35
Kalbfleisch y Prentice (1980)	26
Cox (1966)	23
Lee (1980)	12
Bishop, Fienberg y Holland (1975)	11

[9] La mayor parte de los libros mencionados han tenido muchas ediciones. En el caso de Snedecor y Cochran, por tomar un ejemplo, la de 1980 es la séptima de Ames, Iowa State University, pero muchas de las citas corresponden a la edición de 19.57 o posteriores. Para los listados confeccionados se ha elegido una de las citas para cada libro en caso de que existan más.

Resulta muy significativo que ambos listados contengan, en esencia, los mismos libros: en el caso de los primeros siete, aunque en un orden diferente para una y otra revista, la coincidencia es absoluta. En definitiva, uniendo los dos listados, se identifica que los libros de estadística con los que se trabaja de manera más o menos consistente se reducen virtualmente a 12. Prácticamente todos son libros de perfil amplio, publicados mucho tiempo atrás (18 años como promedio respecto de la fecha del artículo en que se cita [10]) y que contienen los métodos más generales.

Varios investigadores biomédicos me han pedido que les sugiera un material bibliográfico del que fiarse en materia estadística: algún libro especializado que les ayude con el manejo de la estadística en sus tareas de investigación y que cumpla a la vez las condiciones de ser claro, eficiente, relativamente abarcador y no excesivamente técnico. Este es en mi opinión un afán completamente justificado, pues dentro de la profusión mencionada, hay cientos, y quizás miles, de libros muy similares en la línea sugerida por títulos como «Introducción a la inferencia estadística» o «Estadística aplicada a...»

El grado de especialización alcanzado en la actualidad conspira contra la identificación de un libro único; consiguientemente, es virtualmente imposible responder de modo categórico a una pregunta como esa. Adicionalmente, es muy difícil hallar un material uniformemente mejor que los demás; siempre habrá temas mejor tratados en unos que en otros. Finalmente, inevitablemente mediará la propia sensibilidad (y el conocimiento) del que se pronuncia, tal y como ocurre cuando se elige la mejor película del año o el gol más notable del mundial de fútbol. No obstante, intentando satisfacer la demanda, mencionaré algunos materiales que considero altamente recomendables. El criterio general que he tenido en cuenta para esta selección ha sido el de la claridad y el carácter «amistoso» de la exposición. El resultado de este análisis se resume en los 5 libros siguientes:

1) *Statistics in operation* de Castle (1979) es un pequeño y magistral libro que intenta, y consigue, repasar con rigor y sin aridez alguna una amplia gama de problemas prácticos en el manejo de la estadística para la investigación clínica y epidemiológica ***sin hacer uso de una sola fórmula.***
2) Como libro general de estadística, sugiero tener a mano el texto titulado *Statistical methods* de Snedecor y Cochran (1980), cuya primera edición ronda el medio siglo de vida, pero que no casualmente es el más utilizado por los autores de *New England Journal* ***of*** *Medicine.* Quizás no sea suficientemente elemental para algunos colegas principiantes; constituye, sin embargo, un representante muy completo y claro de los enfoques estadísticos clásicos.
3) *Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica,* de Riegelman y Hirsch (1992), es un libro singular, más actualizado

[10] Para hacer este cómputo se tuvo en cuenta la fecha de la edición citada en el artículo.

el lenguaje y en los conceptos que el anterior. Se especializa en el ambiente clínico y su estructura se aleja de toda ortodoxia, pues se centra en la literatura publicada, con lo cual asegura un sentido práctico, que consigue consolidarse a través de ejercicios originales y estimulantes.

4) En relación con los métodos estadísticos para el trabajo de investigación específicamente en epidemiología, sugiero sin dudar el excelente libro *Statistical methods in epidemiology* de Khan y Sempos (1989), material que aborda con acierto diversos contenidos de mediana complejidad.

5) Para el caso particular de los estudios de casos y controles, forma dominante de la investigación epidemiológica, el libro más didáctico y transparente es en mi opinión *Case-Control studies* de Schlesselman (1982).

12.7. Lo que no se comunica

12.7.1. Sesgo de publicación

El llamado *sesgo de publicación,* así denominado por primera vez por Smith (1980), consiste en la tendencia a favorecer la publicación de ciertos trabajos en función de los resultados obtenidos y no necesariamente de sus méritos (pertinencia de la pregunta planteada y transparencia metodológica). Desde hace unos años se ha venido tomando creciente conciencia sobre el problema, pero llegó a tener una entidad tal que llegó a ser irónicamente identificado (Rosenthal, 1979) como el *file drawer problem:* las revistas se llenan con el 5% de los estudios que se realizan (aquellos en que se ha producido el error de Tipo 1) mientras que las gavetas de los laboratorios se llenan con el restante 95% de los trabajos (aquellos para los que no se obtuvo significación).

La existencia de tal sesgo ha sido ampliamente fundamentada (véanse los trabajos de Sterling, 1959; Smart, 1964; Coursol y Wagner, 1986; Shadish, Doherty y Montgomery, 1989; Dickersin, 1990; Easterbrook *et al.,* 1991), pero particularmente sugestivo es el experimento que llevó adelante Mahoney (1977) y que se bosqueja parcialmente a continuación [11].

Se tomaron 75 árbitros de una misma revista con los cuales se conformaron 5 grupos mediante asignación aleatoria. A cada árbitro se le entregó un artículo científico y se le recomendó la tarea de evaluar varios aspectos en una escala de 0 a 6. Entre ellos se hallaban los siguientes:

MÉTODOS
PRESENTACIÓN DE LOS DATOS

[11] La información está tomada de Dickersin (1990).

CONTRIBUCIÓN CIENTÍFICA
MÉRITOS PARA SER PUBLICADO.

Se prepararon 4 artículos idénticos tanto en su ***Introducción*** como en su sección de ***Métodos,*** pero con diferentes contenidos en las otras dos: ***Resultados*** y ***Discusión.*** En uno de ellos los resultados eran «positivos» (se demostraba la hipótesis en discusión); en otro eran «negativos»; en el tercero y el cuarto se pusieron resultados mixtos (tanto positivos como negativos), pero en un caso con una discusión que conducía a «conclusiones positivas», y en el otro una discusión con «conclusiones negativas». Cada uno de estas cuatro versiones fue asignada a uno de los grupos de árbitros; al quinto grupo se le entregó el artículo sin incluir las últimas dos secciones; a los árbitros de este último, naturalmente, no se le pedían criterios sobre la presentación de resultados.

Los resultados, como podrá apreciar el lector en la Tabla 12.3, son sumamente elocuentes del prejuicio prevaleciente entre los profesionales llamados a decidir si se produce o no la publicación.

Tabla 12.3 ***Evaluación de manuscritos con diferentes presentaciones***

Presentación	Número de árbitros	Métodos	Presentación de datos	Contribución científica	Méritos para publicarse
Resultados positivos	12	4,2	4,3	4,3	3,2
Resultados negativos	14	2,4	2,6	2,4	1,8
Resultados mixtos y conclusión positiva	13	2,5	1,3	1,6	0,5
Resultados mixtos y conclusión negativa	14	2,7	2,0	1,7	1,4
Métodos solamente	14	3,4	—	4,5	3,4

El más grave problema que este sesgo trae consigo es el de que distorsiona la visión que tenemos de la realidad y compromete la objetividad. Cada trabajo puede ser objetivo, pero la mirada global deja de serlo si se suprime una parte de la realidad. Naturalmente, una de las áreas que se ve más seriamente limitada por este concepto es el ***metaanálisis*** cuyo fin declarado es el de formalizar la identificación del consenso.

í2.7.2. **Sesgo de ética**

Existe una forma de no comunicación que no se genera por inducción de los editores ni por la propia autocensura del autor sino, directa y simplemente, como resultado de transgresiones éticas que quieren ocultarse.

Recientemente, un profesor de física teórica de la Universidad Autónoma de Madrid, reaccionaba en el periódico madrileño *El País* con gran preocupación por una serie de acontecimientos «científicos» que no fueron publicados y que sólo fueron conocidos como resultado de infidencias o indagaciones periodísticas (Sánchez, 1994).

Entre tales acontecimientos, menciona la noticia que daba cuenta de experimentos secretos realizados a finales de los años cuarenta y hechos públicos por la *Secretaría de Energía* de Estados Unidos. La experiencia consistió en inyectar plutonio a 18 personas que ignoraban su condición de conejillos de indias. No fue un acontecimiento aislado. En 1949, la Reserva Nuclear de Hanford liberó experimentalmente grandes volúmenes de material radiactivo (xenón-133 y yodo-131) hacia la atmósfera para determinar el patrón de diseminación de dichas sustancias. Como resultado se produjo una enorme contaminación cuya magnitud podría superar, según Sánchez, en 1.000 veces la del famoso escape radiactivo de Pennsylvania en 1979; los detalles de aquella experiencia se mantienen clasificados como secretos por el Gobierno de Estados Unidos. Los habitantes de la zona contaminada sólo conocieron el hecho en 1986, gracias a expertos medioambientalistas que obtuvieron algunos informes del Departamento de Energía. *The Albuquerque Tribunes,* rotativo de Nuevo México, Estados Unidos, dio a conocer en 1994 el caso de al menos dos mil personas que fueron sometidas sin saberlo a experimentos con radiactividad entre 1920 y 1989.

Mucho más conocido es el experimento de Tuskegee, en que se dejó sin tratamiento en la década del 30 a cientos de negros de Alabama que padecían de sífilis con el fin de monitorizar la evolución de la enfermedad, experiencia que se prolongó durante 4 décadas, hasta que un periodista dio a la publicidad la historia (Jones, 1981); no menos estremecedores son los experimentos de Willowbrook, en que deliberadamente se indujo hepatitis a niños retrasados para probar una vacuna.

Estimo que estos testimonios hablan con elocuente crudeza de un fenómeno que evidentemente desborda el ámbito de la ciencia y se interna en el del respeto a la condición humana. Se trata de un asunto de máxima actualidad, como revelan el creciente interés despertado por la bioética y el intenso debate que se produce en torno a la genética. La exigencia de que toda experiencia científica sea transparentemente comunicada ha de estar en el ánimo de todos. La reflexión y la toma de posiciones en torno a la ética científica y tecnológica, en sus innumerables y a veces sutiles expresiones, parece ser una gran avenida que sólo se está empezando a recorrer con verdadero rigor.

Bibliografía

Altman DG, Gore S, Gardner J, Pocock SJ (1986). *Statistical guidelines for contributors to medical journals* British Medical Journa1286: 1489-1493.

Altman DG (1994). ***The scandal of poor medical research: we need less research, better research, and research done for the right reasons.*** British Medical Journal 308: 283-284.

Armitage P (1971). ***Statistical methods in medical research.*** Blackwell Scientific Publications, Oxford.

Bailar III JC, Mosteller F (1988). ***Guidelines for Statistical Reporting in Articles for Medical Joumals.*** Annals of Interna1 Medicine 108: 266-273.

Benach J, Tapia JA (1995). ***Mitos o realidades: a propósito de la publicación de trabajos científicos***. Mundo Científico 15: 124-131.

Bishop YMM, Fienberg SE, Holland WP (1975). ***Discrete multivariate analysis: theory andpractice.*** The MIT Press, Cambridge.

Breslow NE, Day NE (1980). ***Statistical methods in cancer research.*** IARC, Lyon.

Broad W, Wade N (1982). ***Betrayers of the truth: Fraud and deceit in the halls of science.*** Simon and Schuster, Inc., New York.

Casino G (1996). ***Bytes médicos.*** El Mundo, Sección Salud, Madrid, p. 7, ll de febrero.

Castle WM (1979). ***Statistics in operation.*** Churchill Livingstone, Edinburgh.

Cole JR, Cole S (1972). ***The Ortega hypothesis.*** Science 178: 368-375.

Comité Organizador del XIII Congreso semFYC de la Coruña (1993). ***Se invita a la reflexión sobre los congresos.*** Atención Primaria 14: 763-764.

Coursol A, Wagner EE (1986). ***Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias.*** Professional Psychology Research Practice 17: 136-137.

Cox DR (1966). ***Analysis of binary data.*** Methnen, London.

Dickersin K (1990). ***The existence of publication bias and risk factors for its occurrence.*** Journal of the American Medical Association 263: 1385-1389.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR (1991). ***Publication bias in clinical research.*** Lancet 337: 867-872.

Eichorn P, Yankaner A (1987). ***Do authors check their references? A survey of accuracy of references in threepublic health Joumals. 77:*** 1011-1012.

Feyerabend P (1974). ***Contra el método.*** Ariel Quincenal, Barcelona.

Ehrenberg ASC (1982). ***Writing technical papers or reports.*** The American Statistician 36: 326-329.

Fleiss JL (1972). ***Statistical methods for rates and proportions.*** Wiley, 1.ª ed, New York.

Fowler HW (1965). ***A dictionary of modem English usage. 2.ª*** ed, Oxford University Press, New York.

Freiman JA et al (1978). ***The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of*** 71 ***«negative» trials New*** England Journal of Medicine 299: 600-694.

Gardner J, Machin D, Campbell MJ (1986). ***Use of check lists in assessing the statistical content of medical studies.*** British Medical Journa1292: 810-812.

Gore S, Jones IG, Rytter EC (1977). ***Misuses of statistical methods: critica1 assessments of articles in BMJ from January to Mar-ch, 1976.*** British Medical Journal 1:85.

Graham CD Jr (1957). ***A glossary for research reports.*** Metal Progress 71: 75.

Huth EJ (1995). ***La publicación electrónica en ciencias de la salud.*** Boletín de la Oficina Sanitaria Panamericana 118: 529-536.

International Statistical Institute (1995). ***Short Book Reviews.*** Vo1 15, No 2, Voorburg.

Jones JH (1981). ***Bad blood: The Tuskegee syphilis experiment.*** The Free Press, New York.

Kalbleisch JD, Prentice RL (1980). ***The statistical analysis of failure time data.*** Wiley, New York.

Khan HA, Sempos CT (1989). ***Statistical methods in epidemiology.*** Oxford University Press, New York.

King J (1987). ***A review of bibliometric and other sciences indicators and their role in research evaluation.*** Journal of Information Science 13: 261-276.

Kleinbaum DG, Kupper LL, Morgenstern H (1982). ***Epidemiologic research: principles, and quantitative methods.*** Lifetime Learning Publications, Belmont, California.

Lee ET (1980). ***Statistical methods for survival data analysis.*** Lifetime Learning Publications, Belmont, California.

López JA, Díaz S (1995). ***Problemas y tendencias actuales de la información científicomedica.*** Revista Cubana de Salud Pública 21: 119-125.

Mahoney MJ (1977). ***Publication prejudices: an experimental study of confirmatory bias in thepeer review system.*** Cognitive and Therapy Research 1: 161-175.

McIntyre N, Popper K (1983). ***The critica1 attitude in medicine: the need for a new ethics.*** British Medical Journa1287: 1919-1923.

Piñero JM, Terrada ML (1992). ***Los indicadores bibliométricos y la evaluación de la actividad médico-científica. (III) Los indicadores de producción, circulación y dispersión, consumo de la información y repercusión.*** Medicina Clínica 98: 142-148.

Relman AS (1977). ***Publish orperish -or both. New*** England Journal of Medicine 297: 724-725.

Riegelman RK, Hirsch RP (1992). ***Cómo estudiar un estudioyprobar unaprueba: lectura crítica de la literatura médica.*** Publicación Científica n.º 531, Organización Panamericana de la Salud, Washington.

Rosen MR, Hoffman BF (1978). ***Statistics, biomedical scientists, and Circulation research.*** (editorial) Circulation Research 42: 739.

Rosenthal R (1979). ***The file drawerproblem' and tolerance for null results.*** Psychological Bulletin 86: 638-641.

Ross OB Jr (1951). Use ***of controls in medical research.*** Journal of the American Medical Association 145: 72-81.

Sánchez M (1994). ***La ciencia y el infierno de Dante. El País, 20*** de octubre, Madrid.

Schlesselman JJ (1982). ***Case-Control studies.*** Oxford University Press, New York.

Schor S, Karten 1 (1966). ***Statistical evaluation of medical joumal manuscripts.*** Journal of the American Medical Association 195: 1123-1128.

Shadish WR, Doherty M, Montgomery LM (1989). ***How many studies in the file drawer?: an estimate from the family/marital psychotherapy literature.*** Clinical P chology Review 9: 589-603.

Sheehan TJ (1980). ***The medical literature: Let the reader be aware.*** Archives of Internal Medicine 140: 472-474.

Siegel S (1956). ***Nonparametric statistics for the behavioral sciences*** McGraw-Hill, New York.

Silva GA (1990). ***La autoría múltiple y la autoria injustificada en los artículos científicos.*** Boletín de la Oficina Sanitaria Panamericana, 108: 141-1.52.

Silva LC, Pérez C, Cuéllar I (1995). ***Uso de métodos estadísticos en la investigación publicada en dos revistas médicas con alto factor de impacto.*** Gaceta Sanitaria — en prensa.

Smart RG (1964). ***The importance of negative results in psychological research.*** Canadian Psychology 5: 225-232.

Smith ML (1980). ***Publication bias and meta-analysis.*** Evaluation and Education 4: 22-24.

Snedecor GW, Cochran WC (1980). ***Statistical methods.*** The Iowa State University Press, Ames.

Spinak E (1996). ***Los análisis cuantitativos de la literatura científica y su validez para juzgar la producción lationamericana.*** Boletín de la Oficina Sanitaria Panamericana 120: 139-145.

Sterling TD (1959). ***Publications decisionsand their possible effects on inferences drawn from tests of significance, or viceversa.*** Journal of the American Statistician Association 54: 30-34.

Winer BJ (1971). ***Statistical Principles in Experimental Design.*** Second Ed., McGraw-Hill, New York.

Zolla-Pazner S (1994). ***Profesor, universidad e industria.*** Investigación y Ciencia 211: 84.

13

Superstición y pseudociencia: la estadística como enemigo

Cuanto menos razón tiene un hombre para suponerse en lo cierto, tanto mayor vehemencia emplea para afirmar que no hay duda alguna de que posee la verdad absoluta.

BERTRAN DRUSSELL

Trataremos aquí el fenómeno de la pseudociencia. Como su nombre sugiere, puede considerarse como tal a cualquier sistema de ideas que, aparentando -explícita o implícitamente- ser una expresión de la ciencia, dimana del pensamiento fantástico de sus creadores o paladines, antes que de un vínculo verdadero con la realidad objetiva.

Además del interés intrínseco de examinar la pseudociencia con el fin de caracterizarla y contribuir a su desenmascaramiento, ocurre que ella tiene una curiosa relación con la estadística. Por una parte, debido a que ésta ha sido ocasionalmente usada como escudo por algunos de quienes se empeñan en defender los resultados de aquella; por otra parte y sobre todo, por contraste, debido a la naturaleza profundamente incompatible entre la forma de pensar que caracteriza a los cultores de una y otra.

Conviene que este examen sea lo más objetivo y desapasionado posible con el fin de no incurrir en los mismos vicios que se quieren denunciar. En el afán de desenmascarar un engendro que, con toda razón, suele producir irritación en cualquier mente científicamente cultivada, se han hecho caracterizaciones de la pseudociencia que resultan estrechas e incompletas. Tal es el caso de la que ofrecen Bueno, Hidalgo e Iglesias (1987); según ellos:

> ***Las pseudociencias pueden definirse como un conjunto de creencias y prácticas, cuyos cultivadores desean, ingenua o maliciosamente, hacer pasar por ciencia, sobre la base de acceso privilegiado a ciertos fenómenos y fuentes secretas de poder que se les escapan al común de los mortales.***

Esa definición resulta estrecha porque, si bien los adivinos profesionales o los creadores de cartas astrales afirman poseer, en efecto, dones suprahumanos, los pseudocientíficos no necesariamente se declaran usufructuarios activos de tales «poderes»; por ejemplo, algunos individuos actúan como portavoces de las teorías parapsicológicas y las defienden como presuntas verdades científicas sin considerarse a sí mismos entes capaces de usar recursos como la telepatía o la telequinesia.

La definición es por otra parte incompleta, ya que excluye a otro tipo de pseudocientíficos: aquellos que, como los divulgadores de la teoría de biorritmos o de la magnetoterapia, no se presentan como elegidos por la providencia sino que sustentan sus puntos de vista a partir de presuntas leyes científicas que otros no conocen o no han alcanzado a comprender. La prédica de estos últimos es, a mi juicio, más inquietante, porque no sólo explota la credulidad de la gente llana sino que puede confundir a profesionales de la ciencia.

Al decir de Bunge (1978):

> ***Lo malo de la seudociencia es no sólo ni precisamente el que sea básicamente falsa puesto que todas nuestras teorías factuales son, a lo sumo, parcialmente verdaderas, sino que no puede fundamentar sus doctrinas porque rompe totalmente con nuestra herencia científica.***

De las múltiples expresiones que tiene esta manifestación del intelecto, las supersticiones y sus innúmeras formas específicas configuran el exponente más primitivo; lamentablemente, son también las que gozan de mayor crédito entre amplias franjas de población, en buena medida porque a esa perpetuación del oscurantismo contribuyen activamente la prensa y la televisión. Shamos (1995) da cuenta de que, de 1.600 periódicos que se publican en los Estados Unidos, alrededor de 1.400 contienen una columna diaria con el horóscopo; de ellos, menos de 50 consignan que éste carece de fundamento científico y se presenta exclusivamente con el fin de entretener a los lectores.

La creencia en adivinadores, mediums e iluminados diversos reposa casi siempre en anécdotas que se difunden de boca en boca y anestesian el discernimiento crítico de los azorados receptores. Sin reparar en que tales personajes hacen uso de toda una serie de recursos basados en su penetración psicológica, los clientes dan muestras del asombroso potencial, probablemente congénito, que posee el ser humano para la ingenuidad.

Según recoge Hernández (1993), la organización española ***Alternativa Racional a las Pseudociencias*** ha clasificado a los videntes en tres categorías: el lógico, el sensacionalista y el generalizador. La caracterizaciones respectivas se exponen a continuación:

Lógico: Utiliza, sobre todo, el sentido común. Dispone de buena información general y con cada cliente va creando un

auténtico archivo de datos. Es el de más éxito y el más sistemático en su trabajo. Debe ser buen psicólogo y observador para captar las emociones de sus clientes y saber cuándo sus predicciones gustan a quien le paga.

Sensacionalista: No le importa hacer predicciones sobre cualquier cosa; el caso es hacerse famoso. Es muy poco riguroso, pero, a pesar de que sus equivocaciones son sonadas, suele conseguir popularidad.

Generalizador: Es el más clásico. Predice o adivina en abstracto: habla de un problema familiar, un viaje, un dolor de cabeza, algo que sucede a todos un montón de veces. Sus clientes, muy sugestionables, escuchan en sus abstracciones las predicciones que quieren oír.

Puesto que por lo general tales sujetos se conducen según pautas oportunistas y sus profecías no se ajustan al lenguaje y las categorías de la ciencia, es casi imposible encarar sus manejos con recursos formales.

No es nuestro propósito hacer un inventario exhaustivo de ese tipo de supercherías, tarea de magnitud enciclopédica, dada la impresionante proliferación de viejos y nuevos artificios que en esta materia padece la sociedad contemporánea. Por otra parte, sería algo esencialmente estéril, pues quien presta oídos a los charlatanes es típicamente sordo a todo razonamiento que se oponga a sus doctrinas. La creencia está por encima de toda razón para ellos; no en balde Murray Gell-Mann, ganador del premio Nobel de Física en 1969 señalaba (Collazo, 1995) que «la pseudociencia es la disociación entre la creencia y la evidencia». La pseudociencia que discutiremos ahora es la que se apropia indebidamente del prestigio de la ciencia e intenta ocupar su espacio.

Puesto que para conseguir sus resultados no demanda de cuidadosos estudios, ni exige comprobaciones rigurosas -a las que, por el contrario, teme- esta forma de la pseudociencia procura por todos los medios no someterse al dictamen de instrumentos objetivos como la estadística; quizás por ello el escrutinio detallado de sus paradigmas sea, paradójicamente, una de las vías más transparentes de ilustrar el modo de pensar de los estadísticos y del modo en que operan sus técnicas.

En materia de investigación la estadística contribuye a identificar los resultados útiles al conocimiento y a separarlos de los que no lo son; con más razón será eficiente entonces para desenmascarar aquellos que, simplemente, nacen de un pensamiento ajeno a la ciencia. Complementariamente, el contenido de este capítulo procura ayudar a conjurar la ingenua propensión hacia la credulidad que parece embargar a algunos colegas.

13.1. Pseudociencia y medicina al final del siglo

Desde siempre, una de las esferas preferidas por la pseudociencia ha sido la medicina. Como es bien sabido, la mayoría de los mitos surgieron históricamente como reacción al desamparo que generaba el desconocimiento de los fenómenos naturales. El organismo humano y las leyes que rigen su funcionamiento han planteado desde la antigüedad un apremiante desafío para los hombres: cómo enfrentar las desviaciones de la salud. Por razones obvias, éstas configuran uno de los dominios que promueve el interés universal, y la ignorancia al respecto producía y sigue generando un caldo de cultivo natural para la conformación de una mitología propia.

En su excelente libro ***El hombre anumérico,*** el divulgador científico norteamericano John Allen Paulos dedica un capítulo completo al tema de la pseudociencia. En él se plantea (Paulos, 1990):

> ***La medicina es un terreno fértil para las retensiones seudocientíficas por una razón muy sencilla. La mayoría de las enfermedades y estados físicos, o bien mejoran por sí solos, o bien remiten espontáneamente o, aun siendo mortales, rara vez siguen estrictamente una espiral descendente. En todo caso, cualquier tipo de intervención, por inútil que sea, puede parecer sumamente eficaz. Esto resulta más claro si uno se pone en el lugar de alguien que practica a sabiendas una forma de falsa medicina. Para aprovechar los altibajos naturales de cualquier enfermedad (o el efecto placebo), lo mejor es empezar el tratamiento inútil cuando el paciente esté empeorando. Así, cualquier cosa que ocurra será conveniente. Si el paciente mejora, se atribuye todo el mérito al tratamiento; si se estaciona, la intervención ha detenido el proceso de deterioro; si el paciente empeora, es porque la dosis o intensidad del tratamiento no fueron suficientemente fuertes; y si muere, es porque tardaron demasiado en recurrir a él.***

El efecto placebo a que alude Paulos es la modificación, muchas veces fisiológicamente constatable, que se produce en el organismo como resultado del estímulo psicológico inducido por la administración de un material inerte, de un fármaco o, más generalmente, de un tratamiento.

Cualquier médico inteligente ha utilizado este recurso (con o sin participación de fármacos), cuyo papel los textos académicos raramente exaltan. El efecto placebo puede ser el único que produce el tratamiento, o bien puede actuar ***además*** del que éste produzca por conducto bioquímico o físico. Al evaluar tecnologías terapéuticas se ha prestado muy poca atención a la separación de una y otra parte del efecto. El tema sin embargo es de máxima importancia, no sólo por motivaciones cognoscitivas sino por intereses prácticos.

Imaginemos que se valora la efectividad de un fármaco conformado a partir de cierto principio activo de tipo esteroideo para el tratamiento del asma. Como cualquier otra droga, ésta comporta ciertos riesgos, en este caso para el sistema cardiovascular; si se pudiera probar que el efecto del fármaco es esencialmente de tipo pla-

cebo, no haría falta someter al organismo a la agresión del esteroide y se obtendrían los mismos dividendos aplicando un tratamiento similar pero sin participación de ese principio activo.

Como se verá, esta circunstancia es de extremo interés para el examen de las pseudociencias médicas.

13.1.1. **Al rescate de la medicina medieval**

Desentendiéndose de la medicina altamente tecnificada, un número creciente de ciudadanos de Occidente elige el tratamiento con técnicas terapéuticas alternativas como la homeopatía y la quiropráctica.

Cada vez en mayor proporción, las personas enfermas (y también muchas que no lo están) buscan la recuperación o el mantenimiento de su salud en las tradiciones del Lejano Oriente, o en otros recursos esotéricos. Eisenberg *et al.* (1993) desarrollaron un estudio nacional en Estados Unidos para estimar la prevalencia y los costos correspondientes al uso de las prácticas médicas alternativas. Sus resultados son, ciertamente, impresionantes: uno de cada tres estadounidenses había consultado a alguno de estos proveedores de salud en el último año. Se estima que en total se produjeron 425 millones de visitas de este tipo en 1990; como elemento de referencia, cabe reparar en que la atención primaria había producido solo 388 millones de consultas; por otra parte, estos pacientes habían pagado por los servicios de medicina alternativa un total de 10.300 millones de dólares, monto similar al desembolsado ese mismo año por ciudadanos norteamericanos por concepto de hospitalizaciones, ascendente a 12.800 millones.

Las anécdotas (nuevamente ellas, antítesis de la valoración estadística) que se relatan en favor de tales prácticas llegan a ser impactantes. En su mayoría, cuando no constituyen exageraciones o, simplemente, falsedades, son obra de la casualidad [1] o pueden atribuirse al antiquísimo «efecto placebo», del que se beneficia, como ya se apuntó, casi cualquier intervención terapéutica, no sólo las de tipo alternativo.

Un artículo publicado en 1994 por ***Der Spiegel*** incluía el siguiente relato, presuntamente verídico. Durante un viaje de negocios a Nueva York, un hombre de California sufre una súbita y aguda dolencia bucal como consecuencia de la cual se ve obligado a visitar con urgencia a un estomatólogo. Este diagnostica una grave inflamación periodontal e indica un tratamiento de urgencia.

[1] La intervención del omnipresente azar se ve, por otra parte, favorecida por nuestra memoria selectiva. Los sucesos que se han vivido se retienen con más frecuencia (el llamado efecto *Jeane Dixon)* y las coincidencias inesperadas se toman como regla; los hechos menos llamativos, tienden a olvidarse. Como señala Paulos (1990), el número de muertos por el tabaco en un año equivale aproximadamente a tres aviones Jumbo estrellándose cada día. El SIDA, por muy trágico que sea, palidece si lo comparamos con los efectos de la malaria. Pero un solo desastre aéreo y el drama de los enfermos de SIDA parecen más trascendentes que los males cotidianos.

Antes de recibir acción terapéutica alguna, el enfermo decide consultar telefónicamente a su ***faith heder*** que se halla en el otro extremo del país. El sanador se informa detalladamente del caso y coordina con su cliente una consulta a distancia: exactamente a las ocho de la noche, en la habitación de su hotel, el enfermo habría de concentrarse y entrar en «contacto» con el sanador. El terapeuta, por su parte, estaría exactamente a esa misma hora enviando energías sanadoras al paciente. El hombre de negocios hace lo que se le ha indicado y, unas horas después, el dolor desaparece totalmente. A la mañana siguiente regresa a ver al dentista para que lo examine. El especialista neoyorquino primero no puede creer lo que oye, pero después no puede creer lo que ve: no hay rastros de la infección. ¿Se ha producido una cura milagrosa? Presa del asombro, decide llamar a California y consultar al responsable de tan portentoso efecto psicosomático para preguntarle qué diablos había hecho el día anterior a las cinco de la tarde (ocho de la noche, hora de la costa éste).

«¿Ayer a las cinco?» preguntó, sin comprender, el sanador. «Sí: ¿cómo procedió para tratar a distancia esa gingivitis?». La respuesta fue: «¿Curación a distancia? ¡Demonios, se me olvidó por completo esa cita!».

La pregunta relevante es: ¿se produjo un efecto placebo o se iba a verificar la curación de todos modos? Sólo la experimentación controlada y rigurosa puede responder genéricamente a esta pregunta. Por otra parte, cabe cuestionarse: si se hubiera verificado el intento telepático y el paciente no se hubiera curado, ¿se hubiera acaso difundido el hecho como reguero de pólvora? Y finalmente, si el terapeuta no hubiera olvidado «la cita», ¿no se le hubiera «apuntado» este éxito?

En general llama la atención cómo la suspicacia contra la superchería se adormece más fácilmente cuando atañe a la esfera médica. Cuando en medio de una contienda bélica se filtra una información secreta, nunca se piensa en la posibilidad de una exitosa interferencia telepática del enemigo: se trata, sin más, de identificar al traidor o al espía que lo hizo posible. Al menos así son las cosas en el siglo xx [2].

13.1.2 Experimentos para no viajar a China

Según Skrabanek y McCormick (1989), el actual interés por la acupuntura en Occidente data de la visita del presidente de Estados Unidos a China en 1970. Nixon y su séquito presenciaron un verdadero «espectáculo» mediante el cual se hizo creer a los visitantes que la aplicación de agujas en el pabellón auricular podía pro-

[2] Como anécdota lateral, cabe la siguiente nota histórica. Durante la guerra hispano-francesa del siglo XVI los españoles usaban una complejísima clave de 600 caracteres, que se consideraba imposible de descifrar. François Vieta (1540-1603), genial matemático aficionado al servicio del rey Enrique IV, recibió la encomienda de intentar la decodificación de un despacho militar interceptado, cosa que consiguió en breve plazo. Cuando Felipe II descubrió que sus secretos militares eran dominados por el enemigo, se quejó ante el Papa de los franceses con la acusación de que apelaban a la hechicería contra España en un acto de desacato a las doctrinas de la iglesia católica.

ducir poderosos efectos anestésicos. Esa fuente afirma que, en realidad, los pacientes que participaron en la demostración habían sido cuidadosamente escogidos y se les había administrado medicación previa a la operación. Ackerknecht (1974) ha señalado (Skrabanek y McCormick, 1989) que la ola de interés por la acupuntura que siguió al encuentro Nixon-Mao fue la quinta en llegar a Occidente desde el siglo XVII y que todas las anteriores finalizaron con la conclusión de que la acupuntura posee, en el mejor de los casos, una notable capacidad para sugestionar a los pacientes.

Como consecuencia del gran interés despertado por la acupuntura tanto en altos niveles de gobierno de los Estados Unidos, como en el Instituto Nacional de Salud y algunas universidades de ese país, esta forma de medicina alternativa ha sido allí la más exhaustivamente investigada. En 1977, la OMS estableció un Programa de Medicina Tradicional, que incluye la acupuntura junto con la fitoterapia (Shang, 1996). El efecto más importante que se le atribuye (véase, por ejemplo, el artículo de corte casi publicitario debido a Wong y Fung, 1991, en la importante revista ***Family Practice)*** es de carácter analgésico, especialmente en casos de osteoartritis, lumbago y migraña. Casualmente, se trata de uno de los efectos más difíciles de medir objetivamente, más dependientes del testimonio del paciente y, por ende, de su grado de sugestionabilidad. Se han consignado, además, presuntos efectos benéficos en el tratamiento de diversos trastornos cardiovasculares, respiratorios y nerviosos.

Numerosos ensayos controlados sobre el tema se han desarrollado en los últimos años. Una gran cantidad de ellos arrojan resultados compatibles con el hecho de que este procedimiento no tiene efecto alguno, o lo tiene equivalente al de un placebo.

Entre los trabajos recientes que examinan el papel de la acupuntura se hallan los de Thomas, Arner y Lundeberg (1992) sobre desórdenes dolorosos de origen idiopático, de Yentis y Bissonnette (1991) sobre la tendencia a vomitar en niños sometidos a una tonsilectomía, y de Lewis ***et al.*** (1991) por una parte y Yentis y Bissonnette (1992) por otra, ambos sobre esta misma tendencia pero tras una operación de estrabismo. Ninguno de ellos consiguió hallar evidencias a favor de la acupuntura. Ballegaard, Meyer y Trojaborg (1991) prueban que la acupuntura correctamente administrada sobre pacientes anginosos es equivalente a un tratamiento con «falsa» acupuntura (es decir, una aplicación en que las agujas se usaban en puntos elegidos al azar y no en los señalados por la teoría). Rogvi ***et al.*** (1991) en un completo estudio sobre ***oftalmopatía de Graves*** no hallaron mejorías significativas en ninguno de varios parámetros después de haberse practicado un tratamiento con agujas a portadores de esta dolencia. Tavola ***et al.*** (1992) niegan toda ventaja a la acupuntura por encima de un placebo a los efectos de la cefalea tensional. Ekblom ***et al.*** (1992) hallan experimentalmente que el efecto de aplicar acupuntura combinada con analgésicos en operados de la muela cordal da peores resultados que el uso exclusivo de analgésicos. Varios artículos de Petr Skrabanek (Skrabanek, 1984; Skrabanek, 1985; Skrabanek, 1986) desarrollan una fuerte crítica a la acupuntura y citan múltiples trabajos que se ubican en la misma línea.

Ciertamente, no faltan publicaciones que registran el éxito de la acupuntura en

sus diversas variantes. Sin embargo, existen motivos para temer por la falta de rigor metodológico que pudiera estar comprometiendo seriamente sus resultados. Eso hace pensar, por ejemplo, el estudio de Kleijnen, ter Riet y Knipschild (1991), publicado por la prestigiosa revista *Thorax. Se* examinaron 13 publicaciones sobre ensayos clínicos que abordan la eficiencia de la acupuntura para el tratamiento del asma. Para cada uno de ellos se examinaron 18 aspectos metodológicos referentes a tres importantes áreas: la población estudiada, la intervención realizada y el procedimiento de medición de los efectos. La calidad de los trabajos resultó ser mediocre; sólo 8 alcanzaron puntajes superiores al 50% y el máximo fue de 72%. Los pocos resultados que proclaman que la acupuntura es tanto o más eficiente que los métodos clásicos o que un placebo, proceden justamente de los artículos con más baja calidad. Estos autores comunicaron similares resultados en ***British Medical Journal*** cuando estudiaron 107 trabajos publicados sobre homeopatía (Kleijnen, Knipschild y ter Riet, 1991).

Además de este contradictorio panorama, todo el ambiente de la acupuntura está signado por claves místicas. Si los esotéricos «meridianos», el «mar de canales Ying y Yang» y los misteriosos «cinco elementos», a los que se alude constantemente en una gran cantidad de los artículos [3], no fueran más que invenciones mitológicas, entonces es harto probable que de las terapias basadas en tales conceptos sólo quepa esperar, en el mejor de los casos, un efecto placebo, y que lo que cura o mejora es la sugestión aparejada. De ser así, lo que cabría concluir no es que tales terapias sean inútiles, pero sí que lo único que procede retener de la acupuntura es la pantomima. En tal caso, la conclusión inevitable sería que para aprender la técnica no hace falta ir a China (ni a ninguna otra parte) a estudiar los complejos mapas humanos con que opera esta disciplina, ni mucho menos los confusos principios teóricos que la envuelven. Y tal conclusión no sería de importancia marginal: basta tener en cuenta que en la década pasada viajaron a ese país 6.000 alumnos procedentes de más de 100 naciones a estudiar estas técnicas y su entorno filosófico (Shang, 1996).

Una corriente de la misma familia afirma que las agujas pueden ser suplidas o complementadas por un proceso de combustión (la llamada «moxibustión») de una planta llamada moxa *(artemisa sienensis)* con el que se estimulan los puntos de acupuntura. Y para poner a tono todo esto con la fiebre tecnologista de hoy, no podía menos que aparecer la síntesis armónica entre la medicina medieval y el futurismo, entre la sutil sapiencia oriental y la supertecnología de Occidente: la ***laserpuntura,*** aplicación de rayos láser en los susodichos puntos.

Otra variante alternativa es la de aplicar presión con los dedos sobre los puntos de acupuntura *(digitopuntura o acupresión).* Un ejemplo de esta última alternativa, citado por Skrabanek y McCormick, se ofrece en un libro de dígitopuntura (Vora, 1984), cuyo prefacio fue escrito por Morarji Desai, primer ministro de la India

[3] Fundamentalmente por autores de origen chino en revistas como ***Acupunture and Electrotherapy Research y Journal of Traditional Chinese Medicine.*** Véanse, por ejemplo, Liao (1992) y Liu ***et al., (1992),*** que sitúan su origen en la filosofía medieval de aquel país.

entonces y fallecido en 1995; en él se sugiere el siguiente tratamiento para la sífilis: aplicar acupresión sobre el tendón de Aquiles en un tobillo mientras se da masaje con orina hervida «a la parte afectada»; además, se recomiendan otras opciones terapéuticas como la magnetoterapia y que el enfermo beba su propia orina.

En esta materia, el rosario de excentricidades y patrañas pseudocientíficas es asombroso: recientemente leí, por ejemplo, una nota aparecida en un boletín informativo para el extranjero de la República Popular China (Xianyang Health Products, 1993) el anuncio de la creación de una «gorra salutífera» debida al profesor Lai Huiwu. Según consigna textualmente la publicación, cuando esta gorra se aplica a los «debidos puntos acupunturales de los canales de la cabeza», produce los beneficios que reproduzco textualmente a continuación:

> ***Nutre la energía vital y el «qi», equilibra el «yin» y el «yan», refuerza los factores antipatógenos, fortalece el cerebro, desobtura los orificios especiales del cuerpo humano, cultiva la inteligencia, calma los nervios, agudiza el oído y la vista, ennegrece el cabello y consolida los dientes.***

Un acupunturista francés inventó una nueva variante conocida como ***acupuntura auricular,*** basada en la insólita convicción de que todos los órganos del cuerpo y sus funciones se encuentran «proyectados» en el lóbulo de la oreja. Tal proyección formaría un pequeño hombre en posición fetal y con la cabeza hacia abajo; de modo que para tratar, por ejemplo, un dolor de cabeza habría que actuar sobre el punto en que se hallaría la cabecita de ese hombrecito imaginario [4].

Pretensiones igualmente descabelladas, tanto para el proceso diagnóstico como para el terapéutico se consolidaron históricamente en varias culturas de Asia. Por ejemplo, Lad (1995) nos ilustra acerca de cómo «se puede conocer el estado funcional de todos los órganos internos mediante la mera observación de la lengua, la cual es un espejo de las vísceras y refleja sus condiciones patológicas», procedimiento inscrito dentro de una teoría profundamente mística de origen hindú (el ***ayurveda).*** Las variantes competidoras eligen otras partes del cuerpo (manos, plantas de los pies o iris del ojo) en las que igualmente se «reproduce» todo el organismo humano. Mi sentido común me lleva a preguntarme si los actuales divulgadores de estas panaceas estarán en su sano juicio, o integran un proyecto para burlarse de la ingenuidad humana.

A mi juicio urge intensificar los estudios para separar la paja del trigo (si es que lo hay) en esta materia.

Un estudio sobre opiniones y convicciones de médicos de familia cubanos realizado por Silva y González (1991) revela que la inmensa mayoría (el 88%) de estos profesionales confiere a la acupuntura un sólido basamento científico. Ésta es la

[4] Como el ojo de esta criatura resulta ser el punto donde usualmente se perfora la oreja para colocar aretes, ello condujo a que un renombrado acupunturista británico, G.T. Lewith, rápidamente hiciera notar que ésta sería la razón por la cual los piratas utilizaban aretes, lo cual daría explicación ala leyenda de sus poderes para observar barcos a distancia, capacidad de la que otros carecían.

situación en un momento en que los argumentos expuestos en favor de la existencia de los «puntos de acupuntura» son sumamente confusos, a la vez que, objetivamente, su existencia no ha sido inobjetablemente demostrada por medio de la anatomía, neurofisiología u otra rama de la biología. Ninguno de los encuestados consideró el efecto placebo como posible mecanismo a través del cual la acupuntura pudiera ser efectiva. Por el contrario, excepto un 3% que marcó la opción «no sé», estos médicos se inclinaron por la afirmación de que «los puntos donde se colocan las agujas tienen importancia fundamental».

Una dificultad para dirimir si hay o no un efecto adicional al placebo dimana del hecho de que los acupunturistas suelen ser muy reacios a los ensayos clínicos controlados. En varias ocasiones, cuando he propuesto llevarlos adelante, he recibido negativas, y he tenido la impresión de que no pocos cultores de esta disciplina se sienten amenazados, como si un experimento con enmascaramiento doble *(double blind)* fuera un enemigo, y como si la propuesta de usar un tratamiento placebo fuera un insulto. En una oportunidad se me dijo que la aplicación de un pseudotratamiento de acupuntura a un grupo de pacientes de sacrolumbalgia aleatoriamente elegido no sería ético, pues privaría a dichos pacientes de la posibilidad de mitigar sus dolores. En este tipo de situaciones uno recuerda a los monjes que se negaban a mirar por el telescopio que les ofrecía Galileo con el argumento de que no hacía falta hacer la observación, ya que la teoría de Tolomeo afirmaba algo diferente a lo que supuestamente dicha observación revelaría.

13.2. La estadística: un enemigo peligroso

Uno de los trabajos clásicos escritos por Ronald Fisher (véase el segundo capítulo de Fisher, 1951), padre de la moderna inferencia estadística, está destinado a explicar en tono divulgativo el papel de la teoría de probabilidades en la evaluación de hipótesis por vía experimental.

El texto plantea el problema de una dama que afirma tener la capacidad de dirimir, después de probar un té con leche, cuál de los dos ingredientes fue vertido en primer lugar dentro de la taza. Con ese *leitmotiv* discurre el trabajo, que detalla la esencia del enfoque probabilístico como recurso valorativo de la posible veracidad de una hipótesis, asentado sobre el diseño experimental.

Es fácil advertir el fuerte hálito pseudocientífico del tema elegido por Fisher para ilustrar el *modus operandi* de la inferencia estadística. De hecho, la mayoría de las afirmaciones de la pseudociencia podrían ser valoradas según pautas muy similares a las explicadas por Fisher si no fuera por dos importantes barreras: primero, que tales propuestas suelen formularse de manera mucho más borrosa que la que hace la hipotética dama de Fisher; segundo, que los paladines de las profecías, la telepatía o el zodíaco se niegan, tanto como pueden, a someterse a la experimentación formal según los cánones universalmente aceptados.

Por otra parte, en ocasiones se ha usado la estadística como escudo para la pseudociencia; aquel que no exhiba una conducta ética o metodológicamente rigurosa en su trabajo suele desembocar en la búsqueda a ultranza de aval estadístico para ver corroborados sus deseos. A partir de ahí, puede pasar cualquier cosa, ya que usará entonces la «técnica» a que se refería Kitiagorodski (1970) cuando escribía:

> ***El estilo habitual de trabajo de un fanático que desea demostrar su razón recurriendo a la estadística consiste en omitir los datos que son, a su modo de ver; desafortunados y, tener en cuenta, en cambio, los que condicen con aquella,***

Exactamente ésta es la conducta que siguen, por ejemplo, algunos laboratorios farmacéuticos; no actúan en ambientes esotéricos ni al margen de los entornos científicos; muy por el contrario, se mueven en el abierto mundo asistencial y académico, pero se caracterizan por una conducta sinuosamente ajena a los preceptos científicos. Ponen cuantiosos recursos en función de demostrar (y difundir) la eficiencia de los fármacos que producen, pero no hacen absolutamente nada para evaluarlos críticamente, mucho menos para que se comuniquen las eventuales insuficiencias que pudieran detectarse.

Puede darse el caso de que el propio investigador llegue a creer en la objetividad de lo que ha engendrado. En la Sección 5.6 se describió el caso del «craneólogo» norteamericano Samuel Morton y su prueba, en la primera mitad del siglo XIX, de la supremacía intelectual de los blancos sobre los negros. El examen detallado de sus manejos estadísticos reveló que no se trataba, aparentemente, de un fraude deliberado sino de una secuencia de errores estadístico-computacionales, quizás debidos parcialmente a la impericia de Morton pero, según Gould (1981), sobre todo al afán subconsciente de demostrar a toda costa sus tesis, de las cuales estaba ciegamente persuadido de antemano. Tras diseccionar el andamiaje numerológico de Morton, Gould concluye que «las teorías se construyen a partir de la interpretación de los números, y los intérpretes quedan atrapados a menudo en su propia retórica».

Se atribuye a Julio César haber dicho que los hombres creen gustosamente aquello que se acomoda a sus deseos. Esta tendencia, es importante recalcarlo, no es privativa ni de los estafadores intelectuales ni de los profesionales de la pseudociencia. El mejor intencionado de los investigadores puede actuar pseudocientíficamente como resultado del manejo incorrecto de los instrumentos de que dispone, la estadística entre ellos.

13.3. La teoría de los biorritmos contra el ji cuadrado

Para examinar más estrechamente algunos de los puntos de interés, consideremos otra interesante representación del pensamiento pseudocientífico: la llamada *teoría de biorritmos. Se* trata de un sistema que, desde el punto de vista de la formu-

lación, no está flagrantemente divorciado del lenguaje científico debido a que, a diferencia del discurso de los zahoríes o la parapsicología, expresiones pseudocientíficas que empiezan por formular sus teorías de manera borrosa, la teoría de los biorritmos es relativamente clara en su pretensión.

En tal sentido se aproxima más, por tanto, a una de las exigencias de Popper, no impugnada por nadie en lo que a ese punto concierne: que una teoría o hipótesis, para ser científica, ha de ser «falseable» (es decir, susceptible de ser rechazada como resultado de su contrastación con la práctica). En particular, ilustra también cómo opera el modo de pensar del estadístico en esta materia.

13.3.1. Teoría BBB: origen y formulación

El comportamiento rítmico de algunas variables biológicas -sobre todo de aquellas asociadas al ser humano- ha sido intensamente investigado a través de un esfuerzo que se remonta, como ocurre con casi todas las ramas del saber, a la antigüedad. El interés se ha dirigido históricamente tanto a la identificación cualitativa de sus rasgos como a su caracterización cuantitativo-paramétrica.

La actividad metabólica y hormonal, la fatiga intelectual, la depresión, la temperatura corporal y la tensión arterial son ejemplos de procesos fisiológicos cuyo comportamiento periódico ha sido exitosamente estudiado. Múltiples referencias sobre esta rama de la ciencia (la llamada *cronobiología*), pueden encontrarse, por ejemplo, en el libro ***Ritmos biológicos y comportamiento humano*** de Colquhom (1971) o, entre muchos otros artículos, en uno especialmente exhaustivo debido a McConnell(l978).

Un sistema relacionado con los ritmos biológicos, inspirado en ideas que se remontan a finales del siglo XIX, resurgió con gran vitalidad hacia 1970. Se trata de la llamada ***teoría de los biorritmos,*** que relaciona la fecha de nacimiento de cada individuo con los acontecimientos físicos y psicobiológicos que se producirán a lo largo de su vida.

Para distinguir esta teoría de la cronobiología, Englund y Naitoh (1980) la han bautizado como ***teoría BBB*** (del inglés ***birthdate based biorhythm),*** expresión que, por comodidad, se utilizará en lo que sigue. Por diversas razones que se analizan más adelante, esta teoría alcanzó - y aún tiene- considerable popularidad, y a mediados de la década de los 70 concitó cierta atención de los círculos científicos.

Su nacimiento data del siglo pasado, cuando el cirujano y otorrinolaringólogo alemán Wilhelm Fliess inició lo que Gardner (1966) llamara «uno de los más extraordinarios y divertidos episodios en la historia de la pseudociencia numerológica». Notable él mismo, en parte gracias a su íntima y tormentosa amistad con Sigmund Freud -que hubo de extenderse a lo largo de los últimos diez años del siglo-, Fliess logró dar cierta notoriedad a una teoría que atribuía místicas propiedades a los números 23 y 28, que él creía asociados al hombre y la mujer, respectivamente.

Con el paso del tiempo, el médico berlinés llegó a convencerse de que todos los fenómenos naturales -desde el nivel celular al planetario- se regían por leyes vinculadas de una u otra forma con esos dos números. Así, sostenía que a los 51 años (un número igualmente connotado según Fliess, por ser el resultado de sumar 23 y 28) todo ser humano corría enormes riesgos, y llegó a vaticinar que el propio Freud moriría a esa edad [5].

Obsesionado por estos números y como una prueba más de la ubicuidad de que estaban dotados, publicó una tabla en que cada uno de los primeros 28 números naturales aparecía como una combinación lineal de 23 y 28. Es decir, consiguió hacer un listado como el siguiente:

$$
\begin{aligned}
1 &= 11 \cdot 23 - 9 \cdot 28 \\
2 &= 22 \cdot 23 - 18 \cdot 28 \\
3 &= 5 \cdot 23 - 4 \cdot 28 \\
&\cdots\cdots \\
&\cdots\cdots \\
27 &= 45 \cdot 23 - 36 \cdot 28 \\
28 &= 28 \cdot 23 - 22 \cdot 28
\end{aligned}
$$

Ensimismado en sus propias lucubraciones y limitado por la vasta ignorancia matemática que poseía, Fliess desconocía que ***todo número entero positivo puede escribirse como combinación lineal de cualquier pareja de números con tal de que éstos, como ocurre con 23 y 28, sean primos relativos.***

Su obra clave, un aparatoso volumen de casi 600 páginas, se tituló ***Los ritmos de la vida: bases para una biología exacta.*** La «exactitud» de Fliess se vertebraba alrededor de sistemáticas alusiones a múltiplos de 23 y de 28, y a números tales como el duplo de 28 al cubo o el producto de 23 y 28.

En ese contexto, Fliess planteaba que todo ser humano está sometido a la influencia de dos corrientes cíclicas que comienzan con el nacimiento y mantienen su efecto a lo largo de toda la vida. Uno de tales ciclos se repite cada 23 días y dirige los «aspectos masculinos» del individuo (o sea, según la concepción de Fliess, rasgos tales como la fuerza o la resistencia física); el otro ciclo ejercería su periódico influjo en lapsos de 28 días y concierne a «rasgos femeninos», tales como la sensibilidad y la intuición. Cada ciclo consta de una fase positiva (primera mitad de los días) y de otra negativa (segunda mitad). La teoría afirma que durante las fases positivas, los dominios gobernados por el ciclo en cuestión tendrían un comportamiento favorable al individuo y que durante los días que dure la segunda mitad tales aspectos estarían deprimidos.

[5] El fallecimiento del psicólogo vienés se produjo a la edad de 73 años.

A esa altura ya Freud había comprendido que todo aquello no era otra cosa que un amasijo de fantasías numerológicas. Ante las primeras insinuaciones de Freud en ese sentido, Fliess -un hombre patológicamente sensible a la menor crítica- rompió la amistad que habían cultivado.

Sin sospechar el auge que sobrevendría pocos años después, Gardner (1966) escribía:

> ***Increíble pero cierto: el sistema de Fliess aún tiene una pequeña pero devota banda de discípulos en Alemania y Suiza. Hay médicos en varios hospitales suizos que continúan usando los ciclos de Fliess a fin de determinar los días favorables para sus operaciones.***

Winstead, Schwartz y Bertrand (1981) dan cuenta de que, alrededor de 1920, la teoría habría de completarse con la incorporación de un nuevo ciclo biorrítmico: Alfred Teltscher afirma haber detectado que la capacidad intelectual (comprensión, creatividad, concentración, etc.) de sus alumnos variaba cíclicamente según periodos de 33 días.

A partir de estos elementos se conforma la teoría BBB de nuestros días, cuya formulación puede sintetizarse del siguiente modo:

1. Existen tres ciclos que comienzan para cada individuo el día de su nacimiento y mantienen su vigencia de manera vitalicia. Se trata del «ciclo físico», el «ciclo emocional» y el «ciclo intelectual» que duran 23, 28 y 33 respectivamente para todos los seres humanos, con independencia del sexo, la edad o cualquier otra circunstancia social o fisiológica del sujeto.
2. Cada ciclo se divide en dos fases exactamente iguales: durante la primera, las potencialidades del individuo están desarrolladas, mientras que a lo largo de la segunda, se hallan atrofiadas.
3. Aquellos días en que se transita de una fase positiva a una negativa o viceversa, son calificados como ***críticos***. Se plantea que durante los *días críticos* el individuo está en situación especialmente vulnerable: accidentes, bajo rendimiento y percances de índole diversa se presentan con probabilidad mucho más alta que en los restantes. Los riesgos aumentan durante el día en que dos de los ritmos cambian de fase, y son particularmente agudos aquellos días en que tal cambio ocurre para los tres ciclos simultáneamente.

La Figura 13.1 refleja gráficamente lo que se ha descrito. Nótese que el eje de ordenadas carece de unidades. No es una omisión; el lector no debe perder su tiempo buscando esas unidades en otra fuente. No las hallará, pues lo que se «mide» para cada día en cada ciclo es una noción totalmente vaga.

Un detalle profundamente absurdo que, hasta donde conozco, nadie ha señalado y que despierta vivamente mi desconcierto es el siguiente: si la primera mitad de

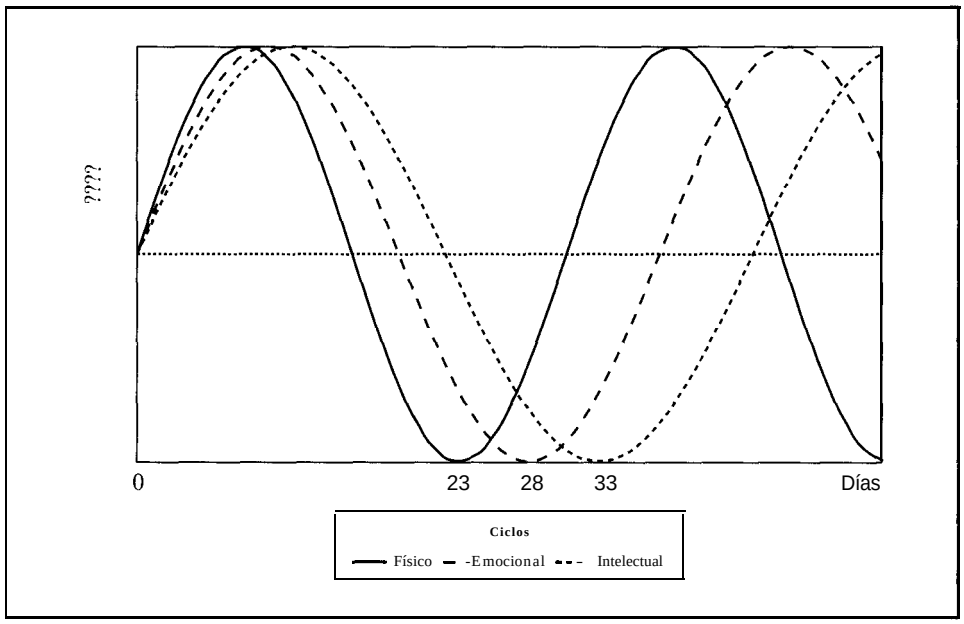


Figura 13.1. ***Curvas correspondientes a los ciclos biorritmos.***

cada ciclo es favorable y la segunda desfavorable, ¿no debería ser «neutro» el punto que separa a dichas fases? Sin embargo, la teoría dice arbitrariamente que tales días, los llamados críticos, son los peores.

13.3.2. La teoría de los biorritmos bajo el enfoque estadístico

Por la sencillez de su planteamiento, por la exaltación con que la han favorecido la prensa y otros dispositivos publicitarios y por la vigencia de ciertos mecanismos motivacionales, esta teoría llegó a conseguir no pocos adictos en distintos momentos del presente siglo.

A su favor se han señalado hechos y experiencias diversas, entre las que se menciona que George Gershwin, Pío XII, Henry Ford, Richard Strauss, Winston Churchill y Marilyn Monroe murieron en días críticos, así como otros acaecimientos que, aunque por lo general se inscriben en la esfera anecdótica, tienen un morboso atractivo. En esa línea se ha citado reiteradamente el ejemplo del boxeador Benny Paret, que recibió su golpiza fatal un día en que sus tres ciclos atravesaban el momento crítico.

Sobre «pruebas» de este tipo, Lyon, Dyer y Gary (1978) advierten que la existencia de hechos trágicos ocurridos en días críticos no constituye indicio alguno, ya que siempre hay una posibilidad de que algo malo ocurra a alguien alguna vez; no es difí-

cil, por tanto, hallar que uno de tales hechos haya ocurrido a una persona a lo largo de uno de sus días señalados como críticos por el engendro BBB. Dicho de otro modo, el análisis está viciado en la medida que no se computen accidentes o percances acaecidos en días *no* críticos, o en la fase positiva de los ciclos, y se proceda a compararlos con *tusas* de accidentes en días críticos o en la fase negativa de los ciclos. Éste es el modo riguroso inherente al correcto pensamiento epidemiológico; aquél, el que corresponde a los explotadores de la sugestión y el anumerismo. Cabe recordar a los crédulos que hasta un reloj dibujado en un papel da la hora correcta dos veces por día. Al analizar la derrota de Jimmy Connors en la final del Campeonato Mundial de Tenis de 1975 a manos de Manuel Orantes, se ha señalado que éste estaba ese día en los «momentos más altos de sus ciclos físico e intelectual», en tanto que el primero estaba en «la fase baja del ciclo intelectual y aproximándose a un día crítico del ciclo físico». Desde luego, si el vencedor hubiera sido Connors, se hubiera argüido que estaba en sus fases positivas para los ciclos físico y emocional, mientras que Orantes estaba en crisis en materia emocional. Como señala Bunge (1978):

> *El seudocientífico, igual que el pescador; exagera sus presas y oculta o disculpa todos sus fracasos. La seudociencia tiende a interpretar todos los datos de modo que sus tesis queden confirmadas ocurra lo que ocurra.*

Otros relatos muy similares pueden hallarse (véase Thommen, 1973) en el libro editado varias veces con el título *«¿Es éste su día?»* que escribió un tal G.S. Thommen, presidente de una compañía productora de «biocalculadoras», «tablas biorrítmicas» y otros objetos similares que permiten la rápida determinación de los días favorables y desfavorables, y contribuyeron al también rápido enriquecimiento del Sr. Thommen.

Más allá de ejemplos aislados y de naturaleza personal, también se registran referencias de otra índole. Un caso típico es el de la compañía de ferrocarriles japonesa Ohmi, a la cual se atribuye una reducción significativa de los accidentes lograda mediante un programa preventivo basado en los días críticos de sus conductores. La referencia dada por Willis (1972) dice que, mientras el 41% de los accidentes se verificaba en los días críticos antes de la puesta en práctica del programa, después de su implantación, la compañía logró 2 millones de km sin accidentes. Lamentablemente, Willis no advirtió que resulta imposible comparar kilómetros con porcentajes.

Pero haciendo honor a la afirmación del famoso matemático inglés Tippet (1968) de que «los estadísticos suelen aparecer ante sus colegas científicos como aniquiladores profesionales de conclusiones interesantes», Khalil y Kurucz (1977) demuestran que, aun creyendo en la veracidad de los datos iniciales, éstos no producen evidencias inequívocas, ya que el porcentaje de accidentes que pueden ocurrir en días críticos de al menos uno de los ciclos sería 27% en caso de que solamente el azar influyese en los hechos; luego, mediante una prueba estadística de

hipótesis, corroboran que la diferencia entre uno y otro porcentajes (41% y 27%) dista de ser significativa.

Generalmente, casi todas las referencias de que tenemos conocimiento en las que se respalda la teoría BBB, padecen de una formulación oscura o vaga, y carecen del tratamiento estadístico adecuado.

Tal es el caso, por ejemplo, del trabajo de Messer (1978) en que se discute el efecto de los biorritmos sobre los trastornos postoperatorios y el de Romero (1983) sobre traumatismos militares. Ocurre incluso que, en los pocos casos favorables donde se ofrecen los datos explícitamente, de manera que pueda hacerse un análisis correcto, éste conduce a la comprobación de que, en realidad, la teoría BBB no resulta mejor que el azar para predecir los acontecimientos.

Con respecto al auge que sobre estas bases estaba alcanzando la teoría BBB, en 1978 la revista *Time* (Anónimo, 1978) escribía que «los científicos no saben si burlarse o encolerizarse, y la inmensa mayoría ha preferido no dignificar la teoría mediante su investigación formal». Sin embargo, lo cierto es que son muchos los trabajos que, en ocasión de la última oleada mística al respecto, abordaron científicamente el tema.

Las áreas de estudio son diversas, pero el método de análisis estadístico es en casi todos los casos básicamente el mismo: se calculan las frecuencias esperadas según «tipo de día» para una muestra de sucesos, bajo el supuesto de que el momento en que ocurren depende exclusivamente del azar, y éstas se comparan con las frecuencias observadas mediante una prueba de bondad de ajuste no paramétrica. Por ejemplo [6], si se toma una muestra de accidentes, éstos pueden clasificarse según hayan ocurrido en un:

a) día crítico simple,
b) día crítico doble,
c) día crítico triple,
d) día no crítico [7].

A través del cálculo de la probabilidad de que un accidente caiga en cada una de estas clases, supuesto que la teoría BBB no rige, se pueden computar las cuatro frecuencias esperadas, y compararlas mediante una prueba estadística con la distribución observada: si las diferencias no son significativas, no habrá motivos para pensar que los biorritmos influyen en el acaecimiento de tales desgracias.

Al examinar la edad exacta (medida en días) que tenía en el momento de su infarto de miocardio cada uno de 1.795 pacientes consecutivamente registrados con ese diagnóstico en los servicios de urgencia de 6 hospitales de Ciudad de la Habana, entre 1986 y 1990, pude conformar la distribución muestra1 según tipo de días (Tabla 13.1).

[6] Hay infinidad de maneras análogas de conformar las categorías.
[7] En el que ninguno de los ciclos cambia de fase.

Tabla 13.1. ***Distribución teórica y observada de 1.795 infartos de miocardio consecutivos registrados en 6 hospitales de Ciudad de La Habana entre 1986 y 1990 según se tratara de días «críticos» o «no críticos» según la teoría de los biorritmos***

Tipo de día	Probabilidad*	Frecuencia esperada*	Frecuencia observada
Crítico simple	0,1885	338,4	341
Crítico doble	0,0147	26,4	20
Crítico triple	0,0004	0,7	1
No crítico	0,7964	1.429,5	1.433
Total	l,0000	1.795,0	1.795

* Calculada sobre el supuesto de que la teoría BBB es falsa.

La prueba de x^2 que corresponde es innecesaria, ya que la concordancia entre lo que debería ocurrir si la teoría BBB fuera falsa y lo que realmente ocurrió es notable, como revelan las dos últimas columnas de la Tabla 13.1. Lo que actuó no fue ningún sistema de biorritmos sino, simplemente, la ley de los grandes números descubierta por Bernoulli en el siglo XVII.

Los resultados de los estudios serios [8] son virtualmente unánimes y la conclusión fundamental es la misma: los ciclos biorrítmicos no existen y la teoría BBB es simplemente falsa.

La ausencia de asociación entre los biorritmos y los accidentes se ha demostrado en decenas de artículos entre los que se pueden mencionar los de Dolan (1976), Hirsh (1976), Latinan (1977), Wolcott ***et al.*** (1977), Carvey y Nibler (1977), Khalil y Kurucz (1977), Persinger, Cooke y Janes (1978) y Shaffer, Schmidt y Zlotowitz (1978).

Asimismo, pueden citarse multitud de estudios, tales como los de Halverson (1976), King (1976), Fix (1976), Simon (1977), Siegel (1978), Quigley (1981), Wright (1981) y Floody (1981), relacionados con otras esferas (desempeño deportivo, infartos, resultados académicos, incidentes policiales, complicaciones quirúrgicas, etc.). Todos reflejan igualmente que no hay prueba alguna de que los tales biorritmos existan.

De especial interés metodológico resulta el de Kunz (1984) por ser uno de los pocos de naturaleza experimental. A dos grupos de alumnos se les pidió que fueran

[8] Entiendo por *serios* los que aparecen en publicaciones científicas arbitradas, no los que se usan para llenar páginas de algunas revistas policromas que inundan los kioscos e idiotizan al prójimo.

clasificando sus próximos 5 días en una escala de 1 a 10 (de mejor a peor) para cada una de las 3 esferas involucradas en la teoría BBB. Al primer grupo (187 alumnos) no se les comunicó nada adicional; a los integrantes del segundo (153 estudiantes) se les explicó la teoría BBB y se les proveyó de unas tarjetas en que aparecían los valores teóricos que corresponderían a cada día para cada sujeto, obtenidos a partir de uno de los tantos programas informáticos que dan un valor en dependencia de su fecha de nacimiento para cada ciclo. Se les solicitó, no obstante, que la calificación entre 1 y 10 que fueran dando a cada día se basara en lo que realmente ocurriese y no en lo que *debería* haber pasado según los valores que se desprenden de la teoría y que ellos conocían.

Luego se computaron los coeficientes de correlación entre los valores teóricos y las autocalificaciones para ambos grupos. Los resultados, que se reproducen en la Tabla 13.2, son sumamente elocuentes y dan cuenta del intenso efecto placebo que se presenta en situaciones como éstas.

Tabla 13.2. ***Coeficientes de correlación entre las prediciones de la Teoría BBB y las autocalificaciones dadas por 2 grupos de estudiantes***

	Ciclos		
	Físico	Emocional	Intelectual
Grupo control	0,019	0,173	-0,005
Grupo experimental	0,372	0,291	0,487

Debe señalarse, finalmente, que existen algunos estudios aparentemente rigurosos que informan del hallazgo de una frecuencia de accidentes en días críticos significativamente mayor que la que cabe esperar como consecuencia exclusiva del azar. He podido «rastrear» dos ejemplos: los trabajos de Pittner y Owens (1975) y Schwartz (1976).

Tratando de hallar explicación para estas curiosas «disidencias», Chaffin y Skadburg (1979), en una muestra de aguda perspicacia, conjeturan primero y demuestran luego la existencia de un sesgo de medición cuando para la identificación de los días críticos se sigue el método popular de observar las curvas impresas por la computadora, en lugar de guiarse por los resultados numéricos.

Finalmente, hay que subrayar que aun cuando la imputación de Chafflin y Skadburg no fuese para todos estos casos atinada, no cabe sorprenderse de la existencia de tales resultados. En efecto, como señalé en otro lugar (Silva 1984), la interpretación que tiene el que un resultado sea significativo, digamos, al nivel $p = 0{,}05$, es, precisamente, que se admite el riesgo de que una de cada 20 veces se declare como significativamente diferente de cero una correlación que realmente no existe. Por lo tanto, lo insólito sería que, en el contexto de una profusa afluencia de estudios sobre el tema,

no hubiese algunos que apuntasen en la dirección equivocada. Quiere esto decir que la comunicación de una indagación que, habiendo sido desarrollada de manera científicamente correcta, no rechace la teoría BBB, es en principio legítima; sacar de esa experiencia aislada una conclusión categórica en su aval, constituiría sin embargo un error metodológico grave. Pero una defensa de la teoría basada en anécdotas y excepciones, no es otra cosa que la conducta típica de los incautos o los farsantes.

13.3.3. **Una valoración final**

Por las lecciones generales que puede dejar para los investigadores, concluyamos el tema con algunas consideraciones teóricas.

En primer lugar, la exacta regularidad de los ciclos y su idéntica influencia en todos los seres humanos, hacen que la teoría esté matizada por un determinismo ingenuo y rígido, que adquiere por esa vía un carácter místico. El ser humano, como todos los sistemas naturales, es abierto, no autónomo y por consiguiente está sujeto a las alteraciones inducidas por su entorno físico y social, a la vez que exhibe un alto grado de adaptabilidad. Consecuentemente, es ridículo admitir que el conocimiento de sus condiciones iniciales como sistema permita la predicción unívoca de sus estados subsiguientes.

El carácter absurdo de la teoría BBB se torna más claro cuando se repara en que «las condiciones iniciales» se reducen a un solo dato: la fecha de nacimiento del sujeto. En realidad todo el planteamiento es de una notable puerilidad, pues la mera idea de que los rendimientos físico, emocional e intelectual de un individuo puedan ser independientes entre sí contradice el sentido común.

Una limitación más debe ser señalada, que no por ser la última, es la menos importante: la ausencia de un marco teórico dentro del cual quede inscrita la hipótesis fundamental de este planteamiento [9]. Después de la presunta observación empírica de los tres ciclos biológicos por parte de Fliess y Teltscher, no parece que se haya siquiera intentado encontrar un respaldo teórico que los explique. Sin embargo, los éxitos de la biología científica han estado y están determinados por el estudio de las leyes que rigen en la naturaleza animada. En el caso que nos ocupa, al no existir una formulación teórica que sirva de puente entre la percepción viva y la confirmación práctica, el tránsito dialéctico que conduce al conocimiento ha quedado truncado de antemano.

Debe notarse que el cúmulo de resultados estadísticos en contra de la teoría que establece la existencia de tres ritmos dependientes del día de nacimiento de cada

[9] Esta carencia es típica del pensamiento pseudocientífico. Algunas corrientes terapéuticas como la homeopatía proclaman explícitamente que no hay ninguna necesidad de explicar los mecanismos según los cuales «funcionan» (véase, por ejemplo, Carlston, 1995).

individuo, a pesar de ser ciertamente notable, cobra su verdadera dimensión sólo cuando se aprecia en el contexto de un debate teórico mucho más amplio, a la luz del conocimiento ya consolidado y gracias a una interpretación integral que no desdeña el sentido común como herramienta. Eso mismo ocurre en cualquier contexto: los resultados estadísticos en general sólo contribuyen al avance científico cuando se interpretan a la luz de un marco teórico integrador.

Bibliografía

Ackerknecht EH (1974). ***Zur Geschichte der Akupunktur*** Anaesthetist 23: 37-38.

Anónimo (1978). ***Those biorhythms and blues.*** Time ll 1: 50-5 1.

Ballegaard S, Meyer CN, Trojaborg W (1991). ***Acupuncture in angina pectoris: does acupuncture have a specific effect?*** Journal of Interna1 Medicine 229: 357-362.

Bueno G, Hidalgo A, Iglesias C (1987). ***Symploke.*** Júcar, Gijón.

Bunge M (1972). ***La investigación científica.*** Ciencias Sociales, La Habana.

Carlston M (1995). ***The mechanism of homeopathy? All that matters is that it works.*** Alternative Therapies in Health and Medicine 1: 95-96.

Carvey DW, Nibler RG (1977). ***Biorhythmic cycles and the incidence of industrial accidents.*** Personnel Psychology 30: 447-454.

Collazo C (1995). ***Perfil de Murray Gell-Mann.*** Muy Interesante 169: 117-118.

Colquhom WP (1971). ***Biological rhythms and human performance.*** Academic Press, New York.

Chaffin R, Skadburg J (1979). ***Effect of scoring set on biorhythm data.*** Journal of Applied Psychology 64: 213-217.

Dolan MH (1976). ***Biorhythms and accidents school children.*** Abstracts of Hospital Management Studies 13(2).

Eisenberg DM, Kessler RC, Foster C, Norlock FE, Calkins DR, Delbanco TL (1993). ***Unconventional medicine in the United States. Prevalence, costs, and pattems of use. New*** England Journal of Medicine 328: 246-252.

Ekblom A, Hansson P, Thomsson M, Thomas M (1991). ***Increasedpostoperative pain and consumption of analgesics following acupuncture.*** Pain 44: 241-247.

Englund CE, Naitoh P (1980) ***An attempted validation study of the birthdate based biorhythm (BBB) hypothesis.*** Aviation Space and Environmental Medicine 15: 583-590.

Fisher RA (1951). ***Statistical methods for research workers.*** ll.ª ed, Oliver and Boyd, Edinburgh.

Fix AJ (1976). ***Biorhythms and sports performance.*** The Zelectic 1: 53-57.

Floody Dr (1981). ***Further systematic research with biorhythms.*** Journal of Applied Psychology 66: 520-521.

Gardner M (1966). ***Freud's friend Wilhelm Fliess and his theory of male and female life cycles.*** Scientific American 215: 108-112.

Gould SJ (1981). ***The mismeasure of man.*** Norton, New York.

Halverson SG (1976). ***The effect of biorhythms on thepatient with a myocardial infraction.*** Abstracts of Hospital Management Studies 13(2).

Hernández A (1993). ***Científicos contra videntes.*** Conocer 131: 4-9.

Hirsh T (1976). ***Biorhythm. Or is it a critical day?*** National Safety News: 41-44, Feb.

Khalil TM, Kurucz CN (1977). ***The influence of «biorhythm» on accident occurrence and performance.*** Ergonomics 20: 389-398.

King KB (1976). ***A comparison of biorhythm cycles and surgical complications.*** Abstracts of Hospital Management Studies 13 (2).

Kitiagorodski A (1970). ***Lo inverosímil no es un hecho.*** Mir, Moscú.

Klijnen J, ter Riet G, Knipschild P (1991). ***Acupuncture and asthma: a review of controlled trials.*** Thorax 46(ll): 799-802.

Klijnen J, ter Riet G, Knipschild P (1991). ***Clinical trials of homeopathy.*** British Medical Journal 302: 316-323.

Kunz PR (1984). ***Biorhythms: an empirical examination.*** Omega 14: 291-298.

Kurucz CN, Kahhlil TM (1977). ***Probability models for analyzing the effects of biorhythms on accident occurrence.*** Journal of Safety Research 9: 150-158.

Lad V (1995) An introduction to ayurveda. ***Alternative Therapies in Health and Medicine.*** 1: ***57-63.***

Latinan N (1977). ***Human sensitivity, intellectual, and physical cycles and motor vehicle accidents.*** Accident Analysis and Prevention 9: 109-112.

Lewis IH, Pryn SJ, Reynolds PI, Pandit UA, Wilton NC (1991). ***Effect of P6 acupressure on postoperative vomiting in children undergoing outpatient strabismus correction.*** British Journal of Anesthesiology 67: 73-7 8.

Liu Y, Tougas G, Chiverton SG, Hunt RH (1992). ***The effect of acupuncture on gastrointestinal function and disorders.*** American Journal of Gastroenterology 87: 1372-1381.

Liao SJ (1992). ***The origin of the five elements in the traditional theorem of acupuncture: a preliminary brief historic enquiry.*** Acupuncture and Electrotherapy Research 17: 7-14

Lyon WS, Dyer FF, Gary DC (1978). ***Biorhythm: imitation of science.*** Chemistry 51: 5-7.

McConnell JV (1978). ***Biorhythms: a report and analysis.*** Journal of Biological Psychology 20: 13-24.

Messer MS (1978). ***Correlation of biorhythm cyclephases with incidence of postoperative infection and requirement forpostoperative analgesia.*** Militar-y Medicine 197: 308-310.

Paulos JA (1990). ***El hombre anumérico.*** Alfaguara, Madrid.

Persinger MA, Cooke WT, Janes JT (1978). ***No evidence for a relationship between biorhythms and industrial accidents.*** Perceptual and Motor Skills 46: 423-426.

Pittner ED, Owens D (1975). ***Chance or destiny? A review and test of the biorhythm theory.*** Professional Safety 20: 42-46.

Quigley BM (1981). ***«Biorhythms» and Australian track and field records*** Journal of Sports Medicine 21: 81-89.

Rogvi B, Perrild H, Christensen T, Detmar SE, Siersbaek K, Hansen JE (1991). ***Acupuncture in the treatment of Graves' ophthalmopathy. A blinded randomized study.*** Acta Endocrinological of Copenhagen 124: 143-145.

Romero RE (1983). ***Biorritmo y trauma en miembros de las Fuerzas Armadas.*** Revista de Medicina Militar 2: 79-86.

Schwartz GR (1976). ***A look at the matter of susceptibility to work errors as related to biorhythm.*** Professional Safety 21: 34-39.

Shaffer JW, Schmidt CW, Zlotowitz HL (1978). ***Biorhythms and highway crashes: are they related?*** Archives of General Psychiatry 35: 41-46.

Shamos MH (1995). ***The myth of scientific literacy.*** Rutgers University Press, New Jersey.

Shang X (1996). ***Traditional medicine and WHO.*** World Health 49: 4-5.

Siegel D (1978). ***Biorhythms: are they useful in predicting athletic performance?*** Joper 35.

Silva LC (1984). ***Teoría de los biorritmos: ¿resurgimiento de un mito?*** Revista Cubana de Administración de Salud 10: 333-340.

Silva LC, González M (1991). ***Conocimientos y criterios del médico de familia sobre esencialidad de medicamentos y medicina alternativa. La Habana, 1991.*** Documento inédito. Facultad de Salud Pública de La Habana.

Simon LR (1977). ***The effect of biorhythms on surgical complications.*** Abstracts of Hospital Management Studies 13(3).

Skrabanek P (1984). ***Acupuncture and the age of unreason.*** Lancet i: 1169-1171.

Skrabanek P (1985). ***Acupuncture: past, present and future.*** En: ***Examining Holistic Medicine*** Stalker D, Glymour G (editores) Prometheus Books, Buffalo, New York.

Skrabanek P (1986). ***Acupuncture - needless needles.*** Editorial Irish Medical Journal 79:334-335.

Skrabanek P, McCormick J (1989). ***Follies and fallacies in Medicine.*** The Tarragon Press, Glasgow.

Tavola T, Gala C, Conte G, Invernizzi G (1992). ***Traditional Chinese acupuncture in tension-type headache: a controlled study.*** Pain 48: 325-329

Thomas M, Arner S, Lundeberg T (1992). ***Is acupuncture an altemative in idiopathic pain disorder?*** Acta of Anesthesiology of Scandinavia 36: 637-642

Thommen GS (1973). ***Is this your day?*** Drown Publishers Inc, New York.

Tippet LHC (1968). ***Statistics. 3ª*** Ed Oxford University Press, London.

Vora D (1984). ***Health in your hands: acupressure therapy. 3.ª*** ed, Gala Publishers, Bombay.

Willis HR (1972). ***Biorhythms and its relationship to human error*** Memorias de la 16.ª Reunión Anual de la Human Factor Society: 274-282.

14

Sobre ordenadores y ordenados

Si además de dedicarte a la enseñanza, consumes algún tiempo reflexionando sobre el sentido y los impactos de la ciencia y de la tecnología, te pueden tachar de filósofo, lo que resultará honroso si no fuera porque en un contexto de ingeniería y técnica tal calificación sugiere la sospecha de que te has pasado a una escala suprema en el orden de la divagación inútil.

FERNANDO SÁEZ VACA

14.1. La herramienta múltiple cambia el mundo

Para concluir esta reflexión sobre el papel de los métodos cuantitativos en la investigación biomédica y epidemiológica contemporánea, he seleccionado un tema ineludible: las computadoras y su proliferación universal.

Mis ideas al respecto son en buena medida tentativas y, verosímilmente, pudieran llegar a ser obsoletas en breve plazo. No puede ser de otro modo, habida cuenta de la vertiginosa movilidad tecnológica que no dejamos de presenciar en esta materia. Los conceptos, el lenguaje, las potencialidades y, sobre todo, los recursos de *soft* y *hardware* se renuevan caleidoscópicamente; los usuarios disfrutamos y padecemos un proceso que nos desborda. Grampone (1992) llamaba incisivamente la atención sobre ello:

> ***Las computadoras son máquinas que en tres años pasan de jóvenes pujantes a adultas decadentes, inservibles, que no valen la energía eléctrica que gastan ni el lugar que ocupan (...) Algunos, inocentemente, compran computadoras en 36 cuotas (o más). En realidad las están arrendando; cuando las terminan de pagar, las deben cambiar por una nueva.***

Por añadidura, hecha la constatación cíclica de que hay que renovarse, quien explota regularmente este recurso se verá ante la dificultad (¿o imposibilidad?) de optimizar su decisión en virtud de lo que Sáez (1994) enuncia en forma de ***ley compumalthusiana:*** «la oferta informática crece en proporción geométrica, en tanto que los clientes lo hacen en proporción aritmética». Un solo dato, mencionado en Dormido (1995), permite hacerse una idea sobre la intensidad de ese proceso: solo en Alemania se desechan por año alrededor de 800.000 toneladas de equipamiento computacional.

Durante mucho tiempo la computadora se relacionó en la mente popular con el acto de calcular, de ahí el nombre. En 1621 el clérigo inglés Willian Oughtred construyó un sistema de dos líneas de números ubicados en respectivas escalas que se deslizaban una sobre otra; había nacido así la primera regla de cálculo, distintivo emblemático del ingeniero durante siglos. Empezó entonces un largo peregrinar por ingeniosos adminículos mecánicos hasta la irrupción del mítico ENIAC *(Electronic Numerical Integrator and Calculator),* voluminoso artefacto compuesto por miles de válvulas electrónicas antecedentes inmediatas de los portentosos transistores. Esta invención, que le valiera a Bardeen, Brattain y Chockley el Nobel en 1956, fue el paso clave para la aparición de la ubicua computadora, nombre inercia1 que ha prevalecido aunque el fenómeno tecnológico que nos ocupa sea muchísimo más que un recurso para hacer cómputos; se trata de un instrumento poseedor de una versatilidad tan prodigiosa que aún aguarda por una denominación suficientemente abarcadora.

Resulta asombroso, e inédito en la historia de la humanidad, que contemos con una única herramienta para encarar tareas tan radicalmente diferentes entre sí. Si hace sólo 30 años alguien hubiera vaticinado que cualquier profesional dispondría en su casa de un instrumento capaz de solucionar problemas tan disímiles como escribir una carta, simular un proceso industrial, enfrentar con éxito a un gran maestro de ajedrez, computar una integral, consultar una receta de cocina, enviar un mensaje intercontinental, diseñar un puente y desencriptar un mensaje cifrado, se le habría conceptuado como un sujeto en estado delirante.

Un recurso con tal capacidad, que además está al alcance de cualquiera, no puede menos que dar lugar a un cambio cualitativo esencial en el mundo que nos rodea y en nuestra manera de interactuar socialmente. Pero vale la pena tener presente que, en cualquier caso, siempre corresponderá al ser humano determinar los perfiles de ese mundo nuevo. En algunos ambientes hispanoparlantes se ha dado en llamar ***ordenador*** (galicismo proveniente de la voz ***ordinator)*** a este instrumento. Tal vocativo resulta no menos chocante que su sinónimo de «computadora»; más concretamente, es conceptualmente contraproducente, ya que puede hacer que algunos olviden que es el hombre quien da las órdenes, en tanto que las computadoras se circunscriben a cumplirlas. El término, sin embargo, ha hecho fortuna; de manera que es comprensible que «computadora» y «ordenador» se usen indistintamente.

Como el manejo de grandes masas de datos estadísticos, lejos de escapar a la esfera de acción del ordenador, es uno de sus más tradicionales aplicaciones, su universalización y en especial el advenimiento de las computadoras personales han

venido a redibujar medularmente el escenario estadístico y sus zonas conexas. Ello ha impuesto una inaplazable readaptación a quienes utilizamos la estadística en alguna de sus modalidades, no sólo, como es obvio, en el dominio estrictamente computacional, sino incluso en el terreno teórico.

Por ejemplo, un área de la estadística que está siendo convulsionada por las computadoras es la de las encuestas, en especial desde la aparición de los ordenadores portátiles. Se trata de registrar directamente los datos y respuestas en soportes magnéticos; es decir, sin intermediarios como el clásico cuestionario impreso. La recolección de datos directamente asistida a través de este medio ha sido bautizada como *CADAC (Computer Assisted Data Collection).* La posibilidad de hacer exámenes de consistencia de las respuestas simultáneamente con su registro (y por ende, de realizar *in situ* las enmiendas que procedan), además del ahorro de papel, personal y tiempo que supone, ha llevado a la entusiasmada adopción del método por la mayoría de las agencias especializadas y por cada vez más equipos investigadores. Los efectos de tal procedimiento están siendo estudiados desde finales de los años 80 (Saris, 1989; Weeks, 1992; Nicholls, Baker y Martin, 1996). De hecho se han manejado dos variantes fundamentales: el uso de las computadoras para el desarrollo de entrevistas (Baker, 1992; Couper y Burton, 1994) y su aplicación en los cuestionarios autoadministrados (O'Reill *et al.*, 1994). Cabe esperar que este proceso renovador se extienda en la medida que se generalicen otras tecnologías como el correo electrónico, que despertó interés a estos efectos desde su surgimiento (Kiesler y Sproull, 1986) en virtud de la espectacular posibilidad de realizar encuestas sin necesidad de contactar físicamente a los encuestados.

Los lazos entre la estadística y la computación son tan intensos que cada vez son más frecuentes los textos de estadística que resultan casi indistinguibles de los manuales de determinados paquetes estadísticos; tal es el caso -por citar uno entre muchos otros- del libro de Venables y Ripley (1994) en que la enseñanza de esta disciplina ya viene asociada al paquete *S-Plus. Y* también viceversa; no es extraño, por ejemplo, que en calidad de cita bibliográfica se halle actualmente el manual de alguno de los grandes paquetes estadísticos tales como el *SPSS.*

14.2. El aprendizaje informático para acceder al presente

En la Sección 2.4.1 hice varios reparos importantes a la enseñanza pasiva, especialmente a la que se concreta a través del intercambio postal de materiales (textos, indicaciones y exámenes) por cheques bancarios. El aprendizaje específicamente del manejo de las computadoras exhibe algunos rasgos singulares; en este caso, no es que me oponga a los cursos informáticos «unilaterales» (que serían casi tan absurdos como aprender a nadar o a bailar por correspondencia), sino que, salvo alguna situación excepcional, reniego *de los cursos propiamente dichos,* por muy bilaterales o interactivos que sean.

Para comprender esta idea, lo primero que ha de interiorizarse es que cuando se llega a dominar una aplicación informática relativamente compleja, ya se domina el 40 por ciento de cualquier otra. No importa para qué sea el sistema que ya se maneja, ni el propósito del que ahora se enfrenta, porque dominar una aplicación es mucho más que adquirir un sistema de habilidades específicas; es entrar en contacto con una filosofía, con un modo de pensar y con un sinnúmero de nociones que son comunes a todo el ambiente de los ordenadores. De modo que para aquel que ya se ha adentrado en dicho ambiente, los cursos casi siempre resultarán esencialmente superfluos.

Pero consideremos el caso de un profesional de la salud que, del mismo modo que no puede prescindir del teléfono, comprende que no es razonable seguir manteniéndose al margen de un universo tecnológico que hace ya tiempo dejó de pertenecer al futuro; pero ignora cuál es la ruta crítica para integrarse en un entorno que aún le resulta personalmente lejano y quizás hostil. La primera sugerencia que le haría es que, salvo que tenga especial interés en desperdiciar entre el 80 y el 90 por ciento de su tiempo, *no* asista a cursos formales de computación.

Los requisitos fundamentales que a mi juicio han de cumplirse para optimizar el proceso de incorporación al mundo informático y en especial, el de dominar una aplicación informática concreta, son cinco:

1. Tener libre acceso a la computadora propiamente dicha.
2. Tener al menos un problema *real* que resolver con alguna aplicación (programa). Por ejemplo, nada mejor que *tener* que presentar en breve un informe técnico para llegar a dominar un sistema gráfico, o *tener* que redactar un documento que ha de entregarse dentro de una semana para iniciarse en la explotación de procesadores de textos (siempre que uno se empecine, claro está, en hacerlo por esa vía, aun cuando sienta por momentos que el método convencional sería más rápido).
3. Poseer un manual de la aplicación en cuestión. Usualmente, hay que bordear el retraso mental para no hallar en él una guía suficiente y adecuada; existen ejemplos, incluso, que son tan machacones que lo más recomendable sería evitarlos. Los programas tutoriales pueden ser un excelente sucedáneo.
4. Procurarse a alguien que «rompa la inercia» en que se halla inicialmente la relación del aprendiz con el programa; al sostener un contacto tripartito (aprendiz-«profesor»-computadora) durante un lapso que puede extenderse de una a cuatro horas (la duración depende del tema y del dominio previo que se tenga sobre el ambiente de la informática), se comprenderá lo que el sistema procura resolver, y se captarán panorámicamente las claves básicas que rigen su funcionamiento.
5. Contar con un amigo a quien acudir -quizás telefónicamente- para esclarecer dudas acuciantes que permitan, en un momento dado, escapar de

algún bloqueo ocasional que se produzca en la comunicación entre el aprendiz y el ordenador [1].

El hechizo tecnológico de las computadoras personales suele operar sobre cualquier espíritu dinámico. Una vez que se han dado los primeros pasos, ya es innecesario procurarse incentivos externos para continuar: el impulso autodidacta cobra vida propia. Tanto es así que hay que cuidarse del peligro de la dependencia, que ha llegado a convertirse en una genuina entidad sicopatológica, como alertaba recientemente Howard Shaffer, director adjunto de la División de Adicciones de la Universidad de Harvard (Verdú, 1995).

14.3. Saturación de los canales de entrada

Nunca como ahora se generó tanta información por unidad de tiempo; como compensación, nunca como ahora se ha contado con las actuales posibilidades de organizarla, recuperarla y transmitirla. El espacio dentro del que navegaban los cirbernantas de INTERNET era en octubre de 1995 de tal dimensión (100 bigabytes) uno de los grandes desafíos consiste en el diseño de programas que hagan las veces de mapa, brújula y vehículo al mismo tiempo. Sin embargo, el mayor de todos nace del hecho de que la «posibilidad de acceder a grandes volúmenes indiscriminados de información no es ninguna ganga, sino todo lo contrario; la utilidad de la información decrece en cuanto ella abunda. No necesitamos información que sea muy abundante, sino información muy relevante. Lo que necesitamos no es información, sino conocimiento, cosa que ninguna fuente de datos, por extensa que sea, puede darnos» (Carbó, 1995).

La cantidad de información que genera cotidianamente la actividad médica en particular es asombrosamente grande; no toda esa producción, sin embargo, es útil. La polución informativa que contamina el ambiente sanitario es ciertamente inquietante. Todd-Pokropek (1987) estimaba hace ya muchos años (en esta materia diez años son muchos) que la cantidad de información almacenada por día en un hospital moderno de 700 camas está cerca de un gigabyte (mil millones de caracteres alfanuméricos) y el volumen de información allí archivado crece en progresión geométrica de razón dos cada siete años. No es nada sorprendente si se tiene en cuenta que, según declaraciones de S. Kaihara, director del Hospital Computer Centre de la Universidad de Tokio [2], el 40% del tiempo laboral del personal sanitario en los grandes hospitales se emplea en tareas de elaboración, transmisión y archivo de información.

[1] Lo ideal es que sea suficientemente amigo como para despertarlo por la madrugada si fuera menester.
[2] Citado en Sáez (1994).

En cuanto a la atención primaria de salud, desde muy temprano se produjo un extraordinario y creciente interés por el registro computarizado de la información procedente de este nivel de atención. El tema fue motivo de reflexión y debate desde que se dieron los primeros pasos informáticos en esta esfera (Barnett, 1984; Sheldon, 1984). Pero los resultados han sido menos alentadores de lo esperado. Ritchie (1990) apuntaba:

> ***Los médicos de la atención primaria hacen diariamente miles de anotaciones en las historias clínicas de sus pacientes, generando así información inestimable sobre la salud de la nación. Tristemente, la mayor parte de esta información se pierde para la posteridad: los métodos tradicionales de obtención de datos son engorrosos y desproporcionados para la magnitud de la tarea, y el potencial epidemiológico permanece oculto.***

En la época en que se iniciaba el acceso universal a las computadoras, Farrell y Worth (1982) señalaban a su vez que «aparentemente los médicos de atención primaria son tan lentos como los demás médicos para emplear la tecnología informática». Transcurrido un lapso prolongado desde entonces, la situación ha cambiado, aunque mucho menos de lo deseable.

En este contexto de saturación se produce la irrupción de algunas tecnologías como el CD-ROM, medio que introduce facilidades tan atractivas como inquietantes. Cada semana el receptor es bombardeado con la oferta de decenas de nuevos discos compactos; cada uno de ellos contiene información tal que, solamente para enterarse del contenido (ya no para asimilarlo o explotarlo), hacen falta muchas horas frente al monitor. El problema es que esta oferta audiovisual ya no pertenece a la escala humana. Recientemente vi en una exposición una aplicación de ***Windows*** que tenía varias ventanas abiertas simultáneamente: en una aparecía un partido de fútbol que transmitía en ese instante la televisión; en otra se podía ver la película ***Lo que el viento se llevó;*** en una tercera había una hoja de cálculo... un prodigio tecnológico. En aquel momento tomé conciencia de que solamente poseo dos ojos y un único cerebro. Entonces sentí una aguda nostalgia por lo que Umberto Eco llama «la reflexión tranquila del texto».

14.4. Ganar tiempo, perder el tiempo

Las posibilidades que brinda la informática como recurso individual para ahorrar tiempo son tan notables y obvias que quizás no valga la pena extenderse en argumentos para persuadir al lector. No obstante, para ilustrar el asunto con un ejemplo muy simple de la práctica cotidiana en la investigación epidemiológica, consideremos el caso en que se quieren estimar los parámetros del modelo logístico por el método de máxima verosimilitud en una situación en que se trabaje con 2.000 sujetos y 15 variables independientes, problema de magnitud media, que una

computadora personal no muy avanzada resuelve en no más de cinco minutos. He estimado que un preso, trabajando 12 horas diarias en su celda con una calculadora de bolsillo, para resolver tal problema sin errores, necesitaría una condena de 4 años.

Es interesante, incluso, el hecho de que un problema planteado hace dos siglos, cuya solución era teóricamente conocida en esa misma época, no pudo ser resuelto ni en aquel momento ni en los 200 años siguientes por la simple razón de que no apareció nadie dispuesto a destinar varios años de su vida a la aplicación del método en cuestión. Se trataba de dirimir a cuál de dos personajes correspondía la autoría de un documento histórico; la aplicación de un algoritmo basado en el uso recursivo del teorema atribuido al reverendo Thomas Bayes [3] y presentado públicamente en 1763 tuvo que esperar por la computadora para que actuara en la interfase que iba de una enorme base de datos -textos de uno y otro personaje cuya autoría estaba fuera de dudas- a la discriminación estilística del documento. Hollingdale y Tootill (1965) dan cuenta de un trabajo similar en que dos eruditos británicos en cuestiones bíblicas de la Universidad de Glasgow (el profesor G. McGregor y el reverendo A. Morton), usando un ordenador de segunda generación, resolvieron hace ya decenios un problema de autoría relacionado con las epístolas de San Pablo, debatido durante siglos.

Para dar una idea tangible del crecimiento que ha experimentado la velocidad de cálculo Dormido (1995) señalaba que si hubiese habido en los últimos 100 años el mismo aumento en la velocidad del transporte que el que ha existido en la de cálculo, se podría viajar de Madrid a Los Angeles en 0,002 segundos o ir de la Tierra a la Luna en 0,l segundos o al sol en 6,5 minutos.

Pero esa «ganancia de tiempo» no está exenta de riesgos: el asunto estriba en cómo se use el «tiempo ganado». Bunge (1985) llega a decir:

> ***El ordenador puede servir de taparrabos para ocultar la indigencia intelectual. Antaño, cuando el intelectual carente de ideas originales quería publicar algo, pergeñaba un artículo o un libro en lenguaje que, por ser oscuro, causaba impresión de profundidad; o bien acumulaba estadísticas al tuntún y calculaba a mano numerosos promedios o coeficientes de correlación. Hoy día, el impostor intelectual puede fabricar basura cultural mucho más rápidamente con ayuda de un ordenador***

La optimización de tiempo que puede conseguirse en el tratamiento de datos estadísticos ha llegado como un bálsamo para muchos ‘profesionales y como una patente de corso para otros. Sáez (1994) ha advertido que sólo lo que cuantificamos numéricamente «nos parece concreto y profesional (...) aunque detrás del efecto

[3] Por cierto, hace unos años se han abierto dudas acerca de la responsabilidad por el famoso «teorema de Bayes». En un ingenioso trabajo histórico-detectivesco, Stigles (1983) da cuenta de un «sospechoso», el matemático inglés Nicholas Sunderson, sobre quien pesan serios indicios «acusatorios».

mágico y adormecedor de los números hay todo un escenario de infiabilidad estadística e intereses. Nos engañamos unos a otros, y a nosotros mismos». Dalla1 (1988), por ejemplo, pone en evidencia en qué medida es esto cierto para algunos «paquetes estadísticos» que circulan.

Ahora bien, ¿pueden estos fabulosos ahorradores de tiempo contribuir también a que lo perdamos? La respuesta, lamentablemente, es positiva. Y no me refiero, por cierto, a los videojuegos, que no implican necesariamente una pérdida de tiempo. Algunos son sumamente instructivos; otros contribuyen a la estructuración lógica del pensamiento. Por otra parte, juegos como *VIDA*, basado en la seductora teoría de los autómatas celulares creada por el matemático John H. Conway (Millium, Reardon y Smar, 1978) no sirven -al menos hasta ahora- para nada, pero entrañan una fascinante propuesta estética [4]. Me refiero a que el magnetismo que ejercen las tecnologías puede hacernos extraviar la brújula que habría de guiarnos a economizar ese recurso, el menos renovable de todos.

Hace un par de años estaba preparando un informe que, por su naturaleza, debía contener tres -a lo sumo cuatro- gráficos estadísticos basados en los datos de que disponía. Puesto que tenía a mi alcance un magnífico sistema para la confección de gráficos, fui creando archivos, cada uno de los cuales contenía una figura. Para algunas de ellas hice hasta cuatro versiones; ya me ocuparía luego de elegir las dos o tres figuras más elocuentes para incluir en el informe. Para acopiar la treintena de preciosos gráficos con que acabé la sesión, había invertido unas cuatro horas. Diez años antes, si ante la misma situación me hubiera empeñado en conseguir esa colección pictórica inicial, hubiera tenido que acudir a un dibujante que destinara al asunto unas tres semanas de dedicación exclusiva; es decir, algo impensable. No hubiera tenido más remedio en tal caso que dedicar dos horas a pensar cuáles serían los 3 o 4 gráficos que habría de pedir al dibujante. Es decir, a ***pensar en el problema,*** en lugar de estar cuatro horas «pensando» en cómo usar los dedos sobre el teclado.

La experiencia ilustra la perversión en que se puede incurrir si ponemos sumisamente nuestro intelecto al servicio de la computadora en lugar de usarla para incrementar las potencialidades de aquel. El hecho me recuerda el caso de un alumno que, después de apretar las teclas debidas en su calculadora de bolsillo, obtuvo (y registró en un examen) que ¡el logaritmo de 1 era 0.0000001!

En general, de lo que no quedan dudas es de que resulta muy fácil perder el tiempo con las computadoras, en especial aprendiendo cosas antes de detenernos a reflexionar cuan inútiles nos van a resultar. Umberto Eco (Schemla, 1993) declaraba recientemente:

[4] En cualquier caso, no veo razones para conceptuar como inútil ninguna elección para el esparcimiento, aunque confieso que una buena sesión aniquilando marcianos, al menos a mí, me deja en un estado de abatimiento (de magnitud proporcional al tiempo utilizado) típico de quien ha estado deambulando por el reino de la intrascendencia.

> ***Hay idiotas de la computadora así como hay idiotas del walkman que uno ve retorcerse y gritar en los conciertos de rock. ¿Pero son acaso más idiotas que las personas que se flagelaban en la Edad Media? Las formas de autodestrucción cambian a través de las épocas.***

Lo cierto es que no resulta difícil llegar a convertirse en un apéndice de la máquina. Miller (1993) hace una observación en cuya elocuente simplicidad recomiendo que se repare:

> ***La mayoría de la gente que conozco usa las computadoras de alguna manera. Los más productivos no son los «expertos» en computación. Los productivos no tienen la más mínima idea acerca de cuán poderosos son sus equipos. Ellos no tienen la menor preocupación acerca de utilidades para mejorar la productividad, ni sobre memoria RAM o el tamaño del disco duro. Ellos usan los dos o tres programas que necesitan; ellos, simplemente, hacen su trabajo.***

Cabe advertir, sin embargo, en que el riesgo de desperdiciar el tiempo no es un fenómeno exclusivo del entorno computacional sino propio de un momento tecnológico singular. En este sentido Galeano (1993) alertaba:

> ***El automóvil, el televisor; el video, la computadora personal, el teléfono celular; y demás contraseñas de la felicidad, máquinas nacidas para ganar tiempo o para pasar el tiempo, se apoderan del tiempo (...) en resumidas cuentas, las personas terminan perteneciendo a las cosas y trabajando a sus órdenes.***

14.5. ¿Mayor equivale a mejor?

El número de «paquetes» para el tratamiento de datos estadísticos que inundan el mercado no sólo es cuantioso sino que se incrementa sin cesar. Entre los sistemas aplicativos más connotados (sin contar las hojas de cálculo y los sistemas de bases de datos con funciones estadísticas incorporadas), pueden mencionarse los siguientes: SAS, BMDP, SPSS, EGRET, S-PLUS, STATGRAPH, SYSTAT, EPIINFO, EPILOG, STATA, MICROSTAT y MINITAB. Así podría construirse una lista de casi una centena; de hecho, esa es la cifra que registran publicaciones especializadas, como la guía de ***software*** para esta rama compilada por Koch y Haag (1992).

Una observación informal pero consistente en esta materia me ha persuadido de que tal espectro de ofertas estadístico-computacionales ha favorecido el uso inadecuado de los métodos. Como siempre, la culpa no la tienen las tecnologías en sí mismas, aunque algunas -envueltas en determinadas propuestas publicitarias- pueden inducir a que se usen de manera acrítica. Lo cierto es que muchos investigadores de segundo nivel destinan más tiempo a aprender el manejo de cada vez más productos computacionales que a procurarse el conocimiento detallado de las

propiedades y las condiciones para la aplicación de los procedimientos que abarcan. La computadora es una «herramienta obediente»: hace lo que le exigen. Pero ocurre que también obedece aunque no se cumplan los presupuestos teóricos de los métodos programados.

En el caso del *software* estadístico, al crecer el número de opciones, se va consolidando la convicción de que es mucho lo que se ignora y, por ende, mucho lo que se tiene que llegar a conocer para hacer una investigación de excelencia. Lo primero es sin duda cierto: el cúmulo de recursos potenciales que un investigador normal no domina es enorme; pero lo segundo, posiblemente, no pase de ser un mito alimentado por intereses comercialistas y por la creciente fascinación ante la tecnología, que es mayor y menos reflexiva cuanto más frívolos sean sus usuarios. Lo que hay que advertir es que entrar en esa febril «carrera» tecnológica es absurdo por la simple razón de que está perdida de antemano.

De hecho, la situación es muy similar a la que exhibe la práctica de prescripción de medicamentos: aunque hay miles de productos disponibles -algunos de altísima especificidad, propios de la atención terciaria- con unas pocas decenas de ellos se resuelve el 90% de las demandas farmacoterapéuticas individuales. El símil, por lo demás, es integral pues, como ocurre con los fármacos, muchos métodos estadísticos son innecesarios, otros crean adicción y se corre el riesgo de abusar de ellos, no raras veces generan efectos adversos, y a menudo se pasan por alto sus contraindicaciones.

La proliferación de sistemas de *software* estadístico trae consigo una oferta aturdidora, hecho que ayuda a que un profesional de la salud crea que lo mejor es aprender el más sofisticado o el más reciente. Téngase en cuenta que muchos productos altamente publicitados en la actualidad -especialmente, y no por casualidad, los más caros- buscan su promoción y avalan su precio en la enorme gama de opciones que ofrecen y exaltando la inclusión de las técnicas más avanzadas y modernas. Procede entonces preguntarse más formalmente: ¿le basta a un investigador (no hablamos de un estadístico profesional) con manejar un programa que se circunscriba a las técnicas estadísticas básicas para encarar la mayoría de los problemas de investigación contemporánea?, ¿existe tal producto?, ¿hay motivos para pensar que algunas de las ofertas están marcadamente lejos de poseer de tales rasgos?

El problema clave reside, en definitiva, en responder cuál sería el procesador estadístico más recomendable. La respuesta a esta inocente interrogante puede despertar pasiones. He presenciado intensas discusiones sobre estas alternativas, defendidas y vituperadas como si se tratara de políticos o cantantes de la música pop.

En cierta ocasión en que varios jóvenes investigadores indagaban sobre la conveniencia de aprender el sistema EPJINFO, disentí de un colega que recomendaba que se adhirieran al SPSS, sistema que a su juicio era «mucho mejor». El argumento fundamental de su sugerencia consistía en que dicho sistema es muy completo, incluye técnicas multivariadas y, para cada problema, ofrece gran cantidad de alternativas. Mi argumento fundamental para recomendar, en cambio, EPIINFO era, sorpren-

dentemente, el mismo: que no incluye técnicas multivariadas (con excepción de la regresión múltiple), ni es muy completo, ni da muchas alternativas de solución al mismo problema, rasgos que considero altamente ventajosos para los nóveles profesionales con que dialogábamos.

Afirmar que un paquete dado es mejor porque es más abarcador es tan absurdo como argumentar que un avión *jet* es siempre un medio de transporte mejor que un pequeño automóvil, por ser más rápido, capaz de transportar a más pasajeros y poder cubrir distancias mucho más largas.

En cualquier caso, antes de profundizar en el tema, vale la pena recordar que hablamos de una herramienta, de un recurso complementario a muchos otros para hallar respuestas de interés y que, como señalan Schoolman ***et al.,*** (1968), «las buenas respuestas se generan a partir de buenas preguntas más que de análisis esotéricos».

Para determinar en qué medida estos dos recursos de ***software*** estadístico bastan para resolver los problemas que realmente se encaran y en qué grado ofrecen procedimientos que desbordan las necesidades prácticas reales, en un estudio desarrollado por Silva y Pérez (1995) se cotejaron los procedimientos estadísticos usados en los artículos publicados entre 1986 y 1990 en dos revistas de alto factor de impacto (***New England Joumal of Medicine*** - N E J M - y ***American Journal of Epidemiology*** -AJE-) con las posibilidades potenciales de estos dos conocidos paquetes estadísticos [5], representantes de respectivas filosofías: una «maximalista» (la del SPSS), la otra orientada a proveernos de los recursos más elementales (la del EPIINFO).

Para cada revista examinada se intentó conocer hasta dónde bastaban estas dos ofertas informáticas para encarar los tratamientos de datos que se planteaban sus autores en materia estadística. Más concretamente, se computó el número de artículos cuyas demandas estadísticas podían cubrirse completamente con los métodos incluidos en cada uno de los dos paquetes. En principio, para un sistema estadístico específico, los artículos pueden clasificarse en tres tipos:

1) Los que no usan estadística.
2) Aquellos cuyas demandas estadísticas pueden ser cubiertas con el sistema.
3) Aquellos que hicieron uso de al menos un procedimiento estadístico no contenido en el sistema en cuestión.

En el caso de NEJM, la primera categoría abarcaba 226 trabajos (17%); para AJE, éstos fueron 41 (4%). La Tabla 14.1 resume los resultados hallados para los trabajos restantes.

[5] EPIINFO no es propiamente un *software* de estadística, pero contiene varios módulos especializados en el tema, que fueron los que se tuvieron en cuenta.

Tabla 14.1. ***Distribución de artículos que usan métodos estadísticos en AJE y NEJM (1986 - 1990) según el grado de satisfacción que darían EPIINFO y SPSS***

	NEJM (1.115 artículos)				AJE (1.004 artículos)			
	EPIINFO		SPSS		EPIINFO		SPSS	
	N.º	%	N.º	%	N.º	%	N.º	%
Quedan cubiertos	711	64	695	62	584	58	391	39

Los datos son elocuentes; los paquetes son similares en el caso de NEJM, pero la supremacía de EPIINFO es apreciable para AJE. SPSS, con sus numerosas posibilidades, no consigue responder a las necesidades del 61% de los trabajos de esta influyente publicación [6]. De lo que no hay duda es de la falta de correspondencia existente entre algunos productos comerciales (y en alguna medida entre los que, como el EPIINFO, no lo son) y la práctica objetiva, al menos en lo que concierne a un importante segmento de la investigación en el campo de la salud.

Por otra parte, y esto es lo más interesante, se pudo constatar que, mientras EPIINFO no contiene métodos ajenos a las necesidades objetivas de los investigadores que han conseguido publicar en alguna de las dos revistas a lo largo del quinquenio, SPSS ofrece una apreciable cantidad de posibilidades que dichos investigadores, al margen de cuál sea la razón, no usan (véase Sección 2.6.1).

Los grandes paquetes estadístico-computacionales de aliento comercial no parecen constituir una solución metodológica más eficiente que la que ofrecen herramientas más elementales a los efectos de conseguir una producción científica de alta calidad.

14.6. La simulación abre sus puertas

Al contar con aliados tan poderosos como las computadoras, los científicos han potenciado su capacidad de acción en todas las fases de la investigación, desde la indagación bibliográfica en que descansa la pertinencia de sus preguntas hasta los procedimientos para comunicar las respuestas halladas. Sin embargo, la etapa de trabajo más radicalmente revolucionada es aquella en que se procesan los datos disponibles, sea en procura de modelos que los expliquen o del proceso inverso, más típico del tratamiento estadístico, consistente en la valoración de hipótesis o teorías a través de la aplicación de modelos teóricos al material obtenido.

Desde que el profesor J. C. Kendrew, de la Medical Research Unit de Cambridge, desarrollara su famosa investigación sobre la mioglobina, compleja proteína

[6] Esto se debe, básicamente, a que este sistema no computa ***odds ratios*** ni otras mediciones epidemiológicas conexas.

que se produce en el músculo animal y que le valiera el premio Nobel de Química de 1962, el papel de la computadora consolidó su presencia en una nueva dimensión: ya no se trataba de un aparato veloz y eficiente para ahorrarnos laboriosos cálculos. Se convirtió también en un desbrozador de caminos, un recurso que no sólo permite transitarlos con más eficiencia sino que es capaz de determinar de antemano cuáles avenidas son inútiles y cuáles promisorias.

Las moléculas de mioglobina, como ocurre con todas las proteínas, poseen miles de átomos. Para comprender su comportamiento es imprescindible determinar, además de la estructura química, lo que Kendrew llamó su «arquitectura molecular»: el modo en que los átomos se organizan tridimensionalmente en el espacio. El examen del astronómico número de posibilidades fue posible gracias a la computadora que en la década del 50 poseía la Universidad de Cambridge, y aun así llevó varios años concluirlo exitosamente. El modelo metodológico de Kendrew constituye un paradigma, hoy desarrollado con muchísima más eficiencia debido a la prodigiosa elevación de la velocidad y a los notables avances de los sistemas de programación de que se ha llegado a disponer. El investigador actual puede establecer un fluido «diálogo» con su ordenador; la dinámica de propuestas y respuestas, de ensayos y constataciones, se desarrolla en ciclos tan breves como lo permita su creatividad. El mecanismo retroalimentador implícito configura un espacio metodológico completamente novedoso.

En ese papel retroalimentador un lugar prominente corresponde a las técnicas de simulación. Los procesos simulados (biológicos, epidemiológicos, sociales o clínicos) son con frecuencia de naturaleza probabilística, y las conclusiones que de ellos se derivan, típicamente estadísticas [7]. Ello completa, como se puede apreciar, una fecunda simbiosis entre tres componentes: la estadística, la simulación y la computadora. Esta triple combinación ha hallado una concreción tangible en un sistema computacional programable denominado *Resampling Stuts*. Este atractivo y original producto, que ha tenido diversas versiones desde 1989, se debe al profesor Julian L. Simon, coautor -por cierto- de un interesante libro de metodología de la investigación (Simon y Burstein, 1978). El programa ha sido diseñado para aprender y enseñar la teoría básica de probabilidades y estadística a partir de la técnica de *bootstrap* (Efron, 1982) y se erige a partir de una concepción cuya línea teórica fundamental puede hallarse en los trabajos de Simon, Atkinson y Shevokas (1976) y de Efron (1983). Por su carácter interactivo, su extrema sencillez y su capacidad para hacernos palpar el misterio de la probabilidad, recomiendo vivamente que se tome contacto con este ingenioso «juego» estadístico.

[7] Técnicas como la de anticiparse computacionalmente al resultado de una cirugía plástica, no caen dentro de lo que clásicamente se ha entendido como «simulación» aunque de hecho se está simulando una operación.

Bibliografía

Baker RP (1992). ***New technology in survey research: Computer assisted personal interviewing (CAPI).*** Social Science Computer Review 10: 145-157.

Barnett GO (1984). ***The application of computer-based medical-record systems in ambulatory practices.*** New England Journal of Medicine 310: 1643-1645.

Bunge M (1985). ***Seudociencia e ideología.*** Alianza, Madrid.

Carbó JM (1995). ***Al otro lado del modem.*** Byte, octubre: 118-125.

Couper MP, Burt G (1994). ***Interviewer reactions to alternative hardware for computer assisted personal interviewing.*** Journal of Official Statistics 8: 201-210.

Dalla1 M (1988). ***Statistical microcomputer - Zike it is.*** The American Statistician 42: 212-216

Dormido S (1995). ***La revolución del conocimiento.*** Muy especial 20: 16-24.

Efron B (1982). ***The jackknife, the bootstrap, and other resamplingplans.*** SIAM.

Efron B (1983). ***Computer intensive methods in statistics.*** Scientific American 232: 116-130.

Farrell DL, Worth RM (1982). ***Implementation of a computer assisted record system in the family practice office.*** Hawaii Medical Journal 41: 90-93.

Galeano E (1993). ***Ser como ellos.*** Semanario Brecha 9 (422): 32, Montevideo.

Grampone J (1992). ***Yo, hombre: tú, computadora.*** La Flor del Itapebí, Montevideo.

Hollingdale SH, Tootill GC (1965). ***Computadores electrónicos.*** Alianza, Madrid

Kiesler S, Sproull LS (1986). ***Response effects in electronic surveys.*** Public Opinion Quarterly 50: 402-413.

Koch A, Haag U (1992). ***The Statistical software guide' 92/93.*** Statistical Software Newsletter.

Miller MC (1993). ***Does more power mean more productivity?*** Medical Decision and Computing 10: 208-209.

Millium J, Reardon J, Smar P (1978). ***Life with your computer:*** Byte ***3: 45-50.***

Nicholls WL II, Baker RP, Martin J (1996). ***The effects of new datacollection technologies on survey data quality.*** En: Lyberg L ***et al.*** (editores) ***Survey measurement and processquality*** Wiley, New York.

O'Reill JM, Hubbard ML, Lessler JT, Biemer PP, Turner CF (1994). ***Audio and video computer assisted self-interviewing: preliminary tests of new technologies for data collection.*** Journal of Official Statistics 2: 197-214.

Ritchie LD-(1990). ***Ordenadores en atención primaria.*** Díaz de Santos, Madrid.

Sáez F (1994). ***El hombre y la técnica.*** América Ibérica, Madrid.

Saris WE (1989). ***A technological revolution in data collection.*** Quality and Quantity 23: 333-349.

Schemla E (1993). ***La computadora es masturbatoria.*** Axxon ***42: 23-26.***

Schoolman HM, Becktel JM, Best WR, Johnson AF (1968). ***Statistics in medical***

research: principles versus practices. Journal of Laboratory and Clinical Medicine 71: 357-367.

Sheldon MG (1984). ***Computers in general practice: a personal view.*** Journal of the Roya1 college of General Practitioners 34: 647-648.

Silva LC, Pérez C (1995). ***Suficiencia e insuficiencia del software estadístico ante las demandas de dos revistas biomédicas de alto factor de impacto.*** Inédito.

Simon JL, Atkinson DT, Shevokas E (1976). ***Probability and Statistics: Experimental Results of a Radically Different Teaching Method.*** The American Mathematical Monthly 83: 733-739.

Simon JL, Burstein P (1978). ***Basic research methods in social science.*** Random House, New York.

Stigler SM (1983). ***Who discovered Bayes' theorem.*** The American Statistician 37: 290-296.

Todd-Pokropek A (1987). ***Medical imaging.*** Computer Bulletin 3(4).

Venables WN, Ripley BD (1994). ***Modern applied statistics with S-Plus.*** Springer-Verlag, Berlin.

Verdú V (1995). ***«Colgados» por la informática. El País, 12 de marzo de 1995***, Madrid

Weeks MF (1992). ***Computer-Assisted Survey Information Collection: A review of CASIC methods and their implication for survey operations.*** Journal of Official Statistics 4: 445-466.

Anexos

Anexo N.º 1

MODELO DE ENCUESTA PAR4 LA EVALUACIÓN DE CONOCIMIENTOS FARMACOLÓGICOS DEL MÉDICO DE ATENCIÓN PRIMARIA

1. ¿En cuáles de estos casos utilizaría usted el CLORAMFENICOL como droga de elección?

-Absceso dental: SÍ ____ NO ____ NO TENGO OPINIÓN ____
-Fiebre tifoidea: SÍ ____ NO ____ NO TENGO OPINIÓN ____
-Amigdalitis: SÍ ____ NO ____ NO TENGO OPINIÓN ____
-Estreptocóccica «B»: SÍ ____ NO ____ NO TENGO OPINIÓN ____
-Infección urinaria: SÍ ____ NO ____ NO TENGO OPINIÓN ____

2. ¿En cuál de los siguientes casos cabría considerar el uso conjunto de AMPICILINA y CLORAMFENICOL?

-Peritonitis por apendicitis perforada: SÍ____ NO____ NO SÉ ____
-Sepsis meningocóccica: SÍ ____ NO ____ NO SÉ ____
-En determinadas infecciones hospitalarias: SÍ ____ NO ____ NO SÉ ____
-Neumonía por Haemophilus: SÍ ____ NO ____ NO SÉ ____

3. Para el tratamiento sintomático de la tos, ¿cuál de los siguientes fármacos alternativos tendría en cuenta para sustituir al COSEDAL?

-EPINEFRINA: SÍ ____ NO ____ NO SÉ ____
-AMINOFILINA: SÍ____ NO ____ NO SÉ ____
-NOSCAPINA: SÍ ____ NO ____ NO SÉ ____
-AMPICILINA: SÍ ____ NO ____ NO SÉ ____

4. ¿Que procedimiento de prescripción le parece el adecuado ante una infección de vías aéreas superiores, con fiebre moderada? Marcar una de las 5 opciones.

_____ Antibioticoterapia parenteral y oral durante 7 días, más antitérmicos.
_____ Antibioticoterapia oral durante 10 días más antitérmicos.
_____ Antibioticoterapia más inmunoglobulinas inespecíficas.
_____ Solamente antitérmicos.
_____ Ignoro la respuesta.

5. Cuáles de las siguientes reacciones secundarias cree que pueden manifestarse tras la administración de BENADRILINA:

-Somnolencia e incoordinación de ideas: SÍ____NO____NO SÉ____
-Debilidad muscular: SÍ____NO____NO SÉ____
-Anorexia: SÍ ___ NO____NO SÉ____
-Hipersalivación: SÍ ___ NO____NO SÉ____

6. Señale cuáles de las siguientes entidades constituyen contraindicaciones para la administración de BUTACIFONA:

-Diabetes Mellitus: SÍ ____ NO ____ NO SÉ ____
-Historia de úlcera péptica: SÍ ____ NO __ NO SÉ ____
-Historia de discrasias sanguíneas: SÍ ____ NO ____ NO SÉ ____
-Hipotensión arterial: SÍ ____ NO ____ NO SÉ ____

7. Señale cuáles de las siguientes afirmaciones acerca de la CIPROHEPTADINA son, a su juicio, correctas :

-No presenta reacciones secundarias relevantes:
CORR _____ INCORR _____ NO SÉ _____
-Está contraindicado en pacientes con glaucoma:
CORR _____ INCORR _____ NO SÉ _____
-No administrar de forma continuada durante más de 6 meses:
CORR _____ INCORR _____ NO SÉ _____
-Está compuesto sólo por vitaminas:
CORR _____ INCORR _____ NO SÉ _____

8. ¿Cuáles de las siguientes afirmaciones cree usted que son correctas acerca del NULIP?

-Una indicación aceptada para su uso es la depresión leve:
CORR _____ INCORR _____ NO SÉ _____

-En su composición entra el Sulfato de Dextroanfetamina:
CORR _____ INCORR _____ NO SÉ _____
-Está contraindicado en el hipertiroidismo:
CORR _____ INCORR _____ NO SÉ _____
-Puede utilizarse libremente en pacientes con trastornos psiquiátricos previos:
CORR _____ INCORR _____ NO SÉ _____

9. De los siguientes hipotensores, ¿cuáles cree que pueden provocar depresión como reacción secundaria?

-METILDOPA: SÍ _____ NO _____ NO SÉ _____
-CLORTALIDONA: SÍ _____ NO _____ NO SÉ _____
-RESERPINA: SÍ _____ NO _____ NO SÉ _____
-FUROSEMIDA: SÍ _____ NO _____ NO SÉ _____

10. ¿Cuáles de las siguientes consideraciones cree que son verdaderas y cuáles falsas acerca de la TETRACICLINA?

-No se debe administrar en embarazadas ni en menores de 7 años:
VERDAD _____ FALSA _____ NO SÉ _____
-Es muy eficaz en el tratamiento de las Salmonelosis:
VERDAD _____ FALSA _____ NO SÉ _____
-Es más bacteriostático que bactericida:
VERDAD _____ FALSA _____ NO SÉ _____
-La ingestión conjunta con leche o derivados interfiere en su absorción:
VERDAD _____ FALSA _____ NO SÉ _____

ll. ¿Cuáles de los siguientes antimicrobianos se consideran potencialmente perjudiciales para su uso durante el embarazo?

-METRONIDAZOL: SÍ _____ NO _____ NO SÉ _____
-PENICILINA: SÍ _____ NO _____ NO SÉ _____
-ERITROMICINA: SÍ _____ NO _____ NO SÉ _____
-SULFAPRIM: SÍ _____ NO _____ NO SÉ _____

12. Ante una «diarrea transicional del recién nacido», ¿cuáles de las siguientes opciones serían a su juicio valorables?

-Indicar FURODONE: SÍ _____ NO _____ NO SÉ _____
-Quitarle el pecho: SÍ _____ NO _____ NO SÉ _____
-Prescribir KAOENTERIN: SÍ _____ NO _____ NO SÉ _____
-No dar tratamiento farmacológico: SI _____ NO _____ NO SÉ _____

13. ¿ Qué otro u otros fármacos emplearía para sustituir al CORINFAR en el caso de que éste no se encontrara en el mercado? Márquelos con una cruz.

_ ATENOLOL _ VERAPAMILO _ PROPANOLOL _ NITROSORBIDO _ NO SÉ

14. Ante un caso con síntomas de depresión y melancolía, ¿cual o cuales de los siguientes medicamentos podrían ser adecuados? Márquelos con una cruz.

_ NITRAZEPAM _ DIAZEPAM- IMIPRAMINA- MEPROBAMATO- NO SÉ

ANEXO N.º 2

INDICADORES DE PARTIDA SELECCIONADOS PARA CONSTRUIR EL IDSC

1. Tasa de mortalidad del menor de cinco años (TMMS)

Es reconocida su capacidad para medir el nivel de los cambios del estado de bienestar. Tiene entre sus virtudes las siguientes: mide los resultados finales del proceso de desarrollo en lugar de los factores intermedios, es el resultado de una amplia variedad de componentes como la salud nutricional y conocimientos básicos de salud de la madre, cobertura de inmunización, acceso de los servicios de atención materno-infantil, nivel de ingresos, acceso a agua potable y saneamiento eficaz, y grado de seguridad del medio ambiente infantil.

2. **Índice de bajo peso al nacer (IBP)**

Mide de forma indirecta el estado nutricional y la atención prenatal recibida. Por otra parte, se ha demostrado que los niños que nacen con bajo peso tienen mayor probabilidad de enfermar o morir, así como que, los que sobreviven, padecen con mayor frecuencia dificultades en el aprendizaje.

3. **Mortalidad por enfermedades diarreicas (MDI)**

Ésta es, a escala mundial, una de las causas más importantes de muerte, tanto del lactante como de personas de avanzada edad. La tasa refleja la calidad de los servicios de atención primaria (tanto en lo relacionado con la prevención de esta enfermedad, como con el diagnóstico oportuno y el tratamiento adecuado). Se relaciona con los programas de protección de salud mediante el saneamiento del medio

ambiente (eliminación de excretas, agua potable, control de vectores) y, por otra parte, se vincula con el nivel de escolaridad de la madre y la alimentación del niño. Fue considerada preferible a las medidas de morbilidad por este concepto debido al consabido subregistro (en algunos contextos, muy acusado) que se produce en relación con las dolencias con este origen.

4. **Tasa bruta de mortalidad por enfermedades del corazón (COR)**

Es un importante trazador para el conjunto de las enfermedades crónicas degenerativas, las cuales se vinculan directamente con la estructura poblacional, la atención preventiva realizada en un territorio, y son reflejo indirecto de hábitos y conductas insaludables, así como del nivel de vida de la población.

5. Consultas estomatológicas por cada mil habitantes (EST)

Da idea de la capacidad de prestación de servicios que tiene una comunidad y, además, muestra el desarrollo de los servicios de estomatología y del nivel de conocimientos higiénicos al respecto alcanzado por la población. Puesto que es conocido que virtualmente todos los sujetos necesitan de alguna forma de atención estomatológica, no hay dudas de que, a mayor valor del indicador, mayor será el nivel de salud de la comunidad.

6. Enfermeras por cada mil habitantes (ENF)

Refleja los recursos humanos destinados a salud en dicha población.

7. Tasa de incidencia de sífilis (VEN)

La sífilis es una enfermedad venérea bastante difundida y estrechamente vinculada a normas o conductas sociales. Este tipo de enfermedad, por su frecuencia y por el daño que causan, representan un problema sanitario de importancia. Por otra parte, se relaciona directamente con la labor de educación para la salud realizada en el territorio y traducen, por tanto, la educación sanitaria prevaleciente.

8. Tasa de abortos por cada 1.000 mujeres entre 12 y 49 años (ARO)

Aunque el aborto no es un método contraceptivo, se utiliza como forma de regulación de la fecundidad. Se relaciona, por otra parte, con la proporción de mujeres

que desean parir, su estado marital y, de forma indirecta, con su nivel cultural. Guarda claro vínculo, finalmente, con el equilibrio psicosocial de la comunidad.

9. **Tasa bruta de divorcialidad (DIV)**

Refleja indirectamente el grado de estabilidad psicosocial de la población, y también, de forma indirecta, se vincula con el problema de la vivienda, el cual influye en la estabilidad del matrimonio.

10. Tasa bruta de mortalidad por suicidio (SUI)

Su magnitud refleja el equilibrio psíquico de los integrantes de la sociedad. Regularmente traduce el grado de estrés en que vive el hombre y su capacidad de adaptación. Involucra asimismo la oportunidad y efectividad del apoyo que las redes sociosanitarias prestan a la comunidad.

11. Porcentaje de población servida con agua suministrada por acueductos (AGU)

Este indicador se relaciona con el medio físico y en cierto sentido es un trazador del marco ecológico en que se inscribe la comunidad; constituye el más importante medio de difusión de enfermedades, en especial en las de índole transmisible.

12. Porcentaje de niños menores de 2 años adecuadamente vacunados (IMN)

Se relaciona íntimamente con el trabajo preventivo realizado en una comunidad; es por tanto un representante de la gestión de promoción de salud.

13. Porcentaje de alfabetismo

Capta la capacidad de comunicarse y participar en la vida de la comunidad. Tanto es así que se considera que el alfabetismo es un requisito básico para poder adquirir información del medio donde se desarrolla el hombre.

14. Consumo per cápita de electricidad

Es un indicador relacionado con el grado de bienestar que disfruta la población. Refleja de forma clara y directa la situación concreta de ese territorio en relación con el nivel y la calidad de vida. Es un componente que refleja en términos generales los recursos disponibles y en buena medida el nivel adquisitivo de las personas.

Notas

1. Cada una de las 15 secciones del libro (la presentación y los 14 capítulos) está presidida por una cita; a estos pensadores estoy expresando por esa vía mi modesto homenaje. En su mayoría son autores nacidos el siglo pasado cuya obra está marcada por un espíritu humanista de hondo calado. A continuación hago un breve comentario sobre el autor y la fuente de cada una de ellas.

Presentación: Aníbal Ponce (1898-1938) , fallecido trágicamente antes de cumplir los 40 años, fue uno de los grandes pensadores marxistas latinoamericanos. La cita procede de su conferencia ***Los deberes de la inteligencia,*** pronunciada en junio de 1930 en la Universidad de su Buenos Aires natal.

Capítulo 1: La cita del epistemólogo argentino Mario Bunge, nacido en 1919, puede hallarse en su libro ***Seudociencia e ideología.***

Capítulo 2: La fecunda obra del insigne ensayista madrileño José Ortega y Gasset (1883-1955) llegó a mí mucho más tarde de lo que hubiera deseado. La cita elegida resume una de las más interesantes ideas de su libro ***En tomo a Galileo.***

Capítulo 3: Esta cita se halla en el libro ***Ensayos,*** debido a uno de los intelectuales más interesantes de Colombia, el gran humanista antioqueño Baldomero Sanín Cano (1861-1957).

Capítulo 4: El pensamiento pertenece a John Stuart Mill (1806-1873), filósofo y economista inglés cuya obra, lejos de envejecer, me parece clave para encarar los actuales tiempos signados por una crisis de valores. Su ***Lógica deductiva e inductiva*** merece el más detenido estudio.

Capítulo 5: John Allen Paulos, matemático norteamericano quien ha destinado parte de su tiempo y su intelecto a hacernos notar en ***El hombre anumérico*** que el anumerismo es tan triste y embrutecedor para el hombre moderno como el analfabetismo.

Capítulo 6: A la aguda penetración que de la naturaleza de las cosas y las personas tenía Paul Valéry (1871-1945), poeta y ensayista fran-

cés, se deben múltiples aforismos, citados una y otra vez. El que se ha elegido es uno de ellos. Ignoro su origen exacto.

Capítulo 7: Gaston Bachelard (1884-1962) filósofo francés, fue el autor de ***La formación del espíritu científico,*** origen de la cita.

Capítulo 8: José Ingenieros (1877-1925), sociólogo positivista argentino, maestro moral de varias generaciones de nuestra América; la cita se encuentra en su influyente libro ***Las fuerzas morales.***

Capítulo 9: Ivan Illich, sacerdote y ensayista nacido en Austria en 1926, sacudió al mundo sanitario con su célebre ***Némesis médica,*** obra de la cual se extrajo la cita.

Capítulo 10: Ernesto Sábato, notable y polémico novelista argentino nacido en 1911. La cita procede de su ensayo ***Heterodoxia.***

Capítulo ll: Realmente ignoro si Galileo pronunció alguna vez las palabras citadas. En realidad fue el dramaturgo alemán Bertolt Brecht (1898-1956) quien se las atribuye en la obra teatral ***Galileo Galilei.*** Lo importante es que la cita resume el espíritu del pensamiento y la obra del gran italiano.

Capítulo 12: El médico español ***Moisés Ben Maimón*** (Maimónides) (1135-1204), es el autor de esta cita -tomada de sus ***Aforismos de Medicina-*** que refleja cuán antiguas son algunas de las inquietudes que nos aquejan cada día.

Capítulo 13: Bertrand Russell (1872-1970), filósofo y matemático inglés, ganador del Nobel en 1950, autor de ***Una mirada científica,*** de la que procede el texto citado.

Capítulo 14: Al salir de una visita a una feria informática en Madrid en 1994, me descubrí cargando con decenas de revistas, plegables publicitarios y materiales similares: el 90% de ese material era simple contaminación informativa. Del proceso depurador que subsiguió, afortunadamente se salvó un pequeño libro de notable interés, debido a Fernando Sáez Vaca, periodista español contemporáneo, especializado en temas socioinformáticos: ***El hombre y la técnica.*** En el capítulo 14 suscribo varias de las ideas en él contenidas.

2. Este libro bien puede ser acusado de emplear un lenguaje sexista, ya que hace uso del plural masculino omnicomprensivo (los lectores, en lugar de los lectores y lectoras) y de términos con desinencia masculina a pesar de que tal vez pudieran existir alternativas más equitativas. En el primer capítulo, por ejemplo, dice:

> ***Muchos investigadores biomédicos (...) se muestran ansiosos por «aderezan» sus análisis con técnicas estadísticas. Algunos de ellos están persuadidos de que lo ideal sería recurrir a las más intrincadas.***

aunque podría (o quizás, debería) decir:

> ***Muchos/as investigadores/as biomédicos/as (...) se muestran ansiosos/as por «aderezar», sus análisis con técnicas estadísticas. Algunos/as de ellos/as están persuadidos/as de que lo ideal sería recurrir a las más intrincadas.***

Aunque tal solución es una de las que se ha sugerido, tiene en mi opinión algunos inconvenientes: por una parte, la inclusión de tantas barras podría terminar siendo en extremo fatigante para el lector (o lectora), más de lo que el libro propiamente dicho pudiera resultar; por otra parte, la solución es discutible, pues en tal caso siempre se estaría colocando la alternativa masculina en primer lugar; por otra parte, la variante de invertir el orden ***(Muchas/os investigadoras/es biomédicas/os . ..)*** no mejoraría las cosas.

La variante de usar el símbolo @ como vocal ambivalente (investigadores para aludir a investigadores e investigadoras) no me convence porque cancela la posibilidad de usar el singular (el investigador y la investigadora) y porque uno tendría la impresión de que el libro está lleno de direcciones electrónicas.

Otras alternativas, tales como elegir al azar cada vez cuál de los dos géneros iría en primer lugar, o cuál de ellos habría de usarse, me parecen absurdas. En vista de mi incapacidad para resolver satisfactoriamente el problema, opté por hacer uso de los recursos sintácticos convencionales, sin que ello traduzca la menor intención discriminatoria para ninguno de los dos sexos.

3. Varios colegas tuvieron la gentileza de dar una lectura crítica a algunos segmentos del libro y hacerme notar parte de sus insuficiencias. Dejo constancia de mi agradecimiento a todos ellos: Humberto Fariñas Seijas, Rosa Jiménez Paneque, José Tapia Granados, Antonio Pareja Bejares, Armando Seuc Jo, Frank Alvarez Li, Carlos Campillo Artero y Javier Nieto.

4. Se atribuye al ilustre matemático francés Jean D'Alambert la afirmación de que «sin una mujer, no podría siquiera enunciar las propiedades del triángulo escaleno». Eso me hace recordar a Regla, Diana, Odalys, Andrea y Zeida. A ellas les debo la paciencia, el café, la laboriosidad y, sobre todo, el afecto cotidiano, elementos presentes en cada página de este libro.

5. Agradezco especialmente a Joaquín Vioque y Antonio Vila, editores de «Díaz de Santos», por su amistad, su apoyo y su confianza irrestrictos.

Indice de autores

Indice de materias