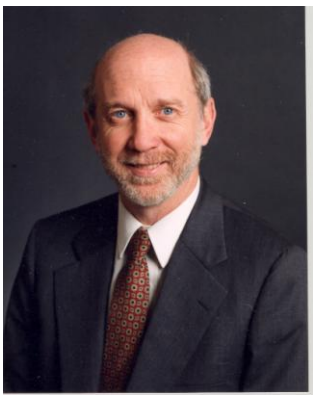
MEDICIÓN, MUESTREO Y RELACIONES CAUSALES[1]

Este capítulo está destinado al juez o abogado no experto en cuestiones científicas a fin de proporcionar una idea accesible de la manera en que los científicos adquieren conocimiento acerca de cómo funciona el mundo. En nuestro país no existe una tradición legal en tal sentido, aplicando la lógica al problema de cómo debe ser observado un fenómeno empírico de modo de permitir obtener conclusiones válidas sobre el mismo. Como ya hemos señalado, el derecho no es otra cosa que una rama particular de aplicación de la lógica y, en sus aspectos empíricos, de las ciencias sociales.

Ahora bien, si la temática del método científico es fundamento necesario de toda disciplina que trate en forma seria de obtener conocimiento sobre el mundo mediante la investigación empírica sistemática – de la misma forma que las ciencias físicas, biológicas, de conducta y sociales – y ha sido descripto como tal en innumerables manuales escolares hasta tratados de postgrado, ¿por qué sigue siendo un objeto extraño para la mayoría de abogados y jueces? Se trata de un tema en el cual son casi analfabetos. Por tal motivo, es necesario analizar el método de las ciencias en forma simple y directa, como introducción a un tratamiento más elaborado.

1) Cómo la ciencia halla respuestas a los problemas


Daniel L. Rubinfeld

¿Cómo se halla una respuesta científica a una cuestión dada? Ejemplos: ¿Qué tratamientos diversos para el cáncer o la inflación son mejores? ¿Protege la vitamina C contra el resfrío, una técnica de cirugía es mejor que otra, qué métodos de enseñanza o de cultivo son mejores? O, pasando a cuestiones menos "aplicadas" y más "básicas": ¿Cuál es la naturaleza del movimiento? ¿Qué aspectos están incluidos en los rasgos adquiridos? ¿Cómo funciona la memoria?

Dios no susurra la respuesta a los científicos, como a miembros de un sacerdocio moderno. La única forma en que un científico puede obtener una respuesta a una cuestión empírica es realizando una investigación empírica, lo cual implica la observación del fenómeno en que uno está interesado, aunque habitualmente de modo disciplinado.

Tratando de responder a la cuestión planteada, supongamos que alguien hace la pregunta ¿cuántos dientes tiene un caballo? Un filósofo platónico hallaría la respuesta mediante deducción o debate acerca de cuántos dientes debería tener, en tanto que el instinto inmediato de un científico moderno es buscar la respuesta en la boca del caballo. Pero este enfoque puede ser defectuoso, ya que el caballo podría tener un número distinto de dientes, por ejemplo por heridas; o el número podría cambiar a medida que cambia la edad; o depender de distintas cruzas o sexos, con lo cual lo más probable es que reporte el promedio y el rango de variabilidad de subgrupos de

[1] Ver David H. Kaye & David A. Freedman, Reference Guide on Statistics, in Reference Manual on Scientific Evidence, 2nd ed., Federal Judicial Center (2000), http://ebour.com.ar/index.php?option=com_weblinks&task=view&id=13485&Itemid=0 pp. 83-178; Daniel L. Rubinfeld, Reference Guide on Multiple Regression, in Reference Manual on Scientific Evidence, 2nd ed., Federal Judicial Center (2000), pp. 179-227; David L. Faigman, Michael J. Saks, Joseph Sanders, and Edward K. Cheng, Modern Scientific Evidence: The Law and Science of Expert Testimony, Vol. 1 (2nd ed., 2002, West Publishing) http://www.judges.org/pdf/ev-webinar/Mod.%20Two%20--%20Faigman%20--%20Mod.%20Sci.%20Ev.%20Ch.%201.pdf; Edward J. Imwinkelried, A New Era in the Evolution of Scientific Evidence - A Primer on Evaluating the Weight of Scientific Evidence, 23 Wm. & Mary L. Rev. 261 (1981), http://scholarship.law.wm.edu/wmlr/vol23/iss2/4

caballos. Éste es un ejemplo de que aún la pregunta más simple requiere ser abordada de manera sistemática y razonada a efectos de evitar respuestas incorrectas. Hay cuestiones que sólo pueden ser abordadas usando métodos disciplinados y especializados, comparando cosas bajo distintas condiciones. Por ejemplo, hace un siglo los cirujanos creían que el mejor tratamiento del cáncer de mamas era extirpar toda la mama y un tejido adicional considerable de alrededor, para lograr reducir las células cancerosas extirpándolas de raíz (así lo estipulaba la teoría). La sugerencia de que podría existir una cirugía menos destructiva tan apropiada como aquella era enfrentada mediante la defensa basada en la fe, más que en la evidencia, de que sólo las *mastectomías* radicales facilitan más protección contra la difusión del cáncer. A menudo la mastectomía se realizaba durante la misma operación en la que se hacía la biopsia para confirmar el diagnóstico. Hoy, la decisión de hacer una mastectomía se basa usualmente en una biopsia previa. También hay una tendencia a un tratamiento más conservativo con el cáncer de seno. La práctica ha cambiado, por una parte, debido a las mejoras en radioterapia y tratamiento coadyuvante (quimioterapia y terapia hormonal) y por otra parte en un reconocimiento más temprano de la metástasis del cáncer. En contraste, en una *lumpectomía*, sólo una porción de tejido es extirpada, lo que vino dándose a partir de los 1970s. Se hizo una muestra aleatoria de pacientes de cáncer de mamas y a algunos se les practicó la mastectomía tradicional, mientras que los restantes recibieron el tratamiento más conservador. Éste demostró que era tan bueno como el primero en detener al cáncer. Todo ello ilustra el valor de una comparación experimental, en la que el valor de un tratamiento es comparado con el de otro. La escisión radical no impedirá tumores secundarios posteriores que ocurran como resultado de micro-metástasis ocurridas antes de descubrir el cáncer. En los países más desarrollados sólo una minoría de los nuevos casos de cáncer de mama son tratados con mastectomía.

Para un científico en serio un hecho de la vida real sólo es tan bueno como los métodos empleados para hallarlo. El método científico es la lógica mediante la cual se practican las observaciones. Un método bien diseñado permite obtener observaciones que conducen a respuestas válidas, útiles e informativas. Para un científico, la palabra clave de “método científico” es *método*. La metodología – que es la lógica del diseño de la investigación, de las medidas y de los procedimientos – es la máquina que genera conocimiento científico. Mientras que para los abogados o jueces lo que resulta clave es la credibilidad para hacerse una idea si los testigos están diciendo la verdad o mintiendo, para los científicos la clave es tener una idea de cuál de diversos estudios contradictorios sea más probablemente correcto analizando su metodología.

Tengan en cuenta que, sin embargo, hay muchos científicos que no usan el método científico. O sea, no han sometido sus creencias a una dócima empírica sistemática.[2] Sus creencias están basadas en información casuística, o en la intuición, fe o autoridad de generaciones pasadas en su campo que ejercían *su* intuición. Al dirigir la atención de los tribunales federales desde la opinión consensuada hacia la ciencia subyacente, hubo un caso famoso en US, Daubert v. Merrell Dow Pharmaceuticals, que implicó que muchas áreas tuvieran que ser reexaminadas.

[2] Una dócima estadística es un método para tomar decisiones estadísticas por medio de datos, ya sea a partir de un experimento controlado o de un estudio de observaciones (no controlado). El concepto de “test de significación” fue acuñado por Ronald Fisher: “Tests críticos de este tipo pueden ser llamados dócimas de significación, y si están disponibles estos tests se puede descubrir si una segunda muestra es diferente en términos significativos de la primera.” (R. A. Fisher (1925). Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd, 1925, p.43). Un uso habitual de las dócimas es decidir si los resultados experimentales contienen suficiente información como para arrojar dudas sobre la sabiduría convencional. http://en.wikipedia.org/wiki/Statistical_hypothesis_testing Recomiendo leer la introducción de Ronald Fisher a este libro, en http://www-history.mcs.st-and.ac.uk/Extras/Fisher_Statistical_Methods.html

Daubert v. Merrell Dow Pharmaceuticals[3] Entre las múltiples discusiones a que dio lugar este caso, la Corte Suprema hizo énfasis en la Regla Federal de Evidencia 702[4] a fin de subrayar que el "testimonio de un experto debe ser *conocimiento científico*". La Corte explicó que *científico* implica tener basamento en los métodos y procedimientos de la ciencia, en tanto que *conocimiento* denota algo más que una creencia subjetiva o especulativa. Al discutir la naturaleza del conocimiento científico, la Corte citó la definición del diccionario Webster y de dos testimonios (*amicus curiae*) – uno de un grupo de científicos, otro de autoría de la American Association for the Advancement of Science y de la National Academy of Science – que advertían que la "ciencia no es un cuerpo de conocimiento enciclopédico sobre el universo, sino un proceso de proponer y refinar explicaciones teóricas sobre el mundo, sujetas a ser refinadas y docimadas". La Corte no definió en forma explícita lo que consideraba ciencia, ni tampoco diseñó un test o listado que debería ser cumplido a tal efecto. En su lugar, ofreció ciertas observaciones generales que los jueces de primera instancia deberían tener in mente al evaluar la evidencia de los expertos. En lugar de pretender que la aceptación general fuera el único estándar – como en el caso Frye – los jueces de la época post-Daubert deben evaluar los méritos de los testimonios con respecto a cuatro pautas generales: (1) *falsación*, o sea si la teoría o la técnica pueden ser (y han sido) docimados; (2) *el error conocido o potencial* asociado con una técnica científica particular, y la existencia y mantenimiento de estándares que controlan la técnica operativa; (3) si la teoría o técnica fueron sujetas a su *revisión por los pares y su publicación*; y (4) la *aceptación general* del testimonio propuesto dentro de la comunidad científica. A partir de la decisión Daubert, los tribunales de US aplicaron estos criterios de "validez científica" para chequear una amplia gama de testimonios, aunque con gran variabilidad con relación a campos como la psicología y otras ciencias "blandas". El documento de Gatowski y otros (2001) presenta evidencia sobre los criterios de Daubert, su utilidad como pautas para tomar decisiones, y su aplicación en diferentes campos que requieren el conocimiento de expertos.[5]

A veces hay una región de conocimiento científico genuino dentro de un campo, pero alguno(s) de sus miembros sale(n) de esa región y efectúa(n) afirmaciones que exceden el conocimiento empíricamente probado en ese campo. O, alternativamente, responden a preguntas basadas en parte en conocimiento bien docimado y en parte en especulación. Antes de que se realizaran investigaciones comparadas de la eficacia relativa de las estrategias quirúrgicas de cáncer de senos, los cirujanos que hacían mastectomías radicales no aplicaban conocimiento científico sobre la efectividad de las mismas, simplemente porque ese conocimiento no existía. Pero con relación a otras técnicas quirúrgicas, pudieron estar aplicando conocimiento testeado por métodos científicos. Una revisión de los tests de las técnicas quirúrgicas más populares halló que alrededor de una tercera parte eran de hecho efectivas, otra tercera parte inútiles, y la tercera parte restante producían más daño que otras técnicas alternativas que estaban disponibles. *Luego, no ayuda tanto preguntarse si la persona que asevera algo es "un científico" (por ejemplo, un investigador en química o en psicología) o si "está aplicando ciencia" (como lo haría un investigador químico o*

[3] http://en.wikipedia.org/wiki/Daubert_v._Merrell_Dow_Pharmaceuticals. El texto del caso puede consultarse en http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=509&invol=579

[4] "Si el conocimiento científico, técnico, o especializado ayuda al juez a entender la evidencia o a determinar un hecho en discusión, un testigo calificado como experto por su conocimiento, capacidad, experiencia, entrenamiento, o educación, podrá brindar su testimonio al efecto bajo la forma de una opinión." Luego, un testigo podrá testificar como experto si se reúnen tres condiciones: (1) tiene un conocimiento diferenciado, (2) que será de auxilio al jurado, y (3) el testigo es un experto calificado. Véase http://www.law.cornell.edu/rules/fre/rule_702

[5] Sophia Gatowski, Shirley A. Dobbin, James T. Richardson, Gerald P. Ginsburg, Mara L. Merlino, and Veronica Dahir, Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-Daubert World, Law and Human Behavior, Vol. 25, No. 5, Oct. 2001 http://www.toxicologysource.com/law/daubert/Resources/Asking%20the%20Gatekeepers_%20.pdf

un psicólogo clínico), sino que es más útil preguntarse qué tan bien basada está la afirmación de que los estudios empíricos que lleva a cabo están bien diseñados y practicados a fin de poner a prueba las proposiciones contenidas en el aserto.

Cuestiones empíricas y normativas En el capítulo I (Parte 1) he descripto en forma sencilla la distinción que cabe hacer entre *análisis positivo, análisis normativo* y *análisis normativo condicional.* En el fondo se trata de distinguir entre dos mundos intelectuales (el mundo del "ser" y el mundo del "debe ser") y la responsabilidad de conducirse con ellos, que corresponde a distintas profesiones. Los científicos, ingenieros y terapeutas se ubican en el primero; filósofos, teólogos, y escritores literarios suelen manejarse con el mundo del "deber ser". Pero, como indica la existencia de un análisis normativo condicional, las distinciones no son tajantes y pueden existir conexiones importantes entre ambos.

A veces se toman decisiones basándose en cuestiones empíricas cuyas respuestas han sido presupuestas. Un abogado tiene un objetivo, como por ejemplo "queremos salvar vidas cerrando la brecha entre la oferta y la demanda de órganos de trasplante". Que ello se pueda lograr y cómo son cuestiones empíricas. Los métodos elegidos pueden causar efectos contrarios a los buscados – algo que sucede en la práctica. En tal situación, una organización de donación de órganos puede cuestionar la ley ante los tribunales porque fracasa en cumplir con una prueba de relación racional. Los abogados pueden desear limitar la disponibilidad de pornografía porque creen que causa daños. Si lo hace, si no tiene efectos, o si hace bien es una cuestión empírica abierta. *Las decisiones normativas no pueden usualmente liberarse de un conocimiento apropiado acerca de cómo funciona el mundo.*

Todo un tema de discusión es el relativo a los efectos de la pornografía. Dejando de lado el tema de la *pornografía infantil*, que ha dado lugar a trabajos de diversos organismos internacionales en pro de limitarla (por ejemplo, del International Centre for Missing and Exploited Children),[6] existen opiniones encontradas sobre el tema, por ejemplo la de Ligia Vera-Gamboa: "Quizá ésta sea una pregunta que muchos de nosotros nos hemos hecho alguna vez, sin embargo es conveniente aclarar lo que se entiende por pornografía. Etimológicamente, esta palabra deriva del griego *pornoi*: prostituta y *grafos*: tratado; entonces pornografía sería el tratado de las prostitutas o de la prostitución, lo que se confirma al revisar la definición del diccionario Larousse, el cual, además añade: "carácter obsceno de obras literarias o artísticas". Puede observarse que este término es impreciso, aunque la connotación que comúnmente se le atribuye es la de "aquello que molesta o

[6] International Centre for Missing & Exploited Children, Pornografía Infantil: Modelo de Legislación & Revisión Global, 2008, Quinta edición. Este documento afirma que "La vida de los niños explotados mediante la pornografía infantil queda afectada para siempre, no sólo como consecuencia de los abusos sufridos por esos niños, sino también por el registro permanente que queda como consecuencia de esa explotación. Una vez que la explotación sexual tiene lugar, el perpetrador de esa violación sexual puede documentar esas violaciones en película o en video. Dicha documentación puede convertirse entonces en una permanente "amenaza" para chantajear al niño por el resto de su vida y obligarlo así a someterse a continuar con ese tipo de relación y mantenerla en secreto. Más aún, esas imágenes documentadas también hacen posible que los violadores de niños "vuelvan a disfrutar interminablemente" sus fantasías sexuales. Cada vez existe una mayor cantidad de violadores de niños que utilizan la tecnología de computación para organizar, mantener y aumentar el tamaño de sus colecciones de pornografía infantil. Las imágenes ilegales de niños, personalmente manufacturadas, tienen un valor especial en la Internet y a menudo dichos violadores se dedican a intercambiar imágenes de sus propias "proezas sexuales". Cuando esas imágenes llegan al espacio cibernético, ya es muy tarde para retraerlas y pueden continuar circulando para siempre, condenando así al niño a ser víctima perenne de imágenes que pueden ser vistas una y otra vez." http://polis.osce.org/library/f/3648/2804/GOV-USA-RPT-3648-ES-Pornografía%20infantil_%20Modelo%20de%20legislación%20y%20revisión%20global%20(Quinta%20edición).pdf

disgusta a una persona". Sin embargo, es obvio que existe dificultad para interpretar la definición, ya que no todas las personas podemos estar de acuerdo para definir lo que es comercial, artístico o pornográfico, algo similar a lo que ocurre con el concepto de belleza. Esta dependerá de cada persona, y aún más dependerá del contexto socio histórico en el que se desarrolle. Por ejemplo, un hecho conocido es que Hamlet produjo escándalo en su época y no ahora y por el contrario algunas obras de Aristófanes fueron vistas con naturalidad en su época y en la actualidad han causado controversia... Los materiales sexualmente explícitos presentan algunas ventajas como permitir a algunas personas enriquecer su vida sexual y desde esta óptica aceptar y respetar que existen aspectos de la sexualidad humana diferentes a los personales. También se han reportado algunas desventajas como el hecho que han reducido a la mujer a un objeto de placer y la reducción de las relaciones sexuales a sólo un acto físico ajeno al contexto de una relación. Asimismo, reduce nuestra corporalidad y sexualidad a la genitalidad y finalmente lleva a la creación de estereotipos. Es frecuente que los adolescentes en esta etapa de la vida busquen este tipo de materiales en la búsqueda de satisfacer la curiosidad sexual propia de esta etapa, curiosidad nacida de la ignorancia sobre sexualidad, por lo que algunos autores mencionan que este tipo de materiales pueden producir en niños y jóvenes una imagen distorsionada de la sexualidad, especialmente los materiales de porno-dura y/o violenta. Por esto, es fundamental mantener una línea de comunicación abierta entre padres e hijos, maestros y alumnos, adolescentes y adultos, que incluya también la sexualidad (incluidos los materiales gráficos sexualmente explícitos) para contrarrestar esta posible situación... Así, con la información conocida hasta hoy, no existen argumentos científicos que avalen que estos materiales sean dañinos o nocivos para una persona y/o sociedad; sin embargo, es de reconocer que si a algunas personas no les agrada este tipo de materiales esto es válido y debe de ser respetado. La presente comunicación no tiene como propósito defender la "pornografía", sino sólo presentar el estado del conocimiento científico de los escasos estudios existentes (dada la dificultad para realizarlos) sobre los efectos de las representaciones gráficas de la sexualidad y resaltar la necesidad de una educación sexual formal que nos permita, como individuos, evitar prejuicios y derribar mitos y tabúes sobre nuestra propia sexualidad y la de otros."[7]

La otra cara de la moneda es que los supuestos normativos pueden orientar, o bien desorientar la investigación empírica. Si, por ejemplo, realizamos un test de dos tratamientos médicos para ver cuál funciona "mejor", los investigadores probablemente medirán por cuánto tiempo adicional viven quienes recibieron los distintos tratamientos. Se infiere – incorrectamente – que aquel tratamiento que extienda la vida en mayor medida es superior.[8]

Contextos de la investigación Una investigación puede tener lugar en diversos contextos: en el laboratorio, como trabajo de campo, o mediante una simulación o modelos de distintos tipos. No hay uno que sea el mejor para estudiar un fenómeno interesante, porque al elegir hay compromisos. Como siempre, la cuestión es saber si las circunstancias de la investigación son apropiadas para el objetivo del estudio, y si las conclusiones que se extraen son prudentes a la luz de los datos recogidos.

En un *laboratorio*, la investigación tiene la ventaja de ejercer un máximo control sobre las influencias extrañas que afectan al fenómeno estudiado. También es más conveniente y de menor costo que otras alternativas. La investigación de laboratorio presupone a veces una simulación del fenómeno en el que uno está interesado. Este tipo de estudios tiene la desventaja de que puede

[7] Ligia Vera-Gamboa, La pornografía y sus efectos: ¿Es nociva la pornografía?, Rev. Biomed. 2000; 11:77-79. http://www.revbiomed.uady.mx/pdf/rb001119.pdf

[8] Un análisis y crítica de este argumento puede apreciarse en Enrique A. Bour, Tratado de Microeconomía, 2009, Capítulo XXX, pág. 968-971.

implicar una instancia más artificial del fenómeno interesante, y de que por lo tanto puede ser menos generalizable a la situación en que el investigador – y el derecho – están interesados. Más aún, los estudios de laboratorio de los fenómenos biológicos pueden ser realizados *in vitro* (por ejemplo en un tubo de ensayo) o *in vivo* (en un organismo viviente, como un animal de laboratorio). Aunque en ambos casos se trata de investigaciones en laboratorio, la diferencia existente entre estudios in vitro e in vivo es, naturalmente, muy grande.

La *investigación de campo* es una solución al problema de asegurarse que el fenómeno estudiado se comporta en forma similar a las versiones naturales a las que el investigador desearía generalizar sus hallazgos. Ha sido desarrollada originariamente en antropología y a veces es conocida como investigación *del participante*, o etnografía en antropología. Ha sido adaptada también al mundo comercial e industrial, con referencia a la tarea de coleccionar o crear nueva información fuera del laboratorio. También es conocida como *trabajo de campo*, un término originado en granjas y plantaciones, y que a veces se usa para las fortificaciones temporarias construidas antes de una batalla. Un problema de la investigación de campo es que es más difícil aislar la variación de influencias extrañas al fenómeno, enmascarando así los efectos estudiados.[9]

Las *simulaciones* de un fenómeno pueden conducir a una investigación más eficiente, pero también más dudosa, según que la simulación capte o no las características esenciales del fenómeno. Puede hacerse una simulación mediante modelos computarizados (representaciones matemáticas del fenómeno en el que uno está interesado), modelos animales (estudiar algo en los animales antes de hacerlo en seres humanos), elaborar un modelo físico de un proceso u objeto a estudiar (p.ej., testear las características de un nuevo diseño de aeronave), simular situaciones sociales (p.ej. cómo reacciona la gente en situaciones de emergencia), juegos (como en ciencias políticas, comerciales, o estudios económicos de la interacción humana), y así sucesivamente.

Finalmente, están las *investigaciones de encuesta* realizadas por encuestadores o mediante cuestionarios pidiendo a la gente que responda a preguntas sobre su propia conducta pasada (p.ej. cuánto adquirieron de un producto particular), o futura (p.ej. si cierto comercio es puesto en marcha, ¿usted nos visitará?), o sobre sus actitudes y creencias (p.ej. si piensan que el acusado ante los tribunales de su comunidad es culpable o no), o sobre sus reacciones a algo que les presentan al encuestarlos (p.ej., qué piensan de ciertas manufacturas de productos que conocen).

Estos estudios le proporcionan al investigador muchos beneficios. Es más simple preguntarle a alguien con qué frecuencia maneja intoxicado su auto que hacerlo persiguiéndolo y observando en forma directa su conducta. Pero este ejemplo nos indica el precio que a veces se paga: una encuesta más fácil puede disminuir la precisión de los datos. La gente puede responder en forma mentirosa con el fin de ser socialmente más aceptada. O también, sus memorias, por sinceras que sean, pueden ser imperfectas. Desplazándonos en el párrafo precedente desde arriba hacia abajo nos indica la clase de preguntas que pueden dar lugar a información más imprecisa (el recuerdo del pasado y la predicción de la conducta en el futuro) hasta las que son probablemente las respuestas más confiables (mostrarle algo y hacerle una pregunta relativamente inocua al respecto). Por consiguiente, es necesario escrutar el contenido de los requerimientos de información.

[9] Una revisión de investigaciones comparativas de estudios de cambio de actitudes en contextos de laboratorio versus estudios de campo halló que al responder a la misma cuestión investigada, ambos contextos tienden a diferenciarse en que a menudo un fenómeno detectado en el laboratorio (con un control elevado) es incapaz de ser detectado en el trabajo de campo (con bajo control).

Como sabemos, incluidos abogados y jueces, la forma en que está estructurada una pregunta, la elección de las palabras y otras características pueden imprimir un sesgo a las respuestas que se obtienen. Por ejemplo, la gente proporciona distintas respuestas cuando se pregunta sobre los derechos de los “no-natos” versus los “productos de la concepción”.

2) Definir y medir

Esto de definir lo que se observa puede parecer algo terriblemente básico, pero es la primera oportunidad en que la investigación puede fracasar estrepitosamente.

Definiciones Conceptuales y Operativas Las primeras son enunciados abstractos sobre los fenómenos de interés. Las operativas son procedimientos concretos a llevar a cabo a fin de observar las cosas que son discutidas. Hablar de los conceptos en abstracto es una cosa (p.ej. agresividad, inteligencia, decisiones razonables, discapacidad). Lo que de lejos es más difícil es definir en forma precisa qué observaciones deben ser tenidas en cuenta como instancias del concepto y cuáles no a los fines de la investigación. Lo que es más, el mundo puede aparecer como un lugar muy distinto usando distintas definiciones de algo.

Ejemplos Se puso en marcha un estudio para hallar cuánta violencia ocurría dentro de cada familia en un escenario y la encontró epidémica. En contraste, otro estudio concluyó que la violencia familiar era bastante poco frecuente. Por definición conceptual, ambos estudios estaban interesados en lo mismo. Su discrepancia puede ser comprendida examinando sus definiciones operativas respectivas. El primer estudio definió violencia familiar como incluyendo todo lo que pasara de un grito en voz alta. El segundo consideró que violencia sólo incluía situaciones que requieren hospitalización.

El laboratorio de un remedio contra el resfrío trató una vez de convencer a los consumidores de que “no todas las aspirinas son la misma cosa” enfatizando que ese laboratorio era “mejor” que la competencia y exhibiendo un comunicado en la pantalla con la dirección a la que los televidentes podrían enviar un informe sobre los detalles que confirmaban. El informe reveló que el producto era superior en las siguientes dimensiones: claridad de la etiqueta, precisión del número de píldoras en un envase de 100 unidades, adherencia de la cola de la etiqueta, número de píldoras rotas en el envase, etc. Como la definición operativa de la propaganda de “mejor aspirina” incluía este tipo de características, los consumidores potenciales hicieron el supuesto de que el fabricante les estaba informando sobre la eficacia farmacológica de la aspirina.

Hace unos años, una administración presidencial buscó mejorar su tratamiento de los derechos civiles de un día al otro redefiniendo un “distrito escolar desagregado” como el que tiene al menos una escuela por distrito (antes era la mitad de las escuelas en un distrito). En forma similar, los índices de crimen, costo de vida, crecimiento económico, capacidad educativa, corrupción y casi cualquier índice pueden parecer que experimentan un cambio por el mero cambio de la definición operativa utilizada, aunque el mundo no haya cambiado un ápice.

A veces los cambios de las definiciones operativas pueden ser muy sutiles, como en el caso siguiente. Unos investigadores habían concluido que la progresividad aparente del cáncer es ilusoria, porque los aumentos de las tasas de supervivencia a 5 años no reflejaban más que mejoras en una detección más temprana. Esto es, haciendo arrancar el reloj más temprano, hay más gente que llega al punto cinco, aunque la historia natural de su enfermedad no cambie y muera la misma cantidad de gente al mismo tiempo. Este ejemplo puntualiza un cambio que no se ha notado en la definición operativa de cuándo “comienza” el cáncer, que altera la definición de si uno “sobrevivió” o no.

Medir Una vez aceptada una definición operativa, el objeto observado debe ser medido. Hay cuatro escalas de medición que pueden ser utilizadas. Su importancia radica en que los procedimientos estadísticos a utilizar dependen del caso. Un resultado válido requiere la utilización de una escala de medida apropiada.

1. La medición de la escala *nominal* implica solamente "nombrar", colocando a menudo al objeto dentro de una categoría u otra. El color del cabello, el género, el diagnóstico de una enfermedad, los números sobre las remeras de los atletas, los números telefónicos son ejemplos de datos medidos en escala nominal. Con este tipo de medición, a lo sumo se puede contar el número de objetos dentro de una categoría, utilizar como medida de tendencia central el *modo* (es decir, la categoría que ocurre con mayor frecuencia), y emplear instrumentos estadísticos que operan con datos categóricos[10] con los cuales se pueden emplear herramientas estadísticas como los tests de χ^2 (léase "chi-cuadrado").

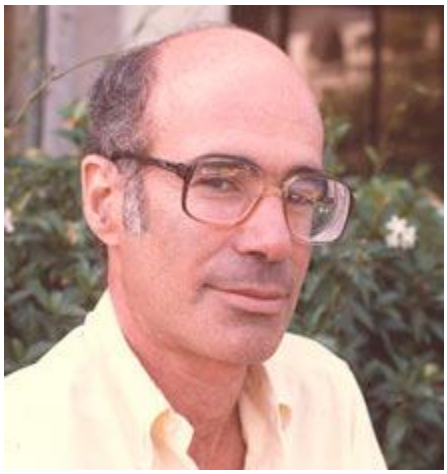
2. Las escalas de medición *ordinal* agregan un orden al nombre, es decir, nociones de mayor o menor en cierta dimensión. *Una escala ordinal implica un ranking*. Los corredores de carreras terminan primero, segundo, etc. Sabemos que el que terminó primero fue más veloz que el segundo, pero no podemos decir a partir del ranking cuánto más veloz fue el primero que los demás. La diferencia entre el primero y el segundo pudo haber sido de una hora, mientras que entre el segundo y el tercero pudo ser de sólo un minuto. Con medidas ordinales se pueden calcular medianas (es decir, el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor) y usar instrumentos estadísticos como la correlación de rangos, además de los usados en las mediciones de escala nominal.

3. Las escalas de medición por *intervalos* añaden cantidad al orden. Las mediciones por intervalos implican tasación en lugar de ranking. Nos dicen que las distancias entre las puntuaciones consisten, p.ej., de intervalos iguales. Una persona con un nivel de CI de 130 es tan inteligente con relación a una persona de 120 como otra de 60 lo es a otra de 50. Las escalas de actitud (método de medir la actitud basado en el supuesto de que poseer una actitud conduce a respuestas consistentes ante personas, objetos o ideas particulares), los índices de gravedad de las enfermedades, y los índices de rendimiento de los empleados son otros ejemplos. Pero estas escalas carecen de un verdadero punto 0, de modo que no se puede afirmar que una persona cuyo rendimiento es 100 se desempeñe dos veces mejor que alguien con un índice de 50. Un enunciado tal distorsionaría lo que es posible con escalas de medición por intervalos. Con la medición por intervalos se pueden calcular medias, desvíos estándar, correlaciones de Pearson (p.ej. la correlación positiva entre la estatura física de padres e hijos, y la correlación –usualmente negativa -- entre la cantidad demandada de un producto y su precio), y usar tests de significatividad estadística como t-tests y F-tests. También se puede aplicar cualquier técnica estadística permitida para datos nominales u ordinales.

[10] Medir en una escala categórica consiste en observar el resultado de un experimento y asignarle una clase o categoría, de entre un número finito de clases posibles. Esta escala es no numérica, y puede ser categórica ordinal, es decir, sus categorías tienen un orden natural, o en caso contrario la escala es categórica nominal. P.ej. la raza, el sexo, el grupo etáreo y el nivel educativo. Aunque las dos últimas variables también pueden ser consideradas usando valores exactos para la edad y el mayor nivel educativo alcanzado, a menudo resulta más informativo ubicarlas en un relativamente pequeño número de grupos.

4. La medición por *razones* implica una escala a la que se añade un verdadero origen cero a lo anterior que permite efectuar enunciados de razones. Con escalas por razones es posible decir, por ejemplo, que una persona que corre 1 km en 3 minutos es doblemente más rápida que otra que lo hace en 6 minutos. Las cantidades como el peso, el costo, la distancia, y el tiempo son ejemplos de mediciones en escala por razones.

El gobernador de una provincia argumentó una vez que había mejorado el gasto educativo, poniendo como evidencia el hecho de que, en comparación con el año anterior, el gasto educativo había aumentado considerablemente. Un dirigente opositor dijo que la provincia, en realidad, había hecho un mal manejo del gasto educativo, ante la evidencia de que su rango entre todas las provincias no había crecido, sino disminuido. Es muy probable que los dos tuvieran datos correctos, pero estaban usando distintas escalas de medición; luego, podían llegar a conclusiones diferentes (la medida de la razón de los gastos absolutos puede aumentar al mismo tiempo que disminuye el rango relativo).



David A. Freedman (1938–2008)

En un comité, el director pidió a sus miembros que ordenaran las preferencias por determinadas opciones de política que enfrentaban. Luego el director computó el rango medio de cada opción. Pero los datos ordenados no pueden ser utilizados de forma válida para calcular medias, porque el proceso matemático empleado requiere que los datos sean al menos intervalos. Cuando el tema fue analizado en forma apropiada, la mediana (la puntuación de la mitad) de los datos ordenados, reflejó un conjunto de preferencias entre los miembros del comité distinto que la otra técnica, incorrecta, de promediar puntuaciones.

Fiabilidad y Validez Ésta es una distinción habitual en materia científica. En Daubert v. Merrell Dow, el juez Blackmun tuvo problemas en rechazar esta diferencia para la ley de la evidencia, combinando tanto la fiabilidad y la validez en lo que denominó, conjuntamente con otros jueces y abogados, la fiabilidad de la evidencia. Éste es un problema que va más allá de lo semántico, y tal vez sería conveniente usar tres términos para referirse a los conceptos. Por ejemplo, podría decirse que las mediciones deben ser “confiables” o “fidedignas” antes de que se deposite demasiada confianza en ellas, siendo componentes principales de este carácter fidedigno su carácter repetitivo (fiabilidad) y su exactitud (validez). Ambas se relacionan con la bondad de la medición, y nos dicen cosas diferentes acerca de la misma.

Para un científico o un estadístico, *fiabilidad* se refiere a la capacidad de una medición de dar el mismo resultado cada vez que se aplica a una misma cosa. La fiabilidad implica consistencia o carácter reproducible. Si cada vez que una persona se sube a una balanza en el baño lee algo diferente (si su peso no cambia) en tal caso la balanza carece de fiabilidad. Ésta se puede poner a prueba, por ejemplo, con 50 personas que pesan distinto subiéndose dos veces cada una, y luego comparando los 50 pares de lecturas. Si los dos conjuntos de lecturas guardan una elevada correlación positiva (esto es, la primera lectura de una persona es altamente predictiva de la segunda lectura), en tal caso se puede decir que la balanza proporciona una medición fiable. Otro ejemplo: un método de identificación de ADN requiere que el laboratorio determine la longitud de los fragmentos de ADN. Haciendo mediciones duplicadas de fragmentos de ADN, el laboratorio puede determinar la probabilidad de que dos mediciones difieran en una cantidad específica. Pero ser fiable es condición necesaria, aunque no suficiente para que se tenga una buena medición. Una medición podría ser fiable sin ser válida. Supongan que a alguien se le ocurre usar la balanza

del baño para medir la inteligencia. Aunque la balanza fuera perfectamente fiable no guardaría correlación con lo que llamamos inteligencia. La medición de la inteligencia mediante la balanza del baño tendría fiabilidad perfecta pero carecería de validez. La *validez* es, por lo tanto, hasta qué punto algo mide lo que intenta medir. Una medición no puede tener más validez que lo que tiene de fiable (p.ej., si la aguja de la balanza del baño la lectura rebota en círculo -baja fiabilidad- nunca podremos saber qué lectura es la correcta –baja validez).

La fiabilidad y la validez de las mediciones generadas por el juicio humano subjetivo, de la misma forma que las producidas por instrumentos mecánicos, tests de laboratorio, y tests de computación, pueden ser evaluadas mediante estudios apropiados. Por ejemplo, hubo estudios de la fiabilidad (y a veces de la validez) del juicio clínico de psicólogos, maestros, jueces, jurados, radiólogos, operadores de sonar, etc. Hay interesantes ejemplos de una completa divergencia de fiabilidad y validez, por ejemplo en la identificación de los textos escritos a mano (un estudio en el que todos los examinadores de un documento llegaron a la misma respuesta a un problema (fiabilidad perfecta) pero donde todos estaban equivocados (validez cero).

Rol de las Variables Las variables desempeñan distintos roles en un estudio, según la naturaleza del estudio y las preguntas planteadas por la investigación. Algunos estudios están diseñados simplemente para medir y describir algo, pero no para explicarlo o predecirlo. Cuando la cuestión planteada por la investigación es de causa y efecto, la variable causa es denominada *variable independiente* y la que resulta como efecto que responde a la causa la *variable dependiente*. Hay veces en que esta relación es muy compleja, con otras variables intermedias entre la causa y el efecto, que son llamadas *variables intervinientes*. Las variables extrañas que de por sí influyen sistemáticamente sobre la variable dependiente, creando la ilusión de una relación causa-efecto entre una variable independiente y la variable dependiente, son a veces llamadas *variables de confusión o confusivas*. En un estudio meramente predictivo, sin la aspiración de explicar causas y efectos, las variables que desempeñan el rol de predecir son llamadas *predictores* y las variables que son predichas serán llamadas o bien *variables criterio* u, otra vez, *variables dependientes*.

No hay nada inherente en las variables que lleve a usar estos términos. Se trata del papel desempeñado en distintos estudios lo que determina la designación. La confusión se presenta en estudios epidemiológicos serios. Por ejemplo, las mujeres que tienen herpes son aquellas en las que es más probable que tengan cáncer cervical que las no expuestas al virus. Se concluyó que el herpes es causa de cáncer, es decir que se trata de una asociación causal. Investigaciones ulteriores sugieren que el herpes es meramente un marcador de actividad sexual. Las mujeres que tienen múltiples parejas sexuales están probablemente más expuestas, no solamente al herpes sino al virus papiloma humano. Algunas cepas del virus papiloma parecen ser causa del cáncer cervical, no así otras. Parece que la asociación entre herpes y cáncer cervical no es factor causal sino que es producto de otras variables. La elección en forma aleatoria tiende a equilibrar los grupos con respecto a las posibles variables confusivas; los resultados que quedan pueden ser evaluados usando técnicas estadísticas. Luego, las inferencias basadas en experimentos randomizados bien ejecutados son más seguras que las basadas en estudios de observaciones.

Podemos ilustrar estos puntos mediante el siguiente ejemplo. Hay varios médicos que piensan que la toma de aspirinas impide el ataque al corazón, pero hay controversias. La mayoría de los que toman aspirina no tiene ataques cardíacos; ésta es una evidencia de un efecto protector, pero demuestra muy poco. Después de todo, mucha gente no tiene ataques cardíacos – tomen o no aspirinas. Un estudio cuidadoso debe comparar los ataques cardíacos para dos grupos: los que toman aspirinas (*el grupo de tratamiento*) y los que no lo hacen (*el grupo de control*). Un estudio de las observaciones podría hacerse fácilmente, pero entonces los que toman aspirinas serán distintos que los controles. Si, por ejemplo, los controles son más saludables, el estudio estaría

sesgado en contra de esa medicación. Un experimento randomizado con aspirinas es más difícil de ser llevado a cabo, pero proporciona mejor evidencia. Son los experimentos que demuestran un efecto protector.

Una vez que las variables han sido definidas operativamente, que han sido elegidas mediciones fiables y válidas de las mismas, y que los objetos del estudio han sido medidos, los datos pueden ser analizados usando las herramientas estadísticas. En lugar de ser un estudio que termina con una larga lista de números, se computan estadísticos descriptivos para facilitar un resumen de su distribución. La tendencia central (o "promedio") de la distribución puede ser calculada como la media –aritmética, geométrica o armónica-, la mediana, o el modo, según la escala de medición usada y la forma de la distribución. La variabilidad (la dispersión) de la distribución puede ser expresada mediante la varianza, el desvío estándar (o desvío típico), o el rango, entre otros.

3) Muestreo

Estudio de datos La mayoría de los estudios estadísticos que se presentan en un tribunal son de observaciones, y no experimentales. Por ejemplo, tomen la siguiente cuestión: ¿Disuade la pena capital el asesinato? Para realizar un experimento controlado randomizado, la gente debería ser asignada en forma aleatoria a un grupo de control y a uno de tratamiento. Los del primer grupo sabrían que a ellos no se les aplicaría la pena de muerte por asesinato, mientras que los asignados al grupo de tratamiento sabrían que podrían ser ejecutados. Se observaría entonces la tasa de asesinatos cometidos en cada grupo, pero este experimento es inaceptable – tanto política, ética como legalmente.

Sin embargo, han sido llevados a cabo varios estudios del efecto disuasivo, todos basados en observaciones, y algunos atrajeron la atención de los jueces. Los investigadores catalogaron la incidencia del asesinato en estados de los US con y sin pena de muerte, y analizaron los cambios de la tasa de homicidio y las tasas de ejecución a través del tiempo. En estos estudios de observaciones, los investigadores pueden hablar de grupos de control (estados sin la pena capital) y del control de variables potencialmente confusivas (p.ej., peores condiciones económicas). Empero, asociación no implica causalidad, y las inferencias causales que pueden ser extraídas de estos análisis descansan en fundamentos menos sólidos que los de un experimento randomizado controlado. Por otra parte, cabe citar un estudio del Instituto Criminológico Australiano[11] que concluye lo siguiente: "Cuando esos delitos se cometen al calor de la pasión, con intención premeditada de matar, el delincuente a menudo hace escaso intento de evitar la detección. En estas circunstancias, está claro que el efecto disuasorio es mínimo. Por otro lado, cuando los homicidios son premeditados, la amenaza de la pena de muerte puede ser un elemento de disuasión menor que el riesgo de ser atrapado. Más allá de esto, la pena de muerte puede crear un efecto de brutalización, realmente inspirar los actos de violencia y así disminuir en lugar de aumentar el efecto disuasorio de la pena capital. Hasta la fecha las pruebas no han podido establecer que la pena de muerte sea más eficaz que el encarcelamiento en disuadir el crimen." Concluyen: "Aunque las encuestas de opinión públicas generalmente indican que una mayoría de la comunidad está a favor de la pena de muerte para determinados delitos, muchos dirían que tiene poco valor disuasorio real por encima de la prisión. Quienes abogan por la pena de muerte a

[11] Ivan Potas and John Walker, Capital punishment, February 1987, Australian Institute of Criminology. http://www.aic.gov.au/en/publications/current%20series/tandi/1-20/tandi03/view%20paper.aspx. Cabe notar que Argentina es un país de abolición relativamente tardía de la pena de muerte (2008, Código de Justicia Militar) si bien el año en que se produjo la última ejecución fue 1916. http://en.wikipedia.org/wiki/Use_of_capital_punishment_by_country#cite_ref-54

causa de que al menos el asesino es eliminado permanentemente de la sociedad, tienen también que tener en cuenta el hecho de que en la práctica la pena de muerte a menudo es administrada caprichosamente y que hay siempre una posibilidad de que un inocente pueda ser ejecutado."

Pocas veces los investigadores recogen datos sobre todas las instancias individuales de los objetos estudiados (*censo*). Lo que hacen usualmente es un *muestreo* de los mismos. Los investigadores agrícolas sacan una muestra de maíz en el campo; no miden cada espiga una por una. Lo mismo sucede en otro tipo de investigaciones. Tomar muestras no sólo es menos costoso y lleva menos tiempo; *en la mayoría de las circunstancias es más preciso que un censo*. Mediante un diseño muestral apropiado, los recursos pueden ser destinados a recolectar datos más precisos sobre una menor cantidad de individuos, cosas, o eventos. En efecto, en US y en otros países los demógrafos evalúan la bondad de los censos a nivel del país comparando los resultados del censo con muestras.

Las Unidades de Análisis El primer paso para practicar una muestra es decidir qué será muestreado, es decir la unidad de análisis y cuál el nivel de agregación. Por ejemplo ¿estamos recolectando datos sobre gente individual o sobre agregados tales como ciudades o naciones? ¿Sobre trabajadores, organizaciones, o industrias? ¿Sobre rocas, planetas, o sistemas solares? Algunas cosas sólo existen a elevados niveles de agregación. Por ejemplo, la forma en que está organizada una empresa no puede discernirse examinando sólo los individuos, sino las relaciones estructurales de grupos de individuos con otros grupos. Estas decisiones pueden afectar los análisis estadísticos realizados y las conclusiones extraídas. A veces, cuando el fenómeno de interés puede ser estudiado observando cosas a distintos niveles de agregación, se tienen diferentes conclusiones que usando una unidad de análisis en lugar de otra.

Tipos de Muestreo Típicamente el objetivo de un muestreo es aprender sobre una población completa de cosas observando un subconjunto de ellas. La clave consiste en seleccionar una muestra que sea *representativa* de la población. Cabe mencionar que para que el muestreo sea válido y se pueda realizar un estudio adecuado (que consienta no sólo hacer estimaciones de la población sino estimar también los márgenes de error correspondientes a dichas estimaciones), debe cumplir ciertos requisitos. Nunca podremos estar enteramente seguros de que el resultado sea una muestra representativa, pero sí podemos actuar de manera que esta condición se alcance con una probabilidad alta. En tal caso, lo que se aprenda de la muestra es probablemente cierto también de la población. Los métodos utilizados para hacerlo son colectivamente conocidos como muestreo probabilístico. El *muestreo probabilístico* involucra seleccionar casos de la población de tal manera que existe una probabilidad conocida de que cualquier caso aparezca en la muestra. Esto permite usar teoría de la probabilidad para extraer inferencias sobre la naturaleza de la población. Algunos tipos comunes de muestreo probabilístico son los siguientes.

El *muestreo probabilístico simple* implica extraer una muestra de la población relevante de forma tal que cada miembro de la población tenga similar oportunidad de ser escogido en la muestra. Por ejemplo, si se desea medir la incidencia de cierta enfermedad entre los estudiantes de una escuela, se puede elegir una muestra aleatoria extrayendo números de un saco o bien mediante un computador que genera un subconjunto aleatorio de estudiantes. En tal caso la muestra puede ser consultada por cualquier necesidad de hacer un test. El *muestreo sistemático* es similar, con la diferencia de que en lugar de seleccionar al azar, se elige un punto de partida aleatorio y entonces, p.ej., cada 10º estudiante del directorio de estudiantes es elegido uno (lo que facilita

una muestra del 10% de los estudiantes). Estos métodos funcionan si cada miembro de la población es conocido y la población es homogénea.[12]

Una *muestra estratificada* es una en la que han sido especificados por adelantado subgrupos de la población, a partir de lo cual tiene lugar un muestreo dentro de cada estrato. Este método es de utilidad cuando algunos grupos de la población registran números pequeños y el investigador desea asegurarse de que un número suficiente sea extraído, de modo que se pueda armar una muestra suficientemente grande como para obtener inferencias fiables sobre los subgrupos de la población. Se podrían extraer muestras proporcionales al tamaño de cada estrato. Por ejemplo, 40% de la muestra correspondería al estrato que contiene 40% de la población y 60% de la muestra al estrato que contiene 60% de la población. En forma alternativa, se puede extraer una "muestra estratificada desproporcionada". Así, se podría tener una muestra de 250 de una minoría del 10% y otros 250 de una mayoría del 90%. Un muestreo aleatorio simple hubiera significado sólo 50% de la minoría, lo que podría ser muy escaso. Las dos sub-muestras de esta muestra estratificada luego deberían ser ponderadas (una recibiendo nueve veces la ponderación de la otra) de modo que, al final, los enunciados realizados sobre la población fueran precisos.

Según la cantidad de elementos de la muestra que se han de elegir de cada uno de los estratos, existen dos técnicas de muestreo estratificado: 1) la *asignación proporcional*: el tamaño de la muestra dentro de cada estrato es proporcional al tamaño del estrato dentro de la población; 2) la *asignación óptima*: la muestra recoge más individuos de aquellos estratos que tienen más variabilidad. Para ello es necesario un conocimiento previo de la población.

Los dos métodos anteriores pueden combinarse de otras maneras para tratar con circunstancias más complejas, como cuando la población es muy amplia y heterogénea, y la identidad de los elementos (habitualmente, el nombre de cada uno) es desconocida. Esto conduce al *muestreo multi-etápico de grupos*. Supongan que deseamos estudiar la salud de la gente en nuestro país. No tenemos un listado de individuos para elegir aleatoriamente, y aunque lo tengamos sería ineficiente tratar de visitar a los seleccionados, dispersos en todo el territorio. Mas disponemos de una "estructura muestral" de todos los partidos y departamentos en todas las provincias. Se saca una lista representativa de departamentos de la lista; inclusive, ésta podría estar estratificada de alguna manera, como por región del país. (Éste es el primer estadio del diseño muestral.) Con los departamentos seleccionados, podemos extraer una muestra aleatoria de gente para efectuar el test, quizá por medio de un llamado aleatorio a números telefónicos, invitando a la gente a efectuar el test. (La selección de estos individuos es el segundo estadio.) Por lo tanto, la gente que resulta muestreada termina "agrupada" en departamentos seleccionados en todo el país.

Se extraen dos lecciones de las técnicas muestrales descriptas. Puede hacerse un diseño particular muestral adaptado a la naturaleza de lo que se muestrea, combinando distintos métodos

[12] Homogéneo significa, en el contexto de la estratificación, que no hay mucha variabilidad. Los estratos funcionan mejor cuanto más homogéneos son cada uno de ellos respecto a la característica a medir. Por ejemplo, si estudiamos la estatura de una población, es bueno distinguir entre los estratos mujeres y hombres porque se espera que, dentro de ellos, haya menos variabilidad, es decir, sean menos heterogéneos. Es decir, no hay tantas diferencias entre unas estaturas y otras dentro del estrato que en la población total. Por el contrario, la heterogeneidad hace inútil la división en estratos. Si se dan las mismas diferencias dentro del estrato que en toda la población, no hay por qué usar este método de muestreo. En los casos en los que existan grupos que contengan toda la variabilidad de la población, se construyen conglomerados, que ahorran trabajo que supondría analizar toda la población. En resumen, los estratos y los conglomerados funcionan bajo principios opuestos: los primeros son mejores cuanto más homogéneo es el grupo respecto a la característica a estudiar y los conglomerados, si representan fielmente a la población, esto es, contienen toda su variabilidad, o sea, son heterogéneos.

muestrales (aleatorio y estratificado, en diferentes estadios). Y cada diseño será una muestra probabilística al conocerse la probabilidad de seleccionar cualquier elemento en todo estadio.

Se han desarrollado varios métodos de muestreo *no probabilístico*. Uno es el *muestreo con un propósito*, que refleja que el investigador tiene in mente un propósito determinado por la manera en que se elige la muestra. P. ej., en un estudio para descubrir cómo los médicos conocen nuevas medicaciones, los investigadores comenzaron con registros farmacéuticos que mostraban qué médicos prescribían la medicación en una comunidad dada. Luego mantuvieron entrevistas con los médicos, preguntándoles, entre otras cosas, con qué otros amigos profesionales alternaban, y preguntaron a estos amigos quiénes eran sus amigos profesionales. Este método es llamado "muestreo por bola de nieve". A medida que la muestra crecía, los investigadores pudieron detectar la difusión de la conciencia de que existía una nueva medicación por medio de la red de amigos emergente.[13] Resulta claro que, en este caso, algunos de los métodos de muestreo probabilísticos tradicionales no hubieran sido útiles para responder a la cuestión investigada.

El Sesgo de Selección El mayor defecto de un proyecto muestral es que fracase en seleccionar los elementos representativos de la población interesada. Lo más frecuente es que ello ocurra por el sesgo de selección.

Este sesgo se refiere a una muestra que ha sido extraída de una forma que no la hace representativa de la población con respecto a la cual deben hacerse las inferencias. El tema es más fácil de apreciar mediante varios ejemplos. Ustedes pueden preguntarse por qué es probable que las conclusiones a partir de las siguientes muestras sean engañosas (entre paréntesis se han incluido posibles respuestas).

> Un criminólogo se puso a estudiar criminales realizando extensas entrevistas con una muestra aleatoria de presos en una prisión. (Lo que pudo haber hallado nos habla sólo de los criminales que fueron apresados y encarcelados).
>
> Los doctores se enteraron de la histoplasmosis[14] estudiando a pacientes que llegaban al hospital con la enfermedad, y concluyeron que se trataba de una enfermedad infrecuente que casi siempre era fatal. (En contraste, investigadores en salud pública practicaron una muestra del público en general y hallaron que la enfermedad era mucho más común y que pocas veces dañaba seriamente. Los doctores de los hospitales sólo veían a los pocos pacientes que sufrían la enfermedad de modo suficientemente serio como para precisar atención médica).
>
> Durante la II Guerra Mundial se hizo un estudio para determinar si un blindaje adicional aplicado a los aeroplanos los protegería del fuego anti-aéreo. Para hacerlo, las aeronaves que regresaban de sus misiones fueron examinadas a efectos de indagar si habían recibido los disparos del fuego enemigo. (Las aeronaves que habían sido derribadas pero

[13] Se halló, por ejemplo, que era más probable que los doctores se enteraran de los nuevos medicamentos por medio de sus amigos que leyendo revistas de medicina.

[14] Micosis sistémica, caracterizada por lesiones necrogranulomatosas, que afecta a carnívoros, equinos y humanos por la infección con una de las tres subespecies del hongo dimórfico *Histoplasma capsulatum*. No es una enfermedad contagiosa que se pueda transmitir entre personas o animales. Su manifestación en personas inmunocompetentes suele ser asintomática. Puede cursar con cuadros parecidos a los de una neumonía con fiebre, distrés respitatorio, y en un 20% aproximadamente de los pacientes se llega a producir un shock séptico, fallo renal y coagulopatía. http://es.wikipedia.org/wiki/Histoplasmosis

pudieron haber sido salvadas mediante blindaje adicional eran las que proporcionaban más información, pero, por supuesto, éstas no retornaron).

Un comité organizador de una clase de graduados envió un cuestionario anónimo a sus miembros antes de su 25ª reunión. Incluía una pregunta sobre el ingreso del graduado. Cuando se promediaron las respuestas que fueron devueltas, el comité quedó sorprendido de saber cuán exitosos habían sido los miembros de su clase de graduados en la vida. (En realidad, una "tasa de respuestas" menor que el 100%, aunque fuera baja, no sería en sí un problema. Como siempre, la cuestión radica en saber si los que respondieron eran representativos de la población. En este caso, si es probable que los alumnos más exitosos sean los que devuelvan sus cuestionarios, el ingreso medio de la clase aparecerá más elevado que lo que es).

Aspectos estadísticos. Tamaño de la muestra ¿Cuán amplia debería ser una muestra para que los resultados obtenidos sean válidos? La respuesta es algo contraria a la intuición. En primer lugar, lo importante es el tamaño absoluto de la muestra, no el tamaño de la muestra con relación al tamaño de la población (más adelante serán analizados algunos aspectos matemáticos involucrados). En segundo término, habitualmente el investigador tiene mejor idea del tamaño necesario de la muestra *después* de recoger los datos en lugar de antes. En tercer término, a mayor tamaño de la muestra, tanto mejor. Sin embargo, el hecho que enfrentan los investigadores es que, si bien se puede obtener un beneficio real en precisión cuando se pasa de n=10 a n=25 o n=100, a medida que son agregados más datos a la muestra, el beneficio marginal en precisión se encoge en forma considerable. Luego un investigador tiene que preguntarse si el gasto adicional de agregar 100 o 500 vale la pena en términos de la ganancia de precisión obtenida.

El tamaño óptimo de una muestra depende de tres factores:

1. La heterogeneidad de la variable que debe ser medida en la población. Supongan por ejemplo que un almacén que tiene latas de sopa se inundó y que todas las etiquetas quedaron borrosas. Pero se sabe que el almacén tenía sólo latas de sopa de un solo gusto. ¿Cuántas latas deberían ser abiertas para responder a la pregunta sobre qué gusto tenía la sopa de todas las latas del almacén? La respuesta obvia es: una (si el almacén hubiera tenido latas de diversos gustos, deberían haber sido abiertas más latas). La homogeneidad de la variable frecuentemente no es posible conocerla hasta que los datos sean recolectados.
2. Cuán próxima a cero debería ser la respuesta del investigador. Esto es, el investigador desea establecer "límites de confianza" alrededor del enunciado estadístico que debe hacer sobre una variable. (A veces más o menos 10 o 20% será correcto; otras veces no).
3. Cuán confiado debe estar el investigador en que el rango obtenido alrededor de los parámetros de la población sea correcto – ello depende lógicamente del tipo de preguntas de la investigación, aunque es típico que los investigadores sigan la convención de buscar una confianza en torno al 99% o al 95%.

Límites de confianza Una muestra es recogida usualmente para efectuar algún tipo de inferencia estadística. Supongan que un funcionario debe calcular la cantidad promedio de monóxido de carbono descargada por automóvil en la ciudad de Buenos Aires. Los investigadores hacen una muestra de autos y de días. Supongan que hallan que la cantidad promedio de la muestra es de "30 mil toneladas anuales". Conociendo (a) la variabilidad o heterogeneidad de la muestra, (b) el tamaño de la muestra, y (c) el nivel de confianza deseado por el investigador, los investigadores pueden luego hallar que el verdadero parámetro poblacional oscila entre 30 mil toneladas más o

menos 8 mil toneladas, con 95% de confianza. Es decir, que existe un 95% de probabilidad de que la verdadera media poblacional caiga en un punto comprendido entre 22 mil y 38 mil toneladas.

4) Relaciones entre variables

Hasta este punto hemos visto cuestiones metodológicas vinculadas con la medición de una sola variable por vez, sin buscar vinculaciones entre una y otra. Pero es frecuente que uno esté interesado en asociar variables como en estos casos: ¿Qué tratamiento es más probable que cure una enfermedad? ¿Qué variables son las mejores predictoras de quién será bueno en la universidad, quién hará el mejor entrenamiento deportivo, etc.? ¿Qué técnicas administrativas hacen que los trabajadores sean más productivos? ¿Qué políticas dan lugar al crecimiento económico? ¿Qué programas es más probable que reduzcan el crimen? ¿Qué métodos educativos son más efectivos? ¿Qué métodos de comunicar las ideas las hace más convincentes? ¿Cuál es la causa de los tornados? La vida está llena de preguntas como éstas, e investigadores de todo tipo están a la búsqueda de respuestas.

Condiciones mínimas que deben darse para inferir una relación A fin de extraer una inferencia acerca de una relación entre variables, debemos tener datos sobre por lo menos dos niveles de dos variables. Menos que esto torna imposible decir algo sobre una relación. Las tablas siguientes ilustran estas condiciones mínimas.

Tabla A

	Buen Desempeño	Mal Desempeño	
Inteligencia			
Alta	260	240	
Baja	225	275	1000

Tabla B

	Buen Desempeño	Mal Desempeño	
Inteligencia			
Alta	225	275	
Baja	300	200	1000

Tabla C

	Buen Desempeño	Mal Desempeño	
Inteligencia			
Alta	400	100	
Baja	400	100	1000

Supongan que la cuestión es si un buen resultado de un test de inteligencia está vinculado con el desempeño de un individuo en un determinado trabajo. Se practica un estudio sobre 1000 trabajadores que desempeñan una determinada tarea, y que mediante una medición válida y fiable se evalúa su desempeño y se mide su inteligencia. En la tabla A, los datos tienden a agruparse en las dos celdas en que (a) la inteligencia es alta y el desempeño fue bueno, y (b) la inteligencia es baja y hubo un pobre desempeño. Luego, se podría inferir a partir de este patrón de evidencia que la mayor inteligencia estuvo asociado con un mejor desempeño en la tarea. Aún así, las diferencias entre estas celdas y las restantes son pequeñas, por lo cual se trata de una débil relación: la inteligencia no tiene demasiado que ver con esta tarea.

En la tabla B la relación es opuesta: la elevada inteligencia está asociada con un pobre desempeño. Más aún, la relación es aún más fuerte que la previa: en esta tarea, ser muy inteligente es un impedimento para un buen desempeño.

En la tabla C los datos revelan que no existe ninguna relación entre inteligencia y desempeño en el trabajo (entre los trabajadores inteligentes, la relación entre los que trabajan bien y los que lo hacen mal es de 4 a 1, exactamente como para los poco inteligentes).

Ahora bien. Observen que una menor cantidad de datos de estas dos medidas en dos variables no nos permitiría extraer ninguna inferencia sobre la relación entre ambas variables. Sin embargo, a menudo la gente termina creyendo en afirmaciones que carecen de sostén empírico. Esto nos conduce al problema de los *patrones de datos faltantes*.

Una situación en la que hay una sola celda y se trata de obtener una inferencia es la siguiente:

Tabla D

	¿Existe comercio primario?	
	Si	No
Guerra Si	8	
Guerra No		

En un debate sobre el valor de establecer una fuerte asociación comercial con otro país enemigo de larga data, con el fin de reducir la eventualidad de que surja una guerra armada entre ambos, un asistente ofreció como evidencia contra la propuesta un listado de ocho pares de estados-nación que habían comerciado en productos primarios antes de entrar en guerra entre sí (tabla D). Durante el debate, esta idea tuvo un fuerte impacto contra la afirmación de que establecer relaciones comerciales reduce las chances de un conflicto armado. ¡Pero un examen de la tabla D muestra que, como ignoramos todo acerca de las celdas siguientes, en realidad no sabemos nada sobre la relación entre comercio y conflicto armado! En otras palabras, el individuo ofreció un patrón de datos faltantes con una sola celda, y se las arregló para convencer a la mayoría de su audiencia de que estaba diciendo algo importante sobre la relación en cuestión, cuando en realidad no estaba diciendo nada.

A veces se ofrece un argumento algo más rico, que consiste de un patrón de una sola fila o columna de la matriz básica de 2x2 que constituye el *minimum minimorum* para extraer una inferencia sobre una relación.

Tabla E

	¿Fumó marihuana en (t-2)?	
	Si	No
Adicto a la heroína en (t-1)		
Si	300	200
No		

Por ejemplo, se ha sugerido que fumar marihuana, aunque no cause daño de por sí, es peligrosa porque conduce (por caminos farmacológicos, psicológicos o sociológicos) al uso de drogas más duras. Se ha dicho que si se hubiera hallado un número sustancial de adictos a la heroína consumiendo marihuana cuando eran jóvenes, se confirmaría la hipótesis. La tabla E indica el

patrón que estos comentarios tienen in mente. Los datos (hipotéticos) del cuadro muestran que 60% de una muestra de adictos a la heroína fumaron marihuana cuando eran más jóvenes. Como estos datos configuran un patrón de toda una línea faltante, no pueden revelar si existe o no una relación. Específicamente, no es posible decir si menos que el 60%, la misma proporción o más del 60% de la gente que no son adictos a la heroína fumaron marihuana. Y la existencia de una relación depende de este tipo de comparaciones.

Alternativamente, supongan que alguien propuso que tomar leche cuando uno es un niño conduce a ser un adicto a la heroína de adulto. ¡En este caso se aplica la misma tabla! ¿Significa esto que debemos condenar a la leche? Nuevamente, si se carece de datos para las otras celdas, es imposible inferir relación alguna.

El patrón final que resta analizar es cuando solamente se dispone de datos de celdas en la diagonal.

Tabla F

	Comportamiento dañino	
	Si	No
Peligrosidad Si	50	
Peligrosidad No		200

Supongamos que a un experto en predecir la peligrosidad se le pide informar su *track record*, y su respuesta es que en 50 oportunidades predijo que individuos que serían peligrosos provocaron daño, mientras que en 200 ocasiones en que predijo que los individuos no serían peligrosos éstos no provocaron ningún daño (tabla F). Ahora ustedes se darán cuenta de que, hasta que no tengamos datos sobre las celdas que están fuera de la diagonal, será imposible saber a partir de estos escasos datos si el experto es muy preciso o impreciso. Si el experto hizo otras 2.000 predicciones, acomodadas en las celdas restantes, entonces sabríamos si estamos ante un experto que yerra en cerca de 90% de sus predicciones.

Relaciones predictivas Cuando están dadas las condiciones mínimas, puede discernirse una relación predictiva siempre que haya alguna relación entre las variables. Las relaciones predictivas simples – relaciones de correlación o hallazgos que son el producto de estudios de observaciones – nos informan si una variable está asociada con otra, y de cuán fuerte es el margen de asociación existente. *No nos están informando que los cambios de una variable causen los de la otra*. Luego hablaremos de cómo establecer relaciones causales, que es una tarea más difícil.

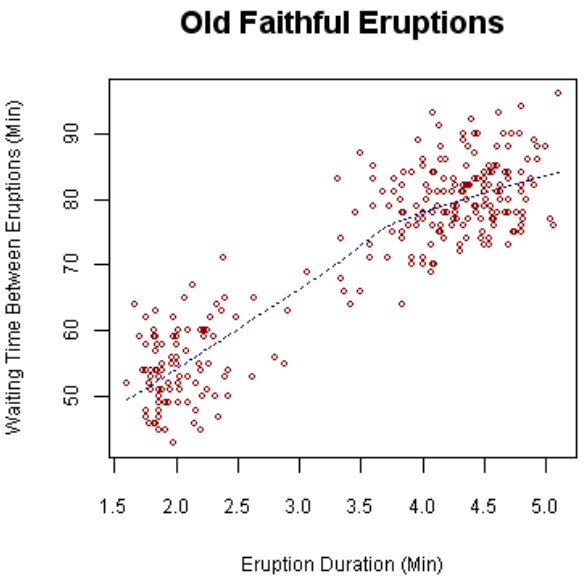
He aquí un listado de relaciones predictivas, para brindar luz sobre el tema de que la correlación no demuestra causalidad, aunque a veces se piense en tal sentido.

> Piojos, barbas, y correlaciones espurias: Los investigadores observaron una vez que las barbas de algunos indígenas del tercer mundo mostraban una tendencia a tener piojos, no así las de otros hombres, y que los hombres con piojos tendían a ser más saludables que los que no tenían. Se halló una correlación (positiva) entre piojos y estado sanitario bueno. A partir de estos únicos datos, son posibles numerosas explicaciones: los piojos son buenos para la salud, lo que haga la gente para ser saludables también promueve los piojos, la gente enferma que superó su enfermedad también contrajo piojos, la gente sana atrae a los piojos, etc. Las correlaciones no permiten distinguir entre distintas posibilidades.

El matrimonio, el crimen, y la dirección causal: Viendo que es más probable que los solteros cometan crímenes que los casados, un comentarista recomendó que la sociedad trate de casar la mayor cantidad posible de gente para combatir el crimen y otros problemas que esta gente representa para la sociedad. Es más probable que los solteros estén enfermos, tengan menor ingreso, menor educación, y otros problemas. Pero la dirección causal puede ir también en sentido contrario. Por ejemplo, los criminales sufren más probablemente problemas de salud, pobreza, etc. y por consiguiente les resulta menos probable hallar compañeras para casarse, lo que conduce a una tasa más baja de matrimonios de ese grupo.

Fumar durante el embarazo e inteligencia de los niños, y el problema de la "tercera variable": Un estudio halló que mujeres que fumaban durante su embarazo tenían niños con un IQ más bajo en promedio, que las que no lo hacían. Este resultado dio lugar a la predicción de que la descendencia de las mujeres embarazadas fumadoras tendría menor inteligencia. Hay una explicación plausible alternativa: es probable que las madres con bajo IQ fumen (lo que parece cierto) y que también tengan hijos con menor IQ (también es cierto). Luego, la tercera variable (el IQ de la madre) es causa de las otras dos variables (fumar e IQ de sus hijos).

Un estudio halló que las provincias con más ventas de pickles también eran aquellas donde los estudiantes alcanzaban puntajes más elevados en las pruebas educativas. Los investigadores concluyeron que los pickles mejoran el rendimiento en la escuela, y recomendaron que los boliches aledaños sirvieran más pickles a los estudiantes. Pero existe una explicación más probable: la situación económica de la provincia es responsable tanto de la venta de pickles (suponiendo que es un bien superior) como de la performance de los estudiantes (la buena situación económica permite una mejor base tributaria, más gasto educativo y mejores escuelas).



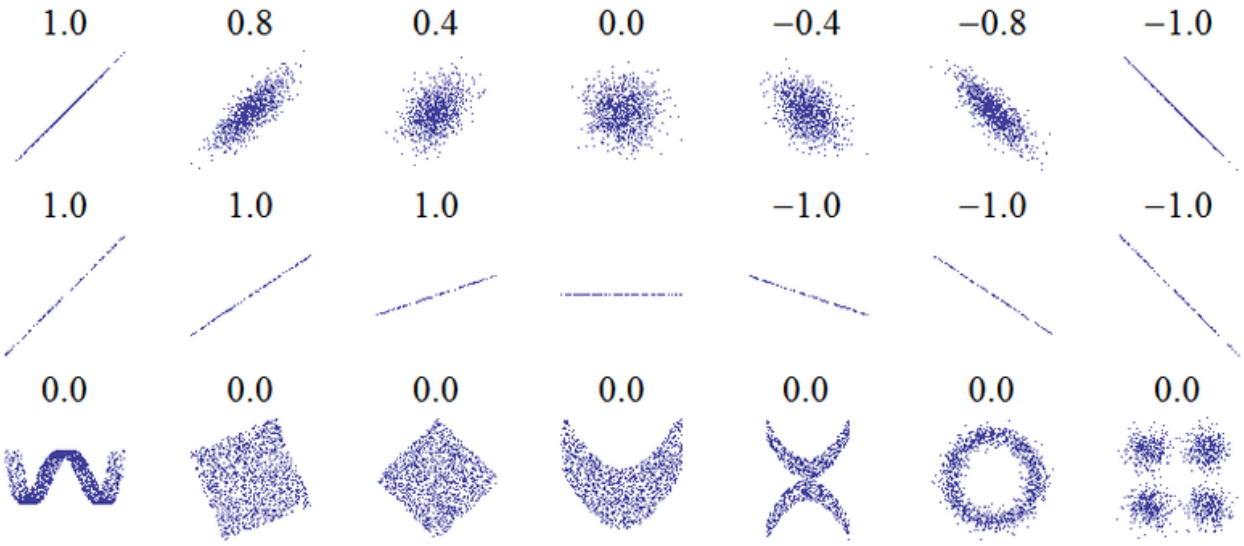
Un observador de Marte que estudia la Tierra observa que los autos tienden a girar hacia la izquierda cada vez que una luz del lado izquierdo del auto comienza a parpadear, y que tienden a girar a la derecha cuando una luz del lado derecho del auto comienza a parpadear. Usando la misma lógica defectuosa que a veces usan los terráqueos, el marciano saca la conclusión de que la luz es causa de que los autos giren (en lugar de que exista una "tercera variable" – el conductor, que es la causa de ambos eventos). *Se halló una buena relación predictiva, pero el observador tiene un diagnóstico erróneo de causalidad*.

Aspectos estadísticos Las relaciones predictivas pueden examinarse de manera gráfica usando "diagramas de dispersión" y "tabulaciones cruzadas" (Las tablas A, B y C proveen ejemplos de este último caso). El diagrama de dispersión anterior proporciona el tiempo de espera entre las

erupciones y la duración de la erupción del géiser Old Faithful en el Parque Nacional Yellowstone, Wyoming, US. Este gráfico sugiere que por lo general hay dos tipos de erupciones: una de corta espera y corta duración y otra de larga espera y larga duración. Un diagrama de dispersión es un tipo de diagrama matemático que utiliza coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos. Los datos se muestran como un conjunto de puntos, cada uno con el valor de una variable que determina la posición en el eje horizontal y el valor de la otra variable determinado por la posición en el eje vertical. Un diagrama de dispersión es llamado también gráfico de dispersión.

Pueden introducirse medidas utilizando distintos estadísticos de correlación: p.ej. el coeficiente de Pearson adopta valores comprendidos entre -1 y +1. Cabe establecer una diferencia entre la correlación *poblacional*, dada por la fórmula:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$



Distintos conjuntos de puntos *(x, y)* indicándose el coeficiente de correlación de Pearson de cada uno

definida como la covarianza de ambas variables dividida por el producto de sus desvíos estándar, y la correlación *muestral*, que se obtiene sustituyendo en esta fórmula por la covarianza y los desvíos estándar de la muestra:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Observen en la figura anterior que la correlación es un índice del ruido y de la dirección de una relación lineal (fila superior), *pero no del valor de la pendiente* de esa relación (fila del centro), ni de varios aspectos de las relaciones no-lineales (abajo). La figura del centro tiene pendiente igual a 0 pero su coeficiente de correlación no está definido, porque la varianza de Y es 0.

Varios autores han propuesto pautas de interpretación de un coeficiente de correlación. Empero, se ha dicho[15] que todos estos criterios son, en cierta forma, arbitrarios y que no deberían ser observados en forma demasiado estricta. La interpretación de un coeficiente de correlación depende del contexto y de los propósitos perseguidos. Una ρ= 0.9 puede resultar demasiado baja si se está verificando una ciencia física usando instrumentos de alta calidad, pero muy elevada en las ciencias sociales donde hay típicamente mayor contribución de factores confusivos.

Cálculo Vean el siguiente ejemplo. Un investigador desea saber si la altura de la gente está asociada con su sentimiento de auto-estima (aquí no hay un problema de decir cuál es el sentido de la causalidad, dado que podemos descartar que la auto-estima influya sobre la altura de las personas). Hemos recopilado datos sobre veinte individuos (todos varones, ya sabemos que la altura promedio difiere entre ambos sexos, de modo que sólo usamos datos de varones). La altura está medida en pulgadas. La auto-estima, extraída de un test psicológico, viene dada como el promedio de 1 a 5 de diversos ítems de calificación (los números más elevados corresponden a una mayor auto-estima). Los datos de los 20 casos están reportados en la tabla siguiente (no tomen demasiado en serio esta información, sólo ha sido preparada para ilustrar de qué se trata una correlación):

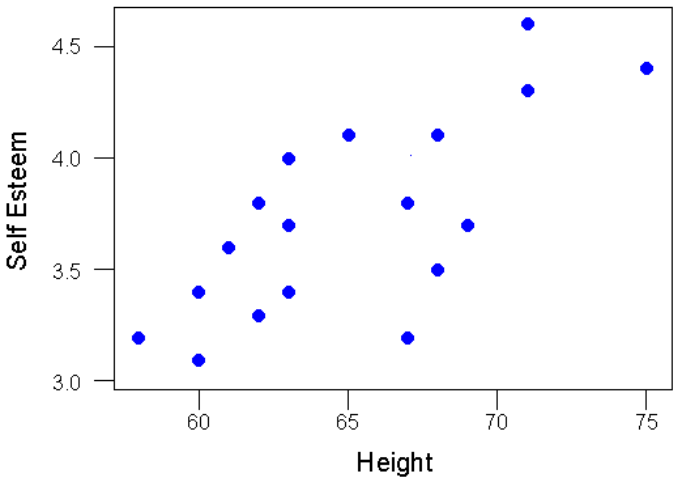
Persona	Altura (pulgadas)	Auto-estima
1	68	4.1
2	71	4.6
3	62	3.8
4	75	4.4
5	58	3.2
6	60	3.1
7	67	3.8
8	68	4.1
9	71	4.3
10	69	3.7
11	68	3.5
12	67	3.2
13	63	3.7
14	62	3.3
15	60	3.4
16	63	4.0
17	65	4.1
18	67	3.8
19	63	3.4
20	61	3.6

[15] J. Cohen, (1988). Statistical power analysis for the behavioral sciences (2nd ed.).

Las estadísticas descriptivas de estos datos han sido dispuestas en la tabla siguiente:

Variable	Promedio	Desvío Estándar	Varianza	Suma	Mínimo	Máximo	Rango
Altura	65.4	4.40574	19.4105	1308	58	75	17
Auto-estima	3.755	0.426090	0.181553	75.1	3.1	4.6	1.5

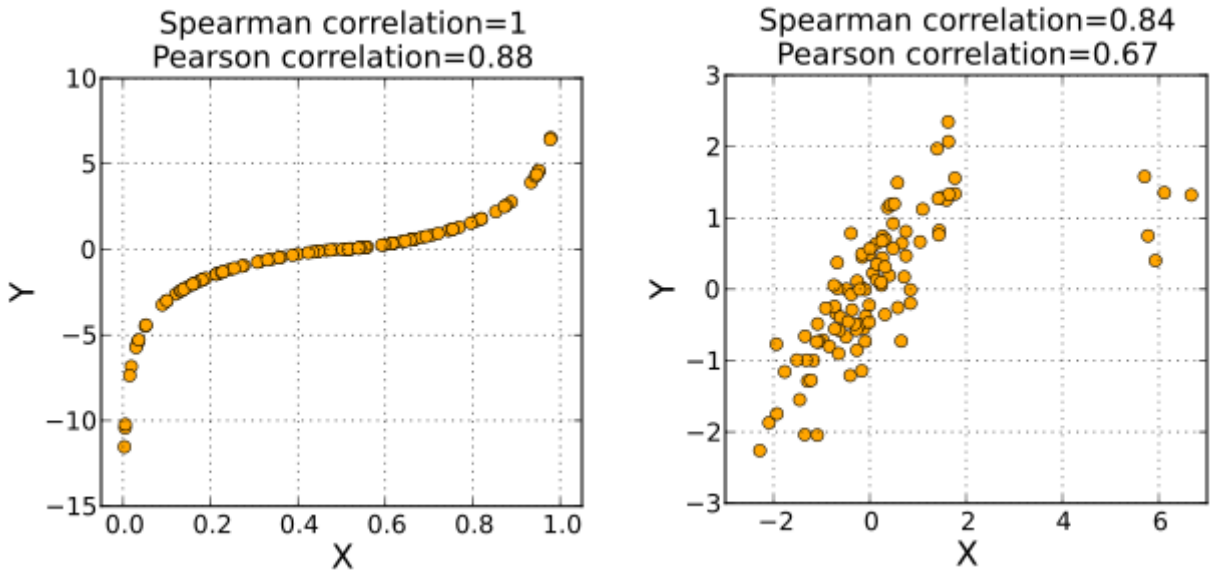
El gráfico de la derecha muestra el diagrama de dispersión.[16] Por lo que podemos ver, estamos frente a una asociación *positiva* de ambas variables. Luego, podemos esperar que el cálculo del coeficiente de correlación arroje una magnitud positiva. *Una asociación positiva significa que, en general, valores más elevados de una variable tienden a estar asociados con valores más elevados de la otra, y que valores más bajos de una variable tienden a estar asociados con valores más bajos de la otra*.



Ahora podemos calcular la correlación. Usando la fórmula para *r*, dado que estamos ante datos de una *muestra*, se obtiene *r*= 0.73. Se está en presencia, por ende, de una correlación positiva bastante elevada.

El *coeficiente de correlación de rangos de Spearman* sirve para establecer si una relación entre dos variables puede ser descripta mediante una función monótona[17] (creciente o decreciente). Si no hay datos repetidos, un coeficiente "perfecto" de +1 o -1 significa que una de las variables es una función perfectamente monótona de la restante. Este coeficiente de correlación de Spearman se define mediante el coeficiente de correlación de Pearson calculado sobre las variables en términos de rango.[18]

[16] En la leyenda del gráfico, *Self Esteem* es la traducción de Auto-estima; *Height* es la traducción de Altura.
[17] Una función entre conjuntos ordenados se dice monótona (o *isótona*) si conserva el orden dado. Las funciones de tal clase surgieron primeramente en cálculo, y fueron luego generalizadas al entorno más abstracto de la teoría del orden. Aunque los conceptos generalmente coinciden, las dos disciplinas han desarrollado una terminología ligeramente diferente; mientras en cálculo se habla de funciones monótonamente crecientes y monótonamente decrecientes (o simplemente crecientes y decrecientes), en la teoría del orden se usan los términos monótona y antítona, o se habla de funciones que conservan e invierten el orden.
[18] Por ejemplo, la serie 3.4, 5.1, 2.6, 7.3 es sustituida por 2, 3, 1, 4 respectivamente.



Es útil observar que la correlación de Spearman es menos sensible que la correlación de Pearson a los puntos atípicos en las colas de ambas muestras. Del primer gráfico de la muestra resulta una correlación de Spearman = 1 si las dos variables comparadas están vinculadas monótonamente, aunque la relación no sea lineal. Por oposición, no se tiene en este caso una correlación perfecta en sentido de Pearson. El gráfico del costado muestra que la correlación de Spearman es menos sensible que la correlación de Pearson a puntos atípicos muy marcados de las colas de ambas muestras. Hay maneras algo diferentes de calcular correlaciones cuando una o ambas variables están medidas usando la escala nominal, ordinal, o por intervalos.

El cuadro siguiente incluye una tabulación de correlaciones en varias áreas de interés, que cubren una amplia gama de ejemplos:

Aspirinas y ataques cardíacos	.033
Efectividad de la psicoterapia	.320
Grado alcanzado en el 1º año s/ LSAT[19]	.410
Efectividad del detector de mentiras	.670
Premios a jurados civiles de los médicos	.714
Distancia del hoyo al punto de tiro en competencias de golf	.940

Lo que es una relación predictiva puede ser descripto de forma más apropiada usando Análisis de Regresión. Este tipo de análisis permite describir la relación entre dos variables con la fórmula de una línea recta. Resulta de interés particular su pendiente: cuanto más empinada sea, tanto más cambio se “produce” en una variable por un cambio de la otra. Tener semejante fórmula permite una predicción directa y literal del número de una variable sabiendo el valor de la otra. Por ejemplo, ¿cuán bien le puede ir a un estudiante en el 1º año escolar predicho a partir del LSAT de ese estudiante? La correlación y la predicción permiten este tipo de análisis con cierta precisión.

A menudo, diversas variables deben ser usadas como predictores de un resultado. Por ejemplo ¿cuán bien le puede ir a un estudiante en el 1º año escolar sabiendo la edad del estudiante de

[19] Por Law School Admission Test, que se toma en U.S., Canadá y Australia, diseñado para evaluar la capacidad lógica y de razonamiento verbal. http://en.wikipedia.org/wiki/Law_School_Admission_Test

grado, el promedio de edades y el resultado del LSAT? La técnica de regresión múltiple permite combinar varias variables predictoras a fin de mejorar la exactitud de la predicción. Un estadístico importante adicional que acompaña un análisis de regresión múltiple es el coeficiente llamado *R cuadrado múltiple*, simbolizado como R^2. Se trata del cuadrado de la correlación entre los resultados *predichos* usando los predictores y los valores reales *observados* de los mismos. Nos proporciona la proporción de varianza predicha (o explicada) por las variables predictoras. A su raíz cuadrada se la suele denominar *coeficiente de correlación múltiple*.

Hay quienes creen que "la proporción de varianza explicada" da la impresión de una relación que subestima la magnitud de las relaciones. Han sugerido la conversión de correlaciones y correlaciones múltiples en un "despliegue del tamaño del efecto binomial" o "DTEB", que sería más significativo en términos intuitivos. La tabla G siguiente es una ilustración:

Tabla G

	r = 0.32		
	Mejora sustancial	Sin mejora sustancial	
Psicoterapia	66	34	100
Sin psicoterapia	34	66	100
	100		100

Los números pueden ser interpretados como %'s. Hay una correlación r=0.32 entre recibir o no psicoterapia y exhibir al menos alguna mejora sustancial de la condición del paciente en un gran número de estudios. Con ese coeficiente de correlación, los investigadores se sentirían tentados a decir que ésta es una mejora modesta. Tomando el cuadrado de *r*, a fin de hallar la proporción de la variación en los síntomas mejorados por la psicoterapia, la relación es aún más modesta: 10% de varianza de los síntomas aparece vinculada al tratamiento. Empero, la tabla muestra que esta correlación (0.32) es equivalente a cambiar la tasa de curación del 34% al 66%.

Hay investigaciones como las realizadas en epidemiología que se manejan con relaciones más sutiles. Por ejemplo, consideren la correlación entre tomar o no aspirinas y tener o no un ataque cardíaco. La figura de la tabla H tiene datos sobre los que se basa esta correlación.

Tabla H

	RR=0.5499		Ataques
	Ataque	Sin ataque	per 1.000
Aspirinas	104	10,993	9.42
Placebo	189	10,845	17.13

Una mayoría de gente estudiada no tuvo ataques, tomaran o no aspirinas. La correlación fue .033 y la proporción de la varianza explicada fue de .001 (un décimo de 1%). Aún DTEB muestra que hay sólo un pequeño cambio en la probabilidad de supervivencia (una reducción de los ataques de 52.6% a 47.4%). Pero en esta área de investigación, encontramos que las relaciones expresadas en términos de "riesgo relativo" (RR), esto es, la verosimilitud de que una persona en un grupo expuesto sufra esa condición en comparación con un miembro del grupo no expuesto, es RR=.55, usado como una medida de protección (la exposición reduce la probabilidad de la condición). Si los datos son puestos de esta forma, se puede afirmar que la aspirina reduce el riesgo de un ataque cardíaco a la mitad, lo que suena bastante distinto a decir que está asociado a la reducción

en un 1% de la varianza de los ataques cardíacos por la ingesta de aspirina. Empero, ambos enunciados son precisos.

5) Relaciones causales

Lo más importante que debe recordarse de lo que vimos hasta aquí es que *la correlación no implica causa*. Empero, el primer paso para establecer una relación causal a menudo es extraer una inferencia causal de los datos de correlación.

La primera reacción suele ser examinar la correlación, que si es particularmente grande nos lleva a argumentar en términos de causa-efecto. Este argumento puede fracasar. Se han encontrado muchas correlaciones espurias, debidas frecuentemente a "otras variables". Recíprocamente, relaciones realmente causales presentan distinto grado de fuerza y por consiguiente pueden tener correlaciones de distinta magnitud. ¿Cómo establecer la verdadera causalidad?

Han sido desarrolladas diversas técnicas estadísticas para ayudar a extraer una mejor inferencia causal a partir de datos de correlación. Estas técnicas implican "controlar las variables que pueden ser potencialmente confusivas", incluyendo correlaciones parciales,[20] correlación de paneles cruzados, y otras. No Vamos a analizarlas aquí. Basta con decir que requieren un análisis cuantitativo sofisticado que apunta a extraer los efectos de las variables posiblemente confusivas. Involucran la expectativa de que han sido medidas todas las variables importantes extrañas de manera que sus efectos puedan ser estadísticamente controlados (es decir, eliminarlos de las influencias), para apreciar sólo los efectos reales de las variables independientes interesantes. Al fin y al cabo, las inferencias causales obtenidas de datos de correlación siempre deben ser tratadas con respeto, por la posibilidad de que los ajustes estadísticos hayan sido incorrectos, o porque las verdaderas variables causales hayan sido omitidas del modelo.

Problemas de validez interna Por diversos motivos la solución más simple y poderosa del problema de sacar una inferencia causal es diseñar un estudio de manera de inferir la causa de modo directo, sin recurrir a un arsenal de soluciones. De hecho, los diseños de investigación varían en la medida que permiten extraer inferencias no confusas sobre las causas reales de la variable dependiente. La estructura lógica de un diseño de investigación es conocida como su "validez interna". Los diseños que minimizan los problemas de validez interna permiten extraer inferencias más claras de causalidad. Los que registran escasa validez interna, por el contrario, no permiten extraer inferencias causales razonables.

Si existen diseños con validez interna ¿por qué los investigadores no los utilizan siempre, evitando así la necesidad de manipulaciones matemáticas para eliminar la confusión en sus datos? Cuando las circunstancias lo permiten los investigadores competentes están ansiosos de aplicar tales diseños. Pero puede haber razones de practicidad o éticas que hagan que los mejores diseños de investigación no sean posibles. Los diseños que permiten las inferencias más fuertes son los verdaderos experimentos, que exigen que la gente o las cosas sean aleatoriamente asignadas a condiciones de tratamiento muy distintas. Pero por ejemplo los astrónomos no tienen la capacidad de asignar en forma aleatoria distintos cuerpos celestes a los grupos experimental y de control. Los deben tomar tal como están. Los investigadores de los efectos biológicos de sustancias tóxicas no pueden indicar a la gente que pase su vida expuesta a ciertas sustancias, y a otra

[20] El *coeficiente de correlación parcial*, denotado como $r_{AB.C}$, permite conocer el valor de la correlación entre dos variables A y B, si la variable C ha permanecido constante para la serie de observaciones consideradas. Dicho de otro modo, el coeficiente de correlación parcial $r_{AB.C}$ es el coeficiente de correlación total entre las variables A y B cuando se les extrae su mejor explicación lineal en términos de C.

gente que no se exponga. Por otro lado, los investigadores en campos tales como la medicina, la física (no la astrofísica o geofísica), la psicología, y la agricultura, entre otros, se pueden dar el lujo de tener diseños experimentales más poderosos.

Si hay problemas de validez interna, nos estamos refiriendo a eventos a los que la gente o las cosas están expuestos en forma diferencial además de a la variable independiente. Como no es posible descartar que otras hipótesis rivales sean verdaderas, las inferencias de causalidad son ambiguas. ¿Cuáles son las amenazas eventuales a la fuerza y debilidad de distintos diseños, tomando en cuenta su capacidad relativa de permitir inferencias causales exentas de ambigüedad? Para entender las amenazas a la validez interna, podemos ejemplificar con un estudio en el que a un grupo de pacientes se le administra un nuevo tratamiento.

[1] *Historia* La historia amenaza la validez interna referida a eventos a los cuales la gente o las cosas han estado sujetas, en forma adicional a la variable independiente. P.ej., si los pacientes que reciben el tratamiento tienen una dieta especial y determinada terapia física, no es posible inferir sin ambigüedad qué afecta los resultados de los pacientes: si la dieta o la terapia física.

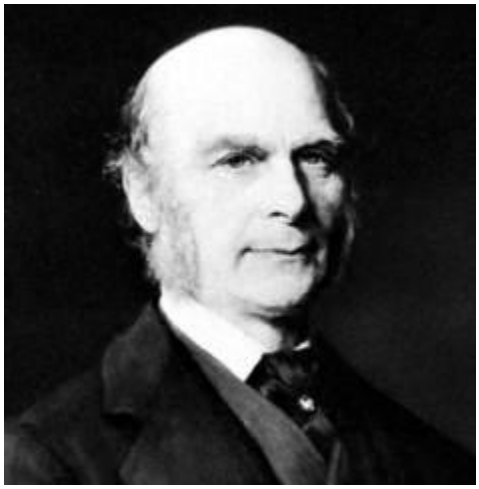
[2] *Maduración* Son procesos que se desarrollan dentro de quienes responden, tales como ser más viejo, estar más cansado o más hambriento. P. ej., los procesos naturales curativos del cuerpo se confunden con los aplicados (de lo cual proviene el dicho de que los médicos deben apurarse a tratar al paciente antes de que la enfermedad se cure de por sí).

[3] *Testeo* La segunda vez que una persona pasa un test será distinta de lo que hubiera sido, por el mero hecho de los cambios en la experiencia del test de la persona. Los cambios debidos a estos efectos no deben ser confundidos con los cambios entre el primero y el segundo test debidos a la incidencia de la variable independiente.

[4] *Instrumentación* La instrumentación se refiere a cambios que tienen lugar en los instrumentos de medición, incluyendo a los observadores humanos, que pueden ser interpretados en forma errónea por cambios de la gente o de las cosas que se observan. P.ej., los pacientes que son atendidos por enfermeras en un turno pueden ser evaluados de manera distinta por las de un turno distinto; las diferencias pueden no originarse en los pacientes sino en las enfermeras. Un área entera de investigación psicológico-social se ha desarrollado ilustrando estos “artefactos sujeto-experimentador”. Hay otro ejemplo vinculado con sesgos en las expectativas, en el cual los observadores tienden a utilizar lo que se les da como una expectativa que tenían de verla. Además de afectar sus percepciones, las expectativas cambian su conducta hacia las personas o animales estudiados, y la misma persona estudiada puede cambiar debido a las expectativas sobre los efectos del tratamiento. Un clásico ejemplo de este problema (conocido como el *efecto Hawthorne*) se produjo en una planta industrial. El término se refiere a una serie de experimentos llevados a cabo con trabajadores administrativos de una empresa en 1923-24 por la Western Electric Company of Chicago. No existe un significado preciso para este término, ya que los resultados fueron un rompecabezas para los que hicieron el experimento original, y de vez en cuando se continúan debatiendo distintas interpretaciones. Las referencias a este experimento conciernen al resultado de que los participantes son conscientes de estar sometidos a una intervención. Hay distintos mecanismos posibles, y todos pueden ser importantes en instancias particulares. Lo que está fuera de cuestión es que aquí hubo una cuestión importante, que debe estar reflejada mediante algún término. Una definición posible del efecto es: un efecto experimental en la dirección esperada, pero no por las razones esperadas; esto es, un efecto positivo significativo que resulta no tener base causal en los motivos teóricos para la intervención, pero que se debe aparentemente al efecto que sobre los participantes tiene el saber que son

estudiados en conexión con los resultados medidos. Parsons[21] lo definió de la forma siguiente: *A partir de una generalización de la situación de Hawthorne, definiría al Efecto Hawthorne como la confusión que tiene lugar si los que practican el experimento no logran apreciar cómo afectan las consecuencias de los resultados de los sujetos lo que hacen estos sujetos*. Sin embargo, éste es un efecto de la motivación y del aprendizaje, y por tal motivo no es necesario un término nuevo, ya que la propensión universal de los humanos a aprender es una "amenaza constante".[22] Lo que sucedía es que en parte los empleados prosperaban en sus tareas poniendo atención, y en parte porque compartían las expectativas de los investigadores de que los cambios introducidos mejorarían las circunstancias y el trabajo de los empleados. A causa de estos efectos de expectativa varios estudios son realizados "a ciegas", es decir sin que la persona que recibe un tratamiento experimental (por ejemplo, una medicina) sea informada si lo que está tomando es la sustancia activa o un placebo. Cuando las circunstancias lo permiten, los estudios son realizados de manera "doblemente ciega": no solamente son "cegados" quienes experimentan el tratamiento, sino también los investigadores que lo administran.

[5] *Regresión Estadística* Ésta ocurre cuando los objetos estudiados, incluyendo la gente, se seleccionan a causa de sus valores extremos en alguna medida. Si no se hace ningún tratamiento de los datos, éstos tienden a una "regresión" hacia la media de la distribución de la cual son tomados. Por consiguiente, los cambios observados debidos al efecto de regresión pueden ser erróneamente tomados como cambios debidos al tratamiento. Sir Francis Galton, primo de Charles Darwin, indicó la necesidad de recurrir a métodos estadísticos para contrastar teorías; en 1889 introdujo el concepto de "línea de regresión" en un estudio comparando estaturas de padres e hijos.[23] En el análisis descriptivo de los datos Galton observó que los padres altos tenían hijos altos pero no tan altos en promedio y que los padres bajos tenían hijos bajos pero no tan bajos en promedio; se producía lo que él denominó una regresión a la media. Galton generalizó esta tendencia bajo la *ley de la regresión universal*: Cada peculiaridad en un hombre es compartida por sus descendientes, pero en promedio, en menor grado.


Sir Francis Galton (1822-1911)

[6] *Selección* La selección confusiva se produce cuando alguna gente u objetos son colocados en condiciones experimentales y de control que difieren de alguna forma al comienzo. Por ejemplo, supongan que la gente relativamente más saludable es elegida para un tratamiento médico experimental (porque se cree que pueden soportar mejor los rigores del tratamiento) y la gente menos saludable que tiene la misma enfermedad es utilizada como grupo de comparación. Distinguir los efectos del tratamiento de diferencias en la salud será difícil o imposible.

[7] *Mortalidad experimental* La mortalidad experimental se refiere a la pérdida diferencial de participantes en diferentes condiciones experimentales. Introduce el mismo problema en un estadio más tardío que un artificio de selección introduce desde el principio de un estudio.

[21] H.M. Parsons, (1974) What happened at Hawthorne? Science vol.183, pp.922-932.
[22] Esto es lo que puntualiza Stephen W. Draper, The Hawthorne, Pygmalion, Placebo and other effects of expectation: some notes, 2009. http://www.psy.gla.ac.uk/~steve/hawth.html
[23] Francis Galton, Natural Inheritance, London: MacMillan, 1889. http://galton.org/books/natural-inheritance/pdf/galton-nat-inh-1up-clean.pdf

[8] *Interacciones* Algunas confusiones ya definidas pueden operar en forma conjunta y oscurecer por consiguiente aún más los resultados. Una es la posible interacción entre selección y maduración: la gente seleccionada en distintas condiciones dentro de un estudio difiere en las condiciones experimentales y de control porque la de un grupo “madura” a una velocidad diferente o en menor grado que la del otro grupo. Ejemplos: los pacientes de diferente edad o condición es más probable que se curen a una velocidad diferente, los niños de edades diferentes aprenderán a una velocidad diferente, las naciones en diferentes estadios de desarrollo económico pueden reaccionar a nuevos desafíos con una tasa de éxito diferente.

[9] *El azar* El azar es un mecanismo que genera fluctuaciones aleatorias en el muestreo y la medición. Dos grupos pueden ser idénticos en todos los aspectos, y al fin de cuentas no habrá idénticas mediciones efectuadas en el estudio. Una diferencia al azar puede ser erróneamente tomada como una diferencia debida a la variable independiente. Si bien los resultados pueden ser protegidos en todos los mecanismos anteriores, el azar es el único que no puede serlo. Para conducirse con el mecanismo del azar se inventaron los tests de significatividad.

[10] *Distorsiones de las Variables Independientes y Dependiente* Al tratar de las cosas que interfieren con la capacidad de investigación para permitir extraer inferencias válidas sobre los efectos posibles de una variable independiente sobre una variable dependiente, no deben ser dejados de lado los problemas con estas últimas variables ([1] a [9] definían tipos de confusiones originados en *variables distintas*).

Definiciones operativas defectuosas Recordar que si no se operacionaliza significativamente la variable, los hallazgos producidos no darán respuesta a la pregunta que el usuario de la investigación pensaba hacer.

Errores en inducir la manipulación experimental Incluso una variable bien operativa puede no alcanzar a quienes responden o puede no haberles sido presentada, o si lo fue, puede no haber sido percibida de la manera esperada. Ejemplos: en un estudio sobre los efectos de sustancias tóxicas, se presupone que un grupo estuvo expuesto a las mismas cuando en realidad no lo fue. Un programa de rehabilitación de prisioneros pudo no haber sido llevado a cabo. Un estudio puede tratar de examinar cómo reacciona la gente a los defensores “ricos” de una sociedad, en tanto que la gente estudiada no percibe que los defensores sean más ricos que el promedio. En esas instancias, es inexacto concluir que la variable independiente no tuvo efectos si esa variable independiente, en realidad, no fue testeada. Los investigadores que se preocupan por estos problemas hacen a menudo “chequeos de manipulación” para ver si las variables independientes lograron el valor inducido buscado.

Efectos piso y techo Esta situación se da cuando hay una manipulación tan extrema de la fuerza o debilidad tal que no hay variación posible entre grupos – es decir, que gente de todos los grupos responde a la variable de la misma forma – resultando imposible testear los efectos de alguna variable sobre los mismos. Por ejemplo, todo estudio de los efectos sutiles sobre los veredictos de innumerables variables que tienen efecto sobre las decisiones del juez o del jurado sería incapaz de detectar diferencias si los hechos del caso básico son tan extremos que cualquiera que analizara el caso alcanzara el mismo veredicto. Lo que haría un buen investigador es tratar de efectuar una prueba previa de sus procedimientos para asegurarse que no haya efectos piso o techo y que exista una variabilidad apropiada de la variable dependiente.

Problemas de validez externa Una vez satisfechos de que un estudio sea internamente válido, la cuestión siguiente es si lo es externamente, lo que significa si es posible generalizarlo más allá de sí. Esto no es arbitrario: sin validez interna no habría nada que generalizar.

La validez externa se refiera a la *representatividad* del estudio. Un estudio externamente válido puede ser generalizado a otras poblaciones (de gente, objetos, organizaciones, momentos, lugares, etc.). Lo usual es hacer una investigación es un momento y un lugar específico sobre una población particular, pero se espera que sea posible generalizar los hallazgos más allá de la gente y circunstancias inmediatas del estudio.

Ejemplos: ¿Es posible hallar un estudio hecho en Córdoba que sea generalizable a la Capital Federal? La eficacia informativa de un estudio escrito ¿es generalizable a la misma información presentada en forma verbal, en video o en computadora? Un estudio sobre la organización industrial ¿es aplicable a la industria de servicios financieros? Si se expone a la gente a un conjunto de tratamientos ¿actuará cualquiera de los mismos sobre gente no expuesta a los restantes? ¿Es posible generalizar a humanos los hallazgos sobre los efectos de una medicación probada en ratas de laboratorio? ¿O en monos? ¿O de adultos sobre los niños? ¿O de los machos sobre las hembras?

Como ilustración de este problema, vamos a analizar un caso particular: los efectos reactivos son los efectos inductores de cambios en una persona que es testeada. P.ej. supongamos que una empresa desea testear los efectos de una campaña publicitaria sobre la actitud de la gente. Si comienza testeando las actitudes del público y encara llamar a los que responden luego de la campaña publicitaria por cierto período, la prueba previa puede dar lugar a que los que responden presten más atención a la publicidad que otra gente del público que no había sido testeada en forma previa. Por consiguiente, los hallazgos sólo pueden ser generalizados a gente que fue entrevistada acerca del tópico en forma previa a la publicidad emitida.

La lógica de la investigación científica es de gran ayuda para evaluar la validez interna de la investigación, pero la validez externa no puede ser tratada en forma ligera. Con relación a cualquiera de las dos preguntas planteadas más arriba, uno puede ejercer su intuición: ¿cuán similares creemos que son distintos contextos o tipos u organismos (con respecto a las variables independiente y dependiente interesantes)? Pero se trata de meras conjeturas. Al fin y al cabo, la única manera de saber con rigurosidad si un efecto observado en Córdoba también será válido en la capital, o si el efecto observado en las ratas se mantendrá en los humanos, es poder replicarlos. Ésta es una de las razones por las que los investigadores depositan demasiada confianza en un único estudio, o aún un único tipo de estudios. Entre otros motivos, en razón de la generalidad, prefieren apreciar hallazgos replicados en otros lugares, usando otros participantes, bajo otras condiciones. A mayor diferencia de circunstancias bajo las cuales sea replicado un fenómeno, y cuanto mayor sea su generalidad, tanta mayor confianza tendrán los investigadores y los consumidores de la misma en el fenómeno.

6) Diseños de investigación

Varios diseños de investigación son más o menos vulnerables a problemas de validez interna y externa. Sin el objetivo de agotar la cuestión se mencionan algunos diseños que permiten apreciar la capacidad de un diseño para proporcionar una respuesta, y la debilidad de otros.

Supongan que el proyecto de investigación es estudiar si la Vitamina C cura o no el resfrío. A continuación se enuncia una variedad de enfoques que podría adoptar un investigador para

enfrentar esta cuestión empírica. Mediante estos ejemplos veremos la fuerza y la debilidad de los diferentes diseños de investigación. Se pondrá énfasis en los problemas de validez interna.

Diseños pre-experimentales En primer lugar, supongan que el investigador emplea un *estudio de casos*. A una persona resfriada se le suministra Vitamina C, y poco después desaparece el resfrío. ¿Es ésta una prueba convincente del poder curativo de la Vitamina C? Piensen en las variables confusivas que pueden haber actuado durante el estudio: otros factores podrían haberlo curado, tales como la sopa de caldo de pollo o el reposo en la cama que el sujeto pudo haber tenido (historia). En lugar de la Vitamina C, el paciente bien pudo haber sido curado por su sistema inmunológico (maduración). A éstos debemos sumar la falta de generalidad al tratar con un estudio de tamaño muestral de n=1.

Entonces, supongan que el investigador reúne a 100 personas para realizar esencialmente el mismo estudio. Se trata de un *diseño de un grupo pre-test post-test*. Al principio todos están resfriados, se les suministra vitamina C, y una semana después se comprueba que sólo un 40% continúa resfriado. ¿La reducción de 100% a 40% se debe a la Vitamina C? Las variables confusivas que había antes no han desaparecido. Adicionalmente, acaso los síntomas de resfrío se volvieron tan insensibles a las mismas desde t_1 hasta t_2 en que fue preciso registrar peores síntomas para considerar que uno está resfriado (instrumentación). Acaso aquellos cuyo resfrío se transformó en una neumonía y ahora pasaron a estar internados en un hospital o que abandonaron el estudio son los que no pudieron ser hallados, exagerando así la tasa de curación (mortalidad). Por las confusiones de diseño, el estudio deja de convencernos.

Nuestro persistente investigador trata a continuación de realizar una *comparación estática de grupos*. En ese estudio, a la gente la encuentra tomando en forma regular vitamina C o no, y sus resfríos son monitoreados a lo largo del tiempo. Supongan que halla una relación indicativa de que los que toman vitamina C tienen menos resfríos y que los resfríos que sufren son más benignos y más cortos. ¿Podemos deducir que la vitamina C cura (y previene) los resfríos? Ustedes pueden reconocer en éste un estudio muy primitivo de correlaciones (o de observaciones) ya discutido. Aún resulta vulnerable a un conjunto de problemas de validez interna. Los que toman vitamina C en forma regular deben diferir en forma sistemática de los que no lo hacen en cuanto a otras cosas hechas: nutrición, ejercicio, descanso (la historia confunde). Quienes toman vitamina C en forma regular pueden ser, a causa de su constitución u otros hábitos sanitarios, básicamente más sanos que los demás (selección), o tener un sistema inmune que mata al virus de la gripe más efectivamente (interacción de selección y maduración). Observen que, en este ejemplo, la instrumentación no aparece como un problema, porque ambos grupos están siendo examinados aproximadamente al mismo tiempo, por lo cual los problemas serían similares para ambos.

Experimentos verdaderos Comparemos ahora el diseño anterior con un verdadero experimento. ¿Qué es un experimento? En el habla cotidiana, un experimento es simplemente algo que es probado o ensayado. Para un científico, *experimento* es un procedimiento mediante el cual se trata de comprobar (confirmar o verificar) una o varias hipótesis relacionadas con un determinado fenómeno, mediante la manipulación de la/s variables que presumiblemente son la causa. En un experimento se consideran todas las variables relevantes que intervienen en el fenómeno, mediante la manipulación de las que presumiblemente son su causa, el control de las variables extrañas y la aleatorización de las restantes. Estos procedimientos pueden variar mucho según las disciplinas (no es igual en física que en psicología, por ejemplo), pero persiguen el mismo objetivo: excluir explicaciones alternativas (diferentes a la variable manipulada) en la explicación de los resultados. Cada repetición del experimento se llama prueba o ensayo. Para los científicos sociales, de la conducta y biólogos, llevar a cabo un experimento requiere *asignar en forma*

aleatoria los participantes del estudio a condiciones experimentales y de control de modo de maximizar la probabilidad de que ambos grupos *no difieran*.

A los experimentos verdaderos se los conoce con diversos sinónimos. Pueden ser llamados, simplemente, experimentos. En medicina a menudo son denominados “ensayos controlados randomizados”, o “ensayos controlados”, o “ensayos clínicos”. Cualquiera sea su nombre, significa que la gente es asignada en forma aleatoria a exponerse a diferentes condiciones o tratamientos, de manera que las variables confusivas son eliminadas y los grupos sólo difieren con respecto a la variable independiente. Entonces, las inferencias de causa y efecto no serán ambiguas.

Veamos cómo funciona con la investigación sobre la vitamina C. Supongan que 200 enfermos de resfrío son asignados aleatoriamente a dos grupos. Uno de esos grupos recibe vitamina C, en tanto que el otro no recibe nada, o simplemente un placebo. (Sería ideal que la vitamina C y el placebo sean administrados de manera doblemente ciega, para que ni la persona que administra ni la persona que recibe la droga sepa cuál es la vitamina y cuál el placebo. Los números de identificación luego serían decodificados por los investigadores). Una semana después la gente es examinada hallándose, por ejemplo, que 50% del grupo de control aún está resfriada, pero sólo un 40% del grupo de la vitamina C lo está. Ante esta situación, la diferencia de un 10 por ciento sugiere que la vitamina C hizo mejor que el placebo.

Aún más: como se utilizó un verdadero diseño experimental, las variables confusivas fueron eliminadas. *Historia*: Como ambos grupos se crearon aleatoriamente, la proporción de los que tomaban sopa de pollo y se quedaban en cama es probablemente similar en ambos grupos. *Maduración*: La respuesta del sistema inmunológico será similar, en promedio, en ambos grupos. Se observa que aún cuando el grupo de control haya perdido a la mitad de sus enfermos por remisión espontánea (sanación no debida a tratamientos médicos convencionales), ello no pudo distinguirse del supuesto efecto de la vitamina C en el grupo de diseño pre-test post-test, y la diferencia de 10 puntos porcentuales representa el efecto de la vitamina C por encima y más allá de la maduración. *Instrumentación*: Si los pacientes fueron enviados para exámenes al azar antes y después de administrar la variable independiente, todo cambio de percepción de los examinadores se distribuirá de modo uniforme[24] entre los grupos experimental y de control y no será causa de diferencia de hechos. *Selección*: Si los adultos jóvenes hubieran sido ubicados en el grupo de la vitamina C y los más viejos en el grupo de control, o si a la gente se le hubiera dado la opción de ubicarse por sí misma en los grupos de la vitamina C y de control, cualquier diferencia observada podría haberse originado en las diferencias de la gente en ambos grupos. Mas la asignación aleatoria de la gente maximiza la probabilidad de que, en cualquier dimensión, se tendrá la misma proporción de gente en un grupo como en el otro, y estas diferencias no podrían causar un efecto sobre un grupo pero no sobre el otro. *Mortalidad*: En forma similar, se espera que la gente que sale de uno de los dos grupos lo haga a la misma velocidad y no en forma diferencial. Esto será cierto para cualquier característica que identifiquemos.

Cuasi-experimentos Éstos son diseños de investigación que no son tan limpios y directos como los experimentos verdaderos, pero que facilitan suficiente control, cierta randomización, o que pueden ser complementados o corregidos para eliminar muchos, y ocasionalmente todos, los problemas de validez. El término “experimento” usualmente implica un experimento controlado, pero hay veces que el control resulta prohibitivamente difícil o imposible. En tal caso los investigadores recurren a experimentos naturales o cuasi-experimentos.

[24] Una distribución uniforme (discreta) es una distribución de probabilidad discreta que se caracteriza por el hecho de que todos los valores de un conjunto finito de valores posibles son equi-probables. http://mathworld.wolfram.com/UniformDistribution.html

Los experimentos naturales descansan exclusivamente en observaciones de las variables del sistema estudiado, en lugar de la manipulación de sólo una o algunas pocas variables como sucede en los experimentos controlados. En lo posible, se trata de recoger datos del sistema de modo que pueda determinarse la contribución de todas las variables, y cuándo los efectos de la variación de algunas variables permanecen aproximadamente constantes de modo de poder discernir los efectos de las demás variables. El grado en que esto sea posible depende de la correlación entre variables independientes en los datos observados. Si las variables independientes no guardan demasiada correlación entre sí, los experimentos naturales pueden alcanzar la potencia de los experimentos controlados. Empero, es usual que haya cierta correlación entre estas variables, lo que reduce la confiabilidad de los experimentos naturales con relación a lo que podría concluirse si pudiera realizarse un experimento controlado. Además, como los experimentos naturales habitualmente suceden en entornos no controlados, las variables de fuentes no detectadas no pueden ser medidas ni mantenerse constantes, lo que puede producir correlaciones ilusorias de las variables estudiadas.

Hay mucha investigación en importantes disciplinas científicas, incluyendo economía, ciencia política, geología, paleontología, ecología, meteorología, y astronomía, que descansan en cuasi-experimentos. Por ejemplo, en astronomía resulta claramente imposible, al testear la hipótesis “los soles son nubes colapsadas de hidrógeno”, comenzar la investigación con una nube gigante de hidrógeno, realizando a continuación el experimento de esperar algunos miles de millones de años hasta que se forme una estrella (sol). Sin embargo, observando varias nubes de hidrógeno en distintos estados de colapso, además de otras implicancias (p.ej., la presencia de distintas emisiones de los espectros de luz estelar) podemos recolectar los datos requeridos para sostener la hipótesis. Un ejemplo temprano de este tipo de experimento fue la primera verificación en los años 1600s de que la luz no se traslada de un lugar a otro en forma instantánea, sino a determinada velocidad. La observación de las caras de las lunas de Júpiter se retrasaba levemente cuando Júpiter se hallaba más alejado de la Tierra, en oposición a cuando estaba próximo a la misma; y este fenómeno fue utilizado para demostrar que la diferencia en el tiempo de aparición de las lunas era consistente con una velocidad medible. Los estudios de casos controlados, que se hacen en investigación epidemiológica, pueden ser vistos como miembros de la clase de diseños de investigación cuasi-experimentales. Para cada caso del grupo expuesto a la sustancia de interés, se elige un caso de comparación similar en diversas características, con la excepción de que el caso de control no fue expuesto a la sustancia. Debe resultar obvio que éste es un intento de reducir los efectos de las posibles variables confusivas, tratando de aproximarse a un experimento verdadero.

En economía y otras ciencias sociales se han realizado últimamente con frecuencia experimentos naturales. Estos estudios examinan los resultados de observaciones en grupos intervenidos y los comparan con grupos que no han estado expuestos al tratamiento. En estos estudios existe variación exógena en las variables que determinan la asignación de la intervención. Generalmente son cambios políticos (Dynarski, 1999)[25] u otros eventos administrativos los que permiten al investigador obtener variación exógena en las principales variables explicativas (F. Barceinas, J. Oliver, J. Raymond, & J. Roig, 2000).[26] Este fenómeno es especialmente útil en situaciones donde las estimaciones están sesgadas debido a variaciones endógenas provenientes de variables

[25] Susan M. Dynarski, Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion, NBER, November 1999. http://www.nber.org/papers/w7422

[26] Fernando Barceinas, Josep Oliver Alonso, José L. Raymond and José L. Roig, Rendimiento público de la educación y restricción presupuestaria, Papeles de Economía Española, 86, 2000, 236-248. http://www.etla.fi/PURE/Rendimiento.pdf

omitidas o sesgos de selección. Estas estrategias están siendo usadas muy extensamente para analizar un amplio rango de problemas (Garruto et al.;[27] Petticrew et al.,[28]). Otras situaciones favorables para ser utilizarlas son los cambios en las políticas de gobierno, que con frecuencia son aplicadas en algunos grupos de la población pero no en otros. Hay una reseña reciente de las aplicaciones de experimentos naturales en economía (M. Rosenzweig & K. Wolpin).[29]

Los estudios de observaciones no son experimentos. Por definición, estos estudios carecen de la propiedad de manipulación requerida por los experimentos que requería Francis Bacon.

Los estudios de observaciones también carecen de las propiedades estadísticas de los experimentos randomizados. En un experimento randomizado, el método especificado en el protocolo experimental guía el análisis estadístico, que usualmente también está especificado en el protocolo experimental. Con un experimento randomizado, la propia randomización proporciona el modelo estadístico utilizado en la inferencia.[30] Si se carece de un modelo estadístico que refleje una randomización apropiada, el análisis estadístico descansa en un modelo subjetivo.[31] Las inferencias derivadas a partir de un modelo subjetivo no son confiables ni a nivel teórico ni práctico. De hecho, hay varios casos en que los estudios de observaciones bien hechos dan resultados incorrectos, o sea que los estudios de las observaciones son inconsistentes y también difieren de los resultados experimentales. Por ejemplo, los estudios epidemiológicos de una alimentación con brócoli y cáncer de colon hallan en forma consistente resultados benéficos, en tanto que los experimentos no hallan beneficio alguno (Freedman, ob. cit., ch. 1).

Otro problema es que los estudios de observaciones tienen grandes dificultades en lograr una comparación equitativa de tratamientos (o exposiciones), porque los grupos que los reciben difieren mucho según la variable independiente. En contraste, la randomización implica que, para cada nivel de la variable independiente, se espera que la media de cada grupo sea la misma. Para un ensayo randomizado, se espera alguna variabilidad de la media, naturalmente, pero la propia randomización asegura que los grupos experimentales tienen medias próximas entre sí, debido al Teorema Central del Límite y a la desigualdad de Markov (temas sobre los que volveremos a hablar). Como no hay randomización, a veces se da el caso de que los estudios de observaciones exhiben elementos confusivos. Esto es, la variación sistemática de las variables independientes entre distintos grupos de tratamiento (o de exposición) torna difícil separar el efecto del tratamiento (exposición) de los efectos de otras variables independientes, que en su mayor parte pueden no haber sido medidos. En resumen, un estudio de observaciones carece de garantías de una equivalencia probabilística entre los grupos expuestos.

Los resultados de los estudios de observaciones son considerados menos convincentes que los de experimentos diseñados porque están expuestos al sesgo de selección. Los investigadores tratan de reducir los sesgos de los estudios de observaciones mediante métodos estadísticos

[27] R. M. Garruto, M. A. Little, G. D. James, and D. E. Brown, Natural experimental models: The global search for biomedical paradigms among traditional, modernizing, and modern populations, 1999, Proceedings of the National Academy of Sciences of the United States of America 96(18), 10536-10543. http://www.pnas.org/content/96/18/10536.full

[28] M. Petticrew, S. Cummins, C. Ferrell, A. Findlay, C. Higgins, C. Hoy, A. Kearns, & L. Sparks, Natural experiments: an underused tool for public health? Public Health (2005) 119, 751–757. http://mres.gmu.edu/pmwiki/uploads/Main/Petticrew2005.pdf

[29] M. Rosenzweig & K. Wolpin, (2000). Natural “natural experiments” in economics. Journal of Economic Literature. 38, 827-874. http://www.uh.edu/~adkugler/Rosenzweig&Wolpin.pdf

[30] David A. Freedman et al., Statistics, 4th ed. (W.W. Norton & Company, 2007).

[31] David A. Freedman (ob.cit); Klaus Hinkelmann and Oscar Kempthorne (2008), Design and Analysis of Experiments, Volume I: Introduction to Experimental Design (Second ed.). Wiley

complicados, como los métodos de apareamiento de las tasaciones de propensión, que requieren amplias muestras de sujetos y gran información sobre las variables independientes.

Aspectos Estadísticos Volvamos a los casos de gripe que ya hemos visto. En este caso, el factor confusivo del azar no puede menospreciarse, aún en el más ideal de los experimentos. La diferencia entre 40% (pacientes del grupo experimental que todavía tienen gripe) y 50% (pacientes del grupo de control que todavía tienen gripe) ¿es una diferencia *real* o se trata simplemente de una fluctuación aleatoria? A esta pregunta se la puede responder mediante el *test de hipótesis estadísticas*. Los datos permiten al investigador calcular la probabilidad de cuál es, si se rechaza el supuesto de partida de que no hay diferencia entre los grupos experimentales y de control (llamada hipótesis *nula*), la probabilidad de que el investigador cometa un error de Tipo I (rechazar erróneamente una hipótesis nula verdadera). Por convención, a menos que esta probabilidad caiga por debajo de .05 (o sea, menos de 5 oportunidades sobre 100 de cometer un error de Tipo I), el investigador se abstiene de rechazar la hipótesis nula (el tema será tratado más adelante).

7) Conclusión

La ciencia no es algo mecánico ni mágico. Es un proceso de extraer inferencias de la evidencia disponible. Esta evidencia es generada por la investigación que emplea necesariamente una selección de un método de investigación. Como se dijo antes, un resultado es sólo tan bueno como los métodos utilizados para hallarlo. No hay una forma óptima de estudiar un fenómeno interesante. Cada elección metodológica implica intercambiar algo por otra cosa (un *trade-off*). La cuestión será siempre si la metodología de investigación es apropiada para las preguntas planteadas en el estudio, y si las conclusiones extraídas son justificables a la luz de los datos recolectados y lo que se sabe de los procesos de generación de los datos. La elección del método de investigación requiere pensar en forma cuidadosa, tanto de parte de los investigadores como de los consumidores de esa investigación. El propósito de este capítulo introductorio ha sido brindar una visión a vuelo de pájaro a los consumidores legales de investigación científica por medio de conceptos que facilitarán su apreciación crítica y detenida.