TEORÍA ESTADÍSTICA[1]

1. Introducción

La estadística es el arte y la ciencia de obtener información a partir de los datos. A efectos estadísticos, un dato significa una observación o medición, expresada en un número. Una estadística puede referirse a un determinado valor numérico derivado de los datos. Por ejemplo, las estadísticas del fútbol de 1ra división de Argentina consisten en el estudio de datos sobre ese juego; en cambio, el promedio de tiros al arco contrario de un equipo de fútbol es un estadístico (también llamado estadígrafo). La estadística incluye tres campos: métodos para 1) recopilar los datos; 2) analizarlos, y 3) obtener inferencias a partir de los mismos. La evaluación estadística es muy relevante en diversos casos, que van desde las leyes y regulaciones anti-monopolio hasta los derechos políticos de una población. Razonar en términos estadísticos puede resultar crucial para interpretar tests (o contrastes) psicológicos, estudios epidemiológicos, el tratamiento diferencial a los empleados de una empresa, y la toma de huellas dactilares de ADN, por mencionar algunas aplicaciones.

En este capítulo, siguiendo a Kaye y Freedman, se describen elementos del pensamiento estadístico. De tal manera, se espera permitir a los jueces y abogados que trabajan con evidencia científica que entiendan la terminología, ubiquen dentro del contexto apropiado la evidencia, apreciando sus fortalezas y debilidades, y apliquen la doctrina legal que regula el uso de esa evidencia. Analizaremos en primer término cuán admisibles y qué peso debería tener este tipo de estudios, los tipos de estudios estadísticos y los límites de la experiencia en la materia, así como cuáles son los procedimientos que pueden contribuir a otorgar mayor credibilidad al testimonio estadístico. Repasaremos los aspectos centrales de un caso paradigmático (el caso Daubert). Luego nos haremos la pregunta acerca de cómo fueron recopilados los datos, y a renglón seguido analizaremos la distinta forma en que pueden ser presentados. Luego se analizará qué inferencias pueden ser extraídas de los datos, lo cual conduce a un análisis de distintos *estimadores*, de sus errores estándar y de sus intervalos de confianza. A fin de obtener conclusiones sobre significación entraremos en el análisis de los *p*—valores. Recién entonces se hablará sobre los tests (o contrastes) de hipótesis y de las probabilidades posteriores. Finalmente, trataremos de entrar al campo de los análisis de correlación, diagramas de dispersión, y de las líneas de regresión, intentando comprender los conceptos de pendiente y de ordenada al origen; lo cual nos abrirá el camino al dominio de los modelos estadísticos muy usados en ciencias sociales y litigios. Dejamos para un Apéndice una nota técnica sobre probabilidades e inferencia estadística, errores estándar, la función de distribución normal y los niveles de significación.

Admisibilidad y Ponderación de los Estudios Estadísticos Los estudios estadísticos bien diseñados pueden ser de gran ayuda en Derecho, y de hecho en US son admitidos por las Reglas Federales de Evidencia. La *invalidez del testimonio de oídas*[2] pocas veces constituye

[1] Ver David H. Kaye and David A. Freedman, Reference Guide on Statistics, in Reference Manual on Scientific Evidence, 2nd ed., Federal Judicial Center (2000), pp. 83-178 http://www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf; David W. Barnes, A Common Sense Approach to Understanding Statistical Evidence, SSRN; San Diego Law Review, Vol. 21, p. 809, 1984; Seton Hall Public Law Research Paper No. 899773 http://ebour.com.ar/index.php?option=com_weblinks&task=view&id=13718&Itemid=0; Palmer Morrel-Samuels and Peter D. Jacobson, Using Statistical Evidence to Prove Causality to Non-Statisticians, SSRN, July 2007 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=995841; Michael I. Meyerson, Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges, SSRN, 2010 http://ebour.com.ar/index.php?option=com_weblinks&task=view&id=13802&Itemid=0 Recomiendo ver la página Statistical Evidence in Litigation, del Dr. Will Yancey, un contador especializado en litigios, con un amplio acceso a recursos en internet. http://www.willyancey.com/statistical_evidence.htm

[2] Este testimonio es información recogida por una persona de otra con respecto a algún evento, condición o cosa sobre los cuales la primera no tuvo experiencia directa. Al ser sometida como

una barrera para la presentación de un estudio estadístico, dado que estos estudios pueden ofrecerse para explicar la base de la formulación de un experto, o admitidos bajo tratados de excepción de la invalidez del testimonio de oídas. Además, como muchos métodos estadísticos usados en los tribunales figuran en libros de texto y artículos de *journals* y pueden dar lugar a resultados útiles cuando son aplicados de modo cuidadoso y razonable, satisfacen en general aspectos importantes del requerimiento de "conocimiento científico" articulado en el caso *Daubert v. Merrell Dow Pharmaceuticals, Inc.* Naturalmente, un estudio en particular puede utilizar un método adecuado, pero tan mal aplicado que sea inadmisible. También podría darse que el método no sea el adecuado para tratar el problema que debe encararse. Finalmente, el estudio puede descansar en datos que no son confiables para los expertos estadísticos. Empero, frecuentemente la discusión no es tanto sobre la admisibilidad del estudio como sobre la importancia o suficiencia de la evidencia estadística.

2. Apreciación del caso Daubert

Vamos a repasar brevemente este caso, ya introducido en el capítulo previo, que es paradigmático para nuestro tratamiento de la ciencia y el derecho. Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993) es un caso de la Corte Suprema de Estados Unidos que determina el estándar a ser seguido para admitir el testimonio de expertos en tribunales federales. El tribunal Daubert sostenía que la promulgación de las Reglas Federales de Evidencia implícitamente daba vuelta el estándar Frye;[3] al estándar que articuló el tribunal se lo conoce como estándar Daubert.[4] Jason Daubert y Eric Schuller habían nacido con serios defectos de nacimiento. Ellos y sus padres demandaron a Merrell Dow Pharmaceuticals Inc., una subsidiaria de Dow Chemical Company, ante un tribunal del estado de California, afirmando que la medicación Bendectin era causa de los defectos de nacimiento. Merrell Dow llevó el caso ante el tribunal federal, y desde allí buscó un juicio sumario porque sus expertos pusieron a consideración documentación que mostraba que no había estudios científicos publicados desmostrando un vínculo entre Bendectin y los defectos de nacimiento. Daubert y Schuller sometieron evidencia de expertos propia sugiriendo que Bendectin podría causar defectos de nacimiento. La evidencia de Daubert y Schuller, empero, estaba basada en estudios *in vitro* y animales *in vivo*, estudios farmacológicos, y estudios re-publicados, y estas metodologías aún no habían ganado aceptación dentro de la comunidad científica general.

El tribunal de distrito otorgó juicio sumario a Merrell Dow, y Daubert y Schuller apelaron al Noveno Circuito. Éste halló que el juicio sumario otorgado era correcto, porque los demandantes ofrecieron evidencia que aún no había sido aceptada como técnica confiable por los científicos que habían tenido oportunidad de analizarla y de verificar los métodos usados por aquellos científicos. Además, el Noveno Circuito era escéptico con respecto a que la evidencia de los demandantes apareciera como generada para el litigio. Sin la evidencia ofrecida, el Noveno Circuito dudaba de que los demandantes pudieran demostrar en juicio que el Bendectin había causado en realidad los defectos de nacimiento que constituían el objeto de la demanda. Los demandantes solicitaron a la Corte Suprema que revisara la decisión del Noveno Circuito, lo que ésta terminó haciendo.

---

evidencia, es llamada *evidencia de testimonio de oídas*. Legalmente, tiene un significado más estrecho que apunta al uso de esa información como evidencia para probar la verdad de lo que se afirma, y como tal no es aceptado en general por los tribunales. Por ejemplo, un testigo dice "Juan me contó que Pedro estaba en la ciudad". Como el testigo no pudo verlo a Pedro en la ciudad, este enunciado sería invalidado como testimonio de oídas, pero sí sería admitido como evidencia que Juan *le contó* al testigo que Pedro estaba en la ciudad.

[3] Este estándar, Frye test, o test de aceptación general, es un test a efectos de determinar la admisibilidad de evidencia científica en los tribunales de US. Establece que la opinión de expertos basada en técnicas científicas sólo es admisible cuando la técnica sea generalmente aceptada como confiable en la comunidad científica relevante.

[4] http://en.wikipedia.org/wiki/Daubert_v._Merrell_Dow_Pharmaceuticals

Tres disposiciones de las Reglas Federales de Evidencia regulaban cuándo el testimonio de un experto podría ser admitido en un tribunal.[5] La *primera* era que el testimonio debe ser de naturaleza científica, y que dicho testimonio debe estar basado en "conocimiento". Por supuesto, la ciencia no reclama para sí conocer algo con certeza absoluta; la ciencia "representa un proceso de proponer y refinar las explicaciones teóricas sobre el mundo que son tema de dócimas y refinamiento adicional". El "conocimiento científico" contemplado por la Regla 702 era uno al que debía llegarse mediante el método científico.

La *segunda* era que el conocimiento científico debe ayudar al juez (o al jurado en un juicio) a entender la evidencia o a comprender los hechos en la cuestión del caso. El juez del caso es a veces un jurado. Pero pueden existir otros investigadores entre las reglas federales de evidencia. Para resultar de utilidad al juez o a jurado, debe haber una "conexión científica válida con la investigación pertinente como prerrequisito para que sea admisible". Si bien está en el área de competencia del conocimiento científico saber si a la noche la luna estaba llena, puede no ser de gran ayuda para el juez o el jurado a fin de determinar si una persona estaba cuerda cuando cometió un acto determinado.

En *tercer* término, las Reglas permitían en forma expresa que el juez establezca la fijación del umbral a partir del cual determinado conocimiento científico podrá asistir a dicho juez (o al jurado) de la forma contemplada por la Regla 702. "Esto implica una evaluación preliminar acerca de si el razonamiento o metodología que subyace al testimonio es científicamente válido y si puede ser aplicado/a a los hechos en disputa." Esta evaluación preliminar puede depender de si algo pasó un test, si una idea fue sometida a revisión por sus pares o publicada en periódicos científicos, cuál es el margen de error involucrado, e incluso de su aceptación general, entre otros factores. *Se ciñe a cuestiones de metodología y de principios, no a las últimas conclusiones generadas*.

[5] Hay todo un cuerpo normativo denominado Federal Rules of Evidence publicado por la Cornell University (Legal Information Institute, Law School http://www.law.cornell.edu/rules/fre) en el que figura la famosa Regla 702. Esta regla incluye algunas notas que vale la pena transcribir: "A menudo es difícil o imposible realizar una evaluación inteligente de los hechos si no se aplica algún conocimiento científico, técnico, o especializado. La fuente más habitual de este tipo de conocimientos es el testimonio de expertos, aunque hay otras técnicas para ofrecerlo. En buena parte de la literatura se supone que los expertos brindan su testimonio sólo mediante opiniones. Ello no tiene fundamento lógico. La regla por consiguiente reconoce que un experto en su estrado puede brindar una disertación o exposición de principios científicos u otros relevantes al caso en cuestión, dejando al abogado que los aplique a los hechos. Como buena parte de la crítica al testimonio de expertos se ha centrado alrededor de la cuestión hipotética, parece sabio reconocer que las opiniones no solamente no son indispensables y alentar el uso de testimonios de expertos bajo la forma de no-opiniones si el fiscal cree que el abogado defensor puede extraer la inferencia requerida por cuenta propia. Ello no significa eliminar el uso de opiniones, ya que se permitirá a los expertos dar pasos adicionales en pro de sugerir las inferencias que podrían ser extraídas de aplicar el conocimiento especializado a los hechos. (Ver Reglas 703 a 705). Que la situación resulte apropiada para usar el testimonio de un experto debe ser determinado en base a ayudar al abogado defensor... El uso de estos testimonios tuvo un gran incremento a partir de la promulgación de las Reglas Federales de Evidencia. Éste era el resultado buscado por quienes escribieron la regla, que así respondían a preocupaciones de que las restricciones impuestas con anterioridad al testimonio de los expertos fueran artificiales y fueran un impedimento para echar luz sobre cuestiones técnicas en disputa. En tanto que ahora son presentados muchos testimonios de expertos que resultan iluminadores y útiles, no es así con todos. Todo significa un gasto, ya para el que lo propone, ya para el adversario. En particular, en litigios civiles con elevados montos financieros, se volvió un lugar común invertir grandes sumas en testimonios de expertos útiles sólo marginalmente. El recurso al testimonio de expertos en ocasiones es usado como una técnica judicial para vencer la resistencia de los adversarios. En resumen, en tanto que el testimonio de los expertos puede ser deseable si no crucial en varios casos, no puede dudarse de que se han cometido excesos y que deben ser limitados."

La corte subrayó que el nuevo estándar de la Regla 702 tenía raíces en el proceso judicial y que se esperaba que fuera algo diferente y separado de la búsqueda de la verdad científica. “Las conclusiones científicas están sometidas a una revisión permanente. Por otra parte, el Derecho debe resolver las disputas de modo rápido y concluyente. Un proyecto científico avanza mediante la consideración de una multitud de hipótesis, ya que de aquellas que sean incorrectas se demostrará su falsedad, lo cual constituye en sí un adelanto.” La Regla 702 fue pensada para poner término a las disputas legales, y por consiguiente debía ser interpretada conjuntamente con otras reglas de evidencia y otros medios legales de terminar con las disputas. Dentro del proceso entre adversarios, el examen cruzado es apropiado para ayudar a los que deben tomar las decisiones para lograr una culminación eficiente de las disputas. “Se reconoce, en la práctica, que un juez que asume un rol de guardián, aunque sea flexible, a veces impedirá al jurado conocer los puntos de vista auténticos y las innovaciones. Sin embargo, éste es el contrapeso que imponen las Reglas de Evidencia que han sido diseñadas no para la investigación exhaustiva de una comprensión cósmica sino para la resolución particular de disputas legales.”

Después del *affaire* Daubert, se esperaba que el rango de evidencia de opiniones científicas usado en los tribunales se expandiera. Empero, los tribunales siguieron aplicando en forma estricta los estándares de Daubert, y en general actuaron con éxito al excluir “ciencia basura” o “pseudo-ciencia”, así como técnicas e investigaciones nuevas o experimentales que hubieran podido ser consideradas como admisibles. Cabe decir que no todas las consideraciones del caso Daubert deben ser reunidas para que sea admitida la evidencia. Sólo se precisa que la mayoría de las pruebas sea superada de forma sustancial.[6]

Durante la discusión de un panel en esa conferencia, los defensores de una de las partes respondieron a los críticos con estos argumentos: *Los que trabajamos en este campo sabemos que es correcto lo que hacemos, si bien no podemos demostrarlo a otros de afuera*. *Ustedes, los críticos, han concentrado su ataque sobre un lunar débil, que es la carencia de datos acerca de lo que sostenemos*. Quien conozca modestamente cómo funciona y se testea el conocimiento científico, se dará cuenta de que estas “defensas” deben reconocerse como la admisión de que la ciencia está ausente en esta discusión.

En la decisión de la Corte Suprema de U.S. de 1993 sobre el caso Daubert, la Corte se concentró en resolver de por sí, de una vez y para siempre, el nudo gordiano de la demarcación de la ciencia de la pseudo-ciencia. Más aún, adoptó la decisión de permitir que cada juez federal resolviera este problema al decidir si el testimonio de todo testigo experto científico debía ser admisible. A la luz de todas las incertidumbres que serán discutidas en este capítulo, cabe decir que se trató de un objetivo ambicioso de ser puesto en práctica.[7] La

[6] El principio establecido en Daubert fue ampliado en Kumho Tire Co. v. Carmichael, en cuyo caso la evidencia en cuestión provenía de un técnico y no de un científico. El técnico testificó que la única causa posible de estallido de una llanta tenía que ser un defecto de fabricación, ya que no podía establecer ninguna otra causa. La corte de Apelaciones había admitido la evidencia bajo el supuesto de que Daubert no era aplicable a evidencia técnica sino solamente a evidencia científica. La Corte Suprema revocó el fallo, admitiendo que el estándar Daubert podía ser aplicado a la simple evidencia técnica, y que en este caso, la evidencia del experto propuesto no era suficientemente confiable.

[7] El titular de la Corte de Justicia Rehnquist, al responder a la opinión mayoritaria en Daubert, fue el primero en expresar su inquietud con la tarea asignada a los jueces federales de esta forma: “No me siento obligado en mi confianza hacia los jueces federales, pero sí con problemas en saber qué se quiere decir con que el status científico de una teoría dependa de su “falsabilidad‟, y sospecho que algunos de ellos también los tendrán.” 509 U.S. 579, 600 (1993) (Rehnquist, C.J., coincidiendo en parte y disintiendo en parte). Su preocupación se hizo eco en el Juez Alex Kozinski cuando el caso fue reconsiderado por la Corte de Apelaciones de US del Noveno Circuito luego de una devolución de la Corte Suprema. 43 F.3d 1311, 1316 (9th Cir. 1995) (“Nuestra responsabilidad, por lo tanto, a menos que estemos malinterpretando la opinión de la Corte Suprema, es resolver las disputas entre científicos respetados y de buen crédito con arreglo a su experiencia, en aquellas áreas en las que no

presentación de evidencia científica en un alegato es una especie de matrimonio forzado entre dos disciplinas. Las dos están obligadas en cierta medida a ceder frente a los imperativos centrales con que la restante suele manejarse, y resulta probable que ninguna muestre su mejor cariz.

La decisión Daubert fue un intento – y ciertamente no el primero – de regular ese encuentro disciplinario. A los jueces se les pide que decidan sobre la "confiabilidad de la evidencia" del testimonio en cuestión, basándose no sobre las conclusiones ofrecidas, sino sobre los métodos utilizados para llegar a las mismas.

Variedades y Límites de la Experiencia Estadística Es conveniente dividir a la estadística en tres campos: Probabilidad, Estadística teórica, y Estadística Aplicada.

La estadística teórica estudia las propiedades matemáticas de los procedimientos estadísticos, p.ej. las tasas de error; la teoría de la probabilidad desempeña un papel central en este contexto. Los resultados pueden ser utilizados por los estadísticos aplicados que se especializan en recopilar tipos particulares de datos, como los estudios de encuesta, o en tipos especiales de análisis, como los métodos multi-variados. El conocimiento estadístico no sólo es requerido por los graduados en estadística. Como el razonamiento estadístico está detrás de toda investigación empírica, los investigadores de casi todos los campos del saber tienen que dominar las ideas básicas de la estadística. Expertos graduados en ciencias físicas, médicas y sociales – y algunos en ciencias humanas – deben ser formalmente entrenados en estadística. Hay especialidades como la bio-estadística, la epidemiología, la econometría, y la psicometría que son primariamente estadísticas, con énfasis en los métodos y problemas de la disciplina vinculada más importante.

Es probable que la gente especializada en el uso de los métodos estadísticos – y cuyas carreras profesionales demuestran esa orientación – aplique correctamente los procedimientos e interpreten en forma adecuada los resultados obtenidos. Por otra parte, los científicos y técnicos forenses dan testimonio a menudo de probabilidades o estadísticas derivadas a partir de estudios compilados por otros, aunque carezcan del entrenamiento o conocimiento requeridos para entender y aplicar la información. El caso "El Estado v. Garrison" (US) ilustra el problema.[8] En una causa por asesinato que implicaba la evidencia de marcas de mordeduras, un dentista que debía prestar testimonio dijo que "la probabilidad de que dos conjuntos de marcas dejadas por dientes sean idénticas en un caso como éste, es aproximadamente igual a 8 en un millón", aunque "estaba inseguro sobre qué fórmula pudo usarse para llegar a ese número si no fuera mediante 'computación'".

También, al mismo tiempo, elegir qué datos hay que examinar, o cómo modelar de la mejor forma un proceso, puede requerir experiencia con el tema de la que carece el estadístico. Los estadísticos a menudo asesoran a expertos sobre procedimientos para recolectar datos y analizan los datos recolectados por otra gente. De lo cual resulta que los casos que implican evidencia estadística son (o deberían ser) casos de testimonio entrelazado de "dos expertos". Por ejemplo, un economista laboral puede definir al mercado laboral relevante del cual el empleador elige a sus empleados, y el experto estadístico puede contrastar el origen

exista consenso científico con respecto a lo que constituye "buena ciencia‟ y a lo que no lo es, y rechazar en forma ocasional este testimonio de expertos porque no fue "derivado mediante el método científico‟. Conscientes de nuestra posición dentro de la jerarquía del poder judicial, respiremos hondo y pongamos manos a la obra con esta dura tarea.")

[8] Ver Paul C. Giannelli, Bite Mark Analysis, Case Legal Studies Research Paper No. 08-06; SSRN. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1086763

provincial de los mismos con la composición por origen del mercado laboral. Naturalmente, el valor del análisis estadístico depende del conocimiento económico subyacente.[9]

Procedimientos que Enaltecen el Testimonio Estadístico

Autonomía Profesional Idealmente, los expertos que llevan a cabo investigaciones en litigios deberían actuar con la misma objetividad con la que actúan en otros contextos. Luego, los expertos que atestiguan (o que producen resultados utilizados en el testimonio por otra gente) deberían responsabilizarse de hacer todo análisis requerido para conducirse de una manera responsable en las cuestiones litigiosas. Cuestiones de libertad de investigación concedida a los expertos que atestiguan, así como el alcance y la profundidad de de sus investigaciones, pueden revelar algunas de las limitaciones de los análisis presentados.

Revelar Otros Análisis Los estadísticos analizan los datos utilizando una variedad de modelos y métodos estadísticos. Hay mucho que decir a favor de mirar los datos de modos distintos. Para permitir una evaluación balanceada del análisis en que el estadístico se coloca, sin embargo, el testimonio de expertos puede explicar la historia que subyace al desarrollo del enfoque de la estadística final.[10] De hecho, algunos comentaristas han instado a que los abogados que saben de otros conjuntos de datos o análisis que no apoyan la posición del cliente deben revelar este hecho a la Corte, en lugar de tratar de engañarla mediante la presentación de resultados que sólo le son favorables.[11]

Revelar Datos y Métodos Analíticos Antes del Juicio La recopilación de datos, a menudo es costosa, y los conjuntos de datos suelen contener, al menos, errores u omisiones. Una cuidadosa exploración de modos alternativos de análisis también puede ser costosa y lleva mucho tiempo. Para minimizar la posibilidad de que se den debates de distracción en el juicio sobre la exactitud de los datos y la elección de las técnicas de análisis, y para permitir los debates de expertos informados sobre el método, deben utilizarse procedimientos previos al juicio, en particular respecto a exactitud y ámbito de aplicación de los datos, y para descubrir los métodos de análisis. Se dispone de procedimientos sugeridos a lo largo de estas líneas.

Presentación del Testimonio de los Expertos Estadísticos El formato más común de la presentación de pruebas en un juicio es secuencial. Los testigos de la parte demandante

[9] En Vuyanich v. Republic National Bank (USA), 505 F. Supp. 224, 319 (N.D. Tex. 1980), vacated, 723 F.2d 1195 (5th Cir. 1984) http://openjurist.org/723/f2d/1195, el experto del acusado criticó el modelo estadístico del demandante por un supuesto implícito, aunque restrictivo, sobre los salarios de los hombres y las mujeres. El tribunal del distrito en que se trataba el caso aceptó el modelo porque el experto del demandante tenía una “conjetura muy sólida” sobre el supuesto, y su experiencia incluía tanto economía laboral como estadística. Resulta dudoso, en todo caso, que el conocimiento económico arroje mucha luz sobre el supuesto, y hubiera sido más sencillo realizar un análisis menos restrictivo. En este caso, el tribunal pudo haberse dejado impresionar por un único experto que aunaba experiencia sustancial y destreza estadística. Una vez que la cuestión es definida mediante el conocimiento sustantivo y legal, algunos aspectos del análisis estadístico sólo terminarán siendo consideraciones estadísticas, y la destreza en cualquiera otra área no resultará pertinente.

[10] Ver por ejemplo, Mikel Aickin, Issues and Methods in Discrimination Statistics, in Statistical Methods in Discrimination Litigation 159 (David H. Kaye & Mikel Aickin eds., 1986); Kingsley R. Browne, The Strangely Persistent Transposition Fallacy: Why Statistically Significant Evidence of Discrimination May Not Be Significant, 14 Lab. Law. 437 (1998-1999). http://faculty.law.wayne.edu/browne/Documents/Articles/Transposition%20Fallacy_Browne.pdf

[11] El Grupo de Expertos en Estadística de las evaluaciones como Evidencia en los tribunales también recomienda que "si una parte proporciona datos estadísticos a diferentes expertos competitivos para el análisis, este hecho debe revelarse al testimonio de expertos, si los hubiere." Cuándo y en qué circunstancias un análisis estadístico en particular podría estar tan imbuido de ideas y teorías del abogado del caso que debe recibir protección como producto del trabajo de abogado es una cuestión que está más allá del alcance de este capítulo.

son llamados en primer lugar, uno por uno, sin interrupción, excepto en el caso de repreguntas, y su testimonio es en respuesta a preguntas específicas y no mediante una narración ampliada. Aunque tradicional, esta estructura no es la obligada por las Reglas Federales de Evidencia (US). Se han propuesto algunas alternativas que podrían ser más eficaces con los testimonios estadísticos importantes. Por ejemplo, cuando los informes de los testigos van de la mano, el juez podría permitir combinar sus presentaciones y que los testigos sean interrogados como un panel en lugar de secuencialmente. Podrían permitirse más testimonios narrativos, y el experto podría ser autorizado a dar una breve clase sobre estadística como fase previa de algunos testimonios. En lugar de permitir a las partes presentar a sus expertos en medio de todas las pruebas, el juez podría llamar a los expertos de las partes oponentes a declarar al mismo tiempo. Algunos tribunales, especialmente en los ensayos sin jurado, pueden tener a ambos expertos bajo juramento y, en efecto, permitírseles participar en un diálogo. Con semejante formato, los expertos serán capaces de decir si concuerdan o no en cuestiones específicas. El juez y el abogado pueden intercambiar preguntas. Estas prácticas tienden a mejorar la comprensión del juez y a reducir las tensiones asociadas con el rol contradictorio de los expertos.

3. Modalidad de recopilación de datos

El análisis sólo es tan bueno como los datos sobre los que descansa. En gran medida, el diseño de un estudio determina la calidad de los datos. Por lo tanto, la interpretación correcta de los datos y de sus implicancias comienza con una comprensión del diseño del estudio y diseños diferentes ayudan a responder a preguntas diferentes.[12] En muchos casos, las estadísticas se presentan para demostrar la causalidad. ¿Los inversores potenciales se comportarían de otra manera al obtener información adicional en un prospecto de divulgación de títulos-valores? ¿Tiende la pena capital a disuadir la delincuencia? ¿Los aditivos de alimentos causan cáncer? El diseño de estudios encaminados a demostrar causalidad es el primero y tal vez el tema más importante de esta sección.

Otra cuestión es el uso de datos muestrales para caracterizar una población: la población es toda la clase de unidades que son de interés, la muestra es un conjunto de unidades elegidas para el estudio detallado. Inferencias desde la parte al todo, sólo se justifican cuando la muestra sea representativa, y ése es el segundo tema de esta sección.

Por último, es importante verificar la exactitud de la recopilación de datos. Los errores pueden surgir en el proceso de toma y registro de las mediciones de las unidades individuales. Este aspecto de la calidad de los datos es el tercer tema en esta sección.

Diseño Apropiado para Investigar la Causalidad Tipos de Estudio Cuando es cuestión de causalidad, los abogados usan tres tipos fundamentales de informaciones: evidencia anecdótica, estudios de observaciones, y experimentos controlados. Como veremos, los informes anecdóticos pueden facilitar alguna información, pero resultan más útiles para estimular investigaciones que por ser una base para establecer asociación. Los estudios de observaciones pueden establecer que un factor está asociado con otro factor, pero todavía falta un largo trecho para cruzar el puente entre asociación y causalidad.[13] Los experimentos controlados son ideales para inferir causalidad, pero pueden ser difíciles de realizar.

[12] Para un tratamiento introductorio a la recopilación de datos, ver, p.ej., David Freedman et al., Statistics (3d ed. 1998); Darrell Huff, How to Lie with Statistics (1954); David S. Moore, Statistics: Concepts and Controversies (3d ed. 1991); Hans Zeisel, Say It with Figures (6th ed. 1985); Angie Vázquez Rosado, Reseña/ Resumen de Libro “How to lie with Statistics”, http://kalathos.metro.inter.edu/Num_2/resenahowtoliewithstatistics.pdf

[13] Por ejemplo, los fumadores tienen tasas más altas de cáncer al pulmón que los no fumadores; por consiguiente, fumar y tener cáncer de pulmón son fenómenos asociados.

La “evidencia anecdótica” significa dar informes de un tipo de evento subsiguiente a otro. Es típico que los informes sean obtenidos al azar o en forma selectiva, pero la lógica del *post hoc, ergo propter hoc* no basta para demostrar que el primer evento sea la causa del segundo. Luego, si bien la evidencia anecdótica puede ser sugerente,[14] también puede ser engañosa.[15] Por ejemplo, niños que viven cerca de líneas eléctricas desarrollan leucemia, pero ¿es la exposición a campos eléctricos y magnéticos la causa de esta enfermedad? La evidencia anecdótica no es convincente ya que también la leucemia ocurre entre niños con una exposición mínima a esos campos. Hay que comparar las tasas de enfermedad entre los que están expuestos y los que no lo están. Si la exposición provoca la enfermedad, la tasa debería ser más alta entre los expuestos, más baja entre los no expuestos. Por supuesto, los dos grupos pueden diferir en otros aspectos cruciales que no son su exposición. Los niños que viven cerca de las líneas de potencia pueden pertenecer a familias más pobres y estar expuestos a otros riesgos ambientales. Estas diferencias pueden dar la sensación de una relación causa-efecto, o bien pueden estar ocultando una relación real. Las relaciones causa-efecto son bastante sutiles, y se requieren estudios cuidadosamente diseñados para extraer conclusiones válidas.[16]

Es típico que un estudio bien diseñado compare resultados para sujetos que están expuestos a algún factor o *grupo de tratamiento* con los de otros sujetos que no lo están – el *grupo de control*. Hay que distinguir entre experimentos controlados y estudios de observaciones. En un experimento controlado, los experimentadores son los que deciden qué sujetos están expuestos al factor de interés y cuáles van a parar al grupo de control. En muchos estudios de observaciones, son los propios sujetos los que eligen su exposición. Con motivo de esta auto-selección es posible que los grupos de tratamiento y de control difieran con relación a otros factores importantes que no son el factor de interés primario[17] (a

[14] En medicina, la evidencia de la práctica clínica es frecuentemente el punto de partida para demostrar un efecto causal. Un ejemplo famoso fue la exposición de madres alemanas al sarampión durante el embarazo, lo que fue seguido por la ceguera de sus hijos. N. McAlister Gregg, Congenital Cataract Following German Measles in the Mother, 3 Transactions Ophthalmological Soc’y Austl. 35 (1941), reprinted in The Challenge of Epidemiology 426 (Carol Buck et al. eds., 1988). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2272051/pdf/epidinfect00028-0013.pdf

[15] Algunos tribunales han sugerido que el intento de inferir causalidad a partir de informes anecdóticos es inadmisible en el caso Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). Ver Haggerty v. Upjohn Co., 950 F. Supp. 1160, 1163–64 (S.D. Fla. 1996) http://www.law.cornell.edu/supct/html/92-102.ZS.html (donde se dice que los informes a la Food and Drug Administration que involucraban a la droga Halcion e “informes de casos anecdóticos que aparecen en la literatura médica... pueden ser usados para generar hipótesis sobre causalidad, pero no conclusiones sobre la misma” porque “las determinaciones de causa-efecto científicamente válidas dependen de ensayos clínicos controlados y de estudios epidemiológicos”); Cartwright v. Home Depot U.S.A., Inc., 936 F. Supp. 900, 905 (M.D. Fla. 1996) (donde se excluye la opinión de un experto de que la pintura al látex sea causa del asma del demandante, en parte porque “los informes de casos... no son un sustituto de una investigación científicamente diseñada y realizada”).

[16] Tómese un clásico ejemplo en epidemiología. En una época, se pensaba que el cáncer de pulmón era causado por el vapor de alquitrán de las carreteras, porque muchos pacientes de cáncer de pulmón vivían cerca de carreteras que habían sido recientemente pavimentadas. Esto es mera evidencia anecdótica. Pero su lógica es bastante incompleta, porque muchos pacientes sin cáncer de pulmón también estaban expuestos al vapor de alquitrán. Se necesita una comparación de tasas. Un estudio cuidadoso demostró que los pacientes de cáncer al pulmón tenían una exposición similar al vapor de alquitrán que otra gente; la diferencia real era la exposición al humo de cigarrillo. Richard Doll & A. Bradford Hill, A Study of the Aetiology of Carcinoma of the Lung, 2 Brit. Med. J. 1271 (1952). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2022425/pdf/brmedj03472-0009.pdf

[17] Para nuestros propósitos, una *variable* es una característica numérica de unidades en un estudio. Por ejemplo, en una encuesta de habitantes, la unidad de análisis es la persona, y las variables podrían ser el ingreso (medido en pesos por año) y el nivel educativo (años completados de escolaridad). En un estudio de las escuelas de un distrito, la unidad de análisis es el distrito, y las variables podrían ser el ingreso promedio familiar de los residentes y los resultados obtenidos por los estudiantes en la escuela. Cuando se investigan relaciones de causa y efecto, la variable que

estos factores se los llama variables confusivas).[18] En estudios de los efectos de las líneas de potencia eléctrica sobre la salud, la estructura familiar podría ser una variable confusiva, como también el estar expuesto a otros riesgos.[19]

Experimentos Controlados Al Azar En experimentos al azar controlados, los investigadores asignan a los sujetos a grupos de tratamiento o de control en forma aleatoria. En tal caso es probable que los grupos sean comparables – excepto en cuanto al tratamiento. La elección al azar tiende a equilibrar los grupos con respecto a las posibles variables confusivas; el efecto de los desbalances residuales puede ser evaluado mediante técnicas estadísticas. Por lo tanto, las inferencias basadas en experimentos al azar bien hechos son más seguras que las basadas en estudios de observaciones.[20] El ejemplo anterior sobre la relación aspirinas-ataques cardíacos también proporciona una idea de que los experimentos al azar, aunque son mucho más difíciles de llevar a cabo, son los que producen mejor evidencia.

Resumiendo: 1º) Resultados provenientes de un grupo de tratamiento que carece de un grupo de control dicen en general muy poco y pueden ser engañosos. *Es esencial poder comparar*. 2º) Si el grupo de control fue obtenido por medio de una asignación al azar antes del tratamiento, la diferencia de resultados entre los grupos de tratamiento y de control puede ser aceptada, dentro de los límites del error estadístico, como la medición verdadera del efecto del tratamiento.[21] Empero, si el grupo de control fue armado de otra forma, las diferencias entre grupos antes del tratamiento pueden contribuir a diferencias de resultados, o a enmascarar otras diferencias que hubieran sido observadas. Por consiguiente, los estudios de observaciones tienen éxito en la medida que sus grupos de tratamiento y de control sean comparables – dejando aparte el tratamiento.

Los Estudios de Observaciones La mayoría de los estudios estadísticos vistos en un tribunal son observacionales, no experimentales. Tomen la cuestión de si la pena capital disuade el

---

caracteriza al efecto es denominada *variable dependiente*, ya que depende de las causas; también son denominadas *variables de respuesta*. Por otro lado, las variables que representan las causas son denominadas *variables independientes*, y también *factores* o *variables explicativas*.

[18] Una variable confusiva está correlacionada tanto con las variables independientes como con la variable dependiente. Si las unidades estudiadas difieren en las variables independientes, también es probable que difieran en las confusivas. Luego, estas últimas – y no las variables independientes – podrían ser responsables de las diferencias observadas en la variable dependiente.

[19] La confusión se presenta aún en cuidadosos estudios epidemiológicos. Recordar lo que se dijo antes sobre la asociación herpes femenino-cáncer cervical.

[20] Pero los experimentos no siempre pueden ser puestos en práctica, como en el caso de las líneas eléctricas. Ver por ejemplo Colin Begg, Mildred Cho, Susan Eastwood, Richard Horton, David Moher, Ingram Olkin, Roy Pitkin, Drummond Rennie, Kenneth F Schulz, David Simel y Donna F Stroup, Mejora de la calidad de los informes de los ensayos clínicos aleatorios controlados. Recomendaciones del grupo de trabajo CONSORT, Rev. Esp. Salud Pública, 1998, vol.72, n.1. http://www.scielosp.org/pdf/resp/v72n1/consort.pdf A fines estadísticos, la aleatorización puede lograrse usando algún método objetivo, tal como la generación de números al azar o un computador; una asignación caótica y desordenada puede ser insuficiente.

[21] Por supuesto, nunca puede descartarse que ambos grupos no sean comparables de manera reconocible. Sin embargo, la asignación al azar permite al investigador computar la probabilidad de observar una gran diferencia de resultados cuando el tratamiento en realidad no tiene efecto alguno. Si esta probabilidad es pequeña, se dice que la diferencia de respuesta es “estadísticamente significativa”. Ver más adelante en este mismo capítulo. Al usar métodos al azar para los sujetos en los grupos de tratamiento y de control se sientan bases sólidas de los test de significación estadística (David Freedman et al., Statistics, 3d ed. 1998); pp. 503–24, pp. 547–78. Lo que es más importante, el azar también asegura que la asignación de personas a los grupos de tratamiento y de control esté libre de la manipulación, consciente o inconsciente, de los investigadores o de los sujetos. El tratamiento al azar no es la única forma de asegurar dicha protección, pero resulta ser “la forma más simple y mejor comprendida de certificarlo” Philip W. Lavori et al., Designs for Experiments—Parallel Comparisons of Treatment, in Medical Uses of Statistics 61, 66 (John C. Bailar III & Frederick Mosteller).

asesinato. Para hacer un experimento aleatorio controlado, la gente tendría que ser asignada al azar a un grupo de control y a un grupo de tratamiento. Los del grupo de control sabrían que no recibirían la pena de muerte por asesinato, mientras que los del grupo de tratamiento sabrían que podrían ser ejecutados. La tasa de ulteriores asesinatos cometidos por los sujetos de estos grupos sería entonces observada. Este experimento es inaceptable, tanto en términos políticos, éticos, y legales.[22]

Sin embargo, hay estudios realizados sobre los efectos disuasivos de la pena de muerte, todos basados en observaciones, y hay algunos que han atraído la atención judicial.[23] Los investigadores catalogaron diferencias en la incidencia del asesinato en estados (o provincias) con y sin pena de muerte, y analizaron los cambios de las tasas de homicidio y las tasas de ejecución a lo largo del tiempo. En estos estudios de observaciones, los investigadores pueden hablar de grupos de control (como los estados que no tienen la pena capital) y de controlar la incidencia de variables potencialmente confusivas (p.ej. peores condiciones económicas).[24] Sin embargo, como la asociación no implica causalidad, las inferencias causales que pueden ser extraídas de estos análisis descansan sobre fundamentos menos sólidos que los provistos por los experimentos al azar controlados.[25]

Los estudios de observaciones pueden ser muy útiles. La evidencia de que fumar causa cáncer de pulmón en los seres humanos, aunque provenga de observaciones, es convincente. Los estudios de observaciones proveen evidencia poderosa en las siguientes circunstancias:

• Se observa una asociación en estudios de distintos tipos entre grupos diferentes. Esto reduce la probabilidad de que la asociación observada se deba a un defecto de algún tipo de estudio o a una peculiaridad de un grupo de personas.

• Se mantiene la asociación al tomarse en cuenta los efectos de variables confusivas plausibles mediante técnicas estadísticas apropiadas, como por ejemplo comparar grupos más pequeños relativamente homogéneos con respecto al factor.[26]

---

[22] El Federal Judicial Center tiene una publicación de 1981: Experimentation in the Law: Report of the Federal Judicial Center Advisory Committee on Experimentation in the Law sobre el asunto. http://www.fjc.gov/public/pdf.nsf/lookup/experlaw.pdf/$file/experlaw.pdf

[23] Ver en general Hans Zeisel, The Deterrent Effect of the Death Penalty: Facts v. Faith, 1976 Sup. Ct. Rev. 317. También, Stephen Nathanson, Does it Matter if the Death Penalty is Arbitrarily Administered? Philosophy and Public Affairs, Vol. 14, No. 2 (Spring, 1985), pp. 149-164. http://orgs.law.ucla.edu/LTW/Documents/Week%201%20-%20Readings/BR%202-Stephen%20Nathanson-Does%20it%20Matter%20if%20the%20Death%20Penalty%20is%20Arbitrarily%20Administered.pdf

[24] El proceso usado con frecuencia para controlar la incidencia de las variables confusivas es la regresión múltiple, acerca de la cual hablaremos más adelante.

[25] Ver David Freedman et al., Statistics (3d ed. 1998): Los grupos seleccionados no al azar casi siempre diferirán de una forma sistemática que no es su exposición al programa experimental. Las técnicas estadísticas pueden eliminar el azar como una posible explicación de las diferencias,... pero si no se ha practicado una elección al azar no hay métodos certeros de determinar si las diferencias observadas entre grupos no se deben en realidad a una diferencia sistemática, pre-existente... La comparación sistemática entre distintos grupos implicará ambigüedades cuando una diferencia sistemática dé lugar a una explicación plausible de los efectos aparentes del programa experimental.

[26] La idea es controlar la influencia de una variable confusiva haciendo comparaciones separadamente dentro de los grupos, para los cuales la variable confusiva se mantiene prácticamente constante y por consiguiente tiene escasa influencia sobre las variables de interés primario. Por ejemplo, es más probable que los fumadores tengan cáncer de pulmón que los no fumadores. Son variables confusivas la edad, el género, la clase social, y la región de residencia, pero si estas variables son controladas no se altera la relación entre tasas de fumadores y de cáncer. Hay diferentes estudios que confirman el vínculo causal. Éste es el motivo por el cual la mayoría de los expertos cree que fumar causa cáncer de pulmón y varias otras enfermedades. Para revisar la

• Existe una explicación plausible del efecto de las variables independientes; luego, el vínculo causal no sólo depende de la asociación observada. Hay otras explicaciones que vinculan la respuesta con las variables confusivas que deberían ser menos plausibles.[27]

Cuando estos criterios no se cumplen, los estudios observacionales pueden producir legítimo desacuerdo entre los expertos, y no existe un procedimiento mecánico para comprobar cuál es correcto. Al final, decidir si las asociaciones son causales no es una cuestión de estadísticas, sino una cuestión de buen juicio científico, y las preguntas que deben plantearse con respecto a los datos ofrecidos en la cuestión de causalidad se pueden resumir como sigue:

• ¿Existió un grupo de control? Si no fue así, el estudio poco puede decir en términos de causalidad;
• Si hubo un grupo de control, ¿a cuántas personas le fue asignado el tratamiento o control? ¿Mediante un proceso controlado por el investigador (un experimento controlado) o un proceso fuera del control del investigador (un estudio observacional)?
• Si el estudio se trató de un experimento controlado, la asignación ¿fue realizada mediante un mecanismo al azar (aleatorización) o dependió del juicio del investigador?
• Si los datos provienen de un estudio observacional o de un experimento controlado no aleatorio ¿cómo se conformaron los sujetos al tratamiento o en grupos de control? ¿Son comparables ambos grupos? ¿Qué grupos están confundidos en el tratamiento? ¿Qué ajustes fueron tomados para tener en cuenta la confusión? ¿Fueron sensibles?[28]

¿Pueden ser Generalizados los Resultados? Todo estudio debe ser realizado sobre un determinado número de personas, en cierto momento y lugar, utilizando determinados tratamientos. En estos aspectos, el estudio debe ser convincente. Debe existir un control adecuado de las variables confusivas, y una inequívoca gran diferencia entre los grupos de tratamiento y de control. Si es así, la validez interna del estudio no será discutida: Para los sujetos del estudio, el tratamiento tuvo efectividad. Pero aún existe una cuestión de validez externa: extrapolar desde las condiciones del estudio a circunstancias más generales siempre suscita problemas.

Por ejemplo, los estudios sugieren que las definiciones de locura dadas por los miembros del jurado influyen en decisiones en los casos de incesto.[29] ¿Tienen esas definiciones un efecto similar en casos de asesinato? Otros estudios indican que las tasas de reincidencia

---

literatura, International Agency for Research on Cancer (IARC), IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans: Tobacco Smoking (2007). http://monographs.iarc.fr/ENG/Monographs/vol89/mono89.pdf

[27] A. Bradford Hill, The Environment and Disease: Association or Causation? 58 Proc. Royal Soc'y Med. 295 (1965); http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/pdf/procrsmed00196-0010.pdf Alfred S. Evans, Causation and Disease: A Chronological Journey 187 (1993).

[28] Estas preguntas han sido adaptadas de Freedman et al., *supra*. Para discusiones de la admisibilidad o ponderación de estudios que pasan por alto otras variables confusivas posibles, ver People Who Care v. Rockford Board of Education, 111 F.3d 528, 537–38 (7th Cir. 1997) http://www.projectposner.org/case/1997/111F3d528 ("La literatura científica social sobre rendimiento educativo identifica un cierto número de variables además de la pobreza y la discriminación que explican las diferencias de rendimientos en la escuela, como el nivel educativo de los padres y en qué medida se involucran en la escolaridad de sus hijos... No puede suponerse que estas variables estén distribuidas de forma aleatoria a lo largo de los distintos grupos raciales y étnicos en Rockford, o que estén perfectamente correlacionadas con la pobreza...").

[29] Max Cohen, Rita James Simon, The Jury and the Defense of Insanity, 2 Val. U. L. Rev. 398 (1968). http://scholar.valpo.edu/vulr/vol2/iss2/10; Julie E. Grachek, The Insanity Defense in the Twenty-First Century: How Recent United States Supreme Court Case Law Can Improve the System, Indiana Law Journal, Vol. 81, 2006. http://www.law.indiana.edu/ilj/volumes/v81/no4/14_Grachek.pdf

de los ex-convictos no se ven afectadas por un apoyo financiero temporario después de ser liberados.[30] ¿Sucede lo mismo bajo otras condiciones del mercado laboral?

La confianza en lo apropiado de una extrapolación no puede provenir del experimento en sí,[31] sino de conocimiento sobre los factores externos que podrían afectar, o no, los resultados.[32] A veces los diversos experimentos u otros estudios apuntan todos en la misma

[30] Para un experimento sobre sustento del ingreso y reincidencia, ver Peter H. Rossi et al., Money, Work, and Crime: Experimental Evidence (1980). La interpretación de los datos ha sido objeto de controversia. V. Hans Zeisel, Disagreement over the Evaluation of a Controlled Experiment, 88 Am. J. Soc. 378 (1982) (con su comentario); Shari Seidman Diamond and Hans Zeisel, Sentencing Councils: A Study of Sentence Disparity and its Reduction, 43 U. Chi. L. Rev. 109 (1975-1976). http://www.law.northwestern.edu/faculty/fulltime/diamond/papers/sentencingCouncils.pdf

[31] Supongan que se realiza un estudio epidemiológico sobre la relación entre una sustancia tóxica y una enfermedad. La tasa de incidencia de la enfermedad sobre un grupo de personas expuestas a la sustancia es comparada con la tasa del grupo de control, y la tasa del grupo expuesto resulta ser más del doble que la del grupo de control. (En términos algo más técnicos, el riesgo relativo es superior a 2). ¿Implican estos datos que el demandante que estuvo expuesto a la sustancia tóxica y contrajo la enfermedad no la habría contraído de no haberse expuesto? Si suponemos que la sustancia es la causa de la enfermedad y que se tuvo en cuenta todas las variables confusivas (juicio difícil de sostener), luego podemos concluir que alrededor de la mitad de los casos de enfermedad del grupo expuesto no hubieran existido de no ser por su exposición. Pero aplicar esta aritmética a una persona resulta problemático. Por ejemplo, el riesgo relativo resulta un promedio sobre toda la gente incluida en el estudio. El grado de exposición y la susceptibilidad a la misma no son, ciertamente, uniformes, y la exposición del demandante y su susceptibilidad no pueden conocerse a partir del estudio. Sin embargo, varios tribunales y comentaristas han aseverado que un riesgo relativo mayor que 2 demuestra que existe causación directa, o recíprocamente, que un riesgo relativo igual o menor que 2 impiden sacar una conclusión sobre causalidad. P.ej. DeLuca v. Merrell Dow Pharms., Inc., 911 F.2d 941, 958–59 (3d Cir. 1990) http://ftp.resource.org/courts.gov/c/F3/9/9.F3d.958.92-5089.html; Marder v. G.D. Searle & Co., 630 F. Supp. 1087, 1092 (D. Md. 1986) http://ftp.resource.org/courts.gov/c/F2/911/911.F2d.941.89-5572.html ("un incremento duplicado del riesgo equivale. . . al requisito legal de una prueba – demostrar causalidad por la preponderancia de la evidencia o, en otras palabras, una probabilidad mayor al 50%"), aff'd *sub nom*. Wheelahan v. G.D. Searle & Co., 814 F.2d 655 (4th Cir. 1987) http://openjurist.org/814/f2d/655/55-uslw-2568-7-fedrserv3d-568; Bert Black & David E. Lilienfeld, Epidemiologic Proof in Toxic Tort Litigation, 52 Fordham L. Rev. 732, 769 (1984); Michael D. Green, D. Michal Freedman, and Leon Gordis, Reference Guide on Epidemiology, Federal Judicial Center, 2000. http://www.fjc.gov/public/pdf.nsf/lookup/sciman06.pdf/$file/sciman06.pdf Algunos criticaron duramente este razonamiento. Steven E. Fienberg et al., Understanding and Evaluating Statistical Evidence in Litigation, 36 Jurimetrics J. 1, 9 (1995); Allan G. King, "Gross Statistical Disparities" as Evidence of a Pattern and Practice of Discrimination: Statistical versus Legal Significance, 22 The Labor Lawyer (2007) http://www.americanbar.org/content/dam/aba/publishing/lel_flash/LL_king.authcheckdam.pdf; Diana B. Petitti, Reference Guide on Epidemiology, 36 Jurimetrics J. 159, 168 (1996) (review essay); D.A. Freedman & Philip B. Stark, The Swine Flu Vaccine and Guillain-Barré Syndrome: A Case Study in Relative Risk and Specific Causation, 23 Evaluation Rev. 619 (1999); James Robins & Sander Greenland, The Probability of Causation Under a Stochastic Model for Individual Risk, 45 Biometrics 1125, 1126 (1989) http://www.hsph.harvard.edu/~robins/prob-caus-89.pdf; Melissa Moore Thompson, Comment, Causal Inference in Epidemiology: Implications for Toxic Tort Litigation, 71 N.C. L. Rev. 247 (1992).

[32] Estos juicios son más fáciles de hacer en las ciencias físicas y de la vida, pero aún en estos casos se presentan problemas. Por ejemplo, puede ser difícil inferir las reacciones humanas a sustancias que afectan a los animales. En primer término, a menudo hay inconsistencias entre los contrastes a distintas especies: un elemento químico puede ser carcinogénico en los ratones pero no en las ratas. Una extrapolación de los roedores a los humanos es aún más problemática. En segundo lugar, para tener efectos medibles en los experimentos con animales se requiere administrar los componentes químicos a dosis muy elevadas. Los resultados son posteriormente extrapolados – utilizando modelos matemáticos – a las dosis muy reducidas que preocupan a los seres humanos. Empero, hay varios modelos de respuesta a las dosis que pueden ser utilizados y se carece de una base sólida para elegir entre los mismos. En general, los diferentes modelos producen estimadores radicalmente diferentes de lo que es una "dosis virtualmente segura" en los humanos. Ver D. A. Freedman and H.

dirección, cada uno con sus propias limitaciones. Éste es el caso, p.ej., con ocho estudios que indican que es más probable que los jurados que aprueban la pena de muerte establezcan una condena en un caso de pena capital.[33] Estos resultados convergentes sugieren que la generalización tiene gran validez.

Encuestas Descriptivas y Censos Luego de la lógica de los estudios para investigar la causalidad, pasemos al segundo tópico – que es el muestreo, es decir, elegir las unidades del estudio. Un censo trata de medir alguna característica de cada unidad de la población de individuos u objetos. El censo de una población estadística consiste, básicamente, en obtener el número total de individuos mediante diversas técnicas de recuento. El censo es una de las operaciones estadísticas que no trabaja sobre una muestra, sino sobre la población total. Uno de los casos particulares de censo y, al mismo tiempo, uno de los más comunes, es el *censo de población*, en el cual el objetivo es determinar el número de personas humanas que componen un grupo, normalmente un país. En este caso, la población estadística comprendería a los componentes o habitantes del grupo o país. En general, un censo de población puede realizar algunas actividades extra que no se corresponden específicamente con la operación censal estadística. Se trata de calcular el número de habitantes de un país de territorio delimitado, correspondiente a un momento o período dado, pero se aprovecha igualmente para obtener una serie de datos demográficos, económicos y sociales relativos a esos habitantes. La exactitud de la información recogida en un censo o una encuesta depende de la forma en que fueron elegidas las unidades, qué unidades han sido medidas en realidad, y de cómo son practicadas las mediciones.[34]

El esquema metodológico de una encuesta científica es más complicado que el de un censo. En encuestas que usan métodos de muestreo probabilísticos, se crea una estructura muestral – esto es, un listado explícito de individuos de la población. A continuación son seleccionadas unidades individuales mediante una especie de lotería, y las mediciones son tomadas sobre esa muestra. Por ejemplo, un abogado defensor encargado de un crimen notorio que está buscando modificar la sede del juicio puede encargar una encuesta de opiniones para demostrar que la opinión del público es tan adversa y arraigada que resultará difícil seleccionar y poner en funciones un jurado imparcial. La población consiste de todos

---

Zeisel, From Mouse-to-Man: The Quantitative Assessment of Cancer Risks, Statist. Sci. Volume 3, Number 1 (1988), 3-28. http://www.econ.canterbury.ac.nz/personal_pages/john_fountain/Teaching/HealthEcon/econ338/mousetoman88small.pdf Por estas razones, muchos expertos – y algunos tribunales en casos de litigios por tóxicos – han concluido que la evidencia a partir de experimentos con animales es en general insuficiente para establecer una relación de causalidad. Ver en general Bruce N. Ames et al., The Causes and Prevention of Cancer, 92 Proc. Nat'l Acad. Sci. USA 5258 (1995) http://www.pnas.org/content/92/12/5258.full.pdf; Susan R. Poulter, Science and Toxic Torts: Is There a Rational Solution to the Problem of Causation?, 7 High Tech. L.J. 189 (1993) http://btlj.org/data/articles/vol7/Poulter.pdf (se requiere evidencia epidemiológica en humanos). Ver también Committee on Comparative Toxicity of Naturally Occurring Carcinogens, National Research Council, Carcinogens and Anticarcinogens in the Human Diet: A Comparison of Naturally Occurring and Synthetic Substances (1996); Committee on Risk Assessment of Hazardous Air Pollutants, National Research Council, Science and Judgment in Risk Assessment 59 (1994) ("Hay razones basadas tanto en principios biológicos como en observaciones empíricas para sostener la tesis de que muchas formas de respuestas biológicas, incluyendo las respuestas tóxicas, pueden ser extrapoladas entre los mamíferos, incluyendo al *Homo Sapiens*, pero no hay base científica rigurosa para permitir una generalización amplia y definitiva").

[33] Phoebe C. Ellsworth, Some Steps between Attitudes and Verdicts, en Inside the Juror 42, 46 (Reid Hastie ed., 1993). Sin embargo, en Lockhart v. McCree, 476 U.S. 162 (1986), la Corte Suprema sostuvo que la exclusión de los que se oponen a la pena de muerte en la fase de culpabilidad de un juicio capital no es violatoria de los requerimientos constitucionales de un jurado imparcial. http://supreme.justia.com/cases/federal/us/476/162/case.html

[34] El Manual de Referencia tiene una sección especial de Seidman Diamond, Reference Guide on Survey Research, http://www.au.af.mil/au/awc/awcgate/fjc/survey_rese_ref.pdf, que recomiendo leer.

los que en la jurisdicción podrían ser llamados a constituir un jurado. Los funcionarios, en tal caso, podrían tener una lista de estas personas.[35] En tal caso, el ajuste entre estructura muestral y población sería excelente.[36]

Hay otras situaciones donde la estructura de la muestra no alcanza a cubrir a la población. P.ej., en un caso de obscenidad, el sondeo del abogado sobre los estándares de la comunidad[37] debería identificar a la población de la comunidad legalmente relevante, cosa que en general no es posible. Si se usan los nombres de un directorio telefónico, a la gente con números no incluidos se la excluye de la estructura de la muestra. Si esta gente, considerada grupalmente, tiene opiniones distintas que las incluidas en la estructura muestral, el sondeo no reflejará esta diferencia, aunque los individuos sean sondeados y por buenas que sean las respuestas obtenidas.[38] La medición del sondeo de la opinión de la comunidad estará sesgada, aunque este sesgo puede no ser importante.

No todas las encuestas utilizan una selección aleatoria. En algunas disputas comerciales sobre marcas registradas o publicidad, la población de potenciales compradores es difícil de identificar. Algunos encuestadores suelen recurrir entonces a algún subgrupo accesible de la

[35] Si no se convoca a la lista del jurado en forma apropiada a partir de las fuentes adecuadas, el juicio podría ser objetado. Ver David Kairys et al., Jury Representativeness: A Mandate for Multiple Source Lists, 65 Cal. L. Rev. 776 (1977).

[36] En forma similar, en investigaciones sobre estupefacientes, puede lograrse con facilidad que la estructura de la muestra para verificar el contenido de frascos, bolsos, o paquetes incautados por la policía esté apareada con la población de todos los ítems incautados en un solo caso. Como verificar ítem por ítem puede llevar mucho tiempo y ser muy costoso, a menudo los químicos extraen una muestra probabilística, analizan el material de esa muestra, y utilizan el porcentaje de drogas ilícitas hallado en la muestra para determinar la cantidad total de drogas ilícitas en todos los ítems incautados. P. ej., United States v. Shonubi, 895 F. Supp. 460, 470 (E.D.N.Y. 1995) (citing cases), rev'd on other grounds http://tillers.net/shonubi4.htm, 103 F.3d 1085 (2d Cir. 1997) http://openjurist.org/103/f3d/1085. Para discusiones de las estimaciones estadísticas en estos casos, C.G.G. Aitken et al., Estimation of Quantities of Drugs Handled and the Burden of Proof, 160 J. Royal Stat. Soc'y 333 (1997); Dov Tzidony & Mark Ravreby, A Statistical Approach to Drug Sampling: A Case Study, 37 J. Forensic Sci. 1541 (1992) http://library-resources.cqu.edu.au/JFS/PDF/vol_37/iss_6/JFS376921541.pdf; Johan Bring & Colin Aitken, Burden of Proof and Estimation of Drug Quantities Under the Federal Sentencing Guidelines, 18 Cardozo L. Rev. 1987 (1997).

[37] Hay discusión sobre cuán admisibles son estos sondeos, ver Saliba v. State, 475 N.E.2d 1181, 1187 (Ind. Ct. App. 1985) ("Aunque el sondeo . . . [no pidió] a los entrevistados . . . informar si la película en cuestión era obscena, el sondeo fue relevante para aplicar los estándares comunitarios"), y United States v. Pryba, 900 F.2d 748, 757 (4th Cir. 1990) ("Preguntarle a alguien en una entrevista telefónica si está ofendidp por la desnudez, está lejos de mostrar el material en cuestión. . . y preguntarle entonces si es ofensivo", por cuyo motivo la exclusión de los resultados de este sondeo fue apropiada).

[38] Un clásico ejemplo de sesgo de selección fue el sondeo de 1936 del Literary Digest. Luego de haber pronosticado el ganador de cada elección presidencial en US a partir de 1916, el Digest utilizó réplicas de unos 2.4 millones de encuestados para predecir que Alf Landon ganaría por un margen de 57% a 43%. En realidad, Franklin Roosevelt ganó por un voto aplastante de 62% a 38%. Ver Freeman et al., nota 8, pp. 334-35. En parte el Digesto quedó tan lejos porque eligió nombres de la guía telefónica, de listados de clubs y asociaciones, directorios de la ciudad, listas de votantes registrados, y listas de correo (ídem, 335, A-20 n.6). En 1936, cuando sólo un encuestado sobre cuatro tenía teléfono, la gente que aparecía en esos listados tendía a ser gente más acomodada. Las listas que tenían una representación más que proporcional habían funcionado bien en las últimas elecciones, cuando ricos y pobres votaron según líneas similares, pero el sesgo de la estructura muestral probó ser fatal cuando la Gran Depresión hizo que la economía pasara a ser una consideración saliente de los votantes. Ver Judith M. Tanur, Samples and Surveys, en Perspectives on Contemporary Statistics 55, 57 (David C. Hoaglin & David S. Moore eds., 1992). Hoy en día, los organismos que realizan sondeos lo hacen por teléfono, pero la mayoría de los votantes tiene teléfono, y las organizaciones seleccionan los números a ser llamados al azar en lugar de obtener muestras de los nombres de las guías telefónicas.

población, como los comerciantes minoristas.[39] Estas muestras de conveniencia pueden estar sesgadas por la discrecionalidad del entrevistador – que es una especie de sesgo de selección – y el rechazo de algunos entrevistados a participar – sesgo por ausencia de respuesta.[40] Se presenta un agudo sesgo de selección cuando los votantes escriben a sus representantes, los oyentes llaman a las radios en programas de entrevistas, los grupos de interés recogen información a partir de sus miembros[41] o los abogados eligen los casos para ir a juicio.[42] El sesgo de selección también afecta los datos de los servicios informativos del jurado que recopila la información a partir de las fuentes disponibles.

Hay varios procedimientos disponibles para tratar el sesgo de selección. P. ej., los métodos de muestreo probabilístico están idealmente adaptados para evitarlo. Una vez que la población conceptual ha sido reducida a una estructura muestral tangible, las unidades que se miden son seleccionadas mediante una lotería que proporciona a cada unidad de la estructura muestral una probabilidad conocida (≠0) de ser elegida. La selección de acuerdo con una tabla de dígitos aleatorios[43] no da lugar a sesgo de selección. En US, estos procedimientos son utilizados rutinariamente para seleccionar individuos como jurados,[44]

[39] Por ejemplo, R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc., 511 F. Supp. 867, 876 (S.D.N.Y. 1980) (se cuestionó cuán apropiado era basar un "porcentaje estadístico a nivel nacional" sobre un estudio de centros suburbanos comerciales).

[40] Analizaremos este último sesgo más adelante.

[41] P.ej., Pittsburgh Press Club v. United States, 579 F.2d 751, 759 (3d Cir. 1978) (exención de impuestos a la encuesta del correo de sus miembros para mostrar que el escaso patrocinio de uso de infraestructura generadora de ingresos era testimonio de oídas inadmisible porque "no era ni objetivo, ni científico, ni imparcial"), rev'd on other grounds, 615 F.2d 600 (3d Cir. 1980) http://bulk.resource.org/courts.gov/c/F2/615/615.F2d.600.79-1492.html

[42] Ver In re Chevron U.S.A., Inc., 109 F.3d 1016 (5th Cir. 1997). http://openjurist.org/109/f3d/1016/in-re-chevron-usa-inc En este caso, el tribunal decidió tratar 30 casos para resolver cuestiones comunes o establecer los daños en 3,000 reclamos que surgían por la eliminación de sustancias peligrosas por una supuesta decisión inapropiada de Chevron. El tribunal exigió a los comparecientes elegir 15 casos cada uno. Empero, seleccionar 30 casos extremos es muy distinto a extraer una muestra aleatoria de 30 casos. Por tanto, la corte de apelaciones dijo que, si bien el muestreo aleatorio habría sido aceptable, la corte no hubiera podido utilizar los resultados en los 30 casos extremos para resolver cuestiones de hecho o establecer los daños en los casos no sometidos a juicio. Id. en 1020. Advirtió que estos casos no eran "casos calculados para representar al grupo de 3.000 damnificados".

[43] Forman parte de este tipo de muestreo todos los métodos en los que puede calcularse la probabilidad de extracción de cualquiera de las muestras posibles. Este conjunto de técnicas de muestreo es el más aconsejable, aunque en ocasiones no es posible optar por él. En este caso se habla de muestras probabilísticas, pues no es rigurosamente correcto hablar de muestras representativas dado que, al no conocerse las características de la población, no es posible tener certeza de que tal característica se haya conseguido. *Sin reposición de los elementos*: Cada elemento extraído se descarta para la subsiguiente extracción. Por ejemplo, si se extrae una muestra de una "población" de bombitas de luz para estimar la vida media de las bombitas que la integran, no será posible medir más que una vez la bombita seleccionada. *Con reposición de los elementos*: Las observaciones se realizan con reemplazo de los individuos, de tal forma que la población es idéntica en todas las extracciones. En poblaciones muy grandes, la probabilidad de repetir una extracción es tan pequeña que el muestreo puede considerarse sin reposición aunque, realmente, no lo sea. *Con reposición múltiple*: Cada elemento extraído se descarta para la subsiguiente extracción. Para realizar este tipo de muestreo, y en determinadas situaciones, es muy útil la extracción de números aleatorios mediante computadoras, calculadoras o tablas construidas al efecto.

[44] Antes de 1968, la mayoría de los distritos federales usaba el sistema de "hombre clave" para compilar listados de jurados elegibles. Los individuos que se creía tenían contactos extensos dentro de la comunidad debían sugerir nombres de los posibles jurados, y el jurado calificado estaría constituido por estos nombres. A fin de reducir el riesgo de discriminación asociado a este sistema, el Jury Selection and Service Act de 1968, 28 U.S.C. §§ 1861–1878 (1988) http://uscode.house.gov/download/pls/28C121.txt sustituyó el principio de "selección aleatoria del nombre de los jurados por las listas de votantes del distrito o división en que está situado el tribunal". S. Rep. No. 891, 90th Cong., 1st Sess. 10 (1967), reprinted in 1968 U.S.C.C.A.N. 1792, 1793. Ver especialmente Andrew D. Leipold, Constitutionalizing Jury Selection in Criminal Cases: A Critical

pero también han sido usados para elegir casos de "indicadores de tendencia" en los casos de juicios representativos para resolver las cuestiones de todos los casos similares.[45]

¿Qué se mide de las unidades seleccionadas? Aunque la probabilidad asegure que, dentro de los límites del azar, la muestra será representativa de la estructura muestral, está la cuestión de saber qué unidades serán medidas. Cuando objetos como los recibos son muestreados en una auditoría, o la vegetación es muestreada para un estudio de la ecología de una región, todas las unidades podrían ser examinadas. Los seres humanos son más problemáticos. Algunos pueden negarse a responder, y la encuesta debería reportar la tasa de no-respuesta. Una tasa muy elevada de no-respuesta advierte sobre la presencia de sesgo[46] aunque los que no responden no difieran de forma sistemática de los que responden con respecto a las características interesantes[47] o puede permitir que los datos faltantes sean imputados.[48]

Evaluation, Georgetown Law Journal, Feb. 1998; Nancy Jean King, The American Criminal Jury, Law and Contemporary Problems, Vol. 62, No. 2, The Common Law Jury (Spring, 1999), pp. 41-67. http://www.law.duke.edu/journals/62LCPKing.

[45] Hilao v. Estate of Marcos, 103 F.3d 767 (9th Cir. 1996) http://www.utexas.edu/law/faculty/lmullenix/info/hilao_v_estate1996.pdf; Cimino v. Raymark Indus., Inc., 751 F. Supp. 649 (E.D. Tex. 1990); cf. Laurens Walker and John Monahan, Sampling Evidence at the Crossroads, Southern California Law Review, Vol. 80, 2007 http://clp.usc.edu/why/students/orgs/lawreview/documents/Walker_Laurens_80_5.pdf; cf. In re Chevron U.S.A., Inc., 109 F.3d 1016 (5th Cir. 1997) http://openjurist.org/109/f3d/1016/in-re-chevron-usa-inc. Si bien los juicios en una muestra aleatoria de casos puede producir estimaciones razonables de los daños promedio, se ha debatido la propiedad de impedir juicios individuales. Comparar Michael J. Saks & Peter David Blanck, Justice Improved: The Unrecognized Benefits of Aggregation and Sampling in the Trial of Mass Torts, 44 Stan. L. Rev. 815 (1992) http://www.ebour.com.ar/index.php?option=com_weblinks&task=view&id=13490&Itemid=0, con Chevron, 109 F.3d at 1021 (Jones, J., concurring); Robert G. Bone, Statistical Adjudication: Rights, Justice, and Utility in a World of Process Scarcity, 46 Vand. L. Rev. 561 (1993).

[46] El Literary Digest de 1936 ilustra el peligro (ver nota 38). Sólo 24% de 10 millones de personas que recibieron cuestionarios los devolvieron. La mayoría de quienes respondieron probablemente tenía claro cuáles serían sus candidatos, y tal vez en su mayoría objetaron el programa económico del Presidente Roosevelt. Es posible que esta auto-selección haya sesgado el sondeo. Ver Maurice C. Bryson, The Literary Digest Poll: Making of a Statistical Myth, 30 Am. Statistician 184 (1976) http://math.sierracollege.edu/Staff/nWai/statistical_myth.pdf; Freedman et al., ob.cit, pp. 335–336. En United States v. Gometz, 730 F.2d 475, 478 (7th Cir. 1984) (en banc) http://www.projectposner.org/case/1984/730F2d475, el Séptimo Circuito reconoció que "una baja tasa de respuesta al cuestionario del jurado podría conducir a la sub-representación de un grupo que tiene derecho a ser representado por un jurado más calificado". Sin embargo, el tribunal sostuvo que según la Jury Selection and Service Act of 1968, 28 U.S.C. §§ 1861–1878 (1988), el secretario no abusó de su discreción al no dar los pasos necesarios para incrementar una tasa de respuesta del 30%. Según la corte, "el Congreso deseaba que fuera posible que toda persona calificada pueda servir como jurado, lo que es distinto a obligar a toda persona calificada a que esté disponible para dicho servicio". Aunque sea "positivo saber por qué hay personas que no responden a un cuestionario del jurado", la corte interpretó que el Congreso "no debería preocuparse por cuestiones tan esotéricas como el sesgo de no-respuesta".

[47] Aunque las características demográficas de la muestra concuerden con las de la población, sin embargo, hay que tener cuidado. En los 1980s, un investigador de la conducta envió 100,000 cuestionarios con el objeto de averiguar cómo las mujeres veían a sus relaciones con los hombres. Shere Hite, en Women and Love: A Cultural Revolution in Progress (1987) recogió una colección enorme de cartas anónimas de miles de mujeres desilusionadas con el amor y el matrimonio, y escribió que estas respuestas eran la "protesta" de feministas "en contra de varias injusticias del matrimonio – la explotación de las mujeres en términos financieros, físicos, sexuales, y emocionales... justa y adecuada" (p. 344). En realidad, la protesta puede ser justificada, pero esta investigación no lo demostró. Cerca de 95% de 100,000 cuestionarios no tuvo respuesta. Las que no lo hicieron pueden haber tenido experiencias con menor estrés con los hombres y, por lo tanto, no sintieron la necesidad de escribir cartas autobiográficas. Aún más, se espera que esta diferencia sistemática se produzca dentro de cada clase demográfica y ocupacional. Luego, argumentar que las respuestas de la

En resumen, una buena encuesta define una población adecuada, utiliza un método insesgado para seleccionar una muestra, registra una elevada tasa de respuesta, y recoge información precisa sobre las unidades de esa muestra. En estos casos, la muestra tiende a ser representativa de la población: las mediciones dentro de la muestra describen de modo imparcial las características de la población. También es posible que, pese a todas las precauciones tomadas, la muestra, lejos de ser exhaustiva, no sea representativa. El análisis estadístico ayuda a calcular la magnitud del riesgo asumido, por lo menos para muestras probabilísticas. Es obvio que las muestras pueden ser útiles aunque no cumplan con todas las exigencias impuestas, pero en ese caso se requieren argumentos adicionales para justificar las inferencias.

Mediciones Individuales: Confiabilidad del Proceso de Medición Hay dos aspectos principales de la exactitud de las mediciones – fiabilidad y validez. En ciencia, "fiabilidad" significa que los resultados son reproducibles.[49] Un instrumento fiable siempre produce mediciones consistentes de la misma cantidad. Una balanza, por ejemplo, resulta fiable si informa siempre el mismo peso de un objeto. Puede que no sea precisa – puede informar siempre un peso demasiado alto o uno demasiado bajo – pero esta balanza perfectamente fiable siempre informa el mismo peso para el mismo objeto. Si tiene errores, serán sistemáticos; siempre apuntan en la misma dirección.

La fiabilidad puede establecerse midiendo la misma cantidad varias veces. Por ejemplo, un método de identificación de ADN requiere que un laboratorio calcule la longitud de los fragmentos de ADN. Haciendo varias veces mediciones de los fragmentos de ADN, el laboratorio puede determinar la verosimilitud de que dos mediciones difieran en cantidades específicas.[50] Estos resultados son necesarios cuando debe decidirse si la discrepancia entre la muestra obtenida en un crimen y la muestra de un sospechoso es suficiente para excluir al sospechoso.[51]

---

muestra son representativas porque "las que participaron según su edad, ocupación, religión, y otras variables conocidas en la mayoría de los casos de la población norteamericana reflejaba la de la población femenina norteamericana" no es convincente (pág. 777). De hecho, los resultados de esta muestra no aleatoria difieren dramáticamente de otros sondeos con mejores tasas de respuesta. Ver Chamont Wang, Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety 174–76 (1993). Una crítica adicional de este estudio fue hecha por David Streitfeld, Shere Hite and the Trouble with Numbers, 1 Chance 26 (1988). http://www.davidstreitfeld.com/archive/controversies/hite01.html

[48] Métodos para "imputar" datos faltantes son discutidos en Judith M. Tanur, Samples and Surveys, in Perspectives on Contemporary Statistics 55, 57 (David C. Hoaglin & David S. Moore eds., 1992) y en Howard Wainer, Eelworms, Bullet Holes, and Geraldine Ferraro: Some Problems with Statistical Adjustment and Some Solutions, 14 J. Educ. Stat. 121 (1989) (con su comentario). El caso más simple es cuando la tasa de respuesta es tan elevada que aún si todos los que no responden hubieran respondido de manera opuesta al que hizo la encuesta, la conclusión sustancial no resulta alterada. En cualquier otro caso, la imputación puede ser problemática.

[49] En US, los tribunales usan el término "fiable" para indicar "algo en lo que se puede confiar" para obtener cierto propósito, tal como establecer la causa probable o dar crédito a un testimonio de oídas cuando el declarante no se presenta. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 590 n.9 (1993), por ejemplo, distingue entre "fiabilidad de la evidencia" de la fiabilidad en el sentido técnico de producir resultados consistentes. Usaremos "fiabilidad" en el segundo sentido.

[50] Committee on DNA Forensic Science: An Update, National Research Council, The Evaluation of Forensic DNA Evidence 139–41 (1996). http://www.pnas.org/content/94/11/5498.full.pdf+html

[51] Ver Committee on DNA Tech. in Forensic Science, National Research Council, DNA Technology in Forensic Science 61–62 (1992); David H. Kaye & George F. Sensabaugh, Jr., Reference Guide on DNA Evidence http://www.fjc.gov/public/pdf.nsf/lookup/sciman09.pdf/$file/sciman09.pdf, Reference Manual of Scientific Evidence, Federal Judicial Center, 2000.

En muchos estudios la información descriptiva viene dada mediante números. A efectos estadísticos, es posible que la información haya sido volcada numéricamente – un proceso llamado "codificación". Debe considerarse cuán fiable es este proceso de codificación. Por ejemplo, en el estudio de una condena a muerte en Georgia, los evaluadores entrenados en cuestiones legales examinaron breves resúmenes de casos y los ordenaron según la culpabilidad del acusado.[52] Hay dos tipos distintos de fiabilidad que cabe considerar. Primero, la variabilidad de los juicios "de un mismo observador" debería ser reducida – el mismo evaluador debería tasar esencialmente casos idénticos de la misma forma. Segundo, la variabilidad "entre observadores" debería ser reducida – los distintos evaluadores deberían aplicar la misma tasa al mismo caso.

Validez del proceso de medición La fiabilidad, que es condición necesaria, no es suficiente para asegurar exactitud. Además se requiere "validez". Un instrumento de medición válido mide lo que se supone que debe hacer. Un detector de mentiras mide ciertas respuestas fisiológicas a estímulos. Puede cumplir con esta función de manera fiable. Sin embargo, no será válido como detector de mentiras a menos que aumentos del pulso, de la presión sanguínea, y otros más estén correlacionados con un engaño consciente. Otro ejemplo es el MMPI (Minnesota Multiphasic Personality Inventory), una prueba con lápiz y papel que, según concuerdan varios sicólogos, mide aspectos de la personalidad o del funcionamiento sicológico. Puede cuantificarse su fiabilidad, pero no hacer de la misma una prueba válida de desvío sexual.[53] El Minnesota Multiphasic Personality Inventory es uno de los diagnósticos de la personalidad más usados en cuestiones de salud mental. Es utilizado por profesionales entrenados para asistir en identificar la estructura de la personalidad y su psicopatología.

Cuando se dispone de una forma independiente y razonablemente exacta de medir la variable de interés, puede hacerse una validación del sistema de medición. Las pruebas de alcoholemia pueden ser validadas comparando los niveles de alcohol hallados en muestras de sangre. Las mediciones efectuadas en las pruebas de empleo pueden validarse comparando el desempeño laboral. Para medir la validez se puede calcular el coeficiente de correlación entre criterio (desempeño laboral) y variable predictiva (la prueba de empleo).[54]

Registro correcto Juzgar si la recopilación de datos es adecuada puede implicar examinar el proceso por el cual se registran estas mediciones. Las respuestas a las entrevistas ¿fueron codificadas correctamente? ¿Se incluyeron todas las respuestas a la encuesta? ¿Hay datos faltantes o errores que distorsionen los resultados?[55]

---

[52] David C. Baldus et al., Equal Justice and the Death Penalty: A Legal and Empirical Analysis pp. 49–50 (1990).

[53] Ver People v. John W., 229 Cal. Rptr. 783, 785 (Ct. App. 1986) (se sostuvo que como usar el MMPI para el diagnóstico de un desvío sexual no era generalmente aceptado como procedimiento válido en la comunidad científica, un diagnóstico parcialmente basado en el MMPI era inadmisible) y Allen N. Cowling, The Penile Plethysmograph and the Abel Assessment In False Allegation Cases, Cowling Investigations, Inc. http://www.allencowling.com/false13.htm

[54] Por ejemplo, Washington v. Davis, 426 U.S. 229, 252 (1976) http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=426&invol=229; Albemarle Paper Co. v. Moody, 422 U.S. 405, 430–32 (1975) http://supreme.justia.com/cases/federal/us/422/405/case.html.

[55] Ver p.ej. McCleskey v. Kemp, 753 F.2d 877, 914–15 (11th Cir. 1985) http://www.law.cornell.edu/supct/html/historics/USSC_CR_0481_0279_ZO.html (la corte del distrito no quedó convencida por un análisis estadístico de condena a muerte, en parte por varias imperfecciones del estudio, incluyendo discrepancias y datos faltantes; la opinión concurrente y en disenso concluye que los hallazgos de la corte distrital sobre datos faltantes o mal registrados fueron claramente erróneos porque los posibles errores no eran tan amplios como para afectar los resultados globales; para una exposición del estudio y de la respuesta a las críticas, ver Baldus et al., obra citada, aff'd, 481 U.S. 279 (1987) http://supreme.justia.com/cases/federal/us/481/279/case.html; G. Heileman Brewing Co. v. Anheuser-Busch, Inc., 676 F. Supp. 1436, 1486 (E.D. Wis. 1987) ("varios errores de codificación... afectaron los resultados de la encuesta"); EEOC v. Sears, Roebuck & Co.,

Presentación de los datos Luego de recopilar los datos, deben ser presentados de tal manera que sean inteligibles. Los datos pueden resumirse mediante números o gráficos. Sin embargo, se puede llegar a conclusiones erróneas si se hizo un resumen inapropiado.[56] Hablamos de tasas o porcentajes, con ejemplos de resúmenes que pueden resultar capciosos, y el tipo de preguntas que podrían ser consideradas cuando son presentados resúmenes numéricos en un tribunal. Los porcentajes son usados para demostrar asociación estadística, que es el tópico tratado a continuación. Luego son considerados resúmenes gráficos de los datos, y finalmente discutimos algunos estadísticos descriptivos básicos usados en los litigios, como la media, la mediana y el desvío estándar.

Interpretación de los datos La presentación selectiva de información numérica es como citar a alguien fuera de contexto. Una publicidad televisiva del Investment Company Institute (una asociación de comercio de fondos mutuos) dijo que una inversión de $10.000 hecha en 1950 en un fondo de acciones mutuo habría aumentado hasta $113,500 a fines de 1972. Por otra parte, según el *Wall Street Journal*, la misma inversión distribuida sobre todas las acciones que conforman el New York Stock Exchange Composite Index hubiera aumentado a $151,427. Pero en su totalidad, los fondos mutuos tuvieron una peor performance que el mercado accionario.[57] En este ejemplo, como en muchas otras situaciones, es de gran ayuda ver más de un único número hacia algún punto de referencia que ponga a la cifra aislada en perspectiva.

Mecanismos de Recopilación Los cambios del proceso de recopilación de datos pueden generar problemas de interpretación. La estadística criminal ofrece muchos ejemplos. La cantidad de robos menores informados en Chicago más que se duplicó entre 1959 y 1960 – no porque hubo una ola criminal abrupta, sino porque un nuevo funcionario policial introdujo un sistema de información más avanzado.[58] En los 1970s, los oficiales policiales en Washington, D.C. "demostraron" el éxito de la campaña del Presidente Nixon ley+orden valorando los objetos robados en $49, por debajo del umbral de $50 que requiere su inclusión en el Uniform Crime Reports del Federal Bureau of Investigation (FBI).[59]

Los cambios de los procedimientos de recopilación de datos no se limitan a las estadísticas criminales.[60] En realidad, casi todos los números que abarcan varios años están afectados por cambios de definiciones y métodos de recopilación. Cuando un estudio incluye una serie

---

628 F. Supp. 1264, 1304, 1305 (N.D. Ill. 1986) http://lilt.ilstu.edu/teeimer/Court%20Cases/EEOCvS.htm ("Errores de codificación mecánica de datos del EEOC –organismo de gobierno de US que aplica las leyes federales por discriminación racial, sexual, etc. – a partir de solicitudes de muestras de empleados y desocupados también hacen que el análisis estadístico basado en estos datos sea menos fiable". El EEOC "codificó la experiencia previa de manera que las mujeres con menor experiencia son consideradas como con la misma experiencia y aún mayor experiencia que hombres más experimentados" y "cometió tantos errores de codificación que su base de datos no refleja de forma equitativa las características de los solicitantes de puestos de comisiones de ventas en Sears.") aff'd, 839 F.2d 302 (7th Cir. 1988), Dalley v. Michigan Blue Cross-Blue Shield, Inc., 612 F. Supp. 1444, 1456 (E.D. Mich. 1985) ("si los demandantes demuestran que hubo errores de codificación, aún les queda mostrar que esos errores son tan generalizados y omnipresentes que todo el estudio carece de validez").

[56] Ver en gral. Freedman et al., Huff, Moore, mencionados en nota 1.

[57] Moore, nota 9, pág. 161.

[58] Moore, *ibidem*, pág. 62.

[59] James P. Levine et al., Criminal Justice in America: Law in Action 99 (1986).

[60] Por ejemplo, la mejoría de la tasa de supervivencia de los pacientes de cáncer puede resultar de mejores terapias. O también puede significar que ahora a los pacientes se les detecta el cáncer en forma más temprana, por mejores técnicas de diagnóstico, con lo cual parece que los pacientes vivieran más tiempo. Ver Richard Doll & Richard Peto, The Causes of Cancer: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today pp. 1278–79 (1981).

de tiempo, es útil preguntarse sobre cambios y buscar saltos bruscos, que pueden ser indicativos de tales cambios.[61]

David H. Kaye

¿Son Apropiadas las Categorías? También pueden producirse resúmenes engañosos mediante la elección de las categorías a ser comparadas. Philip Morris and R.J. Reynolds[62] obtuvieron una orden judicial para detener la publicidad de cigarrillos Triumph con bajo contenido de nicotina que afirmaba que los participantes en una prueba nacional preferían Triumph a otras marcas. Los demandantes alegaron que afirmaciones como "Triumph es un ganador a nivel nacional en las preferencias" o que "Triumph barre a las otras marcas" eran falsas y engañosas. A continuación se incluye una tabla producida por el acusado (los símbolos » indican mucho mejor que, ≥ algo mejor que, ≈ aproximadamente igual a, ≤ algo peor que, « mucho peor que):

Tabla usada por el demandado para refutar la tesis de una falsa campaña publicitaria del acusado

	Triumph » Merit	Triumph ≥ Merit	Triumph ≈ Merit	Triumph ≤ Merit	Triumph « Merit
Número	45	73	77	93	36
Porciento	14%	22%	24%	29%	11%

Solamente 14% + 22% = 36% de la muestra prefería Triumph a Merit, y 29% + 11% = 40% prefería Merit a Triumph. Mediante una combinación selectiva de categorías, empero, el acusado intentó crear una impresión diferente. Un 24% halló que ambas marcas eran aproximadamente similares, y 36% prefería a Triumph, el acusado reclamaba para sí que había una mayoría clara (36% + 24% = 60%) que encontraba a Triumph "al menos mejor que Merit". El tribunal se resistió correctamente a aceptar esta chicana, viendo que los números del acusado no aportaban a su causa publicitaria.

Hubo una distorsión similar en reclamos por la precisión de una prueba de embarazo en el hogar.[63] El fabricante publicitó que la prueba era 99.5% exacta en condiciones de laboratorio. Los datos tras la demanda están en la tabla siguiente:

Resultados de un Test de Embarazo

	Embarazada realmente	Realmente No Embarazada
Test = Embarazada	197	0
Test = No Embarazada	1	2
Total	198	2

Esta tabla refleja sólo un error sobre 200 evaluaciones, es decir una exactitud del 99.5%. También deja implícito que el test puede producir dos tipos de errores –le puede decir a una

[61] Moore, *ibidem*, p. 162.

[62] Philip Morris, Inc. v. Loew's Theatres, Inc 511 F. Supp. 855 (S.D.N.Y. 1980) (ver http://www.leagle.com/xmlResult.aspx?xmldoc=19801366511FSupp855_11218.xml&docbase=CSLWAR1-1950-1985) y R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc. (http://www.leagle.com/xmlResult.aspx?xmldoc=19801378511FSupp867_11229.xml&docbase=CSLWAR1-1950-1985).

[63] Este incidente fue informado por Arnold Barnett, How Numbers Can Trick You, Tech. Rev., Oct. 1994, p. 38, 44–45. http://reocities.com/CapitolHill/4834/barnett.htm

embarazada que no lo está (falso negativo), o le puede informar a una embarazada que no lo está cuando la respuesta es afirmativa (falso positivo). La exactitud del 99.5% oculta un hecho crucial, a saber que la empresa no tenía virtualmente datos con que medir la tasa de falsos positivos.[64]

Base del Porcentaje Si bien las tasas y porcentajes proporcionan resúmenes efectivos de los datos, a veces pueden ser mal interpretados. Un porcentaje realiza una comparación entre dos números: la base y otro número comparado con la base. Si la base es pequeña, los números reales serán más reveladores que los porcentajes. Un ejemplo: los diarios informaron en 1982 sobre una ola de crímenes de los más ancianos. El Uniforms Crime Reports[65] decía que casi se había triplicado la tasa de crímenes cometidos por gente mayor desde 1964, y que los crímenes cometidos por gente más joven sólo se habían duplicado. Pero la gente por encima de los 65 años sólo participó en menos del 1% de toda la gente detenida. En 1980, por ejemplo, sólo hubo 151 detenciones por robo sobre 139,476 arrestos totales por robo.[66]

Comparaciones Por fin, está el tema de cuáles son los números comparados. Los investigadores a veces eligen entre comparaciones alternativas. Puede resultar interesante preguntarse por qué lo hicieron así. Si hubieran hecho otra comparación ¿habrían llegado a una presentación distinta? Por ejemplo, un organismo público puede querer comparar los servicios prestados con respecto a los de años precedentes – pero ¿deberían ser esos años precedentes la base de comparación? Si se usa el primer mecanismo es posible esperar un amplio incremento porcentual a causa del problema del punto de partida.[67] Si se usa el último año como base ¿formaría parte de la tendencia, o se trata de un año inusualmente pobre? Si el año base no es representativo de otros años, entonces el porcentaje puede no representar la tendencia de forma imparcial.[68] No hay una única pregunta que pueda detectar estas distorsiones, pero puede ayudar preguntarse cuáles fueron los números con los que se obtuvieron los porcentajes; preguntar sobre la base también puede resultar útil. Sin embargo, en definitiva, reconocer qué números están vinculados con cuáles temas requiere en general un pensamiento claro que no puede reducirse fácilmente a una lista de comprobación.[69]

Uso de una Medida de Asociación Hay casos que implican una asociación estadística. La promoción de un empleado ¿tiene un efecto de exclusión que dependa del género? La incidencia del crimen ¿cambia con la tasa de ejecución de asesinos convictos? ¿Dependen las compras de un producto de la presencia o ausencia de una publicidad sobre el producto? En esta sección vamos a discutir tablas y estadísticas basadas en porcentajes que son frecuentemente presentadas para responder a estas cuestiones.[70]

Frecuentemente se usan porcentajes para describir la asociación entre dos variables. Supongan que se acusa a una universidad de discriminar en contra del sexo femenino en sus admisiones en dos facultades, las técnicas y las facultades de economía. La universidad

[64] Sólo dos mujeres de la muestra no estaban embarazadas; el test produjo resultados correctos en ambos casos. Si bien lo ideal es una tasa de falsos positivos igual a cero, no lo es una estimación basada sobre una muestra de sólo dos mujeres.
[65] http://en.wikipedia.org/wiki/Uniform_Crime_Reports
[66] Mark H. Maier, The Data Game: Controversies in Social Science Statistics 83 (1991). Ver también Alfred Blumstein and Jacqueline Cohen, Characterizing Criminal Careers, Science 28 August 1987: Vol. 237 no. 4818, pp. 985-991. http://www.soc.umn.edu/~uggen/Blumstein_SCI_87.pdf
[67] Ver Michael J. Saks, Do We Really Know Anything About the Behavior of the Tort Litigation System— And Why Not?, 140 U. Pa. L. Rev. 1147, 1203 (1992).
[68] Jeffrey Katzer et al., Evaluating Information: A Guide for Users of Social Science Research 106 (2d ed. 1982).
[69] Para una ayuda en cómo manejarse con porcentajes, ver Zeisel, ob. cit., pp. 1-24.
[70] Hablaremos de regresión y correlación más adelante, en esta sección y en un capítulo especial.

admite a 350 de cada 800 varones postulantes; en comparación, sólo admite a 200 mujeres de cada 600 postulantes. Estos datos pueden ser presentados en la siguiente tabla:

Admisiones por género

Decisión	Varón	Mujer	Total
Admitir	350	200	550
Rechazar	450	400	850
Total	800	600	1600

Como lo indica la tabla, 350/800 = 44% de los varones son admitidos, en comparación con 200/600 = 33% de mujeres. Una manera de expresar esta disparidad es restar ambos números entre sí, 44% - 33% = 11 puntos porcentuales. Si bien restar porcentajes es un procedimiento que se practica a menudo en casos de discriminación de jurados,[71] es inevitable que la diferencia sea pequeña si ambos porcentajes están próximos a 0. Si la tasa de selección de los varones es 5% y la de las mujeres 1%, la diferencia alcanza sólo a 4 puntos porcentuales. Empero, las mujeres tienen sólo 1/5 de la probabilidad que tienen los varones de ser admitidos, y ésa puede ser la preocupación real.[72]

En la tabla anterior, la tasa de selección (utilizada por la Equal Employment Opportunity Commission (EEOC) en su "regla del 80%")[73] es 33/44 = 75%, lo que significa que, en promedio, las mujeres tienen el 75% de probabilidad de ser admitidas que los varones.[74] Pero la tasa de selección no deja de tener problemas, En el último ejemplo, si las tasas de selección fueran 5% y 1%, las tasas de exclusión serían 95% y 99%. La relación correspondiente sería 99/95 = 104%, lo que significa que las mujeres, en promedio, tienen el 104% del riesgo de los varones de ser rechazadas. Se trata de los mismos hechos, pero esta formulación no parece tan preocupante.[75]

---

[71] Ver p.ej. D.H. Kaye, Statistical Evidence of Discrimination in Jury Selection, in Statistical Methods in Discrimination Litigation.

[72] United States v. Jackman, 46 F.3d 1240, 1246–47 (2d Cir. 1995) Aquí se sostiene que el bajo porcentaje de minorías en la población torna "inapropiado" utilizar "un número absoluto" o un enfoque de "impacto absoluto" para determinar la baja representatividad de estas minorías dentro de la lista de jurados potenciales.

[73] La EEOC considera en general que cualquier procedimiento que seleccione candidatos del grupo menos exitoso a una tasa inferior al 80% de la tasa del grupo más exitoso tendrá un impacto adverso. EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (1993). La regla fue diseñada para descubrir ejemplos de prácticas muy discriminatorias, y la comisión pide a los empleadores que justifiquen cualquier procedimiento que produzca una tasa de selección del 80% o inferior (regla de "cuatro quintas partes"). http://law.justia.com/cfr/title29/29-4.1.4.1.8.0.21.4.html

[74] En epidemiología se utiliza un estadístico análogo llamado *riesgo relativo*. Los riesgos relativos son mencionados usualmente como decimales más que como porcentajes; por ejemplo, una tasa de selección del 75% corresponde a un riesgo relativo de 0.75. Hay una variante, la diferencia relativa de proporciones, que expresa en qué proporción se reduce la probabilidad de selección (David C. Baldus & James W.L. Cole, Statistical Proof of Discrimination § 5.1, at 153, 1980 & Supp. 1987) (allí son listadas varias razones que pueden utilizarse para medir disparidades).

[75] El Departamento de Seguridad del Empleo de Illinois intentó explotar esta característica de la tasa de selección en Council 31, Am. Fed'n of State, County and Mun. Employees v. Ward, 978 F.2d 373 (7th Cir. 1992) http://caselaw.findlaw.com/us-2nd-circuit/1202543.html. En enero de 1985, el departamento despidió 8.6% de negros en sus dependencias en comparación con 3.0% de blancos. Como reconoció que estos despidos colisionaban con la regla del 80% (pues 3.0/8.6= 35%, es muy inferior a 80%), en lugar de ello el departamento presentó una tasa de selección para ser retenido (pp. 375-76). Como los empleados negros fueron retenidos en 91.4/97.0 = 94% de la tasa de los blancos, las tasas de retención no exhibieron un impacto adverso con la regla del 80% (p. 376). Al haber una ola subsiguiente de despidos acusada de discriminatoria, el departamento argumentó que "su tasa de retención es el enfoque apropiado para el caso y… muestra de forma concluyente que no tuvo un impacto dispar". El Séptimo Circuito no estuvo de acuerdo y, cuando revirtió una orden en un

El *ratio* de probabilidades es más simétrico. Si 5% de los postulantes varones son admitidos, la probabilidad de que un varón sea admitido es 5/95 = 1/19; para las mujeres es 1/99. El ratio de probabilidades es (1/99) / (1/19) = 19/99. El ratio de probabilidades de rechazo, en lugar de ser aceptado, es semejante, excepto que se invierte el orden.[76] Si bien el ratio de probabilidades tiene propiedades matemáticas deseables, su significado puede resultar más oscuro que el del ratio de selección o la diferencia simple.

Los datos que muestran impactos dispares en general son obtenidos por agregación – reuniendo estadísticas de varias fuentes. A menos que el material de origen sea razonablemente homogéneo, la agregación puede distorsionar los patrones en los datos. Vamos a ilustrar este problema usando los datos de la tabla precedente, pero ahora clasificando no sólo por género y admisión sino también por escuela, como sigue:

Admisiones por género y escuela

	Facultades técnicas		Facultades de economía	
Decisión	Varón	Mujer	Varón	Mujer
Admitir	300	100	50	100
Rechazar	300	100	150	300

Las celdas de esta última tabla totalizan las celdas de la página anterior. Técnicamente, esta tabla se obtiene sumando los datos de la última tabla. Sin embargo, no hay asociación entre género y admisión en ninguna facultad; los varones y las mujeres son admitidos a tasas idénticas. Combinando dos facultades no asociadas da lugar a una facultad en la que el género está fuertemente asociado con la admisión. Explicación de la paradoja: la facultad de economía, a la que se postuló la mayoría de las mujeres, admite relativamente pocos postulantes; a la facultad industrial, a la que se postuló la mayoría de los varones, es más fácil acceder. Este ejemplo ilustra un problema frecuente: la asociación puede surgir de combinar material estadístico heterogéneo.[77]

Gráficos Los gráficos son apropiados para revelar características críticas de un conjunto de números, tendencias a lo largo del tiempo, y relaciones entre las variables.

Tendencias Los gráficos que trazan valores a lo largo del tiempo son útiles para visualizar las tendencias. Sin embargo, hay que prestar atención a las escalas de los ejes. En general, una tendencia es un patrón de comportamiento de los elementos de un entorno particular durante un período de tiempo. En términos del análisis técnico, la tendencia es simplemente la dirección o rumbo del mercado. Pero hay que tener una definición más precisa para trabajar. Es importante entender que los mercados y otros fenómenos no se mueven en línea recta en ninguna dirección. Los movimientos en los precios se caracterizan por un movimiento zigzagueante. Estos impulsos tienen el aspecto de olas sucesivas con sus

juicio sumario a los demandados en otras materias, indicó que la corte distrital “decida qué método de prueba es más adecuado”.

[76] Para las mujeres, la probabilidad de rechazo es de 99 a 1; para los varones, 19 a 1. El cociente de estas probabilidades es 99/19. Asimismo, el cociente de probabilidades para que un postulante sea varón en lugar de ser un postulante varón rechazado también es 99/19.

[77] Estas dos últimas tablas son hipotéticas, pero siguen el patrón de un ejemplo real. Ver P. J. Bickel, E. A. Hammel, and J. W. O'Connell, Sex Bias in Graduate Admissions: Data from Berkeley, 187 Science 398 (1975) http://www.unc.edu/~nielsen/soci708/cdocs/Berkeley_admissions_bias.pdf. Ver también Freedman et al.; y Moore. Las tablas son un ejemplo de la “Paradoja de Simpson”. Ver en general Myra L. Samuels, Simpson’s Paradox and Related Phenomena, 88 J. Am. Stat. Ass’n 81 (1993). Puede ser de utilidad tener otra perspectiva sobre la tabla de la página anterior. La escuela a la que se postula un estudiante es una variable confusiva. En el contexto actual, a las variables confusivas se las suele llamar “variables omitidas”.

respectivas crestas y valles. La dirección de estas crestas y valles es lo que constituye la tendencia del mercado, ya sea que estos picos y valles vayan al alza, a la baja o tengan un movimiento lateral.

Representación de las distribuciones Una forma común de representar una distribución es mediante su histograma, que es un gráfico de frecuencias tabuladas, indicadas mediante barras. Representa qué proporción de casos cae dentro de cada una de las distintas categorías. Uno de los ejes representa los números, y el otro indica cuántos de estos números caen dentro de intervalos especificados (llamados "intervalos de clase").

Figura 1. El Euro/Dólar tuvo una tendencia bajista de 1999 a 2000 (A), así como durante 2005 (D). Desde fines de 2000 a 2002 mantuvo una tendencia neutral (B). Se observan dos periodos de tendencia alcista en la cotización, el primero entre 2002 y 2004 (C) y el segundo a partir de enero de 2006 (E).

Las categorías se representan usualmente mediante intervalos no traslapados de alguna variable. Las categorías (barras) deben ser adyacentes. Los intervalos (o bandas) en general son del mismo tamaño, pero esto último no es necesario. Los histogramas son utilizados para graficar la densidad de los datos, y a veces la estimación de la función de distribución de probabilidad de la variable subyacente. El área total de un histograma utilizado para graficar la densidad de probabilidad siempre se normaliza igual a la unidad. Entonces, si la longitud de los intervalos del eje de las x es 1, el histograma es idéntico a un gráfico de frecuencias relativas. En el diagrama de la Figura 2 se incluye un histograma que muestra la frecuencia de las llegadas por minuto de un cierto medio de transporte de pasajeros. Hay histogramas donde se agrupan los datos en clases, y se cuenta cuántas observaciones (frecuencia absoluta) hay en cada una de ellas. En algunas variables (variables cualitativas) las clases están definidas de modo natural, p.ej. sexo con dos clases: mujer, varón, o grupo sanguíneo con cuatro: A, B, AB, 0. En las variables cuantitativas, hay que definir las clases explícitamente (intervalos de clase).

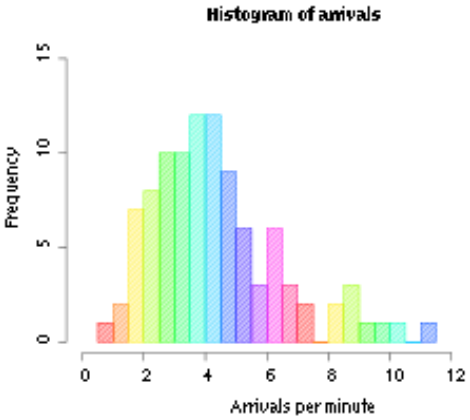


Figura 2. En este histograma se representa la frecuencia de las observaciones dentro de cierto rango de valores, y en base al mismo se puede estimar la distribución de probabilidad de una variable.

Centro de la Distribución Tal vez el estadístico descriptivo más familiar sea la media o el promedio (o "media aritmética"). La obtenemos sumando todos los números y dividiendo por cuántos números hay. En comparación, la mediana se define de tal forma que la mitad de los números sean mayores que la mediana, y la mitad restante inferiores.[78] Un tercer

[78] Técnicamente, al menos la mitad de los números están en la mediana o son mayores; y al menos la mitad se hallan en la mediana o son menores. La mediana coincide con el percentil 50, con el segundo cuartil y con el quinto decil. Para una distribución simétrica, la media es igual a la mediana. Pero los valores son distintos en distribuciones asimétricas, o sesgadas. La distinción entre mediana y media resulta crítica para interpretar la Ley llamada Railroad Revitalization and Regulatory Reform Act, 49 U.S.C. § 11503 (1988), que prohíbe fijar impuestos a la propiedad de los ferrocarriles a una mayor tasa que a otras propiedades comerciales e industriales. A fin de comparar los impuestos, las autoridades tributarias a menudo usan la media, mientras que los ferrocarriles prefieren la mediana.

estadístico es el modo, que es el número más frecuente de un conjunto de números.[79] Aunque son estadísticos todos diferentes uno del otro, no siempre se los distingue claramente.[80] La media significa tomar en cuenta todos los datos – porque involucra la totalidad de números; sin embargo, sobre todo con pocos datos, unos escasos números muy grandes o pequeños pueden influir demasiado sobre la media. En cambio, la mediana es más resistente a estos valores extremos.

Para ilustrar la distinción entre media y mediana, tomemos el caso de un informe acerca de que la indemnización "media" por casos de mala praxis subió de $220.000 en 1975 a más de $1 millón en 1985.[81] La indemnización mediana fue ciertamente muy inferior a $1 millón,[82] y el crecimiento aparentemente explosivo puede resultar de unas pocas indemnizaciones muy grandes. Pero si la cuestión es determinar si los aseguradores experimentaron más costos por los veredictos del jurado, el estadístico más apropiado es la media: ya que las indemnizaciones totales están directamente vinculadas con la media, no con la mediana.[83]

La elección que se realice tiene consecuencias financieras importantes, de lo cual resultaron muchos litigios. Ver David A. Freedman, The Mean Versus the Median: A Case Study in 4-R Act Litigation, 3 J. Bus. & Econ. Stat. 1 (1985).

[79] Hablamos de una distribución bi-modal de los datos cuando encontramos dos modos, es decir, dos datos que tengan la misma frecuencia absoluta máxima. En una distribución tri-modal de los datos hallamos tres modos. Si todas las variables tienen la misma frecuencia diremos que no hay modo.

[80] En lenguaje común, la media aritmética, la mediana y el modo parecen referirse de modo indistinto al "promedio". En estadística, cuando decimos media se trata de media aritmética. Hay un ejemplo para sacar a la luz las diferencias entre estos conceptos: ¿Cuán grande sería el error cometido si todos los números de una canasta fueran reemplazados por el "centro" de la canasta? El modo minimiza el *número de errores*, pues todos se cuentan igual, cualquiera sea su tamaño. La mediana minimiza un tipo distinto de error – la suma de las diferencias entre el centro y los puntos; no se toman en cuenta los signos al computar esta suma, de modo que las diferencias positivas y negativas son tratadas de forma similar. La media minimiza la suma de los cuadrados de las diferencias.

[81] Kenneth Jost, Still Warring Over Medical Malpractice: Time for Something Better, A.B.A. J., May 1993.

[82] Un estudio de casos de North Carolina informó sobre una indemnización "promedio" (media) de $368,000, y una indemnización mediana de sólo $36,000 (p. 71). En TXO Production Corp. v. Alliance Resources Corp., 509 U.S. 443 (1993), http://www.leagle.com/xmlResult.aspx?page=16&xmldoc=1993952509US443_1939.xml&docbase=CSLWAR2-1986-2006&SizeDisp=7 los resúmenes que describían el sistema de daños punitivos informaron que las indemnizaciones punitivas promedio eran diez veces mayores que las indemnizaciones medianas descriptas en informes que defendían al sistema existente de daños punitivos. Ver Michael Rustad & Thomas Koenig, The Supreme Court and Junk Social Science: Selective Distortion in Amicus Briefs, 72 N.C. L. Rev. 91, 145–47 (1993). La media difiere tan dramáticamente de la mediana porque la media toma en cuenta (en la práctica, está muy influida por) las magnitudes de unas pocas indemnizaciones muy grandes; la mediana las oculta. Por supuesto, lo difícil es hallar datos representativos de los veredictos y de las indemnizaciones. Un estudio que utilizó muestras probabilísticas de casos es el de Carol J. DeFrances, Steven K. Smith, Patrick A. Langan, Brian J. Ostrom, David B. Rottman, y John A. Goerdt, Civil Jury Cases and Verdicts in Large Counties, Bureau of Justice Statistics, Special Report, July 1995, NCJ-154346. http://bjs.ojp.usdoj.gov/content/pub/pdf/cjcavilc.pdf

[83] Para obtener las indemnizaciones totales, lo único que hay que hacer es multiplicar la media por el número de indemnizaciones; en contraste, el total no puede ser computado a partir de la mediana. (El número más pertinente para la industria aseguradora no es el total de indemnizaciones otorgadas por los tribunales, sino la experiencia de reclamos reales que incluyen estos acuerdos; naturalmente, aún el riesgo de una indemnización elevada puede tener un impacto considerable). Para continuar con el tratamiento de ésta y otras cuestiones vinculadas, ver Theodore Eisenberg & Thomas A. Henderson, Jr., Inside the Quiet Revolution in Products Liability, 39 UCLA L. Rev. pp. 731, 764–72 (1992); Scott Harrington & Robert E. Litan, Causes of the Liability Insurance Crisis, 239 Science pp. 737, 740–41 (1988).

Medida de Variabilidad Localizar el centro de un conjunto de números no dice nada sobre las variaciones que exhiben estos números.[84] Las medidas estadísticas de variabilidad incluyen el rango, el rango ínter-cuartil, y el desvío estándar. El rango es la diferencia entre el número más grande del conjunto y el más pequeño. Es un concepto natural, que indica la brecha máxima entre los números, pero en general es muy inestable porque depende de los valores más extremos.[85] El intervalo ínter-cuartil es la diferencia en los percentiles 25º y 75º.[86] El intervalo ínter-cuartil contiene el 50% de los números y resulta resistente a cambios de los valores extremos. El desvío estándar es una especie de desvío de la media.[87] No hay reglas sólidas ni rápidas para saber qué estadísticos son los mejores. En general, cuanto mayor sean estas medidas de desvío, más dispersos estarán los números. En particular en pequeños conjuntos de datos, el desvío estándar puede estar muy influido por unos pocos valores extremos. Para eliminar esta influencia, se pueden re-computar la media y el desvío estándar sacando los valores extremos. Más allá, los estadísticos pueden ser complementados con una cifra que indique la mayoría de los datos.[88]

4. Inferencias y Estimación

Las inferencias que puedan extraerse dependerán de la calidad de los datos y del diseño del estudio. Como se discutió previamente, los datos pueden no estar vinculados con lo que se intenta investigar, pueden estar errados sistemáticamente, o puede ser difícil interpretarlos por la presencia de variables confusivas. Ahora analizaremos una cuestión adicional – los errores aleatorios.[89] ¿Es el patrón de los datos resultado del azar? ¿Podríamos limpiar ese patrón mediante la recopilación de datos adicionales?

---

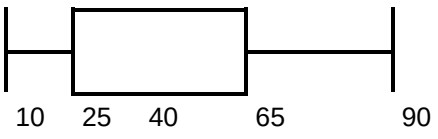
[84] Los números 1, 2, 5, 8, 9 tienen a 5 como media y mediana. Lo mismo sucede con los números 5, 5, 5, 5, 5. En el primer conjunto, los números varían en forma considerable alrededor de la media; en el segundo, no hay ningún tipo de variación.

[85] Es típico que el rango aumente con el tamaño de la muestra, e.d. el número de unidades muestreadas.

[86] Por definición, 25% de los datos están abajo del 25º percentil, 90% abajo del 90º percentil, etc. Luego, la mediana es el 50º percentil.

[87] Como se verá en el Apéndice, cuando la distribución sigue una ley normal, 68% de los datos se hallará en la proximidad de un desvío estándar de la media, y 95% de dos desvíos estándar de la media. Para otras distribuciones, las proporciones de datos dentro de un número especificado de desvíos estándar será distinta. Técnicamente, el desvío estándar es la raíz cuadrada de la varianza; la varianza es la media de los desvíos de la media al cuadrado. Por ejemplo, si la media es 100, el dato 120 está desviado de la media en 20, y su cuadrado es 20^2=400. Si la varianza (es decir, la media de todos los desvíos al cuadrado) es 900, luego el desvío estándar es la raíz cuadrada de 900, es decir $\sqrt{900}$=30. Entre otros aspectos, al tomarse la raíz cuadrada se corrige el hecho de que la varianza está en una escala diferente que las propias mediciones. Por ejemplo, si las mediciones de longitud están en metros, la varianza estará en metros cuadrados; al tomar la raíz cuadrada se vuelve a estar en metros. Para comparar distribuciones en distintas escalas, puede utilizarse el *coeficiente de variación*, igual al desvío estándar, expresado como porcentaje de la media. Sea por ejemplo el conjunto de números 1, 4, 4, 7, 9. La media es 25/5=5, la varianza es (16+1+1 +4+16)/5= 7.6, el desvío estándar es $\sqrt{7.6}$=2.8. El coeficiente de variación es 2.8/5=56%.

[88] Por ejemplo, el "resumen de cinco números" proporciona una lista del valor más reducido, el 25º percentil, la mediana, el 75º percentil, y el valor más elevado. Este resumen puede ser presentado como una caja. Si los cinco números fueran 10, 25, 40, 65 y 90, la caja tendría la siguiente apariencia:



Hay muchas variantes de esta idea, donde las fronteras de la caja, o los "bigotes" que se extienden a partir de ella, representan números levemente diferentes de la distribución de números.

[89] El error aleatorio también es denominado error muestral, error al azar, o error estadístico. Los econometristas usan el concepto paralelo de término de perturbación aleatoria.

Las leyes probabilísticas son fundamentales para analizar los errores aleatorios. Mediante su aplicación, el estadístico puede evaluar el impacto posible de un error aleatorio, usando "errores estándar", "intervalos de confianza", "probabilidades significativas", "test de hipótesis" o "distribuciones de probabilidad posteriores". El ejemplo siguiente ilustra estas ideas: Un empleador planifica usar un examen estándar para seleccionar aprendices de un *pool* de 5,000 varones y 5,000 mujeres postulantes. Este pool de 10,000 postulantes es la "población" estadística. Según el Título VII de la Ley de Derechos Civiles de US de 1964,[90] si el examen propuesto excluye a una cantidad desproporcionada de mujeres, el empleador está obligado a demostrar que el examen está vinculado con el empleo.[91]

Para ver si hay un impacto dispar, el empleador administra un examen a una muestra de 50 varones y 50 mujeres extraídos al azar de la población de postulantes al cargo. En esta muestra, 29 varones pasan la prueba, pero sólo lo hacen 19 mujeres; las tasas de éxito muestral son por consiguiente 29/50=58% y 19/50=38%. El empleador anuncia que de cualquier modo utilizará un examen, y varios postulantes llevan adelante una acción bajo el Título VII. Parece claro que existe un impacto dispar. La diferencia de tasas de éxito es de 20 puntos porcentuales: 58%-38% = 20%. Pero el empleador argumenta que la disparidad podría deberse a un error muestral. Después de todo, sólo una pequeña fracción de gente hizo la prueba, y ésta pudo haber incluido un número más que proporcional de varones con una puntuación elevada y damas de baja puntuación. Está claro que, aunque no haya diferencias entre las tasas de éxito de los postulantes varones y femeninas, en algunos casos los varones podrán superar el puntaje de las últimas. En general, hay que tener en cuenta que una muestra no es un perfecto microcosmos de la población; los estadísticos llaman a las diferencias entre la muestra y la población, sólo por el azar de elegir una muestra, el "error muestral" o "error aleatorio". Cuando se evalúa el impacto del error aleatorio, un estadístico debe considerar los tópicos siguientes:

Estimación. Los demandantes utilizan la diferencia de 20 puntos porcentuales entre los varones y las damas de la muestra para estimar la disparidad entre todos los postulantes varones y mujeres. ¿Es buena esta estimación? La precisión puede expresarse usando los conceptos de "error estándar" o de "intervalo de confianza".

Significación estadística Supongan que el demandado está en lo cierto, y que no hay impacto dispar: en la población de 5,000 varones y 5,000 mujeres postulantes, las tasas de éxito son iguales. *¿Cuán probable resulta que una muestra aleatoria de 50 varones y 50 mujeres dé lugar a una disparidad de 20 puntos porcentuales o más*? A esta probabilidad se la conoce como el *p*-valor. La significación estadística se determina con referencia al *p*-valor, y el "contraste (o test) de hipótesis" es la técnica para computar *p*-valores o para determinar los niveles de significación.[92]

Probabilidades posteriores. Dada la disparidad observada de 20 puntos porcentuales de la muestra, ¿cuál es la probabilidad de que – considerando a toda la población – hombres y mujeres tengan tasas de éxito similares? Esta pregunta resulta de interés directo para los tribunales. Para un estadístico subjetivista, las probabilidades posteriores pueden ser

[90] http://en.wikipedia.org/wiki/Civil_Rights_Act_of_1964

[91] El caso seminal case es Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971) http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=401&invol=424. Los requisitos y procesos de validación de los exámenes pueden ir más allá de demostrar que existe una vinculación con el trabajo. Ver p.ej. Richard R. Reilly, Validating Employee Selection Procedures, in Statistical Methods in Discrimination Litigation, p. 133; Michael Rothschild & Gregory J. Werden, Title VII and the Use of Employment Test: An Illustration of the Limits of the Judicial Process, 11 J. Legal Stud., 261 (1982).

[92] Al "test de hipótesis" también se lo llama "test de significación". En el Apéndice veremos un ejemplo.

computadas utilizando la “regla de Bayes”. Sin embargo, dentro del marco de la teoría estadística clásica, este cálculo carece de significado.[93]

Aplicabilidad de modelos estadísticos La inferencia estadística – ya sea realizada mediante intervalos de confianza o probabilidades significativas, métodos objetivos o subjetivos – depende de la validez de los modelos estadísticos para los datos. Si los datos han sido recolectados en base a una muestra de probabilidad o a un experimento randomizado, habrá modelos estadísticos que calcen muy bien con la situación, y las inferencias que se obtengan con estos modelos serán bastante sólidas. En otro caso, los cálculos estarán basados en general en un razonamiento por analogía: este grupo de gente es como si fuera una muestra aleatoria, aquel estudio de observaciones es como un experimento randomizado. Entonces el ajuste entre el modelo estadístico y los datos puede requerir un examen adicional: ¿es aceptable esta analogía?

Estimación Un estimador es un *estadístico* computado a partir de los datos muestrales para estimar ciertas características numéricas de la población.[94] Por ejemplo, hemos usado la diferencia de las tasas de éxito en una muestra de hombres y mujeres para estimar la disparidad correspondiente en la población de todos los postulantes. En la muestra las tasas de éxito era 58% y 38%; la diferencia de tasas de toda la población se estimó en 20 puntos porcentuales: 58%-38% = 20%. En problemas más complejos, los estadísticos deben optar entre diversos estimadores. En general, se prefieren los estimadores que tienden a registrar errores más pequeños. Esta idea, no obstante, puede formularse de modo más preciso de varias formas,[95] lo que deja cabida para el juicio al elegir un estimador.

Error Estándar e Intervalo de Confianza Un estimador basado en una muestra es probable que no dé en el blanco, al menos por escaso margen, debido al error aleatorio. El error estándar proporciona la magnitud probable de este error aleatorio.[96] Toda vez que sea posible, *un estimador debería estar acompañado por su error estándar*. En este ejemplo, el error estándar está alrededor de 10 puntos de porcentaje: el estimador de 20 puntos de porcentaje es probable que esté errado en unos 10 puntos porcentuales, o algo así, en cualquier dirección.[97] Como no conocemos en realidad las tasas de éxito de los 5,000 varones y las 5,000 mujeres, no podemos decir exactamente cuán alejado está el estimador, pero 10 puntos porcentuales proporciona una magnitud verosímil del error.

Los intervalos de confianza dan una idea más precisa. Los estadísticos que dicen que las diferencias poblacionales caen entre más-menos 1 error estándar de las diferencias muestrales estarán diciendo lo correcto un 68% de las veces. Dicho en forma más

[93] El contexto clásico también es denominado “objetivista” o “frecuencialista”, en contraste con el enfoque “Bayesiano” o “subjetivista”. Dicho en forma breve, los estadísticos *objetivistas* consideran que las probabilidades son propiedades objetivas del sistema estudiado. Los *subjetivistas* ven a las probabilidades como si midieran grados de creencia subjetivos. Más adelante explicamos por qué las probabilidades posteriores se excluyen del cálculo clásico, y también se discute brevemente la posición subjetivista. Para consideraciones adicionales, véase David Freedman, Some issues in the foundation of statistics, Foundations of Science, Volume 1 (1995/6), Number 1, 19-39. http://mangellabs.soe.ucsc.edu/sites/default/files/16/freedman_antibayes.pdf

[94] http://en.wikipedia.org/wiki/Estimator

[95] Además, reducir el error en un contexto puede aumentarlo en otros: también puede existir un compromiso o *trade-off* entre precisión y sencillez.

[96] Al “error estándar” también se lo llama “desvío estándar”, y (en US) los tribunales y varios autores prefieren esta última denominación.

[97] El error estándar depende de las tasas de éxito de los varones y de las chicas en la muestra, y del tamaño de la muestra. Con muestras grandes, el error al azar será más reducido, con lo cual el error estándar decrecerá a medida que aumente el tamaño de la muestra (“Tamaño de la muestra” es la cantidad de individuos incluidos en la muestra). Más sobre este punto en el Apéndice. En general, la fórmula del error estándar debe tomar en cuenta tanto el método usado para extraerla como la naturaleza del estimador. Elegir la fórmula correcta requiere experiencia estadística.

compacta, abreviaremos error estándar como “SE”. Un intervalo de confianza al 68% es el rango

Estimador – 1 SE al estimador + 1 SE.

En el ejemplo, el intervalo de confianza al 68% va desde 10 a 30 puntos porcentuales. Si se desea tener un mayor nivel de confianza, el intervalo de confianza deberá ser ampliado. El intervalo de confianza al 95% es alrededor de

Estimador – 2 SE al estimador + 2 SE.

Este intervalo va desde 0 a 40 puntos porcentuales.[98] Si bien los intervalos de confianza al 95% son usados en forma frecuente, no hay nada especial en 95%. Por ejemplo, también podría usarse un intervalo de confianza al 99.7%:

Estimador – 3 SE al estimador + 3 SE.

Este intervalo va de -10 a 50 puntos porcentuales.

Hasta este punto, hemos llegado a que un estimador basado en una muestra diferirá del valor exacto de la población, debido al error aleatorio; el error estándar mide el tamaño probable del error aleatorio. Si el error estándar es pequeño, el estimador probablemente nos está diciendo la verdad. Si el error estándar es amplio, el estimador puede estar seriamente equivocado. Los intervalos de confianza son una suerte de refinamiento técnico, y “confianza” es un término artístico.[99] A determinado nivel de confianza, un intervalo más estrecho indica un estimador más preciso. Un elevado nivel de confianza de por sí no

[98] Como veremos en el Apéndice, los niveles de confianza son leídos habitualmente a partir de la curva normal. (Técnicamente, el área por debajo de la curva normal entre -2 y +2 está más próxima a 95.4% que 95%: por dicho motivo, los estadísticos utilizan con frecuencia la notación ±1.96 SE para un intervalo de confianza al 95%.) Empero, la curva normal sólo proporciona una aproximación de las probabilidades relevantes, y el error de esa aproximación será a menudo mayor que la diferencia entre 95.4% y 95%. Para simplificar, hablamos de una confianza al 95% utilizando ±2 SE. De la misma forma, usaremos ±1 SE para una confianza al 68%, aunque el área por debajo de la curva entre -1 y +1 está más próxima a 68.3%. La curva normal proporciona buenas aproximaciones cuando el tamaño muestral es grande; para muestras pequeñas, se deben usar otras técnicas.

[99] Dentro de la teoría estándar de la estadística frecuencialista, no es permitido efectuar enunciados de probabilidad sobre las características de la población. Ver, por ejemplo, David Freedman et al., Statistics (3d ed. 1998), pp. 383-386. En consecuencia, es impreciso sugerir que “un intervalo de confianza al 95% significa que existe una probabilidad de 95% de que el verdadero riesgo relativo caiga dentro del intervalo”. Ver también DeLuca v. Merrell Dow Pharms., Inc., 791 F. Supp. 1042, 1046 (D.N.J. 1992), aff’d, 6 F.3d 778 (3d Cir. 1993). http://caselaw.findlaw.com/tx-supreme-court/1013264.html A causa del significado limitado que tiene el término “confianza”, se ha sostenido que el término es equívoco y que debería ser reemplazado por otro más neutro, como “coeficiente de frecuencias”, en las presentaciones en los tribunales. Ver David H. Kaye, Is Proof of Statistical Significance Relevant?, 61 Wash. L. Rev. 1333, 1354 (1986); SSRN http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1411859. Hay otro malentendido, a saber que el nivel de confianza suministra la probabilidad de que estimadores repetidos caigan dentro del intervalo de confianza. P.ej. Turpin v. Merrell Dow Pharms., Inc., 959 F.2d 1349, 1353 (6th Cir. 1992) (“un intervalo de confianza de 95 por ciento entre 0.8 y 3.10... significa que la repetición aleatoria del estudio debería dar como resultado, el 95% del tiempo, un riesgo relativo comprendido en algún punto entre 0.8 y 3.10”); United States ex rel. Free v. Peters, 806 F. Supp. 705, 713 n.6 (N.D. Ill. 1992) (“Un intervalo de confianza al 99%, por ejemplo, indica que si el experimento fuera repetido 100 veces bajo idénticas condiciones, 99 veces de esas 100 el estimador puntual derivado de la experimentación repetida caería dentro del intervalo inicial del estimador...”), rev’d in part, 12 F.3d 700 (7th Cir. 1993). *El nivel de confianza no proporciona el porcentaje de veces que estimadores repetidos caen dentro del intervalo, sino el porcentaje de veces que intervalos de muestras repetidas abarcan el valor verdadero*.

significa demasiado,[100] pero un elevado nivel de confianza para un intervalo pequeño es impresionante,[101] lo que indica que el error aleatorio del estimador muestral es reducido.

Los errores estándar e intervalos de confianza se derivan usando modelos estadísticos del proceso que generó los datos.[102] Si los datos provienen de una muestra probabilística o de un experimento controlado al azar, el modelo estadístico puede vincularse en forma estrecha con el proceso de recopilación de datos. En otros casos, usar el modelo puede ser equivalente a suponer que una muestra de conveniencia constituye una muestra al azar, o que un estudio de observaciones es un experimento al azar, o parecido.

Los errores estándar e intervalos de confianza ignoran en general los errores sistemáticos como el sesgo de selección y el sesgo por ausencia de respuesta; en otros términos, se supone que estos sesgos son despreciables. Por ejemplo, un tribunal – revisando estudios acerca de si una medicación particular causaba defectos de nacimiento – observó que era más probable que las madres de otros niños con defectos de nacimiento recordaran haber tomado la medicación durante el embarazo que mujeres con niños normales.[103] Esta memoria selectiva imprimiría un sesgo a comparaciones de las muestras de los grupos de

[100] Los enunciados sobre confianza en una muestra sin mencionar el intervalo de confianza carecen prácticamente de sentido. En Hilao v. Estate of Marcos, 103 F.3d 767 (9th Cir. 1996), por ejemplo, "un experto estadístico testificó que... una muestra aleatoria de 137 demandas lograría 'una probabilidad del 95% de que la misma demanda válida entre las demandas examinadas fuera aplicable a la totalidad [9,541 casos] de demandas efectuadas." (p.782). Desafortunadamente, no existe una "probabilidad estadística" de 95% de que un porcentaje computado con una muestra aleatoria sea "aplicable" a la población. Se puede computar un intervalo de confianza a partir de una muestra aleatoria y estar 95% confiado de que el intervalo abarque algún parámetro. Esto puede hacerse con muestras de cualquier tamaño, siendo las más grandes las que proporcionan intervalos menores. Lo que le faltó a la opinión fue discutir la extensión de los intervalos relevantes.

[101] Recíprocamente, un amplio intervalo es señal de que el error aleatorio es sustancial. En Cimino v. Raymark Industries, Inc., 751 F. Supp. 649 (E.D. Tex. 1990) http://www.leagle.com/xmlResult.aspx?xmldoc=19901400751FSupp649_11283.xml&docbase=CSLWAR2-1986-2006, la corte distrital extrajo muestras aleatorias de más de 6,000 casos pendientes, elevó a juicio los casos, y utilizó los resultados para estimar la indemnización total a pagar a los demandantes de los casos pendientes. La corte fijó entonces una audiencia para determinar si las muestras eran suficientemente grandes para proporcionar estimadores precisos. El experto de la corte, un psicólogo educativo, testificó que los estimadores eran precisos porque las muestras estaban apareadas con la población en características tales como la raza y el porcentaje de los demandantes que estaban aún con vida (p. 664). Sin embargo, el apareamiento sólo ocurría en sentido de que las características de la población caían dentro de intervalos de confianza a los 99% muy amplios computados con las muestras. El tribunal pensó que apareamientos de intervalos de confianza al 99% demostraban más que los intervalos al 95% (Id.) Lamentablemente, es al revés. *Ser correcto en unos pocos casos al 99% de confianza no resulta demasiado difícil – por definición, estos intervalos son suficientemente amplios como para asegurar la cobertura el 99% de las veces.*

[102] En general, los modelos estadísticos permiten al analista computar la probabilidad de los distintos resultados posibles. Ejemplo: el modelo puede contener parámetros, es decir, constantes numéricas que describen la población de la cual fueron extraídas las muestras. Éste es nuestro caso presente, donde un parámetro es la tasa de éxito de 5,000 postulantes varones, y otro parámetro es la tasa de éxito de 5,000 postulantes mujeres. Como está explicado en el Apéndice, estos parámetros pueden ser utilizados para computar la probabilidad de que se obtenga una diferencia muestral dada. El uso de modelos con parámetros conocidos para hallar la probabilidad de un resultado dado (o uno semejante) es común en los casos en que se alega discriminación en la selección de jurados (P.ej. Castaneda v. Partida, 430 U.S. 482, 496 (1977) http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=430&invol=482; ver Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 311 n.17 (1977) http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=us&vol=433&invol=299 (usa el cómputo de probabilidades para seleccionar maestros negros). Pero si el valor de los parámetros es desconocido, el estadístico deberá inferirlos usando datos muestrales. Éste es el tipo de inferencia estadística descripto en esta sección.

[103] Brock v. Merrell Dow Pharms., Inc., 874 F.2d 307, 311–12 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989).

mujeres. El error estándar de la diferencia estimada de uso de la medicación entre ambos grupos ignora este sesgo. Otro tanto sucede con el intervalo de confianza.[104] En forma similar, el error estándar no toma en cuenta problemas inherentes a las muestras de conveniencia en lugar de las muestras aleatorias.

Nuestro ejemplo está basado en una muestra al azar, lo que justificó los cálculos estadísticos.[105] Hay contextos donde elegir un modelo estadístico apropiado no resulta obvio.[106] Cuando un modelo no se ajusta suficientemente bien a los datos, los estimadores y errores estándar probarán menos.[107]

---

[104] En el caso Brock, la corte estableció que el intervalo de confianza toma en cuenta el sesgo (bajo la forma de memoria selectiva) así como el error aleatorio. 874 F.2d at 311–12. http://openjurist.org/874/f2d/1136 Los autores del Manual disienten. "Aunque no hubiera error muestral – tal sería el caso si se pudiera entrevistar a todas las mujeres que tuvieron hijos durante el período en que la medicación estuvo disponible - la memoria selectiva produciría una diferencia de los porcentajes de exposición a la medicación de madres de niños con defectos de nacimiento y los niños normales. En esta situación hipotética, el error estándar se anularía. Por consiguiente, el error estándar no podría revelar nada sobre el impacto de la memoria selectiva. Lo mismo es válido en presencia de error muestral."

[105] Se verá en el Apéndice que las muestras grandes pueden dar lugar a ciertos estadísticos que están normalmente distribuidos. En parte, debido a que la Corte Suprema usó un modelo de este tipo en el caso Hazelwood y Castaneda, los tribunales y los abogados descreen de análisis que den lugar a otros tipos de variables aleatorias. Ver p.ej. EEOC v. Western Elec. Co., 713 F.2d 1011 (4th Cir. 1983), discutido en David H. Kaye, Ruminations on Jurimetrics: Hypergeometric Confusion in the Fourth Circuit, 26 Jurimetrics J. 215 (1986) y SSRN http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1411758. Pero ver también Branion v. Gramly, 855 F.2d 1256 (7th Cir. 1988) http://openjurist.org/855/f2d/1256/branion-v-b-gramly (se cuestiona un supuesto aparentemente arbitrario de normalidad), discutido en David H. Kaye, Statistics for Lawyers and Law for Statistics, 89 Mich. L. Rev. 1520 (1991) (se defiende el uso de la aproximación normal); Michael O. Finkelstein & Bruce Levin, Reference Guide on Statistics: Non Lasciare Esperanza, 36 Jurimetrics J. 201, 205 (1996) (ensayo de revisión) ("El tribunal rechazó correctamente la distribución normal. . . ."). Que una variable esté normalmente distribuida es una cuestión estadística o empírica, y no del derecho.

[106] Ver más adelante. Para ejemplos de interés legal, ver p.ej., Mary W. Gray, Can Statistics Tell Us What We Do Not Want to Hear?: The Case of Complex Salary Structures, 8 Stat. Sci. 144 (1993) http://users.stat.umn.edu/~sandy/courses/8801/articles/Law/graylaw.pdf; Arthur P. Dempster, Employment Discrimination and Statistical Science, 3 Stat. Sci. 149 (1988). Un estadístico planteó la cuestión en los siguientes términos: "Los datos existentes pueden ser vistos desde más de una perspectiva, y representarse mediante un modelo de más de una forma. Es bastante común que no exista un modelo único que sea el "verdadero" o el correcto; justificar una conclusión fuerte puede requerir conocimiento del que simplemente se carece. Luego, es raro que un conjunto de datos sea analizado de formas aparentemente distintas. Si las conclusiones concuerdan en términos cualitativos, ello puede ser visto como una base para confiar en las mismas. Pero es frecuente que se aplique un solo modelo, y que los datos sean analizados de acuerdo con ese modelo... Luego es frecuente que un conjunto de datos se analice desde varios puntos de vista. Si las conclusiones concuerdan en los aspectos cualitativos, ello se ve como una base para adjudicarles una confianza adicional. Pero es frecuente que se aplique un único modelo, y que los datos sean analizados de acuerdo con él... Las características deseables incluyen (i) que sea manejable, (ii) su parsimonia, y (iii) realismo. Que exista cierta tensión entre estas exigencias no debe sorprender. *Que sea manejable*. Un modelo es tratable en un primer sentido si es fácil de entender y de explicar. Que sea tratable desde el punto de vista computacional también puede ser ventajoso, pero si existe computación barata no debe ponderarse este requerimiento por demás. *Parsimonia*. La sencillez, como que sea manejable, debe ser también evaluada en forma positiva, no ignorada en forma olímpica – pero tampoco debe ser sobrevaluada. Si hay varios modelos plausibles y algunos de ellos se adaptan en forma adecuada a los datos, entonces al elegir entre ellos un criterio sería preferir aquel modelo que resulte más simple que los demás. *Realismo*: ...En primer término, ¿refleja bien el modelo cómo funciona el proceso real [el proceso generador de datos]? Esta pregunta, en realidad, es un abanico completo de preguntas, algunas sobre las distribuciones de los errores aleatorios, otras sobre las relaciones matemáticas entre [variables y] parámetros. A este segundo aspecto a veces se

Los p-valores En el ejemplo, 50 varones y 50 mujeres fueron extraídos al azar de 5,000 varones y 5,000 mujeres postulantes. Se les tomó un examen, y en la muestra, los porcentajes de éxito de los varones y de las chicas fueron 58% y 38%, respectivamente. La diferencia muestral de tasas de éxito fue 58%-38% = 20%. El *p*-valor trata de responder a la siguiente pregunta: Si las tasas de éxito de los 5,000 postulantes masculinos y las 5,000 postulantes femeninas fueran idénticas, ¿cuán probable sería hallar una discrepancia tan alta o mayor que el 20% observada en la muestra? La pregunta es delicada, porque las tasas de éxito de la población son desconocidas – y por tal motivo se tomó una muestra.

La afirmación de que las tasas de éxito de la población son todas iguales es llamada la *hipótesis nula*. La hipótesis anula asevera que no hay diferencia entre varones y mujeres en la población – las diferencias en la muestra son un puro resultado del azar. El *p*-valor es la *probabilidad de tener datos tan extremos o más extremos que los actuales*, suponiendo que la hipótesis nula es cierta:

p= Probabilidad (datos extremos | hipótesis nula del modelo)

En nuestro ejemplo, *p* = 5%. Si la hipótesis nula es cierta, sólo hay una chance del 5% de obtener una diferencia entre las tasas de éxito de 20 por ciento o más.[108] El *p*-valor de la discrepancia observada es 5%, o .05.

En tales casos, pequeños *p*-valores son evidencia de un impacto dispar, mientras que amplios *p*-valores son evidencia en contra de un impacto dispar. Aquí hay involucrados múltiplos negativos. *Un test estadístico es, en esencia, un argumento por contradicción*. La "hipótesis nula" asevera que no hay diferencias en la población – es decir, que no hay un impacto dispar. Los *p*-valores reducidos hablan en contra de la hipótesis nula – existe un impacto dispar, porque la diferencia observada es difícil de ser explicada sólo mediante el azar. *A la inversa*, *p*-valores *amplios indican que los datos son compatibles con la hipótesis nula: la diferencia observada es fácil de explicar recurriendo al azar*. En este caso, pequeños *p*-valores funcionan a favor de los demandantes, mientras que *p*-valores grandes funcionan a favor de la defensa.[109]

Es fundamental tener en cuenta que el *p*-valor está basado en el supuesto de la hipótesis de partida (hipótesis nula). Se rechaza la hipótesis nula si el *p*-valor asociado al resultado observado es igual o menor que el nivel de significación establecido, convencionalmente 0.05 o 0.01, valor que se llama *potencia del contraste*. Es decir, el *p*-valor nos muestra la probabilidad de haber obtenido el resultado que obtuvimos suponiendo que la hipótesis nula es cierta. Si el *p*-valor es inferior a la potencia del contraste nos indica que lo más probable es que la hipótesis de partida sea falsa. Sin embargo, también es posible que estemos ante una observación atípica, por lo que estaríamos cometiendo el error estadístico de rechazar la hipótesis nula cuando ésta es cierta basándonos en que hemos tenido la mala suerte de encontrar una observación atípica. Este tipo de errores se puede subsanar rebajando el *p*-valor; un *p*-valor de 0.05 es usado en investigaciones habituales sociológicas mientras que *p*-valores de 0.01 se utilizan en investigaciones médicas, en las que cometer un error puede acarrear consecuencias más graves. También se puede tratar de subsanar dicho error

---

lo llama carácter robusto. Si el modelo es falso en algunos aspectos, ¿en qué medida quedan afectados los estimadores, los resultados de los test de confianza, etc. basados en el modelo defectuoso? (Lincoln E. Moses, The Reasoning of Statistical Inference, en Perspectives on Contemporary Statistics, 1992.)

[107] En tal caso, aún puede ser útil considerar al error estándar, tal vez, como un estimador mínimo de la incertidumbre estadística de la cantidad considerada.

[108] Este aspecto será tratado en el Apéndice.

[109] Naturalmente, hay otros factores que deben ser tenidos en cuenta, como el tamaño de la muestra.

aumentando el tamaño de la muestra obtenida, lo que reduce la posibilidad de que el dato obtenido sea casualmente raro.

El *p*-valor es un valor probabilístico por lo que oscila entre 0 y 1. Así, decimos que valores altos del *p*-valor no permiten rechazar la H_0 o hipótesis nula. De igual manera, valores bajos de valor P rechazan la H_0. Es importante recalcar que un contraste de hipótesis nula no permite aceptar una hipótesis, *simplemente la rechaza o no la rechaza*, es decir que la tacha de *verosímil* (lo que no significa obligatoriamente que sea cierta, simplemente lo más probable es que sea cierta antes que falsa) o *inverosímil*, por lo que se rechaza.

Veamos otro ejemplo. Supongan que dos amigos están en un bar y uno le dice al otro que es capaz de distinguir, sin lugar a dudas, un whisky barato de uno caro. Como el otro amigo no le cree deciden hacer una prueba. El amigo bravucón dice que acierta qué tipo de whisky está tomando el 90% de las veces, ya que a veces los cubitos de hielo le distorsionan la degustación. Deciden hacerle probar 20 whiskys (en días distintos) y obtienen el resultado de que acertó sobre el contenido del vaso que estaba probando en 14 noches. Dado que nuestro amigo dijo que acertaría el 90% de las veces y sólo acertó el 70% de ellas (14 de 20 noches), ¿podemos creer a nuestro amigo, o nos está engañando? ¿Es posible que fallara por mala suerte, pero si le dejamos seguir intentándolo a la larga acertará el 90%? Está claro que si hubiera acertado todas las noches, o 19 de ellas le creeríamos sin lugar a dudas, también si hubiera fallado todas o casi todas le desmentiríamos sin dudar, pero con 14 sobre 20 es algo dudoso. Esto es lo que podemos medir con el *p*-valor.

Si suponemos que la hipótesis nula es cierta, esto quiere decir que las degustaciones de nuestro amigo se distribuyen según una binomial de parámetro 0,9, esto es, para que se entienda, como una moneda que saliera cara el 90% de las veces y cruz el 10%. ¿Cuál es la probabilidad de que una distribución binomial[110] de parámetro 0,9 repetida 20 veces nos dé como resultado 14 caras y 6 cruces? Calculando esa probabilidad nos queda *p*=0,0088. Si a este valor le sumamos la probabilidad de que acierte sólo 13 veces, más la probabilidad de que acierte sólo 12 veces y así hasta la probabilidad de que no acierte ninguna vez, es decir la probabilidad de que acierte 14 o menos veces esto nos da *p*=0,01125, y éste es el *p*-valor. ¿Qué significa esto? Significa que si suponemos que nuestro amigo acierta el 90% de las veces que prueba una copa y ha probado 20 copas, la probabilidad de que acierte 14 o menos copas es 1,125%. Por tanto, si damos una potencia de contraste usual de 0,05, que significa que aceptamos equivocarnos el 5% de las veces si repitiéramos el experimento, como el *p*-valor es inferior a la potencia del contraste rechazamos la hipótesis nula, y decimos que nuestro amigo es un fanfarrón. Estadísticamente, esto lo hacemos porque el resultado observado (14 aciertos de 20 intentos) es muy poco probable si suponemos que acierta el 90% de las veces, por lo tanto deducimos que no era cierta la hipótesis nula.

¿Qué pasaría si hubiera acertado las 20 veces? En ese caso el *p*-valor saldría muy alto, ya que es muy probable que una distribución binomial de parámetro 0,9 repetida 20 veces nos dé 20. Por tanto *no rechazaríamos la hipótesis nula, que no es lo mismo que decir que la aceptaremos*. Diríamos que es verosímil que acierte 90% de las veces, es posible que tenga razón, no tenemos evidencias en contra de ello. Es importante decir que no se acepta la hipótesis nula ya que también sería lógico aceptar que acierta el 100% de las veces y, o bien acierta el 90% o bien acierta el 100% pero ambas no pueden ser válidas a la vez.[111]

[110] Introduciremos brevemente las propiedades de la distribución binomial en otro capítulo.

[111] Jonathan A C Sterne and Davey Smith, Sifting the evidence—what's wrong with significance tests?, BMJ. 2001 January; 322(7280). La figura de la página siguiente pertenece a este artículo. http://www.ebour.com.ar/index.php?option=com_weblinks&task=view&id=18607&Itemid=0

Como el *p*-valor resulta afectado por el tamaño de la muestra, no sirve para medir la importancia de la diferencia.[112] Volviendo al ejemplo anterior, supongan que los 5,000 postulantes varones y las 5,000 postulantes mujeres difieren en sus tasas de éxito, pero solamente en un punto porcentual. La diferencia podría no ser suficiente para tener un impacto dispar, pero si se incluyen suficientes varones y mujeres en la muestra, los datos podrían terminar dando un *p*-valor muy bajo. Este *p*-valor confirmaría que los 5.000 hombres y las 5,000 mujeres tienen tasas de éxito distintas, pero no mostraría que la diferencia es sustancial.[113] En definitiva, el *p*-valor no mide la fuerza o importancia de una asociación.

Significación estadística En estadística, un resultado se denomina estadísticamente significativo cuando no es probable que haya sido debido al azar. Una "diferencia estadísticamente significativa" solamente significa que hay evidencias estadísticas de que hay una diferencia; no significa que la diferencia sea grande, importante, o significativa en el sentido estricto de la palabra.

El nivel de significación de un test es un concepto estadístico asociado a la verificación de una hipótesis. En pocas palabras, se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula cuando ésta es verdadera (decisión conocida como error de Tipo I, o "falso positivo").[114] La decisión se toma a menudo utilizando el *p*-valor: si el *p*-valor es inferior al nivel de significación, entonces la hipótesis nula es rechazada. Cuanto menor sea el *p*-valor, más significativo será el resultado. En otros términos, el nivel de significatividad

[112] Hay quienes consideran a los *p*-valores como sinónimos de disparidades “brutas” o “sustanciales”. P.ej. Craik v. Minnesota St. Univ. Bd., 731 F.2d 465, 479 (8th Cir. 1984) http://openjurist.org/731/f2d/465. Otros tribunales han puesto énfasis en la necesidad de decidir si los estadísticos de la muestra subyacente revelan una amplia disparidad. P.ej. McCleskey v. Kemp, 753 F.2d 877, 892–94 (11th Cir. 1985), aff’d, 481 U.S. 279 (1987). http://openjurist.org/753/f2d/877/mccleskey-v-kemp

[113] Ver Frazier v. Garrison Indep. Sch. Dist., 980 F.2d 1514, 1526 (5th Cir. 1993) (rechazo del intento de discriminación intencional usando el examen de competencia del profesor que fue resultado de tasas de retención superiores al 95% en todos los grupos). http://openjurist.org/980/f2d/1514/frazier-v-garrison-isd-isd-isd-isd

[114] Los “falsos positivos” han sido de conocimiento público en nuestro país, a partir del diagnóstico erróneo de la afección de la glándula tiroides de la presidenta. El falso positivo, en tal caso, resulta de una prueba que indica que una persona padece una enfermedad o afección determinada cuando, en realidad, no la padece. Nelson Castro, Los dos falsos positivos de la Presidenta, TN, Enero de 2012. http://tn.com.ar/opinion/nelson-castro/00079298/los-dos-falsos-positivos-de-la-presidenta

de un test de hipótesis es una probabilidad P tal que la probabilidad de tomar la decisión de rechazar la hipótesis nula cuando ésta es verdadera no es mayor que P. Si se halla una diferencia observada en el medio de la distribución esperada bajo la hipótesis nula, no hay sorpresas. Los datos de la muestra son del tipo que a menudo serían vistos si la hipótesis nula fuera verdadera: la diferencia no es significativa, y la hipótesis nula no puede ser rechazada. Por otro lado, si la diferencia muestral está alejada del valor esperado – de acuerdo con la hipótesis nula – la muestra es atípica, y decimos que la diferencia es "significativa", y rechazamos la hipótesis nula.

En nuestro ejemplo, los 20 puntos de diferencia porcentual de las tasas de éxito de los varones y las damas, cuyo *p*-valor era cercano a .05, puede ser considerado significativo al nivel de .05. Si el umbral fuera más reducido, por ejemplo .01, el resultado no resultaría significativo.

En la práctica, los analistas estadísticos usan a menudo ciertos niveles de significatividad pre-establecidos – típicamente .05 o .01.[115] Una referencia a resultados "altamente significativos" significa probablemente que *p* sea inferior a .01.[116]

Como el término "significativo" es meramente una etiqueta adosada a cierto tipo de *p*-valores, está sujeto a las mismas limitaciones que los propios *p*-valores. Los analistas pueden referirse a una diferencia como "significativa", indicando de esta manera que el *p*-valor se halla debajo de algún umbral. La significación depende no sólo de la magnitud del efecto, sino también del tamaño muestral (entre otras cosas). Luego, las diferencias significativas son una evidencia de que hay algo más que error aleatorio, pero no son evidencia de que este "algo" sea legal o prácticamente importante. Los estadísticos distinguen entre significación "estadística" y "práctica" para plantear el punto. Cuando se carece de significación práctica – cuando la diferencia o la correlación son despreciables – no hay motivo alguno para darle importancia a la significación estadística.[117]

Como se dijo antes, es fácil confundir al *p*-valor con la probabilidad de que no haya diferencias. Asimismo, si los resultados son significativos a nivel del .05, es tentador concluir que la hipótesis nula tiene sólo una chance de 5% de ser correcta.[118] *Deben resistirse a esta*

---

[115] Implícitamente, la Corte Suprema de US se refirió a esta práctica en Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977) http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=430&invol=482 y en Hazelwood School District v. United States, 433 U.S. 299, 311 n.17 (1977) http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=433&invol=299. En notas a pie de página, la Corte describió la hipótesis nula como "sospechosa para un científico social" cuando un estadígrafo de grandes muestras cae más lejos de "dos o tres desvíos estándar" respecto a su valor esperado bajo la hipótesis nula. Aunque la Corte no lo dijo, estas diferencias producen *p*-valores cercanos a .05 y .01 cuando el estadístico tiene una distribución normal. Los "desvíos estándar" de la Corte son nuestros "errores estándar".

[116] Hay quienes han sugerido que datos no "significativos" al .05 no sean considerados. P.ej. Paul Meier et al., What Happened in Hazelwood: Statistics, Employment Discrimination, and the 80% Rule, 1984 Am. B. Found. Res. J. 139, 152, reprinted in Statistics and the Law.

[117] P.ej., Waisome v. Port Auth., 948 F.2d 1370, 1376 (2d Cir. 1991) http://openjurist.org/999/f2d/711/waisome-v-port-authority-of-new-york-and-new-jersey-a ("si bien se halló que la disparidad era estadísticamente significativa, era de magnitud limitada"); cf. Thornburg v. Gingles, 478 U.S. 30, 53–54 (1986) http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=478&invol=30 (repite la explicación de la corte distrital de por qué "la correlación entre raza del votante y elección de ciertos candidatos por los votantes [no solamente] era estadísticamente significativa", sino además notable por serlo de manera sustancial, en el sentido de que los resultados de la elección individual hubieran sido otros si hubiera sido calculada sólo entre los votantes blancos o sólo entre los votantes negros").

[118] P.ej. Waisome, 948 F.2d at 1376 ("Los científicos sociales consideran significativo un hallazgo de dos desvíos estándar, lo que significa que existe una chance sobre 20 de que la explicación de un

tentación. Desde el punto de vista frecuencialista,[119] las hipótesis estadísticas son o bien verdaderas o bien falsas; las probabilidades son de las muestras, no de los modelos e hipótesis. *El nivel de significación indica lo que es probable que suceda si la hipótesis nula es correcta; no nos puede decir la probabilidad de que la hipótesis sea correcta. La significación no expresa la probabilidad de que la hipótesis nula sea válida más que el p-valor subyacente*.

Evaluación de Tests de Hipótesis: Potencia de un Contraste Si un *p*-valor es elevado, los resultados hallados no son significativos, y la hipótesis nula no es rechazada. Lo cual sucede al menos por dos motivos:

1. No hay diferencias dentro de la población – la hipótesis nula es verdadera; o bien
2. Hay alguna diferencia dentro de la población – la hipótesis nula es falsa – pero, por la incidencia del azar, sucedió que los datos sean del mismo tipo que los esperados bajo la hipótesis nula.

Cuando la "potencia" de un test (o contraste) es baja, puede resultar plausible la segunda explicación. *La potencia es la probabilidad de que un test estadístico declare que hay un efecto cuando existe tal efecto.*[120] Esta probabilidad depende del tamaño del efecto y del tamaño de la muestra. Discernir diferencias sutiles en la población requiere muestras grandes; aún así, las pequeñas muestras pueden detectar diferencias verdaderamente sustanciales.[121]

---

desvío pueda ser aleatoria; es decir, podría decirse con 95% de certeza que el evento no es meramente una casualidad…"); Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982); cf. Ken Feiberg, Scientific illiteracy among the judiciary http://schachtmanlaw.com/scientific-illiteracy-among-the-judiciary/; Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 272 (N.D. Tex. 1980) ("Si se utiliza un nivel de significación de 5%, un valor del estadístico t suficientemente grande indica que la chance de que el verdadero coeficiente sea en realidad cero es menor que una en 20"),vacated, 723 F.2d 1195 (5th Cir. 1984); Sheehan v. Daily Racing Form, Inc., 104 F.3d 940, 941 (7th Cir. 1997). Cf. Nathan Schachtman, Judicial Innumeracy and the MDL Process http://schachtmanlaw.com/judicial-innumeracy-and-the-mdl-process/

[119] Ver Bradley Efron, Modern Science and the Bayesian –Frequentist Controversy, Stanford, 2005. http://stat.stanford.edu/~ckirby/brad/papers/2005NEWModernScience.pdf

[120] Más precisamente, potencia es la probabilidad de rechazar la hipótesis nula cuando es correcta la hipótesis alternativa. Es típico que esta probabilidad dependa de parámetros desconocidos, así como del nivel de significación pre-establecido α. Luego, no hay un sólo número que proporcione la potencia de un contraste. Se pueden especificar valores particulares para los parámetros y el nivel de significación y computar con arreglo a los mismos la potencia del contraste. En el Apéndice veremos un ejemplo. La potencia puede ser denotada mediante la letra griega β. Aceptar la hipótesis nula cuando es verdadera la alternativa es llamada una "falsa aceptación" de la hipótesis nula o "error de Tipo II" (también: "falso negativo" o "señal errónea"). La probabilidad de un falso negativo puede ser computada a partir de la potencia, como 1 – β. La hipótesis frecuencialista mantiene el riesgo de un falso positivo a un nivel específico (digamos α=.05) y busca reducir al mínimo la probabilidad de un falso negativo (1 – β) para dicho valor de α. (Cabe aclarar que esta notación no es totalmente aceptada por los estadísticos). Hay quienes han expresado que el nivel de corte de la significación debería elegirse para igualar la chance de un falso positivo y un falso negativo, en base a que este criterio corresponde a la carga de la prueba en términos "más que probables". Pero el argumento es falaz, porque α y β no proporcionan las probabilidades de las hipótesis nula y alternativa. Ver D.H. Kaye, Hypothesis Testing in the Courtroom, in Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon 331, 341–43 (Alan E. Gelfand ed., 1987).

[121] Para simplificar, los ejemplos numéricos de inferencia estadística de este capítulo suponen que trabajamos con muestras grandes. Algunos tribunales de US expresaron su descontento con los estimadores o análisis basados en pequeñas muestras; de hecho, algunos llegaron a rechazar considerar tales estudios o procedimientos estadísticos formales de manejar pequeñas muestras. Ver p.ej. Bunch v. Bullard, 795 F.2d 384, 395 n.12 (5th Cir. 1986) http://ftp.resource.org/courts.gov/c/F2/795/795.F2d.384.84-4793.85-4423.html (12 sobre 15 blancos y sólo 3 sobre 13 negros que pasaron un test de promoción policial crearon *prima facie* un caso de

Cuando un estudio de bajo contraste no logra exhibir un efecto significativo, es más apropiado decir que los resultados no son concluyentes en lugar de negativos: la prueba es débil porque la potencia es baja.[122] Por otra parte, cuando los estudios tienen buenas chances de detectar una asociación significativa, no obtener significatividad puede constituir evidencia persuasiva de que no hay efecto alguno.[123]

Contrastes En muchos casos el test estadístico puede ser hecho a una (unilateral) o a dos colas (bilateral). El segundo método da lugar a un *p*-valor que es el doble del primer método. Como los *p*-valores reducidos son evidencia en contra de la hipótesis nula, un test a una cola parece producir evidencia más fuerte que otro a dos colas. Sin embargo, esta distinción es, en buena medida, ilusoria.[124]

---

impacto dispar; sin embargo, "el tribunal del distrito no aplicó, ni tampoco lo hacemos nosotros, las teorías de la probabilidad a un tamaño muestral tan reducido como éste" porque "los análisis estadísticos avanzados pueden ser de escasa ayuda para determinar cuán significativas son esas disparidades"); United States v. Lansdowne Swim Club, 713 F. Supp. 785, 809–10 (E.D. Pa. 1989). Ver también Jennifer L. Peresie, Toward a Coherent Test for Disparate Impact Discrimination, Indiana Law Journal, 84 (2009) http://www.repository.law.indiana.edu/ilj/vol84/iss3/1. Otros tribunales han sido más aventurados. P.ej. Bazemore v. Friday, 751 F.2d 662, 673 & n.9 (4th Cir. 1984) (la corte de apelaciones aplicó sus propios test-t en lugar de la curva normal al ordenamiento de cuartiles a fin de tomar en cuenta el tamaño muestral de nueve), 478 U.S. 385 (1986). http://supreme.justia.com/cases/federal/us/478/385/case.html

[122] En nuestro ejemplo, si α = .05, la potencia para detectar una diferencia de 10 puntos porcentuales entre los postulantes varones y mujeres es de sólo 1/6 (Ver Apéndice). Si no se observa en tal caso una diferencia "significativa" sólo se suministra una prueba débil de que la diferencia entre hombres y mujeres es menor que 10 puntos porcentuales. Preferimos los estimadores acompañados por los errores estándar de los test porque los primeros parecen dejar más en claro el estado de la evidencia estadística: La diferencia estimada es 20 ± 10 puntos porcentuales, lo que indica que una diferencia de 10 puntos en por ciento resulta compatible con los datos.

[123] Hay procedimientos formales para agregar resultados de distintos estudios. Ver In re Paoli R.R. Yard PCB Litig., 916 F.2d 829 (3d Cir. 1990) http://openjurist.org/916/f2d/829/paoli-railroad-yard-pcb-litigation-brown-v-monsanto-company. En principio, la potencia de los resultados colectivos será mayor que la potencia de cada estudio por separado. Ver The Handbook of Research Synthesis pp. 226–27 (Harris Cooper & Larry V. Hedges eds., 1993); Larry V. Hedges & Ingram Olkin, Statistical Methods for MetaAnalysis (1985); Jerome P. Kassirer, Clinical Trials and Meta-Analysis: What Do They Do for Us?, 327 New Eng. J. Med. pp. 273, 274 (1992) ("El meta-análisis acumulativo representa un enfoque prometedor"); National Reseach Council, Combining Information: Statistical Issues and Opportunities for Research (1992); Symposium, Meta-Analysis of Observational Studies, 140 Am. J. Epidemiology 771 (1994). Lamentablemente, estos procedimientos tienen sus propias limitaciones. Ver Diana B. Petitti, Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine (2d ed. 2000); Michael Oakes, Statistical Inference: A Commentary for the Social and Behavioural Sciences 157 (1986).

[124] En el ejemplo de los éxitos en el examen, el *p*-valor del test viene aproximado por un área debajo de la curva normal. El contraste unilateral utiliza el "área de la cola" por debajo de la curva a la derecha de 2, lo que produce un *p*-valor = .025 (aprox.). El contraste bilateral contempla el área a la izquierda de -2, así como el área a la derecha de 2. Ahora tenemos dos colas, y *p*=.05. Ver Freedman et al., pp. 549-52. Según la teoría de la estadística formal, elegir entre ambos contrastes puede hacerse viendo cuál es la forma exacta de la "hipótesis alternativa". En el ejemplo, la hipótesis nula es que las tasas de éxito de los varones son iguales a las de las mujeres en toda la población de postulantes. La hipótesis alternativa puede excluir *a priori* la posibilidad de que las mujeres tengan una tasa de éxitos más elevada y sostener que habrá más muchachos que mujeres pasando bien el examen. Esta alternativa asimétrica sugiere realizar un contraste unilateral. Por otro lado, la hipótesis alternativa puede estipular simplemente que las tasas de éxitos de ambos grupos son desiguales. Esta alternativa asimétrica admite la posibilidad de que las damas tengan un mayor puntaje que los muchachos, y permite un test a dos colas. Ver Freeman et al., p.551. Hay expertos que piensan que la elección entre contrastes a una y dos colas a menudo puede hacerse considerando la forma exacta que tienen las hipótesis nula y alternativa.

Hay tribunales que expresan preferencia por los test bilaterales,[125] pero no se requiere una regla rígida si los *p*-valores y los niveles de significación se usan como pistas más que como reglas mecánicas de pruebas estadísticas. Los contrastes unilaterales hacen más fácil lograr un umbral como .05, pero tengan en cuenta que si no usan este valor como una línea divisoria mágica, en tal caso la elección entre contrastes unilaterales y bilaterales no será tan importante – siempre que la elección y su efecto sobre el *p*-valor se hagan explícitos.[126]

Cantidad de contrastes Realizar contrastes repetidos complica la interpretación de los niveles de significación. Con suficientes comparaciones, el error aleatorio garantiza que en alguna oportunidad se tendrá un hallazgo "significativo", aunque no exista. Consideremos el problema de decidir si una moneda está sesgada. La probabilidad de que una moneda razonable produzca 10 caras al ser arrojada 10 veces es $(½)^{10}$= 1/1,024= 0,000976563. Si uno observa 10 caras al arrojarla las primeras 10 veces, luego, habría evidencia fuerte de que la moneda está sesgada. Empero, si una moneda "razonable" es arrojada unas pocas miles de veces, siempre es probable que aparezca al menos una serie de 10 caras consecutivas. El test – consistente en buscar una corrida seguida de 10 caras – puede ser repetido muchas veces.

Estos experimentos son moneda corriente. Como las investigaciones que no llegan a producir resultados no se publican, las revisiones de la literatura pueden llegar a producir una cantidad enorme de estudios que encuentran evidencia estadística.[127] Todo investigador suele buscar tantas relaciones diferentes que algunas surgirán con significación estadística por mera casualidad. *Casi todos los conjuntos de datos – aún páginas enteras de tablas de números al azar – llegan a contener algún patrón inusual que puede ser descubierto mediante una investigación diligente*. Una vez detectado el patrón, el analista puede realizar un contraste estadístico, ignorando sin gracia el esfuerzo de investigación. A lo cual seguirá la significatividad estadística. Diez caras obtenidas al arrojar las primeras diez veces una moneda significa una cosa; diez caras seguidas ubicadas en algún lugar de una cadena de miles de veces que la moneda ha sido arrojada significa algo bastante distinto.

Hay métodos estadísticos para tratar con visiones múltiples de los datos, que permiten el cálculo de *p*-valores significativos en ciertos casos.[128] Sin embargo, no existe una solución general disponible, y los métodos existentes serían de poca ayuda en el caso típico en que los analistas han contrastado y rechazado una variedad de modelos de regresión antes de llegar al que consideran como más satisfactorio. En tales casos, los tribunales no deberán sentirse impresionados por afirmaciones de que los estimadores son significativos. En su lugar, deberían preguntar a los analistas cómo desarrollaron sus modelos.[129]

---

[125] David C. Baldus & James W.L. Cole, Statistical Proof of Discrimination § 5.1, p. 153 (1980 & Supp. 1987); The Evolving Role of Statistical Assessments as Evidence in the Courts, pp. 38–40 (donde se cita a EEOC v. Federal Reserve Bank, 698 F.2d 633 (4th Cir. 1983), rev'd on other grounds *sub nom*. Cooper v. Federal Reserve Bank, 467 U.S. 867 (1984)); David H. Kaye, The Numbers Game: Statistical Inference in Discrimination Cases, 80 Mich. L. Rev. 833 (1982) (cita a Hazelwood Sch. Dist. v. United States, 433 U.S. 299 (1977)). Argumentos para realizar contrastes unilaterales fueron discutidos por Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers (1990), pp. 125-26; Richard Goldstein, Two Types of Statistical Errors in Employment Discrimination Cases, 26 Jurimetrics J. 32 (1985).

[126] Los test unilaterales al .05 son considerados como evidencia débil – es usual que no sean utilizados estándares más débiles en la literatura técnica.

[127] Ver Stuart J. Pocock et al., Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals, 317 New Eng. J. Med. 426 (1987).

[128] Por ejemplo, ver Rupert G. Miller, Jr., Simultaneous Statistical Inference (2d ed. 1981).

[129] Ver p.ej. On Model Uncertainty and Its Statistical Implications: Lecture Notes in Econometric and Mathematical Systems (Theo K. Dijkstra ed., 1988); Frank T. Denton, Data Mining As an Industry, 67 Rev. Econ. & Stat. 124 (1985). http://www.andrew.cmu.edu/course/88-301/introduction/data_mining.pdf La intuición puede sugerir que cuantas más variables sean incluidas en el modelo, tanto mejor. Sin embargo, esta idea parece estar equivocada. Los modelos complejos

Estimación por Intervalos La significación estadística depende del *p*-valor, y éstos dependen del tamaño de la muestra. Luego, un efecto "significativo" bien podría ser pequeño. A la recíproca, un efecto "no significativo" podría ser amplio. Al preguntarse sobre la magnitud de un efecto, las cortes pueden evitar ser confundidas con los *p*-valores. Para concentrar la atención donde se requiere – sobre el tamaño real de un efecto y la fiabilidad del análisis – la estimación por intervalos puede ser valiosa.[130] Apreciar un rango de valores plausibles de la variable de interés ayuda a describir la incertidumbre estadística del estimador.

En nuestro ejemplo, el intervalo de confianza al 95% de la diferencia entre las tasas de éxito de los varones y las damas iba desde 0 a 40 puntos porcentuales. Nuestro mejor estimador de la tasa de éxito de los varones es 20 puntos porcentuales superior al de las damas; y la diferencia podría llegar a ser plausiblemente tan escasa como 0 o tan abultada como 40 puntos. El *p*-valor no proporciona esta información. El intervalo de confianza contiene más información que la de un test de significatividad.[131] En el ejemplo, cero está en el extremo inferior del intervalo de confianza al 95%, luego hay evidencia "significativa" de que la verdadera diferencia en los éxitos de los exámenes de los postulantes varones y femeninos no es cero. Pero hay valores muy próximos a cero *dentro* del intervalo.

Por otro lado, supongan que un test de significación no rechaza la hipótesis nula. El intervalo de confianza puede impedir que se cometa el error de pensar que hay evidencia positiva para la hipótesis nula. P. ej., cámbiese levemente el ejemplo: digamos que 29 hombres y 20 mujeres pasaron el test. El intervalo de confianza va desde -2 a 38 puntos porcentuales. Como una diferencia de cero cae dentro del intervalo de confianza al 95%, la hipótesis nula – de que la verdadera diferencia es cero – no puede ser rechazada a un nivel del .05. Pero el intervalo se extiende 38 puntos porcentuales, lo que indica que la diferencia poblacional podría ser sustancial. La carencia de significatividad no excluye esta posibilidad.[132]

Hipótesis Rivales El *p*-valor de un test estadístico se computa basándose en un modelo de los datos – la hipótesis nula. Es usual practicarlo para sostener la hipótesis alternativa – otro modelo. Pero si se lo ve más de cerca, ambos modelos puedan no resultar razonables.[133] Un *p*-valor reducido indica que algo está pasando, además del error aleatorio; la hipótesis alternativa podría ser considerada como una explicación posible – entre varias – de los datos.[134]

---

puede que reflejen sólo aspectos accidentales de los datos. Los test estadísticos usuales ofrecen poca protección contra esta posibilidad cuando el analista estuvo probando una variedad de modelos antes de llegar a la especificación final.

[130] Un estimador por intervalo puede estar compuesto por un estimador puntual – tal como la media muestral usada para medir la población muestral – en forma conjunta con su error estándar; o bien, el estimador puntual y el error estándar pueden combinarse para formar un intervalo de confianza.

[131] Por tal motivo, se ha sostenido que los tribunales deben solicitar intervalos de confianza (cuando pueden ser computados) sin los tests de significatividad y los *p*-valores explícitamente.

[132] Se han usado intervalos bilaterales, que corresponden a test a dos colas. También pueden usarse intervalos unilaterales (test a una sola cola) que también están disponibles.

[133] A menudo las hipótesis nula y alternativa son enunciados sobre rangos posibles de valores de los parámetros de un modelo estadístico común. El cómputo de los errores estándar, los *p*-valores, y la potencia tiene lugar dentro de los confines de este modelo básico. El análisis estadístico se fija en la plausibilidad relativa de valores competitivos de los parámetros, pero no hace una evaluación global de cuán razonable es el modelo básico.

[134] Paul Meier & Sandy Zabell, Benjamin Peirce and the Howland Will, Journal of the American Statistical Association, Vol. 75, No. 371. (Sep., 1980), pp. 497-506. http://ben-israel.rutgers.edu/711/Meier-Zabell.pdf (explicaciones competitivas en un caso de falsificación). Fuera de la esfera legal, hay muchos ejemplos intrigantes de la tendencia a pensar que pequeños *p*-valores son una demostración definitiva de una hipótesis alternativa, aunque haya otras explicaciones plausibles de los datos. Ver p.ej. Freeman et al., pp. 562-63; C.E.M. Hansel, ESP: A Scientific Evaluation (1966).

En Mapes Casino, Inc. v. Maryland Casualty Co.,[135] por ejemplo, el tribunal reconoció la importancia de explicaciones dejadas de lado por quien había propuesto la evidencia estadística. En esta acción de cobro de una póliza de seguro, Mapes Casino buscaba cuantificar el monto de sus pérdidas por malversación de fondos de un empleado. El casino sostuvo que algunos empleados usaban un intermediario para hacerse de fondos en fichas de otros casinos. Estableció que a lo largo de un período de 18 meses, el porcentaje de ganancia en sus mesas de dados fue de 6%, en comparación con un valor esperado de 20%. La corte reconoció que las estadísticas mostraban el hecho de que algo andaba mal en las mesas de dados – la discrepancia era demasiado importante como para ser resultado del azar. Pero no se dejó convencer por la hipótesis alternativa del demandante. El tribunal apuntó a otras explicaciones posibles (actividades como "timar" o "sacar el jugo") que podrían haber dado cuenta de la discrepancia sin implicar a los empleados sospechosos.[136] En resumen, el rechazo de la hipótesis nula no coloca a la hipótesis alternativa como la única explicación viable de los datos.[137]

Probabilidades Posteriores Los errores estándar, los *p*-valores, y los test de significación son técnicas comunes para evaluar un error aleatorio. Estos procedimientos descansan en datos muestrales, y se justifican en términos de las "características operativas" de los procedimientos estadísticos.[138] Sin embargo, el enfoque frecuencialista no permite al estadístico computar la probabilidad de que una hipótesis en particular sea correcta, dados los datos.[139] Por ejemplo, un frecuencialista puede postular que una moneda es insesgada: tiene una probabilidad 50-50 de caer cara, y las tiradas sucesivas son independientes; esto se considera un enunciado empírico – potencialmente falsable – sobre la moneda. Sobre esta base, resulta sencillo calcular la probabilidad de que la moneda salga cara en las próximas diez tiradas,[140] la respuesta es 0,000976563. Por lo tanto, observar diez caras de seguido pondría en serios aprietos la hipótesis de que no hay sesgo. Rechazar la hipótesis de una moneda insesgada cuando han salido diez caras en diez tiradas sucesivas da un resultado erróneo - cuando la moneda es insesgada – solamente 1 vez en 1,024 veces. Éste es un ejemplo de lo que sería una característica operativa de un procedimiento estadístico.

---

[135] Véase Palmer Morrel-Samuels and Peter D. Jacobson, Using Statistical Evidence to Prove Causality to Non-Statisticians, July, 2007, SSRN. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=995841

[136] En otras palabras, la corte parece haber pensado que era el propio casino que se estafaba a sí mismo, o que pudo haber otros estafadores además de los empleados particulares identificados en el caso. Al menos, la evidencia estadística del demandante no excluía tales posibilidades.

[137] Comparar con EEOC v. Sears, Roebuck & Co., 839 F.2d 302, 312 & n.9, 313 (7th Cir. 1988) (los estudios de regresión de EEOC que indicaban diferencias significativas no establecían responsabilidad porque la encuesta y los testimonios sostenían la hipótesis rival de que las mujeres estaban menos interesadas en los puestos de ventas a comisión), con EEOC v. General Tel. Co., 885 F.2d 575 (9th Cir. 1989) (la hipótesis rival no sustanciada de "falta de interés" en tareas "no tradicionales" era insuficiente para rebatir *prima facie* un caso de discriminación de géneros). También es útil consultar el artículo de Mark S. Brodin, Behavioral Science Evidence in the Age of Daubert: Reflections of a Skeptic, 2004, Boston College Law School Faculty Papers. http://lawdigitalcommons.bc.edu/lsfp/24

[138] Las "características operativas" son el valor esperado, el error estándar de los estimadores, las probabilidades de error de los test estadísticos, y cantidades vinculadas.

[139] Ver *infra* Apéndice. Por lo tanto, cantidades como los *p*-valores o los niveles de confianza no pueden ser comparados directamente a números como .95 o .50 que uno piensa podrían cuantificar la carga de convicción en casos criminales o civiles. D.H. Kaye, Apples and Oranges: Confidence Coefficients and the Burden of Persuasion, 73 Cornell L. Rev. 54 (1987).

[140] Expresado de modo algo más formal, si la moneda es insesgada y cada resultado es independiente (hipótesis), entonces la probabilidad de observar diez caras (datos) es Pr (datos$|H_0$) = $(½)^{10}$ = 1/1,024 = 0,000976563, donde H_0 representa la hipótesis de que la moneda es insesgada.

¿Qué puede decirse de la probabilidad recíproca (si una moneda cae cara diez veces de seguido, cuál es la probabilidad de que sea insesgada)?[141] Para computarlas, es necesario que las probabilidades iniciales de la moneda sean insesgadas, así como probabilidades de ausencia de sesgo en diversos grados. Todo ello está más allá del alcance de la estadística frecuencialista.[142]

En el enfoque Bayesiano o subjetivista, las probabilidades representan grados de creencia subjetiva más que hechos objetivos. La confianza del observador en la hipótesis de que la moneda está insesgada, por ejemplo, se expresa como un número entre cero y uno (donde "confianza" tiene el significado habitual que se le otorga, no una interpretación técnica aplicable a un "intervalo de confianza" frecuencialista. Por consiguiente, puede relacionarse con la carga de la convicción. El observador debe cuantificar sus creencias cuantitativas de las chances de que la moneda esté sesgada en diversos grados – todo antes de ver los datos.[143] Estas probabilidades subjetivas, como todas las probabilidades que gobiernan el movimiento de la moneda, están ahí obedeciendo los axiomas de la teoría de la probabilidad. Las probabilidades de las distintas hipótesis sobre la moneda, especificadas antes de recoger los datos, son llamadas probabilidades a priori. En este caso, las

[141] Kaye y Freedman explican que ésta es llamada *probabilidad recíproca* porque se escribe de forma Pr (H_0|datos) en lugar de Pr (datos|H_0); a veces se usa una frase equivalente, "probabilidad inversa". Hay una tendencia a pensar en Pr (datos|H_0) como si fuera la probabilidad inversa Pr (H_0|datos) conocida como *falacia de transposición*. Por ejemplo, la mayoría de los senadores de US son hombres, pero pocos hombres son senadores. Luego, existe una elevada probabilidad de que un individuo senador sea un hombre, pero la probabilidad de que un individuo hombre también sea senador es prácticamente cero. El *p*-valor frecuencialista, Pr (datos | H_0) no es en general una buena aproximación a la probabilidad bayesiana Pr (H_0| datos); la última también incluye consideraciones de potencia y de números base. Más adelante nos referiremos a aspectos de la estadística bayesiana.

[142] A veces la probabilidad de un evento del que depende un caso puede ser computada con métodos objetivos. Empero, estos eventos deben ser resultados medibles (como la cantidad de caras en una serie de tiradas de una moneda) más que hipótesis sobre el proceso que generó los datos (como la hipótesis de que la moneda sea insesgada). P.ej. en United States v. Shonubi, 895 F. Supp. 460 (E.D.N.Y. 1995), rev'd, 103 F.3d 1085 (2d Cir. 1997), un experto del gobierno estimó para una sentencia la cantidad total de heroína que un demandado nigeriano que vivía en Nueva York había traído de contrabando (tragándose globos llenos de heroína) durante ocho viajes desde y hacia Nigeria. Aplicó un método conocido como resampling o bootstrapping (http://en.wikipedia.org/wiki/Bootstrapping(statistics)). Obtuvo 100,000 muestras simples independientes de tamaño siete de una población de cargas distribuidas como datos aduaneros sobre otros 117 tragadores de globos apresados en el mismo aeropuerto durante el mismo período; descubrió que para un 99% de estas muestras, su peso total era al menos de 2090.2 gramos. 895 F. Supp. at 504. El investigador terminó expresando que existe un 99% de probabilidad de que Shonubi trajo consigo 2090.2 gramos de heroína en los siete viajes previos..." Id. Empero, el Segundo Circuito revirtió este hallazgo requiriendo "evidencia específica sobre lo que había hecho Shonubi". 103 F.3d at 1090. Aunque no resulta clara la base lógica de esta "evidencia específica", hay una dificultad con el análisis del experto. La inferencia estadística en general implica una extrapolación desde las unidades de la muestra a la población de todas las unidades. Por consiguiente, la muestra debe ser representativa. En Shonubi, el gobierno usó una muestra de cargas, una por correo en cada viaje en el que el correo fue atrapado. Buscó extrapolar desde estos datos a varios viajes hechos por un solo correo – viajes en los que el otro correo no fue atrapado. Ver Mark Colyvan and Helen M. Regan, Legal Decisions and the Reference-Class Problem http://homepage.mac.com/mcolyvan/papers/legaldec.pdf

[143] Por ejemplo, sea *p* la probabilidad desconocida de que la moneda aterrice cara: ¿Cuál es la probabilidad de que $p \geq .6$? El estadístico bayesiano debe estar preparado a responder a preguntas de este tipo. A los procedimientos bayesianos se los defiende a veces sobre la base de que las creencias de un observador racional deben conformarse con las reglas bayesianas. Sin embargo, cabe notar que la definición de "racional" es puramente formal. Ver Peter C. Fishburn, The Axioms of Subjective Probability, 1 Stat. Sci. 335 (1986) http://www.ie.boun.edu.tr/~taner/ie544/papers/Fishburn_1986.pdf; David Kaye, The Laws of Probability and the Law of the Land, 47 U. Chi. L. Rev. 34 (1979).

probabilidades a priori pueden ser actualizadas utilizando la “regla de Bayes”, una vez que se tienen datos sobre cómo cayó la moneda. Esta regla muestra cómo una probabilidad condicional (p.ej. la probabilidad de una hipótesis dada la evidencia observada) depende de su inversa (la probabilidad de que se produzca esa evidencia dada la hipótesis).

La idea clave es que la probabilidad de un evento A (p.ej. tener cáncer de mamas) dado el evento B (tener un mamograma positivo) depende no sólo de la relación entre A y B (es decir, de la precisión de los mamogramas) sino además de la probabilidad absoluta de A independiente de B (es decir, la incidencia del cáncer en general), y de la probabilidad absoluta de B independiente de A (es decir, la posibilidad de tener un mamograma positivo). Por ejemplo, si se sabe que las mamografías tienen una precisión del 95%, ello puede deberse a un 5% de falsos positivos, a un 5% de falsos negativos (fallas), o a una mezcla aleatoria de falsos positivos y falsos negativos. La regla de Bayes nos permite calcular en forma precisa la probabilidad de tener cáncer de mamas dada una mamografía positiva en cualquiera de los tres casos, porque la probabilidad de B (un mamograma positivo) sería distinta en cada caso. Nótese que, si el 5% de los mamogramas da resultados positivo, luego la probabilidad de que un individuo con resultado positivo tenga cáncer es bastante reducida, ya que la probabilidad de cáncer está próxima a 1%. La probabilidad de un resultado positivo entonces es 5 veces superior a la probabilidad del mismo cáncer. Esto demuestra el valor de entender y aplicar en forma correcta el teorema de Bayes.

Más técnicamente, el teorema expresa la probabilidad posterior (es decir, luego de que fue observada la evidencia E) de una hipótesis H en términos de las probabilidades a priori de H y E, y de la probabilidad de E dada H. Implica que la evidencia posee un fuerte *efecto confirmatorio* si era más implausible antes de que fuera observada.[144] El teorema de Bayes es válido en todas las interpretaciones corrientes de la probabilidad, y es aplicable tanto en ciencia y en ingeniería.[145] Pero hay desacuerdos entre estadísticos frecuencialistas y subjetivistas bayesianos con respecto a la implementación apropiada y a qué validez tiene el teorema de Bayes.

Resumiendo, los estadísticos bayesianos pueden computar probabilidades posteriores de distintas hipótesis sobre la moneda, con los datos.[146] Si bien estas probabilidades posteriores pueden responder directamente a hipótesis de interés legal, son necesariamente subjetivas, porque no sólo reflejan los datos sino además hipótesis subjetivas sobre la moneda antes de tenerlos.[147]

---

[144] Howson, Colin; Peter Urbach (1993). Scientific Reasoning: The Bayesian Approach. Open Court.

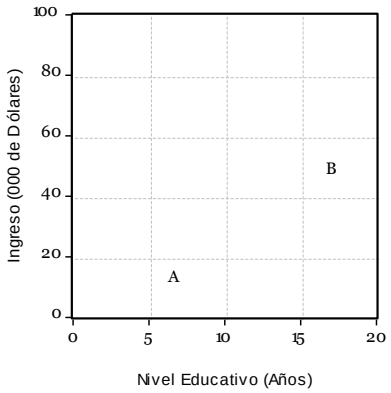
[145] Jaynes, Edwin T. (2003). Probability theory: the logic of science. Cambridge University Press.

[146] Ver en general George E.P. Box & George C. Tiao, Bayesian Inference in Statistical Analysis (Wiley Classics Library ed., John Wiley & Sons, Inc. 1992) (1973). En cuestiones de aplicaciones legales, ver, p.ej., Aitken et al., obra citada, pp. 337–48; David H. Kaye, DNA Evidence: Probability, Population Genetics, and the Courts, 7 Harv. J.L. & Tech. 101 (1993) http://jolt.law.harvard.edu/articles/pdf/v07/07HarvJLTech101.pdf

[147] Dentro de este contexto, surge una pregunta: usaremos creencias, pero ¿de quién? ¿Del estadístico o del investigador oficial? Ver p.ej., Michael O. Finkelstein & William B. Fairley, A Bayesian Approach to Identification Evidence, 83 Harv. L. Rev. 489 (1970). Estos autores proponen que los expertos proporcionen probabilidades posteriores para una amplia gama de probabilidades a priori, a fin de permitir que los jurados utilicen sus propias probabilidades a priori o que sólo juzquen el impacto de los datos sobre los valores posibles de las probabilidades a priori. Pero Laurence H. Tribe (Trial by Mathematics: Precision and Ritual in the Legal Process, 84 Harv. L. Rev. 1329 (1971)), sostiene que los esfuerzos para describir el impacto de la evidencia sobre las probabilidades subjetivas de los jurados podrían impresionar de forma indebida a los jurados y menoscabar la presunción de inocencia y otros valores legales. Ver también Timothy Huang and Stuart Russell, Object Identification in a Bayesian Context, Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97), 1997. http://www.cs.middlebury.edu/~huang/publications/ijcai1997.pdf

Este tipo de análisis ha sido utilizado pocas veces en los tribunales[148] y la cuestión sobre su valor forense ha sido aireada primariamente dentro de la literatura académica.[149] Hay estadísticos que están a favor de los métodos bayesianos[150] y también comentaristas legales que han propuesto usarlos en ciertos casos bajo determinadas circunstancias.[151]

Figura 3. Diagrama de dispersión. El eje horizontal indica el nivel de educación y el eje vertical indica el ingreso anual.



5. Correlación y Regresión

Los modelos de regresión son usados frecuentemente para inferir causalidad a partir de la asociación; por ejemplo, a menudo son usados para demostrar el tratamiento dispar en casos de discriminación, o para estimar los daños emergentes de acciones anti-monopolísticas. Vamos a explicar las ideas básicas y algunos escollos. Al principio se incluye material preliminar, vinculado con los diagramas de dispersión, los coeficientes de correlación y las líneas de regresión a fin de resumir relaciones entre variables. Posteriormente estos temas serán más desarrollados en un capítulo especial.

Diagramas de dispersión Las relaciones entre dos variables pueden ser graficadas en un diagrama de dispersión. Un ejemplo son los datos sobre ingreso y educación de una muestra de 350 personas, de edades comprendidas entre los 25 y los 29 años, que residen en Buenos Aires. Cada persona de la muestra corresponde a un punto del diagrama. Como indica la Figura 3, el eje horizontal representa el nivel educativo de una persona, y el eje

[148] Hay una excepción: los litigios sobre asuntos de paternidad. Cuando las pruebas genéticas indican paternidad, es común testimoniar con relación a una "probabilidad posterior de paternidad". Ver, p.ej., David L. Faigman, David H. Kaye, Michael J. Saks, and Joseph Sanders, Modern Scientific Evidence: The Law and Science of Expert Testimony, 2009-2010 ed., Sección 19-2.5.

[149] Ver, p.ej., Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism (Peter Tillers & Eric D. Green eds., 1988); Symposium, Decision and Inference in Litigation, 13 Cardozo L. Rev. 253 (1991). Probablemente el contexto bayesiano haya sido más aceptado al explicar conceptos legales como la relevancia de la evidencia, la naturaleza de le evidencia perjudicial, el valor probatorio, y la carga de la convicción. Ver p.ej. Richard D. Friedman, Assessing Evidence, 94 Mich. L. Rev. 1810 (1996) (revisión del libro); Richard O. Lempert, Modeling Relevance, 75 Mich. L. Rev. 1021 (1977); Íd., The Significance of Statistical Significance: Two Authors Restate an Incontrovertible Caution - Why a Book?, SSRN (May 2008) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1142865; Íd., Low Probability/High Consequence Events: Dilemmas of Damage Compensation, SSRN (April, 2009) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1371784; D.H. Kaye, Clarifying the Burden of Persuasion: What Bayesian Decision Rules Do and Do Not Do, 3 Int'l J. Evidence & Proof 1 (1999); Kevin M. Clermont, Standards of Proof Revisited, (2009). Cornell Law Faculty Publications. Paper 13. http://scholarship.law.cornell.edu/facpub/13

[150] Donald A. Berry, Inferences Using DNA Profiling in Forensic Identification and Paternity Cases, 6 Stat. Sci. 175, 180 (1991); Stephen E. Fienberg & Mark J. Schervish, The Relevance of Bayesian Inference for the Presentation of Statistical Evidence and for Legal Decisionmaking, 66 B.U. L. Rev. 771 (1986). Sin embargo, muchos estadísticos cuestionan la aplicabilidad general de las técnicas bayesianas: los resultados de los análisis pueden estar influidos sustancialmente por las probabilidades a priori, que son en definitiva bastante arbitrarias. Ver David Freedman, Some Issues in the Foundation of Statistics, 1 Found. Sci. 19 (1995), reimpreso en Topics in the Foundation of Statistics 19 (Bas C. van Fraasen ed., 1997).

[151] Por ejemplo, Joseph C. Bright, Jr. et al., Statistical Sampling in Tax Audits, 13 L. & Soc. Inquiry 305 (1988); Ira Mark Ellman & David Kaye, Probabilities and Proof: Can HLA and Blood Group Testing Prove Paternity?, 54 N.Y.U. L. Rev. 1131 (1979); Finkelstein & Fairley, supra note 174; Kaye, supra note 173.

vertical su ingreso anual. La persona A completó 8 años de escolaridad y alcanzó un ingreso de $19,000. La persona B completó 16 años de escolaridad y llegó a un ingreso anual de $38,000.

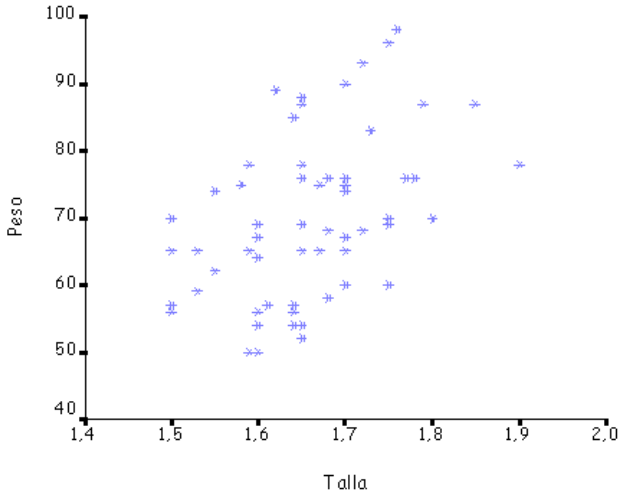


Figura 4

Ya en el Capítulo XXII usamos otro diagrama de dispersión que mostraba el tiempo de espera entre las erupciones y la duración de la erupción del géiser Old Faithful en el Parque Nacional Yellowstone, Wyoming, US. Naturalmente, a medida que aumenta la cantidad de observaciones, los puntos observados se hacen más abigarrados, como en la Figura 4, que indica que a medida que aumenta el peso (en kg.) de una persona, el tamaño de la toalla de baño requerida tiende a ser mayor en (m^2).

*Coeficientes de Correlación D*os variables están correlacionadas positivamente cuando sus valores tienden a aumentar o a disminuir en forma conjunta.[152] El ingreso anual y el nivel educativo de la Figura 4 facilitan un ejemplo con estas características. El coeficiente de correlación (denotado usualmente mediante la letra r) es un solo número que refleja la fuerza de una asociación.

Un coeficiente r=0 indica que no existe asociación lineal entre las variables, mientras que r=+1 indica una relación lineal perfecta: todos los puntos del diagrama de dispersión caen sobre una línea recta orientada en sentido ascendente. Éste es el máximo valor que puede adoptar r. A veces existe una relación negativa entre variables: aumentos de una de ellas tienden a estar acompañados de descensos de la otra. Un ejemplo es la antigüedad de un automóvil y la economía en combustible en miles de litros. Una asociación negativa se indica mediante valores negativos de r. El caso extremo es r=-1, que indica que todos los puntos del diagrama de dispersión están ubicados sobre una recta con pendiente negativa.

Las asociaciones moderadas son la regla general en ciencias sociales; correlaciones superiores a, p. ej., 0.7 son bastante atípicas en muchas áreas. Por ejemplo, la correlación entre grado universitario y primer año de resultados en las facultades de derecho de US está por debajo de 0.3 en la mayoría de las facultades de derecho, mientras que la correlación entre los resultados del test LSAT se sitúa en general en 0.41.[153] La correlación entre la altura de hermanos mellizos es alrededor de 0.5, mientras que la correlación entre la altura de gemelos idénticos está en torno de 0.95. Pero el coeficiente de correlación no puede captar toda la información subyacente. Hay varias cuestiones que pueden presentarse, que consideraremos a continuación.

Asociación Lineal El coeficiente de correlación está pensado para medir una asociación lineal. La Figura 6 muestra un patrón fuertemente no lineal con un coeficiente de correlación

[152] Muchos estadígrafos y gráficos están disponibles para investigar la asociación de variables. Los más comunes son el coeficiente de correlación y el diagrama de dispersión.

[153] Linda F. Wightman, Predictive Validity of the LSAT: A National Summary of the 1990–1992 Correlation Studies 10 (1993); Linda F. Wightman & David G. Muller, An Analysis of Differential Validity and Differential Prediction for Black, Mexican-American, Hispanic, and White Law School Students 11–13 (1990). Al combinar LSAT con la media puntual del estudiante no graduado se obtiene una mayor correlación con los resultados del primer año de las facultades de derecho que tomándolos por separado. Típicamente, el coeficiente de correlación múltiple está en torno de 0.5.

próximo a cero. Si el diagrama de dispersión revela un patrón no lineal muy marcado, el coeficiente de correlación puede no resultar un estadístico sumario útil.

Como hemos visto también en el Capítulo XXII, también es utilizado otro coeficiente de correlación (denominado de Spearman), que en general mide el carácter monótono de una relación (a diferencia del presente coeficiente, denominado coeficiente de Pearson).

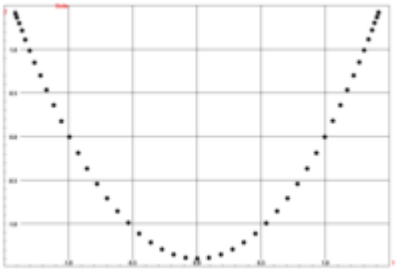


Figura 6. El coeficiente de correlación sólo mide la asociación lineal. El diagrama de dispersión exhibe una fuerte asociación no lineal con un coeficiente de correlación prácticamente nulo.

Valores Atípicos y Coeficiente de Correlación El coeficiente de correlación puede estar distorsionado por valores atípicos – unos pocos puntos alejados de la mayoría de los datos. El panel de la izquierda de la Figura 7 muestra que un valor atípico (en el rincón inferior derecho) puede reducir una correlación perfecta a casi nada. Recíprocamente, el panel de la derecha muestra un valor atípico (el del extremo superior derecho) que lleva la correlación de ser prácticamente nula a un valor cercano a uno.

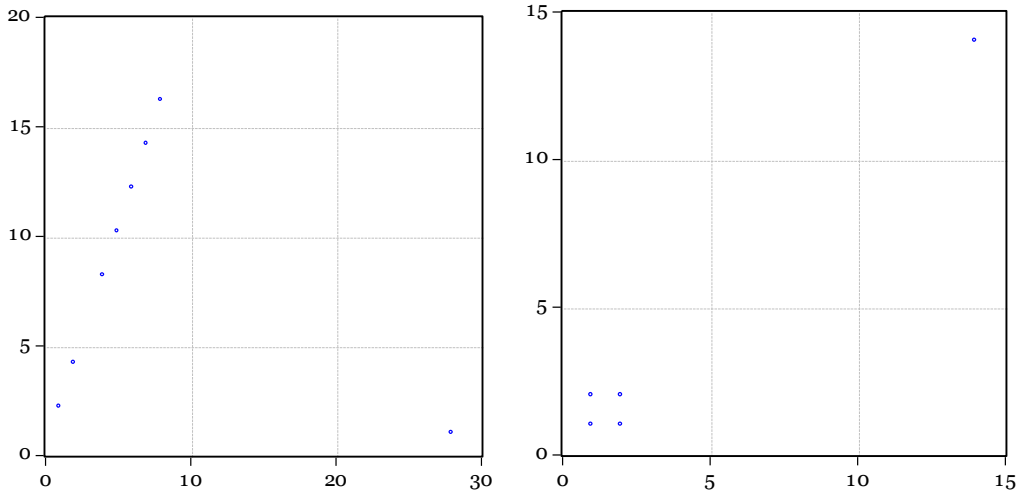


Figura 7. Coeficiente de correlación distorsionado por valores atípicos. A la izquierda, un valor atípico destruye una correlación perfecta. A la derecha, el valor atípico cambia r~0 hasta r~1.

Variables Confusivas El coeficiente r mide la asociación entre dos variables. En general, los investigadores – y los tribunales – están más interesados en la causalidad. Asociación no significa lo mismo que causalidad. Como hemos visto, la asociación entre dos variables puede ser accionada en gran medida por una "tercera variable" omitida del análisis. Un ejemplo sencillo: en los chicos de la escuela el tamaño del calzado y el de su vocabulario están asociados. Pero esto no significa que aprendiendo más palabras sus pies sean más grandes, ni que los pies hinchados hagan que los chicos articulen mejor su vocabulario. En este caso, la tercera variable es fácil de ser identificada – es la edad. Pero en ejemplos más realistas, podemos encontrarnos con casos en que la tercera variable sea más difícil de identificar.

Los métodos básicos de tratar a las variables confusivas implican experimentos controlados o aplicar, mediante la técnica de regresión múltiple "controles estadísticos".[154] Hay ejemplos en que la asociación refleja en realidad causalidad, pero un coeficiente de correlación grande no es suficiente para garantizarlo. Que r sea grande sólo significa que la variable dependiente se mueve en tándem con la variable independiente – sea por cualquier razón, desde la causalidad a la confusión.[155]

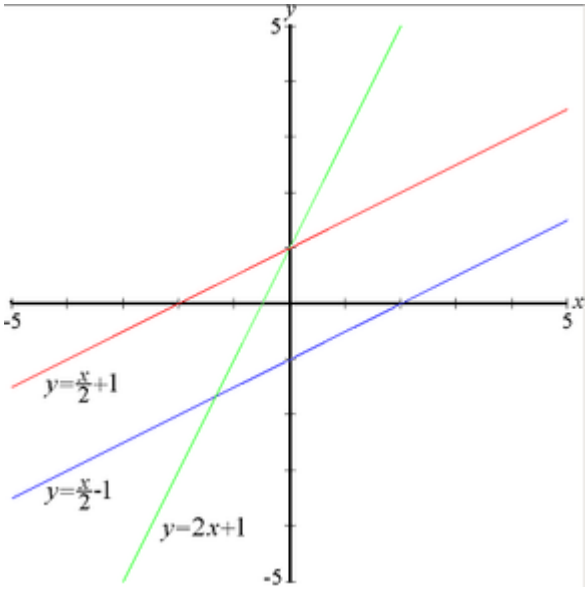


Figura 8. Tres líneas rectas — Las líneas roja y azul poseen la misma pendiente (m) que en este ejemplo es ½, mientras que las líneas roja y verde interceptan al eje *y* en el mismo punto, por lo que poseen idéntico valor de ordenada al origen (b) que en este ejemplo es el punto *x=0, y=1.*

Líneas de Regresión Una línea de regresión puede ser usada para describir una tendencia lineal de los datos. Por ejemplo, las líneas de regresión de la Figura 8 podrían describir la conducta del ingreso mensual medio (*y*) para determinado nivel educativo (*x*) que corresponde a períodos bianuales de nivel educativo alcanzados. En el caso de la línea más empinada (de color verde) el ingreso bianual promedio de gente con 10 años de estudio sería de $ 5,000, indicado por la altura de la línea para el nivel 2. El nivel medio de ingresos de la gente con 10 años de estudio, en otra jurisdicción (línea roja) más pobre sería de $ 2,000.

Veamos un ejercicio realizado a fin de establecer tablas referenciales del Flujo Espiratorio Pico en niños y adolescentes sanos en la provincia Ciudad de la Habana, consistente en indicar al individuo que realice 3 hiperventilaciones antes de la prueba; para comenzar debe realizar una inspiración profunda en la que trate de tomar la mayor cantidad de aire posible y luego realice una espiración forzada, expulsando todo el aire contenido en sus pulmones, cuidando que no se escape fuera de la boca. Se hicieron 3 mediciones y se escogió el mejor resultado. En la Figura 9 se grafican los valores del Test de flujo espiratorio pico[156] cubano del sexo femenino obtenidos en la investigación que muestra el diagrama de dispersión y la regresión entre la *talla* y los *resultados*, donde la mayoría de las mediciones están concentradas o agrupadas sobre la línea mostrando poca dispersión, lo que demuestra favorablemente el resultado de la prueba.

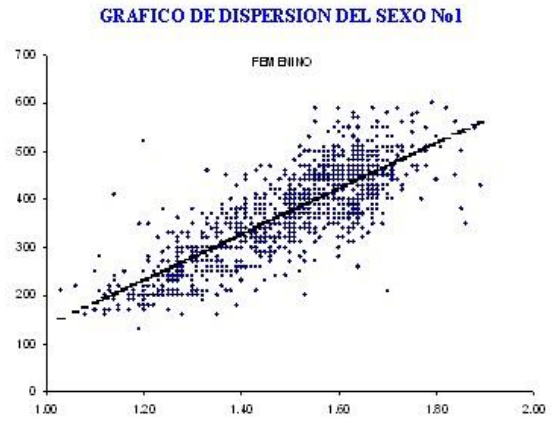


Figura 9. Test de flujo espiratorio pico

[154] Por los motivos ya expuestos, los esfuerzos de aislar las variables confusivas en los estudios observacionales en general son menos convincentes que en los experimentos al azar controlados.

[155] Al cuadrado del coeficiente de correlación, r^2, a veces se lo llama la proporción de varianza "explicada". Empero, "explicada" está dicho en un sentido totalmente técnico, y un amplio valor de r^2 no significa que exista una explicación causal.

[156] Se obtiene determinando la diferencia de presión entre los alvéolos y la boca por unidad de flujo aéreo y se mide con el pletismógrafo corporal. También se mide a partir de la presión intra-pleural obtenida desde el globo intra-esofágico, pero entonces se incluye también la resistencia consecuencia de la viscosidad tisular.

Pendiente y Ordenada al Origen La línea de regresión puede ser interpretada en términos de su pendiente y su ordenada al origen.[157] En la Figura 8, hay dos líneas de pendiente igual a ½ y una tercera de pendiente igual a 2. En la Figura 9, la pendiente es 666,66 ml de flujo espiratorio pico por cm. de altura (se ha representado sólo el caso femenino, ya que existe un diagrama similar para individuos del otro sexo). Es decir, que el aumento de talla de la persona de 10 cm. viene acompañado de un aumento del flujo espiratorio pico de aprox. 6670 ml por año. En la Figura 9 se observa que la ordenada al origen es aproximadamente 140 ml/año. Este estimador no es demasiado bueno, porque 1º) la persona estaría muerta, y carece de sentido su cálculo; 2º) adicionalmente, estas observaciones están muy alejadas del centro del diagrama. En general, las predicciones basadas en líneas de regresión resultan menos confiables a medida que nos alejamos de la masa de datos.

La pendiente tiene idénticas limitaciones que el coeficiente de correlación al tratar de medir el grado de asociación:[158] (1) Sólo mide relaciones lineales; (2) puede estar influida por valores atípicos; y (3) no controla el efecto de otras variables. Considerando los valores de la Figura 8, la asociación entre educación e ingreso es causal sólo parcialmente, porque hay otros factores a considerar, incluyendo la estructura familiar de la gente de la muestra. En cuanto a (1), la pendiente de ½ por cada bienio presenta a cada año de educación adicional como si tuviera el mismo valor, pero algunos años de escolaridad serán más valiosos que otros. Por ejemplo, la gente con grado escolar proviene probablemente de familias más ricas y con mejores estudios que los que abandonan después de tomar un curso. Los graduados tienen otras ventajas además de la educación extra. Factores como éstos seguramente influyen sobre el ingreso ganado. Por tal motivos, los estadísticos cualifican su lenguaje de "en promedio" y "asociado/a con".

Unidad de Análisis Si resulta de interés la asociación entre las características de los individuos, estas características deben ser medidas en los individuos propiamente dichos. A veces los datos individuales no están disponibles, pero se dispone de tasas de variación o de promedios; a las correlaciones computadas a partir de tasas o de promedios se las llama "ecológicas". Empero, las correlaciones ecológicas en general sobre-estiman la fuerza de una asociación. Ejemplo: la correlación entre ingreso y educación de todos los varones de los US es sólo de 0,44.[159] Pero no son los estados los que asisten a las escuelas y obtienen ingresos por su trabajo, sino la gente. La correlación de los promedios de los estados sobre-estima la correlación de los individuos, lo que constituye una tendencia común de las correlaciones ecológicas.[160] Estas correlaciones son usadas con frecuencia en ciencias políticas y en sociología; por lo tanto ¡a tener cuidado!

---

[157] Como toda recta, la línea de regresión tiene una ecuación que responde a la fórmula $y = mx + b$. Aquí, m es la pendiente, el cambio de y por cambio unitario de x. La pendiente es la misma en cualquier lugar de la línea. Esto distingue a las líneas rectas de las curvas. La ordenada al origen b es el valor que asume y cuando x es cero. La pendiente de una línea es similar a la pendiente de una ruta; la ordenada al origen proporciona la elevación inicial. En la Fig. 8, la línea de regresión estima el ingreso bianual medio en $10,000 de los que tienen 10 años de educación. Esta cifra se puede computar a partir de la pendiente y de la ordenada al origen como sigue: ($2,000 cada año) x 2 períodos de 1 año cada uno + $ 1,000 = $ 5,000 por año = $ 10,000 (total).

[158] El coeficiente de correlación es la pendiente de la línea de regresión cuando las variables están "normalizadas", es decir medidas en términos de desvíos estándar a partir de la media.

[159] El organismo encargado de computarla es el Bureau of the Census, Department of Commerce, para la March 1993 Current Population Survey.

[160] Una correlación ecológica utiliza solamente datos promedio, pero dentro de cada estado o provincia hay mucha dispersión en torno al promedio. La correlación ecológica pasa por alto esta variación individual. Para un ejemplo, consultar Epidemiología General y Demografía Sanitaria, que tiene una discusión sobre el uso de este instrumento en epidemiología. https://cv2.sim.ucm.es/moodle/file.php/13035/Estudios_Ecologicos.pdf

Modelos Estadísticos Los modelos estadísticos son muy utilizados en las ciencias sociales y en las contiendas judiciales (pero la frecuencia de su uso no implica que constituyan siempre la mejor opción frente a un problema particular). Por ejemplo, si el censo de población sufre un recuento de individuos inferior al real, más serio en algunos lugares que en otros, de acuerdo con ciertos modelos estadísticos, si confiamos en ellos este error de conteo podría ser corregido cambiando bancas en el Congreso y millones de pesos anuales de los programas de ayuda social con fondos del gobierno. Hay otros modelos que tratan de levantar el velo del secreto de la urna electoral, permitiendo que los expertos determinen cómo votaron distintos grupos sociales de clase media o de cualquier otro tipo (mujeres, gente analfabeta, etc.) – lo que constituye un paso crucial en los litigios para otorgar validez a los derechos de voto. Ahora discutiremos la lógica estadística de los modelos de regresión, dejando un estudio más detallado para un capítulo próximo.

Un modelo de regresión intenta combinar los valores de ciertas variables (llamadas variables independientes) al efecto de obtener valores esperados para otra variable (llamada variable dependiente). El modelo puede expresarse como una ecuación de regresión. Una ecuación de regresión simple sólo tiene una variable independiente, mientras que una ecuación de regresión múltiple tiene varias variables independientes. Los coeficientes de la ecuación a menudo serán interpretados como indicando los efectos de cambiar las variables correspondientes. Por ejemplo, la ley de elasticidad de Hooke o ley de Hooke, originalmente formulada para casos de estiramiento longitudinal, establece que el alargamiento unitario que experimenta un material elástico es directamente proporcional a la fuerza aplicada F:

[1] $\delta / L = F / (A.E)$

siendo δ el alargamiento, L la longitud original, E el módulo de Young, A la sección transversal de la pieza estirada.[161] La ley se aplica a materiales elásticos hasta un límite denominado límite elástico. Habrá cierto número de observaciones de una cuerda. Imaginen que, en cada observación, el físico cuelga un peso de la cuerda, y mide simultáneamente su longitud. Un estadístico podría aplicar un modelo de regresión a estos datos; para una amplia variedad de pesos:[162]

[1'] Longitud = $\alpha + \beta$. Fuerza + ξ.

El término que representa el error, denotado con la letra griega psi (ξ) es necesario porque la longitud medida no será exactamente igual a [α+ β. Fuerza]. Si no hay nada más, el error de medición debe ser reconocido como tal. Modelamos a ξ como si fuera una extracción aleatoria con reemplazo de una urna de tickets. Cada ticket muestra un error potencial que se realizará si se extrae ese ticket. El promedio de todos los errores en la urna se supone que es cero. En términos más estadísticos, se supone que los errores ξ de las distintas observaciones están "independiente e idénticamente distribuidos, con media cero".[163]

[161] http://es.wikipedia.org/wiki/Ley_de_elasticidad_de_Hooke

[162] La variable dependiente de la ecuación [1] es la longitud δ de una cuerda de sección transversal dada, del lado izquierdo de la ecuación. Hay una variable independiente o explicativa en el segundo miembro – el peso (ya que el módulo de Young E es meramente un *parámetro* que caracteriza el comportamiento de un material elástico, según la dirección en la que se aplica una fuerza (http://es.wikipedia.org/wiki/M%C3%B3dulo_de_Young). En general, un parámetro como éste (también llamado "módulo de elasticidad longitudinal") puede ser tabulado en un cuadro; para ver el valor del módulo de elasticidad de distintos materiales hay tablas específicas, como el de las constantes elásticas de diferentes materiales http://es.wikipedia.org/wiki/Anexo:Constantes_el%C3%A1sticas_de_diferentes_materiales. Como hay una sola variable explicativa (F), la ecuación [1] es una ecuación de regresión simple.

[163] Para ciertos fines, también se suele suponer que estos errores siguen una distribución normal. Observen que si la media de los errores fuera una constante c, positiva o negativa, podríamos sumarla a la constante α, dejando la media de los errores igual a 0, sin ningún otro cambio.

En esta ecuación [1'], a y b son parámetros, constantes *desconocidas* de la naturaleza que son características de cada cuerda: a es la longitud de la cuerda si no hay carga, y b es la elasticidad, o aumento de la longitud unitaria por unidad de incremento del peso o fuerza ejercida. Estos parámetros no son observables[164] pero pueden ser estimados por el "método de mínimos cuadrados", un método desarrollado por Adrien-Marie Legendre (francés, 1752–1833) y Carl Friedrich Gauss (alemán, 1777–1855) para ajustar las órbitas de los planetas alrededor del Sol. En notación estadística, los estimadores se denotan con letras griegas; así, *a* es el estimador de *α*, y *b* el estimador de *β*. Los valores de *a* y *b* son elegidos para minimizar la suma de los "errores de predicción" elevados al cuadrado.[165] A estos errores se los llama también "residuos", ya que miden la diferencia entre la longitud real y la longitud predicha de la cuerda, siendo esta última a + b. Fuerza.[166]

[2] *residuo = Longitud real – a – b. Fuerza.*

Obviamente, nadie imagina que haya una urna de tickets oculta en la cuerda. Empero, en varias pero no en todas las circunstancias la variabilidad de las mediciones físicas se parece en forma notable a la variabilidad de extracciones de una urna.[167] En resumen, *el modelo estadístico se corresponde en forma bastante estrecha con los fenómenos empíricos.*

Ejemplo en Ciencias Sociales Ahora volcaremos nuestra atención a una aplicación a las ciencias sociales del tipo que podría observarse en cuestiones litigiosas. Estudiar un caso llevaría muchas páginas, pero un ejemplo estilizado del análisis de regresión usado para demostrar la discriminación sexual en materia salarial puede brindar una idea apropiada. Veremos un tratamiento más extenso de estos conceptos en un capítulo posterior. Utilizaremos un modelo de regresión para predecir los salarios (en dólares/año) de los empleados de una empresa usando tres variables explicativas: la educación (años completados de escolaridad), experiencia (años trabajando en la empresa), y una variable dummy para género, que adopta valor=1 si es hombre y =0 si es mujer.[168] Supongan que la ecuación estimada es la siguiente:[169]

[3] *Salario predicho = $7,100 + $1,300.Educación + $2,200.Experiencia + $700.Género*

---

[164] Da la sensación de que en realidad a es observable; después de todo siempre es posible medir la longitud de una cuerda sin pesos. Pero como la medición está sujeta a errores, lo que uno observa en realidad no es α sino [α+ξ]. Los parámetros α y β pueden ser estimados, y aún muy bien estimados, pero *no pueden ser observados en forma directa*.

[165] Dados valores ensayados para α y β, se computan los residuos como en la ecuación [2], y entonces se calcula la suma del cuadrado de estos residuos. Los estimadores a y b son los valores de α y β que minimizan esta suma de cuadrados. Estos valores de mínimos cuadrados pueden ser fácilmente computados a partir de los datos mediante fórmulas matemáticas. Son la ordenada al origen y la pendiente de la recta de regresión.

[166] Observen que los residuos son observables, pero como los estimadores a y b son solamente aproximaciones de los parámetros α y β, un residuo es una aproximación al término de error ξ de la ecuación [1']. Se usa el término "valor predicho" en sentido especial, porque también se dispone de los valores reales de las variables; los estadísticos suelen referirse a "valor ajustado" en lugar de "valor predicho", a fin de evitar errores de interpretación.

[167] Éste es el término que usaba Gauss para referirse al error de medición.

[168] Una variable dummy ("muda") sólo adopta dos valores (p.ej., 0 y 1) y sirve para identificar dos categorías exhaustivas que se excluyen entre sí.

[169] En esta ecuación [3], la variable del primer miembro, el salario, es la variable de respuesta. Del lado derecho están las variables explicativas – educación experiencia, y la variable dummy del género. Como hay varias variables explicativas, se trata más de una ecuación de regresión múltiple que de regresión simple. Esta ecuación [3] es sugerida, en cierta forma, por la "teoría del capital humano". Empero, persiste una incertidumbre considerable acerca de qué variables entran en la ecuación, qué forma funcional tiene ésta y cómo se comportan los errores. Agregar más variables no siempre es una panacea.

Es decir, a = $7,100, b =$1,300, etc. Según la ecuación [3], cada año adicional de educación significa en promedio $1,300; en forma similar, cada año adicional de experiencia agrega en promedio otros $2,200; y, lo que es más importante, la empresa otorga a los hombres una prima salarial de $700 por encima de las mujeres que tienen la misma educación y la misma experiencia, siempre hablando en promedio. Por ejemplo, un empleado varón con 12 años de educación (estudios secundarios) y 10 años de experiencia tendría un salario predicho de $7,100 + $1,300 x 12 + $2,200 x 10 + $2,200 x 10 + $700 x 1 = $7,100 + $15,600 + $22,000 + $700 = $45,400. Una empleada mujer de situación similar ganaría $7,100 + $1,300 x 12 + $2,200 x 10 + $2,200 x 10 + $700 x 0 = $7,100 + $15,600 + $22,000 + $0 = $44,700.

Hay un dato clave al demostrar discriminación, que consiste en establecer que el coeficiente estimado de la variable dummy sea estadísticamente significativo. Lo cual depende de los supuestos incorporados en el modelo. Por ejemplo, se supone que cada año de educación vale lo mismo (en promedio) para todos los años de experiencia que uno tenga, tanto para las mujeres como para los hombres. En forma similar, cada año adicional de experiencia se supone que vale lo mismo a lo largo de todos los años educativos, tanto para hombres como mujeres. Además, la prima pagada a los hombres no depende sistemáticamente de la educación o de la experiencia. La capacidad, la calidad de la educación o la calidad de la experiencia se suponen no tener influencia sistemática sobre las predicciones del modelo.[170]

Los supuestos realizados sobre el término de errores – que están independiente e idénticamente distribuidos entre personas del mismo conjunto de datos – resultan ser claves para computar los *p*-valores y demostrar la significación estadística. Los modelos de regresión que no producen coeficientes estadísticamente significativos no serán probablemente usados para establecer que existe discriminación, y la significación estadística no puede establecerse a menos que se hagan supuestos estilizados sobre los términos de error no observables.

El típico modelo de regresión se basa en una multitud de supuestos semejantes; si no fueran hechos, no se podrían obtener inferencias a partir de los datos. Con la ley de Hooke – ecuación [1] – el modelo descansa en supuestos relativamente sencillos de ser validados experimentalmente. La validación del modelo de discriminación salarial – ecuación [3] – es más difícil. La corte o el abogado pueden preguntar: ¿Cuáles son los supuestos que están detrás del modelo, y cómo se aplican al asunto discutido en el tribunal? Al respecto, es importante distinguir entre situaciones donde (1) la naturaleza de las relaciones entre variables es conocida y la regresión se usa para obtener estimadores cuantitativos, y (2) la naturaleza de la relación es desconocida en gran parte y la regresión se usa para determinar la naturaleza de la relación – e inclusive si existe alguna. La base estadística de la teoría de la regresión fue desarrollada para manejarse con situaciones del tipo (1), y la ley de Hooke constituye un ejemplo. La base del segundo tipo de aplicación es analógica, y la tensión de la analogía resulta una cuestión crítica.

Errores Estándar, estadísticos t, y Significatividad Estadística La prueba estadística de discriminación ahora depende de cuán significativo sea *d* (el coeficiente estimado del género); la significación se determina mediante un test *t*, usando el error estándar de *d*. El error estándar de *d* mide la diferencia probable entre *d* y δ, originada por la presencia del término aleatorio en la ecuación [3]. El estadístico *t* es igual a *d* dividido por su error estándar. Por ejemplo, en esa ecuación, d= $700. Si el error estándar de *d* es $325, en ese caso t = %700/ $325=2.15. Este resultado es *significativo*, lo cual implica que es difícil de ser explicado como simple resultado del azar. Bajo la hipótesis nula de que δ=0, existe sólo un 5% de probabilidad de que el valor absoluto de *t* (denotado como |*t*|) sea mayor que 2.

[170] Técnicamente, se supone que estas variables omitidas no guardan correlación con el término de error de la ecuación.

Luego, un valor de $t>2$ demostrará la significatividad estadística.[171] Por otra parte, si el error estándar fuera \$1,400, en tal caso t= \$700/ \$1400 = 0.5, en cuyo caso la discrepancia pudo deberse meramente al azar. Naturalmente, el parámetro δ es sólo un constructo de un modelo. Si el modelo es erróneo, el error estándar, el estadístico t, y el nivel de significación serán bastante difíciles de interpretar.

Aún si el modelo es aceptado, hay una cuestión ulterior: el 5% es una probabilidad para datos del modelo, o sea, P ($|t| > 2 = 0$). Sin embargo, el 5% a menudo es mal interpretado como P (δ=0 | datos). Este error es frecuente en la literatura de ciencias sociales, y suele aparecer como describiendo el testimonio de expertos. Para un estadístico frecuencialista, P (δ=0 | datos) no tiene sentido, ya que los parámetros no tienen variaciones aleatorias. Para un estadístico subjetivista, P (δ=0|datos) tiene sentido, pero calculado mediante el test t podría ser erróneo, porque las probabilidades a priori de δ=0 no se tienen en cuenta.[172]

Resumen Las principales ideas de la modelación mediante regresión pueden captarse con un hipotético intercambio entre un demandante que busca probar la existencia de discriminación salarial y una empresa que niega semejante acusación. El intercambio podría funcionar de la manera siguiente:

1. El demandante alega que la empresa acusada paga más a los empleados varones que a las mujeres, lo que da lugar *prima facie* a discriminación.

2. La empresa responde que a los hombres se les paga más porque están más educados y tienen más experiencia.

3. El demandante trata de refutar la teoría de la empresa ajustando una ecuación de regresión como la [3]. Aún luego de ajustar por diferencias de educación y experiencia, los hombres ganan \$700 anuales más que las mujeres, en promedio. Esta diferencia de pagos confirma la discriminación.

4. La empresa argumenta que una diferencia tan reducida como \$700 podría ser un resultado azaroso, y que no es prueba de discriminación.

5. El demandante replica que el coeficiente de "género" en la ecuación [3] es estadísticamente significativo, por cuyo motivo el azar no es una explicación adecuada de los datos.

La significación estadística se determina con referencia al nivel observado de significación usualmente abreviado como p. El p-valor depende no solamente de la muestra, sino del tamaño de la misma, entre otros factores.[173] A mayor tamaño de la muestra, a igualdad de otras condiciones, tanto más reducido será p – y más perentorio el argumento del demandante de que la disparidad no puede ser explicada por el azar. A menudo se utiliza una tasa de corte de 5%; si p resulta inferior al 5%, la diferencia es "estadísticamente significativa".

Hay casos en los cuales el p-valor fue interpretado como la probabilidad de que los acusados sean inocentes de discriminación. Pero esta interpretación es errónea: p

[171] Cabe notar que el valor de corte de 2 se aplica a muestras grandes. Las muestras pequeñas requieren umbrales más elevados.

[172] Para un objetivista, la barra vertical en "|" en P ($|t| > 2|$ δ= 0) significa "habiendo sido computada bajo el supuesto de". Para un subjetivista, la barra significa una probabilidad condicional.

[173] El p-valor depende del valor estimado del coeficiente y de su error estándar. Estas cantidades pueden computarse a partir de (1) el tamaño de la muestra, (2) las medias y los errores estándar de las variables, y (3) de las correlaciones entre pares de variables. El cómputo es bastante intrincado.

representa sólo la probabilidad de obtener un valor de un estadístico muy grande, suponiendo que el modelo es correcto y que el verdadero coeficiente de "género" es cero. Luego, aunque el modelo no esté sometido a discusión, un *p*-valor menor que 50% no demuestra necesariamente una "preponderancia de la evidencia" en contra de la hipótesis nula. En efecto, un *p*-valor menor que 5% o que 1% podría no satisfacer el estándar de preponderancia. En casos de discriminación en el empleo, y también en otros contextos, son utilizados una gran variedad de modelos. Lo cual no sorprende, dado que la ciencia no dicta ecuaciones específicas. Por consiguiente, en un caso muy discutido, es probable que el diálogo continúe con un intercambio acerca de cuál es el mejor modelo. Aunque de tanto en tanto los supuestos estadísticos son discutidos ante los tribunales[174] los argumentos más comunes están alrededor de la elección de las variables. Un modelo puede ser cuestionado porque omite variables que *deberían ser incluidas*[175] – por ejemplo, los niveles de capacidad o evaluaciones realizadas previamente; otro modelo puede ser desafiado porque incluye variables "contaminadas" que reflejan conductas pasadas discriminatorias de la empresa.[176] Es frecuente que cada parte prepare sus propias ecuaciones y tenga su propio equipo de expertos; en esos casos, el tribunal debe decidir cuál de los modelos – si es que hay alguno – es satisfactorio.[177]

[174] Un ejemplo de supuesto estadístico es que el término de error sea *estadísticamente independiente* entre las observaciones en la ecuación [3]; otro ejemplo es que los errores tengan media cero y varianza constante.

[175] Ejemplos: Smith v. Virginia Commonwealth Univ., 84 F.3d 672 (4th Cir. 1996) http://openjurist.org/84/f3d/672/smith-iii-v-virginia-commonwealth-university (disputa acerca de si las variables omitidas impiden un juicio sumario). Comparar Bazemore v. Friday, 478 U.S. 385 (1986), on remand, 848 F.2d 476 (4th Cir. 1988) http://openjurist.org/848/f2d/476/pe-bazemore-v-c-friday-pe-bazemore y Sobel v. Yeshiva Univ., 839 F.2d 18, 34 (2d Cir. 1988) http://openjurist.org/839/f2d/18 (la falla de incluir variables de productividad escolar no pervierte las diferencias salariales del estudio de regresión de los demandantes porque "los expertos de Yeshiva no ofrecieron motivos, ni de evidencia ni analíticas, para concluir que están correlacionadas con el sexo") con Penk v. Oregon State Bd. of Higher Educ., 816 F.2d 458, 465 (9th Cir. 1987) ("Las partes faltantes de la interpretación de la demanda de las ecuaciones de toma de decisión incluían factores tan determinantes de la calidad y la productividad como la calidad, la comunidad y el servicio institucional, la calidad de la investigación y de la enseñanza... todas ellas deben tener una influencia significativa sobre las decisiones salariales") y Chang v. University of R.I., 606 F. Supp. 1161, 1207 (D.R.I. 1985) (la regresión del demandante carece de peso sustancial porque el analista "excluyó variables importantes, aunque sabía que lo eran"). Los mismos problemas surgen también en modelos estadísticos más simples, como los usados para evaluar la diferencia entre dos proporciones. Ver p.ej. Sheehan v. Daily Racing Form, Inc., 104 F.3d 940, 942 (7th Cir. 1997) http://www.projectposner.org/case/1997/104F3d940 ("Fue completamente ignorada la más que remota posibilidad de que la edad esté correlacionada con una calificación legítima al trabajo, tal como la familiaridad con computadoras. Todos saben que la gente más joven se siente más cómoda con las computadoras que la gente de mayor edad, como esta última está más cómoda en general con los autos con cambios manuales que la gente más joven").

[176] Michael O. Finkelstein, The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases, 80 Colum. L. Rev. 737 (1980).

[177] Por ejemplo, Chang, 606 F. Supp. at 1207 ("a la corte le resulta claro que el modelo del demandado incluye instrumentos mejores, más útiles y confiables que la contraparte"); Presseisen v. Swarthmore College, 442 F. Supp. 593, 619 (E.D. Pa. 1977) ("Cada parte hizo un trabajo soberbio en desafiar el análisis de regresión de la otra, pero sólo hizo un trabajo mediocre al tratar de defender el propio... y la corte se queda sin nada en definitiva"), aff'd, 582 F.2d 1275 (3d Cir. 1978). Ver también http://courtlistener.com/ca2/3FCX/roberta-ottaviani-individually-and-on-behalf-of-al/

Apéndice

Probabilidad e Inferencia Estadística

La teoría matemática de la probabilidad consiste de teoremas derivados a partir de axiomas y definiciones. Lo que no está en controversia es el razonamiento matemático, sino cómo debería aplicarse la teoría; es decir, los estadísticos difieren sobre la interpretación adecuada en distintas aplicaciones. Hay dos interpretaciones principales. Para un estadístico subjetivista, las probabilidades representan *grados de creencia*, dentro de una escala comprendida entre 0 y 1. Si el estadístico es un objetivista, las probabilidades no son creencias, sino propiedades inherentes de un experimento. Si el experimento puede repetirse, entonces, a largo plazo, la frecuencia relativa de un evento tiende hacia su probabilidad. Por ejemplo, si se arroja una moneda insesgada, la probabilidad de cara es ½. Si repetimos el experimento, la moneda caerá cara aproximadamente la mitad del tiempo. Si un dado insesgado es echado a rodar, la probabilidad de sacar un as en una tirada es $^1/_6$; si el dado es arrojado varias veces, saldrá 1 cerca de una sexta parte de las veces.[178] A los estadísticos objetivistas se los llama frecuencialistas, mientras que los subjetivistas son Bayesianos, por el apellido del reverendo Thomas Bayes, Inglaterra, 1701-1761.[179]

Detalles Técnicos sobre el Error Estándar, la Curva Normal, y los Niveles de Significación

Recordemos el ejemplo del examen tomado a una población de 5,000 hombres y 5,000 mujeres entre los postulantes. Supongan que las tasas de éxito de estos hombres y mujeres fueron 60% y 35% respectivamente. La diferencia "poblacional" es 60%-35% = 25 puntos porcentuales. Elegimos a 50 hombres al azar, y a otras 50 mujeres. Resulta que en la muestra la tasa de éxito de los hombres es 58% y la de las mujeres 38%, de manera que la diferencia muestral es 58%-38% = 20 puntos porcentuales. En otra muestra, podríamos haber obtenido tasas de éxito de 62% y 36%, con una diferencia muestral de 26 puntos porcentuales. Y así sucesivamente.

En principio, podemos considerar el conjunto de todas las muestras posibles de la población, y hacer una lista de las diferencias correspondientes. Se trataría de una lista muy larga. En efecto, la cantidad de muestras distintas de 50 hombres y 50 mujeres que puede formarse es inmensa – cerca de 5×10^{240}, es decir un 5 seguido por 240 ceros, que es mayor que el objeto denominado googol (10 elevado a una potencia de cien, superior al número de átomos del universo que sería de un orden comprendido entre 10^{72} y 10^{87} – sin contar la llamada "materia oscura"). La diferencia muestral fue elegida al azar de esta lista. La teoría estadística nos permite formular algunos enunciados precisos sobre la lista, y por consiguiente sobre las chances del procedimiento muestral.

• El promedio de la lista – es decir, el promedio de diferencias sobre las 5×10^{240} muestras posibles – resulta igual a la diferencia entre las tasas de éxito de todos los 5,000 hombres y 5,000 mujeres. En lenguaje más técnico, *el valor esperado de la diferencia muestral es igual a la diferencia poblacional*. Más lacónicamente, *la diferencia muestral es un estimador insesgado de la diferencia poblacional*.

[178] Las probabilidades pueden ser estimadas mediante las frecuencias relativas, pero la probabilidad en sí constituye una idea más sutil. Por ejemplo, supongan que una computadora imprime una sucesión de 10 letras H y T (por cara y cruz), que alternan como sigue: H T H T H T H T H T. La frecuencia relativa de caras (H) es 5/10 o 50%, pero no resulta obvio que la chance de tener H en la próxima posición sea 50%.

[179] No hablaremos aquí de la teoría axiomática. Pueden consultar la obra de E. T. Jaynes, Probability Theory: The Logic of Science, Washington University, 1995. http://bayes.wustl.edu/etj/prob/book.pdf Los axiomas de probabilidad son condiciones mínimas que deben verificarse para que una función definida sobre un conjunto de sucesos determine consistentemente sus probabilidades. Fueron formulados por Kolmogórov en 1933.

• El desvío estándar (SD) de la lista – es decir, el desvío estándar de todas las diferencias a lo largo de las 5 x 10^{240} muestras posibles – es igual a[180]

[4] $\sqrt{\{(5{,}000 - 50) / (5{,}000 - 1)\}} \times \sqrt{\{[P_h (1 - P_h)]/50 + [P_m (1 - P_m)]/50\}}$

En [4], P_h representa la proporción de los 5,000 hombres postulantes que pasarían el examen, y P_m la de la correspondiente a las mujeres. Con las cifras postuladas de 60% y 35%, el desvío estándar de las diferencias muestrales sería de 9.6 puntos porcentuales:

[5] $\sqrt{\{(5{,}000 - 50) / (5{,}000 - 1)\}} \times \sqrt{\{[.60 (1 - .60)]/50 + [.35 (1 - .35)]/50\}} = .096$

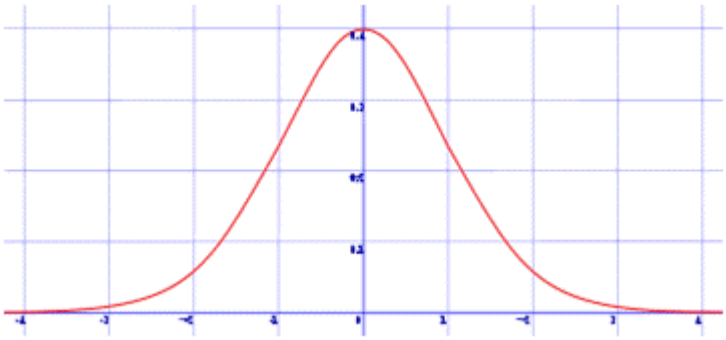


Figura 10. Función de densidad normal (0, 1)

La Figura 10 muestra el histograma de las diferencias muestrales, aproximado por la función normal.[181] El "teorema central del límite" dice que un histograma de diferencias muestrales, seguirá en forma aproximada a la curva normal (ver más adelante).

Pero en general no conocemos las tasas de éxito de la población de hombres y mujeres. ¿Qué hará un estadístico? Usará las tasas de éxito obtenidas en la muestra (58% y 38%) para estimar las tasas de éxito en la población. Sustituyendo en la ec. [4] tenemos

[6] $\sqrt{\{(5{,}000 - 50) / (5{,}000 - 1)\}} \times \sqrt{\{[.58 (1 - .58)]/50 + [.38 (1 - .38)]/50\}} = .097$.[182]

Algunas propiedades de la función de densidad normal (Figura 11):

1.- Es simétrica respecto a su media, μ;

---

[180] El desvío estándar de la diferencia muestral es igual al desvío estándar de la lista de todas las posibles diferencias muestrales, lo que establece una conexión entre el error estándar y el desvío estándar. Si sacamos dos muestras al azar, la diferencia entre las mismas estará en el orden de 2 ≈ 1.4 veces el desvío estándar. En tal caso, el error estándar puede usarse para medir la reproducibilidad de los datos muestrales.

[181] La curva normal es la famosa curva en forma de campana de la estadística, de ecuación

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R},$$

escrita en forma normalizada, donde μ es la media de la distribución y σ la desviación estándar (σ^2 es la varianza). Para apreciar cuán precisamente aproxima en este caso la curva normal estándar (es decir, cuando μ=0 y σ=1) a la distribución de diferencias muestrales de las tasas de éxito cuando P_h=60% y P_m=35%, ver la Figura 11.

[182] Observen que hay escasa diferencia entre [5] y [6] – los errores estándar no dependen demasiado de las tasas de éxito.

2.- Distribución de probabilidad alrededor de la media en una distribución N (μ, σ). La moda y la mediana son ambas iguales a la media, μ;

3.- Los puntos de inflexión de la curva se dan para x = μ − σ y x = μ + σ.

4.- Distribución de probabilidad en un entorno de la media:

- en el intervalo [μ - σ, μ + σ] está comprendido, aprox., el 68,26% de la distribución;
- en el intervalo [μ - 2σ, μ + 2σ] se encuentra, aprox., el 95,44% de la distribución;
- por su parte, en el intervalo [μ -3σ, μ + 3σ] se halla, aprox., el 99,74% de la distribución. Estas propiedades son de gran utilidad para establecer intervalos de confianza. Por otra parte, el hecho de que prácticamente la totalidad de la distribución se encuentre a tres desvíos estándar de la media justifica los límites de las tablas empleadas habitualmente en la normal estandarizada.

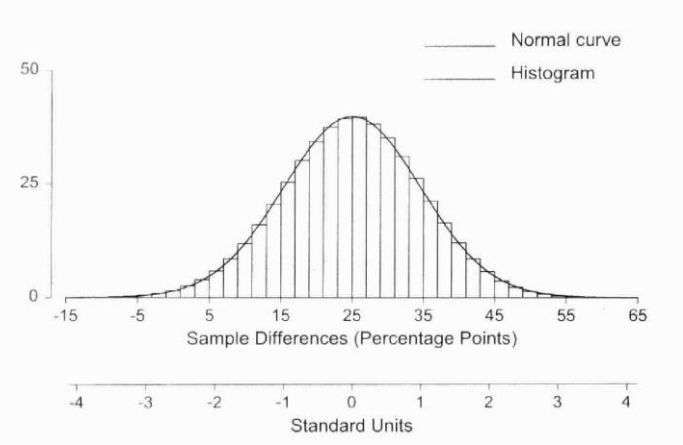


Figura 11. Función de densidad e histograma de las diferencias muestrales (%)

5.- Si X ~ N (μ, σ^2) y a y b son números reales, entonces (aX + b) ~ N (aμ+b, $a^2\sigma^2$). La notación "~" está indicando que la variable X se distribuye como una función normal (que en este caso tiene media=0 y varianza = σ^2).

Pasemos ahora a los *p*-valores. Sea la hipótesis nula de que hombres y mujeres de la población tienen las mismas tasas de éxito globales. En tal caso, las diferencias muestrales están centradas en cero, porque $P_h - P_m = 0$. Como la tasa global de éxitos de la muestra es 48%, usamos este valor para estimar P_h y P_m en la fórmula [4]:

[7] $\sqrt{\{(5{,}000 - 50) / (5{,}000 - 1)\}} \times \sqrt{\{[.48\ (1 - .48)]/50 + [.48\ (1 - .48)]/50\}} = .099$

De nuevo, el error estándar (SE) es de 10 puntos porcentuales. La diferencia observada de 20 puntos porcentuales es 20/10 = 2.0 SE. Como se aprecia en la Figura 11, diferencias de este orden de magnitud, o mayores, sólo tienen una chance del 5% de ocurrir. Aprox. un 5% del área ubicada por debajo de la curva normal llega más allá de ±2.[183]

Calculamos finalmente la potencia. Practicamos un contraste a dos colas al nivel de .05. En lugar de la hipótesis nula, suponemos como válida la alternativa: dentro del conjunto de postulantes, 55% de los hombres tendrían éxito, y 45% de las mujeres. Luego existe una diferencia de 10 puntos porcentuales entre las tasas

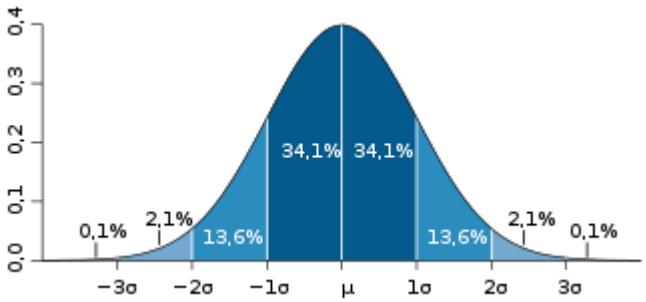


Figura 12. Áreas de la función normal.

[183] Técnicamente, el *p*-valor es la probabilidad de acceder a datos tan extremos o más extremos que los que se tiene a mano. Lo que significa es la chance de tener una diferencia de 20 puntos porcentuales o más a la derecha, junto con la chance de tener -20 o menos a la izquierda. (Esta chance es igual aproximadamente al área del histograma arriba de 19 junto con el área a la izquierda de -19). A su vez, el área debajo del histograma puede ser representada por el área de la curva normal más menos 1.9, que es aproximadamente 5.7%.

de éxito. La distribución de las diferencias muestrales ahora puede centrarse en 10 puntos porcentuales (Ver Figura 13). De nuevo vemos las diferencias muestrales se comportan con arreglo a la curva normal. El verdadero SE está en 10 puntos porcentuales de la ecuación [1] y el SE estimado resulta ser aprox. el mismo. Sobre esta base, sólo las diferencias muestrales mayores que 20 puntos porcentuales o menores que 20 puntos porcentuales serán declaradas significativas.[184] Luego, la potencia del test en contra de la hipótesis alternativa es sólo de alrededor de 1/6. Volveremos al problema de los errores cometidos en los test de hipótesis en un próximo capítulo.

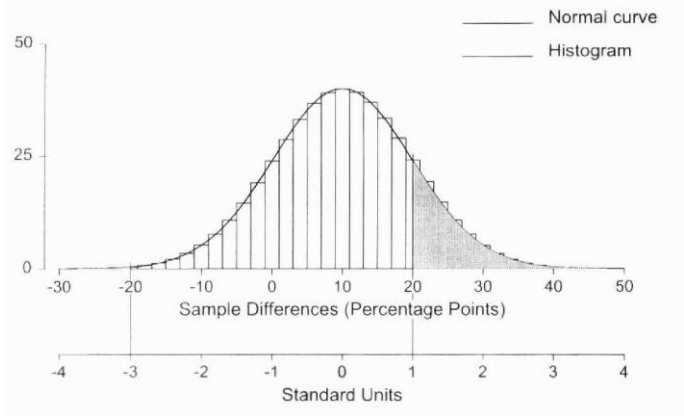


Figura 13

Glosario

El documento de David H. Kaye and David A. Freedman contiene un rico glosario de términos utilizados (pp. 160-177). En internet hay distintas alternativas, entre las cuales menciono un breve glosario de Fernando Valdés, de Comprensión y Uso de la Estadística, que contiene varios enlaces a distintos términos. Incluye traducciones de los términos en inglés y en francés.[185]

Bibliografía

Kaye, David H. and David A. Freedman, Reference Guide on Statistics, in Reference Manual on Scientific Evidence, 2nd ed., Federal Judicial Center (2000), pp. 83-178.[186] Puede resultar útil también analizar cuáles son las mayores dificultades del aprendizaje de estadísticas.[187]

Hay diversos textos en español que pueden ser bajados. Cabe mencionar a Juan Martínez de Lejarza, Grupo Consolidado de Acción Docente, Estadística,[188] de la Universidad de

[184] La hipótesis nula afirma que la diferencia es cero. En la Figura 13 de la Reference Guide (p. 157), 20 puntos porcentuales se hallan a 2 SE a la derecha del valor esperado bajo la hipótesis nula; asimismo, -20 está 2 SE a la izquierda. En cambio, la Figura 14 (p. 159) adopta la hipótesis alternativa como válida; sobre dicha base, el valor esperado es 10 en lugar de 0, de modo que 20 está 1 SE a la derecha del valor esperado, mientras que -20 está a 3 SE a la izquierda. Cerca de $^1/_6$ del área por debajo de la curva normal de la Figura 14 de la Reference Guide yace en esta región. Pongamos *t*= diferencia muestral/SE, estimando al SE a partir de los datos, como en [7]. Una versión formal del test rechaza la hipótesis nula cuando $|t| \geq 2$. Para hallar la potencia, reemplazamos el SE estimado por el SE verdadero, computado como en [7], y reemplazamos al histograma de frecuencias por la curva normal. Las dos aproximaciones son bastante buenas. El tamaño puede aproximarse de la misma forma, dado un valor común de las tasas de éxito de ambas poblaciones. También son posibles cálculos más exactos. En la figura, el área sombreada corresponde a la potencia. Las Figuras 12, 13 y 14 tienen una forma semejante, dado que es válido el teorema central del límite. Pero los histogramas tienen centros diferentes, porque los valores de P_h y P_m son distintos en los tres casos. La figura 12 (p. 156) está centrada en 25 puntos porcentuales, dado que refleja los valores ilustrativos de 60% y 35% de las tasas de éxito. La figura 13 (p. 157) está centrada en cero, porque fue dibujada según la hipótesis nula. La figura 14 (p. 159) está centrada en 10, porque se usa la hipótesis alternativa para calcular el centro, no la hipótesis nula.

[185] http://web.cortland.edu/flteach/stats/glos-sp.html

[186] http://www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf

[187] C. Batanero, J. D. Godino, A. Vallecillos, D.R. Green and P. Holmes, Errors and Difficulties in Understanding Elementary Statistical Concepts, International Journal of Mathematical Education in Science and Technology, Volume 25, Issue 4, 1994. http://www.ugr.es/~batanero/ARTICULOS/errors.PDF

Valencia, que incluye problemas resueltos y apuntes; a Luis Salvarrey, Curso de Estadística Básica,[189] Salto, R.O. Uruguay, (2000), un texto accesible; y Violeta Alicia Nolberto Sifuentes y María Estela Ponce Aruneri, Estadística Inferencial Aplicada, Lima, 2008.[190]

[188] http://www.uv.es/lejarza/syv/
[189] http://guajiros.udea.edu.co/descriptiva/articulos/Curso%20de%20EstadIstica%20Basica.pdf
[190] http://www.unmsm.edu.pe/educacion/postgrado/estadistica.pdf