InDret
REVISTA PARA EL ANÁLISIS DEL DERECHO
WWW. INDRET.COM

Probabilidad, estadística, grandes bases de datos y abogacía*

Albert Satorra
Facultad de Ciencias Económicas y Empresariales
Universitat Pompeu Fabra

Pablo Salvador Coderch
Facultad de Derecho
Universitat Pompeu Fabra

BARCELONA, ABRIL 2016

* Una versión preliminar de este trabajo fue presentada por los autores en la oficina de Madrid de Cuatrecasas, Gonçalves Pereira el 16 de febrero de 2016.

Abstract

La Sentencia 353/2005, de 25 de mayo, de la Sección 2ª de la Audiencia Provincial de Málaga (JUR 2005\195477, ponente Lourdes García Ortiz) -acaso el caso español más clamoroso de denegación de justicia por error judicial en lo que llevamos de siglo- ocupa el centro de un trabajo que pretende incorporar la estadística a la aplicación del derecho en España. Un estadístico y un jurista aúnan sus esfuerzos para analizar el caso mencionado y otros similares, españoles y extranjeros, en el intento de mostrar cómo muchas de nuestras intuiciones más arraigadas deberían de desaparecer del acervo de prácticas injustificadamente asociadas a las reglas de la sana crítica.

Así, los autores recomiendan a las personas y organizaciones dedicadas a la política jurídica, a la legislación, y a la aplicación, judicial y extrajudicial, del derecho que desconfíen de sus intuiciones y que piensen en términos estadísticos antes de proponer, aprobar o aplicar una regla o un principio de derecho, al tiempo que formulan guías para fundamentar estadísticamente propuestas y soluciones normativas.

En España, la tradición jurídica ha sido, para decir lo menos, refractaria a la Estadística, aunque este fenómeno es desgraciadamente universal: la desconsideración del parámetro de prevalencia, el desconocimiento de las probabilidades condicionales, el desapego de casi todos a la hora de tener en cuenta el tamaño de la base de datos, el desorden de quienes ignoran la distinción entre falsos positivos y falsos negativas, así como otras preteriticiones estadísticas han llevado a los sistemas judiciales de muchos países - igual o más desarrollados científica y tecnológicamente que el nuestro- a participar en el desfile de los horrores judiciales que los lectores reconocerán en este artículo, un alegato desgarrado y desazonador en pro de la recta aplicación de la Estadística en la formación y aplicación de cualquier juicio –no solo de las resoluciones judiciales- y en contra de la liviandad de los prejuicios de quienes creen que la intuición es fuente del derecho y de su recta aplicación.

The May 25th 353/2005 decision by Section 2 of the Audiencia Provincial de Málaga (JUR 2005\195477, reporting judge Lourdes García Ortiz) -perhaps the most resounding Spanish case this century regarding miscarriage of justice (denial of justice due to judicial error)- is the focus of a paper that intends to incorporate statistics to the application of law in Spain. A statistician and a jurist join their efforts in analyzing the aforementioned case and other similar ones, both Spanish and foreign, and attempt to show how many of our most rooted intuitions should disappear from the heritage of practices that are unjustifiably associated to the rules of reasoned judgment.

Thus, the authors' recommendation to people and organizations dedicated to judicial politics, legislation, and the judicial and extrajudicial application of law, is to distrust their intuitions and to think in statistical terms before proposing, adopting or applying a legal rule or principle, all the while formulating guides to statistically justify proposals and normative solutions.

In Spain, judicial tradition has been, to say the least, refractory to Statistics, although this phenomenon is unfortunately universal: the thoughtlessness towards the prevalence parameter, the lack of knowledge of conditional probabilities, the disregard by most of us of the databases' size, the disarray caused by those who ignore the distinction between false positives and false negatives, as well as other statistical omissions, InDret 2/2016 Albert Satorra y Pablo Salvador-Coderch 3 have led the judicial systems of many countries -

at least as scientifically and technologically developed as ours- to engage in displays of judicial horrors that readers will notice in this article, a torn and concerned plea in favor of the correct application of Statistics in the development and application of any trial -not only regarding judicial resolutions- and against the lightness of the prejudices of those who believe in intuition as a source of law and its correct application.

Title: Probability, statistics, big data base and law.

Palabras clave: Estadística aplicada al derecho, Falacia del fiscal. Parámetro de prevalencia. Regla del producto. Regla del complementario. Probabilidades independientes. Probabilidades condicionales. Teorema de Bayes. Denegación de justicia. Error judicial. Falsos positivos. Falsos negativos. Perfiles estereotípicos. Intuición y estadística. Grandes centros de datos. Ex ante y ex post. Carga de la prueba.

Keywords: Statistics applied to law. Prosecutor´s fallacy. Base rate. Product rule. Complement rule. Independent probabilities. Conditional probabilities. Bayes theorem. Miscarriage of Justice. Judicial error. Denial of Justice. False positives. False negatives. Profiling. Intuition and Statistics. Big data. Ex ante and ex post. Burden of proof.

Sumario

Introducción: derecho y estadística

En este trabajo, sus autores pretendemos acercar la estadística al derecho, tanto a su establecimiento como a su aplicación extrajudicial y judicial. Uno de nosotros (Albert Satorra) es matemático y profesor de estadística; el otro (Pablo Salvador) es jurista y profesor de derecho. La idea originaria proviene de dos juristas (Ward Farnsworth y Josep Santdiumenge) a quienes los autores quedamos muy agradecidos.

En el artículo recomendamos a las personas y a las organizaciones dedicadas a la política jurídica, al establecimiento de nuevo derecho y a su aplicación que desconfíen de sus intuiciones y que piensen en términos estadísticos antes de proponer o aplicar una regla (*rule*) o un principio de derecho (*standard*). En las páginas que siguen tratamos de justificar esta recomendación y, además, formulamos guías para fundamentar estadísticamente propuestas y soluciones normativas. Los lectores encontrarán en ellas las buenas razones que grandes matemáticos y estadísticos han ido construyendo en los útimos cuatro siglos.

Los autores tratamos de mostrar por qué muchas de nuestras intuiciones más aparentemente fecundas son, con harta frecuencia, engañosas: sí, casi todo el mundo es buena gente, pero el derecho y quienes tienen el encargo de aplicarlo han de tener en cuenta qué ocurre cuando sabemos que una de cada cien personas u organizaciones, o una de cada mil, o una de cada cien mil no es buena gente. O, ¿por qué hay sabiduría en el refrán que reza "Tanto va el cántaro a la fuente que al fin se rompe"?. O, también, ¿por qué ponemos énfasis en el tamaño o dimensión del grupo sobre el cual proyectamos nuestro cálculo?, ¿es que *size matters*?. Tanto es así que el trabajo pone en guardia a sus lectores para prevenirles contra conclusiones precipitadas: si la base de datos es lo suficientemente amplia, las coincidencias raras, rarísimas, inexplicables se darán y nadie tendrá la culpa de ello: así es *Big Data*. Y sin necesidad de llegar tan lejos, existen falsos inocentes (falsos negativos), pero no hay, que sepamos, técnica analítica que no dé falsos culpables (falsos positivos): por esto es siempre aconsejable no prejuzgar, no acusar, no imputar, ni siquiera señalar, antes de realizar un segundo y hasta un tercer análisis al primero que ha dado positivo: el derecho de verdad existe porque una de cada diez, de cada cien o de cada mil veces nos estaremos equivocando.

Esto último es muy importante, pues, como se explica en al artículo, la liviandad en la acusación, la ligereza de la primera impresión, la superficialidad de la sombra de una sospecha, sobre todo, cuando se ve reforzada por la propensión casi irrefrenable del encargado de decidir en derecho a no asumir las consecuencias de generar un falso inocente tiene resultados devastadores para las personas y organizaciones afectadas –piensen, solo por ejemplo, en las sentencias de conformidad, en la cara oscura del *plea bargain*-. Este es un trabajo cuyos autores tienen muy presente la imperiosa necesidad de evitar el prejuicio o, lo que es aún peor, la condena injustificada tras un solo análisis que dio positivo o que ni siquiera era correcto, sino que carecía de toda base estadística: la denominada falacia del fiscal (*Prosecutor´s Fallacy*) ocupa el lugar central de este artículo y el triple abordaje que se hace a su desgarradora realidad creemos que servirá para ayudar a más de una persona inocente[1].

Y es que la recomendación básica del artículo debería llegar a todas aquellas personas que tienen a su cargo las cosas del derecho: por favor, les rogamos, antes de señalar a nadie, vuelvan ustedes a analizar los datos con un método distinto e independiente al que han utilizado la primera vez; tengan ustedes siempre –siempre- en cuenta el parámetro de prevalencia (*base rate*), sepan muy mucho de qué estan ustedes hablando cuando afirman que el x por ciento de las personas o de las organizaciones adolecen de tal o cual defecto, vicio o carencia: les instamos a que, al mismo tiempo o inmediatamente a continuación se pregunten por el (1-x) por ciento restante; les exigimos que piensen en qué ocurre cuando multiplicamos un número menor que uno varias, bastantes o muchas veces por sí mismo, les urgimos a que no olviden nunca nunca que, en 1763, dos años después de la muerte del gran Thomas Bayes, el mundo empezó a saber con conocimiento de causa que muchas probalidades son condicionadas, es decir, que no son independientes: la probabilidad de que ustedes vayan a TED y vea y oigan a Peter Donnelly hablando de *"How Juries are fooled by statistics"*, no es independiente de que, antes, hayan leído este artículo. O, por lo menos, esta introducción.

1. *El profesional desleal, el estudiante deshonesto y el parámetro de prevalencia*

Uno de nosotros (Pablo Salvador) planteó la cuestión al otro (Albert Satorra), más o menos, del modo siguiente:

Albert, en una empresa de prestación de servicios de alto valor añadido a las empresas trabajan mil asociados y quinientos técnicos. La inmensa mayoría de los profesionales que trabajan en ella son honestos, trabajadores y leales para con la compañía, para con sus compañeros y para con sus clientes y la sociedad entera.Pero nunca se puede descartar que, entre cien profesionales haya un garbanzo negro. Ocurre, sin embargo que no siempre somos conscientes de las consecuencias de este hecho que aquí asumimos, pues nuestra intuición es que, sea cual sea el número de profesionales que trabajan en la empresa, seguirá siendo muy raro, tremendamente infrecuente encontrarnos con un profesional desleal. ¿Verdad que es muy improbable? ¿O no, o es bastante probable, hasta muy probable?

Lo mismo sucede en nuestra Universidad (La Universitat Pompeu Fabra): la inmensa mayoría de sus estudiantes no copian nunca, pero no se puede excluir que alguno o que algunos pocos lo hagan una rara vez, o hasta dos o tres. ¿Pero, verdad que es, que sigue siendo muy improbable? ¿O no es así?

La probabilidad, contesta uno de nosotros (Albert) de que un profesional sea desleal o de que un estudiante sea deshonesto (copiar en un examen) es ciertamente muy baja; supongamos que es 0,01 (1 de cada 100). A pesar de ello, la probabilidad de que no haya ningún estudiante deshonesto en la clase (un curso grande, pongamos n estudiantes) también será muy baja.

Veamos: Denotemos por A_i el suceso que el estudiante i (pongamos el estudiante JX) es

[1] No incluimos fotografías o grabaciones en video de las personas que protagonizaron los casos analizados en este artículo, pues, aunque ayudan a percibir realidades devastadoras, los lectores interesados pueden acceder fácilmente en la red (al menos hoy por hoy) a toda la información gráfica relevante.

deshonesto. Sabemos que $P(A_i) = 0{,}01$, de manera que por la Ley del Complementario, la probabilidad que sea honesto será muy alta:

$$P(\overline{A_i}) = 1 - P(A_1) = 1 - 0{,}01 = 0{,}99$$

y por la Ley del Producto de sucesos independientes, la probabilidad de que en el grupo de n no haya ningún estudiante deshonesto, será:

$$P(\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_n}) = 0.99 \times 0.99 \times \ldots \times 0.99 = 0.99^n$$

Como la probabilidad 0,99 es estrictamente menor que 1, como en cualquiera de las probabilidades de la vida, la potencia n de dicho número decrecerá rápido con n. Veamos la tabla de valores $0{,}99^n$, de probabilidad de 0 deshonestos en el grupo, para tamaños de grupo n que va de 2 a 500, cuando el parámetro de prevalencia de deshonesto es 0,01:

n	P(0 deshonestos)
2	0.98
10	0.90
50	0.60
100	0.37
200	0.13
300	0.05
500	0.01

Por ejemplo, en un examen al que se presentans 300 estudiantes, $0.99^{300} = 0{,}049$, la probabilidad de que no haya ningún estudiante deshonesto es baja, inferior al 5%. La Figura 1 muestra el gráfico de dichas probabilidades en función de n.

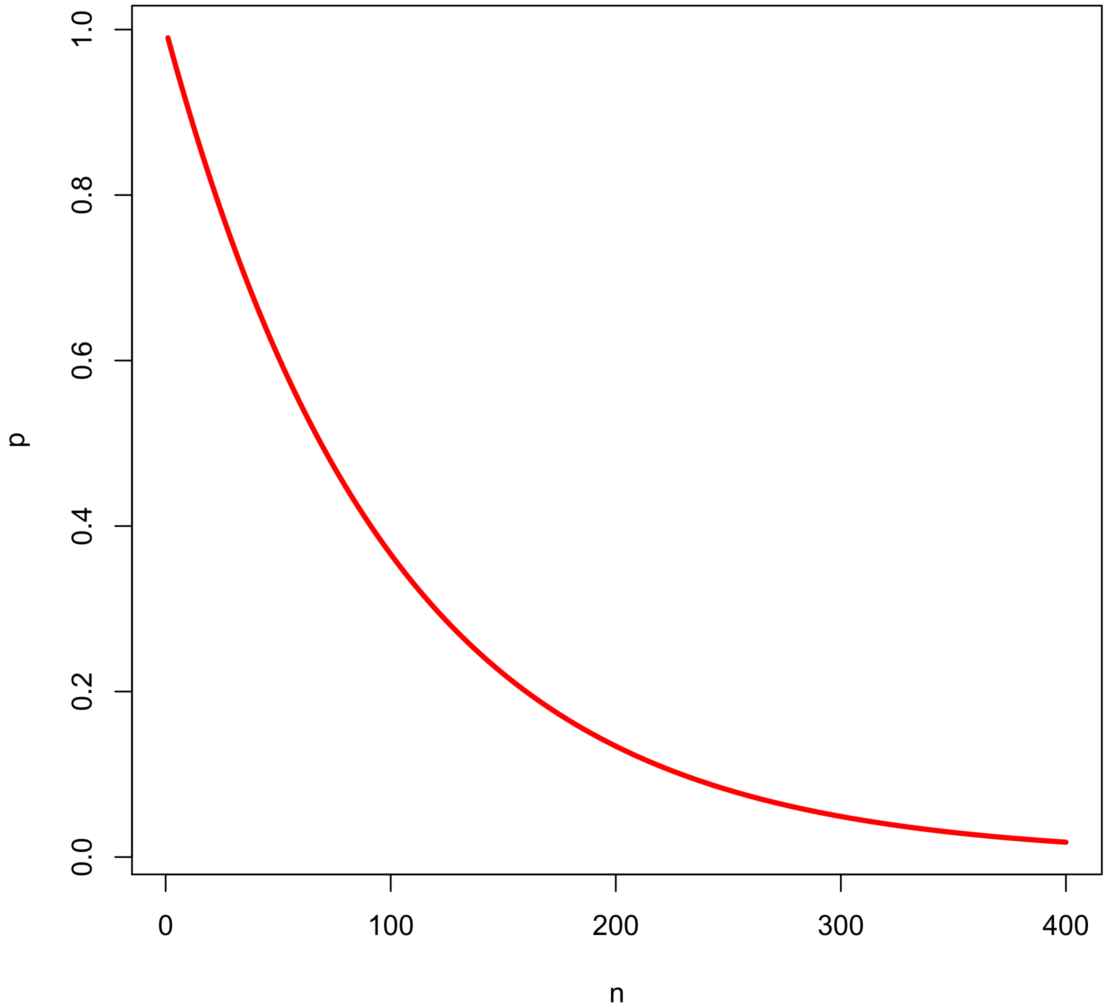

Figure 1: Probabilidad de que no haya ningún estudiante deshonesto en una clase de n estudiantes, dado un valor de prevalencia de 0,01

Vemos que la probabildad que no haya ningún estudiante deshonesto decrece rapidamente cuando crece el tamaño n del colectivo; a partir de tamaño de grupo n = 200, la probabilidad es muy pequeña.

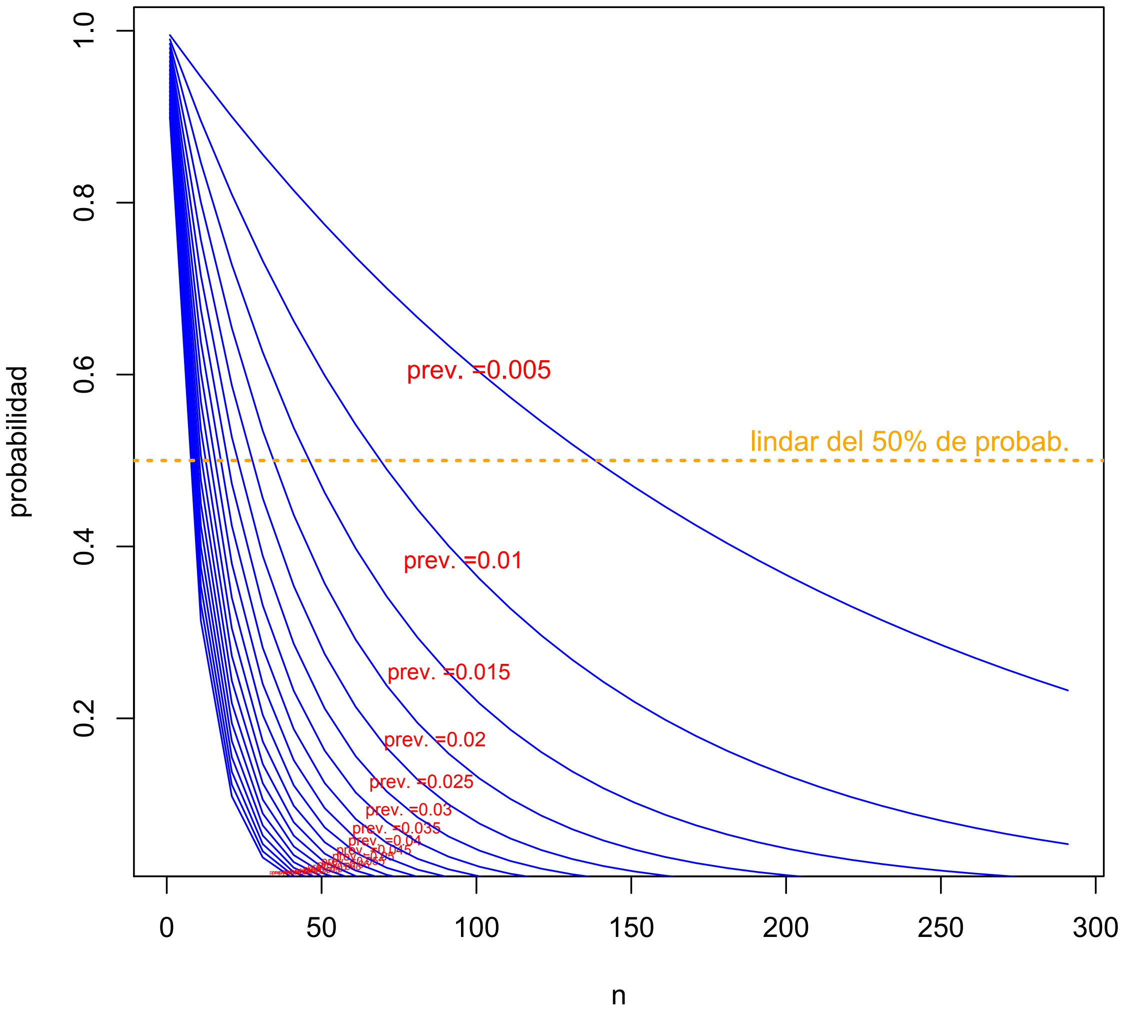

Figure 2: Probabilidad de ningún estudiante deshonesto en una clase de n estudiantes cuando variamos el parámetro de prevalencia ("prev")

La Figura 2 muestra las curvas de la probabilidad de que no haya ningún estudiante deshonesto en función de n, para diferentes valores del parámetro de prevalencia. Comprobamos entonces que la prevalencia (un parámetro que de seguro variará según las universidades, sus países, etc.) tiene un papel clave en la determinación del tamaño de grupo n a partir del cual la probabilidad de ningún estudiante deshonesto es muy baja.

2. *Accidente de ferrocarril en una línea de alta velocidad muy frecuentada (Alvia, 24.7.2013)*

En el trayecto de una línea de alta velocidad (plantea de nuevo uno de nosotros, Pablo) por la cual circula mucho tráfico de ferrocarril, entre cuatro y doce trenes cada día del año, hay una curva peligrosa. Antes de llegar a ella, el maquinista ha de moderar la velocidad del convoy de 275 km/h a 80 km/h. El manual de la vía, que el maquinista ha de dominar, y un doble sistema de señales –una visual y la otra auditiva– le advierten además de que está obligado a hacerlo así. Las señales se activan cuando el tren alcanza un punto determinado de la vía antes de la curva peligrosa.

Sin embargo, en otras líneas, la compañía gestora de las infraestructuras ya ha instalado un sistema de seguridad más avanzado que detiene automáticamente el tren si el maquinista no percibe el doble juego de señales, ni, por tanto, reacciona adecuadamente reduciendo la velocidad del tren.

La regulación aplicable al caso es confusa, pues no especifica claramente cuál de los dos sistemas de seguridad es imperativo, es obligatorio. De hecho, en 2013, en los tendidos de alta velocidad de España convivían dos sistemas similares a los que acabamos de resumir. Los criterios de integración de los deberes de diligencia – de comportarse precavidamente -de individuos y organizaciones tampoco son necesariamente claros en nuestras leyes y nuestras prácticas judiciales. Piénsese entonces que, según cual sea el criterio aplicable o el aplicado por el tribunal al caso concreto, la respuesta a la pregunta sobre si el individuo o la organización demandados o acusados actuaron con negligencia puede ser muy distinta. Así, si rigiera un estándar *BAT (Best Available Technology)*, la compañía de infraestructuras ferroviarias estaría obligada a instalar y mantener en buen funcionamiento la mejor tecnología accesible (no experimental) y si no lo hizo y hubo un accidente, incurriría en responsabilidad; en cambio, si el criterio fuera *ALARA (As Low As Reasonably Achievable)*, el órgano encargado de evaluar la situación emplearía un canon normativo de razonabilidad cercano al tradicional de definición de la negligencia entendido como infracción del canon de comportamiento al que se ajusta una persona *razonable*; mientras que si el estándar fuera de *Cost/Benefit*, el análisis habría de ser exclusiva o predominantemente económico. Luego, si el canon es muy exigente, pero solo se aplica a las infraestructuras construidas después de la entrada en vigor del establecimiento de aquel o solo a partir de los dos, tres o diez años de la entrada en vigor de la ley que lo impone (*Grandfathering, Grandfather Clause*: es muy frecuente que la obligatoriedad de instalación de los nuevos dispositivos de seguridad solo se aplique a las instalaciones, maquinaria o automóviles construidos o fabricados a partir de una fecha determinada: a lo mejor el viejo automóvil del lector está así apadrinado por tal o cual ley). Por último, puede regir un canon ultraexigente de diligencia extrema (*Utmost Care*), lo cual, en la práctica, implica un estándar de responsabilidad objetiva o sin culpa (algo que plantea muchas dificultades en un proceso penal, pues la responsabilidad objetiva quiebra el principio de culpabilidad).

Sea como fuere, un día aciago ocurre un accidente muy grave: el maquinista estaba utilizando su *smartphone*, descuidó las indicaciones del manual, y, además, no atendió a las dos señales, visual ni auditiva, no redujo la velocidad, el tren descarrilló, muchas personas murieron o resultaron muy gravemente heridas.

En el litigio posterior, se discute, entre otras cosas, la concurrencia de culpa o negligencia por parte de los responsables de la compañía de infraestructuras y de esta misma. Los abogados de la compañía alegan que las regulaciones aplicables no eran claras en relación con la cuestión de que si había de haberse instalado o no el sistema más seguro, aquel que detenía automáticamente el tren si su maquinista no atendía a las señales.

Un factor relevante para apreciar la responsabilidad por negligencia es la estimación de la probabilidad *ex-ante* de que ocurra un accidente por las causas descritas. Y lo es porque, tradicionalmente, la negligencia se estima o pondera en función del riesgo generado por la conducta hipotéticamente imprudente. Entonces: ¿no es verdad que resulta muy improbable que, en mi próximo viaje en AVE, el maquinista, avezado a la vía, desatienda el doble sistema de señales? ¿O no, no lo es?

Efectivamente (afirma Albert), el sistema de seguridad convencional, el ASFA, con que operaba el Alvia de Santiago, precisa de señales visibles en la via que sean percibidas e interpretadas por el maquinista. Aquí tenemos que entrar en consideraciones sobre la probabilidad de distracción del maquinista en un día concreto, y su impacto en la probabilidad de que como mínimo ocurra un accidente en repeticiones continuadas del viaje.

Sea $a = P(A_i) = 0.001$, la probabilidad de distracción del maquinista en día i. Probabilidad de accidente en n repeticiones del viaje será

$$P(A_1 \cup A_2 \ldots \cup A_K) = 1 - P(\overline{A_1 \cup A_2 \ldots \cup A_K}) = 1 - (1-a)^n = 1 - 0,999^n$$

Hemos utilizado la Ley de Morgan que dice que

$$\overline{A_1 \cup A_2 \ldots \cup A_K} = \overline{A}_1 \cap \overline{A}_2 \ldots . \cap \overline{A}_K$$

de manera que (Ley del Producto)

$$P(\overline{A_1 \cup A_2 \ldots \cup A_K}) = P(\overline{A_1}) \times P(\overline{A_2}) \times \ldots \times P(\overline{A_K}) = (1-\alpha)^K$$

La probabilidad que pase alguna vez en 1000 repeticiones del trayecto es

$1-(1-a)^{1000} = 1 - 0,37,$ del 63%; en 3000 repeticiones, la probabilidad que pase alguna vez $1-(1-a)^{3000} = 1 - 0.0497 = 0.9503$, muy alta, por encima del 95%!

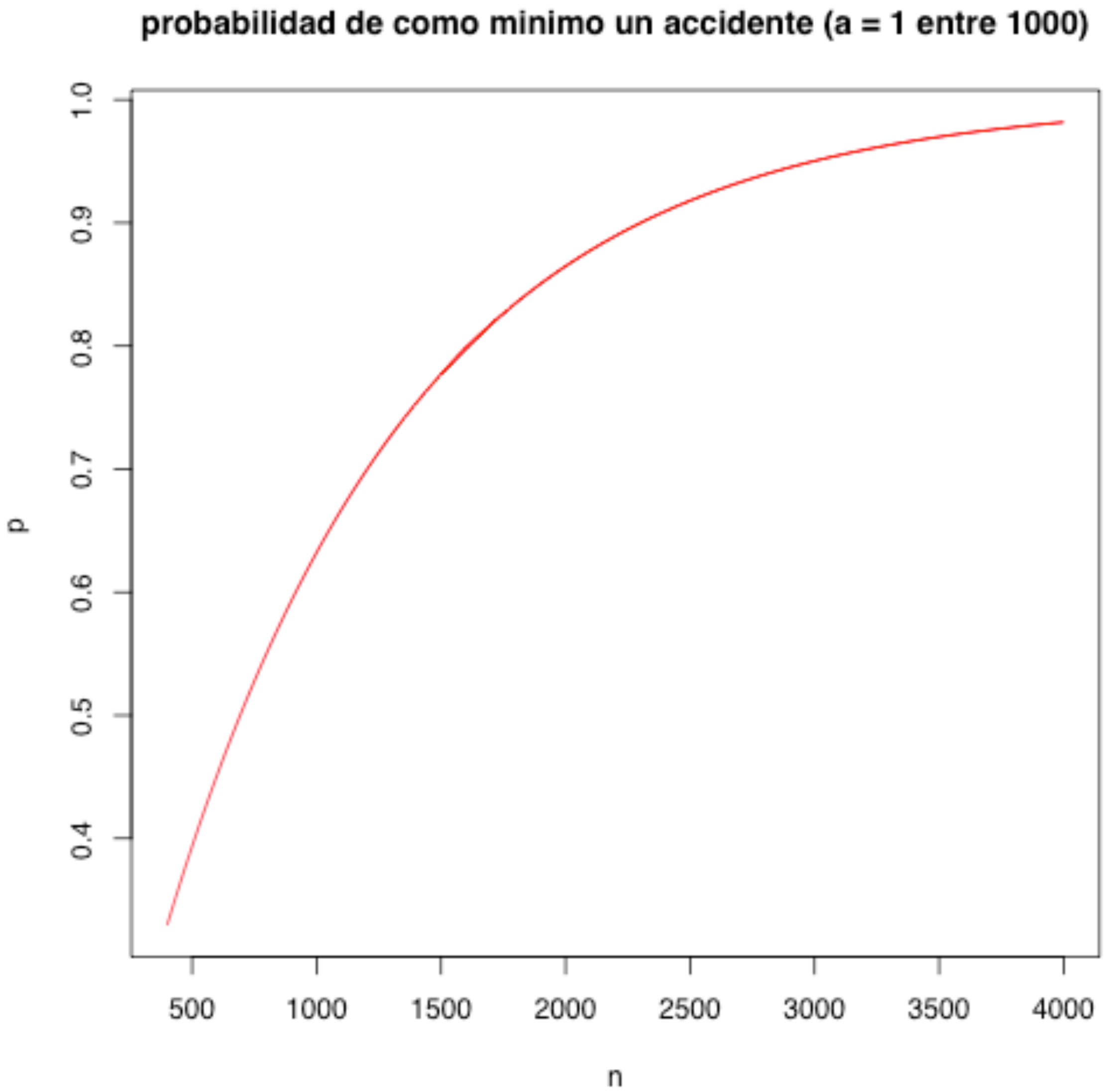

Figura 3: Es muy improbable que el suceso ocurra en un día concreto, pero al mismo tiempo es muy improbable que no pase nunca, en sucesivas repeticiones del viaje. ¿Eran necesarias medidas de seguridad adicionales? ¿Se podía dejar toda la seguridad a merced del maquinista.

Supongamos un protocolo de muchos items, $A_1, A_2, \ldots, A_K$. Cada uno de ellos tiene probabilidad muy alta de no fallar (pensemos por ejemplo en una operación que se descompone en componentes. Por ejemplo, el protocolo de cuidado de un enfermo de Ébola. Denotamos con A_k el suceso que indica que el ítem i no falla. Hay una probabilidad alta de A_k, pongamos $P(A_k) = 0.99$. ¿Cual es la probabilidad de que no falle el protocolo.

Interesa la probabilidad de la intersección: $P(A_1 \cap A_2 \cap A_3 \ldots \cap A_K)$. Una hipótesis razonable es independencia entre los A_i

$$P(\text{Éxito}) = P(A_1 \cap A_2 \cap A_3 \ldots \cap A_K) = P(A_k)^K = 0.99^K$$

Una probabilidad que de nuevo tiende a cero cuando K aumenta. El gráfico de la probabilidad de éxito depende de K y es como el de la Figura 1, en donde en lugar de n tenemos K, el número de componentes del protocolo.

3. *Probabilidad de coincidencia de aniversario en un grupo de n personas: Size Matters*

Otro ejemplo escolar que muestra cuán engañosa puede ser la intuición es el problema clásico de coincidencia de fecha de cumpleaños. En un grupo de n personas ¿cual es la probabilidad que haya como mínimo una coincidencia de cumpleaños? A primera vista, dado que hay 365 días posibles de aniversario (hemos simplificado, dejando fuera los años bisiestos), parecería que, para que la probabilidad de coincidencia fuera, como mínimo, del 50%, se necesitaría un n grande. Tal es la intuición. Un momento.Veamos: si denotamos por C el suceso que corresponde a como mínimo una coincidencia, entonces, aplicando de nuevo la ley del complementario y una variante de la ley del producto, tenemos:

$$P(C) = 1 - P(\overline{C}) = 1 - \frac{365 \times 364 \times \ldots \times (365 - n + 1)}{365^n};$$

Por ejemplo, en el caso de un colectivo de 25 personas la probabilidad de coincidencia de aniversario es de P(C) = 0.57, es decir, mayor del 50%. El mensaje es, pues, que resulta muy improbable que haya coincidencia de aniversario en un grupo de dos personas, pero muy improbable que no lo haya en un grupo grande de n personas!; cuando $n = 25$, dicha probabilidad supera ya el típico cara y cruz.

La idea tiene importancia en cualquier contexto más general. Por ejemplo, la probabilidad de coincidencia de personas de una misma escuela (o mismo barrio, o club de gimnasio, etc.) cuando formamos un panel, un tribunal o un jurado para un premio, una circunstancia que nos obliga a hilar muy fino cuando tal coincidencia pueda implicar un conflicto de intereses, algo que quienes entendemos que la justicia es ciega queremos evitar. Lo mismo ocurre con la probabilidad de que en un colectivo de personas haya como mínimo una coincidencia en el mismo perfil de ADN. Esto tiene importancia trascendental en criminología. Nos dice que **el valor n tamaño de la base de datos es fundamental para no incurrir en error de coincidencia**. En la Figura 4 se muestra la variación de la probabilidad de coincidencia de aniversario en función de n. Vemos que la probabilidad se acerca pronto a 1 a poco que aumentemos el tamaño de colectivo.

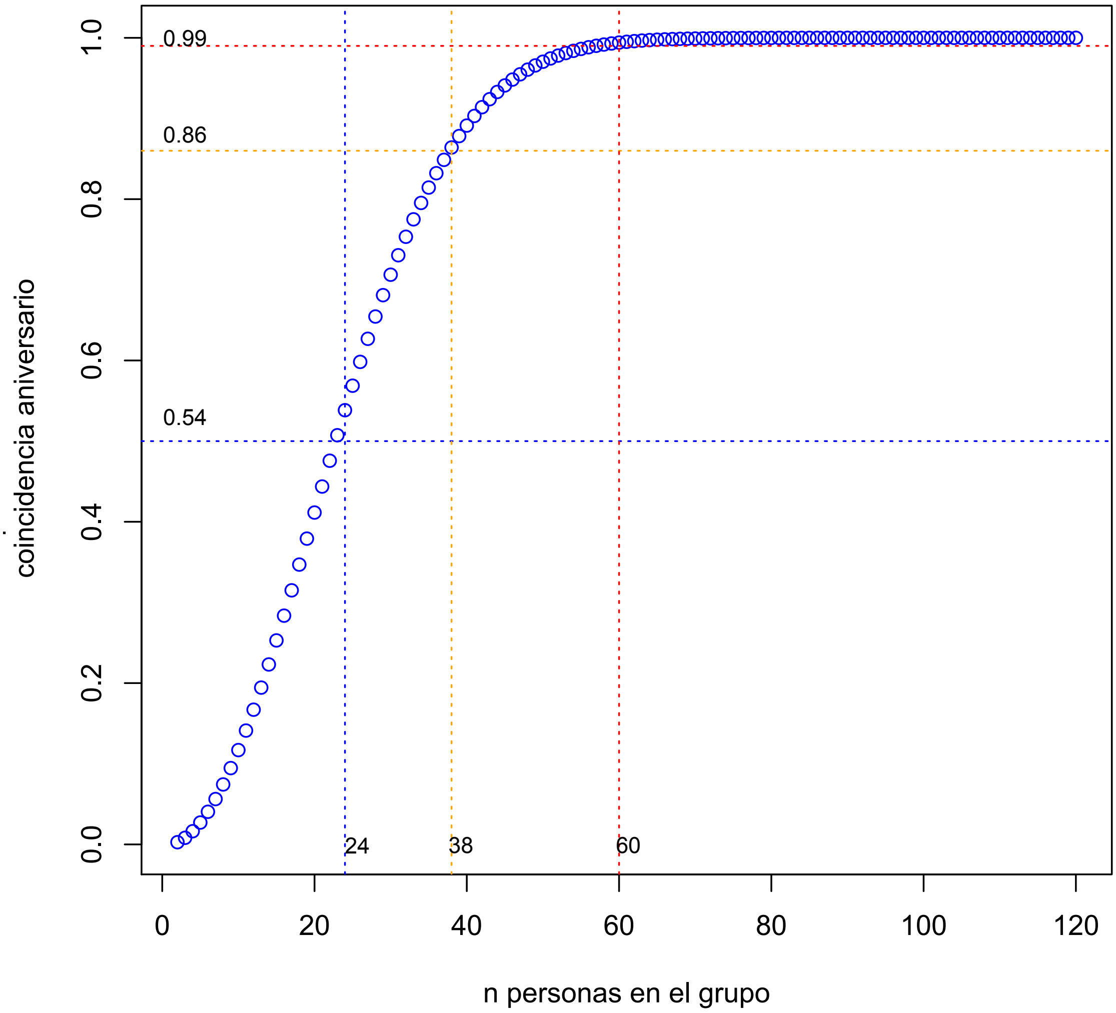

Figure 4: Probabilidad de coincidencia de aniversario en función del tamaño n del grupo

4. *Detección de drogas, profiling y falsos positivos*

En el ejemplo que sigue, y que está tomado con leves modificaciones del extraordinario libro de Ward Farnsworth, titulado *The Legal Analyst. A Toolkit For Thinking About The Law* (The University of Chicago Press, 2007, págs. 281 y ss.), se plantea esta cuestión:

En el Dulles Int'l Airport, el aeropuerto internacional de Washington D.C. (USA), la policía aduanera utiliza perros Beagle para detectar la presencia de drogas en el equipaje o en las ropas de pasajeros procedentes de determinados orígenes geográficos. En particular, la policía sabe que uno de cada cien pasajeros procedentes de una determinada capital lleva drogas, llamémosle *mule*.

La estadística nos dice que, si un pasajero lleva drogas, la probabilidad de que el olfato de un perro Beagle (500 veces más sensible que el ofato humano) que se llama Merlín las detecte es uno, cien por cien. Beagle está entrenado a ladrar cuando detecta drogas: las huele y entonces ladra, avisa.

Sin embargo, si el pasajero no lleva drogas, el Beagle Merlín normalmente no ladrará, pero lo hará en el 10% de estos; es decir, su tasa de acierto es de un noventa por ciento en los pasajeros que no llevan droga.

Entonces, en una ocasión, en un vuelo de 100 pasajeros, uno de sus pasajeros, John Smith, sale del avión, pasa por el control de Merlín, el perro le olfatea y ladra indicando así la probabilidad de que Smith sea un *mule* y lleve drogas.

Albert, ¿Cuál es la probabilidad de que Smith lleve drogas suponiendo que el Beagle Merlín le ha ladrado? ¿es, acáso, un noventa por ciento?

Antes de entrar en la respuesta a la pregunta, me veo obligado (responde Albert) a introducir el concepto de probabilidad condicionada. Ahora, como en todos los casos de aplicación de técnicas analíticas a conjuntos de datos, tenemos dos sucesos, pongamos los sucesos A y B, que un sujeto tenga una determinada carácterística y que el análisis al cual se le somete dé positivo en relación con la existencia de tal característica. Por ejemplo, un sujeto (el señor JX) se somete a un test concreto para detectar si tiene una enfermedad grave concreta. Hay dos sucesos relevantes:

A = "el sujeto JX tiene la enfermedad"

B = "el sujeto JX da positivo en el test"

Una probabilidad relevante es la probabilidad que JX tenga la enfermedad cuando se sabe que ha dado positivo en el test. Dicha probabilidad la llamamos probabilidad de A condicionada a B escrita $P(A \mid B)$.

Siempre hay que hacer notar que $P(A \mid B)$ no es lo mismo que $P(A)$; esta última es la probabilidad que JX tenga la enfermedad cuando no se sabe que ha dado positivo en test o análisis practicado. Más complicado aún: $P(A \mid B)$ no es lo mismo que $P(B \mid A)$, pues la segunda es la probabilidad que el test dé positivo si JX tiene la enfermedad.

Conviene recordar el teorema de Bayes:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})}$$

La fórmula dice que la probabilidad $P(A \mid B)$ se puede obtener de las probabilidades sin condicionar $P(A)$ y $P(\overline{A})$ y las probabilidades condicionadas $P(B \mid A)$ y $P(B \mid \overline{A})$.

Falsos positivos y, de nuevo, la importancia del parámetro de prevalencia (denominado

también *"base rate"*)

Supongamos que:

M = Merlín ladra ante la presencia de droga

$\overline{M}$ = Merlín no ladra la presencia de droga

D = El Sr. Smith lleva droga

$\overline{D}$ = El Sr. Smith no lleva droga

Los cálculos. En el supuesto que Merlin ladre, aplicando el teorema de Bayes, la probabilidad que el Sr. Smith lleve droga es:

$$P(D \mid M) = \frac{P(M \mid D) \times P(D)}{P(M \mid D) \times P(D) + P(M \mid \overline{D}) \times P(\overline{D})}$$

$$= \frac{1 \times 0{,}01}{1 \times 0{,}01 + 0{,}1 \times 0{,}99}$$

$$\approx 1/11$$

Afortunadamente para el Sr. Smith, la policía concluye que la probabilidad que lleve droga es baja, por debajo del 10%, aunque en cualquier caso no es tan baja como para descartar de entrada que no lleva droga. Una inspección adicional no vendrá de más, pero seria un error tremendo dar al Sr. Smith el trato de inculpado (y estas cosas ocurren).

En el cálculo hemos efectuado la inversión de probabilidad condicionada, hemos calculado $P(D \mid M)$ a partir de $P(M \mid D)$. Pues, ¡sí!, sorprendentemente, las dos probabilidades condicionales no solamente no son iguales, sino que pueden ser muy distintas. La fórmula que efectúa el *flip* de estas dos probabilidades condicionadas se conoce como el Teorema de Bayes (Reverendo Thomas Bayes, 1701-1761, Inglaterra), uno de los resultados más importantes del cálculo de probabilidades, y en la base de algoritmos estadísticos implementados por doquier en nuestra organización moderna (por ejemplo, en el servicio de urgencias de un gran hospital, el algoritmo que determina la prioridad de atención de un paciente recién llegado al servicio).

La confusión de la igualdad de las dos probabilidades condicionadas anteriores, está también en la base de la Falacia del Fiscal que discutimos más adelante.

Otra forma de hacer los cálculos

El valor $\frac{1}{11}$ obtenido anteriormente, lo podíamos haber conseguido de esta otra forma. De 1000 pasajeros, 10 lleva droga. Merlín ladrará con seguridad en estos 10 pasajeros (ya que llevan droga) y ladrará también en 99 (el 10% de los 990 restantes) pasajeros que no llevan droga. Es decir Merlín ladra en 109= 10 + 99 pasajeros; el Sr. Smith es uno de ellos. Cual es la probabilidad que sea de los 10 que llevan droga? La Regla de Laplace (casos favorables partido por casos posibles) nos conduce a la probabilidad de

$$\frac{10}{109} \approx \frac{1}{11}$$

el mismo valor que nos daba la aplicación del Teorema de Bayes.

La importancia del parámetro de Prevalencia (*"base rate"*)

Supongamos que el parámetro de prevalencia disminuye, y en lugar 1 de cada 100, suponemos que es 1 de cada 1000. ¿Cómo afecta esto a la probabilidad $P(D \mid M)$ (la probabilidad de que el pasajero lleve droga suponiendo que Merlin ladre)?

Véamos

$$P(D \mid M) = \frac{P(M \mid D) \times P(D)}{P(M \mid D) \times P(D) + P(M \mid \overline{D}) \times P(\overline{D})}$$

$$= \frac{1 \times 0{,}001}{1 \times 0{,}001 + 0{,}1 \times 0{,}999} = 0{,}0099$$

una probabilidad menor que 1 de cada 100, a pesar que Merlín a señalado ladrando al Sr. Smith. En este caso, es altamente probable que la señal de ladrar de Merlín sea un falso positivo.

Efectivamente, el valor de prevalencia es un parámetro fundamental del problema. Si es muy baja, Merlín ayuda muy poco; es por ello que utilizaremos a Merlín solamente cuando el avión provenga de zona de riesgo y cuando el parámetro de prevalencia (la probabilidad que un pasajero tomado al azar lleve droga) no sea demasiado baja.

Las consideraciones anteriores son muy relevantes a la hora de diseñar políticas regulatorias y, sobre todo, a la de aplicar las regulaciones de que se trate: hay circunstancias que aconsejan un segundo examen, un nuevo control, pero que claramente no permiten una detención, un encarcelamiento preventivo o una ilación de culpabilidad. Para evitar el desastre de un prejuicio equivocado, **conviene suspender el juicio y trabajar más y mejor**: realizar nuevos análisis, idealmente, con técnicas independientes que permitan afinar mejor y descartar falsos positivos. Ni los ciudadanos, ni los funcionarios encargados de la aplicación de la ley, ni los jueces, que tienen encomendada la tarea de velar por la corrección de su aplicación, son siempre conscientes de tal problema. La regla de oro es, en todo caso, **suspender el juicio y trabajar más**.

5. *Identificación de un criminal por el grupo sanguíneo*

Supongamos (dice Pablo) que el sospechoso, quien fue visto en las proximidades del lugar del crimen al poco de haber ocurrido este, tiene el mismo grupo sanguíneo que el criminal, el cual, arañado por la víctima, había dejado rastros de su sangre en el cuerpo de esta última. ¿Bastan esos indicios? ¿Qué indican?

En relación al sospechoso (escribe Albert), consideremos los siguientes sucesos:

- A+ ="ser de grupo sanguíneo A+"
- B = "ser inocente"

Supongamos las siguientes probabilidades:

1. $P(A+ \mid B) = 0$, probabilidad de grupo sanguineo A+, cuando se es inocente

2. $P(A+ \mid \overline{B}) = 1$, probabilidad de ser del grupo sanguineo A+ en caso de criminal ($\overline{B}$, no inocente)

3. $P(B) = a$, la probabilidad de ser inocente sin consideraciones de grupo sanguineo

4. $P(B \mid A+)$ la probabilidad de ser inocente a partir de las pruebas (coincidencia de grupo sanguineo con el del criminal); dicha probabilidad es

$$P(B \mid A+) = \frac{P(A+ \mid B)P(B)}{P(A+ \mid B)P(B) + P(A+ \mid \overline{B})P(\overline{B})}$$

$$= \frac{.1 \times a}{.1 \times a + 1 \times (1-a)}$$

Si a = .4 (el 40%)

$$P(B \mid A+) = \frac{.1 \times .4}{.1 \times .4 + 1 \times (1-0.4)} = 0.0625$$

De manera que la probabilidad de culpable es $P(\overline{B} \mid A+) = 1 - 0.0625 = 0.938$; 94% de probabilidades que el sospechoso sea culpable!, el hallazgo de restos de sangre del grupo sanguíneo $A+$ ha disparado la probabilidad de criminal.

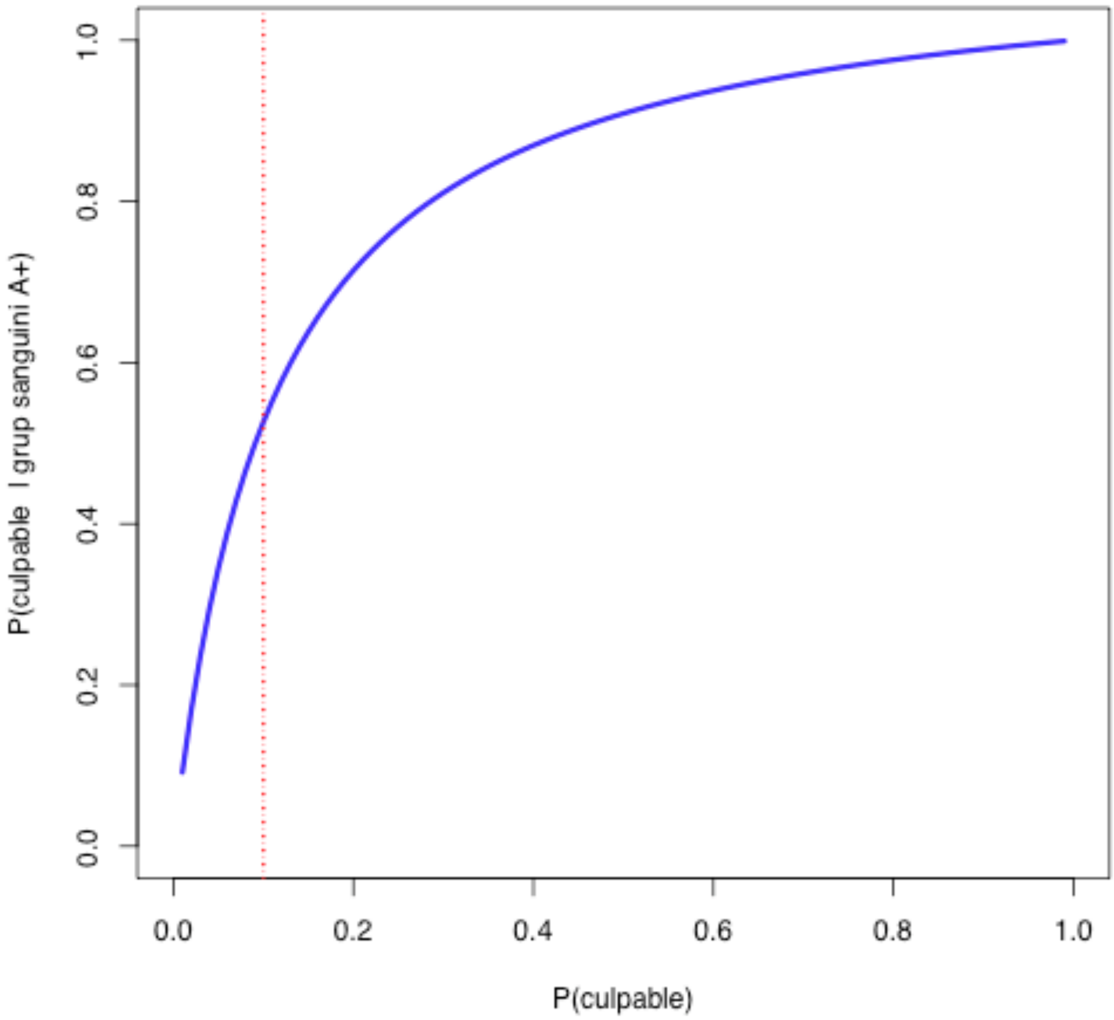

Figure 5: Probabilidad de que el sospechoso sea culpable bajo diferentes valores de la "base rate" a (la probabilidad a priori de culpabilidad)

La coincidencia de grupo sanguíneo del sospechoso altera la probabilidad de que el sospechoso sea culpable, pero solamente en el caso que la probabilidad a priori de ser culpable (parámetro de prevalencia) no sea muy baja. Ello se ilustra en la Figura 5 que muestra en el eje de ordenadas la probabilidad a posteriori (después de observar la coincidencia en grupo sanguíneo) para diferentes valores de la probabilidad a priori de culpabilidad.

Dichos cálculos se podrían aplicar al caso del ciudadano holandés residente en España Romano van der Dussen (Romano vdD), quien pasó "11 años encarcelado pese a las pruebas de ADN" (El País, 10 de mayo, 2015. Extensamente: "Falso culpable", El País Semanal, núm. 2058, 6 de marzo de 2016, portada y págs 32 a 39) por tres intentos de violación ocurridos en una misma noche en Fuengirola, Malaga, el 10 de agosto de 2003. Posteriormente, las pruebas de ADN involucraron a un británico, Mark Philip Dixie (Mark PD), preso en Reino Unido tras haber sido condenado por asesinato y violación y que residía en Málaga cuando se cometieron los hechos por los que se había condenado a Romano vdD. Con los cálculos apuntados se puede discernir las probabilidades de autoría de los dos implicados, Romano vdD. vs. Mark PD. Los sucesos relevantes son los siguientes:

A= "Coincidencia de secuencia de DNA del sospechoso en semen del lugar del crimen"

B= "Sospechoso es inocente"

Para construir la Figura 7 que pasamos a comentar, las probabilidades relevantes són las siguientes: p = P(B), la probabilidad de inocencia sin tener en cuenta la información relativa al DNA; $P(\overline{A} \mid B)$, la probabilidad de que no haya coincidencia de DNA cuando se es inocente (dicha probabilidad parece razonable suponerla alta, en el gráfico la fijamos a 0.99); $P(\overline{A} \mid \overline{B})$, la probabilidad de que no haya coincidencia de DNA cuando se es culpable (en el gráfico es el parámetro a); $P(A \mid B)$, la probabilidad de coincidencia de DNA cuando se es inocente (en el gráfico es el parámetro d); finalmente, $P(A \mid \overline{B})$, la probabilidad de que haya coincidencia de DNA cuando se es culpable (en el gráfico está fijada a 0.7).

Aplicando Bayes, en el caso de Romano vdD (en el que no se observa coincidencia de DNA), la probabilidad a posteriori de inocencia será:

$$P(B \mid \overline{A}) = \frac{P(\overline{A} \mid B)P(B)}{P(\overline{A} \mid B)P(B) + P(\overline{A} \mid \overline{B})P(\overline{B})} = \frac{0.99p}{0.99p + a(1-p)}$$

que varia con p y a. En el caso de Mark PD (en el que sí se observa coincidencia de DNA), la probabilidad a posteriori de inocencia será:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})} = \frac{dp}{dp + .7(1-p)}$$

que varia con p y d. Para diferentes valores de los parámetros a y d, la Figura 6 muestra las curvas de probabilidad $P(B \mid \overline{A})$ y $P(B \mid A)$. Dichas curvas, en azul las de Romano vdD y en rojo las de Mark PD, son las probabilidad de inocencia a posteriori de Romano vdD y Mark PD a la vista de la información sobre DNA obtenida, para diferentes valores de los parámetros a y d.

Dicha Figura 6 muestra que bajo diferentes magnitudes de los parámetros que no conocemos con exactitud, los parámetros a y d, la información sobre coincidencia o no en DNA inducen una probabilidad de inocencia muy alta para Romano vdD, y una probabilidad de inocencia muy baja para Mark PD, salvo que hubiese una probabilidad a priori muy baja de inocencia para Romano vdD, o una probabilidad a priori de inocencia muy alta para Mark PD.

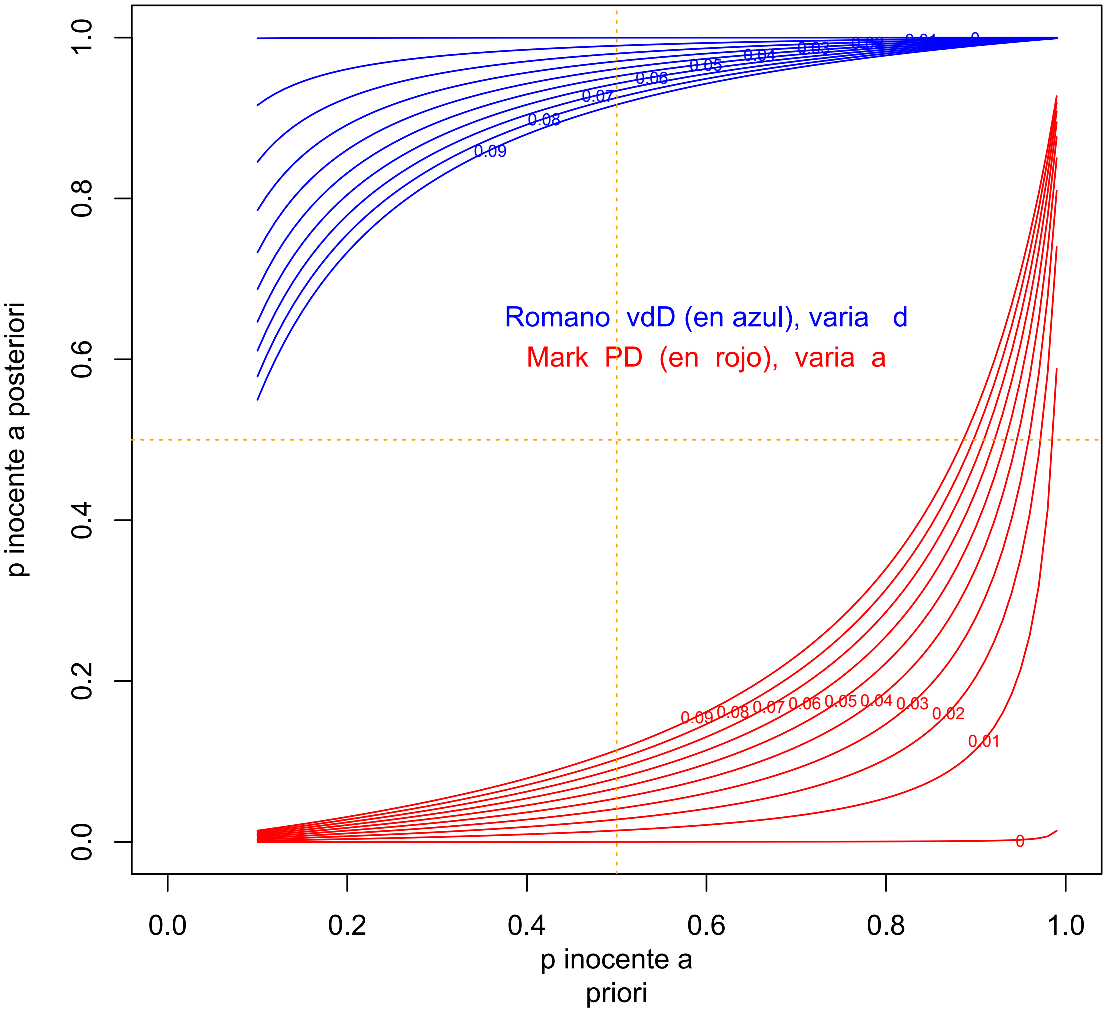

Figura 6: Probabilidad de inocencia a posteriori de Romano vdD (en azul) y de Mark PD (en rojo) bajo diferentes valores de los parámetros d (probabilidad de coincidencia del DNA cuando se es culpable) y a (probabilidad de coincidencia en DNA cuando se es inocente).

El caso de Romano van der Dussen muestra bien a las claras los costes de los falsos positivos y, en particular, los efectos devastadores de la superposición de malas prácticas policiales y judiciales posibilitadas -cuando no incentivadas- por el horror de los encargados de aplicar el derecho a los falsos negativos.

Es sabido, desde hace mucho tiempo, que las identificaciones de las víctimas, las ruedas de reconocimiento, la confianza en una única prueba, la declaración de la víctima –cuando no concurren en ella circunstancias anteriores indiciarias de posible falsedad, es verosímil, reiteradamente coherente y prestada sin fisuras ni vacilaciones- y el canon de la valoración conjunta (no pormenorizada) de las pruebas generan falsos positivos.

Pero tal no es el método de una investigación empírica. Por decirlo de manera clara: tanto el análisis de los datos como las conclusiones obtenidas por el analista han de poder reproducirse por otros analistas independientes[1]. Cualquier investigación policial y toda valoración judicial han de respetar el método científico básico -el elemental-. Mas, como veremos, esto no ocurrió en el caso: la compatibilidad de unos restos orgánicos encontrados en un objeto –un peine- de una de las víctimas con el DNA de quien muchos años más tarde reconocería su posible involucración en los hechos ya constaba en autos aunque nadie paró en mientes sobre las implicaciones del dato de hecho.

En las líneas que siguen extractaremos las dos resoluciones judiciales básicas sobre el caso de Romano van der Dussen. Por supuesto, recomendamos a los lectores escépticos la lectura del texto integral de las sentencias de que se trata, pues nuestra selección puede estar afectada por el sesgo de la retrospección (*hindsight bias*) y, de hecho, por nuestra apreciación de las buenas –y malas- prácticas en investigaciones empíricas. De todos modos, las cosas hablan por sí mismas:

La sentencia 353/2005, de 25 de mayo, de la Sección 2ª de la Audiencia Provincial de Málaga (JUR 2005\195477, ponente Lourdes García Ortiz), que condenó al acusado Romano van der Dussen por tres delitos de agresión sexual, dos de lesiones y dos de robo con violencia, dejó escrito en su Fundamento de Derecho Segundo:

"[A]unque el acusado negó haber atacado a chica alguna e incluso negó su presencia en Fuengirola el día de los hechos, alegando que se encontraba con unos amigos en Torremolinos, en ningún momento ha dado pie a concretar a lo largo de la investigación ni en el acto del juicio su versión exculpatoria".

"[L]a testigo [y víctima] Amparo [nombres supuestos en todos los casos] manifestó que oyó gemidos de auxilio, que vio a un hombre de raza blanca, fuerte, de 1,75 cms. de estatura, 30 años, pelo claro que se detuvo unos 10 o 15 minutos. Se ratificó en sus anteriores declaraciones y reconoció en el acto del juicio al acusado sin duda alguna como el hombre al que vio ese día en las inmediaciones de su domicilio … En sus declaraciones policial y judicial detalló que las inmediaciones del lugar no había nadie más". "Declaró" también "que se encontraba en la CALLE000 y que alguien se le abalanzó y la tiró al suelo. No vio a la persona que lo hizo, ni recordaba que le quitara las bragas, rompiendo a llorar en ese momento. A continuación añadió que tiene amnesia, y que no recordaba lo que le dijo a la policía".

"Concepción [otra de las tres víctimas] … reconoció al agresor en su día sin ninguna duda y en el plenario también lo reconoció cuando sin ser vista pudo verlo, insistiendo en que era él, que no tenía duda y mostrando signos de una gran afectación ante la visión del acusado, añadió que era un "hijo de perra" y que "le lleven al infierno", pidiendo después perdón por sus palabras".

"La testigo [y víctima] Sara declaró también, como las anteriores, sin ser vista … que reconoció la camiseta y el pantalón del agresor y a este. Cuando le reconoció en el plenario, sin ser vista, se echó a llorar y manifestó que sin duda era la persona que intentó abusar de ella y además le quitó la cartera".

"Estas testificales, tan reveladoras y determinantes para considerar desvirtuada la presunción de inocencia que asiste al acusado en los tres supuestos fácticos descritos en el relato de hechos

[1] Clara Atienza Jiménez, graduada en criminología, estudiante de derecho y colaboradora del área de derecho civil de la facultad de derecho de la UPF ha revisado independientemente el texto que sigue y ha sugerido algunas modificaciones que hemos introducido. Le damos las más sinceras gracias.

probados aparecen corroborados por datos objetivos[:]

“[A]sí el dato más importante es el reconocimiento por parte de Amparo…”

“[Otra de las víctimas] no lo ha podido identificar y en el plenario manifestó que no recordaba lo que había manifestado a la policía, constando en el atestado policial que les dio unos datos de su agresor coincidentes en gran medida con los aportados por el resto de los testigos …”

“El ataque contra su libertad sexual [la de Amparo] no nos ofrece duda alguna en la medida de que las lesiones que presentaba y el estado de su ropa revelan claramente que el agresor la agredió con ánimo de satisfacer sus deseos sexuales, empleando para ello no solo la fuerza sino una especial agresividad, brutalidad y crueldad”.

“Esa particular forma de proceder la puso en práctica esa misma madrugada en tres ocasiones, pues en menos de dos horas atacó también se similar manera a Concepción y a Sara. Ellas no han dudado en ningún momento, cuando una vez más lo reconocieron en el plenario, de que fue el acusado y no otro, el que las sometió con su brutalidad a la situación de terror que con su forma de agredirlas les produjo …”

“Los informes médicos de las tres víctimas son ciertamtente reveladores del tipo de agresión que sufrieron y constituyen datos objetivos que vienen a corroborar de manera determinante lo que las tres tuvieron que soportar, de tal manera que contamos con suficientes pruebas directas, en el caso de Concepción y Sara [dos de las víctimas] y con suficientes datos fácticos o indicios acreditados, en el caso de Amparo [tercera víctima] de la comisión por parte del acusado de los delitos de que ha sido acusado [agresiones sexuales y lesiones] y por ello debe responder en concepto de autor de cada uno de ellos…”

***

Casi once años después, la sentencia 75/2016, de 10 de febrero, de la Sala Segunda del Tribunal Supremo (Id Cendoj: 28079120012016100029, ponente Andrés Martínez Arrieta) dice en su Fundamento de Derecho Segundo:

“[O]bligado resulta estimar el recurso [de revisión], que cuenta con el apoyo del Ministerio Fiscal, ya que existen nuevos elementos de prueba que permiten acreditar de modo indubitado la inocencia del acusado respecto del delito de agresión sexual y de lesiones cometido contra Clemencia [nombres siempre supuestos: primera víctima denominada “Amparo” en la sentencia de la Audiencia Provincial de Málaga antes extractada]…”

“La [Unidad Central de Análisis Científicos de la] Comisaría General de Policía Científica … concluyó que en los restos celulares de un peine de carey se habían obtenido dos perfiles genéticos compatibles con los correspondientes a Clemencia y a Benito [Mark PD]. Además Benito [Mark PD] en su manifestación prestada en la prisión de Frankland ante la Abogada en Holanda del promovente [Romano vdD] admitió su posible implicación en el delito cometido el día 10 de agosto de 2003 en la calle Miguel Bueno de Fuengirola”.

“Consecuentemente los resultados de estas nuevas pruebas genéticas revelan datos nuevos y posteriores a la sentencia y la prueba de ADN tiene un carácter técnico e identificador de superior valor a las pruebas en que la sentencia condenatoria cuya revisión se solicita se basó…”

La sentencia del Tribunal Supremo declaró la nulidad parcial de la sentencia dictada por la Audiencia de Málaga anteriormente extractada y acordó que el Juzgado de Instrucción que corresponda instruyera de nuevo la causa. Romano vdD fue liberado.

En el informe de la Unidad Central de Análisis Científicos de la Comisaría General de la Policía Científica se lee (Antecedente 3 de la sentencia del tribunal Supremo):

"En los restos celulares de un peine de carey … se ha obtenido una mezcla de al menos, dos perfiles genéticos, compatible para los marcadores genéticos analizados, con el perfile de Clemencia [una de las tres víctimas] y el de Benito [Mark PD]. La evaluación estadística se expresa en forma de Coeficiente de Verosimilitud o LR, que, en este caso, tiene un valor de 11 157.595 751.396 100.000.000.000. Esto significa que es algo más de once cuatrillones de veces más probable que la mezcla obtenida de los restos celulares del peine de carey presente estos genotipos si [se] ha producido por Clemencia y Benito que si se ha producido por dos personas cualesquiera escogidas al azar de la población española".

***

Concluimos con un breve resumen de datos extractados del Antecedente 2 de la Sentencia del Tribunal Supremo:

En 14.2.2012, el Tribunal Supremo (Sala Segunda) había acordado no autorizar a la representación de Romano vdD interponer recurso de revisión. En 12.11.2015, el Tribunal volvió sobre sus pasos y autorizó la interposición del recurso.

En 23.3.2007, el Jefe de la Unidad Central de Análisis Científicos de la Dirección General de la Policía y de la Guardia Civil, Comisaría General de la Policía Científica, de la Unidad Central de Análisis Científicos había informado que los perfiles genéticos de la víctima Clemencia [una de las victimas] y de Benito [Mark PD] eran compatibles con la mezcla de perfiles "que se obtuvo en la muestra 7.3.2 (restos orgánicos adheridos a un peine) del informe NUM000 y obtenida con ocasión de la agresión con lesiones graves y violación a Clemencia. Al folio (1.101) consta ampliación del Informe Pericial nº NUM000 de fecha 22 de junio de 2004, en le cual exponen "que aunque no conste en el informe pericial de referencia NUM000 de fecha 22 de junio de 2004, el perfil de Mariano [Romano vdD], fue cotejado con la mezcla obtenida de los restos orgánicos adheridos al peine (muestra 7.3.2), NO SIENDO COMPATIBLE CON LA MISMA [énfasis en el original]. **Obra en la causa, folios 1.107 a 1.116 del testimonio presentado, informe de Pericial de 22-06-2004 nº NUM000** [énfasis añadido], e informe pericial de 23 de marzo de 2007". El informe mencionado en este párrafo señalaba también que:

En 8.11. 2006 la OCN-INTERPOL España "difundía un mensaje de INTERPOL-LONDRES, donde se afirmaba que el ciudadano británico Benito [Mark PD] había sido detenido y acusado en el Reino Unido de la violación y homicidio de una joven británica, igualmente afirmaba que dicha persona había residido en el sur de España, concretamente en la zona de Fuengirola … desde finales de 2002 hasta octubre de 2003".

6. *Falacia del fiscal: People v. Collins (68 Cal. 2d 319 (1968))*

Recordemos, al inicio de este epígrafe, que una de las víctimas en el caso de Romano van der Dussen había declarado -citamos de la sentencia de la Audiencia de Málaga- "que oyó gemidos de auxilio, que vio a un hombre de raza blanca, fuerte, de 1,75 cms. de estatura, 30 años, pelo claro que se detuvo unos 10 o 15 minutos", a quien luego reconoció en el acto del juicio como el acusado, a pesar de que padecía de amnesia en relación con el hecho concreto de las agresiones indudablemente sufridas.

A veces, escribe Ward Farnsworth, en su libro citado, *The Legal Analyst* (pp. 276 y ss) se habla de falacia del fiscal (*Prosecutor's fallacy*) para ilustrar lo que ocurrió en un caso famoso, *People v. Collins*: en la ciudad de Los Angeles, una mujer fue víctima de un tirón: había sido empujada en la calle y le habían robado el bolso. En el juicio que siguió y en el que fueron acusados un hombre y su pareja, un testigo ocular declaró que había visto a la presunta ladrona huir de la escena del crimen, que esta era una mujer rubia peinada con una cola de caballo y que, siguió diciendo el testigo, se había subido a un automóvil de color amarillo, conducido por un hombre afroamericano con barba y bigote.

En la noche del mismo día del robo, Malcolm Ricardo Collins y su mujer Janet Collins fueron arrestados bajo acusación de robo. Ella trabajaba como criada en San Pedro, era rubia, peinaba cola de caballo y su marido, que era afro-americano con bigote y barba, conducía un coche amarillo. No tenían mucho dinero, ingresaron al instante en la categoría de sospechosos y, al poco, en la de los acusados.

En el juicio, el fiscal interrogó a un perito, un profesor de matemáticas, sobre las consecuencias de aplicar la regla del producto al caso enjuiciado. El fiscal suministró al profesor las siguientes asunciones acerca de las probabilidades de cada característica o rasgo de la descripción del testigo:

La probabilidad de que un automóvil fuera amarillo era $\frac{1}{10}$.

La de que un hombre llevara bigote, $\frac{1}{10}$.

La de que llevara barba, $\frac{1}{10}$.

La de que una mujer se recogiera el pelo con una cola de caballo, $\frac{1}{10}$.

La de que una mujer fuera rubia, $\frac{1}{3}$.

La de que un hombre afroamericano estuviera en un coche acompañado por una mujer blanca, $\frac{1}{100}$.

El fiscal y el profesor explicaron al jurado que las probabilidades de que todas estas circunstancias coincidieran en un inocente eran muy reducidas, pero podían ser estimadas multiplicando sus probabilidades respectivas (regla del producto) y que el resultado del producto era 1 entre 12.000.000. Había, pues, solo una posibilidad entre bastantes millones de que los acusados, que reunían todas estas características, fueran inocentes. Por ello, concluyeron, debían ser considerados culpables mucho más allá de toda duda razonable. El jurado estuvo de acuerdo con las apreciaciones anteriores y dictó un veredicto de culpabilidad.

Recurrida la sentencia, sigue escribiendo Farnsworth, el Tribunal Supremo de California anuló la decisión.

¿Por qué que hizo eso, cuando todo pareceía indicar que los acusados tenían en su contra bastantes indicios claros?

Estamos ante el caso típico de aplicación errónea del cálculo de probabilidades. El ejemplo que se comenta involucra varios errores, uno de ellos es el que se conoce como la Falacia del Fiscal, que ahora conviene explicar.

7. *Falacia del Fiscal y confusión de probabilidades*

Suponemos

- A: Una coincidencia rara, $P(A) \approx 0$ (uno sobre mil, o sobre un millón, etc.)
- B: el sospechoso es inocente

Aparecen cuatro probabilidades:

1. $P(B)$, la probabilidad que el sospechoso sea inocente
2. $P(A)$, probabilidad de la coincidencia rara, por su propia rareza, dicha probabilidad es un número muy pequeño
3. $P(A \mid B)$, probabilidad de coincidencia rara si se es inocente
4. $P(B \mid A)$, probabilidad de ser inocente si se produce la coincidencia rara

La falacia se produce como resultado de la siguiente cadena desafortunada de igualdades, cada una de ellas errónea:

$$P(B) = P(B \mid A) = P(A \mid B) = P(A)$$

Como A es una coincidència rara, $P(A) \approx 0$, y como de la confusión de probabilidades se ha concluido que $P(B) = P(A)$, el fiscal concluye, erróneamente, que la probabilidad de inocencia $P(B) = P(A)$es muy pequeña; por tanto, por la ley del complementario, la probabilidad $P(\overline{B})$ de que el sospecho sea culpable será el complementario sobre 1 de un número muy pequeño,

$$P(\overline{B}) = 1 - P(B) = 1 - P(A) \approx 1$$

Bajo el efecto de la Falacia del Fiscal, se concluye una probabilidad de culpabilidad muy alta para el sospechoso, una probabilidad próxima a 1.

En el ejemplo descrito de People v. Collins (California, 1968).

B ="Los Collins son culpables".

A="mujer rubia, peinada con cola de caballo, con marido afro-americano con bigote y barba, coche de color amarillo".

El perito realizó los cálculos de $P(A)$ utilizando la regla del producto, multiplicó las probabilidades de cada una de las coincidencias, llegando a la conclusión que $P(A)$ era muy pequeña; por tanto, la probabilidad de inocencia de los Collins, $P(B)=P(A)$ también era muy pequeña.

En este ejemplo, no solamente se da el error de la Falacia del Fiscal, la igualdad $P(B)=P(A)$, sino que además se incurre en el error de calcular la probabilidad de coincidencia $P(A)$ utilizando la regla del producto, una utilización errónea ya que, en este caso, no se cumple el supuesto de independencia de las coincidéncias elementales (llevar bigote no es independiente de llevar barba, por ejemplo).

Otro error cometido en el caso, fue no haber tenido en cuenta que la búsqueda de presuntos culpables se efectuaba en un colectivo muy grande, la ciudad de Los Angeles. Como se ha visto en anteriores discusiones, aunque la probabilidad P(A) de coincidencia descrita sea baja, la probabilidad que hayan dos parejas con estás características en toda la ciudad de Los Angeles no es baja, puede ser muy alta.

8. *Falacia del Fiscal y sucesos independientes: el caso de Sally Clark*

En Inglaterra, la abogada Sally Clark fue juzgada y condenada, en 1999, por el homicidio de dos de sus hijos. Uno de ellos había muerto un año antes que el otro, al parecer, de muerte súbita.

El fiscal presentó a un perito ante el jurado, un distinguido consultor pediátrico, Sir Roy Meadow, quien declaró que había una probabilidad contra 73 millones de dos muertes súbitas consecutivas.

En 1999, Clark fue condenada a cadena perpetua y la sentencia fue confirmada en una primera apelación.

Más tarde la Royal Statistical Society declaró que la condena no tenía base estadística. Sally Clark fue exonerada y excarcelada en 2003. Pero su vida estaba ya tronchada: moriría en 2007 víctima de la tristeza y de una intoxicación de alcohol (los lectores interesados pueden consultar: *[2003] EWCA Crim 1020*.

Albert (pregunta Pablo), ¿por qué la Sociedad de Estadística británica afirmó que la condena estaba estadísticamente injustificada?

Meadow había escrito recientement sobre *SIDS* (*sudden infant death syndrome*), calculado las probabilidades en función de características de si la madre era fumadora, de nivels económicos bajo, madre prematura, etc. Sally Clark no fumaba, tenía una posición social acomodada, y su edad superaba los 27 años, de manera que determinó que estadísticamente la probabilidad de muerte súbita en este caso era de 1 sobre 8543. Pero Meadow no se paró ahí. Especuló sobre la probabilidad de dos muertes súbitas en poco más de un año. Para ello elevó al cuadrado la probabilidad anterior (¡supuso independencia de probabilidades!) obteniendo una probabilidad de 1.37 sobre 100 millones. Meadow escribió

“In England, Wales and Scotland, there are about say 700,000 births a year, so it is saying by change that happening will occur about once every hundred years".

Era una probabilidad pequeñisima, *ergo* no había otra explicación razonable que el crimen. De nuevo nos encontramos con la Falacia del Fiscal, si B="Sally Clark es inocente", A="dos muertes súbitas"

$$P(A) = P(B)$$

de manera que la probabilidad de Sally Clark culpable,

$$P(\overline{B}) = 1 - P(B) \approx 1$$

En este caso, la probabilidad de dos muertes subitas en la familia $P(A_1 \cap A_2)$ se había resuelto mediante la regla del producto para sucesos independientes, es decir el

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

Sin tener en cuenta la regla del producto más general (para sucesos dependientes) que seria

$$P(A_1 \cap A_2) = P(A_1 \mid A_2)P(A_2)$$

Dicho cálculo sería más complicado ya que probabilidad de la segunda muerte súbita, $P(A_2 \mid A_1)$, puede ser muy distinta de $P(A)$. Admitir la posibilidad de dependencia, hace posible un efecto genético en esta enfermedad, que Meadow ignoró por completo. Resolver la probabilidad de la intersección como producto es válido solamete en el caso de sucesos independientes, es decir $P(A \cap B) = P(A \mid B)P(B)$ vale solamente en el caso de A y B independientes.

9. *El catedrático y abogado Alan Dershowitz y el caso de O.J. Simpson.*

Alan Dershowitz (1938) fue –a los 28 años de edad–el catedrático de Harvard más joven de la historia de su facultad de derecho. Como abogado, ganó 13 de los 15 casos de asesinato y tentativa de asesinato que llevó o en cuya defensa participó. Dershowitz, quien todavía vive, ha sido, pues, una estrella en dos galaxias, en la académica y en la profesional.

En el (primer) caso de O.J. Simpson, un proceso penal finalizado en 1995, Dershowitz asesoró a la defensa de Simpson, un antiguo jugador profesional de fútbol americano que fue acusado de haber asesinado a su ex mujer y a un camarero. A la postre, Simpson obtuvo un veredicto de no culpabilidad por parte de un jurado étnicamente dividido.

En el juicio se discutía, entre otras cosas, sobre la relevancia de evidencia anterior de maltrato por parte de Simpson.

Como de seguro saben los lectores, la prueba de la conducta anterior de una persona acusada de un crimen, puede ser con frecuencia (en la mayoría de los sistemas legales democráticos) y por razones que ahora no cabe detallar, normativamente improcedente en el enjuiciamiento de hechos posteriores: si una persona, ya condenada por un homicidio veinte años atrás, es juzgada por un nuevo posible homicidio, un tribunal rechazará en principio todo intento del fiscal de presentar la evidencia anterior como evidencia del comportamiento que ahora se enjuicia. La idea intuitiva es que, en nuestros sistemas legales, se juzgan actos, no actores, pero, como digo, no corresponde ahora extenderse sobre esta cuestión. Parecidamente, el modo de obtención de las pruebas por parte de los funcionarios policiales es relevante: en el caso, la defensa incidió en este aspecto de la cuestión con éxito indudable. Por último, el lugar en el que tiene lugar el juicio y la procedencia del *pool* del que saldrá el jurado también es relevante. De nuevo, la defensa supo aprovechar esta circunstancia en el caso.

En el juicio penal, Simpson obtuvo un veredicto de no culpabilidad aunque luego sería condenado a pagar una elevada indemnización en un proceso civil posterior y, mucho después, sería juzgado y condenado en otro proceso penal por robo. Actualmente está en prisión.

En todo caso, algunos expertos en estadística no siempre ven las cosas como lo hacen las gentes de la ley: así como hay malos usos de la estadística por parte de los abogados, en ocasiones, el estadístico puede utilizar opinablemente las herramientas legales, o puede prescindir de ellas y sugerir que la evidencia de un homicidio anterior incrementa la probabilidad de que el antiguo homicida vuelva a matar, o la de que haya más homicidas entre los homicidas (o entre los maltratadores) que entre quienes no lo son y que esto ha de ser relevante en derecho, son cosas discutibles acaso, pero que conviene conocer.

¿Cómo verías tú el caso, Albert?

Tal como has comentado (responde Albert), Alan Dershowitz razonó que el historial de maltrato del acusado no era relevante en el juicio de asesinato de su mujer, pero, al parecer, razonó que no es que no fuera *normativamente* impertinente, sino que era *factualmente* irrelevante: Argumentó, en efecto, que a partir de las estadísticas de maltrato en años recientes en USA solamente una de 2500 mujeres maltratadas han acabado asesinadas por su maltratador.

Dershowitz se apoyó en la siguiente Estadística de Criminalidad (USA, 1992): población de mujeres = 125 millones; mujeres asesinadas = 4936; mujeres maltratadas = 3.5 millones; mujeres asesinadas por sus maltratadores = 1432.

Consideremos los siguientes sucesos:

B ="mujer maltratada"

M= "mujer asesinada"

C = "mujer asesinada por su maltratador"

A partir de las estadísticas de crimen y maltrato mencionadas, Alan Dershowitz obtiene $\frac{P(C)}{P(B)} = \frac{1432}{3.5 \times 10^6} = \frac{1}{2500}$, el dato que le lleva a la afirmación de que solamente una de cada 2500 mujeres es asesinada por su maltratador.

Dicho dato impresionó al jurado el cual, de nuevo, fue confundido por probabilidades mal calculadas. De hecho, el dato calculado por Alan Dershowitz no es una probabilidad, sino que es un cociente de dos probabilidades que no representa la probabilidad que engañosamente indica.

El cálculo correcto habría sido el siguiente. Hay dos hechos incontestables: la mujer sufría maltrato (B era cierto); la mujer fue asesinada (M era cierto). La probabilidad buscada era la probabilidad de C (mujer asesinado por su maltratador) a sabiendas de B y M. El cálculo correcto sería el de $P(C \mid B \cap M)$, que, en base a las estadísticas mencionadas es simple, se necesitaba calcular

$$P(C \mid B \cap M) = 1432/4936 \approx 0.30$$

Vemos que no es una probabilidad tan pequeña como engañosamente indicaba Dershowitz. De hecho, la probabilidad correcta, da visos de verosimilitud a que el maltratador, en este caso O.J. Simpson, fuese el criminal[1] (aquí hemos criticado a Alan M. Dershowitz, pero los lectores pueden contrastar su versión en su libro titulado *Reasonable Doubts: The Criminal Justice Systm and O.J. Simpson Case*, New York Simon and Schuster, 1996. Hay docenas de libros sobre este caso, accesibles en Amazon).

[1] El artículo Good, I.J. (1995) When batterer turns murderer. *Nature*, 375 ,269-326, cuestiona también la probabilidad calculada por Alan Dershowitzl

10. Big Data I: el caso de Lucia de B(erk)

Lucia de Berk, una enfermera que trabajaba un hospital infantil de La Haya, protagonizó uno de los casos de error judicial más notorios en la historia de Holanda: en 2003 fue condenada a cadena perpetua (sin posibilidad de libertad condicional) por siete asesinatos y tres tentativas de asesinato. Reabierto el caso en 2008 a instancias del Tribunal Supremo holandés, de Berk fue finalmente exonerada en 2010.

¿Qué había ocurrido?

En 2001, como consecuencia del fallecimiento inesperado de un bebé en el Juliana Kinderziekenhuis (Hospital Infantil de La Haya), muertes anteriores y resucitaciones cardiopulmmonares de otros recién nacidos fueron investigadas. A resultas de la investigación, nueve incidentes de este tipo que habían sido inicialmente consideradas casuales se recalificaron como sospechosos y el Hospital denunció a Lucía de Berk, la enfermera encargada de suministrar medicación a los bebés en la sección del Hospital donde habían tenido lugar los fallecimientos y las resucitaciones. En el juicio fue muy relevante la afirmación de un experto según la cual la probabibilidad de que, en la misma sección, se hubieran dado las coincidencias en cuestión por azar era de solo 1 contra 374.000.000.

Ahora, el jurista de entre nosotros (Pablo) pregunta al estadístico (Albert), cómo habría analizado este caso.

Este (responde Albert) es un tema complejo que conecta con los avances más recientes de la estadística, en especial con lo que se denomina *big data*. En este tema tiene relevancia la reflexión de (Richard Gill, March 2014)[1] sobre la frase de Mark Twain:

"If you tell the truth, you don't have to remember anything" (?)

Mas, como comenta Gill, esto no es totalmente cierto en estadística: si solo cuentas una parte de la verdad, puedes estar contando una gran mentira, pues hay que ver por qué fuiste a contar la verdad de solo una parte de los datos y cómo lo hiciste. Y es que el mensajero es parte del mensaje.

La acusación criminal contra Lucía de Berk se apoyó en la consulta a una base de datos muy grande sobre el desempeño de los profesionales sanitarios en grandes hospitales de Holanda. La base de datos incluía a Lucía de Berk. La consulta de dicha base de datos señalaba a Lucía de Berk. con una desviación significativa del patrón de rendimiento normal, en términos del número de muertes en el conjunto de sus casillas de actuación. Una casilla recoge el desempeño del profesional en una unidad determinada de tiempo y lugar, por ejemplo el turno de 7 a 9 de la noche en la planta de Pediatría del Hospital X, en el día D. Previo control estadístico de factores que inciden sobre el *output* de una casilla, se calibra el rendimiento del profesional objeto de inspección; se calibra si hay desviación significativa. Valor significativo es aquel que es poco probable en condiciones de actuación no delictiva. Por necesidad del método estadístico, siempre hay un porcentaje, pequeño, pongamos 5%, de falsos positivos; es decir valores que son significativos a pesar de que no hay actuación delictiva. Ningún método estadístico puede

[1] Los lectores interesados pueden consultar el video https://www.youtube.com/watch?v=cbkdhD6BsoY o la presentación Gill, R. (2014) "Murder by Numbers" http://www.math.leidenuniv.nl/~gill/RDG_TEDxAntwerpen_Naked.pdf

reducir dicho porcentaje a 0. Entonces en un contexto de *Big Data,* el problema es que podemos someter a test a la totalidad de la población de individuos y de manera reiterada en el tiempo. Y en el caso de un parámetro de prevalencia ("*base rate*", probabilidad delictiva de cada individuo) pequeña, este tipo de inspección puede arrojar demasiados falsos positivos. Es el mismo problema que se produce cuando realizamos un test masivo sobre toda la población de un país para detectar una enfermedad rara. No es lo mismo una desviación significativa para un individuo (en situación de riesgo) al que ya se le ha sometido a test, que un individuo con desviación positiva en el conjunto de individuos que fueron sometidos a test. Volvemos a la Falacia del Fiscal: no es suficiente una probabilidad muy pequeña de la coincidencia como fuente de acusación.

Sean:

A ="desviación significativa"

C ="Actuación criminal"

Los cálculos relevantes son

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A \mid C)P(C) + P(A \mid \overline{C})P(\overline{C})} \approx \frac{P(C)}{P(C) + P(A \mid \overline{C})}$$

Si la prevalencia, $P(C)$ es muy pequeña, pongamos $1/1000$ y $P(A \mid \overline{C}) = 0.01$ entonces $P(C \mid A) = 0.09099$, probabilidad pequeña. En la Figura 7 ilustramos en rojo los falsos positivos en muchas pruebas de test en que no hay actuación criminal. Big data maneja datos de muchos individuos de la población y permite señalar aquellos con desviación significativa de un cierto patrón. Es fácil caer en una trampa de metodología estadística, en donde datos conseguidos para un tipo de objetivo (como gestión y previsión de actividades en servicio hospitalario), son utilizados después para un objetivo muy distinto, como el acusatorio de crimen. Un ejemplo de ello es la tragedia de Lucia Berk: en big data las coincidencias raras se dan (los lectores interesados pueden consultar webs como *Five-O Police Rating App 1.2* o *The Innocence Project* y, en ambas, más información sobre otras fuentes).

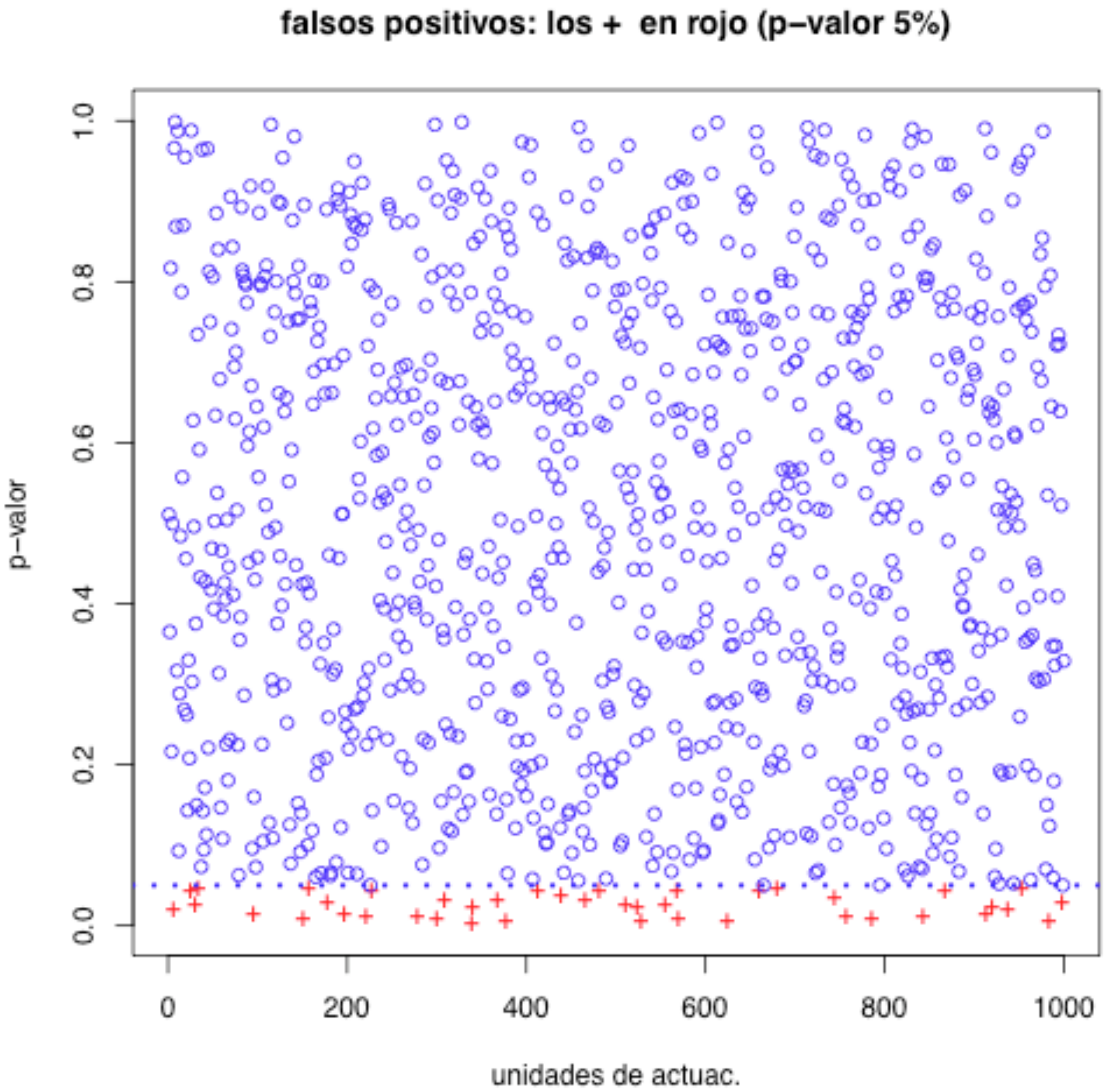

Figure 7: El caso Lucía de Berk seria uno cualquiera de las cruces en rojo.

Bibliografía

EL PAÍS SEMANAL (2016) "Falso culpable", núm. 2058, 6 de marzo de 2016 http://elpais.com/elpais/2016/03/03/eps/1457026285_449993.html

Alan, DERSHOWITZ (1996), "Reasonable Doubts: The Criminal Justice System and O.J. Simpson Case, New York Simon and Schuster" New York Simon and Schuster, 1996

Peter, DONNELLY, «How Juries are fooled by statistics», https://www.ted.com/talks/peter_donnelly_shows_how_stats_fool_juries/transcript

Ward, FARNSWORTH (2008) "The Legal Analyst. A Toolkit For Thinking About The Law", The University of Chicago Press, 2007, págs. 281 y ss.

Richard, GILL (2014), "Murder by Numbers" http://www.math.leidenuniv.nl/~gill/RDG_TEDxAntwerpen_Naked.pdf

Isadore G., GOOD (1995), "When batterer turns murderer", *Nature,* 375, 269-326, 15 junio 1995