Qué creer y cuánto

Una breve introducción al pensamiento estadístico



JOSE BLANCA

Qué creer y cuánto

Una breve introducción al pensamiento estadístico

JOSE BLANCA

Índice

Una nota sobre por qué y cómo se escribió este libro 1
Esto no es una pipa 2
Entender el valor p 13
Qué salió mal con el valor p 21
Máxima verosimilitud 31
La probabilidad inversa: la probabilidad que de verdad queremos 37
La inferencia bayesiana 44
La alternativa frecuentista 53
La defensa bayesiana 59
Una tregua pragmática 67
Pluralismo y oficio 71

Una nota sobre por qué y cómo se escribió este libro

Durante muchos años he enseñado bioinformática y he usado la estadística en el laboratorio, y en todo ese tiempo he visto los mismos dos problemas una y otra vez. El primero es que mucha gente que usa la estadística no entiende realmente lo que está haciendo. El segundo es que la estadística puede dar una falsa sensación de objetividad: un valor p en un artículo hace que un resultado parezca científico, tanto si el razonamiento que hay detrás se sostiene como si no.

Creo que entender estas herramientas es esencial para cualquiera que las use: biólogos, psicólogos, médicos y muchos otros. Y, sin embargo, la mayoría no las entiende. En parte, porque las ideas son genuinamente difíciles. Pero también porque la estadística suele enseñarse como un conjunto de recetas, sin el marco conceptual que explica qué hace cada método y por qué. Este libro intenta ofrecer ese marco.

Hay también una razón más personal. Cuando empecé, era bayesiano, y quería explicar por qué el enfoque frecuentista y el valor p estaban equivocados. Sigo pensando que la forma en que se suele usar el valor p es errónea. Pero por el camino dejé de ser bayesiano. Hoy me consideraría pluralista y, en cierto modo, este libro recorre ese viaje.

Este libro se ha escrito a cuatro manos. Las ideas y la perspectiva son principalmente mías, y soy responsable del resultado final, errores incluidos. Pero Claude Opus 5.5, un modelo de IA creado por Anthropic, ha trabajado conmigo de principio a fin como editor, profesor, verificador de datos y corrector de inglés. Sin sus aportaciones, este libro no habría sido posible.

Esto no es una pipa

¿Qué crees que significa un valor p?

Imagina que estás investigando la eficacia de una terapia: tratas a algunas personas, dejas sin tratar a un grupo control y luego comparas los resultados de los dos grupos. Tras analizar los datos obtienes un valor p de 0,01, muy por debajo del umbral habitual de 0,05. Ahora piensa en qué significa este valor p. ¿Cuáles de estas afirmaciones son verdaderas?

1. Has refutado de forma absoluta la hipótesis nula (que no hay diferencia entre las medias poblacionales).
2. Has hallado la probabilidad de que la hipótesis nula sea cierta.
3. Has demostrado de forma absoluta tu hipótesis experimental (que hay una diferencia entre las medias poblacionales).
4. Puedes deducir la probabilidad de que la hipótesis experimental sea cierta.
5. Si decides rechazar la hipótesis nula, sabes la probabilidad de que estés tomando la decisión equivocada.
6. Tienes un hallazgo experimental fiable, en el sentido de que, si el experimento se repitiera un gran número de veces, obtendrías un resultado significativo el 99 % de las veces.

Elige tus respuestas antes de seguir leyendo.

La respuesta correcta es que las seis son falsas. No te preocupes si has fallado: poca gente entiende de verdad el significado del valor p. El valor p no es la probabilidad de equivocarte al rechazar la hipótesis nula. Tampoco es la probabilidad de que los resultados se deban al azar, ni uno menos la probabilidad de que tu hipótesis sea cierta. De hecho, el valor p no es la probabilidad de que ninguna hipótesis sea verdadera o falsa. Entonces, ¿qué

es? Es algo que tiene que ver con los datos que esperarías ver si la hipótesis nula fuera cierta. Más adelante veremos exactamente qué.

Michael Oakes presentó estas afirmaciones a psicólogos académicos británicos en 1986. En 2002, Heiko Haller y Stefan Krauss se las presentaron a tres grupos de seis universidades alemanas: 44 estudiantes de psicología que habían aprobado sus asignaturas de estadística, 39 profesores de psicología y 30 personas que enseñaban estadística y contraste de la hipótesis nula. Absolutamente todos los estudiantes creyeron que al menos una de las afirmaciones era verdadera. También el 90 % de los profesores de psicología y el 80 % de los profesores de estadística. El error más frecuente fue la afirmación 5, la idea de que el valor p es la probabilidad de equivocarse. Se la creyó alrededor del 70 % de cada grupo alemán y el 86 % de los psicólogos británicos. Estudios similares con psicólogos académicos de España, Italia y Chile han encontrado los mismos errores. En Israel, Ruma Falk y Charles Greenbaum hicieron a estudiantes el mismo tipo de pregunta. Previamente, los estudiantes habían leído un artículo clásico que advertía precisamente contra estos errores, y se les ofrecía de forma explícita la opción «ninguna de las afirmaciones es correcta». Aun así, el 87 % eligió al menos una interpretación errónea.

¿Para qué molestarse?

La estadística es fundamental para la ciencia moderna, y el valor p es una de las herramientas estadísticas más comunes. Sin embargo, la mayoría de los científicos no entiende lo que está haciendo cuando lo aplica. La mayoría aprendimos la estadística como una colección de recetas. Tienes dos grupos: haces una prueba t. Tienes una tabla de contingencia: haces una chi cuadrado. Si el número que sale es menor que 0,05, te pones contento, colocas un asterisco sobre la barra de tu gráfico y sigues adelante. Y si alguien te pregunta qué significa ese número… bueno, esperemos que nadie pregunte.

El físico Richard Feynman tenía un nombre para esto: *ciencia de culto al cargo*. Después de la Segunda Guerra Mundial, algunos isleños del Pacífico que habían visto aterrizar aviones cargados de mercancías construyeron pistas de aterrizaje y torres de control de bambú, y esperaron a que los aviones volvieran. Todo parecía correcto, pero los aviones nunca aterrizaron. Hacer estadística a base de recetas se parece mucho a eso. Seguimos el ritual (recoger los datos, hacer la prueba, comprobar el umbral) sin entender para qué sirve el ritual. Pero la ciencia no consiste en recoger datos y hacer

pruebas para obtener resultados. Consiste en entender el mundo. Y esto no es solo un problema teórico. El mal uso de las herramientas estadísticas, y del valor p en particular, se ha convertido en un problema práctico real para la ciencia. Es una de las causas de las crisis de replicación, junto con otras como el sesgo de publicación, los incentivos de la carrera académica y los estudios demasiado pequeños para detectar nada. Entender lo que hacemos no arreglará todo eso, pero es un comienzo necesario.

Si marcaste como verdadera alguna de las afirmaciones del test y crees que entender lo que hacemos importa, puede que este libro te sea útil. No voy a explicar cómo hacer ninguna prueba. Lo que quiero es entender qué hacemos cuando hacemos estadística, y por qué la hacemos como la hacemos. Te aviso de antemano: buena parte de lo que sigue es filosofía. Detrás del valor p, y detrás de cualquier otra herramienta estadística, hay filosofía. Y hay profundos desacuerdos filosóficos por los que los estadísticos llevan un siglo peleándose.

Jugadores y aseguradores

Empecemos por el principio.

La teoría de la probabilidad nació en la mesa de juego. Los primeros en pensar seriamente sobre las probabilidades querían conocer sus posibilidades con las cartas y los dados. Gerolamo Cardano escribió un libro sobre juegos de azar hacia 1564 (en latín, por si te apetece una lectura ligera). A Galileo le preguntaron por los dados hacia 1620. En 1654 Pascal y Fermat intercambiaron cartas sobre cómo repartir las apuestas de una partida interrumpida, y Huygens publicó un pequeño tratado en 1657. Todos ellos se hacían el mismo tipo de pregunta: si sé cómo funciona el juego, ¿cuál es la probabilidad de cada resultado? Eso es una especie de p(datos | hipótesis).

Después, en 1713, Jacob Bernoulli se hizo la pregunta inversa: si he observado ciertas frecuencias, ¿hasta qué punto puedo estar seguro de la probabilidad subyacente? Esa es la probabilidad inversa, la que de verdad queremos, que aparece ya a principios del siglo XVIII. Un ejemplo de esta probabilidad inversa: si estoy jugando a los dados con un amigo, dados los resultados que he visto, ¿cuál es la probabilidad de que mi amigo esté haciendo trampas? También bastante pronto, De Moivre descubrió que la campana de Gauss, la distribución normal, aparece como aproximación a la binomial cuando lanzas una moneda muchas veces.

En paralelo aparecieron las primeras estadísticas. La propia palabra viene de *estado*: originalmente, la estadística consistía en recoger datos sobre el estado, como cuánta gente vive aquí, cuánta tierra y cuánta riqueza hay… y, por supuesto, cuántos impuestos se pueden recaudar. No es el origen más romántico, pero sí uno práctico. Y aquellas primeras estadísticas trataban de gente que nacía y gente que moría. En 1662 John Graunt convirtió los registros semanales de defunciones de Londres en las primeras tablas estadísticas. En 1693 Edmond Halley, el del cometa, usó una tabla de mortalidad para calcular el precio de las rentas vitalicias. ¿Quién crees que estaba muy interesado en conocer la probabilidad de que la gente muriera o enfermara? Las aseguradoras. Si le vendo a esta persona un seguro de vida, ¿cuánto me va a costar? Las aseguradoras estuvieron entre las primeras en trabajar con estadísticas.

En 1710, John Arbuthnot se dio cuenta de que en Londres, durante 82 años consecutivos, se habían bautizado más niños que niñas. Si fuera cuestión de puro azar, 50-50 cada año, la probabilidad de eso sería $(½)^{82}$, un número ridículamente pequeño. Es, posiblemente, la primera prueba de significación de la historia: se rechazó la hipótesis del 50-50. La conclusión de Arbuthnot, por cierto, fue que aquello era una prueba de la Divina Providencia.

Modelos

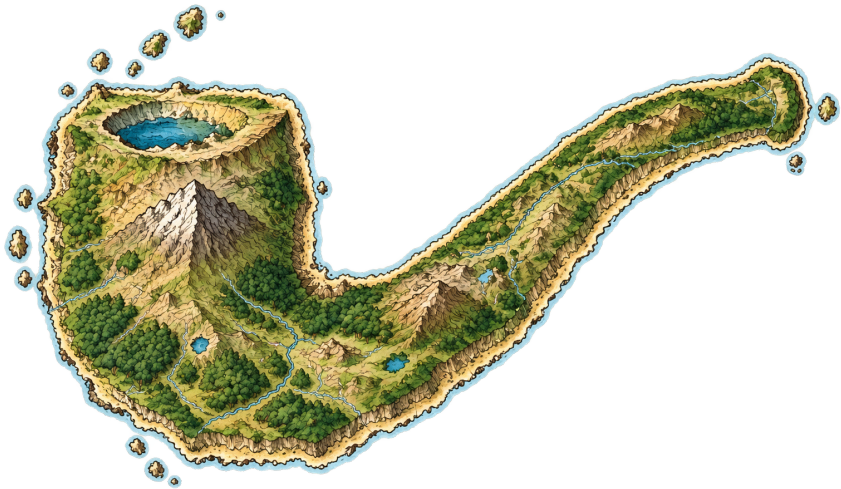
¿Qué hacemos cuando hacemos inferencia estadística? En su forma más básica: tenemos unos datos y queremos acabar con modelos del mundo que se ajusten aproximadamente a esos datos.

Imagina que salimos a la calle y medimos la altura de todo el que pasa. Esos son nuestros datos, una larga lista de números. Después representamos la distribución de esas alturas, quizá una para las mujeres y otra para los hombres, y concluimos: así se distribuye la altura humana. Esa distribución es nuestro modelo. Hemos empezado con datos y hemos acabado con una idea de cómo es el mundo. Al menos, eso parece. Más adelante veremos que el modelo no salió solo de los datos.

Esto no es una pipa

Hay un cuadro famoso de René Magritte. Muestra una pipa y, debajo, la leyenda *Ceci n'est pas une pipe*, «esto no es una pipa». La primera reacción es: ¿cómo que no es una pipa? ¡Claro que es una pipa! Pues no. Es un cuadro de una pipa.

Puede parecer una broma tonta, pero es una de las ideas más importantes de todo este libro, y volveré a ella una y otra vez. Existe la realidad, y existen las representaciones, los modelos, las ideas que nos hacemos de la realidad. No son lo mismo. Otra forma de decirlo: **el mapa no es el territorio**.



Esto no es una isla. Es el mapa de una.



El territorio.

Un mapa es una representación simplificada. No muestra cada piedra. Puede mostrar un río, pero cuando llegas allí quizá solo encuentres un cauce seco. Esta idea, que el modelo no es la realidad, está detrás de algunos de los desacuerdos más profundos que han tenido los estadísticos durante más de un siglo.

Confundimos la realidad con nuestras ideas sobre la realidad constantemente, en nuestra experiencia cotidiana del mundo. Si te digo «ahí hay una silla», ¿hay una silla? La mayoría diría que sí, evidentemente. Pero un físico podría decir que lo que hay es un montón de átomos. ¿Quién tiene razón? Los dos. «Silla» y «átomo» son categorías que usamos para representar la realidad, y son buenas categorías porque son útiles. Si te digo que hay una silla, puedes sentarte. Si te digo que hay un montón de átomos, no sabrás qué hacer con esa información, pero un químico sí. En ese sentido, las sillas y los átomos son ambos reales: son representaciones útiles de la propia realidad. Pero son mapas, no el territorio. Muchos pensaréis que esta es una discusión inútil, que solo interesa a filósofos a los que no les importan las cuestiones prácticas. Pero, a menos que te la tomes en serio, te costará entender la estadística y lo que podemos hacer con ella: entender las enfermedades, probar un nuevo fármaco o, como veremos más adelante en el libro, incluso ayudar a ganar una guerra.

Qué hace bueno a un modelo

La ciencia ni siquiera pretende crear modelos exactos de la realidad; los modelos científicos ignoran a propósito muchos aspectos de la realidad. Cuando estudiaste la bola que rueda por el plano inclinado en clase de física, probablemente ignoraste la rotación de la bola y la resistencia del aire. ¿Era erróneo el modelo? En sentido estricto, sí. ¿Era útil? Mucho. Fuiste al laboratorio e hiciste algunas predicciones con tu modelo simplificado. Si coincidían más o menos con las mediciones, eso era todo lo que querías.

Un buen modelo, entonces, es una analogía que funciona:

- representa algo del mundo real, y sus predicciones se pueden comparar con lo que observamos,
- pero es lo bastante sencillo como para entenderlo y usarlo.

Cualquiera de nuestras ideas sobre el mundo, cualquier hipótesis, teoría o idea común, debería ser así: una simplificación que funciona. Un estadístico,

George Box, lo expresó así: *todos los modelos son erróneos, pero algunos son útiles.*

Los modelos tienen dos propiedades clave. Primera, **un modelo hace predicciones**. Si estudio la física de Newton, puedo predecir cuánto tardará la bola en caer. Segunda, **esas predicciones se pueden comprobar**. Ten presentes estas dos propiedades, porque volverán. Como lo del mapa y el territorio, puede parecer sencillo, pero recuérdalo: es uno de los cimientos del pensamiento estadístico y científico. Puedo ir al laboratorio, medir y ver si el modelo se ajusta a los datos o no. Si me imagino que la puerta está ahí, puedo intentar salir de la habitación. Puede que acierte, o puede que me dé contra una pared. Así de simple.

Modelos estadísticos

Hemos estado hablando de modelos científicos. Los modelos estadísticos son un tipo especial de modelo científico: incluyen alguna idea de aleatoriedad.

Imagina que tengo una moneda en la mano. No una moneda real: una imaginaria. Puedo imaginar que esta moneda tiene una probabilidad de 0,5 de salir cara. Ese es un modelo estadístico típico. Suponiendo ese modelo, ¿podría predecir cuántas caras saldrían si lanzara la moneda 100 veces? No exactamente. No puedo predecir el resultado de un único experimento. Pero sí puedo predecir la distribución de resultados si repito el experimento muchas veces: unas veces saldrán 50 caras, otras 45, otras 48, otras 55. Así que, a partir de un modelo estadístico, puedo calcular la distribución de resultados que espero ver: qué probabilidad hay de obtener 50 caras, o 45, o 70.

Fíjate en la dirección de ese cálculo. Partimos del modelo y obtenemos la probabilidad de los datos. El modelo nos dice qué probabilidad tienen los datos, no qué probabilidad tiene el modelo. Tenlo presente: intercambiar ambas cosas es uno de los errores más comunes con los valores p.

Además, los modelos estadísticos suelen tener parámetros. Podría imaginar una moneda que sale cara con una probabilidad desconocida h, que podría ser 0,5, o 0,4, o 0,3. Encontrar el valor de h que se ajusta a los datos es una de las cosas típicas que querremos hacer. Esto se llama **estimación**.

Los modelos también suelen hacer supuestos: por ejemplo, que cada lanzamiento es independiente del anterior.

No tenemos acceso directo a la realidad. Solo tenemos un sustituto: los datos que reunimos mediante experimentos y observaciones. Con esos datos podemos estimar los parámetros de nuestros modelos, comprobar si se ajustan al mundo y compararlos entre sí.

Los modelos no salen solo de los datos

¿De dónde salen los modelos? Necesitamos algún método para crearlos, y aquí hay un punto importante: los modelos no se pueden crear solo a partir de los datos.

Recuerda el ejemplo de las alturas. Parecía que el modelo hubiera salido directamente de los datos. Pero fíjate en todo lo que aportamos a los datos antes de tener una sola medición. Elegimos medir la altura, y no la talla de zapato o el color de ojos. Decidimos que separar los datos por sexo era buena idea. Y cuando dijimos «así se distribuye la altura», probablemente ya estábamos imaginando la conocida campana de Gauss. Los datos no nos dijeron nada de eso. Lo pusimos nosotros.

Es análogo a lo que ocurre con tus sentidos. Abres los ojos, entra luz y aparece un gato en tu mente. Podrías pensar que la imagen del gato se crea solo a partir de la luz, pero no es así. La crea una máquina muy compleja, tu sistema visual, que usa esa luz. Sin el sistema visual no hay gato en tu mente. Puede parecer un mero detalle, hasta que piensas en las veces que tu sistema visual te engaña.

No es una idea nueva. Los filósofos de la ciencia la llaman la *carga teórica* de la observación: nunca observamos datos «en bruto», porque lo que miramos, lo que medimos y cómo lo interpretamos dependen ya de nuestras ideas sobre el mundo. El aprendizaje automático se ha topado con el mismo problema desde el lado de la ingeniería y lo llama *sesgo inductivo*. Un algoritmo de aprendizaje necesita algunos supuestos incorporados sobre qué tipo de patrón buscar, y sin supuestos aprender de los datos es sencillamente imposible. Así que, para construir modelos a partir de datos, necesitas un método, y todo método lleva consigo supuestos.

Para qué sirve la estadística

La estadística hace cosas distintas con los datos.

La primera es **describir** los datos. Lo más sencillo que podemos hacer es resumirlos. ¿Cuántas caras y cuántas cruces han salido? ¿Cuál es la distri-

bución de las alturas? Esto es la estadística descriptiva, y para ello podemos usar, por ejemplo, gráficos. Es extremadamente útil, y nunca deberías saltártela. En cierto modo, los gráficos y los resúmenes ya son modelos sencillos de los datos.

La segunda es **construir** modelos. Una vez elegido el tipo de modelo, por ejemplo una moneda que sale cara con cierta probabilidad h, construirlo consiste sobre todo en estimar: ajustar sus parámetros hasta que el modelo se ajuste a los datos lo mejor posible. ¿Cuánto vale h? ¿Qué tamaño tiene el efecto de la terapia? ¿Y cuánta incertidumbre tenemos sobre esos valores? La estimación también puede verse como una elección entre modelos. Una moneda con $h = 0{,}5$, otra con $h = 0{,}4$, otra con $h = 0{,}3$: todos son modelos que compiten entre sí y que solo se diferencian en el valor de un parámetro, y estimar significa escoger el que mejor se ajusta a los datos.

La tercera es **comprobar** los modelos frente a los datos. Comprobar es poner a prueba si un modelo se ajusta a los datos. Para hacerlo, comparamos las predicciones del modelo con los datos, y el valor p es una forma de hacerlo.

Por último, cuando tenemos un modelo en el que confiamos, podemos **usarlo** para hacer predicciones sobre el mundo: cuántas caras saldrán en los próximos 100 lanzamientos, o cómo responderá un nuevo paciente a la terapia. La predicción es una de las principales razones para construir modelos, y es la mayor parte de lo que hace el aprendizaje automático.

Un adelanto filosófico

Detrás de las ideas que vamos a explorar hay dos grandes ideas filosóficas y una trampa.

La primera idea es que existe una profunda **asimetría entre rechazar y aceptar** una hipótesis. Rechazar es más directo que aceptar: podemos ver millones de cisnes blancos y aun así equivocarnos al aceptar que todos los cisnes son blancos, mientras que basta un solo cisne negro para rechazar la hipótesis.

La trampa es pensar que la probabilidad de tener una enfermedad dado un resultado positivo en un chequeo es la misma que la probabilidad de un resultado positivo si estoy enfermo. Suenan parecidas. Son radicalmente distintas. La primera es la probabilidad de la hipótesis dados los datos; la segunda es la probabilidad de los datos dada la hipótesis. La primera se llama **probabilidad inversa**.

Y la segunda idea filosófica tiene que ver con la probabilidad de la hipótesis: **¿puede una probabilidad representar un grado de creencia?** Si digo «después del experimento, estoy seguro al 90 % de que el fármaco funciona», ¿significa algo ese 90 %?

Esta segunda idea merece un pequeño anticipo, porque ha provocado mucho debate y es el desacuerdo que subyace a dos enfoques distintos del análisis estadístico. Imagina que tengo una moneda real. Puedo decir que tiene cierta probabilidad de salir cara, y puedo medirla: la lanzo muchísimas veces y calculo la proporción de caras. Este enfoque parece bastante poco polémico. Pero ¿qué pasa con esta idea: «creo que la probabilidad de que mañana llueva es 0,8»? ¿Cómo mides eso? Mañana solo ocurre una vez. Lloverá o no lloverá. ¿Dónde está la frecuencia? La postura de los frecuentistas era que no se deberían asignar probabilidades a las hipótesis, y que no se puede medir la probabilidad de algo que solo ocurre una vez. Los bayesianos, en cambio, aceptan plenamente que podemos asignar probabilidades a las hipótesis.

El problema es que, si aceptas la postura de los frecuentistas, nunca puedes hablar de la probabilidad de una hipótesis. Una hipótesis es verdadera o falsa, igual que mañana llueve o no llueve. Lo único que puedes decir es: si la hipótesis fuera cierta, esperaría ver esto. Por eso la estadística tradicional se centra tanto en el valor p, que es una probabilidad de los datos, no de la hipótesis. Si las hipótesis no pueden tener probabilidades, p(datos | hipótesis), la probabilidad de los datos dada la hipótesis, es lo único que queda, y confundirla con p(hipótesis | datos) es la trampa. Mientras tanto, todos los hombres y mujeres del tiempo de la tele te dicen tan tranquilos que la probabilidad de lluvia mañana es del 80 %, y parecen creer que eso significa algo. Esta discusión puede parecer teórica y sin consecuencias prácticas, pero volverá una y otra vez, y tiene profundas implicaciones para cómo hacemos estadística, e incluso para cómo decidimos si un tratamiento contra el cáncer funciona.

Un libro para leer junto a este

Si quieres una buena introducción a la estadística, te recomiendo *The Art of Statistics: Learning from Data* (2019), de David Spiegelhalter. No es una lista de recetas. Intenta explicar de qué trata la estadística, y lo hace muy bien. Es la perspectiva de un estadístico en activo y con experiencia, llena de ejemplos reales. Mi perspectiva será algo más filosófica, así que creo que ambos se complementan.

Lecturas adicionales

- Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606. La fuente del test y de sus resultados.
- Haller, H. & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online*, 7(1).
- Oakes, M. (1986). *Statistical inference: a commentary for the social and behavioural sciences*. Wiley.
- Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5(1), 75–98.
- Badenes-Ribera, L. et al. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology*, 7, 1247.
- Greenland, S. et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350.

Entender el valor p

Coge una moneda y lánzala 100 veces. Te salen 60 caras. ¿Es una moneda equilibrada? Haces la prueba binomial exacta y el programa te da p = 0,057. Justo por encima de 0,05, así que no es significativo. ¿Significa eso que la moneda está equilibrada? ¿Que está casi trucada? ¿Y qué significa realmente ese 0,057?

Has usado valores p, y probablemente los seguirás usando el resto de tu vida científica. Así que intenta responder a esa última pregunta. La respuesta más común que recibo es: «es la probabilidad de que la hipótesis sea cierta». Si le dijeras eso a Ronald Fisher, el hombre que popularizó el valor p, no le haría ninguna gracia. En un valor p no hay probabilidad de ninguna hipótesis. Ninguna.

No te sientas mal. Como vimos con el test, casi nadie entiende el valor p, incluidos muchos de los que lo enseñan. Parte del problema es que su definición es un trabalenguas: el valor p es la probabilidad de obtener unos datos al menos tan extremos como los observados, si la hipótesis nula fuera cierta.

Los valores p, como otros conceptos estadísticos, no son ideas fáciles. Fisher trabajaba en una estación de investigación agrícola, rodeado de científicos que necesitaban métodos prácticos, no demostraciones matemáticas. Así que escribió libros para ellos, llenos de ejemplos resueltos y con muy poca teoría: *Statistical Methods for Research Workers* y *The Design of Experiments*. Tuvieron un éxito enorme, y muchos de sus lectores, y después muchos libros de texto, los usaron como recetas que aplicar en lugar de ideas que entender. Lo que te enseñaron en tu asignatura de estadística son, más o menos, esas recetas. Intentemos entenderlas.

¿Estoy viendo lo que espero ver?

El valor p descansa sobre dos ideas:

1. Si supongo una hipótesis, un modelo del mundo, puedo calcular lo que espero ver.
2. Si lo que veo es muy distinto de lo que esperaba, rechazo la hipótesis.

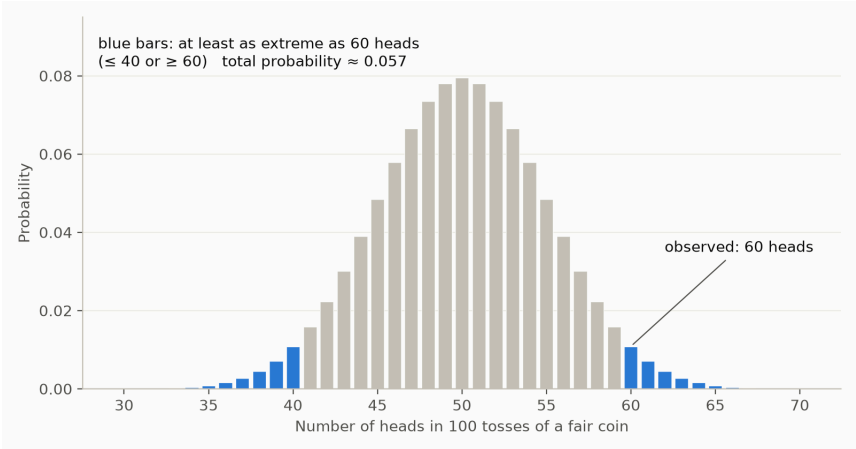
Veamos un ejemplo sin ninguna probabilidad. Imagina que quiero comprobar la hipótesis «no hay estudiantes en el aula». Si fuera cierta, ¿qué esperaría ver al abrir los ojos? A nadie. Abro los ojos y veo gente sentada. Rechazo la hipótesis, y lo hago con total seguridad.

Eso es exactamente lo que hace una prueba de significación. La única diferencia es que los modelos estadísticos incluyen aleatoriedad, así que sus predicciones no son un único resultado, sino una distribución de resultados.

Volvamos a la moneda imaginaria que tengo en la mano. Supongamos que sale cara con probabilidad 0,5. Si la lanzara 100 veces, ¿qué esperaría ver? La mayoría de las veces, algo en torno a 50 caras: 45, 48, 52, 55. Puedo calcular esa distribución con precisión. Ahora cojo una moneda real, la lanzo 100 veces y obtengo 100 caras. Eso sería extremadamente raro para una moneda equilibrada, así que rechazo la hipótesis de que la moneda está equilibrada.

Eso es el valor p: si la hipótesis fuera cierta, ¿qué probabilidad habría de observar algo tan extremo como lo que he observado, o más? Si esa probabilidad es muy pequeña, los datos son muy raros bajo la hipótesis, y empiezo a sospechar de la hipótesis. Para 100 caras de 100, esa probabilidad es de unos 0,0000000000000000000000000000002. Raro, sin duda.

Volvamos ahora al caso menos extremo del principio: 60 caras. La figura muestra lo que nos daría una moneda equilibrada si repitiéramos los 100 lanzamientos muchísimas veces.



La mayoría de las veces obtendríamos algo entre 40 y 60 caras. Exactamente 50, el resultado más probable, solo sale alrededor del 8 % de las veces. Por eso no preguntamos «¿qué probabilidad hay de obtener exactamente 60 caras?»: cualquier resultado concreto es improbable. Lo que hace sospechosas las 60 caras es lo lejos que están de lo que esperábamos. Así que sumamos la probabilidad de todos los resultados que están al menos igual de lejos de 50, en cualquiera de las dos direcciones: 60 caras o más (unos 0,028) y 40 o menos (otros 0,028). En total, unos 0,057. Ese es el valor p, y eso es lo que significa «al menos tan extremo».

¿Es 0,057 lo bastante raro como para rechazar la hipótesis de que la moneda está equilibrada? Según la convención habitual, 0,05, no lo es, aunque por muy poco. Volveremos a esa convención.

Rechazar hipótesis

Cuando usamos valores p hacemos algo bastante peculiar. Si quiero saber si hay estudiantes en el aula, no planteo esa hipótesis. Planteo la contraria: que no hay estudiantes. Si quiero saber si un fármaco mejora el tratamiento del cáncer, el procedimiento estándar es suponer que no lo hace (la *hipótesis nula*) e intentar rechazarlo. ¿Por qué no planteo directamente que el fármaco funciona e intento aceptarlo? La razón es una idea filosófica: la asimetría entre rechazar y aceptar hipótesis. Rechazar puede ser una decisión tajante, pero aceptar es mucho más arriesgado.

Para entender por qué ponemos a prueba la hipótesis en la que *no* creemos, necesitamos un pequeño rodeo por la filosofía: cisnes, vacas voladoras,

Newton y un filósofo escocés que dijo cosas que nadie quería oír. Ten paciencia; volveremos al valor p con una idea mucho más clara de lo que hace.

Cisnes negros y vacas voladoras

El ejemplo clásico es el de los cisnes. Durante siglos los europeos vieron cisnes, y todos eran blancos. Miles de observaciones, siempre blancos. Concluyeron, con gran seguridad: «los cisnes son blancos». Luego alguien fue a Australia y encontró cisnes negros. Miles de observaciones no habían bastado para demostrar la hipótesis, pero una sola observación bastó para destruirla.

Esta es la **asimetría lógica**: es mucho más fácil que la evidencia hable de forma decisiva en contra de una hipótesis que que la establezca.

Otro ejemplo. Supongamos que quiero demostrar que las vacas vuelan. Es muy difícil. Voy al campo, miro las vacas y no están volando. Pero quizá estén descansando. Quizá vuelen de noche. Nunca puedo estar seguro de que no vuelan solo porque no haya visto a ninguna hacerlo. En cambio, si mi hipótesis es «las vacas no vuelan», refutarla es fácil: una vaca voladora y asunto zanjado.

Fisher escribió en *The Design of Experiments* (1935) que la hipótesis nula «nunca se demuestra ni se establece, pero posiblemente se refuta», y que todo experimento existe «solo para dar a los hechos la oportunidad de refutar la hipótesis nula». Como no puedes aceptar hipótesis, solo rechazarlas, siempre planteas la hipótesis de la que te quieres librar. Si quieres saber si un fármaco funciona, la nula es que no funciona. Si estás haciendo un estudio de asociación de genoma completo (GWAS) y quieres saber si un locus está asociado con un fenotipo, la nula es que no lo está. ¿Y si quieres comprobar que las vacas no vuelan? Entonces la hipótesis nula es que las vacas sí vuelan. Siempre lo contrario de lo que te interesa.

La idea de Fisher era: mira si los datos son raros bajo tu hipótesis. Si lo son, sospecha de la hipótesis. Si no lo son, quizá la hipótesis esté bien, aunque no puedes estar seguro de ello. Como veremos, lo que pasó después es otra historia: la enseñanza de la estadística lo convirtió en un juego del teléfono escacharrado.

El método científico que no existe

Esta idea está relacionada con lo que te enseñaron en el colegio como *el método científico*, más propiamente llamado método hipotético-deductivo: formulas una hipótesis, deduces lo que deberías observar si es cierta, vas a mirar y después ves si la hipótesis se ve respaldada o refutada. (No existe un único método científico, pero dejemos eso para otro día).

¿Puedes llegar a estar seguro de algo de esta manera? Piensa en Newton. Te enseñaron las leyes de Newton, la gravedad, las manzanas y los planetas. ¿Es verdadera la física newtoniana? Durante dos siglos todos los experimentos encajaron: los planetas, las mareas, las balas de cañón. Y entonces alguien miró con atención la órbita de Mercurio, y su perihelio no se desplazaba como predecía Newton. Llegó Einstein con una teoría que sí encajaba. Son otra vez los cisnes: creíamos que era verdad, y las observaciones demostraron que no lo era. Así que ahora sabemos que la física newtoniana es errónea. Pero ¿son verdaderas la relatividad y la mecánica cuántica? Hasta ahora ninguna observación las ha contradicho. ¿Son verdaderas, entonces?

El filósofo Karl Popper llevó esta asimetría hasta el límite. Para él, la ciencia consiste en rechazar hipótesis. Nunca se puede demostrar que una teoría es verdadera, pero se puede poner a prueba y, posiblemente, refutar. E hizo de eso el rasgo que define la ciencia: la **falsabilidad**. Una hipótesis solo es científica si, en principio, la observación podría refutarla. Una idea científica tiene que jugarse algo: «predigo esto, y si no ocurre, estoy equivocado».

La falsación tiene una propiedad preciosa: en principio solo requiere lógica deductiva válida. Si la hipótesis implica X, y X no ocurre, la hipótesis es falsa. Pero también tiene serias limitaciones. En ciencia queremos saber qué hipótesis son verdaderas, no solo cuáles son falsas. Cuando voy al laboratorio, está bien saber que me equivocaba en algo, pero también quiero saber si la hipótesis con la que trabajo ahora tiene sentido. Popper te diría que es imposible saber si la relatividad es verdadera. Mientras tanto, los ingenieros construyen con ella los satélites del GPS.

Por debajo de todo esto hay un problema más antiguo. David Hume, en el siglo XVIII, señaló que el sol ha salido todos los días de tu vida, pero nada garantiza lógicamente que vaya a salir mañana. Filósofos posteriores plantearon lo mismo con los cisnes: mil cisnes blancos no garantizan lógicamente que todos los cisnes sean blancos. Este es **el problema de la inducción**: el razonamiento inductivo no es lógicamente válido. El filósofo

C. D. Broad expresó la esperanza de que algún día «el razonamiento inductivo, que ha sido durante mucho tiempo la gloria de la ciencia, deje de ser el escándalo de la filosofía».

El contraste de hipótesis: falsación con dados

Aquí acaba el rodeo, y ahora podemos juntar las piezas. **Las pruebas de significación son el rechazo de hipótesis aplicado a hipótesis probabilísticas.**

Si nuestras hipótesis hicieran predicciones exactas, podríamos rechazarlas con pura lógica deductiva, como en el aula vacía. Pero los modelos estadísticos incluyen aleatoriedad, así que nuestras predicciones son solo probabilísticas. El procedimiento es:

1. Plantea una hipótesis probabilística: «el fármaco no funciona».
2. Calcula la distribución de resultados que esperarías si la hipótesis fuera cierta.
3. Haz el experimento y mira los datos. Si el resultado cae dentro del rango esperado, no hagas nada. Si es demasiado raro, rechaza la hipótesis.

¿Cuánto es «demasiado raro»? La convención habitual es 0,05, es decir, uno entre veinte. Si usas 0,05 estás diciendo: si repitiera este experimento veinte veces y la hipótesis nula fuera cierta, más o menos una vez vería algo tan extremo como esto, o más, y diecinueve veces no.

Volvamos a las vacas. En sentido estricto, nunca podría rechazar la hipótesis de que las vacas vuelan solo por ver vacas echadas en un prado: quizá estaban descansando. Pero el valor p no trata con certezas. Me dice con qué frecuencia vería algo tan extremo si la hipótesis fuera cierta. Si eso solo pasa más o menos una vez de cada veinte, estoy dispuesto a rechazarla. Hagamos probabilística la hipótesis de las vacas voladoras: las vacas vuelan, pero solo una hora al día; las otras 23 descansan (pobrecillas, volar es agotador). Si voy a una granja con una vaca, lo más probable es que la vea descansando. Eso no demuestra nada. ¿Y si veo 26 vacas, todas descansando? Suponiendo que cada vaca decide independientemente si vuela, la probabilidad de eso es $(23/24)^{26} \approx 0{,}33$. Nada raro, así que no puedo rechazar la hipótesis de las vacas voladoras. Pero si encuentro una granja con 71 vacas y ninguna está volando, la probabilidad baja a unos 0,05. Ahora sí es lo bastante raro: rechazo la hipótesis de que las vacas vuelan.

Eso sí, una advertencia. Esto no es exactamente lo mismo que el aula vacía. Allí la lógica era impecable: si no hay estudiantes, no veo a nadie; veo gente; luego la hipótesis es falsa. Con probabilidades, el argumento pasa a ser: si la hipótesis es cierta, estos datos serían improbables; he visto estos datos; luego la hipótesis es improbable. Suena igual de razonable, pero no es válido. Pruébalo con este ejemplo: si una persona es ciudadana estadounidense, es muy improbable que sea miembro del Congreso de Estados Unidos. Esta persona es miembro del Congreso. Por tanto, probablemente no es ciudadana estadounidense. Evidentemente erróneo.

Así que, cuando «rechazamos» una hipótesis porque el valor p es pequeño, estamos tomando una decisión: los datos son demasiado raros como para seguir tomándonos en serio la hipótesis. *No* hemos calculado que la hipótesis sea improbable. El valor p es una probabilidad de los datos, no de la hipótesis, y para pasar de una a otra necesitaremos más ingredientes. Esa es la trampa del adelanto filosófico, y volveremos a ella.

Recuerda: rechazar, nunca aceptar

La regla que se enseña en las asignaturas de estadística es: *un resultado estadísticamente significativo puede llevarte a rechazar la hipótesis nula; un resultado no significativo no te permite aceptarla.*

Así que los estadísticos nos dicen: solo puedes rechazar. Haces tu ensayo clínico y rechazas la hipótesis de que tu terapia contra el cáncer no funciona. Pero ¿funciona la terapia? Normalmente suponemos que sí: «rechazo que el fármaco no funcione, luego el fármaco funciona». Es como si los estadísticos dijeran «solo rechazar» mientras te dan un codazo y te guiñan un ojo: *ya me entiendes, ¿eh?, ¿eh?* Y luego, cuando resulta que el fármaco no funciona, los estadísticos te dicen: «Te dije que solo rechazaras. Aceptaste la alternativa; eso es cosa tuya». Hemos llegado a una situación bastante absurda. ¿Cómo hemos llegado hasta aquí?

Lecturas adicionales

- Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.
- Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson. (Publicado originalmente en alemán en 1934). (Traducción española: *La lógica de la investigación científica*, Tecnos).
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. (Publicado originalmente como *Philosophical Essays Concerning Human Unders-*

tanding; retitulado en 1758). (Traducción española: *Investigación sobre el entendimiento humano*).
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606.

Qué salió mal con el valor p

Hemos visto qué es un valor p y la lógica que hay detrás: supones una hipótesis, calculas lo que esperarías ver y rechazas la hipótesis si los datos son demasiado raros. Y acabamos en una situación absurda. Los estadísticos nos dicen que solo podemos rechazar, y todo el mundo acepta. ¿Cómo hemos llegado hasta aquí? ¿Y tiene importancia?

El hijo bastardo

El valor p que te enseñaron es una quimera, un monstruo mitológico hecho con partes de animales distintos. Es el hijo bastardo de dos ideas, y ninguno de sus padres lo habría reconocido.

La prueba de significación de Fisher. Propones una hipótesis, la nula. El objetivo es ver si los datos desacreditan seriamente esa hipótesis. El valor p mide lo raros que son los datos bajo la hipótesis. Fisher sugirió 0,05 como una convención cómoda, pero más tarde insistió en que ningún científico usa un nivel de significación fijo en todos los casos: para él, un valor p pequeño era un motivo de sospecha, no un veredicto.

El contraste de hipótesis de Neyman y Pearson. Jerzy Neyman y Egon Pearson (el hijo de Karl Pearson, el Pearson de la chi cuadrado) pensaban más como ingenieros. En su enfoque hay dos hipótesis bien definidas, y el objetivo es elegir entre ellas. Fijaban de antemano dos tasas de error: α, la probabilidad de rechazar una hipótesis nula verdadera (error de tipo I), y β, la probabilidad de aceptar una hipótesis nula falsa (error de tipo II; $1 - \beta$ es la *potencia*). Esos umbrales se eligen antes de ver los datos, en función de lo costoso que sea cada tipo de error, y describen con qué frecuencia se equivoca el procedimiento a la larga.

¿Por qué importarían los costes? Yo genotipo tomates. Si me equivoco con un genotipo, mi análisis de estructura poblacional sale un poco más ruidoso, pero no se muere nadie. Un genetista clínico que le dice erróneamente a un paciente que porta una variante que determina su tratamiento contra el cáncer está en una situación muy distinta. La idea de Neyman y Pearson era que decides de antemano cuántos errores puedes tolerar, teniendo en cuenta lo que está en juego.

Fíjate en la pregunta que están respondiendo. No es «¿qué debo creer?», sino «¿qué debo hacer?»: qué regla de decisión mantiene mis errores en un nivel aceptable, dado lo que está en juego. Fisher, en cambio, se preguntaba qué dicen los datos sobre la hipótesis nula. Volveremos a esta distinción, porque es una de las claves de todo el debate.

NHST (*null hypothesis significance testing*, prueba de significación de la hipótesis nula), el híbrido que te enseñaron: hay una hipótesis nula, como en Fisher, con una alternativa vaga que nadie se molesta en definir. El objetivo declarado es rechazar la nula, como en Fisher, pero en la práctica usas el resultado para elegir la alternativa, como en Neyman–Pearson. Calculas el valor p a partir de los datos, como en Fisher, pero luego lo comparas con un umbral fijo, como α en Neyman–Pearson. Y β, la potencia, suele olvidarse. Ni Fisher, ni Neyman, ni Pearson habrían aceptado esto.

	Fisher	Neyman–Pearson	NHST (lo que te enseñaron)
Hipótesis	una, la nula	dos, ambas bien definidas	la nula, más una alternativa vaga
Umbral	ninguno, usa tu criterio	α y β, fijados de antemano según los costes de cada error	0,05, siempre
Qué obtienes	un grado de sospecha sobre la nula	una decisión, con tasas de error a largo plazo conocidas	«significativo», luego el efecto es real

El psicólogo Gerd Gigerenzer lo llamó **el ritual nulo**:

1. Plantea una hipótesis nula de «ningún efecto» y no digas qué predice tu propia hipótesis.

2. Usa el 5 % como convención para rechazarla. Si p < 0,05, acepta tu hipótesis de investigación.
3. Hazlo siempre, sea cual sea el problema.

Piénsalo un segundo. Aceptas una hipótesis que ni siquiera has enunciado y sobre la que no has calculado nada. La única hipótesis que has definido y sobre la que has calculado algo es la nula, y lo único que te estaba permitido hacer con ella era rechazarla. Pero es un ritual, y los rituales no están para entenderlos. Todas las revistas quieren la gráfica de barras con los asteriscos, así que pones la gráfica de barras con los asteriscos.

¿Cómo pasó esto? Nadie se sentó a proponerlo. Nadie escribió nunca un artículo en una revista de estadística defendiendo que así es como deben hacerse las cosas. Fue un proceso sociológico, y la culpa es sobre todo de los autores de libros de texto y de los profesores de estadística de los años cuarenta y cincuenta. Pensaban que, si explicaban las cosas como es debido, nadie las entendería. Además, nunca hay tiempo para la filosofía cuando hay un temario que cubrir. Así que mezclaron ambos enfoques en una receta sencilla: p < 0,05, verdadero. Y una idea sencilla se propaga bien. A los editores de las revistas les encantó. Compara el «tenemos datos que hacen que la hipótesis de que el fármaco no funciona parezca bastante extraña» de Fisher con «p < 0,05, el fármaco funciona». Mucho más fácil de explicar. Los profesores se copiaron unos a otros, y aquí estamos, setenta años después, enseñándolo todavía.

¿Tiene esto consecuencias prácticas? Ya lo creo.

La crisis de replicación

Hagamos un experimento mental. Imagina que alguien inventa una terapia contra el cáncer y hace un ensayo clínico con 100 pacientes. El valor p no es significativo: la terapia no parece funcionar. Luego otra persona prueba otra terapia. Tampoco es significativo. Ahora imagina que cien grupos de investigación, cada uno con su propia idea, hacen cada uno su propio ensayo. Y supón que, en realidad, ninguna de las terapias funciona en absoluto.

¿Qué pasa? Aproximadamente uno de cada veinte de esos ensayos dará p < 0,05 por pura casualidad. Quizá los pacientes del grupo tratado resultaron ser un poco más jóvenes, o estar más sanos, o hacer más ejercicio. Nadie hizo nada mal. Y esos cinco grupos afortunados publicarán que su terapia funciona.

Ahora añade un segundo efecto. De esos cien grupos, noventa y cinco escriben un artículo que dice «probamos esta molécula y no hace nada». Cinco escriben «probamos esta molécula y cura el cáncer». ¿Qué artículos crees que se publican? Esto es el **sesgo de publicación**: las revistas seleccionan, de entre todos los experimentos que se hicieron, precisamente los que tuvieron suerte.

Junta los dos efectos. Se publican cinco artículos, y los cinco están equivocados. La literatura publicada sobre estas terapias es falsa al 100 %, y nadie hizo trampa.

¿Es una preocupación teórica? En 2005 John Ioannidis publicó un artículo con el provocador título *Why Most Published Research Findings Are False* («Por qué la mayoría de los resultados de investigación publicados son falsos»). Hizo cálculos de este tipo, algo más sofisticados que los míos, y causó un escándalo. Su argumento clave era que la proporción de resultados publicados que son falsos depende de cuántas de las hipótesis que se ponen a prueba eran verdaderas de entrada. Si la mayoría de las ideas que ponemos a prueba son apuestas improbables, como en nuestro experimento mental, la mayoría de los resultados significativos serán chiripas. Esa es una pregunta sobre lo probables que eran las hipótesis *antes* del experimento, y volverá a aparecer cuando hablemos de probabilidades inversas y de probabilidades a priori. También examinó 49 estudios médicos muy citados; 45 habían afirmado que una terapia era eficaz. De ellos, 7 fueron contradichos después por otros estudios, y otros 7 habían afirmado efectos más fuertes que los que se encontraron más tarde.

Y la valoración que hizo Ioannidis del problema resultó ser demasiado optimista. En 2012 Glenn Begley contó que su equipo en una empresa farmacéutica había intentado reproducir 53 estudios preclínicos de referencia sobre el cáncer. Pudieron replicar 6. Un once por ciento. Más tarde, el Reproducibility Project: Cancer Biology completó 50 experimentos de replicación de 23 artículos: solo el 46 % de los efectos se replicaron con éxito, y para los efectos positivos la mediana del tamaño del efecto era un 85 % menor que en los artículos originales.

En psicología, en 2011, Brian Nosek organizó una colaboración para replicar 100 estudios publicados en revistas destacadas. Ya conoces el tipo de resultado que le encanta a la prensa generalista: sostener una taza caliente hace que juzgues a la gente como más cálida, una «postura de poder» hace

que te sientas más seguro de ti mismo. 97 de los 100 artículos originales habían informado de un efecto significativo. Los replicadores acordaron de antemano los métodos con los autores originales, para que nadie pudiera decir que se había hecho mal, y no hay motivo para pensar que los autores originales hicieran trampa. Consiguieron replicar 35. Y, en el conjunto de las replicaciones, los efectos tenían de media la mitad del tamaño publicado originalmente.

No todo el mundo se tomó bien el movimiento de la replicación. En los acalorados debates que lo rodearon, el psicólogo Daniel Gilbert compartió un texto que describía a los replicadores como «la policía de la replicación» y «pequeños matones desvergonzados». Cuestionar el valor p y la manera de hacer ciencia puede enfadar a la gente.

P-hacking: el jardín de los senderos que se bifurcan

Hasta ahora he supuesto que todo el mundo lo hace todo según las reglas. Ahora seamos realistas.

Haces un ensayo con 100 pacientes en el grupo control y 100 en el grupo tratado. Al final, descubres que a un paciente del grupo tratado lo atropelló un autobús. Evidentemente lo quitas del análisis; que te atropelle un autobús no tiene nada que ver con tu fármaco. Y probablemente ni lo mencionas. ¿O… quizá lo atropelló el autobús porque tu fármaco lo mareaba muchísimo?

Una situación muy común en los ensayos clínicos: 20 de tus 100 pacientes tratados dejan de venir a las visitas de seguimiento. ¿Qué haces? ¿Dices que trataste a 80? Eso no es lo que supone la prueba estadística: dijiste que tratarías a 100. Y quizá esos 20 se fueron porque los efectos secundarios eran horribles y no se molestaron en decírtelo.

Los ensayos clínicos bien hechos tienen una respuesta estándar para ambos problemas: el análisis por **intención de tratar**. Analizas a todos los pacientes que aleatorizaste, en el grupo al que fueron asignados, con autobús o sin él, con abandono o sin él. Pero no todos los estudios son tan cuidadosos, y fuera de los ensayos clínicos estas decisiones a menudo se toman sin decir nada.

Otra más. Pruebas un fármaco contra el cáncer de páncreas en 200 personas y no obtienes ningún efecto significativo. Pero espera, en la muestra hay hombres y mujeres. ¿Quizá funcione mejor en las mujeres? Analizas hombres y mujeres por separado. No es significativo en las mujeres, pero sí

en los hombres, así que publicas que el fármaco funciona en los hombres. ¿Es correcto?

No es que los fármacos no puedan funcionar de manera distinta en grupos distintos; muchos lo hacen. El problema es estadístico. Uno de cada veinte es la tasa de error *para una prueba.* Acabas de hacer tres pruebas (todos juntos, hombres, mujeres). Si sigues, niños de 0 a 10 años, adolescentes, adultos de 20 a 40, personas mayores, fumadores, no fumadores… cada prueba es otra tirada del dado de uno entre veinte. Sigue tirando y acabará saliendo. Y ese es el que publicas.

O no decides de antemano cuándo vas a parar el ensayo. Miras los datos cada pocas semanas y paras en cuanto la prueba sale significativa. Es como un partido de fútbol en el que, en lugar de jugar 90 minutos, jugamos hasta que mi equipo va ganando. ¿Quién crees que va a ganar? Hay maneras legítimas de mirar los datos a medida que llegan, llamadas diseños secuenciales, pero hacen el umbral más estricto en cada mirada para compensar. El problema es mirar sin ajustar.

Cuando todo esto se hace buscando un resultado significativo, se llama ***p--hacking***. Pero el estadístico Andrew Gelman señala que el problema aparece incluso cuando nadie está pescando a propósito, y tiene un nombre precioso para ello: **el jardín de los senderos que se bifurcan**. Cada decisión parece inocente, incluso científicamente sensata:

- ¿Excluimos este dato raro? Parece un error de medida, así que quitarlo parece limpiar los datos.
- ¿Controlamos por edad o por sexo? Es plausible que la edad afecte al resultado.
- ¿Transformamos la variable? Tomar el logaritmo hace que la distribución tenga mejor pinta, así que parece mejor estadística.
- Esta prueba no ha salido significativa; ¿probamos otra?
- ¿Cuándo dejamos de recoger datos?

Cada sendero que tomas es otra prueba, aunque al final solo hagas «una». Como dijo la American Statistical Association en una declaración que veremos al final: una inferencia adecuada requiere informar de todo y transparencia. Si los lectores no saben cuántos senderos exploraste, no pueden juzgar tu valor p.

Ahora añade los incentivos. Imagina que terminas la carrera y te pones a trabajar. ¿Se pondrá tu jefe igual de contento si tu experimento muestra que

la cosa no funciona que si muestra que sí? Si eres estudiante de doctorado, un resultado significativo puede marcar la diferencia entre publicar y terminar la tesis o no. Si eres una farmacéutica cuya cotización en bolsa depende de que un nuevo fármaco supere un ensayo clínico, ¿qué incentivo tienes para buscar pacientes «raros» en el grupo tratado que estén hundiendo los resultados? Una empresa podría incluso hacer muchos ensayos pequeños en lugar de uno grande. Un ensayo grande te da un valor p; muchos pequeños te dan muchas tiradas de dados, y es muy fácil que una de ellas salga significativa. La acción sube. Más tarde resulta que el fármaco no funciona, pero mientras tanto alguien ha ganado mucho dinero. Y se ha tratado a gente con fármacos que no funcionaban.

Cuando más datos empeoran las cosas

Hay un problema más. Supón que tu resultado es $p = 0{,}06$. ¡Qué cerca! La idea natural es hacer un estudio más grande, con más pacientes. Más datos significan más potencia para detectar cosas que son verdad, ¿no?

Con los valores p, más datos también te dan más potencia para detectar cosas que no importan. Cuantos más datos tienes, más pequeña es la diferencia que resulta significativa. Con suficientes datos, puedes rechazar cualquier hipótesis nula, porque en el mundo real no hay dos grupos que sean nunca exactamente idénticos. Y si rechazar la nula significa aceptar tu hipótesis, entonces con suficientes datos puedes aceptar cualquier hipótesis que quieras.

Un ejemplo. Un estudio con unos 19.000 participantes se preguntó si las parejas que se habían conocido por internet estaban más satisfechas con su matrimonio que las que se habían conocido fuera de internet. El resultado fue significativo con $p < 0{,}001$. No 0,05: ¡0,001! Ultrasignificativo. ¿Y la diferencia? En la escala de satisfacción, 5,64 frente a 5,48. Aunque esa diferencia sea real, es ridícula. Y puede que ni siquiera sea real, porque con tantos datos cualquier sesgo minúsculo en la metodología (por ejemplo, que la gente que se conoció por internet estuviera un poco más dispuesta a responder un cuestionario en línea) basta para que el resultado sea significativo.

El resultado es estadísticamente significativo pero no significa nada en la práctica. Con los enormes conjuntos de datos que tenemos hoy, esto está por todas partes. Cuantos más datos tienes, más cuidado tienes que tener.

Entonces, ¿por qué lo seguimos usando?

Esa es la pregunta que los estudiantes siempre hacen llegados a este punto. Si el valor p causa tantos problemas, ¿por qué se sigue usando?

Hay varias razones. Una es histórica: los libros de texto están escritos así, y los profesores tienden a seguir los libros. Otra es el temario: explicar todo esto quita tiempo a otros temas. También hay una razón filosófica, que veremos más adelante: hay alternativas, pero algunas personas, personas muy inteligentes, tienen razones de principio para rechazarlas.

Pero creo que la razón principal es que ansiamos la objetividad. Un regulador que aprueba un fármaco necesita decir «había evidencia objetiva de que funciona, la mirara quien la mirara». Imagina a la empresa diciendo «bueno, teníamos algo de evidencia que parecía sugerir que podría funcionar; me salió un 0,04, y sé que no debería tomarlo como la verdad, pero es una evidencia más». Eso no cuela. Así que la presión no va en la dirección de que la gente deje de usar el valor p. Va en la contraria, porque el valor p parece objetivo. Cuando mandas un artículo a una revista, los revisores preguntan: «¿qué análisis estadístico objetivo respalda esta conclusión?». Pero en ciencia esa objetividad no es real: sopesamos distintas evidencias. Cuando un efecto es enorme, cuando una barra es mucho más alta que la otra, todo el mundo está de acuerdo y ni siquiera hace falta una prueba. Cuando las barras son casi iguales y necesitas una prueba sofisticada para distinguirlas, la conclusión honesta es: diga lo que diga el valor p, desconfía mucho.

Se están probando algunos remedios. En los ensayos clínicos hay ahora presión para **prerregistrar**: antes de empezar el ensayo declaras cuántos pacientes vas a reclutar, qué grupos vas a comparar, qué análisis vas a hacer, incluso si vas a mirar por grupos de edad. Luego haces el experimento y rellenas los números. Eso elimina los senderos que se bifurcan, y para los estudios confirmatorios, como un ensayo clínico que va a decidir si se aprueba un fármaco, es una gran idea. Pero el prerregistro no es la respuesta a todos los problemas. Mucha ciencia es exploratoria: «esto ha salido raro, voy a añadir aquí un grupo metilo a ver qué pasa». El prerregistro no prohíbe ese tipo de trabajo; lo separa de las pruebas confirmatorias que deberían venir después. La exploración sigue necesitando criterio, y ningún formulario de registro puede proporcionarlo. No hay un conjunto objetivo

de reglas que garantice la buena ciencia. Eso sería un método científico y, como dijimos, no existe.

Lo que dicen los estadísticos

En 2016 a la American Statistical Association le preocupaba lo suficiente el mal uso que se estaba haciendo de los valores p como para publicar una declaración oficial con seis principios. A estas alturas deberías reconocerlos todos:

- **1.** Los valores p pueden indicar lo incompatibles que son los datos con un modelo estadístico especificado.
- **2.** Los valores p no miden la probabilidad de que la hipótesis estudiada sea verdadera, ni la probabilidad de que los datos se hayan producido solo por azar.
- **3.** Las conclusiones científicas y las decisiones empresariales o políticas no deberían basarse solo en si un valor p supera un umbral determinado.
- **4.** Una inferencia adecuada requiere informar de todo y transparencia.
- **5.** Un valor p, o la significación estadística, no mide el tamaño de un efecto ni la importancia de un resultado.
- **6.** Por sí solo, un valor p no proporciona una buena medida de la evidencia respecto a un modelo o una hipótesis.

Quiero dejar clara una cosa: el valor p no es malvado. Usado como lo usaba Fisher, como una evidencia más que te hace sospechar más o menos de una hipótesis, es perfectamente razonable. El problema es creer que $p < 0{,}05$ significa que el fármaco funciona. Si entiendes lo que estás haciendo, la conclusión es: «aquí hay una evidencia más que parece sugerir que el fármaco funciona».

Pero fíjate en lo que hemos estado haciendo todo el tiempo. Todos los cálculos hasta ahora han sido de la forma «si la hipótesis fuera verdadera, ¿qué probabilidad tendría lo que he visto?». Esa probabilidad tiene nombre, la verosimilitud, y a continuación veremos que se puede usar de otra manera: en lugar de preguntar si los datos son demasiado raros para una hipótesis, podemos preguntar cuál de todas las hipótesis posibles explica mejor los datos.

Lecturas recomendadas

- Wasserstein, R. L. & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129–133.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606. El origen de «el ritual nulo».
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2), 218–228.
- Begley, C. G. & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483, 531–533.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.
- Errington, T. M. et al. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10, e71601.
- Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460–465.
- Cacioppo, J. T. et al. (2013). Marital satisfaction and break-ups differ across on-line and off-line meeting venues. *PNAS*, 110(25), 10135–10140.

Máxima verosimilitud

Hemos dedicado mucho tiempo al valor p y a los problemas que surgen de malinterpretarlo. Pero debajo del valor p había una idea perfectamente sólida, y vamos a seguir usándola: si tengo una hipótesis bien definida, siempre puedo calcular lo que espero ver. También podemos usar esa idea para buscar la hipótesis que mejor explica los datos.

Volvamos a mi moneda imaginaria. La tengo en la mano y, como me la estoy imaginando, puedo decidir que sale cara con probabilidad 0,5. Si lanzara esta moneda imaginaria 10 veces, ¿con qué frecuencia obtendría 0 caras? ¿1 cara? ¿2? … ¿10? Todo eso lo puedo calcular. Si la lanzara 100 veces, ¿qué probabilidad habría de obtener exactamente 50 caras y 50 cruces? Eso también lo puedo calcular. Si mi hipótesis es lo bastante precisa, siempre puedo predecir la probabilidad de lo que observaría en el mundo que me he imaginado.

Esa probabilidad la escribimos así:

p(datos | hipótesis)

La barra vertical se lee «dado» o «suponiendo»: la probabilidad de los datos, suponiendo que la hipótesis es verdadera. Todos los cálculos que hicimos con valores p eran de este tipo: la probabilidad de observar ciertos datos, suponiendo una hipótesis concreta.

Una pregunta distinta

Con el valor p nos preguntábamos: *¿es lo que estoy viendo tan raro que más me vale rechazar la hipótesis?* Lanzo una moneda real 100 veces, obtengo 100 caras y me pregunto lo raro que sería eso para una moneda equilibrada.

La máxima verosimilitud hace una pregunta distinta: *de todas las hipótesis que estoy considerando, ¿cuál explica mejor los datos?*

¿Qué significa «todas las hipótesis»? Piensa en todas las monedas imaginarias que podría tener en la mano. Podría imaginar una moneda que siempre sale cara. U otra que siempre sale cruz. U otra que sale cara con probabilidad 0,5, o 0,9, o 0,8, o 0,1234. Ahora supón que cojo una moneda real, la lanzo 100 veces y obtengo 100 caras. De todas esas monedas imaginarias, de todos esos modelos distintos, ¿cuál hace más probable obtener 100 caras de 100?

Una moneda con una probabilidad de 0,1 de salir cara podría producir 100 caras; sería posible, solo que extraordinariamente improbable. Si la moneda tuviera una probabilidad de 0,9, 100 caras serían mucho más plausibles. Y con una moneda que siempre sale cara, probabilidad 1, obtener 100 caras no solo es plausible: es seguro. Así que la moneda que mejor explica mis datos es la que siempre sale cara.

Eso es la máxima verosimilitud. De todas las hipótesis que estamos considerando, elige la que hace más probables los datos que has observado. Fíjate en lo de «que estamos considerando». La máxima verosimilitud no inventa hipótesis: elige entre las que nosotros proponemos. Como explicamos antes, los modelos no salen solo de los datos. Decidimos de antemano qué tipo de modelo tiene sentido, y luego los datos eligen la mejor versión de ese modelo.

Verosimilitud

La probabilidad de los datos suponiendo una hipótesis se llama **verosimilitud**. (En sentido estricto, no siempre es una probabilidad: para datos continuos usamos densidades de probabilidad. Pero la idea es la misma.)

Si supongo que las vacas vuelan, ¿cuál es la verosimilitud de salir al campo y ver una vaca volando? Si supongo que no vuelan, ¿cuál es la verosimilitud de ver una vaca volando? Ninguna. Si supongo una moneda equilibrada, ¿cuál es la verosimilitud de obtener dos caras en dos lanzamientos? Esa también es fácil: 0,5 del primer lanzamiento por 0,5 del segundo, $0{,}5^2 = 0{,}25$.

Ahora fíjate en una cosa. Ese 0,25 depende del 0,5, la probabilidad de salir cara. Si mi moneda imaginaria tuviera una probabilidad de 0,4, la verosimilitud de dos caras sería $0{,}4^2 = 0{,}16$. Si fuera 0,9, sería 0,81. Si la moneda nunca saliera cara, la probabilidad de obtener dos caras sería 0. Y si siempre saliera cara, sería 1.

Así que dejemos de escribir 0,5 y llamemos h a la probabilidad de cara. Ahora el modelo tiene un *parámetro*, y la verosimilitud es una función de

ese parámetro. Para dos caras observadas, es h^2. Eso se llama *función de verosimilitud*:

$$L(\theta) = f(\text{datos} \mid \theta)$$

donde θ representa los parámetros del modelo (aquí, θ es simplemente h), y los datos se mantienen fijos. Una vez hecho el experimento, los datos son simplemente lo que observamos; lo que variamos es la hipótesis.

Encontrar el máximo

Supón que lanzo una moneda tres veces y obtengo tres caras. La función de verosimilitud es h^3. Si $h = 0{,}5$, vale 0,125. Si $h = 0{,}1$, vale 0,001: muy improbable. Si $h = 0{,}9$, vale 0,729. Si represento h^3 para todos los valores de h entre 0 y 1 y busco el punto más alto, lo encuentro en $h = 1$. Con esos datos, tres caras en tres lanzamientos, la estimación de máxima verosimilitud es que la moneda siempre sale cara.

¿Apostarías a que esta moneda siempre sale cara? Probablemente no. Sabes cómo suelen comportarse las monedas, y tres lanzamientos no son mucha evidencia. Pero la máxima verosimilitud solo mira los datos, no lo que ya sabemos, así que con pocos datos puede pecar de exceso de confianza. Ten presente esa incomodidad; volveremos a ella más adelante.

Ahora supón que la lanzo una cuarta vez y sale cruz: tres caras y una cruz. La función de verosimilitud es ahora $h^3(1 - h)$. Si $h = 1$, la verosimilitud ahora es cero: una moneda que siempre sale cara nunca podría producir esa cruz. Si represento la función y busco el pico, está en 0,75.

Eso es lo que hacen los programas estadísticos cuando calculan estimaciones de máxima verosimilitud: recorren los valores posibles de los parámetros y eligen el que hace más probables las observaciones.

A veces podemos hacerlo analíticamente. Encontrar el máximo de una función es un problema de cálculo: derivas, igualas a cero y despejas. Para la moneda, la probabilidad de obtener x caras en n lanzamientos es la binomial:

$$p(x \mid h, n) = n! / (x!\,(n - x)!) \cdot h^x\,(1 - h)^{n-x}$$

Derivas con respecto a h, igualas a cero y obtienes:

$$\hat{h} = x / n$$

El sombrerito sobre la h es como los estadísticos marcan una estimación de máxima verosimilitud. Cuando veas un sombrero en un artículo o en un libro

de texto, eso es lo que significa. Y el resultado es: cuenta las caras y divide por el número de lanzamientos. Exactamente lo que te decía la intuición. En este caso, toda esta maquinaria estadística parece matar moscas a cañonazos.

En muchos modelos complicados la derivada no se puede calcular, así que los programas buscan numéricamente: prueban valores, miran si la verosimilitud sube o baja y siguen subiendo hasta llegar a la cima.

Cuando falla la intuición

Para la moneda, la máxima verosimilitud da la respuesta obvia. ¿Es siempre así de aburrida? No.

Imagina que queremos saber cuánto viven, de media, los pacientes después de que se les diagnostique un tipo concreto de cáncer. Esto importa mucho a los pacientes y a sus familias, a los médicos que eligen un tratamiento y a cualquiera que compare una terapia nueva con la antigua. Así que seguimos a 10 pacientes desde el momento del diagnóstico. No podemos esperar eternamente, así que el estudio termina a los cinco años. Para entonces:

- 6 pacientes han muerto, a los 0,5, 1, 1,5, 2,5, 3 y 4,5 años,
- 4 pacientes siguen vivos.

¿Cuál es el tiempo medio de supervivencia tras el diagnóstico?

La respuesta intuitiva es promediar los tiempos que tenemos: $(0{,}5 + 1 + 1{,}5 + 2{,}5 + 3 + 4{,}5) / 6 \approx 2{,}2$ años. Pero eso está mal, y es demasiado pesimista. ¿Qué hacemos con los cuatro pacientes que siguen vivos? No podemos ignorarlos sin más, porque nos están diciendo algo importante: algunos pacientes viven más de cinco años. Y no podemos asignarles el valor cinco, porque sabemos que viven más.

A esto se le llama datos *censurados*, y están por todas partes: pacientes que siguen vivos al final de un estudio, máquinas que aún no han fallado, plantas que no han florecido cuando se acaba el experimento.

La máxima verosimilitud maneja esto con naturalidad. Proponemos un modelo de cómo se distribuyen los tiempos de supervivencia y preguntamos: ¿qué tiempo medio de supervivencia hace más probables nuestras observaciones? Y «nuestras observaciones» incluyen ambos hechos: seis pacientes murieron en esos momentos concretos, y cuatro seguían vivos al cabo de cinco años.

Un modelo sencillo es la distribución exponencial. Tiene un parámetro, la tasa λ, y el tiempo medio de supervivencia es $1/\lambda$. Supone que el riesgo de morir se mantiene igual año tras año. Es una simplificación, pero basta para mostrar la idea. La verosimilitud tiene dos partes:

- para cada paciente que murió en el tiempo t_i, la densidad de probabilidad de morir exactamente entonces: $\lambda e^{(-\lambda t_i)}$;
- para cada paciente que seguía vivo en $T = 5$ años, la probabilidad de sobrevivir al menos ese tiempo: $e^{(-\lambda T)}$.

Multiplicas todas, tomas el logaritmo, derivas, igualas a cero y obtienes:

tiempo medio de supervivencia = (suma de los tiempos de supervivencia observados + años de seguimiento de los pacientes que siguen vivos) / número de pacientes que murieron = $(13 + 4 \times 5) / 6 = 33 / 6 =$ **5,5 años**

La estimación de máxima verosimilitud de la media es 5,5 años: más del doble de la media ingenua, y más allá del final de nuestro estudio. Tiene sentido, ya que casi la mitad de los pacientes seguían vivos al cabo de cinco años.

Comparar modelos

Hay otra cosa más que podemos hacer con las verosimilitudes. Una vez que tenemos las estimaciones de máxima verosimilitud de los parámetros de un modelo, podemos calcular la verosimilitud de las observaciones bajo ese modelo. Así que, si tenemos varios modelos alternativos, podemos repetir el proceso con cada uno: ajustar sus parámetros por máxima verosimilitud y luego calcular lo probables que son los datos bajo la mejor versión de ese modelo. Después elegimos el modelo con mayor verosimilitud.

Aunque hay una trampa. Un modelo con más parámetros siempre se ajustará a los datos al menos tan bien como uno más sencillo, así que, si solo miráramos la verosimilitud, siempre ganaría el modelo más complejo, incluso cuando sus parámetros adicionales solo están ajustando ruido. Esto se llama *sobreajuste*. En la práctica usamos criterios como el AIC o el BIC, que tienen en cuenta la verosimilitud pero penalizan cada parámetro adicional.

La verosimilitud como evidencia

Las verosimilitudes también pueden usarse para comparar directamente dos hipótesis concretas. Vuelve a las tres caras en tres lanzamientos. Una moneda que siempre sale cara da ese resultado con probabilidad 1; una moneda

equilibrada, con probabilidad 0,125. La razón es 8: los datos favorecen a la moneda que siempre sale cara frente a la equilibrada por un factor de 8.

Algunos estadísticos, los *verosimilitudistas*, piensan que esta razón es la mejor respuesta a la pregunta «¿qué dicen los datos?». Mide la evidencia, y es la misma para todo el mundo, crea lo que creyera antes cada cual. Acuérdate de la apuesta: probablemente seguirías sin apostar a que la moneda siempre sale cara. Para un verosimilitudista eso no es ninguna contradicción, porque lo que crees, o lo que decides hacer, son preguntas distintas de lo que dicen los datos. Volveremos a estas tres preguntas. El enfoque funciona mejor cuando solo hay dos hipótesis precisas que comparar; con más, las cosas se complican.

¿Qué hemos calculado en realidad?

Parémonos un segundo a pensar en lo que hemos hecho. Queríamos saber la probabilidad de que la moneda salga cara, dado lo que hemos observado. Pero ¿qué hemos calculado? La probabilidad de lo que hemos observado, dada cada moneda posible. Y luego hemos elegido la moneda que hacía más probable nuestra observación.

No es lo mismo. «¿Qué hipótesis hace más probables los datos?» es una pregunta distinta de «¿qué hipótesis es más probable, dados los datos?». Quizá pienses que esto es un simple juego de palabras. No lo es. Como veremos, confundir esas dos probabilidades puede llevarte a creer que tienes el VIH cuando probablemente no lo tienes, y puede mandar a la cárcel a una persona inocente.

La probabilidad inversa: la probabilidad que de verdad queremos

Lo que de verdad queremos saber

Imagina que queremos saber **si un fármaco funciona**, a la vista de los datos experimentales. ¿Qué probabilidad calcularíamos con el valor p? La probabilidad de observar que cierto número de personas mejora *si suponemos que el fármaco no funciona.*

Cuando usamos la máxima verosimilitud para estimar la probabilidad de que una moneda salga cara, a partir de lo que la habíamos visto hacer, ¿qué calculamos? La probabilidad de observar las caras y cruces que observamos *si suponemos que la moneda tiene una determinada probabilidad de salir cara.*

Cuando un médico quiere saber si tienes una enfermedad, a la vista de los resultados de tus pruebas, ¿qué se puede calcular fácilmente? La probabilidad de obtener esos resultados *si tienes la enfermedad.* Si estoy resfriado, ¿qué probabilidad hay de que tenga tos y mocos? Alta. Ahora bien, si tengo tos y mocos, ¿qué probabilidad hay de que esté resfriado? Esa es otra pregunta. Quizá sea una alergia. Quizá sea la gripe. Quizá sea otra cosa.

¿Ves el patrón? En la estadística clásica siempre hemos trabajado con

p(evidencia | hipótesis)

pero lo que de verdad queremos es

p(hipótesis | evidencia)

Parecen lo mismo escrito al revés, y nuestra mente tiende a tratarlas como más o menos equivalentes. Pero no lo son, y pueden ser tremendamente distintas. Cuando pienso «tengo tos y estornudo, así que estoy resfriado»,

estoy razonando sobre p(resfriado | síntomas). Pero toda la estadística que hemos visto hasta ahora solo me da p(síntomas | resfriado).

Has dado positivo en VIH

Vas paseando por la calle Mayor, hay una campaña de concienciación y aceptas hacerte una prueba como parte de un estudio sobre la prevalencia del VIH en tu comunidad. Te sacan sangre. Unos días después te llaman: la prueba ha dado positivo.

También te dicen que la prueba es bastante buena. Su tasa de falsos positivos es muy pequeña. Esa es la probabilidad de dar positivo si no estás infectado: p(positivo | no infectado).

Entonces, ¿qué probabilidad hay de que estés infectado?

Cuando se hace esta pregunta a la gente, la mayoría, incluidos muchos médicos, responde algo como «muy alta, quizá del 98 % o el 99 %». Al fin y al cabo, la prueba casi nunca da positivo en alguien que no está infectado.

Sin embargo, la probabilidad de estar infectado no se puede calcular con la información que te han dado. Nos falta algo esencial. Y, como veremos, lo más probable es que no estés infectado, aunque hayas dado positivo.

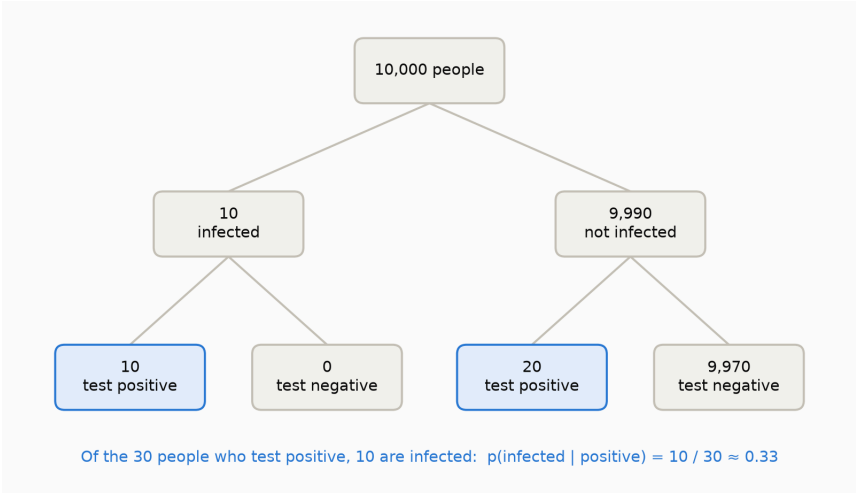
Piensa en todas las personas que participaron en la campaña. Si nadie hubiera estado infectado, unas cuantas habrían dado positivo de todos modos: los falsos positivos. Si todo el mundo hubiera estado infectado, casi todas habrían dado positivo. Lo que quieres saber es: *de todas las personas que dieron positivo, ¿cuántas están infectadas?* Y eso depende de cuántas personas infectadas había para empezar.

Pongamos algunos números realistas. Supongamos que:

- la prevalencia del VIH en la población es de alrededor del 0,1 %: una persona de cada mil,
- la probabilidad de un falso positivo es del 0,2 %,
- la probabilidad de un falso negativo es muy pequeña, así que la ignoraremos.

Toma 10.000 personas al azar y hazles la prueba a todas.

- 10 de ellas están infectadas (una de cada mil). Todas dan positivo.
- 9.990 no están infectadas. De ellas, el 0,2 % da positivo de todos modos: unos 20 falsos positivos.



Así que hay unos 30 resultados positivos, y solo 10 corresponden a personas infectadas. Cuando te dicen que has dado positivo, eres uno de esos 30. La barra vertical en p(infectado | positivo) significa exactamente eso: ahora nos limitamos al grupo de 30. Y la probabilidad de estar infectado es de 10 entre 30: alrededor de 0,3. Una prueba con una tasa de falsos positivos del 0,2 % y, aun así, si das positivo, es el doble de probable que *no* estés infectado.

Ahora cambia una sola cosa: la prevalencia. Imagina una población en la que el 1 % de las personas está infectada. De 10.000, 100 están infectadas y dan positivo. De las 9.900 que no lo están, unas 20 dan positivo de todos modos. Ahora, si das positivo, la probabilidad de estar infectado es de 100 entre 120: alrededor de 0,83. La misma prueba, la misma tasa de falsos positivos, y una conclusión radicalmente distinta.

Por eso también importa que te hicieran la prueba al azar, paseando por la calle. Si te la hubieras hecho porque habías estado expuesto al virus, pertenecerías a un grupo en el que la infección es mucho más frecuente, y un resultado positivo sería mucho más preocupante. Y por eso, en la práctica, un positivo en una prueba de cribado del VIH siempre se confirma con una segunda prueba distinta.

Es tan fácil equivocarse con esto que el error tiene nombre: la falacia de la tasa base. En este caso, la tasa base es la prevalencia: lo frecuente que era aquello antes de mirar ninguna evidencia. Si la ignoras, tus conclusiones pueden ser un disparate.

La falacia del fiscal: el caso de Sally Clark

Cuando el mismo error ocurre en un tribunal, recibe otro nombre: la falacia del fiscal. En un juicio se presentan pruebas y el jurado debe decidir: a la vista de estas pruebas, ¿es culpable el acusado? La falacia consiste en argumentar que, como las pruebas serían muy improbables si el acusado fuera inocente, es muy improbable que el acusado sea inocente. Una vez más: p(inocente | evidencia) no es lo mismo que p(evidencia | inocente). Veamos un caso real.

En 1996, Sally Clark, una abogada inglesa, dio a luz a un niño aparentemente sano. Menos de tres meses después, el niño murió. Ella estaba sola con él; dijo que dejó de respirar después de acostarlo. Por desgracia, a veces los bebés mueren así, de repente y sin ninguna causa aparente. Se llama síndrome de muerte súbita del lactante, SMSL.

En 1997 tuvo otro bebé. Murió antes de cumplir ocho semanas, en circunstancias parecidas. Esta vez el bebé mostraba señales de traumatismo. Clark dijo que el traumatismo se debía a que ella y los sanitarios habían intentado reanimarlo.

La acusaron de asesinar a sus dos hijos. Y no había muchas más pruebas que estas. Así que llamaron a declarar a un pediatra, el profesor Roy Meadow, para que explicara lo probable que era que aquello hubiera ocurrido por azar. Su razonamiento fue más o menos este:

- La incidencia del SMSL en una familia sana y no fumadora como los Clark es de 1 entre 8.543.
- Por tanto, la probabilidad de dos muertes por SMSL en la misma familia es de 1 entre 8.543 × 8.543: **1 entre 72.982.849**. Una entre 73 millones.

Antes de seguir leyendo, ponte en el lugar del jurado. ¿Es correcto el cálculo? Y, a la vista de estas pruebas, ¿es culpable?

Fallo 1: el propio número

La cifra de 1 entre 8.543 se había obtenido escogiendo el subgrupo de menor riesgo. La tasa estimada de SMSL en aquella época rondaba 1 entre 1.300. Con ese número, el mismo razonamiento da alrededor de 1 entre 1,7 millones. Sigue siendo pequeño, pero 40 veces mayor.

Fallo 2: las cosas raras pasan continuamente

Supón que algo ocurre una vez de cada millón. Si juegas un millón de veces, ¿ocurrirá? Muy probablemente. Si juegas dos millones de veces, todavía más probablemente. Ahora bien, ¿cuántas familias con dos hijos hay en el Reino

Unido? Millones. Así que algo con una probabilidad de uno entre un millón le ocurrirá a alguna familia, en algún sitio, de vez en cuando. Que le ocurriera a *esta* familia es mucho menos sorprendente de lo que parece.

Fallo 3: la independencia

¿Cuándo podemos multiplicar probabilidades? Si tiro un dado, la probabilidad de sacar un seis es 1/6. Si lo tiro dos veces, la probabilidad de sacar dos seises es 1/6 × 1/6 = 1/36. Esto funciona porque las dos tiradas son independientes: la primera no influye en la segunda.

p(A y B) = p(A) × p(B), solo si A y B son independientes.

¿Eran independientes las dos muertes? Si hay un factor genético que predispone al SMSL, o uno ambiental compartido por el hogar, la primera muerte hace más probable la segunda. Una familia que ya ha perdido un hijo por SMSL no es una familia cualquiera. Si las muertes no son independientes, la probabilidad de una segunda muerte podría ser de 1 entre 100, o de 1 entre 1.000. No lo sabemos, porque no entendemos lo bastante bien las causas del SMSL. Pero desde luego no podemos suponer la independencia, y el cálculo de Meadow lo hizo. Como vimos antes, todo modelo tiene supuestos, y la independencia es uno muy común. Aquí, ignorar uno ayudó a mandar a la cárcel a una mujer inocente.

Fallo 4: la probabilidad inversa ataca de nuevo

Pero el problema más profundo persiste aunque arreglemos todo lo anterior. Supongamos que todo se hubiera calculado correctamente. ¿Qué se ha calculado?

p(dos muertes | SMSL)

La probabilidad de que mueran dos bebés, suponiendo que murieron por causas naturales. ¿Es eso lo que necesita el jurado? No. El jurado necesita

p(asesinato | dos muertes) = 1 − p(SMSL | dos muertes)

Dado que han muerto dos bebés, ¿qué probabilidad hay de que los matara su madre? Y p(SMSL | dos muertes) no es p(dos muertes | SMSL). Es el mismo tipo de problema que teníamos con la prueba del VIH.

Piensa en lo que haría falta para responder a la pregunta de verdad. Hay dos hipótesis sobre la mesa: los bebés murieron por causas naturales, o la madre los mató. Un doble SMSL es extremadamente raro. Pero que una madre asesine a dos de sus bebés también es extremadamente raro. La evidencia (dos bebés muertos) es muy improbable con ambas hipótesis, así que la

pregunta es cuál de las dos explicaciones, ambas muy raras, es menos rara. Demostrar que una explicación es muy improbable no te dice nada sobre la culpabilidad a menos que también tengas en cuenta lo improbable que es la otra.

Podemos contar, como hicimos con la prueba del VIH. En 2004, el matemático Ray Hill hizo esta comparación usando las tasas de familias como los Clark. Estimó que el doble SMSL es entre 4,5 y 9 veces más frecuente que el doble asesinato. Así que piensa en todas las familias como los Clark que pierden dos bebés. De cada 10, aproximadamente 8 o 9 los perdieron por SMSL, y solo 1 o 2 por asesinato. Las dos muertes por sí solas hacían posible la culpabilidad de Sally Clark, pero no probable: entre 1 de cada 10 y 1 de cada 5. Muy lejos de una entre 73 millones.

Sally Clark fue condenada en 1999. La Royal Statistical Society emitió un comunicado en el que criticaba el mal uso de la estadística en su juicio. En 2003 su condena fue anulada en apelación. Había salido a la luz que se habían ocultado pruebas de una infección en su segundo hijo. Nunca se recuperó de aquella experiencia, y murió en 2007.

Lo que de verdad queremos

Todo lo que hemos visto en la estadística clásica, desde el valor p hasta la máxima verosimilitud, se basa en p(datos | modelo). Podemos calcularla siempre que tengamos un modelo bien definido: imaginamos cómo funciona el mundo y calculamos lo que esperaríamos ver.

Pero lo que de verdad queremos saber es si deberíamos fiarnos del modelo, a la vista de los datos que hemos observado: p(modelo | datos). Es decir, la probabilidad de que el fármaco funcione, de que la moneda sea equilibrada, de que el paciente esté enfermo o de que el acusado sea culpable.

Los estadísticos solían llamar a esto *probabilidad inversa*; hoy se suele llamar *probabilidad a posteriori*. Tiene su gracia, si lo piensas: llamaron «inversa» precisamente a la probabilidad que nos interesa. ¿Inversa respecto a qué? Respecto a la que es fácil de calcular a partir del modelo.

Entonces, ¿cómo podemos calcular la probabilidad de una hipótesis a la vista de la evidencia, cuando lo único que sabemos calcular es la probabilidad de la evidencia dada la hipótesis? Ya hemos visto una pista en el ejemplo del VIH: necesitábamos un ingrediente más, la prevalencia: lo probable que era la hipótesis *antes* de ver la evidencia. Hay un teorema que nos dice

exactamente cómo combinar esos ingredientes. Como veremos, usarlo con este fin desencadenó una de las peleas más enconadas de la historia de la estadística.

Lecturas recomendadas

- Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You*. Simon & Schuster. (Publicado en el Reino Unido como *Reckoning with Risk*.)
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53–96.
- Royal Statistical Society (2001). Royal Statistical Society concerned by issues raised in Sally Clark case. Nota de prensa, 23 de octubre de 2001.
- Hill, R. (2004). Multiple sudden infant deaths: coincidence or beyond coincidence? *Paediatric and Perinatal Epidemiology*, 18(5), 320–326.

La inferencia bayesiana

Hemos llegado a una conclusión incómoda. Toda la estadística clásica que hemos visto calcula p(datos | hipótesis), pero normalmente lo que nos interesa es p(hipótesis | datos): la probabilidad de que el fármaco funcione, de que el paciente esté enfermo, de que el acusado sea culpable, a la vista de lo que hemos observado. ¿Cómo pasamos de una a otra? La respuesta está en un teorema sencillo.

Un pastor y su amigo

En 1763 se publicó un teorema que nos permite calcular la probabilidad que de verdad queremos. Thomas Bayes (1701–1761) era un pastor presbiteriano aficionado a las matemáticas. Tras la muerte de Bayes, su amigo Richard Price, también pastor, encontró entre sus papeles un ensayo sobre cómo razonar desde los sucesos observados hacia sus causas. Price lo editó y lo publicó en 1763. Ese es el origen de lo que hoy llamamos teorema de Bayes, aunque el ensayo no se parece a la fórmula que escribimos hoy. Está escrito con palabras y geometría.

¿Por qué les interesaba esto a dos pastores? A Price le interesaba pensar racionalmente sobre la existencia de Dios, y el ensayo ofrecía un método riguroso para razonar desde la evidencia hacia las causas: desde lo que observamos en el mundo, como el orden y el diseño que vemos en la naturaleza, hacia lo que lo produjo. Piensa en un gato: garras, colmillos, reflejos rápidos, perfectamente construido para cazar. ¿Qué probabilidad hay de que alguien lo diseñara, a la vista de lo que vemos? Ese es el tipo de pregunta que estas nuevas matemáticas parecían responder. Así que el teorema que está en el corazón de la inferencia bayesiana nació, en parte, de la teología.

Quien de verdad lo convirtió en una herramienta de trabajo fue Pierre-Simon Laplace, el gran matemático francés. Llegó al teorema de forma

independiente en 1774, escribió las fórmulas en la forma que usamos hoy y las aplicó a problemas estadísticos. Uno de los primeros análisis bayesianos de la historia fue suyo: a la vista del número observado de niños y niñas nacidos, ¿qué probabilidad hay de que la probabilidad subyacente de que nazca un niño sea mayor que un medio? Encontró que estaba tan cerca de 1 que, en la práctica, era una certeza. Laplace era un pragmático, y también usaba lo que hoy llamaríamos métodos frecuentistas cuando le convenía.

El teorema de Bayes

$p(A \mid B) = p(A) \cdot p(B \mid A) / p(B)$

Ese es el teorema de Bayes. Es muy fácil de demostrar; se deduce casi directamente de la definición de probabilidad condicionada. Y fíjate en lo que hace: tiene $p(A \mid B)$ a un lado y $p(B \mid A)$ al otro. Es precisamente el puente entre una probabilidad y su inversa.

Ahora sustituye A y B por hipótesis y datos:

$p(\text{hipótesis} \mid \text{datos}) = p(\text{hipótesis}) \cdot p(\text{datos} \mid \text{hipótesis}) / p(\text{datos})$

Eso es la *inferencia bayesiana*. A la vista de la evidencia, ¿qué probabilidad debo dar a mi hipótesis? Dado que la radiografía muestra una sombra, ¿qué probabilidad hay de que mi paciente tenga cáncer? Dado que han muerto dos bebés, ¿qué probabilidad hay de que los matara su madre?

Aquí hay un matiz importante. Nadie discute la primera fórmula. Es un teorema. Pero la segunda, la que sustituye A y B por hipótesis y datos, se ha discutido ferozmente a lo largo de todo el siglo XX, y todavía hoy hay gente que no está contenta con ella.

Las piezas de la fórmula

Ya hemos visto *la verosimilitud*, $p(\text{datos} \mid \text{hipótesis})$. Si supongo que la hipótesis es cierta, si supongo que mi paciente tiene cáncer, ¿qué probabilidad hay de que observe los datos que he observado?

La probabilidad de los datos, $p(\text{datos})$, es la probabilidad de ver lo que hemos visto, teniendo en cuenta todas las hipótesis. Si la madre es una asesina, ¿qué probabilidad hay de que mueran los bebés? ¿Y si no lo es? Si mi paciente está enfermo, ¿qué probabilidad hay de que vea una sombra en la radiografía? ¿Y si está sano? Ponderamos cada hipótesis según lo mucho que creíamos en

ella y lo sumamos todo. (En la inferencia bayesiana siempre consideramos todas las hipótesis, no solo una, como hacíamos con la hipótesis nula.)

El cociente entre ambas, p(datos | hipótesis) / p(datos), nos dice con cuánta fuerza favorecen los datos a esta hipótesis frente a todas las demás. Llamémoslo la *fuerza de la evidencia*. Aquí va un ejemplo de evidencia sin ninguna fuerza. Supón que traigo una botella de agua de mi casa y te digo: esta es el agua de mi grifo, y si te la bebes, vivirás muchos años. Te la bebes, esperamos y, efectivamente, la mayoría de las personas que la beben viven muchos años. ¿Deberíamos concluir que el agua de mi grifo es milagrosa?

No. ¿Por qué no? Porque esas personas también habrían vivido muchos años sin ella. Las dos hipótesis, «el agua es milagrosa» y «el agua es solo agua», predicen lo mismo. Cuando calculamos el cociente, sale aproximadamente 1: estos datos no apoyan una hipótesis más que la otra. Así que me quedo con lo que creía antes de empezar. Y antes de empezar, ¿qué probabilidad le daba a que el agua de mi grifo fuera milagrosa? No conozco muchas aguas milagrosas. Así es exactamente como funciona la homeopatía. «Tómate esto para el resfriado y te pondrás mejor.» Y te pones mejor. Pero también te habrías puesto mejor sin ello.

La probabilidad a priori

Y así llegamos a *la probabilidad a priori*: p(hipótesis). Es la probabilidad de la hipótesis antes de tener en cuenta los datos.

No es opcional. Si quiero saber qué probabilidad hay de que mi paciente esté enfermo después de ver los resultados de las pruebas, entonces, antes de que el paciente entre en la consulta, ya necesito una probabilidad de que esté enfermo. Antes de ir al juicio de Sally Clark, necesito una probabilidad de que sea una asesina. Eso es lo que faltaba en el ejemplo del VIH: la prevalencia era la probabilidad a priori.

Para ver por qué la verosimilitud sola no basta, aquí va un ejemplo. Llegas a casa de un amigo y te encuentras una ventana rota y una piedra en el suelo. ¿Qué ha pasado? Propongamos dos hipótesis:

1. Un elefante estaba jugando en el jardín y rompió la ventana. Luego, para encubrir al elefante, su amigo, un marciano, puso la piedra en el suelo y los dos salieron corriendo.
2. El niño que vive en la casa estaba jugando a tirar piedras en el jardín, y una rompió la ventana sin querer.

¿Cuál es la verosimilitud de la hipótesis 1? Si suponemos que lo hicieron el elefante y su amigo marciano, ¿qué probabilidad hay de que encontremos una ventana rota y una piedra en el suelo? Muy alta: la hipótesis predice exactamente eso. ¿Y la verosimilitud de la hipótesis 2? También alta. Las verosimilitudes son parecidas. Los datos no distinguen entre ellas.

Y, sin embargo, nadie duda ni un segundo. ¿Por qué? Lo que marca la diferencia no es la verosimilitud, sino la probabilidad a priori: ¿qué probabilidad hay de que un elefante con un amigo marciano esté jugando en el jardín de tu amigo, comparada con la de un niño tirando piedras? La evidencia por sí sola no puede zanjar la cuestión. Lo que la zanja es lo que sabíamos antes. Y hasta que nos encontramos con las probabilidades inversas, todo lo que hacíamos en estadística se basaba en la verosimilitud, en la probabilidad de los datos dada una hipótesis. Si lo que quieres es la probabilidad de una hipótesis, esa es una laguna fundamental en la mayor parte de la estadística que te enseñaron: ignora lo probables que eran las hipótesis para empezar.

Así que la inferencia bayesiana tiene tres partes:

- *Grado de creencia a priori*, p(hipótesis): lo que creía antes.
- *Fuerza de la evidencia*, p(datos | hipótesis) / p(datos): cuánto favorecen los datos a cada hipótesis.
- *Grado de creencia a posteriori*, p(hipótesis | datos): lo que debería creer ahora.

Un médico hace esto constantemente: prevalencia de la enfermedad (probabilidad a priori) + resultados de las pruebas (evidencia) ⟶ diagnóstico (probabilidad a posteriori).

Pero hay un precio que pagar. Para hacer todo esto, tenemos que aceptar que nuestro grado de creencia en una hipótesis se puede representar con una probabilidad. Nadie discute la verosimilitud: si imagino una moneda con una probabilidad de cara de 0,5, puedo calcular la probabilidad de sacar diez caras en diez lanzamientos. Pero decir que una *hipótesis* tiene una probabilidad es algo que sí se discute. Los estadísticos conocidos como frecuentistas lo rechazaron. Es una postura legítima, y más adelante veremos qué proponen en su lugar quienes la defienden.

¿Qué obtenemos a cambio de pagar ese precio? Un razonamiento correcto con premisas inciertas. El teorema de Bayes no te da buenas creencias; hace avanzar las que tienes. Si tus grados de creencia de partida siguen las reglas de la probabilidad y los actualizas con el teorema de Bayes, tus nuevos

grados de creencia también seguirán esas reglas. Tus creencias seguirán encajando entre sí o, como dicen los filósofos, seguirán siendo coherentes. No acabarás sosteniendo creencias que se contradicen entre sí.

Y los nuevos grados de creencia no solo serán coherentes: habrán tenido en cuenta la evidencia de la manera correcta, cada pieza exactamente con el peso que merece. Ni siquiera importa en qué orden llegue la evidencia; acabas en el mismo sitio. De hecho, si tus grados de creencia son probabilidades, el teorema de Bayes es la única forma de actualizarlos que se mantiene consistente. Como en la lógica, no hay garantía de que tus conclusiones sean verdaderas, pero habrás llegado a ellas mediante inferencias válidas. También volveremos sobre esto.

Todo esto depende de aceptar que los grados de creencia deben ser probabilidades. Algunos estadísticos no lo aceptan, y volveremos sobre ello.

La apuesta de la moneda

Hagamos un ejemplo completo. Un amigo te propone una partida de ajedrez y, para decidir quién juega con blancas (las blancas mueven primero, así que tienen ventaja), lanza una moneda: si sale cara, él juega con blancas. Te huele un poco raro, así que le pides que la lance tres veces más. Cara, cara, cara. Cuatro caras seguidas.

¿Qué probabilidad hay de que la moneda no sea equilibrada?

Un razonamiento tentador: si la moneda fuera equilibrada, la probabilidad de sacar cuatro caras sería $0{,}5^4 = 0{,}0625$, así que hay 1 posibilidad entre 16 de que la moneda sea equilibrada. ¿Qué te parece este razonamiento?

Es otra vez el error de la probabilidad inversa. 0,0625 es p(4 caras | moneda equilibrada), no p(moneda equilibrada | 4 caras).

Veámoslo con cada uno de los enfoques que hemos visto.

Con un valor p: supongamos que la moneda es equilibrada. Cuatro caras en cuatro lanzamientos no es tan raro; ocurre más de una vez de cada veinte. No podemos rechazar la hipótesis de que la moneda es equilibrada. Con cinco caras todavía no es lo bastante raro (p bilateral $\approx 0{,}06$); con seis, sí ($p \approx 0{,}03$). Pero incluso entonces, lo único que podemos decir es que rechazamos la moneda equilibrada. Nunca obtenemos una probabilidad de nada.

Con máxima verosimilitud: ¿qué probabilidad de cara hace más probable obtener cuatro caras de cuatro? Una moneda que siempre sale cara. La estimación de máxima verosimilitud de la probabilidad de cara es 1.

Con Bayes: primero tenemos que plantear las hipótesis y sus probabilidades a priori. «Moneda equilibrada» no basta: ¿cuáles son las alternativas? Probemos varias opciones.

Opción a: todas las monedas son igual de probables. Antes de ir a casa de mi amigo no tengo ni idea, así que supongo que cualquier probabilidad de cara entre 0 y 1 es igual de probable. Es una probabilidad a priori plana: no representa ningún conocimiento previo. Si calculamos la probabilidad a posteriori con esta probabilidad a priori… tiene exactamente la misma forma que la verosimilitud. ¿Por qué? Porque la probabilidad a priori es la misma para todas las hipótesis, así que no cambia nada: la probabilidad a posteriori es simplemente la verosimilitud, normalizada. La conclusión es la misma que con la máxima verosimilitud.

Opción b: la mayoría de las monedas son equilibradas. Antes de ir a casa de tu amigo, ¿de verdad crees que una moneda que siempre sale cara es tan probable como una moneda normal? Si tuvieras que apostar 100 euros, ¿apostarías por eso? Claro que no. La mayoría de las monedas del mundo son bastante equilibradas. Quizá no exactamente 0,5, pero casi. Así que usemos una probabilidad a priori concentrada alrededor de 0,5. Ahora la verosimilitud sigue siendo la misma (los datos no han cambiado), pero la probabilidad a posteriori solo se ha desplazado ligeramente hacia 1. Con cuatro caras, empiezo a sospechar un poco: si tuviera que apostar en el siguiente lanzamiento, ahora la cara es algo más probable que la cruz. Y si siguen saliendo caras, la probabilidad a posteriori sigue desplazándose hacia «esta no es una moneda normal».

Opción c: a mi amigo le encanta Batman. ¿Te acuerdas de la película de Batman con Dos Caras? Lanza una moneda para decidir el destino de la gente, y siempre sale del mismo lado. ¿Por qué? Porque su moneda tiene dos caras. Supón ahora que mi amigo es un bromista y un fan de Batman. Entonces mis hipótesis son otras: o la moneda es normal (probabilidad de cara 0,5), o es una moneda de broma de dos caras (probabilidad 1).

La evidencia es la misma en todos los casos: cuatro caras. Fíjate en lo que pasa con distintas probabilidades a priori:

Probabilidad a priori de «moneda equilibrada»	Probabilidad a posteriori de «moneda equilibrada» tras 4 caras
0,999 (la mayoría de las monedas son equilibradas)	≈ 0,98
0,5 (mi amigo es un bromista)	≈ 0,06
0,2 (mi amigo es un fan acérrimo de Batman)	≈ 0,015

Los mismos datos, tres conclusiones muy distintas. En una, la moneda es casi seguro equilibrada. En otra, es casi seguro que mi amigo me está tomando el pelo.

Entonces, ¿qué conclusión es la correcta? ¿Qué debería pensar una persona razonable?

¿Podemos tener verdades distintas?

En este punto siempre hay alguien que dice: «¡Consigue más evidencia! ¡Mira la moneda!». Y es buena idea, pero a menudo la evidencia que tienes es la que tienes. Has hecho una radiografía y has visto una sombra. Puedes pedir una biopsia, pero hay lista de espera. ¿Empiezas el tratamiento o no? Yo he analizado muestras tan valiosas que no había forma de conseguir más material. La conclusión depende de los datos, pero también de lo que creías al principio. Y si dos personas creían cosas distintas al principio, llegan a conclusiones distintas.

Esta es una de las cosas que menos gustan de la estadística bayesiana. Y no es solo una cuestión teórica. Imagina a dos personas razonables, Elena y Pablo, que examinan al mismo paciente. Pablo piensa: esta persona fuma mucho, así que a priori es bastante probable que tenga cáncer. Elena piensa: dada la prevalencia para alguien de esa edad, no es muy probable. Ven la misma radiografía, los dos razonan correctamente con el teorema de Bayes y llegan a conclusiones distintas. ¿Significa esto que cada uno tiene derecho a su propia verdad?

Cuando lo pregunto en clase, la mayoría responde: sí, supongo que cada uno tiene su verdad. Pero aquí está la cuestión. El paciente o tiene cáncer o no lo tiene. Esa es una pregunta. Otra pregunta distinta es qué *debería pensar* Elena, una persona razonable, a la vista de su evidencia y de lo que creía antes. Son dos cosas distintas, y ya lo hemos visto antes. Recuerda: *el*

mapa no es el territorio. En el mundo, el territorio, el paciente tiene cáncer o no lo tiene. En la cabeza de Elena, el mapa, hay un grado de confianza. La inferencia bayesiana no le dice a Elena qué es verdad sobre el mundo. Le dice cómo debería dibujar su mapa y cómo debería actualizarlo. Si el mapa es correcto es otra cuestión.

Algunas consecuencias prácticas

La probabilidad a posteriori es la probabilidad a priori multiplicada por la fuerza de la evidencia. Ese hecho tan sencillo tiene algunas consecuencias muy prácticas.

Busca la evidencia que separa tus hipótesis. No todos los datos son igual de útiles. Unos datos que son igual de probables con todas las hipótesis no pueden cambiar nada: si tanto un resfriado como la gripe te hacen toser, la tos no te ayuda a elegir entre ellos. Lo que mueve la probabilidad a posteriori es la evidencia que es mucho más probable con una hipótesis que con las demás. Vuelve a la moneda: cada cara es solo el doble de probable con la moneda de dos caras que con una equilibrada, así que las caras nos empujan poco a poco. Pero una sola cruz es imposible con la moneda de dos caras, y zanja la cuestión de golpe. Así que cuando diseñes un experimento o pidas una prueba, pregúntate: ¿qué resultado se vería muy distinto según cuál sea la hipótesis verdadera?

Si tus probabilidades a priori son fuertes, necesitarás una evidencia fuerte para cambiarlas. Antes de cualquier experimento, ¿qué probabilidad le das a la hipótesis de que el agua de mi grifo cura cualquier enfermedad? Muy baja. Así que no deberías conformarte con el nivel de evidencia habitual. Beberse el agua y no morirse no basta; deberías exigir una cantidad enorme de evidencia. Como solía decir Carl Sagan: *las afirmaciones extraordinarias requieren evidencias extraordinarias*. Una persona racional hace exactamente eso.

Si tus probabilidades a priori son muy fuertes, puede que la evidencia no te haga cambiar de opinión en absoluto. Si partes muy convencido de algo, será muy difícil que ninguna evidencia te mueva. Eso no es irracional en sí mismo, pero es algo que deberías saber de ti mismo.

Y en el extremo, se vuelve imposible. Si le das a una hipótesis una probabilidad a priori de exactamente cero, ninguna cantidad de evidencia te hará cambiar de opinión jamás: la probabilidad a posteriori es la probabilidad a

priori multiplicada por la fuerza de la evidencia, y cero por cualquier cosa es cero. Lo mismo ocurre con una probabilidad a priori de uno, porque estar completamente seguro de que una hipótesis es cierta significa dar cero a todas las demás. El estadístico Dennis Lindley lo llamó la regla de Cromwell, por la súplica de Oliver Cromwell a la Iglesia de Escocia: «Os suplico, por las entrañas de Cristo, que penséis que es posible que estéis equivocados». Nunca des una probabilidad a priori de exactamente 0 o 1 a nada que no sea lógicamente imposible o lógicamente seguro. Puedes decir que una hipótesis es muy improbable a priori. Pero ¿cómo de improbable? Según lo pequeña que sea, podrías mandar a alguien a la cárcel o no.

La alternativa frecuentista

No a todo el mundo le gusta la inferencia bayesiana

Si la estadística bayesiana es tan bonita, ¿por qué a los estudiantes se les enseñan sobre todo métodos basados en p(datos | modelo), como los que se apoyan en el valor p? Todo el mundo acepta el teorema de Bayes:

$$p(A \mid B) = p(A) \cdot p(B \mid A) / p(B)$$

Eso son matemáticas, y nadie lo discute. La inferencia bayesiana es otra cosa:

$$p(\text{hipótesis} \mid \text{datos}) = p(\text{hipótesis}) \cdot p(\text{datos} \mid \text{hipótesis}) / p(\text{datos})$$

¿Por qué muchos estadísticos rigurosos rechazaron durante tanto tiempo la inferencia bayesiana? Había dos problemas filosóficos principales:

1. *¿Qué es la probabilidad?* ¿Tiene sentido hablar de la probabilidad de una hipótesis, o de que llueva mañana? ¿Es la probabilidad una propiedad del mundo o de nuestro conocimiento sobre el mundo?
2. *Las probabilidades a priori son subjetivas.* Si mis conclusiones dependen de mis probabilidades a priori, y mis probabilidades a priori son mías, ¿dónde queda la objetividad de la ciencia?

Había también un problema práctico: la inferencia bayesiana suele requerir mucho cálculo, y hasta los años ochenta los ordenadores no estaban a la altura.

Las probabilidades son frecuencias

En el siglo XIX, algunos matemáticos y filósofos, como Robert Leslie Ellis, George Boole y John Venn, criticaron la forma en que Laplace asignaba probabilidades a las hipótesis. Para ellos, una probabilidad debía ser una frecuencia. Aproximadamente desde la década de 1840 hasta los años ochenta

del siglo XX, la estadística estuvo dominada por lo que hoy llamamos el enfoque frecuentista.

¿Qué significa que la probabilidad de que una moneda salga cara sea 0,5? Parece una pregunta fácil. En 1866, John Venn dio la respuesta clásica: una probabilidad es la proporción de veces que algo ocurre a la larga, si pudiéramos repetir el experimento indefinidamente. Lanza la moneda muchísimas veces, un número infinito de veces si pudieras, y la probabilidad de cara es la proporción de caras que obtendrías. Esta es la *interpretación frecuentista*. En ella, la probabilidad es una propiedad objetiva del mundo: de la moneda, del dado, de la población.

Esto tiene una consecuencia clara. ¿Cuál es la probabilidad de que una hipótesis sea cierta, por ejemplo la hipótesis nula? Para un frecuentista, la pregunta no tiene sentido. La hipótesis es verdadera o falsa; no hay una larga serie de repeticiones en la que a veces sea cierta y a veces no. Por eso, después de una prueba de significación, un frecuentista no te dirá cuánto debería cambiar tu confianza en la hipótesis nula, o en la alternativa.

Ahora puedes ver por qué los frecuentistas están encantados de rechazar hipótesis, pero no de aumentar o disminuir nuestro grado de creencia en ellas. Para un frecuentista, no tiene sentido preguntar lo probable que es una hipótesis. Mañana lloverá o no lloverá; tu paciente tiene la enfermedad o no la tiene; la hipótesis es cierta o no lo es. La probabilidad no trata de nuestras hipótesis ni de lo que creemos.

Los biometristas

A finales del siglo XIX, un grupo de investigadores se propuso estudiar la herencia y la evolución midiendo la variación en poblaciones enteras: Francis Galton (primo de Darwin), Karl Pearson y Walter Weldon, entre otros. Inventaron muchas herramientas que todavía se usan muchísimo: la correlación, la regresión y, en 1900, la prueba de chi cuadrado de Pearson.

Sus objetivos eran sobre todo descriptivos, y sus muestras eran grandes: miles de individuos. Con tantos datos, en la práctica la probabilidad a priori apenas les importaba. De hecho, Karl Pearson no estaba en absoluto en contra del razonamiento bayesiano. Aceptaba la probabilidad inversa y la usaba él mismo, por ejemplo en un artículo de 1920 sobre lo que llamó «el problema fundamental de la estadística práctica». El verdadero ataque contra las probabilidades a priori llegaría más tarde, de la mano de Fisher.

Las probabilidades a priori no son objetivas

El segundo problema eran las probabilidades a priori. Si quieres decir lo probable que es una hipótesis, necesitas una probabilidad a priori; el teorema de Bayes no deja escapatoria. ¿Cuál es la probabilidad a priori de que el Barça gane al Madrid en su próximo partido? ¿O de que un fármaco nuevo, antes de ningún experimento, cure una enfermedad? ¿Qué probabilidad hay de que alguien sea un asesino, antes de tener ninguna evidencia? Pregunta a varios expertos con la misma información y te darán números distintos. Nadie ha encontrado una regla que decida cuál de ellos tiene razón. Si mis conclusiones dependen de mis probabilidades a priori, y mis probabilidades a priori son mías, ¿dónde queda la objetividad de la ciencia?

Las probabilidades a priori también pueden ser peligrosas, porque las tasas base pueden arrastrar prejuicios. Imagina que usas la proporción de una minoría en las cárceles de un país como probabilidad «a priori» de que un joven de esa minoría sea un delincuente. Esas tasas base son a su vez el producto de una historia de discriminación, y usarlas como probabilidades a priori puede disfrazar nuestros sesgos de datos racionales. Una probabilidad a priori es una premisa, y tienes que poder defenderla.

Fisher

Ronald Fisher (1890–1962) fue probablemente el estadístico más influyente del siglo XX. Desarrolló la verosimilitud y la máxima verosimilitud, el análisis de la varianza (ANOVA) y los principios del diseño experimental, como la aleatorización, y popularizó el valor p.

Fisher también hizo aportaciones clave a la biología. Hasta 1918, mucha gente pensaba que la genética de Mendel y la evolución de Darwin eran incompatibles. La genética mendeliana dice que la herencia es discreta: Mendel veía guisantes amarillos y guisantes verdes, nada intermedio. Pero la evolución darwiniana necesita variación continua, todos los matices intermedios, para pasar gradualmente de una forma a otra. El artículo de Fisher de 1918 demostró que si un carácter está controlado por muchos genes mendelianos, cada uno con un efecto pequeño, se obtiene variación continua. Aquel artículo reconcilió a Mendel y Darwin, y de él nacieron tanto la genética de poblaciones como la genética cuantitativa.

Hizo algo igual de importante. Tomó estas herramientas, las suyas y las de otros, y las difundió por toda la comunidad científica: la prueba t de William Gosset (publicada con el seudónimo «Student»), la prueba de chi cuadrado

de Pearson, la regresión de Galton. Fisher trabajaba en una estación de investigación agraria (Rothamsted), rodeado de científicos que necesitaban métodos prácticos más que teoría matemática. Escribió dos libros para ellos: *Statistical Methods for Research Workers* (1925) y *The Design of Experiments* (1935). Están llenos de ejemplos resueltos y contienen muy pocas matemáticas. Tuvieron un éxito enorme, y llevaron su enfoque práctico, y su perspectiva frecuentista, a biólogos, agrónomos, psicólogos y médicos de todo el mundo. Después, estadísticos y autores de manuales tomaron esos libros, a menudo eliminaron los matices del propio Fisher y los convirtieron en un dogma: la ortodoxia frecuentista que todavía se enseña en muchas clases.

Y Fisher estaba firmemente en contra de la inferencia bayesiana. En *Statistical Methods for Research Workers* escribió que «la teoría de la probabilidad inversa se basa en un error y debe rechazarse por completo». La autoridad de Fisher, junto con el marco de Neyman y Pearson, que no tenía ningún hueco para las probabilidades de las hipótesis, convirtió la inferencia bayesiana en una postura minoritaria en la estadística académica durante medio siglo.

Neyman y Pearson no necesitaban a Bayes

¿Necesita la ciencia probabilidades de las hipótesis? Los frecuentistas dicen que no. Lo que necesita la ciencia, en su opinión, son procedimientos que rara vez nos lleven por mal camino, y podemos juzgar un procedimiento por frecuencias: si lo sigo, ¿con qué frecuencia acertaré y con qué frecuencia me equivocaré? Jerzy Neyman y Egon Pearson formalizaron esta forma de pensar. En lugar de preguntarte lo seguro que estás de esta hipótesis concreta, estableces un procedimiento, con un umbral para decidir hacer una cosa u otra.

Por ejemplo, creas un conjunto de reglas para tratar a los pacientes según sus síntomas y los resultados de sus pruebas, y luego compruebas en cuántos casos las reglas dieron la respuesta correcta. Podemos aplicar la misma forma de pensar incluso al sistema judicial. Un tribunal no puede saber con certeza si este acusado concreto es culpable, y no necesita ponerle una probabilidad. Lo que hace es seguir una regla, «más allá de toda duda razonable», diseñada para que, a lo largo de miles de juicios, pocos inocentes acaben en la cárcel. Neyman llamó a esto *comportamiento inductivo*. La pregunta no es «¿qué debo creer?», sino «¿qué regla debo seguir, sabiendo que la aplicaré una y otra vez?».

En el enfoque de Neyman y Pearson se te permite aceptar una hipótesis: no como una creencia, sino como la decisión de actuar como si fuera cierta, sabiendo con qué frecuencia se equivoca ese tipo de decisión. El médico no necesita conocer la probabilidad de que el paciente tenga cáncer para seguir una regla que, considerando todos los pacientes, mantenga bajo control los dos tipos de errores.

Los médicos ya usan reglas diseñadas desde el punto de vista del sistema. Los umbrales de cribado, los valores de referencia de un análisis de sangre, las guías clínicas: todos ellos se eligieron por cómo funcionan con muchos pacientes, cuántos enfermos detectan y cuántas personas sanas mandan a hacerse pruebas innecesarias. Cuando un médico sigue una guía, está aplicando una regla al estilo de Neyman, aunque piense en su paciente de forma bayesiana.

Los tribunales a veces eligen a propósito el punto de vista del sistema. La presunción de inocencia pide al juez que deje de lado cualquier probabilidad a priori sobre el tipo de personas que acaban en el banquillo. Puede parecer que así se tira información, pero hay una buena razón para ello. Como vimos, las tasas base pueden arrastrar prejuicios, y una probabilidad a priori puede ser un disfraz muy fino para ellos: «la gente como él suele ser culpable» suena a probabilidad, pero puede no ser más que un sesgo con un número pegado. Un sistema jurídico puede preferir un procedimiento que proteja a los inocentes, sean cuales sean las probabilidades a priori del juez.

La filósofa Deborah Mayo ha desarrollado más las ideas de Neyman y las ha conectado con Popper. Para ella, una afirmación está bien respaldada cuando ha superado una *prueba severa*, una que muy probablemente habría fallado si la afirmación fuera falsa. Superar una prueba fácil no significa nada; superar una difícil significa mucho.

Esta visión tiene un verdadero punto fuerte: explica algunos de los problemas que vimos con el valor p. ¿Por qué está mal el *p-hacking*? Si pruebas muchos análisis, o sigues recogiendo datos hasta que $p < 0{,}05$, el resultado que finalmente publicas puede parecer exactamente igual que uno honesto. Lo que ha cambiado es el procedimiento: habría «encontrado» un efecto aunque no lo hubiera. La prueba no era severa, y las tasas de error que afirmas no son las reales. La estadística del error lo condena directamente. A un bayesiano estricto, que solo mira los datos realmente observados, le

cuesta más explicar por qué el jardín de los senderos que se bifurcan es un problema.

La defensa bayesiana

Los bayesianos académicos

El razonamiento bayesiano era una postura minoritaria en el mundo académico, pero nunca llegó a desaparecer del todo. Unos pocos estadísticos y filósofos siguieron discutiéndolo y desarrollándolo. Harold Jeffreys, un geofísico, discutió con Fisher por escrito y publicó su *Theory of Probability* en 1939. Frank Ramsey y Bruno de Finetti se ocuparon de averiguar qué podía significar que una probabilidad fuese un grado de creencia. Y en los años cincuenta y sesenta, Leonard Savage, Dennis Lindley e I. J. Good, entre otros, sentaron las bases de la estadística bayesiana moderna.

Jeffreys intentó responder directamente a la objeción de la subjetividad. Creía en las probabilidades a priori objetivas: personas racionales con el mismo conocimiento deberían coincidir en la probabilidad a priori, y la tarea consistía en encontrar las reglas que la proporcionan. Hoy pocos defienden ese programa tan ambicioso, por el motivo que ya vimos: expertos con la misma información dan números distintos. Pero las probabilidades a priori elegidas mediante reglas convencionales se siguen usando mucho, y a menudo se las llama probabilidades a priori «objetivas».

Los bayesianos propusieron además otra interpretación de la probabilidad. En 1926, Frank Ramsey, un joven filósofo de Cambridge, definió las probabilidades como *grados de creencia*. Cuando digo que la probabilidad de que mi paciente tenga la enfermedad, o de que llueva mañana, es 0,9, estoy describiendo el estado de mi conocimiento, no a mi paciente ni a las nubes. Pero ¿cómo se puede medir un grado de creencia? La respuesta de Ramsey: por las apuestas que aceptarías. Si pagarías hasta 90 céntimos, y no más, por un boleto que paga un euro si llueve, tu grado de creencia es 0,9.

El matemático italiano Bruno de Finetti, que trabajó durante años como actuario, llevó esta idea hasta sus últimas consecuencias. «La probabilidad no existe», escribió. No está ahí fuera, en el mundo; solo está en nuestra cabeza. La probabilidad trata del mapa, no del territorio. También mostró por qué los grados de creencia deberían seguir las reglas de la probabilidad. Si no las siguen, alguien podría ofrecerte un conjunto de apuestas que, una a una, te parecen justas, pero que juntas te hacen perder pase lo que pase. En el mundo de las apuestas, esto se llama «apuesta holandesa» (*Dutch book*). Veamos un caso sencillo. Pensemos en una final de copa, en la que alguien tiene que ganar, y supongamos que estás seguro al 90 % de que ganará el Madrid, y también seguro al 90 % de que ganará el Barça. Entonces deberías estar encantado de pagar 90 céntimos por un boleto que paga un euro si gana el Madrid, y otros 90 céntimos por un boleto que paga un euro si gana el Barça. Has pagado 1,80 euros y, pase lo que pase, solo uno de los boletos paga: recuperas un euro. Pierdes 80 céntimos con total seguridad. Si estás seguro al 90 % de que ganará el Madrid, tienes que estar seguro al 10 % de que ganará el Barça. Esa es la coherencia que vimos con la inferencia bayesiana. Fíjate en que la coherencia no señala un único conjunto de creencias: muchos conjuntos distintos pueden ser coherentes. Es coherente pensar que el Madrid ganará con una probabilidad de 0,9 y el Barça con una de 0,1, o al revés. Lo que no puedes hacer es creer que ganarán los dos con una probabilidad de 0,9.

Así pues, para los frecuentistas la probabilidad vive en el territorio: es una propiedad de las monedas y de las poblaciones. Para los bayesianos vive en el mapa: en quien razona, y describe lo que sabe. Este asunto del mapa frente al territorio, que parecía una distinción bastante peculiar, está en el centro de la pelea entre estadísticos. Unos dicen que la estadística trata del mundo. Otros dicen que la estadística trata de lo que deberíamos pensar sobre el mundo. Eso es pura filosofía de la estadística, y ha alimentado muchas discusiones entre profesionales.

Al final, los frecuentistas tampoco son objetivos

La gran objeción a la inferencia bayesiana es que las probabilidades a priori son subjetivas. Los frecuentistas intentaron escapar de esa subjetividad explícita, pero podría decirse que simplemente la escondieron debajo de la alfombra. Por ejemplo, a los frecuentistas les gusta ver sus tasas de error como algo objetivo. Pero las tasas de error también dependen de decisiones:

el modelo y sus supuestos, el umbral de significación (una convención) e incluso el plan de análisis, es decir, lo que habrías hecho si los datos hubieran sido distintos.

Ambos enfoques descansan sobre decisiones. Los frecuentistas las toman sobre el modelo y el procedimiento; los bayesianos, además, sobre lo creíbles que son las hipótesis. Los estadísticos Andrew Gelman y Christian Hennig han propuesto abandonar por completo las palabras «subjetivo» y «objetivo». En su lugar, deberíamos exigir las virtudes que realmente queremos, como la transparencia (expón abiertamente tus supuestos), el consenso (¿podrían otras personas estar de acuerdo con ellos?) y la conciencia de hasta qué punto tus conclusiones dependen de esas decisiones.

Las hipótesis no son objetivas

En la inferencia estadística, incluso en el enfoque frecuentista, se supone a menudo que empezamos con una hipótesis, o un modelo, y luego miramos los datos. Pero ¿de dónde salió la hipótesis? Tendemos a suponer que las hipótesis se construyen a partir de los datos, pero en ciencia muchas veces no es eso lo que ocurre.

Durante siglos, los astrónomos pudieron predecir las posiciones de los planetas con gran precisión sin ninguna hipótesis física sobre lo que hacían los planetas. La astronomía se consideraba una rama de las matemáticas, no de la física: la física trataba de cómo es realmente el mundo, mientras que las matemáticas solo lo describían. Los astrónomos tenían modelos matemáticos, círculos sobre círculos, y no les interesaba mucho nada más. Cuando se publicó el libro de Copérnico en 1543, venía con un prefacio escrito no por Copérnico, sino por el teólogo Andreas Osiander. El prefacio presentaba el modelo exactamente de esa manera: como una herramienta de cálculo, no como una descripción de la realidad. La Tierra no tenía por qué moverse de verdad; simplemente, las fórmulas eran más fáciles de usar. Luego Galileo armó un escándalo al insistir en que la Tierra se mueve de verdad. Para él, era una hipótesis física.

La historia de Kepler muestra lo lejos que puede estar una hipótesis de los datos. Los datos eran de Tycho Brahe, las observaciones astronómicas más precisas de la época. Tycho era todo un personaje. Había perdido parte de la nariz en un duelo y llevaba una prótesis metálica. También tenía un alce domesticado que, según cuenta la historia, bebió demasiada cerveza en una

cena, se cayó por las escaleras y murió. Contrató a Kepler para que analizara sus observaciones, pero las guardaba celosamente, quizá por miedo a que Kepler sacara las grandes conclusiones y se llevara el mérito. Cuando Tycho murió en 1601, Kepler se aseguró de que los datos no se le escaparan. En sus propias palabras: «Confieso que, cuando Tycho murió, aproveché rápidamente la ausencia, o la falta de circunspección, de los herederos, tomando las observaciones a mi cuidado, o quizá usurpándolas».

Pero Kepler no leyó la elipse en los datos. Era platónico, y para un platónico el círculo es el movimiento perfecto, así que los planetas tenían que moverse en círculos. Pasó años probando una combinación de círculos tras otra, y ninguna se ajustaba a las observaciones de Tycho sobre Marte. Solo entonces probó otras formas, primero óvalos, hasta que dio con la elipse. Al final, la evidencia sí ganó. Pero la hipótesis ganadora no salió de los datos; salió de Kepler, una vez que su favorita había fracasado.

Arquímedes describió su propio método así: consigue tu idea como puedas, incluso con un razonamiento mecánico tosco, y luego demuéstrala con rigor; si la demostración funciona, da igual de dónde haya salido la idea. (El texto en el que lo explica, *El método*, estuvo perdido durante siglos. Apareció en 1906 en un palimpsesto: un libro de oraciones medieval escrito encima de una copia más antigua de su obra). Y el químico August Kekulé afirmaba que la estructura en anillo del benceno se le ocurrió en un sueño en el que una serpiente se mordía la cola.

Los filósofos llaman a esto la diferencia entre el *contexto de descubrimiento* y el *contexto de justificación*. El descubrimiento es cómo se nos ocurre una hipótesis: una corazonada, un sueño, un prejuicio platónico, un intento fallido. La justificación es cómo la ponemos a prueba frente a los datos. Mientras una hipótesis supere la prueba, no nos preocupa mucho de dónde haya salido.

Un frecuentista podría responder: «Exacto. Me da igual de dónde haya salido tu hipótesis, porque mi prueba la juzga. Pero tu probabilidad a priori no se queda en el contexto de descubrimiento: entra en el propio veredicto». Es una objeción justa. Una justificación objetiva sería estupenda, si fuera posible. Pero, como acabamos de ver, la justificación frecuentista también descansa sobre decisiones. E incluso una prueba perfectamente objetiva no resolvería el problema principal, porque una prueba solo puede juzgar las hipótesis que hemos elegido poner a prueba. Al elegir unas hipótesis igno-

ramos todas las demás, y ahí es donde la ciencia se equivoca con frecuencia. Un ensayo puede concluir que un fármaco no funciona cuando solo se probó en la población equivocada: quizá funciona en mujeres mayores y nadie miró ahí. La prueba fue impecable; la elección de la hipótesis, no.

Y los frecuentistas eligen sus hipótesis, sus modelos y sus supuestos igual que los bayesianos; como vimos, los modelos no salen solo de los datos. Así que la objetividad de la estadística frecuentista no es tan completa como parece, y la subjetividad bayesiana no es tan mala como parece.

La evidencia suficiente acaba ganando

Por suerte, hay una salida, al menos a veces. Con suficiente evidencia y las mismas hipótesis, dos personas que razonen con el teorema de Bayes acabarán llegando a las mismas conclusiones, sean cuales sean sus probabilidades a priori iniciales.

Volvamos a la moneda. Con cuatro caras, la persona que empezaba con 0,999 para la moneda equilibrada seguía pensando que era equilibrada. Pero si el amigo sigue lanzando y saca quince caras seguidas, incluso la probabilidad a posteriori de esa persona para la moneda equilibrada cae hasta aproximadamente 0,03. Ahora todo el mundo está de acuerdo en que algo pasa. Con suficientes datos, las probabilidades a priori quedan borradas.

Esta propiedad es más profunda de lo que parece. Aunque empecemos con un mapa equivocado, con probabilidades a priori muy alejadas de la realidad, siempre que sean coherentes y no descarten la verdad asignándole una probabilidad a priori de cero, actualizar con los datos permite que el territorio vaya remodelando el mapa. La estadística bien hecha, bayesiana o frecuentista, deja que la evidencia corrija nuestro punto de partida. Esa es una de las grandes virtudes de estas herramientas.

Pero remodelar el mapa tiene un precio: datos. Cuanto más lejos del territorio esté nuestro mapa de partida, más evidencia necesitaremos para llegar a la misma conclusión. En el ejemplo de la moneda, la persona que empezaba con 0,999 para la moneda equilibrada necesitó quince caras seguidas para bajar hasta aproximadamente 0,03. Alguien que hubiera empezado con 0,5 habría llegado ahí con solo cinco. Lo que más datos cuesta es una probabilidad a priori que está equivocada con mucha seguridad: una probabilidad a priori vaga, como 0,5, cuesta poco aunque esté desviada, mientras que una firme, como 0,999, cuesta mucho. Así que unas buenas probabilidades

a priori no son solo una cuestión de principios. Una probabilidad a priori que representa bien el territorio nos lleva a buenas conclusiones con menos datos, y en ciencia los datos cuestan tiempo y dinero.

La otra cara de la moneda es que las probabilidades a priori importan mucho cuando no tenemos muchos datos. Y eso es útil saberlo. Ser conscientes de que nuestros datos son débiles, de que no favorecen una hipótesis frente a otra, es una información valiosa en sí misma.

¿Lloverá mañana?

Para un ingeniero que diseña un proceso de producción, o para alguien que gestiona un sistema sanitario o un sistema judicial, la pregunta de Neyman es la natural: a lo largo de miles de casos, ¿cuántos errores cometerá mi regla? Pero piensa en los pacientes y los médicos. Para el médico, la pregunta natural es «¿qué probabilidad hay de que esta persona esté enferma?». Sería muy raro que un médico tratase a un paciente sin pensar en lo probable que es que este paciente concreto tenga la enfermedad, y que solo pensara «estoy siguiendo un procedimiento que, considerando a todos mis pacientes, tiene una tasa de error de tipo I del 5 % y una tasa de error de tipo II del 10 %». Lo mismo vale para un juez ante un acusado concreto: lo que el juez tiene que decidir es si esta persona es culpable.

Podríamos decir que el enfoque bayesiano adopta el punto de vista del caso, y el frecuentista, el punto de vista del sistema. Pensar en el sistema judicial en su conjunto a la manera de Neyman tiene todo el sentido; pedirle a un juez concreto que piense así, no. Y si quieres llegar a una conclusión sobre la probabilidad de una hipótesis, no hay forma de evitarlo: la inferencia bayesiana es la única manera consistente. Solo tenemos que aceptar que los grados de creencia pueden representarse con probabilidades, y entonces la inferencia bayesiana nos da probabilidades consistentes para las hipótesis.

Bayes funciona

Fuera del mundo académico, había gente que usaba ideas bayesianas con fines prácticos, incluso cuando esas ideas no eran populares en las universidades. Dos ejemplos: el dinero y la guerra.

Los actuarios de seguros necesitaban ponerles precio a los riesgos, y los debates académicos no les importaban demasiado. A principios del siglo XX, los actuarios estadounidenses desarrollaron lo que llamaron *teoría de la*

credibilidad. Para fijar el precio de una póliza para una empresa concreta, combinaban el historial de siniestros de la propia empresa con la experiencia de todo el sector, y daban más peso a los datos de la empresa cuantos más hubiera. Eso es razonamiento bayesiano en todo salvo en el nombre: una probabilidad a priori procedente del sector, actualizada con los datos de la empresa. Más tarde, los actuarios demostraron que, en efecto, era bayesiano.

Durante la Segunda Guerra Mundial, las fuerzas armadas alemanas cifraban sus mensajes con la máquina Enigma. Unos matemáticos polacos ya habían logrado avances notables para descifrar versiones anteriores de Enigma, y poco antes de la guerra compartieron lo que sabían con los británicos y los franceses. Pero la configuración cambiaba cada día y, sin ella, los mensajes eran un galimatías. En Bletchley Park, al norte de Londres, los británicos reunieron a un equipo de matemáticos, lingüistas y aficionados a los crucigramas para descifrar el código.

Uno de ellos era Alan Turing. En matemáticas, Turing es un gigante: en 1936 había definido, en abstracto, qué es una máquina de computar, y había sentado así las bases teóricas de la informática. En Bletchley diseñó las máquinas electromecánicas (las *bombes*) que se usaban para encontrar la configuración de Enigma, y desarrolló métodos estadísticos para decidir qué hipótesis sobre la configuración merecía la pena comprobar. Esos métodos eran bayesianos: sopesaba la evidencia a favor de cada hipótesis, e incluso inventó una unidad para medir la fuerza de la evidencia, el *ban*. Funcionaron, y la información obtenida fue crucial, por ejemplo, en la batalla contra los submarinos alemanes en el Atlántico. (Bletchley Park ha sido restaurado y hoy es un museo). Pero todo esto era secreto de Estado, y siguió siéndolo durante décadas, hasta los años setenta.

El final de la historia de Turing es triste. Turing era homosexual y no lo ocultaba. En 1952 fue procesado por «indecencia grave» y, para evitar la cárcel, aceptó la castración química mediante inyecciones de hormonas. Nadie sabía entonces lo que había hecho por su país. En 1954, dos semanas antes de cumplir 42 años, lo encontraron muerto por envenenamiento con cianuro, con una manzana a medio comer junto a la cama. La manzana nunca se analizó, y todavía no está claro si fue un suicidio o un accidente (hacía experimentos de química en casa). Le fascinaba la escena de *Blancanieves* de Disney en la que la bruja envenena la manzana. Uno de los matemáticos más importantes del siglo XX, una figura clave para ganar la guerra, destruido

por aquellos a quienes había servido. El Gobierno británico pidió disculpas oficialmente en 2009, y Turing recibió un indulto real en 2013.

Seguramente habrás oído que el logotipo de la manzana mordida de cierta empresa de ordenadores es un homenaje a Turing. Es una historia preciosa. Por desgracia, el diseñador del logotipo ha dicho que no es cierto, y la empresa nunca lo afirmó. El escritor y actor Stephen Fry le preguntó una vez a Steve Jobs por ello y, según cuenta Fry, Jobs respondió: «No es verdad, pero, Dios, ojalá lo fuera». Es el tipo de leyenda que merece existir.

Una tregua pragmática

Preguntas distintas

Entonces, ¿qué deberíamos hacer? ¿Qué enfoque deberíamos aceptar? ¿Deberíamos rechazar el bayesianismo o el frecuentismo?

En las últimas décadas, cada vez más estadísticos han adoptado una postura pragmática: usa la herramienta que se ajusta a la pregunta y entiende lo que hace cada una. Podríamos aceptar que ambas interpretaciones de la probabilidad son legítimas, y pensar en las dos escuelas como enfoques para preguntas distintas. El estadístico Richard Royall señaló que, cuando miramos unos datos, podemos estar haciéndonos tres preguntas diferentes:

1. *¿Qué dicen los datos?* ¿Cuánto favorecen una hipótesis frente a otra? Solo necesitas los datos y las hipótesis.
2. *¿Qué debería creer?* Eso es Bayes: además, necesitas una probabilidad a priori.
3. *¿Qué debería hacer?* Eso es Neyman y Pearson: además, necesitas tasas de error, o los costes de cada tipo de error.

La primera pregunta es la que se hacen los verosimilitudistas, Royall entre ellos. Miden la evidencia con la razón de verosimilitudes. Volvamos a la moneda de la partida de ajedrez y sus cuatro caras. Una moneda de dos caras da ese resultado con probabilidad 1; una moneda equilibrada, con probabilidad 0,0625. La razón es 16: los datos favorecen la moneda de dos caras frente a la equilibrada por un factor de 16. Ese número es el mismo para todo el mundo, creyera lo que creyera antes.

Los bayesianos están de acuerdo con esta forma de medir la evidencia. Con dos hipótesis, el teorema de Bayes dice exactamente eso: las probabilidades relativas (*odds*) después de ver los datos son las de antes multiplicadas por

esta razón. Así que la respuesta a la primera pregunta forma parte de la respuesta a la segunda: el bayesiano toma la evidencia y la multiplica por una probabilidad a priori. Y si la pregunta es «¿qué debería creer?», la forma más directa de responderla es un análisis bayesiano.

Cada escuela responde bien a su propia pregunta, y muchas de las peleas entre estadísticos vienen de responder a una pregunta con la herramienta de otra escuela. El ritual nulo es el caso extremo: aplica un procedimiento de decisión, pensado para la tercera pregunta, y luego interpreta el resultado como evidencia y como creencia: «p < 0,05, el fármaco funciona».

Además, en la práctica, los métodos clásicos y los bayesianos llegan a menudo a las mismas conclusiones. Eso suele ocurrir cuando:

- tienes muchos datos,
- el resultado es claro, no está en el límite,
- tus hipótesis están bien definidas (por ejemplo, una nula y una alternativa) e, idealmente, la alternativa indica el tamaño del efecto que te interesa,
- las hipótesis son, a priori, más o menos igual de probables,
- haces una sola prueba, o corriges adecuadamente las comparaciones múltiples y evitas el *p-hacking*,
- y entiendes lo que estás haciendo: el valor p no es la probabilidad de ninguna hipótesis, y ni siquiera un valor p minúsculo dice nada sobre lo grande que es el efecto.

Lo de «no está en el límite» importa. Podrías pensar que, con muchos datos, los dos enfoques siempre coinciden, pero hay una excepción muy conocida. Con una muestra enorme, un valor p justo por debajo de 0,05 corresponde a un efecto minúsculo. El valor p dice «rechaza la nula», mientras que un análisis bayesiano de los mismos datos dice que la nula ha ganado apoyo. Esto se conoce como la paradoja de Lindley. Con muchos datos, fíjate en lo grande que es el efecto, no solo en si p ha cruzado un umbral.

Los frecuentistas también usan probabilidades a priori

Los frecuentistas también usan probabilidades a priori y el teorema de Bayes, en algunos casos. Cuando existe una tasa base, no hay ningún desacuerdo. Piensa en la prevalencia en la prueba del VIH. Ambas escuelas ponen el mismo número en la misma fórmula, el teorema de Bayes. El bayesiano lo llama probabilidad a priori; el frecuentista lo llama tasa base,

e insiste en que debe ser una frecuencia medida. Pero usan las mismas matemáticas, la misma teoría y los mismos números.

Podemos verlo en el enfoque de Neyman y Pearson. Cada una de sus dos tasas de error se calcula suponiendo un estado del mundo: α es la probabilidad de un resultado positivo *si* el paciente está sano, y β, la probabilidad de un resultado negativo *si* el paciente está enfermo. Ninguna de las dos necesita la prevalencia. Estas tasas de error describen el procedimiento, no un resultado concreto. Así, Neyman puede garantizar que solo se trate a una pequeña proporción de las personas sanas y que solo se escape una pequeña proporción de las enfermas. Pero, para saber cuántos errores cometerá un hospital en total, necesitas saber cuántos enfermos hay: α por el número de pacientes sanos más β por el número de enfermos. Necesitas la prevalencia, que no es más que una probabilidad a priori con otro nombre.

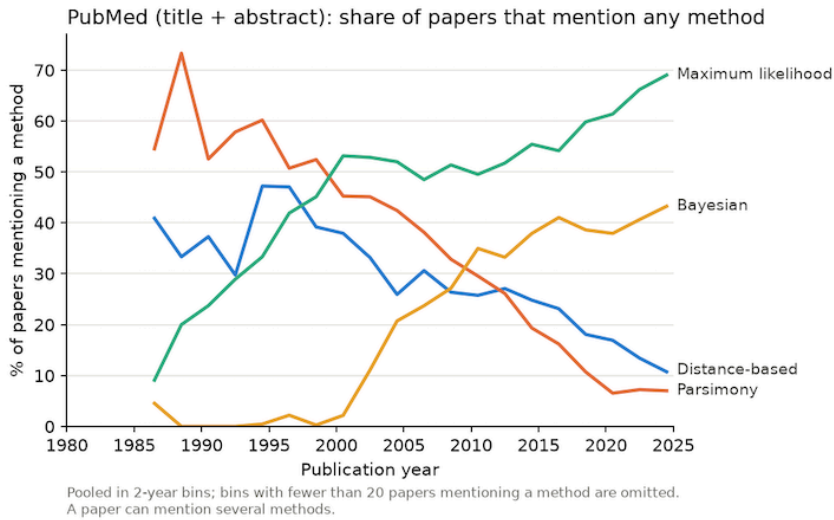
El verdadero desacuerdo entre frecuentistas y bayesianos se limita a las hipótesis que no tienen detrás ninguna frecuencia, como la culpabilidad de Sally Clark o la verdad de una teoría científica, como la relatividad general. Ahí, el frecuentista se niega a dar un número y el bayesiano tiene que darlo. Para un bayesiano, el problema de dejar fuera la probabilidad a priori es este: un campo puede usar $\alpha = 0{,}05$ con total fidelidad y, aun así, que la mayoría de sus hallazgos significativos sean falsos, si la mayoría de las hipótesis que pone a prueba eran improbables de entrada. Eso es parte de lo que está pasando en la crisis de replicación.

Bayes tuvo que esperar a los ordenadores

Aunque ambos enfoques sean útiles, durante mucho tiempo solo uno de ellos podía usarse en la práctica. Para la mayoría de los problemas reales, la probabilidad a posteriori no se puede calcular a mano. El denominador, p(datos), exige sumar sobre todas las hipótesis posibles y, con modelos que tienen muchos parámetros, ese es un cálculo que nadie puede hacer en papel. Durante la mayor parte del siglo XX, el debate entre bayesianos y frecuentistas fue en gran medida una discusión entre estadísticos y filósofos, con pocas consecuencias prácticas: los análisis bayesianos de problemas realistas estaban, sencillamente, fuera de nuestro alcance. Incluso la máxima verosimilitud, salvo en unos pocos casos sencillos, era demasiado sin ordenadores.

Los algoritmos clave existían mucho antes de que se usaran. El algoritmo de Metropolis, una forma de explorar una distribución de probabilidad con una especie de paseo aleatorio guiado, lo publicaron en 1953 unos físicos de Los Álamos, y W. K. Hastings lo generalizó en 1970. Pero no fue hasta 1990, cuando Alan Gelfand y Adrian Smith mostraron lo que estos métodos de simulación podían hacer por la estadística, en una época en la que los ordenadores se estaban abaratando, cuando la estadística bayesiana se volvió práctica. Estos métodos, conocidos como Monte Carlo basado en cadenas de Markov (MCMC, por sus siglas en inglés), son los que ejecutan programas de filogenia como MrBayes cuando te dan la probabilidad a posteriori de un árbol.

Puedes ver esta historia en mi propio campo, en cómo se han construido los árboles filogenéticos. En los años ochenta, casi todo el mundo usaba métodos de distancias y de parsimonia, que son rápidos. Los árboles de máxima verosimilitud no se generalizaron hasta los noventa, cuando los ordenadores pudieron con ellos. Los árboles bayesianos apenas existían hasta alrededor del 2000. Y las universidades cambian despacio: métodos que se volvieron prácticos hace solo una generación todavía se están abriendo paso en los cursos de estadística. Eso es parte de la respuesta a la pregunta que nos hicimos antes: por qué a ti te enseñaron, en cambio, el valor p.



Pluralismo y oficio

Yo era bayesiano

Durante años me consideré bayesiano, y mi principal motivo era que rechazaba la forma en que se suelen usar los valores p. Sigo rechazando esa forma de usarlos, pero ahora me doy cuenta de que, como crítica a la estadística frecuentista, era injusta: un hombre de paja. Como vimos, ni Fisher ni Neyman habrían aceptado el ritual nulo, y los frecuentistas rigurosos de hoy, como Deborah Mayo, también lo rechazan. Criticar la estadística frecuentista por culpa del ritual es como criticar la estadística bayesiana porque alguien podría elegir la probabilidad a priori que le diera la gana y obtener la respuesta que quisiera. Los bayesianos rigurosos no defienden probabilidades a priori arbitrarias, y los frecuentistas rigurosos no defienden el ritual.

¿Y qué hay de los argumentos frecuentistas contra los bayesianos? En realidad son dos objeciones. La primera dice que los grados de creencia no pueden representarse mediante probabilidades. Veamos por qué, como argumento formal, no se sostiene.

Kolmogórov: una teoría formal de la probabilidad

En 1933, el matemático ruso Andréi Kolmogórov hizo algo que, visto en retrospectiva, hace que gran parte de la pelea entre las dos escuelas parezca menos necesaria. En su formulación, una probabilidad es cualquier función que asigna números a sucesos y cumple tres reglas:

1. Las probabilidades nunca son negativas.
2. La probabilidad de que ocurra algo es 1.

3. Si dos sucesos no pueden ocurrir a la vez, la probabilidad de «uno u otro» es la suma de sus probabilidades. Por ejemplo, cara y cruz en un único lanzamiento: p(cara o cruz) = p(cara) + p(cruz).

Todo lo demás, incluido el teorema de Bayes, se deriva de estas reglas. Y aquí viene lo importante: Kolmogórov no dijo qué *es* la probabilidad. Construyó una estructura matemática, y cualquier cosa que siga esas reglas puede estudiarse con ella. La teoría no se ocupaba de lo que significa la probabilidad ni de cómo usarla en ciencia. Como la lógica, según veremos a continuación, es una teoría *formal*: nos dice cómo deben relacionarse las probabilidades entre sí, no qué significan.

Las frecuencias a largo plazo cumplen las tres reglas de Kolmogórov. Los grados de creencia coherentes también las cumplen; eso es exactamente lo que demuestra el argumento de las apuestas de De Finetti. Así que ambos son usos legítimos de las mismas matemáticas. Y esa es mi postura: todo lo que se comporte según las reglas de Kolmogórov puede representarse con probabilidades. Las frecuencias pueden, y también los grados de creencia coherentes. Esta postura es una especie de pluralismo sobre la interpretación de la probabilidad.

La segunda objeción dice que los grados de creencia no pueden medirse, o al menos no de forma objetiva. Esta sobrevive, no como un argumento demoledor, sino como una advertencia práctica. Las apuestas de Ramsey solo miden las creencias de forma aproximada, y una probabilidad a priori puede ser perfectamente coherente y estar muy equivocada, o ser un fino disfraz del prejuicio. Una probabilidad a priori ya es un mapa, y deberíamos intentar que fuera un buen mapa, midiéndola cuando podamos y justificándola abiertamente cuando no. Pero no hay ninguna garantía de que dos personas razonables partan del mismo mapa. En general, las probabilidades a priori no pueden ser objetivas.

Así que no podemos tener probabilidades a priori completamente objetivas. Pero eso no significa que cualquier probabilidad a priori valga. Si queremos que nuestras conclusiones se parezcan al territorio, nuestras probabilidades a priori tienen que representar el territorio suficientemente bien: una prevalencia medida en la población adecuada, los resultados de experimentos anteriores, lo que sabemos sobre cómo funciona el mundo.

La probabilidad como lógica

La inferencia bayesiana también puede verse como una forma generalizada de inferencia lógica. La lógica es la disciplina que estudia las reglas del razonamiento válido. La inventó Aristóteles, más o menos él solo. Quería una manera de distinguir el razonamiento sólido de los argumentos ingeniosos pero vacíos de los sofistas, los debatientes profesionales de su época. Un argumento es *válido* si, una vez que aceptamos las premisas como verdaderas, la conclusión se sigue necesariamente:

1. Todas las personas son mortales.
2. Sócrates es una persona.
3. Por tanto, Sócrates es mortal.

La validez es una propiedad de la forma del argumento, no de su contenido. «Si P, entonces Q; P; por tanto, Q» es válido sean lo que sean P y Q. La verdad, en cambio, es una propiedad de cada enunciado individual. La pregunta central de la lógica es «¿se sigue la conclusión de las premisas?», no «¿son verdaderas las premisas?».

Por eso la validez no garantiza la verdad. Este argumento es perfectamente válido:

1. Todos los gatos pueden volar.
2. Luna es un gato.
3. Por tanto, Luna puede volar.

La validez conserva la verdad, pero solo si la verdad ya estaba en las premisas. Y también ocurre lo contrario: un argumento no válido puede tener una conclusión verdadera.

1. Sócrates es mortal.
2. Platón es mortal.
3. Aristóteles es mortal.
4. Por tanto, todas las personas son mortales.

La conclusión es verdadera, pero este argumento no la justifica. Y fíjate en qué tipo de argumento es: una inducción, de algunos casos a todos. Como vimos con Hume y la salida del sol, las inducciones no son lógicamente válidas. Pero la ciencia no puede vivir sin ellas, porque necesitamos conclusiones que vayan más allá de los datos que tenemos.

Así que la inferencia bayesiana puede verse como una lógica para razonar con premisas inciertas. Las premisas son las hipótesis, sus grados de creencia

a priori y la evidencia; las conclusiones son los grados de creencia a posteriori. En 1946, el físico Richard Cox demostró que razonar con grados de plausibilidad tiene que seguir las reglas de la probabilidad, si se cumplen unos pocos requisitos básicos de consistencia. Más tarde, el físico Edwin Jaynes construyó todo un enfoque de la estadística sobre ese resultado, y lo llamó *la teoría de la probabilidad: la lógica de la ciencia*. Si tus grados de creencia no siguen estas reglas, tu razonamiento es inconsistente, del mismo modo que concluir «P» a partir de «si P, entonces Q» y «Q» no es válido.

Coherencia y verdad

Pero cumplir las reglas de Kolmogórov hace que un conjunto de números sea coherente, no necesariamente verdadero. Un grado de creencia en «Sally Clark es culpable» puede cumplirlas a la perfección y, aun así, ser una descripción pésima del mundo. Esta es, de hecho, la versión más fuerte de la objeción frecuentista: satisfacer los axiomas no basta; tiene que haber algo en el mundo a lo que respondan los números.

Recuerda a Elena y Pablo. Miraron la radiografía del mismo paciente y ambos razonaron correctamente con el teorema de Bayes, pero llegaron a conclusiones distintas porque partían de probabilidades a priori distintas. ¿Tiene cada uno derecho a su propia verdad? Para responder, necesitamos decir qué es la verdad.

Aristóteles dio la definición clásica de verdad: una idea es verdadera cuando se corresponde con lo que hay en el mundo. «El paciente tiene cáncer» es verdadero si el paciente tiene cáncer. Parece trivial, pero no lo es; se llama *teoría de la verdad como correspondencia*. El mundo en sí no es ni verdadero ni falso; el mundo simplemente *es*. Lo que es verdadero o falso son nuestras creencias, las cosas que tenemos en la cabeza: son verdaderas cuando encajan con el mundo, y falsas cuando no. Otra vez la pipa de Magritte: el cuadro es un cuadro bueno o malo de una pipa, pero no es una pipa.

Así que *las conclusiones válidas y la verdad no son lo mismo*. La inferencia bayesiana, como la lógica deductiva, es una máquina formal: garantiza que tus conclusiones se siguen correctamente de tus premisas (tus hipótesis, tus probabilidades a priori, tu evidencia). No dice nada sobre si tus premisas son correctas. Se nos permite tener conclusiones válidas distintas; no se nos permite tener verdades distintas. La verdad depende de la realidad, no de mis probabilidades a priori.

Hay una alternativa a la teoría de Aristóteles: la *teoría de la verdad como coherencia*, según la cual una creencia es verdadera cuando encaja bien con el resto de nuestras creencias. El problema de esta teoría es que no está anclada en la realidad. Una buena novela puede ser perfectamente coherente y completamente falsa. Un mapa puede ser perfectamente consistente, con todas las carreteras bien conectadas entre sí, y aun así no corresponderse con el territorio. La inferencia bayesiana garantiza la coherencia, y eso es valioso, pero no basta. Es el mismo problema que tendríamos si intentáramos estudiar el mundo solo con la lógica deductiva. La lógica garantiza que unas premisas verdaderas llevan a conclusiones verdaderas. Pero si partes de premisas que no describen el mundo real, un argumento perfectamente válido puede llevarte a una conclusión perfectamente equivocada.

Si queremos que nuestras conclusiones se correspondan con el mundo, la coherencia no basta. Si queremos que nuestros mapas se parezcan al territorio, necesitamos premisas que representen el mundo suficientemente bien: buenos modelos, probabilidades a priori no demasiado alejadas de la realidad y buena evidencia. Y todavía mejor si la evidencia es lo bastante fuerte como para borrar nuestras probabilidades a priori: entonces, incluso unas probabilidades a priori poco realistas llevan a conclusiones aproximadamente verdaderas.

Cuidado: no puedes ser objetivo

Eugenesia y genética

Los biometristas que conocimos antes tenían también una cara mucho menos presentable. Galton y Pearson eran abiertamente racistas. Buena parte de sus mediciones pretendían demostrar la superioridad de unos grupos sobre otros: medían cráneos, por ejemplo, para mostrar que los ingleses blancos acomodados tenían mejores cabezas que todos los demás. Galton acuñó la palabra *eugenesia* en 1883. Y el vínculo entre la nueva estadística y la eugenesia era institucional, no solo personal. En 1911, Pearson se convirtió en el primer catedrático Galton de Eugenesia del University College de Londres. Allí, su departamento de estadística aplicada creció junto al laboratorio de eugenesia de Galton. La estadística moderna nació, en parte, en las mismas instituciones que la eugenesia, y la dirigieron las mismas personas.

Soy genetista, y a los genetistas nos recuerdan a menudo la eugenesia, con razón. Muchos de los primeros genetistas destacados la apoyaron: Charles

Davenport dirigía la Eugenics Record Office en Estados Unidos. Y la eugenesia no era solo una idea de derechas: el premio Nobel Hermann Muller, genetista y socialista convencido, defendió su propia versión. Otros, como Thomas Hunt Morgan, J. B. S. Haldane y Lancelot Hogben, estuvieron entre sus primeros críticos. Pero los genetistas no estaban solos. Como hemos visto, los fundadores de la estadística moderna estaban en el centro del movimiento. Y Fisher, eugenista, genetista y estadístico a la vez, muestra cómo ambas disciplinas contribuyeron a estas ideas horribles.

¿Por qué contar esta historia en un libro sobre estadística? No porque haga que sus métodos sean erróneos. La correlación, la regresión y el análisis de la varianza funcionan igual de bien los inventara quien los inventara, y Fisher fue uno de los científicos más brillantes del siglo XX. La lección trata de las personas, y también de los métodos. Estaban entre los científicos más rigurosos de su época. La objetividad era su ideal; de hecho, fue uno de los principales motivos por los que Fisher rechazó las probabilidades a priori bayesianas. Y estaban convencidos de que estaban siendo objetivos. Sin embargo, tenían sesgos profundos, sobre todo cuando razonaban sobre asuntos ajenos a su especialidad, como la raza y la clase social. Ahí compartían los prejuicios de su época y de su círculo social.

Fisher nos da otro ejemplo, de una etapa posterior de su vida. En los años cincuenta, cuando los primeros grandes estudios relacionaron el tabaco con el cáncer de pulmón, Fisher sostuvo públicamente que no se había demostrado que la relación fuera causal: quizá los mismos genes que hacían que la gente quisiera fumar también la predisponían al cáncer. Era un argumento ingenioso y, en principio, una pregunta legítima. Pero Fisher era un gran aficionado a la pipa, y trabajaba como asesor para la industria tabacalera. Ahí estaba el mayor estadístico de su tiempo, convencido de que estaba siendo riguroso, defendiendo el bando equivocado en una de las cuestiones de salud pública más importantes del siglo.

Fisher era extraordinariamente inteligente. Pero era un hombre de su tiempo, con el racismo y las ideas eugenésicas habituales entre la gente como él, y con sus propios sesgos personales, como todos nosotros. Esa es la lección incómoda: la inteligencia y el rigor no nos protegen de los sesgos. Es la misma advertencia que vimos con las probabilidades a priori, donde una premisa que parece un número neutral puede ser un fino disfraz del prejuicio. Y se aplica a todos nosotros: todos tenemos sesgos.

Ninguna metodología nos salvará de esto. Esta gente tenía las mejores herramientas estadísticas de su época, muchas de las cuales habían inventado ellos mismos, y las herramientas no los protegieron. Puedes tener un taller magnífico, pero las herramientas solo producen un trabajo excelente en manos de un buen artesano. Hace falta oficio. Ninguna herramienta crea una gran obra por sí sola.

Entonces, ¿qué ayuda? En mi opinión, la mejor protección que tenemos es el diálogo con gente que no piensa como nosotros. Un diálogo de verdad, en el que estemos dispuestos a escuchar a las personas con las que no estamos de acuerdo, y a aprender de ellas. Nuestros propios sesgos son muy difíciles de ver desde dentro, y a menudo resultan evidentes para alguien que no los comparte.

Escuchar no es fácil

Dos de los estadísticos más importantes del siglo XX, Fisher y Karl Pearson, se odiaban. ¿Por qué? Cuando Fisher envió su gran artículo de genética de 1918 a la Royal Society, no se lo aceptaron, supuestamente por las malas revisiones de Pearson. Las revisiones son anónimas, pero estas cosas siempre acaban sabiéndose, y Fisher nunca se lo perdonó.

La venganza de Fisher fue exquisita. La mayor contribución de Pearson a la estadística era la prueba de chi cuadrado. Así que Fisher la estudió con detenimiento y descubrió que Pearson se había equivocado: no había tenido en cuenta correctamente los grados de libertad. Imagínate: llega alguien y le cuenta al mundo que el inventor de la chi cuadrado no sabe calcular la chi cuadrado. Después de aquello, no volvieron a hablarse nunca.

Los dos eran racistas y los dos destacaban en el movimiento eugenésico. Galton fundó la eugenesia, la idea de que a los «no aptos» (los pobres, por ejemplo) se les debía disuadir de tener hijos, o impedírselo, para que la sociedad mejorase. Pearson fundó la revista *Annals of Eugenics*. Siendo estudiante en Cambridge, Fisher ayudó a fundar la sociedad eugenésica de la universidad y, en 1933, sucedió a Pearson como director de los *Annals* y como catedrático Galton de Eugenesia. Lo que pudo haber sido una bonita amistad entre racistas, arruinada por una revisión por pares.

El odio duró generaciones. El hijo de Pearson, Egon, desarrolló el contraste de hipótesis, con su α y su β, junto con Jerzy Neyman. Acabó dirigiendo un departamento en el mismo edificio que Fisher. Casi nunca se hablaban. Y Fisher combatió las ideas de Neyman y Egon Pearson durante el resto de su

vida. Aquí está la ironía: el valor p que te enseñaron es el hijo bastardo de la prueba de significación de Fisher y del contraste de hipótesis de Neyman y Pearson. Sus autores no se soportaban, y los autores de los libros de texto fusionaron los dos enfoques en algo que ninguno de ellos habría aceptado.

Hazte artesano de los datos

Después de tanta filosofía, aquí va un consejo práctico, y no depende de a qué escuela pertenezcas. Como dijimos al principio de todo: describe tus datos, represéntalos gráficamente y no te saltes nunca ese paso.

Los gráficos muestran lo que los números esconden. En 1973, el estadístico Francis Anscombe publicó cuatro pequeños conjuntos de datos con la misma media, la misma varianza, la misma correlación y la misma recta de regresión. Representados gráficamente, tienen un aspecto completamente distinto: uno es una nube razonable de puntos alrededor de una recta, otro es una curva, otro es una recta perfecta con un único valor atípico, y en el último todos los puntos menos uno están sobre la misma línea vertical. Cualquier cálculo que solo mirara esos resúmenes trataría los cuatro conjuntos de datos como idénticos. (Una versión moderna y más divertida, el *Datasaurus*, esconde un dinosaurio entre conjuntos de datos con estadísticos idénticos). Los gráficos son también la forma de comprobar algunos de los supuestos de los que depende cualquier análisis, bayesiano o frecuentista. ¿La relación es realmente una recta? ¿Hay valores atípicos haciendo todo el trabajo? ¿Los puntos son realmente independientes?

Los estadísticos tienen un viejo chiste, atribuido a Joseph Berkson: la mejor prueba es la *prueba del traumatismo interocular*, aquella en la que el resultado te golpea justo entre los ojos. Si un efecto es evidente en un gráfico, la estadística confirmará, en general, lo que ya ves. Si necesitas un cálculo sofisticado para encontrarlo, y los datos no hablan con claridad en los gráficos, ten especial cuidado, diga lo que diga tu valor p o tu probabilidad a posteriori.

Eso sí, una advertencia: nuestros ojos también pueden engañarnos. Vemos caras en las nubes y patrones en el ruido, y mirar docenas de gráficos hasta que aparece algo no es más que otro jardín de los senderos que se bifurcan. Hay una forma bonita de ponerte a prueba, llamada *rueda de reconocimiento* (*lineup*): esconde el gráfico de tus datos reales entre varios gráficos hechos con datos en los que no pasa nada, y mira si alguien es capaz de distinguir el

real. Si no puede, el patrón que viste quizá sea ruido. Es una prueba de significación hecha a ojo. Así que mirar y calcular no son rivales. Los gráficos te dicen qué sospechar y qué comprobar; los cálculos te dicen cuánto fiarte de lo que viste.

La estadística es una de las herramientas más potentes que tenemos para construir mapas del mundo. Pero es una herramienta para razonar bien, no una máquina que produce verdades. El mapa no es el territorio, y ningún método, por elegante que sea, puede hacer que lo sea. Lo que sí puede hacer la buena estadística es ayudarnos a dibujar mejores mapas, a saber cuánto fiarnos de ellos y a cambiarlos cuando el territorio nos dice que estábamos equivocados.

Y, como vimos con Fisher y Pearson, ninguna herramienta hará eso por sí sola. Las buenas herramientas necesitan un buen artesano. Llegar a serlo requiere práctica y comprensión, y no es fácil. E incluso los buenos artesanos tienen puntos ciegos. Por eso creo que lo mejor que podemos hacer, como científicos y como personas, es seguir hablando con quienes no piensan como nosotros, dispuestos a escucharlos y a aprender de ellos.