

Un simulador para asistir en la enseñanza del teorema de Bayes*

María Jesús Bárcena¹, María Araceli Garín¹, Ana Martín¹, Fernando Tusell^{1,3} y Aitziber Unzueta²

¹Dpto. Economía Aplicada 3 (Econometría y Estadística), UPV-EHU

²Dpto. Matemática Aplicada, Estadística y Computación, UPV-EHU

³Autor al que dirigir correspondencia

Abstract

Teaching some basic concepts in Statistics greatly benefits from individual practice with immediate feedback. In order to provide such practice to a large number of students we have written a simulator described below, first of a planned series of teaching aids. It is designed to assist in the teaching of Bayesian concepts and, in particular, to instill in students the idea that probabilities can change as evidence accumulates.

Keywords: *Bayesian statistics, Statistics, simulator.*

Resumen

La enseñanza de algunos conceptos básicos de Estadística se beneficia de prácticas individuales con realimentación inmediata. Para poder llevarlas a cabo con un gran número de alumnos hemos escrito un simulador que se describe a continuación, primero de una serie que planeamos. Su propósito es incrementar la familiaridad de los estudiantes con conceptos básicos de la estadística bayesiana; en particular, familiarizarles con la idea de que las probabilidades cambian a medida que la información aumenta.

Keywords: *estadística bayesiana, Estadística, simulador.*

*Este proyecto ha contado con una subvención del SAE de la Universidad del País Vasco (UPV-EHU), en la convocatoria 2015-2017 de Programas de Innovación Educativa.

1 Introducción

El teorema de Bayes, expresable en su forma más simple mediante la igualdad

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1)$$

permite “invertir” probabilidades: de la probabilidad condicional de B dado A y las respectivas marginales permite pasar a la probabilidad condicional de A dado B .

Muchos ejemplos pueden ser utilizados para imbuir en los estudiantes la importancia práctica de (1). El de la diagnosis médica es uno: la probabilidad $P(A|B)$ de padecer la enfermedad A dada la presencia del síntoma B puede obtenerse en términos de la probabilidad del síntoma dada la enfermedad, $P(B|A)$, y de las respectivas probabilidades marginales de enfermedad y síntoma, $P(A)$ y $P(B)$.

Adquirir familiaridad con los conceptos involucrados —en particular, asimilar que en el enfoque bayesiano la probabilidad es un grado subjetivo de creencia, que cambia a la luz de la evidencia disponible— requiere interacción repetida con (1), más allá de la que un ejemplo en clase permite adquirir. El asignar ejercicios en clase para ser realizados por los alumnos y luego corregidos es una posibilidad, pero muy costosa en tiempo para el profesorado y sólo proporciona realimentación diferida a los alumnos.

El simulador que se describe a continuación proporciona una alternativa. Construido sobre una historia real que admite un tratamiento simple, obliga al estudiante a familiarizarse con la actualización de distribuciones a priori a medida que nueva información se hace disponible, le proporciona realimentación inmediata cuando lo hace mal, y ello sin una carga agobiante sobre el profesorado.

La Sección 2 describe la historia que hemos utilizado, en una recreación simplificada, como motivación de un juego en que los estudiantes deben encontrar un submarino perdido. La Sección 3 describe la implementación, incluyendo cuestiones como coste, escalabilidad y seguridad. La Sección 5 incluye algunos comentarios sobre el papel que atribuimos a este tipo de simuladores en nuestra enseñanza.

2 La pérdida del USS Scorpion

2.1 La historia

El USS SCORPION fue un submarino nuclear de ataque en la Armada de los Estados Unidos. Desapareció el 22 de mayo de 1968, en las proximidades del archipiélago de las Azores, cuando retornaba a su base en Norfolk tras una misión. Las causas son todavía inciertas: se especuló con una explosión, la accidental activación de un torpedo, un encuentro con buques soviéticos y diversos fallos mecánicos.

Cuando transcurrieron varios días sin contacto y se hizo evidente que algo había acontecido, se inició la búsqueda del submarino. La zona contigua a la última posición conocida se dividió en 400 sectores, a cada uno de los cuales se asignó una probabilidad *a priori* de albergar el buque, basada en la opinión de expertos acerca de las posibles causas, reacciones probables del capitán ante cada una, régimen de corrientes, etc.

La búsqueda se realizó utilizando métodos bayesianos, con asesoramiento de estadísticos. Tales métodos habían sido ya puestos en práctica con éxito en la búsqueda de una bomba nuclear accidentalmente caída cerca de Palomares, en España, en 1966.

Tras varios meses de búsqueda, el USS SCORPION fue localizado en octubre de 1968, a más de 3000m de profundidad y a unas 400 millas náuticas de las Azores.

(Cressie y Wikle 2011) y (McGrayne 2012) ofrecen descripciones de la búsqueda del USS SCORPION con atención a los aspectos estadísticos. La Wikipedia también ofrece una historia detallada y enlaces a otras fuentes de información. Una referencia interesante, más técnica, es (Davey y col. 2016), mostrando las técnicas de búsqueda bayesianas empleadas en el caso de vuelo Malaysian Air Lines MH370, desaparecido en 2014. Nuestra presentación de la teoría relevante sigue estrechamente la primera referencia.

2.2 Aproximación bayesiana a la búsqueda

Sea Y_i una variable aleatoria con dos estados: $Y_i = 0$ significa “El submarino está presente en el sector i .” mientras $Y_i = 1$ significa lo contrario.

Análogamente, sea X_i una variable aleatoria codificando el resultado de rastrear el sector i : $X_i = 0$ si el submarino no se localiza en dicho sector y $X_i = 1$ cuando sí se localiza.

Claramente X_i depende de Y_i :

- Si el submarino no está presente en el sector i -ésimo, no lo podemos localizar allí:

$$P(X_i = 1|Y_i = 0) = 0.$$

- Si, por el contrario, sí que está en el sector i -ésimo, la probabilidad de localizarlo allí es p :

$$P(X_i = 1|Y_i = 1) = p.$$

Un rastreo del sector correcto no está garantizado que tenga éxito: $p < 1$: podemos no localizar el submarino en un rastreo del sector i -ésimo aunque realmente esté allí.

Supongamos que la probabilidad *a priori* de encontrar el submarino en el sector i -ésimo es π_i ; si rastreamos dicho sector sin éxito, la probabilidad *a posteriori* de que el submarino esté allí es, utilizando (1):

$$\begin{aligned} P(Y_i = 1|X_i = 0) &= \frac{P(X_i = 0|Y_i = 1)P(Y_i = 1)}{P(X_i = 0)} \\ &= \frac{P(X_i = 0|Y_i = 1)P(Y_i = 1)}{P(X_i = 0|Y_i = 0)P(Y_i = 0) + P(X_i = 0|Y_i = 1)P(Y_i = 1)} \\ &= \frac{(1-p)\pi_i}{(1-\pi_i) + (1-p)\pi_i} \\ &= \pi_i \frac{(1-p)}{1-p\pi_i} \end{aligned} \tag{2}$$

Llamamos la atención de los estudiantes sobre el hecho de que una búsqueda infructuosa del sector i -ésimo no excluye la posibilidad de que realmente esté allí, pero hace la probabilidad *a posteriori* más pequeña que la probabilidad *a priori*: el ratio

$$\frac{(1-p)}{1-p\pi_i}$$

que multiplica a π_i en (2) es menor que 1, y tanto menor cuanto mayor sea p . Como consecuencia de un rastreo infructuoso del sector i las probabilidades de los demás también se modifican. Para $j \neq i$ tenemos:

$$\begin{aligned} P(Y_j = 1 | X_i = 0) &= \frac{P(X_i = 0 | Y_j = 1)P(Y_j = 1)}{P(X_i = 0)} \\ &= \frac{1 \times \pi_j}{1 - p\pi_i} \\ &= \pi_j \frac{1}{1 - p\pi_i} \end{aligned} \tag{3}$$

De nuevo como la intuición sugiere, el hecho de que el submarino no ha sido encontrado en un rastreo del sector i incrementa nuestra creencia de que puede encontrarse en algún otro sector j , $j \neq i$, pues el factor que multiplica la probabilidad *a priori* π_j en (3) es mayor que 1.

3 El simulador: uso e implementación

3.1 Objetivos condicionando el diseño

El objetivo no es construir una herramienta para *enseñar*, sino para adquirir familiaridad con algo *ya aprendido*, aunque quizá no completamente asimilado e internalizado. En consecuencia, la escasa teoría relevante (básicamente el contenido de la Sección precedente) se presenta en clase y previamente al uso del simulador se entrega a los estudiantes un guión describiendo su funcionamiento.

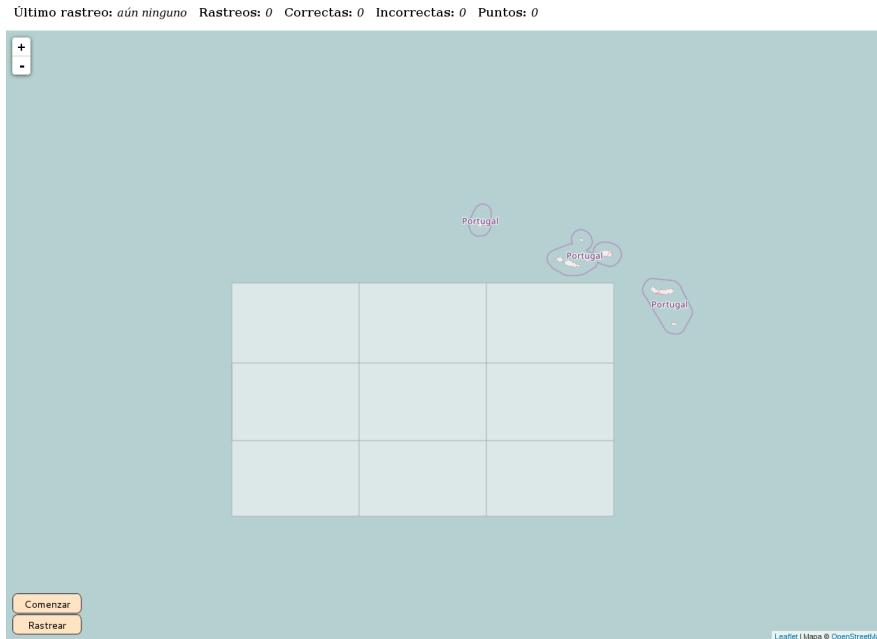
Un requerimiento era que el simulador había de construirse con los recursos disponibles, sin necesitar mano de obra especializada o hardware adicional. En la práctica, ello supone que había de ser programado por nosotros y debía funcionar sobre alguno de los servidores del Departamento, sin cargar en exceso los mismos ni perturbar sus otras funciones. Fue preciso algún aprendizaje de JAVASCRIPT por nuestra parte, al concluirse tempranamente que ésta era la herramienta adecuada.

Finalmente, otro de los requerimientos condicionando el diseño es que había de servir a un número elevado de estudiantes, geográficamente dispersos. Esto reduce las opciones a un simulador basado en una página web, que no requiera instalación de ningún software en la máquina cliente y pueda correr en cualquier navegador con JAVASCRIPT activado.

3.2 Implementación

Se ha programado un simulador que enfrenta al estudiante a las mismas decisiones que hubo de tomar el equipo de búsqueda del USS SCORPION, pero en una situación simplificada: en lugar de los 400 sectores en que se dividió la zona de búsqueda, sólo nueve se presentan por el simulador sobre un mapa (ver Figura 1).

Fig. 1: Pantalla inicial del simulador



Para arrancar el simulador, el estudiante debe pinchar el botón **COMENZAR** en el margen inferior izquierdo del mapa. El simulador entonces genera una nueva instancia del juego asignando probabilidades *a priori* a los nueve sectores rastreables, los representa en un mapa y coloca la silueta de un buque de búsqueda junto a la esquina sudeste del área rastreable (Figura 2).

Un *click* en cualquiera de los sectores proporciona información sobre su probabilidad *a priori* en cualquier momento. Esta probabilidad se codifica también como saturación de color en una paleta de verdes¹.

La primera búsqueda es trivial: simplemente basta seleccionar el sector con la mayor probabilidad *a priori*. Para ello, el estudiante sólo debe arrastrar el buque de búsqueda al sector correspondiente y hacer *click* sobre él, o sobre el botón **RASTREAR** en la esquina inferior izquierda: esta última alternativa se ha mostrado necesaria en algunas tabletas con pantallas pequeñas.

Después de cada elección del estudiante, el simulador actualiza las cabeceras con el último sector rastreado, número de rastreos “correctos” e “incorrectos” y puntos

¹Una diferente paleta apta para daltónicos está también disponible, si la necesidad surge.

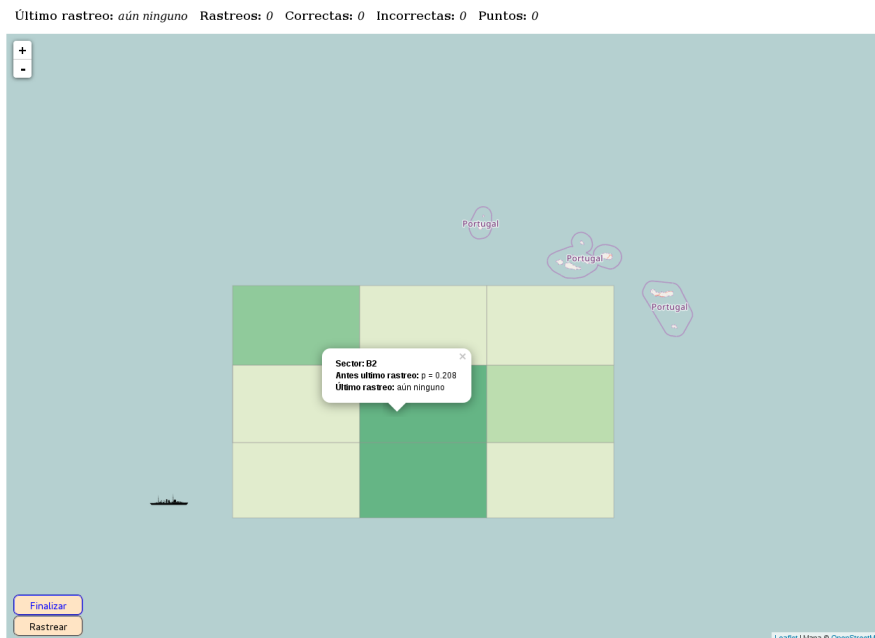
obtenidos. Un rastreo es “correcto” si se realiza sobre el sector que más probablemente puede contener el submarino.

El simulador informa al estudiante cuándo el submarino es encontrado, lo que da fin al juego, o actualiza las probabilidades de cada sector y permite al estudiante seguir la búsqueda. Estas probabilidades actualizadas, sin embargo, no se muestran ni se modifica la coloración de cada sector en la pantalla de inmediato. **La pantalla siempre muestra las probabilidades a priori ANTES de realizado el último rastreo. Es misión del estudiante el actualizarlas mediante las fórmulas (2)-(3).**

Puede indicarse a los estudiantes, o dejar que lo descubran por si mismos, que en cada etapa tienen sólo dos posibilidades potencialmente correctas: el último sector rastreado —que incluso después de una búsqueda infructuosa podría seguir siendo el más prometedor— o aquél de los no rastreados que previamente al último rastreo tenía máxima probabilidad *a priori* —pues la actualización bayesiana opera multiplicando todas las probabilidades *a priori* de los sectores no rastreados por el mismo factor, y por tanto preserva su orden; véase (3)—.

El juego finaliza con el hallazgo del submarino y la oferta al estudiante de guardar la puntuación obtenida (cociente de rastreos correctos entre rastreos efectuados en una escala de 10, o como se determine). La puntuación no depende de lo larga que sea la búsqueda, sino de la corrección de las decisiones tomadas. Es opción del profesor dar a los estudiantes un sólo acceso al juego o (como nosotros preferimos) darles opción de jugar repetidamente, conservando sólo su última nota.

Fig. 2: Pantalla del simulador, iniciado el juego y antes de la primera búsqueda. El barco de búsqueda puede verse en la esquina sudeste. El usuario ha hecho *click* sobre uno de los sectores, lo que abre una lengüeta mostrando su probabilidad *a priori*. Tonalidades de verde más oscuro codifican mayores probabilidades *a priori*.



3.3 Aspectos de la implementación

Cuando por primera vez pensamos en este proyecto, concebimos una página web en PHP², principalmente porque esta era la técnica empleada en otros proyectos previos. En nuestros intentos de representar un mapa y situar objetos en él, pronto reparamos en la simplicidad y elegancia de la librería LEAFLET³, escrita en JAVASCRIPT. El uso de LEAFLET libera de las complejidades asociadas a la representación espacial, proyecciones, zoom, etc. y permite descargar teselas de lugares públicos sin coste. Esto nos permite concentrarnos en los aspectos puramente estadísticos de la simulación. Adicionalmente, JAVASCRIPT es un lenguaje maduro, expresivo y desarrollado, que proporciona todas las facilidades requeridas para un proyecto de esta naturaleza (generador de números aleatorios, coma flotante, etc.).

Hay una ventaja adicional en el uso de JAVASCRIPT. A diferencia de PHP, que se ejecuta íntegramente en el servidor, JAVASCRIPT es interpretado y ejecutado en las máquinas clientes, sin requerir más que un navegador razonablemente actual y con JAVASCRIPT activado. Esto hace que nuestra aplicación sea escalable prácticamente sin límites. Podemos servir a números muy grandes de estudiantes, incluso con recursos de servidor y caudal de red moderados. La única carga significativa sobre el servidor se produce al inicio, con la descarga de la página web y librerías de soporte (que pueden descargarse también de lugares alternativos y en todo caso permanecen en cache para sucesivas ejecuciones). Al final del juego, las puntuaciones de los estudiantes deben ser guardadas, pero esto es una demanda infrecuente (e insignificante) sobre el servidor.

Hay un posible inconveniente en el uso de JAVASCRIPT. El hecho de que es un lenguaje interpretado que se ejecuta en la máquina cliente significa que el programa puede ser inspeccionado por el usuario. Esto puede dificultarse ofuscando la fuente o cargándola en forma pre-procesada (en un fichero `.js`); pero es cierto que un usuario técnicamente competente y con la suficiente determinación y tenacidad puede acceder a la fuente si se lo propone.

Esto probablemente no representa un problema en una aplicación como la nuestra. Las vulnerabilidades de seguridad sólo son de temer cuando los incentivos exceden el coste de del ataque preciso. Dado el muy modesto esfuerzo que requiere la actividad que este simulador propone al alumnado, no es verosímil que nadie intente invertir mucho tiempo y energía en vulnerar el código, máxime en presencia de una política como la nuestra de “juega cuanto quieras hasta que tu nota sea óptima”. No obstante, la posibilidad existe.

En algunos detalles de la implementación se ha hecho “trampa”. El submarino no está desde el principio en ningún sitio. Lo que aleatorizamos es el número de rastreos que el estudiante deberá efectuar: esto es para evitar la posibilidad de un hallazgo al primer intento, que privaría al estudiante de la oportunidad de practicar lo que queremos que practique —y a nosotros de la oportunidad de comprobar que lo ha hecho con corrección—. En todo evento, el estudiante deberá al menos realizar seis rastreos. Esto es fácil de cambiar.

²Véase por ejemplo <https://secure.php.net>. Visitada el 8 de marzo de 2017.

³Ver <http://leafletjs.com>. Visitada el 8 de marzo de 2017.

Otro aspecto que el profesor puede desear ajustar es el valor del parámetro p —la probabilidad de éxito al rastrear el sector correcto. Se sigue del párrafo precedente que no tiene ninguna influencia sobre la duración del juego; pero si la tiene —y muy grande— sobre la actualización de las probabilidades *a priori*. Si p se fija muy grande, un rastreo infructuoso reduce drásticamente la probabilidad *a posteriori* de que el sector rastreado albergue el submarino. La elección subsiguiente por tanto será casi invariablemente el sector con mayor probabilidad *a priori* antes del último rastreo. Los estudiantes no tardarán en detectar la pauta, que permite jugar sin hacer recurso a las fórmulas (2)–(3) —lo que desvirtúa la finalidad didáctica del simulador—. Es por ello aconsejable fijar p no muy alto. Un valor en torno a 0,4 es adecuado.

El simulador es accesible en <http://www.et.bs.ehu.es/bayes> y la fuente está disponible en <https://github.com/FernandoTusell/BayesSim.git>, para su utilización o modificación por terceros sin mas limitaciones que las contenidas en la licencia GPL. Agradecemos correcciones y modificaciones, particularmente en forma de *pull requests*.

4 Recursos similares

Previamente a iniciar nuestro proyecto rastreamos Internet en búsqueda de trabajo similar al realizado. Aunque no encontramos nada cubriendo el tema preciso que buscábamos, hay abundantes recursos que utilizan juegos o simulaciones de alguna especie con finalidad didáctica. La mayoría de ejemplos que hemos encontrado están relacionados con diseño experimental y tienen una historia que en ocasiones se remonta a los trabajos pioneros (Mead y Stern 1973) y (Pike 1974); véase (Stern, Latham y Stern 2009), por ejemplo.

Más cercano a nuestros objetivos de introducir a los estudiantes a los rudimentos del planteamiento bayesiano es (Erickson 2017), que propone dos ejemplos de actividades, con énfasis en gráficos de mosaico como ayuda a la intuición.

Aunque no directamente relevantes a nuestro propósito, es de interés una espléndida colección de juegos interactivos disponibles en la página web de Statistics New Zealand⁴; la utilización de Flash PlayerTM unida a un contenido bien concebido proporciona un resultado muy atractivo. La orientación es diferente de la nuestra, en cuanto son juegos más dirigidos a enseñar que a practicar lo aprendido.

5 Discusión

La enseñanza en Estadística, como en otras especialidades, requiere diferentes estrategias para diferentes temas y audiencias en diferentes momentos. Pensamos que no hay reemplazo para la interacción directa entre un profesor motivador e involucrado y sus estudiantes de cara a la adquisición de ideas de relativa complejidad o profundidad.

Hay sin embargo nociones no particularmente difíciles, pero que requieren tiempo y práctica para permear las mentes de los estudiantes. Este es el ámbito en el que

⁴At http://www.stats.govt.nz/tools_and_services/schools_corner/Activities/Interactive-Games.aspx, visitada el 10-marzo-2017.

situamos ayudas a la enseñanza como la presentada, que permiten hacer trabajar a un grupo numeroso de estudiantes sobre instancias únicas de un mismo problema y obteniendo realimentación inmediata a sus aciertos y errores.

Planeamos extender nuestro trabajo a la construcción de simuladores orientados a este tipo de temas, que requieren para su asimilación práctica sobre ejemplos. No deja de ser interesante que algunos de los que más naturalmente se prestan a implementación en un formato similar al empleado en el trabajo aquí presentado ya aparecen anticipados en (Mead y Stern 1973) y (Pike 1974). Es una muestra del gigantesco avance en la disponibilidad de recursos educativos basados en ordenador que lo que era en aquel momento casi visionario pueda hoy ser implementado en tiempo razonable, por no especialistas y con esfuerzo y recursos muy modestos.

Referencias

- Cressie, Noel y Christopher K. Wikle (2011). *Statistics for Spatio-Temporal Data*. Wiley.
- Davey, Samuel y col. (2016). *Bayesian Methods in the Search for MH370*. Springer-Verlag.
- Erickson, Tim (2017). “Beginning Bayes”. En: *Teaching Statistics* 39.1, págs. 30-35.
- McGrayne, S.B. (2012). *The Theory That Would Not Die: How Bayes’ Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale Univ. Press.
- Mead, R. y R.D. Stern (1973). “The use of the computer in the teaching of statistics (with discussion)”. En: *Journal of the Royal Statistical Society, Ser. A* A136, págs. 191-225.
- Pike, D.J. (1974). “Statistical games as teaching aids”. En: *Journal of the Royal Statistical Society, Ser. D (The Statistician)* D25, págs. 109-115.
- Stern, David, Sonya Latham y Roger Stern (2009). “Statistical games to support problem-based learning”. IBS SUSAN Conference Proceedings. Retrieved from http://www.reading.ac.uk/ssc/resources/Docs/SUSAN/Statistical_Games_to_supportproblem-based_learning.pdf.