Probabilidad, canicas en urnas y estadística de la (bio)diversidad

J. Andrés Christen

Si alguna vez hemos tomado un curso de probabilidad o, en general, de matemáticas, gran parte de dicho curso, si no es que todo, nos la pasamos diciendo: "y todo esto, ¿para qué?" Lo que recordamos de algún curso de probabilidad son planteamientos como: "Si las placas de los coches se identifican con tres letras y tres números, ¿cuál es el número total de las placas posibles si ninguna letra puede usarse más de una ocasión en la misma placa?, ¿cuál es el número total sin esta restricción?" También se presentan los casos clásicos de barajas: "¿Cuántas personas deben escoger una carta, cada una de diferente baraja, para tener una probabilidad mínima de 0.9 de que por lo menos se escoja un as?"; o los tradicionales de canicas: "Se tiene una urna con canicas rojas, verdes y azules. Se sacan tres canicas, una a la vez, regresando cada una a la bolsa: ¿Cuál es la probabilidad de que la primera sea verde, la segunda azul y la tercera roja?"

Sí, ciertamente estos problemas parecen no tener sentido. Y más nos lo parecen cuando abundamos en el estudio de la probabilidad o la estadística y nos damos cuenta de que el resto del curso no tiene nada que ver (aparentemente) con barajas y canicas en urnas. Sin embargo, aun cuando artificiales, estos ejercicios constituyen las bases para resolver problemas más complejos, de los que llamamos "de a de veras". Un caso importante es el siguiente: "Se tienen canicas de varios colores en una urna y se puede sacar una canica a la vez regresando ésta a la urna: ¿Cuántos diferentes colores de canicas hay en la urna?"

En apariencia es un problema tonto. Podríamos simplemente sacar canicas hasta que hayan salido *todos* los colores. Pero, claro, el meollo del asunto es que *no* sabemos cuántos colores hay. La pregunta entonces es: ¿cuándo dejamos de sacar canicas?

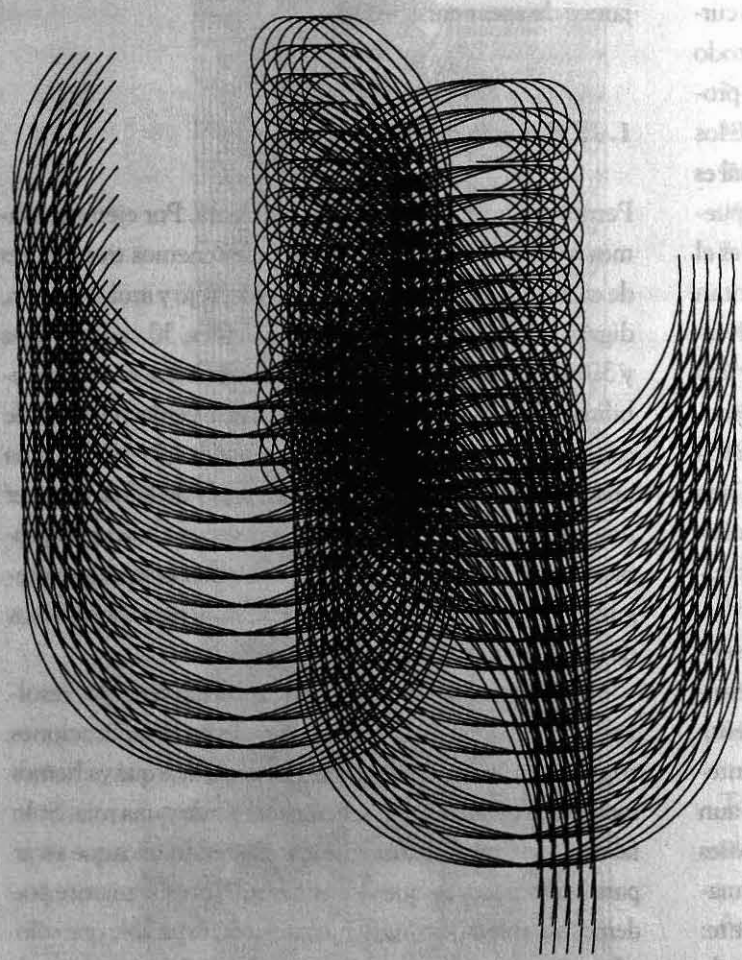
1. *El ejercicio de clase*

Pensemos un poco más en el problema. Por ejemplo, vamos a suponer por el momento que tenemos tres colores de canicas en la urna, digamos, verde, rojo y azul. Más aún, digamos que tenemos 30 canicas verdes, 30 canicas rojas y 30 azules. Entonces, es claro que tendremos una probabilidad de 1/3 de obtener una canica verde (el total de canicas es 90, y hay 30 verdes, entonces $1/3=30/90$), una probabilidad de 1/3 de obtener una roja y de 1/3 de obtener una azul. Sacamos canicas y las regresamos a la urna; el problema entonces lo podríamos resumir con la siguiente pregunta: ¿cuántas canicas debemos de sacar para asegurarnos de que hayan salido los tres colores?

Aun cuando éste parece un problema sencillo, resolverlo no es nada trivial. Para evitar caer en complicaciones técnicas y para agilizar la discusión, pensemos que ya hemos sacado dos canicas y que salieron una verde y una roja. Sólo faltaría una azul. ¿Cuántas canicas más tendríamos que sacar para asegurarnos de que salga una azul? Intuitivamente podemos ver que, si sacamos dos o tres más, es posible que sólo salgan canicas verdes o rojas. La probabilidad de que salga azul es 1/3, entonces la probabilidad de no sacar una canica azul es $1-(1/3)=2/3$. La probabilidad de no sacar una canica azul en dos intentos es $(2/3)\times(2/3)$ (la multiplicación, por ser eventos independientes), etcétera. Entonces, pensando en general, la probabilidad de no sacar ninguna canica azul

en n intentos seguidos es $(2/3)^n$. Esta probabilidad es aproximadamente 0.66, 0.44, 0.29, 0.19, 0.13, para n=1, 2, 3, 4, 5 intentos respectivamente. Nótese que dicha probabilidad se va haciendo pequeña, lo cual coincide con la idea intuitiva de que mientras más canicas saquemos, más probable es que terminemos por sacar una canica azul. Sin embargo, y esto es lo importante, $(2/3)^n$ es siempre un número mayor que cero. Es decir, por más canicas que saquemos (por más grande que sea n), siempre existe la posibilidad de que *no* salga una azul. Imaginemos por un momento que no sabemos que hay tres colores y que sacamos 10 canicas, y todas salen verdes y rojas (¡esto es posible para nuestra urna!) ¿Nos detendríamos ahí y diríamos que sólo hay canicas verdes y rojas?

El caso de nuestra urna es el caso simple, en el que hay igual número de canicas para cada color, el caso homogéneo. Podría ser mucho más complejo si a nuestra urna le agregáramos, por ejemplo, una sola canica blanca (el caso no-homogéneo); sería muy difícil sacar dicha canica y fácilmente pensaríamos que sólo había tres colores. Por dificultades técnicas que están más allá de los alcances de este artículo, no vamos a explicar con mayor detalle la solución de este problema. En los párrafos anteriores sólo quisimos dar una descripción somera de los elementos implícitos en dicha solución. En las siguientes secciones presentaremos el problema real y cómo se aborda, explicando brevemente cuál es una solución del mismo.



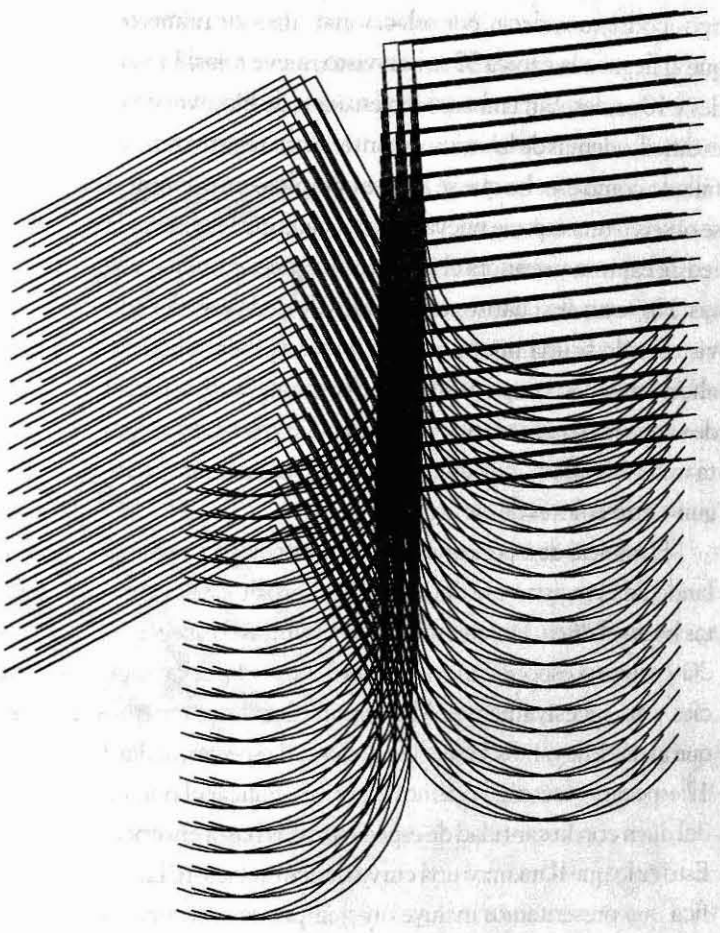
2. *El problema real*

Hasta ahora parece ser que estamos en un curso tradicional de probabilidad y estadística, hablando de canicas en urnas y esas cosas. Pero, ¿qué tal si hablamos de vampiros en la selva de Chiapas? ¿Cuántas diferentes especies de vampiros hay? Este tipo de preguntas es común en estudios de biodiversidad, donde se requiere, por ejemplo, saber sobre la ecología y el número y tipos de especies que habitan en una cierta región. Esto es muy importante hoy en día para el manejo y conservación de reservas naturales y para estudios de impacto ambiental, entre otras muchas cosas. Bueno, para saber cuántas diferentes especies de vampiros habitan cierta región, lo que se nos ocurre en primera instancia es poner redes de captura y clasificar a los vampiros por especie. Cada vez que aparece una especie nueva la anotamos y regresamos los ejemplares a su medio (no es necesario matarlos). ¿Nos suena esto familiar?

¿Qué tal si pensamos en la región de estudio como si fuera una urna y en cada vampiro como una canica? Cada color para las canicas será entonces una especie diferente, y sacar una canica será capturar un vampiro. En pocas palabras, nuestro problema de clase relativo a las canicas en urnas se ha vuelto un problema con una aplicación en la realidad. Lo que es más, hemos tomado como ejemplo un problema de biodiversidad, pero hay muchos otros ámbitos donde aparecen problemas de diversidad. Por ejemplo, tenemos un programa de computadora (un *software*) y queremos saber cuántos errores tiene. Lo *corremos* y vamos encontrando errores: ¿cuándo sabemos que ya no hay más errores? ¿Cuándo está listo para entregárselo a nuestro cliente? En este caso, los errores son los colores de las canicas y sacar una canica es toparse con un error. Otra aplicación tiene lugar en el campo de la lingüística, cuando se requiere

conocer el número total de palabras usadas por un autor. Hay además otras muchas aplicaciones en actividades tan diversas como la numismática, la astronomía, la sociología, etcétera.

Notamos que fue "fácil" traducir el problema de urnas y canicas a nuestro problema real de (bio)diversidad. Sin embargo, el lector podrá notar que se presenta una serie de simplificaciones al llevarse a cabo dicha traducción. Por ejemplo, podemos capturar muchos vampiros al mismo tiempo, mientras que no es permitido sacar muchas canicas simultáneamente. Puede ser el caso, también, que algunas especies se escondan en días de lluvia, supongamos, y otras no. Esto es como si de repente algunas canicas se escondieran entre las otras dentro de la urna y nuestra mano sólo pudiera sacar canicas de ciertos colores. Sin embargo, el problema de las canicas es lo que es: un modelo, relativamente sencillo, para un problema real, mucho más complejo.



Si pensamos en un modelo a escala de una casa (una maqueta), éste será muy diferente de la casa real. En lugar de paredes de ladrillos y cemento se usarán paredes de cartón, en lugar de ventanas con vidrios tendremos cuadritos con celofán, no será del tamaño real, etcétera. Pero, precisamente, lo que *no* queremos es la casa en sí, sino un modelo de ella. El modelo es mucho más fácil y rápido de hacer, lo podemos ver con facilidad desde cualquier ángulo y podemos modificarlo a nuestro antojo. No es la casa real, pero se le parece lo suficiente como para que podamos darnos cuenta de cómo será cuando esté terminada. Y si no nos gusta la maqueta, pues cambiamos el diseño de la casa modificando los errores del diseño original. Un modelo matemático es parecido a una maqueta. No es el problema real, sino una gran simplificación de éste.

Así pues, a nuestro problema de (bio)diversidad le hemos hecho una "maqueta". Nos imaginamos que capturar especies es como sacar una canica e identificar su color. Así se simplifican muchas cosas, pero, a la vez, tenemos la posibilidad de avanzar, de utilizar la información disponible de manera coherente y hasta de predecir cosas. Éste es un proceso común en la ciencia, donde utilizamos modelos matemáticos para abordar los problemas en estudio. Nótese, sin embargo, que el problema del cual estamos hablando tiene una naturaleza eminentemente aleatoria. Nunca sabremos con exactitud de qué color será la siguiente canica que saquemos de la urna. Aun así es posible hacer modelos matemáticos que incluyan esa aleatoreidad. Éste es el ámbito de los modelos probabilísticos.

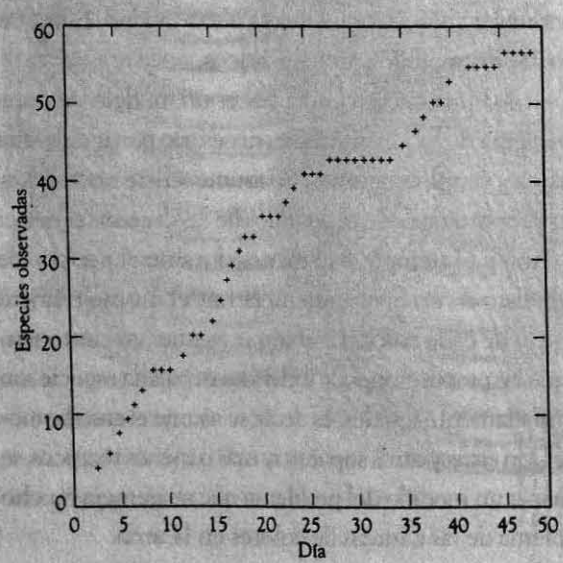
Como dijimos, para poder hacer un modelo de nuestro problema de biodiversidad es necesario partir de varios supuestos y simplificaciones. Se asume, entre otras cosas, que las proporciones de individuos de las diversas especies no varían en el tiempo, o al menos durante el periodo de estudio. Esto es, en la urna siempre hay el mismo número de canicas de cada color. También se asume (en cierto sentido) que las proporciones de individuos de cada especie son aproximadamente iguales; es decir, se asume el caso homogéneo. Con estos y otros supuestos, más o menos técnicos, se logra hacer un modelo del problema que se asemeja mucho al problema de las canicas de colores en la urna.

3. *Curvas de acumulación*

Hay algunos casos en estudios de (bio)diversidad en que se mantiene un registro del número de individuos vistos

de cada especie. A título de ejemplo de un registro de este tipo, podríamos decir, por seleccionar algunos números, que al llegar a la canica 30 se han visto nueve rojas, 11 verdes y 10 azules. Sin embargo, en estudios de biodiversidad es difícil, además de laborioso, mantener un registro tan detallado como ése. Lo que se hace es simplemente anotar si se observó una especie nueva. Esto es, cada día se observa la red de captura y se anota el número de especies nuevas vistas. No se anota cuántos individuos de cada especie son vistos, sólo si una nueva especie ha sido observada. Esto ahorra mucho tiempo y esfuerzo: distinguir entre especies de vampiros (o de otro tipo de seres) no es trivial, representa una gran labor, muy diferente de la que entraña distinguir entre colores de canicas.

El número de especies nuevas, vistas por día, se va acumulando en el registro y se obtiene el número de especies vistas hasta el día *n*. Por ejemplo, el día 1 vimos 5 especies, el día 2 vimos 3 especies nuevas, el día 3 vimos 4 nuevas especies y así sucesivamente. Al acumular los datos tenemos que al día 1 llevamos 5 especies, al día 2, 8 especies, al día 3, 12 especies, etcétera. Podemos entonces graficar el número del día *n* con la cantidad de especies vistas hasta entonces. Esto es lo que llamamos una curva de acumulación. La gráfica que presentamos incluye un ejemplo de una curva de acumulación (los datos no son reales; se ofrecen únicamente para ilustrar el problema).



Tenemos entonces que toda la información disponible para estimar el número total de especies está contenida en la curva de acumulación. De la gráfica se desprende que para el día 45 o 50 ya casi no encontramos especies nuevas. La curva se está "aplanando". Por otro lado, vemos que la curva creció mucho, casi a un ritmo constante, del día 1 al día 15 o 20. Entonces pensamos lo siguiente: si la curva está creciendo, aún faltarán especies por ver. Si la curva ya se estabilizó, ya no habrá muchas más especies por ver. Por lo tanto, importa no sólo el total de especies vistas, o sea, hasta dónde llegó la curva, sino la forma de ésta y la manera en que creció.

Nosotros (el autor y sus colegas) hemos abordado este problema de la manera siguiente: se plantea un modelo probabilístico que describe los saltos en la curva de acumulación, esto es, el número de especies nuevas vistas cada día. Tomando varios supuestos, se llega a un modelo, no muy complejo,[1] que describe dichos saltos. Uno de los parámetros del modelo es el total de especies. Dicho parámetro es estimado a través de métodos estadísticos, con lo que obtendremos una estimación del número total de especies, esto es, las especies vistas más una estimación de las que faltan por ver. La estimación toma en cuenta la manera en que evolucionó la curva de acumulación, lo cual coincide perfectamente con lo expuesto anteriormente, sin embargo, dados los alcances de este artículo, no abundaremos más en los detalles técnicos de esta estimación.

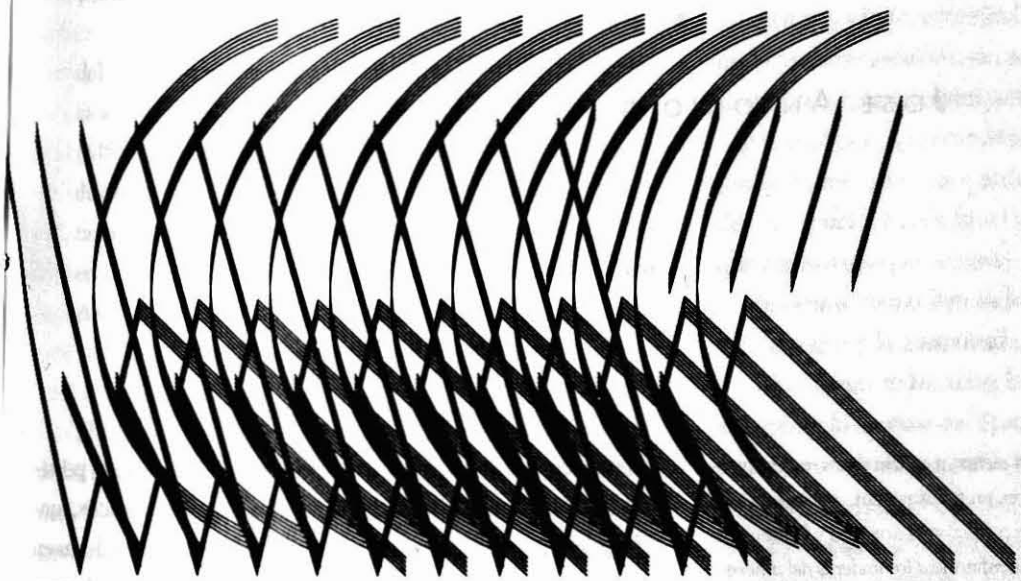
La estimación del número total de especies (por el tipo de estadística usada) es una probabilidad para cada número posible del total, a partir del número de especies vistas. Esto es, en el caso de la curva de acumulación presentada en la gráfica, se observaron 57 especies distintas. La probabilidad de tener un número total de especies menor que 57 es 0 (obviamente). La estimación de la probabilidad de tener un número de especies de 57 a 63, resulta ser (redondeando) de 0.02, 0.05, 0.07, 0.09, 0.1, 0.09 y 0.08. Esta probabilidad sube y luego baja, situándose el máximo en 61 especies (esto es, 0.1). Después de 80 especies resulta que la probabilidad es menor que 0.004. Si sumamos las probabilidades de tener desde 57 hasta 72 especies, acumulamos más de 0.9. Lo anterior significa que con una probabilidad de más de 0.9 (90%) esperamos que existan entre 57 y 72 especies en nuestra región (15 especies más de las observadas hasta ahora).[2]

Por lo explicado antes en relación con el problema de las canicas en la urna, siempre cabe la posibilidad de que no

[1] Es decir, "no muy complejo" en relación con los modelos probabilísticos en general; desde luego que dicho modelo es demasiado técnico para ser explicado aquí.

[2] Los cálculos pertinentes para las estimaciones fueron llevados a cabo en una computadora personal usando el lenguaje LISP.

hayamos visto algún color de canica, esto es, alguna especie (o varias). Esto se refleja en nuestros resultados, pues siempre hay una probabilidad positiva para cualquier total de especies, por grande que éste sea. De hecho, es necesario que establezcamos un máximo para el total de especies (cosa que es posible en estudios de biodiversidad, porque el número de especies de cualquier índole es siempre finito).



4. *¿Cuándo parar?*

Aún queda por plantear la pregunta fundamental: ¿cuándo parar? Usando nuestro modelo, y diversas técnicas estadísticas, que no explicaremos en detalle, hemos podido hacer una estimación del número de especies nuevas por ver en futuros días. Por ejemplo, refiriéndonos a la curva de acumulación presentada en la gráfica podemos preguntarnos qué pasaría si nos quedamos cinco días más buscando especies nuevas. Resulta entonces que la probabilidad de observar cero especies nuevas es (redondeando) de 0.3. Para 1, 2, 3, 4 y 5 especies nuevas dicha probabilidad es (redondeando) de 0.3, 0.2, 0.1, 0.03 y 0.01, respectivamente. Vemos entonces que tenemos una probabilidad aproximada de 0.9 de ver tres especies nuevas, o menos, en cinco días más de observación.[3]

¿Qué tan caro, o qué costo representa, quedarnos cinco días más en la selva para encontrar unas dos o tres especies más de vampiros? Ésa es la cuestión principal aquí. Dadas las probabilidades obtenidas para el número de especies nuevas por ver, ¿qué tan costeable es que sigamos muestreando o no? La pregunta ¿cuándo parar?, no se puede contestar en abstracto, tendrá que formularse en términos de lo que representa observar nuevas especies, esto es, el costo de hacerlo. En el caso de la urna, cuesta muy poco sacar una canica, anotar su color y regresarla a la urna. Dará casi lo mismo sacar cinco que 10 más. Sin embargo, en estudios de (bio)diversidad muestrear puede ser muy costoso. Entonces, puede y debe hacerse un análisis de costo-beneficio para tener un punto óptimo para detenerse en la búsqueda de nuevas especies, en caso de que la información extra que se podría obtener sobre el total de especies no justifique su costo.

Como vemos, el problema de clase sobre las canicas de colores en urnas representó la base para resolver un problema real, complejo e importante. A nosotros siempre nos interesa llevar al salón de clase aplicaciones donde se usen las técnicas en estudio. La probabilidad y la estadística son disciplinas indispensables para abordar problemas relevantes, ya sea de la ciencia y la técnica o de la vida cotidiana. Los cursos de probabilidad y estadística nos abren las puertas a un mundo poderoso, de conceptos profundos y grandes repercusiones, aun cuando los ejercicios de dados, barajas y canicas no nos lo muestren de golpe.

Desde luego que en esta exposición solamente hemos planteado el problema de la estimación del número de especies. Si el lector requiere más detalles técnicos, le sugerimos que nos contacte. Este trabajo aún está en desarrollo y por el momento sólo hemos atacado el problema dentro de una región y en un tiempo limitados. Un aspecto muy importante es observar el cambio de la biodiversidad en el espacio y el tiempo para saber más acerca de migraciones, cambios temporales y/o locales de especies, entre otras cosas. Además, sólo hemos considerado problemas de biodiversidad y es posible que el lector esté enfrentando otros problemas de diversidad que rebasan ese campo. Nos gustaría saber de ellos. ◆

[3] Véase la nota 2.