Selección de una muestra ordenada con semillas y algoritmos de números aleatorios.

Gabriel Camaño y Juan José Goyeneche
Instituto de Estadística
Facultad de Ciencias Económicas y de Administración
Universidad de la República

28 de febrero de 2011

Resumen

Se plantea seleccionar una muestra ordenada de tamaño n de una lista de N participantes, asegurando las condiciones de: equidad, transparencia y simplicidad del proceso. Se describe un procedimiento informático que, en combinación con *semillas* aleatorias seleccionadas por la Dirección Nacional de Loterías y Quinielas, asegura una muestra que cumple las condiciones requeridas. El algoritmo proporciona una lista ordenada de las N personas, de modo que se puede extender la muestra de n de ellas en caso de ser necesario.

Se presenta un ejemplo en el cual hay que seleccionar 2.000 personas de un total de 39.751 inscriptos para un determinado Organismo.

1. Introducción.

En el Instituto de Estadística de la Facultad de Ciencias Económicas y de Administración de la Universidad de la República (IESTA) se ha diseñado un método informático para ser aplicado en sorteos que implican seleccionar un número elevado de participantes de una lista de inscriptos. El IESTA puso especial énfasis en que el método desarrollado poseyera tres características fundamentales:

1. Transparencia.
2. Igualdad de oportunidades (equiprobabilidad).
3. Simplicidad.

TRANSPARENCIA En este tipo de sorteos la transparencia implica que los resultados sean reproducibles, en el sentido de que el ordenamiento de los individuos a sortear sea invariante dada las mismas condiciones iniciales. De esta

manera se garantiza que los resultados, dadas las mismas condiciones iniciales, son exactamente los mismos independientemente de que persona u organismo aplique el método.

EQUIPROBABILIDAD Es de vital importancia que todos los individuos a sortear tengan exactamente la misma probabilidad de ser seleccionados. Si esto no fuera cierto se estaría violando las expectativas de los inscriptos a participar en un proceso limpio y justo. Por lo tanto, el método a utilizar debe evitar que haya individuos que tengan probabilidad muy baja o alta de ser seleccionados. Más aún, debe evitar los casos extremos: casos que tengan probabilidad cero de ser seleccionados. Un individuo que tenga probabilidad cero de ser elegido nunca será seleccionado, tornando inútil que se haya anotado para participar en el sorteo.

SIMPLICIDAD El método a utilizar debe ser fácil de comprender por el público en general y por aquellos interesados en conocerlo. Se debe utilizar un método que sea fácilmente entendible por una persona que no sea experta en la materia

Un método de sorteo que aúne las tres características previamente detalladas asegura que el participante tiene la garantía plena de que se están respetando todos sus derechos.

2. Método de muestreo

El método utilizado por el IESTA tiene dos pasos, los cuales se implementan conjuntamente por medio de un solo programa de computación, pero que se detallan separadamente con el propósito de claridad:

1. El Organismo en cuestión proporciona la lista de aspirantes. Esta lista podría haber sido ordenada por el personal de dicho Organismo de acuerdo a un criterio específico con fines administrativos; por ejemplo, podría estar ordenada alfabéticamente por apellido, o por número de cédula, o por número de inscripción. Dicho ordenamiento, lo haya o no, es desconocido para los investigadores del IESTA.

   Para eliminar cualquier ordenamiento específico que podría estar presente, se utiliza un primer número elegido al azar que será utilizado para "mezclar" o "desordenar" la lista de aspirantes. Esto se hace utilizando el primer número sorteado entre 00000 y 39.999 por la Dirección Nacional de Loterías y Quinielas, como "semilla" de un algoritmo computacional que asignará números al azar entre 0 y 1 (sin repeticiones) a cada uno de los integrantes de la lista. Paso seguido, se ordena la lista de menor a mayor de acuerdo a la magnitud de los números generados[1]. Esto resulta en que

[1] En el supuesto caso que el primer número sorteado sea el 00000, se utilizará como semilla el número 40.000.

el ordenamiento de los aspirantes sea totalmente aleatorio, no sigue ningún criterio prefijado. A esta lista resultante le llamaremos lista "mezclada".

2. Se sortea un segundo número entero mayor o igual a 1 pero menor o igual al número total de aspirantes, en el caso presentado, un número entre 00001 y 39.751. Este número indica la posición que ocupa el primer aspirante seleccionado en la lista "mezclada". A partir de esta persona, se seleccionan las 39.750 que le "siguen" en la lista, lo que determinará en definitiva las personas seleccionadas. Si k es el segundo número sorteado, el método produce una nueva lista de N individuos que consiste en las personas ubicadas en los lugares $k, k+1, \ldots, N$ en la lista "mezclada" y continúa con las personas que ocupan los lugares $1, 2, \ldots, k-1$.

   En este punto hay que hacer una observación importante: Dado que el número sorteado en segundo lugar indica la posición en la lista "mezclada" de la primera persona seleccionada, este número tiene que ser un número entre 0001 y 39.751 (que es la cantidad de aspirantes en nuestro ejemplo). En un sorteo de estas características, la Dirección Nacional de Loterías y Quinielas utiliza cinco bolilleros, donde en el primero hay bolillas del 0 al 3, y en los cuatro restantes hay bolillas numeradas del 0 al 9. Por lo tanto, podría darse el hecho de que el número sorteado fuera el 00000 (cinco ceros) o un número mayor a 39.751 (entre 39.752 y 39.999). En ese caso, se desecha ese segundo número sorteado y se procede a repetir el sorteo del segundo número hasta que se obtenga uno mayor o igual a 00001 y menor o igual a 39.751.

3. Semillas

Dos interrogantes respecto a la selección de las *semillas*.

1. ¿Qué se entiende por "*semilla*"?

2. ¿Por qué los números que serán utilizados como "*semillas*" son sorteado por la Dirección Nacional de Loterías y Quinielas?

Se entiende por *semilla* al valor inicial que se introduce en el programa de computación para que el algoritmo utilizado genere la serie de números aleatorios.

Como los números aleatorios que se utilizarán para hacer el ordenamiento inicial de los candidatos son generados por un programa de computación, su "aleatoriedad" depende justamente que el número que se la da para el arranque (la *semilla*) sea un número aleatorio. Es esa la razón por la cual el sorteo para esa semilla se hace ante Escribano Público en la Dirección Nacional de Loterías y Quinielas.

El algoritmo utilizado por el IESTA para generar número seudo-aleatorios es uno de los más sofisticados que existen en la actualidad, el algoritmo Mersenne

Twister. Detalles técnicos acerca del mismo escapan los objetivos de este documento, pero se pueden encontrar en el trabajo de Matsumoto y Nishimura (disponible en http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html).

4. Conclusiones

Que la Dirección Nacional de Loterías y Quinielas sortee el número a ser utilizado como semilla para inicializar un algoritmo computacional de generación de números aleatorios, que determinará cuales son las personas de una lista de aspirantes que serán seleccionadas no es garantía de que el sorteo sea válido desde el punto de vista de las condiciones requeridas en este artículo. A lo sumo generará la percepción, errónea, de validez.

La generación de un número "semilla" por medio de un proceso puramente aleatorio, como es el utilizado por la Dirección Nacional de Loterías y Quinielas, aunado a su utilización en un método que garantice tanto la equiprobabilidad de selección de los participantes como la posibilidad de que se puedan reproducir sus resultados por actores diferentes a los que inicialmente lo llevaron a cabo, es lo que realmente valida el proceso de selección.

El método utilizado por el IESTA, aquí presentado, cumple las tres condiciones fundamentales enunciadas al principio de este documento: transparencia, igualdad de oportunidades (equiprobabilidad) y simplicidad. Como ventaja adicional, el método no sólo selecciona la muestra de tamaño n, sino que proporciona una lista ordenada de forma aleatoria de los N candidatos presentados, que puede usarse para seleccionar los n requeridos y sus posibles suplentes.

5. Bibliografía

Matsumoto, M. y Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, **8**, 1, January pp.3-30.