TEMat

Los sorteos que utilizan las primeras letras de los apellidos como criterio de selección son injustos

Ramiro Martínez Pinilla
Graduado en Matemáticas (UPC)
Estudiante del Master in Advanced Mathematics and Mathematical Engineering (UPC)
ramiro.martinez@estudiant.upc.edu

Resumen: Este estudio pretende cuantificar las diferencias de probabilidad que se producen al seleccionar un grupo de personas mediante un sorteo en el que se obtiene una pareja de letras al azar y se asignan las plazas por orden alfabético de los apellidos. La distribución de las primeras letras de los apellidos no es uniforme en la población, por lo que este sistema no garantiza que todos los participantes tengan la misma probabilidad de ser seleccionados. Este es un hecho conocido, pero hasta el momento no se habían calculado estas probabilidades y estos sorteos se siguen utilizando.

Para calcular cada probabilidad de ser elegido realizaremos simulaciones de este tipo de sorteos. Analizaremos el margen de error de los resultados y comprobaremos que, efectivamente, se dan estas diferencias y las desigualdades que se producen son significativas.

Finalmente se propone una alternativa justa en la que todos los solicitantes tienen la misma probabilidad de ser seleccionados.

Abstract: This paper tries to quantify the inequalities that are produced when you select a group of people by means of a draw in which a pair of letters is obtained at random and the posts are assigned in alphabetical order. The distribution of the firsts letters of surnames is nonuniform, hence this scheme does not guarantee that all participants have the same probability of being selected. This is a well-known fact but, as far as we know, those probabilities have not been computed and this kind of draw is still being used.

To compute each probability of being selected we simulate many draws of this kind. We analyze the error of the results and we check that those differences arise and the inequalities are significant.

Finally, we propose an alternative fair scheme in which all participants have the same probability of being selected.

Palabras clave: probabilidad, estadística, simulación, intervalos de confianza, sorteos, apellidos.

MSC2010: 62P25.

Recibido: 12 de febrero de 2017.
Aceptado: 23 de octubre de 2017.

Agradecimientos: Agradezco a la ANEM por la creación de esta revista y especialmente a los revisores, que me ayudaron a mejorar notablemente el artículo y me animaron a ampliarlo a partir de su versión inicial.

Referencia: Martínez Pinilla, Ramiro. «Los sorteos que utilizan las primeras letras de los apellidos como criterio de selección son injustos». En: *TEMat*, 2 (2018), págs. 1-13. ISSN: 2530-9633. URL: https://temat.es/articulo/2018-p1/.

1. Introducción

En este estudio analizaremos un tipo de sorteo en el que, de una lista de N aspirantes, se pretende asignar una plaza a n de ellos, con $n \leq N$. En los sorteos que estudiamos se ordenan alfabéticamente sus apellidos y se obtiene al azar una pareja de letras. A partir de la posición en la lista marcada por estas dos letras se comienzan a adjudicar las plazas por orden alfabético, teniendo en cuenta que si se llega a la ZZ se continúa por la AA.

Estos sorteos tienen una serie de problemas. Aunque la elección de la pareja de letras sea uniformemente aleatoria, la distribución de las primeras letras de los apellidos en la población no es homogénea, y tampoco lo será en un grupo concreto de solicitantes. Esto hace que, dada una lista, si realizamos el sorteo, no todos tengan la misma probabilidad de obtener plaza, y el sorteo sea manifiestamente injusto.

Observación 1. Durante todo el artículo trataremos con diferentes espacios de probabilidad. Cuando se escogen las dos letras se hace de forma equiprobable, pues se eligen uniformemente al azar, pero esto no se traduce en que la distribución de las plazas sea también equiprobable, pues se trata de un espacio de probabilidad distinto. En el resultado final del sorteo, al escoger n de los N candidatos el término *equiprobable* podría interpretarse como que cada subconjunto de n elementos del conjunto N tenga la misma probabilidad de ser escogido. Sin embargo, no es necesaria una condición tan fuerte, pues es suficiente con que cada candidato individualmente tenga una probabilidad de n/N de ser escogido, pero no necesitamos que sean eventos independientes. Es por ello que nos fijaremos en esta última condición: todo candidato ha de tener la misma probabilidad de ser elegido al participar en un sorteo en el que el resto de candidatos se escogen uniformemente al azar del resto de la población. Esto es lo que denominaremos de forma abreviada como sorteo *justo*, pues es lo que realmente esperaríamos de este tipo de sorteos. ◂

La distribución de los apellidos en la muestra de solicitantes reflejará la distribución de los apellidos dentro de la población, no siendo tampoco homogénea. En un sorteo justo la distribución de las plazas debería reflejar la de los apellidos dentro de dicha población. Lo importante es que cualquier solicitante tenga la misma probabilidad de ser admitido, independientemente de las iniciales de su apellido.

Sin embargo, una persona cuyo apellido comience por una letra posterior a una muy frecuente tendrá una menor probabilidad de obtener plaza que otra cuya inicial siga a letras menos frecuentes. Simplificando el sorteo con una sola letra podemos ver fácilmente un ejemplo. Alguien cuyo apellido comience por A obtendrá plaza si en el sorteo se extrae la letra A. Pero, como las letras W, X, Y, Z son muy poco frecuentes en castellano, es improbable que haya algún otro solicitante cuyo apellido empiece por alguna de ellas. Por eso, si la escogida por el sorteo es W, X, Y o Z, la persona con apellido que comienza por A tiene también bastante probabilidad de obtener la plaza. Por otra parte alguien cuyo apellido comience por la letra N obtendrá plaza cuando sea esta letra la que salga del sorteo, pero tiene pocas posibilidades de obtenerla si el resultado es una letra anterior, puesto que la probabilidad de que alguno de los otros solicitantes tenga un apellido que comience por M es elevada. Este mismo razonamiento es igual de válido cuando se consideran dos letras en vez de una, pues lo único que se hace es aumentar el número de variables. Este es un hecho conocido y se ha publicado en blogs [4], artículos especializados [6] y en la prensa generalista [8] desde hace décadas.

Un caso extremo sucedería si a un sorteo de una única plaza se presentaran dos personas con apellidos diferentes pero que comenzaran por las mismas dos primeras letras. Aquella con el apellido anterior por orden alfabético siempre precederá a la otra en el resultado del sorteo, y esta segunda persona nunca podrá obtener plaza independientemente de las letras seleccionadas.

A pesar de ello, este sistema de selección en el que se utilizan las primeras letras de los apellidos para asignar plazas está muy extendido y se utiliza, entre otros, en la admisión del alumnado en los centros docentes de algunas comunidades autónomas [1, 3]. Es importante, por tanto, ir un paso más allá y comprobar si se trata de una mera curiosidad teórica o si en casos reales estas desigualdades son significativas y realmente se está discriminando a una parte de la población.

Para cuantificar las posibles diferencias no es suficiente con analizar una muestra concreta, como se ha hecho en el artículo de la revista de matemáticas *SUMA* [6] y en el artículo del diario *El País* [4]. Por ello, en este trabajo, para calcular la probabilidad de obtener plaza con un apellido concreto, programaremos

la simulación de un número suficientemente grande de sorteos utilizando datos de la población española. De esta forma, empleando herramientas estadísticas, conseguiremos cuantificar estas diferencias.

Es necesario hacer esto porque no podemos descartar *a priori* el caso de que, aunque en cada ejemplo concreto el sorteo sea injusto, globalmente estos desequilibrios puedan compensarse y la esperanza para cada inicial del apellido sea la misma. Podría ser que con una lista de solicitantes concreta un apellido que comience por XY se vea perjudicado, mientras que otra lista le beneficie, y la esperanza podría ser la de un resultado justo. En ocasiones, quienes plantean este tipo de sorteos interpretan que, como es aleatorio, unas veces perjudicará y otras veces beneficiará y se acabará compensando, confundiendo aleatorio con uniformemente aleatorio. Descartaremos esta posibilidad pues, fijados N y n, veremos que hay parejas de iniciales con una probabilidad de obtener plaza significativamente menor que otras.

También habría que tener en cuenta la posibilidad de que la dependencia respecto a los parámetros N y n fuera caótica, y pequeñas variaciones en ellos dieran lugar a distribuciones de probabilidad completamente distintas, en las que de nuevo unas parejas de iniciales fueran perjudicadas en unos casos y beneficiadas en otros. Calcular sistemáticamente los porcentajes de probabilidad concretos para cada caso nos permitirá comprobar que esto no es lo que sucede, y que las desigualdades generalmente persisten para diferentes valores de n y N.

Analizaremos en último lugar la dependencia geográfica de los resultados, pensando sobre todo en las comunidades con dos lenguas oficiales, repitiendo los experimentos tomando como población únicamente las comunidades autónomas de Castilla y León y del País Vasco para así identificar si existen grandes diferencias entre ellas. Veremos que, aunque en casos específicos los resultados son diferentes, en líneas generales obtendremos probabilidades similares.

Para finalizar, propondremos un método justo en el que todos los participantes tengan la misma probabilidad de ser escogidos independientemente de su apellido, teniendo en cuenta que sea sencillo de realizar y publicar, discutiendo las posibles dificultades técnicas a la hora de implementarlo.

2. Método

Consideraremos como población todos los habitantes de España a fecha de 1 de enero de 2016 [1]. Para la realización de este trabajo se han solicitado las primeras letras de los apellidos de la población al Instituto Nacional de Estadística [2] [2]. Ordenamos alfabéticamente estos apellidos para obtener una lista de 46 524 505 habitantes.

Fijado un número de plazas n y un número de solicitantes $N \geq n$, queremos calcular cuál es la probabilidad $p_{\mathtt{XY}}$ de que un solicitante cuyo apellido comience por las letras XY obtenga una plaza, siendo XY la pareja de letras que queramos estudiar en cada momento. Modelamos cada sorteo como una variable aleatoria de Bernoulli X que toma valor 1 si obtiene plaza y 0 en caso contrario. Idealmente, en un sorteo justo $p_{\mathtt{XY}} = n/N$ independientemente de XY.

Para obtener una muestra aleatoria simple de estas variables aleatorias programamos un *script* (algoritmo 1) en C++ que simule sorteos como los que estamos estudiando.

Queremos calcular la probabilidad de que, si una persona cuyas iniciales comienzan por XY se presenta al sorteo, obtenga plaza. Para ello, el *script* selecciona una persona control cuyo apellido comience por XY, simula todo el sorteo y anota si el candidato control ha obtenido o no plaza en el sorteo.

Utilizaremos la notación habitual en la que $variable \xleftarrow{\$} \{conjunto\}$ significa que se asigna a la variable *variable* un elemento de forma uniformemente aleatoria de entre los del conjunto $\{conjunto\}$. Cuando simplemente utilicemos el símbolo $\leftarrow$ se trata de una asignación determinista.

Hemos de tener en cuenta que, como los solicitantes siempre se ordenan alfabéticamente, no todos aquellos cuyo apellido comience por XY tienen la misma probabilidad de obtener la plaza. En el caso de

[1] En las simulaciones se han excluido los apellidos con caracteres especiales (*ñ, ç, '*) o de una sola letra, para facilitar el tratamiento informático.

[2] El INE es la fuente del dato primario, el grado de exactitud o fiabilidad de la información derivada por elaboración propia del autor es de la exclusiva responsabilidad de este.

Algoritmo 1 (Sorteo estudiando a un sujeto XY, con N solicitantes y n plazas).

subrutina SORTEO
▷ Elegimos a un candidato control cuyo apellido empiece por las letras que queremos estudiar
sujeto control $\xleftarrow{\$}$ {*población con apellidos que comienzan por* XY}
{*candidatos*} $\leftarrow$ *sujeto control*
▷ Elegimos al resto de candidatos uniformemente al azar entre el resto de la población
para $i \leftarrow 2, \ldots, N$ **hacer**
aux $\xleftarrow{\$}$ {*población*} $\setminus$ {*candidatos*}
{*candidatos*} $\leftarrow$ {*candidatos*} $\cup$ *aux*
fin para
▷ Escogemos uniformemente al azar una pareja de letras
letras $\xleftarrow{\$}$ {*parejas de letras*}
▷ Encontramos la posición que determinan en la lista y asignamos las plazas
encuentra la posición de la lista {*candidatos*} que señala *letras*
asigna plaza a los n primeros candidatos a partir de esta posición
▷ Comprobamos si el sujeto control ha recibido plaza
si *sujeto control* ha recibido plaza, **entonces**
$X \leftarrow 1$
en caso contrario
$X \leftarrow 0$
fin si
fin subrutina

que coincidan en el mismo sorteo dos personas con apellidos diferentes que comiencen por las mismas dos primeras letras, aquella cuyo apellido sea anterior por orden alfabético siempre obtendrá su plaza antes que la otra, independientemente del resultado del sorteo, como ya hemos mencionado previamente. Por lo tanto, hemos de precisar que llamamos p_{XY} al promedio de las probabilidades de obtener una plaza de cada una de las personas cuyo apellido comienza por XY. En consecuencia, elegiremos el solicitante control aleatoriamente entre todas las personas cuyo apellido comience por XY.

La pareja de letras que se obtenga del sorteo no determina directamente si una persona cuyo apellido comienza por XY obtendrá o no plaza. Esto dependerá también de los apellidos del resto de personas que hayan solicitado esa misma plaza. Por ello completamos la lista con otros $N-1$ solicitantes elegidos completamente al azar de la lista que contiene a toda la población. Podría haber más solicitantes cuyo apellido comience también con XY, pero nos aseguramos de que todos ellos sean distintos entre sí y del primero.

Por último, elegimos aleatoriamente un par de letras y anotamos si el individuo inicial ha obtenido una de las n plazas ($X = 1$) o no ($X = 0$). Repetimos este experimento m veces para tener una muestra aleatoria simple $X_1, \ldots, X_m$ de tamaño m.

Con esta muestra aleatoria simple, nuestro estimador de p_{XY} será $\widehat{p_{\mathrm{XY}}} = \frac{\sum_{i=1}^{m} X_i}{m}$. Si m es suficientemente grande, gracias al teorema central del límite (teorema 1) podemos considerar que

$$\widehat{p_{\mathrm{XY}}} \sim N\left(p_{\mathrm{XY}}, \sqrt{\frac{p_{\mathrm{XY}}(1-p_{\mathrm{XY}})}{m}}\right).$$

Teorema 1 (teorema central del límite [7]). *Si las variables independientes idénticamente distribuidas X_i tienen todas la misma distribución y media μ y desviación típica $\sigma \neq 0$ finitas, entonces la variable*

$$Y_m = \frac{X_1 + \cdots + X_m - m\mu}{\sigma\sqrt{m}}$$

es asintóticamente normal $N(0,1)$, es decir, la función de distribución $F_m(z)$ de Y_m verifica para todo z la relación

$$\lim_{m\to\infty} F_m(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x.$$

En nuestro caso, la media es p_{XY} y la desviación típica de las variables de Bernoulli es $\sqrt{p_{XY}(1-p_{XY})}$. Sustituyendo en la fórmula obtenemos

$$N(0,1) \sim \frac{X_1 + \cdots + X_m - m p_{XY}}{\sqrt{p_{XY}(1-p_{XY})}\sqrt{m}} = \frac{\frac{\sum_{i=1}^m X_i}{m} - p_{XY}}{\sqrt{\frac{p_{XY}(1-p_{XY})}{m}}}$$

$$N\left(p_{XY}, \sqrt{\frac{p_{XY}(1-p_{XY})}{m}}\right) \sim \frac{\sum_{i=1}^m X_i}{m} = \widehat{p_{XY}}.$$

Queremos que el estimador $\widehat{p_{XY}}$ nos permita asegurar, con un alto nivel de confianza, que el valor real se encontrará en un cierto entorno del valor del estimador. Para ello tendremos que tomar un valor suficientemente grande de m.

Elegimos como entorno $\left[\widehat{p_{XY}} - 0{,}005, \widehat{p_{XY}} + 0{,}005\right]$ y como nivel de confianza, un 99 %. Es decir, queremos asegurar que con una probabilidad de un 99 % el valor real de la probabilidad p_{XY} satisface que

$$p_{XY} \in \left[\widehat{p_{XY}} - 0{,}005, \widehat{p_{XY}} + 0{,}005\right]. \tag{1}$$

Modificamos la expresión (1) para normalizar el estimador y obtenemos

$$\frac{-0{,}005}{\sqrt{\frac{p_{XY}(1-p_{XY})}{m}}} \leq \frac{\widehat{p_{XY}} - p_{XY}}{\sqrt{\frac{p_{XY}(1-p_{XY})}{m}}} \leq \frac{0{,}005}{\sqrt{\frac{p_{XY}(1-p_{XY})}{m}}}. \tag{2}$$

Ahora $\frac{\widehat{p_{XY}} - p_{XY}}{\sqrt{\frac{p_{XY}(1-p_{XY})}{m}}} \sim N(0,1)$, y queremos que el intervalo definido por (2) abarque al menos un 99 % de probabilidad. Consultamos los cuantiles de la normal y obtenemos que ha de contener al intervalo $[-2{,}575829, 2{,}575829]$.

Para una m concreta, la longitud del intervalo (2) dependerá del valor de p_{XY} y será mínima con $p_{XY} = 0{,}5$. Sustituimos p_{XY} por 0,5 para calcular m, pues así garantizamos que en cualquier otro caso con esa m tenemos una confianza superior al 99 %. De esta manera tenemos que para $m \geq 66\,349$ obtendremos una estimación de p_{XY} con la que, efectivamente, podremos asegurar que el valor real se encontrará en el intervalo (1), con un 99 % de confianza.

Este proceso se ha de realizar para cada pareja de letras XY y para todos los pares (n, N) que queramos estudiar. Escogeremos diferentes valores de N que sean representativos de las posibles situaciones reales en las que se utilizan este tipo de sorteos.

3. Resultados

Para obtener las figuras 1 a 3 se ha calculado, para cada letra, la media ponderada de las probabilidades obtenidas para las parejas que comienzan por esa letra. Se ha tenido en cuenta el peso específico en la población de cada pareja de letras entre las que comienzan por la misma inicial.

La figura 1 corresponde al caso $n = 10$ y $N = 20$. En un sorteo justo esperaríamos que cada solicitante tuviera una probabilidad de un 50 % de obtener plaza. Observamos, sin embargo, que existen diferencias significativas entre las probabilidades de las distintas iniciales, que llegan a superar 15 puntos porcentuales.

Si analizamos el caso $n = 20$ y $N = 40$ en la figura 2, la probabilidad deseada es de nuevo un 50 %, pero el resultado experimental sigue sin ser el de un sorteo justo. Observamos que la distribución es muy similar a la del caso anterior, aunque no coincide exactamente.

Completamos este primer análisis con el caso $n = 10$ y $N = 40$ (figura 3). Podemos apreciar que las diferencias relativas entre las iniciales beneficiadas y las perjudicadas se han incrementado respecto a las figuras 1 y 2, en las que había una misma proporción *plazas*/*solicitantes*, que no se mantiene en esta figura. En este caso hay letras que tienen el doble de probabilidad que otras.

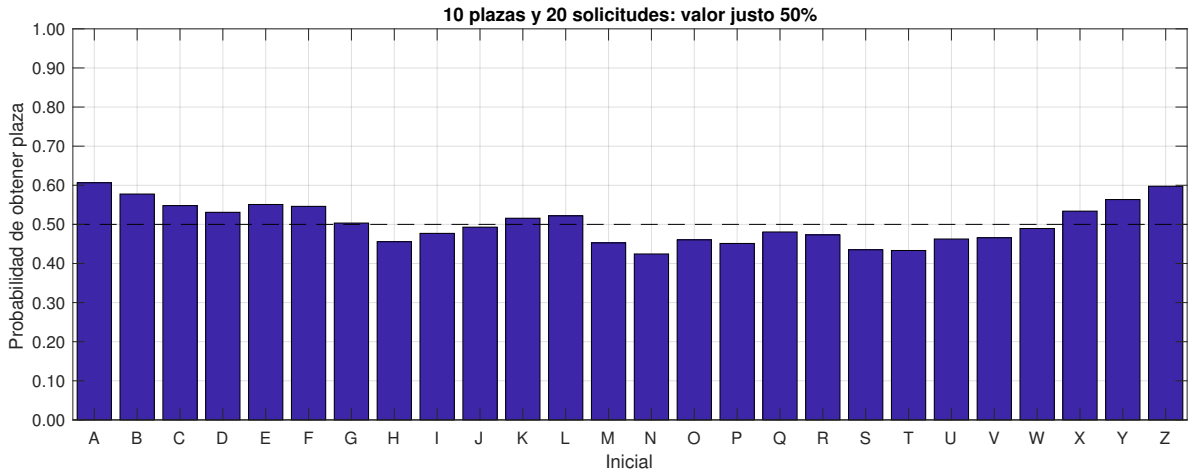

Figura 1: Probabilidad para cada inicial ($n = 10, N = 20$).

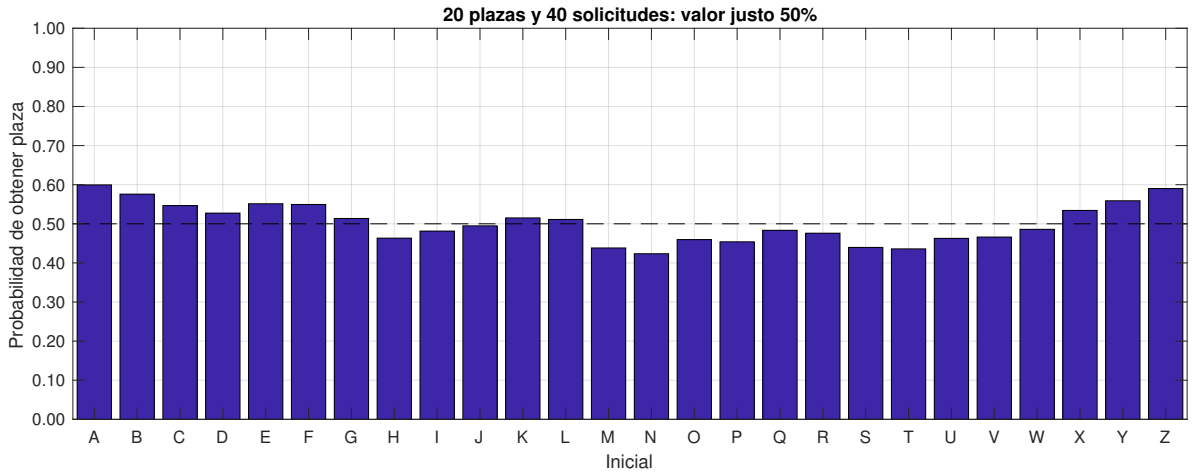

Figura 2: Probabilidad para cada inicial ($n = 20, N = 40$).

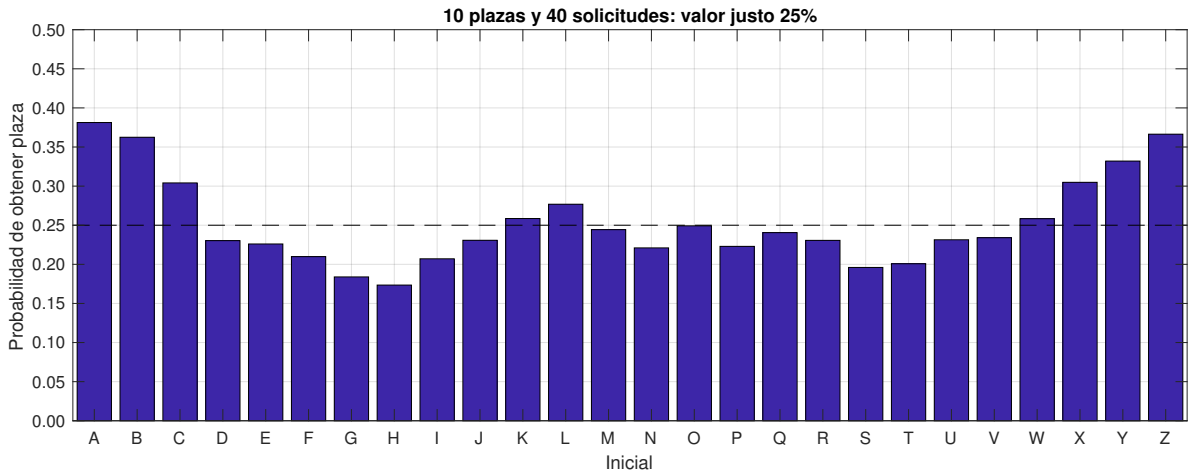

Figura 3: Probabilidad para cada inicial ($n = 10, N = 40$).

Las figuras 1 a 3 muestran que este sistema, basado en las iniciales de los apellidos, da lugar a unas probabilidades no uniformes, con las que no todos los apellidos tienen las misma probabilidad de obtener una plaza.

En las figuras 4 y 5 representamos la probabilidad de obtener plaza para los apellidos que comienzan por AL, CI, HE, KE y MU. Hemos seleccionado estas cinco parejas de iniciales porque representan los diferentes comportamientos que se pueden dar.

Estudiamos distintos casos dejando el número de plazas fijo y observamos el comportamiento de la probabilidad, a medida que incrementamos el número de solicitantes.

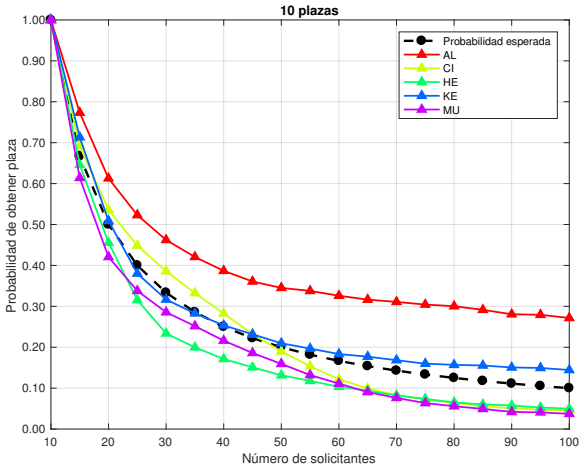

Figura 4: Diez plazas

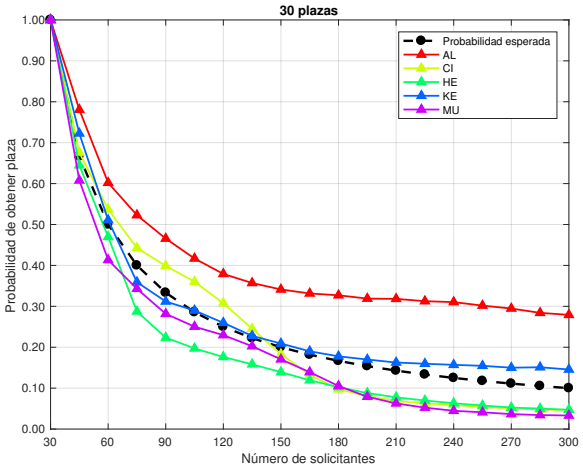

Figura 5: Treinta plazas

En el caso $n = 10$ (figura 4), aquellos cuyo apellido comience por AL se verán claramente beneficiados, con cualquier número de solicitantes. Aquellos cuyo apellido comience por HE o MU siempre se verán perjudicados por este tipo de sorteo. En algunos casos, como los apellidos que comienzan por KE, la probabilidad real se aproxima bastante bien a la teórica para los primeros valores del número de solicitantes N, mientras que, a medida que aumentan los solicitantes, resultan beneficiados con una probabilidad superior a la deseada. Por último, es también interesante el ejemplo de los apellidos que comienzan por CI, pues se ven beneficiados con un número bajo de solicitantes mientras que a partir de 50 solicitantes su probabilidad de obtener una plaza cae por debajo de lo esperado en un sorteo justo.

En la figura 5 hemos representado el caso $n = 30$. Observamos que el comportamiento es cualitativamente muy similar al de la figura anterior, aunque con mayores fluctuaciones, por ejemplo, en los apellidos que comienzan por CI.

Los cuadros 1 y 2 reflejan la probabilidad de obtener una plaza de aquellos cuyo apellido comience por las parejas de iniciales más beneficiadas y perjudicadas en cada caso, en el primer cuadro cuando se adjudican 10 plazas y en la segunda cuando se adjudican 20.

Cuadro 1: Parejas de iniciales con menor y mayor probabilidad de obtener una plaza ($n = 10$).

N	15	20	25	30	35	40	45	50	55	60	65	70
prob. teórica	67 %	50 %	40 %	33 %	29 %	25 %	22 %	20 %	18 %	17 %	15 %	14 %
iniciales mín.	SE	MV	HG	HG	GV	GV	GW	HG	DK	DJ	DJ	MV
probabilidad	56 %	41 %	31 %	23 %	19 %	17 %	15 %	13 %	11 %	9 %	8 %	7 %
iniciales máx.	AJ	AE	AI	AI	AK	AJ	AI	AK	AK	AK	AK	AK
probabilidad	78 %	62 %	53 %	47 %	43 %	40 %	37 %	35 %	34 %	33 %	32 %	32 %

Finalmente, en la figura 6 dejamos fijo el número de plazas ($n = 10$) y variamos de nuevo el número de solicitantes. Para cada número de solicitantes y cada valor de la probabilidad, el color blanco significa que

Cuadro 2: Parejas de iniciales con menor y mayor probabilidad de obtener una plaza ($n = 20$).

N	30	40	50	60	70	80	90	100	110	120	130	140
prob. teórica	67 %	50 %	40 %	33 %	29 %	25 %	22 %	20 %	18 %	17 %	15 %	14 %
iniciales mín.	SF	MV	HF	GV	GV	GV	DJ	DJ	DJ	DJ	NB	MV
probabilidad	56 %	41 %	29 %	22 %	19 %	17 %	15 %	12 %	10 %	9 %	8 %	6 %
iniciales máx.	AK	AK	AK	AA	AI	ZZ	AJ	AJ	AK	AK	AK	AK
probabilidad	79 %	62 %	53 %	48 %	43 %	39 %	37 %	35 %	34 %	33 %	33 %	32 %

en un sorteo con esas condiciones nadie tiene esa probabilidad de obtener plaza. El color negro significa que toda la población tiene esa probabilidad (es lo que sucede con 10 solicitantes: toda la población tiene probabilidad 1). Finalmente, los colores intermedios representan, en escala logarítmica, la fracción de la población que tiene esa probabilidad de obtener una plaza. En este caso se ha representado con línea discontinua verde la probabilidad deseada.

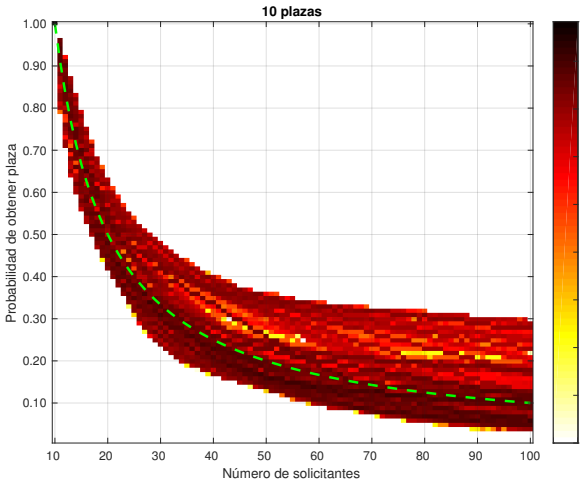

Figura 6: Diez plazas

3.1. Análisis de los resultados

Únicamente hemos mostrado una serie de ejemplos en las figuras y los cuadros, pero estos son suficientes para extraer las conclusiones que buscábamos.

Podemos concluir, en primer lugar, que, fijados el número de plazas y el de solicitantes, el sorteo no es justo, pues la probabilidad de obtener una plaza no es la misma para todos los solicitantes, sino que depende de su apellido. Además, como se puede ver en los cuadros 1 y 2, así como en las figuras 4 y 5, estas desigualdades no son anecdóticas, sino que las diferencias entre los beneficiados y los perjudicados son muy significativas.

En el caso $n = 10$ vemos que, cuando el número de solicitantes es 15, ya hay apellidos con 22 puntos porcentuales más de probabilidad de obtener una plaza que otros y, si el número de solicitantes es 30, hay personas cuya probabilidad de obtener plaza es el doble que la de otras.

Lo mismo sucede si el número de plazas ofertadas es 20, pues con 30 solicitantes una persona cuyo apellido comience por AK tiene 23 puntos porcentuales más de probabilidad de obtener una plaza que otra cuyo apellido comience por SF. Si el número de solicitantes es superior a 60, vuelve a haber apellidos con más del doble de probabilidad de obtener una plaza que otros. La existencia de estos casos tan desiguales es suficiente para replantearse el uso de este tipo de sorteos.

Es interesante comprobar que estas desigualdades no son un caso particular, sino que se mantienen para diferentes valores de n y N. Es lo que podemos observar en el comportamiento suave de las figuras 4 y 5 y en las similitudes entre ellas. Notamos también que, cuando la proporción entre plazas y solicitudes es la misma, los resultados obtenidos para cada pareja de letras son muy similares, con algunas fluctuaciones. Esta apreciación se pone de manifiesto tanto en las figuras 1 y 2 como en las figuras 4 y 5.

Aunque por cuestiones de espacio no hemos podido incluir el resto de iniciales ni más valores para n y N, hemos estudiado otros casos y en todos ellos se observan comportamientos similares a los descritos. Hemos estudiado el comportamiento global de la población en la figura 6 y lo que se observa concuerda con lo descrito anteriormente. En un sorteo justo toda la población debería tener la probabilidad deseada y, sin embargo, vemos cómo hay una franja de probabilidades por encima y por debajo que no es anecdótica. Esto significa que hay personas que (por su apellido) siempre se ven perjudicadas en este tipo de sorteos, independientemente del número de plazas y de solicitantes.

3.2. Dependencia geográfica de los resultados

Hasta ahora hemos analizado la dependencia respecto al número de plazas y de solicitantes, pero para conocer la probabilidad real en un caso concreto también habría que tener en cuenta la población que se considera a la hora de obtener los datos de las frecuencias de los diferentes apellidos.

Unos apellidos son más comunes en unas regiones que en otras, y tiene interés por sí mismo el analizar cómo repercute esto en este tipo de sorteos. Como ejemplo, repetiremos el mismo experimento tomando como poblaciones los habitantes de Castilla y León y el País Vasco respectivamente.

Figura 7: Probabilidad para cada inicial ($n = 10$, $N = 20$, Castilla y León).

En las figuras 7 y 8 representamos la probabilidad de obtener una plaza en un sorteo con 10 plazas y 20 solicitantes con un apellido que comience por cada una de las letras del abecedario, obtenida tras calcular la probabilidad para cada pareja de letras y luego ponderar su peso, tal como hemos hecho anteriormente en la figura 1.

Observamos que existen diferencias entre ambas gráficas y también respecto al caso estatal, con iniciales que en un caso se ven perjudicadas y en otro beneficiadas y viceversa. Comprobamos también lo que venimos recalcando en todo este estudio, y es que existen diferencias significativas entre unas iniciales y otras. Observamos que las diferencias entre comunidades son menores que las diferencias que existen entre unas letras y otras, y que en general se mantiene la forma cualitativa de las gráficas.

Recalcamos que, aunque únicamente mostremos el caso $n = 10, N = 20$, las mismas conclusiones se deducen del resto de casos, salvo alguna excepción anecdótica. Por ejemplo, al disminuir la proporción entre plazas y solicitudes tanto a nivel estatal como en Castilla y León, aquellos cuyo apellido comienza por B van perdiendo poco a poco su ventaja respecto a otros apellidos, pero incluso con una proporción

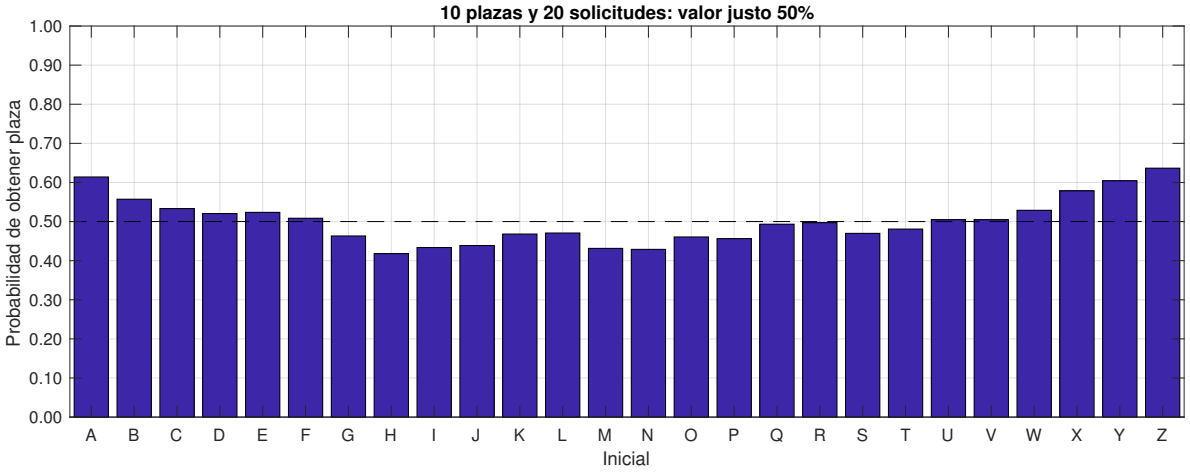

Figura 8: Probabilidad para cada inicial ($n = 10$, $N = 20$, País Vasco).

de solicitudes respecto a las plazas de 10 a 1 todavía siguen siendo beneficiados. Esta caída es mucho más rápida en el País Vasco, donde con una proporción de 6 a 1 los apellidos que comienzan por B ya pasan a verse perjudicados.

Este tipo de diferencias pueden observarse mejor en las figuras 9 y 10, en las que analizamos la evolución de parejas de letras concretas en un sorteo con 10 plazas cuando incrementamos el número de solicitudes.

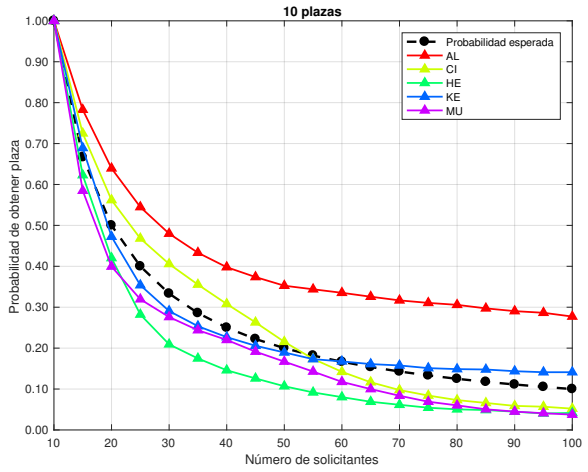

Figura 9: Diez plazas (Castilla y León).

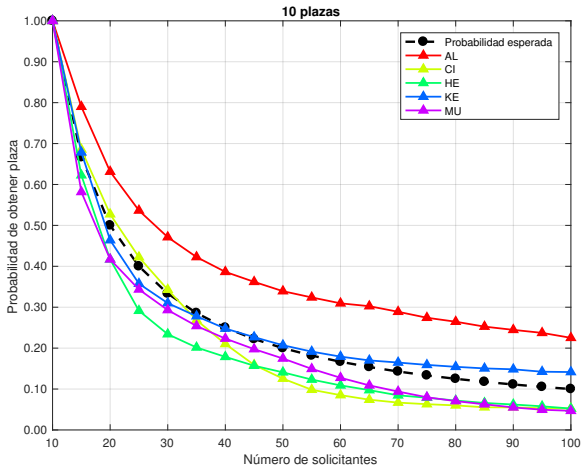

Figura 10: Diez plazas (País Vasco).

Las cinco parejas de letras escogidas a nivel estatal representan bastante bien lo que sucede al comparar diferentes regiones. Vemos que, a grandes rasgos, los resultados son similares, pero existen algunas particularidades específicas. Los apellidos que comienzan por CI se ven ampliamente beneficiados en Castilla y León cuando hay pocos solicitantes; no obstante, esta ventaja es menor y casi inexistente en el País Vasco, donde a su vez la desventaja de los apellidos que comienzan por HE es menos pronunciada. Esto es lo que veríamos comparando el resto de parejas de letras, la mayoría siguen la misma tendencia que a nivel estatal pero con algunas diferencias locales.

Análogamente a la figura 6, representamos la distribución de la población en las diferentes probabilidades en las figuras figuras 11 y 12. Vemos que el resultado es similar en ambas, con diferencias ligeramente más pequeñas en el País Vasco.

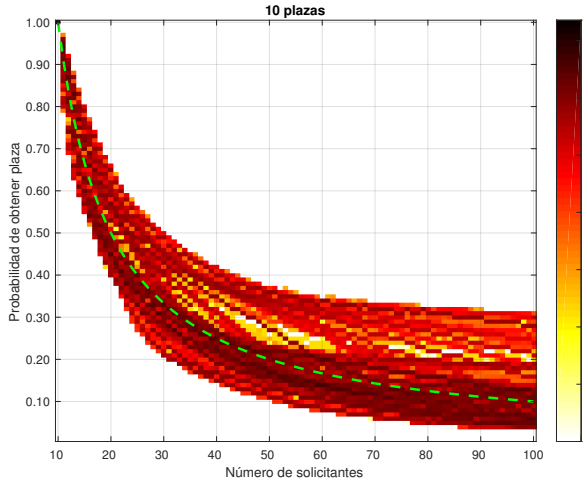

Figura 11: Diez plazas (Castilla y León).

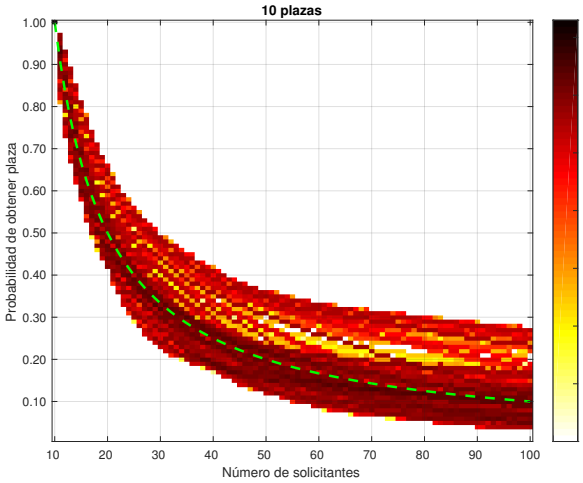

Figura 12: Diez plazas (País Vasco).

El método de sorteo es injusto independientemente de la región, y no se esperan mayores diferencias en otras comunidades autónomas a las ya estudiadas.

4. Propuesta alternativa de sorteo

Nuestro objetivo es encontrar una forma de seleccionar n candidatos de un conjunto de N con la que cada uno de ellos individualmente tenga una probabilidad de n/N de ser escogido.

Para ello lo más natural sería ordenar a los candidatos (por ejemplo, por orden alfabético) y escoger uniformemente al azar un entero $m \xleftarrow{\$} [1, N]$. Entonces asignaríamos plazas a los candidatos $m, m+1, \ldots, m+(n-1) \mod N$. Es inmediato comprobar que para cualquier candidato la probabilidad de obtener plaza es exactamente n/N.

Esto nos lleva a preguntarnos por qué no se utiliza directamente este método en lugar del actual. Una de las razones esgrimidas por la administración es que el sorteo ha de ser justo pero también fácil de publicar y de ser reutilizado en múltiples casos. Según su argumento, una pareja de letras `XY` publicada en el boletín oficial de la comunidad autónoma les sirve para asignar plazas en diferentes centros, mientras que con el método natural sería necesario realizar un sorteo específico para cada centro, dada la dependencia con N, que será diferente en cada caso.

Se trata de un argumento falaz, porque es posible diseñar un sistema que sea a la vez sencillo de auditar y que otorgue la misma probabilidad de obtener plaza a todos los solicitantes. Si pensamos desde el punto de vista de una implementación informática, requeriríamos un generador de números pseudoaleatorios que tome como entrada una N y produzca como salida un entero módulo N de forma uniformemente aleatoria. Lo único que sería necesario publicar en el boletín oficial de la comunidad autónoma sería la descripción del generador utilizado y una semilla (el estado interno del generador, que una vez especificado lo transforma en una función determinista).

Habitualmente, un generador de números pseudoaleatorios produce como salida un entero en un intervalo $[1, RAND\_MAX]$ (dependiendo de la implementación, el intervalo puede comenzar en el 0, pero se trata de casos equivalentes). Si escogemos uniformemente un entero de ese intervalo y nos quedamos con él, módulo N, esto no garantiza que la distribución sea equiprobable. Si $RAND\_MAX$ no es un múltiplo de N lo podemos escribir como $RAND\_MAX = cN + r$, con $c = \lfloor RAND\_MAX/N \rfloor$ y $r = RAND\_MAX - N\lfloor RAND\_MAX/N \rfloor$. Los enteros $m \in [1, r]$ tendrán una probabilidad $(c+1)/(RAND\_MAX)$, mientras que los enteros $m \in [r+1, N]$ tendrán únicamente una probabilidad $c/(RAND\_MAX)$.

Si $RAND\_MAX$ es suficientemente grande podemos hacer este error tan pequeño como queramos. Si no es suficiente tolerar pequeños errores sino que queremos que exactamente todo el mundo tenga la misma probabilidad, entonces la estrategia más común sería forzar que $RAND\_MAX$ fuera un múltiplo de N.

Habitualmente esto se consigue ejecutando el generador de números pseudoaletorios y aceptando la salida únicamente si pertenece a un intervalo que sí sea múltiplo de N, por ejemplo $[1, cN]$. Esto nos lleva a descartar las salidas $m \in [cN + 1, RAND\_MAX]$. Si rechazamos la salida, volvemos a ejecutar el generador hasta obtener un número en $[1, cN]$, lo que garantiza que al hacer módulo N todos los resultados tengan la misma probabilidad. Se trata de un compromiso entre el error que estamos dispuestos a asumir y el número de ejecuciones del generador que podamos utilizar (la probabilidad de que en la i-ésima ejecución no hayamos obtenido una salida válida decae exponencialmente en i).

Nos encontramos ahora con que estos descartes que nos vemos obligados a realizar para obtener una salida realmente equiprobable dependen nuevamente de N. Una primera aproximación sería forzar que el *RAND_MAX* efectivo después de hacer los rechazos sea ahora un múltiplo del mínimo común múltiplo entre $1, 2, \ldots, N\_MAX$, donde N_MAX sea una cota al número máximo de solicitantes que esperamos. De esta forma, al hacer módulo por cualquier natural menor que N_MAX tendríamos una distribución equiprobable. Esto no es implementable, puesto que la sucesión de mcm $(1, 2, 3, ..., n)$ crece demasiado rápido como para que sea práctico utilizarlos, tal como se puede comprobar en *La Enciclopedia On-Line de las Secuencias de Números Enteros* (OEIS, por sus siglas en inglés) [5].

La solución más sencilla sería, entonces, publicar únicamente la semilla que se va a utilizar con el generador de números pseudoaleatorios y, en cada caso, realizar el número de ejecuciones necesarias, tal como se ha explicado, para obtener un entero uniformemente distribuido en $[1, N]$. Como la semilla se ha fijado desde un primer momento, el número de rechazos necesarios es determinista y sigue siendo posible auditar el sorteo.

Esta es una posible implementación de un sorteo que cumpla las condiciones explicitadas en la observación 1, pero podría implementarse de otra forma, siempre que respetara esas condiciones y no acabase discriminando a parte de la población.

5. Conclusiones

Al plantear y mantener este sorteo como criterio de desempate, da la impresión de que los legisladores supusieron que el hecho de que el sorteo sea aleatorio implica que todos los participantes han de tener la misma probabilidad de obtener una plaza. Sin embargo, este estudio demuestra que este no es el caso. Solo con tener unas nociones básicas sobre probabilidad podemos aplicarlas para detectar una situación injusta, en la que, sin pretenderlo, se está discriminando a una parte de la población.

Ante situaciones como esta se pone de manifiesto que, aunque el conjunto de la población no va a tener que utilizar habitualmente contenidos matemáticos de un nivel avanzado, sí que es imprescindible haber adquirido la competencia matemática necesaria para entender problemas como el que aquí se plantea. Por ejemplo, para la mayoría de la población no es necesario conocer los pormenores del teorema central del límite, pero sí tener suficientemente claro el concepto de probabilidad, para no llegar a la conclusión errónea de que el sorteo es justo a partir del hecho de que todas las parejas de letras tienen la misma probabilidad.

Una de las razones por las que posiblemente este tipo de sorteos se mantienen, a pesar de ser injustos, podría ser la dificultad para calcular analíticamente cada una de las probabilidades, debido a la gran cantidad de variables a considerar. Como el resultado depende de la población concreta, del número de plazas y del número de solicitudes, no hay una fórmula sencilla que exprese las probabilidades para poder evidenciar ante las administraciones que el procedimiento es manifiestamente injusto. Sin embargo, aunque no sea fácil obtener una solución analítica al problema, cualquier ordenador personal tiene capacidad para realizar todos los cálculos necesarios, y la estadística nos proporciona las herramientas para asegurar qué nivel de fiabilidad tienen los resultados obtenidos. De esta forma, combinando matemáticas, estadística y la potencia de cálculo de un ordenador, podemos no solo afirmar que este tipo de sorteos discriminan a una parte de la población, sino también cuantificar las diferencias que se producen.

Referencias

[1] Gobierno del Principado de Asturias. «RESOLUCIÓN de 19 de febrero de 2014, de la Consejería de Educación, Cultura y Deporte, por la que se aprueba el procedimiento de admisión del alumnado en centros docentes no universitarios públicos y privados concertados del Principado de Asturias». En: *Boletín Oficial del Principado de Asturias* 46 (25 de feb. de 2014), págs. 1-19. issn: 1579-4180. url: https://sede.asturias.es/bopa/2014/02/25/2014-03363.pdf.

[2] Instituto Nacional de Estadística. *Estadística del Padrón Continuo a fecha 01/01/2016*. 2016.

[3] Junta de Castilla y León. «RESOLUCIÓN de 15 de enero de 2016, de la Dirección General de Política Educativa Escolar, por la que se concreta la gestión del proceso de admisión del alumnado en los centros docentes de Castilla y León para cursar en 2016-2017 enseñanzas sostenidas con fondos públicos de segundo ciclo de Educación Infantil, Educación Primaria, Educación Secundaria Obligatoria o Bachillerato». En: *Boletín Oficial de Castilla y León* 16 (26 de ene. de 2016), págs. 4962-4981. issn: 1989-8959. url: http://bocyl.jcyl.es/html/2016/01/26/html/BOCYL-D-26012016-13.do.

[4] Murcia, Joseángel. «¿Por qué el sorteo por letra es el más injusto?» En: *Verne* (2016). url: http://verne.elpais.com/verne/2015/12/21/articulo/1450708343_196121.html.

[5] OEIS Foundation Inc. *A003418. Least common multiple (or LCM) of {1, 2, ..., n} for n >= 1, a(0) = 1*. En: *The On-Line Encyclopedia of Integer Sequences*. url: http://oeis.org/A003418.

[6] Pérez Porcel, José Antonio. «¿Son justos los sorteos de tribunales basados en las letras de los apellidos?» En: *SUMA, revista para la enseñanza y el aprendizaje de las matemáticas* 50 (nov. de 2005), págs. 65-68. url: http://revistasuma.es/revistas/50-noviembre-2005/son-justos-los-sorteos-de.html.

[7] Ríos, Sixto. *Métodos Estadísticos*. Ediciones del Castillo, 1977. isbn: 978-84-219-0154-0.

[8] Simón, Federico y González Olaya, Vicente. «El teatro de la Zarzuela reconoce que el sorteo para adjudicar los abonos de la ópera es injusto». En: *El País* (1 de dic. de 1993). url: http://elpais.com/diario/1993/12/01/madrid/754748666_850215.html.