# ¡Qué difícil es la aleatoriedad! (extracciones y azar)*

**F. Montes† & A. Corberán†. Universitat de València**

## 1 Introducción

El azar forma parte de nuestra vida cotidiana y se nos manifiesta de forma espontánea a través de múltiples circunstancias y fenómenos; pero en otras muchas ocasiones somos nosotros quienes necesitamos convocarlo. La obtención de una secuencia aleatoria de números es, sin duda, la más corriente de estas necesidades.

Estas notas pretenden ocuparse de la dificultad de generar con éxito una secuencia aleatoria que merezca tal nombre y lo hacen mediante el análisis de una lotería muy popular, la Primitiva, y sus familiares cercanos, el Bono Loto y el Gordo, y un sorteo relacionado con un llamamiento al servicio militar en los EUA. El orden de presentación de los ejemplos no es casual, están ordenados de *mejor* a *peor* atendiendo al éxito en alcanzar el objetivo de aleatoriedad perseguido. En el primero de ellos, la lotería Primitiva y sus variantes, las combinaciones ganadoras cumplen con los requisitos de aleatoriedad exigibles a un juego social de azar tan popular (¿podría haber sido de otra manera?). En el segundo, el sorteo del servicio militar del año 70 en los EUA, las cosas son peores porque a pesar de las precauciones que se tomaron el resultado no pasó con éxito ninguno de los tests de aleatoriedad a que fue sometido. Un ejemplo más cercano, en el que un planteamiento erróneo inicial condujo a un mal resultado, fue el sorteo para determinar los excedentes de cupo del reemplazo del 98 en España del que ya nos ocupamos en [2].

## 2 La Primitiva

La Primitiva y el resto de loterías conocidas genéricamente como 6/49, consisten en la extracción al azar, sin reemplazamiento, de 6 números entre los 49 primeros números naturales. Estos 6 números, cuyo orden de extracción es irrelevante, constituyen lo que se denomina *combinación ganadora* y su acierto es el objetivo (*soñado*) de los apostantes, que efectúan sus apuestas mediante la elección de 6 números en el correspondiente boleto.

La dificultad de acertar la combinación ganadora, premio de *1ª categoría*, se palia otorgando premios de menor categoría basados en el acierto parcial de dicha combinación. El premio de *2ª categoría* se otorga cuando se aciertan 5 de los 6 números de la combinación ganadora más el llamado *complementario*, número que se ha extraído al azar de entre los 43 que no forman parte de dicha combinación. La *3ª categoría* de premios consiste en acertar 5 de los 6 números ganadores y en las *categorías 4ª y 5ª* han de acertarse, respectivamente, 4 y 3 de los 6 números

*Este texto recoge la conferencia impartida en el curso *Un cambio histórico: el lenguaje de las matemáticas en sus aplicaciones*, que tuvo lugar del 11 al 15 de septiembre de 2000 en la UIMP de Santander.

†Dirección: *Departament d'Estadística i Investigació Operativa. Universitat de València. E-46100 Burjassot. Spain.* e-mail: `francisco.montes@uv.es, angel.corberan@uv.es.`

de la combinación ganadora. Veamos hasta qué punto somos benevolentes calificando sólo de difícil la obtención de un premio de 1ª categoría.

## 2.1 El espacio muestral de los posibles resultados

El experimento consistente en extraer los 6 números en las condiciones antes descritas es un experimento aleatorio que da lugar a un espacio muestral finito e equiprobable, $\Omega$, con $\binom{49}{6} = 13.983.816$ resultados. La obtención de la probabilidad de ocurrencia de cualquier suceso $A$ se lleva a cabo aplicando la fórmula de Laplace,

$$P(A) = \frac{\text{casos favorables a la ocurrencia de } A}{\text{casos posibles}}.$$

La tabla recoge la probabilidad de acierto para cada categoría.

| **categoría** | **aciertos** | **favorables** | **probabilidad** |
|---|---|---|---|
| primera | 6 | $\binom{6}{6} = 1$ | $7,15 \times 10^{-8}$ |
| segunda | 5+C | $\binom{6}{5} \times \binom{1}{1} = 6$ | $4,29 \times 10^{-7}$ |
| tercera | 5 | $\binom{6}{5} \times \binom{42}{1} = 252$ | $1,80 \times 10^{-5}$ |
| cuarta | 4 | $\binom{6}{4} \times \binom{43}{2} = 13.545$ | $9,69 \times 10^{-4}$ |
| quinta | 3 | $\binom{6}{3} \times \binom{43}{3} = 246.820$ | $1,76 \times 10^{-2}$ |
| casos posibles | | $\binom{49}{6} = 13.983.816$ | |

## 2.2 Aleatoriedad de las extracciones

Probabilidades tan pequeñas no impiden que estas loterías gocen de gran popularidad. Dos parecen ser las razones de esta gran aceptación: la primera, sin duda, los suculentos premios que reparte, especialmente en aquellas ocasiones en las que hay *bote*, acumulación a los premios de 1ª categoría de premios de la misma categoría que no fueron otorgados en sus correspondientes sorteos por no haber habido ningún acertante de la misma. Al fin y al cabo, como señalan algunas de las cuñas publicitarias del ONLAE (**O**rganismo **N**acional de **L**oterías y **A**puestas del **E**stado), soñar cuesta apenas veinte duros (¿o son ya treinta con la inflación?). Pero siendo importante, esta razón no sería suficiente de no estar acompañada por una segunda: la confianza de los apostantes en que las extracciones se llevan a cabo verdaderamente al azar. ¿Está justificada esta confianza?

El estudio que presentamos a continuación responde a la pregunta anterior analizando las combinaciones ganadoras desde distintos puntos de vista. El estudio se basa en los 4.024 sorteos de las tres loterías celebrados desde el 17 de octubre de 1985, fecha en la que tuvo lugar el primer sorteo de la Primitiva, hasta el 9 de julio de 2000 (los sorteos celebrados con posterioridad no han de cambiar el significado de las conclusiones). La base de datos utilizada para ello proviene parcialmente de la información que el ONLAE pone a disposición pública en su web `http://onlae.terra.es`, pero en su mayor parte proviene de [9], de donde se han extraído también algunos resultados.

Los puntos de vista utilizados para analizar los resultados de los 4.024 sorteos se corresponden con diferentes variables aleatorias y familias de sucesos asociadas al espacio muestral. A saber,

**Número extraído.-** Se trata de una variable aleatoria cuyo valor coincide con el número extraído.

**Figura.-** Un criterio de clasificación de la combinación ganadora es el que atiende a la cantidad de números consecutivos que contiene. Cada figura[1] define un suceso aleatorio y el conjunto de todos ellos forman una partición del espacio muestral que estudiaremos.

**Suma.-** Es una variable aleatoria que para cada combinación ganadora toma el valor de la suma de los números que la constituyen.

Si el mecanismo de extracción funciona aleatoriamente, el comportamiento probabilístico teórico de las variables y sucesos anteriores será conocido y podremos compararlo con su comportamiento experimental observado a partir de la muestra de 4.024 sorteos. La medida de la discrepancia entre ambos nos informará acerca de la bondad del mecanismo. Antes de comenzar con los análisis específicos debemos señalar que las posibilidades de estudio y análisis que las loterías del tipo 6/49 ofrecen no se agotan, ni con mucho, en los tres elementos anteriores. La propia web del ONLAE y la base de datos que [9] mantiene son buen ejemplo de ello. Para el objetivo que nos hemos marcado, a medio camino entre el rigor y la diversión, estos tres aspectos son suficientes.

## 2.3 Números

La variable $N$ que toma como valor el número obtenido en cada una de las seis extracciones que componen un sorteo, no se obtiene a partir del espacio muestral antes descrito. El espacio muestral ligado a cada extracción, $\Omega'$, es mucho más sencillo. Bajo el supuesto de aleatoriedad, $\Omega' = \{1, 2, \ldots, 49\}$ y es equiprobable. La variable $N$ puede definirse sobre él como la identidad y su distribución es la uniforme sobre el conjunto de los 49 valores de su soporte, $N \in \{1, 2, \ldots, 49\}$. Así pues,

$$P(N = k) = \frac{1}{49}, \quad k = 1, \ldots, 49. \tag{1}$$

El lector estará seguramente de acuerdo con (1) si piensa en la primera de las extracciones de cada sorteo, pero quizás le surjan dudas al pensar en alguna de las 5 restantes. Intentaremos disipárselas. Para ello consideremos el espacio muestral resultante de llevar a cabo las extracciones teniendo en cuenta el orden en que se han producido. El número de posibles resultados[2] será $V_{49,6}$, las variaciones de 49 elementos tomados de 6 en 6. Para obtener la probabilidad de que $N$ tome el valor $k$ en la extracción *i-ésima*, $\{N_i = k\}$, consideramos el suceso $A =$*{conjuntos de 6 extracciones en las que el número* k *ocupa la posición* i*}*. Evidentemente $P(N_i = k) = P(A)$. Los casos favorables a $A$ se obtienen fácilmente excluyendo $k$ de los 49 números, extrayendo 5 de los 48 restantes e insertando a continuación $k$ en la posición deseada, lo que dará lugar a un total de $V_{48,5}$ casos. Por lo tanto, como decíamos antes, tenemos

$$P(N_i = k) = p = \frac{V_{48,5}}{V_{49,6}} = \frac{1}{49}.$$

Los 4.024 sorteos suponen 24.144 extracciones. Estamos ante una muestra aleatoria de tamaño $n = 24.144$ de la variable $N$. Si las extracciones han sido realizadas al azar, la frecuencia

[1] El nombre de *figura*, tomado directamente de la ONLAE, quizás no sea el adecuado porque como puede observarse poco o nada tiene que ver con la disposición geométrica de los números en el boleto.

[2] Este espacio muestral es diferente del $\Omega$ manejado en 2.1 porque ahora tenemos en cuenta el orden de las extracciones

esperada de cada uno de los 49 números es

$$\text{frecuencia esperada de } k = np = \frac{24.144}{49} = 492,7 \quad k = 1, 2, \ldots, 49.$$

La tabla nos muestra ambas frecuencias, observada y esperada, para cada número a lo largo de las 24.144 extracciones.

| **N** | **f_obs (o)** | **f_esp (e)** | $\frac{(o-e)^2}{e}$ | **N** | **f_obs (o)** | **f_esp (e)** | $\frac{(o-e)^2}{e}$ |
|---|---|---|---|---|---|---|---|
| **1** | 503 | 492,7 | 0,2139 | **25** | 491 | 492,7 | 0,0061 |
| **2** | 470 | 492,7 | 1,0490 | **26** | 520 | 492,7 | 1,5087 |
| **3** | 501 | 492,7 | 0,1386 | **27** | 508 | 492,7 | 0,4729 |
| **4** | 488 | 492,7 | 0,0455 | **28** | 471 | 492,7 | 0,9587 |
| **5** | 454 | 492,7 | 3,0450 | **29** | 481 | 492,7 | 0,2795 |
| **6** | 493 | 492,7 | 0,0001 | **30** | 512 | 492,7 | 0,7532 |
| **7** | 491 | 492,7 | 0,0061 | **31** | 516 | 492,7 | 1,0985 |
| **8** | 489 | 492,7 | 0,0283 | **32** | 500 | 492,7 | 0,1071 |
| **9** | 486 | 492,7 | 0,0920 | **33** | 517 | 492,7 | 1,1950 |
| **10** | 484 | 492,7 | 0,1548 | **34** | 512 | 492,7 | 0,7532 |
| **11** | 490 | 492,7 | 0,0152 | **35** | 494 | 492,7 | 0,0032 |
| **12** | 485 | 492,7 | 0,1214 | **36** | 501 | 492,7 | 0,1386 |
| **13** | 485 | 492,7 | 0,1214 | **37** | 468 | 492,7 | 1,2417 |
| **14** | 496 | 492,7 | 0,0216 | **38** | 512 | 492,7 | 0,7532 |
| **15** | 458 | 492,7 | 2,4486 | **39** | 541 | 492,7 | 4,7278 |
| **16** | 483 | 492,7 | 0,1923 | **40** | 465 | 492,7 | 1,5611 |
| **17** | 501 | 492,7 | 0,1386 | **41** | 506 | 492,7 | 0,3571 |
| **18** | 472 | 492,7 | 0,8725 | **42** | 500 | 492,7 | 0,1071 |
| **19** | 491 | 492,7 | 0,0061 | **43** | 480 | 492,7 | 0,3291 |
| **20** | 463 | 492,7 | 1,7944 | **44** | 497 | 492,7 | 0,0369 |
| **21** | 503 | 492,7 | 0,2139 | **45** | 533 | 492,7 | 3,2904 |
| **22** | 509 | 492,7 | 0,5369 | **46** | 469 | 492,7 | 1,1433 |
| **23** | 507 | 492,7 | 0,4130 | **47** | 529 | 492,7 | 2,6691 |
| **24** | 445 | 492,7 | 4,6244 | **48** | 489 | 492,7 | 0,0283 |
| | | | | **49** | 485 | 492,7 | 0,1214 |

Estamos en presencia de un fenómeno que se comporta aleatoriamente y no cabe por tanto esperar que las frecuencias observadas de cada número y las esperadas coincidan. La cuestión es hasta qué punto las diferencias entre unas y otras son debidas al azar o son consecuencia de un mecanismo defectuoso no compatible con la hipótesis de extracciones al azar. Existe una familia de tests estadísticos, conocidos genéricamente como *test de la bondad del ajuste* [10], que permiten contrastar si los datos reales se ajustan a la distribución de probabilidad de $N$ obtenida bajo dicha hipótesis. En el lenguaje habitual del contraste de hipótesis, pretendemos contrastar

**$H_0$:** la variable $N$ se distribuye uniformemente en {1,2,...,49},

**$H_A$:** la variable $N$ no se distribuye uniformemente en {1,2,...,49},

o sus equivalentes,

**$H_0$:** las extracciones se realizan al azar,

**$H_A$:** las extracciones no se realizan al azar.

El funcionamiento del test está basado en medir la discrepancia entre las distribuciones teórica y empírica. El que nosotros vamos a utilizar, el de la $\chi^2$, usa como medida de dicha discrepancia la diferencia, adecuadamente corregida, entre los valores observados de cada número, $o_i$, y los esperados, $e_i$. El correspondiente estadístico se obtiene a partir de la expresión,

$$\chi_e^2 = \sum_{i=1}^{49} \frac{(o_i - e_i)^2}{e_i}.$$

cuya distribución asintótica (cuando el tamaño de la muestra crece a infinito) es una $\chi^2$ con 48=49-1 grados de libertad. Valores altos del estadístico, mayores que cierto umbral, implican una elevada discrepancia de los datos observados con la hipótesis establecida en $H_0$ y, por tanto, nos llevan a rechazarla y a aceptar la hipótesis $H_A$.

Para los datos de las 24.144 extracciones se obtiene $\chi_e^2 = 39,9354$ y el umbral del 5% se establece en 65,1707. Como nuestro estadístico no supera este último valor concluimos que no hay discordancia manifiesta entre los datos observados y $H_0$, lo que nos lleva a *aceptar* (*no rechazar* estaría mejor dicho) la hipótesis de que las extracciones han sido realizadas al azar.

## 2.4 Figuras

El concepto de *figura* hace referencia a la cantidad de números consecutivos que hay en la combinación ganadora. Existen figuras con 0, 2, 3, 4, 5 y 6 números consecutivos, o las posibles agrupaciones que con ellos puedan darse. La tabla siguiente contiene las distintas figuras que pueden presentarse, las diferentes *formas* con las que cada una de ellas se manifiesta y el número de combinaciones favorables a cada forma y figura, necesarias para obtener la correspondiente probabilidad. Si suponemos que la combinación extraída está ordenada de menor a mayor, las *formas* indican la posición que los números consecutivos ocupan en ella. Hemos añadido también una columna con ejemplos de cada figura y forma.

| **Figura** | **Formas** | **Ejemplo** | **Favorables** | **Totales** |
|---|---|---|---|---|
| **111111** | 111111 | *1:3:5:7:9:11* | 7.059.052 | **7.059.052** |
| **21111** | **2**1111 | **1**:**2**:*4:6:8:10* | 1.086.008 | |
| | 12111 | *1:3:4:6:8:10* | 1.086.008 | |
| | 11211 | *1:3:5:6:8:10* | 1.086.008 | |
| | 11121 | *1:3:5:7:8:10* | 1.086.008 | |
| | 11112 | *1:3:5:7:9:10* | 1.086.008 | **5.430.040** |
| **2211** | **22**11 | **1**:**2**:**4**:**5**:*7:9* | 135.751 | |
| | 2121 | *1:2:4:6:7:9* | 135.751 | |
| | 2112 | *1:2:4:6:8:9* | 135.751 | |
| | 1221 | *1:3:4:6:7:9* | 135.751 | |
| | 1212 | *1:3:4:6:8:9* | 135.751 | |
| | 1122 | *1:3:5:6:8:9* | 135.751 | **814.506** |
| **222** | 222 | *1:2:4:5:7:8* | 13.244 | **13.244** |
| **3111** | 3111 | *1:2:3:5:7:9* | 135.751 | |
| | 1311 | *1:3:4:5:7:9* | 135.751 | |
| | 1131 | *1:3:5:6:7:9* | 135.751 | |
| | 1113 | *1:3:5:7:8:9* | 135.751 | **543.004** |

| Figura | Formas | Ejemplo | Favorables | Totales |
|---|---|---|---|---|
| **321** | 321 | *1:2:3:5:6:8* | 13.244 | |
| | 312 | *1:2:3:5:7:8* | 13.244 | |
| | 231 | *1:2:4:5:6:8* | 13.244 | |
| | 213 | *1:2:4:6:7:8* | 13.244 | |
| | 132 | *1:3:4:5:7:8* | 13.244 | |
| | 123 | *1:3:4:6:7:8* | 13.244 | **79.464** |
| **33** | 33 | *1:2:3:5:6:7* | 946 | **946** |
| **411** | 411 | *1:2:3:4:6:8* | 13.244 | |
| | 141 | *1:3:4:5:6:8* | 13.244 | |
| | 114 | *1:3:5:6:7:8* | 13.244 | **39.732** |
| **42** | 42 | *1:2:3:4:6:7* | 946 | |
| | 24 | *1:2:4:5:6:7* | 946 | **1.892** |
| **51** | 51 | *1:2:3:4:5:7* | 946 | |
| | 15 | *1:3:4:5:6:7* | 946 | **1.892** |
| **6** | 6 | *1:2:3:4:5:6* | 44 | **44** |

La obtención del número de combinaciones o casos favorables a cada uno de los sucesos que las formas representan es sencilla pero puede llegar a ser farragosa, particularmente para las figuras/formas con pocos o ningún número consecutivo[3]. Como ilustración obtendremos a continuación los casos favorables a la forma *321*.

**Casos favorables a la forma 321.-** Elegiremos en primer lugar el menor de los números que forman la terna consecutiva, que para esta forma será también el menor elemento de la combinación ordenada. Este número puede ser cualquiera de los 42 primeros, puesto que un número posterior impediría que una forma como la propuesta pudiera aparecer. Designémoslo por $i$. Los dos números que le siguen deben ser $i+1$ e $i+2$. Por tanto, el primero de los dos números consecutivos siguientes, $j$, podrá ser uno cualquiera de los comprendidos entre $i+4$ i 46 ($i+4 \leq j \leq 46$). El último número de la combinación, $k$, habrá de verificar $j+3 \leq k \leq 49$, lo que supone $49-(j+3)+1 = 47-j$ posibilidades. Si sumamos ahora para todos los valores de $i$ y $j$ tendremos

$$\begin{aligned}
\sum_{i=1}^{42}\sum_{j=i+4}^{46}(47-j) &= \sum_{i=1}^{42}\{47(46-(i+4)+1)-\frac{1}{2}(46+i+4)(46-(i+4)+1)\} \\
&= \sum_{i=1}^{42}\{47\times 43-47i-\frac{1}{2}(50+i)(43-i)\} \\
&= \sum_{i=1}^{42}47\times 43-\sum_{i=1}^{42}47i-\frac{1}{2}\sum_{i=1}^{42}(50\times 43-7i-i^2) \\
&= 13.244
\end{aligned}$$

En las 4.024 combinaciones estudiadas, la distribución de las frecuencias de cada una de las figuras es la que mostramos en la tabla. La tabla contiene también las probabilidades para cada figura y las frecuencias esperadas que de ellas se derivan. Con toda esta información podremos

[3]La dificultad puede obviarse recurriendo a un PC, puesto que un sencillo programa en cualquiera de los lenguajes habituales contará rápidamente el número de combinaciones asociadas a cada forma y figura. Los autores deben confesar que han hecho uso de este recurso. ¿No habíamos quedado que la tecnología informática era también para esto?

obtener el valor del correspondiente estadístico $\chi^2_e$ y efectuar el pertinente test de bondad del ajuste[4].

| **figura** | **f_obs** | **prob** | **f_esp** | $\frac{(o-e)^2}{e}$ |
|---|---|---|---|---|
| 111111 | 1994 | 0,504802 | 2031,32 | 0,6857 |
| 21111 | 1610 | 0,388309 | 1562,55 | 1,4406 |
| 2211 | 232 | 0,058246 | 234,38 | 0,0242 |
| 3111 | 138 | 0,038831 | 156,26 | 2,1328 |
| 321 | 33 | 0,005683 | 22,87 | 4,4906 |
| 411 | 15 | 0,002841 | 11,43 | 1,1126 |
| resto | 2 | 0,001288 | 5,18 | 1,9564 |
| *222** | *2* | *0,000947* | *3,81* | |
| *42** | *0* | *0,000135* | *0,54* | |
| *51** | *0* | *0,000135* | *0,54* | |
| *33** | *0* | *0,000068* | *0,27* | |
| *6** | *0* | *0,000003* | *0,01* | |
| **totales** | **4.024** | | **4.024** | **11,8429** |

Al valor $\chi^2_e = 11,8429$ le corresponden en esta situación 6=7-1 grados de libertad y el valor umbral correspondiente al 5% se establece en 12,5916, que al no ser superado por nuestro valor nos conduce, como antes, a aceptar la hipótesis de que las extracciones han sido realizadas al azar.

## 2.5 Sumas

La suma de los números que componen la combinación ganadora es una variable aleatoria cuya distribución de probabilidad, bajo el supuesto de extracciones al azar, puede conocerse. Podemos comprobar si la distribución empírica que se deriva de los 4.024 sorteos concuerda con aquélla.

La variable suma, $S$, es una variable discreta cuyo soporte es el conjunto {21,22,...,279} y puede definirse de dos formas equivalentes:

**Def 1.-** Si $N_i$ denota el número obtenido en la extracción i-ésima,

$$S = \sum_{i=1}^{6} N_i, \quad N_i \in \{1, 2, \ldots, 49\}, \quad N_i \neq N_j, \quad i \neq j.$$

**Def 2.-** Si ordenamos de forma creciente los números extraídos y designamos por $N_{(i)}$ el que ocupa el i-ésimo lugar,

$$S = \sum_{i=1}^{6} N_{(i)}, \quad 1 \leq N_{(1)} < N_{(2)} < \ldots < N_{(6)} \leq 49.$$

La esperanza y la varianza de $S$ pueden obtenerse fácilmente a partir de la primera de las definiciones. Recordemos que las $N_i$ se distribuyen todas ellas uniformemente en {1,2,...,49},

[4]Para evitar problemas con la convergencia del estadístico $\chi^2_e$, es conveniente que la frecuencia esperada no sea inferior a 5 en ninguna de las categorías. Por esta razón hemos agrupado en una sola categoría denominada *resto* las figuras con menor probabilidad, que aparecen en cursiva y señaladas con un * en la tabla.

si bien no son independientes. Como es sabido,

$$E(N_i) = \frac{1+49}{2} = 25, \quad var(N_i) = \frac{49^2-1}{12}, \quad \forall i = 1, \ldots, 6$$

Tendremos pues para $S$,

$$E(S) = \sum_{i=1}^{6} E(N_i) = 6 \times 25 = 150,$$

y puesto que no hay independencia,

$$var(S) = \sum_{i=1}^{6} var(N_i) + \sum_{1 \le i \ne j \le 6} cov(N_i, N_j) = 6 \times var(N_1) + 30 \times cov(N_1, N_2). \qquad (2)$$

La covarianza entre $N_1$ y $N_2$ la obtendremos a partir de la expresión $cov(N_1, N_2) = E(N_1 N_2) - E(N_1)E(N_2)$, de la que nos falta conocer la esperanza del producto.

$$E(N_1 N_2) = \sum_{k \ne l} k \cdot l \cdot P(N_1 = k, N_2 = l) = \frac{1}{49 \times 48} \sum_{k=l}^{49} \sum_{\substack{l=1 \\ l \ne k}}^{49} k \cdot l = \frac{25 \times 1192}{48}.$$

La covarianza valdrá

$$cov(N_1, N_2) = \frac{25 \times 1192}{48} - 25^2 = -\frac{25}{6}.$$

Sustituyendo en (2),

$$var(S) = 6 \times 200 + 30 \times \frac{-25}{6} = 1.200 - \frac{30 \times 25}{6} = 1.075.$$

La función de probabilidad de $S$ vale,

$$f(s) = P(S = s) = \begin{cases} \frac{||A_s||}{V_{49,6}}, & 21 \le s \le 279, \\ 0 & \text{fuera}, \end{cases}$$

donde $||A_s||$ denota el cardinal del conjunto $A_s$ definido mediante

$$A_s = \{(n_1, n_2, \ldots, n_6);\ \sum_{i=1}^{6} n_i = s,\ 1 \le n_i \le 49,\ n_i \ne n_j\}.$$

La obtención de los distintos valores de $f(s)$ es larga y tediosa, por lo que convendrá recordar cuanto se decía en una anterior nota a pie de página. El recurso a un PC nos permite obtener sin gran esfuerzo los valores de $||A_s||$ y $f(s)$. Corresponde ahora contrastar la distribución empírica obtenida a partir de los 4.024 sorteos con la teórica.

La tabla siguiente y la figura 1 resumen la información de ambas distribuciones. La tabla muestra los valores de $||A_s||$ y la figura permite comparar visualmente las gráficas de $f(s) = P(S = s)$ y de *f_rel(s)*, la frecuencia relativa observada para $S = s$. La impresión visual de concordancia que muestran ambas gráficas hemos de corroborarla con el test de bondad del ajuste. Esta concordancia se manifiesta también a través de la media y de la varianza de las sumas observadas:

$$\overline{x} = 151,00, \quad s^2 = 1072,26.$$

| s | $\|\|\mathbf{A_s}\|\|$ | s | $\|\|\mathbf{A_s}\|\|$ | s | $\|\|\mathbf{A_s}\|\|$ | s | $\|\|\mathbf{A_s}\|\|$ | s | $\|\|\mathbf{A_s}\|\|$ | s | $\|\|\mathbf{A_s}\|\|$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **21** | 1 | **64** | 4935 | **107** | 74781 | **150** | 165772 | **193** | 74781 | **236** | 4935 |
| **22** | 1 | **65** | 5426 | **108** | 77624 | **151** | 165732 | **194** | 71928 | **237** | 4494 |
| **23** | 2 | **66** | 5940 | **109** | 80542 | **152** | 165490 | **195** | 69161 | **238** | 4070 |
| **24** | 3 | **67** | 6506 | **110** | 83440 | **153** | 165176 | **196** | 66388 | **239** | 3692 |
| **25** | 5 | **68** | 7097 | **111** | 86412 | **154** | 164654 | **197** | 63706 | **240** | 3331 |
| **26** | 7 | **69** | 7748 | **112** | 89348 | **155** | 164062 | **198** | 61031 | **241** | 3009 |
| **27** | 11 | **70** | 8423 | **113** | 92350 | **156** | 163273 | **199** | 58446 | **242** | 2702 |
| **28** | 14 | **71** | 9163 | **114** | 95311 | **157** | 162410 | **200** | 55875 | **243** | 2432 |
| **29** | 20 | **72** | 9933 | **115** | 98324 | **158** | 161354 | **201** | 53402 | **244** | 2172 |
| **30** | 26 | **73** | 10769 | **116** | 101285 | **159** | 160236 | **202** | 50944 | **245** | 1945 |
| **31** | 35 | **74** | 11637 | **117** | 104295 | **160** | 158923 | **203** | 48586 | **246** | 1729 |
| **32** | 44 | **75** | 12579 | **118** | 107235 | **161** | 157554 | **204** | 46253 | **247** | 1540 |
| **33** | 58 | **76** | 13552 | **119** | 110215 | **162** | 156004 | **205** | 44016 | **248** | 1360 |
| **34** | 71 | **77** | 14603 | **120** | 113119 | **163** | 154397 | **206** | 41809 | **249** | 1206 |
| **35** | 90 | **78** | 15690 | **121** | 116048 | **164** | 152617 | **207** | 39703 | **250** | 1057 |
| **36** | 110 | **79** | 16856 | **122** | 118889 | **165** | 150794 | **208** | 37625 | **251** | 931 |
| **37** | 136 | **80** | 18059 | **123** | 121751 | **166** | 148800 | **209** | 35648 | **252** | 811 |
| **38** | 163 | **81** | 19349 | **124** | 124507 | **167** | 146771 | **210** | 33706 | **253** | 709 |
| **39** | 199 | **82** | 20673 | **125** | 127274 | **168** | 144587 | **211** | 31860 | **254** | 612 |
| **40** | 235 | **83** | 22087 | **126** | 129930 | **169** | 142370 | **212** | 30051 | **255** | 532 |
| **41** | 282 | **84** | 23540 | **127** | 132581 | **170** | 140008 | **213** | 28340 | **256** | 454 |
| **42** | 331 | **85** | 25082 | **128** | 135109 | **171** | 137629 | **214** | 26663 | **257** | 391 |
| **43** | 391 | **86** | 26663 | **129** | 137629 | **172** | 135109 | **215** | 25082 | **258** | 331 |
| **44** | 454 | **87** | 28340 | **130** | 140008 | **173** | 132581 | **216** | 23540 | **259** | 282 |
| **45** | 532 | **88** | 30051 | **131** | 142370 | **174** | 129930 | **217** | 22087 | **260** | 235 |
| **46** | 612 | **89** | 31860 | **132** | 144587 | **175** | 127274 | **218** | 20673 | **261** | 199 |
| **47** | 709 | **90** | 33706 | **133** | 146771 | **176** | 124507 | **219** | 19349 | **262** | 163 |
| **48** | 811 | **91** | 35648 | **134** | 148800 | **177** | 121751 | **220** | 18059 | **263** | 136 |
| **49** | 931 | **92** | 37625 | **135** | 150794 | **178** | 118889 | **221** | 16856 | **264** | 110 |
| **50** | 1057 | **93** | 39703 | **136** | 152617 | **179** | 116048 | **222** | 15690 | **265** | 90 |
| **51** | 1206 | **94** | 41809 | **137** | 154397 | **180** | 113119 | **223** | 14603 | **266** | 71 |
| **52** | 1360 | **95** | 44016 | **138** | 156004 | **181** | 110215 | **224** | 13552 | **267** | 58 |
| **53** | 1540 | **96** | 46253 | **139** | 157554 | **182** | 107235 | **225** | 12579 | **268** | 44 |
| **54** | 1729 | **97** | 48586 | **140** | 158923 | **183** | 104295 | **226** | 11637 | **269** | 35 |
| **55** | 1945 | **98** | 50944 | **141** | 160236 | **184** | 101285 | **227** | 10769 | **270** | 26 |
| **56** | 2172 | **99** | 53402 | **142** | 161354 | **185** | 98324 | **228** | 9933 | **271** | 20 |
| **57** | 2432 | **100** | 55875 | **143** | 162410 | **186** | 95311 | **229** | 9163 | **272** | 14 |
| **58** | 2702 | **101** | 58446 | **144** | 163273 | **187** | 92350 | **230** | 8423 | **273** | 11 |
| **59** | 3009 | **102** | 61031 | **145** | 164062 | **188** | 89348 | **231** | 7748 | **274** | 7 |
| **60** | 3331 | **103** | 63706 | **146** | 164654 | **189** | 86412 | **232** | 7097 | **275** | 5 |
| **61** | 3692 | **104** | 66388 | **147** | 165176 | **190** | 83440 | **233** | 6506 | **276** | 3 |
| **62** | 4070 | **105** | 69161 | **148** | 165490 | **191** | 80542 | **234** | 5940 | **277** | 2 |
| **63** | 4494 | **106** | 71928 | **149** | 165732 | **192** | 77624 | **235** | 5426 | **278** | 1 |
| | | | | | | | | | | **279** | 1 |

El gran número de valores que toma la variable $S$ y la necesidad de efectuar agrupaciones para que los valores esperados de las categorías resultantes sean mayores o iguales que 5, como una correcta aplicación del test de la $\chi^2$ exige, hacen aconsejable recurrir a otro test de la bondad del ajuste concocido como *test de Kolmogorov-Smirnov* (test K-S), que utiliza como medida de discrepancia la diferencia en valor absoluto entre las funciones de distribución teórica,

$$F(x) = \sum_{s \leq x} f(s),$$

y empírica,

$$F_n(x) = \sum_{s \leq x} f\_rel(s) = \frac{||\{s_i,\ s_i \leq x,\ i = 1, \ldots, n\}||}{n},$$

donde $n$ es el número de observaciones y $s_i$ la i-ésima observación de la variable $S$ (en nuestro caso $n$ será el número de sorteos). El correspondiente estadístico se obtiene mediante la expresión

$$D_n = \sup_{x \in R} |F_n(x) - F(x)|,$$

cuya distribución permite establecer el umbral del 5% a partir del cual aceptaremos o rechazaremos nuestra hipótesis. Para $n$ suficientemente grande, como es nuestro caso, dicho umbral vale $1,36/\sqrt{n}$. Para los datos de los 4.024 sorteos $D_n = 0,0159$, menor que el valor del umbral $0,0214 = 1,36/\sqrt{4.024}$. Aceptaremos también ahora que las extracciones se han efectuado al azar.

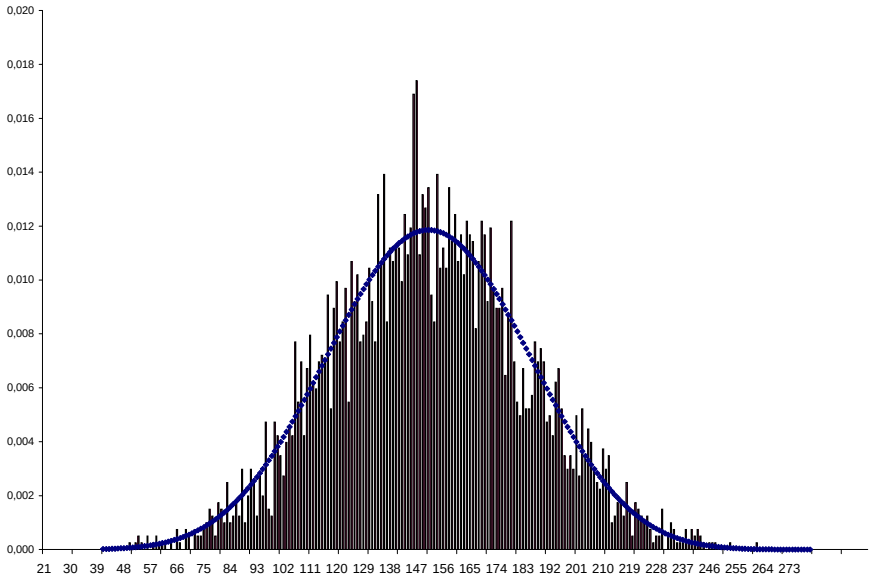


**Figura 1.-** Gráficas de $f(s)$ (linea de puntos) y de la $f\_rel(s)$ (barras verticales).

## 2.6 ¿Cómo juegan los apostantes?

Es difícil concluir esta sección sin caer en la tentación de mirar el juego desde la perspectiva del comportamiento de los apostantes. Y ésto aún a pesar del doble inconveniente que supone, de una parte, situarnos fuera del guión que nos hemos marcado y, de otra, la carencia de información directa sobre las apuestas. El empeño se justificaría porque lo que en un principio nos parecieron sólo llamativas anécdotas, son en realidad las manifestaciones de un comportamiento por parte de los apostantes alejado del puro azar. En efecto, se trata de corroborar algo ya sabido: que la gente traslada al boleto sus preferencias, manías, supersticiones y quimeras; actitudes que tienen como resultado una apuesta casi siempre alejada de lo que entendemos por una elección al azar de los seis números.

Una comprobación indirecta de este hecho, puesto que las apuestas son inaccesibles directamente, la proporciona el análisis conjunto de la combinación ganadora y del número de acertantes de cada categoría para los 4.024 sorteos. Conocemos la probabilidad de acierto para

cada categoría, $p_i$, y el número $n_a$ de apuestas jugadas en cada sorteo, podemos pues conocer $e_i = p_i n_a$, número esperado de acertantes la categoría $i$ bajo el supuesto de que las apuestas han sido hechas al azar. El contraste mediante el test de la $\chi^2$ de la hipótesis de que las apuestas se han realizado al azar en cada uno de los 4.024 sorteos, se resume en la tabla.

| **Hipótesis de apuestas al azar** | |
|---|---|
| **aceptada** | **rechazada** |
| 66 | 3958 |
| *1,64%* | *98,36%* |

Dos comentarios alusivos a la obtención de la tabla:

- las dos primeras categorías de premios han sido agrupadas en una sola porque el número de apuestas no siempre daba lugar al mínimo esperado de 5 acertantes en la 1ª categoría,
- los 66 sorteos en los que la elección al azar de las apuestas es aceptada son una cota máxima que debe ser en realidad menor. En efecto, el contraste de la bondad del ajuste debe realizarse sobre todas las categorías de acertantes, lo que incluye a los que aciertan 2, 1 y 0 números de la combinación ganadora, pero sobre éstos no se tiene información alguna. Por esta razón, la apuesta al azar en cada sorteo la hemos rechazado cuando el valor de $\chi^2_e$ obtenida a partir de las categorías cuyos acertantes conocemos superaba el umbral del 5% obtenido a partir de una $\chi^2_6$, puesto que el número de grados de libertad cuando se incluyen las categorías desconocidas es $6 = 7 - 1$.

Este análisis depara una última sorpresa. Los 66 sorteos en los que hemos aceptado la hipótesis de que las apuestas han sido elegidas al azar, se distribuyen entre los tres tipos de loterías de forma aparentemente distinta a como lo hace el total de sorteos (4.024). La tabla recoge ambas distribuciones y el test de la bondad del ajuste, que en este contexto se denomina *test de homogeneidad*, corrobora la diferencia con un valor $\chi^2_e = 20,5996$ frente a un umbral $\chi^2_2 = 5,9915$. Todo parece indicar que la presencia de 58 Bonolotos excede, por encima de lo que la aleatoriedad explicaría, los 42 esperados, y ello en detrimento de la presencia de Primitivas. ¿Son más amantes del azar los jugadores de Bonoloto que los de la Primitiva? Vaya Ud. a saber.

| | **Bonoloto** | **Gordo** | **Primitiva** |
|---|---|---|---|
| **4.024 sorteos** | 2.576 | 192 | 1.256 |
| **66 al azar** | 58 | 5 | 3 |
| *esperados* | *42,25* | *3,15* | *20,60* |

### 2.6.1 ¿Y los de 6 aciertos?

El conocimiento completo de la apuesta de los acertantes de 1ª categoría (6 aciertos) permite hacer una incursión más detallada en dichas apuestas. Como comprobará más adelante el lector el esfuerzo merece la pena. Comencemos con una estadística de su número a lo largo de los 4.024 sorteos.

En la tabla, $a_i$ es el número de apuestas con 6 aciertos y $n_i$ la frecuencia (número de sorteos) con la que ha aparecido $a_i$.

| $\mathbf{a_i}$ | $\mathbf{n_i}$ | % | $\mathbf{a_i \times n_i}$ | $\mathbf{a_i}$ | $\mathbf{n_i}$ | % | $\mathbf{a_i \times n_i}$ |
|---|---|---|---|---|---|---|---|
| **0** | 2379 | 59,12 | 0 | **11** | 1 | 0,02 | 11 |
| **1** | 974 | 24,20 | 974 | **12** | 4 | 0,10 | 48 |
| **2** | 374 | 9,29 | 748 | **15** | 2 | 0,05 | 30 |
| **3** | 146 | 3,63 | 438 | **17** | 1 | 0,02 | 17 |
| **4** | 63 | 1,57 | 252 | **18** | 1 | 0,02 | 18 |
| **5** | 31 | 0,77 | 155 | **19** | 1 | 0,02 | 19 |
| **6** | 18 | 0,45 | 108 | **20** | 1 | 0,02 | 20 |
| **7** | 8 | 0,20 | 56 | **23** | 1 | 0,02 | 23 |
| **8** | 7 | 0,17 | 56 | **24** | 1 | 0,02 | 24 |
| **9** | 4 | 0,10 | 36 | **56** | 1 | 0,02 | 56 |
| **10** | 5 | 0,12 | 50 | **114** | 1 | 0,02 | 114 |
| | | | | **totales** | **4024** | **100** | **3253** |

Un par de comentarios a la tabla:

- De los 4.024 sorteos aproximadamente el 60% no ha tenido acertantes de 6 y han generado por tanto un *bote* que se ha acumulado a los premios de 1ª categoría de otros sorteos.
- ¿Cómo explicar los 114 acertantes de la combinación ganadora en un mismo sorteo? Quizás atendiendo a la popularidad de ciertos números o a la atracción que ejercen entre la gente. O quizás la explicación sea mucho más sencilla y esté en la estructura del boleto. O quizás una mezcla de ambas razones. Para que el lector decida, ésta fue la combinación ganadora de aquel sorteo,

  **5 15 25 26 36 46**

  y éste el boleto con la combinación marcada sobre él:

| | | | | |
|---|---|---|---|---|
| | 10 | 20 | 30 | 40 |
| 1 | 11 | 21 | 31 | 41 |
| 2 | 12 | 22 | 32 | 42 |
| 3 | 13 | 23 | 33 | 43 |
| 4 | 14 | 24 | 34 | 44 |
| 5 | 15 | 25 | 35 | 45 |
| 6 | 16 | 26 | 36 | 46 |
| 7 | 17 | 27 | 37 | 47 |
| 8 | 18 | 28 | 38 | 48 |
| 9 | 19 | 29 | 39 | 49 |

**Figura 2.-** Combinación ganadora del día 28.10.88 que tuvo 114 acertantes.

Los acertantes de 6 aciertos parecen tener todavía más agudizada la *azarofobia* del apostante. La figura 3 apoya esta opinión. En ella hemos representado la frecuencia esperada para cada uno de los 49 números en las $19.518 = 3.253 \times 6$ elecciones de números llevadas a cabo en las 3.253 apuestas que aquellos jugaron (ver tabla anterior). Como este valor esperado es el mismo para todos ellos, 398,33, aparece como una linea continua paralela al eje de abcisas. Las frecuencias observadas para cada número se han representado mediante puntos, que hemos unido a la recta anterior para que se aprecien mejor las diferencias entre lo esperado y lo observado. El valor del estadístico $\chi^2_e = 623,25$ supera con creces el umbral del 5% de la $\chi^2_{48} = 65,17$.

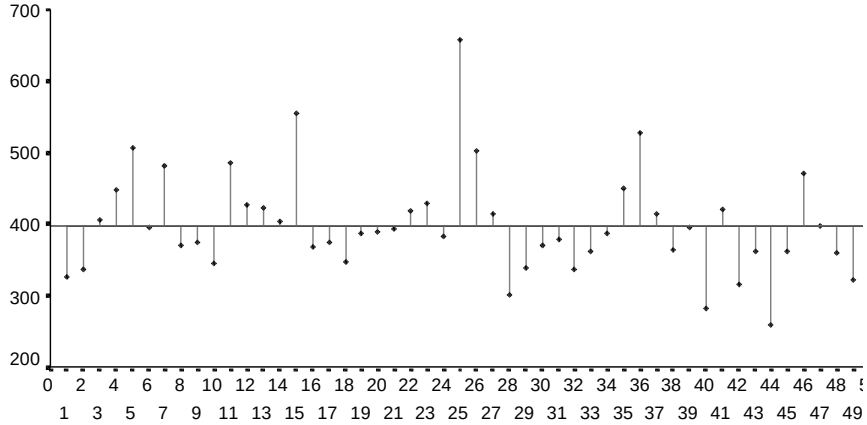


**Figura 3.-** Frecuencias de los 49 números en las apuestas con 6 aciertos.

Finalicemos con el que podemos denominar *boleto-robot* de 6 aciertos. Podemos obtenerlo eligiendo un código de colores y aplicándolo al boleto, de manera que la casilla de cada núnero se tiña con un color adecuado. Un criterio para elegir el color puede ser asociar los valores *más frecuentes* con los colores *más oscuros*, adaptando el rango a una escala de grises. Obtendríamos un resultado como el que se muestra en la figura 4.


**Figura 4.-** Boleto robot de los 6 aciertos.

## 2.7 Dos conclusiones y un comentario

### Conclusión acerca de las extracciones

El buen funcionamiento del mecanismo de extracción de los números en todos los sorteos de las loterías 6/49 ha de aceptarse en base a las características aleatorias analizadas. No arriesgaríamos gran cosa si afirmáramos que con el estudio de otras características el resultado hubiera sido el mismo. No esperábamos otra cosa y habría que señalar dos razones para ello. La primera, y más importante, es la sencillez del mecanismo utilizado, un bombo y 49 bolas homogéneas numeradas, que nos recuerda el comienzo de cualquier enunciado clásico de un problema de probabilidad: *De una urna que contiene bolas numeradas del ....* La segunda razón, de índole ético, se basaría en la creencia de que un organismo estatal de loterías está obligado a hacer las cosas bien aunque sólamente sea

por el dinero que hay en juego, no tanto el de los premios como el que ingresa la hacienda pública.

**Conclusión acerca de las apuestas**

Nada que añadir al comentario que hacíamos al comienzo del párrafo 2.6. Los distintos análisis y anécdotas allí descritos confirman la ausencia de azar en las apuestas de los jugadores. Pero, ¿qué ocurrirá a medida que transcurra el tiempo y pueda generalizarse el uso por parte de los jugadores de las máquinas automáticas de apuestas? Quizás entonces sea el generador pseudoaleatorio de apuestas el que merezca una revisión crítica.

**Y un comentario bibliográfico**

Para cerrar esta sección dedicada a la Primitiva hemos de insistir en que la miscelánea anterior no agota, ni lo pretende, las posibilidades que este juego ofrece. Los juegos de azar son fuente inagotable de inspiración para aquellos que tienen el azar como profesión o pasión, una *o* no necesariamente exclusiva, y han dado lugar por ello a abundante literatura. En la científica, la que aquí nos atañe, son muchos los textos con títulos inequívocamente alusivos al tema, especialmente abundantes entre los precursores de la Teoría de la Probabilidad, lo que da fe de un origen ligado a los juegos de azar. Entre ellos: Huygens [8], Montmort [11], De Moivre [4], ... Pero también existen textos muy recientes que se ocupan del tema con amenidad y rigor, es el caso de los libros de Rao [12] y Haigh [7]. Cerrar esta incursión biliográfica sin citar a dos autores clásicos como Feller [5] y Rényi [13] nos parecería una desconsideración hacia el lector por no haberle hecho partícipe de nuestra pasión por estos dos textos, referencias obligadas cuando se habla de probabilidad discreta (Feller) o de teoría de la probabilidad (Rényi).

# 3 El sorteo del servicio militar del año 70 en los EUA

A finales del año 69 un decreto del Presidente de los EUA, en aquella época Richard Nixon, establecía que los llamamientos para el servicio militar del año 70 debían llevarse a cabo mediante una selección aleatoria basada en la fecha de nacimiento de los implicados[5]. La promulgación del decreto tiene su pequeña historia por las dificultades y prejuicios que hubo de vencer. En ocasiones anteriores se habían utilizado sorteos para determinar la incorporación a filas, adquiriendo algunos de ellos, especialmente el del año 40, una merecida notoriedad debido a la falta de rigor con la que se llevaron a cabo, lo que supuso que las condiciones de equiprobabilidad e independencia necesarias no se satisfacieran.

Se tomaron toda clase de precauciones para que el sorteo del 70 garantizara la igualdad de oportunidades. En un ameno artículo publicado al año siguiente en Science, Fienberg [6] describe con detalle, desde la forma y tamaño de las cápsulas que contenían las 366 fechas de nacimiento (el 29 de febrero estaba también contemplado), hasta la manera en la que las cápsulas eran introducidas en una caja de madera, donde se mezclaban previamente antes de ser alojadas definitivamente en el recipiente del que serían extraídas. El proceso de mezcla merece un comentario más extenso, por ser probablemente el origen de lo insatisfactorio del resultado. Se comenzó alojando las 31 fechas correspondientes a enero en el fondo de la caja, a continuación se hizo otro tanto con las 29 de febrero y se procedió a mezclar ambas. El proceso se repitió con cada uno de los restantes meses, mezclando las bolas después de cada introducción. Este proceso supone que mientras las bolas de enero fueron mezcladas en 11 ocasiones, las de noviembre y diciembre sólo lo fueron dos y una vez, respectivamente. Digamos

[5] De este sorteo ya nos ocupamos en [2], pero analizaremos aquí un nuevo aspecto que consideramos interesante y que proporciona una perspectiva diferente.

por último que una vez las cápsulas fueron arrojadas en el recipiente de la extracción, una gran copa de cristal, ya no fueron agitadas.

| **día** | **ene.** | **feb.** | **mar.** | **abr.** | **may.** | **jun.** | **jul.** | **agos.** | **sept.** | **oct.** | **nov.** | **dic.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 305 | 86 | 108 | 32 | 330 | 249 | 93 | 111 | 225 | 359 | 19 | 129 |
| **2** | 159 | 144 | 29 | 271 | 298 | 228 | 350 | 45 | 161 | 125 | 34 | 328 |
| **3** | 251 | 297 | 267 | 83 | 40 | 301 | 115 | 261 | 49 | 244 | 348 | 157 |
| **4** | 215 | 210 | 275 | 81 | 276 | 20 | 279 | 145 | 232 | 202 | 266 | 165 |
| **5** | 101 | 214 | 293 | 269 | 364 | 28 | 188 | 54 | 82 | 24 | 310 | 56 |
| **6** | 224 | 347 | 139 | 253 | 155 | 110 | 327 | 114 | 6 | 87 | 76 | 10 |
| **7** | 306 | 91 | 122 | 147 | 35 | 85 | 50 | 168 | 8 | 234 | 51 | 12 |
| **8** | 199 | 181 | 213 | 312 | 321 | 366 | 13 | 48 | 184 | 283 | 97 | 105 |
| **9** | 194 | 338 | 317 | 219 | 197 | 335 | 277 | 106 | 263 | 342 | 80 | 43 |
| **10** | 325 | 216 | 323 | 218 | 65 | 206 | 284 | 21 | 71 | 220 | 282 | 41 |
| **11** | 329 | 150 | 136 | 14 | 37 | 134 | 248 | 324 | 158 | 237 | 46 | 39 |
| **12** | 221 | 68 | 300 | 346 | 133 | 272 | 15 | 142 | 242 | 72 | 66 | 314 |
| **13** | 318 | 152 | 259 | 124 | 295 | 69 | 42 | 307 | 175 | 138 | 126 | 163 |
| **14** | 238 | 4 | 354 | 231 | 178 | 356 | 331 | 198 | 1 | 294 | 127 | 26 |
| **15** | 17 | 89 | 169 | 273 | 130 | 180 | 322 | 102 | 113 | 171 | 131 | 320 |
| **16** | 121 | 212 | 166 | 148 | 55 | 274 | 120 | 44 | 207 | 254 | 107 | 96 |
| **17** | 235 | 189 | 33 | 260 | 112 | 73 | 98 | 154 | 255 | 288 | 143 | 304 |
| **18** | 140 | 292 | 332 | 90 | 278 | 341 | 190 | 141 | 246 | 5 | 146 | 128 |
| **19** | 58 | 25 | 200 | 336 | 75 | 104 | 227 | 311 | 177 | 241 | 203 | 240 |
| **20** | 280 | 302 | 239 | 345 | 183 | 360 | 187 | 344 | 63 | 192 | 185 | 135 |
| **21** | 186 | 363 | 334 | 62 | 250 | 60 | 27 | 291 | 204 | 243 | 156 | 70 |
| **22** | 337 | 290 | 265 | 316 | 326 | 247 | 153 | 339 | 160 | 117 | 9 | 53 |
| **23** | 118 | 57 | 256 | 252 | 319 | 109 | 172 | 116 | 119 | 201 | 182 | 162 |
| **24** | 59 | 236 | 258 | 2 | 31 | 358 | 23 | 36 | 195 | 196 | 230 | 95 |
| **25** | 52 | 179 | 343 | 351 | 361 | 137 | 67 | 286 | 149 | 176 | 132 | 84 |
| **26** | 92 | 365 | 170 | 340 | 357 | 22 | 303 | 245 | 18 | 7 | 309 | 173 |
| **27** | 355 | 205 | 268 | 74 | 296 | 64 | 289 | 352 | 233 | 264 | 47 | 78 |
| **28** | 77 | 299 | 223 | 262 | 308 | 222 | 88 | 167 | 257 | 94 | 281 | 123 |
| **29** | 349 | 285 | 362 | 191 | 226 | 353 | 270 | 61 | 151 | 229 | 99 | 16 |
| **30** | 164 | | 217 | 208 | 103 | 209 | 287 | 333 | 315 | 38 | 174 | 3 |
| **31** | 211 | | 30 | | 313 | | 193 | 11 | | 79 | | 100 |

El resultado del sorteo se muestra en la tabla anterior, apareciendo al lado de cada fecha el orden en que fue extraída, orden en el que serían llamados a filas los que nacieron ese día. Este detalle es relevante porque los llamamientos se harían en función de las necesidades, pudiendo suceder que después de determinado llamamiento no se efectuara ninguno más. De hecho, fuentes del Ministerio de Defensa y de la propia Casa Blanca reconocieron por aquellas fechas que, agrupando los 366 llamamientos en tres tercios de 122 fechas cada uno, las necesidades del servicio hacían prever que muy probablemente los del tercer tercio no serían llamados y tampoco parte del segundo.

En el artículo citado de Fienberg y en el de Corberán & Montes [2] se describen con detalle varios métodos para contrastar si el resultado del sorteo era compatible con la extracción al azar de las fechas de nacimiento. Recordaremos aquí uno de ellos y completaremos una sugerencia de Fienberg.

## 3.1 Estudio de las medias mensuales del orden de extracción

El análisis consiste en obtener los valores medios del orden de extracción o llamamiento que han correspondido a las fechas de nacimiento de cada mes y representar gráficamente las parejas

(*mes, valor medio*). Si las extracciones fueron hechas verdaderamente al azar, los puntos en dicha gráfica no deberían presentar ningún tipo de patrón, pero tal como se observa en la figura 5, la tendencia lineal decreciente es clara, especialmente para los meses de mayo a diciembre. La ecuación de la recta de regresión es

$$y = 229,37 - 7,05x,$$

y el coeficiente de correlación vale $r_{xy} = 0,8660$. Semejante valor del coeficiente de correlación nos lleva a aceptar la hipótesis de que las medias se sitúan a lo largo de una recta, lo que resulta incompatible con la supuesta aleatoriedad de las extracciones.

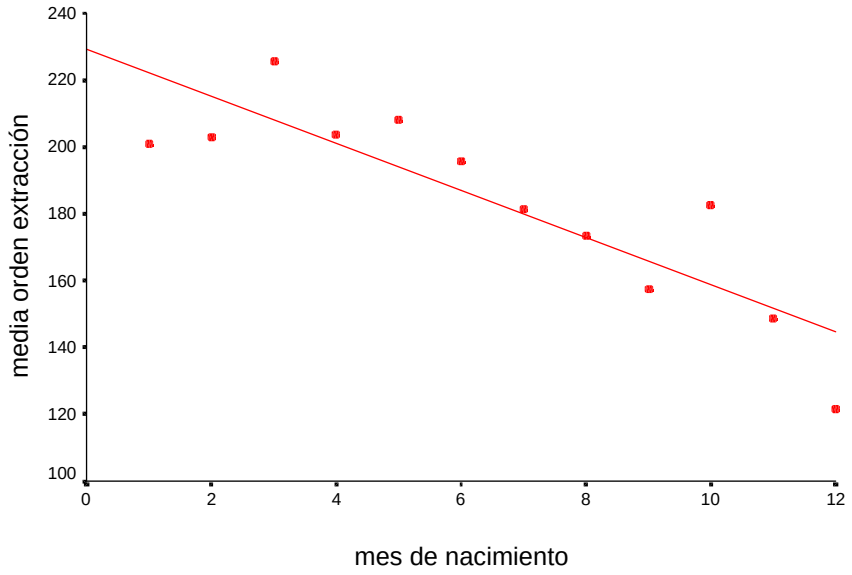


**Figura 5.-** Regresión de la media del número de extracción sobre el mes correspondiente

Los valores de las medias mensuales se muestran en la tabla siguiente y confirman las primitivas sospechas que el método empleado para mezclar las cápsulas hacía temer. Los primeros 5 valores son aproximadamente iguales, la razón puede estar en que la mezcla entre ellos fue más homogénea, pero a medida que aumentaban los meses las cápsulas iban siendo sometidas a menos mezclas. El bajo valor de la media del mes de diciembre refuerza este argumento.

| mes | media orden extracción |
|---|---|
| enero | 201,1 |
| febrero | 203,0 |
| marzo | 225,8 |
| abril | 203,7 |
| mayo | 208,0 |
| junio | 195,7 |
| julio | 181,5 |
| agosto | 173,5 |
| septiembre | 157,3 |
| octubre | 182,5 |
| noviembre | 148,7 |
| diciembre | 121,5 |

## 3.2 Un patrón espacial de puntos

Los resultados del sorteo admiten otro tipo de representación gráfica, la que hemos llevado a cabo en la figura 6. Se trata de un gráfico de dispersión en la que hemos representado el orden de extracción (eje Y) frente a la fecha de nacimiento (eje X).

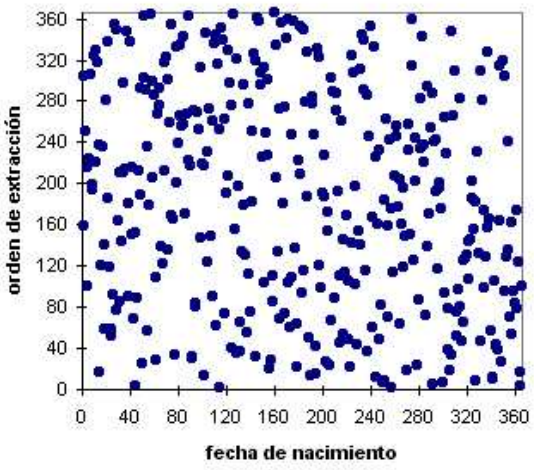


**Figura 6.-** Gráfica del orden de extracción respecto de la fecha de nacimiento

En su artículo, Fienberg analiza visualmente el gráfico y sugiere una cierta escasez de puntos en los extremos de la bisectriz del primer cuadrante. Observa también una tendencia a la agregación de puntos, que tiene como consecuencia la aparición de áreas en blanco relativamente extensas. Pero su análisis no va más allá de apreciaciones subjetivas que en ningún caso tienen el valor de un contraste de hipótesis.

Estamos en presencia de un patrón espacial de puntos sobre un retículo. Si las extracciones han sido realizadas al azar, el patrón debe ser *completamente aleatorio*. ¿Cómo contrastar la hipótesis de aleatoriedad completa? La obtención y exposición rigurosa de semejante test de hipótesis va más allá del objetivo de estas líneas. Hay que añadir, además, que el problema no es trivial debido a las dificultades que conlleva la obtención de la distribución de probabilidad de

los estadísticos asociados con la aleatoriedad espacial completa. Lo que en su lugar haremos es desarrollar un método intuitivo y práctico basado en uno de ellos, la distancia entre los puntos (sucesos) del patrón.

### 3.2.1 Distancia entre sucesos

Una característica numérica que resume bien los rasgos del patrón espacial de $n$ sucesos observados en un recinto acotado, es la *distancia* entre dos cualesquiera de ellos. Más concretamente, la *función de distribución empírica (fded)* de las $\frac{1}{2}n(n-1)$ distancias que generan. Esta es, en realidad, una herramienta propia de los procesos puntuales espaciales definidos sobre $R^2$, pero nada impide su utilización en el campo discreto. En uno u otro soporte, a un patrón espacial completamente aleatorio podemos oponerle alternativas en dos direcciones: un patrón *agregado*, en el que los sucesos se nos muestran formando clusters de mayor o menor tamaño, o un patrón *regular* o *de rechazo*, cuyos sucesos muestran una disposición más regular originada por el rechazo entre ellos. La figura 7 nos muestra patrones que representan las tres situaciones descritas. El de la *izquierda* corresponde a las localizaciones de 62 plantones de secuoias en un cuadrado de 23 ms. de lado, y la agregación que nos muestra podría explicarse por el hecho de que cada grupo de plantones ha crecido alrededor de una sequoia madre que no aparece reresentada en la imagen. En la gráfica *central* los sucesos son las localizaciones de 63 brotes de pinos negros japoneses. El patrón regular que nos muestra la gráfica de la *derecha* responde a la localización de los centros de un grupo de células humanas. Las imágenes están tomadas de [3].

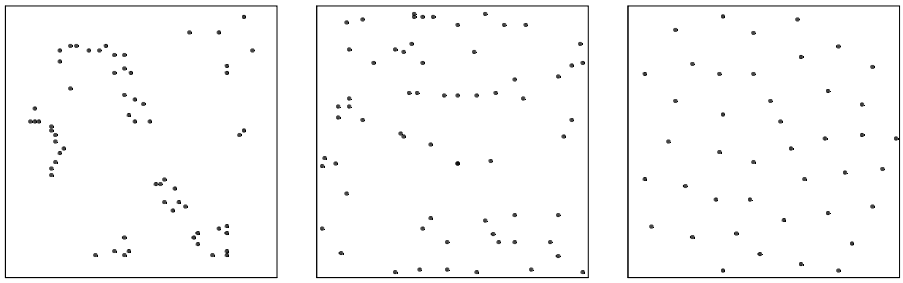

**Figura 7.-** Patrones espaciales de sucesos: agregado (izqda.), aleatorio (centro) y regular (dcha.).

Para ver la capacidad discriminadora de la *fded* observemos la figura 8. Se muestran en ella los histogramas y la *fded* de las distancias entre sucesos de los patrones simulados que aparecen bajo cada gráfica. En todos los casos las simulaciones se han efectuado en un recinto circular de radio 1 y el número de sucesos es similar: 48 para el patrón agregado y 50 para los otros dos.

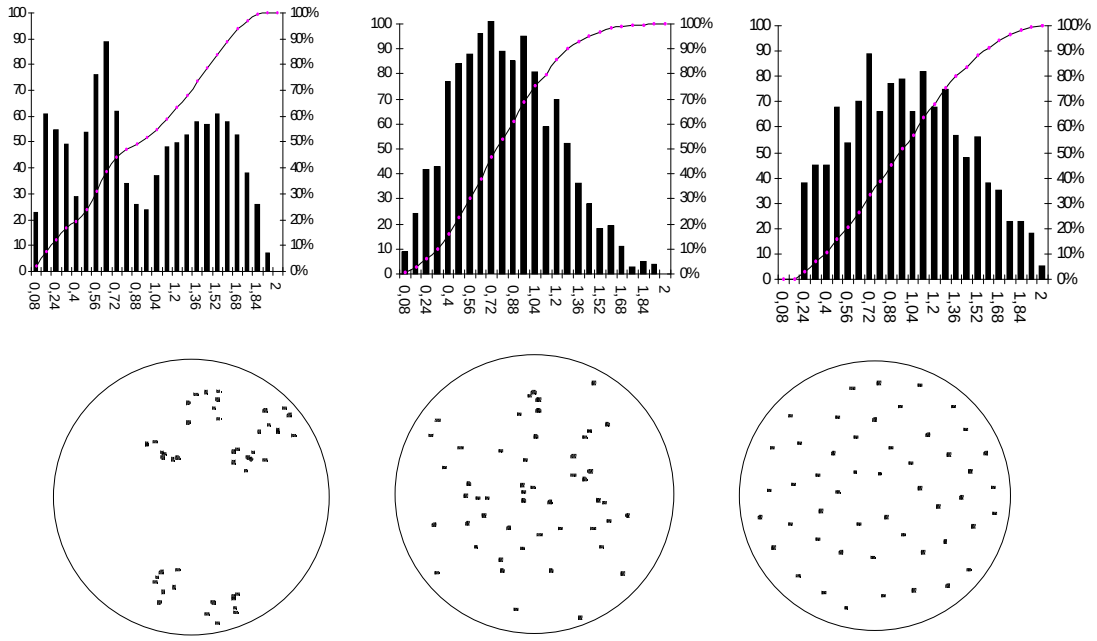

**Figura 8.-** Histogramas y *fded* para patrones agregado (izqda.), aleatorio (centro) y regular (dcha.).

Puede observarse que:

- en el patrón *agregado*, hay una mayor proporción de distancias pequeñas y grandes y una frecuencia mucho menor de distancias intermedias, lo que da lugar a una *fded* con pendiente pronunciada al inicio,
- en el patrón *regular* las distancias pequeñas son inexistentes o mucho menos frecuentes que en el caso agregado, mientras que las intermedias tienen mayor presencia con un cierto equilibrio entre ellas, y
- en el patrón *completamente aleatorio* la distribución, aunque un poco sesgada, recuerda la de una normal, con colas (valores extremos) ligeras y parecidas, y una gradación de valores intermedios con mayor presencia de los centrales.

La obtención de la *fded* es muy sencilla a partir de su definición: en cada punto $x$, es la proporción de distancias que no superan a $x$,

$$\hat{F}(x) = \frac{||\{d_{ij},\ d_{ij} \leq x\}||}{\frac{1}{2}n(n-1)}.$$

Si conociésemos $F(x)$, la función de distribución teórica de las distancias, una primera aproximación al contraste de aleatoriedad podría consistir en representar la ordenada $\hat{F}(x)$ frente a la abscisa $F(x)$, lo que daría lugar, bajo la hipótesis de aleatoriedad completa, a algo muy parecido a la bisectriz del primer cuadrante. Se trata de un método empírico y subjetivo de comprobar dicha aleatoriedad. Excepto en aquellos casos extremos en que los puntos estén alineados a lo largo de la recta $y = x$ o muestren una gráfica claramente distinta de ella, la decisión es siempre difícil de tomar.

El recurso a métodos convencionales de Inferencia Estadística exige conocer la distribución de $\hat{F}(x)$ bajo la hipótesis de aleatoriedad completa. Desgraciadamente ello es imposible o, como mínimo, muy complicado. Cualquier alternativa pasa por utilizar métodos basados en la simulación de patrones completamente aleatorios de sucesos semejantes al nuestro y comparar las funciones de distribución empíricas, o alguna característica asociada, de unos y otro. Veamos dos de estos métodos y apliquémoslos a los datos del sorteo.

### 3.2.2 Método de las envolturas superior e inferior de la *fded*

Decir que el resultado del sorteo es compatible con la extracción al azar de las fechas de nacimiento, equivale a decir que el orden de extracción es una permutación aleatoria de los 366 primeros naturales. Lo que haremos será generar $n-1$ de estas permutaciones y obtener para cada una de ellas su *fded*, $\hat{F}_i(x)$, $i=2,\ldots,n$ y compararlas con la $\hat{F}_1(x)$ derivada del sorteo.

Una forma sencilla de llevar a cabo la comparación es construir las *envolturas superior* y *inferior* de las simuladas mediante

$$S(x)=\max_{i=2,\ldots,n}\hat{F}_i(x), \qquad I(x)=\min_{i=2,\ldots,n}\hat{F}_i(x).$$

Si los datos del sorteo son completamente aleatorios, $\hat{F}_1(x)$ no debería distinguirse de las $\hat{F}_i(x)$, $i=2,\ldots,n$ y se verificaría

$$P(\hat{F}_1(x)>S(x))=P(\hat{F}_1(x)<I(x))=\frac{1}{n}, \quad \forall x. \tag{3}$$

Si sistemáticamente la $\hat{F}_1(x)$ está por encima de $S(x)$ o debajo de $I(x)$, tendremos entonces un claro indicio en contra de la hipótesis establecida.

Para analizar los datos del sorteo hemos generado 99 permutaciones aleatorias del 1 al 366, lo que significa que las probabilidades en (3) valen 0.01, y es por tanto muy improbable que $\hat{F}_1(x)$ supere a $S(x)$ o sea inferior a $I(x)$. La gráfica conjunta de las tres funciones no es muy informativa debido a que sus diferencias son muy pequeñas, pero si representamos conjuntamente las gráficas $(x,\hat{F}_1(x)-I(x))$ y $(x,\hat{F}_1(x)-S(x))$, como hemos hecho en la figura 9, veremos que mientras la primera es siempre positiva, como era de esperar, la segunda es positiva entre $x=50$ y $x=128$, en contra de lo esperado. Habría que concluir que existen claros indicios de que la extracción no fue al azar.

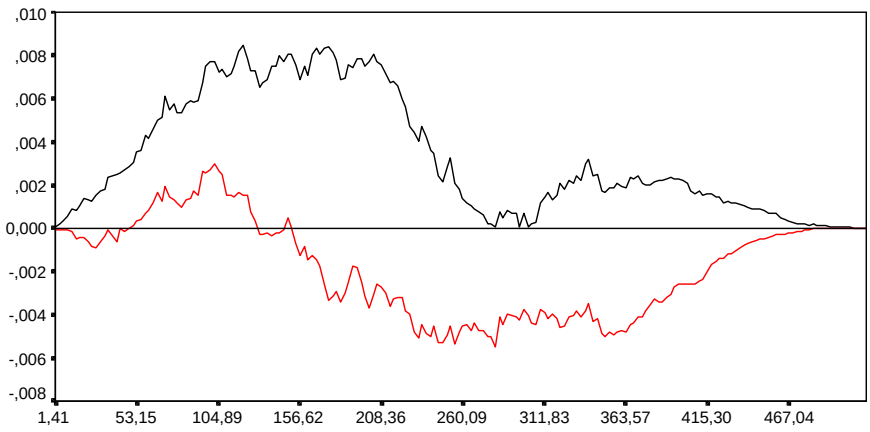


**Figura 9.-** Comparación de $S(x)$, $\hat{F}_1(x)$ y $I(x)$.

### 3.2.3 Test de Montecarlo para el intervalo intercuartílico

El otro método de contraste propuesto se basa en una característica numérica asociada a la *fded* de las distancias, y forma parte de una familia de métodos conocidos genéricamente como *Tests de Montecarlo*, introducidos por Barnard en 1963 [1]. En esencia, se trata de comparar la característica numérica estudiada en el patrón observado con los valores que esa misma característica toma en los $n-1$ patrones completamente aleatorios que hemos simulado. Si la hipótesis de aleatoriedad completa es satisfecha, al ordenar conjuntamente el valor observado y los simulados, aquél podrá ocupar cualquier posición con la misma probabilidad, $1/n$. Al igual que en un contraste de hipótesis clásico, posiciones extremas, originadas por valores extremos[6], son evidencias en contra de la hipótesis nula asumida.

El método permite establecer el nivel de significación a partir del cual rechazaremos la hipótesis nula. Así, si en un contraste bilateral rechazamos la hipótesis cuando el valor observado se encuentra entre los $k$ mayores o los $k$ menores, el nivel de significación del test es $\alpha = 2k/n$. Una elección adecuada de $n$ nos permitirá fijar el $\alpha$ deseado.

Volvamos al resultado del sorteo. La característica numérica que vamos a observar es el llamado *intervalo intercuartílico*, $IIQ$, de la distribución de las distancias entre sucesos. Se trata de una medida de dispersión definida mediante

$$IIQ = p_{75} - p_{25},$$

donde $p_{75}$ y $p_{25}$ son, respectivamente, el *tercer* y el *primer cuartiles* que, recordemos, son los valores que dejan a su izquierda el 75% y el 25% , respectivamente, de las distancias observadas.

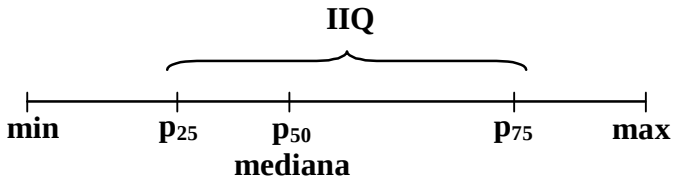


**Figura 10.-** El intervalo intercuartílico.

El intervalo intercuartílico de las $66.795 = \frac{1}{2} \times 366 \times 365$ distancias obtenidas a partir del resultado del sorteo vale $IIQ = 139,5$. Para las 99 simulaciones los valores de sus $IIQ$ varían entre 138,4 y 136,4. Nuestro valor es el mayor de todos ellos La conclusión es obvia: rechazaremos la hipótesis de aleatoriedad completa puesto que la probabilidad de encontrar un valor en cualquiera de los extremos es $2 \times 0,01$ (el contraste es bilateral porque la hipótesis alternativa es la ausencia de aleatoriedad completa, que puede manifestarse indistintamente con valores altos o bajos).

## 3.3 Conclusión

Todo parece indicar que las precauciones que se tomaron no fueron suficientes. El resultado del sorteo no parece compatible con las condiciones requeridas por las extracciones al azar, como se deduce de las pruebas anteriores y de otras complementarias que pueden consultarse en los artículos de Fienberg [6] y Corberán & Montes [2].

El resultado de los tests de aleatoriedad completa nos conducen al rechazo de ésta, pero nos informan también del sentido de la desviación respecto de aquélla. En efecto, los patrones

[6]Según que el contraste sea *unilateral* o *bilateral* tendremos en cuenta uno o los dos extremos

agregados tienen mayor dispersión que los regulares (una nueva ojeada a la figura 8 puede merecer la pena) y el elevado valor del $IIQ$ observado estaría indicándonos la presencia de un patrón agregado. El resultado de la regresión va también en este sentido.

## Agradecimientos

Los autores quieren manifestar su agradecimiento por la ayuda recibida en el apartado de la Primitiva a D. Pedro R. López Torres, informático vehemente, para quién este juego sólo guarda un secreto: acertar la combinación ganadora.

## Referencias

[1] G.A. Barnard. Contribution to the discussion of Professor Bartlett's paper. *J.R.Statist.Soc. B* , 25:294, 1963.

[2] A. Corberán & F. Montes. Perversiones y trampas de la probabilidad. *La Gaceta de la RSME*, 3:198:229, 2000.

[3] P.J. Diggle. *Statistics Analysis for Spatial Point Patterns.* Academic Press, London, 1983.

[4] A De Moivre. *The Doctrine of Chances. 3rd. Edition.* A. Millar, London, 1756. (Reeditado por Chelsea Pub. Comp., New York, 1967)

[5] W. Feller. *An Introduction to Probability Theory ans Its Applications. Vol I. 3rd. Edition.* John Wiley, New York, 1968.

[6] S.E. Fienberg, Randomization and Social Affairs: The 1970 Draft Lottery. *Science*, 171:255–261, 1971.

[7] J. Haigh. *Taking Chances.* Oxford University Press, Oxford, 1999.

[8] C. Huygens. *Tractatus de Rationiis in Aleae Ludo.* In *Exercitationes Mathematicae.* Ed. F. van Shooten, Amsterdam, 1657.

[9] P.R. López Torres. Comunicación personal. Albacete, 2000.

[10] R.J. Larsen and M.L. Marx. *An Introduction to Mathematical Statistics ans Its Applications. 3rd. Edition.* Prentice Hall College Div., New Jersey, 2000.

[11] P.R. Montmort. *Essai d'Analyse sur les Jeux d'Hazard.* Ed. Jacques Quillau, Paris, 1713. (Reeditado por Chelsea Pub. Comp., New York, 1980)

[12] C.R. Rao. *Estadística y Verdad: Aprovechando el Azar.* PPU, Barcelona, 1994.

[13] A. Rényi. *Probability Theory.* North-Holland, Amsterdam, 1970.