

Verificación y monitoreo de la aleatoriedad de los juegos de números de d dígitos

Testing and Monitoring the Randomness of d Digit Number Game

HUMBERTO GUTIÉRREZ-PULIDO^a, JUAN MANUEL GARCÍA^b

DEPARTAMENTO DE MATEMÁTICAS, CUCEI, UNIVERSIDAD DE GUADALAJARA, GUADALAJARA, MÉXICO

Resumen

El interés de este trabajo se centra en el problema de probar la aleatoriedad de los resultados de los juegos de números de d dígitos. Es usual que este problema se aborde con pruebas aproximadas del tipo χ^2 y otras pruebas de independencia de resultados sucesivos. Pero estas pruebas, tienen entre otras limitantes, el hecho de que requieren muestras grandes. Como alternativa, en este trabajo se detalla una prueba bayesiana basada en el modelo multinomial. Además, para monitorear los resultados de este tipo de juego de azar y detectar en forma oportuna patrones y resultados no aleatorios, se propone la utilización de una carta de control geométrica. Se hace un breve estudio Monte Carlo para comprender mejor las características de la carta propuesta. Como caso práctico se analizan los resultados de 500 sorteos de la lotería mexicana Tris y se detectan problemas de falta de aleatoriedad, tanto con la prueba bayesiana como con la carta de control.

Palabras clave: carta de control, distribución geométrica, distribución multinomial, distribución Dirichlet, lotería, prueba ji-cuadrada.

Abstract

this work is centered on testing the randomness of d -digit number game. It is usual that this problem is studied by the χ^2 test and other tests for independence of successive draws. However, these tests require of large sample sizes. As an alternative, it is proposed a Bayesian methodology based on the multinomial model. This methodology does not depend on asymptotic results. Besides, for monitoring the results of this type of game, it is proposed a geometric control chart. Monte Carlo study is carried out to analyze this chart. As a practical case, the data of 500 draws of mexican lottery Tris were analyzed, and problems of lack of randomness are detected.

Key words: Control chart, Dirichlet distribution, Geometric distribution, Lottery Chi-square test, Multinomial distribution.

^aProfesor titular. E-mail: humberto.gutierrez@cucei.udg.mx

^bProfesor titular. E-mail: jmgarcia@prodigy.net.mx

1. Introducción

La fascinación y afición por los juegos y apuestas de azar tienen profundo arraigo en la historia de la humanidad, y están documentadas desde los primitivos juegos de dados en la antigüedad hasta los actuales juegos de póker y apuestas en línea (ver Schwartz 2003, 2006). El gran interés por los juegos de azar motivó, sobre todo a partir del siglo XVII, el desarrollo teórico del cálculo de probabilidades. En la actualidad, con el incremento del tiempo de ocio y la aparición de “casinos electrónicos”, los juegos de azar son una práctica para mucha gente, lo que se debe reconocer como un riesgo, porque hay personas que alcanzan una adicción patológica a este tipo de juegos. A esta adicción se le conoce como ludopatía, y ha llevado a que en diferentes épocas y países, los juegos y apuestas de azar hayan sido prohibidos. Sin embargo, la tendencia ha sido a liberalizar y reglamentar lo que hoy sin duda es una industria mundial en la que participan los propios gobiernos, con la aparición de las loterías nacionales.

En este contexto, uno de los mayores retos es darle confianza y certeza al jugador y a la sociedad en cuanto a que los resultados de los sorteos son aleatorios. Hay dos formas básicas con las que se ha buscado dar tal confianza y certeza; la primera, controlando el proceso, y para ello se buscan mejores equipos para realizar los sorteos, mayor uniformidad entre las esferas utilizadas, elección aleatoria de urnas, sorteos en vivo, etc. La otra forma, es mostrar evidencias estadísticas de que los resultados se pueden considerar aleatorios. Sobre esto último, en el Reino Unido, la Comisión Nacional de la Lotería (National Lottery Commission, NLC) reguladora de la Lotería Nacional, tiene un programa de investigación para comprobar si hay elementos de no aleatoriedad en los diferentes juegos de la Lotería Nacional del Reino Unido (ver Royal Statistical Society 2000, 2002 y University of Salford 2004, 2005*a*, 2005*b*).

Adicionalmente, en la literatura especializada en estadística hay diferentes trabajos que abordan problemas de la aleatoriedad de los juegos de azar y aspectos relacionados. Unos enfocados a estudiar juegos de números de d dígitos, en los que el número ganador se determina por la extracción al azar de d esferas, usualmente cada una proveniente de una urna que contiene diez esferas numeradas con los dígitos: 0, 1, 2, ..., 9. Otros trabajos se enfocan a los juegos de lotería del tipo k de N , en los que la combinación ganadora se determina por la extracción al azar de k esferas sin remplazo de una urna que contiene esferas similares numeradas del 1 al N . Sobre este tipo de juegos, en Joe (1987) se estudian las posibles distribuciones de un conjunto de k -tuplas, dadas las frecuencias marginales de los números del 1 al N , y se analizan datos de la lotería Lotto 6/49 de Canadá. En Stern & Cover (1989) se estima la distribución de máxima entropía de los boletos comprados en la lotería Lotto 6/49 de Canadá, buscando determinar números impopulares para el público. En Joe (1990) se cuestiona la estrategia de comprar un boleto con una k -tupla de números impopulares. Por su parte, Johnson & Klotz (1993) desarrollan métodos numéricos para calcular las estimaciones de máxima verosimilitud de las probabilidades para los diferentes números de una lotería, y dan el estadístico de razón de verosimilitud generalizado $-2\ln()$ para probar que los números son equiprobables; estos métodos los aplican a un conjunto de datos de

la lotería Lotto América. En Joe (1993) se obtiene la modificación del estadístico de prueba ji-cuadrada para probar la uniformidad de los números del 1 al N , dado que los números se extraen sin remplazo, y se analizan datos de la lotería Lotto 6/49 de Canadá. Finkelstein (1995) estima la distribución de frecuencia de los números a los que se les apuesta en la lotería de California, y prueba la hipótesis de que los números pequeños son más populares. En Genest, Lockhart & Stephens (2002) se discute el uso del estadístico χ^2 para bondad de ajuste, para probar la equiprobabilidad de la ocurrencia de números de una lotería, el cual no sigue una distribución χ^2 simple porque los números se extraen sin remplazo, y muestran que su distribución asintótica es la de una suma ponderada de variables aleatorias χ^2 ; este resultado lo utilizan para comprobar la uniformidad de números ganadores en la lotería Lotto 6/49 de Canadá. Koning & Vermaat (2002) calculan las probabilidades de ganar premios y el pago esperado en la Lotto holandesa; también prueban la hipótesis de que los números y colores se extraen aleatoriamente. Percy (2006) describe una metodología bayesiana para probar la aleatoriedad de los sorteos de lotería, y considera el caso de datos de la Lotería Nacional del Reino Unido. En Coronel-Brizio, Hernández-Montoya, Rapallo & Scalas (2008) se abordan también las loterías del tipo k/N mediante la obtención de la media y varianza de la variable aleatoria que representa a los números extraídos; también usan la distribución hipergeométrica, y sus resultados los aplican a la lotería mexicana Melate y a la lotería italiana. De esta revisión se desprende que los estudios estadísticos para los juegos de lotería del tipo k de N son variados; y algunos han propuesto metodologías para verificar la aleatoriedad de los resultados, y otros han estudiado la forma en que la gente elige los números a los que les apuestan.

En cuanto a los juegos de números de d dígitos, en Clotfelter & Cook (2008) se usan datos del juego de números de la lotería de Maryland para probar la hipótesis de que los jugadores de lotería creen que la probabilidad de un evento disminuye cuando ese evento ha ocurrido recientemente (“la falacia del jugador”), y lo comprueban mostrando cómo las apuestas por los números que han ganado recientemente disminuye en los sorteos siguientes. En Teo & Leong (2002) se estudia el diseño de un mecanismo de control para el manejo de las apuestas en un juego de números de cuatro dígitos, esto es, para decidir si las apuestas deben ser aceptadas o rechazadas por el operador que ofrece el juego. En términos generales, el tema de juegos de números de d dígitos ha recibido mucha menos atención en la literatura especializada; sin embargo el tipo de herramientas que se utilizan para probar la aleatoriedad de estos tipo de sorteos presentan algunos problemas. Por ejemplo, usualmente para verificar la igualdad de la frecuencia de los dígitos o grupos de ellos se aplican pruebas aproximadas, como la tradicional prueba χ^2 , que entre otros inconvenientes requiere tamaños de muestra grandes para lograr buenas aproximaciones (Cai & Krishnamoorthy 2006); otro problema es que para probar la independencia se aplican usualmente pruebas que precisan el supuesto de normalidad.

El objetivo del presente artículo es proponer alternativas de solución a los dos problemas anteriores. Para el primero, en la sección 2 se propone utilizar el procedimiento bayesiano basado en la distribución multinomial que no requiere resultados asintóticos. Para el segundo problema, en la sección 3 se propone el uso

de una carta geométrica como alternativa para estar monitoreando en tiempo real la aleatoriedad de los resultados de un juego de d dígitos. Adicionalmente, en la sección 4 se presentan los resultados de un estudio de simulación mediante el cual se analizan las propiedades de la carta de control propuesta. En la sección 5 se aplican los procedimientos propuestos a un conjunto de resultados del sorteo Tris, un juego popular en México, promovido por una institución pública. Mediante ambos procedimientos se detectan problemas significativos en la aleatoriedad de dicho sorteo. Por último, en la sección 6 se hace la discusión y se plantean las conclusiones del trabajo.

2. Igualdad de frecuencia de los dígitos

Una primera forma de probar la aleatoriedad de los resultados de los juegos de números de d dígitos es ver si hay igualdad de frecuencia de los diferentes dígitos, ya sea considerando cada urna por separado o todas a la vez. Esto se puede expresar de manera general, considerando n ensayos independientes (n sorteos), donde cada esfera tiene probabilidad p_i de ser extraída, con $p_i > 0$ y $\sum_{i=1}^m p_i = 1$, y si X_i es el número de veces que se extrae la esfera i , entonces el vector $X = (X_1, \dots, X_m)$ sigue una distribución multinomial con parámetro n y $\mathbf{p}$, donde $\mathbf{p} = (p_1, \dots, p_m)$. La densidad de probabilidades multinomial está dada por

$$f(x_1, \dots, x_m; n, \mathbf{p}) = \frac{n!}{x_1! \dots x_m!} \prod_{i=1}^m p_i^{x_i} \quad (1)$$

con $n = \sum_{i=1}^m x_i$. Es de interés probar la siguiente hipótesis

$$\begin{aligned} H_0 : p_1 &= p_{1,0}, \dots, p_{m-1} = p_{m-1,0} \\ H_a : p_i &\neq p_{i,0} \text{ para algún } i = 1, 2, \dots, m-1 \end{aligned} \quad (2)$$

donde en el caso del juego de números de d dígitos, $m = 10$ y $p_{i,0} = 0.1$.

2.1. Algunas pruebas basadas en la distribución ji-cuadrada

En Cai & Krishnamoorthy (2006) se analizan varias pruebas para (2) donde se usa de alguna forma la distribución χ^2 . La primer prueba, que es la usual y que originalmente desarrolló Pearson en 1990, se basa en el estadístico dado por:

$$\chi_0^2 = \sum_{i=1}^m \frac{(X_i - np_{i,0})^2}{np_{i,0}} \quad (3)$$

Bajo H_0 este estadístico sigue en forma aproximada una distribución χ_{m-1}^2 con $m-1$ grados de libertad (Cai & Krishnamoorthy 2006), por lo que la hipótesis nula en (2) se rechaza, con un nivel de significancia de α , si $\chi_0^2 \geq \chi_{\alpha, m-1}^2$.

Por su parte, la prueba de la razón de verosimilitud se basa en el estadístico

$$R = 2 \sum_{i=1}^m X_i \ln \left(\frac{X_i}{np_{i,0}} \right) \quad (4)$$

que sigue en forma aproximada una distribución χ_{m-1}^2 con $m-1$ grados de libertad (Cai & Krishnamoorthy 2006). Nass (1959) propuso otra aproximación a la distribución de (3), con

$$c\chi_0^2 \sim \chi_v^2 \quad (5)$$

donde $c = 2E(\chi_0^2)/Var(\chi_0^2)$ y $v = cE(\chi_0^2)$. Estas constantes se obtienen (ver Cai & Krishnamoorthy 2006), a partir de

$$E(\chi_0^2) = m-1 \quad y \quad Var(\chi_0^2) = 2(m-1) - (m^2 + 2m - 2)/n + \sum_{i=1}^m (np_{i,0})^{-1} \quad (6)$$

Una variante propuesta por Nass (1959), de especial interés en este trabajo, es cuando se tiene igual probabilidad en las celdas de la distribución multinomial, es decir, cuando $p_{i,0} = 1/m$ para todo i . En este caso se hace una corrección por continuidad para que el estadístico (3) tome la forma siguiente:

$$\chi_c^2 = \frac{\sum_{i=1}^m X_i^2 - 1}{(n/m)} - n \quad (7)$$

con lo que la varianza de χ_c^2 en (6) toma la forma $2(m-1)(n-1)/n$.

Una problemática de estas pruebas para validar la aleatoriedad de los juegos de azar de d dígitos es que son aproximadas, y requieren valores grandes en el número de sorteos, lo que restringe su aplicación. En Cai & Krishnamoorthy (2006) se muestra que las pruebas aproximadas pueden tener un inadecuado comportamiento para tamaños de muestra pequeños y para $p_{i,0}$ pequeños, que serían las situaciones de mayor interés en el caso de juegos de número de d dígitos, ya que para empezar $p_{i,0} = 0.1$.

Finalmente es importante agregar otras variantes sobre las que existen muchos trabajos en la literatura, que usan la aproximación χ^2 para probar la hipótesis en (2) por el método de intervalos de confianza. Estas variantes consisten en estimar intervalos de confianza simultáneos para las m proporciones p_i del modelo multinomial (ver por ejemplo Sison & Glaz 1995, Hou, Chiang & Tai 2003). El primer trabajo sobre el particular es el de Quesenberry & Hurst (1964), y una mejora de este, bastante fácil de implementar, es la propuesta por Goodman (1965), que señala que un intervalo simultáneo aproximado, con una confianza de $100(1-\alpha)$ para cada una de las p_i se obtiene con:

$$\frac{B + 2x_i \pm \{B[B + 4x_i(n - x_i)/n]\}^{1/2}}{2(n + B)} \quad (8)$$

donde B es el $100 \times (1 - \alpha/m)$ percentil superior de la distribución χ^2 con un grado de libertad.

2.2. Alternativa bayesiana

Dado el modelo multinomial (1), la verosimilitud correspondiente está dada por

$$L(p_1, \dots, p_m; n, X) \propto \prod_{i=1}^m p_i^{x_i} \quad (9)$$

A priori, para el modelo multinomial se puede utilizar su conjugada (Bernardo & Smith 1994), que es la distribución Dirichlet (m, α) :

$$\pi(p_1, \dots, p_m) = \frac{1}{B(\alpha)} \prod_{i=1}^m p_i^{\alpha_i-1} \quad (10)$$

donde $p_i > 0$, $\sum_{i=1}^m p_i = 1$, $\alpha_i > 0$, $\alpha = (\alpha_1, \dots, \alpha_m)$ y

$$B(\alpha) = \frac{\prod_{i=1}^m \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^m \alpha_i)}$$

Con $\alpha_i = 1$ para todo i , se logra una distribución a priori no informativa. Una propiedad interesante es que si $\mathbf{p}$ tiene una distribución Dirichlet (m, α) , entonces la distribución marginal de p_i es una distribución Beta $(\alpha_i, \alpha_0 - \alpha_i)$, con $\alpha_0 = \sum_{i=1}^m \alpha_i$. Además la distribución marginal de cualquier sub-vector de una Dirichlet es también Dirichlet; por ejemplo la distribución marginal de $(p_i, p_j, 1 - p_i - p_j)$ es Dirichlet $(\alpha_i, \alpha_j, \alpha_0 - \alpha_i - \alpha_j)$.

De (9) y (10) se concluye que la distribución posterior está dada por

$$\begin{aligned} \pi(p_1, \dots, p_m | X) &\propto \prod_{i=1}^m p_i^{x_i} \prod_{i=1}^m p_i^{\alpha_i-1} \\ &\propto \prod_{i=1}^m p_i^{\alpha_i+x_i-1} \end{aligned} \quad (11)$$

que, salvo por una constante, corresponde a una distribución Dirichlet (m, α^*) , con $\alpha^* = (\alpha_1 + x_1, \dots, \alpha_m + x_m)$. De esta manera, con base en (11) se puede hacer inferencia para $\mathbf{p} = (p_1, \dots, p_m)$, lo que permite explorar la hipótesis (2) o el caso particular de interés en los juegos de número de d dígitos, donde $m = 10$ y $p_{i,0} = 0.1$. Para explorar la distribución posterior resulta conveniente obtenerla por simulación; por ejemplo en el software libre R, que se puede bajar de <http://cran.r-project.org> junto con el paquete MCMCPack, que al cargarlo puede generar la distribución posterior con la función *MCmultinomdirichlet* $\mathbf{x}$, α^* , mc , donde $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_m)$ y mc es el número de realizaciones que se quiere de $\mathbf{p} | \mathbf{X}$. También puede ser de utilidad la función *ddirichlet* $(\mathbf{p}, \alpha)$.

Alternativamente a la simulación se puede recurrir a las expresiones analíticas. Por ejemplo, de acuerdo con lo que se dijo antes, la distribución marginal posterior de cualquiera de las p_i es una Beta $(\alpha_i + x_i, \alpha_0^* - (\alpha_i + x_i))$, con $\alpha_0^* = \sum_{i=1}^m (\alpha_i + x_i)$. Así, para cualquier p_i se tiene que

$$\pi(p_i | X) = \frac{\Gamma(\alpha_0^*)}{\Gamma(\alpha_i + x_i) \Gamma(\alpha_0^* - (\alpha_i + x_i))} p_i^{\alpha_i+x_i-1} (1 - p_i)^{\alpha_0^* - (\alpha_i + x_i) - 1} \quad (12)$$

Para explorar esta distribución se puede graficar para alguna i y ver qué tantas posibilidades se le da a $p_i = 0.1$, ya sea calculando un intervalo de probabilidad para p_i con los cuantiles $q_{\gamma/2}$ y $q_{1-\gamma/2}$ de la distribución $\pi(p_i | X)$, o bien calculando $\Pr(p_i > 0.1 | X)$ o $\Pr(p_i < 0.1 | X)$.

El procedimiento descrito también se puede ajustar fácilmente para plantear la aleatoriedad de grupos de dígitos. Basta hacer los ajustes en el modelo multinomial correspondiente. El procedimiento descrito no requiere aproximaciones asintóticas, por lo que se puede aplicar en cualquier momento, además de ser más informativo al obtener la distribución posterior para $\mathbf{p}$ o las p_i de interés.

3. Independencia y monitoreo de resultados

Puede pasar que con base en las pruebas de la sección anterior se declare que la frecuencia con la que ocurre cada dígito es similar; sin embargo esto no es suficiente para asegurar la aleatoriedad de los resultados, ya que pudieron darse rachas de sorteos que favorecieran a ciertos dígitos, pero que al combinar toda la información y resumir las frecuencias, estas rachas queden ocultas. Por ello se requiere también verificar la independencia entre los resultados sucesivos.

La forma tradicional de verificar esta independencia es ver los resultados de una urna a través de los diferentes sorteos como una serie de tiempo, y calcular la autocorrelación para medir la correlación entre los valores de la serie (ver Royal Statistical Society 2002). Sea x_j una serie de n valores ($j = 1, 2, \dots, n$), el coeficiente de autocorrelación simple r_k con retardo k , típicamente se estima con

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \quad (13)$$

donde $\bar{x}$ es la media aritmética de la variable x en la serie. Es usual calcular r_k para varios valores de k , usualmente pequeños. Una prueba típica de independencia de las series es la de Box & Pierce (1970), que se basa en el siguiente estadístico:

$$Q(r) = n \sum_{k=1}^t r_k^2 \quad (14)$$

que asintóticamente tiene una distribución χ^2 con t grados de libertad, por lo que se rechaza la independencia de la serie si $Q(r) > \chi_{1-\alpha}^2$. En Ljung & Box (1978) se propone una modificación a la prueba de Box-Pierce, que tiene un mejor desempeño cuando se satisface el supuesto de normalidad. La propuesta se basa en el estadístico

$$L(r) = n(n+2) \sum_{k=1}^t (n-k)^{-1} r_k^2 \quad (15)$$

que para n grande se distribuye como χ_t^2 . El que estas pruebas requieran normalidad, muestras grandes y el hecho de aportar poca evidencia gráfica, son limitantes importantes para su aplicación práctica en juegos de lotería.

Una alternativa a las pruebas de independencia sería monitorear los resultados de los sorteos en tiempo real y no un año o dos después, cuando todo es historia. Es reconocido que las cartas de control son una de las mejores formas de monitorear procesos repetitivos (Gutiérrez & de la Vara 2009), por lo que dado que cada sorteo

de un juego de azar es una repetición en la realización de un proceso, se propone monitorear la aleatoriedad de los juegos de azar mediante cartas de control.

Una forma práctica e intuitiva de monitorear los resultados de un juego de números de d dígitos es viendo el dígito que resultó en el más reciente sorteo y analizar hace cuántos sorteos (ensayos) había aparecido tal dígito; de esta manera, si Y es el número de ensayos o sorteos que transcurrieron para que apareciera nuevamente tal dígito, entonces bajo el supuesto de aleatoriedad, Y tiene una distribución geométrica con parámetro $p = 0.1$, por lo que

$$f(y | p) = \Pr(Y = y | p) = (1 - p)^{y-1} p \quad (16)$$

con $y = 1, 2, \dots$

La distribución acumulada está dada por

$$F(y | p) = \Pr(Y < y | p) = p \sum_{i=1}^y (1 - p)^{i-1} = 1 - (1 - p)^y \quad (17)$$

El límite de control superior (LCS) e inferior (LCI) indican dónde se espera esté Y bajo el supuesto de aleatoriedad. Si se fija una probabilidad de error tipo I igual a γ , entonces estos límites se obtienen con $\Pr(Y > LCS | p) = \gamma/2$ y $\Pr(Y < LCI | p) = \gamma/2$. Sin embargo, con la distribución geométrica con $p = 0.1$ se obtiene que $P(Y = 1) = 0.1$, por lo que dado que usualmente los valores de γ son pequeños (el valor más comúnmente usado en control de procesos es $\gamma = 0.0027$), entonces no se debe usar LCI , por lo que $\Pr(Y > LCS | p) = \gamma$. De aquí que

$$\begin{aligned} \Pr(Y > LCS | p) &= 1 - \Pr(Y \leq LCS | p) \\ &= 1 - F(LCS + 1 | p) \\ &= (1 - p)^{LCS+1} = \gamma \end{aligned}$$

Despejando a LCS de esta última igualdad, se obtiene que el valor aproximado para el LCS está dado por

$$LCS = \frac{\ln(\gamma)}{\ln(1 - p)} - 1 \quad (18)$$

Nótese que para la aplicación específica de la carta es necesario partir de que $p = 0.1$; si no fuera así, entonces la incertidumbre sobre p la podría aportar una distribución Beta, y hacer inferencia con la distribución posterior correspondiente. En Gutiérrez (2006) se presenta una versión bayesiana para una carta de control bajo el supuesto anterior. La versión frecuentista de la carta geométrica y algunos detalles de su caracterización se pueden consultar en Yang, Xie, Kuralmani & Tsui (2002).

Si $\gamma = 0.0027$, entonces el límite de control superior LCS se puede tomar igual a 56, ya que $\Pr(Y > 56 | p = 0.1) = 0.002739$. Además de los límites de control es usual establecer otras zonas que ayudan a detectar otros patrones no aleatorios en el comportamiento del proceso (ver Gutiérrez & de la Vara 2009). Sin embargo,

en el caso de la carta de control geométrica no es posible utilizar las zonas típicas debido a la forma de la distribución geométrica, que tiene una forma exponencial donde acumula las probabilidades más altas a los números enteros más pequeños.

Como alternativa se propone generar cinco zonas con una cobertura aproximada de 0,20 en la probabilidad para los correspondientes valores de Y . Estas zonas se detallan en la tabla 1, y se han obtenido con el apoyo de la distribución acumulada (17). Como en el juego de números de d dígitos lo que se quiere garantizar es la aleatoriedad, entonces esto será equivalente a observar un proceso (juego) bajo control estadístico en una carta geométrica, en donde la proporción de puntos en las diferentes zonas a lo largo de la carta de control sea consistente con la probabilidad indicada en la tabla 1. La separación de las zonas de la tabla 1 puede darse con el valor medio que las separa, así $l_A = 2,5$, $l_B = 5,5$, $l_C = 9,5$, $l_D = 16,5$.

TABLA 1: Zonas en la carta de control geométrica.

Zona	Valores de Y	Probabilidad
A	$\{1, 2\}$	0,19000
B	$\{3, 4, 5\}$	0,21951
C	$\{6, 7, 8, 9\}$	0,20307
D	$\{10, \dots, 16\}$	0,20212
E	$\{Y \geq 17\}$	0,18530

De esta forma, para monitorear el comportamiento de Y en un juego de número de d dígitos, la observación se puede centrar en cada una de las urnas con las que usualmente se generan los dígitos del juego. Monitoreando en forma separada cada urna y llevando un registro (carta) particular para cada dígito -para ello cada vez que se realice un sorteo se ve qué dígito fue extraído de la urna bajo análisis-, se analiza cuántos sorteos fueron necesarios para que volviera a aparecer tal dígito, con lo que en cada sorteo se tiene el valor de Y . Para aplicar la carta desde el primer sorteo, se puede suponer que en el sorteo cero salieron todos los dígitos.

Si se viera un punto fuera del LCS , se deberá ver qué pasó, particularmente qué dígito i es, porque esto identificará la esfera que ha tardado demasiados sorteos en salir, es decir, esto sugerirá que $p_i < 0,1$. Al considerar que $\sum_{i=1}^{10} p_i = 1$, y si $p_i < 0,1$ para algún i , entonces esto implica en contraste que $p_j > 0,1$ para una o más $j = 1, \dots, 10$, con $j \neq i$. Esto, trasladado en la carta de control, se reflejará en que mientras los valores de Y para la esfera i , Y_i , tenderán a ser más grandes, los valores de Y_j tenderán a ser más pequeños. Ambas cosas se detectarán en las zonas de la carta de control. Esta estructura de dependencia plantea la necesidad de hacer un monitoreo simultáneo e individualizado para todas las esferas. Simultáneo en el sentido que se tenga claro que a quien se está monitoreando es a Y (número de ensayos o sorteos que transcurrieron para que saliera nuevamente el dígito que acaba de aparecer), pero individualizado porque se propone que el valor de Y se registre en la celda (carta) que le corresponde a cada dígito.

De esta manera en cada carta, para cada dígito, se debe vigilar visualmente que aproximadamente la quinta parte de los valores de Y caigan en cada una de

las zonas de la carta. Si hay sospechas de que esto no está ocurriendo para un dígito, porque por ejemplo hay una racha larga de puntos donde los valores de Y tienden a caer en las últimas zonas de la correspondiente carta de control; esto también se reflejará en forma opuesta en por lo menos otra zona de otro dígito, donde se tendrán muchos puntos en las primeras zonas de la carta de control. Identificado esto, para verificar formalmente las posibles rachas de puntos que siguen un patrón no aleatorio se tienen dos posibilidades. En la primera se parte de un modelo multinomial para el dígito i con los cinco sucesos dados por las zonas de la carta, es decir, un modelo multinomial con $\mathbf{p} = (p_A, p_B, p_C, p_D, p_E)$, donde se observan x_i veces la esfera i ; entonces la distribución marginal posterior para cualquiera de las p_k (con $k = A, \dots, E$), se obtiene a partir de (12), para precisar mejor que quede lo indicado como la distribución a priori no informativa, $\alpha_k = 1$. Específicamente la distribución es Beta($1 + w_k, 5 + x_i - (1 + w_k)$):

$$\pi(p_k | x_i, w_k) = \frac{\Gamma(5 + x_i)}{\Gamma(1 + w_k) \Gamma(5 + x_i - (1 + w_k))} p_k^{1+w_k-1} (1 - p_k)^{5+x_i-(1+w_k)-1} \quad (19)$$

donde w_k es el número de puntos que cayeron en la zona k de la carta de control para el dígito i . A partir de esta distribución marginal se puede calcular un intervalo de probabilidad para ver si éste incluye al correspondiente valor de p_k dado en la tabla 1. Adicionalmente se pueden combinar dos o más de las zonas de la tabla 1, y hacer los análisis correspondientes mediante (19).

El otro criterio para verificar rachas de valores de Y que no sigan un patrón aleatorio es la aplicación de la distribución binomial. Supóngase que hay una racha sospechosa donde h de t puntos consecutivos cayeron en la zona k de la carta geométrica, considerando que la zona k puede ser alguna de las indicadas en la tabla 1, o bien una combinación de algunas zonas de esta tabla u otras que reúnan algún otro criterio. Entonces es claro que bajo el supuesto de aleatoriedad, la probabilidad de que tal racha ocurra está dada por la distribución binomial (t, p_k) , donde p_k es la probabilidad de la zona k construida como se ha indicado. Por tanto

$$\Pr(h | t) = \binom{t}{h} p_k^h (1 - p_k)^{t-h} \quad (20)$$

Por ejemplo, supóngase que en la zona A de la carta de control geométrica para algún dígito caen tres puntos consecutivos, lo que significa que tal dígito salió como resultado en tres sorteos consecutivos o casi consecutivos. Para calcular la probabilidad de esto, se tiene que $h = 3$, $t = 3$ y $p_k = 0,19$; y de acuerdo con (20), la probabilidad de tal racha es igual a 0,006859. Esta probabilidad es suficientemente baja como para pasar por alto la situación; por tanto lo que se sugiere hacer es evaluar de cerca lo que está pasando con el proceso de aleatorización, como por ejemplo se podrían simular sorteos para corroborar la sospecha con más datos. En la sección 5, donde se analizan datos de un caso práctico, se aplica (20) para varias rachas sospechosas.

Para analizar el funcionamiento de la carta propuesta se hace el siguiente estudio, y además más adelante se verá el caso práctico del sorteo Tris.

4. Estudio Monte Carlo

Para analizar y visualizar el funcionamiento de la carta geométrica propuesta en la sección previa se hizo un estudio Monte Carlo. Las rutinas fueron programadas con S-Plus. Se buscó simular lo que ocurre en una urna que contiene diez esferas numeradas con los dígitos, y en cada sorteo se extrae una. Se asignó una probabilidad de extracción p_i para cada esfera y se simularon 300 mil repeticiones del sorteo. Además de cuantificar la proporción de veces que resultó cada esfera, se registró el número de sorteos que tuvieron que pasar para que volviera a salir el mismo resultado, es decir se cuantificó Y , junto con la proporción de veces, respecto al total de sorteos en que el valor de Y fue superior al límite de control superior ($LCS = 56$). Además, para cada esfera se registró la proporción de veces en que el valor de Y cayó en cada una de las zonas de la carta para cada dígito (tabla 1). En la tabla 2 se reportan los resultados para dos casos. En el primero se supuso que dos esferas, la 3 y 4, tenían probabilidades de ser extraídas de $p_3 = 0,07$ y $p_4 = 0,13$, y a las restantes se les asignó una probabilidad de 0,1. Se observa que la proporción en la que cada esfera fue el resultado del sorteo fue muy similar a su correspondiente probabilidad; por ejemplo, el 7,0 % de las ocasiones la esfera 3 fue el resultado del sorteo, y la cuatro un 13,0 % de las veces. Por otro lado, en cuanto a los valores de Y en las zonas A a E para cada esfera, se ve que en el caso de la esfera 3, las proporciones para las primeras tres zonas fue de 0,138, 0,168 y 0,172, que son visiblemente menores a las esperadas, de alrededor de 0,20; en contraste, en la última zona, la E , tiene una proporción de 0,313, contra una esperada de 0.185. En otras palabras la esfera tres tardaba demasiados sorteos en salir. En cuanto a la esfera 4, los resultados son opuestos: más veces cayó Y en las primeras zonas de la carta y pocas en las últimas; por ejemplo, en la zona E la proporción fue de 0,107, que es casi la mitad de la esperada bajo aleatoriedad (0,185). En los restantes dígitos los resultados en las diferentes zonas son muy similares a los esperados bajo aleatoriedad.

Para el segundo caso de la tabla 2, se asignó a la esfera 2 una probabilidad de ser extraída de $p_2 = 0,145$, y este aumento respecto a 0,10 se quitó de forma equitativa al resto de las esferas, con lo que sus probabilidades de extracción fueron de 0,095. Se observa que la proporción con la que salió cada esfera es muy similar a la probabilidad asignada, por ejemplo, un 14,5 % de las ocasiones la esfera 2 fue el resultado del sorteo, mientras que las proporciones de los valores de Y para tal esfera que caen en las zonas A y B son de 0,271 y 0,270 respectivamente, ambas muy por arriba de 0,20. Estos aumentos contrastan con los datos observados para la zona E , a la que le corresponden sólo el 8,0 % de los valores de Y para la esfera 2. Respecto del resto de las esferas, se nota una ligera disminución de la proporción de valores de Y para la zona A .

Con los resultados de las simulaciones de la tabla 2 se ilustra que un incremento en el porcentaje de extracciones de un dígito en particular se puede detectar, además de en el porcentaje mismo, en la carta de control mediante un incremento de la proporción de valores de Y que caen en las primeras zonas de la carta y una baja proporción de valores de Y en las últimas zonas, principalmente en la zona E , mientras que cuando se tienen proporciones de extracción menores a 0,1, se

TABLA 2: Resultados de estudio de simulación para dos casos seleccionados.

		Resultado del sorteo	Puntos fuera del <i>LCS</i>	Proporción de puntos en cada zona				
				<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
Prob. teórica $p_i = 0,1$			0,100	0,00027	0,190	0,220	0,203	0,202
Resultados para los dígitos, con $p_3 = 0,07$, $p_4 = 0,13$, y los restantes igual a 0,1	1	0,100	0,00027	0,188	0,221	0,203	0,206	0,183
	2	0,100	0,00024	0,192	0,219	0,203	0,201	0,186
	3	0,070	0,00119	0,138	0,168	0,172	0,209	0,313
	4	0,130	0,00006	0,245	0,255	0,214	0,179	0,107
	5	0,100	0,00025	0,191	0,217	0,204	0,202	0,185
	6	0,100	0,00032	0,191	0,220	0,200	0,202	0,186
	7	0,100	0,00035	0,193	0,218	0,203	0,203	0,183
	8	0,099	0,00028	0,191	0,221	0,201	0,196	0,191
	9	0,101	0,00026	0,190	0,219	0,206	0,202	0,183
	0	0,100	0,00032	0,189	0,219	0,199	0,208	0,185
Resultados para los dígitos, con $p_2 = 0,145$ y los restantes igual a 0,095	1	0,095	0,00035	0,182	0,211	0,197	0,206	0,204
	2	0,145	0,00001	0,271	0,270	0,216	0,163	0,080
	3	0,095	0,00033	0,184	0,211	0,198	0,206	0,201
	4	0,095	0,00034	0,180	0,216	0,199	0,202	0,203
	5	0,095	0,00037	0,181	0,215	0,197	0,206	0,201
	6	0,095	0,00038	0,183	0,215	0,196	0,208	0,198
	7	0,095	0,00037	0,184	0,208	0,203	0,204	0,201
	8	0,094	0,00036	0,179	0,215	0,197	0,203	0,206
	9	0,095	0,00034	0,182	0,214	0,196	0,203	0,204
	0	0,095	0,00038	0,197	0,207	0,202	0,207	0,202

detecta en la carta, con puntos fuera del *LCS*, mayores proporciones de valores de Y que caen en las zonas D y E , y una menor proporción para las primeras zonas de la carta. Las sospechas deben ser validadas con las pruebas descritas en la subsección 2.2.

En relación con el uso del *LCS* para detectar cambios, primero es importante puntualizar la probabilidad del error tipo I, que consiste en declarar que ha habido un cambio cuando en realidad no fue así, por lo que el correspondiente punto fuera del *LCS* será una falsa alarma. En la tabla 2, se ha estimado la proporción de falsas alarmas debido a puntos fuera del *LCS*. Se ha hecho respecto al total de sorteos, porque al final de cuentas se está monitoreando sólo un proceso, y los resultados se reportan de forma separada para cada dígito como una estrategia para separar fuentes de variabilidad atribuible y así lograr un mejor monitoreo del proceso. Ante esto, el *LCS* se ha establecido con $\gamma = 0,002739$, como es usual en cartas de control, de tal forma que en condiciones de control estadístico (aleatoriedad) se esperaría que entre todos los dígitos sólo se tuviera esa proporción de falsas alarmas. Si a 0,002739 se le divide entre diez, entonces se esperaría que cada dígito aporte 0,00027 puntos fuera del *LCS*. En la tabla 2 se puede ver que en las esferas que tuvieron $p_i = 0,1$, su proporción de puntos fuera está muy cerca de lo esperado, mientras que para las esferas en que $p_i < 0,1$ se tienen más puntos fuera, porque tarda más la esfera en salir como resultado; por ejemplo, para la esfera 3 en el caso 1, se tiene que la proporción de puntos fuera fue de 0,00119, en cambio, si $p_i > 0,1$, la proporción de puntos fuera es mucho menor, como se ve en la esfera 4 del caso 1, donde tal proporción fue de 0,00006. Por tal razón el tipo de

problemas de falta de aleatoriedad que se detectan con el LCS son aquellos donde $p_i < 0,1$, y no así cuando $p_i > 0,1$, esto último debido a que la carta geométrica no tiene LCI .

Con lo que se ha dicho respecto al error tipo I, queda claro que la probabilidad de este error γ , que es con la que se calcula el LCS , se mantendrá respecto al total de sorteos bajo el supuesto de aleatoriedad, a pesar de que se lleve el registro por separado de los diferentes dígitos. Además, dadas las limitaciones del LCS para detectar cambios, es necesario el uso de las cartas para cada dígito junto con las zonas propuestas en la tabla 1. Esto incrementa la potencia para detectar problemas mediante la identificación de rachas o patrones con sospecha de falta de aleatoriedad. Pero para mantener en un nivel bajo la posibilidad de declarar cambios cuando no los ha habido, error tipo I global, entonces cuando haya sospechas de una racha no aleatoria en una carta de control para un dígito, esto deberá ser corroborado con criterios probabilísticos, como los sugeridos en la sección anterior, con las expresiones (12), (19) y (20).

5. Aplicación al sorteo Tris de México

El sorteo Tris es un juego de d dígitos de la empresa Pronósticos para la Asistencia Pública del gobierno mexicano. Inicialmente el participante seleccionaba tres dígitos, posteriormente fue cambiado para que la selección fuera de cuatro dígitos, y en el año 2007 se transformó en un juego de cinco dígitos, con lo cual el participante forma un número entre 00000 y 99999. El sorteo se realiza en forma pública, y para extraer las esferas que definen el número premiado se utiliza un dispositivo electromecánico (urna). Los números ganadores del sorteo Tris están disponibles en la página www.pronosticos.gob.mx. Para este trabajo se tomaron los resultados del dígito que formaba las unidades de millar, a la que se identifica como la urna 4, para un total de 500 sorteos Tris entre el 20/07/1996 (sorteo 2913) y el 28/03/1997 (sorteo 3412). La frecuencia resultante para cada uno de los dígitos en estos 500 sorteos en la urna 4 se muestra en la tabla 3.

TABLA 3: Resultados de 500 sorteos Tris para la urna 4 (unidades de millar).

Dígito	1	2	3	4	5	6	7	8	9	0
Frecuencia	48	56	45	52	49	57	56	43	60	34
Proporción	0,096	0,112	0,09	0,104	0,098	0,114	0,112	0,086	0,12	0,068

Es evidente que bajo la hipótesis de selección aleatoria de las esferas, cada dígito tiene una probabilidad teórica de ser seleccionado de 0,1. Sin embargo, a primera vista los resultados de la tabla 3 parecen contradecir esta hipótesis, debido particularmente a los dígitos cero y nueve. Este último salió casi el doble de veces que el primero. Para confirmar estas sospechas se procede a probar la hipótesis (2) con $m = 10$ y $p_{i,0} = 0,1$, mediante las diferentes pruebas aproximadas descritas en la subsección 2.1. Para empezar, la prueba usual basada en el estadístico (3),

donde se tiene que

$$\chi_0^2 = \sum_{i=1}^{10} \frac{(x_i - 50)^2}{50} = 11,2 \quad (21)$$

como el valor crítico es $\chi_{0,05,9}^2 = 16,919$, no se rechaza H_0 . Este no rechazo se ve también con el valor- p para (21), que es de 0,2622. Así, esta prueba no aporta evidencia significativa en contra de la aleatoriedad de los resultados sintetizados en la tabla 3. Algo similar ocurre con la prueba de la razón de verosimilitud del estadístico (4), ya que el valor de $R = 11,703$, lleva a no rechazar H_0 , puesto que el valor- p correspondiente es igual a 0,2306. Por su parte, con la prueba de Nass basada en el estadístico (7), se tiene que $\chi_c^2 = 11,18$, y dado que los grados de libertad para la distribución χ^2 de referencia son $v = 9,02$, entonces esta prueba tampoco rechaza H_0 . En cuanto a los intervalos simultáneos para las diez p_i , basados en (8), con $\alpha = 0,05$, todos incluyen al 0,1; por ejemplo los intervalos para p_1, p_9 y p_{10} están dados, respectivamente, por $[0,065, 0,139]$, $[0,085, 0,167]$ y $[0,043, 0,107]$, con p_{10} la probabilidad para el dígito cero. Como todos los intervalos incluyen a 0,1, no se rechaza (2) con esta metodología. Es de señalar que el intervalo para el dígito cero es el que estuvo más cercano de no incluir al 0,1.

Para aplicar el procedimiento bayesiano descrito en la subsección 2.2, se tiene que $n = \sum_{i=1}^{10} x_i = 500$ y $\alpha_0^* = 5 + 500$. Un primer asunto de interés es explorar la distribución conjunta posterior de la probabilidad del dígito cero, p_{10} , y la del nueve, p_9 . Para ello se hizo una corrida de dos mil simulaciones de la distribución posterior de $\mathbf{p}$, utilizando el MCMCPack de R, como se indicó en la subsección 2.2. En la parte superior de la figura 1 se ve una muestra de la distribución conjunta posterior de (p_{10}, p_9) , de donde se aprecia la poca posibilidad que se le da a que $p_{10} \geq 0,1$; de hecho, de los dos mil puntos sólo 14 están a la derecha de la línea vertical $p_{10} = 0,1$, lo que en proporción significa 0,007. Estas evidencias refuerzan el hecho de que $p_{10} < 0,1$. Por su parte, la posibilidad de que $p_9 > 0,1$ es un poco mayor, ya que de los dos mil puntos 165 están abajo de la línea horizontal $p_9 = 0,1$, lo que significa una proporción de 0,0825.

Una conclusión similar a la anterior se llega al observar la parte baja de la figura 1 que muestra la gráfica de contornos con diez cortes de la distribución posterior de $\mathbf{p}_m = (p_{10}, p_9, 1 - p_{10} - p_9)$, que de acuerdo con lo establecido en la subsección 2.2, es $\text{Dirichlet}(\mathbf{p}_m, \alpha_m)$, con $\alpha_m = (1 + 34, 1 + 60, 510 - 2 - 34 - 60)$. En esta figura se ve la poca posibilidad que se le da a la región $p_{10} \geq 0,1$.

De la misma manera se puede obtener la distribución posterior marginal (12) para cualquiera de los dígitos, que según lo establecido en la subsección 2.2 es una distribución $\text{Beta}(1 + x_i, 510 - 1 - x_i)$. Como es de particular interés hacer inferencia sobre los dígitos nueve y cero, de la tabla 3 se ve que éstos tuvieron una frecuencia de $x_9 = 60$ y $x_{10} = 34$, respectivamente. En la figura 2 se muestra la distribución posterior para estos dos casos, de donde es claro que en el caso del dígito cero se debe rechazar la idea de que $p_{10} = 0,1$, como requiere el supuesto de aleatoriedad. Esto se respalda si se calculan intervalos de probabilidad para p_{10} con los cuantiles $q_{\gamma/2}$ y $q_{1-\gamma/2}$ de la distribución $\pi(p_{10} | X)$. Tomando $\gamma = 0,05$, se obtiene que estos intervalos están dados por $[0,0929, 0,1491]$ y $[0,0484, 0,0921]$, para p_9 y p_{10} , respectivamente.

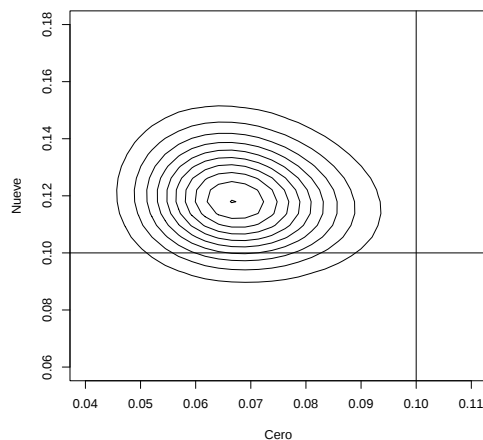


FIGURA 1: Distribución posterior conjunta para los dígitos cero y nueve.

Como en el caso del dígito nueve, el valor $p_9 = 0,1$ está dentro del correspondiente intervalo, en este caso no se alcanza a rechazar la hipótesis de aleatoriedad, cosa que si ocurre en el caso del dígito cero, ya que $p_{10} = 0,1$ está fuera del correspondiente intervalo de probabilidad. De hecho la probabilidad de ver este evento o uno peor respecto a aleatoriedad, se puede obtener calculando el área en la cola de la distribución marginal posterior, es decir calculando $\Pr(p_{10} > 0,1 \mid X)$ que es igual a 0,0056. Esto muestra que se tiene una fuerte evidencia contra la aleatoriedad del sorteo Tris en el período analizando. Mientras que en relación al dígito nueve se tiene que $\Pr(p_9 < 0,1 \mid X) = 0,0807$, lo que da una cierta evidencia en contra de la aleatoriedad, aunque no concluyente como ya se había puntualizado en la figura 1.

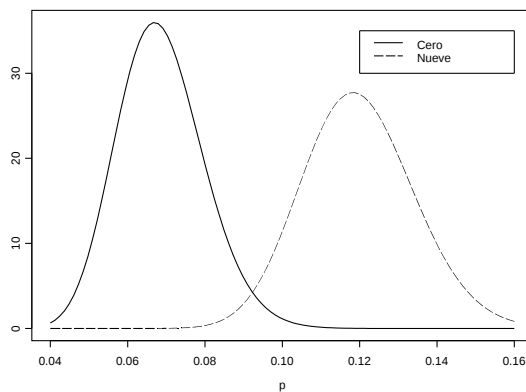


FIGURA 2: Distribución marginal posterior para los dígitos cero y nueve.

Si se quisiera un estimador puntual para p_9 y p_{10} , en lugar de los estimadores de intervalo que se han dado antes, se puede calcular la moda de las correspondientes distribuciones marginales posteriores. Estas modas están dadas por $\hat{p}_9 = 0,1181$ y $\hat{p}_{10} = 0,0669$.

Por lo anterior, contrario a las pruebas tradicionales basadas en aproximar diferentes estadísticos con la distribución χ^2 , el procedimiento bayesiano sí detecta que hay problemas en la aleatoriedad de los resultados del sorteo Tris, como intuitivamente se esperaba a partir de la tabla 2, con la ventaja adicional de que se aporta evidencia de la distribución posterior, conjunta o marginal, para los p_i con problemas.

Para probar la independencia de los resultados del sorteo Tris, con los datos de la urna del millar de los 500 sorteos se calculó el coeficiente de autocorrelación r_k , para $k = 1, \dots, 24$, y se obtuvo que el estadístico Box-Pierce de la expresión (14) toma el siguiente valor:

$$Q = 500 \sum_{i=1}^{24} r_i^2 = 36,74$$

que tiene un valor- $p = 0,0464$, por lo que con una significancia de $\alpha = 0,05$ se rechaza la independencia de los resultados de la serie. Al aplicar la prueba Ljung-Box con el estadístico (14), se obtiene que $L(r) = 38,00$, que tiene un valor- $p = 0,0346704$, por lo que también se rechaza la independencia de la serie, aunque la evidencia gráfica de la falta de independencia, no reportada en este trabajo, para toda la serie es poco clara.

Con el fin de aportar una mejor evidencia gráfica y monitorear el comportamiento a través del tiempo de los resultados del sorteo Tris, primero, con base en cada dato de la serie original se calculó Y , es decir se identificó qué dígito fue el resultado y se contó cuántos sorteos fueron necesarios para que volviera a aparecer tal dígito. Para aplicar la carta desde el primer sorteo, se supuso que en el sorteo cero salieron todos los dígitos. En la tabla 4 se muestra un extracto de los valores de Y .

TABLA 4: Muestra de los valores de Y para los primeros 20 sorteos de Tris.

Sorteo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dígito	9	9	5	2	3	5	7	1	8	9	9	0	2	5	3	0	2	4	6	1
Y	1	1	3	4	5	3	7	8	9	8	1	12	9	8	10	4	4	18	19	12

Con los valores de Y para cada dígito se obtuvieron las correspondientes cartas de control geométricas. Por ejemplo en la figura 3 se muestran las cartas de control para los dígitos nueve y cero, que ya se vio tienen problemas de cumplimiento del supuesto de aleatoriedad, por lo que es de interés ver cómo lo reflejan las cartas de control. En el caso del dígito cero, el punto 26 con valor de $Y = 60$ está fuera del LCS , lo que indica que tuvieron que pasar 60 sorteos para que volviera a salir la esfera del dígito cero en la urna del millar, lo que no es congruente bajo el supuesto de aleatoriedad en el sorteo. Además, visualmente parece haber demasiados puntos en la parte alta de la zona E de la carta. En particular parece sospechoso el que

9 de 34 valores de Y sean mayores que 22. Para corroborar esta sospecha (racha) mediante la aplicación de (20), se tiene que $h = 9$, $t = 34$, y como la zona bajo sospecha en la carta está dada por aquellos puntos de Y que son mayores que 22, entonces para calcular la probabilidad de que un punto caiga en esta zona se puede usar la distribución geométrica acumulada $F(y)$ de la expresión (17), con $p = 0,1$; por tanto $p_k = \Pr(Y \geq 23) = 1 - F(23 | p = 0,1) = 0,0886$, entonces aplicando (20) se obtiene que la probabilidad de la racha referida es igual a 0,00174, lo que es una probabilidad muy baja. Esto confirma la idea de que $p_{10} < 0,1$ para el dígito cero. Aquí la carta de control hubiese resultado un buen instrumento de monitoreo del sorteo, ya que hubiera avisado que con frecuencia pasaban demasiados sorteos en donde la esfera cero no salía como resultado.

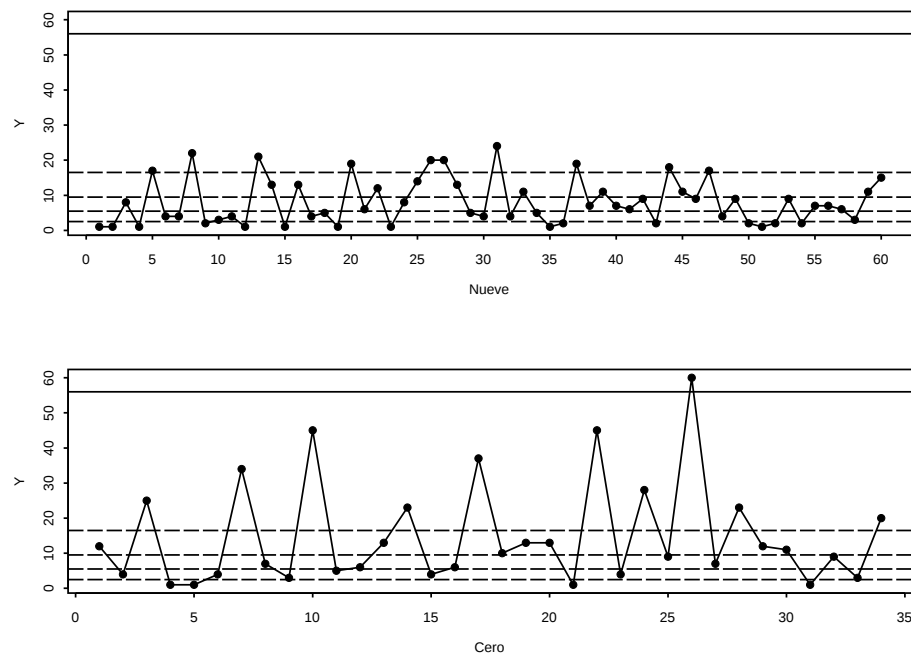


FIGURA 3: Cartas geométricas para los dígitos cero y nueve del sorteo Tris.

En la parte baja de la figura 3 se muestra la carta de control para el dígito nueve, que presentó una situación opuesta a la del dígito cero, ya que los valores de Y tienden a ser más pequeños; por ejemplo, hubo nueve veces en que $Y = 1$, lo que implica que en nueve ocasiones hubo dos sorteos consecutivos donde se tuvo como resultado al nueve. Además, de las 60 veces que salió el número 9, en todas $Y < 25$, y como de acuerdo con (17) se tiene que $P(Y < 25) = F(25 | p = 0,1) = 0,9282$, al aplicar (20) con $h = 60$, $t = 60$ y $p_k = 0,9282$, se tiene que la probabilidad de que en los sesenta sorteos ocurra que $Y < 25$ es igual a 0,011, lo que es una baja probabilidad. Lo anterior se suma a la sospecha de que en el período analizado, el

número nueve salía con más frecuencia de lo esperado. Esto confirma la utilidad de la carta, y en el caso del dígito nueve se tiene que $p_9 > 0,1$.

Igualmente se analizó cuántos valores de Y cayeron en cada zona de la correspondiente carta de control para cada dígito. Los resultados de esto se muestran en la tabla 5. Como por diseño la frecuencia con la que los valores de Y caen en cada zona para cada dígito es más o menos similar, resaltan los casos de las esferas 3 y 8 debido a que las primeras zonas de sus respectivas cartas de control incluyen una cantidad pequeña de puntos. Para evaluar con mayor detalle esto, se utiliza (19). Si se parte del supuesto de que $p = 0,1$ y si se junta la información de las zonas A y B , se tiene que $p_{AB} = p_A + p_B = 0,41$ (ver tabla 1); entonces en el caso del dígito 3 se tendrá que $x_i = 45$ y $w_k = 5 + 6$, y por lo que la distribución posterior de p_{AB} es Beta con parámetros (12, 37). Al calcular con esta distribución un intervalo de probabilidad para p_{AB} con una cobertura de 0,95, se obtiene que el intervalo es $[0,1364, 0,3731]$, que no incluye a $p_A + p_B = 0,41$, lo que es una evidencia contra el supuesto de que $p_3 = 0,1$ para la esfera 3; esto se ratifica si se calcula $\Pr(p_{AB} > 0.41 \mid x_i, w_k) = 0,0068$, lo que sugiere que a través del tiempo Y para la esfera 3 no siguió una distribución geométrica con $p_3 = 0,1$, y que más bien parece que $p_3 < 0,1$ por lo menos por un período largo, lo que es una evidencia más contra la aleatoriedad del sorteo Tris.

En la parte superior de la figura 4 se muestra la carta de control para el dígito 3, y la gráfica sugiere dos periodos: el primero con $p_3 > 0,1$, que va del punto 1 hasta aproximadamente el 20, donde los valores de Y tienden a que Y tenga valores pequeños; por ejemplo del punto 14 al 18, se da una racha donde 4 de 5 puntos consecutivos caen en la zona A de la carta; luego aplicando (20) con $h = 4$, $t = 5$ y $p_k = 0,19$, se tiene que bajo el supuesto de aleatoriedad la probabilidad de tal racha es de apenas 0,005. Por el contrario, en la segunda parte de la carta parece que los valores de Y tienden a ser más grandes; por ejemplo, en esta segunda parte se da una racha desde el punto 19 hasta el 45 donde ningún valor de Y cayó en la zona A , lo que sugiere que en esta segunda parte de la carta más bien $p_3 < 0,1$. Al mezclarse estos dos periodos hacen que la frecuencia total con la que resultó la esfera 3 no suscite ninguna sospecha de problemas de aleatoriedad. Lo que ha ocurrido con la esfera 3 muestra la utilidad de la carta de control geométrica propuesta para monitorear la aleatoriedad de este tipo de juegos.

En la carta de control para el dígito 8, que no se presenta aquí y que es otro caso donde hay pocos valores de Y en las zonas A y B , ocurre algo relativamente similar a lo del dígito 3: en los primeros 19 puntos de la carta, 18 caen fuera de las zonas A y B , lo que bajo el supuesto de aleatoriedad tiene una probabilidad muy baja de ocurrir de apenas 0,0006. En cambio, en la segunda parte de la carta los valores de Y fluctúan de la manera esperada.

En la parte inferior de la figura 4 se muestra la carta de control para el dígito 6, que presenta un caso opuesto al dígito 3, ya que hay pocos valores de Y en las zonas D y E , lo que indica que la esfera 6 tendía a salir demasiado pronto, por lo que pareciera que $p_6 > 0,1$ en algunos periodos. Por ejemplo del punto 15 al 38 se presentó una racha donde de 24 puntos sólo uno cayó en la zona E , lo que tiene una probabilidad de ocurrir de 0,04. Aunque no es una evidencia contundente,

sí apoya un tanto que $p_6 > 0,1$ durante ese periodo. En apoyo a esto se destaca visualmente que tres puntos consecutivos de Y cayeron en la zona A (del 10 al 12), y la probabilidad de que esto ocurra bajo el supuesto de aleatoriedad es de $(p_A)^3 = (0,19)^3 = 0,00686$, que se puede considerar baja.

TABLA 5: Frecuencia de los valores de Y por zona de la carta para el sorteo Tris.

Dígito	Zona					Puntos fuera	
	A	B	C	D	E	Total	del LCS
1	9	10	10	8	11	48	0
2	11	10	16	11	8	56	0
3	5	6	13	11	10	45	0
4	13	11	11	7	10	52	0
5	8	9	12	13	7	49	0
6	10	14	16	8	9	57	0
7	12	13	9	17	5	56	0
8	5	6	9	15	8	43	0
9	15	12	13	10	10	60	0
0	4	7	6	7	10	34	1

En suma, más allá del posible incremento del error tipo I al rechazar la hipótesis de aleatoriedad en el proceso de extracción de las esferas de la urna del millar en el sorteo Tris al evaluar la aleatoriedad de diferentes rachas, lo que importa mostrar con las cartas de control de la figura 4 es que para probar la aleatoriedad de juegos de d dígitos no basta probar la igualdad de la frecuencia con la que los diferentes dígitos aparecen como resultado, como se hizo en el caso del sorteo Tris para el dígito 0 y 9 a través de las figuras 1 a 3, sino además es fundamental monitorear la rapidez con la que los diferentes dígitos aparecen, y así poder identificar rachas no aleatorias con apoyo de las cinco zonas de la carta geométrica para cada dígito, o que permitirá actuar con mayor oportunidad y de manera más preventiva. Pero como ya se ha dicho, es importante evaluar formalmente la aleatoriedad de las rachas sospechosas, para lo cual se han propuesto las opciones que representan las expresiones (19) y (20).

6. Discusión y conclusiones

Verificar la aleatoriedad y legalidad de los resultados de los juegos de azar es un asunto que toma cada día mayor importancia social, dada la existencia de estos juegos y sorteos tanto en formato tradicional como en formato electrónico o en internet. Por ello se justifica profundizar en metodologías estadísticas que ayuden a verificar tal aleatoriedad. En la revisión bibliográfica hecha para este trabajo se destaca que los juegos de números del tipo k de N han recibido más atención por parte de los estadísticos, aunque varias de las pruebas propuestas en la literatura están basadas en el estadístico χ^2 , que requiere tamaños de muestra grandes para lograr buenas aproximaciones, por lo que sería deseable repensar procedimientos que no dependen de aproximaciones asintóticas, como las opciones

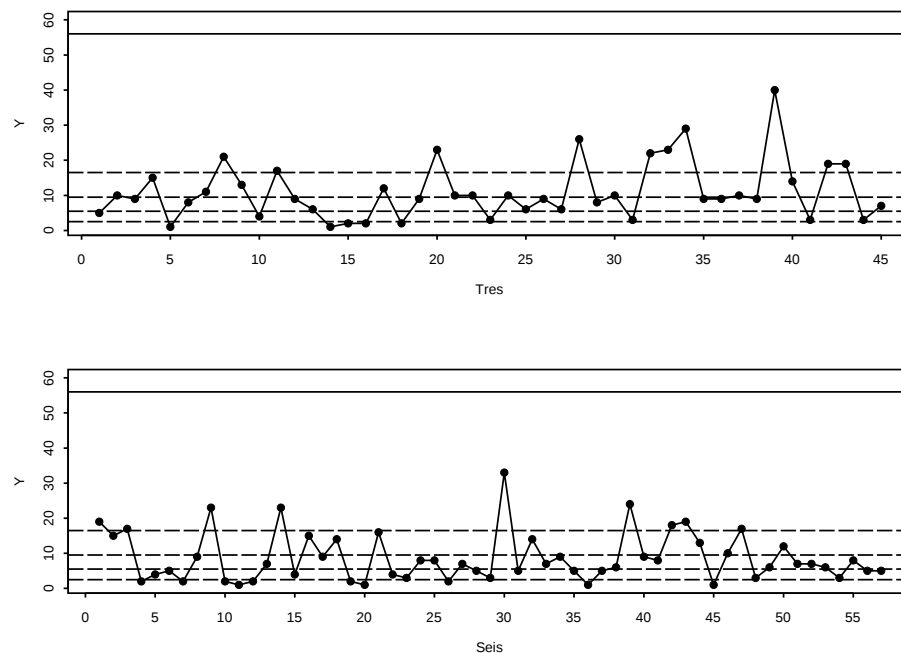


FIGURA 4: Cartas geométricas para los dígitos 3 y 6 del sorteo Tris.

bayesianas. Además es importante pensar en herramientas para monitorear en forma cotidiana este tipo de sorteos.

Por su parte, los juegos de números de d dígitos, preocupación central de este trabajo, han recibido menos atención, pero también requieren mejores formas de verificar su aleatoriedad, ya que la prueba χ^2 , que es el procedimiento estándar para verificar la igualdad en la frecuencia con la que aparece como resultado cada dígito, depende de aproximaciones asintóticas y precisa tamaños de muestra grandes. Por ello, como alternativa se propuso en este artículo utilizar el procedimiento bayesiano basado en un modelo multinomial, lo que permite hacer inferencia con base en la distribución posterior $\pi(\mathbf{p} \mid X)$, que es fácilmente obtenida por simulación. Además se pueden obtener distribuciones posteriores marginales de una o varias proporciones que se quieran analizar con mayor detalle. Esta alternativa, aparte de no requerir aproximaciones asintóticas, es un procedimiento sencillo e intuitivo que permite generar inferencia sobre el valor de p_i .

El caso práctico analizado en este trabajo sobre los resultados del sorteo Tris de México ha mostrado claramente lo anterior, ya que a pesar de que la frecuencia de los diferentes dígitos parecía no ser igual (ver tabla 3), ninguna de las pruebas tradicionales de la subsección 2.2 fue capaz de evidenciar la falta de aleatoriedad, cosa que sí se logra con base en $\pi(\mathbf{p} \mid X)$ (ver figuras 1 y 2).

Para verificar la aleatoriedad de un juego de números de d dígitos, también es necesario verificar la aleatoriedad a través del tiempo. En la literatura se encontró que esto se hace con base en pruebas de independencia entre ensayos sucesivos, y que un procedimiento utilizado se basa en un estadístico χ^2 obtenido a partir de los coeficiente de autocorrelación (ver Royal Statistical Society 2002) y Box & Pierce (1970). Sin embargo, estas pruebas dependen de supuestos distribucionales y de resultados asintóticos que usualmente requieren tamaño de muestra grande. Además, cuando estas pruebas detectan falta de independencia aportan pocos elementos para diagnosticar el origen del problema. Por ello, y como alternativa se ha propuesto en este trabajo el uso de la carta de control geométrica con cinco zonas para monitorear la aleatoriedad de los juegos de número de d dígitos, ya que esto ratifica que no sólo es importante verificar la igualdad de la frecuencia de los dígitos, sino analizar la rapidez con la que aparece en los diferentes dígitos. La carta de control es intuitiva, no requiere aproximaciones asintóticas ni supuestos distribucionales adicionales; además permite actuar oportunamente cuando se detecten resultados o patrones que evidencien la falta de aleatoriedad.

De esta manera, el hecho de monitorear la variable Y , que es igual al número de sorteos para que vuelva a aparecer como resultado un mismo dígito, y llevar un registro separado para cada dígito permite detectar evidencias del desempeño del proceso, e identificar específicamente cuál es la esfera (dígito) que está teniendo problemas. Para facilitar la interpretación de la carta geométrica e incrementar la potencia para detectar rachas de puntos que se desvíen de la aleatoriedad, se propuso dividir la carta en cinco zonas con una cobertura aproximada de 0,20 cada una (ver tabla 1).

La aplicación de la carta geométrica propuesta a los datos del sorteo Tris aportó evidencia adicional contra la falta de aleatoriedad de los dígitos 0 y 9. Además se detectaron otros problemas para otro par de dígitos (el 3 y el 6), que aunque en sus frecuencias totales no parecía que tuvieran problemas, la velocidad con la que aparecieron mostró rachas, donde en un periodo aparecían más rápido y en el otro fue más lento. Esto se validó con las pruebas estadísticas propuestas en las expresiones (19) y (20).

De esta manera, la carta geométrica propuesta resulta un instrumento útil para monitorear la aleatoriedad de juegos de azar del tipo d dígitos, tanto por sus propiedades teóricas como por su sencillez y sentido práctico.

En cierto sentido, los problemas detectados en la aleatoriedad de los resultados del sorteo Tris sorprenden porque la empresa responsable de este sorteo es una empresa pública que goza de un buen prestigio en México. Pero es claro que requiere mejores formas de asegurar la aleatoriedad de sus sorteos, como también lo concluye Coronel-Brizio et al. (2008), al detectar problemas de aleatoriedad en el sorteo Melate de esta misma empresa, que es un juego del tipo k/N .

[Recibido: julio de 2009 — Aceptado: septiembre de 2010]

Referencias

- Bernardo, J. M. & Smith, A. F. M. (1994), *Bayesian Theory*, John Wiley, Chichester.
- Box, G. E. P. & Pierce, D. A. (1970), 'Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models', *Journal of the American Statistical Association* **65**, 1509–1526.
- Cai, Y. & Krishnamoorthy, K. (2006), 'Exact Size and Power Properties of Five Tests for Multinomial Proportions', *Communications in Statistics-Simulation and Computation* **35**, 149–160.
- Clotfelter, C. T. & Cook, P. J. (2008), 'The "Gambler's Fallacy" in Lottery Play', *Management Science* **39**(12), 1521–1525.
- Coronel-Brizio, H., Hernández-Montoya, A., Rapallo, F. & Scalas, E. (2008), 'Statistical Auditing and Randomness Test of Lotto k/N -Type Games', *Physica A* **387**(25), 6385–6390.
- Finkelstein, M. (1995), 'Estimating the Frequency Distribution of the Numbers bet on the California Lottery', *Applied Mathematics and Computation* **69**(2-3), 195–207.
- Genest, C., Lockhart, R. A. & Stephens, M. A. (2002), ' χ^2 and the lottery', *Journal of the Royal Statistical Society: Series D (The Statistician)* **51**(2), 243–257.
- Goodman, L. A. (1965), 'On Simultaneous Confidence Intervals for Multinomial Proportions', *Technometrics* **7**, 247–254.
- Gutiérrez, H. (2006), 'Cartas de control bayesianas para atributos y el tamaño de subgrupo grande en la carta p ', *Revista Colombiana de Estadística* **29**(2), 163–180.
- Gutiérrez, H. & de la Vara, R. (2009), *Control estadístico de calidad y seis sigma*, segunda edn, McGraw-Hill, México.
- Hou, C. D., Chiang, J. & Tai, J. J. (2003), 'A Family of Simultaneous Confidence Intervals for Multinomial Proportions', *Computational Statistics & Data Analysis* **43**, 29–45.
- Joe, H. (1987), 'An Ordering of Dependence for Distribution of k -tuples, with Applications to Lotto Games', *The Canadian Journal of Statistics* **15**(3), 227–238.
- Joe, H. (1990), 'A Winning Strategy for Lotto Games?', *The Canadian Journal of Statistics* **18**(3), 233–244.
- Joe, H. (1993), 'Tests of Uniformity for Sets of Lotto Numbers', *Statistics & Probability Letters* **16**(3), 181–188.

- Johnson, R. & Klotz, J. (1993), 'Estimating hot Numbers and Testing Uniformity for the Lotto', *Journal of the American Statistical Association* **88**(422), 662–668.
- Koning, R. H. & Vermaat, M. B. (2002), A Probabilistic Analysis of the Dutch Lotto, Technical report, University of Groningen. Research report.
- Ljung, G. M. & Box, G. E. P. (1978), 'On a Measure of Lack of Fit in Time Series Models', *Biometrika* **65**(2), 297–303.
- Nass, C. A. G. (1959), 'The 2-test for Small Expectations in Contingency Tables, with Special Reference to Accidents and Absenteeism', *Biometrika* **46**, 365–385.
- Percy, D. F. (2006), Bayesian Methods for Testing the Randomness of Lottery Draws, Technical report, University of Salford, Centre for Operational Research and Applied Statistics. Research report.
- Quesenberry, C. P. & Hurst, D. C. (1964), 'Large-Sample Simultaneous Confidence Intervals for Multinomial Proportions', *Technometrics* **6**, 191–195.
- Royal Statistical Society (2000), Reports on the Randomness of U. K. Lottery Games, Technical report. Obtenidos en septiembre de 2007.
*<http://www.natlotcomm.gov.uk>
- Royal Statistical Society (2002), Reports on the Randomness of U. K. Lottery Games, Technical report. Obtenidos en septiembre de 2007.
*<http://www.natlotcomm.gov.uk>
- Schwartz, D. G. (2003), *Suburban Xanadu: The Casino Resort on the Las Vegas Strip and Beyond*, Routledge, New York.
- Schwartz, D. G. (2006), *Roll the Bones: The History of Gambling*, Gotham Books, New York.
- Sison, C. P. & Glaz, J. (1995), 'Simultaneous Confidence Intervals and Sample size Determination for Multinomial Proportions', *Journal of the American Statistical Association* **90**(429), 366–369.
- Stern, H. & Cover, T. M. (1989), 'Maximum Entropy and the Lottery', *Journal of the American Statistical Association* **84**, 980–985.
- Teo, C. P. & Leong, S. M. (2002), 'Managing Risk in a Four-Digit Number Game', *SIAM REVIEW* **44**(4), 601–615.
- University of Salford (2004), Randomness of the Lotto Draws: Summary of Findings, Technical report, Centre for the Study of Gambling. Obtenido en septiembre de 2007.
*<http://www.natlotcomm.gov.uk>

University of Salford (2005*a*), Randomness of the Lotto Lucky Dip, Technical report, Centre for the Study of Gambling. Obtenido en septiembre de 2007.
*<http://www.natlotcomm.gov.uk>

University of Salford (2005*b*), Randomness of thunderball draws, Technical report, Centre for the Study of Gambling. Obtenido en septiembre de 2007.
*<http://www.natlotcomm.gov.uk>

Yang, Z., Xie, M., Kuralmani, V. & Tsui, K. L. (2002), ‘On the Performance of Geometric Charts with Estimated Control Limits’, *Journal of Quality Technology* **34**(4), 448–459.