Urnas de Polya como Proceso Estocástico en Tiempo Discreto

Martín Prado

June 4, 2024

Universidad de los Andes − Bogotá Colombia

1 Modelo Simple de Urnas de Polya

Las urnas de Polya son una familia de modelos de muestreo cuyo objetivo es proponer una generalización de otros modelos más conocidos como el modelo hipergeométrico (muestreo sin reemplazo) y el modelo del experimento de Bernoulli (muestreo con reemplazo). El caso que se va a trabajar dentro de este proyecto es el más sencillo, tenemos una urna con $n = r + g$ bolas, de las cuales r son rojas y g son verdes. Cada que se selecciona una bola de la urna, se añaden I bolas del color seleccionado a la urna.

Identifiquemos al color verde con el número 0 y al color rojo con el número 1. Sea X_i la variable que denota el color de la bola escogida aleatoriamente dentro del modelo, tenemos las siguientes observaciones:

Observación. Algunos casos particulares de las urnas de Polya son:

- Si $I = 0$, entonces $X_i \sim \text{Ber}\left(\frac{r}{n}\right)$. Por lo tanto, $\sum_{i=1}^{m} X_i \sim \text{Bin}(m, \frac{r}{n})$, en donde Bin es la distribución binomial.
- Si $I = -1$, entonces $X_m \sim \text{Ber}\left(\dfrac{r - \sum_{i=1}^{m-1} X_i}{n - i}\right)$. Por lo tanto, $\sum_{i=1}^{m} X_i \sim \text{HG}(n, r, m)$, en donde HG es la distribución hipergeométrica.

Definición 1.1. La formula general de permutaciones se define de la siguiente forma. Para $r, s \in \mathbb{R}$ y $j \in \mathbb{N}$,

$$r^{(s,j)} = r(r+s)(r+2s)\dots(r+(j-1)s) = \prod_{i=0}^{j-1} r + is.$$

Observación. Algunos casos particulares a resaltar de la fórmula anterior son:

- $r^{(0,j)} = r^j = \underbrace{r \times \cdots \times r}_{j \text{ veces}}$.
- $r^{(-1,j)} = r^{(j)} = r(r-1)\cdots(r-j+1)$.
- $r^{(1,j)} = r^{[j]} = r(r+1)\cdots(r+j-1)$.
- $r^{(r,j)} = j!r^j$.
- $1^{(1,j)} = j!$.

2 Distribución Beta-Binomial

Vamos a exponer algunas funciones que aparecen dentro de las distribuciones con las que estamos trabajando al usar los modelos de urnas. En primer lugar, tenemos la función Γ (Gamma) que funciona como una generalización del factorial para números reales.

Definición 2.1 (Función Gamma)**.** Para cualquier $x \in \mathbb{R}$, la función Gamma se define como

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

Teorema 2.1. Para cualquier $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$. En particular, si $n \in \mathbb{Z}^+$, entonces $\Gamma(n) = (n-1)!$

Proof. Usando integración por partes obtenemos que

$$\begin{aligned}
\Gamma(x) &= \int_0^\infty t^{x-1}e^{-t}dt \\
&= \cancel{\left[-e^{-t}t^{x-1}\right]_0^\infty} \to 0 + \int_0^\infty (x-1)t^{x-2}e^{-t}dt \\
&= (x-1)\int_0^\infty t^{x-2}e^{-t}dt \\
&= (x-1)\Gamma(x-1).
\end{aligned}$$

Note que para $x = 1$,

$$\Gamma(1) = \int_0^\infty t^0 e^{-t}dt = 1 = 0!$$

Por lo que para cualquier número natural $n \geq 1$,

$$\Gamma(n) = (n-1)(n-2)\cdots 1 \cdot \Gamma(1) = (n-1)!$$

□

Observación. También, se puede crear con la función Gamma una generalización de las formulas de permutación generalizadas.

$$\begin{aligned} r^{(s,j)} &= \prod_{i=0}^{j-1} r + is \quad = s^j \cdot \prod_{i=0}^{j-1} \frac{r}{s} + i \\ &= s^j \cdot (r/s)^{[j]} = s^j \cdot \frac{\Gamma(r/s + j)}{\Gamma(r/s)} \end{aligned}$$

Otra función que posee algunas similitudes con la función combinatoria es la función β (Beta).

Definición 2.2 (Función Beta)**.** Para cualquier $a, b \in \mathbb{R}$ la función Beta se define como

$$\beta(a, b) = \int_0^1 p^{a-1}(1-p)^{b-1} dp.$$

Teorema 2.2. Para cualquier $a, b \in \mathbb{R}$,

$$\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

. □

De aquí en adelante, denotaremos Y_m como la cantidad de bolas rojas obtenidas en el modelos simple de Polya después de m intentos.

Definición 2.3.

$$Y_m = \sum_{i=1}^{m} X_i.$$

La meta de los siguientes teoremas es encontrar una distribución para Y_m usando las herramientas que hemos desarrollado en esta parte del proyecto.

Teorema 2.3. Para $X_1, \ldots, X_m$ obtenidas del proceso de Polya que definimos anteriormente, tenemos una sucesión $(x_i)_{i \leq m} \in \{0, 1\}^m$ que define qué color se sacó en cada turno, y a lo largo del experimento se saca un total de $y = \sum_{i=1}^{m} x_i$ bolas rojas. Entonces,

$$\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m\} = \frac{r^{(I,y)} g^{(I,m-y)}}{n^{(I,m)}}.$$

Proof. Dentro de los m turnos, se sacan en total y bolas rojas y $m - y$ bolas verdes. Eventualmente,

- En algunos turnos la urna va a tener $r, (r + I), \ldots, (r + (y-1)I)$ bolas rojas.

- En otros turnos $g, (g+I), \ldots, (g+(m-y-1)I)$ bolas verdes.
- Por otra parte, el total de bolas, independientemente de cuál color saquemos va a ser en cada turno $n, (n+I), \ldots, (n+(y-1)I)$ respectivamente.

Por principio de multiplicación, después reordenar la formula de probabilidad obtenemos

$$\begin{aligned}\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m\} &= \frac{r(r+I)\cdots(r+(y-1)I) \times g(g+I)\cdots(g+(m-y-1)I)}{n(n+I)\cdots(n+(m-1)I)} \\ &= \frac{r^{(I,y)}g^{(I,m-y)}}{n^{(I,m)}}.\end{aligned}$$

□

Corolario 2.3.1. Para $0 \leq y \leq m$, y una sucesión $x_1, \ldots, x_m$ tal que $\sum_{i=1}^{m} x_i = y$,

$$\mathbf{P}\{Y_m = y\} = \binom{m}{y}\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m\} = \binom{m}{y}\frac{r^{(I,y)}g^{(I,m-y)}}{n^{(I,m)}}.$$

. □

Esta puede parecer una respuesta satisfactoria ante la pregunta de cuál es la distribución Y_m. Sin embargo, podemos mejorar aún más la respuesta introduciendo la siguiente función de distribución.

Definición 2.4 (Distribución Beta-Binomial)**.** Se dice que una variable aleatoria discreta Y tiene la distribución Beta-Binomial con parámetros $m \in \mathbb{N}$, $a, b \in \mathbb{R}^+$, o en otras palabras, $Y \sim \text{BetaBin}(m, a, b)$ cuando

$$\mathbf{P}\{Y = y\} = \binom{m}{y}\frac{\beta(y+a, m-y+b)}{\beta(a,b)}.$$

Para hacer uso de la función de distribución anterior, debemos hacer los siguientes cambios a la ecuación del corolario anterior

$$\frac{r^{(I,y)}g^{(I,m-y)}}{n^{(I,m)}} = \frac{[r^{(I,y)}/I^y][g^{(I,m-y)}/I^{m-y}]}{[n^{(I,m)}/I^m]} = \frac{(r/I)^{[y]}(g/I)^{[m-I]}}{(n/I)^{[m]}}.$$

Por ende,

Teorema 2.4. Para $I \neq 0$, Y_m tiene una distribución Beta-Binomial con parámetros m, r/I y g/I. Es decir,

$$Y_m \sim \text{BetaBin}\left(m, \frac{r}{I}, \frac{g}{I}\right)$$

Proof. Usando la fórmula anterior, y propiedades de la función Γ obtenemos

$$\begin{aligned}
\mathbf{P}\{Y_m = y\} &= \binom{m}{y} \frac{(r/I)^{[y]}(g/I)^{[m-I]}}{(n/I)^{[m]}} \\
&= \binom{m}{y} \frac{\frac{\Gamma(\frac{r}{I}+y)}{\Gamma(\frac{r}{I})} \cdot \frac{\Gamma(\frac{g}{I}+m-y)}{\Gamma(\frac{g}{I})}}{\frac{\Gamma(\frac{n}{I}+m)}{\Gamma(\frac{n}{I})}} \\
&= \binom{m}{y} \frac{\Gamma(\frac{r}{I}+y)\cdot\Gamma(\frac{g}{I}+m-y)}{\Gamma(\frac{n}{I}+m)} \cdot \frac{\Gamma(\frac{n}{I})}{\Gamma(\frac{r}{I})\cdot\Gamma(\frac{g}{I})} \\
&= \binom{m}{y} \frac{\beta\left(\frac{r}{I}+y, \frac{g}{I}+m-y\right)}{\beta\left(\frac{r}{I}, \frac{g}{I}\right)}
\end{aligned}$$

□

3 Propiedades y Ejemplos

En esta sección vamos a explorar propiedades interesantes del proceso de Polya. Lo primero que vamos a observar es que la distribución límite de

Definición 3.1 (Variable Intercambiable)**.** Para una sucesión de variables aleatorias $X_1, \ldots, X_m$ con densidad conjunta f, se dice que $X_1, \ldots, X_m$ son intercambiables si para cualquier reordenamiento de los índices $\sigma \in S^m$,

$$f(x_1, \ldots, x_m) = f(x_{\sigma(1)}, \ldots, x_{\sigma(m)}).$$

Observación. Note que de la definición se sigue que las densidades $f_1, \ldots, f_m$ de $X_1, \ldots, X_m$ respectivamente son iguales debido a que

$$\begin{aligned}
f_1(x_1) &= \int_{\Omega^{m-1}} f(x_1, \ldots, x_m) dx_2 \ldots dx_m \\
&= \int_{\Omega^{m-1}} f(x_{\sigma(1)}, \ldots, x_{\sigma(m)}) dx_2, \ldots, dx_m \\
(x_1 \text{ en posición } \sigma(1)) &= f_{\sigma(1)}(x_1).
\end{aligned}$$

Sin embargo, esto no significa que sean independientes como en el caso de la urna de Polya.

Teorema 3.1. Si $X_1, \ldots, X_m$ definen un proceso de Polya, entonces son intercambiables

Proof. De acuerdo al teorema 2.3, la probabilidad conjunta de $X_1, \ldots, X_m$ solo depende de la suma $y = \sum_{i=1}^{m} x_1$:

$$f(x_1, \ldots, x_m) = \mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m\} = \frac{r^{(I,y)} g^{(I,m-y)}}{n^{(I,m)}}.$$

Por lo tanto, para cualquier re-ordenamiento de índices

$$f(x_\sigma(1), \ldots, x_{\sigma(m)}) = \frac{r^{(I,y)} g^{(I,m-y)}}{n^{(I,m)}} = f(x_1, \ldots, x_m)$$

□

Por lo tanto, $X_1 \sim X_j$ para cualquier $j \in \mathbb{N}$, y a partir de esto, será fácil calcular la esperanza y las varianzas del proceso de Polya

Teorema 3.2.

$$\mathbf{E}\left[X_i\right] = \frac{r}{n}, \quad \mathbf{Var}\left[X_i\right] = \frac{rg}{n^2}, \quad , \mathbf{Cov}\left[X_i, X_j\right] = \frac{rgk}{n^2(n+I)^2}$$

Proof. En primer lugar, note que para la primera bola $X_1 \sim \text{Ber}(r/n)$, y por ende,

$$\mathbf{E}\left[X_1\right] = \frac{r}{n}, \qquad \mathbf{Var}\left[X_1\right] = \frac{r}{n} \cdot \frac{g}{n}.$$

Por el teorema anterior, es fácil ver que como $\mathbf{P}\{X_j = 1\} = \mathbf{P}\{X_1 = 1\} = r/n$, entonces, comparten los mismos momentos. Finalmente, como son intercambiables, se sigue que

$$\mathbf{P}\{X_i = 1, X_j = 1\} = \mathbf{P}\{X_1 = 1, X_2 = 1\} = \frac{r}{n} \cdot \frac{r+I}{n+I}$$

Por ende, para la covarianza de las variables,

$$\begin{aligned}
\text{Cov}(X_i, X_j) &= \mathbf{E}\left[X_i X_j\right] - \mathbf{E}\left[X_i\right]\mathbf{E}\left[X_j\right] \\
&= \mathbf{P}\{X_i = 1, X_j = 1\} - \frac{r^2}{n^2} \\
&= \frac{r}{n} \cdot \frac{r+I}{n+I} - \frac{r^2}{n^2} \\
&= \frac{rgk}{n^2(n+I)^2}.
\end{aligned}$$

□

Corolario 3.2.1.

$$\mathbf{E}\left[Y_m\right] = m\frac{r}{n}, \qquad \mathbf{Var}\left[Y_m\right] = \frac{r}{n} \cdot \frac{g}{n} \cdot \frac{m(n+1)}{2n^2}.$$

Teorema 3.3. Para una sucesión $\{x_i\}_{i=1}^m$ y sea $y = \sum_{i=1}^m x_i$, se sigue que

$$\mathbf{P}\{X_{m+1} = 1 \mid X_1 = x_1, \ldots, X_m = x_m\} = \frac{r + ky}{n + km}$$

Proof. De acuerdo a la regla de probabilidad condicional y al teorema 2.3

$$\begin{aligned}\mathbf{P}\{X_{m+1} = 1 \mid X_1 = 1, \ldots, X_m = x_m\} &= \frac{\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m, X_{m+1} = 1\}}{\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m\}} \\ &= \frac{r^{(k,y+1)} g^{(k,m-y)}}{n^{(k,m+1)}} \cdot \frac{n^{(k,m)}}{r^{(k,y)} g^{(k,m-y)}} \\ &= \frac{r + ky}{n + km}.\end{aligned}$$

□

Corolario 3.3.1.

$$\mathbf{P}\{X_{m+1} = 1 \mid Y_m = y\} = \frac{r + ky}{n + km}$$

Proof. Usando el corolario 2.3.1 y el teorema anterior,

$$\begin{aligned}\mathbf{P}\{X_{m+1} = 1 \mid Y_m = y\} &= \frac{\mathbf{P}\{X_{m+1} = 1, Y_m = y\}}{\mathbf{P}\{Y_m = y\}} \\ &= \frac{\binom{m}{y}\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m, X_{m+1} = 1\}}{\binom{m}{y}\mathbf{P}\{X_1 = x_1, \ldots, X_m = x_m\}} \\ &= \frac{r + ky}{n + km}.\end{aligned}$$

□

4 Martingalas

Definición 4.1. A una sucesión $Z_1, Z_2, \ldots$ de variables aleatorias se le considera un martingalas con respecto a una filtración $\mathscr{F} = \{F_j : j \in \mathbb{N}\}$ si, para cada $i \in \mathbb{N}$,

$$\mathbf{E}\left[Z_{m+1} \mid F_m\right] = Z_m.$$

Definición 4.2. A una sucesión $Z_1, \ldots, X_m$ de martingalas se le considera acotada si existe $c_i > 0$, tal que para cualquier $i < m$,

$$|Z_{i+1} - Z_i| \leq c_i,$$

con probabilidad 1.

Teorema 4.1 (Desigualdad de Azuma-Hoeffding)**.** Sea $Z_1, \ldots, Z_m$ un martingalas acotado por una sucesión $\{c_i\}_{i\in\mathbb{N}} > 0$, entonces

$$\mathbf{P}\{Z_m - Z_1 \geq \varepsilon\} \leq \exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^{m} c_i^2}\right).$$

. □

A partir de ahora, suponga que $k \geq 0$ para que el proceso pueda continuar de forma indefinida

Definición 4.3. Llame M_m al promedio muestral del proceso de Polya y Z_m a la proporción de bolas rojas que quedan en la urna después de m turnos.

$$M_m = \frac{1}{m}\sum_{i=1}^{m} X_i = \frac{1}{m}Y_m,$$

$$Z_m = \frac{r + kY_m}{n + km} = \frac{r + kmM_m}{n + km}.$$

Note que en caso de existir un límite en probabilidad para M_m, Z_m tendría el mismo límite.

Ahora, con estas herramientas, el objetivo es probar que la sucesión de variables Z_m converge c.s. a una variable aleatoria Z_∞ con distribución beta. Para esto vamos a probar que son un martingalas y luego usar el teorema de convergencia de Doob para probar que existe la variable aleatoria límite.

Teorema 4.2. Las variables $Z_1, Z_2, \ldots$ definen un martingalas con respecto a la filtración $\mathscr{F}_m = \sigma(X_1, \ldots, X_m)$.

Proof. Para la última linea usamos el teorema anterior

$$\begin{aligned}
\mathbf{E}\left[Z_{m+1} \mid F_m\right] &= \mathbf{E}\left[\frac{r+kY_{m+1}}{n+k(m+1)} \;\middle|\; F_m\right] \\
&= \frac{\mathbf{E}\left[r+k(Y_m+X_{m+1}) \;\middle|\; F_m\right]}{n+km+k} \\
&= \frac{r+k\mathbf{E}\left[Y_m+X_{m+1} \;\middle|\; X_1,\ldots,X_m\right]}{n+km+k} \\
&= \frac{r+k\left[Y_m+\mathbf{E}\left[X_{m+1} \mid X_1,\ldots,X_m\right]\right]}{n+km+k} \\
&= \frac{r+k\left[Y_m+\mathbf{P}\{X_{m+1}=1 \mid X_1,\ldots,X_m\}\right]}{n+km+k} \\
&= \frac{r+k\left[Y_m+\frac{r+ky}{n+km}\right]}{n+km+k}
\end{aligned}$$

Después de simplificar obtenemos

$$= \frac{r+kY_m}{n+km} = Z_m.$$

□

Teorema 4.3 (Teorema de Doob para convergencia de martingalas)**.** Sea $Z_1, Z_2, \ldots,$ un martingalas que satisface

$$\sup_n \mathbf{E}\left[|Z_m|\right] < \infty.$$

Entonces, existe Z_∞ tal que $Z_m \to Z_\infty$ casi siempre y Z_∞ tiene esperanza finita.

. □

Teorema 4.4. M_m converge en distribución a una variable aleatoria P con distribución beta, $P \sim \beta(r/I, g/I)$.

Proof. Usamos teorema 2.4

$$\mathbf{P}\{M_m \le x\} = \mathbf{P}\{Y_m \le mx\} = \sum_{y=0}^{\lfloor mx \rfloor} \binom{m}{y} \frac{\beta\left(\frac{r}{I}+y, \frac{g}{I}+m-y\right)}{\beta\left(\frac{r}{I}, \frac{g}{I}\right)}$$

Luego, usamos la fórmula explícita de la función beta, para abreviar usamos $a = r/I$, $b = g/I$

$$\begin{aligned}
\mathbf{P}\{M_m \leq x\} &= \frac{1}{\beta(a,n)} \sum_{y=0}^{\lfloor mx \rfloor} \binom{m}{y} \beta(a+y, b+m-y) \\
&= \frac{1}{\beta(a,b)} \sum_{y=0}^{\lfloor mx \rfloor} \binom{m}{y} \int_0^1 p^{a+y-1}(1-p)^{b+m-y-1} dp \\
&= \frac{1}{\beta(a,b)} \int_0^1 \left(\sum_{y=0}^{\lfloor mx \rfloor} \binom{m}{y} p^y (1-p)^{m-y} \right) p^{a-1}(1-p)^{b-1} dp
\end{aligned}$$

Defina una variable aleatoria $B_m \sim \text{Bin}(m,p)$ y note que la sumatoria de la expresión anterior es equivalente a $\mathbf{P}\{B_m \leq mx\} = \mathbf{P}\{B_m/m \leq x\}$. Note que B_m/m es promedio muestral de m Bernoulli's y por la ley de grandes números, $\mathbf{P}\{B_m/m \leq x\}$ converge a $\mathbb{1}\{p \leq x\}$. Por ende, usando el teorema de convergencia dominada, finalmente obtenemos

$$\mathbf{P}\{M_m \leq x\} \to \frac{1}{\beta(a,b)} \int_0^1 p^{a-1}(1-p)^{b-1} dp = \text{Beta}(a,b)$$

Por lo tanto, M_n converge en distribución a la distribución $\text{Beta}(a,b) = \text{Beta}(r/I, g/I)$. □

Corolario 4.4.1. la sucesión Z_m converge c.s. a una variable aleatoria Z_∞ con distribución $\text{Beta}(r/I, g/I)$.

Proof. Note que Z_m y M_m tienen el mismo límite en distribución. Por lo tanto, usando el teorema anterior Z_m converge en distribución a la distribution $\text{Beta}(r/I, g/I)$. Por otra parte, $\mathbf{E}[|Z_m|] \leq 1$, por ende, usando el teorema de Doob, probamos que el límite Z_∞ existe, y como este límite es único debe tener distribución $\text{Beta}(r/I, g/I)$. □

References

[Ahlberg(2023)] Daniel Ahlberg. Notes on urn models, 2023. URL https://staff.math.su.se/daniel.ahlberg/notes-urns.pdf.

[Alon and Spencer(2016)] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[Helfand(2013)] NORA Helfand. Polya's urn and the beta-bernoulli process. *University of Chicago REU*, 2013.

[Siegrist(2024)] Kyle Siegrist. Pólya's urn process. https://www.randomservices.org/random/urn/Polya.html, 2024. Accessed: 2024-05-23.