Capítulo 3

La Familia de Distribuciones de Pólya

3.1. La distribución de Pólya

En este capítulo se va a partir del esquema de urna de Pólya expuesto en el capítulo anterior. Con el fin de distinguir los aspectos referentes a la distribución de Pólya, de los ya comentados en los modelos de urnas, y para utilizar una notación que ha sido comúnmente aceptada en la literatura al respecto, se va a plantear el esquema de urna que da origen a la distribución de Pólya de la manera que se muestra a continuación:

Consideremos que una urna finita contiene inicialmente M bolas blancas y $N-M$ bolas negras (o alternativamente, pN bolas blancas y qN bolas negras, siendo N el total de bolas de la urna, $p = \frac{M}{N}$ y $q = 1 - p$). Se extrae una bola, se anota su color y se introduce de nuevo en la urna junto a c bolas más del mismo color que la extraída. El procedimiento se repite n veces. La probabilidad de extraer k bolas

blancas en las n extracciones es

$$P[X=k]=\frac{\binom{-\frac{pN}{c}}{k}\binom{-\frac{qN}{c}}{n-k}}{\binom{-\frac{N}{c}}{n}}$$

donde

$$\binom{-A}{B}=(-1)^B\binom{A+B-1}{B}$$

La forma habitual de expresar esta probabilidad, y a la que se hará referencia en las páginas siguientes es:

$$P[X=k]=\binom{n}{k}\cdot\frac{M^{(k,c)}(N-M)^{(n-k,c)}}{N^{(n,c)}}.$$

Para llegar a esa expresión se parte de la probabilidad de extraer una bola blanca en la primera extracción, $P[b_1]=\frac{M}{N}$. La probabilidad de obtener una bola blanca en la primera extracción, y también bola blanca en la segunda extracción será $P[b_1,b_2]=\frac{M}{N}\cdot\frac{M+c}{N+c}\cdot\frac{M+c}{N+c}$. Del mismo modo, la probabilidad de conseguir tres bolas blancas en las tres primeras extracciones, vendrá dada por:

$$P[b_1,b_2,b_3]=\frac{M}{N}\cdot\frac{M+c}{N+c}\cdot\frac{M+2c}{N+2c}.$$

Y en general, k bolas blancas en las k primeras extracciones:

$$P[b_1,b_2,b_3,\ldots,b_k]=\frac{M}{N}\cdot\frac{M+2c}{N+2c}\cdots\frac{M+(k-1)c}{N+(k-1)c}=\frac{M^{(k,c)}}{N^{(k,c)}}.$$

La probabilidad de que ocurra ésto y además en las $n-k$ extracciones restantes se obtengan bolas negras, será:

$$P[b_1,b_2,b_3,\ldots,b_k,n_{k+1},n_{k+2},n_{k+3},\ldots,n_n]=$$

$$=\frac{M^{(k,c)}}{N^{(k,c)}}\cdot\frac{N-M}{N+kc}\cdot\frac{N-M+c}{N+(k+1)+c}\cdots\frac{N-M+(n-k-1)c}{N+(n-1)c}=\frac{M^{(k,c)(N-M)^{(n-k,c)}}}{N^{(n,c)}}.$$

Como esta probabilidad es la misma sea cual fuera el orden de los colores extraídos, basta multiplicar esta expresión por $\binom{n}{k}$ para obtener la expresión referida.

Bosch(1963) presenta siete formas distintas de expresar esta misma probabilidad. A continuación se recogen dos de ellas, donde la expresión de la función de probabilidad de la distribución de Pólya está en términos de las funciones Gamma y Beta respectivamente, siempre tomando $c \neq 0$:

$$P[X = k] = \binom{n}{k} \cdot \frac{\Gamma\left(\frac{M}{c} + k\right) \Gamma\left(\frac{N-M}{c} + n - k\right) \Gamma\left(\frac{N}{c}\right)}{\Gamma\left(\frac{M}{c}\right) \Gamma(\left(\frac{N-M}{c}\right))\Gamma\left(\frac{N}{c} + n\right)}$$

$$P[X = k] = \binom{n}{k} \cdot \frac{\beta\left(\frac{M}{c} + k; \frac{N-M}{c} + n - k\right)}{\beta\left(\frac{M}{c}; \frac{N-M}{c}\right)}.$$

El parámetro c puede tomar valores negativos, pero de ser así debe verificar la condición $N + c(n-1) > 1$. Además, en este caso, si queremos que el proceso finalice sin dificultad deben ser M y $N - M$ divisibles por c, y

$$\text{máx}\left(0, n + \frac{N-M}{c}\right) \leq k \leq \text{mín}\left(n, \frac{-M}{c}\right)$$

La distribución de Pólya puede definirse, tal como ya se ha planteado, en función de los cuatro parámetros N, M, n y c, que determinan el proceso. El parámetro M puede sustituirse por p, ya que ambos parámetros proporcionan la misma información.

En función de los valores que reciba el parámetro c, Pólya(1954) interpretó los siguientes casos particulares:

(1) $c > 0$: Interpreta que el éxito y el fracaso son contagiosos en el sentido de que un éxito o un fracaso aumenta la probabilidad de éxito o de fracaso, respectivamente.

(2) $c = 0$: Entonces los sucesos son independientes.

(3) $c < 0$: Interpreta que cada extracción va a originar un revés de la fortuna, en el sentido en que el éxito disminuye a su vez la probabilidad de obtener un nuevo éxito (del mismo modo el fracaso disminuye también la probabilidad de un nuevo fracaso).

En función de los valores particulares de los parámetros, se obtienen distribuciones notables como casos particulares de la distribución de Pólya. Así,

(1) $c = -1$: Distribución hipergeométrica.

(2) $c = 0$: Distribución binomial.

(3) $c = 1$: Distribución beta binomial o hipergeométrica negativa.

(4) $c = M = N - M$: Distribución uniforme discreta.

3.2. Distribución inversa de Pólya

También puede realizarse el planteamiento que se ha llamado Inverso, si en lugar de repetir el proceso n veces se establece la regla de parar después de haber obtenido k bolas blancas. De esta forma, en lugar de considerar la probabilidad de obtener k bolas blancas en n extracciones, se consideran cuántas extracciones son necesarias para obtener k bolas blancas. La probabilidad de que se necesiten $k+r$ extracciones, viene indicada por la distribución Inversa de Pólya. La probabilidad de que sean necesarias r extracciones para obtener k bolas blancas viene dada por:

$$P[X = k] = \frac{k}{r} \cdot \binom{-\frac{pN}{c}}{k} \binom{-\frac{qN}{c}}{r-k} \binom{-\frac{N}{c}}{r}$$

Con más frecuencia se utiliza la siguiente expresión:

$$P[X = k] = \binom{k+r-1}{r} \cdot \frac{M^{(k,c)}(N-M)^{(r,c)}}{N^{(k+r,c)}}.$$

La distribución Inversa de Pólya puede definirse como función de los parámetros N, M, k y c, que determinan el proceso. En este caso, como en el anterior, puede sustituirse el parámetro M por el parámetro $p = \frac{M}{N}$.

En función de los valores particulares de los parámetros se obtienen distribuciones notables como casos particulares de la distribución inversa de Pólya. Así, tenemos que:

(1) $c = -1$: Distribución beta binomial o hipergeométrica negativa

(2) $c = 0$: Distribución binomial negativa

(3) $c = 1$: Distribución beta Pascal

(4) $c = 1$; $M = 1$: Distribución geométrica

3.3. La distribución de Markov-Pólya generalizada

A partir del modelo de urnas que desarrollan Janardan y Schaeffer(1977), se deriva la distribución de Markov-Pólya generalizada. El nombre se debe a Janardan y Schaeffer que reivindican la autoría de la conocida como distribución de Pólya, a Markov.

Utilizando la misma notación dada a los parámetros en el modelo de urna de referencia, la función de masa de probabilidad de la distribución de Markov-Pólya generalizada es:

$$P[X = k] = \frac{\frac{a}{a+kt} \cdot (a + kt)^{(k,c)} \frac{b}{b+(N-k)t} \cdot (b + (N - k)t)^{(N-k,c)}}{\frac{a+b}{a+b+Nt} \cdot (a + b + Nt)^{(N,c)}}$$

Casos especiales de la distribución de Markov-Pólya generalizada son:

(1) $c = -1$: Distribución mixtura cuasi-hipergeométrica

$t = 0$: Distribución hipergeométrica

$t = 1$: Distribución hipergeométrica negativa

(2) $c = 0$: Distribución mixtura cuasi-binomial

$t = 0$: Distribución binomial

(3) $c = 1$: Distribución mixtura cuasi-hipergeométrica negativa

$t = 0$: Distribución hipergeométrica negativa

3.4. Momentos

3.4.1. Distribución de Pólya

Parece interesante hacer mención a la existencia de las llamadas *Distribuciones Hipergeométricas Refundidas Generalizadas*, $GHRD$, que presentan función generatriz de probabilidad de la forma:

$$\frac{{}_pF_q[\lambda z + \varepsilon]}{{}_pF_q[\lambda + \varepsilon]}$$

Cuando $\varepsilon = 0$ se obtienen como caso particular las *Distribuciones Hipergeométricas Generalizadas*, $GHPD$, grupo al que pertenece la distribución de Pólya.

El primero en realizar un estudio acerca de los momentos correspondientes a la distribución de Pólya, fue Jordan(1927). El procedimiento que se expone a continuación se basa en la caracterización dada de la distribución de Pólya dentro de las distribuciones hipergeométricas generalizadas.

El r-ésimo momento factorial de la distribución de Pólya, se obtiene de la particularización de la forma general del mismo correspondiente a las distribuciones de $GHPD$,

$$\mu'_{(r)} = \frac{n!\left(\frac{-M}{c}\right)!\left(\frac{-N}{c}-r\right)!}{(n-r)!\left(\frac{-M}{c}-r\right)!\left(\frac{-N}{c}\right)!}$$

Para las distribuciones de Pólya estos momentos existen siempre, y toman el valor cero cuando $r > n$. Además, son finitos para todo r.

A partir de esta expresión se obtienen la esperanza y la varianza de la distribución de Pólya:

$$E[X] = np \ ; \ Var[X] = \mu_2 = \frac{npq(N+nc)}{(N+c)}$$

Del mismo modo, se tiene que el momento de orden 3 es:

$$\mu_3 = \frac{\mu_2(q-p)(N+2nc)}{(N+2c)}$$

Bosch(1963) también obtiene estos momentos sin utilizar la propiedad de las distribuciones hipergeométricas generalizadas.

3.4.2. Distribución Inversa de Pólya

Patil y Joshi(1968) recogen los momentos media y varianza de la distribución inversa de Pólya:

$$E[X] = \frac{kqN}{(pN-c)}; \quad \text{si } M > c,$$

no existiendo la esperanza para otros valores de M.

$$Var[X] = \left(\frac{k(N-M)}{M-c}\right)\left(\frac{N-c}{M-c}\right)\left(\frac{M+(k-1)c}{M-2c}\right)$$

La varianza tiene sentido cuando $M > 2c$, y no existe para otros valores.

3.5. Función generatriz de probabilidad

3.5.1. Distribución de Pólya

La función generatriz de probabilidad de la distribución de Pólya se obtiene a partir de la forma general para las distribuciones $GHPD$, realizando la adecuada particularización de los parámetros:

$$G(z) = \frac{{}_2F_1\left(-n, \frac{M}{c}; -n+1-\frac{(N-M)}{c}; z\right)}{{}_2F_1\left(-n, \frac{M}{c}; -n+1-\frac{(N-M)}{c}; 1\right)}$$

Ollero y Ramos (1995), tomando como base la inclusión de la distribución de Pólya en el sistema de Pearson discreto, establecen, como se verá en el siguiente capítulo, la siguiente expresión de la función generatriz de probabilidad en el caso en que $c < 0$:

$$G(z) = p_m \cdot z^m \cdot {}_2F_1\left(-n+m, \frac{M}{c}+m; -\frac{N-M+c(n-1)}{c}+2m; z\right)$$

donde el soporte de la variable viene dado por $\{m, m+1, ..., s\}$ y p_m es el valor de la función de masa de probabilidad en el entero menor m.

El interés de esta última expresión radica en que se contempla el caso en que el soporte no sea completo, es decir $m > 0$. Por otra parte, en el trabajo antes citado, Ollero y Ramos obtienen asimismo la expresión de la función generatriz de probabilidad para las distribuciones de cierta familia, que ellos llaman $\mathcal{P}_H$, a la que nos referiremos en el capítulo *4*.

Patil y Joshi (1968), indican el valor de la función generatriz de probabilidad en

los siguientes términos:

$$G(z) = \frac{\left(\frac{-qN}{c}\right)^{(n)}}{\left(\frac{-N}{c}\right)^{(n)}} \cdot {}_2F_1\left(-n, \frac{pN}{c}; \frac{-qN}{c} - n + 1; z\right).$$

Acerca de las raíces de la función generatriz de probabilidad de la distribución de Pólya, Ollero y Ramos (1995) probaron el siguiente resultado

Teorema 3.1 *Si $G_p(z)$ es la función generatriz de probabilidad de una distribución de Pólya $P(N, M, n, c)$, con $c < 0$ y rango $\{m, m+1, ..., s\}$, entonces $G_p(z)$ tiene s raíces reales, de las cuales $s - m$ son simples negativas y las restantes son la raíz $z = 0$, con orden de multiplicidad m.*

3.5.2. Distribución Inversa de Pólya

Patil y Joshi (1968) dan la siguiente expresión para la función generatriz de probabilidad de la Distribución Inversa de Pólya:

$$G(z) = \frac{\left(\frac{-qN}{c}\right)^{(k)}}{\left(\frac{-N}{c}\right)^{(k)}} \cdot {}_2F_1\left(k, \frac{qN}{c}; \frac{N}{c} + k; z\right)$$

3.6. Caracterización de la distribución de Pólya

Janardan(1984) obtiene la siguiente caracterización de la distribución de Pólya $P(N, M, n, c)$ basada en su comportamiento frente a mixturas respecto del parámetro n cuando este parámetro sigue una distribución binomial negativa.

Teorema 3.2 *Consideremos la familia de distribuciones $f(k; n)$ indexadas por el parámetro $n : 0, 1, 2, \ldots$, definida cada una sobre un subconjunto de $\{0, 1, ..., n\}$, e independientes de q. Si n sigue una distribución binomial negativa, $BN(n; N/c, q)$, entonces la mixtura resultante con $f(j; n)$ es una distribución binomial negativa*

$BN(k; M/c, q)$, si y sólo si la familia de distribuciones $f(k; n)$ es la familia de distribuciones de Pólya $P(N, M, n, c)$.

Recordemos que dada una familia de distribuciones $F_j(x_1, x_2, \ldots, x_n)$, con $j = \ldots, -1, 0, 1, 2, \ldots$, y una sucesión de parámetros a_j tales que $\sum_{j=-\infty}^{\infty} a_j = 1$, con $a_j \geq 0$, $\forall j$, se llama *composición o mixtura* de F_j respecto de a_j a la distribución obtenida como

$$F(x_1, x_2, \ldots, x_n) = \sum_{-\infty}^{\infty} a_j F_j(x_1, x_2, \ldots, x_n)$$

A partir del teorema anterior, Janardan (1984) obtuvo el siguiente corolario.

Corolario 3.2.1 *Las variables aleatorias independientes X e Y siguen distribuciones $BN(x; M/c, q)$ y $BN(y; (N-M)/c, q)$ respectivamente si y sólo si la distribución condicionada de X dado $X + Y = n$ sigue una distribución de Pólya $P(N, M, n, c)$.*

Ramos y Ollero (1989) obtuvieron caracterizaciones de la distribución hipergeométrica basadas en su comportamiento frente a mixturas con otras distribuciones. Una de las mixturas estudiadas fue con la distribución de Pólya, derivándose de ella el siguiente teorema de caracterización.

Teorema 3.3 *Sea $\{F_M\}$, $M = 0, 1, \ldots, N$, una familia de funciones de distribución que asignan probabilidades no nulas sólo a los enteros $0, 1, ..., n$. Condición necesaria y suficiente para que esta familia sea la formada por la familia de distribuciones hipergeométricas $H(N, M, n)$, es que se verifique:*

(1) Su composición cuando M sigue una distribución de Pólya $P(N', M', N, c)$ es la distribución de Pólya $P(N', M', n, c)$.

(2) Las distribuciones F_M no dependen del valor de N'.

3.7. Distribución de Pólya multivariante

A partir de las extensiones de los modelos de urna que se desarrollaron, puede obtenerse la distribución multivariante de Pólya. Esta distribución fue desarrollada por Steyn (1951) como perteneciente a una familia de distribuciones multivarianes discretas que él mismo definió.

En este caso, a partir del esquema de urna

$$(N, M_1, M_2, \ldots, M_m, c, d, n),$$

se plantea buscar la probabilidad de obtener una configuración de colores: $\mathbf{K} = (K_1, K_2, \ldots, K_m)$ en n extracciones.

La función de probabilidad de la distribución multivariante de Pólya viene dada por

$$P[X = K] = \frac{\prod_{i=1}^{m+1} \binom{-p_i \cdot \frac{N}{c}}{K_i}}{\binom{\frac{-N}{c}}{n}}$$

donde $p_i = \dfrac{M_i}{N}$, y por tanto $p_{m+1} = 1 - \sum_{i=1}^{m} p_i$.
Cuando m vale 1 esta distribución se reduce a la distribución de Pólya.

En el trabajo de Patil y Joshi (1968) presentan los principales momentos correspondientes a la distribución multivariante de Pólya.

Vector de medias $\mu_i = np_i$

Matriz de covarianzas

$$\sigma_{i,j} = \begin{cases} n\left(\frac{a_i}{N}\right)\left(\frac{N-a_i}{N}\right)\left(\frac{N+nc}{n+c}\right); & i = j \\ -n\left(\frac{a_i}{N}\right)\left(\frac{a_j}{N}\right)\left(\frac{N+nc}{n+c}\right); & i \neq j \end{cases}$$

Asimismo se recoge en el citado trabajo la correspondiente función generatriz de probabilidad cuyo valor es

$$G(z_1, z_2, \ldots, z_m) = \frac{\left(\frac{p_{m+1}N}{c} + n - 1\right)^{(n)}}{\left(\frac{N}{c} + n - 1\right)^{(n)}} \cdot A$$

siendo

$$A = \left[{}_{m+1}F_1\left(-n, \frac{p_1 N}{c}, \frac{p_2 N}{c}, \ldots, \frac{p_m N}{c}; -\frac{p_{m+1}N}{c} - n + 1; z_1, z_2, \ldots, z_m\right)\right]$$

Bibliografía

[1] Abramowitz, M. and Stegum, I.A. (1965). *Handbook of Mathematical Functions*, New York: Dover.

[2] Bosch, A.J. (1963). The Pólya distribution, *Statistica Neerlandica*, **17**, 201-213.

[3] Eulacio, N.R. (1985). *La familia de distribuciones de Pólya truncada*. Tesis de Mestría. Instituto de Enseñanza e Investigaciones de Ciencias Agrícolas. Méjico: Chapingo.

[4] Janardan, K.G. (1984). On characterizing the Markov-Pólya distribution, *Sankhya, Series A*, **46**, 444-453.

[5] Janardan, K.G. and Schaeffer, D.J. (1977). A generalization of Markov-Pólya distribution its extensions and applications, *Biometrical Journal*, **19**, 87-106.

[6] Johnson, N.L., Kotz, S. and Kemp, A.W. (1992). *Univariate Discrete Distributions*, second edition, New York: John Wiley & Sons.

[7] Jordan, C. (1927). Sur un cas généralisé de la probabilité des épreuves répétées, *Comptes Rendus, Académie des Sciences, Paris*, **184**, 315-317.

[8] Ollero, J. y Ramos, H.M. (1995). Description of a Subfamily of the Discrete Pearson System as Generalized-Binomial Distributions, *Journal of the Italian Statistical Society*, **2**, 235-249.

[9] Patil, G.P. and Joshi, S.W. (1968). *A Dictionary and Bibliography of Discrete Distributions*, Edinburgh: Oliver and Boyd.

[10] Pólya, G. (1954). *Patterns of Plausible Inference*, Princeton: Princeton University Press.

[11] Ramos, H.M. and Ollero, J. (1989). Teoremas de caracterización de la distribución hipergeométrica, *Cuadernos de Estadística Matemática, Serie A*, **11**, 85-101.

[12] Steyn, H.S. (1951). On discrete multivariate probability functions, *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A*, **54**, 23-30.