# Superstardom modelling using an extended Yule distribution

A. Conde-Sánchez* A. J. Sáez-Castillo A. M. Martínez-Rodríguez
J. Rodríguez-Avi M. J. Olmo-Jiménez

XXXIII Congreso Nacional de Estadística e I.O. Madrid, 2012.

**Resumen**

En este trabajo se utiliza una distribución biparamétrica que extiende a la distribución de Yule para modelizar unos datos relativos al número de apariciones en un lista sobre los mejores analistas de mercado (*security analysts*) en Estados Unidos, y de paso intentar determinar si el éxito en dicho campo se debe a la destreza y capacidad del analista o bien a otros factores como la suerte.

## 1. Introducción

En los últimos años se ha creado cierta polémica sobre la naturaleza del superestrellato, como muestran algunos artículos publicados recientemente. Podemos considerar que "el fenómeno superestrella se refiere a la situación donde unos pocos individuos ganan gran cantidad de dinero y dominan el campo en el que llevan a cabo su actividad". La controversia viene dada porque algunos autores señalan que se debe a diferencias en el talento, mientras otros indican que las superestrellas pueden existir independientemente del talento, simplemente porque la gente tiende a seguir a la multitud, creando de esta forma un efecto bola de nieve.

Una forma de decidir sobre la naturaleza de dicho fenómeno se basa en la distribución de Yule. Algunos autores han considerado dicha distribución como mecanismo de probabilidad subyacente en la elección de productos artísticos por parte del consumidor, en concreto, Chung y Cox (1994), Fox y Kochanowski (2004) y Giles (2006) en el campo de la música popular, o Chung y Cox (1998) en el campo del cine. O bien extrapolando esta idea a otros campos, como Cox y Kleiman (2000) en el sector de las finanzas, o Serenko *et al.* (2011) en el campo de la infometría. Dichos autores consideran que dicho modelo es consistente con la idea de que el fenómeno superestrella podría existir entre individuos con igual talento. Así, si esta distribución no es adecuada para modelizar unos datos relativos al superestrellato, proponen que no existe el efecto bola de nieve.

*Departamento de Estadística e I.O. Universidad de Jaén

Sin embargo, la cuestión no está tan clara, pues dependiendo de los datos considerados se ha llegado a conclusiones contrarias en el campo de la música popular americana. Por otra parte, Spierdijk y Voorneveld (2009) indican que la distribución de Yule "es una aproximación bastante buena de los cuantiles inferiores de la distribución superestrella, pero sobrestima el efecto bola de nieve... En otras palabras, la distribución de Yule captura el estrellato, pero no el superestrellato". Por ello, consideran una distribución generalizada de Yule de dos parámetros, la cual conserva la misma génesis que la distribución de Yule, y que afirman sigue siendo consistente con el efecto bola de nieve.

En esta misma línea, nosotros hemos propuesto una nueva distribución, que extiende a la distribución de Yule y con una génesis similar, manteniendo por tanto sus características más relevantes. Esta distribución incluye un nuevo parámetro que permite controlar la cola superior de la distribución, por lo que recoge adecuadamente el superestrellato. Así, hemos utilizado esta distribución para modelizar los datos analizados por Cox y Kleiman (2000), obteniendo un excelente ajuste a diferencia de lo que sucede con la distribución de Yule. Además, ofrecemos una discusión sobre la existencia de superestrellas sin talento desde otra perspectiva, considerando que dichas distribuciones se pueden obtener como mixtura. Sin embargo, la controversia sobre la naturaleza del superestrellato no queda resuelta de forma objetiva, salvo que hagamos alguna suposición sobre el mecanismo que genera la distribución propuesta.

## 2. La distribución EYule

La distribución de Yule pertenece a la familia de distribuciones generadas por la función hipergeométrica de Gauss ${}_2F_1$, concretamente, dentro de las distribuciones llamadas de Tipo I según la clasificación de Rodríguez et al. (2007). Estas distribuciones se notan $GHD\,\mathrm{I}\,(\alpha,\beta,\gamma,\lambda)$ y tienen como expresión general de su función masa la siguiente:

$$f(r) = \frac{1}{{}_2F_1(\alpha,\beta;\gamma;\lambda)}\frac{(\alpha)_r(\beta)_r}{(\gamma)_r}\frac{\lambda^r}{r!} \tag{1}$$

para $r = 0, 1, 2, ...$, siendo $\alpha, \beta, \gamma > 0$ si $0 < \lambda < 1$ y $\gamma > \alpha + \beta$ si $\lambda = 1$. Si trasladamos el soporte de la distribución mediante el cambio de variable $k = r + 1$, la distribución de Yule de parámetro $\rho > 0$ aparece al considerar $\alpha = \beta = 1$ y $\lambda = 1$, siendo $\rho = \gamma - 2$, por lo que se notaría $GHD\,\mathrm{I}\,(1, 1, \rho + 2, 1)$.

Una generalización de la distribución de Yule dentro de la familia de distribuciones $GHD$ tipo I, se obtiene eliminando la restricción $\lambda = 1$, es decir, considerando la distribución $GHD\,\mathrm{I}\,(1, 1, \rho + 2, \lambda)$, con $0 < \lambda \leq 1$, y cuya función masa viene dada por

$$f(r) = \frac{(\rho+1)\lambda^r B(r+1,\rho+1)}{{}_2F_1(1,1;\rho+2;\lambda)} = \frac{r!\lambda^r}{{}_2F_1(1,1;\rho+2;\lambda)(\rho+2)_r}, \qquad r = 0, 1, 2, \ldots \tag{2}$$

o, considerando el cambio de variable $k = r + 1$,

$$f(k) = \frac{(\rho+1)\lambda^{k-1}B(k,\rho+1)}{{}_2F_1(1,1;\rho+2;\lambda)} = \frac{(k-1)!\lambda^{k-1}}{{}_2F_1(1,1;\rho+2;\lambda)(\rho+2)_{k-1}}, \qquad k = 1,2,\ldots \quad (3)$$

con $\rho > -2$; notemos que dicho parámetro puede ser negativo, al contrario de lo que sucede con la distribución de Yule. Por simplicidad, de ahora en adelante consideramos que el soporte de la variable observada es $r = 0, 1, 2, \ldots$. A dicha distribución la notaremos como $EYule(\rho, \lambda)$. Su función generatriz de probabilidad es

$$g(t) = \frac{{}_2F_1(1,1,\rho+2,\lambda t)}{{}_2F_1(1,1,\rho+2,\lambda)}. \quad (4)$$

Debe ser tenido en cuenta que, en general, la función ${}_2F_1(\alpha,\beta;\gamma;\lambda)$ no tiene expresiones explícitas salvo en el caso $\lambda = 1$, de manera que debe ser aproximada de forma computacional.

## 2.1. Propiedades

Veamos algunas propiedades de la distribución $EYule(\rho, \lambda)$. Una exposición más detallada aparece en Martínez-Rodríguez *et al.* (2011).

El hecho que la distribución $EYule$ sea dentro de la familia $GHD$ una distribución Tipo I le confiere algunas propiedades relevantes (Rodríguez et al, 2007). Por ejemplo, no se requiere ninguna restricción sobre los parámetros para garantizar la existencia de todos sus momentos: por tanto, no es posible el efecto de varianza infinita a diferencia de lo que ocurre con la distribución de Yule. Para dicha distribución la varianza existe si $\rho > 2$.

Para valores de $\rho > -1$ esta distribución puede ser obtenida, al igual que la distribución de Yule, como una mixtura de una distribución geométrica con f.m.p. $P(1-P)^r$, donde $P$ sigue una distribución beta generalizada o, de otra forma, se trata de una distribución de Poisson cuyo parámetro $\mu$ sigue una distribución exponencial de parámetro $V = (1-P)/P$ y donde $P$ sigue la densidad beta generalizada dada por

$$f(p) = \frac{\rho+1}{{}_2F_1(1,1;\rho+2;\lambda)}\frac{1}{\lambda p}\left(1-\frac{1-p}{\lambda}\right)^{\rho}, \qquad 1-\lambda < p < 1, \quad (5)$$

con $\rho > -1$ y $0 < \lambda \leq 1$. Obsérvese que si $\lambda = 1$, (5) pasa a ser una densidad beta clásica y (3) la función masa de una distribución de Yule.

Merece la pena dar una explicación de la mixtura que conduce a la distribución EYule. La distribución de Poisson es adecuada en situaciones donde no hay heterogeneidad. En presencia de heterogeneidad se considera que el parámetro $\mu$ no es constante y sigue una determinada distribución. En este caso se está considerando una doble mixtura, permitiendo distinguir entre heterogeneidad provocada por factores externos al individuo y

por factores internos propios del individuo. Es posible entonces considerar que las variaciones en $\mu$ debidas a factores externos sigan una distribución exponencial, mientras que dicha distribución varía de individuo a individuo según sus características individuales mediante una distribución beta generalizada, si bien esta asignación es arbitraria. Es decir, se puede considerar que la distribución exponencial modeliza los factores internos y la distribución beta generalizada los factores externos.

Así, se puede obtener una descomposición de la varianza debido a su génesis como mixtura para $\rho > -1$ de la siguiente forma:

$$Var\left(X\right) = E_P\left(V\right) + E_P\left(V^2\right) + Var_P\left(V\right). \tag{6}$$

El primero de los términos en esta suma estaría relacionado con aquella parte de la variabilidad debida al azar, el segundo término se interpreta en relación con la variabilidad debida a factores externos que afectan a la población o *riesgo* y el tercero en relación a las diferencias entre los individuos o *propensión* (Rodríguez et al, 2007). Hay que señalar que esta asignación para los dos últimos términos se intercambia si consideramos la otra modelización indicada. Extrapolando al contexto del superestrellato, nosotros consideramos que el factor propensión hace referencia al talento o destreza, mientras que el riesgo estaría relacionada con factores externos que inciden en el éxito del mismo.

Además, dado que $Var_P\left(V\right) < E_P\left(V^2\right)$, la parte debida a la propensión es siempre menor que la debida al riesgo[1]. Veremos las connotaciones que esta interpretación puede tener en las aplicaciones del superestrellato.

Otra consecuencia destacable de este hecho es que esta distribución es sobredispersa para $\rho > -1$ al ser mixtura de una distribución de Poisson. Por su parte, si $\rho \leq -1$, pueden darse situaciones de infradispersión.

Por último, una interpretación del parámetro $\lambda$ surge al considerar que

$$\lim_{k\to\infty} \frac{f(k+1)}{f(k)} = \lim_{k\to\infty} \frac{(1+k)}{(\rho+2+k)}\lambda = \lambda,$$

por lo que está relacionada a la pendiente de la cola. El caso $\lambda = 1$, que corresponde con la distribución de Yule, es un caso límite donde el decrecimiento de la cola es el más lento, por lo que es especialmente apropiado para conjuntos de datos con colas pesadas o varianza infinita. Por su parte, $\lambda < 1$ produce un decrecimiento más brusco en las probabilidades de la cola, que es menos pesada. Este hecho podría resultar relevante en las aplicaciones a datos reales, porque una de las características del superstardom es que se supone que hay pocos artistas con mucho éxito, lo cual se traduce en que las variables consideradas tendrían colas relativamente grandes. Sin embargo, según parece de los resultados obtenidos en las aplicaciones, estas colas no son tan pesadas como para que sean ajustadas bien mediante la distribución de Yule.

[1]O bien la parte del riesgo es siempre menor que la debida a la propensión, ya que como hemos indicado anteriormente es posible considerar otra asignación para el riesgo y la propensión.

### 2.2. Inferencia

La estimación de los parámetros se puede llevar a cabo mediante máxima verosimilitud o bien mediante el método de los momentos. En este trabajo hemos utilizado únicamente la estimación máximo verosímil. La función de log-verosimilitud es:

$$\ln L = n\ln(\rho+1) + \left(\sum_{i=1}^{n} x_i - n\right)\ln\lambda + \sum_{i=1}^{n}\ln B(x_i,\rho+1) - n\ln\,{}_2F_1(1,1;\rho+2;\lambda),$$

donde $x_i$, $i = 1,\ldots,n$ son los datos. Dado que las ecuaciones de verosimilitud no tienen soluciones explícitas, las estimaciones máximo-verosímiles se han obtenido mediante la optimización de la función de log-verosimilitud (función *optim* de R).

Para valorar la bondad de los ajustes hemos utilizado el estadístico $\chi^2$. Teniendo en cuenta que las colas de las distribuciones ajustadas suelen ser muy pesadas, no es conveniente utilizar la distribución $\chi^2$ como distribución asintótica en el muestreo bajo la hipótesis nula, ya que es usual que haya muchas frecuencias esperadas inferiores a 5. En su lugar, hemos considerado el valor del estadístico sin realizar ninguna agrupación de frecuencias, y se ha calculado el p-valor asociado mediante un test de Monte Carlo con $B = 2000$ réplicas.

## 3. Aplicación

En el artículo de Cox y Kleinman (2000) se estudia el *fenómeno superestrella* en el contexto del mecanismo probabilístico subyacente en la selección de analistas de mercado (security analysts) para su inclusión en el *Intitutional Investor's* All-American Research Team (II AART). El II AART es elegido como una medida del prestigio del analista de mercado, siendo un importante determinante de su remuneración y notoriedad. Sin embargo, algunos analistas sostienen en privado que la lista es poco más que un concurso de popularidad y no refleja diferencias en el talento.

Para comprobar este hecho, emplean la distribución de Yule con parámetro $\rho = 1$[2], encontrando que dicha distribución no describe los datos empíricos en el sector del análisis de mercado (security analysis industry). Así, sus resultados confirman la idea de que los analistas de mercado más exitosos logran tales resultados debido a su destreza más que al azar o la suerte, y coinciden con las conclusiones obtenidas con trabajos anteriores.

Los datos para este estudio fueron obtenidos de *II AART first team lists* registradas anualmente en *Intitutional Investor's*, para los años 1972 a 1996, la cual contiene 328 analistas de mercado. El número de analistas en el primer equipo ha variado de 26 en 1972 a 81 en 1996. La distribución de frecuencias de los analistas de mercado se muestra en la tabla 2.

[2]Hay que señalar que no estiman el parámetro.

Hemos ajustado tanto la distribución de Yule como la distribución EYule por máxima verosimilitud. En la tabla 1 se resumen los resultados obtenidos, mientras que en la tabla 2 se muestran las frecuencias esperadas. Se aprecia que efectivamente la distribución de Yule no modeliza adecuadamente los datos mientras que la distribución EYule sí proporciona un buen ajuste.

| Distribución ajustada | $\hat{\rho}$ | $\hat{\lambda}$ | $\chi^2$ | p-value |
|---|---|---|---|---|
| $Yule(\rho)$ | 0.8666 (0.0587) | - | 123.2561 | 0.0005 |
| $EYule(\rho, \lambda)$ | -0.7889 (0.12997) | 0.7806 (0.0257) | 15.4693 | 0.6647 |
| | $\hat{\beta}$ | $\hat{\rho}$ | | |
| $Waring(\beta, \rho)$ | 43.1835 (50.6661) | 16.0099 (17.7132) | 17.9261 | 0.5197 |

Cuadro 1: Estimaciones de los parámetros (errores estandar entre paréntesis) y resultados del test $\chi^2$. Los p-valores han sido calculados mediante el test de Monte Carlo con $B = 2000$ réplicas

Así mismo hemos obtenido la partición de la varianza para la distribución $EYule$ ajustada. Hay que señalar que para la distribución de Yule no existe varianza dado que $\widehat{\rho} < 2$. Los resultados se muestran en la tabla 3, donde se aprecia que la mayor parte de la variabilidad no se debe al azar sino a factores externos y a diferencias propias de los analistas, pero en función de la modelización considerada podemos considerar que o bien la destreza determina esa variabilidad o bien son los factores externos.

Por último, se han comparado los ajustes con los obtenidos con otra distribución biparamétrica, la distribución de Waring de parámetros $\beta$ y $\rho$, y que también pertenece a la familia GHD (Johnson *et. al*, 2005); concretamente, se trata de una $GHDI(1, \beta, \beta + \rho + 1, 1)$ (la distribución de Yule es un caso particular para $\beta = 1$). Como vemos, el p-valor del test de bondad de ajuste (tabla 1) nos indica que dicha distribución se ajusta bastante bien a los datos, si bien tiene el problema de que las estimaciones de los parámetros no son nada precisas, dado el elevado valor de sus errores estándar. Esta distribución también tiene una génesis similar a la EYule, obteniéndose como mixtura de la distribución de Poisson, y verificando una partición de la varianza como la dada para la distribución EYule. Los resultados obtenidos para la misma, no difieren en exceso de los encontrados para la distribución EYule, proporcionando una interpretación similar de la descomposición de la varianza.

## 4. Conclusiones

¿Son las diferencias en el talento las que determinan la existencia de "superestrellas" en el campo del análisis de mercado? ¿O estos existen por otras causas? La partición de la varianza proporcionada por la distribución EYule no es capaz de responder a estas

| | | Esperadas | | |
|---|---|---|---|---|
| Nº de apariciones | Observadas | EYule | Yule | Waring |
| 1 | 95 | 95.08 | 152.28 | 88.71 |
| 2 | 66 | 61.29 | 53.12 | 63.64 |
| 3 | 37 | 43.27 | 27.48 | 45.95 |
| 4 | 33 | 31.56 | 16.94 | 33.38 |
| 5 | 17 | 23.40 | 11.55 | 24.40 |
| 6 | 21 | 17.53 | 8.41 | 17.93 |
| 7 | 18 | 13.22 | 6.41 | 13.25 |
| 8 | 8 | 10.02 | 5.06 | 9.85 |
| 9 | 8 | 7.62 | 4.11 | 7.36 |
| 10 | 7 | 5.81 | 3.40 | 5.52 |
| 11 | 2 | 4.44 | 2.87 | 4.16 |
| 12 | 2 | 3.40 | 2.45 | 3.15 |
| 13 | 3 | 2.61 | 2.12 | 2.40 |
| 14 | 4 | 2.00 | 1.85 | 1.84 |
| 15 | 1 | 1.54 | 1.64 | 1.41 |
| 16 | 3 | 1.19 | 1.45 | 1.09 |
| 17 | 1 | 0.91 | 1.30 | 0.84 |
| 18 | 1 | 0.70 | 1.17 | 0.65 |
| 19 | 1 | 0.54 | 1.06 | 0.51 |
| 20 ó más | 0 | 1.85 | 23.32 | 1.94 |

Cuadro 2: Frecuencias observadas del número de analistas y frecuencias esperadas en el ajuste mediante una $Yule\,(\rho)$ y una $EYule\,(\rho, \lambda)$ por máxima verosimilitud

cuestiones a menos que tomemos partido por una de las dos posibles modelizaciones de dicha distribución. Atendiendo a otros estudios empíricos en este campo de aplicación, parece más razonable considerar que la elección de los analistas de mercado en el II AART se basa en su destreza.

En cualquier caso hemos comprobado que dicha elección no se basa en el azar o la suerte. Sin embargo, la controversia sobre la naturaleza del superestrellato no queda totalmente resuelta de forma objetiva. Si fuera posible asumir que los factores externos son similares para todos los individuos analizados, la parte de la varianza debida al riesgo sería menor que la debida a la propensión, pero esto se basaría en la opinión del experto. No queda claro que dicha suposición se pueda establecer en la aplicación analizada, dado que los datos se han recogido durante 25 años, por lo que aquellos analistas de mayor edad o con más años de dedicación en esta actividad pueden ser seleccionados con una mayor probabilidad, invalidando de este modo la idea de que las condiciones son similares

| | Aleatoriedad | | Riesgo | | Propensión | | |
|---|---|---|---|---|---|---|---|
| | Total | % | Total | % | Total | % | Varianza |
| EYule | 2.8750 | 22.1388 | 9.1885 | 70.7552 | 0.9228 | 7.1060 | 12.9863 |
| Waring | 3.0824 | 21.2317 | 10.4682 | 72.1060 | 0.9672 | 6.6623 | 14.5177 |

Cuadro 3: Partición de la varianza

para todos los individuos.

Sería interesante considerar un modelo de regresión donde sea posible acabar con la indeterminación sobre la asignación del modelo exponencial y beta generalizado al riesgo y la propensión, de forma similar al establecido para el modelo Waring por Rodríguez *et al* (2009).

## Referencias

[1] Chung, K.H. and Cox, R.A.K. (1994). A stochastic model of superstardom: an application of the Yule distribution. *Review of Economics and Statistics* 76, 771-775.

[2] Chung, K.H. and Cox, R.A.K. (1998). Consumer behavior and superstardom. *Journal of Socio-Economics* 27, 263-270.

[3] Cox, R.A.K. and Kleiman, R.T. (2000). A stochastic model of superstardom: evidence from Institutional Investor's All-American Research Team. *Review of Financial Economics* 9, 43-53.

[4] Fox, M.A. and Kochanowski, P. (2004). Models of superstardom: an application of the Lotka and Yule distributions. *Popular Music and Society*, 27(4), 507-522.

[5] Giles, D.E. (2006). Superstardom in the US popular music industry revisited. *Economics Letters* 92, 68-74.

[6] Johnson, N.L., Kemp, A.W, and Kotz, S. (2005). *Univariate discrete distributions.* Wiley, New York. Third edition.

[7] Martínez-Rodríguez, A.M., Sáez-Castillo, A.J., and Conde-Sánchez, A. (2011). Modelling using an extended Yule distribution. *Computational Statistics and Data Analysis* 55(12), 863-873.

[8] Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J. and Olmo-Jiménez, M. J. (2007). A new generalization of the Waring distribution. *Computational Statistics and Data Analysis* 51(12), 6138-6150.

[9] Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J. and Olmo-Jiménez, M. J. (2009). A Generalized Waring Regression Model for Count Data. *Computational Statistics and Data Analysis* 53(10), 3717-3725.

[10] Serenko, A., Cox, R.A.K., Bontis, N. and Booker, L.D. (2011) The superstar phenomenon in the knowledge management and intellectual capital academic discipline. *Journal of Informetrics* 5, 333–345.

[11] Spierdijk, L. and Voorneveld, M., (2009). Superstars without talent? The Yule distribution controversy, *The Review of Economics and Statistics* 91(3), 648–652.