SOBRE ESTADÍSTICAS DE CITAS A TRABAJOS CIENTÍFICOS

Pablo Kittl Duclout y Jorge Gibert Galassi

RESUMEN

Se discute el clásico trabajo de De Solla Price (1976), junto a otros trabajos sobre cienciometría ó estadística de citas, considerando los aportes en el campo de estudio con el objetivo de llegar al mínimo de hipótesis y fórmulas para representar los datos experimentales. Se reivindica una formulación basada en la clásica distribución de Vilfredo Pareto para establecer una adecuada representación de los datos. Finalmente, se presenta una interpretación discontinua de la citación científica, fundada en el supuesto que existe una dinámica de citación que caracteriza a un grupo A de trabajos en términos de capacidad para generar N citaciones que lo distingue de un grupo B capaz de producir N+1.

ON SCIENTIFIC PAPERS CITATION

Pablo Kittl Duclout and Jorge Gibert Galassi

SUMMARY

The paper focuses on De Solla Price's classic work from 1976, and discusses in relation with others contributions, its role within the field of research. The aim is to arrive to a minimum of hypothesis and formulas to represent experimental data properly. We vindicate a formulation based on classical distribution by Vilfredo Pareto to do so. Finally, a discontinuous interpretation for scientific citation is presented, with foundations in the presumptions that a given group of papers A is different from another group B due to the capacity to produce N citations by one group as well as some different capacity to produce N+1 by another group.

Introducción

Derek de Solla Price (1963), siguiendo la ley de Lotka (1926), estableció que "el número de científicos que producen 'n' trabajos es aproximadamente proporcional a $1/n^2$. Esta ley de la productividad del inverso al cuadrado estima que por cada 100 autores que producen 1 trabajo científico, hay solo 25 que producen 2, 11 que producen 3 y así sucesivamente". Esta formulación, citada por Cole y Cole (1973, p. 218), y otras similares, ha mostrado ser cuestionable en la actualidad, dada la intensidad de la colaboración entre equipos de investigación de múltiples universidades (Jones *et al.,* 2008) y la extensión de la coautoría a un número muy grande de autores, que dificulta dimensionar a la contribución individual.

PALABRAS CLAVE / Citación Científica / De Solla Price / Estadística / Pareto /

Recibido: 06/10/2012. Modificado: 02/04/2014. Aceptado: 05/04/2014.

Pablo Kittl Duclout. Licenciado en Física, Universidad Nacional de Cuyo, Argentina. Profesor, Universidad de Chile.

Jorge Gibert Galassi. Sociólogo y Doctor en Filosofía, Universidad de Chile, Chile. Profesor, Universidad de Valparaíso, Chile. Dirección: Facultad de Ciencias Económicas y Administrativas, UV, Las Heras 6, Valparaíso, Chile. e-mail: jorge.gibert@uv.cl

ESTATÍSTICAS SOBRE CITAÇÃO A TRABALHOS CIENTÍFICOS

Pablo Kittl Duclout e Jorge Gibert Galassi

RESUMO

Discute-se o clássico trabalho de De Solla Price (1976), junto a outros trabalhos sobre cienciometria ou estatística de citações, considerando as contribuições no campo de estudo com o objetivo de chegar ao mínimo de hipóteses e fórmulas para representar os dados experimentais. Reivindica-se uma formulação baseada na clássica distribuição de Vilfredo Pareto para estabelecer uma adequada representação dos dados. Finalmente, se apresenta uma interpretação descontinua da citação científica, fundada no suposto que existe uma dinâmica de citação que caracteriza a um grupo A de trabalhos em termos de capacidade para gerar N citações que o distingue de um grupo B capaz de produzir N+1.

A pesar de ello, en los últimos años, numerosos estudios han mostrado la utilidad de la citación Revista - Revista (JCR) para mostrar la situación de productividad científica de disciplinas, instituciones y personas, así como de las redes existentes y la organización de los campos de estudio (Dorein y Fararo, 1985; Borgman y Rice, 1992; Tijssen, 1992; Cozzens y Leydesdorff, 1993; Besselaar y Leydesdorff, 1996). Las estructuras de citas han mostrado que hay jerarquías de nivel entre revistas al interior de los campos disciplinarios, pero con excepciones, como Nature y Science, cuyo nivel jerárquico es el más alto a pesar de ser revistas misceláneas (Carpenter y Narin, 1973; Burt, 1982; Knoke y Kuklinsky, 1982; Dorein y Fararo, 1985; Doreian, 1986; Leydesdorff, 1986; Tijssen *et al*., 1987; Borgman y Rice, 1992; McCain y Whitney, 1994). Sin embargo, aunque la temática se desarrolla vigorosamente, no se ha discutido la naturaleza ni el alcance de los trabajos clásicos sobre el tópico.

En este trabajo se estudia el 'mecanismo de contagio' que adopta el clásico trabajo cienciométrico de De Solla Price (1976) y se postula que la formulación de Pareto es más simple y adecuada.

Supuestos Distintos, para Llegar a Distintas Formulaciones

Como se sabe, una medida de la importancia de un trabajo o publicación, es el número de veces que este trabajo está citado a través del tiempo. Este punto de vista pudo implementarse desde 1960, cuando Eugene Garfield fundó el *Institute of Scientific Information*, cuya metodología facilitó la captura de los trabajos que se publican en revistas que ellos consideran serias, donde los trabajos publicados fueron juzgados por personas que se supone de reconocida solvencia en los temas que ellos tratan.

En la presente nota veremos solo algunos trabajos publicados sobre el tema y trataremos de analizar en forma específica el trabajo de De Solla Price (1976) y, secundariamente, los trabajos de Dieulefait (1942), De Solla Price (1965), Kittl *et al*. (1995) y Acevedo *et al*. (2007).

En la formulación de Dieulefait (1942) se llega a la estadística con la función B de Euler, que tiene la forma

$$B(a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}=\int_0^1 x^{a-1}(1-x)^{b-1}dx$$

$$\Gamma(a)=\int_0^\infty e^{-x}x^{a-1}dx \qquad (1,1)$$

La distribución B, a la que se llega por un 'mecanismo de contagio' que repite De Solla Price (1976), es el esquema de contagio que se debe originariamente a George Polya. Las citas de un trabajo en otros es una especie de contagio. Pero las expresiones derivadas de (1,1) se simplifican enormemente si en el trabajo de De Solla Price se introduce implícitamente la hipótesis

$$\frac{\frac{dF(P)}{F(P)}}{\frac{dP}{P}}=-\alpha=\text{constante};\ \alpha>0 \qquad (2,1)$$

donde F es proporcional al número de entidades que tienen la cualidad P. La expresión (2,1) significa que a un aumento en la cantidad de la cualidad específica P, o sea dF(P)/F(P), corresponde una disminución en la cualidad específica dP/P que la tiene. Cuanto más condiciones hay que cumplir, más difícil es el ascenso. En Acevedo *et al*. (2007) se partió de este supuesto básico para obtener la ley de Pareto (1896), por una simple integración

$$F=\begin{cases}\dfrac{C}{P^{\alpha}}, & P_o\le P,\ C=\text{constante}\\ 0, & 0\le P<P_o\end{cases} \qquad (3,1)$$

La expresión en (3,1) significa que no hay población en el tramo $0\le P<Po$, lo cual aplicado a las rentas sería que nadie puede vivir con menos de Po. Lo que es muy importante es que la repartición de las rentas tiene una forma muy parecida para todos los países y en todas las épocas porque $\alpha\approx1,5$. En nuestro caso, P es el número n de veces que un trabajo está citado en N trabajos, así que (3,1) se transforma en

$$N=\begin{cases}\dfrac{N_i}{n^{\alpha}}, & 1\le n\\ 0, & 0\le n<1\\ N=N_i, & n=1\end{cases} \qquad (4,1)$$

Una Discusión sobre Aplicaciones Cienciométricas

En el trabajo de De Solla Price (1976) aparece el número N de trabajos citados n veces en el *Science Citation Index* de 1975 y otros. Nos ocuparemos solo de una parte, que corresponde a 6157 trabajos del índice de 1975. Tomando logaritmos en (4,1) tenemos

$$\ln\frac{N}{N_i}=-\alpha\ln n \qquad (1,2)$$

La recta (1,2) pasa por los puntos ($N_i=10^4$; n=1)* y (N_f=2031,8; n_f=2); (N_f=808,8; n_f=3); (N_f=276; n_f=4); (N_f=121,8; n_f=5); (N_f=55,2; n_f=6), para cada uno de los datos se calculó

$$\alpha=-\frac{\ln\frac{N_f}{N_i}}{\ln n} \qquad (2,2)$$

y se obtuvo un promedio α= 2,48. En el caso de una representación geométrica tomando como eje lnN y ln n se obtuvo una recta que pasa entre medio de los puntos con pendiente α= 1,33.

Los valores de la tabla de De Solla Price (1976), normalizados en la Figura 1, permiten apreciar que para los valores observados, frente a sus valores estimados por el método de los mínimos cuadrados, el coeficiente de correlación R^2 es igual a 1.

En el trabajo de Kittl *et al*. (1995) se presenta un modelo probabilístico de tiro de una moneda contando las partidas que terminan cuando luego de n jugadas de secas o caras uno saca una desigual n+ caras o secas; en este caso, α= 1,64. Se estudió también las citas de un libro y se obtuvo α= 2,12. También se encontraron números muy parecidos en Acevedo *et al*. (2007).

La Reivindicación de Pareto y una Propuesta de Interpretación

En el trabajo de De Solla Price (1976) se representa en escala doble logarítmica a N y n. Claramente, las rectas con una quebradura indican la existencia de dos poblaciones. Así que de esto se deduce que hay que separarlas trabajando con rectas determinadas por mínimos cuadrados y por sobre todo aplicar un criterio como el coeficiente de correlación R^2 para estimar la bondad del ajuste. No se explica en el trabajo de De Solla Price qué utilidad tiene la tabla de B (p,q). Bastando para representar el fenómeno la ecuación (3,1).

En la Figura 1 se puede ver como siguen la ley de Pareto. Entre n=1 y n=3 tenemos $\alpha \approx 2{,}29$ y entre n=3 y n=6 es $\alpha \approx 3{,}87$. Se puede decir que en el presente ejemplo a partir de 3 citas es mucho mas difícil tener más citas.

Como se puede observar en las Figuras 2 y 3; se constatan cuatro agrupaciones de trabajos. La primera, trabajos que generan entre 1 y 3 citaciones (Figura 2, izquierda), cuyo coeficiente de correlación es perfecto, esto es, $R^2=1$. La segunda agrupación, constituida por trabajos que generan entre 3 y 7 citaciones, con coeficiente de correlación $R^2=0{,}99$ (Figura 2, derecha). Es decir, podemos hipotetizar que existen dos familias de trabajos entre aquellos que producen entre 1 y 7 citaciones. Hay un quiebre, es decir, hay dos rectas; cuyo límite son las 3 citaciones: aquellos trabajos que pueden sobrepasar ese número adquieren la dinámica de otro grupo, el grupo de aquellos trabajos capaces de generar entre 3 y 7 citaciones. Lo mismo pasa si seguimos el ejercicio: se produce otro quiebre, el de las rectas correspondientes a la agrupación de trabajos que producen entre 8 a 10 citaciones (Figura 3, izquierda) y el de trabajos que generan entre 12 a 14 citaciones (Figura 3, derecha). Entre estos últimos dos grupos no hay límite, sino un salto, que se expresa en que aquellos trabajos que producen 11 citaciones se ubican fuera de las rectas del grupo de 8-10 citaciones y del grupo de 12-14 citas. Siendo así, podríamos conjeturar que este último grupo posee una propiedad cualitativa distinta; mientras que en los tres primeros grupos, se podría hablar de magnitudes de una misma (o mismas) cualidades: sólo expresarían grados de magnitud distintos.

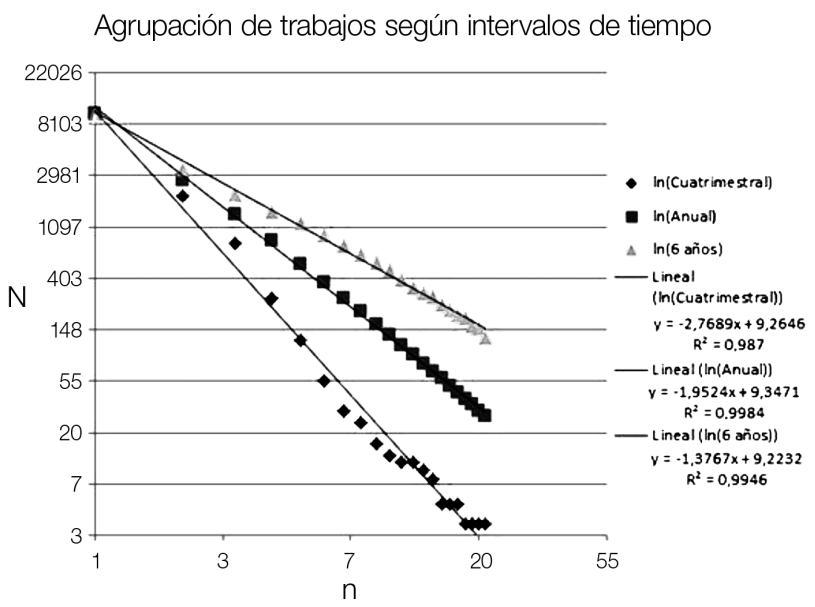

Figura 1: Número de trabajos N que tienen n citas en escala logarítmica, que permite la agrupación de investigadores según intervalos de tiempo (Cuatrimestral; Anual; 6 años)

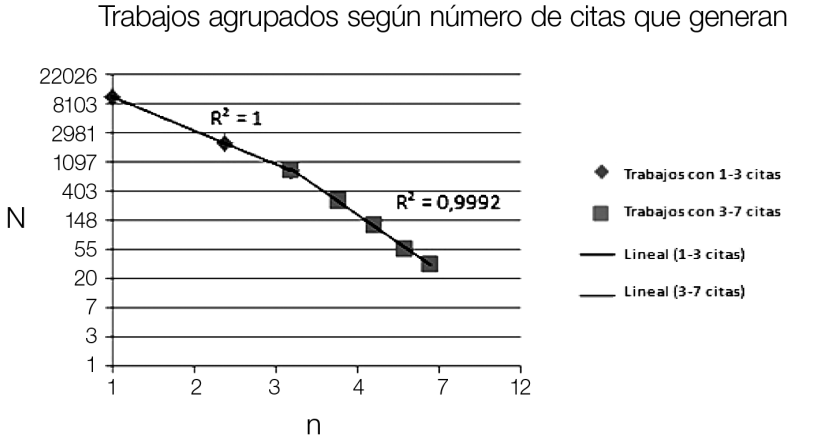

Figura 2: Número de trabajos N que tienen n citas en escala logarítmica, correspondiente a la agrupación de trabajos que producen entre 1 y 3 citaciones (cuatrimestral) y a la agrupación de trabajos que producen entre 3 y 7 citaciones (cuatrimestral).

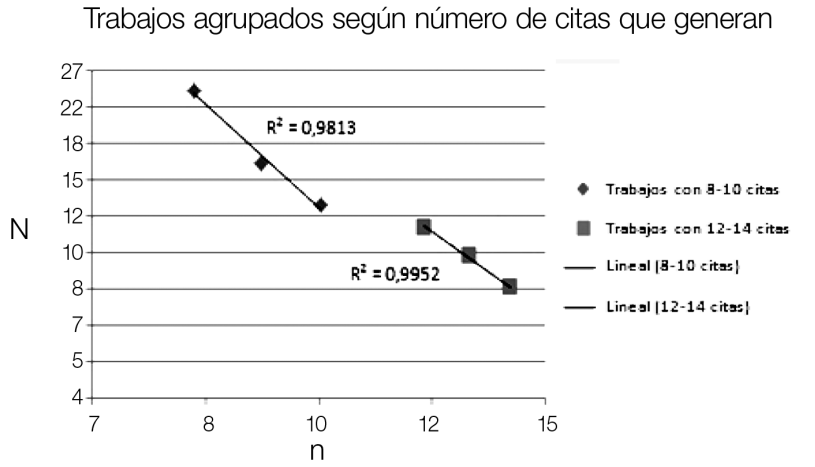

Figura 3: Número de trabajos N que tienen n citas en escala logarítmica, correspondiente a la agrupación de trabajos que producen entre 8-10 citaciones (cuatrimestral) y a la agrupación de trabajos que producen entre 12-14 citaciones (cuatrimestral).

A Modo de Conclusiones

Si el objetivo de una teoría estadística es describir datos experimentales y, en lo posible, predecir otros, parece inútil reemplazar una estadística que describe con un alto grado de correlación datos experimentales, para ensayar otras estadísticas con un mayor número de parámetros. En este caso la estadística de Pareto, con dos parámetros Ni y α se adapta con un coeficiente de correlación muy elevado, la estadística con B(a,b) y Ni tiene tres parámetros y puede adaptarse con un mayor grado de correlación, aunque en este caso no se estudió el coeficiente de correlación ni la prueba de χ^2 en el trabajo de De Solla Price, puesto que parece muy difícil o inútil porque el coeficiente de correlación en la estadística

de Pareto es prácticamente 1. Se puede concluir que De Solla Price no utilizó la formulación más simple para analizar la data. Se podría especular sobre lo que impidió a De Solla Price desarrollar la teoría de una forma más simple. Lo cierto es que no lo hizo.

Debido probablemente a que su motivación era la descripción general del fenómeno, tampoco se percató que la interpretación más adecuada de los datos implicaba reconocer que la mejor conjetura era que no había solo una recta, sino varias rectas, correspondientes a varias familias o agrupaciones de trabajos con diferentes capacidades para generar citaciones. Desde esa conjetura, se podrían haber desarrollado otras teorías. Nuestra posición al respecto es que esta revisión de los datos usados por De Solla Price, así como probablemente otros datos, indican que es plausible la siguiente conjetura: la citación científica es un indicador discontinuo. Es decir, no existe una linealidad en el volumen de citas y, por tanto, hay grupos de trabajos que 'característicamente' producen de 1 a 3 citaciones por año y que no serán capaces de generar 4 o más citaciones, salvo la mediación de factores extraordinarios. Indudablemente, no queremos insinuar que existe una jerarquía como la descrita por A. Huxley en el conjunto de trabajos científicos publicados, pero sí que las citaciones se relacionan con otras características cualitativas (¿la creatividad?) y que ello implica que existen barreras entre un grupo de trabajos y otro. Así, se elabora una conjetura sobre la citación científica, como un fenómeno no lineal: una interpretación discontinua de la citación científica.

REFERENCIAS

Acevedo R, Díaz G, Kittl P (2007) *Statistics of Quotations Reported by the Institute for Scientific Information (ISI). A Working Example of a Chilean Institution*. Universidad Mayor. Santiago, Chile. www.ingenews.cl/web/download/publicaciones/Estadisticas-ISI.pdf

Borgman CL, Rice RE (1992) The convergence of information science and communication: A bibliometric analysis. *J. Am. Soc. Inf. Sci. 43*: 397-411.

Burt RS (1982) *Towards a Structural Theory of Action*. Academic Press. Londres, RU. 381 pp.

Carpenter MP, Narin F (1973) Clustering of scientific journals. . *J. Am. Soc. Inf. Sci. 24*: 425-436.

Cole J, Cole S (1973) *Social Stratification in Science*. University of Chicago Press. Chicago, IL, EEUU. 283 pp.

Cozzens SE, Leydesdorff L (1993) Journal systems as macro indicators of structural change in the sciences. En Van Raan AFJ (Ed.) *Proc. Joint EC/ Leiden Workshop on Science and Technology Indicators*: DSWO Press. Leiden University. Leiden, Holanda. pp. 219-233.

De Solla Price DJ (1963) *Little Science, Big Science*. Columbia University press: Nueva York, EEUU. 119 pp.

De Solla Price DJ (1965) Network of scientific papers. *Science 149*: 510-515.

De Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci. 27*: 292-306.

Dieulefait CE (1942) *Elementos de Estadística Metodológica*. Fasc. 4, Caps IX y X. Instituto de Estadística. Rosario, Argentina.

Doreian P (1986) A revised measure of standing of journals in stratified networks. *Scientometrics 11*: 63-72.

Doreian P, Fararo TJ (1985) Structural equivalence in a journal network. *J. Am. Soc. Inf. Sci. 36*: 28-37.

Jones BF, Wuchy S, Uzzi B (2008): Multi-University research teams: shifting impact, geography, and stratification in science. *Science 322*: 1259-1262.

Kittl P, Díaz G, Gibert J (1995): Lógica y conocimiento científico [1972]. En *El Desarrollo Científico y Tecnológico, Particularmente en Chile*. Mimeo. Santiago. Chile.

Knoke D, Kuklinsky JH (1982) *Network Analysis*. Sage. Beverly Hills, CA, EEUU. 95 pp.

Leydesdorff L (1986) The development of frames of references. *Scientometrics 9*: 103-125.

Lotka AJ (1926) The frequency distribution of scientific productivity. *J. Wash. Acad. Sci. 16*: 17-23.

McCain KW, Whitney PJ (1994) Contrasting assessments of interdisciplinarity in emerging specialties, the case of neural network research. *Knowl. Creat. Diffus. Util. 15*: 285-306.

Pareto W (1896) *Cours d'Economie Politique*. Tomo II, Libro III. Lausanne, Suiza.

Tijssen RJW (1992) *Cartography of Science: Scientometric Mpping with Multidimensional Scaling Methods*. DSWO Press. Leiden University. Leiden, Holanda.

Tijssen RJW, De Leeuw J, Van Raan AFJ (1987) Cuasi-correspondence analysis on square scientometric transaction matrices. *Scientometrics 11*: 347-361.

Van Den Besselaar P, Leydesdorff L (1996) Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence. *J. Am. Soc. Inf. Sci. 47*: 415-436.