

PRESENÇA DE *OUTLIERS* NA LEI DO ELITISMO

Dávilla Veira Odizio Silva (UNIR)

davilla.odizio@gmail.com

Alexandre Masson Maroldi (UNIR)

alexandremaroldi@yahoo.com.br

Luis Fernando Maia Lima (UNIR)

maialima2000@gmail.com

EIXO TEMÁTICO: Métodos Técnicas e Ferramentas para Estudos
Bibliométricos e Cientométricos

MODALIDADE: Apresentação oral

1 INTRODUÇÃO

Ao compulsamos a literatura acerca do que é a ciência, Schwartzman (apud CARVALHO, 2011, p.5) define que a ciência é “um conjunto de conhecimentos a respeito das coisas, conhecimentos que se desenvolvem, acumulam, se transformam e se reestruturam em função de uma lógica própria de uma organização do conhecimento – seu *logos*”

Nesse aspecto acumulativo do conhecimento, vemos que Vanti (2012) assevera que a ciência, para atingir o nível de desenvolvimento e de credibilidade que apresenta nos dias de hoje, buscou, nos números, os alicerces para as suas teorias. A preocupação com a mensuração na ciência vem de tempos remotos. Teve início na Antiguidade, quando os filósofos, matemáticos e astrônomos tentavam dimensionar a distância entre as estrelas para, a partir daí, estabelecer as medidas de tempo em dias, meses, anos e em horas e minutos.

Um dos primeiros autores a pensar a ciência de forma mensurável e aplicar a ela o seu próprio método foi Price (1976, p.3), o qual detectou que o “crescimento da ciência é surpreendentemente rápido” e exponencial, identificando que a mesma quadruplica a cada geração e a literatura científica dobra num período de 10 a 15 anos.

Price (1976) conjecturou, com base na Lei de Lotka, que um critério para separar os autores mais profícuos (elite) dos menos profícuos de determinado campo de conhecimento seria dada pela raiz quadrada do número total de autores, o qual foi titulado Lei do Elitismo ou Lei de Price.

Desta forma, estudar a Lei do Elitismo torna-se parte integrante da quantificação da produtividade científica, que é normalmente mensurada em termos de trabalhos publicados (URBIZAGÁSTEGUI ALVARADO, 2009) de onde se originam os cálculos necessários para a medição desta produtividade dos autores.

Contudo, Ravichandra Rao (1986), corroborado por Nicholls (1988), pondera que a alegação de Price é baseada em poucas evidências, e recomenda testes estatísticos para validar a Lei do elitismo. Uma primeira questão central é que o resultado numérico da raiz quadrada do total de autores pode gerar um número não inteiro, ou seja, caso haja um total de 216 autores, então a raiz quadrada é 14,7 autores, surgindo as seguintes seguintes situações (NICHOLLS, 1988): o bibliometrista deverá arredondar o valor encontrado de 14,7 autores para mais ou para menos, e, além disto, pode ocorrer de o valor teórico obtido (14,7 autores) não seja de fácil identificação aos dados reais.

Considerando-se a distribuição assimétrica da Lei de Lotka (em forma de “J” invertido) que serve de base ao raciocínio de Price, há a ocorrência de valores extremos, que podem ser tratados como *outliers* ou valores discrepantes (DE BELLIS, 2009), pois apresentam comportamento que estão bem distante (fora do padrão geral) dos demais dados (TRIOLA, 2008).

Diante dessas dificuldades de adequar o valor teórico obtido pelo método de Price aos dados reais, e assumindo como hipótese que *outliers* podem ser um indicador de autores profícuos cabe então a seguinte questão: como os *outliers* podem contribuir na Lei do Elitismo?

2 A LEI DO ELITISMO NOS CÁLCULOS BIBLIOMÉTRICOS

Atualmente os estudos bibliométricos podem ser considerados indicadores da produção científica e são necessários para se compreender a dinâmica e a evolução da ciência. Nesse sentido os indicadores bibliométricos vêm sendo utilizados como instrumentos para análise da atividade científica e das suas relações com o desenvolvimento econômico e social, bem como para se obter as características de um determinado campo do conhecimento.

O emprego de indicadores bibliométricos apresenta uma série de vantagens para sua utilização na avaliação científica, por se tratar de dados verificáveis, reproduzíveis, que apresentam resultados objetivos, numéricos, além de poderem ser aplicados a um grande volume de dados. (IGAMI, 2011).

A bibliometria possui três importantes Leis que contribuíram para ao avanço desse saber: Lei de Lotka (mede a produtividade dos pesquisadores); Lei de Bradford (Lei de dispersão do conhecimento) e Lei de Zipf (modelo de distribuição e frequência de palavras de um texto).

A primeira Lei, a Lei de Lotka foi formulada em 1926 e elaborada a partir de um estudo sobre a produtividade de cientistas a partir da contagem de autores presentes no *Chemical Abstracts*, onde posteriormente a esses estudos estabeleceu os fundamentos da lei do quadrado inverso, afirmando que o número de autores que fazem n contribuições em um determinado campo científico é aproximadamente $1/n^2$ daqueles que fazem uma só contribuição e que a proporção daqueles que fazem uma única contribuição é de mais ou menos 60%. (URBIZAGÁSTEGUI ALVARADO, 2002).

Baseado na Lei de Lotka, Price (1976) aperfeiçoou-a a partir de seus estudos realizados. Price (1976) concluiu que 1/3 da literatura é produzida de menos de 1/10 dos autores mais produtivos, levando a uma média de 3,5 documentos por autor e 60% dos autores produzindo um único documento. Dessa forma propõe a Lei do Elitismo onde se k representa o número total de contribuintes numa disciplina, $\sqrt{k}$ representaria a elite da área estudada.

Todavia, a Lei do Elitismo apresenta alguns problemas para a sua aplicação, o qual Nicholls (1988) esclarece: independentemente do valor teórico advindo de “ $\sqrt{k}$ ” (“ k ” total de autores), há o problema de se adequar aos dados reais. Ou seja, qualquer que seja o valor teórico de “ $\sqrt{k}$ ”, mesmo que arredondado (seja para mais, seja para menos), obrigatoriamente encontra-se dentro de umas das faixas da distribuição de frequência, portanto, gerando o problema de escolher quais de fato são os autores mais profícuos.

Nicholls (1988) utiliza em sua pesquisa os dados originados dos estudos de Dresden sobre a produtividade de matemáticos americanos onde há um total de 278 autores. A raiz quadrada deste valor de acordo com a Lei do Elitismo fornece 16,7 autores como a elite. Assim, o primeiro problema que emerge é arredondar para menos (16 autores) ou para mais (17 autores). Mas, ao consultar os dados reais, aparece o problema de adequação do valor teórico: há 15 autores (portanto abaixo de 16 autores) que produziram mais de 13 trabalhos por autor; por outro lado, há 20 autores (logo acima de 17 autores) que produziram mais de 12 trabalhos por autor.

Logo, onde está a elite? Nos 15 autores? Nos 20 autores? E qual critério complementar deve ser usado para caracterizar o 17º autor profícuo, visto que produziu o mesmo tanto do 16º ao 20º autor?

O mesmo problema foi encontrado nos estudos de Urbizagástegui Alvarado (2009), pois há um total de 376 autores que escreveram sobre a produtividade de autores; logo a elite, segundo a raiz quadrada do total de autores é 19,4. A elite é 19 ou 20 autores? Ao confrontar

com seus dados reais, refletindo as dificuldades do uso do critério da raiz quadrada “a cifra mais próxima a essa quantidade é de 17 autores”, o de que fato reflete as dificuldades quando utiliza-se o cálculo de Price. (URBIZAGÁSTEGUI ALVARADO, 2009, p. 74).

3 OUTLIERS APLICADOS À LEI DO ELITISMO

Outliers são valores discrepantes ou valores muito afastados das demais observações coletadas, e uma vez identificados, podem revelar informações importantes (TRIOLA, 2008; BORNMAN, 2008). Lima, Maroldi e Silva (2013) procuram expor que *outliers* podem influir ou representar detalhes relevantes nos cálculos bibliométricos, principalmente, no momento discussão dos dados, uma vez que atualmente os [...] “estudos bibliométricos são mais complexos do que apenas um levantamento estatístico puro”. (FERREIRA, 2010, p.2)

Dentre as diversas possibilidades do aparecimento de um *outlier*, primeiro pode ser o registro errado de uma observação (neste caso, exclui-se automaticamente o *outlier* dos cálculos); segundo, pode indicar a coleta de um dado pertencente à outra população diferente da população de interesse; terceiro, o *outlier* é um valor correto, mas de probabilidade de ocorrência muito baixa (McCLAVE, BENSON, SINCICH, 2009), como no caso de determinar uma elite, ou seja, poucos autores (baixa probabilidade) produzem a maior parte da produção científica. Já Bornmann (2008) utiliza os *outliers* para identificação de periódicos altamente citados bem como os muito poucos citados.

Para a detecção e cálculo dos *outliers*, utiliza-se a Análise Exploratória de Dados (AED), técnica introduzida em 1977 por John Tukey. A AED objetiva extrair dos dados coletados o máximo possível de informações, incluindo os *outliers* (TRIOLA, 2008; LIMA, MAROLDI, SILVA, 2013).

Como estamos interessados nos valores de *outliers* que caracterizam uma elite, e visualizando que valores extremos de produção científica podem estar associados aos *Outliers Superiores Extremos* (O.S.E.), que representam elevada produção científica, sugerimos aos bibliometristas durante os cálculos da Lei do Elitismo a Equação (1), com uso do primeiro quartil (sigla Q1; 25% das observações abaixo dele) e terceiro quartil (sigla Q3; 75% das observações abaixo dele). Em termos matemáticos (TRIOLA, 2008):

$$\text{Equação 1: O.S.E.} > Q1 + 3,0 \cdot (Q3 - Q1)$$

O uso de *outliers* muda o foco de cálculo da elite, pois os quartis (Q1 e Q3) são associados aos trabalhos produzidos, e assim o O.S.E. também se refere aos trabalhos

produzidos. O valor encontrado para o O.S.E. pode até ser um número não inteiro, mas como indica a equação 1, utiliza-se o primeiro número inteiro maior que o resultado encontrado.

Identificado o O.S.E., então o número de autores que compõem a elite é obrigatoriamente um número inteiro, sem necessidade de se adaptar o resultado teórico aos dados reais. E aqui se encontra a principal diferença dos métodos, ou seja, a Lei do Elitismo utiliza o total de autores, gerando um valor teórico de autores em geral não inteiro, além de haver necessidade de se adaptar o valor teórico aos dados observados. Já os *outliers* utilizam no cálculo os trabalhos produzidos, que podem até gerar um número fracionário, mas quando arredondado para mais, gera um número inteiro de autores e que se encaixa aos dados coletados.

Com as ideias dos *outliers* e utilizando os achados de Nicholls (1988), e seguindo a metodologia de cálculo dos quartis proposto por Triola (2008), então o primeiro quartil (Q1) corresponde aos trabalhos produzidos pelo 70º autor, que é de 1 trabalho (Q1 = 1 trabalho). O terceiro quartil (Q3) corresponde aos trabalhos publicados pelo 209º autor, que é de 4 trabalhos (Q3 = 4 trabalhos).

Desta forma, através do uso da equação 1, verifica-se que o O.S.E. (e, portanto, a elite) é quem produz mais de 13 trabalhos, de outra maneira, a elite é quem produz exatamente 14 ou mais trabalhos. E quem produz 14 ou mais trabalhos? São 14 autores.

Portanto, o cálculo da elite via *outliers* permite a caracterização única do número de autores mais profícuos, sem necessidade de arredondamentos (em relação ao total de autores) nem utilização de critérios complementares para o valor teórico obtido aos dados reais.

4 CONSIDERAÇÕES FINAIS

Com o uso da equação para o cálculo de *outliers* extremos que advém da Análise Exploratória de Dados, mostramos sua aplicabilidade para a determinação da elite de qualquer campo de conhecimento, fornecendo outro critério além da raiz quadrada de Price.

Para a produtividade de autores americanos em matemática, a Lei de Price fornece 16,7 autores. Vimos, na Lei do Elitismo que em alguns cálculos para detectar a elite de um determinado conhecimento os números deverão ser arredondados para cima ou para baixo e, nesses casos o bibliometrista deverá adaptar aos dados coletados, ou seja, a elite são 15 ou 20 autores? Já o critério dos *outliers* gerou uma elite de exatamente (detalhe) 14 autores.

Portanto, o critério de Price, além de fornecer valor não inteiro para a elite, ainda possui o problema de adaptação aos dados reais. Já o método dos *outliers* pode até gerar valores fracionados para o valor da quantidade de produção que define a elite (deve ser arredondado para mais), todavia, fornece um valor único para a identificação dos autores que compõem a elite.

Recomendamos novos estudos para validação do método dos *outliers* como identificador de elite de produção científica, bem como para verificar que certamente ambos os métodos podem fornecer valores distintos para a caracterização da elite.

REFERÊNCIAS

BORNMANN, L. et al. Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. Disponível em: <<http://www.int-res.com/articles/esep2008/8/e008p093.pdf>>. Acesso em: 07 fev. 2014.

DE BELLIS, N. Bibliometrics and citation analysis: from the Science Citation Index to Cybermetrics. Maryland: Scarecrow Press, 2009.

CARVALHO, K. Revista científica e pesquisa: perspectiva histórica. In D. A. PROBACIÓN, G. et al. (Org.). Revistas científicas: dos processos tradicionais as perspectivas alternativas de comunicação. Cotia, SP: Ateliê Editorial, 2011.

FERREIRA, A. G. C. Bibliometria na avaliação de periódicos científicos. DataGramaZero, v. 11, n. 3, jun. 2010.

IGAMI, M. P. Z. Elaboração de indicadores de produção científica com base na análise cientométrica das dissertações e teses do IPEN. 180 f. Tese (Doutorado). Instituto de Pesquisas Energéticas e Nucleares. Autarquia da Universidade de São Paulo, São Paulo, 2011.

LIMA, L. F. M.; MAROLDI, A. M.; SILVA, D. V. O. da. Outlier(s) em cálculos bibliométricos: primeiras aproximações. Disponível em: <<http://revista.ibict.br/liinc/index.php/liinc/article/view/522/399>>. Acesso em: 07 de fev. 2014.

McCLAVE, J. T.; BENSON, P. G.; SINCICH, T. Estatística para economia e administração. 10. ed. São Paulo: Pearson, 2009.

NICHOLLS, P. T. Price's square root law: empirical validity and relation to Lotka's law. Information Processing & Management, v. 24, n. 4, p. 469-477, 1988.

PRICE, D. J. S. O desenvolvimento da ciência. Rio de Janeiro: LTC, 1976.

RAVICHANDRA RAO, I. K. Métodos quantitativos em biblioteconomia e ciência da informação. Brasília: Associação dos Bibliotecários do Distrito Federal, 1986.

TRIOLA, M. F. Introdução à estatística. 10. ed. Rio de Janeiro: LTC, 2008.

URBIZAGÁSTEGUI ALVARADO, R. Elitismo na literatura sobre a produtividade dos autores. Ci. Inf., Brasília, v. 38, n. 2, p. 69-79, maio/ago. 2009.

_____. A Lei de Lotka na bibliometria brasileira. Ci. Inf., Brasília, v. 31, n. 2, p. 14-20, maio/ago. 2002.

VANTI, N. A cientometria revisitada à luz da expansão da ciência, da tecnologia e da inovação. Disponível em:
<<http://www.portalseer.ufba.br/index.php/revistaici/article/view/5679/4099>>. Acesso em: 07 fev. 2014.