A lei de lotka: o modelo lagrangiano de poisson aplicado à produtividade de autores

Rubén Urbizagástegui Alvarado

Bibliotecário
Universidade da Califórnia, Riverside
Riverside, CA 97521-5900
USA
ruben@ucrac1.ucr.edu

Descreve a natureza da distribuição Lagrangiana de Poisson, conforme desenvolvida por Janardan e Schaeffer. Oferece as equações específicas para o caso de freqüência de zero observações presentes na amostra coletada. É pouco comum encontrar-se tal distribuição no campo da bibliometria, assim, descreve-se passo a passo a forma de aplicação do modelo, usando-se os dados estudados por Targino & Caldeira sobre a produtividade dos professores da Universidade Federal do Piauí (UFPI), Brasil

Palavras-chave: Lei de Lotka; Distribuição Lagrangiana de Poisson; Produtividade de autores; Bibliometria - Universidade Federal do Piauí, Brasil; Infometría; Cienciometría

Recebido em: 21.01.2003 Aceito em: 16.09.2003

Introdução

Este trabalho constitui um guia didático para análise e teste da lei de Lotka sobre a produtividade científica dos autores, usando-se o modelo Lagrangiano de Poisson. Embora amplamente conhecida na literatura estrangeira de biologia e estatística, esse modelo tem sido pouco utilizado nas práticas bibliométricas latino-americanas. Talvez esse fato seja devido à falta de familiaridade dos bibliotecários latinos com os modelos estatísticos e outros instrumentos matemáticos utilizados para testar dados. Sabe-se que desvendar complexos modelos estatísticos não é o passatempo favorito desses profissionais. Por isso, consideramos importante a elaboração do presente guia, com a esperança de ajudar a compreensão, a adoção e a aplicação do modelo Lagrangiano de Poisson.

Os estudos sobre a produtividade dos autores intelectuais não são privativos da biblioteconomia e da ciência da informação, podendo também ser feitos pelos psicólogos e sociólogos, mas com diferentes objetivos. Os psicólogos estão mais interessados em explorar o mundo da criatividade, os fatores cognitivos que possibilitam a existência dos *gênios* e a *inteligência.* Os sociólogos procuram apontar as condições sociais que permitem a produção estratificada e desigual na ciência. Os bibliotecários, por outro lado, estão mais interessados nas publicações (teses, livros, artigos, trabalhos apresentados em congressos técnico-científicos, capítulos de livros e similares) de caráter didático ou destinadas à divulgação de resultados de pesquisa científica. A distribuição desses documentos e respectivos autores em uma curva de distribuição de freqüências revela, aparentemente, uma relação monotônica negativa entre essas duas variáveis, isto é, quando o número das produções (contribuições) aumenta, o número dos autores (contribuintes) diminui. Observando essa relação monotônica negativa na literatura da física e da química, Lotka (1926) estabeleceu os fundamentos estatísticos do seu modelo, afirmando que o número de autores que totalizam *n* contribuições, em um determinado campo científico, é aproximadamente $1/n^2$ daqueles que fazem uma só contribuição, e que a proporção daqueles que apresentam uma única contribuição é de mais ou menos 60 %. Esta proposição tem sido denominada lei do quadrado inverso ou *lei de Lotka.*

Desde a época em que Lotka estabeleceu esse modelo, muitos estudos têm sido realizados para pesquisar a produtividade dos autores em distintas disciplinas. Até dezembro de 2000, cerca de 250 trabalhos - compreendendo artigos, monografias, capítulos de livros, comunicações em congressos e literatura cinzenta - tinham sido produzidos, criticando, replicando ou reformulando esse modelo bibliométrico (Urbizagástegui & Lane, 2002). Esses autores elaboraram uma exaustiva bibliografia transnacional, procurando referências em diversas bases de dados, como ISA, LISA, *Library Literature,* MAGS (revistas e jornais), *Current Contents,* Eric, PsyInfo, *Compendex,* Agrícola, Biosis, Inspec, Hapi, *Dialog,* Pascal, *Uncover; Sociological Abstracts,* bem como as bases de dados do CINDOC (Espanha), LICI do IBICT (Brasil) e INFOBILA (México).

Apesar das numerosas pesquisas realizadas sobre o assunto, os resultados ainda são contraditórios ou inconclusos e parecem não proporcionar uma clara validade dessa lei. Por exemplo, Oppenheim (1986) afirma que *"deve-se enfatizar que a Lei de Lotka tem sido testada em muitas coleções de*

dados, mas o ajuste nem sempre tem sido bom". Nicholls (1989:383) reclama que " os *resultados desses estudos são incomparáveis, devido à falta de padronização da forma de medição, da estimação dos parâmetros, dos procedimentos do teste, bem como da interpretação do modelo".* Esse ponto não tem sido suficientemente ressaltado na bibliografia já publicada. As revisões do estado-da-arte realizadas por Vlachy (1980) e Potters (1981) têm contribuído para reorientar o interesse dos pesquisadores em direção a outros tipos de distribuições capazes de proporcionar melhores ajustes dos dados observados e calculados. Essa reorientação de interesses levou à adoção de modelos estatísticos experimentados com êxito nas ciências naturais, como a distribuição hiperbólica, a distribuição logarítmica, a distribuição de Yule, a distribuição binomial, a distribuição binomial negativa, a série geométrica, a série logarítmica, a distribuição de Weinbull, a distribuição de Waring, a distribuição de Poisson, a distribuição lognormal de Poisson, a distribuição truncada de Poisson e, finalmente, a distribuição Gauss-Poisson inversa generalizada.

Juntamente com essas críticas, têm aparecido algumas discordâncias relacionadas às três possíveis formas de realizar a contagem relativa à autoria múltipla. A contagem direta, quando somente os autores *sênior* ou principais (os autores nomeados em primeiro lugar) são considerados, ignorando-se os autores secundários (colaboradores); a contagem completa, em que cada autor (principal e/ou secundário) recebe o crédito de uma contribuição; e a contagem ajustada, quando se atribui a cada autor (principal e/ou secundário), uma fração ou porção da contribuição total, ou seja, no caso de cinco autores de um único artigo, cada autor recebe o crédito de 1/5 do artigo. Alguns autores afirmam que as contagens direta e ajustada não produzem diferenças essenciais e que "*os dois meios de contagem produzem a mesma coisa, não havendo, pois, necessidade, de se considerar a contagem ajustada, devendo-se prestar mais atenção à contagem direta"* (Nath & Jackson, 1991: 207).

Pao (1985) ressalta o fato de que *"não existe um método uniforme de coleta e organização dos dados"* para testar a lei de Lotka. Salienta também que "*muitos pesquisadores simplesmente atribuíram o valor 2 a n, sem estimar esse valor dos dados observados".* Essa preferência é atribuída ao fato de os procedimentos de cálculo realizados dessa maneira serem muito simples e fáceis de se realizarem. A mesma autora afirma, ainda, que "*muitos pesquisadores evitaram fazer um teste apropriado do grau de ajuste dos dados observados"*, com relação aos dados esperados. Por isso, Vlachy (1974) teria encontrado sérias discrepâncias entre os dados observados e o ajuste da lei do quadrado inverso.

Além disso, alguns estudos usaram dados referentes a períodos pequenos, de apenas um ou dois anos, enquanto outros estudos cobriam períodos longos da história de um determinado assunto. Esses fatos possibilitaram a Potter (1981) afirmar, após uma extensa revisão da literatura, que "*para períodos de cobertura iguais ou superiores a dez anos, e comunidades de autores definidas amplamente, a produtividade dos autores aproxima-se da distribuição de freqüências observada por Lotka, a qual é conhecida como lei de Lotka".*

Entre 1926 e 1970, as pesquisas sobre a produtividade dos autores intelectuais foram esporádicas, mas, a partir do início dos anos 70, o interesse pela aplicação do modelo da produtividade científica foi retomada por Vlachy (1970), Naranan (1971), Turkeli (1973) Terrada (1973) e Murphy (1973). A partir dessa década, as pesquisas sobre a produtividade dos autores tornaram-se mais sérias e, de certo modo, mais científicas, tanto que o modelo de Lotka

foi testado em muitas áreas que incluem bases de dados de patentes (Oppenheim, 1986), óleos lubrificantes (López Calafi; Salvador e Guardia, 1985), educação superior (Budd, 1988), música popular (Cook,1989), finanças (Chung e Cox, 1990), economia (Cox e Chung, 1991), contabilidade (Chung; Pak e Cox, 1992), indústria musical (Cox; Felton e Chung, 1995), psiquiatria (López-Muñoz e Rubio Valladolid, 1995), psicofisiologia (Sánchez-Hernández; et al., 1996), glândula pineal e melatonina (López-Muñoz; et al., 1996), biblioteconomia e ciência da informação espanhola (Jiménez Contreras e Moya Anegón, 1997), genética (Gupta; Kumar e Rousseau, 1998), bibliometria (Urbizagástegui Alvarado, 1999), antropologia brasileira (Urbizagástegui Alvarado e Oliveira, 2001). Mas os dados dessas pesquisas variam muito, incluindo desde dados tomados de bibliografias exaustivas, como as de entomologia (Gupta,1987), pesquisas sobre batata (Gupta; et al., 1996), dados tomados de um grupo de periódicos (Nath & Jackson, 1991), até dados de um único periódico (Murphy, 1973; Carpintero, et al., 1977; Gisbert Tio & Valderrama Zurian, 1994; Urbizagástegui e Cortés, 2002). Essas discrepâncias na cobertura da literatura recopilada, na forma de medição, na estimação dos parâmetros, na forma do teste e na interpretação do modelo, mostraram a necessidade de normalização do processo. No geral, os autores concordam que, para uma correta aplicação do modelo de Lotka, devem-se seguir as seguintes recomendações:

a) Selecionar um campo específico de produção científica. Quanto mais específico o campo, melhor o resultado;

b) Selecionar uma bibliografia existente ou elaborar uma bibliografia sobre o campo específico cuja cobertura seja exaustiva. Quanto mais extensa e exaustiva melhor. Sugere-se que a cobertura dessa bibliografia seja maior ou igual a dez anos;

c) Contar a produtividade de cada autor, considerando-se também os co-autores. Isso significa que deve-se adotar o método da contagem completa;

d) Ordenar os dados coletados em uma tabela de freqüências para facilitar a visualização dos mesmos;

e) Selecionar o modelo estatístico mais adequadamente sugerido pelos dados tabulados;

f) Calcular os valores esperados ou teóricos, seguindo as especificações do modelo estatístico escolhido;

g) Estabelecer as hipóteses a serem testadas e a região de rejeição dessas hipóteses no nível de significância de $a = 0.05$;

h) Testar a qualidade do ajuste dos dados, usando-se o teste do qui-quadrado ou Kolmogorov-Smirnov.

É necessário ressaltar que, nas últimas décadas, outros pesquisadores têm experimentado diferentes modelos estatísticos para descrever a distribuição da produtividade dos autores, especialmente quando, na distribuição de

freqüências observadas, têm sido coletados autores que não tenham publicado no período estudado, ou seja, quando as amostras incluem produções nulas. Este trabalho objetiva analisar um desses modelos, conhecido como distribuição Lagrangiana de Poisson, que será aplicado à produtividade dos professores da Universidade Federal do Piauí.

Natureza da distibuição langrangiana de Poisson

Na literatura estatística e biológica, sabe-se que a distribuição de Poisson é caracterizada pela igualdade entre a média e a variância. Esse fato faz com que o índice de dispersão seja igual à unidade. Essa constatação tem levado às seguintes suposições: se o índice de dispersão é igual a um, a distribuição de freqüências se ajustará à distribuição de Poisson; se o índice de dispersão é menor que um, a distribuição de freqüências se ajustará à distribuição binomial; se o índice de dispersão é maior que um, a distribuição de freqüências se ajustará à distribuição binomial negativa.

Como se sabe, a distribuição de Poisson é produzida por eventos que ocorrem aleatória e independentemente uns dos outros em um determinado período. Isso significa que a ocorrência ou não ocorrência de um evento, não tem nenhum efeito na ocorrência ou não ocorrência de um evento subseqüente. Se a ocorrência de um evento particular alterasse ou influísse a probabilidade de ocorrência de um evento subseqüente, a distribuição desses eventos poderia exibir subdispersão ou superdispersão em relação à distribuição de Poisson e daria origem à distribuição Lagrangiana de Poisson.

A distribuição Lagrangiana de Poisson proporciona um modelo que se ajusta muito bem a dados experimentais caracterizados pela superdispersão, subdispersão ou, ainda, ausência de dispersão. O índice de dispersão da distribuição Lagrangiana de Poisson pode ser maior, menor ou igual à unidade. Isso não tem nenhuma importância nem peso na distribuição. A distribuição Lagrangiana de Poisson é mais poderosa do que a distribuição de Poisson. Além disso, como todas as distribuições discretas, é bem fácil de ser aplicada. É usada quando a distribuição de Poisson revela um ajuste inadequado aos dados observados. Foi introduzida por Cole (1946) e Cônsul e Jain (1973) na literatura biológica, mas foram Janardan e Schaeffer (1977) e Janardan; Kester e Schaefer (1979) que descreveram essa distribuição na forma da equação (1) seguinte:

$$N_k = N\left[\frac{g_1\left(g_1 + g_2 k\right)^{k-1} e^{-\left(g_1+g_2 k\right)}}{k!}\right] \qquad (1)$$

na qual,
k = freqüência das classes 0,1, 2, 3, ... n
e = base dos logaritmos naturais, 2.718

N = número total dos valores observados

g_1 = *taxa de atração* do processo de Poisson que afeta o movimento das variáveis independentes em direção às variáveis dependentes. Por exemplo, o movimento dos autores, em direção à produção de artigos. Quanto mais autores existam, propensos à produção de artigos, mais artigos teremos. Quanto menos autores propensos à produção de artigos, menos artigos teremos. Então, g_1 é a *taxa de atração* de autores à produção de artigos.

g_2 = uma função complexa da *taxa de competição ou repulsão.* No caso dos autores, seria a taxa de competição ou repulsão em direção à produção de artigos mais atrativos ou de maior visibilidade que ajudassem na criação da autoridade ou competência na área escolhida.

Nessa equação, k! representa o fatorial do valor de $k = 0, 1, 2, 3, ...$ Esse fatorial é calculado da seguinte maneira:

Para k = 0! = 1
Para k = 1! = 1 x 1 = 1
Para k = 2! = 2 x 1 = 2
Para k = 3! = 3 x 2 x 1 = 6
Para k = 4! = 4 x 3 x 2 x 1 = 24
Para k = 5! = 5 x 4 x 3 x 2 x 1 = 120
Para k = 6! = 6 x 5 x 4 x 3 x 2 x 1 = 720
Para k = 7! = 7 x 6 x 5 x 4 x 3 x 2 x 1 = 5040
Para k = 8! = 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 = 40320
Para k = 9! = 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 = 362880
Para k = 10! = 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 = 3628800
e assim sucessivamente para todos os valores de k em estudo.

Como se pode ver na equação (1), a distribuição lagrangiana de Poisson tem somente dois parâmetros desconhecidos, simbolizados pelas letras minúsculas g_1 (taxa de atração) e g_2, (efeito da disperão). Conhecendo-se o valor desses parâmetros, pode-se calcular a probabilidade de toda a distribuição de freqüências.

Aplicação do modelo lagrangiano de Poisson à produtividade de autores

Aplicamos o modelo Lagrangiano de Poisson a um exemplo previamente estudado pela lei de Lotka por Targino e Caldeira (1988). Esses autores analisaram a produtividade científica dos docentes da Universidade Federal do Piauí (UFPI), no Brasil, relativa a um período de dois anos (1984 e 1985). Nesse período, a Universidade Federal do Piauí tinha um total de 958 docentes, mas somente 95 deles tiveram produção científica sob a forma de livros, capítulos de livros, artigos de periódicos, relatórios de pesquisa e comunicações em congressos. Isso significa que 863 professores não produziram um único artigo. Ainda que, na suas análises pela lei de Lotka, aqueles autores não tivessem incluído os docentes sem produtividade científica, decidiu-se incluí-los neste trabalho. Os procedimentos de aplicação do modelo Lagrangiano de Poisson são em seguida detalhados passo a passo, tendo em vista que este trabalho tem também caráter didático.

a) Coleta dos dados e distribuição das freqüências observadas:

A distribuição de autores e artigos coletados para este estudo são mostrados na *TAB. 1.* Nos dois anos estudados, 90% dos professores não mostraram produtividade alguma, logo, os trabalhos só foram realizados por 10% dos docentes. Mais ainda, a tabela mostra que somente 1% dos professores publicaram 5 ou mais trabalhos. Os outros dados indicam que a média de produtividade foi bastante baixa, alcançando apenas 0.2432 trabalhos por docente, com uma variância de 1.2 trabalhos e um desvio padrão de 1.1 trabalhos.

TABELA 1 - Distribuição das freqüências observadas dos artigos produzidos por autor

No. de contri-buições por autor x	No. de autores y	% de autores % y	No, de artigos xy	% de artigos % xy
0	863	90.08	0	0.00
1	44	4.59	44	18.88
2	23	2.40	46	19.74
3	10	1.04	30	12.88
4	9	0.94	36	15.45
5	2	0.21	10	4.29
6	2	0.21	12	5.15
7	2	0.21	14	6.01
13	2	0.21	26	11.16
15	1	0.10	15	6.44
Total	958	100.0	233	100.0

b) Cálculo da média aritmética

Para calcular-se a média aritmética, organizaram-se os dados coletados conforme mostrado na tabela seguinte. Depois, multiplicou-se o número de contribuições por autor (x) pelo número de autores (y), a fim de se obter o número total de artigos publicados. Finalmente, calculou-se a soma acumulada tanto dos autores como dos artigos conforme mostrados na TAB. 2. Essas somas acumuladas atingiram a quantidade total de 958 autores e 233 artigos respectivamente.

TABELA 2: Distribuição das freqüências observadas dos artigos produzidos por autor

No. de contribuições por autor x	No. de autores y	No, de artigos xy
0	863	0
1	44	44
2	23	46
3	10	30
4	9	36
5	2	10
6	2	12
7	2	14
13	2	26
15	1	15
Total	958	233

$$\tilde{x} = \frac{\sum_{i=1}^{n} xy}{n} = \frac{233}{958} = 0.2432$$

c) Cálculo da variância

Para calcular-se a variância, adicionaram-se duas novas colunas aos dados apresentados na TAB. 2. Na quarta coluna desta tabela, apresentam-se os valores do número de contribuições por autor (x) elevados ao quadrado (x^2). Na quinta coluna, estão os valores do número de autores (y) da segunda coluna multiplicados pelos valores da quarta coluna, isto é, x^2y. Finalmente, estimou-se a soma acumulada da segunda, terceira e quinta colunas, conforme mostrados na TAB. 3. Depois, calculou-se a variância, usando-se a fórmula seguinte:

$$\text{var} = \frac{\sum x^2 y - \frac{(\sum xy)^2}{N}}{N - 1}$$

TABELA 3: Distribuição das freqüências observadas dos artigos produzidos por autor

x	y	xy	x^2	x_2y
0	863	0	0	4
1	44	44	1	44
2	23	46	4	92
3	10	30	9	90
4	9	36	16	144
5	2	10	25	50
6	2	12	36	72
7	2	14	49	98
13	2	26	169	338
15	1	15	255	225
Total	958	233		1153

$$\text{var} = \frac{1153 - \frac{(233)^2}{958}}{958 - 1} = \frac{1153 - \frac{54289}{958}}{957} = \frac{1153 - 56.6691023}{957}$$

$$\text{var} = \frac{1096.330898}{957} = 1.145591325 \pm 1.1456$$

d) Cálculo do desvio padrão

$$DS = \sqrt{\text{var}} = \sqrt{1.1456} = 1.070327053 \pm 1.0703$$

e) Cálculo do índice de dispersão

$$ID = \frac{\text{var}}{\tilde{x}} = \frac{1.1456}{0.2432} = 4.71053$$

f) Cálculo do efeito da dispersão

$$g_2 = 1 - \hat{D}^{-0.5}$$

onde D é o índice de dispersão

$$g_2 = 1 + (4.71053)^{-0,5} = 1 - 0.ü46075 = 0.53925$$

g) Cálculo da taxa de atração

$$g_1 = \bar{x}(1 - \hat{g}_2)$$

$$g_1 = 0.2432\ (1 - 0.53925) = 0.2432 \times 0.46075 = 0.112054$$

h) Cálculo da taxa de competição

$$b = \frac{g_1}{g_2}$$

$$b = \frac{0.112054}{0.53925} = 0.207796$$

i) Cálculo dos valores esperados ou teóricos, usando-se a seguinte equação (1):

$$N_k = N\left[\frac{g_1\left(g_1 + g_2 k\right)^{k-1} e^{-\left(g_1 + g_2 k\right)}}{k!}\right] \qquad (1)$$

a) Exemplo de cálculo de k = 0 (número de autores esperados ou teóricos que não produziram nenhum artigo)

$$N_0 = \frac{0.112054(0.112054+(0.53925)(0))^{0-1}\ 2.718^{-(0.112054+(0.53925)(0))}}{0!}$$

$$N_0 = \frac{0.112054\ (0.112054)^{-1}\ 2.718^{-0.112054}}{1}$$

$$N_0 = \frac{0.112054 \times 8.92427 \times 0.894006}{1}$$

$$N_0 = \frac{0.894006}{1} = 0.894006$$

$$N_0 = 0.894006 \times 958 = 856.458$$

$N_0 = 856.5$ (valor arredondado)

b) Exemplo de cálculo de k = 1 (número de autores esperados ou teóricos que produziram 1 artigo)

$$N_1 = \frac{0.112054(0.112054+(0.53925)(1))^{1-1} x 2.718^{-(0.112054+(0.53925)(1))}}{1!}$$

$$N_1 = \frac{0.112054\,(0.651304)^0 \times 2.718^{-0.651304}}{1}$$

$$N_1 = \frac{0.112054 \times 1 \times 0.521401}{1}$$

$$N_1 = \frac{0.0584251}{1} = 0.0584251$$

$$N_1 = 0.0584251 \times 958 = 55.9712$$

$$N_1 = 56.0 \textit{ (valor arredondado)}$$

c) Exemplo de cálculo de k = 2 (número de autores esperados ou teóricos que produziram 2 artigos)

$$N_2 = \frac{0.112054(0.112054 + (0.53925)(2))^{2-1}\, 2.718^{-(0.112054+(0.53925)(2))}}{2!}$$

$$N_2 = \frac{0.112054\,(0.112054 + 1.0785)^1\, 2.718^{-(0.112054 + 1.0785)}}{2}$$

$$N_2 = \frac{0.112054\,(1.190554)^1\, 2.718^{-1.190554}}{2}$$

$$N_2 = \frac{0.112054 \times 1.190554 \times 0.30409}{2}$$

$$N_2 = \frac{0.0405675}{2} = 0.0202838$$

$$N_2 = 0.0202838 \times 958 = 19.4319$$

$$N_2 = 19.4 \textit{ (valor arredondado)}$$

d) e assim sucessivamente, para k = 3, 4, 5, ... n, isto é, para todos os valores esperados ou teóricos dos autores que produziram 3, 4, 5, ..., n artigos, incluindo as freqüências do número de artigos para os quais não se observaram autores produtores mas que estão implicitamente incluídos na amostra analisada. Esse é o caso das freqüências de 8, 9, 10, 11, 12, e 14 artigos, para os quais não se observaram docentes produtores.

j) Estabelecimento das hipóteses e seleção do nível de significância (recomenda-se a = .05)

Desejava-se testar se os valores de k = 1, 2, 3, ... n procediam de uma distribuição do tipo lagrangiana de Poisson, ou seja, a probabilidade de que um elemento incluído na amostra seria igualmente provável para todos os elementos nessa mesma situação. Para isso, estabeleceram-se as hipóteses da seguinte maneira:

Ho = a distribuição representa a contagem de k = 0, 1, 2, 3 ..... n.
Ha[1]a distribuição não representa a contagem de k = 0, 1, 2, 3 ..... n.

k) Cálculo do teste estatístico x^2 (qui-quadrado), usando-se a seguinte equação:

$$\chi^2 = \sum_{1}^{n} \frac{(f_o - f_t)^2}{f_t}$$

na qual,

f_0 = Freqüência observada
f_t = Freqüência teórica, esperada ou calculada.

Para facilitar o cálculo do qui-quadrado, elaborou-se a TAB. 4. É necessário alertar ser imprescindível calcularem-se as freqüências esperadas correspondentes às freqüências observadas com células vazias (neste caso, as freqüências 8 a 12 e 14 que não apresentaram observações nos dados originais, porque não se observaram professores que tivessem produzido 8, 9, 10, 11, 12 e 14 artigos cada um). Dessa maneira, não se terá nenhuma freqüência vazia no momento de se estimarem as freqüências esperadas ou teóricas:

TABELA 4: Cálculo do qui-quadrado sem agrupação das freqüências menores que 5

x	*fo*	*ft*	*(fo - ft)*	$(fo - ft)^2$	$\frac{(fo - ft)^2}{ft}$
0	863	856.5	6.50	42.25	0.0493
1	44	56.0	-12.00	144.00	2.5714
2	23	19.4	3.60	12.96	0.6680
3	10	9.4	0.60	0.36	0.0383
4	9	5.4	3.60	12.96	2.4000
5	2	3.4	-1.40	1.96	0.5765
6	2	2.2	-0.20	0.04	0.0182
7	2	1.5	0.50	0.25	0.1667
8	0	1.1	-1.10	1.21	1.1000
9	0	0.8	-0.80	0.64	0.8000
10	0	0.6	-0.60	0.36	0.6000
11	0	0.4	-0.40	0.16	0.4000
12	0	0.3	-0.30	0.09	0.3000
13	2	0.2	1.80	3.24	16.2000
14	0	0.2	-0,20	0.04	0.2000
15	1	0.1	0.90	0.81	8.1000
Total	958	957.5		$x^2 = 34.1884$	

Também é necessário chamar atenção para o fato de que, na aplicação do teste do x^2 (qui-quadrado), nenhuma das freqüências observadas deve ser menor que 5. Encontrando-se freqüências inferiores a 5, será necessário agrupar as freqüências adjacentes em grupos com freqüências iguais ou superiores a 5, para que, dessa forma, o x^2 (qui-quadrado) seja válido e consistente. Como exemplo, para mostrar as variações encontradas em situações práticas, incluímos aTAB. 5, que é a mais adequada e recomendável. Na TAB. 4, na qual as freqüências menores que 5 não foram agrupadas, o qui-quadrado foi igual a 34.1884, enquanto na TAB. 5, com freqüências agrupadas, o qui-quadrado revelou-se muito menor (6.1919).

TABELA 5: Cálculo do qui-quadrado com agrupacao das freqüências inferiores a 5

x	*fo*	*ft*	*(fo - ft)*	$(fo - ft)^2$	$\frac{(fo - ft)^2}{ft}$
0	863	856.5	6.50	42.25	0.0493
1	44	56.0	-12.00	144.00	2.5714
2	23	19.4	3.60	12.96	0.6680
3	10	9.4	0.60	0.36	0.0383
4	9	5.4	3.60	12.96	2.4000
5-6	4	5.6	-1.60	2.56	0.4571
7-15	5	5.2	-0.20	0.04	0.0077
Total	958	957.5			$x^2 = 6.1919$

O valor de 6.1919 para c^2 (qui-quadrado) pode ser considerado o suficientemente alto para se concluir que os valores da distribuição da produção de trabalhos pelos professores da Universidade Federal do Piauí foram igualmente prováveis para todos os professores, isto é, procederam de uma distribuição do tipo lagrangiana de Poisson?

I) Especificação da região de rejeição das hipóteses ao nível de significância á = .05

a) Determinação dos graus de liberdade:

gl = k - 1 - n =7-1-2 = 4

onde,

k = é o número de pares de dados observados e usados para estimar o qui-quadrado. Esses pares de dados podem se vistos na tabela do qui-quadrado com agrupação de freqüências inferiores a 5 (TAB. 5). Nesse caso, k = 7;

1 = é o número de restrições usadas para os cálculos dos valores esperados. Nesse caso, é igual a 1;

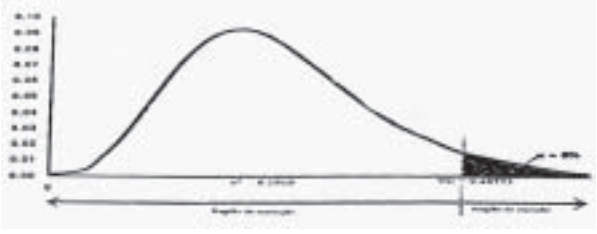
n = é o número de parâmetros usados na solução da equação (1), Nesse caso, n é igual a 2 porque são dois os parâmetros (g_1, taxa de atração e g_2, taxa de competição).

b) Usando-se um nível de significância á = .05, obtem-se na tabela do x^2 reproduzida em qualquer texto estatístico (neste caso TAB. 1 incluída no anexo), o valor do x^2 que define a região de rejeição.

Se o $x^2 > x^2_{.05}$ rejeita-se a hipótese nula (aceitar a Ha)
Se o $x^2 < x^2_{.05}$ aceita-se a hipótese nula (rejeitar a Ha)

Em outras palavras, se o x^2 calculado é maior que o valor crítico do x^2 para um a = .05, obtido da TAB. 1 incluída no anexo, rejeita-se a hipótese nula. Isso significa aceitar a hipótese alternativa. E se o x^2 calculado é menor que o valor crítico obtido para um á = .05, aceita-se a hipótese nula. Isso significa rejeitar a hipótese alternativa.

c) Usando-se a TAB. 1 dos valores críticos do $x^2_{.05}$ correspondente a 4 graus de liberdade, encontra-se um valor igual a 9.48773. Assim, a região de rejeição é a parte situada à direita da linha pontilhada da FIG. 1. A região de aceitação é a parte à esquerda dessa figura.



m) Interpretação:

Como o x^2 calculado (igual a 6.1919) foi menor que o valor crítico do $x^2_{.05}$ (igual a 9.48773), aceitou-se a hipótese nula ao nível de significância de 0.05. Por isso, concluiu-se que os valores das freqüências procediam de uma distribuição do tipo de lagrangiana de Poisson e eram igualmente prováveis para todos os valores de k.

n) Comparação dos valores observados com os valores calculados:

A TAB. 6 comparou os valores observados e os valores calculados. Deve-se observar, nessa tabela quão próximos ou distantes estão os valores correspondentes. Deve-se observar também que os totais são os mesmos, quase os mesmos ou divergentes. Nesse caso, a discrepância mais notável apresentou-se na segunda freqüência. O modelo estimou 56 autores com um trabalho produzido mas observaram-se 44 deles: uma superestimação de 12 autores (27%). Na terceira freqüência houve uma subestimação de 4 autores (17%). Não obstante, estas discrepâncias são minimizáveis face ao tamanho da população de 958 docentes estudados.

TABELA 6 - Freqüências observadas e esperadas

No.de contri-buições por autor x	Freqüências observadas	Freqüências esperadas
0	863	856.5
1	44	56.0
2	23	19.4
3	10	9.4
4	9	5.4
5	2	3.4
6	2	2.2
7	2	1.5
8	0	1.1
9	0	0.8
10	0	0.6
11	0	0.4
12	0	0.3
13	2	0.2
14	0	0.2
15	1	0.1
Total	958	957.5

o) Traçado do gráfico de dispersão das freqüências observadas e esperadas

O ajuste entre as freqüências observadas e as freqüências esperadas da distribuição da produtividade dos autores pode ser melhor observada traçando-se a dispersão dessas duas variáveis, conforme mostrado na FIG. 2.

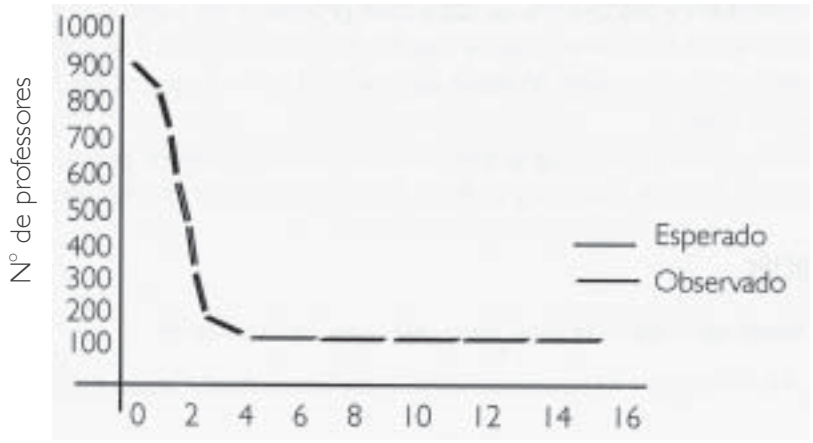


FIGURA 2: Dispersão das freqüências observadas e esperadas

Conclusões

O estudo da produtividade dos docentes da Universidade Federal do Piauí permitiu identificar alta percentagem (90%) de professores que, no período estudado, não mostrou produção intelectual sob a forma de livros, folhetos, capítulos de livros, artigos de periódicos, informes de pesquisa e comunicações em congressos. Também identificou-se um percentual de 8% de pequenos produtores responsáveis por 52% da literatura produzida. Os produtores medianos e grandes, por sua vez, representaram somente 2% dos docentes, mas foram responsáveis por quase a metade da literatura produzida (48%).

Foi encontrado também que 1% dos professores publicou cinco ou mais trabalhos, representando um terço da produção total (33%). A média de produtividade foi bastante baixa, alcançando apenas 0.2432 trabalhos por docente, com uma variância de 1.2 trabalhos, um desvio padrão de 1.1 trabalhos por professor e um índice de dispersão de 4.71053 indicando uma relativa falta de comunicação e colaboração entre os professores incluídos na amostra estudada.

O teste do x^2 foi usado para comparar os valores observados aos valores esperados. Com uma taxa de dispersão $g_2 = 0.53925$ e uma taxa de atração $g_1 = 0.112054$, ao nível de significância 0.05 e 4 graus de liberdade, verificou-se que o valor crítico do x^2 (9.48773), é muito maior do que o valor do x^2 calculado (6.1919). Concluiu-se, então, que essa literatura ajusta-se muito bem ao modelo de Lotka.

Talvez seja pertinente salientar que Targino & Caldeira (1988) não estimaram os parâmetros nem testaram seus dados e apenas parecem sugerir que seus dados truncados ajustam-se ao modelo de distribuição do quadrado inverso de Lotka.

Lotka xs law: the Poisson-Lagrangian model applied to author xs productivity

Describes the nature of the Poisson Lagrangian distribution as developed by Janardan & Schaeffer with specific equations for cases when zero observations are present in the data collected. Since in the bibliometric area it is not very commom to find data with zero frequency distributions, an application process of the Lagrangian Poisson distribution model is described step by step for this particular case. The application is illustrated with process data on literature produced by professors at the Federal University of Piaui.

Key-words: Lotka's law; Lagrangian Poisson model; Author productivity; Bibliometrics; Federal University of Piaui; Brazil; Infometrics; Scientometrics

Referências

BUDD, J. M. A bibliometric analysis of higher education literature. *Research in Higher Education,* v. 28, n. 2, p. 180-190,1988.

CARPINTERO, H.; PEIRÓ, J. H.; QUINTANILLA, I. El Anuário de Psicologia (1969-1974): un estudio estadístico y bibliométrico. *Anuario de Psicología,* v. 16, n. 1, p. 22-34,1977.

CHUNG, K. H.; PAK, H. S.; COX, R. A. K. Patterns of research output in the accounting literature: a study of the bibliometric distributions. *Abacus* , v. 28, n. 2, p. 168-185,1992.

CHUNG, K. H.; COX, R. A. K. Patterns of productivity in the finance literature: a study of bibliometric distributions. *The Journal of Finance.* v. 45, n. 1, p.301-309, Mar. 1990.

COILE, R. C. Lotka and Information Science. *Jasis,* v. 26, n. 2, p. 133, Mar./Apr. 1975.

COLE, La M. C. A theory for analyzing contagiously distributed populations. *Ecology,* v. 27, n. 4, p. 329-341, Oct. 1946.

CONSUL, P. C.; JAIN, G. C. A generalization of the Poisson distribution. *Technometrics,* v. 15, n. 4, p. 791-799, Nov. 1973.

COOK, K. L. Laws of scattering applied to popular music. *JASIS,* v. 40, n. 4, p. 277-283, July 1989.

COX, R. A. K.; CHUNG, K. H. Patterns of research output and author concentration in the economic literature. *The Review of Economics and Statistics,* v. 73, n. 4, p. 740-747, Nov. 1991.

COX, R. A. K.; FELTON, J. M.; CHUNG, K. H. The concentration of commercial success in popular music: an analysis of the distribution of gold records. *Journal of Cultural Economics,* v. 19, n. 4, p. 333-340,1995.

GISBERT TIO, A., VALDERRAMA ZURIAN, J. C. Estúdio bibliométrico de la revista Española de Drogodependencias, 1976-1993. In: JORNADAS DE INFORMACIÓN Y DOCUMENTACIÓN DE CIENCIAS DE LA SALUD, 5.,1994, Palma de Mallorca. *Actas...* Palma de Mallorca, 1994. p. 1-7.

GUPTA, D. K. Lotka's law and productivity patterns of entomological research in Nigeria for the period, 1900-1973. *Scientometrics,* v. 12, n. ½,p. 33-46,1987.

GUPTA, B. M.; KUMAR, S.; SYED, S.; SINGH, K. vir. Distribution of productivity among authors in potato research (1900-1980). *Library Science with Slant to Documentation,* v.33, n. 3, p. 127-134, Sept. 1996.

GUPTA, B. M.; KARISIDDAPPA, C. R. Author productivity patterns in theoretical populations genetics, 1900-1980. *Scientometrics.* v. 36, n. 1, p. 19-41,1996.

GUPTA, B. M.; KUMAR, S.; ROUSSEAU, R. Applicability of selected probability distributions to the number of authors per article in theoretical population genetics. *Scientometrics,* v. 42, n. 3, p. 325-334,1998.

JANARDAN, K. G.; SCHAEFFER, D. J. Models for the analysis of chomosomal aberrations in human leukocytes. *Biometrical Journal,* v. 19, n. 8, p. 599-612,1977.

JANARDAN, K. G.; KESTER, H. W.; SCHAEFFER, D. J. Biological applications of the Lagrangian Poisson distribution. *BioScience* , v. 29, n. 10, p.599-602, Oct. 1979.

JIMÉNEZ CONTRERAS, E.; MOYA DE ANEGÓN, F. Análisis de la autoría en revistas españolas de Biblioteconomía y Documentación, 1975-1995. *Revista Española de Documentación Científica,* v. 20, n. 3, p. 252-266,1997.

LÓPEZ CALAFI, J.; S., A.; GUARDIA, M. de la. Estudio bibliométrico de la literatura científica sobre la determinación de elementos metálicos en aceites lubricantes por espectroscopia de absorción atômica. *Revista Española de Documentación Científica,* v. 8, n. 3, p. 201-213,1985.

LÓPEZ-MUÑOZ, F.; MARÍN, E.; CALVO, J. L . Scientific research on the pineal gland of melatonin: a bibliometric study for the period 1966-1994. *Journal of Pineal Reseach;* v. 20, n. 3, p. 115-124,1996.

LÓPEZ-MUÑOZ, F.; RUBIO VALLADOLID, G. La producción científica española em psiquitría: estudio bibliométrico de las publicaciones de circulación intenacional durante el periodo 1980-1993. *Anales de psiquiatría,* v. 2, n. 2, p. 68-75, 1995.

LOTKA, A. J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences,* v. 16, n. 12, p. 317-323, June 1926.

MURPHY, L. J. Lotka's law in the Humanities? *JASIS*, v. 24, n. 6, p. 461-462, Nov./Dez. 1973.

NARANAN, S. Power law relations in science bibliography: a self consistent interpretation. *Journal of Documentation,* v. 27, n. 2, p. 83-97, June 1971.

NATH, R.; WADE, M. J. Productivity of management information systems researchers: does Lotka's Law apply? *Information processing & Management*, v. 27, n. 2/3. p. 203-209, 1991.

NICHOLLS, P. T. Empirical validation of Lotka's Law. *Information Processing & Management*, v. 22, n. 5, p. 417-419, 1986.

NICHOLLS, P. T. Estimation of Zipf parameters. *JASIS,* v. 38. p. 443-445, 1987.

NICHOLLS, P. T. Bibliometric modeling process and the empirical validity of Lotka's Law. *JASIS,* v. 40, n. 6, p. 379-385, 1989.

OPPENHEIM, C. Use of online databases in bibliometric studies. In: INTERNATIONAL ON-LINE INFORMATION MEETING. 9th, 1985. London. *Proceedings...* Oxford, England; Medford. N.J.: Learned Information, [1985?]. p. 355-364.

PAO, M. L. Bibliometrics and computational musicology. *Collection Management,* v. 3,n. 1, p.79-109,1979.

PAO, M. L. Lotka's law: a testing procedure. *Information Processing & Management*, v. 21, n. 4, p. 305-320,1985.

PAO, M. L. An empirical examination of Lotka's Law. *JASIS,* v. 37, n. 1, p. 26-33, 1986.

POTTER, W. G. Lotka's Law revisited. *Library Trends,* v. 31, p. 21-39,1981.

SÁNCHEZ, H. A.; PEDRAJA, M. J.; QUIÑONEZ-VIDAL, E.; MARTÍNEZ-SÁNCHEZ, F. A historic-quantitative approach to psychophysiological research: the first three decades of the journal Psychophysiology (1964-1993). *Psychophysiology,* v. 33, n. 6, p. 629-636,1996.

TARGINO, M. das G. & CALDEIRA, P. da T. Análise da produção científica em uma instituição de ensino superior: o caso da Universidade Federal do Piauí. *Ciência da Informação,* Brasília, v. 17, n. 1, p.15-25, 1988.

TERRADA, M. L. La productividad de los autores médicos españoles (Ley de Lotka). In: ——. *La literatura médica española contemporánea: estudio estadístico y sociométrico.* Valencia: Cento de Documentación e Informática Médica, 1973. p. 85-92.

TERRADA, M.-L.; NAVARRO, V. La productividad de los autores españoles de bibliografía médica. *Revista Española de Documentación Científica,* v. 1, n. 1, p. 9-19.1977.

TURKELI, A. The doctoral training environment and post-doctorate productivity among Turkish physicists. *Science studies,* v. 3, n. 3, p. 311-318,1973.

TURKELI, A. Doctoral training environments and post-doctorate productivity of Turkish physicists. *Hacettepe Bulletin of Social Sciences and Humanities,* v. 5, n. 1, p. 91-100,1973.

URBIZAGÁSTEGUI ALVARADO, R. La ley de Lotka y la literatura de Bibliometria. *Investigación Bibliotecológica,* México, v. 13, n. 27, p. 125-141, jul./dic. 1999.

URBIZAGÁSTEGUI ALVARADO, R.; OLIVEIRA, M. de. A produtividade dos autores na antropologia brasileira. *DataGramaZero,* v. 2, n. 6, p. 1-17, dez. 2001.

URBIZAGÁSTEGUI ALVARADO, R.; LANE, S. *Lotka's Law*: a bibliography. Riverside, 2002.

URBIZAGÁSTEGUI ALVARADO, R. A lei de Lotka na bibliometria brasileira. *Ciência da Informação*. Brasília, v. 31. n. 2, p. 14-20, maio/ago. 2002.

URBIZAGÁSTEGUI ALVARADO, R.; CORTÉS, M. T. La productividad de autores en la Revista Geológica de Chile. *Ciencia de la Infomación,* Cuba, 2002. No prelo.

URBIZAGÁSTEGUI ALVARADO, R. *Lotka 's law on Lotka's literature:* an exploration. 2002. Datilogr.

VLACHÝ, J. On publication characteristics of research establishments. *Czechoslovak Journal of Physics,* v. B20, p. 1149-1155,1970.

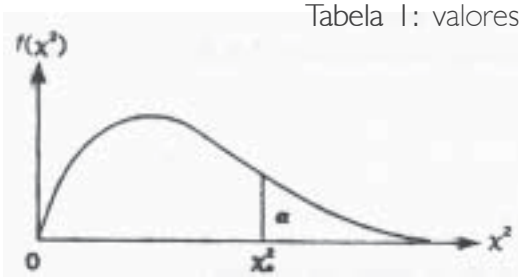
VLACHÝ, J. Distribution patterns in creative communities. In: WORLD CONGRESS OF SOCIOLOGY, 8., 1974, Toronto. *Proceedings...* Toronto, 1974.

VLACHÝ, J. Time factor in Lotka's law. *Probleme de Informare si Documentare*, v.10, n. 2, p. 44-87, mar./apr. 1976.

VOOS, H. Lotka and Information Science. *JASIS,* v. 25, n. 4, p. 270-272, July/Aug. 1974.

Anexo 1

Tabela 1: valores criticos del x^2



DEGREES OF FREEDOM	$X^2_{.995}$	$X^2_{.990}$	$X^2_{.975}$	$X^2_{.950}$	$X^2_{.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581

DEGREES OF FREEDOM	$X^2_{.100}$	$X^2_{.050}$	$X^2_{.025}$	$X^2_{.010}$	$X^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	13.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9796	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169