UNIVERSIDAD EAFIT
Abierta al mundo

Escuela de Economía y Finanzas

Documentos de trabajo
Economía y Finanzas

Centro de Investigación
Económicas y Financieras

No. 16-05
2016

The productivity of top researchers: A semi-nonparametric approach

Cortés, Lina M.; Perote, Javier; Mora-Valencia, Andrés

cief
Centro de Investigaciones Económicas y Financieras
Universidad EAFIT

The productivity of top researchers: A semi-nonparametric approach

Lina M. Cortés*
Department of Finance, School of Economics and Finance. Universidad EAFIT, Medellin, Colombia.
Phone: +5742619500 ext. 9756.
E-mail: lcortesd@eafit.edu.co

Javier Perote
Department of Economics and IME, University of Salamanca, Salamanca, Spain.
E-mail: perote@usal.es

Andrés Mora-Valencia
School of Management, Universidad de los Andes, Bogota, Colombia.
E-mail: a.mora262@uniandes.edu.co

*Corresponding author.

Abstract

Research productivity distributions exhibit heavy tails because it is common for a few researchers to accumulate the majority of the top publications and their corresponding citations. Measurements of this productivity are very sensitive to the field being analyzed and the distribution used. In particular, distributions such as the lognormal distribution seem to systematically underestimate the productivity of the top researchers. In this article, we propose the use of a (log)semi-nonparametric distribution (log-SNP) that nests the lognormal and captures the heavy tail of the productivity distribution through the introduction of new parameters linked to high-order moments. To compare the results, we use research performance data on 140,971 researchers who have produced 253,634 publications in 18 fields of knowledge (O'Boyle and Aguinis, 2012) and show how the log-SNP distribution provides more accurate measures of the performance of the top researchers in their respective fields of knowledge.

Keywords: Research evaluation, Research productivity, Heavy tail distributions, Semi-nonparametric modeling.

JEL Codes: C14, C44, C53

Introduction

In recent years, the evaluation of academic research productivity in different fields of knowledge has been related to the impact of the results of scientific production (Abramo et al. 2008; Sabharwal 2013; Campanario 2015). The motivation for studying productivity lies in the wish to promote academic excellence and render the research from each country as competitive as possible on the global stage (Frandsen 2005; Kocher et al. 2006; Abramo and D'Angelo 2014).

The quality of a research study is determined by a great number of variables, from the personal characteristics of the researcher to national and international policies and trends (Genest 1997; Dundar and Lewis 1998; Williamson and Cable 2003; Seggie and Griffith 2009). However, the criteria for evaluating research performance are combined mainly in two ways. First, the peer review process is assumed as the principal evaluation method, but this in turn is the object of a certain subjectivity level (Abramo et al. 2008, Bornmann 2011; Bertocchi et al. 2015; Day 2015).

Alternatively, another way of evaluating scientific activity in terms of productivity is based on bibliometric analysis. This method consists mainly of quantifying the number of documents published by a country, institution, research group or individual, as well as the citations received by such documents (Broadus 1987; Borokhovich et al. 1995; Abramo et al. 2008; Heberger et al. 2010; Finardi 2013; Bertocchi et al. 2015). The most common bibliometric measurements are those based on publications and citations, and this information comes from different databases such as Web of Science (WoS), Scopus, and Google Scholar, among others.

The majority of research productivity studies are focused on a single field of knowledge. For example, the literature focused on research performance in economics is abundant (Hodgson and Rothman 1999; Coupé 2003; Kocher et al. 2006; Ellison 2013). As a result, and taking into account the existing scientific advancements in each field of knowledge, it becomes relevant to study research productivity not only from the standpoint of measuring scientific production results, but also for the purpose of analyzing differences between the fields of knowledge in question (Sabharwal 2013; Abramo and D'Angelo 2014; Ruiz-Castillo and Costas 2014; Bertocchi et al. 2015).

In addition, studies on research productivity have taken into account different probability distribution functions in order to identify patterns in quantitative relationships between authors and their contributions over a period of time. These studies have determined that bibliometric indicators such as the number of articles published or the number of citations received by an author are characterized by distributions with heavy tails (Lotka 1926; Price 1976; Redner 1998; Chung and Cox 1990; Albarrán et al. 2011; Eom and Fortunato 2011; Da Silva et al. 2012; Ruiz-Castillo and Costas 2014; Campanario 2015).

As a result, the probability distribution models that have been applied the most in the literature on research productivity are those that obey the following laws: Lotka's law (Lotka 1926; Nicholls 1986; Chung and Cox 1990; Kretschmer and Kretschmer 2007), the power law (Price 1976; Egghe 2005; Albarrán et al. 2011; Aguinis et al., 2015) and Bradford's Law (Garfield 1980; Rousseau 1994; Nicolaisen and Hjørland 2007; Campanario 2015). These laws are mainly based on distribution functions such as the exponential or Pareto distributions. However, studies such as those by Kumar et al. (1998), Radicchi et al. (2008), Perc (2010), Eom and Fortunato (2011) and Birkmaier and Wohlrabe (2014) have proposed the application of the lognormal distribution to study research activity.

Nevertheless, all of these distributions have the disadvantage that they depend on very few parameters to capture the entire shape of the productivity distribution, particularly the right tail of the distribution. This makes the productivity measurements obtained very imprecise and comparisons of productivity between different fields of knowledge unreliable. To obtain reliable research productivity estimates, we propose the use of semi-nonparametric (SNP) approximations of productivity distributions based on the Edgeworth and Gram-Charlier expansions. These distributions have been applied in very diverse fields, where the precision of capturing the tails of distributions is important for the correct measurement of the frequency of extreme values (see Blinnikov and Moessner 1998, or Mauleon and Perote 2000, as examples of applications to astronomy or finance, respectively). In this article, we propose their use for the first time to measure research productivity and to determine with a higher degree of accuracy the quantiles that sort the most productive researchers in each field of knowledge as a proxy of the level of difficulty involved in being a star researcher in each field. In particular, we propose logarithmic transformations of an SNP distribution (which we refer to as log-SNP), which are extensions of a lognormal distribution that allow for approximating any empirical distribution through the introduction of additional parameters. Given that bibliometric indicators usually exhibit relatively long tails and multimodality (Guerrero-Bote et al. 2007;

Lancho-Barrantes 2010; Sabharwal 2013), we show that, compared to the lognormal distribution, the log-SNP distribution provides a better fit when characterizing research performance.

The productivity distribution

The characterization of a random variable through its probability density function (pdf) and its fit to the empirical distribution of a series can be achieved using different approaches, from a parametric perspective based on a frequency distribution with a known functional shape to a purely nonparametric approach. An intermediate possibility is the use of SNP approximations in which the functional shape is only partly parametrized, with the rest being an unknown function (Chen 2007). In this study, we consider an SNP approach in which the unknown function is modelled based on an orthogonal polynomial series expansion. In particular, we will analyze Edgeworth and Gram-Charlier expansions that have been shown to be valid asymptotic approximations of any empirical distribution under relatively weak regularity conditions (Sargan 1975; Phillips 1977). Next, we define the SNP distribution based on the Gram-Charlier series, as well as its logarithmic transformation, and analyze its basic properties.

The SNP distribution

Let $\{P_s(x)\}$, $x \in \mathbb{R}$ and $s \in \mathbb{N}$ be a family of orthogonal polynomials with respect to a density function $w(x)$ that satisfies the following relationship[1]

$$\int_{-\infty}^{\infty} P_s(x)P_j(x)w(x)dx = 0, \quad \forall s \neq j,\ s,j = 0,1,2,\dots \tag{1}$$

Within this family, Hermite polynomials (HPs) are those that use a standard normal density distribution, with weight $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. In particular, the HP of order *s*, $H_s(x)$, can be obtained in terms of the derivative of order *s* of the density function of the standard normal distribution, as expressed in equation (2):

$$H_s(x) = \frac{(-1)^s}{\phi(x)} \frac{d^s\phi(x)}{dx^s} \tag{2}$$

Next, we show the first eight HPs:

[1] Different weight functions $w(x)$ can be used; for details, see Abramowitz and Stegun (1972, pp. 774-775). We will consider $P_0(x) = 1$.

$$H_0(x) = 1 \tag{3}$$
$$H_1(x) = x \tag{4}$$
$$H_2(x) = x^2 - 1 \tag{5}$$
$$H_3(x) = x^3 - 3x \tag{6}$$
$$H_4(x) = x^4 - 6x^2 + 3 \tag{7}$$
$$H_5(x) = x^5 - 10x^3 + 15x \tag{8}$$
$$H_6(x) = x^6 - 15x^4 + 45x^2 - 15 \tag{9}$$
$$H_7(x) = x^7 - 21x^5 + 105x^3 - 105x \tag{10}$$
$$H_8(x) = x^8 - 28x^6 + 210x^4 - 420x^2 + 105 \tag{11}$$

It is easy to proof that these polynomials satisfy the mentioned orthogonality property given that $\forall s, j = 0, 1, 2, ...$

$$\int_{-\infty}^{\infty} H_s(x) H_j(x) \phi(x) dx = \begin{cases} 0, & s \neq j \\ s!, & s = j \end{cases} \tag{12}$$

The HPs also constitute the basis of the Edgeworth and Gram-Charlier (Type A) series, which allow, under certain regularity conditions (Cramér 1925), the expression of any pdf, $f(x)$, in terms of an infinite series (Wallace, 1958) as follows[2]

$$f(x) = \sum_{s=0}^{\infty} \delta_s H_s(x) \phi(x), \text{ where } \delta_s = \frac{1}{s!} \int_{-\infty}^{\infty} H_s(x) \phi(x) dx \tag{13}$$

Moreover, thanks to the orthogonality of the HPs, truncating the series to a specific order n of the expansion allows for defining a family of SNP distributions, $g(x; \boldsymbol{d})$, where $\boldsymbol{d} = (d_1, \dots, d_n)' \in \mathbb{R}^n$ denotes the vector of the parameters.[3]

$$g(x; \boldsymbol{d}) = [1 + \sum_{s=1}^{n} d_s H_s(x)] \phi(x) \underset{n \to \infty}{\longrightarrow} f(x) \tag{14}$$

However, the SNP distribution defined in equation (14) is only a density function for a subset of values of $\boldsymbol{d}$ that guarantee $g(x; \boldsymbol{d}) \geq 0$. To solve this problem, different types of restrictions or positivity transformations have been proposed (Gallant and Nychka 1987), even though they involve the introduction of unnecessary complexity for empirical applications that implement maximum likelihood (ML) algorithms (given that in the optimum ML leads to estimations that guarantee positivity).

[2] For more details about the Edgeworth and Gram-Charlier series, see Kendall and Stuart (1977 pp. 167-172).
[3] It must be noted that given a truncating order, the resulting distribution is purely parametric, but the truncating order is flexible to achieve a more accurate approximation to a given distribution. Without loss of generality, we will assume that $d_0 = 1$.

The great advantage of SNP distributions when compared to other parametric specifications lies precisely in the improved fit to empirical data, given their great parametric flexibility. In addition, the resulting higher number of parameters does not involve more complexity in theoretical or empirical terms. For example, the central moments can be easily obtained as linear functions of the distribution parameters (see Appendix A). Note that the even (odd) moment of order *n* depends only on the *n* first even (odd) parameters. This fact allows for the search of initial values for the optimization logarithms through the direct application of the method of moments (MM). A closed expression can also be obtained for the cumulative distribution function (cdf) of the SNP distribution as a function of the normal distribution cdf, as shown in equation (15) (see the proof in Appendix B). This allows for a simple calculation of the probabilities and quantiles of the SNP distribution.

$$G_x(a) = \int_{-\infty}^{a} g(x;\boldsymbol{d})dx$$

$$= \int_{-\infty}^{a} \phi(x)dx - \phi(a)\sum_{s=1}^{n} d_s H_{s-1}(a) \qquad (15)$$

The log-SNP distribution

Ñíguez et al. (2012) define a variable $z > 0$ as (standard) log-SNP if the variable $x = \log(z)$ is SNP distributed and its pdf defined as in equation (14). The resulting distribution inherits all the good properties of the SNP distribution, particularly its flexibility in capturing the extreme values of the distribution. We will go a step further and similarly define a log-SNP distribution, but rather over a linear transformation $y = \sigma x + \mu$.

Definition: We will say that the variable $z > 0$ is log-SNP distributed with location parameter $\mu \in \mathbb{R}$, scale $\sigma^2 \in \mathbb{R}$ and shape parameters $\boldsymbol{d} = (d_1, \dots, d_n)' \in \mathbb{R}^n$ if its pdf can be expressed as

$$h(\mathrm{z};\ \mu, \sigma^2, \boldsymbol{d}) = \left[1 + \sum_{s=1}^{n} d_s H_s\left(\frac{\log(z)-\mu}{\sigma}\right)\right]\left(\frac{1}{z\sigma\sqrt{2\pi}} e^{\frac{-(\log(z)-\mu)^2}{2\sigma^2}}\right). \qquad (16)$$

Defined in this manner, the lognormal distribution is a particular case of the log-SNP (for $d_s = 0,\ \forall s$), which allows for a comparison of the improvements in the fit of the latter to those obtained with the lognormal by using linear restrictions tests such as the likelihood ratio (LR). This article shows that, in effect, the parametric flexibility of the log-SNP allows for significant fit improvements to productivity distributions, as the log-SNP is capable of

representing different shapes (including jumps in the probability mass function and heavy tails) through the incorporation of parameters in addition to those of a traditional parametric distribution, such as the lognormal distribution.

Data and Methodology

Data

To test whether a lognormal or a log-SNP distribution fits the best to the performance distribution of 140,971 researchers who have produced 253,634 publications in 18 fields of knowledge, we used the data from O'Boyle and Aguinis (2012).[4] These authors classified the fields of knowledge based on the Journal Citation Reports (JCR), which provide impact factors (IFs) in different fields of knowledge labeled within the categories of "sciences" and "social sciences".

The authors used the IFs to select the five main journals within each field of knowledge. For these journals, they identified all the authors who published at least one article in one of them during the period that ranges from January of 2000 to June of 2009.[5] With this information they measured the productivity of the researchers as the number of articles published by an author in each of the fields of knowledge during the observation period of 9.5 years.

In addition, we use the JCR of the year 2007[6] for each of the categories sciences and social sciences in order to obtain the Median Impact Factor (MIF) indicator of the five main journals in each of the selected fields of knowledge. The objective behind this indicator is to obtain a broader view of each of the selected fields and to make inferences about their behavior. Table 1 shows the descriptive statistics of the sample selected in this study.

[Insert Table 1]

It can be seen that throughout the 18 fields of knowledge analyzed, the minimum number of researchers is 1,073 for the field of Ethics and the maximum is 30,531 for Dermatology. The publications average in the five top journals for each field and for each of the researchers in the sample varies from 1.42 to 2.26, and the standard deviation has a range

[4] The authors thank Herman Aguinis and Ernest O'Boyle for allowing us to use their database on academic productivity compiled in O'Boyle and Aguinis (2012).
[5] For details about the data treatment, see O'Boyle and Aguinis (2012), p.86.
[6] We took the JCR of the year 2007 to be consistent with O'Boyle and Aguinis (2012), as that was the year used by the authors to select the five main journals within each field of knowledge.

of 0.97 to 3.38 publications. When analyzing the skewness and excess kurtosis of the productivity distribution, it is clear that all the fields exhibit positive skewness and leptokurtosis, with the field of Genetics being the most skewed and leptokurtic of the sample. The maximum number of articles per researcher varies from 13 (Law) to 120 (Genetics), depending on the field considered.

In addition, we find large differences when considering the MIF indicator (of the top five journals in each area), which varies from 0.85 (History) to 18.30 (Genetics). Furthermore, as seen in Table 1, the MIF is related to the maximum number of articles per researcher. As a result, Genetics has the highest MIF and the maximum number of publications per researcher, while the MIF of History places 18th and 17th in number of publications per researcher.

The results show that the behavior of each field of knowledge is different. The fields that belong to the Sciences JCR category have a larger number of researchers and, then, a larger MIF. Therefore, this exploratory analysis suggests that the level of research productivity that can be attained depends on the field of knowledge being studied.

Methodology

This section presents the methodology applied to characterize the research productivity in each field of knowledge based on the log-SNP distribution. Details are provided on the ML estimation methodology and its related goodness of fit measures used to choose between the different pdfs nested on the family of log-SNP distributions (including the lognormal). The pdf of the log-SNP distribution is sequentially estimated up to a truncating order of $n = 8$.

Let z_i be the number of articles published by an author in one of the selected fields of knowledge; the log-likelihood function for a log-SNP$(\mu, \sigma^2, \boldsymbol{d})$ distributed observation truncated to the eighth moment is given by:

$$logL(\mu, \sigma^2, \boldsymbol{d}|z_i) = -\frac{1}{2}\log(2\pi\sigma^2 z_i^2) - \frac{1}{2}\left(\frac{\log(z_i)-\mu}{\sigma}\right)^2 + \log\left[1 + \sum_{s=1}^{8} d_s H_s\left(\frac{\log(z_i)-\mu}{\sigma}\right)\right] \quad (17)$$

The sequential estimation begins with the simplest nested density, the lognormal, and the d_s parameters are recursively added, the initial values of which are selected consistently with their sample moments counterparts. The inclusion of new parameters in the productivity distribution is performed according to accuracy criteria, i.e. the log-likelihood (logL) and the Akaike Information Criterion (AIC), and linear restrictions tests provided by the LR statistic.

Based on these criteria, n=8 was selected as the optimum truncating order, and only the even parameters, d_2, d_4, d_6 and d_8, were selected.

Results

Table 2 presents the ML estimates of the parameters of the performance distributions for each of the fields selected. Panel A shows the estimated parameters for a lognormal distribution, and Panel B shows the estimated parameters for the log-SNP distribution. Panel C displays the LR statistic for comparing the log-SNP and the lognormal distributions.

[Insert Table 2]

The results of the estimation reveal that all the models adequately capture the mean and standard deviation of each of the fields, denoted as parameters μ and σ, respectively. The p-values clearly indicate that these parameters are highly significant for both distributions. However, as shown in Panel B, for the log-SNP distribution, the d_s parameters are also highly significant for the majority of fields of knowledge. When analyzing the AIC (which penalizes log-likelihood value with the inclusion of additional parameters) for the two distributions, we found that this criterion is consistently lower for the log-SNP distribution, which suggests that the modeling based on this distribution is clearly superior. In addition, from the LR statistics included in Panel C, we conclude that for all the selected fields, incorporating the d_s parameters improves the accuracy of the model.

An example of the fit quality obtained for two selected randomly fields, Finance and Dentistry, is captured in Fig. 1. This figure depicts the empirical histogram and pdf values estimated under a lognormal specification and under the log-SNP. In both cases, the log-SNP distributions more adequately capture not just the values around the mean but also the extreme values. Fig. 2 shows in detail the right tails of the distribution, which capture the frequency of the researchers with higher productivity. From these figures it is clear that the log-SNP specification allows the better characterization of the research activity.

[Insert Fig. 1]

[Insert Fig. 2]

Fig. 3 shows the comparison between the fitted densities for Finance and Dentistry in terms of the empirical and theoretical cdfs for both specifications, the log-SNP and the

lognormal. The latter appears to underestimate the cumulative probability (especially for Dentistry) when compared to the log-SNP.

[Insert Fig. 3]

The Fig. 3 shows how the lognormal distribution underestimates research productivity, especially for the more extreme values (under the lognormal distribution, a researcher must publish less articles to be included in the top quantiles of the performance distribution). Table 3 illustrates these effects for the different fields of knowledge by computing the empirical and estimated quantiles under the lognormal and log-SNP for confidence levels of 5%, 1%, 0.1% and 0.05%.[7]

[Insert Table 3]

The values in the table clearly indicate the higher accuracy of the log-SNP distribution fits, particularly in the tails, and the underestimation of the productivity of top researchers obtained from the traditional parametric distributions such as the lognormal. For example, for the field of Agronomy, it can be seen that to belong to the top 0.05% of researchers who publish the highest number of articles in the best journals, 15 publications are empirically required. This limit is much less strict if we assume that the distribution is lognormal (6) as compared to log-SNP (12). These results are consistent with the research by Kumar et al. (1998), Perc (2010) and Eom and Fortunato (2011), who, when applying the lognormal distribution to bibliometric indicators, found that it fell short when modeling series with very heavy tails.

Conclusions

Bibliometric analysis has been shown to be a valuable method for evaluating scientific production and has a growing impact in the academia. However, the literature indicates that in most cases, the distributions commonly used for measuring productivity have been shown to underestimate the behavior of the top researchers, given that their productivity seems to be generated by a distribution with very heavy tails. This fact calls for the search of more appropriate distributions and methodologies.

This study analyzes the research productivity in 18 fields of knowledge belonging to the JCR categories of sciences and social sciences between the years 2000 and 2009. The results

[7] The quantiles of the log-SNP distribution are obtained from the cdf displayed in equation (15) and the Inverse Transform Method (ITM).

show that the level of productivity, as measured by the number of publications per author, depends on the field of knowledge being studied. In particular, the fields that belong to the category of sciences have a higher number of publications per author. In addition, we observe that the MIF indicator is highly correlated to the maximum number of articles per researcher; that is, the greater the number of articles published in top journals by each researcher (usually the most cited), the greater the MIF by field of knowledge.

This study proposes a novel methodology based on the log-SNP distribution for measuring the scientific productivity distribution of top researchers in different fields of knowledge. Such a distribution nests the lognormal and includes new parameters for accurately capturing the heavy tail of the research productivity distribution. Our study shows that the log-SNP provides a better fit of research performance distribution than the lognormal and quantifies the differences in the measures of the top researchers' productivity attached to the distributional hypothesis. We also find that the results are very sensitive to the field of knowledge being studied and thus the productivity of top researchers depends on the field of knowledge.

References

Abramo, G., & D'Angelo, C. A. (2014). Assessing national strengths and weaknesses in research fields. *Journal of Informetrics, 8(3)*, 766–775.

Abramo, G., D'Angelo, A. C., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics, 76(2)*, 225–244.

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications.

Aguinis, H., O'Boyle, E., Gonzalez-Mulé, E., & Joo, H. (2015). Cumulative advantage: Conductors and insulators of heavy-tailed productivity distributions and productivity tars. *Personnel Psychology*, http://dx.doi.org/10.1111/peps.12095 (in press).

Albarrán, P., Juan, A. C., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics, 88(2)*, 385-397.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44(2)*, 451-466.

Birkmaier, D., & Wohlrabe, K. (2014). The Matthew effect in economics reconsidered. *Journal of Informetrics, 8(4)*, 880–889.

Blinnikov, S., & Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and astrophysics Supplement Series, 130(1)*, 193–205.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45(1)*, 199–245.

Borokhovich, K. A., Bricker, R. J., Brunarski, K. R., & Simkins, B. J. (1995). Finance research productivity and influence. *The Journal of Finance, 50(5)*, 1691-1717.

Broadus, R. N. (1987). Toward a definition of 'bibliometrics'. *Scientometrics, 12(5-6)*, 373-379.

Campanario, J. M. (2015). Providing impact: The distribution of JCR journals according to references they contribute to the 2-year and 5-year journal impact factors. *Journal of Informetrics, 9(2)*, 398–407.

Chen, X. (2007). Large Sample Sieve Estimation of Semi-Nonparametric Models. En J. Heckman, & E. Leamer, *Handbook of Econometrics, Vol. 6, Ch. 76, Part B* (págs. 5549-5632). Elsevier.

Chung, K. H., & Cox, R. A. (1990). Patterns of productivity in the finance literature: a study of the bibliometric distributions. *The Journal of Finance, 45(1) 1, 301-309*, 301-309.

Coupé, T. (2003). Revealed performances. Worldwide rankings of economists and economics departments. *Journal of the European Economic Association, 1(6)*, 1309–1345.

Cramér, H. (1925). On some classes of series used in mathematical statistics. *Sixth Scandinavian Congress of Mathematicians*, (págs. 399-425). Copenhagen.

Da Silva, R., Kalil, F., De Oliveira, J. M., & Martinez, A. S. (2012). Universality in bibliometrics. *Physica A: Statistical Mechanics and its Applications, 391(5)*, 2119-2128.

Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. *Research Policy, 44(6)*, 1266–1270.

Del Brio, E. B., & Perote, J. (2012). Gram–Charlier densities: Maximum likelihood versus the method of moments. *Insurance: Mathematics and Economics, 51(3)*, 531-537.

Dundar, H., & Lewis, D. (1998). Determinants of research productivity in higher education. *Research in Higher Education, 39(6)*, 607-631.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics.* Kidlington, UK: Elsevier Academic Press.

Ellison, G. (2013). How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics, 5(3)*, 63-90.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6(9)*, e24926.

Finardi, U. (2013). Correlation between Journal Impact Factor and Citation Performance: An experimental study. *Journal of Informetrics, 7(2)*, 357–370.

Frandsen, T. F. (2005). Geographical concentration. The case of economics journals. *Scientometrics, 63(1)*, 69-85.

Gallant, A. R., & Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica, 55(2)*, 363–390.

Garfield, E. (1980). Bradford's Law and related statistical pattern. *Essays of an Information Scientist, 4(19)*, 476-483.

Genest, C. (1997). Statistics on statistics: measuring research productivity by journal publications between 1985 and 1995. *The Canadian Journal of Statistics, 25(4)*, 427-443.

Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Gomez-Crisostomo, R., & Moya-Anegon, F. (2007). Import–export of knowledge between scientific subject categories: The iceberg hypothesis. *Scientometrics, 71(3)*, 423–441.

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation, 31(1)*, 24-44.

Hodgson, G. M., & Rothman, H. (1999). The editors and authors of economics journals: A case of institutional oligopoly? *The Economic Journal, 109(453),* 165–186.

Kendall, M., & Stuart, A. (1977). *The Advanced Theory of Statistics, Vol. I, 4th ed.* London: C. Griffin.

Kocher, M. G., Luptacik, M., & Sutter, M. (2006). Measuring productivity of research in economics: A cross-country study using DEA. *Socio-Economic Planning Sciences, 40(4)*, 314-332.

Kretschmer, H., & Kretschmer, T. (2007). Lotka's distribution and distribution of co-author pairs' frequencies. *Journal of Informetrics, 1(4)*, 308–337.

Kumar, S., Sharma, P., & Garg, K. C. (1998). Lotka's law and institutional productivity. *Information Processing & Management, 34(6)*, 775–783.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). The iceberg hypothesis revisited. *Scientometrics, 85(2)*, 443–461.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science, 16(12)*, 317-323.

Mauleón, I., & Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth-Sargan density to Student's t. *European Journal of Finance, 6(2)*, 225-239.

Nicholls, T. P. (1986). Empirical validation of Lotka's law. *Information Processing & Management, 22(5)*, 417–419.

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of Bradford's law: a critical examination of the received view. *Journal of Documentation, 63(3)*, 359 - 377.

Ñíguez, T. M., Paya, I., Peel, D., & Perote, J. (2012). On the stability of the constant relative risk aversion (CRRA) utility under high degrees of uncertainty. *Economics Letters, 115(2)*, 244-248.

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65(1)*, 79–119.

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics, 4(2)*, 358–364.

Phillips, P. B. (1977). A general theorem in the theory of asymptotic expansions as approximations to the finite sample distributions of econometric estimators. *Econometrica, 45(6)*, 1517–1534.

Price, D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27(5)*, 292–306.

Radicchi, F., Fortunado, S., & Castellano, C. (2008). Universality of citation distribution: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 105(45)*, 17268–17272.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems, 4(2)*, 131-134.

Rousseau, R. (1994). Bradford curves. *Information Processing and Management, 30(2)*, 267–277.

Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics, 8(4)*, 917–934.

Sabharwal, M. (2013). Comparing research productivity across disciplines and career stages. *Journal of Comparative Policy Analysis: Research and Practice, 15(2)*, 141-163.

Sargan, D. (1975). Gram-Charlier approximation applied t ratios or k-class estimatiors. *Econometrica, 43(2)*, 327-346.

Seggie, S. H., & Griffith, D. A. (2009). What does it take to get promoted in marketing academia? Understanding exceptional publication productivity in the leading marketing journals. *Journal of Marketing, 73(1)*, 122-132.

Wallace, D. L. (1958). Asymptotic approximations to distributions. *Annals of Mathematical Statistics, 29(3)*, 635-654.

Williamson, I. O., & Cable, D. M. (2003). Predicting early career research productivity: The case of management faculty. *Journal of Organizational Behavior, 24(1)*, 25-44.

Appendix A

This appendix lists the first eight d_s parameters in terms of the central moments of the SNP distribution. For more information, see Del Brio and Perote (2012).

$$d_1 = \mu_1 \tag{A.1}$$

$$d_2 = \frac{1}{2}(\mu_2 - 1) \tag{A.2}$$

$$d_3 = \frac{1}{6}(\mu_3 - 3\mu_1) \tag{A.3}$$

$$d_4 = \frac{1}{24}(\mu_4 - 6\mu_2 + 3) \tag{A.4}$$

$$d_5 = \frac{1}{120}(\mu_5 - 10\mu_3 + 15\mu_1) \tag{A.5}$$

$$d_6 = \frac{1}{720}(\mu_6 - 15\mu_4 + 45\mu_2 - 15) \tag{A.6}$$

$$d_7 = \frac{1}{5040}(\mu_7 - 21\mu_5 + 105\mu_3 - 105\mu_1) \tag{A.7}$$

$$d_8 = \frac{1}{40320}(\mu_8 - 28\mu_6 + 210\mu_4 - 420\mu_2 + 105) \tag{A.7}$$

Appendix B

This appendix derives the cdf of the SNP distribution.

$$G_x(a) = \int_{-\infty}^{a} g(x;\boldsymbol{d})dx = \int_{-\infty}^{a} \phi(x)dx + \sum_{s=1}^{n} d_s \int_{-\infty}^{a} H_s(x)\phi(x)dx$$

$$= \int_{-\infty}^{a} \phi(x)dx - \sum_{s=1}^{n} d_s H_{s-1}(x)\phi(x)\Big|_{-\infty}^{a}$$

$$= \int_{-\infty}^{a} \phi(x)dx - \phi(a)\sum_{s=1}^{n} d_s H_{s-1}(a)$$

Given that $\lim_{x\to\pm\infty} H_s(x)\phi(x) = 0 \ \ \forall s \geq 1$, it follows that

$$\int H_s(x)\phi(x)dx = \int (-1)^s \frac{d^s\phi(x)}{dx^s}dx_t = (-1)^s \frac{d^{s-1}\phi(x)}{dx^{s-1}}$$

$$= (-1)^s(-1)^{s-1}H_{s-1}(x)\phi(x) = -H_{s-1}(x)\phi(x).\square$$

Table 1 Descriptive statistics

Field of knowledge	N	Mean	Std	Skew	K	Max	MIF	N (ordinal position)	Max (ordinal position)	MIF (ordinal position)	JCR edition
Agronomy	8,923	1.42	1.16	6.36	72.68	26	2.36	7	13	12	Science
Anthropology	5,755	1.87	1.95	4.49	34.52	30	2.31	9	8	14	Social Sciences
Clinical psychology	10,418	1.89	2.38	10.80	267.22	93	4.68	6	2	3	Social Sciences
Dentistry	12,345	2.26	2.98	6.54	74.62	66	3.37	3	4	6	Science
Dermatology	30,531	2.25	3.38	8.01	113.19	93	3.50	1	2	5	Science
Ecology	5,730	1.71	1.68	7.88	148.90	50	4.82	10	6	2	Science
Economics	3,048	1.62	1.67	7.14	82.10	27	3.69	13	11	4	Social Sciences
Educational psychology	3,032	1.70	1.55	5.41	52.04	27	2.35	14	11	13	Social Sciences
Ethics	1,073	1.65	1.78	6.82	71.24	26	1.31	18	13	16	Social Sciences
Ethnic studies	2,003	1.48	1.38	5.99	50.89	17	0.89	16	15	17	Social Sciences
Finance	3,019	2.14	2.52	4.69	33.93	28	2.99	15	9	8	Social Sciences
Forestry	12,211	1.82	1.80	5.66	68.58	46	2.14	4	7	15	Science
Genetics	16,574	1.71	2.18	26.42	1240.47	120	18.30	2	1	1	Science
History	6,708	1.54	0.97	3.33	25.11	14	0.85	8	17	18	Social Sciences
Law	1,350	1.55	1.24	3.88	24.07	13	3.09	17	18	7	Social Sciences
Linguistics	3,600	1.73	1.78	5.98	59.06	28	2.37	12	9	11	Social Sciences
Mathematics	3,972	1.45	1.02	4.86	41.42	15	2.56	11	16	10	Science
Statistics	10,679	2.08	2.52	6.22	67.39	54	2.97	5	5	9	Science

This table shows the descriptive statistics of the publications in the five top journals for 18 fields of knowledge belonging to the JCR categories of sciences and social sciences between the years 2000 and 2009. N=number of researchers, Std=standard deviation, Skew=skewness, K=excess kurtosis coefficient, Max=maximum score, MIF=Median Impact Factor (five top journals in 2007).

Table 2 Results of the estimation

Field of knowledge	Panel A Lognormal				Panel B Log-SNP								Panel C LR
	μ	σ	logL	AIC	μ	σ	d_2	d_4	d_6	d_8	logL	AIC	
Agronomy	0.2143	0.4368	-3359.52	6723.04	0.1182	0.4771	-0.0786	0.1448	0.0252	0.0042	-1890.49	3792.98	2938.07
	(<.0001)	(<.0001)			(0.000)	(<.0001)	(0.000)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Anthropology	0.3753	0.6024	-3089.70	6183.40	0.1693	0.5438	0.1912	0.2733	0.0408	0.0050	-2259.29	4530.58	1660.83
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Clinical psychology	0.3791	0.5994	-5501.26	11006.52	0.1689	0.5556	0.1535	0.2611	0.0444	0.0055	-4236.31	8484.63	2529.90
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Dentistry	0.4934	0.6763	-6598.224	13200.45	0.2959	0.6913	0.0194	0.1481	0.0157	0.0027	-5740.93	11493.86	1714.58
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.1765)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Dermatology	0.4553	0.6914	-18154.32	36312.64	0.8375	0.4294	1.1923	0.3812	0.1092	0.0179	-7262.16	14536.32	21784.32
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Ecology	0.3335	0.5445	-2736.83	5477.66	0.1653	0.5435	0.0499	0.1708	0.0174	0.0037	-2027.75	4067.50	1418.16
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.0023)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Economics	0.2887	0.5198	-1450.68	2905.37	0.1418	0.5133	0.0538	0.2073	0.0277	0.0041	-935.65	1883.29	1030.08
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.0819)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Educational psychology	0.3404	0.5320	-1356.60	2717.21	0.1764	0.5367	0.0381	0.1614	0.0194	0.0034	-1108.26	2228.51	496.70
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.0900)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Ethics	0.2952	0.5262	-516.72	1037.45	0.1556	0.5301	0.0282	0.2231	0.0351	0.0048	-338.55	689.11	356.34
	(<.0001)	(<.0001)			(0.0028)	(<.0001)	(0.4423)	(<.0001)	(0.0017)	(<.0001)			(<.0001)
Ethnic studies	0.2287	0.4647	-849.22	1702.44	0.1290	0.5045	-0.0854	0.1877	0.0347	0.0050	-511.39	1034.78	675.66
	(<.0001)	(<.0001)			(0.0038)	(<.0001)	(0.0011)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Finance	0.4560	0.6688	-1692.96	3389.92	0.1693	0.5763	0.2975	0.2992	0.0484	0.0060	-1390.41	2792.82	605.10
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)

continues

Table 2 (continued)

Field of knowledge	Panel A Lognormal				Panel B Log-SNP								Panel C LR
	μ	σ	logL	AIC	μ	σ	d_2	d_4	d_6	d_8	logL	AIC	
Forestry	0.3785	0.5755	-5958.31	11920.63	0.1797	0.5490	0.1149	0.1942	0.0232	0.0037	-4879.51	9771.02	2157.61
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Genetics	0.3338	0.5350	-7617.37	15238.74	0.1720	0.5379	0.0399	0.1748	0.0224	0.0037	-6015.27	12042.54	3204.20
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
History	0.3080	0.4570	-2198.69	4401.39	0.1984	0.5112	-0.0776	0.0627	-0.0004	0.0013	-2095.15	4202.29	207.10
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.8251)	(<.0001)			(<.0001)
Law	0.2788	0.4908	-578.29	1160.59	0.1507	0.4953	0.0244	0.1747	0.0163	0.0027	-389.59	791.18	377.40
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(0.6560)	(<.0001)	´(0.0272)	(<.0001)			(<.0001)
Linguistics	0.3307	0.5556	-1801.66	3607.31	0.1558	0.5395	0.0844	0.2007	0.0246	0.0042	-1270.77	2553.54	1061.77
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)
Mathematics	0.2458	0.4342	-1346.20	2696.39	0.1652	0.4945	-0.1013	0.1159	0.0071	0.0019	-971.81	1955.62	748.77
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0210)	(<.0001)			(<.0001)
Statistics	0.4510	0.6390	-5553.69	11111.38	0.2429	0.6251	0.0779	0.1858	0.0253	0.0036	-4758.50	1590.38	1590.38
	(<.0001)	(<.0001)			(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)			(<.0001)

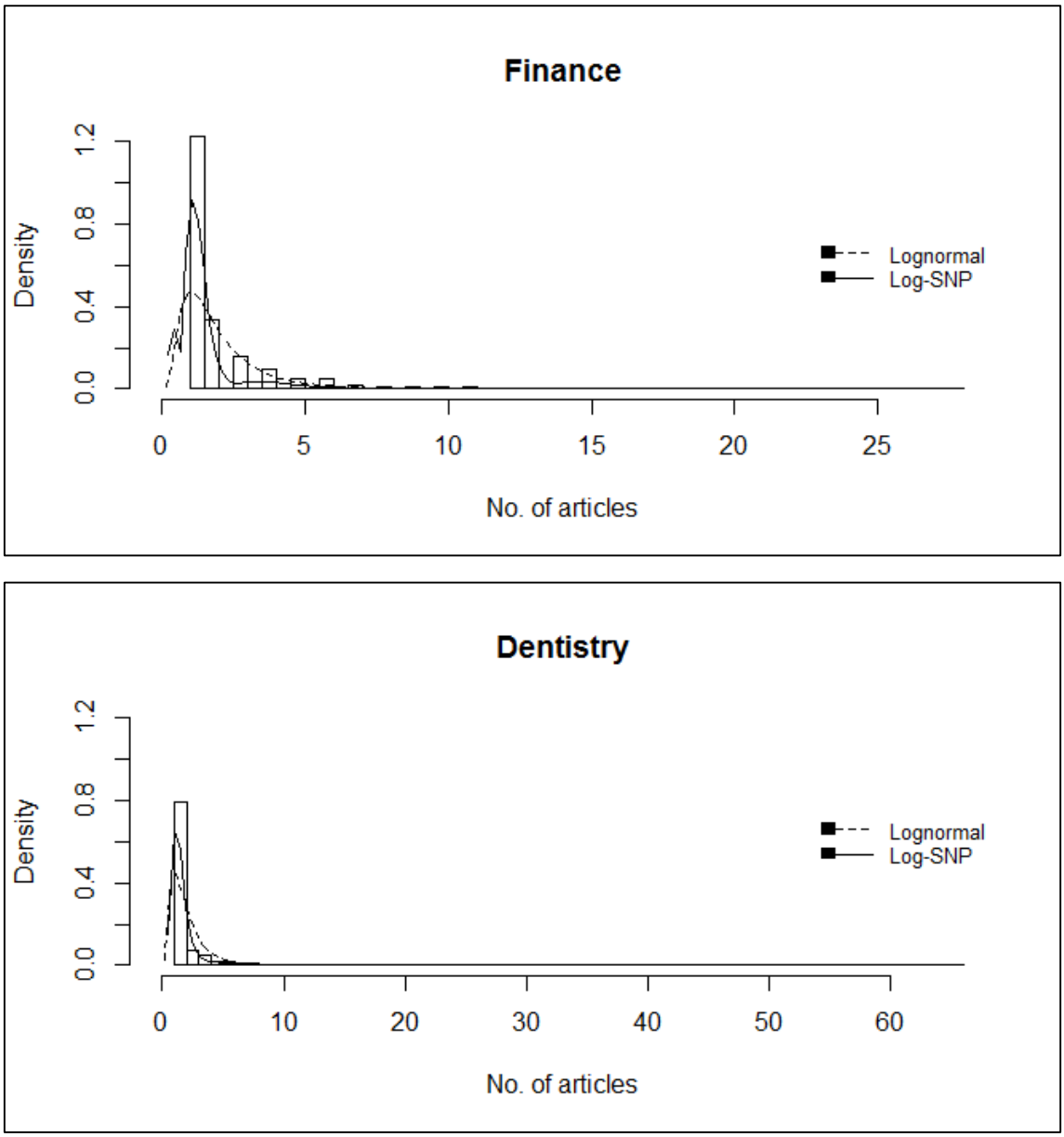
This table reports the ML estimation for each of the fields selected. Panel A shows the estimated parameters for the lognormal distribution. Panel B shows the estimated parameters for the log-SNP distribution. Panel C shows the likelihood ratio applied to both distributions. μ and σ are the location and scale parameters, respectively, and d_2, d_4, d_6 and d_8 are the weight parameters of the Hermite polynomials. logL=log-likelihood, AIC= Akaike Information Criterion, LR=likelihood ratio for testing the log-SNP and lognormal. P-values are shown in parentheses. The study corresponds to 18 fields of knowledge that belong to the JCR categories of sciences and social sciences between the years 2000 and 2009.

Table 3 Number of articles observed empirically versus those expected theoretically under the lognormal and log-SNP

Field of knowledge	N	Observed No. of articles Top				Expected number of articles							
						Lognormal Top				Log-SNP Top			
		5%	1%	0.1%	0.05%	5%	1%	0.1%	0.05%	5%	1%	0.1%	0.05%
Agronomy	8,923	3	7	13	15	3	4	5	6	3	4	10	12
Anthropology	5,755	5	10	19	22	4	6	10	11	4	9	16	17
Clinical psychology	10,418	5	11	27	35	4	6	10	11	4	9	17	19
Dentistry	12,345	7	15	32	36	5	8	14	16	5	11	29	34
Dermatology	30,531	7	16	40	50	5	8	14	16	7	14	20	22
Ecology	5,730	4	8	17	20	4	5	8	9	4	7	14	16
Economics	3,048	4	8	25	26	4	5	7	8	3	7	13	14
Educational psychology	3,032	4	8	18	18	4	5	8	9	4	7	14	16
Ethics	1,073	4	9	24	25	4	5	7	8	3	8	14	16
Ethnic studies	2,003	3	8	16	16	3	4	6	6	3	5	12	14
Finance	3,019	6	13	26	28	5	8	13	15	5	11	19	21
Forestry	12,211	5	9	18	22	4	6	9	10	4	8	15	17
Genetics	16,574	4	8	18	23	4	5	8	9	4	7	14	16
History	6,708	3	5	8	12	3	4	6	7	3	5	8	11
Law	1,350	4	7	13	13	3	5	7	7	3	6	11	12
Linguistics	3,600	5	9	22	23	4	6	8	9	4	7	14	16
Mathematics	3,972	3	6	13	14	3	4	5	6	3	5	10	11
Statistics	10,679	6	13	26	35	5	7	12	13	5	10	22	26

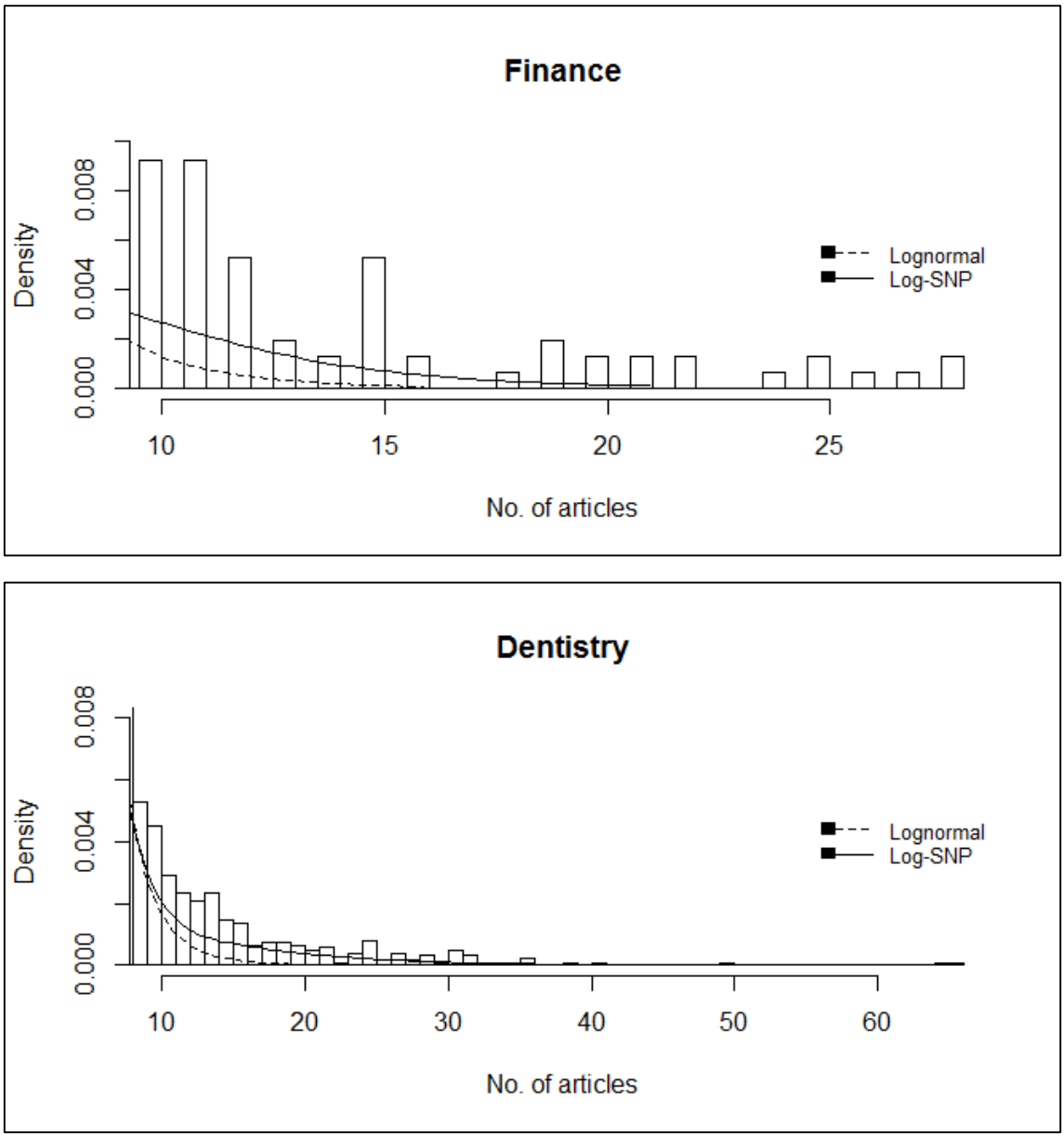
This table compares the number of articles observed empirically in each of the fields with those theoretically expected under the lognormal and log-SNP distributions. N=number of researchers. The values 5%, 1%, 0.1% and 0.05% are distribution percentiles. The study corresponds to 18 fields of knowledge that belong to the JCR categories of sciences and social sciences between the years 2000 and 2009.

Fig. 1 Pdf of research productivity in Finance and Dentistry



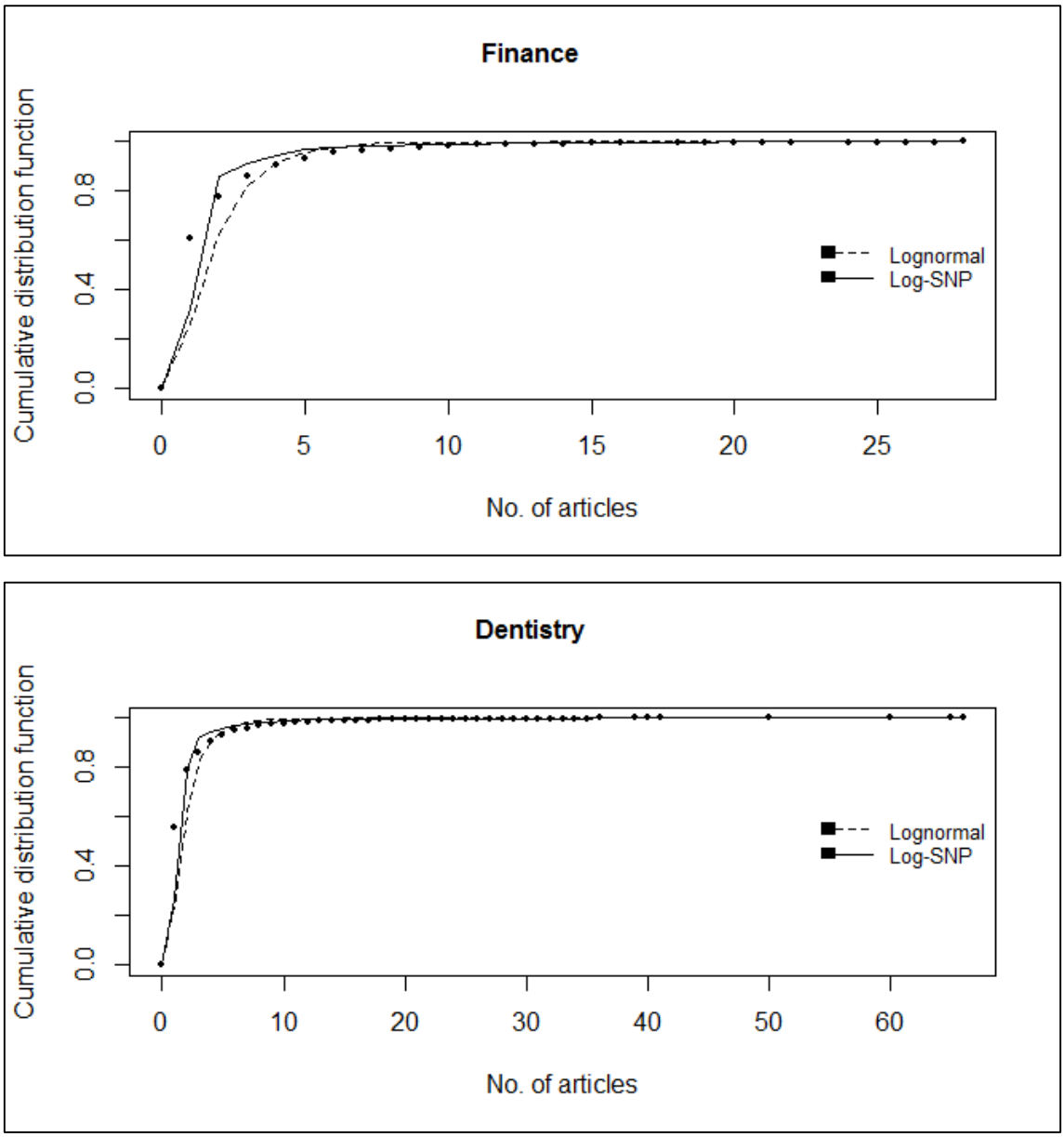
The figure shows the distribution of the empirical frequencies (histogram) of the productivity of the researchers who published in the five top journals (in JCR-2007 terms) in Finance and Dentistry during the period 2000-2009. The estimated pdfs under the lognormal and log-SNP specifications are depicted in dashed line and solid line, respectively.

Fig. 2 Pdf of research productivity in Finance and Dentistry



The figure shows the right tail of the distribution of empirical frequencies (histogram) of productivity of the researchers who published in the five top journals (in JCR-2007 terms) in Finance and Dentistry during the year 2000-2009. The fitted lognormal and log-SNP pdfs are depicted in dashed line and solid line, respectively.

Fig. 3 Cdf of research productivity in Finance and Dentistry



The figure shows the empirical cumulative distribution function of the productivity of the researchers who published in the five top journals (in JCR-2007 terms) in Finance and Dentistry during the period 2000-2009. The fitted lognormal and log-SNP cdfs are depicted in dashed line and solid line, respectively.