Scientometrics (2016) 109:891–915
DOI 10.1007/s11192-016-2072-5

The productivity of top researchers: a semi-nonparametric approach

Lina M. Cortés[1] · Andrés Mora-Valencia[2] · Javier Perote[3]

Received: 24 December 2015 / Published online: 23 July 2016

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. The application uses scientific production data on 140,971 researchers who have produced 253,634 publications in 18 fields of knowledge (O'Boyle and Aguinis in Pers Psychol 65(1):79–119, 2012) and publications in the field of finance of 330 academic institutions (Borokhovich et al. in J Finance 50(5):1691–1717, 1995), and shows that the log-SNP distribution outperforms the lognormal and provides more accurate measures for the high quantiles of the productivity distribution.

Keywords Research evaluation · Research productivity · Heavy tail distributions · Semi-nonparametric modeling

✉ Lina M. Cortés
lcortesd@eafit.edu.co

Andrés Mora-Valencia
a.mora262@uniandes.edu.co

Javier Perote
perote@usal.es

[1] Department of Finance, School of Economics and Finance, Universidad EAFIT, Carrera 49 No 7 Sur-50, Medellin, Colombia

[2] School of Management, Universidad de los Andes, Calle 21 No. 1-20, Bogota, Colombia

[3] Department of Economics and IME, University of Salamanca, Campus Miguel de Unamuno, 37007 Salamanca, Spain

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; Duch et al. 2012; Kaur et al. 2015). However, the criteria for evaluating research productivity 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; Kaur et al. 2015; 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. However, the heterogeneity in publication and citation policies between the different fields of knowledge (Kaur et al. 2013; Ruiz-Castillo and Costas 2014; Mingers and Leydesdorff 2015) make the direct comparison in terms of the number of published articles and cites 'unfair' (Crespo et al. 2012) and raise the need for the search of more appropriate methods of comparison.

The majority of research productivity studies are focused on a single field of knowledge. For example, the literature focused on research productivity 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, mainly based on distribution functions such as the exponential or

Pareto distributions, have been controversial and have generated a strong debate during more than a century. For instance, Newman (2005) asserted that few real-world processes follow a power law over their entire range, and in particular not for smaller values of the variable being measured. Martínez-Mekler et al. (2009) argued that, when real data are used, power laws hold only for an intermediate range of values, whereas the tails of the distributions tend to deviate from the values expected according to the power law. Therefore, the authors suggested that the two-parameter law incorporates the product of two power laws defined over the complete data set: One of these power laws measured from left to right, and the other from right to left.

Other 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, the evidence on the true distribution of scientific production and citation is still inconclusive (Albarrán et al. 2011), which might be a consequence of the use of only one- or two-parameter distributions.

In fact, all of the proposed 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 fact might result in more imprecise productivity measurements and unreliable comparisons of productivity between different fields of knowledge. 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 Mauleón 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 top researcher in each field.

For the purpose of holding the parameter flexibility of Gram–Charlier distributions but restricting the domain to positive values, we propose logarithmic transformations of a 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 et al. 2010; Sabharwal 2013), we show that, compared to the lognormal distribution, the log-SNP distribution provides a better fit when characterizing research productivity in top journals.

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]

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

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 Eq. (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:

$$H_0(x) = 1 \tag{3}$$

$$H_1(x) = x \tag{4}$$

$$H_2(x) = x^2 - 1 \tag{5}$$

$$H_3(x) = x^3 - 3x \tag{6}$$

$$H_4(x) = x^4 - 6x^2 + 3 \tag{7}$$

$$H_5(x) = x^5 - 10x^3 + 15x \tag{8}$$

$$H_6(x) = x^6 - 15x^4 + 45x^2 - 15 \tag{9}$$

$$H_7(x) = x^7 - 21x^5 + 105x^3 - 105x \tag{10}$$

$$H_8(x) = x^8 - 28x^6 + 210x^4 - 420x^2 + 105 \tag{11}$$

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

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

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

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]

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

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 $\boldsymbol{d} = (d_1, \ldots, d_n)' \in \mathbb{R}^n$ denotes the vector of the parameters.[3]

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

However, the SNP distribution defined in Eq. (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).

The advantages of SNP distributions when fitting frequency functions lies in their flexible parametric structure that permits to adjust location and scale with different parameters than those used for skewness, leptokurtosis and even higher order moments. Figure 1 illustrates the allowable shape of the SNP (depicted with 1000 simulated observations) compared with a normal distribution. For the sake of comparison, in both cases we consider the same location and scale parameters, $\mu = 0$ and $\sigma = 1$, but we introduce additional (even) parameters in the SNP. Particularly, Panels (a1) and (a2) incorporate $d_2 = 0.1$ and $d_4 = 0.1$ and Panels (b1) and (b2) $d_2 = 0.1$, $d_4 = 0.01$, $d_6 = 0.001$ and $d_8 = 0.005$. Note also that Panels (a1) and (b1) represent the whole domain but Panels (a2) and (b2) just a detail of the right distribution tails. It is clear from these pictures that the SNP not only captures leptokurtosis but also presents wavy and heavy tails that may adapt the probability pattern of any data generating process.

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 1" section). 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

[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 $d_0 = 1$.

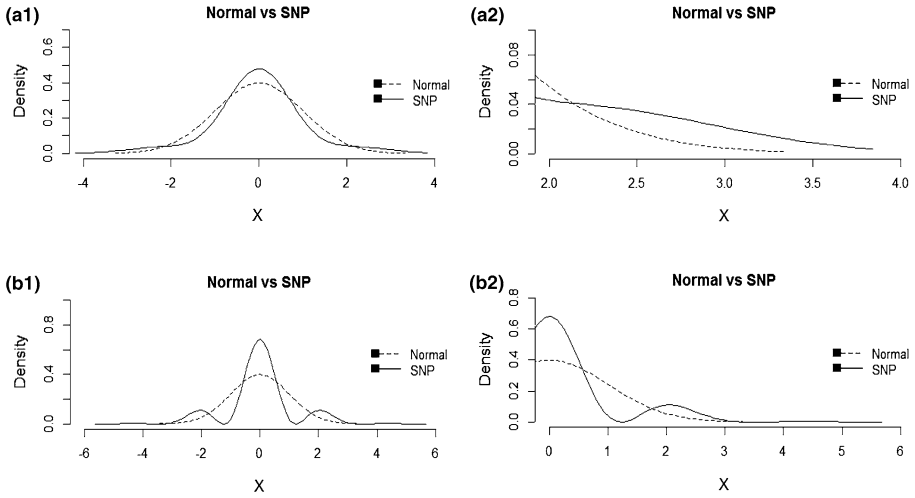

Fig. 1 Pdf of normal versus SNP distribution. Figures compare the shape of both Normal (*dashed line*) and SNP (*solid line*) distributions with location and scale parameters, $\mu = 0$ and $\sigma = 1$, and additional parameters for the latter. Particularly, *Panels* (*a1*) and (*a2*) incorporate parameters $d_2 = 0.1$ and $d_4 = 0.1$ and *Panels* (*b1*) and (*b2*) consider $d_2 = 0.1$, $d_4 = 0.1$, $d_6 = 0.001$ and $d_8 = 0.005$. *Panels* (*a1*) and (*b1*) represent the whole domain, whereas *Panels* (*a2*) and (*b2*) a detail of the right tails of the distributions. Data are simulated through 1000 replications

function of the normal distribution cdf, as shown in Eq. (15) (see the proof in "Appendix 2" section). This allows for a simple calculation of the probabilities and quantiles of the SNP distribution.

$$\begin{aligned} G_x(a) &= \int_{-\infty}^{a} g(x;\boldsymbol{d})dx \\ &= \int_{-\infty}^{a} \phi(x)dx - \phi(a)\sum_{s=1}^{n} d_s H_{s-1}(a) \end{aligned} \tag{15}$$

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 Eq. (14). The resulting distribution inherits all the good properties of the SNP distribution, including its flexibility in capturing the extreme values of the distribution, but the density is defined on $\mathbb{R}^+$, which is required to fit productivity data. 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 $\boldsymbol{d} = (d_1, \ldots, d_n)' \in \mathbb{R}^n$ if its pdf can be expressed as

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

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, as a matter of fact, 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 standard lognormal distribution. These parameters are directly related to the distribution moments[4] and constitute additional degrees of freedom for the estimation procedures. For example, if only d_s parameters are considered for s even skewness depends only on parameter σ, and the larger the expansion the heavier (and possibly wavier) the distribution tail is.

Figure 2 presents an illustration (1000 simulated replications) of the log-SNP allowable shape in comparison with that of the lognormal, both with the same location and scale parameters, i.e. $\mu = 0$ and $\sigma = 1$. Panels (a1) and (a2) depict a log-SNP with additional parameters $d_2 = 0.12$ and $d_4 = 0.11$ and Panels (b1) and (b2) incorporate parameters $d_2 = 0.28$, $d_4 = 0.44$, $d_6 = 0.07$ and $d_8 = 0.009$. In order to emphasize the behavior for the extreme (positive) values Panels (a2) and (b2) display a zoom on the right tails of the distribution. For this case, it is clear that the log-SNP allows more flexibility to capture thick (and wavy) tails. Even more important, biased estimations and misleading results may be obtained when using a single parameter distribution to fit distribution shape and heavy tails.

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). 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". As it is well-known, there are multiple subfields included within one JCR category, but they identified authors across all subfields so that authors publishing in more than one area would have all their publications included.

The authors used impact factors from JCR in 2007 to identify the top five journals within each field.[5] They selected field-specific journals to avoid having the search contaminated by authors from other sciences. Additionally, the authors used the "Publish or

[4] Log-SNP's moments can be directly derived as $E[z^t] = e^{\mu t+\frac{1}{2}t^2\sigma^2}\left[1+\sum_{s=1}^{n} d_s(\sigma t)^s\right]$ (see Ñíguez et al. 2013).

[5] It should be noted that the different size of journals in the JCR categories represents a shortcoming of the selection procedure. Nevertheless, it is not clear if other arbitrary selection method would yield to better results and, anyhow, this issue does not affect the advantages of the methodology proposed in this paper.

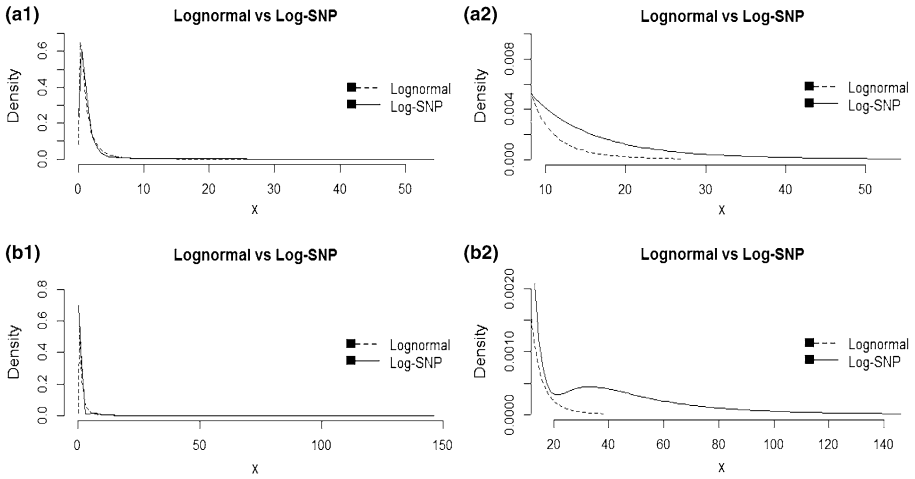

Fig. 2 Pdf of lognormal versus log-SNP distribution. Figures compare the shape of both Normal (*dashed line*) and SNP (*solid line*) distributions with location and scale parameters, $\mu = 0$ and $\sigma = 1$, and additional parameters for the latter. Particularly, *Panels* (*a1*) and (*a2*) incorporate parameters $d_2 = 0.12$ and $d_4 = 0.11$ and *Panels* (*b1*) and (*b2*) consider $d_2 = 0.28$, $d_4 = 0.44$, $d_6 = 0.07$ and $d_8 = 0.009$. *Panels* (*a1*) and (*b1*) represent the whole domain, whereas *Panels* (*a2*) and (*b2*) a detail of the right tails of the distributions. Data are simulated through 1000 replications

Perish" program (Harzing 2008), which relies on Google Scholar, to identify all authors who had published at least one article in one of these journals between January 2000 and June 2009.[6] Their procedure is not absent of the common problems in bibliometric analysis produced by the authors' names ambiguity, e.g. the treatment of synonyms, homonyms, misspellings, and processing errors. Further techniques might have been used to refine the database (Momeni and Mayr 2016; Van den Besselaar and Sandström 2016) although they are not substantial in our methodological framework. Furthermore, focusing on just five top journals and the combined use of Web of Science and Google Scholar databases may help to mitigate such problems (Yang and Meho 2006; Harzing and Van der Wal 2008; Harzing 2014; Harzing and Alakangas 2016). With this information, the productivity of the researchers is measured as the number of articles published by an author in each of the fields of knowledge during the observation period of 9.5 years.

A limitation of using this measure of productivity is the multiple authorship of manuscripts, since multiple authorship may bias the results in favor some authors and, comparatively, in areas where a high number of authors per paper is commonly accepted by the research community (Ruiz-Castillo and Costas 2014; Mingers and Leydesdorff 2015). The literature on scientific production, however, does not discriminate by multiple authorship. This procedure is known as "complete count", i.e. the fact that an article is equally valuable for all its authors regardless the number of authors and their marginal contribution to the manuscript (Nicholls 1989; Ruiz-Castillo and Costas 2014). On the other hand, another important issue when analysing research productivity is the relation between quantity and quality (Kaur et al. 2015). In our study, as well as in O'Boyle and Aguinis (2012), quality of top researchers is imposed by the fact that we only consider publications in the five main journals in every field of knowledge. Then comparisons are

[6] For details about the data treatment, see O'Boyle and Aguinis (2012), p. 86.

Table 1 Descriptive statistics for scientific production

Field of knowledge	*N*	*N* (ordinal position)	Mean (1)	Mean (2)	% authors with publications smaller than or equal to mean (1)	% authors with publications greater than mean (2)	Std	Skew	K	Max	Max (ordinal position)	Median impact factor	Median impact factor (ordinal position)	JCR edition
Agronomy	8923	7	1.42	2.84	77.23	8.13	1.16	6.36	72.68	26	13	2.36	12	Science
Anthropology	5755	9	1.87	3.55	66.06	10.50	1.95	4.49	34.52	30	8	2.31	14	Social sciences
Clinical psychology	10,418	6	1.89	4.88	64.80	6.08	2.38	10.80	267.22	93	2	4.68	3	Social sciences
Dentistry	12,345	3	2.26	5.92	79.18	6.98	2.98	6.54	74.62	66	4	3.37	6	Science
Dermatology	30,531	1	2.25	4.24	61.45	9.23	3.38	8.01	113.19	93	2	3.50	5	Science
Ecology	5730	10	1.71	3.18	67.61	8.53	1.68	7.88	148.90	50	6	4.82	2	Science
Economics	3048	13	1.62	3.20	71.75	6.63	1.67	7.14	82.10	27	11	3.69	4	Social sciences
Educational psychology	3032	14	1.70	3.03	65.60	7.26	1.55	5.41	52.04	27	11	2.35	13	Social sciences
Ethics	1073	18	1.65	3.21	70.74	6.43	1.78	6.82	71.24	26	13	1.31	16	Social sciences
Ethnic studies	2003	16	1.48	3.02	76.49	4.34	1.38	5.99	50.89	17	15	0.89	17	Social sciences
Finance	3019	15	2.14	5.40	77.97	6.69	2.52	4.69	33.93	28	9	2.99	8	Social sciences
Forestry	12,211	4	1.82	3.27	63.80	9.77	1.80	5.66	68.58	46	7	2.14	15	Science
Genetics	16,574	2	1.71	3.12	83.49	3.56	2.18	26.42	1240.47	120	1	18.30	1	Science
History	6708	8	1.54	2.56	65.28	12.00	0.97	3.33	25.11	14	17	0.85	18	Social sciences
Law	1350	17	1.55	2.93	71.48	11.48	1.24	3.88	24.07	13	18	3.09	7	Social sciences

Table 1 continued

Field of knowledge	*N*	*N* (ordinal position)	Mean (1)	Mean (2)	% authors with publications smaller than or equal to mean (1)	% authors with publications greater than mean (2)	Std	Skew	K	Max	Max (ordinal position)	Median impact factor	Median impact factor (ordinal position)	JCR edition
Linguistics	3600	12	1.73	3.32	68.69	8.50	1.78	5.98	59.06	28	9	2.37	11	Social sciences
Mathematics	3972	11	1.45	2.61	72.18	8.06	1.02	4.86	41.42	15	16	2.56	10	Science
Statistics	10,679	5	2.08	5.52	81.04	5.48	2.52	6.22	67.39	54	5	2.97	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, *Mean (1)* mean of publications of researchers in the entire sample, *Mean (2)* mean of publications the researchers with a number of articles above Mean (1), *Std* standard deviation, *Skew* empirical skewness, *K* excess kurtosis, *Max* maximum score, *MIF* median impact factor (JCR's five top journals in 2007)

only done in terms of quantity of manuscripts for a given probability in the "true" distribution (i.e. a quantile), which proxies a level of difficulty taking into account the different practices between the areas.

Table 1 shows the descriptive statistics for the publications of the top researchers included in our sample. In this table we also record the Median Impact Factor (MIF) of the top five journals in each of the selected fields of knowledge, based on the JCR of the year 2007[7] for each of the analyzed categories classified in sciences and social sciences. We provide this citation index to obtain a broader view of each of the selected fields and, particularly, its correlation with scientific production.

It can be observed that throughout the 18 fields of knowledge analyzed, the minimum number of researchers is 1073 for the field of Ethics and the maximum is 30,531 for Dermatology. For each field, we compute two Mean values on scientific production: Mean (1) is the mean of publications for each field and for the entire sample of researchers; Mean (2) is the mean of publications for each field but only for the researchers with a number of articles above Mean (1). Mean (1) varies from 1.42 to 2.26 and Mean (2) from 2.56 to 5.92. Furthermore, we compute the percentage of authors with less/more publications than (or equal) to Mean (1)/Mean (2). We find that, on average, 71.38 % of all researchers have productivity below Mean (1), whilst researchers with productivity above Mean (2) represent the 7.76 %. These results support Ruiz-Castillo and Costas' (2014) findings about the skewness of field productivity distributions, since a large proportion of researchers have below mean productivity and only a small percentage of them account for most of the publications.

Regarding the other statistics in Table 1, the standard deviation of publications has a range of 0.97–3.38 publications and skewness and excess kurtosis reveals that the productivity distributions 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, it is clear that the MIF is positively correlated 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. In general, fields that belong to the Sciences JCR category have a larger number of researchers and, then, a larger MIF.

All in all, the results confirm the existence of wide differences in scientific production in terms of number of articles per author between the different fields of knowledge, which is consistent with other studies, e.g. Abramo and D'Angelo (2014) or Mingers and Leydesdorff 2015). Next we propose a new methodology based on the log-SNP to study how these differences affect the productivity distribution, especially when measuring the productivity of the top researchers.

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

[7] 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.

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[8] for a log-SNP(μ, σ^2, $\boldsymbol{d}$) distributed observation truncated to the eighth moment is given by:

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

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.

The results of the estimations 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. It is noteworthy that the parameter σ, which also capture skewness of the lognormal and the log-SNP provided that odd parameters are not included, remains very stable for all productivity distributions. This evidence is consistent with Ruiz-Castillo and Costas (2014) who found that "in spite of wide differences in production and citation practices across fields, the shape of field productivity distributions is very similar across fields". 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.[9]

[8] The code for the implementation of the maximum likelihood estimation algorithm in R package is available upon request.

[9] Note that we did not include the d_s parameters for s odd, after having tested that they were not significantly different from zero. This result reinforces the fact that the parameter σ captures all relevant features about the skewness. It must be highlighted that the latter does not contradict the fact that the d_s parameters for s even are highly significant, which means that productivity distributions have very thick tails and thus

Table 2 Estimates for the productivity distribution under lognormal and log-SNP

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
	(<0.0001)	(<0.0001)			(0.000)	(<0.0001)	(0.000)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
Clinical psychology	0.3791	0.5994	−5501.26	11,006.52	0.1689	0.5556	0.1535	0.2611	0.0444	0.0055	−4236.31	8484.63	2529.90
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
Dentistry	0.4934	0.6763	−6598.224	13,200.45	0.2959	0.6913	0.0194	0.1481	0.0157	0.0027	−5740.93	11,493.86	1714.58
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(0.1765)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
Dermatology	0.4553	0.6914	−18,154.32	36,312.64	0.8375	0.4294	1.1923	0.3812	0.1092	0.0179	−7262.16	14,536.32	21,784.32
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(0.0023)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(0.0819)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(0.0900)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(0.0028)	(<0.0001)	(0.4423)	(<0.0001)	(0.0017)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(0.0038)	(<0.0001)	(0.0011)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
Forestry	0.3785	0.5755	−5958.31	11,920.63	0.1797	0.5490	0.1149	0.1942	0.0232	0.0037	−4879.51	9771.02	2157.61
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
Genetics	0.3338	0.5350	−7617.37	15,238.74	0.1720	0.5379	0.0399	0.1748	0.0224	0.0037	−6015.27	12,042.54	3204.20
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)

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	
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.8251)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(0.6560)	(<0.0001)	(0.0272)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)
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
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(0.0210)	(<0.0001)			(<0.0001)
Statistics	0.4510	0.6390	−5553.69	11,111.38	0.2429	0.6251	0.0779	0.1858	0.0253	0.0036	−4758.50	1590.38	1590.38
	(<0.0001)	(<0.0001)			(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)	(<0.0001)			(<0.0001)

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 for testing differences between log-SNP and lognormal. 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. *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

logL log-likelihood, *AIC* Akaike Information Criterion, *LR* likelihood ratio test

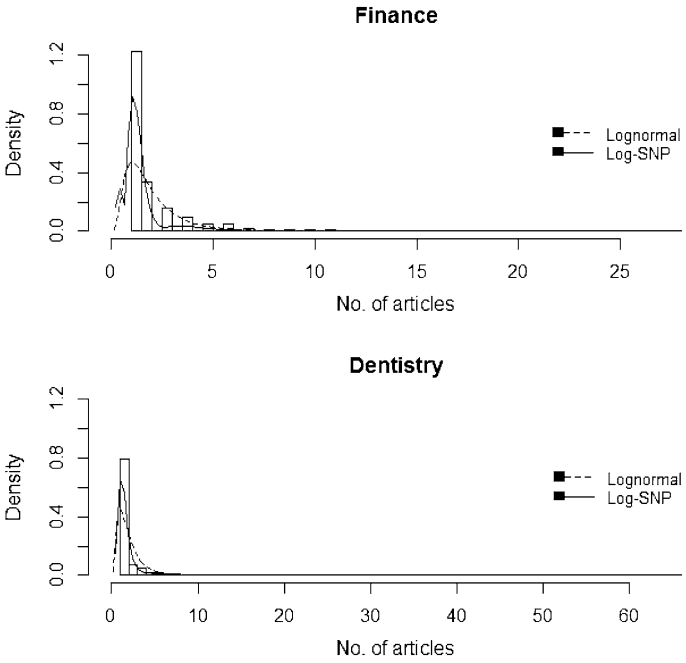

Fig. 3 Pdf of research productivity in finance and dentistry. The figure shows the distribution of the empirical frequencies (histogram) of the productivity of the researchers who published in the five top journals (in JCR-2007 terms) in finance and dentistry during the period 2000–2009. The estimated pdfs under the lognormal and log-SNP specifications are depicted in *dashed line* and *solid line*, respectively

An example of the fit quality obtained for two (randomly selected) fields, Finance and Dentistry, is captured in Fig. 3. 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. Figure 4 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.

Figure 5 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.

Figure 4 illustrates 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 quantifies these effects for the different fields of knowledge by computing the empirical and estimated quantiles under the lognormal and log-SNP for confidence

Footnote 9 continued
require different parameters to provide accurate measures of the "probability of being a very top researcher" in every field.

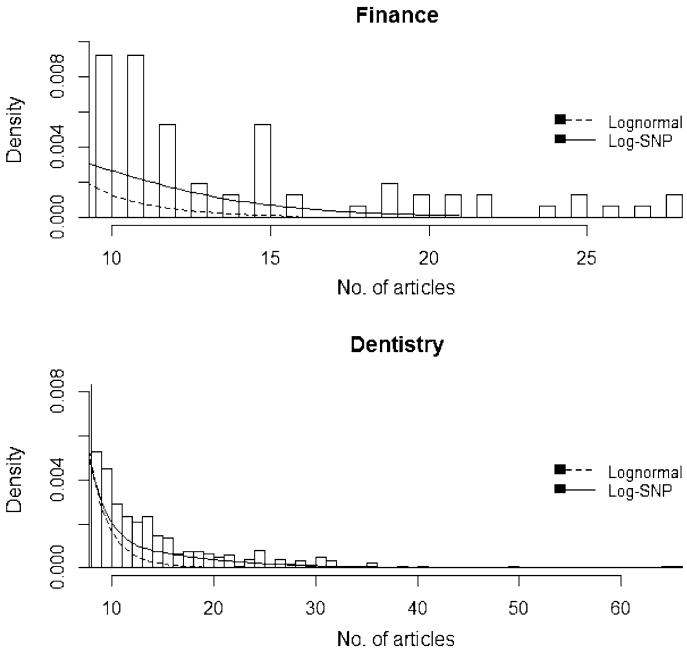

Fig. 4 Pdf of research productivity in finance and dentistry (right tail). The figure shows the right tail of the distribution of empirical frequencies (histogram) of productivity of the researchers who published in the five top journals (JCR-2007 ranking) in finance and dentistry during the year 2000–2009. Fitted lognormal and log-SNP pdfs are depicted in *dashed line* and *solid line*, respectively

levels of 5, 1, 0.1 and 0.05 %.[10] Note that, once the productivity distributions are properly estimated, the definition of a top researcher in every field requires the computation of the corresponding quantile for a given probability. These quantiles represent bounds of performance in terms of number of articles (regardless the number of authors), provided that quality is guaranteed by considering only publications on the top 5 reviews in every field. Furthermore, these quantiles are fairly comparable among different areas.

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 publications) as compared to log-SNP (12 publications). These results are consistent with the research by Kumar et al. (1998), Perc (2010) and Eom and Fortunato (2011), who found that the use of the lognormal distribution for modeling bibliometric indicators underestimates the heavy tails of the distributions.

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

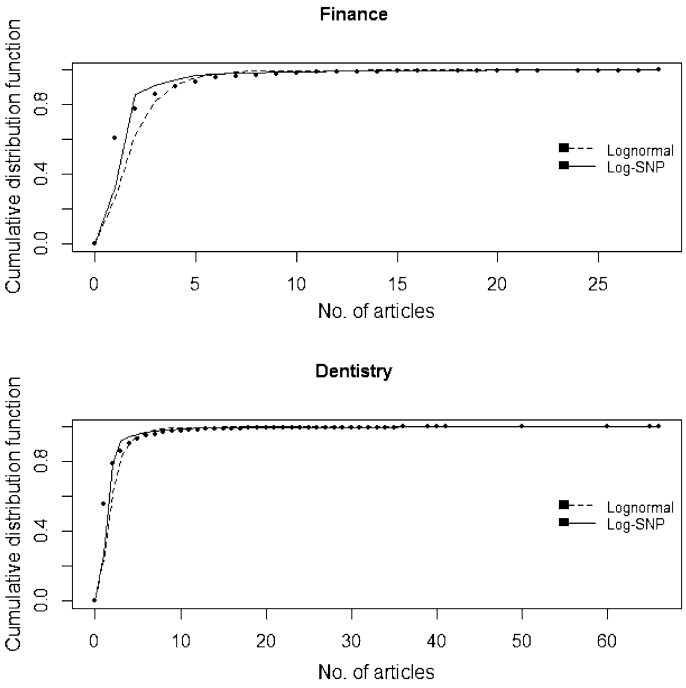

Fig. 5 Cdf of research productivity in finance and dentistry. The figure shows the empirical cumulative distribution function of the productivity of the researchers who published in the five top journals (JCR-2007 ranking) in finance and dentistry during the period 2000–2009. Fitted lognormal and log-SNP cdfs are depicted in *dashed line* and *solid line*, respectively

Further results

This article proposes a new methodology to compute research productivity for top researchers through the quantiles of a new and general distribution called log-SNP. Our main application compares these measures with those of the lognormal with a sample of scientific production in 18 (arbitrarily chosen) fields, finding the outperformance of the log-SNP. Nevertheless, the focus of the paper is done on the technique more than on the particular results. In order to justify that our result is general we replicated the study with the productivity data provided in Borokhovich et al. (1995), which refer to academic institutions (in the field of finance) instead of individual researchers. Particularly, the data accounts for the number of articles published from 1989 through 1993 in a set of 16 finance journals by authors affiliated to different institutions at the time of publication. The journals in finance (excluding real estate and insurance) were selected from those listed in Heck's Finance Literature Index for 1993. Only articles and notes were included in sample. The number of publications attributed to each academic institution was adjusted for the number of authors. For example, for publications with two authors affiliated to different institutions every institution received a credit for 0.5 article. Any proportion of an article that was not attributable to an author affiliated with an academic institution located in the United States or Canada was deleted from the study. A total of 330 institutions were included in this sample.

Table 3 Number of articles empirically observed versus those theoretically expected under 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	8923	3	7	13	15	3	4	5	6	3	4	10	12
Anthropology	5755	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	5730	4	8	17	20	4	5	8	9	4	7	14	16
Economics	3048	4	8	25	26	4	5	7	8	3	7	13	14
Educational psychology	3032	4	8	18	18	4	5	8	9	4	7	14	16
Ethics	1073	4	9	24	25	4	5	7	8	3	8	14	16
Ethnic studies	2003	3	8	16	16	3	4	6	6	3	5	12	14
Finance	3019	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	6708	3	5	8	12	3	4	6	7	3	5	8	11
Law	1350	4	7	13	13	3	5	7	7	3	6	11	12
Linguistics	3600	5	9	22	23	4	6	8	9	4	7	14	16
Mathematics	3972	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 probabilities for which distribution quantiles are computed. 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 4 Institutional research productivity under lognormal and log-SNP

Institutional research productivity	Lognormal				Log-SNP								LR
	μ	σ	logL	AIC	μ	σ	d_2	d_4	d_6	d_8	logL	AIC	
Panel A: estimates of the productivity distribution under lognormal and log-SNP													
Academic institutions	1.102 (<0.0001)	1.4319 (<0.0001)	−950.38	1904.76	1.1574 (<0.0001)	0.61 (<0.0001)	2.2593 (<0.0001)	1.1595 (<0.0001)	0.1979 (<0.0001)	0.011 (0.0018)	−920.19	1852.38	2157.61 (<0.0001)

Institutional research productivity	N	Observed No. of papers				Expected number of papers							
						Lognormal				Log-SNP			
		5 %	1 %	0.10 %	0.05 %	5 %	1 %	0.10 %	0.05 %	5 %	1 %	0.10 %	0.05 %
Panel B: number of articles empirically observed versus those theoretically expected under the lognormal and log-SNP													
Academic institutions	330	30	50	83	86	32	85	252	335	29	45	70	78

The table reports the results for research productivity of academic institutions that published on a set of 16 finance journals between the years 1989 and 1993. Panel A reports the ML estimates for the parameters of both the lognormal and log-SNP distributions and the likelihood ratio for testing the differences between them. 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. P values are shown in parentheses. Panel B compares the number of articles empirically observed with those theoretically expected under the lognormal and log-SNP distributions. The values 5, 1, 0.1 and 0.05 % are probabilities for which distribution quantiles are computed

logL log-likelihood, *AIC* Akaike Information Criterion, *LR* likelihood ratio statistic, *N* number of academic institutions

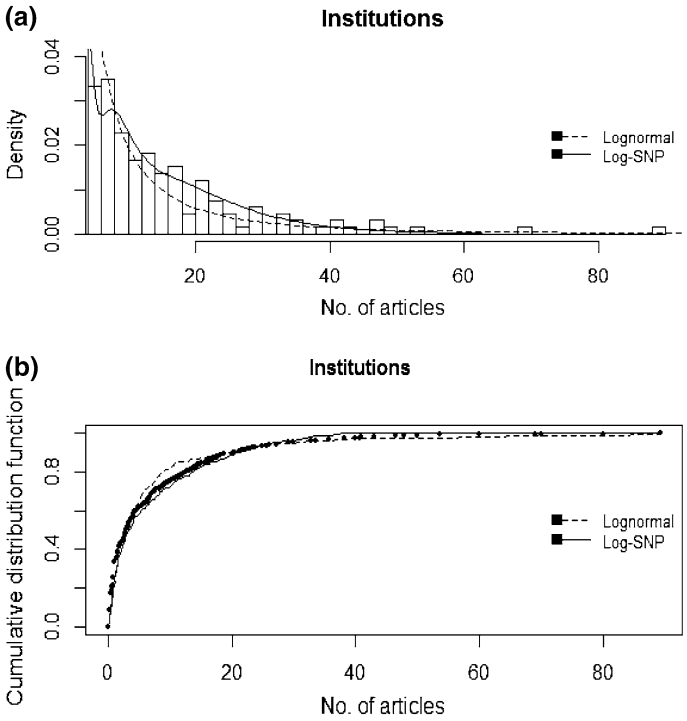

Fig. 6 Pdf and cdf of institutional research productivity. The figure shows: **a** the right tail of the distribution of empirical frequencies (histogram) of research productivity for academic institutions that have published a set of 16 finance journals between the years 1989 and 1993. The fitted lognormal and log-SNP pdfs are depicted in *dashed line* and *solid line*, respectively. **b** The figure shows the empirical cumulative distribution function of the same sample. The fitted lognormal and log-SNP cdfs are depicted in *dashed line* and *solid line*, respectively

Table 4 reports the results of this new estimation. Panel A displays the ML estimates for the research productivity on top finance journals for academic institutions. The results are consistent with those previously obtained for researchers in different fields of knowledge, i.e. the log-SNP outperforms the lognormal according to the LR test and thus the log-SNP parameters are highly significant. Panel B compares the number of articles empirically observed with those theoretically expected under both specifications revealing the outperformance of log-SNP. In this case it seems that the lognormal overestimates the distribution tails, particularly for low confidence levels. This result corroborates the evidence that the use of rigid distributions involve misleading results because are unable to fit different characteristics of the distribution (particularly extreme values) with a single (or two) parameter(s).

Figure 6 illustrates the assessment above showing the best fit of the lof-SNP in terms of the right tails of the pdf (Fig. 6a) and the cdf (Fig. 6b) compared with the empirical distributions.

Discussion and conclusions

Bibliometric analysis has been shown to be a valuable method for evaluating scientific production and has experienced 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 show that the level of productivity, as measured by the number of publications per author, depends on the field of knowledge being studied, which is consistent with previous evidence. 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 computation of the quantiles of a flexible 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, for both researchers and institutions productivity, that the log-SNP provides a better fit of research productivity distribution than the lognormal and quantifies the differences in the measures of the top researchers' productivity attached to the distributional hypothesis. We argue that the log-SNP is an accurate data generating process for the top researchers' productivity, and thus this process is more reliable than that of the lognormal (that is nested in our model) since the log-SNP is more flexible when data are highly skewed and there are possible jumps in the tail due to extreme observations.

Therefore we provide an interesting methodology to measure scientific productivity that can be used when the performance of authors, institutions or fields have to be compared or aggregated so as to implement policies based on academic performance.

Acknowledgments We thank Herman Aguinis and Ernest O'Boyle for allowing us to use their database on academic productivity compiled in O'Boyle and Aguinis (2012). We also thank two anonymous referees for their constructive and valuable suggestions. Financial support from the Spanish Ministry of Economics and Competitiveness, through the project ECO2013-44483-P, FAPA-Uniandes, through the project PR.3.2016.2807, and Universidad EAFIT are also gratefully acknowledged.

Appendix 1

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).

$$d_1 = \mu_1 \tag{18}$$

$$d_2 = \frac{1}{2}(\mu_2 - 1) \tag{19}$$

$$d_3 = \frac{1}{6}(\mu_3 - 3\mu_1) \tag{20}$$

$$d_4 = \frac{1}{24}(\mu_4 - 6\mu_2 + 3) \tag{21}$$

$$d_5 = \frac{1}{120}(\mu_5 - 10\mu_3 + 15\mu_1) \tag{22}$$

$$d_6 = \frac{1}{720}(\mu_6 - 15\mu_4 + 45\mu_2 - 15) \tag{23}$$

$$d_7 = \frac{1}{5040}(\mu_7 - 21\mu_5 + 105\mu_3 - 105\mu_1) \tag{24}$$

$$d_8 = \frac{1}{40320}(\mu_8 - 28\mu_6 + 210\mu_4 - 420\mu_2 + 105) \tag{25}$$

Appendix 2

This appendix derives the cdf of the SNP distribution.

$$\begin{aligned} G_x(a) &= \int_{-\infty}^{a} g(x;\boldsymbol{d})dx = \int_{-\infty}^{a} \phi(x)dx + \sum_{s=1}^{n} d_s \int_{-\infty}^{a} H_s(x)\phi(x)dx \\ &= \int_{-\infty}^{a} \phi(x)dx - \sum_{s=1}^{n} d_s H_{s-1}(x)\phi(x)\Big|_{-\infty}^{a} \\ &= \int_{-\infty}^{a} \phi(x)dx - \phi(a)\sum_{s=1}^{n} d_s H_{s-1}(a) \end{aligned}$$

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

$$\begin{aligned} \int H_s(x)\phi(x)dx &= \int (-1)^s \frac{d^s\phi(x)}{dx^s}dx_t = (-1)^s \frac{d^{s-1}\phi(x)}{dx^{s-1}} \\ &= (-1)^s(-1)^{s-1}H_{s-1}(x)\phi(x) = -H_{s-1}(x)\phi(x) \end{aligned}$$

□

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,*. doi:10.1111/peps.12095.

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. In J. Heckman & E. Leamer (Eds.), *Handbook of econometrics, Ch. 76, Part B* (Vol. 6, pp. 5549–5632). Amsterdam: 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), 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. In *Sixth scandinavian congress of mathematicians* (pp. 399–425). Copenhagen.

Crespo, J. A., Ortuño-Ortín, I., & Ruiz-Castillo, J. (2012). The citation merit of scientific publications. *PLoS ONE, 7*(11), e49156.

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.

Duch, J., Zeng, X. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., et al. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS ONE, 7*(12), e51332.

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: 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.

Harzing, A. (2008). Publish or Perish: A citation analysis software program. http://www.harzing.com/resources.htm.

Harzing, A. W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics, 98*(1), 565–575.
Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804.
Harzing, A. W., & Van der Wal, R. (2008). Google Scholar as a new source for citation analysis? *Ethics in Science and Environmental Politics, 8*(1), 61–73.
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.
Kaur, J., Ferrara, E., Menczer, F., Flammini, A., & Radicchi, F. (2015). Quality versus quantity in scientific impact. *Journal of Informetrics, 9*(4), 800–808.
Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics, 7*(4), 924–932.
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 and 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.
Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE, 4*(3), e4791.
Mauleón, I., & Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth–Sargan density to the Student's t. *European Journal of Finance, 6*(2), 225–239.
Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research, 246*(1), 1–19.
Momeni, F., & Mayr, P. (2016). Evaluating co-authorship networks in author name disambiguation for common names. arXiv:1606.03857.
Newman, M. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323–351.
Nicholls, P. T. (1986). Empirical validation of Lotka's law. *Information Processing and Management, 22*(5), 417–419.
Nicholls, P. T. (1989). Bibliometric modelling processes and the empirical validity of Lotka's law. *Journal of the American Society for Information Science, 40*(6), 379–385.
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.
Ñíguez, T.-M., Paya, I., Peel, D., & Perote, J. (2013). Higher-order moments in the theory of diversification and portfolio composition. Economics Working Paper Series 2013/003. Lancaster University.
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.

Van den Besselaar, P., & Sandström, U. (2016). What is the required level of data cleaning? A research evaluation case. *Journal of Scientometric, 5*(1), 07–12.

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.

Yang, K., & Meho, L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology, 43*(1), 1–15.