Journal of Informetrics 8 (2014) 917–934

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The skewness of scientific productivity

Javier Ruiz-Castillo[a,*], Rodrigo Costas[b]

[a] Departamento de Economía, Universidad Carlos III of Madrid, Spain
[b] Centre for Science and Technology Studies, Leiden University, Netherlands

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 15 August 2014
Accepted 22 September 2014
Available online 14 October 2014

Keywords:
Individual scientist's productivity distributions
Skewness of science
Disambiguation algorithm
Co-authorship

ABSTRACT

This paper exploits a unique 2003–2011 large dataset, indexed by Thomson Reuters, consisting of 17.2 million disambiguated authors classified into 30 broad scientific fields, as well as the 48.2 million articles resulting from a multiplying strategy in which any article co-authored by two or more persons is wholly assigned as many times as necessary to each of them. The dataset is characterized by a large proportion of authors who have their *oeuvre* in several fields. We measure individual productivity in two ways that are uncorrelated: as the number of articles per person and as the mean citation per article per person in the 2003–2011 period. We analyze the shape of the two types of individual productivity distributions in each field using size- and scale-independent indicators. To assess the skewness of productivity distributions we use a robust index of skewness, as well as the Characteristic Scores and Scales approach. For productivity inequality, we use the coefficient of variation. In each field, we study two samples: the entire population, and what we call "successful authors", namely, the subset of scientists whose productivity is above their field average. The main result is that, in spite of wide differences in production and citation practices across fields, the shape of field productivity distributions is very similar across fields. The parallelism of the results for the population as a whole and for the subset of successful authors, when productivity is measured as mean citation per article per person, reveals the fractal nature of the skewness of scientific productivity in this case. These results are essentially maintained when any article co-authored by two or more persons is fractionally assigned to each of them.

1. Introduction

In this paper, we study the size and the mean of individual citation distributions in a given period of time for all authors in a number of scientific fields. Naturally, the size of individual citation distributions, that is, the number of publications per author, is a standard measure of individual productivity. The productivity of individual scientists has been studied extensively since Lotka's (1926) pioneer contribution, in which the probability of an author publishing a certain number of articles in Chemistry was estimated to be an inverse square function of the number of publications (Alvarado, 2012, counts 651 publications concerning the so-called *Lotka's law* from that date until 2010). However, most of these contributions

* Corresponding author. Tel.: +34 91 624 95 88.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

http://dx.doi.org/10.1016/j.joi.2014.09.006
1751-1577/

analyze a relatively small number of scientists and, to the best of our knowledge, do not systematically study productivity distributions using comparable and large datasets for several scientific disciplines.[1]

On the other hand, the mean citation per article per author is a standard (size-independent) measure of the citation impact achieved by any researcher in her field of study. Nevertheless, to simplify the exposition, we will refer to this indicator of citation impact as a second definition of individual productivity. At any rate, we do not know of systematic studies concerning the distribution of this variable within and between representative samples for a variety of scientific disciplines.

As in any other scientific discipline, in Scientometrics we should clearly establish the stylized facts that characterize basic constructs in all fields. Consequently, this paper studies the productivity of individual scientists – in the two senses indicated above – in 30 broad fields using a large dataset, indexed by Thomson Reuters, consisting of 7.7 million distinct articles published in the period 2003–2011 in academic journals. Applying a variable citation window from the publication year until 2012, these articles receive 78.9 million citations.

Regardless of how we measure individual productivity, a study of this type must confront the following four methodological problems: (i) the classification of articles into scientific fields; (ii) the identification of the author(s) of each article, (iii) the allocation of authors to fields, and (iv) the attribution of individual responsibility in cases of multiple authorship. After these problems are solved (see Section 2), we end up with a dataset consisting of 17.2 million authors and 48.2 million articles.

Of course, we know a priori that the between-field variability with respect to several basic characteristics is typically very large. Firstly, the size of productivity distributions, namely, the number of authors per field, is bound to be very different across fields. Secondly, because of well-known differences in production and citation practices, the average number of articles per author, as well as the average mean citation per article per author are also expected to be very different across fields.

Therefore, what we should study is the shape of field productivity distributions abstracting from size and scale differences across fields. To simplify the presentation, we focus on the skewness of productivity distributions. Naturally, the extensive literature on Lotka's law leads us to expect that productivity distributions according to the first definition are highly skewed in all fields, in the sense that a majority of individuals publish very little, while a large proportion of the total number of publications must be attributed to a small number of authors. Finally, if only by analogy with the skewness of science in so many dimensions (see De Solla Price, 1963; Lotka, 1926; Seglen, 1992, to cite only a few classics), we expect that all field productivity distributions according to the second definition are also highly skewed.

In this scenario, the main aim of this paper is to investigate the between-field variation of the skewness of productivity distributions that is expected to be prevalent in each field. For the reasons already explained, we need size- and scale-independent indicators of skewness. We follow two complementary approaches. In the first place, we study the broad features of this phenomenon by simply partitioning productivity distributions into three classes of individuals with low, fair, and very high productivity. For this purpose, we adopt the Characteristic Scores and Scale (CSS hereafter) approach first introduced in Scientometrics by Schubert, Glänzel, and Braun (1987). In the second place, we are interested in summarizing the skewness of productivity distributions with a single scalar. Among the size- and scale-independent skewness measures that are also robust to extreme observations, in this paper we use the one suggested by Groeneveld and Meeden (1984) that has been used before in Albarrán, Perianes-Rodriguez, and Ruiz-Castillo (2014), and Perianes-Rodriguez and Ruiz-Castillo (2014). Finally, for reasons that will be apparent in the sequel, we analyze the shape of productivity distributions in each field for two samples: the entire population, and what we call *successful authors*, namely, the subset of scientists whose productivity is above their field average.

In the Working Paper version of this article, Ruiz-Castillo and Costas (2014), hereafter referred to as RCC, we study a second characteristic of the shape of field productivity distributions: the productivity inequality exhibited both by the entire population and successful authors for the two productivity definitions. A summary of results is presented below in a section on extensions. Also, to facilitate the reading of the text, some statistical information and, in many cases, the numerical results for a variety of field characteristics, are relegated to the Supplementary Material Section (SMS hereafter) of the paper. At the end of each section we include a footnote specifying which aspects of the questions discussed in the text can be found in the SMS.

The rest of this paper is organized into five sections. Section 2 describes the data and discusses our approaches to cope with the four methodological issues. Sections 3 and 4 present the results concerning the characteristics of productivity distributions when individual productivity is measured as the number of publications and as the mean citations per article, respectively, whereas Section 5 summarizes the main results concerning productivity inequality and other issues explored in detail in RCC. Finally, Section 6 summarizes the paper and suggests possible extensions.

[1] Kyvic (1989) compares the productivity between three very broad scientific disciplines – the Medical, the Natural, and the Social Sciences – and the Humanities, using a relatively small dataset. A key exemption is the important contribution by Ioannidis et al. (2014), which studies 15.1 million authors that have published at least one indexed item in the entire Scopus database in the period 1996–2011. See below for a comparison of our methods and results with those of Ioannidis et al. (2014).