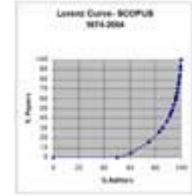


[Gallery: G Prathap](#)

Who's Afraid of Research Assessment? The Skewness of Unfairness and Scientific Productivity



Introduction

Although the field of Scientometrics now offers well-tested procedures for some measure of quantitative assessment of research performance, these are largely left unused in our country when we attempt exercises to assess the performance of individuals or institutions. This is baffling in a country that is so comfortable with its obsession with cricket and cricket statistics. The present analysis is based on data from the SCOPUS database, and this approach has the potential to offer interesting sociological insights into the scientific productivity of individuals, research institutes and of research agencies.

The Skewness of Unfairness and Scientific Productivity

"Life is unfair," said John F Kennedy famously. Human ability ranges very widely and is distributed in a highly non-linear fashion within a population. It has been known that patterns exist, which go beyond conventional Gaussian or normal distribution. Rank-order statistics based on power-law distributions are used to describe this. Scientometrics suggests that an area of intellectual activity that is most easily amenable to quantification is the production of research output as measured by publications in the open literature. The ecology of this enterprise, where a large number of scientists work, about half of them publish, but only a very few account for the highly cited work, is complex. Slowly, soft laws have emerged, where the role of power law distributions is easily seen. Norbert Wiener ("I am a mathematician", Science, 1964) is said to have argued that 95% of the original works is made by less than 5% of all scientists. We examine some of the laws and use some concrete evidence from a study of the performance of a premier institute X (name withheld) with data from SCOPUS (www.scopus.com), an Elsevier product which is all set to become the single largest scientometric database with more than 27 million abstracts and citations covering 14,500 journals from 4,000 publishers, and dating back to 1966. Free access to the beta version of this database made this study possible.

Methodology

The SCOPUS database was interrogated for all records of papers published by scientists from Institute X from 1994 to 2004. Here it is important to make the following distinctions. All databases will naturally indicate those individuals from the institute who have papers which are published in journals covered by the database. This set will vary according to how exclusive or inclusive the database is. SCOPUS is probably the largest scientometric database today and it is particularly generous to Indian journals, covering about 150 or so. We call all individuals (scientists) who appear in this list (i.e. who have at least one paper in the database for this period) authors. Not all of the papers will have received citations (right now the database offers only the citations obtained in the last ten years) and therefore many authors will remain uncited. There will also be a large number of the individuals (scientists) who will not have published during the period, or even if they have published, will not be fortunate to have their papers registered in the database because the journals in which they have published do not belong to the set of 14,500 journals covered by SCOPUS. Thus authors are a subset of scientists. Scientists without papers are called the zero item cases and for obvious reasons, their identities cannot be gleaned from the database. Sociologically, it is equally important to know who of the several hundreds who participate in the intellectual process associated with scientific discovery at Institute X never get to publish! This of course, the scientometric data can never capture. For Institute X, in

addition to all those who have achieved author status during the period covered by this study, an attempt is made to include all the senior scientists (non-entry level, as it is assumed that entry level scientists will take some time to get themselves established) on the latest roster as the zero item cases. For this study to be more complete, it is important that at a future date, the list of zero-paper scientists be obtained from the total population of all scientists who worked at Institute X during the period 1974-2004. For now, one must be satisfied with the present restriction but it is felt that this will still give an indicative idea of the knowledge gathering and dissemination process at Institute X.

Lotka's Law

[Figure 1](#) shows the histogram from a special arrangement of the scientometric data extracted from SCOPUS recently (early October 2004) and displayed in [Table 1](#). It is important to note that the database is dynamic and changes every week! All papers from Institute X during 1974-2004 and all citations received since 1995 are registered. During this period, the number of unique authors, unique papers and whole count papers were:

Unique authors: 184

Unique scientists: 366

Whole count papers: 882

We use the whole count method for computation of the performance instead of fractionating a paper into partial counts. Thus a paper with three authors will be counted as 3 whole counts or contributions. Accordingly, each author, whose name has appeared in a paper is given credit for the paper regardless of the number of co-authors. The data in [Table 1](#) shows (complete data available from the author upon request) that at the low end, 182 scientists wrote no papers, 40 unique authors wrote only 1 paper each in the 30 years from 1974-2004, and another 52 authors contributed to 2 papers each. This distribution is captured in [Figure 1](#). The distribution is bimodal, with a huge peak at 0 papers and another peak at 2 papers. This suggests that in Institute X, there is a very large component of the task force doing only applied work (time-bound, mission-oriented, etc.) and does not publish, and that among those who publish, the peak is at 2 papers. That is, nearly 50% of the actual population of active scientists (a distinction which was made clear earlier) at Institute X produced no papers during this period. Of the 50% that has published, only a small fraction actually published 1 paper, and a larger number published 2 papers. When such distinctions are made, it becomes very difficult to assess this evidence in the light of what Lotka observed in 1926, where the population studied was all authors with names beginning with A or B in Chemical Abstracts covering the years 1907-1916. About 60% of the authors produced only one paper during the period in Lotka's study. From this, he formulated his famous law of scientific productivity, whereby, the number of authors making n contributions is about $1/n^2$ of those making 1.

However, the law is not accurate at the extreme tail of high end scientific productivity. It is perfectly possible that we may find an author with a 100 papers and another with 50 or more papers. This is where Zipf's law comes to the rescue.

Zipf's Law

At the high end of the distribution, we see that the most productive authors produce much more than the average person. Indeed, we see from [Table 1](#) that there are many individuals who have 10 (4 authors), 11, 14, 15, 16 (2 authors), 18 (2 authors), 19 (2 authors), 20, 21 (2 authors), 22, 24 (2 authors), 32, and 56 papers respectively. Zipf's law was the first to record this and Zipf's is the law of rank frequency, which postulates that rank r occurs with a frequency which is inversely related to r . Note that a very large number of variables are hidden in the system, but the rank to frequency relationship is captured in a simple way. Thus, if an author of the first rank has a 100

papers, an author of the second rank may have 50 ($=100/2$) or 25 ($=100/2^2$) papers, depending on the power of the inverse relationship. In this simple relationship that Zipf postulated, some kind of "principle of least effort" was operating.

Lorenz Curve and Pareto's Law

The combined effect of the Lotka law at one end of the distribution and Zipf's law at the other end of the distribution is to confirm the general intuition described so well by Narin and Breitzman³ that "eminence is highly concentrated in a small fraction in the population," and that "scientific creativity and productivity, are very highly concentrated in a population, and in the minds and abilities of a relatively small number of highly talented individuals". Before we demonstrate that this is true for the present data, let us briefly review another variation that expresses this kind of inequality or disparity.

The Italian engineer-turned-economist and political sociologist, Vilfredo Pareto realized that wealth is not evenly distributed⁴. Some of the people have most of the money. In fact, a fairly consistent minority, about 20% of people, controlled about 80% of a society's wealth. A closer examination would indicate that of the top 20% which owns 80% of the wealth, the 80-20 formula still applies reasonably consistently, so that the following pyramid can be set up as shown in [Table 2](#). Thus, less than 1% or so of the population may account for 50% or so of the wealth. In Haiti, which has been very much in the news recently for all the wrong reasons, this is precisely the situation. In advanced countries like Australia, Japan and the United States, the top 1% accounts for 40% or more of the wealth.

That the same distribution is true for many other areas has been frequently noticed and is now termed the Pareto principle. Recently, press reports indicated that the three richest families in the world have as much wealth as the total populations of the poorest 46 countries of the world!

We can display this inequality or disparity very dramatically using a Lorenz curve, a device used by economists to represent inequality of income and wealth distribution in a population. In [Table 1](#), column 3 and column 5 add up the whole count contributions and the authors cumulatively. Columns 4 and 6 show this in percentage terms. This is then displayed as a Lorenz curve in [Figure 2](#). In Institute X, we find that 8 per cent of the most productive scientists is responsible for 50% of the contributions. In the manner of Pareto's law, we see that 78% of the output is contributed by 22% of the authors (active scientists who publish one or more papers). One can interpret this to mean that the inequality in the distribution of scientific productivity at Institute X is only very slightly less acute than indicated by the canonical 80:20 Pareto rule. This curve is incomplete in the sense that it does not incorporate the measure of the number of the scientists who were on the rolls of the Institute prior to 2003 but no longer in 2003 and who should have contributed during the 1974-2003 period but did not (i.e. the number of scientists who had 0 papers) for whatever reason. Only if this elusive group is identified will the Lorenz curve be complete. Perhaps, with this included, Pareto's 80:20 law also will be more closely approached.

To a science assessor such distributions are very valuable. About 10-20% of the scientists in any organization are responsible for about 50% of the papers published and the remaining 80-90% account for the remaining half of the output. Another 10-20% will be trying to move up the ladder (value chain in modern management terminology), and these should be encouraged. There will always be about 60% who will remain at the bottom

The Skewness of Unfairness and Scientific Excellence

So far, we have dealt with the question of scientific productivity alone as evidenced by the number of publications in SCOPUS journals. Since these journals are chosen by some exclusive criteria, this is in itself a measure of quality for scientific output coming from a third world

country. We have seen from Lotka's and Zipf's Laws and the Lorenz curve, how unfairly human ability is distributed across a given population if the measure of ability is restricted to a raw count of the number of papers alone.

A very popular measure of the quality of a paper is the number of citations that it has received in the open literature subsequent to publication. SCOPUS also provides a complete citation database for the last ten years of all papers published during this period 1974-2004. Again, we use the whole count method for computation of the performance instead of fractionating a citation into partial counts. Thus a citation for a paper with three authors will be counted as 3 whole counts of citation. Accordingly, each author, whose name has appeared in a paper is given credit for the citation regardless of the number of co-authors. [Table 3](#) shows how the citations are distributed among the 366 authors. It is now possible to identify a group of hapless scientists (249 in number) who have not managed to collect a single citation for any of their contributions over this 30 year period. This is almost 68% of all the scientists who appear in our list. At this low end, it is seen, and frustratingly so, that Lotka's law could not be applied. One of the problems is the presence of this large group with 0 citations.

Similarly, at the high end of performance, it was not easy to apply Zipf's law. However, it is clear from [Table 3](#) that now approximately 2% of the high-end authors account for 50% of the whole-count citations. The inequality is greater than predicted by the canonical Pareto rule; in fact 12% of the authors account for 88% of the citations. This is because those with 0 citations have been identified; i.e. the 68% of the scientists who have not received a citation during this period. This is very graphically illustrated by the Lorenz curve in [Figure 3](#). What we now see is that when scientific excellence is brought into the picture, the skewness of the unfairness is even more acutely emphasized.

Some sociological deconstruction

The main lesson from this exercise is that science assessment is too important a subject to be left to scientometricians who try to straight-jacket the data to simple formulae like Lotka's law and Zipf's law. We are dealing with complex situations governed by multiplicative random processes and this is why the Lorenz curves with such skewed tails are seen. We see from Institute X's profile that there are two main groups. One serves the mission-oriented, time-bound projects (maybe about 90% of the population), who have 0 or few papers. The smaller group does more academic and open-ended work and the existence of these two separate groups may explain the bimodal distribution seen clearly in the histogram in [Figure 1](#).

We also see that even among the group that is presumably devoted to academic research, there is a very high concentration of excellence in a very small sub-group. Thus 2% of the scientists account for 50% of the highest quality work. However, such are the vagaries of the reward system that often the 2% of the scientists who account for 50% of the awards gathered are not from this deserving population. It is tempting to visualize a Pareto's law of luck whereby 2% of the population has 50% of the luck! This underlines the need to have a more rational assessment procedure that will clearly demarcate the high quality workers and reward them accordingly.

- Lotka, A. J., *J. Washington Acad. Sci.* , 1926, **16**, 317-323.
- Zipf, G., *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambridge , Massachusetts , 1949.
- Narin, F. and Breitzman, A., *Research Policy*, 1995, **24**, 507-519.
- Pareto, V., *Cours d'Economie Politique*, Droz, Geneva , 1896.

Gangan Prathap

Links to

[Table 1. Cumulative list of scientists/authors and their papers](#)

[Table 2. The Pareto principle and the pyramid of wealth distribution](#)

[Figure 1. Histogram showing the distribution of performance in terms of the number of papers published.](#)

[Figure 2. The Lorenz curve shows that 8% of the high-end authors accounts for 50% of the output. In Pareto' terms, 78% of the output comes from the more productive 22% of the authors. Those with 0 papers have been identified, and introduced into the figure, and this helps explain why Pareto's 80:20 law has been more closely approached.](#)

[Table 3. Cumulative list of authors and their citations](#)

[Figure 3. The Lorenz curve for citations shows that less than 2% of the high-end authors accounts for 50% of the whole-count citations. The inequality is greater than the Pareto rule; in fact 12% of the authors accounts for 88% of the citations. This is because those with 0 citations have been identified; in fact 68% of the scientists have not received a citation during this period.](#)

[| About NAL |](#) [| Review of Wings of Fire|](#) [| Review of Science Matters|](#) [| Research Assessment at NAL|](#) [| The FEPACS saga|](#) [| Who's Afraid of Research Assessment?](#)
[| CET Watch 2004|](#) [| K A V Pandalai - An Obituary |](#) [| P N Shankar's columns |](#) [| Review of Imagined Worlds |](#) [| The Dharma of Science |](#) [| Home |](#)

[| Please send us your reaction |](#) [| Write to G Prathap |](#)
posted on 25 Oct 2004