A model-based statistical analysis of life sciences research in India during 1992–94

Arunachalam (*Curr. Sci.*, 1999, **76**, 1191–1203) presented a statistical analysis of 20,046 papers published in 1582 journals in the area of life sciences during 1992–1994, by over 1400 institutions located in over 450 cities/towns in India. It was observed that about 46% of these papers were published in non-SCI journals and further 37.5% in journals with impact factor less than one. Further, the paper also gave the leading 23 institutions ranked according to the total number of publications from them.

For the above data we propose a model for the distribution of impact factor. Based on such a parametric model, we propose a quality index as a function of the parameters occurring in the model. In the proposed model, the papers published in non-SCI journals are assumed to have zero impact factor, although it is possible that such a paper may influence future research as has been pointed out by one of the referees. However, publication in a non-SCI journal is regarded similar to infant mortality in demographic analysis or instantaneous failure in analysis of life time distributions.

We have followed the current practice of the evaluation of quality of research through the impact factor of the journal in which the article is published, as has been done by Arunachalam.

The above model-based analysis of the data of Arunachalam indicated that the quality indices of the leading 23 institutions do not differ from those of the rest of the institutions. We then ranked according to the proposed quality index and observed that Spearman rank correlation coefficient in the two types of ranking is negative.

After the initial classification of journals into non-SCI and SCI, it was observed that the frequency of papers with higher impact factor was decreasing and the rate of decrease is quite fast, a situation commonly occurring in life time distributions in demography or industrial statistics. We thus propose a model such that, if $F(x)$ denotes the distribution of the impact factor, then

$$-\log [1-F(x)] = -\log P(X > x).$$

Since the impact factor is bounded by a constant, a simple model can be taken as

$$F(x) = 1 - \exp\left\{\frac{-ax}{(b-x)}\right\} \quad 0 < x < b,\ a > 0, \tag{1}$$

where b is the maximum possible impact factor.

The rate at which $-\log [1-F(x)]$ decreases is given by the failure rate function

$$\mu(x) = \frac{\mathrm{d}}{\mathrm{d}x}\{-\log[1-F(x)]\} = \frac{1}{\frac{b}{a}\left(1-\frac{x}{b}\right)^2} \quad 0 < x < b, \tag{2}$$

which is an increasing function of x over $0 < x < b$, so that papers published in SCI journals with high impact factors would have small frequencies.

Thus, together with initial probabilities of the event of a paper being published in a SCI journal ($x > 0$) or a non-SCI journal ($x = 0$), the model for the distribution of impact factor is now given by

$$G(x) = (1 - \nu) + \nu F(x),$$

where $P(x > 0) = \nu$ represents the probability that a paper is published in a SCI journal and $P(x = 0) = 1 - \nu$, i.e. the paper is published in a non-SCI journal.

Therefore

$$G(x) = 1 - \nu\exp\left\{\frac{-ax}{(b-x)}\right\} \quad 0 < x < b,\ a > 0,\ 0 < \nu < 1. \tag{3}$$

The parameter ν is estimated by the proportion of papers published in SCI journals, b is estimated by the observed maximum impact factor and the estimate of a is obtained by the method of maximum likelihood using estimated values of ν and b. Thus we obtained $\hat{\nu} = 0.5355$, $\hat{b} = 25.47$, $\hat{a} = 25.79$ and the estimated distribution function as

$$\hat{G}(x) = 1 - \hat{\nu}\exp\frac{(-\hat{a}x)}{(\hat{b}-x)} \quad 0 < x < \hat{b}.$$

We note that the maximum absolute difference between the observed and estimated distribution function for Arunachalam's data is 0.0416. Figure 1 shows that the fitted distribution adequately describes the distribution of the impact factor.

We next attempt to define a quality index Q for the distribution $G(x)$ defined as above. Q should reflect the current evaluation, which prefers publication in SCI journals and particularly in those with high impact factors. Thus larger values of ν and b would be indicative of better quality of research. Further, smaller values of a which corresponds to rate at which high impact factor publications decrease would ensure higher proportion of publications in high impact factor journals. Thus Q must increase in each of the variables ν, b and $1/a$. Looking at the failure rate function given in eq. (2), a simple quality index which satisfies the above requirements can be proposed as

$$Q = \frac{b}{a} + \nu. \tag{4}$$

The above index for the complete distribution then is given by $Q = 1.5229$. We can also work out the quality index for any subgroup in a similar manner by fitting the distribution of type G for that subgroup and estimating (ν, a, b) for the subgroup. For example, quality index for the leading 23 institutions is $Q_t = 1.7938$, whereas for the remaining over 1375 institutions, it is given by $Q_r = 1.3825$. We also note that the estimate of b in both groups was the same, namely 25.47, but $\hat{\alpha}_t = 0.6429$ compared to $\hat{\alpha}_r = 0.4888$, and $\hat{a}_t = 22.13$ whereas $\hat{a}_r = 28.50$.

We also worked out quality indices for the leading 23 institutions in a similar way and obtained Spearman rank correlation between the two rankings based on total number of papers published by these 23 institutions and the ranks based on quality indices. The rank correlation worked out to be –0.6798,

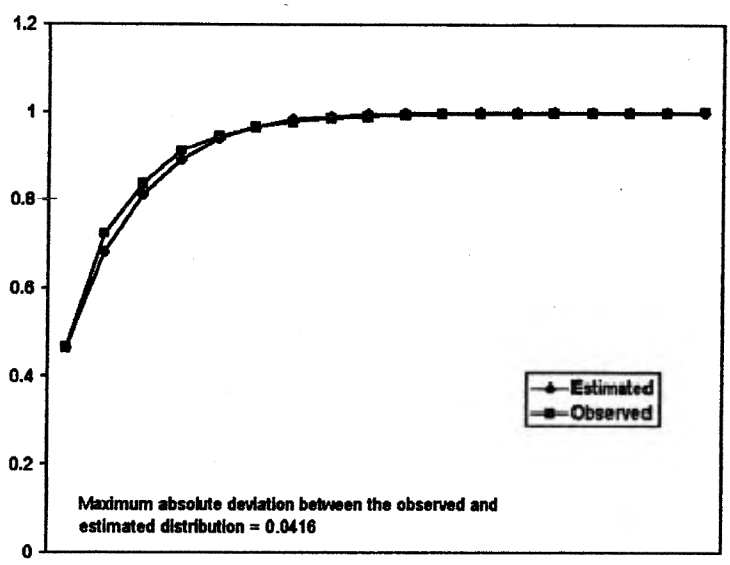

Figure 1. Graph of observed and estimated distribution functions of the impact factor.

indicating that the leading institutions must make stronger efforts to balance quality and quantity of their research publications as measured by publications in SCI and non-SCI journals and the distribution of impact factors.

As pointed out by one of the referees, to treat non-SCI journal publication as having zero impact factor may be incorrect as there are cases, although few and far between, in which important publications, particularly in new areas have been published in non-SCI journals. Like any other index number constructed to represent existing behavioural patterns of income, expenditure, etc. these index numbers may not adequately represent the goals to be achieved.

The above statistical analysis indicates that the distribution of the impact factor is extremely skew (i) over all papers (ii) over papers published by the leading 23 institutions and (iii) over papers published by the rest of the institutions. The quality indices of these three groups is not far apart. Further, the distribution of quality indices of the leading 23 institutions is also skew, with 16 out of 23 institutions having quality index less than 3 and only 3 institutions having quality index more than 6, while the remaining 4 institutions have quality index in the interval (3, 6). The analysis based on quality indices indicates that even among the leading 23 institutions there are three clusters. The first cluster consists of a large number of institutions publishing in non-SCI journals and journals with low impact factors. The second cluster is formed by a small number of institutions publishing in journals with low to medium impact factors and the third cluster of still smaller number of institutions publishing smaller number of papers but in high impact factor journals. The analysis based on Q thus supports the discovery of two clusters by Arunachalam based on total number of publications using percentiles of the observed distribution of the impact factor.

The quality index developed here needs to be augmented by a covariate representing infrastructural facilities available, including the financial support. Such an augmentation is possible by standard techniques commonly used in survival analysis with covariates. However, this would need additional database and much more hard work towards model building. Other quality indices can be defined by various functions of a, b and ✓ which satisfy the basic properties listed above and the choice of a suitable quality index is an open problem. One can get better feel for such indices after they have been worked out for a sufficiently long period of time and over other disciplines as well.

ACKNOWLEDGEMENTS. We thank the Head, Statistics Department and Head, Geography Department, University of Pune for providing necessary research facilities. We also thank the referees for their helpful comments, which improved considerably the earlier draft of the paper. V.Y. thanks DST for providing SRF under DST project 'Quantification of Manpower and Financial Resources in Academic Sector'.

Received 18 February 2000; revised accepted 20 June 2000

B. K. KALE[†,*]
VIDYA YERNENI[‡]

[†]*Department of Statistics and*
[‡]*Department of Geography,*
University of Pune,
Pune 411 007, India
**For correspondence.*
e-mail: bkkale@stats.unipune.ernet.in