The Haitun dichotomy and the relevance of Bradford's law

B.C. Brookes
Department of Information Science, The City University, London EC 1V 0HB, United Kingdom

Received 22 February 1984

In a critical review of all the empirical laws of bibliometrics and scientometrics, the Russian statistician S.D. Haitun has shown that the application of modern statistical theory to social science data is 'inadmissible', i.e. it 'does not work'.

Haitun thus points to the need to develop a wholly new statistical theory for the social sciences in general and for informetrics in particular. This paper discusses the implications of Haitun's work and explains why the older Bradford law still has an important role to play in the development of a new theory.

1. Fifty years on

Harry East [1] has recently reminded readers of the *Journal* that it is exactly 50 years since Bradford first described the peculiar statistical regularity he had observed in the bibliographies prepared at that time by the staff of the Science Museum Library in London. As the relevance of the Bradford law to information work is still not fully appreciated, this is a good moment to try to explain its importance, especially as recent Russian work puts the Bradford law in a wider perspective.

Information scientists often cite the neglect of Mendel's work by his scientific contemporaries when he reported the results of the sweet pea hybrids he had grown in the quietude of his monastery garden. Mendel found that his hybrids conformed quantitatively with the binomial theorem! Ridiculous! But 50 years later, Mendel's statistical oddity helped to found modern genetics. In another 50 years, I expect historians of the social sciences to tell a similar story about Bradford and the results he discovered in the quietude of his Director's office in the Science Museum Library and published in the journal *Engineering* 50 years ago.

Robert Fairthorne knew Bradford and tells me that they talked about the law. Bradford sensed that it had a significance beyond its immediate bibliographical context but what exactly that significance was he could not say.

2. A statistical oddity

The searching of a data-base for information relevant to some specified topic is an activity central to information work. So the fact that a statistical regularity can be found in the outputs of such activities is a surprising discovery which is of continued interest even though the data bases now searched are highly mechanized. As Bradford had been a professional physical chemist before he became a librarian he must have had experience of organizing and analysing experimental data and of conventional statistical techniques. His problem with the statistical regularity he had observed was to express it as *exactly* as possible. But he found that it defied exact expression in conventional terms and he was forced to express it as a *ranked* distribution in which the contributing journals were arranged in descending order of their productivities of relevant papers.

In 1967 Leimkuhler [2] showed that the Bradford law could be expressed in the form

$$F(x) = \frac{\ln(1 + bx)}{\ln(1 + b)} \qquad (1)$$

where b is a parameter and $x = r/n$ where r is the rank and n the total number of journals. This was a helpful clarification though it ignored Bradford's 'nucleus' and so was incomplete. Nevertheless, the analytical form of the law was clearly revealed.

Rank formulations are wholly unconventional. Ranks are *ordinal* numbers. One cannot logically calculate with 1st, 2nd, 3rd, as though they are the real cardinal numbers 1, 2, 3, that we do use for calculation. So the Bradford law has remained a statistical oddity because it could not be expressed

North-Holland
Journal of Information Science 8 (1984) 19–24

exactly as a conventional frequency distribution.

A further snag was that the law could not be derived from consideration of classical probability theory. As it cannot be related to that theory it has been excluded from the text-books of orthodox statistics.

Yet when the mechanized data bases became operational, the print-outs I examined at that time, some very much larger than any bibliographies that Bradford could ever have seen, conformed with the Bradford law.

So though there is an intriguing problem here, the Bradford law has remained a statistical oddity of interest only within the context of bibliography and information retrieval.

3. Other statistical oddities

An empirical law closely related to Bradford's was noted at about the same time by the linguist G.K. Zipf who published his law of vocabulary in 1935. The main difference between the two laws is that Bradford cumulated his data whereas Zipf did not.

I recently showed [3] that the Bradford and Zipf laws are equivalent to a third law introduced by Laplace as long ago as 1774. This law can be expressed as a frequency distribution but the Laplace law also has led to much controversy. It has baffled all attempts by theorists to derive it from classical probability theory even though some of them admit that the law is 'plausible'. It can be derived, however, by adopting the so-called 'principle of indifference' which offers a simple way of allowing for human judgement in cases where human judgement has to be used to get the estimates required. Though the Laplace law has given rise to a very large literature, it too remains a statistical oddity.

Thus, though it was useful to discover the frequency form of the Bradford and Zipf laws, I otherwise showed only that the group of statistical oddities was even wider and that they were closely related. As it is possible to subsume the Lotka law of scientific author productivity under the Bradford law also, it was becoming apparent that there was a large family of empirical laws found in bibliometrics and scientometrics which were all closely related but which could not be captured by orthodox statistical theory. Why not? How could this family of empirical laws be made academically respectable?

4. Haitun's clarification

During 1982, the Russian statistician S.D. Haitun published a series of papers [4] in which he brought together all reported empirical distributions of bibliometrics and scientometrics—over 100 of them—and critically compared them with the distributions of modern statistical theory. What he noted as a result of his comprehensive critical review and by the application of modern statistical theory was that all these statistical oddities do in fact lie outside the scope of that theory and he also discovered why.

First, all examples of orthodox distributions are found in *physical* contexts but all the statistical oddities are found in *social* contexts. Secondly, whereas all the *physical* distributions have as many higher moments as modern statistical theory requires, the *social* distributions have no moments at all. Because, as Haitun shows, modern statistical theory is essentially a theory of distribution moments, it cannot be usefully applied to distributions with no moments at all. In short, modern statistical theory is a *physicalist* theory and the distributions which arise in the *social* sciences lie beyond its reach.

Yet, of course, modern statistical theory *is* widely applied to the social sciences—to economics, for example—and is the analytical instrument which is used in formulating social policies. In the *physical* domain, modern statistical theory 'works' admirably, but in the social domain its success is more doubtful. Is *money* a *physical* or a *social* entity? It has been assumed by economists that money can be regarded as a physical entity but when we see so many competing economic theories and so many hotly disputed economic policies, all justified by the measuring and analysis of monetary transactions by modern statistical theory, the success of that theory is less obvious. The only justification for using it in the social sciences is that at the present time it is the only analytical calculus we have.

So Haitun points to an urgent need to develop a statistical theory designed for the social sciences. It has taken Bradford and Zipf 50 years to make their point.

5. G-type and Z-type statistics

Haitun described all the physical distributions which have all the moments that modern statistical theory requires as 'Gaussian'—I shall call them 'G-type'. When, using statistical techniques, one tries to fit empirical data to some hypothetical distribution, the conventional technique assumes that the empirical data constitute a finite sample from the hypothetical infinite population. If the data are of the G–type, the numerical values of the distribution parameters are found by calculating the first, second, ... moments. Thus, if the data appear to conform with a Poisson distribution, all that is needed is to calculate the mean of the data. The formula of the hypothetical population can then be written down and the corresponding values of the variable can be calculated. The degree of match, or of mis-match, between the data and the corresponding calculated values can then be estimated. To see whether data fit a Normal (i.e. Gaussian) distribution, the mean and the variance of the data distribution need to be calculated because the Normal distribution has two parameters. The technique is now highly sophisticated and works very well.

Haitun described those distributions which 'have no moments at all' as 'Zipfian'—I shall call them Z-type. When it is said of a Z-type distribution that it 'has no moments at all', the orthodox statistician is saying that the hypothetical distribution, from which it is presumed the data are drawn as a sample, has moments which are all infinite. And if all the moments are infinite, modern statistical theory has nothing useful to say about the distributions. All the 'statistical oddities' I have mentioned are of the Z-type.

We therefore need for the social sciences a calculus of Z-type distributions comparable in analytical power with that of the sophisticated calculus of G-type distributions for the physical sciences which has been developed over the past 100 years. It will need time and effort to work it out but Haitun has clarified the problem we face.

6. Frequency and frequency-rank distributions

As Z-type distributions arise only in the social sciences (with one interesting exception to be noted later) it is not surprising that even hypothetical infinities do not arise. All social affairs remain finite. So I seriously considered creating a theory of moments for *finite* distributions, i.e. with n terms though with n remaining large but always finite.

The results were not encouraging. The typical Zipf frequency distribution is expressed as $f(x) = k/x^v$ where the parameter v can have any value greater than 1, though it usually lies between 1.5 and 2.5. Unfortunately there is no convenient algebraic formula for finite sums of such series; they just have to be calculated term by term as required. Even with all possible mechanical and analytical aids the process is very clumsy and demands heavy computation.

But another factor has to be taken account of. In order to express his law *exactly*, Bradford was forced to adopt the unconventional *frequency rank* form of distribution. Though Zipf considered the *frequency* function cited above, he subsequently abandoned it for a *frequency-rank* distribution also.

There is a good reason for this choice. In *social* affairs, though not in *physical* affairs, the people, their artefacts and the other social entities we may be concerned with are distinguished from each other and identified by their *names*. Thus all concerned with the writing and publishing of books and papers—the authors, editors, librarians, readers ..., go to some trouble to give a title to their work, to the journal or publisher, to the lists of references and to all the other elements which documentalists and others have to take note of when one publication among millions of others has to be identified. Without such identifiers the search for the information we seek among the documents accessible to us would be reduced to browsing hopefully in the overwhelming 'noise' of irrelevance. Even when a computer is used to speed the search it has to be fed with discriminating cues—names, words, ...—to give relevance to its search.

Any statistical calculus which ignores such identities is not going to be of much use for information work and other social activities which depend on the searching of files concerned with humans and their social activities. But orthodox statistics, based on and derived from analysis of the anonymities of the physical domain, ignores much of the information about the particulars on which discourse about social matters depends.

I have already shown that *frequency-rank* distri-

)utions retain for analysis information which *frequency* distributions discard automatically as empirical social data are cast into the frequency classes on which they operate. Those social elements which form frequency classes are stripped of some part of their individualities and this loss is measurable [5]. This loss implies that f/r distributions are theoretically more penetrating than the corresponding frequency distributions because no subsequent theoretical device, however sophisticated it may be, can make up for the loss of *empirical* information. [6].

Another serious defect of frequency distribution analysis when applied to the social sciences is that it focuses on those elements of the data set which occur least frequently and relegates to the distant tail those elements which are the most productive. Bradford and Zipf adopted the f/r distribution because, as professionals interested in bibliography and linguistics rather than in statistics, the f/r distribution gave priority to the most productive sources and relegated the least productive journals or the rarely used words to the distant tail. In social affairs, it is usually the most productive or most active elements which are of greatest professional interest.

So though Haitun, as a professional statistician, appears to have adopted the frequency distribution $f(x) = k/x^{v}$ as the typical distribution of the social sciences, I regard this choice as unhelpful. It is of course a 'Zipfian' distribution as he defined the term but it is not the distribution Zipf himself finally adopted. The only advantage it offers is that, by taking logs of both sides, it becomes

$$\log f(x) = \log k - v \log x \qquad (2)$$

so that data sets which conform with it yield a straight line graph of slope $-v$ when the data are plotted on log/log paper. But in theoretical analysis it is intractable. Moreover, it is a frequency distribution and therefore discards empirical data which are often hard won.

7. The Bradford nucleus

Bradford centred the formulation of his law around the concept of the nucleus. Formula (1) which Leimkuhler derived applies only to the peripheral journals beyond the nucleus. With the advantage of being able to study large print-outs from mechanized data bases I found that the journals of the nucleus, which in Bradford's words were 'more specialized in the subject' also conform, in most cases, with the law that applies to the periphery but with different values of its two parameters. Thus one could write:

$$G(r) = \begin{cases} k_1 \ln(1 + r_1/w), & r_1 = 1,2,3,\ldots,n, \\ & \text{the nucleus}, \\ k_2 \ln(1 + r_2/w), & r_2 = 1,2,3,\ldots,p, \\ & \text{periphery}, \end{cases}$$

where $r_2 = r - r_1$. These equations imply that there is a discontinuity which separates the nucleus from the periphery. But the existence of the nucleus also implies that Bradford recognized that bibliographies are usually not homogeneous. Different kinds of categorization are applied to the specialist and the peripheral journals.

In studies of vocabulary, I now apply the cumulative Bradford law in preference to Zipf's own law and find that in vocabularies there is usually a nucleus also. The vocabularies of natural language texts are rarely homogeneous: more usually they are multiply heterogeneous i.e. as analysed by the Bradford law they may form several distinct groups. And authors writing in the same language betray differences of writing style through f/r analysis of their vocabularies.

Regrettably, because Bradford's concept of the nucleus has been ignored by orthodox statisticians dependent on frequency distribution analysis and seeking the 'one true law' which would capture all variants of the statistical oddities, much effort has been wasted in futile dispute about the exact form of 'the one true law' because the implication of the Bradford nucleus has not been recognized.

8. The Bradford law as an analytical instrument

Haitun's clarification of the relation of the Z-type statistical oddities to modern statistical theory will give an impetus to research on Z-type distributions and their application to the social sciences. One of the first needs will be to adopt a distribution which is convenient to use and which can be used to sort out all possible variants. For such exploratory analysis the Bradford f/r law together with its Laplace frequency analogue looks

the most promising of the candidate forms. One advantage of the laws is that they are 'telescopic'. For example, if a very large data set looks as though it conforms with the Bradford law

$$G(r) = k \ln(1 + r/w),$$

then it should also conform with the law

$$G(s) = k \ln(1 + s/w')$$

where the s series consists of the sums of s consecutive terms of the r series and where $w' = w/s$.

Similarly, if for the Laplace law we have a distribution conforming with

$$F(x) = \frac{k}{j} - \frac{k}{j + x}$$

then if consecutive sums of x terms are grouped together s at a time to form a new z series, the z distribution should conform with

$$F(z) = \frac{k'}{j'} - \frac{k'}{j' + z}$$

where $j' = j/s$ and $k' = k/s$.

the judicious use of these telescopic properties makes possible a very quick preliminary scan of the data in the search for any discontinuities that signal a lack of homogeneity.

9. The Bradford law as a descriptor of social affairs

As a statistical descriptor of social affairs the Bradford law is not unduly flattering to the human race. In an early paper [7] I said it described a social mechanism in which 'success breeds success'. The more productive, the richer, the stronger is any social 'element' in the particular social activity described, the greater is the chance of that element becoming even more productive, richer or stronger. Thus it describes very realistically a Darwinian type of struggle for survival in a competitive world.

For example, I have recently been applying the Bradford law to the published data on 'household incomes in the U.K.'. This possibility arises because some of the data are published as a ranked distribution. The Bradford graph of 'original' incomes reveals a nucleus of keen competition followed by a peripheral group of those who are no longer competing—the pensioners, the unemployed, the infirm and those who may prefer to play some other game. The Bradford graph of 'final' household incomes, i.e. after income-tax and social security benefits are taken into account, conforms very closely with the Bradford law except that there remains some misfit at the poorest end to their relative disadvantage. But this close fit to the law suggests that those who collectively manage these matters seem to be imbued with a sense of economic justice which is realistic rather than egalitarian. However, the index used officially to measure the effects of the redistribution of incomes—the Gini coefficient—measures the differences from *equality* of incomes and so implies that this is the official ideal. But the Bradford law is more realistic in its measure of the redistributions and I am presently engaged in noting the changes in the Bradford parameters of the redistributions over recent years.

I mention this application to exemplify the fact that the Bradford law need no longer be confined to the bibliographical field in which Bradford first found it. I believe it is directly applicable to all social populations engaged in some measurable activity and in which at least some of the population are engaged competitively.

10. Z-type distributions and intelligent machines

An exceptional Z-type distribution noted by Haitun is that of the energies of the particles which reach Earth from outer space and are therefore clearly in the physical domain. Physicists regard it as describing the steady entropic decay of the energy sources of the cosmos—the fading echoes of the primal 'big bang'.

The entropy of a physical system is a measure of its statistical disorder. There is a theory that life could emerge in the cosmos, creating its own minute and temporary patches of order, only when the cosmic rate of decay had reached a stage at which it could be reversed—at least locally and on a relatively small scale—by living forms. We build our bodies and maintain them, we build cathedrals —physical and cognitive too—against the flowing tide of steadily increasing physical disorder.

So here we are—all busily trying to bring some order into the chaos that surrounds us. (Or are we? The inimitable Stanislaw Lem seizes on this idea: he ascribes to 'Academician A. Slysz' the comment that "it is quite possible that the Universe with a

weakened tendency to entropy could give rise to very large information systems that turn out to be very stupid". We have been warned!)

As a materialist, Haitun seeks a physical basis for *all* phenomena. Though I accept Haitun's general conclusions I cannot help noting the great void which separates physicalist theories about the emergence of life and the emergence of human knowledge. So Haitun has to speculate. He argues that "dependencies of the type $E(x) = \rho \ln x$ are the only ones to be encountered in the investigation of human society" and he expects that what he calls the 'Zipf' law will prove to be "the fundamental quantitative law of human activity".

Though information transactions necessarily have a physical basis, the cognitive aspects of these transactions far transcend the physical effects and are different in kind. And, as Haitun so clearly shows, the physical and the cognitive phenomena rest on different and opposing entropic bases. When the differences are so clear and so profound, I see methodological advantages in conceding, as Karl Popper has done, an autonomy to the world of information and knowledge.

To take this step also makes it possible to provide the cognitive world with a mental space which is different from that of physical space. There is nothing extreme about such an idea. Whenever mathematicians feel a need for special spaces different from the physical one we all share, they invent whatever they need. Thus I note that in current studies of the packing of spheres they are now working with spaces of more than 100 000 dimensions yet this work has potential applications both theoretical and practical. The mental space I seek would have to accomodate the *idea* of such multi-dimensional spaces and many other imaginative creations too. It is very different from the space of the physical world which accomodates our bodies.

A further advantage of adopting a special mental space is that it can be designed to accomodate Haitun 'dependencies of the type $E(x) = \rho \ln x'$ without difficulty (it would be like the perspective space of landscapes) and so help to maintain the distinction between physical things and mental entities that are still too often confused.

This argument may seem remote from the practical world of information scientists but the Alvey programme's target of creating "intelligent knowledge-based systems " will soon be facing these issues head-on. The hardware and software of these systems are being confidently designed on the basis of the G-type calculus and its associated logics. But if these systems are to simulate human thought (and mechanize the work of information scientists) rather than produce further ingenious machines which humans still have to adapt to, interpret for users, or complement, they will eventually have to work with a Z-type calculus on Z-type phenomena.

The target problem of the IKBS is the same that Bradford explored 50' years ago—that of identifying the relatively few elements relevant to some human need from the background of irrelevance in which they are embedded—a Z-type problem.

The relevance of Bradford is that he discovered the shape of relevance.

Acknowledment

I gratefully acknowledge the support of a Leverhulme Emeritus Research Fellowship while preparing a book which expands ideas outlined in this paper.

References

[1] H.East, Bradford revisited, *Journal of Information Science 7* (1983) 127–129.

[2] F.F. Leimkuhler, The Bradford distribution, *Journal of Documentation 23* (1967) 197–207.

[3] B.C. Brookes, The empirical law of natural categorization, *Journal of Information Science 6* (1983) 147–157.

[4] S.D. Haitun, Stationary scientometric distributions, *Scientometrics 4* (1982); Part I 5–25, Part II 89–104, Part III 181–194.

[5] B.C. Brookes, People versus particles, in: *Theory and Applications of Information Research*, Proc. 2nd International Research Forum on Information Science, Copenhagen, 1977 (Maxwell, London, 1980) 106–119.

[6] B.C. Brookes, *Towards Informetrics: Haitun, Laplace, Zipf, Bradford and the Alvey Programme*, to be published.

[7] B.C. Brookes, Bradford's law and the bibliography of science, *Nature 224* (1969) 953–955.