INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)
Elsevier Science Publishers B.V., 1988

APPLICATIONS OF THE BIBLIOMETRIC DISTRIBUTIONS

A. BOOKSTEIN*

University of Chicago, Graduate Library School
Chicago, Illinois 60637, USA

Abstract

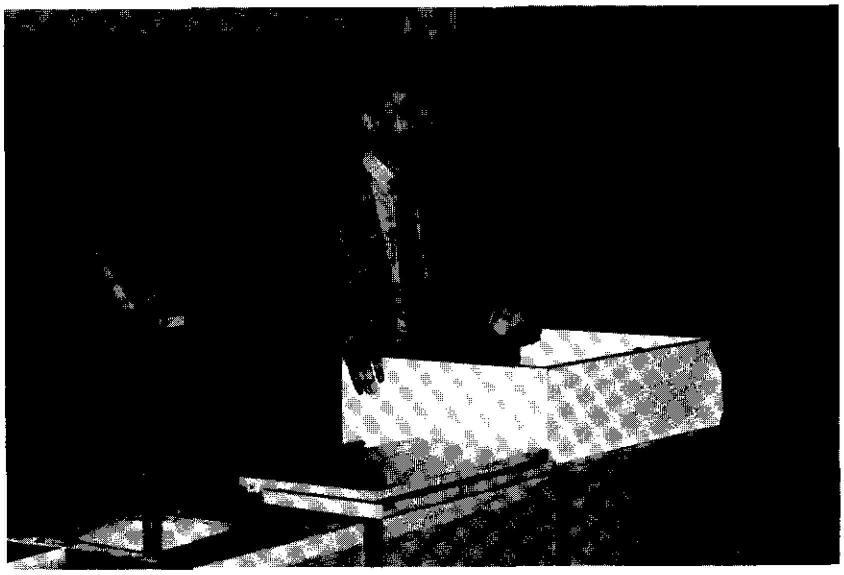
Two questions often asked of researchers in bibliometrics are : 1) Do these distributions really exist ? and 2) How can they be applied ? We here suggest these two questions are related : that one way to defend the reality of these distributions is to demonstrate that one can make reasonable decisions in practical situations by using them. To indicate how they can be applied, a microeconomic approach is taken to analyzing a specific, and in the library literature, often discussed, problem : how to allocate resources over various parts of a library book collection. A conceptualization of this problem is presented in which it is necessary to have a model of book use before we can proceed. The consequences of using the bibliometric distribution is compared with those of other plausible candidates. The differences are discussed.

1. INTRODUCTION

The bibliometric distributions constitute a family of probability distributions that occur very often, in a variety of fields, when people count things. In their analytical forms they provide precise statements of regularities that typically have first been noticed informally and with surprise. Although they have been studied for decades, their origins remain a mystery; perhaps for this reason, they are often treated as curiosities not fully to be believed.

Two questions frequently arise when these distributions are discussed : 1) What good are they ? and 2) Are they really there ? Although this paper will concentrate on the first of these questions, I believe that the two questions are related, and that an understanding of how one can talk about het first provides insight into the second as well. The question of whether a probability distribution "really" describes a phenomenon is a difficult one, here as elsewhere. In the ultimate sense, it is unlikely that any theoretical distributions describes any actual phenomenon. The usual approach to answering the question of whether a particular distribution adequately describes a body of data is to carry out a goodness of fit test. But such tests are sensitive to the amount of data that is available as well as to the correctness of the distributions itself; thus increasing the amount of data being tested may well reveal a distribution previously thought applicable as not fitting after all. Large bodies of data reveal small discrepancies, and there always will be small discrepancies if we have enough data to discover them. Yet people do describe data by theoretic distributions, make decisions on the basis of such descriptions, and often succeed in this endeavor. Of course, the answer to this problem is that the theoretic distributions are approximations, and as such can be quite useful. It is only when one is concerned with distributions that one is not predisposed to believe that this question comes seriously to the fore. The question we are posing, then, is : Once we accept that the bibliometric distributions are approximations,

* This material is based, in part, on work supported by the National Science Foundation under Grant IST-8026564. This support is gratefully acknowledged.



A. Bookstein

in what sense can we argue that they describe a body of data, or that a particular phenomenon is well-modeled by these distributions. I believe two approaches are helpful here :

1. If there are theoretical reasons for expecting a distribution to describe a phenomenon, at least in a limiting or idealized sense, one will feel more comfortable in using it. In the case of the bibliometric distributions, several attempts to find theoretical justifications do exist [1-4].
2. A second approach that we believe could play an important role here is to accept a distribution as describing a body of data - or a mechanism for generating that data - if one can make better decisions using that distribution than one can make using other distributions (using no distribution is in fact using some implied, intuitive distribution). Though the bibliometric distributions have not been discussed from this point of view, it is not entirely new. For example, hypothesis testing often compares a null hypothesis, as emphasized in analyses of the power of statistical tests. However, in that case it is usually members of the same family that are compared, and, in any case, discussions of power are typically not given much attention in traditional data analysis. The approach we are taking in this paper is much closer to that of the Bayesian or Decision Theory school, in which costs and the need to make decisions which have consequences that can be evaluated play a much more central role.

Specifically, suppose F_0 is the family of bibliometric distributions and the set $\{F_i\}$ denotes other families or distributions; in each family, members are distinguished by parameter values. If a decision must be made on the basis of data, D, then we can consider $c_j(F_i,D)$, the expected cost of using family F_i in a situation indexed by j, as the basis of decision making - here c includes the cost of parameter estimation. (The notation indicates that D is a random variable; c_j is also a random variable if we consider the situation to be randomly selected from a space of possible situations.) If for a particular type of data and variety of uses to which knowledge about that data might be used, the expected value $E_j c_j(F_i,D)$ is less for the family F_0 than for other families, including the implicit family of distributions a decision maker might use if a formal analysis isn't made, then it is reasonable to say that F_0 describes that kind of data. The expectation indicated by E_j is over the situations in which a decision is based on this data must be made. Thus, studying how such distributions enter into the decision-making process has theoretical as well as practical content. Below we will examine a specific problem area in which knowledge of distributional properties is important. Our emphasis, however, will be on the character of the decision making process, using the example as an illustration for concreteness. The conclusions reached using the bibliometric family of distributions will be compared with those based on other popular distributions.

2. THE BIBLIOMETRIC DISTRIBUTION

The bibliometric distributions refer to a variety of regularities taken from different fields and exhibiting a variety of forms. We first note that though these distributions differ greatly in appearance, they can be conceptualized as versions of a single regularity, so that we can properly speak of the bibliometric law and its manifestations; the differences in appearance are largely due to different applications finding it natural to cumulate data in different ways, and thus examining the relationship between different variables (this has been observed by a number of people - see [5] for an overview). In a sense, that a single form recurs is not completely surprising. The social sciences in which many, but by no means all, of these regularities are found, do not deal with concepts that are as well defined as do the physical sciences, and the lack of conceptual precision and ambiguity impedes our efforts to find regularities in this domain. Indeed, for a regularity to be findable it is important that it be of a form that is resistant to such ambiguities, and this restriction greatly reduces the class of distributions from which our choice can be made. I have earlier argued that the bibliometric law does satisfy these conditions, often doing so uniquely [5,6]. I have found making these observations reduces the value of these laws for many people, who see this as a "debunking" of the laws. Perhaps this response is a consequence of our having learned to associate scientific laws with causal rather than technical explanations. But this has not been my intention. Rather, I am arguing that the place in statistics of the bibliometric law should be very much the same as that of the normal distributions, which also describes a wide variety of phenomena, and for a similar reason. We would also like to learn how to take advantage of the law when it does appear. The purpose of the following section is to provide an example of such an application; this applications is to a typical library problem, though many other examples are possible. In this example, the ability to accept the bibliometric distribution as describing the phenomena conveys real information. The consequences are useful, and, as shown, distinct from that of other assumptions. An important point to keep in mind while reading this section is that generality does not imply uselessness.

Many decision making problems in information science depend on a frequency distribution describing activity. An example from coding theory is the creation of Huffman codes [7], which depend on the frequency of occurrence of characters making up the source alphabet. The problem I shall describe here is developing a rationale for dividing a library book budget over segments of the collection. Many writers have argued that relative amounts of book use in these segments should be the basis of resource allocation, but no real conceptual foundation has been established for dealing with this problem. I shall describe an approach here, and indicate how assuming a Zipf-type distribution of use will influence the form taken by the selection. Our basic assumptions is that book purchases resemble capital investments: books are purchased today for their expected future benefits. A library provides an output, the magnitude of which depends, in part, on the number and quality of the books it can make available.

In order to be able to concentrate on the critical issues, we shall make several simplifying assumptions.

1. That we can look at the materials budget alone; actually, this is not very restrictive, since its solution for various levels of the book budget will form a component in the larger problem, in which the book budget is itself determined.
2. We shall restrict our discussion to a two-segment collection - e.g., subject A vs. subject B; college vs. research; or branch 1 vs. branch 2. (A similar approach can be taken to guide the allocation of resources among other categories of service - e.g., book purchases vs. public services). An extension to many segments is immediate, though more complex.
3. One of the most difficult tasks in discussing quantitatively the provision of information services is measuring output. It is often noted that library service is intangible. Perhaps, but decisions of a very non-tangible nature must be made one way or the other, and a quantification of output, even if very approximate, can be valuable if a rational decision is to be made. Often, measures such as circulation are used. I will develop the procedure by assuming simply that some function, $U(x_A,x_B)$, exists that gives the benefit of allocation x_A and x_B to components A and B; I shall derive the consequence of making such an assumption. To be specific, one could substitue a measure such as circulation for U. We assume that the budget is allocated in a rational manner; that is, it is allocated in a way that maximizes U [8].

Our problem, then, could be formulated as follow :

$$\max_{x_A,x_B} U(x_A,x_B)$$

subject to the constraint that

$$P_Ax_A + P_Bx_B = \text{Budget} .$$

Here, P_A and P_B are the typical prices of the books in each class. The solution to this problem, for example, using Lagrange multiplier techniques, is

$$\frac{1}{P_A} \cdot \left(\frac{dU}{dx_A}\right) = \frac{1}{P_B} \left(\frac{dU}{dx_B}\right), \tag{1}$$

that is, buy books from classes A and B so that, when done, the last dollar provides the same benefits from each group. This criterion contrasts with others that have been proposed (buy in proportion to numbers of students and faculty in an area; buy according to total circulation alone) or implicitly implemented (buy in proportion to the forcefulness of demands made by departmental library committees or based on historical precedent). Our approach, based on microeconomic reasoning, introduces a benefit/cost criterion and emphasizes the importance of cost per item as well as benefit.

If we were to implement this theoretical approach in an actual decision making situation, it would next be necessary to operationalize the concepts involved, and particularly that of benefit; that is, the function U must be made measurable. However, even without an explicit measure of benefit, it is interesting to analyze the forms taken by the solutions based on models of how utility depends on the quantity of books purchased from each class. In particular, it is interesting to compare the consequences of assuming a bibliometric model with other candidate models. Following the argument given in section 1, we would like to check whether, in a decision-making context, the bibliometric distribution can be distinguished from other plausible distributions.

We begin by making the assumption of additivity :

$$U(x_A,x_B) = aU_A(x_A) + bU_B(x_B),$$

where a and b measure the relative importance of satisfying users of class A as compared to those of class B. To be specific, we can think of U_A and U_B as circulations. Thus, our models are, in effect, models of how book use in a section of books is influenced by the size of the collection - it assesses use at the margins, if we can ignore the effect that the earlier purchases have on the availability and use of the last purchases.

If the utility function is additive; as suggested above, the optimality criterion takes a particularly simple form :

$$\frac{a}{P_A}\frac{dU_A}{dx_A} = \frac{b}{P_B}\frac{dU_B}{dx_B} .$$

Below we will be interested in the implications of this relation when it is combined with specific distributional assumptions. But the relation itself has definite implications for collection management. For, if we are willing to assign values to the parameters a and b, we would measure the performance of the last one thousand dollars, say, of book purchases (in effect determining $\frac{dU}{P}$) for various sections of the collection and test for equality.

In this manner the collection can be tuned. Altèrnatively, one can measure book use and, assuming that resources are being rationally allocated, assess the values of a and b that are implicit in resource allocation. Making these values explict can be very illuminating.

a. Exponential model : Suppose

$$U = aC_A(1 - \exp(-\frac{x_A}{r_A})) + bC_B(1 - \exp(-\frac{x_B}{r_B})), \qquad (2a)$$

where C_A and C_B are the circulations asymptotically approached as the number of items in a class approaches infinity; and r_A and r_B measure the rate at which asymptotic behavior is approached. The criterion, equation 1), under this model is :

$$(\frac{aC_A}{P_Ar_A})\exp(-\frac{x_A}{r_A}) = (\frac{bC_B}{P_Br_B})\exp(-\frac{x_B}{r_B})),$$

or

$$x_A = (\frac{r_A}{r_B})x_B - r_A\ln(\frac{bC_BP_Ar_A}{aC_AP_Br_B}). \qquad (3a)$$

b. Square root model : This model also assumes a declining rate of return on additional stock, but not as radical a decline as model a). This detailed model is,

$$U = aC_A(\sqrt{\frac{x_A}{r_A} + 1} - 1) + bC_B(\sqrt{\frac{x_B}{r_B} + 1} - 1) \qquad (2b)$$

The criterion becomes

$$\frac{aC_A}{P_A r_A}\ \frac{1}{\sqrt{\frac{x_A}{r_A}+1}} = \frac{bC_B}{P_B r_B}\ \frac{1}{\sqrt{\frac{x_B}{r_B}+1}} \qquad \text{(2nd)}$$

or

$$x_A = \left(\frac{a}{b}\ \frac{C_A}{C_B}\ \frac{P_B}{P_A}\right)^2 \frac{r_B}{r_A}\, x_B + \left[\left(\frac{a}{b}\ \frac{C_A}{C_B}\ \frac{P_B}{P_A}\right)^2 \frac{r_B{}^2}{r_A} - r_A\right] \qquad \text{(3b)}$$

c. Bibliometric model : We finally assume the Leimkuhler [9] form of the bibliometric distributions. Then

$$U = aC_A \ln\left(1 + \frac{x_A}{r_A}\right) + bC_B \ln\left(1 + \frac{x_B}{r_B}\right). \qquad \text{(2c)}$$

The criterion, given eqn 1), is,

$$\frac{aC_A}{r_A P_A\left(1 + \frac{x_A}{r_A}\right)} = \frac{bC_B}{r_B P_B\left(1 + \frac{x_B}{r_B}\right)},$$

or

$$x_A = \left(\frac{aC_A P_B r_B}{bC_B P_A} - r_A\right) + \left(\frac{aC_A P_B}{bC_B P_A}\right) x_B\ . \qquad \text{(3c)}$$

The three models provide three rates of decrease in utility per additional item. The above analysis indicates how these affect our conclusions. Perhaps most important, we note that all of the models suggest that x_A increases linearly with x_B; perhaps after establishing some initial stock, we increase stock A as compared to B at a constant ratio. The models differ in how parameters should enter, although, in each, it is the ratios of parameters that is of importance; these are also easier to assess than absolute values. In the bibliometric model, it is the ratio of the importance of satisfaction in the two sections and the ratio of cost per circulation that determines the relative addition of funds and items as we go from one section to the next. This differs from the other models; if we can accept the validity of the bibliometric model, then we gain guidance regarding to the form of the solution, and the rules are distinguished from that of other models.

3. DISCUSSION

Globally, these models are similar in that they suggest that in building a collection we purchase exclusively in one section, the one providing the most utility, until a threshold value is reached, and only then begin purchasing in the second section. These models are not impressed by arguments of balance.

But it is the margins that are most interesting, since it is here where the most difficult decisions must be made. Suppose then that the collection in in utility balance, and an additional amount of items can be purchased, say with Δ B additional funds. Let us first compare just the exponential and bibliometric models. The increase in category A books, Δx_A, given an increase in category B books Δx_B, is given by :

$$\Delta x_A = \frac{aC_A P_B}{bC_B P_A} \Delta x_B$$

for the bibliometric distribution, and

$$\Delta x_A = \frac{r_A}{r_B} \Delta x_B$$

for the exponential distribution.

These two models reflect rather distinct purchasing philosophies. The bibliometric distributions looks at the circulation per dollar ratio of the two collections and the relative utility per circulation. These are ignored by the exponential model, which, at the margins, is influenced only by the ratio at which the distribution approaches its asymptotic value (of course, the other parameters have an influence - but only in determining the threshold at which one begins purchases in the second collection). My own intuition here finds the bibliometric approach more plausible.

The square root model is intermediate between the two, taking both the relaxation rate and other parameters into account. The prescriptions do differ from the bibliometric model. For example, the ratio of additional class A purchases between the two models, for a unit increase of class B items, is

$$\frac{\Delta x_A(\text{square root})}{\Delta x_A(\text{bibliometric})} = \left(\frac{a}{b}\right)\left(\frac{P_B}{P_A}\right)\left(\frac{r_B}{r_A}\right) R,$$

where R denotes the ratio of the parameters C_A and C_B as they enter the two models. The square root model is more sensitive to differences in values put on outputs in the two classes - if we double a compared to b, the square root model will increase the additional A purchases twice as fast as the bibliometric model would. A similar effect is found if we examine the circulation per dollar

parameters $\frac{C_A}{P_A}$ and $\frac{C_B}{P_B}$.

For example, the square root model is more sensitive to price changes : doubling the price of class A books will lower the number of class A items bought twice as fast as the bibliometric model will*.

In the above development, we have implicitly assumed that initially the library has no books, and must construct a collection by choosing from an infinite inventory of possible purchases. Such a model might well approximate the library's task each year of selecting new publications to add to its collection. But some purchases are of previously published materials, so it is of interest to ask, what is the incremental value of continuing purchases for an existing collection.

It would be desirable that the form taken by this function be as simple as possible. Consider, for example, the bibliometric case : suppose, as before, that the utility of x items in a given section is given by $C\ln(1 + \frac{x}{r})$.

Suppose further that the collection already has x_0 items and we wish to compute the utility of adding an additional x items. The utility of the total is given by $C\ln(1 + \frac{x_0 + x}{r})$, which is easily shown to be equal to

* We specify the effect of price here because the parameters, C_A and C_b, like r_A and r_B, are not comparable across models.

$C\ln((1 + \frac{x_0}{r})\ (1 + \frac{x}{x_0+r}))$, or $C\ln(1 + \frac{x_0}{r}) + C\ln(1 + \frac{x}{r+x_0})$. Thus considering the x_0 items already in the collection as a datum, the number of circulations contributed by the additional x items is given by $C\ln\ (1 + \frac{x}{r+x_0})$; we see that the form of the utility function is just as before, with r becoming $r' = r+x_0$. This result is most satisfying : if the utility of the total collection is given by the bibliometric form, then so is the utility of an extension. So in making our computation we need not consider in detail the earlier purchases. Further it is not necessary to assume the collection to be in utility balance when the addition of new items is being contemplated : given the parameters C and r', details of the already existing collection can be ignored. This argument also gives some insight into the meaning of the parameter r : this appears to represent something like the number of existing items before the selection effort begins. (Of course, this interpretation is a fiction, since r will take a non-zero value even when the collection is born : it seems that collections perform as if r books were already available. The parameter can perhaps be interpreted, as the effective number of items available to a person before he uses the library's collection.)

We should note, without going into a similar level of detail, that both the exponential distribution and the square root utility function, share this property. Schematically, we can represent this property as follows : if u(x) is the utility of x items, we would like,

1) $u(0) = 0$; and
2) $u(x + x_0) \approx u(x_0) + u(x)$,

where the squigly equality denotes the intuitive notion of "having the same form", though differing in detail because of the parameters of the model changing values.

We do not yet know whether the above distributions are unusual in having this property. If the interpretation given above (in parenthesis) is valid, however, it would be difficult to see how a function not satisfying this condition can be used to describe the utility function of a specific number of items.

In this paper we examined the impact of three distributional models on decision making. Of course, other possibilities exist and should be investigated. But our purpose here was to present a conceptual approach to dealing with problems in which the distributional models may enter and to emphasize the impact they may have. The point I have most strongly tried to make is that ultimately, when we discuss distributions such as the bibliometric distributions, it is pointless to argue in abstract whether or not they have theoretic or practical value. Our analysis emphasizes that assuming different distributions in decision making contexts do yield different results; that a committment to a particular family of distributions has real consequences; and that the acceptability of these consequences is the main criterion on which these families should be evaluated.

REFERENCES

[1] Yule, G.U., A Mathematical Theory of Evolution Based on the Conclusions of Dr. J.C. Willis, Philosophical Transactions of the Royal Society of London, Series B 213 (1924) p. 2187.

[2] Brillouin, L., Science and Information Theory (2nd ed.), (N.Y., Academic Press, 1962).

[3] Perline, R., Pareto-Mimicking Models of Weakly Harmonic Laws, (Doctoral Dissertation, submitted to the University of Chicago, 1982).

[4] Bookstein, A., Explanations of the Bibliometric Laws, Collection Management, 3 (1979) p. 151-162.

[5] Bookstein, A., Robustness Properties of the Bibliometric Distributions, Report to the National Science Foundation, 1981 (Being prepared for publication).

[6] Bookstein, A., Patterns of Scientific Productivity and Social Change : A Discussion of Lotka's Law and Bibliometric Symmetry, Journal of the American Society for Information Science 28 (1977) p. 418-427.

[7] Huffman, D.A., A Method for the Construction of Minimum Redundancy Codes, Proc. IRE, 40, (1952) p. 1098-1101.

[8] Bookstein, A., An Economic Model of Library Service", Library Quarterly, 51 (4) (1981) p. 410-428.

[9] Leimkuhler, F.F., The Bradford Distribution, Journal of Documentation, 23 (1967) p. 192-207.