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

PROBABILITY DISTRIBUTIONS AND INEQUALITY MEASURES FOR ANALYSES OF CIRCULATION DATA

I.K. RAVICHANDRA RAO

Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile, Mysore Road, R.V. College Post, Bangalore 560 059, India *

Abstract

Bibliometric analysis of circulation data is a useful tool for effective library management. For the purpose of such an analysis, the theoretical basis is provided by the study of the distribution of transactions. The negative binomial distribution is a good approximation to describe the circulation distribution. Thus, the phenomenon of the distribution of the transactions is a manifestation of success-breeds- success phenomenon. However, for highly skewed, long tailed, heterogeneous data, the negative binomial is unlikely to be a good approximation. In such cases, one may derive a distribution function by maximizing the entropy.
Further, various measures of inequality are found to be equally useful in analyzing the data. Amongst these, Gini's index, Lorenz ratio, information measures (such as Theil's measure, entropy, etc.) are mostly used and easily interpretable measures.
Analysis of circulation data helps us in effective physical organization of documents, in file organization, in acquiring multiple copies of documents and in determing the core collection.

1. INTRODUCTION

Since Gutenberg's invention of movable types in the 15th century, the world's publication of documents is increasing in a rapid manner. In fact, due to the evolution of automated printing technology and also due to intensive R&D work, especially since World War II, publication of documents is increasing at an exponential rate. Consequently, investment on books, periodicals, etc. is increasing. In this world of cost efficiency, every investment is accountable especially from a view point of its utility. Bibliometrics has thus grown out of the necessity for better management of document-oriented information systems.
The first law of the late Dr. S.R. Ranganathan states that "Books are for use". Such use can, however, be directed for even its information content. The measure of use of information content on its users becomes equally important and it is within the scope of bibliometrics. Unless we define the term use precisely, it is very difficult for one to collect the relevant data to compute the necessary measures and analyse this data. It is for this purpose (of developing the measures and analysis) that the use of documents may be defined in terms of the number of times they are removed* from the shelves by the library users or for the library users. It is very difficult and time consuming to collect such data, since it is practically impossible to count the number of times a document has been removed from the shelves. We therefore, restrict the scope of the term 'use' only to the number of times the document has been circulated.

* for reading, for reference, etc.
** The terms circulated, issued, borrowed and transacted are used synonymously in this text.

I.K. Ravichandra Rao

In my presentation, I shall, therefore, concentrate only on the analysis of circulation data.

Discussions on library transactions have in recent years occupied a considerable place in the bibliometric literature. Empirical studies are, however, of very recent vintage, being a function of the availability of circulation data. Since Stieg's work [23] in 1943, availability and accuracy of circulation data in industrialized countries have improved continuously, especially because of the increasing number of automated circulation control systems. Empirical studies of the size distribution of library transactions have thus proliferated accordingly. The situation with respect to data in less-developed countries is worse and there is hardly any work being done to generate and evaluate circulation data.

The development of 'measures' reflecting the distribution of transactions has progressed in parallel with both the

i) development of data, and
ii) theorizing about the significance and determinants of the distribution.

2. ANALYSIS OF CIRCULATION STATISTICS

An analysis of circulation data gives rise to information related to the volume of use by type of documents and type of user, use of documents of different age, seasonal variations in transactions etc. Circulation data, as depicted by transaction records of document borrowed could be taken as an indicator of the use of the library resources. Simple circulation statistics suggest that a document borrowed most frequently by the same or different borrowers is more useful than one borrowed less frequently. It demonstrates that circulation data can be used to measure the degree of utility of the library resources. The borrowed document data together with the borrower data provide information which can be used by the management for formulating and adopting policies for the acquisition and processing of documents.

For meaningful results, however, circulation data should be correlated with other variables such as data/information and teaching methods adopted, circulation changes effected and university examination procedures adopted. In addition, it is possible to correlate the circulation data based on characteristics of user and documents, such as sex, status of users, subject of documents, etc. which then would give a picture of a library's use. They can be further enriched by investigating :

1) The pattern of relative frequency distribution of documents in the collection in a given period of time; and
2) The proportion of the circulated documents to the total collection in the library.

Such an investigation would involve :

1) An analysis of the data in terms of its usefulness to library management;
2) Developing a technique for collecting data;
3) Identifying the types of deductions that can be made from this data.

3. SIZE DISTRIBUTIONS OF CIRCULATION PROCESS

Size distributions of use of documents have fascinated several scientists. Using circulation data, Morse [15,16] suggested a stochastic model for predicting future demands for books. Morse's model relates circulation in one year to that of the next year, under the following assumptions :

i) There exists a minimum level of use (α) which is independent of the item
ii) There exists a level of use that tends on the average to be a fixed proportion (β) of the previous year's use;
iii) The random variable (say, N(m)) which represents the number of uses in the following year follows a Poisson Distribution.

Assumptions (i) and (ii) lead to a linear form

$$N(m) = \alpha + \beta m$$

where m is the number of uses of a document in a given year.

Assumption (iii) leads to

$$P\{N(m) = n\} = \frac{(\alpha + \beta m)^n}{n!} \exp\{-(\alpha + \beta m)\}$$

where α and β are defined in assumptions (i) and (ii) respectively; m and n are non-negative integers. This simple Markov model of book use has been modified by many and used in a number of applications in the last two decades. Chen [9,10], for instance, in a similar study extended Morse's work; her model, based on Markov theory, can be used to predict the possible future rate of demand for those documents which were weeded to a less accessible storage area. In her study at Harward's county Medical Library, she observed higher values of α and lower values of β. She, however, observed lower values of α for old documents. Chen modified Morse's model in order to remove the bias that the circulation data generally relates only to active documents; she observed that the data pertaining to active books conform to Morse's model.

Assuming that the average of use of a document decreases exponentially with its age and the average use of a document increases with an increase in the user population, Rouse [21] suggested a Markov model. He showed in his paper that the average use during the period (t + 1) is linearly related to the average use during the period (t) and he insisted that increments in the user population and acquisition policy during the time period t should also be considered.

Assuming that P (n|m) (where n and m are the number of times a document is borrowed in time t+1 and t respectively) follows a conditional negative binomial distribution, Ravichandra Rao [20] made an attempt to fit the conditional negative binomial distribution for Morse's data. The conditional mass (density) function is

$$P(n|m) = \binom{k+n+m-1}{n} \left(\frac{P_2}{1+P_2}\right)^n \left(\frac{1}{1+P_2}\right)^{k+m}$$

$$P(m) = \binom{k+m-1}{m} \left(\frac{P_1}{1+P_1}\right)^m \left(\frac{1}{1+P_1}\right)^k$$

P_1,P_2 and k are constants; m and n are non-negative integers. He observed that P(n/m) as a conditional negative binomial distribution is a better contender than a Poisson model.

The reliability of the basic assumptions underlying the parameter of Morse's model has not been tested until recently. Utilizing 11 years of circulation transactions of Saskatchewan University Library, Beheshti and Tague [1] showed that Morse's model fits approximately 99 % of the data for the whole collection and for three subjects. They have further shown that a (one of the parameters in Morse's model) is time dependent. In another study based on the Ohio State University Library data, Coady [11] tested the Markovity of the circulation data. Test of the time independence of the data, in his study, indicated a time dependence in the data sets. Further, tests of the Markovity of the circulation, where second order Markovity is assumed and 1st order Markovity is tested, revealed all the fourteen data sets to be without Markovity.

Bookstein [2] suggested a single theoretical distribution to describe different bibliometric distributions. His function is

$$f(x) = \frac{k}{x^{\alpha}}$$

$$x = 1, 2, 3, 4, \ldots\ldots$$
$$\alpha, k > 0$$

This function can be used to describe Zipf's, Lotka's and Bradford's law. Ravichandra Rao [20] observed that even this function can be used to describe circulation data; f(x), the number of documents which were borrowed x times is inversely proportional to x^{α}. This model however fits only for the medium size library and for data of homogenous type. For a large data, expecially of highly skewed type, this model does not fit the data very well. In such cases, one may have to try other distributions.

3.1. Log-normal Distribution

The log-normal distribution distinguishes itself from the generalized model in a sense that it is derived from the law of proportionate effect (- the use of a document at any stage is a random proportion of its use in the immediate previous stage -) i.e. if the document is used x_j times at the jth interval, x_{j+1} (the use at the $(j+1)^{th}$ interval) is given by :

$$x_{j+1} - x_j = \varepsilon_j x_j$$

The ε_j's are mutually independent and further they are all independent of all x_j's. After a sequence of n proportionate 'random shocks' use of an individual document will be :

$$x_n = x_0 (1 + \varepsilon_1)(1 + \varepsilon_2)(1 + \varepsilon_3) \ldots\ldots (1 + \varepsilon_n)$$

where x_0 is the initial use at some arbitrary origin of time.

By taking the logarithms, we have

$$\log x_n = \log x_0 + \log (1 + \varepsilon_1) + \log (1 + \varepsilon_2) + \ldots + \log (1 + \varepsilon_n) .$$

For sufficiently large n, from the central limit theorem, it can be shown that $\log x_n$ is normally distributed, as each of the terms on the right hand side of the above equation is an independent random variable.
However, in a real situation, for large n, especially for the heterogenous and highly skewed data the log-normal distribution does not fit the data very well.

3.2. Negative Binomial Distribution

O'Neil [17], Tague and Farradane [24], Ravichandra Rao [20], Burrell [5,6], Brownsey and Burrell [3] and many others have argued that the negative binomial distribution has its applications to library and information work.
The negative binomial distribution is concerned with the same type of 'drawing' as the binomial and geometric distributions. That is, independent 'drawing' with a constant probability at each trial for two kinds of outcomes. The negative binomial distribution gives the probability of the occurrence of kth success at the yth trial ($k \leqslant y$). The density function of the negative binomial distribution is given by :

$$p(x) = \frac{(k+x-1)!}{(k-1)!x!} p^k (1-p)^k$$

$$x = 0,1,2,3,4, \ldots\ldots$$

In our context, p(x) gives the probability that a document circulated x times in time t; the mean and the variance of the distribution are kq/p and kq/p^2 respectively. p(x) can be derived in several ways; for instance, one can derive it under the assumptions of success-breeds-success phenomenon, using the principles of differential equations [20]; it can also be derived as a compound distribution of Gamma and Poisson distributions, under certain assumptions [6]; the negative binomial distribution is a special case of a generalized distribution based on the principles of the Multiple Urn Model [25].

3.2.1. Success-breeds-success Phenomenon

Lex x be a random variable which represents 'the number of times a document has circulated"; the domain of X is the non-negative integers. Let p(x,t) be the probability that the document is circulated x times in time t. Also let us assume that during the time interval (t, t+dt), a document which is circulated x times will be circulated again with the probability f(x,t) dt; dt is so small that a document can be circulated only once during the time interval. In this presentation p(x) is used to mean one or all of the parameters of the probability density function as a function of t.
Let us further assume that the probability that a document which has circulated x times will circulate again during the time interval (t, t+dt) is independent of time and increases linearly with the number of times it has already been borrowed. This is in accordance with the success-breeds-success phenomenon.
That is,

$$f(x,t) = a + bx$$

$$a, b > 0$$
$$x = 0, 1, 2, 3, \ldots\ldots$$

In such a case p(x) follows a negative binomial distribution.

3.2.2. Compound Distribution of Gamma and Poisson

As mentioned above, p(x) can also be derived as a compound distribution of Gamma and Poisson distributions, under the following assumptions :

i) Only a proportion (say, α) of the collection is active;

ii) Each book in the active collection is borrowed at random, according to a Poisson process, with borrowing rate λ ;

iii) λ follows a gamma distribution of the form

$$f(\lambda) = \frac{\beta^{-\nu} \lambda^{\nu-1} \exp(-\lambda \beta^{-1})}{\Gamma(\nu)},$$

$$\lambda > 0$$

From (i), (ii) and (iii), we thus have

$$p(x) = \int_0^\infty \frac{e^{-\lambda} \lambda^x}{x!} f(\lambda)\, d\lambda.$$

Using some simplifications, it can easily be shown that p(x) is the density function of the negative binomial distribution.

For a large set of data, it has been observed that the negative binomial distribution fits the circulation data very well. Therefore, the phenomenon of distribution of transactions is a manifestation of the success-breeds-success phenomenon.

3.3. Other Related Distributions

There are many who have suggested the Poisson distribution, the geometric distribution, ect. to describe circulation data. For instance, in one of his publications in 1980, Burrell [4] pointed out that the distribution of transactions closely approximates a geometric distribution. Morse [16], Kent and others [14] and Cane [8] also suggested a geometric distribution to describe circulation data.

If one tries the rank frequency approach, for small collections, it may be observed that the Bradford distribution also gives rise to fairly satisfying results; in fact this approach will be very useful, if our aim is simply to determine the core collection. On the other hand, analysis of size distributions helps us in computing various probabilities.

Apart form the above mentioned distributions, it has been observed in many cases that an empirical distribution of circulation can be described by the :

i) Waring distribution

ii) Borrel-Tanner distribution

iii) Cumulative Advantage distribution

iv) Truncated versions of Poisson and negative binomial distributions

v) Weibull distribution

vi) Bradford distribution (- rank distribution -).

The author has, however, observed in his research [20] that none of these distributions can be used to describe the highly skewed, long-tailed and heterogenous data. In such cases, it is suggested here to use the principle of maximum entropy to derive a suitable distribution function.

4. AN ENTROPY APPROACH

Let $P = \{p_1, p_2, \ldots\ldots, p_n\}$ be a finite probability distribution; i.e., n real numbers satisfying

$$p_x \geq 0 \text{ for } x = 1, 2, 3, \ldots\ldots, n$$

and

$$\sum_x p_x = 1 .$$

The number p_x may represent the probability of the xth outcome of a 'chance experiment' or the probability of the xth possible value taken by a random variable. For example, p_x may refer to the probability that a document curculated x times.
The entropy as defined by Shannon [22] attached to the probability distribution P is the number :

$$H_n(P) = H_n(p_1, p_2, ..., p_n) = - \sum_{x=1}^{n} p_x \ln p_x$$

In computing $H_n(P)$, we put $0.\ln 0 = 0$ to ensure the continuity of the function $-x\ln x$ at the origin. For $n \geqslant 2$, $H_n(P)$ is a function defined on the set of probability distributions satisfying the above conditions.
From its definition, the entropy is a measure of uncertainty. The uncertainty is maximum when the outcomes are equally likely. The uniform distribution maximizes the entropy; so it contains the largest amount of uncertainty.
Let X be a random variable with $E(X) = \mu$ and $x = 1, 2, 3, ... n$. To compute P_x ($\geqslant 0$, for every x), let us consider the following two simple constraints :

$$E(X) = \sum_x x\ p_x$$

$$\sum_x p_x = 1 .$$

Applying now the principle of maximum entropy, we choose the most uncertain probability distribution (i.e., the probability distribution that maximizes the entropy). That is, maximize

$$H_n(P) = - \sum p_x \ln p_x$$

subject to the above two constraints.
Thus considering the relevant Lagrange functions (to maximize $H_n(P)$), we have :

$$L = H_n(P) - \alpha \left[\sum_x p_x - 1 \right] - \beta \left[\sum_x x p_x - E(x) \right]$$

where α and β are the Lagrange multipliers. Corresponding to the two constraints and putting the first order partial derivatives equal to zero, we get :

$$\frac{\partial L}{\partial p_x} = - \ln p_x - 1 - \alpha - \beta x = 0$$

$$(x = 1, 2, 3, ..., n)$$

$$\frac{\partial L}{\partial \alpha} = 1 - \sum_x p_x$$

$$\frac{\partial L}{\partial \beta} = E(X) - \sum_x x\, p_x$$

Thus the solution is

$$p_x = \frac{e^{-\beta_0 x}}{\Sigma e^{-\beta_0 x}}$$

$$x = 1, 2, 3, \ldots\ldots n$$

where β_0 is the solution of the exponential equation :

$$\sum_x (x - E(X)) \cdot e^{-(x - E(X))} = 0$$

In the above discussion, we have defined only two constraints; similarly, it can be extended for the case where there are more than two constraints. For example, a finite number of constraining relations may be of the type :

$$\sum_{x=1}^{n} x^r\, p(x) = \mu'_r \qquad \text{(A)}$$

$r = 1, 2, 3, 4$.

In order to have a consistent probability interpretation, we must require one of the constraints to be of the form

$$\sum_x p(x) = 1 \qquad \text{(B)}$$

The entropy is given by :

$$H_n(P) = -\sum_x p(x) \ln p(x) \qquad \text{(C)}$$

As has been said earlier, this measure introduced by Shannon gives us an unique unambiguous measure of the amount of uncertainty connected with a discrete probability distribution. It provides us with the best estimate for p(x) on the basis of the information available; i.e., p(x), which maximizes (C) subject to (A) and (B), is given by :

$$p(x) = \exp[-\lambda_0 - \Sigma \lambda_r x^r]$$

where λ_0 is the Lagrange multiplier associated with and determined by equation (B) and the λ_r's are associated with comparable multipliers associated with relations given in equation (A). Since equation (B) is only a normalization condition, it follows immediately that :

$$e^{-\lambda_0} = \sum_x \exp\{-\sum_r \lambda_r x^r\}$$

Interpreting the moments μ'_r as constant in time, the entropy approach yields an equilibrium distribution for p(x); i.e., the one which will hold for all time so long as the system is not disturbed.
In the absence of a suitable theoretical distribution, it is hoped here that an entropy approach is useful to analyze the circulation data.

5. MEASURING INEQUALITY

Any measure of inequality should embody a sensible notion of inequality. Perhaps the simplest definition is to state that there is inequality whenever a document is used more than any other documents. A majority of the measures are concerned only with relative inequality in the sense that a proportionate change in all transactions does not change the measure.
If measures are to be comparable accross countries or over time, they must have some desirable general properties.
They are :

i) Measures should be unit free
ii) Measures should be unaffected by the size of the library collection
iii) Measures should preferably be bounded .

5.1 Categorization of Measures of inequality

Circulation distributions are usually presented in two ways. The first is simply to plot the frequency of the distribution. On the horizontal axis is 'use' (number of times used) and on the vertical axis is the relative frequency of the documents. As with the typical frequency functions the area under the curve is equal to one. A typical frequency curve is shown in Figure 1.
The second way to represent the circulation distributions is to graph the Lorenz curve or the concentration curve. This is a special kind of cumulative frequency graph. Such a graph is used to portray the nature of non-uniformity in the distribution of inherently positive quantities like use, wealth, income, etc. The graph relates the proportion of total frequency of the variable, say, for example, library circulation use to various proportions of the relevant population (documents, users, etc.). The graph is a curve running through the points of cumulative relative/percentage frequency (say, circulation use) and cumulative percentage of items (say, documents). It is presented in a square with both the axes taking values from 0 to 1 (or 0 to 100). The more the concentration curve deviates from the straight line (diagonal line), the greater is the inequality. Figure 2 is an example of a typical concentration curve.
If every document is used an equal number of times, the Lorenz curve would be a straight line (00'). i.e., the first 10 % of the documents circulated 10 % of the total circulation; the first 20 % of the documents circulated 20 % of the total circulation, etc.
The other well-known graph to measure the circulatio distributions is Trueswell's Curve [27]; in his study, based on the "last circulation data", he observed that nearly 90 % of the documents had a last circulation date within the preceeding three months and 10 % of the documents within the past one month. His graph is similar to that of a cumulative frequency distribution. His curve shows the relation of the percentage of circulation(having the last circulation date within cumulative time period) to the specific time period. Turner [28] in his generalization of Trueswell's approach presented a method that he called an identifier method that could be used to develop a rule in many non-library-oriented areas where identification of low-use items is desirable.
Most of the inequality measures which are developed in the recent past are either based on Figure 1 or on Figure 2. The two graphs obviously contain te same information in the sense that one can be derived from the other. One can in fact go from the frequency function to the Lorenz curve by integrating (-summing up) over document and use; to go from Lorenz curve to frequency function, it is very difficult; we require in addition, data on various use levels (i.e. the values of the variable at different points such as for $x = 1, 2, 3, \ldots$).
The notion of inequality may easily be defined in terms of the Lorenz Curve. Equality is defined as equal shares of total use accuring to equal shares of documents. In Figure 2, any deviation from the diagonal line indicates the presence of inequality. In terms of frequency function, perfect equality implies a degenerate distribution concentrated at the point of mean use (- it may hardly be a very satisfying diagram).

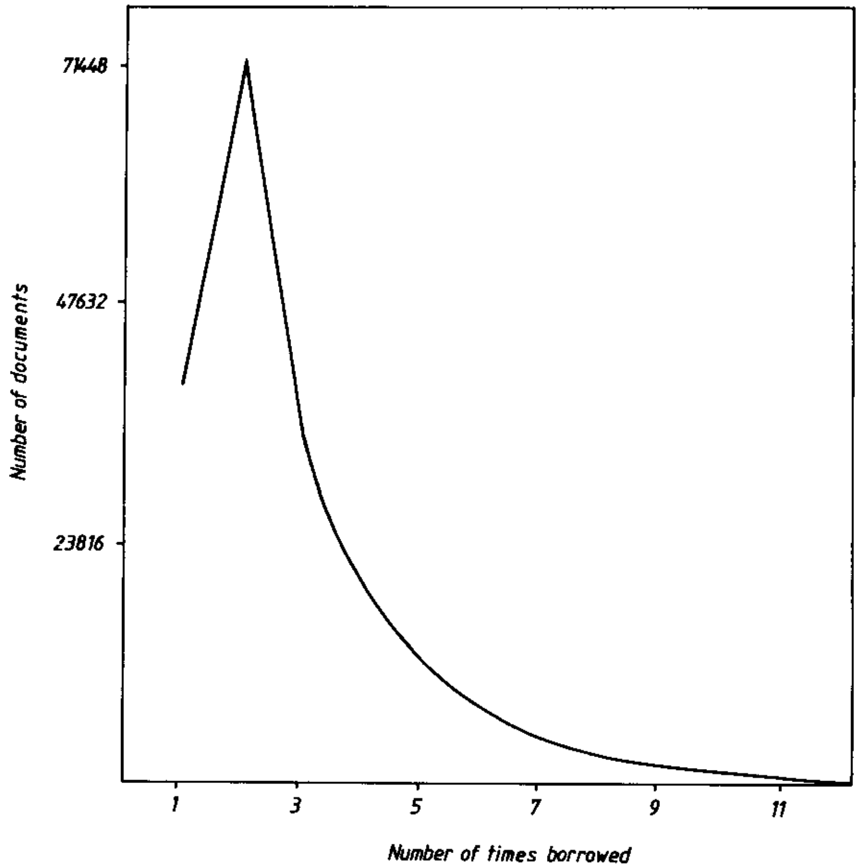

Figure 1 : A typical frequency curve for circulation distributions : distribution of the transactions of documents in the University of Alberta in 1975-76.

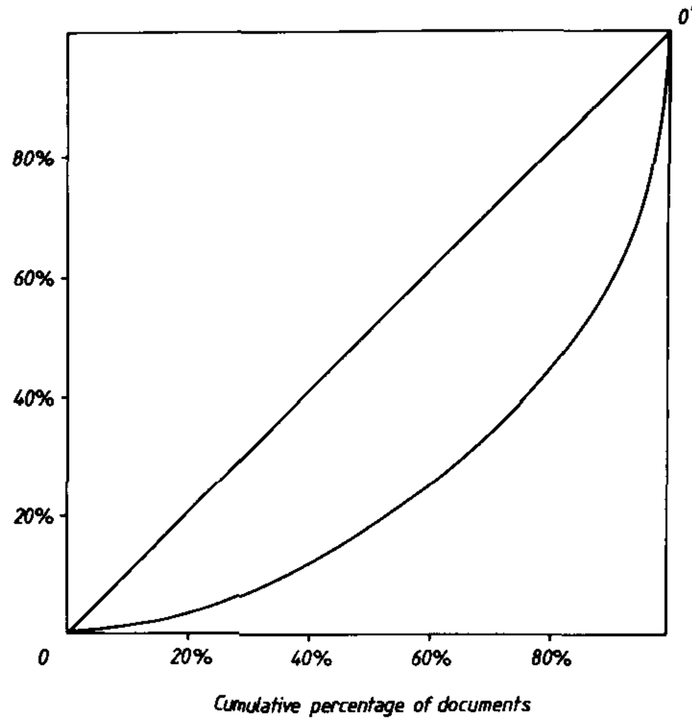

Figure 2 : A Typical concentration curve : The concentration of documents in the University of Waterloo in 1977-78.

However, if one is interested in developing or identifying a theory to describe the circulation distribution, then the frequency function will be very useful. For instance, if the circulation distribution conforms to a Lognormal or Poisson distribution, the parameters of the distribution then become of special interest. Thus, measures of inequality may mainly be grouped into those which describe the frequency function and Lorenz curve.

5.2 Frequency Function Measures

Coefficient of Variation

The variance of a frequency distribution measures the spread of observations and it can be used as a measure of inequality, It is, however, a function of the units in which the variable is measured; comparisons are, therefore, difficult. The coefficient of variation is defined as the standard deviation divided by the mean; it is unit free. It is a measure of relative dispersion.

Pearson Skew Coefficient

The frequency distribution for the circulation data is usually reverse-J shaped, highly skewed to the right and it has a long tail. For a right-skewed distribution, the mean is greater than the median which in turn is greater than the mode. For a left-skewed distribution, the inequalities are exactly opposite to that of the right-skewed distribution. Based on these inequalities, Pearson suggested a measure, it is given by :

$$\text{Skew Coefficient} = \frac{3(\text{mean} - \text{median})}{\text{Standard Deviation}}$$

This measure is unit free; the more skewed the distribution is, the higher the absolute value of this measure. The skew coefficient is not in itself a measure of inequality; regardless of the degree of skewness, a distribution with a very low variance is likely to have less inequality. Thus, the variance and skewness must be considered together to analyse the circulation data.

The α-measure

In the generalized model, as suggested by Bookstein [2] :

$$Y = \frac{K}{x^{\alpha}} \qquad \alpha, K > 0$$

α can be treated as a measure of inequality. For a given sample size (say, N) and for a large value of α, the proportion of highly productive documents (--frequently borrowed documents) will increase; the larger the parameter α, the greater is the gap (in use !) between the frequently and infrequently borrowed groups of documents. In this sense, α is considered as a measure of inequality in circulation distributions.

Variance of the logs of Use

The variance of the logarithm of use may also be used as a measure of inequality in circulation distribution. Let X be a random variable which represents a number of times a document circulated; let n be a sample size; then, log variance is given by :

$$\text{Var}(\log x) = \frac{1}{n} \sum_i (\log x_i - \overline{\log x})^2$$

where $\overline{\log x}$ is the mean of the logarithms of use. If the circulation distribution (i.e. the distribution of X) conforms to a log-normal distribution, the log variance is the maximum likelihood estimate of the variance of the distribution. The log variance is a relative measure and is not affected by proportional changes in total use; it is thus independent of the average use and it is unit free.

5.2 Lorenz Curve Measures

Gini's Index

The Gini coefficient is undoubtedly the most commonly used measure of the Lorenz curve. It is defined as the ratio of the area between the Lorenz curve and the diagonal to the total area under the diagonal. It is an increasing function of inequality and equals zero when the distribution is perfectly equal. For a given N, the Gini index is given by [13] :

$$g = 1 - \frac{1}{N} \sum_i f_i (h_i + h_{i-1}) \text{ , where } h_i = \sum_{k=1}^{i} kf_k \text{ .}$$

Upper and lower bounds of g are given by [13] :

$$\max g = \frac{N-1}{N}$$

$$\min g = 1 - \bar{x} + 2j - \frac{j(j+1)}{\bar{x}}$$

where j is the nearest integer less or equal of the mean. For the sake of convenience we may represent g as :

$$G = \frac{g - \min g}{\max g - \min g}$$

so that, G lies between 0 and 1.

Concentration Ratio

It is defined as [13] :

$$C = 1 - \frac{\sum_i h_i - A}{\bar{x}\ (\sum_i b_i - N)}$$

where $N = \sum_i f_i$, $b_i = \sum_{k=1}^{i} f_k$, $A = \sum_k k\ f_k$ and $h_i = \sum_{k=1}^{i} kf_k$.

It is also given by a formula called Lorenz coefficient.

$$L = 1 - \sum_i (p_i - p_{i-1})\ (q_i + q_{i+1})$$

p_i is the percentage of variable X (say, the number of documents in the ith class) and q_i is the percentage of Y (say, percentage of transactions of p_i documents in the ith class.)

Pratt's Measure

In 1977 Pratt [18] introduced a measure of class concentration; it is know as Pratt's measure. It can be described as follows :

Let us consider a fixed number n of categories or classes. Let us also have a fixed number of objects (say, for example, documents) which are distributed over these n classes (may be, according to some law). Then order the classes in decreasing order of the number of objects as class (i+1). Let the fraction of objects that are in class i be a_i (i = 1, 2, 3, ...n). Thus we have :

$$\sum_{i=1}^{n} a_i = 1$$

Let us define

$$q = \sum_{i=1}^{n} ia_i \ .$$

Pratt's measure is then defined as :

$$C = \frac{2\ [\frac{n+1}{2} - q]}{n - 1}$$

Egghe [12] derived various formulae for C for different distributions (such as, Zipf's distribution, Mandelbrot's distribution, Lotka's distribution, Negative Binomial Distribution.)

Egghe also discussed the relation of C with the 80/20 rule. The 80/20 rule is a symbolic name studying a concentration problem; in the context of circulation distribution, it means 80 % of the documents are hardly used and 20 % of the documents are mostly used. This has been observed by many scientists [7, 12].

5.3 Other Measures

Theil [26] has proposed an inequality measure based on information theory. It is given by :

$$T = \log n - \sum_i y_i \log (1/y_i)$$

where y_i is the share of aggregate use of the ith document and n is the total number of documents; the upper bound of the measure is log n: the larger n is the greater is the amount of possible inequality. T is, however, a function of n. Comparison of measure for two different populations is thus difficult. This can, however, be normalized by dividing it by log n.

The entropy $H_n(P)$, as defined by Shannon, can also be used as a measure of concentration of transactions over documents [19]. It is discussed below :

Let n be the number of documents. Let p_1, p_2, p_3, ... p_n be the probabilities (that a document is used) associated with 1st, 2nd ... and nth document respectively. By definition, we have $0 \leqslant p_i$ and $\sum_i p_i = 1$.

The entropy of the circulation distribution (whose probabilities are given by these p_i's) is :

$$H(x) = - \sum_i p_i \log p_i \quad .$$

H(x) can be regarded as an inverse measure of concentration of transactions over documents. If the probability that one particular book is used is 1 and the probability of all the other books used is zero (i.e. p_i = 1 for some i and p_j = 0 for every j = i), then we have :

$$H(x) = 0.$$

That is, the minimum entropy value and the maximum degree of concentration. If the probability that a book is used is 1/n for every book, we have :

$$H(x) = \log n$$

which is the maximum value of the entropy for a given n and also the minimum degree of concentration of transactions for a given n.

When the number of books increases, while all p_i's are the same, the entropy log n increases too. This is in accordance with the decreasing degree of concentration. When there are 100 books with equal probability of intersections ($p_1 = p_2 = \ldots\ldots = p_{100}$) which account for total transactions, there is much less concentration than the library consists of only 2 books with equal probability of transactions ($p_1 = p_2$) and certainly less concentration than in the case of a library which consists of only one book. In the last case, we have H(x) = 0 and a maximum concentration of transactions. Thus, the entropy is an appropriate inverse measure of concentration of transactions over documents.

The higher the entropy, the lower the degree of concentration of transactions over documents and vice versa.

Table 1 : A few inequality measures

Sl. No.	Inequality Measures		Minimum Value	Maximum Value
1	Coefficient of Variation (CV)	$\sigma / \overline{x}$	0	$\sqrt{n-1}$
2	Pearsonian Skew Coefficient (SK)	$\frac{3(\text{Mean-Median})}{\text{standard deviation}}$	$-\infty$	$+\infty$
3	α -measure	α in $y = k/x$	$\alpha > 0$	
4	Variance of Logarithms (VL)	$\frac{1}{n} \Sigma (\log x_i - \overline{\log x})^2$	> 0	
5	Gini's Index (g)	$1 - \frac{1}{N} \Sigma f_i (h_i + h_{i-1})$	*	$1 - 1/N$
6	Concentration Ratio (C)	$1 - \frac{\sum_i h_i - A}{\overline{x} (\sum_i b_i - N)}$	0	1
7	Lorenz Coefficient (L)	$1 - \sum_i (p_i - p_{i-1})(q_i + q_{i+1})$	0	1
8	Pratt's Measure (C)	$\frac{2 [\frac{n+1}{2} - q]}{n - 1}$		
9	Theil's Measure (T)	$\log n - \sum_i y_i \log 1/y_i$	0	$\log n$

$$* = 1 - \overline{x} - 2j - \frac{j (j+1)}{\overline{x}}$$

6 CONCLUDING REMARKS

Studies on circulation distributions have so far generated many theoretical papers trying to consolidate them or deduce them. Many have identified different theoretical distributions for different sets of data. Most of these studies reveal that circulation distribution are :

i) reverse-J shaped
ii) very positively skewed, and
iii) have a long tail.

It is further observed by many that approximately 80 to 90 % of the documents which are borrowed infrequently contribute roughly 40 % to 75 % of the total transactions. This shows that circulation distributions do not strictly conform to the 80/20 rule.

Among the various theoretical distributions identified, it has been observed that the negative binomial distribution is a good approximation to describe the circulation distribution; thus the phenomenon of distribution is a manifestation of the success-breeds-success phenomenon.
Instead of studying a size-frequency distribution of the circulation data, if one attempts to study the rank-frequency distribution, the Bradford distribution is likely to be a good approximation.
We often find it very difficult to identify a theoretical model to a highly positively skewed, long-tailed, heterogeneous and to a large data. In such cases either we may have to adopt either a graphical method or to use certain inequality measures or derive a distribution function based on a principle of maximum entropy.
Various inequality measures were also used and suggested to analyze the circulation data. Gini's index, Lorenz ratio, α -parameter (of the generalized bibliometric distribution), Pratt's measure, Pearson's skew coefficient are found to be the most useful to analyse the data. The entropy (-- a measure of uncertainty) of a distribution can equally be used to interpret (and accordingly analyse) the data. The entropy in an inverse measure of concentration of transactions over documents. The higher the entropy, the lower the degree of concentration of transactions over documents and vice versa.

6.1 Applications

We usually maintain and publish statistics regarding the collection size (by type of documents) and the number of users. These are hardly useful to measure the effectiveness of library services. On the other hand, probability distributions and inequality measures are very useful in identifying the active collection (-- the one which is frequently borrowed/ used) size. Especially inequality measures are very useful to compare the effectiveness of various library services in several libraries as well as within a library in over a period. However, these measures are static in nature and none of them measures the dynamic aspects of libraries.
Among the various probability distributions suggested in the text, negative binomial distribution is likely to be a most suitable model to describe the circulation data. However, for a heterogeneous and for a large sample, it is unlikely to fit the data; in such cases, one may attempt to identify a theoretical distribution based on the definition of entropy.
Among the various inequality measures suggested in the text, Gini's index and the concentration ratio will be most useful to measure the inequality.
Collection and analyses of circulation distribution thus help the librarians to improve their collection development as well as circulation policies.
Analysis of circulation data together with the data pertaining to the cost of retrieving and shelving of documents will facilitate decision making in regard to different types of storage to be adopted. For example, frequently borrowed documents can be shelved as primary storage in the main building of the library and the infrequently borrowed documents can be shelved as secondary storage, adopting either a compact storage in the campus or in off-campus location.
Further identification of frequently and infrequently borrowed documents will facilitate proper organization of transaction records relating to documents (and also users) in circulation files in an automated circulation system. Circulation studies will also enable us to determine the "core collection", not from the point of view of subjects, but from the point of view of 'book-use'. Further such studies will help us in taking decisions regarding purchase of multiple copies as well as purchase of lost or mutulated books.

REFERENCES

[1] Beheshti, J. and Tague, J.M., Morse's model of book use revisited. Journal of the American Society for Information Science, 35 (1984) p. 259-267.

[2] Bookstein, A., Bibliometric distributions, Library Quarterly, 46 (4) (1976) p. 416-423.

[3] Brownsey, K.W.R. and Burrell, Q.L., Library Circulation Distributions : Some observations on the PLR Sample, Journal of Documentation, 42 (1) (1986) p. 22-45.

[4] Burrell, Q.L., A Simple Stochastic model for library loans", Journal of Documentation, 36 (1980) p. 115-132.

[5] Burrell, Q.L., Alternative models for library circulation data, Journal of Documentation, 38 (1) (1982) p. 1-13.

[6] Burrell, Q.L. and Cane, V.R., The analysis of library data, Journal of the Royal Statistical Society, Series A, Part 4 (1982) p. 439-471.

[7] Burrell, Q.L., The 80/20 rule : library lore or statistical law, Journal of Documentation, 41 (1) (1985) p. 24-35.

[8] Cane, V.R., Making room for books, Library Review, 28 (1979) p. 148-150.

[9] Chen, C.-C., The use patterns of physics journals in a large academic research library, Journal of the American Society for Information Science, 23 (4) (1972) p. 254-270.

[10] Chen, C.-C., Applications of operations research methods to libraries : a case study of the use of monographs in the Francis A. Countway Library, (MIT Press, Cambridge, Mass., 1976).

[11] Coady, R.P., Testing for Markov-Chain properties in the circulation of social science monographs, Behavioural and Social Sciences Librarian, 3 (4), p. 53-68.

[12] Egghe, L., Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule, Journal of the American Society for Information Science, 38 (4) (1987) p. 288-297.

[13] Hustopecky, J. and Vlachy, J., Identifying a set of inequality measures for science studies, Scientometrics, 1 (1) (1978), p. 85-98.

[14] Kent, A and others, (eds.), Use of library materials : the University of Pittsburgh study, (New York, Dekker, 1979).

[15] Morse, P.M., Library effectiveness : a system approach, (MIT Press, Cambridge, Mass., 1968).

[16] Morse, P.M., Measures of library effectiveness, Library Quarterly, 42 (1) (1972) p. 15-30.

[17] O'Neill, E.T., Monographs circulation patterns in academic libraries, Research memorandum 2, School of Information and Library Studies, (State University of New York at Buffalo, Buffalo, 1973).

[18] Pratt, A.D., A measure of class concentration in bibliometrics, Journal of the American Society for Information Science, 28 (5) (1977) p. 285-292.

[19] Ravichandra Rao, I.K., Entropy of probability distribution of transactions/ users : a measure of concentration of transaction/users over documents, Library Science with a slant to Documentation (1980).

[20] Ravichandra Rao, I.K., Document and User distribution in transaction records of Canadian University Libraries, (Ph. D. Thesis, Faculty of Graduate Studies, School of Library and Information Science, the University of Western Ontario, London, Ontario, Canada 1981).

[21] Rouse, W.B., Tutorial : Mathematical modelling of library systems, Journal of the American Society for Information Science, 29 (1978) p. 181-192.

[22] Shannon, C.E., A mathematical theory of communication, Bell System Tech. J. 27 (1948) p. 379-423, 623-656.

[23] Stieg, L., A technique for evaluating the college library book collection, Library Quarterly, 13 (1943) p. 34-44.

[24] Tague, J.M. and Farradane, J., Estimation and reliability of retrieval effectiveness measures, Information Processing and Management, 14 (1978) p. 1-16.

[25] Tague, J.M., Success-breeds-success phenomenon and bibliometric processes, Journal of the American Society for Information Science, 32(4) (1981) p. 280-286.

[26] Theil, H., Economic and Information Theory, (North Holland Publishing Co., Amsterdam, 1969).

[27] Trueswell, R.W., User circulation Statistician vs. Size of Holdings at Three Academic Libraries, College and Research Libraries, 30 (1969) p. 204-213.

[28] Turner, S.J., The identifier method of measuring use as applied to modelling the circulation use of books for a university library, Journal of the American Society for Information Science, 28 (2) (1977) p. 96-100.