Urquhart's Law: Probability and the Management of Scientific and Technical Journal Collections

Part 1. The Law's Initial Formulation and Statistical Bases

Stephen J. Bensman

ABSTRACT. The topic of this paper is a law formulated by Donald J. Urquhart on the use of scientific and technical (sci/tech) journals through interlibrary loan and central document delivery. The paper will be published in three parts. Part 1 discusses the genesis of the law and the probabilistic theory on which it is based. A primary aim of this part is to teach librarians about the probability distributions underlying library use and how to identify these distributions through simple mathematical and graphical techniques.

KEYWORDS. Donald J. Urquhart, scientific journals, interlibrary loan, document delivery, library use, probability, Poisson Process

---

Stephen J. Bensman, MLS, PhD (History), is Technical Services Librarian, LSU Libraries, Louisiana State University, Baton Rouge, LA 70803 (E-mail: notsjb@lsu.edu).

Science & Technology Libraries, Vol. 26(1) 2005
Available online at http://www.haworthpress.com/web/STL

doi:10.1300/J122v26n01_04

INTRODUCTION

This paper analyzes Urquhart's Law of Supralibrary Use. Supralibrary use is defined as the use by a given library's patrons of materials supplied from outside the library through either interlibrary loan or central document delivery. It is contrasted to intralibrary use, which is the use of a library's own materials by its own patrons. Stated in its simplest form, Urquhart's Law specifies that the supralibrary use of scientific and technical (sci/tech) journals is positively correlated with the number of libraries holding these journals in a system and therefore is a measure of their aggregate use within the library system, including their intralibrary use at the individual libraries of the system. The law was formulated by Donald J. Urquhart, who established the National Lending Library for Science and Technology (NLL) that later was merged into the British Library Lending Division (BLLD), now called the British Library Document Supply Centre (BLDSC). Urquhart was the first librarian to investigate scientifically the nature of sci/tech journal use and apply probability to it.

The paper will be published in three parts in three separate issues of this journal. Part 1 discusses the initial formulation of the law and its statistical bases; Part 2 will analyze how Urquhart applied probability to create and manage a central document delivery collection; and Part 3 will be dedicated to the implications of the law for all the libraries of a given library system. Each part will have an introduction, which will lay out not only what it will discuss but will also summarize the discussions and conclusions of the preceding parts. The purpose of these introductions is threefold: (1) to provide the reader with a roadmap of the overall structure and logic of the paper; (2) to enable, as much as possible, each part to be read independently from the others; and (3) to highlight the important points for the reader.

Part 1 shows Urquhart's Law as a natural outgrowth of the Law of Scattering formulated in the early 1930s by S. C. Bradford, director of the Science Museum Library (SML) in London. Bradford's Law describes the distribution of articles on a given scientific subject across journal titles. In doing so, it demonstrates the inability of sci/tech libraries to hold all the titles necessary to their patrons, proving the need of such libraries for document delivery support from either other sci/tech libraries at their level or a comprehensive central scientific library. Bradford aspired to convert the SML into a central document delivery library, and Urquhart fulfilled these aspirations by establishing the NLL. To prepare for the establishment of this library, Urquhart con-

ducted a study of the loans made by the SML to outside organizations in 1956. This study laid the bases for his law of supralibrary use.

The primary focus of Part 1 is the statistical analysis of Urquhart's data on the external loans made by the SML in 1956. These data are utilized here for the didactic purpose of teaching librarians about the probability distributions underlying library use as well as how to identify these distributions through simple mathematical calculations and graphical techniques. Part 1 analyzes the set structure of sci/tech journals arising from Bradford's Law and the stochastic processes underlying these journals' use. Using Lexian analysis, it demonstrates how these stochastic processes are partly a function of this set structure. The binomial and the Poisson processes are discussed, and the greater applicability of the Poisson process to library use is demonstrated. As a component of this, Part 1 explains the crucial importance for library collection management of Bortkiewicz's Law of Small Numbers, which establishes the theoretical basis for the stability and permanence of the low-use classes. There are presented both the simple Poisson distribution and compound Poisson distributions, the key one of which is the negative binomial distribution. Compound Poisson distributions are proven to be the best model of library use due to their ability to capture all the stochastic processes operative in this use. Part 1 concludes with an explanation of the various indices of dispersion, which can be utilized to identify the stochastic processes and resulting probability distributions operative in library use.

GENESIS OF URQUHART'S LAW

Two of the most important libraries in the historical development of library and information science were the Science Museum Library (SML) in South Kensington, London, and the National Lending Library for Science and Technology (NLL), a direct predecessor of the present-day British Library Document Supply Centre (BLDSC), in Boston Spa, Yorkshire. The first was the prototype for the second. Each of these libraries is closely associated with an important bibliometric law formulated by the person serving as its head.

The SML is linked to Bradford's Law of Scattering. This law was derived by S. C. Bradford (1934), who was the chief librarian there for the period 1925-1938. It deals with the distribution of articles on a given scientific subject across journals, positing that such articles concentrate in a small nucleus of journals specifically devoted to the subject, and

then scatter in ever decreasing numbers across other groups or zones of journals. As a result of this phenomenon, Bradford came to the conclusion that special libraries could never collect the complete literature on their subject, and in his classic book *Documentation*, Bradford (1953, 102-122) advocated the establishment of a national central library for science and technology, one of whose major functions would be "external lending" or "the lending of books to research workers and students through the medium of approved institutions and lending agencies" (p. 117). Bradford assiduously worked to convert the SML into such a library. Prior to his tenure as chief librarian, this library had served the Science Museum and the neighboring Imperial College of Science and Technology. Upon taking charge of it, Bradford strove to develop its holdings into a comprehensive collection of the world's scientific and technical (sci/tech) literature, making these resources available to scientists nationwide through approved institutions with which they were associated.

Bradford's vision of a national central library of science and technology was implemented by Donald J. Urquhart. The two men were similar in that they both had science doctorates. Urquhart began his library career at the SML, obtaining a job at this institution in 1938 at the time Bradford retired as its head. His service there was interrupted by World War II, but after the war he returned to the SML, where he stayed until 1948 when he moved to the Department of Scientific and Industrial Research. As its name implies, the Department of Scientific and Industrial Research, or DSIR, was the agency through which the British government supported scientific research and ensured that industry utilized new scientific findings.

Urquhart rose to prominence at the Royal Society Scientific Information Conference of 1948, contributing three papers to this conference. In one of these papers entitled "The Organization of the Distribution of Scientific and Technical Information" Urquhart (1948, 526) proposed a national library that would have "a specific responsibility for organizing the library service for scientific and technical literature." At the end of 1956, DSIR formed a Lending Library Unit headed by Urquhart with a staff of four, of whom the most important was Miss R. M. Bunn. This unit planned the creation of the National Lending Library for Science and Technology (NLL). It was in the establishment and operation of this library that Urquhart developed what will be termed in this paper "Urquhart's Law of Supralibrary Use."

THE ANALYSIS OF 1956 SCIENCE MUSEUM LIBRARY (SML) EXTERNAL LOANS

The first step in the creation of the NLL was an analysis of the loans of sci/tech journals made by the SML to outside organizations during 1956. This analysis was the first large-scale scientific study of journal use, and it must be emphasized that it was a study of supralibrary use. Supralibrary use may be defined as the use by patrons of a given library of materials not owned by that library but supplied from the outside through either some form of centralized document delivery or from other libraries by means of interlibrary loan. It is to be contrasted with intralibrary use, which is the use by the patrons of a given library of materials held by that library. Therefore, the analysis of the SML external loans of sci/tech journals in 1956 was a study of United Kingdom (UK) supralibrary use of these materials in that year.

Urquhart (1959) reported on this analysis in a paper to the International Conference on Scientific Information held in Washington, DC, in November,1958. Another report on the SML analysis was published by Urquhart and Bunn (1959) the following year. The key findings of the study of the 1956 SML external loans of sci/tech journals were two. One was that the distribution of supralibrary use of sci/tech journals is highly skewed. This distributional finding was summarized by Urquhart and Bunn (1959, 21) thus:

> In general it seems that a small percentage of the current serial titles account for a large percentage of the use of all serials. In the Science Museum in 1956 about 350 titles accounted for 50 per cent. of the total use of serials, and about 1200 titles for 80 percent. of the total use. This, despite the fact that in 1956 the Science Museum Library contained 9120 current serials, and possibly an equal number of dead ones. In general terms it appeared that, excluding less than 2000 titles, the total national interlibrary loan use could be satisfied by one loan copy.

By use of this information, it was possible to approximate the shape of the distribution of journals by number of 1956 SML external loans as well as to estimate its key statistical characteristics. These goals were accomplished in the following manner. First, the loan classes from 1 to 382 loans and the number or frequency (f) of titles in these classes were taken from Table VII in Urquhart (1959, 291). The total number of titles actually loaned was 5,632. From the above statement that the SML had

9,120 current serials and an equal number of dead ones, the total SML serial holdings were estimated to be 18,000 titles, and the number of journals in the zero class was calculated by subtracting the number of serials loaned, or 5,632, from 18,000 to arrive at 12,368 titles.

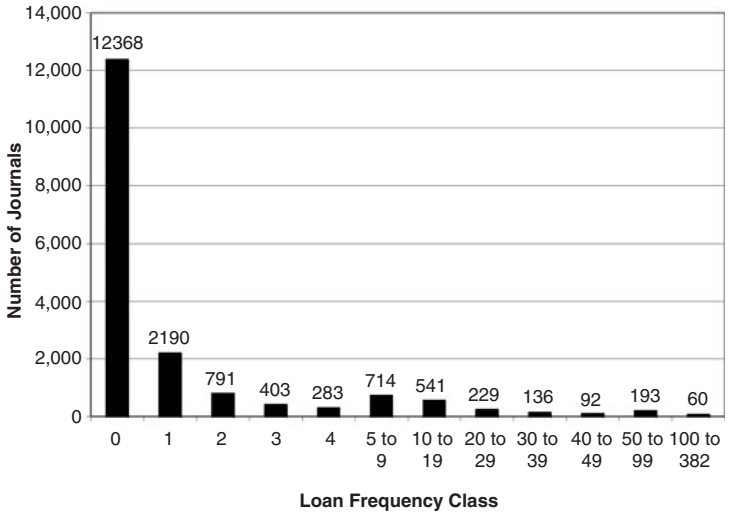
These results are given in Columns 1-2 of Table 1 and graphed by the bar chart in Figure 1.

Estimation of the key statistical characteristics of the distribution began by approximating the average or mean (m) number of loans per title in each class. For classes 0 through 4, the class mean was a given. The estimates of the mean loans per title for the other classes were based on a number of complex factors, including the loan information on 391 titles presented in the appendix to Urquhart and Bunn (1959, 25-37). Column 4 of Table 1 gives the final estimates of the class mean loans per title (m). These estimates were adjusted so that class mean loans per title (m) in Column 4 multiplied by the frequency of titles per class (f) in

TABLE 1. Estimated Observed Frequency Distribution of Scientific Journals over 1956 Science Museum Library (SML) External Loan Classes

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
Loan Class	No. Titles in Class [f]	Loans per Class [x]	Class Mean Loans per Title [m]	Deviation Set Mean from Class Mean [m − M]	(m − M)^2	f*(m − M)^2
0	12,368	0	0.00	−2.96	8.74	108,103.29
1	2,190	2,190	1.00	−1.96	3.83	8,382.61
2	791	1,582	2.00	−0.96	0.91	723.60
3	403	1,209	3.00	0.04	0.00	0.76
4	283	1,132	4.00	1.04	1.09	308.19
5 to 9	714	5,002	7.01	4.05	16.40	11,708.64
10 to 19	541	7,732	14.29	11.34	128.50	69,517.87
20 to 29	229	5,284	23.07	20.12	404.68	92,670.69
30 to 39	136	4,446	32.69	29.74	884.35	120,272.06
40 to 49	92	3,773	41.01	38.05	1,447.94	133,210.64
50 to 99	193	12,386	64.17	61.22	3,747.67	723,300.74
100 to 382	60	8,480	141.33	138.38	19,148.28	1,148,896.80
SUM	**18,000**	**53,216**				**2,417,095.88**
Set Mean Loans per Title (M) = SUM(x)/SUM(f) = 53,216/18,000 =						**2.96**
Variance (VAR) = SUM(f*(m − M)^2/((SUM(f) − 1)) = 2,417,095.88/(18,000 − 1) =						**134.29**
Standard Deviation (STDEV) = SQRT(VAR) = SQRT(134.29) =						**11.59**

FIGURE 1. Frequency Distribution of Scientific Journals by Urquhart's 1956 Science Museum Library (SML) External Loan Classes



Column 2 yielded loans per class (x) in Column 3 that in turn summed to the total number of reported loans of 53,216.

The summary of the key distributional findings by Urquhart and Bunn above suggests two methods of aggregating the frequency distribution of sci/tech journals across 1956 SML external loans into broader loan classes. These methods are shown in Table 2. The first method is to aggregate the distribution into two loan classes: low (0 to 9 loans) and high (100 to 382 loans). This method of aggregation played a key role in Urquhart's management of the NLL journal collection. Inspection of the high loan class reveals that it contained 1,251 titles that comprised 6.95% of the titles and accounted for 79.11% of the loans. This accords well with the statement by Urquhart (1959, 293) in his 1958 conference report that "about 1,250 serials (or less than 10% of those available if the non-current serials are included) are sufficient to meet 80% of the demand for serial literature." The second method is to aggregate the journals into three loan classes: low (0 to 9 loans), high (10 to 39 loans), and super high (40 to 382 loans). It can be seen that the super high class encompassed 345 titles or 1.92% that accounted for 46.30% of the loans. This comes close to the above statement in Urquhart and Bunn that "about 350 titles accounted for 50 per cent of the total use of seri-

TABLE 2. Two Methods of Aggregating 1956 Science Museum Library (SML) External Loan Classes

1. Two classes					
Loan Class	**No. Titles in Class**	**Mean Loans per Title**	**Loans per Class**	**% Titles per Class**	**% Loans per Class**
Low (0 to 9)	16,749	0.66	11,115	93.05%	20.89%
High (10 to 382)	1,251	33.65	42,101	6.95%	79.11%
SUM	**18,000**		**53,216**	**100.00%**	**100.00%**
2. Three Classes					
Loan Class	**No. Titles in Class**	**Mean Loans per Title**	**Loans per Class**	**% Titles per Class**	**% Loans per Class**
Low (0 to 9)	16,749	0.66	11,115	93.05%	20.89%
High (10 to 39)	906	19.27	17,462	5.03%	32.81%
Super High (40-382)	345	71.42	24,638	1.92%	46.30%
SUM	**18,000**		**53,216**	**100.00%**	**100.00%**

als." From these facts it can be seen that the above estimate of the frequency distribution of sci/tech journals across the 1956 SML external loans is a good approximation of the one that was actually observed.

The above approximation can now be utilized to calculate certain key statistical characteristics of the frequency distribution of sci/tech journals across 1956 SML external loans. In making such calculations, Excel spreadsheet notation is utilized in this paper. The first characteristic is the arithmetic mean or average loans per title. AVERAGE is the Excel function for arithmetic mean. This is a measure of central tendency, and for the entire set of SML journals it is derived by dividing the total number of journals into the total number of external loans. In Table 1, the set mean is designated by M to distinguish it from the class means m. Using the terminology of Table 1, the calculation of M is the following:

$$\text{AVERAGE (M)} = \text{SUM(x)/SUM(f)} = 53{,}216/18{,}000 = 2.96$$

The next two statistical characteristics are both measures of the dispersion of the external loans of individual titles around the set mean M. Of these, the basic measure is sample variance. Variance is a measure of the variability or dispersion of the values of a dataset found by averaging the squared deviations about the mean. Table 1 demonstrates a shorthand

method of calculating variance that is of great utility in library research. With this method, one first groups the observations into classes as is done in Columns 1-2. Then, as was done in Column 4, one estimates either the midpoint or mean for each of these classes–in this case, m. The next steps are to subtract the set mean M from each class mean m (Column 5), square these remainders (Column 6), and multiply the squared remainders by the number or frequency of observations f in each class (Column 7). These products are then added, and the resulting sum is then divided by the sum of the observations f. For technical reasons, it is best to subtract 1 from the sum of the observations. The result is the variance. In the Table 1 terms, the calculation is the following:

VAR = SUM(f*(m − M^ 2)/(SUM(f) − 1) = 2,417,095.88/(18,000 − 1) = 134.29

The other measure of dispersion is the sample standard deviation, which is found by taking the square root of the sample variance thus:

STDEV = SQRT(VAR) = SQRT(134.29) = 11.59

It is to be noted that the set variance is much greater than the set mean and that the bulk of the variance–82.97%–derives from the titles in the super high loan class.

In his report to the 1958 scientific information conference, Urquhart (1959, 289-291) related the distribution of sci/tech journals by supralibrary use to the number of library holdings of these journals. The result comprised the second key finding of the analysis of 1956 SML external loans. To do this, he first listed in descending rank order the top 10 journals by 1956 SML external loans and their corresponding library holdings as given by the *British Union Catalogue of Periodicals* (BUCOP). One is struck by the prestigious nature of most of these top 10 journals. The highest one was the *Proceedings of the Royal Society of London (Series A)* with 382 external loans, and among these top journals were *Science* and the *Journal of the Chemical Society*. Their mean number of external loans was 232.5. Urquhart then took samples of 10 journals from those titles with respectively 20, 2, and 0 external loans. He averaged the BUCOP holdings of these four samples and summarized the results in a table that is replicated by Table 3. Here is visible the strong correlation of SML external loans with BUCOP holdings, with the average number of these holdings skewing rapidly downward from 57 for the top ten titles by SML loans, to 22.4 for the titles with 20

TABLE 3. Relationship of Number of 1956 Science Museum Library (SML) External Loans to Number of Holdings Listed in the *British Union Catalogue of Periodicals* (*BUCOP*)

Loan Class	Mean No. BUCOP Holdings
10 Titles Most Frequently Loaned (Mean No. Loans = 232.5)	57
Sample of 10 Titles Loaned 20 Times	22.4
Sample of 10 Titles Loaned 2 Times	4.5
Sample of 10 Titles Loaned 0 Times	2.3

Adapted from: Data from Urquhart 1959, 291, Table VI.

loans, to 4.5 for the titles with 2 loans, and to 2.3 for the titles with 0 loans. Urquhart drew the following conclusion from these data (p. 290):

> External organizations will naturally only borrow from the Science Museum Library scientific literature which they do not hold themselves, or which they cannot obtain from some more accessible collection. Thus the external loan demand on the library is, in general, only a residual demand . . . Nevertheless, possibly because so many external organizations (some 1200) use the Science Museum Library, it appears . . . that the use of the copies of a serial in the library is a rough indication of its total use value in the United Kingdom.

From the above, it is now possible to derive three of the main tenets of Urquhart's Law of Supralibrary Use: (1) the supralibrary use of sci/tech journals is highly skewed and concentrated on a relatively few titles; (2) the supralibrary use of sci/tech journals is highly correlated with the number of libraries holding these journals; and (3) the supralibrary use of sci/tech journals is a rough indicator of the total use value of these journals and therefore of their intralibrary use.

THE PROBABILISTIC BASES OF LIBRARY USE

Sets and Probability in Respect to Sci/Tech Journals

The great statistician, Kendall (1949, 101), distinguished two basic approaches toward the problem of probability. One takes probability as "a degree of rational belief," whereas the second defines probability in

terms of "frequencies of occurrence of events, or by relative proportions in 'populations' or 'collectives.'" In this paper, we will be concerned with the frequency theory of probability. The most cogent development of the frequency theory was done by Von Mises (1957), who based probability on relative frequencies within what he termed the "collective" but may also be considered a "set." Von Mises defined the collective as "a sequence of uniform events or processes which differ by certain observable attributes, say colours, numbers, or anything else" (p. 12), and he admonished, "It is possible to speak about probabilities only in reference to a properly defined collective" (p. 28). Von Mises (pp. 16-18) used as an example of this requirement the fact that a person's probability of dying at a given age is dependent on whether this person is defined as belonging to a collective containing both men and women or only men. Mises' requirement of well-defined sets poses one of the central problems for the application of the frequency theory of probability in library and information science.

Set definition in respect to sci/tech journals is governed by Bradford's Law of Scattering. Starting from the principle of the unity of science by which every scientific subject is related to every other scientific subject, Bradford (1934, 1986) gave the following verbal formulation of his Law of Scattering:

> . . . the law of distribution of papers on a given subject in scientific periodicals may thus be stated: if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus and succeeding zones will be as $1 : n : n^2$ . . .

Bensman (2001, 238) interpreted Bradford's Law as "a mathematical description of a probabilistic model for the formation of fuzzy sets." Classical set theory is based upon the binary "crisp set," whose elements are either clearly members of the set–numerically represented by 1–or not members of the set–numerically represented by 0. In contrast, a "fuzzy set" consists of elements that are not always fully in the set and can have membership grades ranging from 0 to 1.

The relationship of the Law of Scattering to fuzzy set theory can be demonstrated with the data presented by Bradford (1934) on the distribution of articles on the subject Lubrication over a set of 164 journals during the period 1931 through June 1933. These data were compiled

from references to Lubrication articles in a current bibliography being prepared at the Science Museum Library. Bradford aggregated his data into 3 classes: (a) journals producing more than 4 references per year; (b) journals producing more than one and not more than 4 references per year; and (c) journals producing 1 or less references per year. The results are shown in Table 4, and it can be seen that the distribution is very similar to the distribution of journals by 1956 SML external loans observed by Urquhart more than 20 years later. In the terms of Bradford's Law of Scattering, class a is the "nucleus of periodicals more particularly devoted to the subject," whereas classes b and c are the "several groups or zones containing the same number of articles as the nucleus." Thus, class a is comprised of 4.9% of the journals that produced 27.9% of the Lubrication references, and the number of journals in classes b and c has to rise exponentially from 17.7% to 77.4% to produce approximately the same percentage of Lubrication references. Bradford's three classes can also be defined by their decreasing subject membership grade in the following manner: (a) Lubrication; (b) Lubrication/Not Lubrication; and (c) Not Lubrication/Lubrication. Using these definitions, it is possible to construct the following function for quantifying the membership grade of the 164 journals in the Lubrication set:

> If the number of references per year to a journal is greater than 4, then the membership grade of this journal equals 1; but if the number of references per year equals or is less than 4, then the membership grade of this journal equals the number of references per year divided by 4.01.

Applying this function to Bradford's Lubrication data yields the results shown in Table 5. Here is seen a small core of journals fully in the Lubrication set with a membership grade of 1, and outside this core the membership grade of the journals skews rapidly downward from 0.998 to 0.125 as the number of journals skews rapidly upward from 3 to 102. As the proportion of Lubrication articles decreases in the journals, scope opens in the journals for articles on other subjects. A zero class–(d) Not Lubrication–has been added in the table with a question mark for the number of journals in it. The number of journals in the zero class has been deliberately left open, as this is an exceedingly complex question, which Bradford himself never successfully resolved.

Bradford's Law of Scattering mandates that subject sets of sci/tech journals will not be crisp ones but complex composites of various subject subsets. Moreover, the inability to determine the zero class means

TABLE 4. Bradford Journal Classes Derived from References to Lubrication, 1931-June, 1933 (Few 1933 References)

Class	No. Journals	% Journals	No. References	% References
(a) More than 4 References per Year [22 to 9 Total References]	8	4.9%	110	27.9%
(b) 2 to 4 References per Year [8 to 3 Total References]	29	17.7%	133	33.7%
(c) 1 or Less References per Year [2 to 1 Total References]	127	77.4%	152	38.5%
SUM	**164**	**100.0%**	**395**	**100.0%**

TABLE 5. Bradford's Law in Terms of Fuzzy Set Theory: Lubrication, 1931-June 1933 (Few 1933 References)

Classes	No. References per Year (1)	No. Journals Producing References	Membership Grade (2)
(a) Lubrication	11.00	1	1.000
	9.00	1	1.000
	7.50	1	1.000
	6.50	2	1.000
	5.00	2	1.000
	4.50	1	1.000
Classes a/b Boundary	4.01		1.000
(b) Lubrication/ Not Lubrication	4.00	3	0.998
	3.50	3	0.873
	3.00	1	0.748
	2.50	7	0.623
	2.00	2	0.499
	1.50	13	0.374
Classes b/c Boundary	1.01		0.252
(c) Not Lubrication/ Lubrication	1.00	25	0.249
	0.50	102	0.125
Classes c/d Boundary	0.01		0.002
(d) Not Lubrication	0.00	?	0.000

(1) In calculating the number of references per year for Lubrication, Bradford reported that it was assumed that practically all the references related to 1931 and 1932 only and the divisor used was therefore 2.
(2) Membership function: If the number of references per year to a journal is greater than 4, then the member grade of this journal equals 1; but if the number of references per year to a journal equals or is less than 4, then the membership grade of this journal equals the number of references per year to it divided by 4.01.

that there are no clear lines of demarcations between subject sets and subsets. Librarians have long been aware of this characteristic of sci/tech journals. In his standard work on serials, Osborn (1980, 268-288) states that libraries can operate excellently without classifying their periodicals and satisfactorily without providing subject headings for them. He advanced indexing as a better method of providing subject access to journals.

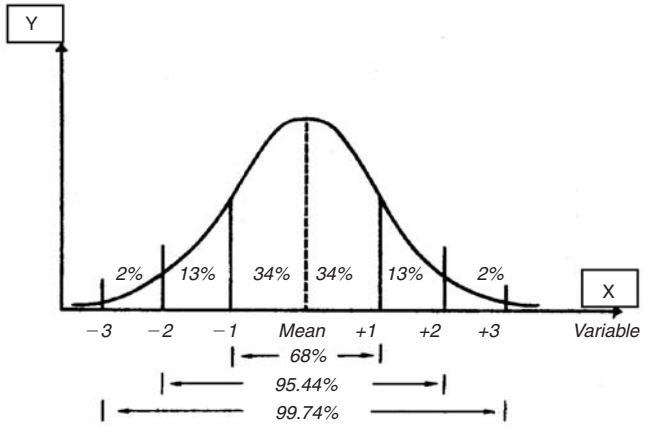
However, while these practical implications of Bradford's Law are well understood, the same cannot be said for its probabilistic consequences. As a result of this law, sci/tech journal distributions are most often complex amalgams of various distributions resulting from the different underlying probabilities of the component subject subsets. This phenomenon is also characteristic of distributions of other library materials. Given the fuzzy nature of these sets and subsets, these distributions are often not amenable to precise mathematical formulation.

The Calculation of Probability and the Normal Distribution

Probability is calculated by a mathematical equation called the probability density or probability mass function that determines what can be described generically as the proportion of members of a given set that have a specific characteristic. For example, this can be the proportion of a set of coins that are heads or–in terms of the question under analysis–the proportions of the sci/tech journal collection of the SML that in 1956 were externally loaned 0, 1, 2, 3, etc., times. A crucial element of the equation is a numerical constant called the parameter, which can be logically known a priori or estimated a posteriori from the data. Probability can be represented visually by means of a curve on a graph, on whose X, or horizontal axis, are the measures of the characteristic and on whose Y, or vertical axis, are the measures of the proportion or number of occurrences of this characteristic. Total probability is numerically defined as 1.00, and this number is assigned to the total area under the mathematical curve. The probability of the characteristic is the proportion or percent of the area under the curve that is above a defined segment of the X axis.

These basics of probability will be demonstrated with the normal distribution, which is graphically represented in Figure 2. The equation for the normal distribution has two parameters–the arithmetic mean and the standard deviation. This equation results in some form of the bell-shaped curve that is shown in Figure 2. Looking at the graph, it can be seen that with the normal distribution the area under the curve on

FIGURE 2. Probability as a Proportion of the Area Under the Normal Curve



A normal distribution indicating placement of 1, 2, and 3 standard deviation units on either side of the mean. Source: Carpenter, Ray L., and Ellen Storey Vasu. 1978. *Statistical methods for librarians*. Chicago: American Library Asoociation, p. 23. Reprinted with permission.

the segment of the X axis between the mean and one standard deviation above the mean is 34% of the total area under the curve and, therefore, contains 0.34 of the observations or members of the set. If one turns to Figure 1 with the bar chart of the frequency distribution of sci/tech journals by SML external loans in 1956 and mentally constructs a curve by connecting the tops of the bars with lines, one can see by comparing curves that this frequency distribution is nowhere near being represented by the normal distribution or any approximation to it.

The normal distribution was developed in the 18th and early 19th century as a law of error in astronomy and geodesy. Eisenhart (1983, 530) defines laws of error as "probability distributions assumed to describe the distribution of the errors arising in repeated measurement of a fixed quantity," and he states that they had the purpose of demonstrating the utility of taking the arithmetic mean of these measurements as a good choice for the value of the magnitude of this quantity. This explains the shape of the normal distribution. The mean is the same as the mode, or the point of the most frequently occurring value in a set of observations, and the symmetrical shape of the curve mandates that there is a 50/50 chance of an observation being on either side of the mean. During the 19th century, the normal distribution was thought to de-

scribe not only the distribution of error but also of all physical and social measurements. But this idea was refuted, and Snedecor and Cochran (1989, 40 and 44-50) state that the single most important reason for use of the normal curve is the central limit theorem, by which the distribution of the means of samples from even a non-normal population tends to become normal as the size of the sample increases.

The Binomial Distribution

The normal distribution is a continuous distribution in that it describes the distribution of variables that can take on any value including fractional ones. However, for the most part, library data consists of discrete integer counts and therefore requires discrete or discontinuous probability distributions. The basic ones of the latter type are the binomial distribution and the Poisson distribution.

Of these two distributions the binomial is historically the most important one, as it was the first probability distribution from which all the others were ultimately derived. The binomial distribution is based upon the repeated drawing of samples of a given size s from a population consisting of two classes (success-failure, yes-no, etc.). Besides the sample size s, its density function has one other constant, the parameter p or probability, which is the proportion of successes in the total population. Concerning the other population class, the proportion of failures is designated by q, so that q = 1 − p and p + q = 1. The distribution itself is calculated by the expansion of the binomial (p + q)^s. Of great importance in the binomial distribution is the close connection of the arithmetic mean with probability. The arithmetic mean is the size of the sample s multiplied by the probability of success p, so that:

$$\text{AVERAGE} = s*p$$

This close connection caused Rietz (1927, 14-16) to equate the mean with "the *mathematical expectation* of the experimenter." It is also important to note for the discussion below that the variance and standard deviation of a binomial distribution can be calculated by the following two equations:

$$\text{VARbinom} = s*p*q$$
$$\text{STDEVbinom} = \text{SQRT}(s*p*q) = \text{SQRT}(\text{VARbinom})$$

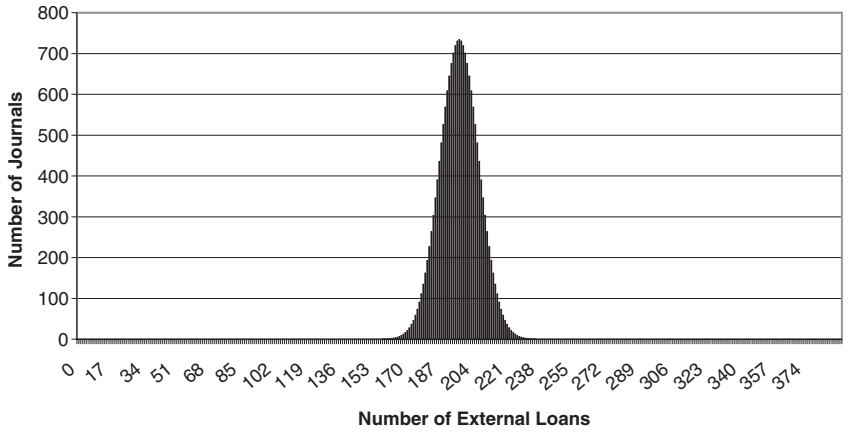
The binomial distribution will now be demonstrated with Urquhart's 1956 SML external loan data on the a priori assumption that p = 0.5 or

the probability of heads on the flipping of a fair coin. A major problem of applying the binomial distribution to library use is that it requires knowledge of not only the number of successes but also the number of failures. However, while it is relatively easy to count the number of times a journal has been loaned, it is not possible to count the number of times a journal has not been loaned. This makes it difficult to determine the size of the binomial sample s and to estimate the parameter p. One way around this difficulty is to utilize the technique suggested by Grieg-Smith (1983, 57-58) and recommended by Elliott (1977, 17). The technique requires that one first define the size of the binomial sample s by determining the maximum possible number of occurrences for any given member of the set. The journal most frequently borrowed by outside organizations from the SML in 1956 was the *Proceedings of the Royal Society of London (Series A)*, which accounted for 382 external loans, and it is logical to use this number for the size of the binomial sample s. From the perspective of binomial theory, each journal now becomes a sample of 382 possible loans. Having done this, it is now possible to make the following calculations:

Number of SML Journals (n) = 18,000
s = 382
p = 0.5
q = 1 − 0.5 = 0.5
Total Possible Loans (Tpos) = n*s = 18,000*382 = 6,876,000
Total Actual Loans (Tact) = Tpos*p = 6,876,000*0.5 = 3,438,000
AVERAGE = Tact/n = 3,438,000/18,000 = 191
AVERAGE = s*p = 382*0.5 = 191
VARbinom = s*p*q = 382*0.5*0.5 = 95.51
STDEVbinom = SQRT(VARbinom) = SQRT(95.51) = 9.77

The two ways of calculating AVERAGE demonstrate the close connection of the arithmetic mean with probability. To provide a further understanding of the binomial distribution, 382 and 0.5 were respectively used as the constants s and p in the binomial density function, and the resulting distribution of journals by number of external loans was both calculated and graphed. Figure 3 graphs the binomial distribution of titles at p = 0.5 over each possible number of 1956 SML external loans from 0 to 382. A look at Figure 3 in conjunction with Figure 2 illustrates the close connection of the binomial distribution with the normal law of error, as both frequency curves have the same symmetrical, bell-shaped form. According to Snedecor and Cochran (1989, 117-119 and 130), as s increases, the discrete binomial distribution approximates

FIGURE 3. Distribution of Scientific Journals by Number of 1956 Science Museum Libary (SML) External Loans on Assumption of Binomial with p of 0.5



more and more the continuous normal distribution. The size of the s required for this approximation is dependent on the value of p, being smallest at p = 0.5, where the approximation is good with s as low as 10.

At p = 0.5, all 18,000 titles in the SML collection would have concentrated in the Loan Class (100 to 382), which is the upper level of the Super High Loan Class in Section 2 of Table 2. However, a glance at Tables 1-2 and Figure 1 demonstrates that this was obviously not the case, because in reality an estimated 16,749 titles or 93.05% were concentrated in the Low Loan Class (0 to 9). Moreover, a comparison of the observed set mean of 2.96 loans per title to the theoretical set mean of 191 loans per title at p = 0.5 is proof that the actual overall probability of titles being loaned was extremely low. Using the same techniques as above but basing ourselves on the total number of 53,216 external loans actually observed, we can calculate a posteriori the binomial characteristics of the distribution of sci/tech journal titles by 1956 SML external loans thus:

Number of SML Journals (n) = 18,000
s = 382
Total Possible Loans (Tpos) = n*s = 18,000*382 = 6,876,000
Total Observed Loans (Tobs) = 53,216
p = Tobs/Tpos = 53,216/6,876,000 = 0.01

$$q = 1 - p = 1 - 0.01 = 0.99$$
$$\text{AVERAGE} = \text{Tobs/n} = 53{,}216/18{,}000 = 2.96$$
$$\text{AVERAGE} = s*p = 382*0.01 = 2.96$$
$$\text{VARbinom} = s*p*q = 382*0.01*0.99 = 2.93$$
$$\text{STDEVbinom} = \text{SQRT(VARbinom)} = \text{SQRT}(2.93) = 1.71$$

These calculations demonstrate that in 1956 the overall probability of SML external loans was only 0.01.

The Poisson Distribution

As probability becomes extremely low, the binomial distribution is transformed into the Poisson distribution. The latter distribution is the most important probability distribution for modeling library use. Its importance in this respect arises not only from its characteristics that make it suitable for this purpose but also from its overall importance. Thus, R. A. Fisher (1970, 54), one of the founders of modern inferential statistics, ranked the normal distribution as the most important of the continuous distributions and the Poisson as the most important of the discontinuous distributions. The name of this distribution is taken from that of Siméon Denis Poisson, who is generally credited with being the first to derive this distribution as a limit to the binomial in an 1837 book on judicial decisions.

What makes the Poisson distribution particularly fit for modeling library use are the following characteristics. First, it is a discontinuous distribution, and library use is measured by integer counts. Second, it arises as a limit to the binomial as p becomes very small, and the probability governing library use is usually very small. As has been seen above, the binomial p of external loans for all sci/tech journals in the SML collection in 1956 was 0.01, and this probability was much reduced in respect to individual titles. For example, the probability of the most highly loaned title, the *Proceedings of the Royal Society of London (Series A)*, was 0.00006. Third, the process by which the Poisson arises is suited to library use. Thus, whereas the binomial distribution is based upon the repetitive taking of samples of a given size containing both successes and failures, the Poisson distribution is based on mean rate of occurrence–technically called lambda (λ)–over some defined continuum such as time or space. With library use, space can be defined in terms of either individual titles or subject classes. Lambda is the only parameter of the Poisson density function, and it is much easier to estimate from library use data than the binomial p. With the binomial p, one

has to make an estimate not only of the number of times items were used but also the number of times items were not used, whereas the Poisson lambda can be estimated by simply counting the number of uses over some observation period and then dividing by the number of items subject to use. The Poisson's versatility in respect to library use is enhanced by the fact that, when p is small, the binomial and the Poisson are equivalent and can be substituted for each other. A key feature of the Poisson distribution is that lambda is equal to both the mean and the variance.

This is expressed by the following identity:

$$\lambda = \text{AVERAGE} = \text{VAR}$$

To demonstrate the Poisson in terms of the distribution of sci/tech journals by 1956 SML external loans, the mean of these loans per title–2.96–was utilized as lambda in the Poisson density function. This mean was selected to make the Poisson equivalent to the binomial with $p = 0.01$. The resulting frequency distribution of these titles was tabulated in terms of Urquhart's 1956 SML external loan classes presented in Table 1, and the tabulation is given in Table 6 next to the observed distribution of titles across these classes. This hypothetical Poisson distribution is graphed by Figures 4A and 4B above in two different ways. Figure 4A shows the correct theoretical shape of the distribution, whereas Figure 4B utilizes the same bar chart structure by Urquhart's classes as Figure 1 to facilitate comparison of the hypothetical Poisson distribution to the distribution actually observed in 1956. Comparing the tabulations of the hypothetical Poisson distribution against the observed distribution in Table 6 and Figure 4B against Figure 1 reveals that the Poisson distribution differs from the observed frequency distribution in two key ways: (1) the observed number of titles in Loan Classes 2, 3, and 4 around the mean of 2.96 is much lower than the number predicted by the Poisson; and (2) the observed number of titles in the loan classes at the two extremes–0 as well as 10 and above–is much higher than the number predicted by the Poisson. It should be noted that equivalent tabulations and graphs of the hypothetical binomial distribution with $p = 0.01$ were virtually identical to those of the Poisson with $\lambda = 2.96$. Moroney (1956, 127) counsels that the Poisson distribution can always be used as an approximation to the binomial distribution whenever p in the binomial is small with the approximation becoming better as p approaches zero.

TABLE 6. Comparison of Observed Frequency Distribution of Scientific Journals over 1956 Science Museum Library (SML) External Loan Classes to Hypothetical Poisson and Negative Binomial Distributions with Parameters Estimated A Posteriori from 1956 SML External Loan Data

Loan Class	Observed Title Frequency Distribution	Hypothetical Poisson Distribution	Hypothetical Negative Binomial Distribution
0	12,368	936	12,368
1	2,190	2,767	1,357
2	791	4,091	728
3	403	4,031	494
4	283	2,980	370
5 to 9	714	3,177	1,060
10 to 19	541	18	852
20 to 29	229	0	356
30 to 39	136	0	179
40 to 49	92	0	97
50 to 99	193	0	128
100 to 382	60	0	13
SUM	**18,000**	**18,000**	**18,000**

The Stochastic Processes of Library Use: Pearson and Asymmetric Distributions

The poor fit of both the binomial and the Poisson distributions to the frequency distribution of sci/tech journals by 1956 external loans indicates that there are stochastic or random processes affecting library use which limit the applicability of these distributions. Since the Poisson is a special case of the binomial, the requirements for these distributions are similar. Thus, in respect to the binomial, Rietz (1927, 24) states two requirements: (1) p must remain constant from sample to sample; and (2) the samples must be mutually independent in that the results of a sample should not depend in any significant degree on the results of previous samples. As for the Poisson distribution, Elliott (1977, 22) lists four such requirements: (1) p must be constant and small; (2) the number of successes per sampling unit must be well below the maximum number that can occur in a sampling unit; (3) the occurrence of a success must not increase or decrease the probability of another success; and (4) the samples must be small relative to the population. Library use violates these requirements in two important ways. The first may be categorized as "heterogeneity," namely, that library collections are com-

FIGURE 4A. Distribution of Scientific Journals by Number of Science Museum Library (SML) External Loans on Assumption of Poisson with Lambda Estimated A Posteriori from 1956 SML External Loan Data

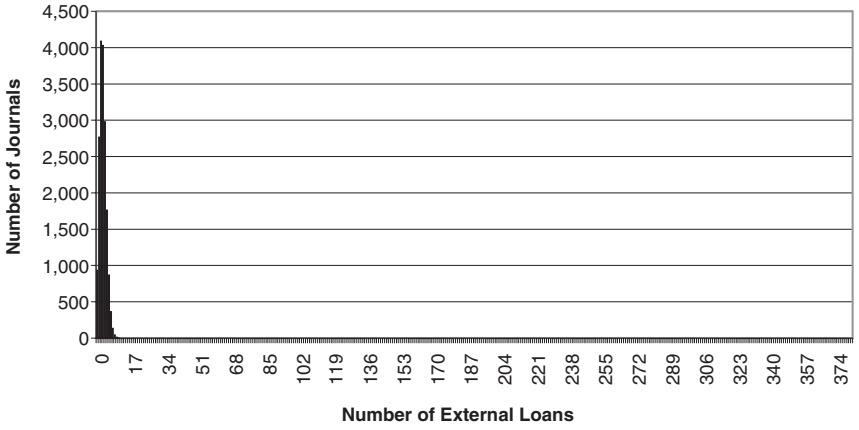
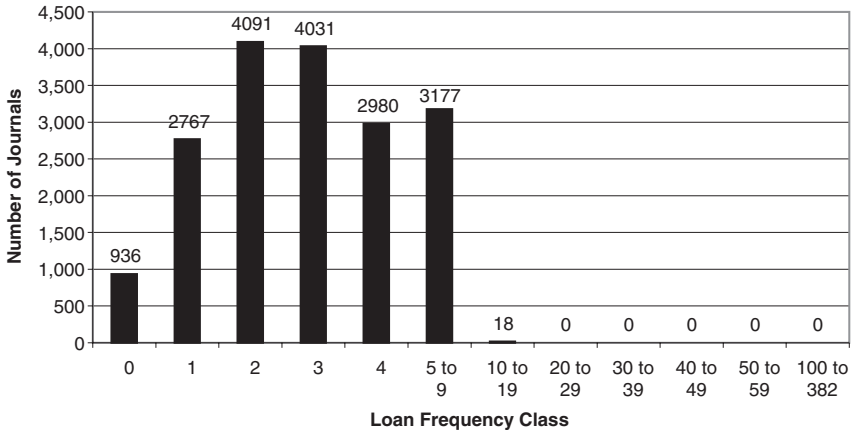


FIGURE 4B. Distribution of Scientific Journals by Urquhart's 1956 Science Museum Library (SML) External Loan Classes on Assumption of the Poisson with Lambda Estimated A Posteriori from 1956 SML External Loan Data



prised of elements, whose probabilities of being read wildly differ from each other. The other may be defined as "contagion" in the sense that library uses are not independent, because the use or non-use of an item affects not only its own future use or non-use but also that of other items.

Karl Pearson (1894; 1895; 1901; 1916) analyzed the stochastic processes causing the asymmetric frequency distributions of the type dominating library use in a series of four memoirs. In the second of these memoirs, Pearson (1895, 344-345) pointed out that asymmetric frequency curves may arise from two distinct classes of causes. The first such class is when the material measured may be heterogeneous in that it consists of a mixture of two or more homogeneous materials. In library terms, such a curve would result from the use of materials in a set comprised of two or more subject subsets with different underlying probabilities. Pearson (1894) had dealt with this type of asymmetric distribution in his first memoir, proving that asymmetric distributions were not solely a function of such mixed sets but could arise even with homogeneous material.

In his second memoir, Pearson (1895, 344) began his analysis of the other class of asymmetric curves that arise "in the case of homogeneous material when the tendency to deviation on one side of the mean is unequal to the tendency to deviation on the other side." In this and the succeeding memoirs, Pearson mathematically modeled this second type of curve with a system of twelve asymmetric frequency curves, which he derived off the hypergeometrical series Pearson (1916, 429-430) stated that he deliberately chose this series because it violated the three fundamental axioms of the normal distribution: (1) the equality in frequency of plus and minus errors of the same magnitude from the mean is replaced by an arbitrary ratio; (2) the number of contributory causes is no longer indefinitely large; and (3) the contributions of these causes are no longer independent but are correlated. The last condition incorporated the concept of contagion through being a direct violation of the binomial and Poisson requirement for independence of trials. Pearson demonstrated that his asymmetric curves better described the types of distributions found in reality than the normal distribution, which he found of little use in this respect. Of Pearson's asymmetric curves, three of the most important proved to be the following: Types I and VI establishing the bases for two forms of the beta distribution; and Type III, also known as the gamma distribution, which is the one best describing the shape of the frequency distributions dominating library use.

As part of his work with distributions, Pearson (1900) developed a method known as the chi-squared goodness of fit test for determining

how well an actual frequency distribution matches a theoretical frequency distribution. This test is based upon the chi-squared distribution, which is a particular case of his Type III or gamma distribution. R. A. Fisher (1966, 195-196) identified the essence of Pearson's chi-squared test as a comparison of the variance estimated from a sample with the true variance.

As a result of Pearson's work, it is possible to summarize the problem of analyzing the asymmetric distributions dominating library use under the following points. First, such distributions may arise not only from heterogeneous sets comprised of various subsets with differing probabilities but also from homogeneous sets whose individual elements may have differing probabilities. Second, the underlying causes may be correlated, and, therefore, a contagious process may be taking place, whereby the occurrence of an event affects its probability of reoccurrence. And, third, the amount of variance is one of the key characteristics that distinguish one type of distribution from another.

The Lexian System of Distributions

The pioneering work on Pearson's first class of asymmetric distributions–those arising from a mixture of two more homogeneous materials–was done by the German economist, Wilhelm Lexis, who expounded his theories in a series of articles and monographs published in the period 1875-1879. Two of the most cogent English expositions of Lexis' ideas were written by Rietz (1924) and A. Fisher (1922, 117-126).

The basis of Lexis' ideas was to test for the structure of a set by comparing its actual variance to its theoretical binomial variance through the Lexis Ratio (L). This is done by first calculating the standard deviation (STDEV) directly off the data as was demonstrated in Table 7 then its theoretical binomial standard deviation (STDEVbinom), and finally dividing the direct standard deviation by the theoretical binomial standard deviation thus:

$$L = STDEV/STDEVbinom$$

Both Rietz (1924) and A. Fisher (1922) use urn models to demonstrate set structure, and this practice will be followed here.

The urn model for the binomial distribution can be a single urn filled with black and white balls in constant proportions, where the drawing of a white ball is considered a success. Samples are drawn from this urn and replaced so that the proportion or probability of white balls from

sample to sample remains constant and the results of one sample does not affect the results of another sample. Under these conditions–homogeneity and independent trials–the Lexis Ratio should equal or approximate 1, indicating that the actual variance directly calculated from the data equals its theoretical binomial variance. Given the binomial's close relationship to the normal distribution, a Lexis Ratio of 1 indicates that the dispersion around the mean is primarily due to random error.

However, a Lexis Ratio greater than 1 means that the actual variance is greater than the theoretical binomial variance, and Lexian theory divides the variance into two components: the "ordinary or unessential" binomial component and the "physical" component (Rietz 1924, 86). The binomial component may be defined as that variance due to random error, whereas the excess variance is interpreted as resulting from the differing probabilities of the component subsets and is a sign of the Lexis distribution. A model for the Lexis distribution is a number of separate urns each containing different proportions of black and white balls. The urns thus represent subsets with different probabilities of white balls. Samples are fully drawn from the various urns and replaced in rotation so that the probabilistic heterogeneity of the urns is emphasized. Replacement of the samples maintains independence of trials. This is actually a good model for the binomial sampling of journal use. Under it, journals can be conceptualized as use samples of size s drawn fully in rotation from urns with differing proportions of uses and non-uses.

According to Lexian theory, the variance of a Poisson distribution is less than the corresponding variance of a binomial, so that a Lexis Ratio significantly less than 1 is indicative of the former distribution. The urn model for the Poisson distribution is the same as that for the Lexis distribution, in that it, too, consists of urns with differing proportions or probabilities of white balls. However, instead of the samples being fully drawn from the urns in rotation, they are constructed by taking 1 ball at a time from each urn, thereby randomizing the heterogeneous probabilities. It should be emphasized that a variance lower than the theoretical binomial variance is not necessarily a sign of the Poisson. Rietz (1924), for example, utilized binomial distributions with an overall $p = 0.5$ to demonstrate that randomizing the heterogeneous probabilities in the above fashion significantly reduces the variance below that theoretically expected with the binomial. The tendency of randomizing the probabilities to reduce the amount of variance plays an important role in a statistical law that is of crucial importance for the modeling of library use–Bortkiewicz's Law of Small Numbers.

TABLE 7. Comparison of Lexis Ratio of Observed Frequency Distribution of Scientific Journals over 1956 Science Museum Library (SML) External Loans with Lexis Ratios of Hypothetical Poisson and Negative Binomial Distributions with Parameters Estimated A Posteriori from 1956 SML External Loan Data

Statistical Measures	Observed Title Frequency Distribution	Hypothetical Poisson Distribution	Hypothetical Negative Binomial Distribution
Direct Standard Deviation from Data	11.59	1.72	8.93
Theoretical Binomial Standard Deviation	1.71	1.71	1.71
Lexis Ratio	6.77	1.00	5.21
Significant at 0.05 Level	Yes	No	Yes

Interpretation of Measures: (1) if the Lexis ratio is 1 or very near to 1 and **not** statistically significant, the actual variance approximates theoretical binomial variance; (2) if the Lexis ratio is below 1 and statistically significant, the actual variance is less than theoretical binomial variance; and (3) if the Lexis ratio is above 1 and statistically significant, the variance higher than theoretical binomial variance, indicating a Lexis distribution.

Ladislaus von Bortkiewicz was a student of Lexis, and he is best known for uncovering the importance of the Poisson distribution. According to Haight (1967, 115), although Poisson discovered the mathematical formula, Bortkiewicz discovered the probability distribution. His Law of Small Numbers is so closely connected with the Poisson distribution that it is often confused with it, but this is a misunderstanding of the Lexian bases of his work. Bortkiewicz set forth his law in a pamphlet published in 1898, and this pamphlet has been analyzed by Winsor (1947), who modernized the mathematical notation and translated key sections of it. To develop his law, Bortkiewicz analyzed the rate soldiers were kicked to death by horses in 14 Prussian Army corps in the 20-year period 1875-1894. These corps had different probabilities of soldiers being killed, so that the mean rate of deaths–or the lambda–differed from corps to corps. Nevertheless, when Bortkiewicz aggregated the data for all the corps, he found that the resulting frequency distribution closely fitted the Poisson. This caused him to formulate his Law of Small Numbers, which can be summarized simply in the following manner: If the field of observation is restricted to a set defined by infrequent occurrences, the resulting frequency distribution will fit the Poisson, whatever the differing probabilities of the elements or subsets comprising that set. The main requirement is that the number of occur-

rences be small out of a large population, and the more this requirement is met, the better the fit to the Poisson. It should be noted that this method of restricting of the field of observation has the effect of randomizing the probabilities since the occurrences are happening haphazardly over elements or subsets with differing probabilities, and Lexian theory dictates that the actual variance should therefore be less than the theoretically expected one.

Bortkiewicz's Law of Small Numbers has enormous implications for the evaluation and management of library collections, for it means that if the set is restricted to those items manifesting low use, no matter what the items or their subject class, one can expect not only that the set will have a low overall mean rate of use but also that the use of any component of this set will not deviate very far from this mean.

Lexian analysis was applied to 1956 SML external loans, and the results are presented in Table 7. Binomial theory required that all 18,000 journals be considered individual urns with differing probabilities of external loans, from which samples of 382 possible external loans are fully drawn in rotation. The theoretical binomial standard deviation of this journal set was calculated to be 1.71 above (p. 19-20). In respect to the distribution actually observed in 1956, the direct standard deviation was found to be 11.59 utilizing the method demonstrated in Table 1. These values yield the following Lexis Ratio for the observed distribution:

$$L = STDEV/STDEVbinom = 11.59/1.71 = 6.77$$

The distribution of scientific journals by 1956 SML external loans was thus a Lexis one, and it is possible to hypothesize that one reason for this excess variance is that Urquhart aggregated the loan figures for the SML journal collection as a whole instead of presenting the loan data in terms of well defined subject subsets, thereby controlling for one source of probabilistic heterogeneity. The direct standard deviation was also calculated for the Poisson distribution, whose λ parameter was estimated a posteriori from the 1956 SML external loan data, and it was 1.72. Dividing it by the theoretical binomial standard deviation of 1.71 results in a Lexis Ratio of 1 indicating the binomial distribution. This experiment demonstrates that Lexian theory is rather problematic on the distinction between binomial and the Poisson, for at the low level of probability, where the Poisson distribution arises, the two distributions tend to be equivalent.

Heterogeneity vs. Contagion

Probabilistic heterogeneity in the use of sci/tech journals and other library materials involves two basic, interacting factors. First, there is the Lexian factor of the various subject classes having different probabilities of being used. This factor is inherent in library use due to Bradford's Law of Scattering, which dictates that virtually every subject set of library materials will contain subject subsets with different underlying probabilities of being read. Second, there is the Pearsonian factor of the probabilistic differences of members of homogeneous subject sets being used due to such causes as importance or quality, size, age, completeness, language, etc. The main vehicle for modeling probabilistic heterogeneity has been the compound distribution, which is a distribution that results when the parameter–p in the binomial case, λ in the Poisson case–has its own distribution sometimes termed the "mixing distribution." One of the chief uses of Pearson's asymmetric distributions has been to serve as such mixing distributions.

Properly conceived, the Lexis distribution is a mixture of binomial distributions, and it was the forerunner of the compound binomial distribution. Moran (1968, 76), as well as Johnson and Kotz (1969, 79), describe the beta distribution, which was pioneered by Pearson, as the "natural" mixing distribution for p in the compound binomial distribution. This form of the compound binomial is sometimes named the beta binomial distribution. However, for a number of reasons, it is the compound Poisson distribution that is more applicable for modeling library use. The most important compound Poisson distribution is the negative binomial distribution (NBD). One reason for the importance of the NBD is that it results not only from the stochastic process of heterogeneity but also from that of contagion. The heterogeneity form of the NBD is a compound Poisson model, which was developed by Greenwood and Yule (1920) on the basis of industrial accidents among British female munitions workers during World War I.

The Greenwood and Yule model can be explained simply in the following manner. Each female worker was considered as having a mean accident rate over a given period of time or her own lambda. Thus, the accident rate of each female worker was represented by a simple Poisson distribution. However, the various female workers had different underlying probabilities of having an accident and therefore different lambdas. Greenwood and Yule posited that these different lambdas were distributed in a skewed fashion described by Pearson's Type III or

gamma distribution, and therefore, certain workers had a much higher accident rate than the others and accounted for the bulk of the accidents. They found that this model fitted the data very well. Given its construction, the Greenwood and Yule form of the negative binomial distribution is called the gamma Poisson model. This form of the negative binomial distribution can be considered as modeling the probabilistic heterogeneities of members of a homogeneous set, and it used Pearson's gamma distribution as the mathematical description of accident proneness.

Eggenberger and Pólya (1984) formulated the contagious form of the NBD in a 1923 paper that analyzed the number of deaths from smallpox in Switzerland in the period 1877-1900. They derived their model off an urn scheme that involved drawing balls of two different colors from an urn and not only replacing a ball that was drawn but also adding to the urn a new ball of the same color. In this way, numerous drawings of a given color increased the probability of that color being drawn and decreased the chance of the other color being drawn.

In a key paper, Feller (1943) stated that Eggenberger and Pólya had independently rediscovered a distribution originally found by Greenwood and Yule. He then analyzed the different stochastic bases of the Pólya-Eggenberger and Greenwood-Yule derivations of the negative binomial. According to Feller, the Pólya-Eggenberger form was a product of "true contagion," because each favorable event increases (or decreases) the probability of future favorable events, while the Greenwood-Yule model represented "apparent contagion," since the events are strictly independent and the distribution is due to the heterogeneity of the population. Given that Greenwood-Yule and Pólya-Eggenberger reached the NBD on different stochastic premises–the first on heterogeneity, the second on contagion–Feller posed the conundrum that one therefore does not know which process is operative when one finds the negative binomial, and he pointed out that this also applies to other types of contagious distributions. Feller's conundrum certainly holds true for library use. Thus, one does not really know whether a given scientific journal circulates more than others, because it is qualitatively or quantitatively different, because patrons have used and recommended it, or because these two factors are operating interactively.

To test the applicability of the NBD as a model of library use, one such distribution was constructed by deriving its parameters off Urquhart's 1956 SML external loan data. This probability distribution has two parameters: the arithmetic mean and a negative exponent s, which Elliott (1977, 23) describes as a measure of the excess variance in a population.

The mean is estimated in the usual fashion, and the method of the observed proportion of zeros of Anscombe (1949; 1950) was utilized to estimate s. Once again, the resulting frequencies were tabulated by Urquhart's 1956 SML external loan classes, and these tabulations were placed in Table 6 together with the other distributional tabulations for comparative purposes. The NBD frequencies were also graphed in the same two ways as before, i.e., by possible number of 1956 external loans from 0 to 382 (Figure 5A) and by Urquhart's 1956 SML external loan classes (Figure 5B). Figure 5A showing the actual shape of the NBD distribution is overwhelmed by the huge zero class, but it does show that the NBD differs from the Poisson distribution with its λ parameter estimated from the data by being more compressed against the left vertical Y axis. Inspection of Table 6 and comparison of Figure 5B to Figure 1, which is the graph of the observed distribution by Urquhart's loan classes, demonstrate that the negative binomial is a fairly good model of library use. The most striking similarities are the high concentration of titles in the loan classes below the mean of 2.96 and the long tail extending to the right that contains the journals accounting for the vast bulk of the external loans. However, the fit of the NBD to the observed distribution is not statistically precise, and there are two main reasons for this. First, the estimates of both parameters are distorted by the rough approximation used for the number of journals in the zero class. This is a common problem in library analyses, where most distributions are truncated on the left due to the difficulties in determining the size of the zero class. Second, due to Urquhart's aggregation of all journals into one distribution regardless of subject class, we are not dealing with a single distribution but a composite of many distributions resulting from the different probabilities of the component subject subsets. Proof of this is seen in Table 7, which shows that the Lexis Ratios of the observed distribution and the hypothetical negative binomial are both significantly above 1–6.77 for the observed, 5.21 for NBD. The complications caused by this multiplicity of subject subsets are increased due to their interactions resulting from their fuzziness and lack of clear demarcating lines. However, as will be seen, mathematical precision is really not required for the practical application of probability models to the evaluation and management of library collections. The main role of such models should be to help determine what are the underlying stochastic processes and therefore the consequences resulting from any given decision.

FIGURE 5A. Distribution of Scientific Journals by Number of 1956 Science Museum Library (SML) External Loans on the Assumption of the Negative Binomial with Parameters Estimated A Posteriori from Actual Data

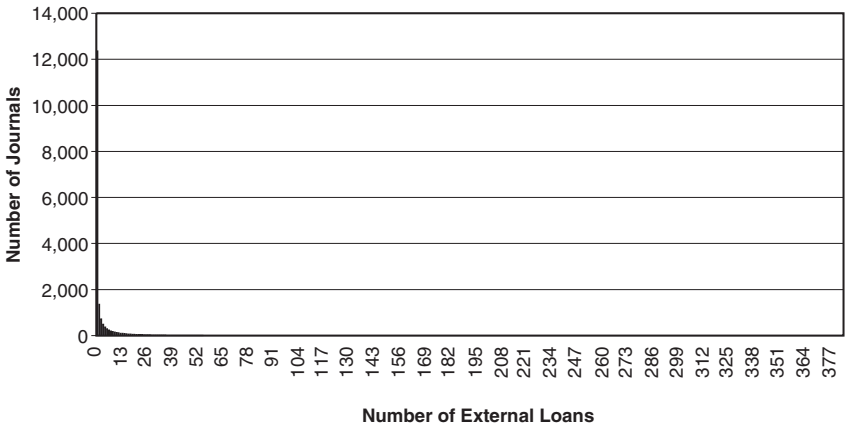
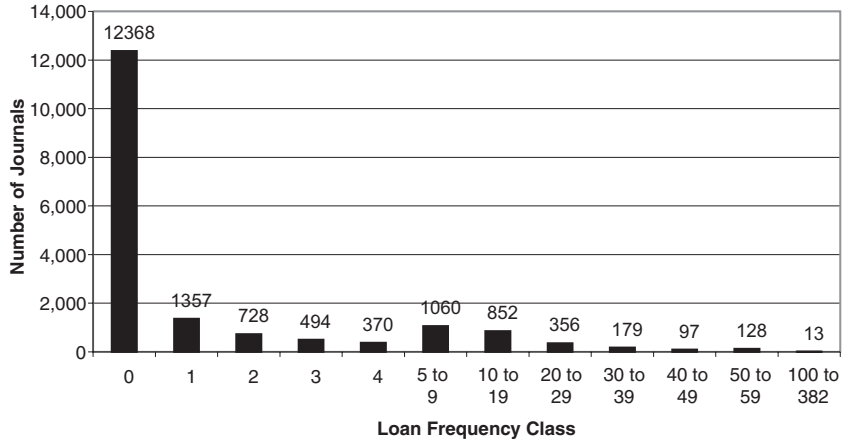


FIGURE 5B. Distribution of Scientific Journals by Urquhart's 1956 Science Museum Library (SML) External Loan Classes on Assumption of the Negative Binomial with Parameters Estimated A Posteriori from Actual Data



The negative binomial distribution has found numerous applications in the biological and social sciences. One of its most important biological applications is in ecology, where Elliott (1977, 50-51) describes it as probably the most useful mathematical model for distributions of species within given geographic areas. In respect to the social sciences, the NBD serves as the model of the zero sum game. This utilization can be demonstrated with Urquhart's 1956 SML external loan data in the following manner. For any given period–say, the year 1956–there can be only so many external loans. If a given journal has a greater probability being loaned, then it can achieve its higher loan rate only at the expense of other journals having a lower or even zero loan rate. The higher the probability of certain journals for being loaned, the lower must be the probability for other journals of being loaned. This results in what is technically called "over-dispersion" or the dispersion of journals away from the mean to both extremes of the distribution. Such a phenomenon is clearly visible in both the observed frequency distribution and the hypothetical NBD with the heavy concentration of titles in the zero class and the long tail to the right. In this respect, both the observed distribution and negative binomial distributions stand in sharp contrast to the hypothetical Poisson distribution, where the journals are heavily concentrated in the loan classes around the mean and there is no long tail to the right. Another name for "over-dispersion" in the social and information sciences is the "Matthew Effect." This term is derived from the gospel of St. Matthew (13:12), which states: "For whoever has, to him shall more be given, and he shall have an abundance; but whoever does not have, even what he has shall be taken away from him."

Indices of Dispersion

In his classic textbook, R. A. Fisher (1970, 57-61, 68-70) presented two tests–one for the binomial, the other for the Poisson–that utilize Pearson's chi-squared distribution as an index of dispersion. These tests comprise the easiest ways to determine the type of probability distribution and stochastic processes governing a set of data. An examination of Fisher's equation for chi-squared in his binomial test reveals it to be based upon a comparison of the actual variance of a set of data to its theoretical binomial variance. The relationship to the Lexis Ratio is obvious, and Fisher himself states (p. 80), "In the many references in English to the method of Lexis, it has not, I believe, been noted that the discovery of the distribution of [chi-squared] in reality completed the method of Lexis." He then outlined a method by which a given

chi-squared could be transformed into its equivalent Lexis Ratio. Fisher's equation for chi-squared in his index of dispersion test for the Poisson is based upon a comparison of the actual variance of a set to the arithmetic mean of the set. However, since under the conditions of the Poisson, the mean is equal to the variance, this test is also a comparison of actual variance to the theoretical variance. Given this identity, the ratio of the variance to the mean serves for the Poisson the same function as does the Lexis Ratio for the binomial, i.e., indicates the hypothesized distribution, if equal or approximate to 1.

Fisher's index of dispersion tests were further developed by Cochran (1954), who placed them within the system of hypothesis testing which is the standard method in statistics today. This system involves null and alternative hypotheses. Given Fisher's linking of his binomial index of dispersion test with the Lexis Ratio, one can define the hypotheses for his binomial index of dispersion test in accordance with Lexian theory and its further development through the compound binomial distribution. This results in a two-tailed test. The null hypothesis is the binomial distribution. If the actual variance is significantly less than the theoretical binomial variance, the alternative hypothesis is that the distribution has the subnormal dispersion indicative of the Poisson distribution; if the actual variance is significantly greater than the theoretical binomial variance, the alternative hypothesis is that the distribution has the supernormal dispersion characteristic of the Lexian distribution or a compound binomial such as the beta binomial. Thus, Fisher's binomial index of dispersion test can be considered from the Lexian viewpoint a test for whether a set is homogeneous or composed of subsets governed by differing probabilities. The latter case is the most frequent one in library use, where Bradford's Law of Scattering mandates that each subject set will be comprised of subsets from various subject fields.

The hypotheses for Fisher's Poisson index of dispersion test have been defined by Elliott (1977, 40-44). In the system presented by him, the null hypothesis is the Poisson distribution. If the variance is significantly less than the mean, Elliott defines the alternative hypothesis as "a regular distribution"; if the variance is significantly greater than the mean, he states the alternative hypothesis as "a contagious distribution." According to Elliott (1977, 46 and 50-51), the positive binomial distribution is the approximate mathematical model for a regular distribution, whereas the negative binomial is the most useful mathematical model for the diverse patterns of contagious distributions. Fisher's index of dispersion test for the Poisson is more applicable to library use than his index of dispersion test for the binomial. The main reason is

that his binomial test requires an estimation of the parameter p–a complex task due to the inability to count non-uses–whereas his Poisson test is based upon the parameter lambda or mean rate of use that is easily estimated from the observed rate of use. Not only is Fisher's Poisson test more easily applied to library use, it has the further advantage that at the low level of probabilities governing library use the Poisson is equivalent to the binomial, thereby making his Poisson test also a test for the binomial. Therefore, the Poisson test captures the effects of heterogeneity, whether it be the Lexian differences in the probability of subject subsets or the Pearsonian differences in the probability of members of homogeneous subject sets. Moreover, given Feller's conundrum, Fisher's Poisson index of dispersion test serves also as a test for the operation of contagion.

To illustrate Fisher's Poisson index of dispersion test, the ratio of the variance to the mean was calculated for the actual distribution of scientific journals by 1956 SML external loans as well as the hypothetical Poisson and negative binomial models of this distribution. The significance of these ratios was determined with Fisher's test. Table 8 presents the results. Here it can be seen that both the actual distribution and the hypothetical negative binomial have variance-to-mean ratios significantly above one–45.42 for the former, 26.95 for the latter–whereas the hypothetical Poisson distribution does not, proving that the actual distri-

TABLE 8. Comparison of Variance-to-Mean Ratio of Observed Frequency Distribution of Scientific Journals over 1956 Science Museum Library (SML) External Loans with Variance-to-Mean Ratios of Hypothetical Poisson and Negative Binomial Distributions with Parameters Estimated A Posteriori from 1956 SML External Loan Data

Statistical Measures	Observed Title Frequency Distribution	Hypothetical Poisson Distribution	Hypothetical Negative Binomial Distribution
Mean	2.96	2.96	2.96
Variance	134.29	2.96	79.69
Variance-to-Mean Ratio	45.42	1.00	26.95
Significant at 0.05 Level	Yes	No	Yes

Interpretation of Measures: (1) if the variance-to-mean ratio is 1 or very near to 1 and **not** statistically significant, the distribution is the Poisson; (2) if the variance-to-mean ratio is below 1 and statistically significant, the distribution is the binomial; and (3) if the variance-to-mean ratio is above 1 and statistically significant, the distribution is the negative binomial or one similarly resulting from inhomogeneity and contagion.

bution is better modeled by the negative binomial than by the Poisson. Comparison of Table 8 to Table 7 containing the results of the Lexis Ratio tests provides interesting insights. First, the results of the variance-to-mean ratio test validate the findings of the Lexis Ratio test that also demonstrated that the NBD is the better model of the actual distribution than the Poisson. Second, whereas the Lexis test indicates that hypothetical Poisson distribution is a binomial distribution due to the ratio being 1, the variance-to-mean test shows the hypothetical Poisson distribution to be a Poisson distribution, because once again the ratio is 1. This is proof of the equivalency of the binomial and Poisson distributions at the low levels of probability governing library use. Taken all together, these tests demonstrate the operation of both stochastic processes of heterogeneity and contagion in the observed distribution, which was affected by probabilistic heterogeneity between subject subsets as well as within subject subsets, where the loan or non-loan of a journal affected its later probability of loan or non-loan.

CONCLUSION

Urquhart's analysis of the 1956 loans of the Science Museum Library (SML) to outside organizations embodied two major advances. The first was distributional and pertained to the history of science as a whole. This aspect of Urquhart's work has been discussed in detail in two articles by Bensman (2000; 2005). During the 19th century, statistical analysis was based upon the normal paradigm, according to which the distributions of all phenomena are governed by the normal law of error. The late 19th century witnessed a scientific revolution, which led to the discovery that, on the contrary, many if not most phenomena–but particularly biological and social ones–have distributions that are highly and positively skewed. Two key figures in this scientific revolution were Wilhelm Lexis and Karl Pearson. Britain became the primary locus of the revolution due to Darwinism that stimulated the development of modern inferential statistics in this country. From this perspective, the studies conducted by Bradford and Urquhart at the SML must be viewed as extensions of this scientific revolution, because these studies were among the first to demonstrate that the probability distributions governing library and information science are also for the most part highly and positively skewed.

The second advance was conceptual and concerned specifically libraries. It resulted from Urquhart's focus on supralibrary use and must

be counted as Urquhart's original contribution to library and information science. The conceptual breakthrough was that there is no sharp distinction between supralibrary use and intralibrary use. As a result of this, in terms of the usage of their materials, all libraries function as part of a unified distributional system. Although this may seem a simple insight, it has manifold and complex ramifications. The next two parts of this paper will analyze these ramifications in respect to a central document delivery library and all the other libraries supported by such a library.

Received: August 23, 2004
Revised: January 13, 2005
Accepted: February 10, 2005

REFERENCES

Anscombe, F.J. 1949. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics* 5(2): 165-173.

_____. 1950. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* 37 (3/4): 358-382.

Bensman, Stephen J. 2000. Probability distributions in library and information science: A historical and practitioner viewpoint. *Journal of the American Society for Information Science* 51 (9): 816-833.

_____. 2001. Bradford's Law and fuzzy sets: Statistical implications for library analyses. *IFLA Journal* 27 (4): 238-246.

_____. 2005. Urquhart and probability: The transition from librarianship to library and information science. *Journal of the American Society for Information Science and Technology* 56 (2): 189-214.

Bradford, S.C. 1934. Sources of information on specific subjects. *Engineering* 137: 85-86.

_____. 1953. *Documentation*. 2d. ed. London: Crosby, Lockwood.

Carpenter, Ray L., and Ellen Storey Vasu. 1978. *Statistical methods for librarians*. Chicago: American Library Association.

Cochran, William G. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics* 10 (4): 417-451.

Eggenberger, F., and George Pólya. 1984. Ûber die Statistik verketteter Vorgänge. In George Pólya, *Probability; Combinatorics; Teaching and learning in mathematics*, edited by Gian-Carlo Rota, 94-104. Collected papers/George Pólya, vol. 4. Mathematicians of our times, vol. 22. Cambridge, Mass.: MIT Press.

Eisenhart, Churchill. 1983. Laws of error, I: Development of the concept. In *Encyclopedia of statistical sciences*, ed. Samuel Kotz and Norman L. Johnson, vol. 4: 530-547. New York: John Wiley.

Elliott, J.M. 1977. *Some methods for the statistical analysis of samples of benthic invertebrates*. 2d ed. Freshwater Biological Association scientific publication, no. 25. Ambleside, England: Freshwater Biological Association.

Feller, William. 1943. On a general class of "contagious" distributions. *Annals of Mathematical Statistics* 14 (4): 389-400.

Fisher, Arne. 1922. *The mathematical theory of probabilities and its applications to frequency curves and statistical methods*. Vol. 1, *Mathematical probabilities, frequency curves, homograde and heterograde statistics*. 2d ed. Translated by Charlotte Dickson and William Bonynge. New York: Macmillan.

Fisher, Ronald A. 1966. *The design of experiments*. 8th ed. Edinburgh: Oliver and Boyd.

_____. 1970. *Statistical methods for research workers.* 14th ed. New York: Hafner.

Greenwood, Major, and George Udny Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* 83 (2): 255-279.

Grieg-Smith, P. 1983. *Quantitative plant ecology*. 3d ed. Studies in ecology, vol. 9. Berkeley: University of California Press.

Haight, Frank A. 1967. *Handbook of the Poisson distribution*. Publications in operations research, no. 11. New York: John Wiley.

Johnson, Norman L., and Samuel Kotz. 1969. *Discrete distributions*. Distributions in statistics. Boston: Houghton Mifflin.

Kendall, Maurice G. 1949. On the reconciliation of theories of probability. *Biometrika* 36 (1/2): 101-116.

Moran, P.A.P. 1968. *An introduction to probability theory*. Oxford: Clarendon Press.

Moroney, M.J. 1956. *Facts from figures*. 3rd ed. Baltimore, Md.: Penguin Books.

Osborn, Andrew D. 1980. *Serial publications: their place and treatment in libraries*. 3rd ed. Chicago: American Library Association.

Pearson, Karl. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, A, 185: 71-110.

_____. 1895. Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, Series A, 186: 343-414.

_____. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (5th Series): 157-75.

_____. 1901. Mathematical contributions to the theory of evolution, X: Supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London, Series A*, 197: 443-59.

_____. 1916. Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London, Series A*, 216: 429-57.

Rietz, Henry Lewis. 1924. Bernoulli, Poisson, and Lexis distributions. In *Handbook of mathematical statistics*, ed. H.L. Rietz, 82-91. Boston: Houghton Mifflin.

_____. 1927. *Mathematical statistics*. The Carus mathematical monographs, no. 3. Chicago: Published for the Mathematical Association of America by Open Court Publishing Co.

Snedecor, George W., and William G. Cochran. 1989. *Statistical methods*. 8th ed. Ames: Iowa State University Press.

Urquhart, Donald J. 1948. The organization of the distribution of scientific and technical information. In *The Royal Society Scientific Information Conference, 21 June-2 July 1948: reports and papers submitted*, 524-527. London: The Royal Society.

_____. 1959. Use of scientific periodicals. In *Proceedings of the International Conference on Scientific Information, Washington, D.C., November 16-21, 1958*, vol. 1, 287-300. Washington, D.C.: National Academy of Sciences, National Research Council.

Urquhart, Donald J., and R.M. Bunn. 1959. A national loan policy for scientific serials. *Journal of Documentation* 15 (1): 21-37.

Von Mises, Richard. 1957. *Probability, statistics and truth*. 2nd rev. English ed. prepared by Hilda Geiringer. New York: Macmillan.

Winsor, Charles P. 1947. Das Gesetz der kleinen Zahlen. *Human Biology* 19 (3): 154-161.