# Probability Distributions in Library and Information Science: A Historical and Practitioner Viewpoint

**Stephen J. Bensman**
*LSU Libraries, Louisiana State University, Baton Rouge, LA 70803. E-mail: notsjb@lsu.edu*

**This paper has a dual character dictated by its twofold purpose. First, it is a speculative historiographic essay containing an attempt to fix the present position of library and information science within the context of the probabilistic revolution that has been encompassing all of science. Second, it comprises a guide to practitioners engaged in statistical research in library and information science. There are pointed out the problems of utilizing statistical methods in library and information science because of the highly and positively skewed distributions that dominate this discipline. Biostatistics are indicated as the source of solutions for these problems, and the solutions are then traced back to the British biometric revolution of 1865–1950, during the course of which modern inferential statistics were created. The thesis is presented that science has been undergoing a probabilistic revolution for over 200 years, and it is stated that this revolution is now coming to library and information science, as general stochastic models replace specific, empirical informetric laws. An account is given of the historical development of the counting distributions and laws of error applicable in statistical research in library and information science, and it is stressed that these distributions and laws are not specific to library and information science but are inherent in all biological and social phenomena. Urquhart's Law is used to give a practical demonstration of the distributions. The difficulties of precisely fitting data to theoretical probability models in library and information science because of the inherent fuzziness of the sets are discussed, and the paper concludes with the description of a simple technique for identifying and dealing with the skewed distributions in library and information science. Throughout the paper, emphasis is placed on the relevance of research in library and information science to social problems, both past and present.**

## Introduction

This paper has a dual character dictated by its twofold purpose. First, it is a speculative historiographic essay, in which I attempt to describe the present state of library and information science in terms of the overall development of science. To do this, I connect the history of library and information science with the history of probability and statistics. Second, the paper is intended to serve as a practical guide to persons doing statistical research in library and information science. Thus, the components comprising the duality of this paper are closely interconnected.

I came to this topic as a result of recently completed research (Bensman, 1996; Bensman & Wilder, 1998) on the market for scientific information. In this research, I wanted to solve what seemed a simple problem: What role does scientific value play in the price libraries pay for scientific journals? To solve this problem, I had to use parametric statistical techniques such as correlation and regression, and these techniques immediately confronted me with what seemed to be an extremely difficult problem. These techniques are based on the assumption of the normal distribution, whereas library and information science data do not conform to the normal distribution but are dominated by horrendously skewed distributions. I realized that there is a need to connect the information science laws with the probability distributions, on which statistics are based, in some easily understandable manner, as an aid to persons conducting statistical investigations of the problems afflicting libraries. This need is becoming especially pressing, as computers are not only making much more data available but also making simpler highly sophisticated statistical analyses through spreadsheets and software such as SAS.

To obtain help in this matter, I contacted the LSU Department of Experimental Statistics, which assigned me as an adviser an ecologist named Jay Geaghan. Jay suggested that I read a manual entitled *Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates* (Elliott, 1977). It was an eye-opener in two respects. First, the manual introduced me to the system of probability distributions, with which biologists model patterns in nature, showing how to test for them and transform them for standard parametric statistical operations. Second, it pointed out that the key model for the skewed distributions dominating biological phenomena is the negative binomial distribution. This jarred a memory of Price (1976) describing the negative binomial distribution as the model for the double-edged Matthew Effect, which Robert K. Merton and

his students, Jonathan and Stephen Cole and Harriet Zuckerman, had placed at the basis of the social stratification of science. In my previous writings (Bensman, 1982, 1985), I had posited the double-edged Matthew Effect as underlying the skewed patterns of library use.

The research on the scientific information market necessitated solving many complex statistical problems. In seeking the solutions for these problems, I noticed a dual pattern. First, most of these problems had already been solved in biostatistics. Second, most of the works presenting these solutions were British. The Elliott manual, for example, was published by the British Freshwater Biological Association, and based on samples of benthic invertebrates in the English Lake District. It also dawned on me that bibliometrics as a discipline had also risen to a great extent in Britain. Being a historian by training, my interest was naturally piqued, and I decided to write a book that would not only present a history of this development but would also be an aid to persons doing statistical research in library and information science. Such an approach seemed particularly beneficial, because, like myself, many persons in the library field understand things better in terms of their historical development than mathematically.

## The Probability Distributions Affecting Library and Information Science

In the analysis of the production, dissemination, use, and evaluation of human knowledge, we are basically dealing with three discrete or counting distributions and two continuous laws of error. The counting distributions are the following: (1) the binomial, which models uniformity and whose characteristic is that the variance is less than the mean; (2) the Poisson, which models randomness and whose characteristic is that variance equals the mean; and (3) the negative binomial, which models concentration and whose characteristic is that the variance is greater than the mean. I hasten to add that the negative binomial is only the most useful of a series of contagious distributions, and, depending on the circumstances, it can change into the beta binomial, Poisson, or logarithmic series.

To help explain the idea of a law of error, I will present to you my concept of a statistical model. A statistical model is a mental construct of reality that is logically designed to test a hypothesis. It is centered on a hypothetical point, from which deviations are measured according to a law of error. Depending on the size of the deviation from the hypothetical point on which the model is centered, one accepts or rejects the hypothesis being tested. In statistical textbooks, the law of error is the normal distribution, and the hypothetical center is the arithmetic mean, from which deviations are measured in standard numbers. Historically, the normal distribution was derived as an approximation to the binomial. However, because of the multiplicative character of many phenomena in the biological, social, and information sciences, the law of error in these disciplines is in numerous cases the lognormal distribution, and the hypothetical center is the geometric mean, from which deviations are measured in logarithmic units. The negative binomial can be transformed into an approximation of the lognormal distribution.

## Library and Information Science within the Context of the Historical Relationship of Probability and Statistics to Science as a Whole

The history of probability and statistics is too complex to be adequately summarized in a paper such as this. Therefore, I will restrict myself to reviewing the theses of two key books on this subject. Together, these books validate the view presented in a two-volume collection of essays published by MIT Press and entitled *The Probabilistic Revolution* (1987): that since 1800, the world has been experiencing a scientific revolution, in which the mathematical theory of probability has been adopted in discipline after discipline. This probabilistic revolution is coming to library and information science, as specific, empirical, bibliometric laws are being replaced by general stochastic models. Of primary importance in this transition has been the seminal work on bibliometric distributions by Bertram C. Brookes and Abraham Bookstein.

For his part, Brookes (1977, 1979, 1984) concentrated on Bradford's Law of Scattering, which he explored theoretically as a very mixed Poisson model. Coming to regard Bradford's Law as a new calculus for the social sciences, he found it almost identical mathematically to other empirical bibliometric laws, suspecting of these laws that "beneath their confusions there lurks a simple distribution which embraces them all but which remains to be identified" (Brookes, 1984, p. 39). He reduced these various laws to a single law, which he modeled two ways as "the Inverse Square Law of frequencies" and "the Log Law of ranks." The main features of Brookes' hypothesis of a single distribution arising from a mixed Poisson process were endorsed by Bookstein. In his work, Bookstein (1990, 1995, 1997) posited through mathematical analysis that the various bibliometric laws together with Pareto's law on income are variants of a single distribution, in spite of marked differences in their appearance. Seeking a way to deal with these distributions, Bookstein (1997) came to the following conclusion:

> I have argued. . .that one important mechanism for surviving in an ambiguous world is to create functional forms that are not too seriously affected by imperfect conceptualization. In this article I pushed this notion further, and looked at suitable random components for the underlying stable expectations. The family of compound Poisson distributions seems uniquely able to provide this service. (p. 10).

The first book to be reviewed is *Contributions to the History of Statistics* by Westergaard (1968). In this work, Westergaard writes that the history of modern statistics has been marked by two lines of evolution, which surprisingly have had little to do with each other. The first was "Political

Arithmetic," which originated in London in 1662 with the publication by John Graunt of a remarkable book, *Natural and Political Observations upon the Bills of Mortality*. This was the first attempt to interpret mass biological and social behavior from numerical data. "Political Arithmetic" was first concerned with questions of mortality and other problems of vital statistics but later turned to economic statistics. Gradually the expression "Political Arithmetic" was replaced by the word "statistics," a term earlier employed for the description of states. Independently from "Political Arithmetic," there evolved what Westergaard calls the "Calculus of Probabilities," which was developed by mathematicians in investigations of a purely abstract character. According to Westergaard's interpretation, the history of modern statistics has been a struggle to merge "Political Arithmetic" with the "Calculus of Probability," so that proper inferences could be drawn from the collection of numbers. A similar struggle is taking place today in library and information science, as librarians conduct exercises in "Library Arithmetic" on the massive amounts of data they collect to solve their problems, whereas information scientists play complex mathematical games without any regard to the abilities of librarians or the basic statistical problems, which they need to solve.

The thesis of the other book to be discussed—*The History of Statistics: The Measurement of Uncertainty before 1900* by Stigler (1986)—is complementary to that of Westergaard in that he, too, traces the development and spread of statistical ideas. However, his treatment is much different. Whereas Westergaard concentrates on the history of the collection of data of interest to Political Arithmetic like state censuses, dealing minimally with the Calculus of Probability, Stigler emphasizes the role of statistics in the assessment and description of uncertainty, accuracy, and variability, by focusing on the introduction and development of explicitly probability-based statistics in the two centuries from 1700 to 1900. According to Stigler, during this period, statistics in his sense underwent what might be described as a simultaneous horizontal and vertical evolution. It was horizontal in that, prior to 1827, probability-based statistics originated in astronomy and geodesy, spreading after that date to psychology, biology, and to the social sciences. It was vertical in that the understanding of the role of probability advanced, as the analogy of games of chance gave way to probability models for measurements, leading finally to the introduction of inverse probability and the beginnings of statistical inference.

Such a perspective makes most interesting Stigler's discussion of the derivation of the normal distribution from the binomial in astronomy and geodesy during the eighteenth century. This process culminated at the beginning of the nineteenth century, when Pierre Simon Laplace and Carl Friedrich Gauss simultaneously combined all the elements of the normal distribution in what Stigler terms "the Gauss-Laplace Synthesis." This synthesis included, among others, the following elements: (a) Jacob Bernoulli's Law of Large Numbers, by which, as the number of observations increases, the relative number of successes must be within an arbitrarily small (but fixed) interval around the theoretical probability with a probability that tends to one; (b) Abraham De Moivre's derivation of the normal probability or bell-shaped curve as an approximation to the probability for sums of binomially distributed quantities lying between two values; (c) Thomas Simpson's justification of the advantage of taking the arithmetic mean of several observations in astronomy over that of a single, well-taken observation; and (d) Adrien Legendre's development of the least squares method for minimizing error. From Stigler's treatment, the normal distribution clearly emerges as what it actually is: the law of error in point estimation in astronomical and geodetic observations.

With this interpretation in mind, it is important to state the three principles that Pearson (1956b, p. 108) emphasized as connoted by the normal curve of errors: (1) an indefinite number of "contributory" causes; (2) each contributory cause is in itself equally likely to give rise to deviation of the same magnitude in excess and defect; and (3) the contributory causes are independent. Under these conditions, there arises the bell-shaped curve, where the mean equals the mode and the observations are symmetrically distributed on both sides of this point.

## The Normal Paradigm

We now come to what I will term the "normal paradigm," i.e., the idea that phenomena in nature and society, if sorted into homogeneous sets, are distributed according to the same law of error as observations in astronomy and geodesy. From personal experience, I can testify that this paradigm still has a powerful hold on the minds of people, at least in Baton Rouge, Louisiana. I am not alone in this observation. Already in 1916, Pearson (1956c), who demolished the normal paradigm in a series of brilliant papers in the 1890s, wrote in exasperation that "to attempt to describe frequency by the Gaussian curve is hopelessly inadequate" and "It is strange how long it takes to uproot a prejudice of that character!" (p. 555).

The main culprit in the rise of the normal paradigm was the Belgian scientist, Adolphe Quetelet. In the period 1823–1832, Quetelet's main project was the establishment of an astronomical observatory in Brussels. As part of this project, he visited Paris, where he came into contact with Laplace. This contact aroused in Quetelet the keen interest in statistical research, based on the theory of probabilities, that became the focus of all his scientific work.

One of Quetelet's main contributions was to extend the use of probability from celestial to terrestrial phenomena, and he is best known for the application of probability in studies of the physical and social attributes of human beings. Quetelet's studies of these attributes were dominated by his concept of the "average man" or the "homme moyen." The basis of this concept was his belief that all naturally occurring distributions of properly collected and sorted data follow a normal curve. He applied this theory

whether he was dealing with the chest sizes of Scottish soldiers or the penchant of humans to commit crimes. This thinking dominated Quetelet's approach to the definition of sets. His reasoning in this matter is summed up by Stigler (1986):

> What was true of astronomical observations would also be true of heights of men, of birth ratios, and of crime rates. Now, if homogeneity implied that observations would follow the normal law, then why not use this device for discerning homogeneity? Simply examine the distribution of a group of measurements. If they fail to exhibit this form, then this is surely evidence of lack of homogeneity—or at least evidence that the primary inhomogeneities are not in the nature of a large number of accidental (independent, random, of comparable force and size) causes. If they do exhibit this normal form, then this is prima facie evidence that the group is homogeneous and susceptible to statistical analysis as a group, without distinguishing the members of the group by identifying labels. (p. 205)

Quetelet used this type of reasoning in analysis of the heights of 100,000 French conscripts, coming to the conclusion of large-scale draft evasion because of the excess of men in the shortest class exempt from service. In another study of the heights of French conscripts—this time from the Department of Doubs in eastern France in the period 1851–1860—Adolphe Bertillon applied the same type of logic. Finding that the heights did not exhibit the usual symmetrical shape but rather had two modal values, Bertillon hypothesized that the population of Doubs consisted of two human types, one short and one tall. His theory seemed confirmed when his colleague Lagneau subsequently found that the inhabitants of Doubs were primarily of two different races, the Celts and the Burgundians. Bertillon's investigations bore an uncanny resemblance to the later work by Weldon that led Pearson to challenge the normal paradigm.

Quetelet's theories were subjected to severe criticism. One of the most devastating came from Bertrand; and Hogben, in his critique of statistical theory, lovingly quotes in full Bertrand's attack in a translation that captures all its Voltairean glory. By his transfer of the normal distribution from astronomy to the study of humans, Quetelet shifted the arithmetic mean from an actual point like a star or a comet in the sky to a hypothetical point in a set of humans, to which no human in the set may actually conform. This opened him—and, indeed, all modern inferential statistics—to charges of Platonism, which Bertrand made against him in the following passages of an analysis of one of Quetelet's findings that the average height of a set of 20,000 soldiers was 1 m 75:

> M. Quetelet. . .would have us accept a precise definition of the word Man, independently of human beings whose particularity can be considered accidental. . . . Our inequalities of height are, in his eyes, the result of inept measurements taken by Nature on an immutable model in whom alone she reveals her secrets. 1 m 75 is the normal height. A little more makes no less a man, but the surplus or deficit in each individual is nature's error, and thus monstrous. . . . (Hogben, pp. 172–173)

After a short disquisition in solid geometry on the relationship of the volume and surface of spheres to their radius, Bertrand drove the dagger home, stating:

> Men's shapes unfortunately can vary, and M. Quetelet profits therefrom. By combining the mean weight of 20,000 conscripts with their mean height, we should produce an absurdly fat man and, whatever Reynolds might have said, a poor model for an artist. . . . (Hogben, pp. 173–174)

Another flaw in Quetelet's thinking is that there is no inextricable link of the homogeneity of sets with the normal distribution, and, as we shall see, homogeneity lies at the basis of quite a different probability distribution. In his classic treatise on probability, Keynes (1921) noted that, because of Quetelet's work, the "suspicion of quackery" had not yet disappeared from the study of probability and that "There is still about it for scientists a smack of astrology, of alchemy" (p. 335).

## The British Biometric Revolution

The normal paradigm was demolished in the British biometric revolution. This revolution lasted roughly from 1865 to 1950, and it led to the creation of modern inferential statistics. The most important figures of the early phase of the revolution were Francis Galton, Karl Pearson, W.F.R. Weldon, George Yule, and William Gosset or "Student," whereas the most important persons of the latter phase were Ronald Fisher, Pearson's son Egon, Jerzy Neyman, and Maurice Kendall. University College, London, was the epicenter of the British biometric revolution, which entered the United States through the work of Snedecor at Iowa State University, a major agricultural research center. Fisher, Kendall, and Neyman visited Iowa State. One of the primary vehicles for the transmission of British statistical methods into the United States was Snedecor's textbook *Statistical Methods*, which went through eight editions from 1937 to 1989. A major reason for the broad influence of this book was that it made statistical methods accessible to persons with little mathematical training. The early editions' full title was *Statistical Methods Applied to Experiments in Biology and Agriculture.* The latter editions were coauthored by Cochran, who came to Iowa State in 1938 from the Rothamsted Experimental Station, the site of Fisher's major work, where agricultural research had been in progress since 1843. As the dates of these editions show, we are still very close historically to the British biometric revolution in terms of the diffusion of ideas. The work of the creators of the British biometric revolution was prolific and wide-ranging, and cannot be summarized in a paper of this nature. Here, I will focus only on the work of the men of its first phase on probability distributions.

*Destruction of the Normal Paradigm*

The British biometric revolution began with an attempt to place Darwin's theory of evolution on firm mathematical bases. It was begun by Galton, whose first cousin was Charles Darwin. The dominant theme in Galton's work from 1865 on was the study of heredity, and his work in this field was distinguished by its statistical nature. As a statistician, he was a direct descendant of Quetelet. In his book, *Hereditary Genius*, first published in 1869, Galton (1978) paid tribute to Quetelet after stating, "The method I shall employ for discovering all this, is an application of the very curious theoretical law of 'deviation from an average'" (p. 26). Galton devoted the appendix of this book to a demonstration of Quetelet's method in astronomy. The worshipful attitude Galton displayed toward the normal distribution is shown by the following oft quoted passage from his 1889 work, *Natural Inheritance*:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. (Galton, 1889, p. 66)

Despite his veneration of the normal distribution, even Galton noticed that it did not correspond to much of reality, particularly in biological and social phenomena. Because of his weakness in mathematics, he enlisted the aid of a Cambridge mathematician named McAlister to help solve this problem. In 1879, Galton presented a paper along with a memoir by McAlister (1879) that worked out the mathematics of his insight. In his paper, Galton (1879) stated that his purpose was "to show that an assumption which lies at the basis of the well-known law of 'Frequency of Error'. . .is incorrect in many groups of vital and social phenomena. . ." (pp. 365–366). He then defined the assumption in the following terms:

> The assumption to which I refer is that errors in excess or in deficiency of the truth are equally possible; or conversely, that if two fallible measurements have been made of the same object, their arithmetical mean is more likely to be the true measurement than any other quantity that can be named. (p. 366)

Galton then referred to the work in sense perception by Fechner, who had developed a law, whose simplest form is: sensation = log stimulus. According to Galton, in such cases, the geometric mean, rather than the arithmetic mean, is the better measure, and he then stated:

> The same condition of the geometric mean appears to characterise the majority of the influences, which, combined with those of purely vital phenomena, give rise to the events with which sociology deals. It is difficult to find terms sufficiently general to apply to the varied topics of sociology, but there are two categories of causes, which are of common occurrence. The one is that of ordinary increase, exemplified by the growth of population, where an already large nation tends to become larger than a small one under similar circumstances, or when capital employed in a business increases in proportion to its size. The other category is that of surrounding influences, or "milieux". . .such as a period of plenty in which a larger field or a larger business yields a greater excess over its mean yield than a smaller one. Most of the causes of those differences with which sociology [is] concerned may be classified under one or the other of these two categories. . . . In short, sociological phenomena, like vital phenomena are, as a general rule, subject to the condition of the geometric mean. (pp. 366–367)

Galton then went on to warn that the ordinary law of the frequency of error, based on the arithmetic mean, could lead to absurdity, when applied to wide deviations, stating that statisticians must confine its application within a narrow range of deviations. McAlister's memoir, entitled "The Law of the Geometric Mean," was a working out of the mathematics of the lognormal distribution.

Although Galton never understood the importance of his breakthrough and made only sporadic use of the lognormal in his own work, the true significance of this distribution for the biological, social, and, in particular, information sciences can be seen in the following two analyses. The first was done by Keynes in his treatise on probability. Commenting upon McAlister's work, Keynes (1921) wrote:

> [McAlister's] investigation depends upon the obvious fact that, if the geometric mean of the observations yields the most probable value of the quantity, the arithmetic mean of the logarithms of the observations must yield the most probable value of the logarithm of the quantity. Hence, if we suppose that the logarithms of the observations obey the normal law of error (which leads to their arithmetic mean as the most probable value of the logarithms of the quantity), we can by substitution find a law of error for the observations themselves which must lead to the geometric mean of them as the most probable value of the quantity itself. (pp. 198–199)

Shortly thereafter, Keynes came to the conclusion:

> . . .the main advantage of. . .Sir Donald McAlister's law of error. . .lies in the possibility of adapting without much trouble to unsymmetrical phenomena numerous expressions which have been already calculated for the normal law of error and the normal curve of frequency. (p. 200)

The other analysis pertains specifically to information science. It was done by Shockley (1957), a co-winner of the 1956 Nobel Prize in physics for his role in the creation of the transistor. Shockley studied the publication rates of

scientists at Los Alamos Scientific Laboratory and other places. He found highly skewed distributions with some individuals publishing at a rate of at least 50 times greater than others did. As a result of this, Shockley decided that it was more appropriate to consider not simply the rate of publication but its logarithm, which appeared to have a normal—or, better, lognormal—distribution over the population of typical research laboratories.

The problem, which led to the collapse of the normal paradigm, has been described by E. Pearson (1970, pp. 328–330), Karl's son, as that of "the double humped curve." It arose in Galton's studies of heredity. Galton (1978, pp. xvii–xix) himself succinctly defined the problem in the preface of the 1892 edition of his book, *Hereditary Genius*, in terms of the concepts of "variations" and "sports." During the course of his work, Galton developed the idea of "regression to the mean," which he defined in the 1892 preface in the following manner:

> It has been shown in *Natural Inheritance* that the distribution of faculties in a population cannot possibly remain constant, if, *on the average*, the children resemble their parents. If they did so, the giants (in any mental or physical particular) would become more gigantic, and the dwarfs more dwarfish, in each successive generation. The counteracting tendency is what I called "regression." The *filial* centre is not the same as the *parental* centre, but it is nearer to mediocrity; it regresses towards the racial *centre*. (Galton, 1978, p. xvii)

Galton viewed "variations" as variance in characteristics that occur around this racial center in accordance with the normal law of error without shifting this center. The case was much different with "sports." Here, according to Galton, a new characteristic appears in a particular individual, causing him to differ distinctly from his parents and from others of his race. In this scheme, "sports" are different from "variations," because, when transmitted to descendants, they establish a new racial center, towards which regression must be measured, thereby marking a new stage in evolution.

The concept of a "sport" lay at the basis of the problem of "the double humped curve," which Weldon handed Pearson, leading to the overthrow of the normal paradigm. A professor of zoology at University College, London, Weldon became convinced that the problem of animal evolution was essentially a statistical problem. Because of Galton's influence, he worked under the assumption that measurements of the physical characteristics in animal populations would be normally distributed within a homogeneous race. Weldon had success with this theory in his early work, but then he obtained an asymmetrical result that did not fit the normal curve in measuring the frontal breadth of a sample of crabs from the Bay of Naples. Weldon (1893, p. 324) hypothesized that the asymmetrical distribution he had obtained arose from the presence, in the sample measured, of two races of individuals clustered symmetrically about separate mean magnitudes. He excitedly wrote to Galton: "Therefore, either Naples is the meeting point of two distinct races of crabs, or a 'sport' is in the process of establishment" (E. Pearson, 1970, p. 328n). The solution of the problem required the dissection of a frequency distribution into two normal components. Lacking the necessary mathematical skills, Weldon turned for help to Pearson, who taught applied mathematics at University College.

The problem posed by Weldon led to a series of statistical memoirs by Pearson entitled "Contributions to the Mathematical Theory of Evolution," the first two of which were published in the *Philosophical Transactions of the Royal Society of London* during 1894–1895. Pearson (1956a, pp. 1–40) analyzed Weldon's problem in the first statistical memoir. As set forth by E. Pearson (1970, p. 329), there were three obvious alternatives to the solution of the problem of the "double humped curve:" (a) the discrepancy between theory and observation was no more than might be expected to arise in random sampling; (b) the data are heterogeneous, composed of two or more normal distributions; and (c) the data are homogeneous, but there is real asymmetry in the distribution of the variable measured. Acceptance of alternative (c) meant rejection of the normal paradigm, and this is precisely what Pearson (1956a) did in a dazzling display of mathematical prowess, coming to the following conclusion: "Professor Weldon's material is *homogeneous*, and the asymmetry of the 'forehead' curve points to real differentiation in that organ, and not to the mixture of two families having been dredged up" (p. 29).

Notwithstanding the first statistical memoir's huge significance, the second memoir by Pearson (1956b, pp. 41–112) was even of greater import for the future. Subtitled "Skew Variation in Homogeneous Material," it was dedicated to those frequency curves that arise in the case of homogeneous material, when the tendency to deviation on one side of the mean is unequal to deviation on the other side. Pearson noted that such curves arise in many physical, economic, and biological investigations, and he described the general type of this frequency curve as varying through all phases from the form close to the negative exponential to a form close to the normal frequency curve. Pearson's plotting of examples of these curves is shown in Figure 1.

To deal with these frequency curves, Pearson (1956c, pp. 529–530) resorted to the hypergeometrical series, because this series abrogates the fundamental axioms on which the Gaussian frequency is based in the following three ways: (1) the equality of plus and minus errors of the same magnitude 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 correlated. In the view of Pearson (1905/1906, pp. 203–204) the "Galton-McAlister Geometrical Mean Law," or the lognormal distribution, had also abrogated the third Gaussian axiom, because it amounted to saying that increments of the variate are correlated with the value of the variate already reached.

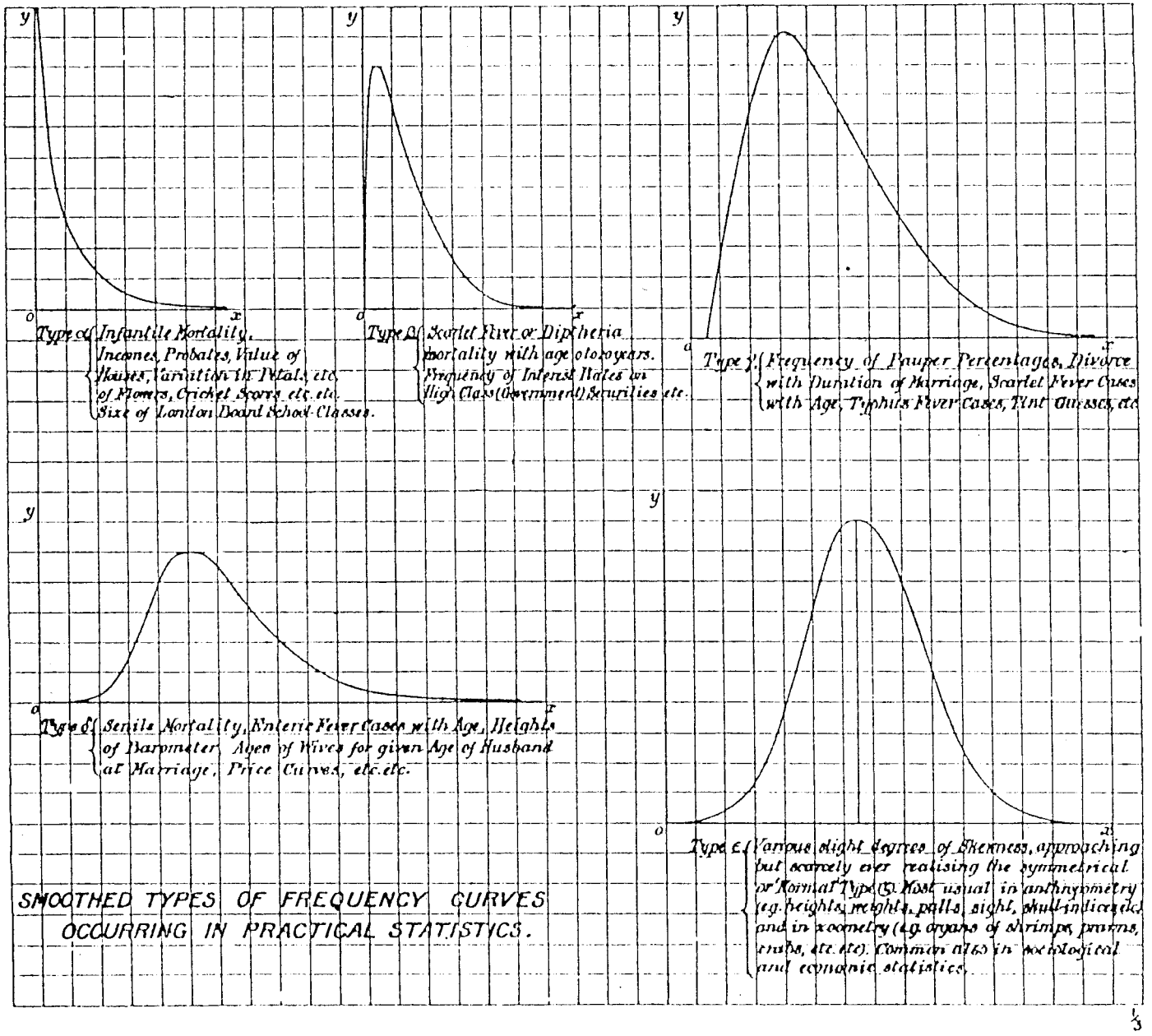


FIG. 1. Pearson's examples of asymmetrical frequency curves occurring in practical statistics.

On the basis of the hypergeometrical series, Pearson constructed in his second statistical memoir a system of five frequency curves, which are given below together with their modern description and Pearson's characterization of them (Pearson, 1956b, p. 58):

Type I (asymmetric beta)—Limited range in both direction, and skewness;
Type II (symmetric beta)—Limited range and symmetry;
Type III (gamma or chi-square)—Limited range in one direction only and skewness;
Type IV (a family of asymmetric curves)—Unlimited range in both directions and skewness; and
Type V (the normal)—Unlimited range in both directions and symmetry.

As E. Pearson (1970, pp. 329–330) noted, his father's derivation of the normal distribution off the hypergeometrical series was revolutionary, because it broke with the 200-year-old tradition of deriving the normal curve as an approximation to the binomial. The final 30 pages of the second statistical memoir were devoted to demonstrating the superiority of this system of curves over the normal distribution in the description of reality, using examples that ranged from barometric pressures, to the heights of St. Louis schoolgirls, to statistics on pauperism. Ultimately, Pearson expanded his system of frequency curves to 12, and he pointed out that "the Gaussian is a mere point in an infinite range of symmetrical frequency curves, and a single point in a doubly infinite series of general frequency distributions" (Pearson, 1956c, p. 550).

As a further development of his work during this period, Pearson (1956d, pp. 339–357) developed his chi-square goodness-of-fit test, which determines how well the observed frequencies of an actual distribution match the expected frequencies calculated from a theoretical distribution. He used this test to demonstrate again the superiority of his skewed distributions over the normal distribution, declaring that, "if the earlier writers on probability had not proceeded so entirely from the mathematical standpoint, but had endeavored first to classify experience in deviations from the average, and then to obtain some measure of the actual goodness-of-fit provided by the normal curve, that curve would never have obtained its present position in the theory of errors" (p. 355).

In an evaluation of Pearson's system of frequency curves, Neyman (1939) stated that their importance is "because of the empirical fact that, it is but rarely that we find in practice an empirical distribution, which could not be satisfactorily fitted by any such curves" (p. 55). Neyman saw as one of the main tasks explaining and mathematically describing the "machinery" producing empirical distributions of a given kind.

Together, the Galton-McAlister papers on the lognormal distribution and the Pearson paper on skew variation in homogeneous material can be regarded as the founding documents for the application of statistical methods in library and information science. The first provided the law of error, whereas the second laid the bases for the actual types of distributions with which this discipline has to deal. Of particular importance in the latter sense is Pearson's Type III, the gamma or chi-square distribution. In his second memoir, Pearson (1956b, pp. 72–73) called special attention to it, pointing out how this distribution in its malleability corresponds to the types of frequency curves found in reality, ranging from close to the negative exponential to close to the normal frequency curve. He graphed seven subtypes of the gamma distribution, and these subtypes are shown in Figure 2.

I can personally testify to the accuracy of Pearson's observations on the relative infrequency of the normal distribution and as to how well the gamma distribution describes those often found in reality. For the research on scientific journals, I ran dozens of datasets on such measures as total citations to journals and academic departments, journal impact factors, faculty ratings of journals, library use of journals, library holdings of journals, prices of journals, etc. Invariably, I obtained a distribution that corresponded to three of Pearson's subtypes of the gamma distribution, in that it was unimodal with the mode at the lower end of the distribution as well as highly and positively skewed to the right. When I showed my statistical adviser computer runs of faculty ratings and prices of journals in several subjects, he just shook his head and commented that all the data seemed to conform to the same distribution. The only time I did not obtain such a distribution was when I was selected by the National Research Council to test the database it had constructed during its 1993 assessment of the quality of U.S. research-doctorate programs. In running the distributions of the peer ratings of chemistry departments on the basis of a questionnaire designed in 1964 and used in that year, 1969, 1981, and 1993, I obtained for all years a symmetrical binomial approximating the normal distribution. This was such an unusual event that I suspected systematic instrument error. With such a distribution, there is a 50/50 chance of being on either side of the mean, but 11 chemistry programs had always been in the top 15 in peer ratings of such programs since 1924, in spite of the ever-increasing number of these programs being evaluated. Moreover, the intercorrelations of the 1964, 1969, 1981, and 1993 ratings ranged from 0.78 to 0.93. Cal Tech, MIT, and the other nine programs were pasted like barnacles on the extreme right side of the distributions and were not shifting back and forth across the mean as would be expected under the conditions of random observational error. I could only come to the conclusion that the 1964 questionnaire was severely flawed. As a historical aside, it is interesting to note that Galton had as a student and profoundly influenced James McKeen Cattell, the psychologist who constructed the first rankings of U.S. universities and academic departments in the sciences on the basis of peer ratings at the opening of the twentieth century.

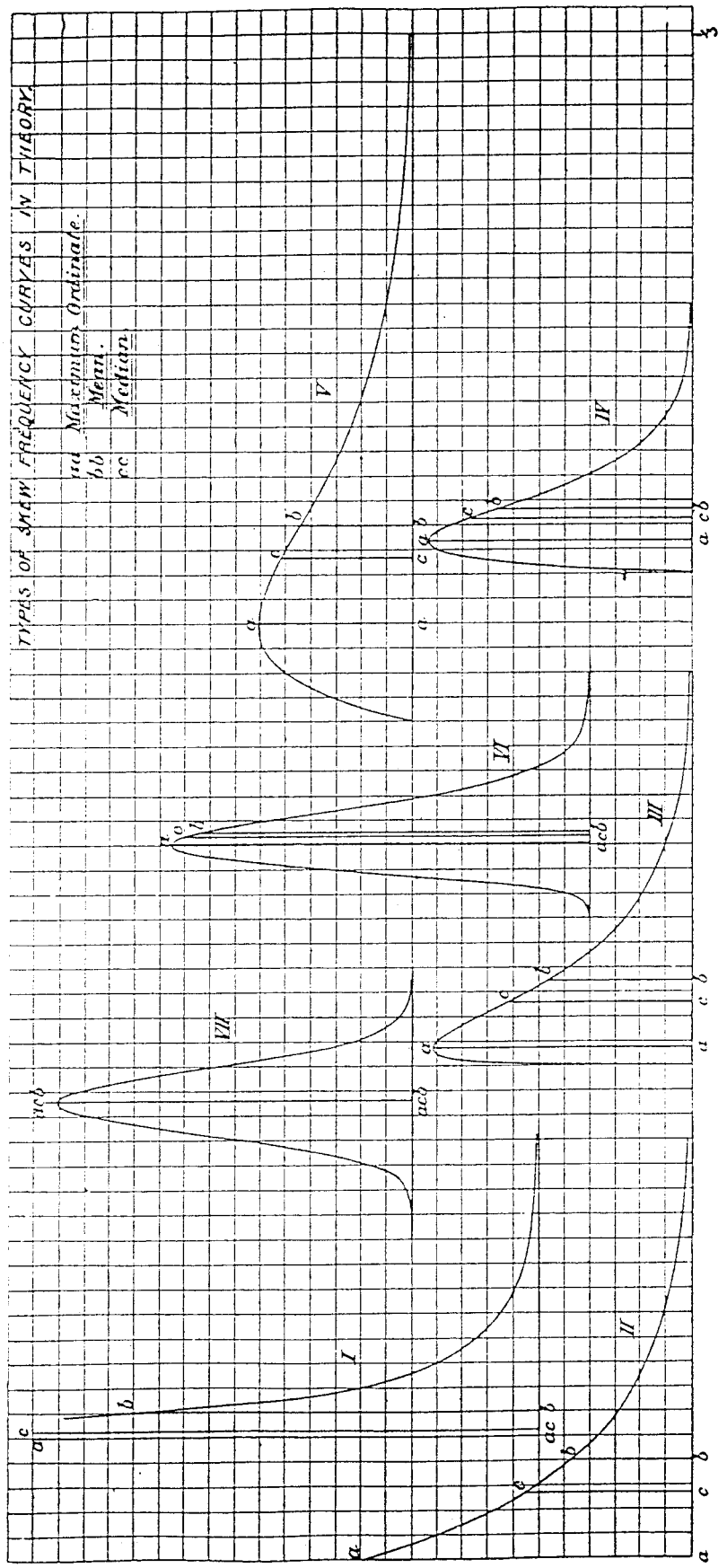


FIG 2. Pearson's seven examples of Type III gamma distributions.

*The Eugenics Connection*

At this point, it is important to emphasize that there is an extremely dark and dangerous side to the subject under discussion. The development of modern inferential statistics was closely intertwined with the creation of eugenics. Galton was the creator of eugenics, which he defined in his 1883 book, *Inquiries into the Human Faculty and Its Development*, in the following manner:

> [Eugenic questions are] questions bearing on what is termed in Greek, *eugenes*, namely, good in stock, hereditarily endowed with noble qualities. This, and the allied words, *eugeneia*, etc., are equally applicable to men, brutes, and plants. We greatly want a brief word to express the science of improving stock, which is by no means confined to questions of judicious mating, but which, especially in the case of man, takes cognisance of all influences that tend in however remote a degree to give to the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had. (Galton, 1883, pp. 24n–25n)

In 1904, he founded a research fellowship in national eugenics at the University of London, which was to develop in a few years into the Galton Laboratory of National Eugenics, with Karl Pearson as its first director. Pearson was succeeded by Fisher. In 1911, the Galton Laboratory of National Eugenics was merged with the Biometric Laboratory founded by Pearson in 1895 to form the first academic department of statistics, that of University College, London, and Pearson was its first professor. MacKenzie (1981) devotes his book, *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge*, to the close connection of the development of modern statistics with eugenics, and he sums up the relationship thus:

> One specific set of social purposes was common to the work of Galton, Karl Pearson, and R.A. Fisher. All were eugenists. They claimed that the most important human characteristics, such as mental ability, were inherited from one generation to the next. People's ancestry, rather than their environment was crucial to determining their characteristics. The only secure long-term way to improve society, they argued, was to improve the characteristics of the individuals in it, and the best way to do this was to ensure that those in the present generation with good characteristics (the 'fit') had more children than those with bad characteristics (the 'unfit'). (p. 11)

As MacKenzie points out, the work of these men in eugenics contributed to a body of notorious and controversial ideas, laying the bases for debates about race, class, and IQ. Eugenics were discredited by the virulent form they took in the racial policies of Nazi Germany, and this bad odor informed the ferocity of the assault on the logical structure of statistics after World War II by the biologist Hogben, who described Galton as "the father of the political cult variously named *eugenics* or *Rassenhygiene*" (p. 106), referred to "Galton's racialist creed" (p. 325), and decried "K. Pearson's flair for ancestor worship" (p. 176).

The dangers inherent in this situation for library and information science can be seen in the strange career of Shockley following his paper described here on the productivity of individual scientists. Lotka's Inverse Square Law of Scientific Productivity stands in apparent contradiction to the usual way intelligence is measured through the Stanford-Binet and other such tests. In their textbook, Brown and Herrnstein (1975, pp. 506–510), two Harvard psychologists, state that the scores of these tests result in the normal distribution, making the following observation: "If a test yields scores that do not match a normal distribution, it is often possible to whip them into the right shape by a suitable transformation" (p. 510). This fact underlies the title of the highly controversial book, *The Bell Curve: Intelligence and Class Structure in American Life*, by Herrnstein and Murray, whose conclusions and policy prescriptions caused Fraser (1995) to declare that "Not since the eugenics craze of the 1920s has this line of thought occupied a serious place on the national agenda" (p. 3). In his rediscovery of Lotka's Law in the lognormal form, Shockley (1957) hypothesized that the rise of this distribution was the manifestation of some fundamental mental attribute based on a factorial process, which enabled some individuals to be much more highly productive, and that the great variation in rate of production from one individual to another could be explained on the basis of simplified models of the mental processes concerned. From here he went on to develop his theory of "dysgenics," using data from the U.S. Army's pre-induction IQ tests to prove that African Americans were inherently less intelligent than Caucasians. Among his more controversial actions was to propose that individuals with IQs below 100 be paid to undergo voluntary sterilization, and he raised eyebrows by openly and repeatedly donating to a so-called Nobel sperm bank designed to pass on the genes of geniuses. Shockley died, regarding his work on race as more important than his role in the discovery of the transistor.

*Development of a New Stochastic Model*

Before one can understand the next development in the British biometric revolution of vast import for library and information science, it is necessary to have some knowledge of the Poisson distribution. Basically, the Poisson distribution arises as a result of random occurrences over time and space. In his *Introduction to Mathematical Sociology*, Coleman (1964, p. 291) states that the advantage of the Poisson process for the social sciences is that, unlike the binomial process, which is based on discrete "trials," the Poisson process occurs continuously over time and, therefore, is suitable for the naturally occurring events studied in the social sciences, where controlled experiments are often not possible. This distribution was first derived by Simeon Poisson as a limit to the binomial in a study published in 1837 on French judicial decisions. However, the first person

to grasp the statistical significance of the Poisson formula was von Bortkiewicz, who in 1898 published a small pamphlet entitled *Das Gesetz von kleinen Zahlen*, containing the classic illustration of the Poisson distribution in an analysis of the rate at which soldiers were kicked to death by horses in 14 Prussian army corps in the period 1875–1894. The literal translation of the title is "The Law of Small Numbers," but Crathorne (1928) suggests as better translations "The Law of the Probability of Rare Events" or "The Law of Large Numbers in the Case of Small Frequencies" (p. 173). An English translation of the key parts of von Bortkiewicz's pamphlet with modern mathematical notation is given by Winsor (1947).

The Poisson distribution has one parameter, which is estimated off the mean ($m$) and also equals the variance ($s^2$). The basic equation for the Poisson is:

$$P(x) = e^{-z}(z^x/x!)$$

where $x$ is the number of successes, $z$ is the parameter estimated, and $e = 2.71828$ and is the base of natural logarithms. As Poisson did, modern textbooks derive the Poisson as a limit of the binomial, and Snedecor and Cochran (1989, pp. 30–31, 110–113, 117–119, 130–133) illustrate the basic logic behind this method. Given a two-class population of successes and failures, where $p$ is the probability of success and is obtained by dividing the number of successes by the total number of trials or $n$, the population mean ($m$) equals $p$. Hence, in repeated trials taking random binomial samples of any size $n$, $m = np$. Using a sample size of eight, Snedecor and Cochran show how the Poisson arises from the binomial. When $p = 0.5$, the probability distribution of success is symmetrical with the mode at four successes, approximating the normal, but when $p = 0.2$, the probability distribution becomes positively skewed with the mode at one success and a concomitant decrease in the probability of the higher numbers of successes. Snedecor and Cochran then show that the binomial tends toward the normal for any fixed value of $p$ as $n$ increases, with the required $n$ being smallest at $p = 0.5$, but this approximation to the normal fails at $p < 0.5$, when the mean $m = np$ falls below 15, even if $n$ is large. The binomial distribution then remains positively skewed, and we are dealing with rare events, requiring the Poisson distribution.

It is often easier to explain what a phenomenon is by demonstrating what it is not, and I propose to do this with the Poisson distribution, using Urquhart's Law as an example. Although relatively unknown, Urquhart's Law is one of the most important library and information science laws. Briefly stated, it posits that the supralibrary use of scientific journals is the same as their intralibrary use, concentrating in the same fashion on the same titles. By supralibrary use, I mean use outside the collections of individual libraries, such as through interlibrary loan or centralized document delivery, whereas intralibrary use refers to the use of materials held in the collections of individual libraries by the patrons of these libraries. Urquhart's Law is the library use counterpart to the citation-based Law of Concentration posited by Garfield (1971, 1972, p. 476). Taken together, these two laws dictate that significant science tends to concentrate within a relatively few journals. Urquhart's Law lies at the basis of the operations of the British Library Document Supply Centre, and failure to understand the implications of this law has led to efforts at interlibrary cooperation in the U.S. resulting in what may be considered expensive failures.

Urquhart's Law was formulated by Donald J. Urquhart as a result of study of the use of scientific journals conducted in 1956 at the Science Museum Library (SML) in the South Kensington section of London in preparation for his establishment in Boston Spa, Yorkshire, of the National Lending Library for Science and Technology, which ultimately became the present-day British Library Document Supply Centre. This was the first major study of library use, and Urquhart's work preceded the use study done at the University of Chicago by Fussler and Simon (1969), as well as the formulation by Trueswell (1969) of his 80/20 Rule of library use. The fact that Urquhart studied journal use at the Science Museum Library is of historical significance, for this was the library once headed by Samuel C. Bradford. Bradford (1934) formulated his Law of Scattering at this library, which he turned into Britain's central interlibrary loan library for scientific journals.

Urquhart (1958; Urquhart & Bunn, 1959) analyzed 53,216 "external loans" or loans made in 1956 by the SML to outside organizations. These loans came from 5632 titles out of an estimated 18,000 titles held by the SML, of which 9120 were current, with the remainder being noncurrent. Of the serials titles used, 2769 were noncurrent. Urquhart found that around 1250 titles—or less than 10% of the titles—were enough to satisfy 80% of the external demand. Moreover, Urquhart compared the external use of the serials at the SML with the holdings of these titles by major U.K. libraries as listed in the *British Union List of Periodicals.* He found a direct correlation of the external use of the titles at the SML with the number of their holdings in U.K. libraries (i.e., the more heavily a serial was borrowed on interlibrary loan, the more widely it was held). As a result, Urquhart (1958) came to the following conclusion:

> External organizations will naturally 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 indicator of its total use in the United Kingdom. (p. 290)

Many years later, Urquhart (1981) formulated his law in the following manner:

> The fact that the heaviest inter-library demand is for periodicals, which are held by a number of libraries is of major

TABLE 1. Chi-square test of the goodness-of-fit of the Poisson distribution to the external loans attributable to the titles held at the Science Museum Library in 1956.[a]

| Number of external loans | Observed number of titles in class | Estimated number of external loans per class | Estimated percent of total external loans per class | Poisson probability per class | Expected number of titles in class | Chi-square |
|---|---|---|---|---|---|---|
| 0 | 12,368 | 0 | 0.0 | 0.052 | 936.1 | 139,615.8 |
| 1 | 2,190 | 2,190 | 4.1 | 0.154 | 2,767.4 | 120.5 |
| 2 | 791 | 1,582 | 3.0 | 0.227 | 4,090.9 | 2,661.8 |
| 3 | 403 | 1,209 | 2.3 | 0.224 | 4,031.5 | 3,265.8 |
| 4 | 283 | 1,132 | 2.1 | 0.166 | 2,979.7 | 2,440.6 |
| 5–9 | 714 | 5,355 | 10.1 | 0.176 | 3,176.7 | 1,909.2 |
| 10–382 | 1,251 | 41,748 | 78.5 | 0.001 | 17.8 | 85,349.3 |
| Totals | 18,000 | 53,216 | 100.0 | 1.000 | 18,000.0 | 235,362.9 |

[a] Poisson distribution rejected at 0.005 level at any chi-square above 16.75. Mean = 2.96 external loans per title. Variance estimated to be 138.1.

> importance in designing inter-library services. To draw attention to this relationship I have called it "Urquhart's law." It means, for instance, that the periodicals in the Boston Spa collection which are rarely used are unlikely to be used to any appreciable extent in a British university. There may be some exceptions to this deduction. . . . Nevertheless, the law is very important in considering the need for a central loan collection. (p. 85)

Up to now, we have been dealing with a brilliant piece of Library Arithmetic. However, Urquhart made an attempt to take the matter one step further and base his findings on a probability model. By doing this, he became one of the first persons to try to apply probability to library use. In the papers presenting these findings, Urquhart (1958, p. 291; Urquhart & Bunn, 1959, p. 21) assumes that the external loans at the SML were random, adopting the Poisson model to construct hypothetical distributions of journal use that would result from this model without actually fitting the Poisson to the SML data. Here, Urquhart is on more shaky ground, as he manifests ignorance about the working of this distribution. This is seen in his book, *The Principles of Librarianship*, in which Urquhart (1981) calls upon librarians to have an understanding of the Poisson distribution and makes the following statement: "Poisson was studying the number of grooms kicked to death by horses in the Prussian army" (p. 76). Urquhart thus attributes to Poisson the analysis done by von Bortkewicz on Prussian soldiers being kicked to death by horses long after Poisson himself was dead. The fact that the same journals dominate both supralibrary and intralibrary use defies the concept of randomness even from the linguistic point of view, much less the statistical.

From the tables and data presented by Urquhart in his papers, I was able to reconstruct a fair approximation of the total distribution of the 1956 external loans at the SML. With 18,000 titles and 53,216 loans, the mean external loans per title was 2.96. Using an antique method to derive the standard deviation off tabular data, I estimated the variance of the distribution at 138.1—needless to say, significantly higher than the mean of 2.96. I then fitted the Poisson distribution to Urquhart's data, using Pearson's chi-square goodness-of-fit test. The results are presented in Table 1. In doing so, I made an interesting discovery. Urquhart's major finding of 1250 titles accounting for 80% of the external uses was a probablistic impossibility by his own theoretical model. Urquhart's high-use class ranges from ten external loans to 382 external loans. However, the denominator of the Poisson equation—based on the factorials of the number of successes—rises exponentially faster than the numerator of the equation—based on the mean to the power of the number of successes, quickly crushing out the probability of any observation varying too far above the mean. At a mean of 2.96, the Poisson probability of ten external loans was 0.00073, and the total probability of all the titles, which accounted for ten external loans or above, was 0.001. The goodness-of-fit test resulted in a chi-square of 235,362.9, and a chi-square at anything above 16.75 meant rejection of the Poisson at the 0.005 level.

The reasons for the failure of Urquhart's data to fit the Poisson can be located in two requirements emphasized by Thorndike (1926) in a study of the applicability of this distribution to practical problems. According to Thorndike, the successful utilization of the Poisson requires that the occurrences constituting the sample under study be the result of "uniform" and "independent" trials. He then clarified these terms in the following manner:

> The term 'uniform' applies, of course, not to the results of the trials (or samples) but to the essential conditions under which they are obtained, and 'independent' is used with the meaning that the result of one trial (or sample) does not affect the occurrence of the event in any other trial (or sample). (p. 607)

With this statement, Thorndike puts his finger on the two stochastic processes—"inhomogeneity" and "conta-

gion"—by which a Poisson process results in the negative binomial distribution. As a matter of fact, so closely related is the Poisson distribution to the negative binomial that Stigler (1982) points out that, although the Poisson distribution appears in Poisson's 1837 study of French judicial decisions as the limit to the binomial, he actually derived his distribution directly as an approximation to the *negative* binomial cumulative distribution.

With respect to the first process, it is especially important for the observations in the set under study to be homogeneous in terms of an equal propensity to produce occurrences, if the distribution is to fit the Poisson. Under the conditions of high homogeneity, events occur randomly over the observations in the set, and this results in the observations clustering tightly around the mean. Von Bortkiewicz was aware of the need for homogeneity, causing him to exclude from his set four Prussian army corps organizationally different from the other ten. The need for homogeneity makes the test for the Poisson a good method in industrial inspection to determine whether manufacturing processes are creating products within required specifications. Inhomogeneity results in what is technically known as "over-dispersion." Using the example of a set of truck drivers, whose mean accident rate is one accident every 3 years, Borel (1943/1962, pp. 44–46), the noted French expert on the Poisson distribution, explains in the manner below how over-dispersion arises, when the truck drivers differ in their probability of having an accident:

> In the language of the calculus of probabilities, we sum up this increase of the proportion of cases where the number of accidents is 0, 2, 3, and the inevitably correlative decrease of cases where the number of accidents is equal to unity, which is the mean, by saying that the observed *dispersion* is greater than the normal dispersion. It is a general law of the calculus of probabilities that, in this case, the material observed is not homogeneous: the probabilities are not equal for all individuals, but above the average for some and therefore below the average for others. (p. 46)

Scientific journals are notoriously inhomogeneous in size, quality, and social influence. The difference in quality and social influence is a function of the fact that scientific talent is not randomly or uniformly distributed but concentrated in a relatively few persons. As Bensman and Wilder (1998) have proven with the help of the National Research Council database, the better scientists tend to concentrate at a few traditionally prestigious institutions and publish in a certain set of journals, of which those published by U.S. associations form an essential part. The resulting inhomogeneity of scientific journals manifests itself in faculty ratings, citation patterns, as well as in both intralibrary and supralibrary use.

Inspection of the table fitting the Poisson distribution to Urquhart's data reveals extreme inhomogeneity and over-dispersion, whose key points are highlighted by where the chi-squares are the greatest. For example, whereas the expected number of titles in the zero class is 936.1, the observed number is 12,368; whereas the expected number of titles in the high-loan class is 17.8, the observed number is 1,251; and there is a corresponding overprediction by the Poisson of the number of titles in the three classes—2 , 3 , and 4 external loans—surrounding the mean of 2.96 such loans. Here we see at work the double-edged Matthew Effect and the zero-sum game underlying the scientific information system.

The classic model of the effect of inhomogeneity in a Poisson process is the gamma Poisson version of the negative binomial distribution, and the construction of this model was made possible by the almost simultaneous breakthroughs by Pearson and von Bortkewicz. Pearson (1905/1906, pp. 208–210, 1915/1917, 1956c, p. 554) himself became intrigued with the potential of the negative binomial of the form $(p - q)^{-n}$, where $p - q = 1$. His interest was shared by other participants in the British biometric revolution. For example, Pearson (1915/1917) reported:

> Thus, if two or more of Poisson's series be combined term by term *from the first*, then the compound will always be a negative binomial. This theorem was first pointed out to me by 'Student' and suggested by him as a possible explanation of negative binomials occurring in material which theoretically should obey the Law of Small Numbers, e.g. 'Student's' own Haemacytometer counts. Of course, the negative binomial may quite conceivably arise from other sources than heterogeneity. . . . (pp. 139–140)

The above passage outlines in an unclear fashion the essential elements of the inhomogeneity model, whose precise mathematical development was accomplished by Pearson's protégé, George Yule, in collaboration with the epidemiologist, Major Greenwood.

The utility of the negative binomial was first brought to the attention of statisticians by Yule (1910) in a paper read before the Royal Statistical Society in December 1909. In this paper, Yule dealt with distributions that arise when the causal effects act cumulatively on homogeneous populations. For such distributions, he derived a law by which the probabilities were calculated by the successive terms of the binomial expansion of $p^r (1 - q)^{-r}$, and he was able to fit this distribution to two sets of data on diseases and one on divorces that were not amenable at all to fitting by an ordinary binomial series. This paper was particularly notable, because in it Yule derived from this negative binomial series a formula for the Type III gamma distribution, which Pearson had originally calculated on the basis of the normal binomial expansion of $(p + q)^r$.

As described by Greenwood (Newbold, 1927, pp. 536–537), it was the needs of the Royal Flying Corps during World War I that led Greenwood and Yule to construct the gamma Poisson version of the negative binomial distribution. At that time, the question arose as to whether to allow a pilot who had suffered a small accident to fly again. The problem emerged whether it would be possible to distin-

guish by analysis of frequency distributions of accidents between three possibilities: (1) that accidents are accidents in the sense of being "simple" chance events; (2) that the distribution of *first* accidents is that of "simple" chance events; and (3) that the distribution whether of first or subsequent accidents differs in a specific way from the "simple" chance scheme. As no hypothesis could be tested from Flying Corps data because of the question of unequal exposure to risk, Greenwood and Yule confined their investigation to the statistics of trivial accidents among female workers in munitions factories.

In solving this problem, Greenwood and Yule (1920) based themselves upon the Poisson, taking advantage of the following characteristic of this distribution that makes it particularly applicable to time and space problems: the sum of a set of numbers, each following a separate Poisson series (about different means), is itself a Poisson series. As a result, time and space units can be split into smaller ones, cells can be divided, periods of exposure can be changed, or the records of separate individuals can be summed into records of groups; and each of the single sets as well as the sum of the whole will yield a Poisson series. The solution, for which Greenwood and Yule opted was based on two assumptions: (1) each individual worker had a different mean rate of accidents that was constant throughout the period, thus forming her own Poisson series; and (2) the mean accident rates of the workers were distributed over the population according to a theoretical curve. For the latter, Greenwood and Yule considered the normal curve of error but rejected it as not being positively skewed enough. Noting that the choice of skew curves was arbitrary, they selected Pearson's Type III gamma distribution, but in the form derived by Yule off the negative binomial series. Greenwood and Yule found that their gamma Poisson model gave reasonable fits to observed accident rates in tests of fourteen sets of data on female munitions workers. In a paper read before the Royal Statistical Society, Newbold (1927) extended and confirmed the work of Greenwood and Yule, and together these papers established the concept of "accident proneness," with the gamma distribution serving as its mathematical description.

As a demonstration of the continuous spread of probabilistic methods to other disciplines, two other papers, which were presented to the Royal Statistical Society on the gamma Poisson distribution, should be mentioned at this point. In 1957, Ehrenberg (1959) read a paper that extended its use to marketing as the model for consumer buying, with the purchases of individual buyers following the Poisson distribution in time and the average purchasing rates of the different consumers being proportional to the Pearson Type III distribution. Twenty-five years later, Burrell and Cane (1982) presented a paper that used this distribution as the basis for the construction of a stochastic model for the circulation of library books, with each book circulating randomly at a given mean rate and the distribution of the mean rate of use over the book collection in the gamma form. Burrell and Cane found that the gamma Poisson model approximated Trueswell's 80/20 Rule under certain conditions.

In the light of the work of Ehrenberg, Burrell, and Cane, the potential of the gamma Poisson model became evident, when a Baton Rouge river boat recently admitted that 80% of its revenues comes from 20% of its customers. Since Bernoulli's Law of Large Numbers dictates that the percentage of loss on the amount wagered must ultimately equal the percentage of the riverboat's gross profit with a certainty of one on the condition of a sufficient number of trials, the admission appears to justify the characterization of the State's gambling revenues by a conservative Louisiana legislator as "a tax on stupidity," the gamma distribution ensuring that the tax is a steeply progressive one.

The other reason Urquhart's data fail to fit the Poisson may be traced to Thorndike's requirement that the trials be "independent" in the sense that "the result of one trial (or sample) does not affect the occurrence of the event in any other trial (or sample)." In my opinion, this condition does not hold for most phenomena in library and information science. In particular, it does not hold for library uses because of the following factor. When library materials are borrowed, the patrons borrowing them form opinions; and these opinions are communicated to other patrons, raising and lowering the probabilities of the materials being borrowed again. The process of one trial affecting the outcome of another trial is encompassed in the concept of "contagion." At this point, I must admit that I am somewhat confused about the logical implementation of contagion because of these reasons: (1) it is sometimes impossible to distinguish the effect of contagion from that of inhomogeneity; and (2) the process of contagion in library and information science is counteracted by the growing obsolescence of literature as it ages, especially in the sciences.

The first problem was posed by Feller (1943) in his classic conundrum about "true contagion" and "apparent contagion." Feller noted that this conundrum first arose in the paper by Greenwood and Yule (1920), in which they developed the gamma Poisson version of the negative binomial. Feller considered this version a case of "apparent contagion," because it was based on the inhomogeneity of the observations in the set, and not upon one trial affecting another. However, in this paper Greenwood and Yule (1920, pp. 258–264) also experimented with a model involving "the assumption that the happening of the event not only improves the prospects of the successful candidates but militates against the chances of those who had hitherto failed" (p. 264). They developed a mathematical solution for this problem, but by their own admission their solution was not in a form very suitable for computation. Describing this model as "true contagion," Feller ascribed the complexity of their formulas to the very general nature of their scheme. He then pointed out that on the basis of a special model of true contagion, which turned out to be the simplest case of the more generalized Greenwood and Yule scheme,

George Pólya was led to exactly the same distribution that Greenwood and Yule had constructed on the basis of inhomogeneity. The contagion model used by Pólya was an urn model with balls of two different colors, where the drawing of a ball of one color is followed by the replacement of this ball along with more balls of the same color, thereby increasing the probability of drawing this color, and decreasing the probability of drawing the other color. As a result of Feller's finding, if one encounters the negative binomial, one does not know from which process it arose—whether inhomogeneity or contagion—and there is a good possibility in library and information science that it could have arisen by both stochastic processes operating simultaneously and interactively.

The difficulties such a situation can cause for the modeling of library use emerges as soon as one approaches the problem of the obsolescence of literature over time. This is evident in the studies of monograph use described below, and the factors taken into account in these studies are also operative in the use of journal backfiles. Basing himself on inhomogeneity in his studies of monograph use, Burrell (1985, 1986, 1987) built into his gamma Poisson model an aging factor, by which the desirability of the monographs measured by the gamma distribution decays exponentially at the same rate for all monographs, leading to stable distributions over time and a growing zero class as other monographs merge with those already there because of their initial zero desirability. Moreover, in their study of monograph use, Fussler and Simon (1969, pp. 5–7, 142–143, 187–188) assumed that, within a specified period of time, the use of a particular book was entirely a random process, independent from its use in a previous time period and dependent only upon the underlying probability estimated by observing the use of a group of books with common characteristics. Fussler and Simon tested for "contagion," (i.e., whether the use of a book in one year substantially raises the probability that it will be used in the next year), leading to a group of books showing progressively greater uses over the years. Their assumption of independence of use from one time period to another seemed to be supported by the data, but they admit that their test was made imperfect by the overall decrease in the use of books. Fussler and Simon also appear to negate their own conclusion with the statement that "a not unexpected though crucial finding was that past use over a sufficiently long period is an excellent and by far the best predictor of future use" (p. 143). As a sign of the possible operation of both inhomogeneity and contagion, Fussler and Simon noted that "the observed distributions do not resemble the Poisson closely but have a much higher variance" (p.187).

All this leaves me scratching my head as to the influence of contagion. Is use of library materials over time in relation to each other determined solely by their initial desirability? Or does contagion act by slowing the inevitable obsolescence of some materials as against others? Or is there a time definition problem, and some materials would show first increasing use because of contagion and then decreasing use because of obsolescence slowed by contagion? Then, too, what creates desirability in the first place—inhomogeneities in quality, as judged independently by individuals, or some form of social communication?

Although contagion must play some sort of role, the problem of obsolescence does seem simpler to handle logically the way Burrell did i.e., on the basis of inhomogeneity without taking into account the influence of contagion. Moreover, this approach can be integrated with the transition from the negative binomial to the Poisson through the views of Borel (1950/1963, pp. 61–66) on entropy. Entropy is a measure of the amount of disorder or randomness in a system, and Borel equated it with homogeneity. In his view, inhomogeneity represents the opposite or order. According to Borel, order is less probable than disorder, as it requires work or energy for its creation, and he declared that "vital phenomena usually consist in the creation of states not considered probable" (p. 64). From this standpoint, he defined the guiding principle of entropy in the following manner: "The principle of entropy. . . states that a closed system—i.e., one that does not receive any energy from an external source—will necessarily evolve by passing from less probable states toward more probable ones" (p. 61). Analyzing this principle on the cosmic scale and rephrasing it in terms of inhomogeneity vs. homogeneity, Borel concluded that "the universe is steadily evolving from the heterogeneous toward the homogeneous, i.e., from a relatively ordered state toward a state that is more and more disordered. . ." (p. 66).

An application of Borel's views on entropy to Burrell's gamma Poisson model leads to the hypothesis that, as the desirability of a set of library materials decreases with age and the set receives less and less outside energy, its use should decline, with an ever diminishing mean and the collapse of the variance toward this mean. When the variance equals the mean, the materials could be considered in a state of equilibrium marked by rare and random events. It should then be feasible to shift these materials to some form of remote storage.

## Final Considerations and a Practitioner Recommendation

I will not discuss in this paper the debates provoked by the introduction of the negative binomial into library and information science. This topic is thoroughly covered in Bensman and Wilder (1998, pp. 161–171). Here I only want to emphasize three general conclusions, at which I arrived as a result of research on the scientific information market and testing the National Research Council database.

First, the skewed distributions found in library and information science and described by empirical informetric laws are not unusual. The discovery of these laws was only a part of a broad process of uncovering the skewed distributions underlying phenomena in many other disciplines

that took place after Pearson's devastating assault on the normal paradigm. As a matter of fact, the discovery of these skewed distributions was taking place even before Pearson. For example, the doctrine of Karl Marx with its concentration of the means of production and impoverishment of the masses can be considered in many respects as the drawing of wrong conclusions from a correct observation that the stochastic processes operative in the negative binomial are operative in human society. In the paper in which he derived the first informetric law—the Inverse Square Law of Scientific Productivity— Lotka (1926) was well aware that he had found nothing unusual. Thus, he wrote:

> Frequency distributions of the general type (1) have a wide range of applicability to a variety of phenomena, and the mere form of such a distribution throws little or no light on the underlying physical relations. (p. 323)

To back up these statements, Lotka cited the work of Corrado Gini on the inequality of income within a population and John Christopher Willis on the distribution of species. Perhaps the distinguishing feature of frequency distributions within library and information science is the fuzzy nature of the sets, within which they arise. This fuzziness is a function of the way disciplines overlap and share the same literature. From this perspective, the two most important informetric laws, which set apart library and information science from other disciplines, are Bradford's Law of Scattering and Garfield's Law of Concentration.

Second, the working out of the probability distributions and stochastic processes underlying library and information science is primarily the accomplishment of the British and those working in the British tradition. Therefore, it is to the scientific history of Britain, to which one must turn in order to gain an insight into the nature of these distributions. As further proof of the need to study British work, it should be noted that Garfield derived his Law of Concentration, which underlies the operations of the Institute for Scientific Information, off Bradford's Law of Scattering by transposing the latter from the level of a single discipline to that of science as a whole. This need holds true for the introduction of the negative binomial into library and information science, which was primarily the work of Burrell at the University of Manchester as well as of Tague and Ravichandra Rao at the University of Western Ontario. The debate over the negative binomial in library and information science was a British and Canadian one, in which the Americans were for the most part silent. I came to the negative binomial through an observation by Price, who was born in London and received his doctorates at the University of London and Cambridge though he taught at Yale, that it was the model for the social stratification of science posited by Robert Merton. As a result of this, the Bensman and Wilder (1998) paper can be regarded as an attempt to merge library and information science with the American sociology of Robert Merton and the British biometrics of Karl Pearson.

And, finally, as a result of my reading about probability distributions and working with them, I have come to adopt what Särndal (1977) has described as the "agnosticism" of the modern statistician in such matters. In a telling passage, Särndal wrote:

> At the end of the nineteenth century a rapid development occurred. The illusion of the normal law as a universal law of errors was shattered. Instead, there emerged conceptions of data distributions that have the characteristics of agnosticism still prevailing today, as shown in our reluctance, or even inability, to state that given data are distributed according to a normal law, or any other narrowly specified law for that matter. If we do adopt a certain distributional law as a working model, we usually acknowledge the fact that this may be just an adequate approximation. Or, as in modern studies of robustness, we may assume only that the true distribution is a member of some family of distributions, or we take yet a further step and assume a nonparametric model. (p. 420)

My statistical adviser at the LSU Department of Experimental Statistics put the case more succinctly: "We have in our computer lab thirty computers, each named after a different distribution, and, if we ever get the money to buy thirty more computers, I am sure that we will have no problem in coming up with another thirty distributions, after which to name the new ones."

I was led to the agnostic view by the realization that there are two major difficulties in fitting library and information science data to precise theoretical distributions. First, subtle changes in conditions can lead to shifts in the underlying stochastic processes being modeled. Second, the numbers we obtain are a function of the logical structure of our sets, and because of their inherent fuzziness, library and information science sets are full of contaminants resulting in outliers that distort the mathematical parameters of any theoretical model. Moreover, much of what I read on probability distributions appeared to deal with mathematical refinements that would be overwhelmed by the faults inherent in library and information science data.

Because of my conversion to agnosticism, I came to the conclusion that all that is necessary for practical statistical research in library and information science is to follow these simple procedures: (1) utilize the index of dispersion test that determines whether you are dealing with the Poisson by comparing the variance to the mean (Elliott, 1977, pp. 40–44, 73–75); (2) if the variance is found significantly greater than the mean—and it almost invariably is—assume that you are dealing with the negative binomial or a distribution closely akin to it; (3) perform the appropriate logarithmic transformations of your data to approximate the correct law of error; and (4) proceed to analyze the questions that really interest you.

## References

Bensman, S.J. (1982). Bibliometric laws and library use as social phenomena. Library Research, 4, 279–312.

Bensman, S.J. (1985). Journal collection management as a cumulative advantage process. College & Research Libraries, 46, 13–29.
Bensman, S.J. (1996). The structure of the library market for scientific journals: The case of chemistry. Library Resources & Technical Services, 40, 145–170.
Bensman, S.J., & Wilder, S.J. (1998). Scientific and technical serials holdings optimization in an inefficient market: A LSU Serial Redesign Project exercise. Library Resources & Technical Services, 42, 147–242.
Bookstein, A. (1990). Informetric distributions, Part I: Unified overview; Part II: Resilience to ambiguity. Journal of the American Society for Information Science, 41, 368–386.
Bookstein, A. (1995). Ambiguity in measurement of social science phenomena. In M.E.D. Koenig & A. Bookstein (Eds.), Fifth International Conference of the International Society for Scientometrics and Informetrics (pp.73–82). Medford, NJ: Learned Information.
Bookstein, A. (1997). Informetric distributions, Part III: Ambiguity and randomness. Journal of the American Society for Information Science, 48, 2–10.
Borel, E. (1962). Probabilities and life (translated by M. Baudin). New York: Dover Publications (originally published in 1943).
Borel, E. (1963). Probability and certainty (translated by D. Scott). New York: Walker (originally published in 1950).
Bradford, S.C. (1934). Sources of information on specific subjects. Engineering, 137, 85–86.
Brookes, B.C. (1977). Theory of the Bradford Law. Journal of Documentation, 33, 180–209.
Brookes, B.C. (1979). The Bradford Law: A new calculus for the social sciences? Journal of the American Society for Information Science, 30, 233–234.
Brookes, B.C. (1984). Ranking techniques and the empirical log law. Information Processing & Management, 20, 37–46.
Brown, R., & Herrnstein, R.J. (1975). Psychology. Boston: Little, Brown.
Burrell, Q.L. (1985). A note on ageing in a library circulation model. Journal of Documentation, 41, 100–115.
Burrell, Q.L. (1986). A second note on ageing in a library circulation model: The correlation structure. Journal of Documentation, 42, 114–118.
Burrell, Q.L. (1987). A third note on ageing in a library circulation model: Applications to future use and relegation. Journal of Documentation, 43, 24–45.
Burrell, Q.L., & and Cane, V.R. (1982). The analysis of library data. Journal of the Royal Statistical Society, Series A (General), 145, 439–471.
Coleman, J.S. (1964). Introduction to mathematical sociology. Glencoe: The Free Press.
Crathorne, A.R. (1928). The Law of Small Numbers. American Mathematical Monthly, 35, 169–175.
Ehrenberg, A.S.C. (1959). The pattern of consumer purchases. Applied Statistics, 3, 26–41.
Elliott, J.M. (1977). Some methods for the statistical analysis of samples of benthic invertebrates (2nd ed.). Freshwater Biological Association scientific publication, no. 25. Ambleside, England: Freshwater Biological Association.
Feller, W. (1943). On a general class of "contagious" distributions. Annals of Mathematical Statistics, 14, 389–400.
Fraser, S. (Ed.) (1995). The bell curve wars: Race, intelligence, and the future of America. New York: BasicBooks.
Fussler, H.H., & Simon, J.L. (1969). Patterns in the use of books in large research libraries. Chicago: University of Chicago Press.
Galton, F. (1879). The geometric mean in vital and social statistics. Proceedings of the Royal Society, 29, 365–367.
Galton, F. (1883). Inquiries into human faculty and its development. New York: Macmillan.
Galton, F. (1889). Natural inheritance. London: Macmillan.
Galton, F. (1978). Hereditary genius: An inquiry into its laws and consequences. London: J. Friedman.
Garfield, E. (1971). The mystery of the transposed journal lists—wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. Current Contents, 3(33), 5–6.
Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178, 471–479.
Greenwood, M., & Yule, G.U. (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 related accidents. Journal of the Royal Statistical Society, 83, 255–279.
Herrnstein, R.J., & Murray, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.
Hogben, L. Statistical theory: The relationship of probability, credibility and error. New York: W.W. Norton.
Keynes, J.M. (1921). A treatise on probability. London: Macmillan.
Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16, 317–323.
MacKenzie, D.A. (1981). Statistics in Britain, 1865–1930: The social construction of scientific knowledge. Edinburgh: Edinburgh University Press.
McAlister, D. (1879). The Law of the Geometric Mean. Proceedings of the Royal Society, 29, 367–376.
Newbold, E.M. (1927). Practical applications of the statistics of repeated events particularly to industrial accidents. Journal of the Royal Statistical Society, 90, 487–547.
Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. Annals of Mathematical Statistics, 10, 35–57.
Pearson, E.S. (1970). Some incidents in the early history of biometry and statistics. In E.S. Pearson & M. Kendall (Eds.), Studies in the history of statistics and probability (Vol. 1, pp. 323–338). London: C. Griffin.
Pearson, K. (1905/1906). "Das Fehlergesetz und seine Verallgemeinerungen": A rejoinder. Biometrika, 4, 169–212.
Pearson, K. (1915/1917). On certain types of compound frequency distributions in which the components can be individually described by binomial series. Biometrika, 11, 139–144.
Pearson, K. (1956a). Contributions to the mathematical theory of evolution. In Karl Pearson's early statistical papers (pp.1–40). Cambridge: Cambridge University Press.
Pearson, K. (1956b). Contributions to the mathematical theory of evolution II: Skew variation in homogeneous material. In Karl Pearson's early statistical papers (pp. 41–112). Cambridge: Cambridge University Press.
Pearson, K. (1956c). Mathematical contributions to the theory of evolution XIX: Second supplement to a memoir on skew variation. In Karl Pearson's early statistical papers (pp. 529–557). Cambridge: Cambridge University Press.
Pearson, K. (1956d). 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. In Karl Pearson's early statistical papers (pp. 339–357). Cambridge: Cambridge University Press.
Price, D.J.d.S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27, 292–306.
The Probabilistic revolution (1987). Cambridge, MA: MIT Press.
Särndal, C.-E. (1977). The hypothesis of elementary errors and the Scandinavian school in statistical theory. In M. Kendall & R.L. Plackett (Eds.), Studies in the history of statistics and probability (Vol. 2, pp. 419–435). New York: Macmillan.
Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. Proceedings of the Institute of Radio Engineers, 45, 279–290.
Snedecor, G.W., & Cochran, W.G. (1989). Statistical methods (8th ed.). Ames: Iowa State University Press.

Stigler, S.M. (1982). Poisson on the Poisson distribution. Statistics & Probability Letters, 1, 32–35.
Stigler, S.M. (1986). The history of statistics: The measurement of uncertainty before 1900. Cambridge, MA: The Belknap Press of Harvard University Press.
Thorndike, F. (1926). Applications of Poisson's probability summation. Bell Technical Journal, 5, 604–624.
Trueswell, R.L. (1969). Some behavioral patterns of library users: The 80/20 Rule. Wilson Library Bulletin, 43, 458–461.
Urquhart, D.J. (1959). The use of scientific periodicals. In Proceedings of the International Conference on Scientific Information, Washington, D.C., November 16–21, 1958 (Vol. 1, pp. 287–300). Washington, D.C.: National Academy of Sciences, National Research Council.
Urquhart, D.J. (1981). The principles of librarianship. Metuchen, NJ: Scarecrow Press.
Urquhart, D.J., & Bunn, R.M. (1959). A national loan policy for scientific serials. Journal of Documentation, 15, 21–37.
Weldon, W. F. R. (1893). On certain correlated variations in *Carcinus moenas*. Proceedings of the Royal Society of London, 54, 318–329.
Westergaard, H. (1968). Contributions to the history of statistics. New York: Agathon Press.
Winsor, C.P. (1947). Das Gesetz der kleinen Zahlen. Human Biology, 19, 154–161.
Yule, G.U. (1910). On the distribution of deaths with age when the causes of death act cumulatively, and similar frequency distributions. Journal of the Royal Statistical Society, 73 (New Series), 26–38.