DOCUMENT RESUME

ED 060 921 52 LI 003 612

AUTHOR Resnikoff, H. L.; Dolby, J. L.
TITLE Access; A Study of Information Storage and Retrieval With Emphasis on Library Information Systems. Final Report.
INSTITUTION R and D Consultants Co., Los Altos, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-8-0548
PUB DATE Mar 72
CONTRACT OEC-0-9-140548-2791(095)
NOTE 280p.; (118 References)

EDRS PRICE MF-$0.65 HC-$9.87
DESCRIPTORS Costs; *Information Retrieval; *Information Storage; *Information Systems; *Library Automation; Library Collections; *Models; Use Studies

ABSTRACT

It is the purpose of this study to provide fresh insight into the nature of library problems by systematically studying the question of size in various information contexts. In the introduction, the role that size plays from the user's point of view is illustrated. In the following sections, a study of the card catalogue, the classification system and various other access mechanisms are presented. Their size characteristics are determined and the impact of these considerations on the creation and use of access mechanisms is shown. The several chapters that follow are devoted to the more extensive statistical and mathematical justifications necessary to provide a solid base for future study improvement, and design of information access systems. (Author/NH)

FINAL REPORT

PROJECT NO. 8-0548

CONTRACT NO. OEC-0-9-140548-2791(095)

ACCESS

A STUDY OF INFORMATION STORAGE AND RETRIEVAL WITH EMPHASIS ON LIBRARY INFORMATION SYSTEMS

H.L. Resnikoff and J.L. Dolby

R & D CONSULTANTS COMPANY

Los Altos, California and Houston, Texas

March 1972

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U. S. Department of

HELATH, EDUCATION, AND WELFARE

Office of Education

Bureau of Research

PREFACE

This monograph describes the work performed by R & D Consultants Company under Contract #OEC-0-9-140548-2791(095) with the Office of Education of the Department of Health, Education, and Welfare. The contract is titled "A Computer-aided Study of Access Management and Collection Management in Libraries"; its principal objectives are the development of a model for information access and storage systems, and the study of the structure of existing access systems with the intent of augmenting them in significantly useful ways by means of automated processing of machine readable data bases.

The specification of such a model naturally requires considerable mathematical and statistical detail that makes for dry reading at best. We have therefore prepared a rather extensive introduction that summarizes the findings with only a minimum of documentation and then provided the necessary backup in the following chapters. In addition to the material contained herein, the contract called for a study of computer programming languages as they apply to problems in the library. At the invitation of the Editor of the Journal of Documentation and with the permission of the contract officer, this study was published in the June 1971 issue of that journal under the title:

> PROGRESS IN DOCUMENTATION: Programming Languages in Mechanized Documentation.

Throughout the course of this study we have been indebted to Mr. Lawrence S. Papier of the Office of Education who has provided many helpful suggestions both with regard to the plan of our research and the problems of documenting the results.

The authors also wish to express their appreciation to Richard O'Keefe and other members of the library staff of the Fondren Library, Rice University for their generous help and cooperation in the selection of the Fondren Index Sample which provides the central data base of this study. We are also indebted to Richard De Gennaro and Foster Palmer of the Harvard University Library for making available information about the contents of the Widener Shelflist which enabled us to determine the dynamic structure of their classification system and also for the five year summary of their circulation statistics; to the late Gerald Mitchell of the Institute for Defense Analysis who aided us in the preparation of the distribution of digraphs; to the Conference Board of the Mathematical Sciences, and particularly the NISIMS Committee, who supported those aspects of this

work particularly concerned with accessing mathematical archives; to John W. Tukey, the Statistical Research Techniques Group of Princeton University and the National Science Foundation who supported the work on algorithmic indexing and made available for this study preliminary output from their permuted title listings of the retrospective file of statistical papers; and to M. L. Puri, Department of Mathematics, Indiana University, for his thoughtful contributions to the study of the mathematical models of access systems.

Finally, we should like to acknowledge the contributions of the staff of R & D Consultants Company; William E. Houchin, particularly for his work on the information theoretic aspects of the problem; Val Forsyth for her invaluable contributions to the overall data handling problems; and to Joan Resnikoff and Rena Wells for their painstaking efforts in analysing in fine detail the index structure of the Fondren Index Sample.

March 1972

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIBRARY ACCESS SYSTEMS

INTRODUCTION

For some years many observers on the information science scene have been commenting on the "information explosion" and the effect this has on the librarian and on the library user. The fundamental assertion is quite simple: libraries grow exponentially. It is easy to show that this phenomena has persisted at least since Gutenburg. As long as the base is very small, exponential growth can be coped with. Sooner or later, however, repeated doubling of even a very small base every 20 or 30 years will lead to a very large base. When it becomes clear that an information base is already so large as to strain our ability to control and direct it, doubling its present size within another 20 or 30 years can only be viewed with considerable concern.

Whether library collections have reached such a stage at the present time is debatable. Some arguments have been put forth in recent years to the effect that a universal collection is now obsolete; that even the largest public and university libraries will soon have to move towards specialization of their collections and increased dependence on each other to achieve comprehensive coverage. Nevertheless, many large libraries continue to grow at their accustomed rate and new libraries of increased capacity continue to be built.

We raise the question here not with the hopes of resolving it, but rather to emphasize a self evident point: the size of a library collection is of fundamental importance. This should not be construed as implying that the quality of a collection is unimportant, but to stress that there are a number of basic problems concerning libraries that depend almost totally on questions of size rather than quality, however quality may be measured.

That collection size is important to a user may be simply illustrated by a rather mundane set of examples. Consider, for instance, a collection of two or three dozen books on a desk. Clearly, the arrangement of such a collection is of no importance whatsoever. The scanning speed of the normal human eye and the recognition mechanism of the brain is so fast that one can locate a desired book even while the arm mechanism is reaching forward to retrieve it. However, if we consider the 800-1,000 books that one can comfortably store on shelves on one wall of a modest size office, some degree of organization becomes necessary. In such a collection, physical size normally plays an important role as grouping of books by size makes for more efficient use of space. But size is also an important visual key for locating a book. Equally, color is useful both for aesthetic considerations and location keys. Size and color are generally not incompatible with a rough subject grouping, particularly if the collection consists of one or more series of publications.

At this level it may also be useful to note that there is a kind of Parkinsonian law in operation: the size of a collection expands rapidly to fill the available shelf space. Thus a gap of open shelf in such a collection is more likely to indicate the absence of a book rather than deliberate planning for future expansion, thereby providing an elemental circulation system including a simple mechanism for determining the spot for returning the book after use. Collection growth is normally handled by temporary storage on desk and table space until there is sufficient incentive to add a new shelf or set of shelves at which time "the new books" are not infrequently all shelved together, thus providing another retrieval key for the personal access system: time of acquisition.

Unprofessional though it may appear in terms of professional librarianship, such a system is both effective and cost-effective. It is designed for the use of only a few people--perhaps only one--and it is presumed that these users are intimately familiar with the system. Periodic rejuggling of the storage positions is not only costly, but detrimental: it breaks down the simple access system that is quite capable of remembering that the needed document is that "medium sized blue book with the red stripe on the fourth shelf near the door." No catalog system in the world can beat that kind of retrieval speed.

When we move up to the 25-30,000 books normally found in a small public or college library, the access system must become more formal if for no other reason than the fact that there will be many more users, including a host of infrequent ones who must operate with reasonably simple instructions. At this level, only a handful of users (and certainly not all of the library staff) will be intimately familiar with the entire collection. This is not to say that personal knowledge of the collection is unimportant or that individual variations in the ways in which the books are shelved do not exist. Every experienced library user knows that the fastest way to determine if a book is in the collection, and if so where it is to be found, is to ask the librarian. In fact this is so well known that all librarians develop subtle and not so subtle techniques for fending off such requests both to preserve their sanity and to give themselves some time to attend to their other duties.

However, a librarian who is never willing to guide a user to a book does not recognize the nature of the system. All modest size libraries vary from standard cataloguing practice in certain ways if only to keep cataloguing costs in line. Standardization is mainly useful to the user who moves about from one library to another over a period of time and does not wish to invest the time necessary to acquaint himself with the vagaries of a particular library the first time he has need of its contents. For such a user, personal direction is of great value.

The progression on to larger and larger collections creates more and more complex access problems. A large university library is so complex that no one librarian is able to personally familiarize himself with all of it. Instead, a staff of reference librarians is maintained, each covering different specialties. The complex access system is now so large that the user may need to consult the librarian to find an item in the access system where in a smaller library he could expect the same effort to provide him with the book itself.

In brief, every library user quickly learns that size is a barrier to access and that his best strategy in trying to locate a book is to head for the smallest collection that is likely to contain or point to that book. According to this standard, a sophisticated user is one who can exercise good judgement in this regard. The primary rule presumably is that the older (and/or rarer) the book, the more likely it is that one will have to go to a large collection. Other properties of the document such as language, place of imprint, subject matter, etc. clearly enter into the exercise of this judgement. It is curious that libraries do not, in general, provide detailed information of this kind about their holdings so that users can exercise this judgement more efficiently. Precise counts from the card catalog would, of course, be costly to obtain and many librarians might be loathe to publish their opinions about the approximate breakdown of their holdings by language, place of publication, etc., but these factors may not outweigh the utility of such descriptive information.

Deriving counts from a machine readable catalogue is quite simple and relatively inexpensive so it is to be hoped that as more libraries shift to machine cataloguing they will follow Harvard's lead in publishing refined descriptions of their holdings by language, date of imprint, subject, etc.

Up to this point we have constrained our discussion to the problem of finding a book within a set of books. Until recently, few would question that this was a fundamental problem in librarianship, if not the fundamental problem. Today some authors would prefer to view all requests placed at libraries as requests for information, many of which could be best served by supplying the information itself rather than by directing the user to a document containing the information. In the frame in which we view the problem this is equivalent to requiring a much larger access system than most libraries could currently afford. However, even assuming an increase in funds for libraries and/or a decrease in costs for access systems, it is still not clear that requests for books will disappear. If one wants to read Oliver Twist presumably nothing else will do and classifying such a request as an "information request" in the interests of obtaining a unified theory does little to change the problem. The user still wants the book. Nor is this phenomenom restricted to fiction. Even such simple requests as "what is the current

population of the United States" or "what is the most recent estimate of the speed of light" are frequently phrased in a context that demands not only a proper definition of the source, but also ancillary information about the methods used to obtain the estimate and the author's own views on the strengths and weaknesses of his procedures. In short, the user will frequently require access to the document containing the information and, in many cases, access to the supporting documents cited.

Nevertheless, there are many proper user requests that are shaped as requests for information rather than for specific documents and it is well to consider the effect of collection size in such a situation. Here again it is clear that a law of parsimony is in operation. In short, one does not approach the Library of Congress to determine the size of a badminton court. Or at least one should not. Not only librarians but many other information sources are continually plagued by questions that could be more efficiently answered by reference to the nearest desk-sized dictionary or one volume encyclopedia. It takes a patient member of a library staff to handle such requests in a manner that is likely to advance the user another step in user sophistication. The education of users is clearly a critical aspect in the effectiveness of any information system.

Even on the basis of elementary arguments it seems reasonable to conclude that size is indeed a critical factor in the evaluation of collections of books and documents. The larger the collection the more likely it will be that the needed information will exist in it, and the more difficult it will be to find it. It is only a short step from such an observation to the hypothesis that the size of the access system will also be of importance, and will also be most effectively used by resort to a law of parsimony. Indeed, it is essential to recognize that the access system itself is typically a collection of pieces of information, not just a set of pointers to an information collection.

It is customary to think of a catalogue card as a container for a collection of information about a book, including information about its location. However, it is more than this. It also contains a subset of the information in the book and, no matter how small, this is in fact information. And it may just be the information the user needs. Titles contain information. Contents notes contain information. It is not unusual to find information about the author which is not contained in the book itself. Further, the collection of catalogue cards provides information that few if any of its books are likely to contain. To the extent that the statistics are available, any description of a library serves also as a description of the community it serves, biased to be sure by the collective decisions of the acquisitions staff over a period of years, but still descriptive of qualities that are very difficult to study from any other source. Where the collection in a particular field is large enough to be considered representative, or even definitive, statistics on the holdings can

be most useful to an author making decisions on what to include in an introductory or expository work.

When one moves on to a study of other access devices such as indexes, abstracts, special bibliographies, permuted title lists, citation indexes, or cumulative lists of tables of contents, it is even easier to argue that each such device plays both roles: a container of information, and a pointer to other information. But for that matter, the book itself plays both roles; it not only contains information; it points to other containers of information through footnotes, citations, appended bibliographies, and remarks in the text itself. Thus a book is an access device as well as an information container.

What then is the fundamental difference between the book and the catalogue card, or the index and the table of contents? Both contain information; both point to information. The user, again operating under a principle of parsimony, goes to the smallest container, or set of containers, that is likely either to contain the information or point to a small set that does contain the information. The entire system operates under the fundamental assumption that the user is only willing to scan a certain amount of material to find what he wants. Both the author of the book and the author of the catalogue, in somewhat different ways, attempt to break down the sum total of knowledge into bite-sized pieces and organize those pieces in various orderings so that the user can thread his way through the maze to the bite that he needs.

Both questions of order and questions of size are of fundamental importance in any formal inquiry into the structure of information systems. A formal investigation into why certain orderings of information are useful and why others are not (or are of marginal utility) would require a much deeper understanding of the structure of information than is presently available. Ordering a library catalogue by author is presumably useful because almost all libraries do it. But trying to decide whether librarians do this because users remember authors or whether users remember authors because they know that authors are a useful access point in most catalogues is not likely, at least for the present, to bring us much closer to a proper understanding of how such systems work. We are therefore forced to take the view that a new ordering is by definition useful if some segment of the community is willing to pay for its initial production and maintenance.

It is in this context where we can see more clearly than in any other the potential impact of the use of computers in libraries. The great cost in the use of computers in this area is the cost of initial programming and the cost of data base acquisition. Marginal costs of producing new orderings of a data base are relatively small compared with the cost of obtaining the first ordering. The more the data base is "exploded," the smaller

the unit cost of material produced. Several examples should help to put the problem in perspective:

> Permuted Titles. The title is keyed once, with associated information about the author, source, etc. and then exploded by a factor of from 5 to 6 to produce access to each significant word in the title.
>
> MARC. The data is keyed once and then exploded by tape copying for use in many libraries and commercial firms, some of which explode the records again, e.g. for producing the several copies necessary for maintenance of their card catalogue.
>
> Widener Shelf List. The shelf list is keyed once, and then exploded at the first level by generation of the shelf list itself together with alphabetical and chronological listings of the same entries. A second level explosion occurs through listing through the machine (first by line printer, more recently by computer typesetting) and the printing of copies through normal book production.
>
> MEDLARS. The material is keyed once for production of Index Medicus and then exploded through on-line and batch processing of information retrieval requests.

Other examples involving the production of book catalogues for county library systems (where in many cases the explosion extends to copies for local schools), citation indexing, and various forms of union lists are now in fairly wide use.

The Widener Library Shelf List is of particular interest because it provides an important example of the interplay between size and ordering. The chronological listing represents a new ordering, at least for a collection of this size, and it will be of interest to assess its utility after a period of time. We shall later make use of this feature to study the dynamics of the classification system. The alphabetic listing is not new; indeed such listings go back to antiquity. Further, special listings for subcollections are probably nearly as old. However, the systematic listing by alphabet for each main category of the classification system for a system of this size is only possible with the machine help. Provision of this information in addition to the alphabetic listing in the public catalogue makes it possible for the user who has reason to believe that the material he is searching for is in, say, the American History class, to go directly to a much smaller collection for his search, with the attendant time savings. In other words, the machine not only provides the possibility of experimentation with new orderings, it also permits one to exercise access judgements of a variety of choices of the size of the traditional listings.

Such reorganization of information is not limited to computer dependent schemes as is evidenced by the recent popularity of the undergraduate library concept in schools with large main library holdings.

It is the purpose of this study to try to provide fresh insight into the nature of library problems by systematically studying the question of size in various information contexts. In this introduction we have tried to illustrate the role that size plays from the user's point of view. In the sections that follow we shall study the card catalogue, the classification system, and various other access mechanisms. We will determine their size characteristics and show the impact of these considerations on the creation and use of access mechanisms. Finally, we shall devote the several chapters that follow to the more extensive statistical and mathematical justification necessary to provide a solid base for future study, improvement, and design of information access systems.

THE CARD CATALOGUE

The primary library access device is the card catalogue. In simplest terms, the catalogue is a set of linear files: the shelf list, the subject heading file, the author or author-title file, the new accessions list, etc. Let us consider the problem of finding a particular entry of known form in one of these linear files. Clearly, if the file is of any length, it will be ordered by some filing rules (with which we will assume the user is familiar) and a superstructure of guides will be imposed to permit the user to move rapidly to the general area in which the required item is to be found.

The natural superstructure for a linear file is hierarchal; in this case taking the form of cabinets which contain drawers which in turn are partitioned into sets of cards further separated by file guides. The user first scans the cabinet labels to locate the right cabinet, then he scans the drawer labels to locate the proper drawer, then he scans the file guides to locate the correct subset of cards, and finally he scans the card headings individually to find the desired card. It is perhaps worth noting that many libraries neglect to provide cabinet labels that can be scanned in the first step, thus requiring the user to scan the relatively small drawer labels in order to locate the proper cabinet.

Several authors have studied the problem of determining optimal strategies for establishing the proper number of file guides, the proper size of a card drawer, etc. (See for example, Shoffner (1) and Lipetz and Song (2)). In the simplest case, if it be assumed that scanning speed (and hence, cost) is the same at every level of the access structure, it is easy to show that the optimal strategy is to design each level of the structure in such a way that it decomposes the next level into a set of file segments of equal size, say K segments, where K is independent of the level. In terms of the card catalogue, this would imply that we should have K cabinets, each consisting of K drawers, each containing K file guides, each of which serve as separators for precisely K catalogue cards. See Chapter III for the technical details.

Determination of the proper value of K is not so easy. If we totally neglect the cost of providing and maintaining the access structure and choose that value of K which minimizes the searcher's costs we find that K should be equal to the natural constant of the calculus, $e = 2.718\ldots$. As catalogue cards are integral units, we are forced to choose K as an integer value, either 2 or 3. The choice $K=2$ corresponds to a binary search, a procedure that is widely used in file searching in computers.

However, it is not reasonable to neglect the cost of providing and maintaining the access system. The smaller the value of K, the greater the cost of the access system. Formally, if S is the size of the linear file, then the size of the optimal level structured access system, A, is related to K and S by the simple formula

$$A = \frac{S - 1}{K - 1} .$$

This leads to a typical problem in optimization: As K decreases towards the natural constant e, scanning time (and hence cost) decreases, but access system costs increase, slowly at first and then rather rapidly. Thus there is presumably an optimal value of K that minimizes the total system cost, i.e. the sum of the user costs and the access system costs. In theory, these costs could be measured and an optimal value for K thereby determined. However, it is not easy to obtain such cost data--particularly those associated with user scan time--so we choose instead to adopt a standard procedure from the field of operations research where such questions occur routinely: we shall assume that current practice is constrained by economic restraints to be close to optimal and determine the value of K currently in use.

From elementary considerations, it is evident that the value of K currently in use is in the neighborhood of 30. File cabinet construction varies, but the most popular size is the 4 by 8 cabinet containing 32 drawers. However, as cabinets are normally placed side by each, the distinction of "cabinet" is largely lost from the visual point of view (perhaps explaining why so many libraries fail to provide large label designations for each cabinet). A more significant measure can be found by determining the average number of cards per drawer. For the Fondren Library at Rice University this was found to be 826 (see (3)). The drawer is, by itself, a two-level file consisting of cards and file guides. The earlier derivation for a N-level file reduces to the following result for a two-level file: the number of file guides should be equal to the square root of the number of cards. This is the main conclusion of the Lipetz and Song study (2). Now the square root of 826 is 28.74, again a number in the vicinity of 30.

Now consider a typical university library. In such a library the shelf list or other simple linear catalogue file* is a full four-level access system: cards, file cards, drawers, and cabinets. The mean size of a university library, computed from those listed in (4), is 757,354 volumes. As this is a four-level system, K is equal to the fourth root of 757,354 or K = 29.50, again a value very close to 30.

* Intermixed subject-title-author files naturally multiply the total number of cards by nearly 3 and modify the details but not the essential result of this argument.

It is tempting to choose 30 as the "natural constant for access systems" because this simplifies the problem of computing higher powers. However, the value

$$K = (2e)^2 = 29.55....$$

fits the data somewhat better and may provide a useful suggestion for an eventual derivation of this constant on information theoretic grounds. Table 1 provides a short table of the powers of K along with the corresponding access levels of a card catalogue:

TABLE 1

Powers of K and Sizes of
Card Catalogue Access Levels

Power or Level	K to that Power = No. of Units in Access Level	Access Level in Card Catalogues
1	29.55	Cabinet
2	873	Drawer
3	25803	File Guides
4	762483	Cards

Not every library fits this pattern precisely; indeed, if a library catalogue system has this particular structure at some time it will almost undoubtedly stay at that point for only a short time as its natural growth carries it beyond the size implied by Table 1. However, before discussing the distribution and dynamics of library size, it is well to consider the applicability of the level structured model we have thus far presented to other questions of information access.

In later chapters we provide the necessary statistical support for the following assertions:

1. An abstract is approximately 1/30th the size (in number of characters) of the technical paper it abstracts.
2. An index is approximately 1/30th the size (in characters) of the book it indexes.
3. The table of contents is typically 1/30th the size of the index.
4. The book title is approximately 1/30th the size of the table of contents.
5. The average number of characters in a book is very close to the average number of books in a university library.

Indeed, libraries themselves tend to appear, at least on the average, in sizes that are very close to the powers of K given in Table 1. At the lowest level, the encyclopedia plays the role of a mini-library of information and is typically the size of approximately 30 average books (though usually packaged in a somewhat smaller number for reasons of printing and binding economy). It is more difficult to measure the size of personal library collections, but it is not unreasonable to suggest that an average of 873 books is a good estimate. At the next level, we find that the average size of the 717 junior (or community) colleges listed in (4) is 22,635 volumes per library. This figure is somewhat smaller than the level 3 value of 25,803 given in Table 1 but that is not surprising in light of the very substantial growth in the number of junior colleges over the past decade. As the number of JC's stabilizes, the average size of JC libraries can be expected to increase to a value not far from the level 3 figure. Finally, we have already remarked, and will substantiate below, that the average size of the university library is very close to the 763,483 given for level 4 in Table 1.

Having assigned level 3 to the community college and level 4 to the university, we are left with the rather intriguing question of where to place the four year college. Some of the oldest and best known institutions of higher learning consider themselves to be four year schools and are so considered by the Office of Education which releases annual reports on various measures of activity in academic circles. However important these institutions may have been in the past and are now for the particular role that they play in the national system of higher education, it can no longer be said that the majority of college students attend four year colleges. Since World War II, state after state has converted its state college system to a state university system. With the enabling legislation passed in 1971, California is now in the process of removing the last few large public colleges (indeed the only ones remaining with more than 12,000 students) from the college category and placing them in the university category.

Similarly, more and more states are adopting, or expanding, systems of community (i.e., junior) colleges to increasingly provide lower division education to their residents in that form. In California this process is, perhaps, most advanced: over 800,000 students attend community colleges and nearly 400,000 students (under the new definitions) attend universities. By comparison, only a handful attend four year schools.

At the same time, the existence of the four year college provides an answer to the question posed earlier: if every library in a given subset is growing at, say, 5% per year, how can the average library size remain constant? The answer is that new institutions are introduced into the set periodically, usually with collections that are substantially smaller than the mean, thereby balancing the general growth of the existing institutional

collections. Eventual conversion of the California State Colleges to universities will increase the number of universities in the country (by present OE definitions) by nearly ten percent. However, at present the largest of these collections is at San Jose State College and that collection is slightly below the mean for all the existing universities.

At the upper end of the scale there exists the Library of Congress whose current holdings, according to the Annual Report of the Librarian, are about 20 times the mean size of a university library and currently growing at a rate that will reach $K^5 = 22,531,361$ in the middle of the next decade. It is not far-fetched to suggest that the largest university libraries are in the process of becoming regionally located national libraries, in character if not in name.

This in turn raises the question of how we are to determine when a library has passed from one category to the next. Mathematically, the natural way to define a boundary between two values on an exponentially increasing scale is to compute the geometric mean of the two values. Table 2 provides the mean for each level together with the upper and lower bounds for each level in Table 1 and the fifth level needed to encompass the Library of Congress, the New York Public Library, and the few university libraries with more than four million volumes.

TABLE 2

Ranges in the Level Structure
(in volumes)

Level n	Type	Number of Volumes Minimum	Mean = K^n	Maximum
1	Encyclopedia	5	30	161
2	Personal	161	873	4747
3	Jr. College	4747	25803	140266
4	University	140266	762483	4144851
5	National	4144851	22531361	122480276

We have reasonably good data to test this definition of boundary only for levels 3 and 4. For level 3, 25 of the community colleges have collections smaller than 4,747 volumes according to (4) and one has a collection slightly larger than 140,266. Thus a total of 26 or 3.6% lie outside the suggested bounds. For level 4, four universities have collections larger than the upper bound and one has a collection smaller than the lower bound yielding 2.5% outside the suggested bounds. Given the dynamic characteristics of the library situation and the inherent difficulty of determining membership in the various classes, this can only be categorized as rather good agreement with the model.

Our reliance to this point on the statistics of college and university libraries rather than public libraries is a result of statistical reporting practice rather than a predilliction for one over the other. In the college and university area there are two essential statistical advantages: the Office of Education provides an authoritative source for categorizing the various institutions under community college, four-year college, and university rubrics, and most reporting institutions provide statistical descriptions of their holdings by library (or, more properly, by campus). Statistics of public library collections are generally reported in terms of the total holdings of each public library system together with the number of branch libraries, which makes it difficult to attempt for the public libraries the kind of analysis made above for academic libraries.

We can, however, add an observation that tends to support the view that an analysis of public library statistics might lead to rather similar results. The largest public library system is the New York Public Library. According to the American Library Directory (1970-71) the then current holdings of the Research Branch of NYPL totaled 4,057,565 volumes - a figure very close to the upper bound given in Table 2 for level 4 collections.

Let us summarize for a moment. We began this discussion of card catalogue structure by following the lead of Shoffner, Lipetz, Song, et. al. and observing that the optimal file structure from the viewpoint of minimizing the time required to find something in a large linear file using hierarchal search techniques is to subdivide each level in the access system by a factor of K. If the cost of the access system is negligible K should be small, equal approximately to 2 or 3. However, access systems in libraries do cost money and although we are unable to supply a precise derivation of the appropriate value of K from cost considerations alone, it is possible to collect sufficient data on card catalogue structure to show that the operative value of K in modern libraries is approximately equal to 30. Since the available data on card catalogues is not as convincing as one might like, we provide further support for the value of K = 30 (the symbol = is a convenient abbreviation for the phrase "is approximately equal to") by first examining the ratios of various access system sizes to the sizes of the information base to which they provide access and then by examining in more detail the structure of the book collection itself. Taken in its entirety, this data provides a reasonably comfortable basis for our assertion that K = 30.

These results can be rephrased more formally. We have concluded that in studying the size of access systems the primary measure is that the size S_n of a n-level access system is given by the expression:

$$S_n = K^n$$

where K = 30. This simple mathematical formula is reminiscent of work in other fields, particularly studies of human response to sense inputs, e.g. sound, touch, etc. In such work it is common practice to replace formulas of the above type by the corresponding expression derived by taking logarithms of both sides. Thus

$$\log S_n = n \log K .$$

As K is a constant, so too is log K. Hence, log S_n is a homogeneous linear function of the level n.*

Figure 1 illustrates the use of logarithmic graph paper (with one scale transformed logarithmically on which the size vs. level relationship corresponds to a straight line. It shows the size breakdowns with maximum and minimum values for each level. Figure 1 also provides a visual demonstration of our earlier argument that the "natural" way to obtain the boundary points was to use the geometric mean: in Figure 1 the boundary points are now equally spaced on the transformed scale.

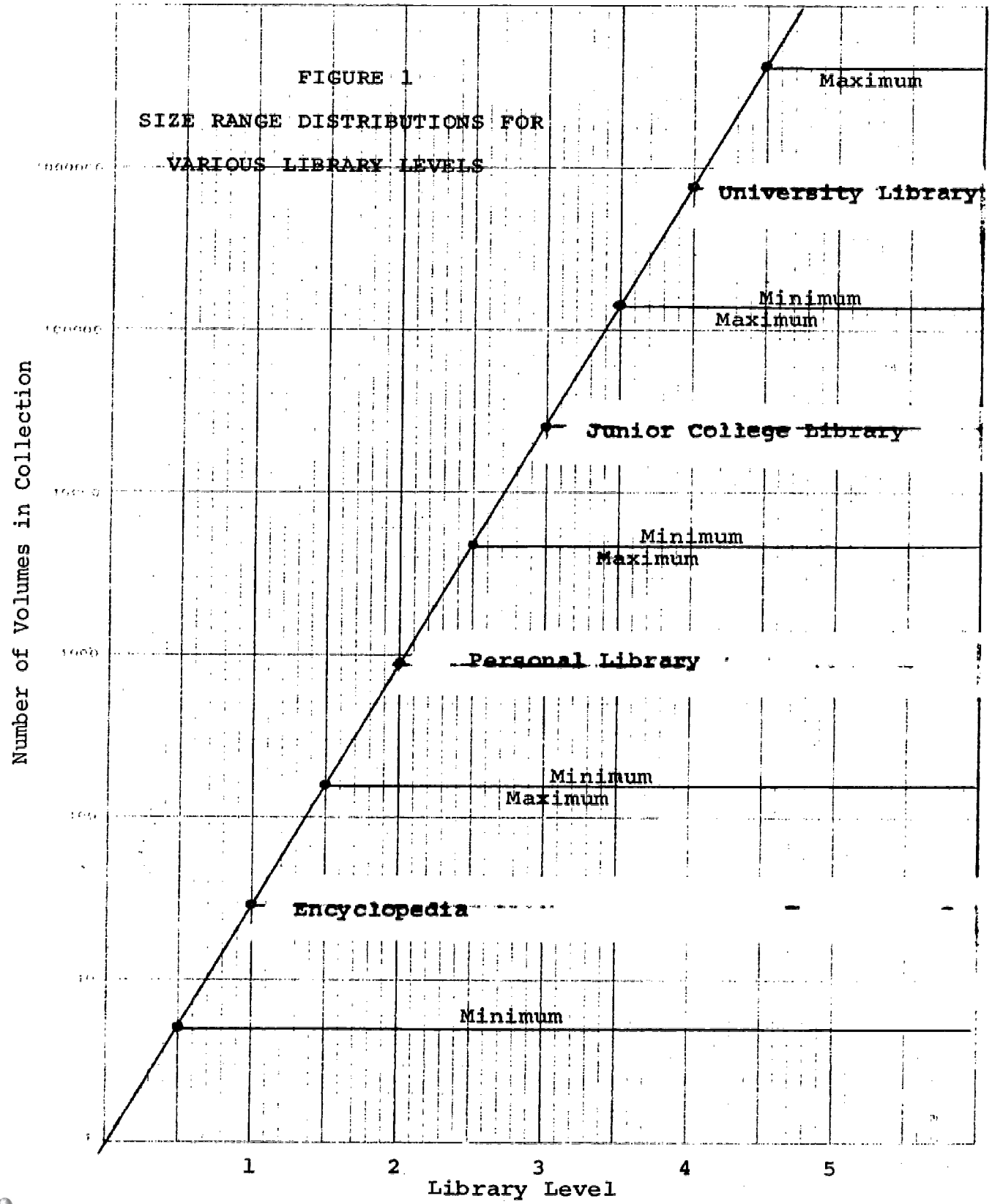
The case for the use of logarithms in sensory studies goes much deeper than this. Indeed, the original basis for the use of logarithmic scales was not that they induce convenient straight line representations (though they do) but rather that they simplify the task of describing the precision with which the human being can estimate the intensity of the sound, or other stimulus, he experiences. Early experiments showed that the human response mechanism operating over wide ranges of intensity of input could judge the intensity up to a fixed percentage of the input. Use of the logarithmic transformation enables one to convert such a statement to absolute rather than percentage terms. Thus, in accoustical work it becomes possible to report that the precision for a particular subject's estimate of sound intensity as plus or minus so many decibels (or fraction thereof) regardless of the intensity level at which the estimate is made.

With these observations in mind we posit that information, in addition to being stored in libraries, must also have a structure that is consistent with the human stimulus-response system. All stimuli are information bearing, but we are here concerned primarily with information that can be, and is, represented in linguistic form which minimizes the importance of the perceptual sense transducers such as the eye, ear, etc. for our concerns.

* In acoustical studies, for instance, it is common to express sound intensity levels in decibels and the decibel scale is nothing more, nor less, than a constant multiple of the logarithm of sound intensity.

FIGURE 1

SIZE RANGE DISTRIBUTIONS FOR VARIOUS LIBRARY LEVELS



The structure that information qua information appears to share with the sensory stimuli is the logarithmic relation noted above. "Intensity" of information is, in this view, to be measured by the number of information bearing units, e.g. by the number S_n of characters, of books, etc. The use of log S_n as a measure of "perceptual" size is then natural from a phychophysical viewpoint.

However, if we are to lean on the analogy with psychophysical results we must also determine if the main property that justified the use of logarithms--constancy of percentage variation rather than a constancy of absolute variation--also holds for information distributions. In order to determine whether this property holds for information distributions it is first useful to determine the mathematical form of the information distribution itself.

The study of information distributions dates back to the late nineteenth century at least and the work of Mendenhall (24), a physicist who devoted a considerable amount of effort to the accumulation of word length distributions for Shakespeare, Bacon, Marlowe, and others to see what light, if any, such studies might shed on the question of authorship of certain of Shakespeare's publications. A generation later Yule (25), although apparently unaware of Mendenhall's work, tried much the same approach to the authorship problem using instead sentence length distributions. Williams (26) summarized the results of these two studies and observed that both Mendenhall's data and Yule's data together with data he had collected could be accurately approximated by the lognormal distribution. In the terms we are using here this is equivalent to showing that the logarithm of size (logsize) is normally distributed.[1]

This result further strengthens our claim that the appropriate way to measure size as it affects users of information bases is

---

[1]According to (5) the first use of the lognormal distribution was made by McAlister (6) at the suggestion of Galton. Galton in turn "had derived his ideas from a consideration of the Weber-Fechner law relating responses to stimuli." McAlister presented his results to the Royal Society of London in 1879, some 8 years before Mendenhall wrote his paper on the "Characteristic curves of composition," and some 77 years before Williams noticed that Mendenhall's distributions were nicely fit by the lognormal, a situation which in and of itself may have something to say about the need for interdisciplinary information access. Indeed, the first mention we have been able to find of the relation between the Weber-Fechner results and the structure of information distributions is a passing remark made by Fairthorne in his summary (7) published in 1969, though the very brevity used in the remark suggests that the idea has been discussed for some time.

to use logsize, because the normal distribution is the first order approximation in the family of probability distributions just as the straight line is the first order approximation in the family of real analytic functions. It would be nice to be able to assert that information distributions share this property with other stimulus distributions. However, the work of Weber and Fechner predated the wide use of the probability distribution in data analysis and, considered from this standpoint, the question has apparently not been of much interest to more recent workers in the field.[2]

If logsize is normally distributed, at least to a first approximation, this permits one to represent a set of information distributions in a particularly simple form. Figure 2 presents a set of information distributions, plotted on log-normal probability graph paper, commercially available graph paper constructed so that a cumulative log-normal distribution will plot as a straight line. First observe that all of the lines are nearly straight, confirming that the distributions are approximately equally spaced, confirming, at least for this data, that the means of the size distributions occur naturally in powers of k. Finally, observe that the straight lines all have nearly the same slope. The slope of the line on lognormal probability paper is a linear function of the standard deviation of the lognormal distribution. Parallel lines correspond to distributions whose variation is constant on the logsize scale.

Figure 2 therefore provides a simple graphical representation of our first order approximation to a mathematical model for information distributions: a set of equally spaced, parallel lines on lognormal probability graph paper. We call this model a "level-structured model for access systems," and reiterate its principal advantages:

1. Logsize is a linear function of level

2. Logsize is normally distributed

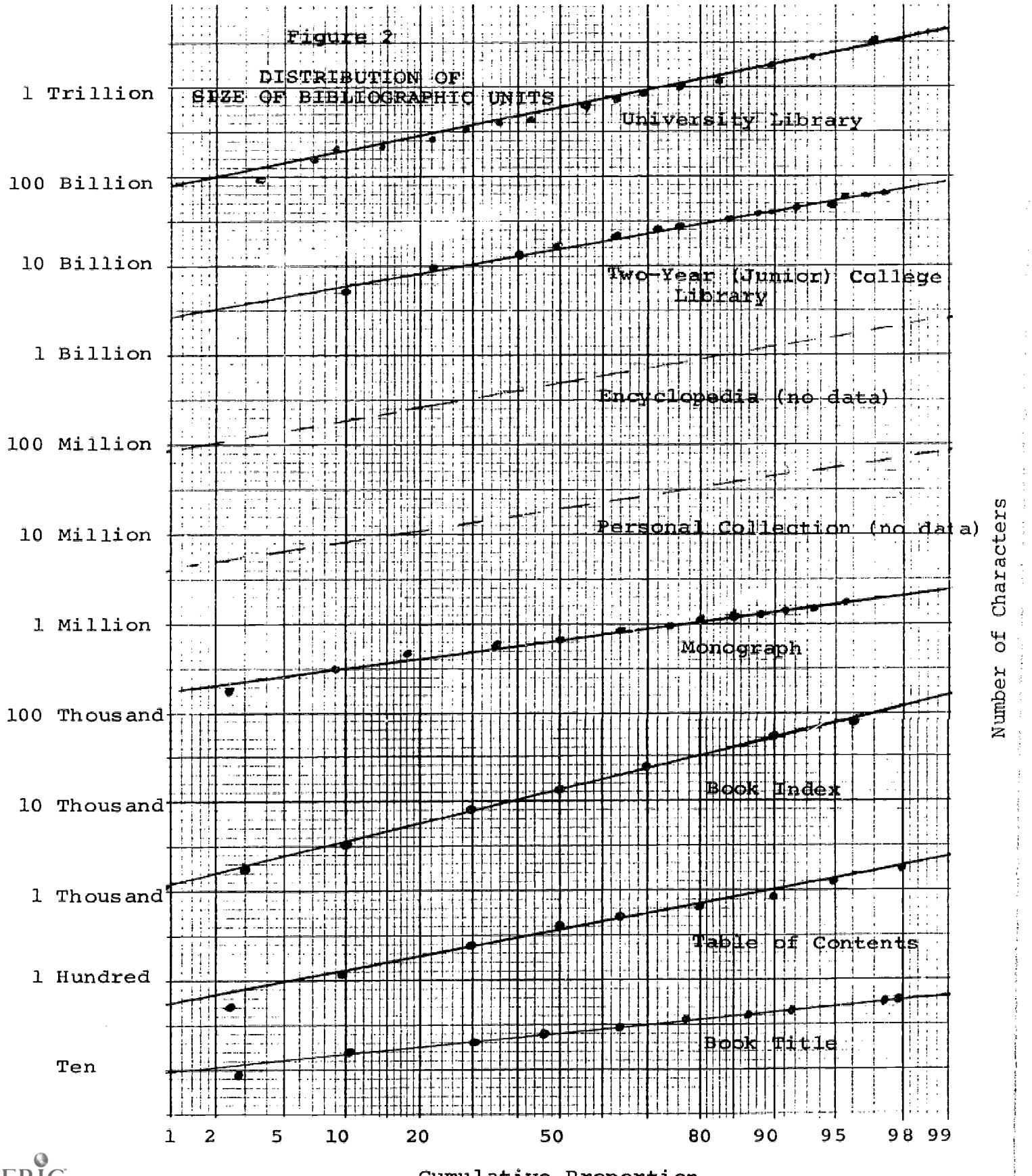
3. Logsize has a constant variance, independent of the mean of the distribution.

---

[2]One of the primary uses for lognormal probability paper is in "probit" analysis, a procedure widely used by biometricians, frequently in situations where one is interested in the effect of various potentially lethal doses of drugs on laboratory animals. Although to the layman it seems a bit curious to consider a lethal drug as a "stimulus," it is clear that the two problems are intimately related.

Figure 2

DISTRIBUTION OF SIZE OF BIBLIOGRAPHIC UNITS



In short, taking logarithms simultaneously linearizes the functions, normalizes the distribution, and stabilizes the variance, a triad of desirable properties that is frequently induced by the proper choice of transformations of data (see Tukey (8)).

Returning, then, to the original consideration of card catalogues, we must next see what, if anything, the logsize model tells us about the catalogue as an access device. The first aspect of the problem is to determine how the size of the catalogue compares with the size of the collection it accesses. This presents certain problems; in particular, how large is a card catalogue? The question is compounded by the mechanical constraints imposed by the card system itself. A catalogue is a set of files, rather than a single file. With modern duplication equipment it is common practice to make (or purchase) a set of identical cards for each book and then "head" these with the various entries (other than the main entry) under which they are to be filed. On the average, full cataloguing produces approximately 450 characters per card. If we count complete duplication of characters, 57 cards would be required for each book for a card catalogue that would provide as much access to each book as does a typical book index (in terms of number of characters). We call such a system a "first order access system". Examples already cited include the index to a book and the abstract of a journal article.

Under such a definition we must conclude that no card catalogue is a first order access system. By the same measure, a "second order access system" would require only two cards. Although estimates of the average number of cards per volume differ, it seems clear that the number must be closer to 2 than to 57, even on the logsize scale (where the boundary is 10 cards.) A book catalogue with full cataloguing for the main entry, short cataloguing for all other entries, and 4.4 cards per book would be an exact 2nd order access system if the short entries contained 132 characters on the average. This is not an unreasonable estimate. We therefore conclude that a card catalogue is a second order access system.

If this is, indeed, the case, one can then sympathize with the user who complains that the card catalogue is not an adequate access device: it is one order of magnitude too small. Of course, this is a simple reflection of the difficulty of expanding the file in a manual system not only in terms of the human labor necessary but also in terms of the sheer physical space it would require; it is clear that the order of 1/30 of the physical space necessary to house the collection itself is required to store the catalog.

Given a machine readable catalogue, it would be possible to generate 57 distinct images of the main entry or the somewhat larger number if we count the main entry as 450 characters and all other entries as a smaller number (such as the 132 characters

mentioned above). However, it is not at all clear that 57 orderings of the catalogue, or some similar explosion through composite use of title words, would really extend the utility of the catalogue by one order of magnitude. As space limitations dictate the use of book-form rather than card form at this level of bulk, the cost of computer printout tends to inhibit even experimental use of such access devices. COM devices alleviate the problem somewhat, but do not eliminate it.

It seems inherently more reasonable to suggest that the catalogue should continue to be viewed in the traditional way: it is a book finding device which is implemented in a manner which allows continued updating to provide near-current access to library holdings. As such, it operates as a second level access device and will apparently continue to do so.

First level access systems are, by definition, more costly by a factor of 30. To offset this substantial increase in price one must give up either continuous updating of the file or specific reference to one's own collection, or both. There is, of course, a long standing precedent for the second alternative in the serials field: abstracts (which are a first level access system) to technical papers are collected and published nationally for use by all libraries even though only a small handful may contain a nearly complete collection of the journals represented in the abstract publication.

Book abstracts are virtually unknown. However, a substantial (and probably increasing) proportion of the books published have indexes. The index provides first level subject access to the book. Consolidation of a set of indexes in a particular subject field would provide first level access to that set of books. If the set of books were so chosen as to represent a 1/K sample from a larger set of books in the same subject area, the consolidated index would be a second level access system for the larger collection. The latter course provides a useful opportunity to study the potential of consolidated indexes, as it reduces the cost of producing them by a factor of K. Should they provide a useful extension to the access system at this level, it would then be possible to experiment with yet larger consolidated indexes to see just how far one should go. We shall return to this question later in the discussion to examine the mechanical problems connected with the production of consolidated, or cumulative, indexes.

level structured model of the library access system provides undation for designing and measuring the performance of access tems. The following short list suggests some of the simpler ential applications.

1. Cabinet labels. All libraries using card catalogues and having sufficient volume to require more than one cabinet of card trays should provide large labels at the top of each cabinet (or section of approximately 30 trays) indicating the range of the alphabet contained in the cabinet. This is a trivial matter, but to the user it is just as important as the much more costly provision of file guides within the trays. Because it can be provided at 1/30th the cost it obviously should be done.

2. With improved transportation and increased urbanization (and suburbanization), most library users have access to several library collections in addition to their own personal collection. In such circumstances they need more information to enable them to determine "the smallest collection that is likely to contain the information currently needed." To satisfy this need, libraries should maintain and regularly disseminate detailed statistical summaries of their holdings in terms of the number of documents by broad subject classification, by language, and by time of publication. The regular publication of a "mini-catalogue" of the most frequently used documents would probably stimulate library use quite considerably. As short form entries would be desirable in such a publication, libraries with automated circulation systems could contruct this publication directly from their machine readable circulation records at very low cost. Such a catalogue should be 1/30th the size of the main entry section of the regular catalogue.

3. The Dewey Decimal System provides, in its most obvious use, a level structured subject access system with K = 10. Larger libraries are increasingly switching to the Library of Congress classification which, at least for the first two levels, provides a level structured subject access system with K equal to 22, which is more compatible with the level spacing of traditional access systems (K $\sim$ 30) which probably optimizes elicited information per unit cost. This observation may have some implications for future modifications of the subject classification system, e.g., through the enlargement of the alphabet from 26 to 30 and/or the further use of alphabetic representations in place of numeric representations to the right of the "decimal point" in present practice.

4. A typical monograph contains, within its covers, a three-level access system in addition to the main body of text:

the title, the table of contents, and the index. Title information, together with other title page specifications, is regularly published by both public institutions and profit-making organizations to enable libraries to build access systems for their holdings. Table of contents information is included in some catalogue entries and, particularly in the case of journal publications, is published both by public and profit-making companies to provide additional access to information stores. It seems only natural to carry the process one step further and produce cumulative indexes to further increase access to the same information stores.

5. Information access and transfer is not limited to libraries, or even to the printed word. It is perhaps not coincidental that in our one study of the distribution of four year college class size (detailed in a later chapter), we find a lognormal distribution with a mean of 29.32 students per class for a sample of over 3,000 classes. Further studies would obviously be desirable to determine the applicability of level-structured education models founded on maximizing educational "information" transfer per unit cost.

THE DYNAMICS OF CLASSIFICATION SYSTEMS

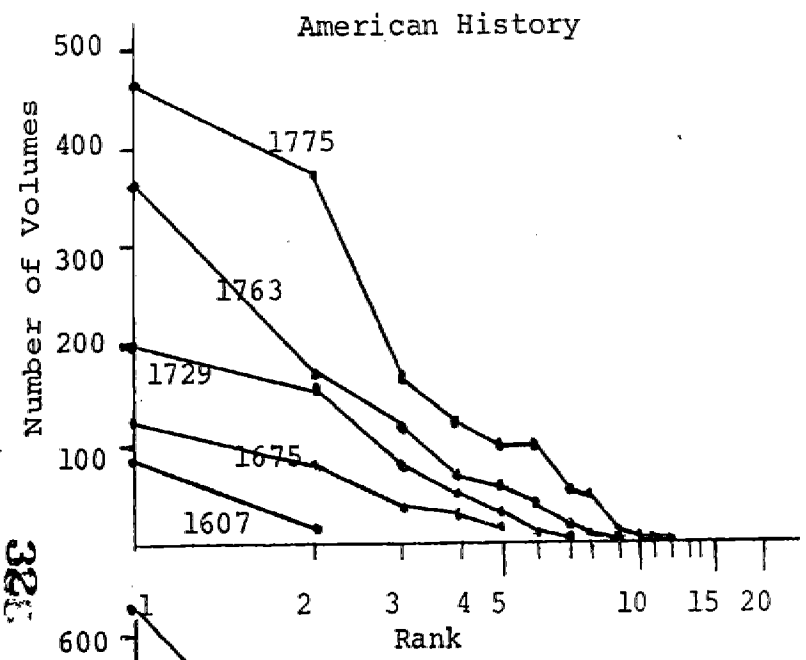
We noted at the beginning that growth is the fundamental problem in many libraries today: there is a continued in-pouring of materials that must be classified, shelved, circulated, and maintained. Clearly, any comprehensive attempt to model a library system must include consideration of growth and the way or ways that libraries have historically tried to cope with it. Many aspects of the problem can be posed in terms of the need for more money: if the shelves are crowded, build more shelves; if the circulation desk cannot cope, hire more clerks and automate the system; and so on. However, the orderly expansion of the classification system is not simply a matter of money (though that helps). It also involves the continuing need to revise and extend its structure so that it remains compatible with the intellectual content of the changing archive.

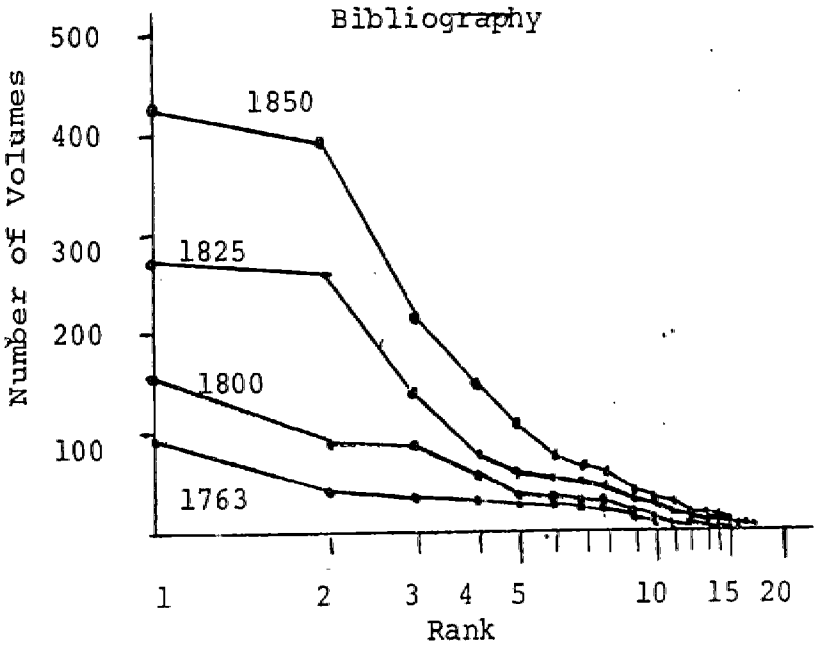
Some insight into the nature of this process can be obtained by examining the growth of the use of the various classification categories used in a particular library over a period of time. As an approximation to this we here refer to the collection of the Widener Library as exemplified in the chronological listings provided in the Widener Shelf List. Time, in this case, is measured by publication date rather than by date of acquisition; the collection itself is considerably older than the present classification system. However, these potential difficulties are not of sufficient significance to outweigh the utility of this kind of analysis.

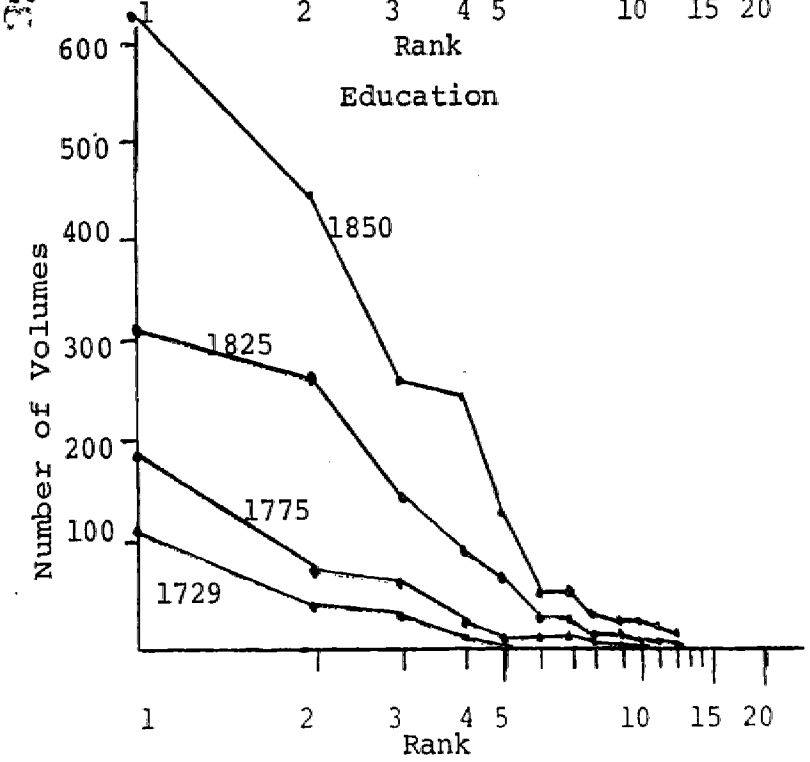
Our first analysis of these chronological listings (9) showed that, at least for the Bibliography and American History classes, the distribution of use of the various broad classes within those major classes could be accurately approximated by the Whitworth distribution, which Good (10) had previously shown to be a useful distribution for approximating the use of the letters (and phonemes) of English. Figure 3 shows the distribution of use of the subclasses for several other broad classes of the Widener List. De Solla Price (11) has since called our attention to the work of Avramesque along similar lines. More recently Krevitt and Griffith (12) (who were kind enough to send us an advance copy of their paper) have studied the use of Whitworth distributions in competition with other candidates of similar form.

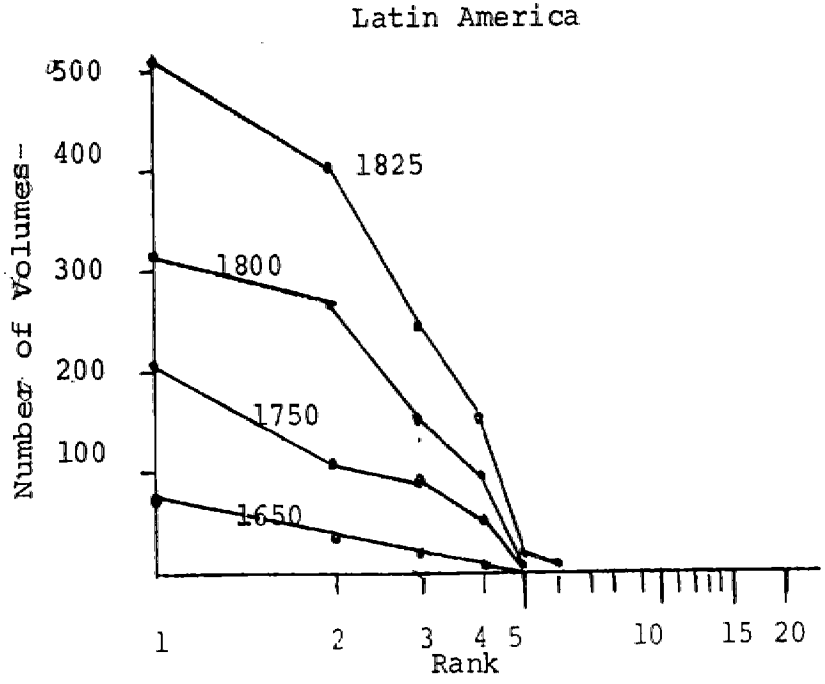
Whitworth (13) originally derived his distribution as the solution to the following problem: suppose separators are introduced on a shelf of fixed length at random; find the expected (i.e. average over a very large number of trials) distance between the two separators that are closest together, the two separators that are next closer together, etc. Now the notion that cataloguers lay out the books on the shelf and then establish class boundaries purely at random is not particularly appealing. Even after observing that not all cataloguers will class a book in the same

Figure 3
Widener Shelf List - Use of Subject Classes









way, it is quite clear that the procedure they use is more nearly a deterministic one than a random one.

What we think happens is this: In any dynamic storage system it does not pay to try and optimize the arrangement at any particular instant in time because the optimum arrangement for a store of size N is not a subset of the optimum arrangement for a store of size N + n. Suppose for instance, that one had a shelf of 1,000 books on a particular subject and, further, that the "optimum" allocation of storage in a fixed environment indicated that each subclass should be of the same size. (In information theory equal use of the classes is optimum under the naive model.) We could then partition the 1,000 books into, say, 25 classes of 40 books each. Now suppose that upon returning to the collection a month later we find 10 new books, all evidently on an entirely new subject when compared with the classes we had already established. In order to return to an "optimal" arrangement it would be necessary to completely reclassify the collection. On the other hand, had we originally partitioned the collection randomly, the introduction of one more random separator would have the effect of splitting one of the existing classes (not necessarily the largest one) at random without changing the mathematical and statistical properties of the system.

In library classification, it is quite clear that the cost of reclassification is prohibitive. As a result the librarian must construct a flexible classification system that can cope with continued growth of the store in a consistent manner. There is no "deterministic" mathematical model that is consistent under constant growth. Therefore we use a "random" model and obtain consistency "statistically." But the distinction between "deterministic" and "random" is nothing more (nor less) than a convenient way of classifying mathematical techniques.

It remains to show that there is some utility in modeling the classification system in its dynamic mode. Examination of the graphs in Figure 3 shows first that although there is variation from the straight line that would represent exact correspondence to the Whitworth model, the straight line approximation is reasonable. Krevitt and Griffith (12) point out that one can approximate this same data, and other data sets, nearly as well with a logarithmic scale rather than a Whitworth scale. The difference between the two is, in fact, quite small and well within the limits of variation for this data. The agreement is, of course, not coincidental. The Whitworth distribution satisfies a difference equation that is the analogue of the differential equation for the logarithm. In other words, the logarithm is the continuous analog of the Whitworth distribution. The difference between the two is greatest at the left side of the graph (small rank, high usage) and is sensible only to about rank five. We prefer the Whitworth form mainly because of its direct connection with the problem of discretely segmenting the library shelf.

In each case we note that the straight line which best approximates the data intersects the horizontal axis (corresponding to zero usage) to the left of many of the data points. These data points correspond, according to the model, to classes that are "not used" (and, in practice, they do have very low usage values). Averaged over all the sets we have examined, we find that about one-third of the classes are "not used" according to the Whitworth model.

Figure 4 shows the subclass distribution for a particular class at various times. The succession of lines corresponds, from left to right, with a succession of times varying from less to more recent. More subclasses are in use for recent times. Some of these represent classes that had existed at an earlier time with near-zero use while others represent classes that came into use during the time interval between two consecutive lines. This shows that the "zero-use" subclasses are those established by the librarian after a few items have actually been received, but before a sufficient number have arrived to enable one, relative to the Whitworth model, to categorize them as being of positive use. Such a procedure enables the librarian to class the books almost immediately rather than having to wait to determine whether the book in hand is an isolated case or the representative of a new potential subclass.

One value of this model is that it enables one to establish an objective measure of the number of classes that are dedicated to "preparing for the future." In this case, we find that approximately one-third of the classes fall in this category. We may then ask whether such a high percentage is cost-effective. In particular, would it be cost-effective in a library having a machine readable catalog with a regular publication schedule for each class in the catalogue?

In such a context it might be more reasonable to class all new material that does not fit the existing structure into a miscellany class until a section of the catalogue is reprinted, at which time the composition of the miscellany class would be studied to determine what new subclasses (if any) should be constructed. Such a decision would ultimately depend on a careful investigation of the cost of maintaining the classification system (including the cost of "knowing" it) as well as a detailing of the reclassification cost, the reshelving cost, etc.

Of deeper significance is the observation that the number of classes grows linearly with time even though the collection is growing exponentially. Here we refer to the number of classes of positive use relative to the Whitworth model; however, as the proportion of classes of near zero use is relatively constant, the observation is also true for the total number of classes. In other words, the number of classification categories is proportional to the logarithm of the size of the main class, providing yet one more illustration of utility of measuring size

as logsize. Indeed, this is just what one should anticipate, for if the size of a collection at time t is

$$s(t) = s(o)\ e^{at}$$

with some constant a (a > o for a growing collection), then the size of an access system for the collection should vary, from what we have said above about the level structured access model, as log s (t), thus, access system size should be proportional to

$$\log s(t) = at + \log s(o).$$

Therefore the total access system should grow linearly with time, as should each of its component subsystems. The classification subsystem should exhibit this growth function, and so it does, as our study of the Widener classification system shows.

THE SUBJECT HEADING SYSTEM

A classification system consists of a controlled vocabulary--a set of words and phrases that is established, and expanded, to serve at any instant in time as a means for partitioning the set of books in a collection into subsets of similar subject matter. It is a hierarchal system. A primary partition is established and maintained for very long periods of time. Each class of the primary partition is subdivided into subclasses; new subclasses are added at a very slow rate (e.g. one every 14 years for American History at the Widener). These subclasses are further partitioned as needed, but at each level both intent and practice insure that terminology is controlled and expansion is deliberately slow.

A degree of semantic flexibility in a classification system is provided by the coding system used to record the classification on book spines and catalogue cards. The codes are minimal and provide no suggestive clues as to content. Thus changes in subject terminology over a period of time can be implemented in the system by introducing the appropriate changes in the definitions of the codes rather than in the codes themselves.

As a means of subject access the classification system is severly limited by the requirement that each book be assigned to a unique class. Books covering several subject classes (e.g. statistics, psychology, Germany) cannot be simultaneously placed in three distinct positions in the collection (unless, of course, three different copies of each such book are purchased.) To cope with this problem, subject headings are used. Any number of subject heading labels can be attached to the record of a particular document and the file can then be exploded so that a copy of the record can be filed at each appropriate point in the alphabetically arranged subject file. The rate of explosion is quite uniformly reported to be about 1.4 subject headings per title.

Changes in terminology in the subject heading file are handled through various reference entries. Thus when the "theory of aggregates" in mathematics came to be known as "set theory," it was only necessary to add a reference entry in the first category to direct users to the second and conversely. This is bothersome for users but has the compensating advantage that it provides a convenient time break in the sequence: books listed under "theory of aggregates" are much more likely to be older books (say before 1950) and books under "set theory" are much more likely to be newer books.

The subject heading and subject classification access systems have a great deal in common: both use a controlled vocabulary; both have an efficient system for coping with changes in vocabulary; both are of the same order of magnitude in size (or logsize) though the subject heading file is about 40% larger; both provide

subject access. The fundamental difference is that the classification system is hierarchal in structure but the subject heading system is not. Thus it made sense to study and discuss the usage of the main classes in the classification system and then study the use of the subclasses within any one of the main classes in a more detailed way. This process can clearly be extended through as many classification levels as seems desirable for a particular collection, but nothing like this can be done with subject headings since no such hierarchy exists for the subject heading system. The latter consists of a single set of alphabetic entries with no natural decomposition into a collection of smaller sets. Indeed, if one could decompose the subject heading system into a hierarchal system one would, in effect, be deriving a classification system; the only difference is that the books could be "classed" in more than one "class" in such a system.

This leads to two observations, one pragmatic, the other relating to the derivation of the appropriate model. Pragmatically, if we wish to study the holdings of a library at a particular moment in time to determine how the holdings are distributed relative to the subject structure, it is clear that we should study the distribution by class number rather than by subject heading. The classification system allows us to deal with the problem at any level in the system. The subject heading system only allows a very coarse measure. In this sense, the statistical properties of the subject heading structure are inherently less interesting. We hasten to add that this comment should in no way be interpreted as a criticism of the existence of the subject heading system or the way it is implemented. It provides useful access to a collection and, given that the fine-structure already exists in the classification system, there would seem to be little point in adding fine structure to the subject heading system. Nevertheless it would seem useful to establish a model appropriate for describing subject heading use distribution in order to provide a simple mechanism for quality control evaluation of the system.

This, in turn raises a mathematical problem: what is the proper mathematical form for the distribution of subject heading usage? At first glance, it might seem appropriate to simply extend the use of the Whitworth distribution to the subject headings. However, although Whitworth provides an adequate description of the classification system where the number of classes is approximately equal to K, it does not work nearly so well for much larger sets. Krevitt and Griffith (12), for instance, compared the utility of the Whitworth distribution with that of the Zipf distribution for four situations where the number of classes in use was of the order of K (English phonemes, Czech phonemes, English letters, and the Widener data on Bibliography) and for these found that the Whitworth distribution was noticeably better in fitting the usage distribution data than the Zipf distribution when the Coefficient of Determination is

used as a measure of goodness of fit. On the other hand, when these two distributions were tested against word counts from the Permuterm Index the situation was reversed. Although we do not know the exact number of terms used in Permuterm, it is clearly considerably in excess of K and more likely in the order of K^3.

This suggests that the mathematical model appropriate for the description of the distribution of usage of a "vocabulary" will depend on the size of the vocabulary. For vocabularies consisting of circa K-30 terms, the Whitworth distribution is appropriate; for large vocabularies such as those of natural languages, the Zipf distribution appears to be appropriate. In short, if we are to extend the model to include larger sets we must generalize the form to take into account the size of the set. Various aspects of this problem are discussed rigorously in Chapter 3. Here we will present the development in a slightly different form and without mathematical details. Before doing so, it is important to emphasize that the problem of determining the mathematical form of usage distributions is, in fact, different from the study of size distributions discussed earlier. For the size problem we were concerned with the determination of the distribution of the number of items of each size (e.g. the number of libraries with 100,000 books) regardless of how much use was made of each library. Now we turn to the question of use of the various elements in such a collection, be it of letters, books, classes, etc. regardless of how large each element is.

Usage distributions have been studied by many prominent researchers in the course of this century. In the present context, Zipf on linguistic questions and Bradford on bibliographic questions are undoubtedly the best known. Zipf's work is better known outside the library community and his empirical observation that a "hyperbolic" distribution accurately approximates word usage distribution in English is generally referred to as "Zipf's Law." Fairthorne (7) and Good (10) provide detailed summaries of the historical development.

Mandelbrot (14) showed that the hyperbolic distribution is a special case of a more general solution to a problem in information theory. He defined information in a store as entropy (following Shannon) and assumed that cost (or effort) was proportional to the logarithm of the rank of the item desired in the store; on this basis he derived the distribution that maximizes the amount of information per expected effort (i.e. the ratio of the two measures). The resulting function is quite simple: if we plot the logarithm of the usage against the logarithm of the rank (where the most frequently used item has rank one, the next most frequently used item rank two, etc.) we obtain a straight line. The Zipf result is the special case where the slope of the line is equal to minus one.

We agree with Mandelbrot's basic notion that one should maximize the amount of information per dollar spent. His measure of

information is open to question (as is any measure) but is backed not only by a good deal of empirical evidence and theoretical work but also by wide application in other problem areas involving the transmission of information, particularly by electrically coded means. Thus, if we are to extend Mandelbrot's result we must concentrate on the shape of the cost function.

In his summary of Mandelbrot's work Good (10) suggests that the logarithmic cost function increases too slowly through the store, if the store is very large. He notes, for example, that the cost of finding the millionth most popular item is only twice as large as the cost for finding the thousandth most popular item. Good then introduces a modification to correct for this deficiency that leads to a complicated and inconvenient generalization of the Zipf-Mandelbrot distribution with which we will have nothing to do here.

We can gain some further insight into this problem by enumerating the distribution that empirically best fits the usage data as a function of the size of the vocabulary being used, thus the form of the cost function that would yield that best fitting distribution using Mandelbrot's derivation.

TABLE 3

Optimal Cost Function for Various Sized Stores

Size of Store	Example	Best Fitting Distribution	Order of Growth of Corresponding Cost Function
K	Widener Classes	Whitworth	log (log x)
K^2	Back of Book Index	"Inverse" Lognormal	$\sqrt{\log x}$
K^3	Permuterm Word Count	Zipf	log x
K^4	Widener Circulation Natural Language Text	Lognormal	$(\log x)^2$

It is tempting, and for that matter relatively easy, to convert the right hand column into a direct mathematical function of the size of the store in the left hand column. However, our data at this time is too limited to warrant this. Our primary interest

in this enumeration is to show that several of the usage distributions that have been found to apply in this area can all be related to one another and to the information theoretic structure of Shannon and Mandelbrot. The pattern in the right hand column is consistent with Good's observation that the logarithmic cost function increases too slowly for large (on our scale, greater than K^3) stores.

The appearance of the lognormal in this organization is worth commenting on. Others have used the lognormal to approximate usage distributions (15,16) with success. Nevertheless, as one surveys the many attempts to "fit" usage data, he cannot help but be struck by the great variety of functions that have been applied with varying degrees of success. This is due, in part, to a lack of firm ground rules in the field of curve-fitting and particularly to the inability to simultaneously achieve linearization and stabilization of variance. As Fairthorne has aptly written:

> "Some years ago I remarked, as have others, that a straight line law connecting any empirical data always can be achieved with the aid of suitably scaled logarithmic paper and a robust conscience. Even more can be achieved if you give yourself the option of declaring the limits of the straight line portion only after you have plotted the data."

In short, if the data exhibits some degree of variability (as do all the examples we have studied in this area), several competing functional forms may fit the data almost equally well. As we noted earlier, Krevitt and Griffith (12) had no difficulty in establishing decisive choices between Whitworth and Zipf distributions for samples with very small vocabularies (where Whitworth was clearly superior) and for samples with much larger vocabularies (where Zipf was superior). The objective of Table 3 is to put this in perspective. One of the implications of its organization is that if Krevitt and Griffith were to apply the lognormal (in the proper form) to all of their data sets they would find that it came in second best in each case and would be the superior choice only if a new set of data were adjoined illustrating the use of a store of size K^2. In other words, Table 3 is arranged in such a way that one could expect to compare any usage distribution with every function in the Table, find a functional form which fits the data better than any other and has the further property that the goodness of fit would become steadily worse as one moved away from that best fitting form in either direction.

Table 3 does not contain every function that has been found useful in this area. It contains only those functions derivable from the Mandelbrot argument using powers of the logarithm of the rank. However, this collection is of sufficient generality to determine a useful collection of approximating functions. The

fact that it is based on both the information theoretic concepts of Shannon and Mandelbrot and the logsize concepts that appear to underlie most of the work in this field lead us to believe that it is quite serviceable, but it is also closely related to a family of curve fitting distributions which have been extensively studied (cf. Dolby (27), Tukey (8)), and which pass into a family of functions which includes those listed in Table 3 by replacing both dependent and independent variables by their logarithms.

We began this brief excursion into the structure of usage distributions by observing that the subject heading vocabulary was two orders of magnitude larger than the classification vocabulary because of its lack of fine structure. Our conclusion of how to treat subject headings is now clear: subject headings should have Zipf-Mandelbrot distributions, though not necessarily with a slope of minus one. We are not aware of data concerning the usage distribution of the circa 90,000 Library of Congress subject headings and therefore have not been able to test this model on that vocabulary, but the Permuterm results are in accord with it, as are other studies on the use of "index terms." (cp. Houston and Wall (16)).

Since the subject heading vocabulary is two powers of K larger than the classification vocabulary, it is natural to ask whether there exists any subject-oriented sets in between. To answer this we turn to the back of the book index.

BACK OF THE BOOK INDEXING

We noted in the previous section that there was a two level jump from the vocabulary of the refined structure of the shelf list classification to the vocabulary of the subject heading structure. Currently, there is no library-maintained subject access system in between. The obvious candidate is the back of the book index. At first glance, it would seem that the back of the book index is essentially different from the subject heading in ways other than size. It is (usually) prepared by a single person, frequently the book's author. It generally is prepared in "free" form where the subject headings are chosen from an authority list. On the surface, at least, an index is closer to being an "extractive" device than the subject headings are in the sense that many, though not all, of the index entries represent linguistic strings that either occur exactly in the text or occur with minor grammatical variations.

Indeed, at the end of the first decade of experimentation in linguistic computation (circa 1963), the prevailing opinion was that it was possible to construct back-of-the-book indexes algorithmically but that it would not be possible to "classify" books algorithmically. Part of this conclusion was based on pure economics: the index was larger (by a factor of K, incidentally) and hence allowed more opportunity for the employment of machine implemented statistical techniques. But it was also thought that classification is an intellectual activity and hence not readily adaptable to machine procedures whereas the intellectual content of indexing, though important, represents a smaller part of the final result.

As a result of our extensive studies, reported in Chapters IV and V, we have concluded that the only fundamental difference between a book index and a subject heading is the substantial difference in size. The average length of a subject heading in the seventh edition of Subject Headings Used in the Dictionary Catalogs of the Library of Congress is 19.17 characters (17). As there are approximately 1.4 subject headings per title, this means that a total of 26.8 characters per title are used, on the average, for subject headings. In other words, subject headings are a first level access mechanism. The average length of a book index was computed for a random sample of some 700 books from the Fondren library and found to be 24,637 characters which is nearly equal to the level 3 size, $K^3 = 25,803$. The average length of an individual index entry, not counting the page location, is 25.47 characters. Thus, on average, the book index is two powers of K larger than the subject heading information.

Other than size, however, there are striking similarities in the way subject headings and indexes are produced. The index is a deep access device and hence requires a substantial human effort for production. This effort acts on the text itself through

processes of intellectual distillation as well as a set of linguistic transformations. The subject headings, being fewer in number than the index entries, represent a still greater distillation (by two powers of K) when compared to the text itself. However, subject headings are not derived directly from the text. Indeed, it is rather obvious that the library community can not afford the cost of requiring the cataloguer to read the entire text. Instead, the cataloguer must restrict his study of the book to the access systems that the book itself provides: the title and any information provided on the title page; the table of contents (including the section headings when these are explicitly presented, and the index.

It is well known that titles are frequently useless and sometimes misleading. However, in fields where titles are usually related to content, they are related in ways that are highly correlated with the derivation of subject headings. For instance, we checked the subject heading "mathematical analysis" in the Stanford University Library Catalogue and found that of the 30-odd books with LC class other than QA (the primary mathematical class), all but one had the word "mathematics" or a linguistic variant in the title. The "mathematical" content phrase in the maverick title was "operations research."

The librarian must act under some rules of parsimony; hence if the title does not contain sufficient information to specify the subject headings, then the next larger access device, the table of contents, must be brought into play. The book index is normally created by a single person at one point in time. As the subject headings are two levels smaller than the index, a cataloguer referring to the table of contents alone could presumably catalogue K^2 (= 873) books with the same effort; we are, of course, only referring to that portion of the cataloguing effort devoted to subject headings. Indeed, we would insist that a professional cataloguer using only the table of contents (and being freed from the constraints of the authority list) could provide the same degree of access to a personal size book collection that the professional indexer would provide through his back of the book index and could do so with approximately the same degree of effort.

But librarians are not primarily concerned with providing indexes (through subject headings) to personal size collections; their task is to provide subject access to larger collections. To do so economically, they must act in concert and the mechanism for accomplishing this is the authority list. The authority list, then, exists primarily to allow the library community to provide the kind of consistent third order access that a single cataloguer (or indexer) provides at the first order. The explicit recognition of this fact not only provides another example of how size dictates the way things are to be done but also may, by simplifying the problem statement, help to clarify the myriad of problems associated with the construction and maintenance of authority lists.

We have gone to some length to establish a plausible basis for the parallelism between subject cataloguing and back of the book indexing to establish a case for a deep look into an activity normally carried out outside the library walls in this report on library access, and also to enable us to move back and forth between the two activities with, we hope, some profit to both. With this in mind we now must look into the following questions:

1. What is the distribution function for back-of-the-book indexes and what implications does it have?

2. To what extent is it possible to automate the creation of book indexes?

3. What problems exist in the amalgamation of several book indexes into a single index covering a specific subject area?

4. What is the distribution function for such amalgamated indexes?

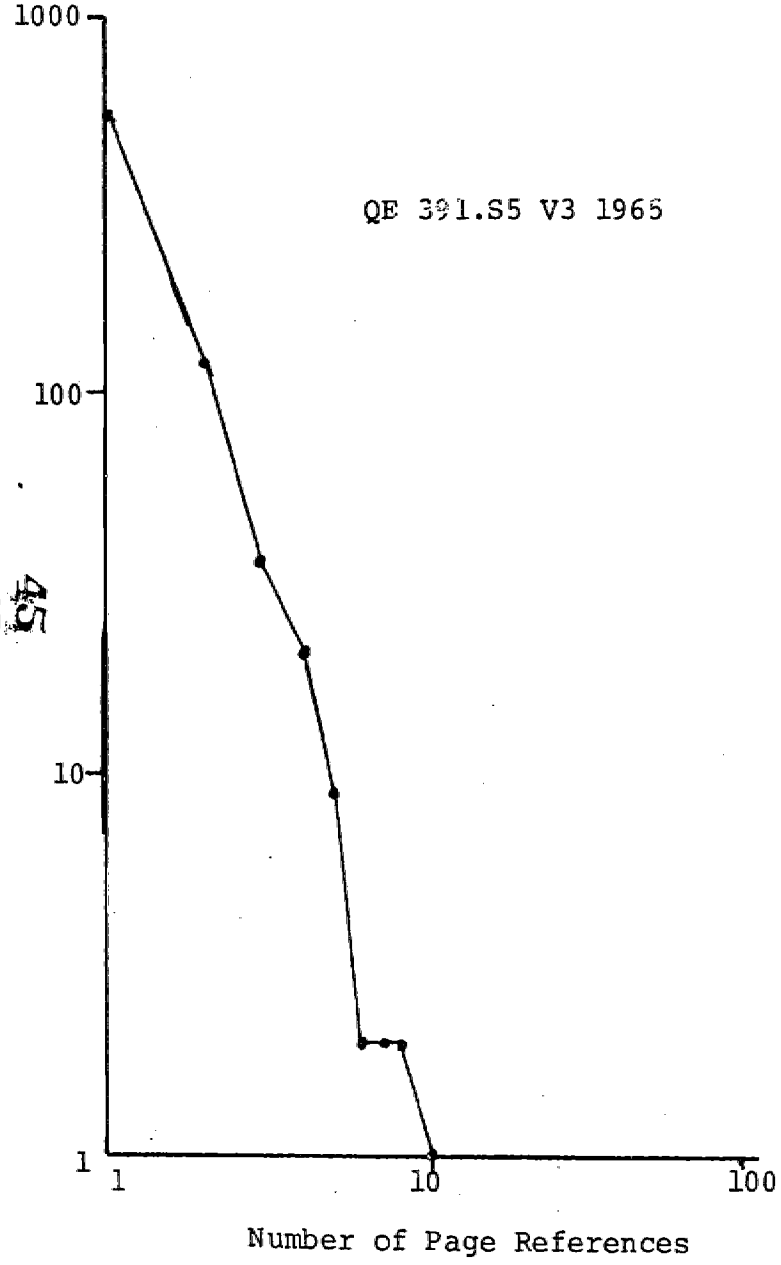
5. What relationship, if any, does the back-of-the-book index have to subject headings?

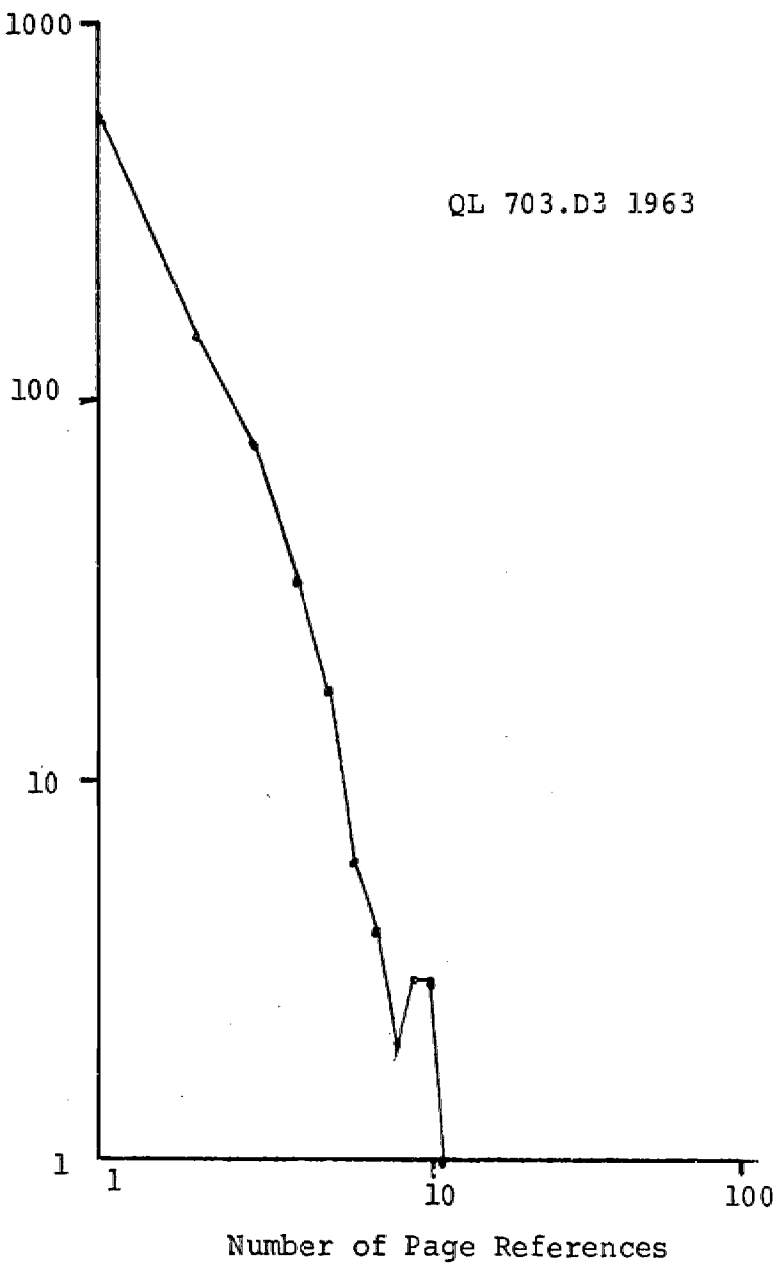
The first question can be answered by studying the behavior of book index distributions in a random sample, in this case a random subsample from the Fondren Sample of books with indexes described in Chapter 4. Appendix II contains the graphs for the sample, two of which are reproduced in Figure 4 for quick reference at this point in the text. The data is plotted on log-log paper to simplify comparison with other information distributions considered in this report.

It will be noted that in Figure 4 (and in the corresponding plots in Appendix II) we have plotted the "number of page references" versus number of index entries" rather than "rank" versus "number of index entries." In the Whitworth plots shown earlier there was little to be gained by this strategem because each rank corresponded to a unique number of items, as is generally the case with first level collections. For book indexes, and also for larger collections, the method now used is superior because it compresses the plot without information loss and simplifies the statistical description.

Figure 4 also provides some insight into the essence of the curve fitting problem discussed earlier: taken individually, each of the "curves" can be sensibly approximated by a straight line, i.e. by a Zipf-Mandelbrot distribution. On closer inspection, however, it is seen that to the extent that curvature is present, it is always of the same type, that is, concave down. It is this consistency of curvature that provides the empirical support for our contention in Table 3 that back-of-the-book

Figure 4
Page Reference Distributions





indexes have distributions that correspond to a "square root of the logarithm" cost function in the Mandelbrot model, and therefore, when properly interpreted, to a lognormal distribution as we show in Chapter 3. This observation shows how essential it is to have a general model that takes size into account throughout the entire range of variation.

These graphs have important implications. We noted earlier that the Mandelbrot derivation of information usage distributions is based on the idea that one should choose that distribution which maximizes the amount of information per unit cost. Having maximized this ratio, its value expressed in terms of the parameters of the distribution provides a measure of the information per unit cost provided by the particular system. Unfortunately, the exact result is mathematically complex for each of the various distributions considered here but for the Zipf-Mandelbrot distribution where the cost function is simply the logarithm (rather than a power of the logarithm) it is simplest. In this case it is possible to show (as we do in Chapter 3) that to a first approximation the ratio is equal to the absolute value of the slope of the line on log-log paper. As all known instances of information usage distributions lead to negative slopes of one or less, the Zipf distribution itself corresponds to the lowest possible information per unit cost ratio, namely unity in our scale of measurement.

Extending this argument to the book index data requires a further approximation: first we approximate the curve by a straight line, and then repeat the argument above for the straight line solution. The relatively slight curvature in Figure 4 justifies the replacement of these distributions by linear approximations. It follows that in the Mandelbrot sense, the "best" book index is the one with the steepest slope. The extreme case occurs when each and every entry in the index refers to one and only one page reference. Dictionaries have this structure, as do certain other types of highly compactified sources of reference information. The "most cost-effective" indexes according to this definition lead the user to a single location to find what he is looking for. This result does not imply that all indexes should have this property; on the contrary the index should point to as many locations as exist in the book if it is to serve its basic function. Nor does it imply that all books should be so arranged as to collect all information about each indexed item in a specific location so as to induce the desired property in the index. The necessity for distributing the reference points in a book is just as pressing (but no more so) than the corresponding necessity for cataloguing all books of the same type into a single sequence in the shelf list. In each case, the author (cataloguer) must grapple with the problem of trying to organize information in a linear string as best he can; and then supplement the linear ordering with multiple reference points (subject headings) to the extent necessary for the particular information collection. The "best" way to do this will depend

on the purpose of the book or other information archive. Thus, a pedagodgical work must necessarily be less "steep" (as reflected by the slope associated with its index) than a reference work, and index slope is consequently one measure of the utility of a particular book for pedagodgical purposes.

AUTOMATIC INDEXING

Experiments in automatic indexing date back at least a decade (13) and, although there is still no clear agreement as to how to properly measure the adequacy or effectiveness of an algorithmically derived index, it is clear that a computer can, at the least, derive and arrange a set of index terms that can be edited by a professional indexer (or the author) at considerably less cost (in people time) than is presently involved in manual indexing. In the jargon of the trade, such an index would be called a "computer-assisted index."

The lengthy delay in obtaining widespread use of such a procedure stems primarily from the very slow rate of increase in the use of computer typesetting for book production. If the cost of putting the text into machine readable form must be born solely by the index operation, there is not only no saving (in people time), but an increase in people time and total cost. The last two years have seen significant decreases in production cost for computer typesetting that tend to make computer typesetting much more attractive. Specifically, per page composition costs have gone down sharply mainly because of the use of improved software and a growing sophistication in the use of the equipment itself. At the input end, the use of optical character recognition devices (OCR) has increased rather spectacularly during the same period, again with significant savings of production costs. Just how fast these developments will signal a substantial shift towards the use of computer typesetting for books remains to be seen. However, it seems safe to say that significant progress is now being made in this direction after several years of relative stagnation.

Let us assume that our objective is to derive a suitable set of potential index entries from a machine readable text which will be presented to the author in page proof form along with the page proof of the text itself. But let us also assume that this is to be done under realistic economic conditions; i.e. we must minimize computer time in the process.

Formally, we can look at the problem from the computing point of view as follows: the text will enter the computer as one long linear string; i.e. one character after the next from the beginning to the end. The problem is, in essence, a recognition problem: we must recognize those sequences which should appear in the index, either as they appear in the text, or in some grammatically modified form. Although there are many variations on the theme of how to do this, most reasonable approaches involve one or the other of the following:

1. A direct attack on the problem wherein every possible sequence is tested either through algorithm, or against a large authority list of "admissible" sequences, or some combination of these two procedures.

2. A two pass system wherein the original linear string is decomposed into a set of (disjoint) substrings which are then tested individually for appropriateness.

The first alternative is inherently uneconomic because it leads to too many string comparisons and hence to too great a computing cost.

The second alternative can also lead to considerable cost unless great care is exercised in the choice of the segmentation rule. Certain "natural" segmentation markers exist in any machine readable text string. Most important amongst these is the end of paragraph symbol, generally a special symbol introduced into the machine readable text at input to explicitly delineate paragraphs for photocomposition. Further refinement of the segmentation following traditional grammatical lines, i.e. through sentence and phrase to word, is possible. However, the available explicitly demarcated symbolic structure is essentially limited to interword spaces and marks of punctuation and there is no thoroughly agreed on efficient algorithmic "parse" of the sentence that we can lean on.

Moreover, a sentence parse provides more information than is required; it generates, for each sentence of text, a transformational or tree structure corresponding to the corresponding grammar that is assumed to underlie the language. Segmentation corresponds to some appropriate horizontal section of the tree structure; the remainder is irrelevant for our purposes.

We have studied simpler procedures with the limited goal of producing sentence segments adequate for indexing and inexpensive enough to compete with human indexing practice. Following a suggestion of Tukey (20), we look to the structure of permuted title indexes. There we find that each word in the title is tested against a (relatively) short "stop list" to determine if the word should appear at the gutter, or center, of the permuted title list, or whether it should be "stopped" from so appearing. Let us call all words (or sequences of characters between spaces) that are not "stopped," "go" words. This procedure segments the text into sequences of consecutive "stop" words, followed by sequences of consecutive "go" words. The utility of the procedure is its computational simplicity: the list of "stop" words is relatively short, consisting, as it does, of the frequently occurring structure words of the language together with other words that carry meaning but are not generally included as index entries. Hence it is only necessary to check each consecutive word to determine if it is contained in the "stop" list in order to determine whether the current sequence of consecutive "stop" words should be continued (if that is the case) or whether the current list of "go" words should be terminated.

This simple procedure can then be modified step by step until a minimal segmentation rule is obtained. Obviously, one will wish to make use of punctuation other than end-of-paragraph markers. There is a certain utility to extension of the "stop" list by "algorithmic stopping," e.g. by systematic removal of all short words (say, less than four letters) and systematic removal of all words with certain terminal strings such as "ly." One can also provide a more sophisticated structure by providing overrides so that, for example, "of" is only considered a stop word when it occurs at the end of a sequence of "go" words, the latter defined by all those rules not involving "of."

The form we have chosen for implementation is discussed in detail in Chapter V. However, except for the fact that we are able to demonstrate that usable segmentation can be obtained from high speed production programs, the actual form is not particularly important. What is important, with regard to this problem as well as to all other problems of linguistic computation, is the way the algorithmic portion of the solution is to be structured. To obtain economic production, there are two essential ground rules. First, it is the nature of natural language text that a few simple rules can be constructed for the solution of almost any problem which will successfully treat a large proportion (say, 80%) of the text to be processed. Each rule that is added to the system after this "easy" fraction has been dealt with will tend to be more difficult to derive and implement and will also successfully process a smaller proportion of the remaining available unprocessed material. In short, the problem solver is faced with a classic problem of marginal utility: at what point does he cease to implement new rules on the grounds that the increase in processing cost is greater than the alternative cost of not implementing the rule. (The latter may be expressed either in terms of increased cost to the user if the "errors" are left in the processed material, or in terms of the cost of removing the errors through a manual editing operation.) The cutoff point will, of course, vary from one problem to the next, but in all problems it will be absolutely necessary to make a careful determination of it if the final product is to be economically viable.

In a sense it almost seems redundant to observe that there is a rule of marginal utility operating in computer programming. However, it needs saying, for there is a particular hazard in this relatively new area that might not be completely obvious. Today, a competent programmer can readily obtain the linguistic information needed so that he can program the steps necessary to accomplish subtle linguistic tasks well beyond the point of marginal utility. He may just find this part of the job the most exciting and be loathe to let the program fall short of including all the refinements that he can think of. Or he may think that his professional reputation is at stake and that he cannot afford to allow a job out the door until it displays his full range of knowledge of transformational grammar or some other important

but always costly and often unnecessary advanced linguistic model.

Any professional who studies the simple segmentation algorithm we provide in Chapter V will immediately see improvements that he could make. So do we. But that is not the point, for improvements must be made only when it is clear that they are cost effective. It will not be easy for the reader to find such additions to the algorithm offered in Chapter V.

Perhaps a specific example will help to bring the problem into perspective. In running text, it is natural to use both singular and plural forms. In indexing, all forms are usually converted into the singular. A naive approach to the problem would be to institute a detailed study of the way plurals are formed in English (and other languages, if it is a multi-lingual data base) and then to construct and program a "complete" plural-to-singular algorithm that would be used on every text word. A less naive approach to the problem is to ignore it until the rest of the system is operative and one has a substantial amount of test material accumulated and put in final form. Examination of such test material will quickly show that the only proglems the plural forms introduce is that they inhibit the agglomeration of entries that are identical except for number. With rare exceptions this will happen only when the last word of the segment (or entry) appears in both plural and singular form. It is wasteful to test every word in the text for number if only the last word of the segment is important. Thus the singular-plural logic should be brought into play at the entry agglomeration stage, not at the text stage. Further, when one studies the problem in detail it becomes clear that it is only necessary to use a final-s rule to resolve 80% of the problems. Further simple improvements such as translating "ies" to "y" and deleting "es" conditionally clean up the problem adequately for almost all actual examples. It would be a rare case where it would be worthwhile to adjust the program to convert irregular plurals such as "men" to "man."

Before leaving this aspect of the problem it might be well to note that these same considerations apply with equal force to another well-known library automation problem: the problem of designing and implementing compùterized filing rules. The economic solution to the filing rule problem requires a judicious mixture of a knowledge of library practice in this area together with detailed statistical information concerning how often each of the potentially useful rules actually occurs in a data base of a given size. Testing every name in the file to determine if it is prefixed by "sir" may be an expensive way to handle three cases in a million.

There is a second basic principle of programming strategy that is of nearly equal importance to the one that requires the observation of economic dictums concerning marginal utility.

This rule is also obvious, but its range of applicability is much wider than is normally expected. The rule, quite simply, is this: whenever the set of objects under consideration can be decomposed into two sets one of which must be stored in the computer for matching purposes, always use the smaller set. For example as we have already observed, in permuted title work the title words are decomposed into "stop" words and "go" words. The number of "stop" words is very small (generally about 100 to 200) whereas the number of "go" words consists of all other character sequences used in the data base. Obviously, the proper strategy is to store the "stop" words and then define any sequence not found in the "stop" list as "go." This procedure is advantageous and meaningful because not all words are exhausted by these two classes in practical application.

This procedure has several advantages. It reduces the programming effort as fewer words need to be inserted (and corrected) during programming and it reduces the size of the required store so that the program can be implemented on smaller (and cheaper) machines. However, the greatest gain comes through the reduction of the number of character matches necessary during operation of the program. In any non-trivial linguistic program (i.e. a program that is not bound by the speed of the input and output devices), the operating cost is primarily determined by the number of character matches. Hence any procedure that reduces character matches is of the greatest importance.

The principle, as we have said, is obvious. Indeed, we know of no permuted title program that attempts to match against the longer "go" list. The problem, then, is to find ingenious ways of using the principle repeatedly to obtain increased gains in program speed. For instance, there are relatively few short words in the English language; most of them appear on most "stop" lists. As the machine must "know" the length of the word as a result of the procedures necessary to determine its boundary points, it is a simple matter to "stop" all words less than a certain length. In the mathematical texts we have studied, the "go" list for words of fewer than 4 letters is one-tenth the size of the stop list. Similarly, in our automatic indexing routine we found that the number of one-word index entries was very small compared to the number of one word segments produced by the rest of the algorithm. Thus we stop one-word entries (with exceptions) with considerable gain in program speed.

With these operational questions in hand, let us now assume that we have a segmentation algorithm in operation with reasonable economic properties (whether it be a variant of the algorithm we have derived, or some entirely different approach to the problem). Let us further assume that the segments have been (machine) sorted into alphabetical order, page locations have been accumulated under each distinct segment, and duplicate page locations have been eliminated. The question then remains: is the resulting organized list sufficiently close to what is needed

to permit the author (or professional indexer) to edit the list to produce a final index at a cost that is not only lower than the cost of manual production but also sufficiently lower to offset the computer costs of producing the list in the first place.

Part of the answer to this question is subjective: the author, or publisher, must determine whether the product is the "proper" kind of index to go with this book. The procedure will be viable only if most of the time the answer is "yes" so that the initial programming cost can be written off against a number of jobs.

Part of the answer lies in the subsequent degree of refinement possible in the automatic generation of inverted entries, insertion of see and see also entries, and use of authority lists for comparison in machine readable form. The mechanical questions involved will be discussed in the following section devoted to the agglomeration of indexes from various books concerned with the same general subject matter.

Nevertheless, as we have now come to expect, a good deal of useful information about the suitability of the automatic indexing procedure can be obtained in terms of the size of the index it produces, and other size related distributions. We know that a first order index should be approximately 1/30th the size of the book it indexes. Although it is somewhat less time consuming to delete from a provisional index entries that are not wanted than to insert entries that were missed, there is an upper limit to the amount of deletion activity that an author will tolerate and this constraint must be met. But note that some of the entries in the algorithmically obtained index will appear because of errors in the text itself. Thus, study of the provisional index will, incidentally, reward the author by drawing his attention to certain kinds of text errors that might not otherwise be caught. We have no way of estimating the potential utility of this by-product of automatic index construction.

In addition to the gross size of an index, we can check its page reference distribution to determine if it is compatible with manually constructed indexes and/or with theoretically predicted distributions of the general form derived in the preceding section. Mean entry length and the entry length distribution can be similarly examined. Presumably, if all of these measures coincide with, or at least approximate, the measure derived for manually derived indexes, we can be assured that the index has the proper statistical "shape," thus providing necessary, although not sufficient, measure of performance if the average manually produced index is used as a standard.

We have applied the automatic segmentation algorithm to one book length text (3) and found that it does in fact closely approximate

the statistical shape defined by manually created indexes, thus demonstrating that even the simple procedures outlined in Chapter V are sufficient to provide an index which is statistically similar in structure to usual indexes.

CUMULATIVE BOOK INDEXES

We observed earlier that a book is three powers of K ~ 30 larger than the subject heading that provides access to it. The book index is only one power of K smaller than the book itself. Thus the user is faced with a third order access device (the subject heading) when he wishes to gain access to the collection, and a first order access device (the index) when he wishes to gain specific access to the book. To improve subject access to the collection we must move up to a second order access device. A complete cumulation of the indexes to all books in the collection would be one level larger. Therefore, the next logical step is to obtain a selection of books from the collection, either by random sampling, or by seeking the guidance of a specialist in each field to select the most useful (or perhaps most widely used) books and then cumulate their index entries. Table 4 shows the size relationships measured in characters of typical systems for book access.

TABLE 4

Size of Subject Access Mechanisms

Mechanism	Average Number of Characters per Book
Book	762,483
Complete Index	25,803
Selected Index	873
Subject Heading	29.55

When considering index cumulation the question naturally arises whether it is feasible to consolidate the various styles of indexing that will naturally occur in a large number of books. To test the difficulty of this problem we selected some 75 books in statistics--a field wherein we could exercise some qualitative judgements as to the utility of the cumulative product index, keyed the indexes into machine readable form and constructed a cumulative index to the set. From a study of the materials prior to input, it was determined that three major problems of format variation required further study:

1. Variations in the citation of personal names;
2. Variations in decisions regarding forward and inverted entries;
3. Variations in the use of see and see also references.
4. Variations of singular and plural forms.

Personal name variations is, of course, a familiar and solvable problem. Reduction of all forms to surname-plus-first-initial form by algorithm leads to a high accuracy solution that leaves

for manual correction only those cases where only the last name is given and those where two or more people have the same surname and first initial. In this particular corpus the only cases spotted of the latter situation involved the Bernoulli's. Variations in the surname that occurred with sufficient frequency to make algorithmic adjustment useful were restricted to the simple procedure necessary to identify De Moivre and Demoivre, and to remove hyphens. (Hyphens always present a special problem in the determination of whether it is better to leave them alone, delete, or delete-and-close. Simple deletion proved most effective here.) Finally, "of," and "off" in terminal position of a proper name entry were systematically converted to "ov." Other variations and counter variations were handled by manual correction. It might be well to note that invocation of the algorithmic procedures before any manual editing is done permits the shortcomings of the algorithms to be treated right along with the shortcomings of the keying operation so that the added cost of correction is strictly a function of the number of entries requiring correction.

The problem of treating forward and inverted entries (e.g. "normal distribution" and "distribution, normal") was studied prior to input (21) and the following simple procedure was adopted: at input all inverted entries were converted to forward form by the keypuncher; inverted forms were systematically machine generated using the following rule:

1. All entries including the word "of" were maintained in forward position (except those treated by rule 3) and repeated in inverted order; e.g. "analysis of variance" occurs as an entry as does "variance, analysis of."

2. A frequency list of last words of each entry was constructed and certain of these were used to generate inverted entries; e.g. "normal distribution" and "distribution, normal."

3. A frequency list of initial entry words was also constructed and from this certain words were used to suppress the normal, or forward form of the entry; e.g. "least squares, method of" occurs as an entry but "method of least squares" does not.

In the original data only five percent of the entries occurred in identical form both as forward and inverted entries. Thus the simple rule, although not ideal from the professional indexer's point of view, generates increased access to the information in a systematic fashion.

Reversion of inverted entries by the keypuncher did present some problems as not all entries in the material were of "correct grammatical form" and the reversion occasionally led to unfortunate sequences. However, the proportion of such entries was small and easily taken care of at the final proofreading stage.

The use of references did not post a significant problem. If one indexer included a see also reference, it may be argued that it should then be included for the whole set. If a see reference is provided in one book, it is only necessary to make an obvious test to determine if this should be converted to a see also reference for the list as a whole. At this writing, no attempt has been made to determine if any of the see also references lead to blind points either through omission of the source material or errors in keying. Nor has any attempt been made to insure that all see also references are inverted; e.g. "Gaussian distribution, see also Normal Distribution" and "Normal Distribution, see also Gaussian Distribution."

Restriction of the reference entries to those provided by the individual indexers does not insure that certain, potentially useful, reference entries will be missed entirely. However, the union of the efforts of 76 individual indexers should at least provide a good first approximation to a thorough system.

The singular and plural form problem is surprisingly persistent. Although we had anticipated a substantial application for singular-plural conversion rules in entries derived from running text, we thought that "reduce almost all forms to the singular" would have been the rule in a cumulation of indexes. However, a check of the first few pages of sorted output shows that approximately half of the entries that occur in more than one book occur both as singular and plural forms. Thus the inclusion of the singular-plural conversion rules derived for the indexing algorithm is mandatory for the cumulative index as well.

Finally, a word is in order about entry length. Indexers obviously vary in their practice of entry concatenation with the result that some entries are quite long. As a simple expedient to "force" conformity on the collection, we instructed the keypuncher to truncate all records at 80 characters (the length of a standard punched card); truncated records were deleted from the sample (unless the truncation occurred in the page location field). The number of entries so deleted was small and the number of page locations lost in the entries left in was insignificant.

THE CUMULATIVE INDEX DISTRIBUTION

The utility of a cumulative index to selected books in specific subject fields can only be determined by making a number of these indexes and distributing them to ultimate users. However, a few remarks are in order on the statistical properties of the one sample we have studied.

Table 5 lists the 50 most frequently occurring entries in the 76 books, where frequency of occurrence is measured in terms of number of books rather than total number of page references over all books.* Anyone familiar with the field of statistics will recognize that all of the entries are, in fact, important concepts or persons in the field. The nine distributions listed--binomial, normal, Poisson, F, chi square, multinomial, bivariate normal, hypergeometric, and exponential--do in fact dominate the the field in terms of utility. ("Conditional distribution" is a generic term rather than the name of a specific distribution.) Fisher, Bartlett, Neyman, "Student," Yates, Cochran, Egon Pearson, Wald, and Cramer are all names to be reckoned with by any scholar in the field. (We, of course, explicitly refrain from trying to draw any conclusions about the relative worth of the work of a man whose name appears somewhat further down the list. A slightly different choice of books might present a different ordering.) Similarly, the most widely used statistics--standard deviation, variance, etc.--and the most widely used procedures--analysis of variance, least squares, etc.--also appear in the list. In short, the high frequency index entries do provide a reasonable picture of what "statistics" is all about, as one would hope.

Originally, 31,232 index entries were keyed and read onto tape. Elimination of duplicate entries introduced by error or through the convention of "reinverting" inverted entries, overly long entries, and the reference entries (which are not included in the counts here) reduced the data base to 27,471 entries. The total number of reference entries was 1,195, including duplicate entries from the various book indexes. Of the 27,471 non-reference entries, 20,388 were unique and 7,083 represented entries occurring in more than one index. The frequency distribution and graph thereof are shown in Table 6 and Figure 5, respectively.

Figure 5 is drawn on log-log paper and, except for the first point (number of entries occurring in only one book), the straight line approximation is very good, as we would expect for a sample of this size. In other words, with 20-odd thousand distinct entries, one should expect the Zipf-Mandelbrot approximation to be quite good, and it is.

* These counts do not reflect the fact that some terms appear in different forms in the list (e.g. normal and gaussian distributions).

TABLE 5

Most Frequently Occurring Entries
Cumulative Index to 76 Books on Statistics

Index Term	Number of Books
BINOMIAL DISTRIBUTION	46
NORMAL DISTRIBUTION	43
CONDITIONAL PROBABILITY	42
STANDARD DEVIATION	39
FISHER; R	37
POISSON DISTRIBUTION	37
CHI SQUARE DISTRIBUTION	35
VARIANCE	34
RANDOM VARIABLE	33
F DISTRIBUTION	32
CENTRAL LIMIT THEOREM	31
ANALYSIS OF VARIANCE	30
DEGREES OF FREEDOM	28
MOMENT	28
STATISTIC	28
COVARIANCE	27
CORRELATION COEFFICIENT	25
MULTINOMIAL DISTRIBUTION	25
T DISTRIBUTION	25
MEDIAN	24
BARTLETT; M	23
NEYMAN; J	23
NULL HYPOTHESIS	23
STUDENT	23
BIVARIATE NORMAL DISTRIBUTION	22
YATES; F	22
EVENT	21
MEAN	21
PARAMETER	21
PROBABILITY	21
SIGN TEST	21
HYPERGEOMETRIC DISTRIBUTION	20
LEAST SQUARES	20
CHARACTERISTIC FUNCTION	19
COCHRAN; W	19
CONDITIONAL DISTRIBUTION	19
CONFIDENCE LIMIT	19
PEARSON; E	19
PERMUTATION	19
POPULATION	19
RANGE	19
WALD; A	19
COMBINATION	18
CORRELATION	18
CRAMER; H	18
EXPONENTIAL DISTRIBUTION	18
HISTOGRAM	18
INDEPENDENT EVENT	18
MOMENT GENERATING FUNCTION	18
NORMAL EQUATION	18

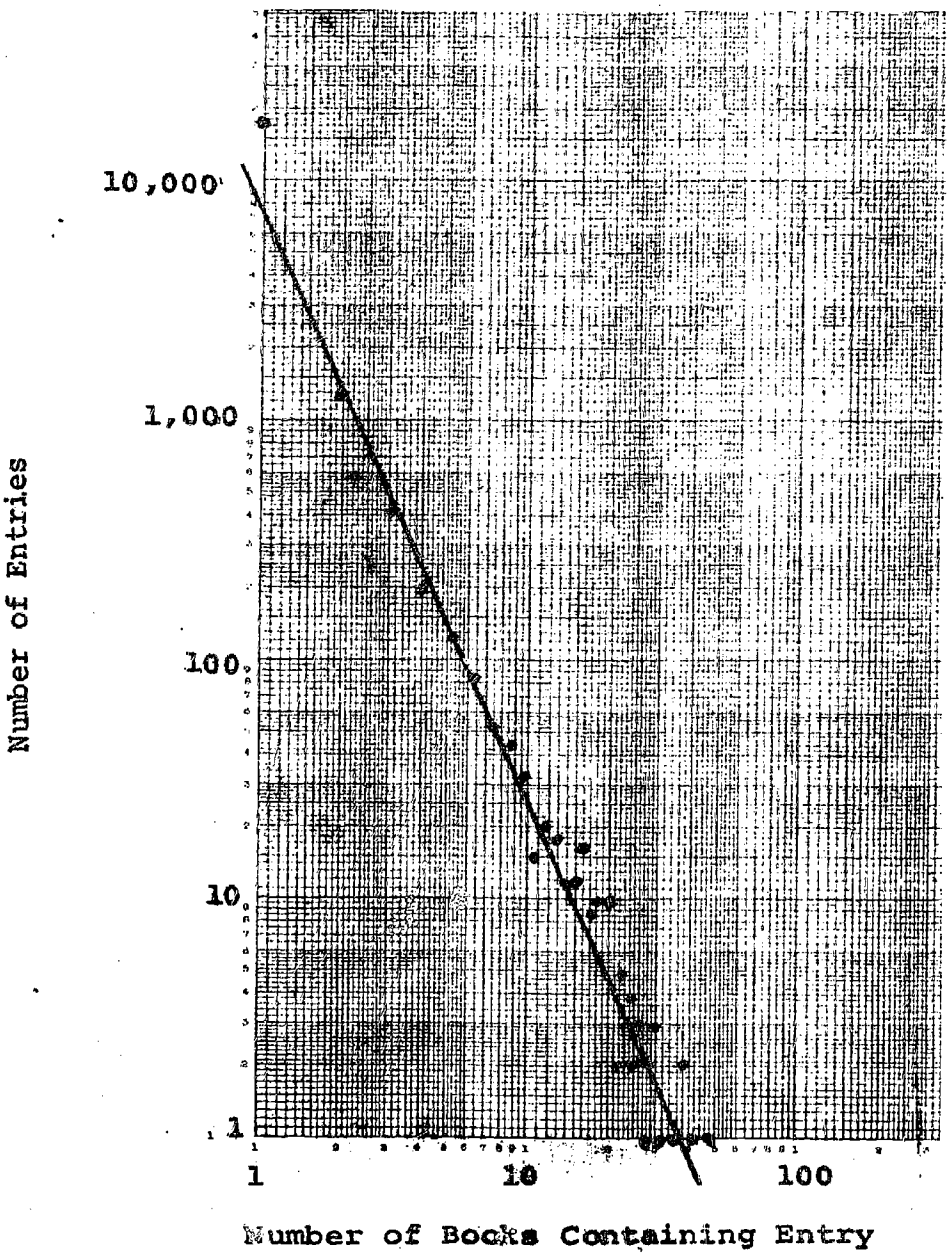
TABLE 6

USAGE DISTRIBUTION OF CUMULATIVE INDEX TERMS
CUMULATIVE INDEX TO STATISTICAL LITERATURE

Number of Times Used	Number of Terms with that Usage
46	1
43	1
42	1
39	1
37	2
35	1
34	1
33	1
32	1
31	1
30	1
28	3
27	1
25	3
24	1
23	4
22	2
21	5
20	2
19	9
18	10
17	10
16	8
15	17
14	12
13	12
12	18
11	21
10	15
9	32
8	44
7	55
6	88
5	126
4	199
3	402
2	1,246
1	18,039

Figure 5

Book Reference Index Distribution
Index Entries ~ 76 Books on
Statistics



The apparently excessive number of singly occurring entries can be explained in part. The counts given here are based on the machine edited sample and thus contain machine corrections but no manual editing. Keying errors almost invariably lead to entries that occur only once in the sample. After correction, some of these entries will continue to be unique. However, with errors removed, some will correspond to other singly or multiply occurring entries, which will slightly decrease the frequency of singly occurring entries and increase the frequency of some of the multiply occurring entries. The result will be a slight increase in the slope and hence a slightly better overall fit of the data by the line.

The slope of the line in Figure 5 is approximately -2.5, almost exactly the average of the Fondren Index sample discussed earlier. Thus, although we have increased the size of the index by almost a factor of K~30, when compared with the size of the average book index we have done so in such a way that the information per unit effort has been maintained at the same level we can expect for individual book indexes. If the books constituting this sample were too redundant, i.e. if all discussed the same few basic ideas, the slope would be reduced and the user would be better off to use any one of the indexes rather than the cumulation.

On the other hand, if the books were totally disjoint, that is, if no entry occurring in one book occurred in any other book, as might happen if we cumulated indexes from a book on American history, a book on statistics, and another on education, the resulting consolidated index would have an infinite slope (every entry occurring just once in the sample). Although this would maximize the information per unit effort ratio, the user would only discover that "analysis of variance" was to be found in a statistics book. "American Revolution" in a history book, and so forth. In other words, such an accumulation would act more as a dictionary than as an access device to the collection of books.

One other comment. The number of entries in the cumulative index is 20,388, which is 5,415 fewer entries than the 25,803 prescribed by the level structured model. It is of interest to learn how many more books would be necessary to bring the collection up to the 25,803 distinct entries representing the mean size of level 3 of the model. This problem is discussed by Good (10); since we already have nearly the desired number of entries we can use the first term of his expansion to estimate the number of additional books required to obtain a full 25,803 distinct entries in the consolidated index. This reduces to the following simple formula: divide the number of new distinct terms needed (5,415) by the number of singly occurring entries in the collection at this point (18,038) and multiply by the number of books in the present collection (76). This yields a result of 23 books which should be added to the collection to reach the desired size.

THE USAGE OF LIBRARY MATERIALS

Up to this point we have been primarily concerned with the structure of the information store and the attendent access system. Now we must turn to the question of how present systems are--and could be--used. There have been many studies of library usage over the years, but two recent studies in The Library Quarterly illustrate certain aspects that we wish to stress.

In the first of these studies (22), Lipetz reports on the usage of the Yale University Library card catalogue. His Table 3 shows that no less than 73% of the catalogue searches had, as an immediate objective, the location of a document. Only 16% were subject searches while 6% were searches for information on authors and the remaining 5% for bibliographic information. Although these results apply only to the circumstances existing at the Yale library--their collection, their catalogue, and their community of users--at a particular interval in time, we think that document searching is the primary activity in the catalogue room of most libraries.

Lipetz's further investigations show that nearly one-third of the users were in fact looking for subject information, even though only half that number actually searched the subject heading list. There are several possible explanations for this apparent anomoly. Perhaps the most obvious is that if the user knows, or has strong reason to believe, that the information he is seeking is in a book that he can find in, say, the author list, it will be wasteful to look for it in the subject heading list. Moreover, as we mentioned earlier, a user does have some interest in the authoritativeness of the source; if he knows of one whose authority he does not question, he will obviously head for that document first.

Even if the user is not sure that the document he has in mind will contain the required information, he can usually use that document as a surrogate for the classification system outline: If it is the right kind of document, location of it will place him in the vicinity of a set of books that are very likely to contain the information he wants. Further, examination of the citations in these books may provide a key to the journal literature that is not provided by the card catalogue.

Further insight into the utility of the subject organization of books on shelves is given by Morse (23). In a study made at the MIT library, Morse notes that mathematicians typically visit the card catalogue once each time they visit the library; that they typically consult two books per visit, and that 40% of the time they end up borrowing a book for further use. The pattern for chemists is different, with less emphasis on the use of the card catalogue (only 30% of the time), greater emphasis on book consultation while in the stacks (typically consulting four books rather than two), and a lower book borrowing rate (20% in place of 40%).

These figures tend to emphasize the utility of "subject" information. The mathematician looks at two books for every catalogue visit; the chemist, nearly 14. Regardless of what this tells us about the differences between mathematicians and chemists (or mathematics and chemistry), it strongly suggests that both are interested in the information that can be gleaned from a "consultation" of a document, or at least willing to take some given that they are already in the library. Both are willing to leave the library without a book in hand more often than not.

Based on these studies and the recent proliferation of subject oriented access tools, we claim that the subject heading serves only a small proportion of the catalogue lookups and even a smaller proportion of the monograph subject searches in a library. This does not imply that maintenance of the subject heading lists should be reduced or eliminated: Lipetz's Yale Library data shows that nearly one percent of the campus library users use the subject portion of the catalogue daily and it performs a significant service function for the users. Rather, it emphasizes what might have been stated from first principles: the subject heading list is not a primary (=order 1) subject access device but a tertiary (=order 3) device designed mainly to mediate the shelf list subject organization which is constrained by the physical limitation that books (of which there are single copies) can only be stored in a linear file.

Earlier in this report we made a case for the creation of cumulative indexes to increase subject access based on the contention that the subject heading list is "too small" according to our measure of size. The Lipetz and Morse data suggest that such an "explosion" of subject access would find use in their libraries. The MIT chemist, according to Morse, consults 22 books for each book he borrows. No doubt some of these consultations are successful, indeed so successful that the consultation itself removes the need to borrow the book for further study. However, one cannot help but suspect that many of these consultations result in a quick look in a book index that is sufficient to show that that book does not contain the required information--at least insofar as its index indicates. A single lookup in a cumulative index, both preferably located on the shelf as well as near the catalogue, would reduce look-up time, and also increase the probability that the information would be found.

Morse also comments on the utility of the Zipf-Mandelbrot-Bradford distribution in analyzing usage distributions. As we have noted earlier, this model is a good approximation for collections of the proper size and a useful one for most collections. For very large collections and very small collections, it is generally not adequate. As a demonstration of this we have studied the circulation of the Widener Library for the period 1965-69, as given to us by Foster Palmer in a private communication. The data is shown in Table 7 with the accompanying graph (drawn on lognormal

probability graph paper) in Figure 6. The excellence of this model is evidenced by the straight line fit in Figure 6 and by the fact that using this approximation we can compute the number of books "used" zero times and hence the size of the whole collection to within approximately 10% of its reported value from data referring only to the 6% of the collection actually used during the interval. See Chapter 3 for a detailed analysis.

Where the Morse and Lipetz data provide information on a "usage-per-user basis," the Widener data allows one to examine the usage of the whole collection. We see, for example, that only 6% of the books were "used," i.e. borrowed, over the five year period. This is consistent with Morse's data; recall that the chemist "consults" 22 books for every book he borrows although the mathematician consults 5 books for every book he borrows. However, this is not really the point. Almost any organization of usage distributions, including the one we suggest here, will imply that the larger the collection, the larger the number of items which will be "unused" in any particular time interval, and, as we show in Chapter 3, is consistent with optimal use of the library's resources when effectiveness is measured by information per unit effort.

The deeper questions involve how large a large collection "should" be and whether some libraries should attempt to collect "everything," and if so, how many such ambitious libraries we should have. As we stated at the beginning, we shall not attempt to answer such questions here. However, we do claim that the level structured model presented above can be used to analyze these larger problems. For instance, the question whether a set of libraries in a particular geographical region should integrate into some sort of network is closely akin to the question of accumulating indexes and can be modeled the same way. It is possible to analyze the holdings of the various libraries to determine if the proposed usage distributions for the network provide an improvement of the distributions for each of the individual libraries. To the extent that these questions will be of importance in the next decade, libraries should be encouraged to accumulate usage information by class and by book so that it will be possible to compare individual and cumulative distributions in considering proposed mergers through networks or other organizational means.

Table 7

Harvard University Library
Circulation Distribution
1965-9

Times Circulated	Number of Books Circulated this Frequently
1	260,878
2	72,911
3	36,022
4	21,179
5	13,560
6	9,409
7	6,666
8	5,136
9	3,752
10	2,886
11	2,255
12	1,700
13	1,322
14	1,086
15	765
16	631
17	479
18	382
19	303
20	189
21	162
22	95
23	81
24	66
25	32
26	34
27	17
28	11
29	11
30	5
31	4
32	3
33	2
34	1
35	4
36	3
40	1
47	1

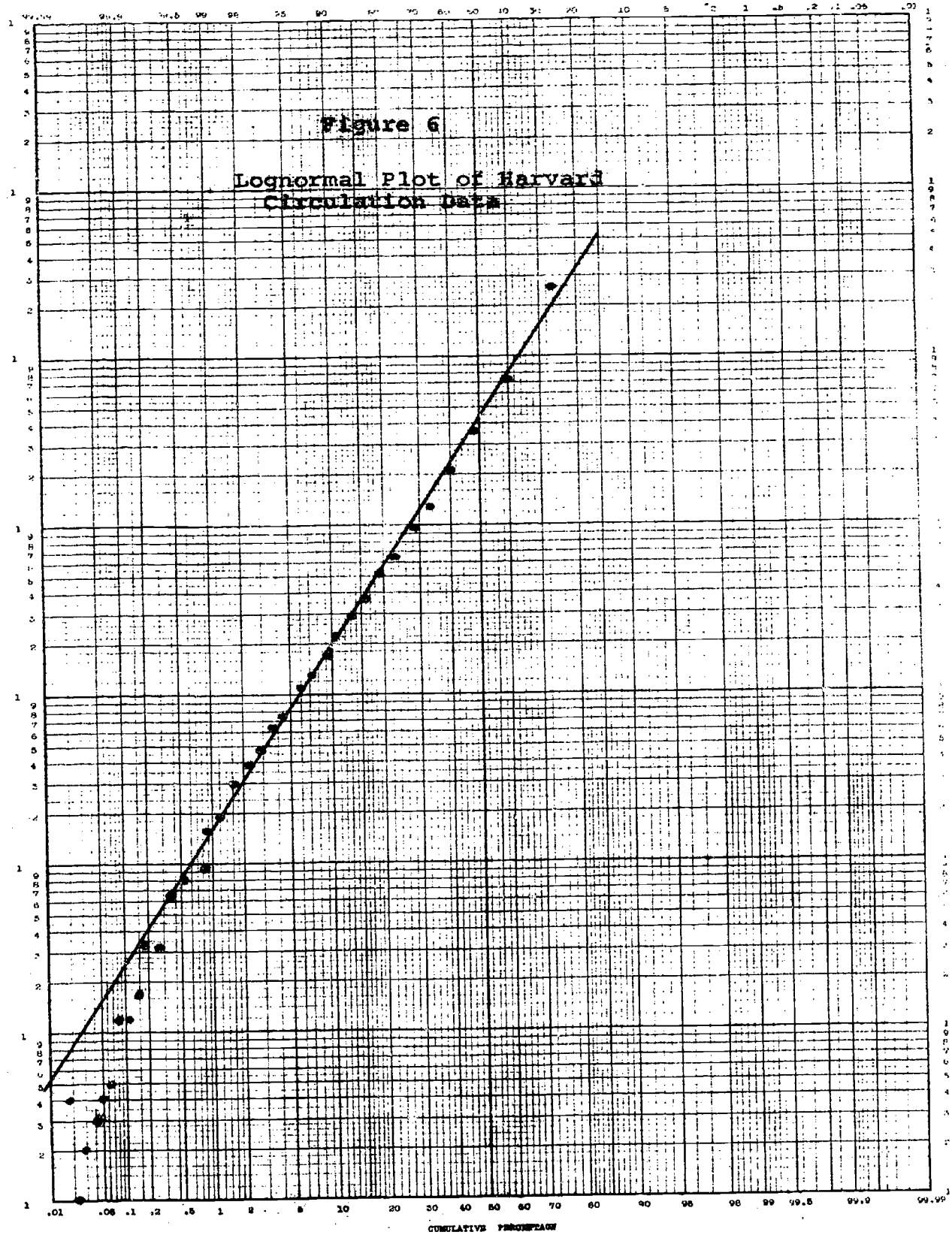


Figure 6

Lognormal Plot of Harvard Circulation Data

THE IMPACT OF DEVELOPMENTS IN THE COMPUTER WORLD

Elsewhere (3) we have noted the long term downwards trend in the cost of computing hardware. From the desktop calculator to the largest computer, hardware cost per operation continues to fall at a rather spectacular rate. Improved peripheral devices, particularly in computer output microfilm (COM), are now commonly available at attractive prices. Mini-computers now exist with speed and capacity that rival the much more costly low end of the big computer line of five years ago. Although there have been minimal advances in computer typesetting hardware other than COM in the last five years, increased usage enhanced by an increasing sophistication in the software area connected with computer typesetting have led to significantly lower prices which in many cases are unrestrictedly competitive with traditional typesetting and page composition methods. Optical character recognition (OCR) has apparently turned the corner at last and there are now a sufficient number of these devices in operation that OCR must be considered a viable competitor to the keypunch for the input of linguistic materials.

Not every library is in a position to take advantage of the most recent gains. Many university and public libraries are secondary computer users, totally dependent on the main computer shop of the parent organization. As such they may not be able to exert significant influence on the choice of equipment. Further, those that are necessarily small users of large equipment may find that they pay a [illegible]rice for the complex operating system that enables machine to handle a plethora of operations in sc[illegible]t[illegible] time sharing or multi-programming mode. The fact that the library's jobs might be handled on a smaller machine at lower cost is immaterial if the organizational ground rules forbid it to use such equipment in the general interest. The notion that a user should use the smallest library that is likely to provide an answer makes sense to any librarian or library user. The corresponding rule applies with equal force in the computing field, but is frequently much more difficult to implement within larger organizational structures. Nevertheless, the library community as a whole must continue to monitor computer cost structure as time passes in order to keep close track of when various interesting and useful computer-based access mechanisms pass over the cost margin and become cost effective.

Time sharing is one area that has long fascinated both librarians and library users — particularly that aspect devoted to on-line interrogation of large files. Many such systems have been designed, implemented, and made available to the public during the last five years. The various problems that plagued the pioneers in this field have now largely been resolved. Several on-line systems are now in routine use on very large data bases. However, the economic viability of such systems, except where substantial government support is available or a high premium

is justified for continuous update and instant information return, still is an unresolved question.

Some light can be shed on this situation by considering, at least in a simple fashion, how an on-line system works. Although details of implementation vary in important ways, all such systems have two basic data structures: the main data base consisting of a set of data elements (say, abstracts of technical papers and reports) and a unique identification number for each element; and associated with this main data base an inverted file which is a set of bibliographic tags each paired with every indentification number that has that tag. In other words, we have a collection of items and an access system for that collection. As new items are added to the main data base each is given its unique identification and the inverted file is completely revised to reflect the additional material.

In use, the potential user is instructed in the use of the "query" language that connects him to the data base by way of the inverted file. Having mastered this, he then uses a typewriter device connected by phone lines to the computer in order to request all documents containing certain information, that is, having certain tags in the inverted file. It is, of course, possible to generate by machine methods a "tag" for every word in the abstract, possibly skipping a list of "stop" words, so that the system is only limited by the size of the tagging system. Most of the more sohpisticated systems provide a semi-automatic cross reference facility that "corrects" spelling errors (through word compression schemes), and some even maintain cross reference files. Having entered the request and navigated past the Scylla of error correction and the Charybdis of inquiry reformation, the user is presented with a count of the number of items in the file satisfying his requirements. (The computer "replies" either by taking control of the typewriter and typing out a message or by displaying the information on a cathode ray tube.) If the resulting count is not too large, the user can insert a command that will cause each of the documents (or items) in the main data base to be brought forth to the viewing area for direct inspection. Once the problem of mastering the query language is solved, the whole system is rather appealing.

There remains, however, the question of cost. Let us consider an alternative to the usual on-line implementation. Suppose we construct the inverted file (which we shall now call an "index") exactly as before and maintain it and the main data base in machine readable form also as before. However, instead of connecting a set of typewriters (or "terminals") to the computer let us process the inverted file and the data base through a COM device and distribute microfilm copies of both to the users. In this mode, a microfilm reader replaces the terminal and the user scans the index to the section of the file of interest. Having found the item numbers corresponding

to his request in the index, he then scans the microfilm data base to get a look at the items themselves.

If the number of item lookups is not large, the two systems are closely competitive. A microfilm reader and a terminal generally cost about the same. Each requires some investment in time for the user to become effective in its use. The on-line procedure will generally provide faster response, but if the number of items is small, this is not important.

For the microfilm system, problems arise when the response to a request yields a large number of items. In the on-line system, the user can immediately devise a strategy to shorten the list by adding further constraints to his request. As the microfilm user cannot physically rearrange the entries on the film, some extra arrangement must be provided for him in the organization of the data base itself. This could be done quite simply by extracting all tags with more than a "reasonable" number of identifications and generating double (or even triple) tags wherein the popular tag is paired with every other word, or tag, in each of the documents containing that tag. If this had to be done for the whole file, it would lead to a prohibitively large index. But it only has to be done the tags with many identification numbers and these, as we know from our study of the structure of information distributions, are infrequent. Thus an index of perhaps twice the size of the on-line inverted file would provide essentially the same access capability for almost every case.

By structuring the problem is this rather peculiar way, a very simple cost comparison can be made. Both systems require the same data base preparation cost and the same cost for generating the inverted file. At the other end, the device necessary for the user is of about the same cost and both devices have the property that they can be used for other things. The cost difference then narrows to comparison of phone-line-plus-computer charges for the on-line user versus the cost of preparing and distributing the microfilm for the microfilm user. If there are not many uses per month, on-line wins out because the computer bill would then be less than the production of the master copy of the microfilm. However, as the number of uses increases, the microfilm approach becomes increasingly attractive and it inevitably becomes more attractive financially. The exact breakeven point depends not only on the current cost of computation, but also on the level of usage, the utility of weekly updates of the file versus monthly or longer cumulations, the options exercised with respect to a single master microfilm index versus a five-year index with more recent material cumulated monthly since the last five-year period, and on similar considerations. However, in most cases involving more than perhaps 25 user centers where instant updating is not required, it seems likely that the microfilm form will be more cost effective and nearly as effective without regard to cost.

Indeed, if the user body is widely dispersed geographically, phone-line charges will accumulate with sufficient rapidity to make the on-line system a loser even when computer time is free. Experts predict major decreases in telecommunications costs as well as in computing cost, so the economic potential for on-line information access will gradually expand during the next decade, even though some of these improvements will also tend to reduce the cost of COM output.

REFERENCES

1. Shoffner, Ralph M., "A Technique for the Organization of Large Files", American Documentation, (Jan. 1962) 95-101.

2. Lipetz, Ben-Ami and Song, C.T., "How Many Cards Per File Guide? Optimizing the Two-Level File," Journal of the American Society for Information Science 5 (March-April 1970), 140-141.

3. Dolby, J. L., Forsyth, V. J., Resnikoff, H. L., Computerized Library Catalogs: Their Growth, Cost, and Utility, M.I.T. Press, 1969.

4. Price, Bronson, Library Statistics of Colleges and Universities, Fall 1969, Data for Individual Institutions, U.S. Office of Education, Washington, D.C., 1970.

5. Aitchison, J. and Brown, J.A.C., The Lognormal Distribution, The University Press. Cambridge, 1963.

6. McAlister, D., "The Law of the Geometric Mean," Proceedings of the Royal Society, 29(1879), 367.

7. Fairthorne, R. A., "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction," Journal of Documentation, 25(1969), 319-43.

8. Tukey, J. W., "On the Comparative Anatomy of Transformation," Annals of Mathematical Statistics, 28, (1957) 602-632.

9. Resnikoff, H.L. and Dolby, J.L., "On Archival Access," ASIS Proceedings, 9(1970), 255-264.

10. Good, I. J., "Statistics of Language: Introduction," Encyclopaedia of Linguistics, Information, and Control, Pergamon Press, London, (1969) 567-81.

11. Price, D. J. De Solla, "Some Remarks on Elitism in Information and the Invisible College Phenomenon in Science," Journal of the American Society for Information Science, (March-April 1971) 74-75.

12. Kravitt, B. and Griffith, B. C., "Applications of Several Theoretical Distributions to Computational Linguistics," to be published.

13. Whitworth, W. A., "Choice and Chance," Cambridge: Deighton and Bell, (1901).

14. Mandelbrot, B., "On the Language of Taxonomy:an Outline of a 'Thermostatistical' theory of Systems of Categories with Willis (natural) Structure," Information Theory; Papers Read at a Symposium on Information Theory, London 1955, Butterworth, (1956) 135-45.

15. Carroll, J. B., "On Sampling from a Lognormal Model of Word Frequency Distribution," Computational Analysis of Present-Day American English (Henry Kucera and W. Nelson Francis), Brown University Press, Providence, Rhode Island, (1967) 406-24.

16. Houston, N. and Wall, E., "The Distribution of Term Usage in Manipulative Indexes," American Documentation, 15(1964), 105-14.

17. Dolby, J. L., Houchin, W. E. and Resnikoff, H. L., On the Compression of Library of Congress Subject Headings, R & D Consultants Co., Oct. 1969.

18. Artandi, S., "Automatic Book Indexing by Computer," American Documentation, 15(1964), 250-257.

19. Kuno, S. and Oettinger, A., "Syntactic Structure and the Ambiguity of English Words," AFIPS Conference Proceedings, Vol. 24, (1963) pp. 397-418.

20. Tukey, J. W., New Approaches to Automatic and Semiautomatic Indexing and a Citation Index for Statistical Methodology, Final Report to the National Science Foundation, Princeton University, 1971.

21. Dolby, J. L., "The Structure of Indexing the Distribution of Structure-Word-Free Back-of-the-Book Entries," ASIS Proceedings 5(20-24 October 1968), 65-72.

22. Lipetz, Ben-Ami, "Catalog Use in a Large Research Library," Library Quarterly, (1972), 129-139.

23. Morse, P. M., "Measures of Library Effectiveness," Library Quarterly, (1972), 15-30.

24. Mendenhall, T. C., "The Characteristic Curves of Composition," Science, 9, (214, supplement) (1887), 237-49.

25. Yule, G. U., The Statistical Study of Literary Vocabulary, Cambridge: The University Press, 1944.

26. Williams, C. B., "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style," Biometrika, 31(1940), 356-61.

27. Dolby, J. L., "A Quick Method for Choosing a Transformation," Technometrics, 5(1963), 317-325.

CONCLUSIONS AND RECOMMENDATIONS

Whereas the authors have attempted to maintain a high standard of objectivity in the presentation and analysis of data and theories in the body of this monograph, this summary of our conclusions and recommendations necessarily reflects our personal interpretation of their reliability and significance as well as our undoubtedly biased but strongly held views about the practical applicability of conclusions derived from the theoretical development. The reader undoubtedly will not need to be reminded of this in what follows. Our recommendations follow naturally from our conclusions; we have therefore thought it simplest to intertwine the two in the order of their natural occurrence.

1. Information stored in large data bases primarily for use in information access systems (such as libraries) is structured so that it approximates the optimal configuration described by the level structured access model. In this model a set of access stores is associated with each primary data store, and each access store is approximately 1/30 the size (measured in characters) of the next larger store. At any fixed time, the distribution of size of the data objects (e.g. books) constituting one access level in the model is lognormal; the lognormal standard deviations are constant for the size distributions corresponding to the various access levels constituting one access system.

> R1.1 We recommend that every access system include an access subsystem for every possible level in order to maximize the cost effectiveness of the access system. If the size of the primary data store is N characters, there should be just $\log_K N$ levels (more precisely, the nearest integer to that number), including the level of the primary store, where $K \doteq 28.54 \doteq 30$ is the level structure constant determined from observations.

2. The cost of construction (including acquisition) and maintenance of each level of an access system should bear approximately the same ratio to the cost of construction and maintenance of the primary information store as the size of the access level store does to the size of the primary store in order to maximize cost effectiveness. Thus, a first order access subsystem (the largest proper subsystem) should cost about 1/30 the cost of the primary store. Moreover, the cost of the total access system, including levels of access of all possible orders greater than zero, should be approximately $1/K + 1/K^2 + 1/K^3 + \ldots \doteq 3.5\%$ the cost of the primary system.

R2.1 We recommend that existing information access systems be analyzed for cost effectiveness by comparing their cost of construction and maintenance with that of the primary data base they access. Excluding special situations wherein the value of the accessed information or of its timely acquisition is exceptionally great (as is the case for certain medical, national security, and other real time applications), access subsystems of order m whose maintenance cost is significantly greater than $1/K^m$ the cost of the primary data base should be eliminated. Conversely, organizations which spend significantly less than 3.5% of the acquisition and maintenance cost of their primary information store on access system construction and maintenance should increase their expenditures. If they cannot, they should consider eliminating their information facility and purchasing information services elsewhere since it is unlikely that their system can be either cost effective or effective.

R2.2 Large information systems frequently maintain several access subsystems which function at the same level. When more than $\sqrt{K} \doteq 5$ such subsystems operate at the same level, they have a cumulative size closer to that of an access subsystem belonging to the next higher level, but they do not normally provide access equivalent to a higher level system. We therefore recommend that organizations monitor the proliferation of access subsystems belonging to one level; more than 5 should not be permitted to operate at one level (with special exceptions related to the time value of information, as noted above). If more than 5 subsystems operating at level n appear necessary, it is likely that one new system of level (n+1) should be constructed to replace all but one of the existing level n systems.

3. The size of a classification system should vary as the logarithm of the size of the collection it classifies. In normal periods of historical duration, collection growth will be exponential, which implies that the classification system should be expanded linearly with time. As analysis of the American History subcollection of the Widener Library shows, one new subclass is introduced every 14 years (average).

R3.1 A study should be undertaken to determine whether the Widener American History classification dynamics is typical of other subject areas and libraries. Moreover, the optimal number of classification categories should be determined, where optimality is determined by search cost effectiveness.

4. Collection size is the principal barrier to access. Consequently, users should be encouraged to go to the smallest collection that is likely to provide the needed information.

> R4.1 We recommend that libraries enhance and extend their efforts to describe their collections to routine and periodic activities which adhere to a national standard for the statistical description of information holdings. Moreover, holdings statistics should be published periodically in a standard format and made broadly available to user populations.

5. Traditional means of providing subject access to library monograph collections require supplementation. Publication of amalgamated indexes to selected books in each subject area with periodic and partially cumulative updating provides one method of accomplishing an expansion of monograph subject access.

6. Algorithmic indexing of full text and abstracts is now cost effective for all documents which have been put into machine readable form for other purposes. Increasingly, publishers of books and abstract journals routinely commit the of their publications to machine readable form in the process of typesetting. They should be encouraged to put indexes in machine readable form (whether the indexes were machine derived or not) to simplify the task of amalgamating indexes from various sources for use in access subsystems.

> R6.1 We recommend the establishment of a standard national format for index material for both books and journal literature. Moreover, there should be a central national authority charged with the responsibility of collecting, standardizing, and distributing index information, and statistical measurements of information value therefrom derived.

7. Progress in linguistic computation has now reached the state where it is usually possible to accomplish more by machine than is necessary. Increased attention should be directed toward the accumulation of operating statistics of the frequency of occurrence of various linguistic structures and entities in the processing of titles, special characters, author variants, transliteration variants, etc., to simplify sound planning of future systems and to provide objective means for the evaluation of the cost effectiveness of existing systems.

8. Library networks and other cooperative methods for increasing access to library materials can be assessed by using the level structured access model, and are subject to the general restraints imposed by it. Because of the large capital investment represented by such cooperative arrangements, careful statistical analysis of both the user population and the library

holdings should be made in order to insure that the resulting agglomerations are cost effective in the sense of the information theoretic measures introduced above as well as in the more usual financial measures.

9. As the operating costs of on-line information systems decrease per unit inquiry, increasing attention should be directed to their potential economic utility. Proposed applications in this area must be compared with other access systems utilizing the same information store that may provide a non-interactive output using microform or traditional printing techniques. Comparable measures of delivered information per unit inquiry cost -- including user education and skill maintenance costs -- should be used to evaluate competing systems.

10. Finally, although the level structured access model introduced in this monograph is derived from theoretical considerations of a fundamental character, and in substantial agreement with a large collection of observational data collected from widely varying sources, we cannot assert and do not believe that it has received sufficient analysis or comparison with existing information systems to be uncriticlly accepted as the natural standard for the design and evaluation of information systems. On the other hand, its full range of applicability has not yet been plumbed. We know, for instance, that it successfully describes the distribution of batch turn-around time in a medium size university computing center, and deviations from its predictions in that case have turned out to be of importance to management in the task of detecting and controlling deviations from optimal effectiveness. But even this application remains untested for other computing centers with other equipment and a different mix of users, procedures and financial constraints.

> R10.1 We therefore recommend that the level structured model of access systems be studied to determine what modifications may be necessary to enable it to describe a broader range of information systems, and that the limits of its applicability be established, in part by comparing its predictions with statistical observations representing various types and sizes of libraries not included in the present study.

LEVELS OF

INFORMATION STORAGE AND ACCESS

In the previous chapter we have stressed the view that the problem of insufficient access is primarily a problem of the great size of the archive to which access is desired. This study is directed toward problems of library archives and in this context it is access to the content of books and collections of books that is of immediate concern although libraries are increasingly becoming archival depositories of other types of information bearing records.

There are technical reasons that make it desirable to restrict attention--at least in a preliminary study such as the present one--to the monograph collection; we will have some useful remarks to make about serials and can also exhibit data supporting the extension of the model that will be proposed to describe the serial collection.

The book is a natural halfway house in the hierarchy of means for storing written information in libraries. Within the book are usually to be found certain standard apparatus which aid in directing the user to the internal location of information with which the book is concerned; these include, in descending order of size, the index, the table of contents, and the title. The library itself is of course a collection of books but it too contains certain apparatus for directing the user to those amongst the many books held that contain information concerning some particular matter; these include, in increasing order of size, the classification system, the reference section, and the card catalog. There are also other types of traditional access means that aid in locating books which contain certain information, including special bibliographies and, too often overlooked, the reference librarians. If indeed size is the predominant factor determining the need for access, then a study of the size of the various natural bibliographic units named above may shed light on the structure, if any, of the traditional access systems and thereby also provide guidelines for those who study the possible ways for increasing and automating the means of access.

We will proceed up the scale of size of the naturally occurring access means associated with books and collections of books, with the intent of determining the statistical distribution of size of each such system; this information will lead in a natural way to the level structured model of access systems briefly described in Chapter I.

Initially limiting our attention to the book itself, there are four systems of interest:

1. Title
2. Table of Contents
3. Index
4. Book Text

In each case we wish to know the mean (average) size of the item in question, measured, let us say, by the number of characters (including the interword space) contained in the item. Moreover, it will turn out to be important to know the distribution of size for each case so that it will be possible to say to what the extent the mean is characteristic of the distribution and also because the distributions will turn out to have an intrinsic connection with the access problem via the intervention of the mathematical discipline known as information theory; this latter aspect of our study will be described in Chapter III.

It is not easy to obtain reliable statistics about the size of bibliographic units; it is especially difficult if general samples that are not restricted to one or a few fields of interest are desired. We have based our book studies on the Fondren Sample, a random sample of 1926 cards drawn from the shelf list catalog of the Fondren Library at Rice University in 1968; it has been described in some detail in Ref. (1). Associated with each shelf list card is one or more monographs; these monographs constitute the sample on which our study is based. It is appropriate to refer to it as a random sample of books from a medium sized university library.

Because we are interested in studying the interaction of the various traditional access systems used in books we have extracted from the Fondren Sample all those books that contain an index (here and throughout all that follows, index will of course mean back of the book index), thus yielding what we have called the Fondren Index Sample, which may reasonably be called a random sample of indexes. There are of course certain

unavoidable biasses present in this index sample: the Fondren Library does not have an adequate collection in medicine or law, for instance; it has an exceptionally fine collection in other areas. But, to the best of our knowledge, these samples are the closest in existence to truly random samples of books and of books with indexes belonging to the complete population of all books ever published.

With these preliminaries in mind we can now turn to study the structure of book titles. Figure 2.1 displays the distribution of the number of characters per book title for books from the Fondren Index Sample drawn on lognormal probability graph paper. The mean number of characters per title is 28.15.

Next consider the size of a table of contents measured by the number of characters it contains.

Although the "structure" of a book title is relatively standardized, the same cannot be said of the table of contents. Some books include phrases such as "Chapter 1", others simply record "1" to designate the first chapter and others do not bother to indicate the chapter ordinal at all. There are tables of contents which include, in addition to a chapter title, relatively extensive descriptions of the text content of a narrative nature; others include section titles. Despite the rather excessive degree of variation that does occur, there are certain components of a table of contents which appear to be nearly invariable in their presence, including the chapter titles and page number designating the beginning of each chapter. We have chosen to define the table of contents as that portion of the material contained in what is normally termed the table of contents that corresponds to the chapter title, excluding from consideration all headings, chapter ordinals, appendices, tables of figures, etc., and page number referents to the location of chapter initial pages. With this convention, a random subsample of 161 tables of contents was selected from the Fondren Index Sample and the number of characters (including interword space characters) was counted for each selected table of contents. It turns out that the mean size of a table of contents defined in this way is 505 characters. Figure 2.2 displays the distribution of table of contents size for this subsample.

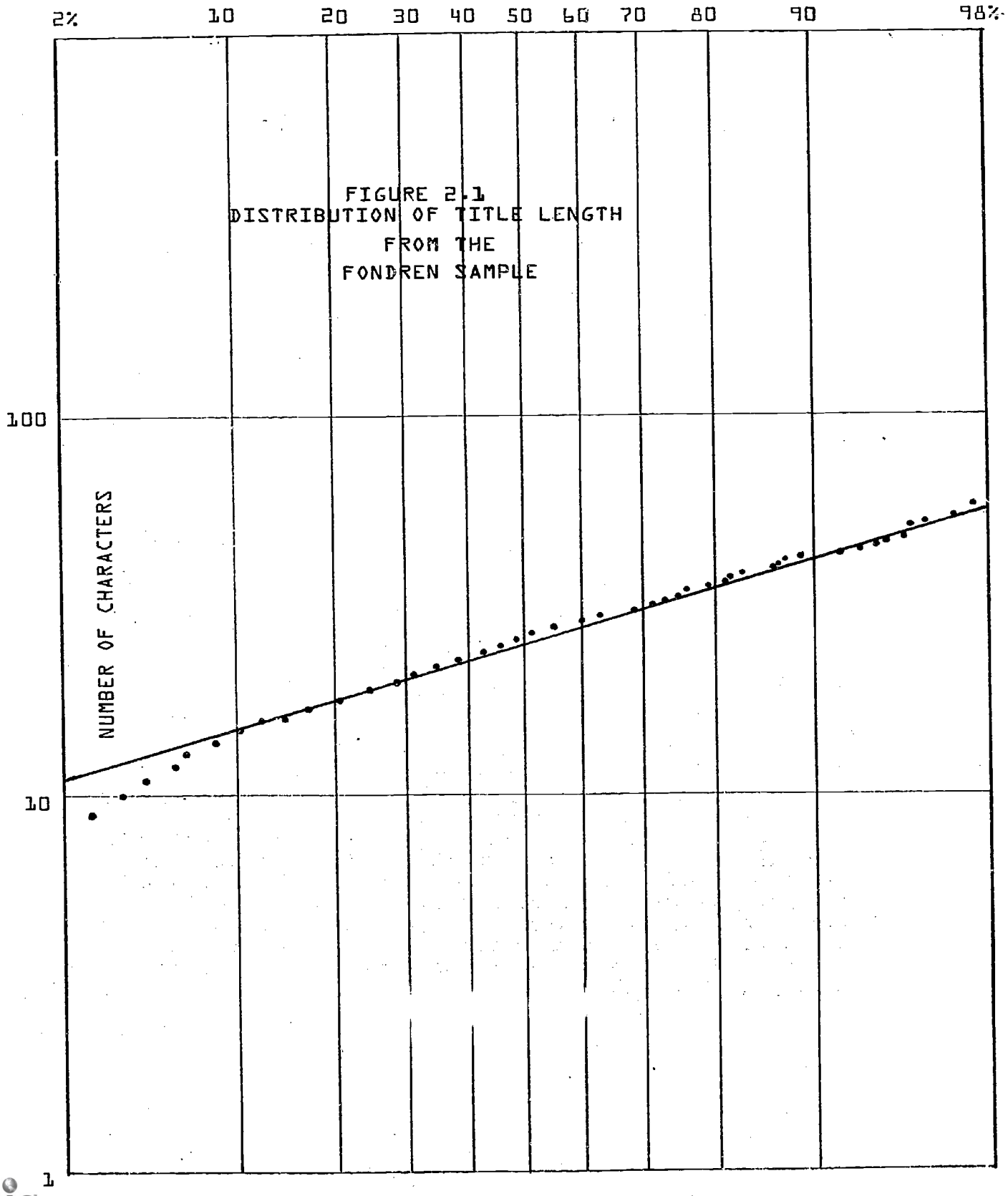


FIGURE 2.1
DISTRIBUTION OF TITLE LENGTH
FROM THE
FONDREN SAMPLE

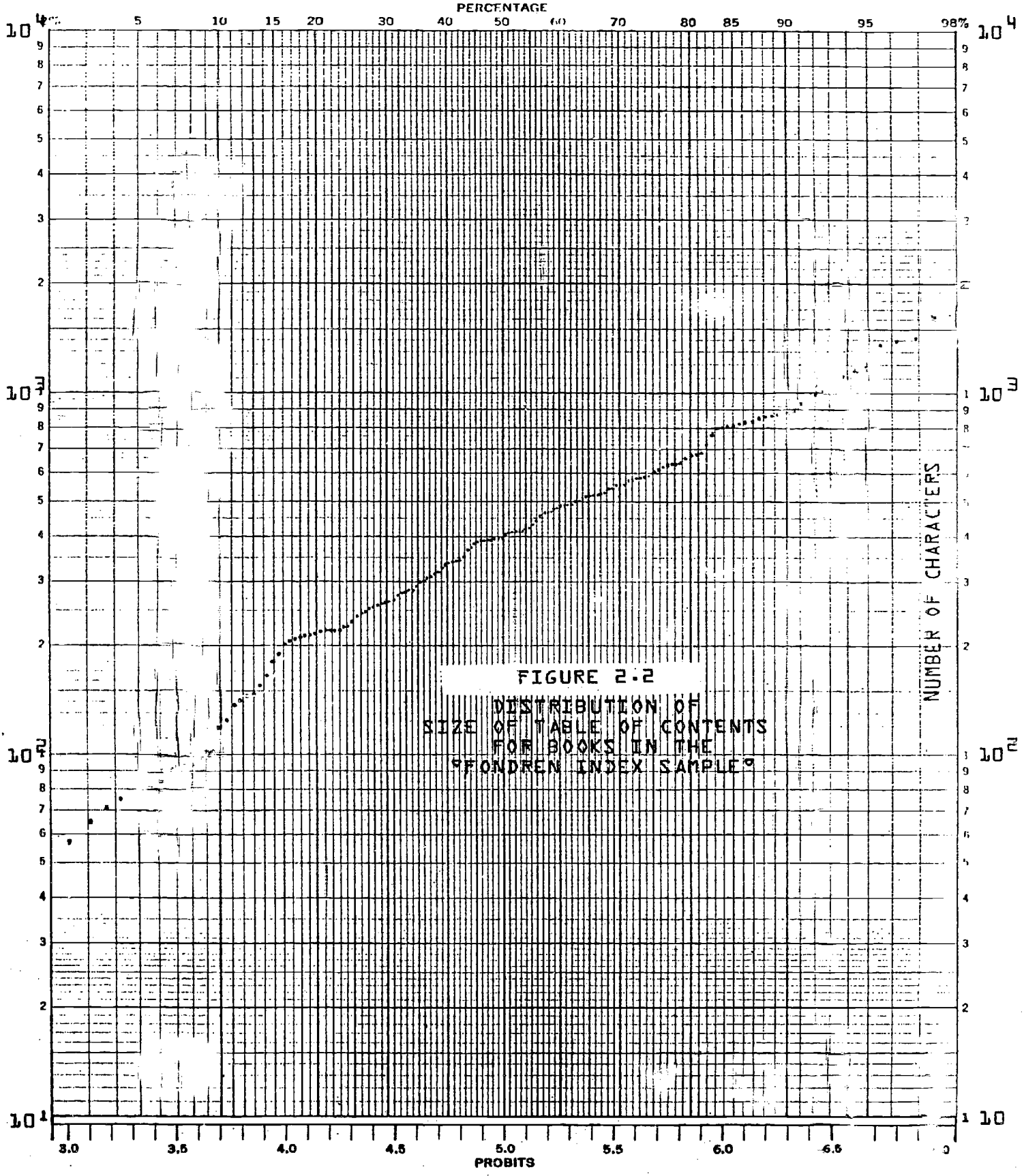


FIGURE 2.2

DISTRIBUTION OF SIZE OF TABLE OF CONTENTS FOR BOOKS IN THE "FONDREN INDEX SAMPLE"

The reader can hardly help but notice that the data exhibited in each of Figures 2.1 and 2.2 fall nearly along a line, and moreover that the two lines have similar slope. The graph paper is so designed that straight lines indicate that the data are drawn from a lognormal distribution, whose properties will be discussed later on in this chapter and extensively in Chapter III; it suffices here to stress that thus far the data indicates that the two lowest levels of distribution of size of book access systems belong to some well known family of statistical distributions and indeed to the same family. We will want to look for this possibility when examining data referring to other access systems.

The index is the next largest access tool traditionally found in books, and from many points of view it is the most important and responsive to the detailed demands of the user. It therefore deserves extensive examination.

The Fondren Index Sample consists of 706 indexes. Chapter IV investigates the relationship of indexed books to the unindexed books in the Fondren Sample and studies such properties of the indexed books as their distribution among the Library of Congress classification categories. Here we are only interested in considerations of size. The mean number of index entries per index is 836.

Figure 2.3 contains the distribution of the number of index entries per book, again on lognormal probability graph paper. It is evident that the data can be accurately approximated by a line and furthermore that the line has a slope which once again is similar to the slope of the lines occurring in the previous two figures. One word of caution: here only the number of index entries is exhibited. Ideally one would wish to measure the size of an index by the number of characters it contains, but it would not be feasible to count the characters in more than half a million index entries. Furthermore, once again the question of which characters to count can not be resolved in a completely unambiguous way. For instance, it is easy to agree whether page reference numbers should be counted, and what to do about consecutive spaces used as separators, but format problems related to multiple entries grouped under a common initial phrase, and inverted order entries demand operational decisions that are not often guided by a clear cut purpose. These problems exist when entries alone are counted, but they are magnified when characters are counted. We have agreed, when counting entries, to count

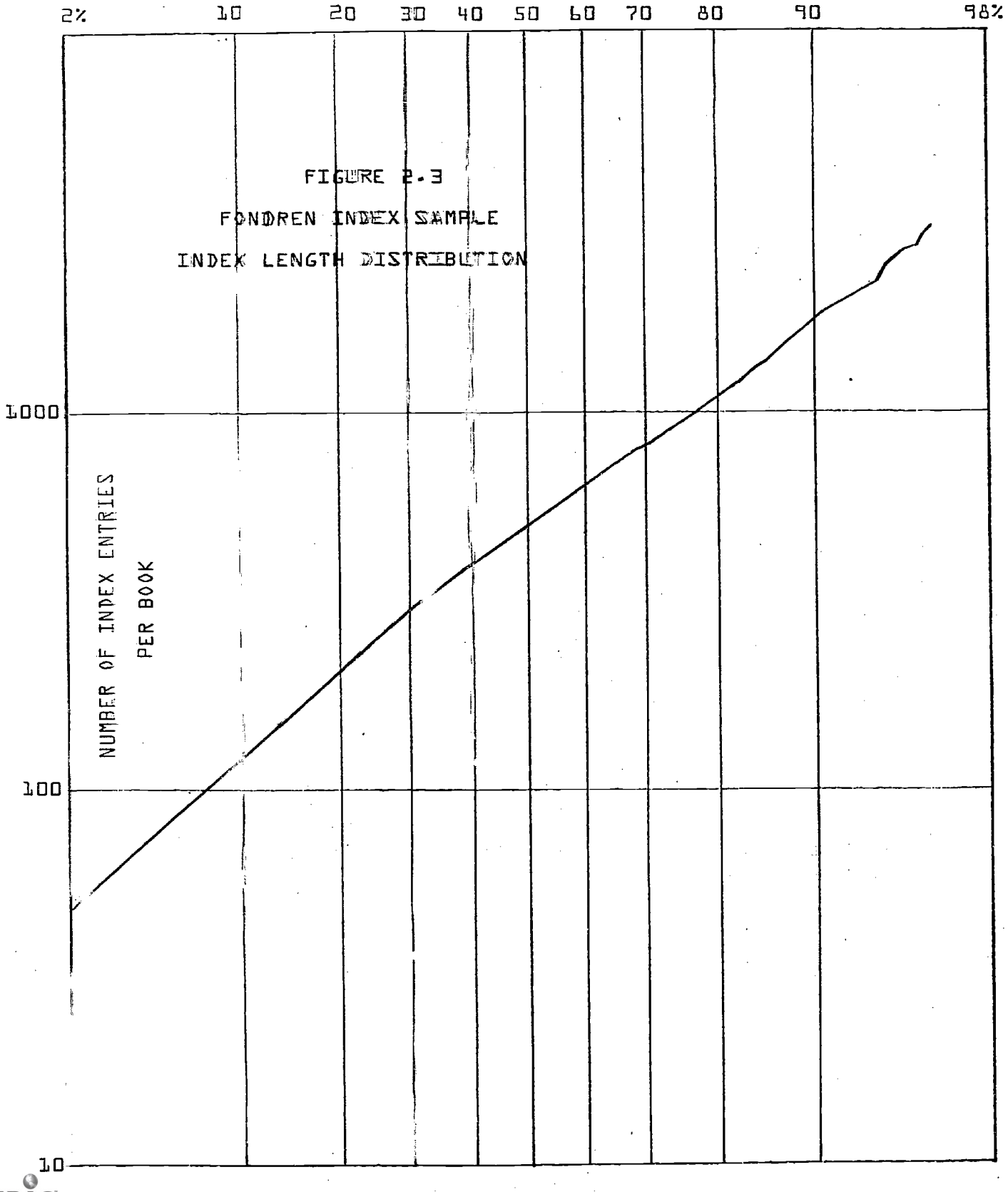


FIGURE 2-3

FONDREN INDEX SAMPLE

INDEX LENGTH DISTRIBUTION

each group of page reference numbers: this defines the index entries, at least as far as their cardinal number is concerned, and provides a relatively clear cut procedure requiring a minimum amount of subjective decision by the persons performing the counting. In order to obtain an approximation to the number of characters contained in an index, a rather indirect procedure was used. We have in a convenient form all of the index entries contained in 80 books in the field of statistics, all printed in a fixed typefont whose characters are of constant width, and printed a fixed number of lines to the page. These characteristics make it possible to count the number of characters in an entry by measuring the length of the entry. This was done for a uniform subsample (comprising about 1.75% of the total Statistical Index Sample of 31,232 index entries). Table 2.1 lists the number of entries consisting of from 1 to 76 characters, and, opposite 77 characters, the number of entries that had at least 77 characters. The mean number of characters per entry, exclusive of page reference numbers but inclusive of interword spaces, is 25.47. Figure 2.4 displays the distribution of size of the entries in the Statistical Index Sample. If we assume that the distribution of size of index entries is independent of the distribution of the number of entries per index, then the average number of characters per index will be the product of the average number of entries by the average number of characters per entry. Using the number for the Statistical Index Sample for the latter, we find that the average number of characters per index (exclusive of page references) is 836 x 25.47 = 21,293. If it be assumed that there are typically three digits and an interword space required to provide the page reference location information, then augmenting the average number of characters per entry by 4 leads to 24,560 characters per index (inclusive of page reference approximation).

The distribution of index entry length for the Statistical Index Sample is, again, lognormal to a high degree of approximation.

Table 2.1

STATISTICAL INDEX SAMPLE

Distribution of Entry Length in Characters
(excluding page references)

No. of Char.	No. of Entries	No. of Char.	No. of Entries
1	0	41	3
2	0	42	3
3	4	43	6
4	2	44	3
5	4	45	5
6	7	46	5
7	8	47	3
8	9	48	4
9	20	49	4
10	22	50	2
11	19	51	4
12	18	52	4
13	14	53	3
14	23	54	6
15	11	55	4
16	28	56	3
17	16	57	2
18	8	--	
19	14	62	1
20	12	63	1
21	17	64	1
22	17	65	1
23	19	--	
24	19	67	1
25	14	--	
26	8	69	1
27	11	--	
28	9	72	1
29	11	73	1
30	7	--	
31	10	75	1
32	9	76	2
33	10	≥77	9
34	5		
35	5		
36	4		
37	5		
38	10		
39	10		
40	9		

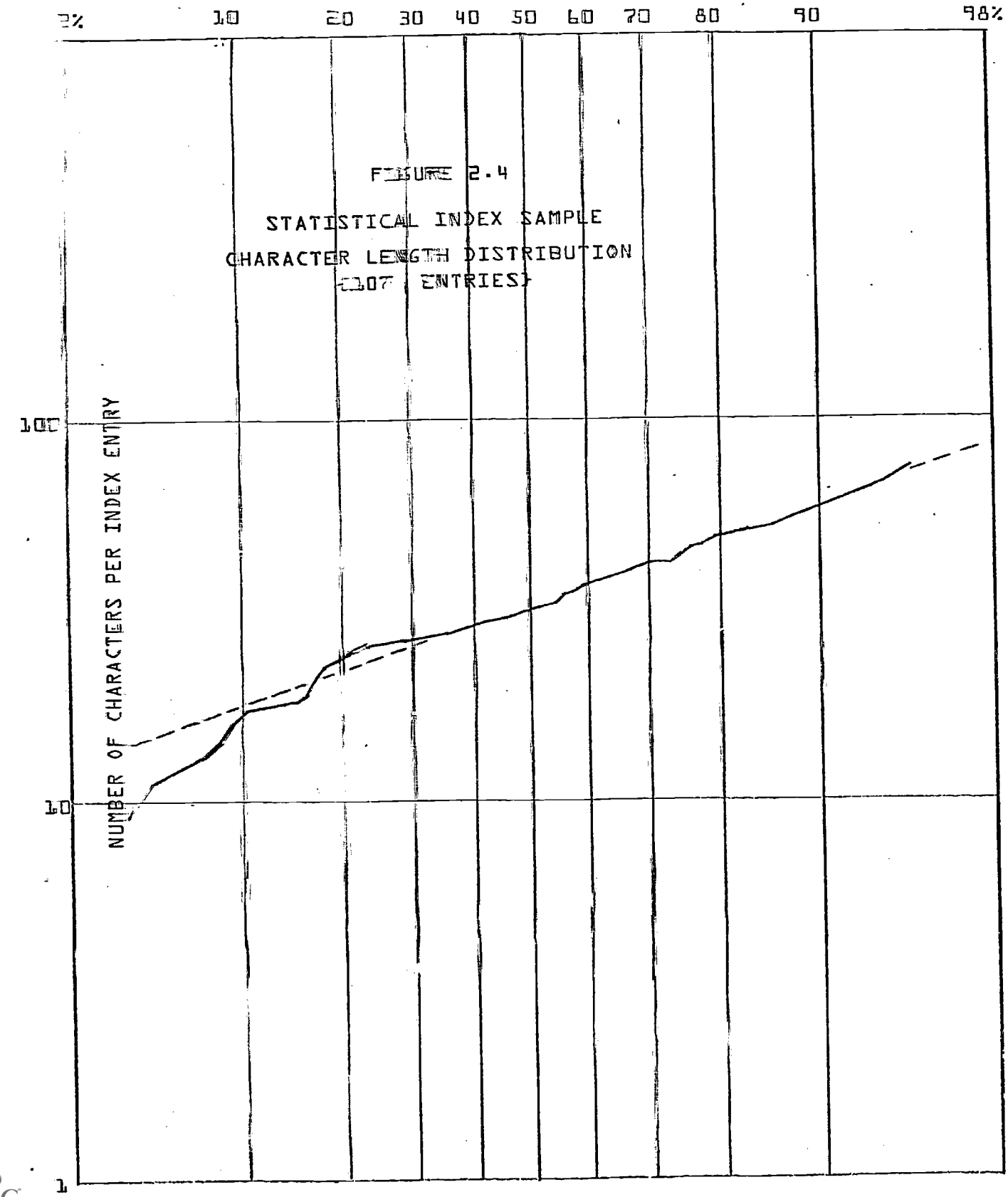


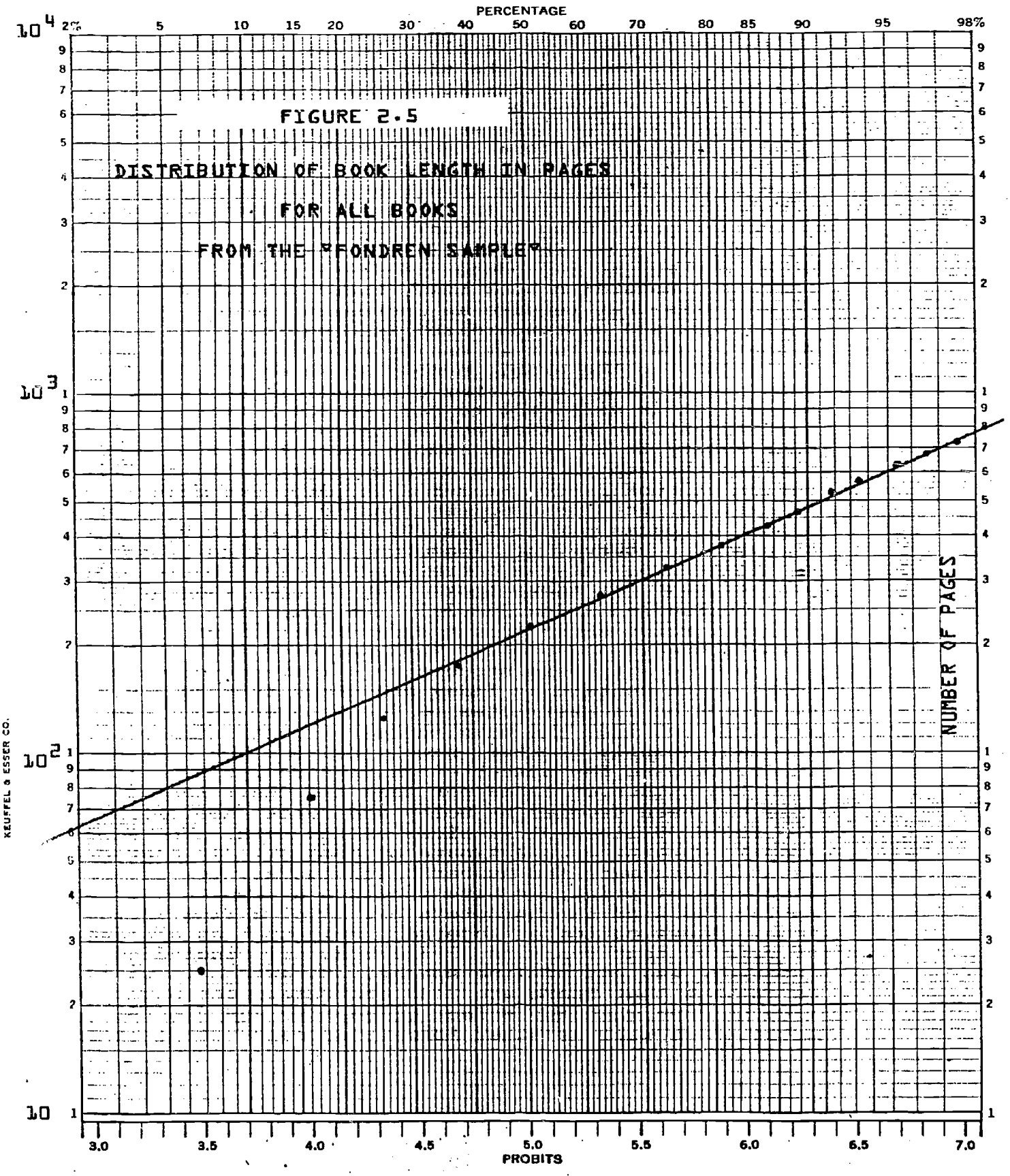
FIGURE 2.4

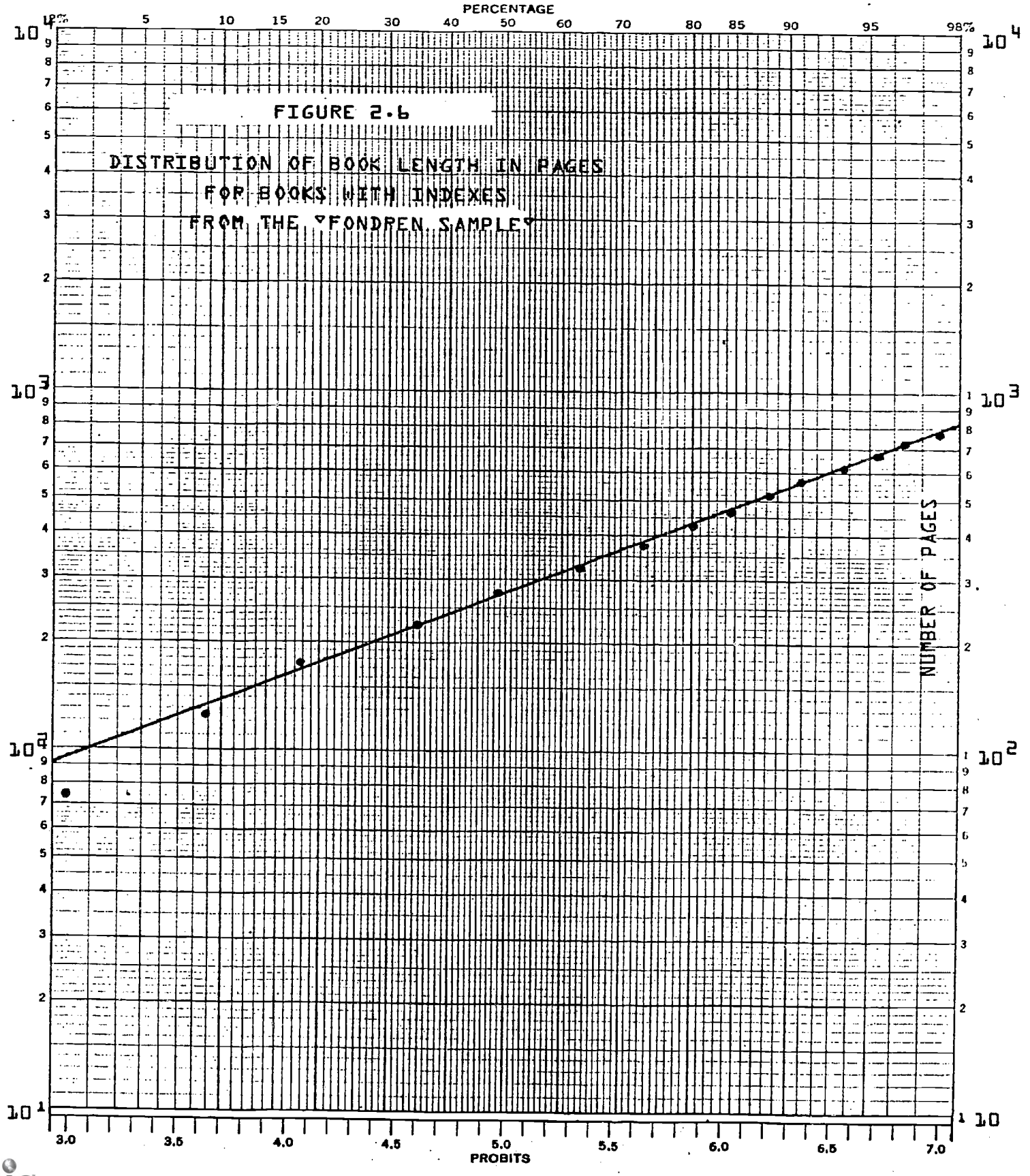
STATISTICAL INDEX SAMPLE

CHARACTER LENGTH DISTRIBUTION

(207 ENTRIES)

The last of the four natural access tools for monographs is the monograph text itself. It will be even more difficult to estimate the size of a book measured by the number of characters it contains because of the variability of type font and page layout supplemented by the presence of tabular and figured material. Although numerous different and justifiable procedures of making such a size estimate are conceivable, we have once again attempted to choose a method that would be simple and insensitive to subjective judgements of the personnel performing the task in order to improve accuracy but more importantly to make it possible for other workers to reproduce (at least nearly) our results. Regarding book text, there are several levels of analysis that require an increasing amount of extraneous and unstandardized information. The simplest measure, and one that it easily reproduced, is simply to transcribe the arabic number shown on the catalog card designating the number of non-front matter pages. It is difficult to say precisely which pages are represented by that number in each case, but it is unnecessary to do so; we simple agree that this number defines the length of the book in pages. The distribution of book length measured in pages was determined in Ref. (1) for the complete Fondren Sample. The mean number of pages per book is 276.6; the distribution of pages is however not lognormal as is readily seen in Figure 2.5. If the corresponding distribution is plotted just for those books that do have indexes (i.e., for the Fondren Index Sample), the graph in Figure 2.6 results, which shows that the distribution of size of these books is lognormal. This suggests that there may be some intrinsic structural difference between books which contain an index and those that do not. If attention is restricted to the Fondren Index Sample, it turns out that the mean number of pages per book is significantly greater, namely 341.5. The next step in determining the number of characters per book is to find the number of lines per page and their length; this has been studied by Dolby and Jones (Ref.(2)), who found 38 lines of 24 picas as the mean. The final step in obtaining an estimate of book size in characters is to approximate the number of characters per 24 pica line of print; we have analyzed a sample of printed matter and find 63 characters per 24 pica line as the mean. These estimates together imply that an average page of printed text contains 2394 characters, including interword and end of line spaces. Hereafter it will be assumed that there are 2400 characters per page. We have no idea what the effect of tabular and figured material as well as other formatting conventions is on


PERCENTAGE
2%
5
10
15
20
30
40
50
60
70
80
85
90
95
98%
FIGURE 2.5
DISTRIBUTION OF BOOK LENGTH IN PAGES
FOR ALL BOOKS
FROM THE "FONDREN SAMPLE"
NUMBER OF PAGES
PROBITS
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
KEUFFEL & ESSER CO.


PERCENTAGE
1%
5
10
15
20
30
40
50
60
70
80
85
90
95
98%
FIGURE 2.6
DISTRIBUTION OF BOOK LENGTH IN PAGES
FOR BOOKS WITH INDEXES
FROM THE "FONDREN SAMPLE"
NUMBER OF PAGES
PROBITS
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0

the estimate of book length in characters; nevertheless, excluding these matters from consideration, we find that the average book in the Fondren Index Sample is 341.5 x 2400 = 819,600 characters in size.

Turning now to collections of books, let us first consider the university library. Here it is essential that the notion "university" be specified in some way so as to enable one to distinguish university libraries from libraries of colleges in a manner consistent with that used for other purposes by governmental agencies and the educational institutions themselves. We implicitly use the definition used by the Office of Education of the Department of Health, Education and Welfare because we use their statistical data book Reference (3) as our source of information about the holdings of college and university libraries.

Unfortunately the data presented in Reference (3) is incomplete; notable ommissions are the University of Chicago and Yale University. Although these ommissions undoubtedly will have some influence on the statistical parameters of the distributions of interest to us, these will most likely be quite minor and in no event can they be expected to change the form of the distribution nor substantially affect its mean or variance.

There is one other defect of the data presented in Reference (3) which is more critical for our concerns. Most state university systems have had their statistics amalgamated; thus it is impossible to determine (from this source) the size of the library of the University of California at Berkeley--only the total number of volumes held in the entire California university system is presented. This unfortunate state of affairs holds for most of the other state systems also and tends both to depress the number of distinct university libraries and inflate the size of those that remain. Two factors permit us to extract useful information from this tabulation despite its amalgamated nature: first, it is easy to obtain lists of all units belonging to a state system (and also for the few private systems that operate more than one campus) and thereby estimate the total number of libraries whose structure must be studied. Second, within state systems there is usually one 'giant' library and a number of much smaller ones; this has the consequence that the departure of the distribution from lognormality, as is shown in Figure 2.7 which we will shortly consider, is diminished when the separate system units are accounted for, and, in view of the smallness of the possible effect, it is not necessary for us to study this difference in detail. Furthermore,

we can easily obtain the mean size from the revised estimate of the number of libraries. By adjusting the number of libraries represented in Reference (3) through deletion of the special dental and medical school branches and addition of all general campuses, a total of 201 university libraries is attained. The total number of volumes held in these institutions is 152,230,163 (nearly one for every inhabitant of the United States, and nearly as many as are held by all public libraries), so the mean number of volumes per university library is 757,364. The range in size may appear remarkable to the reader, ranging as it does from some 100,000 volumes to more than 8 million. Figure 2.7 exhibits the size distribution, which, as we have by now come to expect, is lognormal.

Knowing that the average book contains 819,600 characters and assuming that the distribution of book size is independent of the distribution of university library size, we readily find that there are some 620,735,534,400 = 6.2×10^{11}, or approximately 620 billion characters stored in the average university library.

At this point we have established the mean size and distribution of size for book based bibliographic entities ranging in average size from about 30 characters up to 620 billion characters, entities which differ in size by a factor of 20 billion. Our immediate task is to demonstrate that there is a simple and reasonable model which encompasses the entire range of bibliographic entities in a systematic way, relating those of one size to those of another in a uniform and unvarying manner.

In order to proceed, recall that the book title, table of contents, index, and text are four bibliographic units of increasing average size; let us say that they belong to levels 1,2,3,4 respectively. Let Y_n stand for the base 10 logarithm of the average size of the units belonging to level n; Figure 2.8 displays the points whose coordinates are (n,Y_n) for $n = 1,2,3,4$, and also the point $(8,Y_8)$ where Y_8 is the base 10 logarithm of the mean size of a university library, and the point $(7,Y_7)$ where Y_7 is the base 10 logarithm of the mean size of a two-year college library, obtained by analyzing the first 206 two-year college libraries listed in Reference (3); this procedure is biassed, leading to a slightly high estimate of the mean size of two-year college libraries because the State of California dominates the initial part of the list both in number of two year colleges and in the size of their libraries,

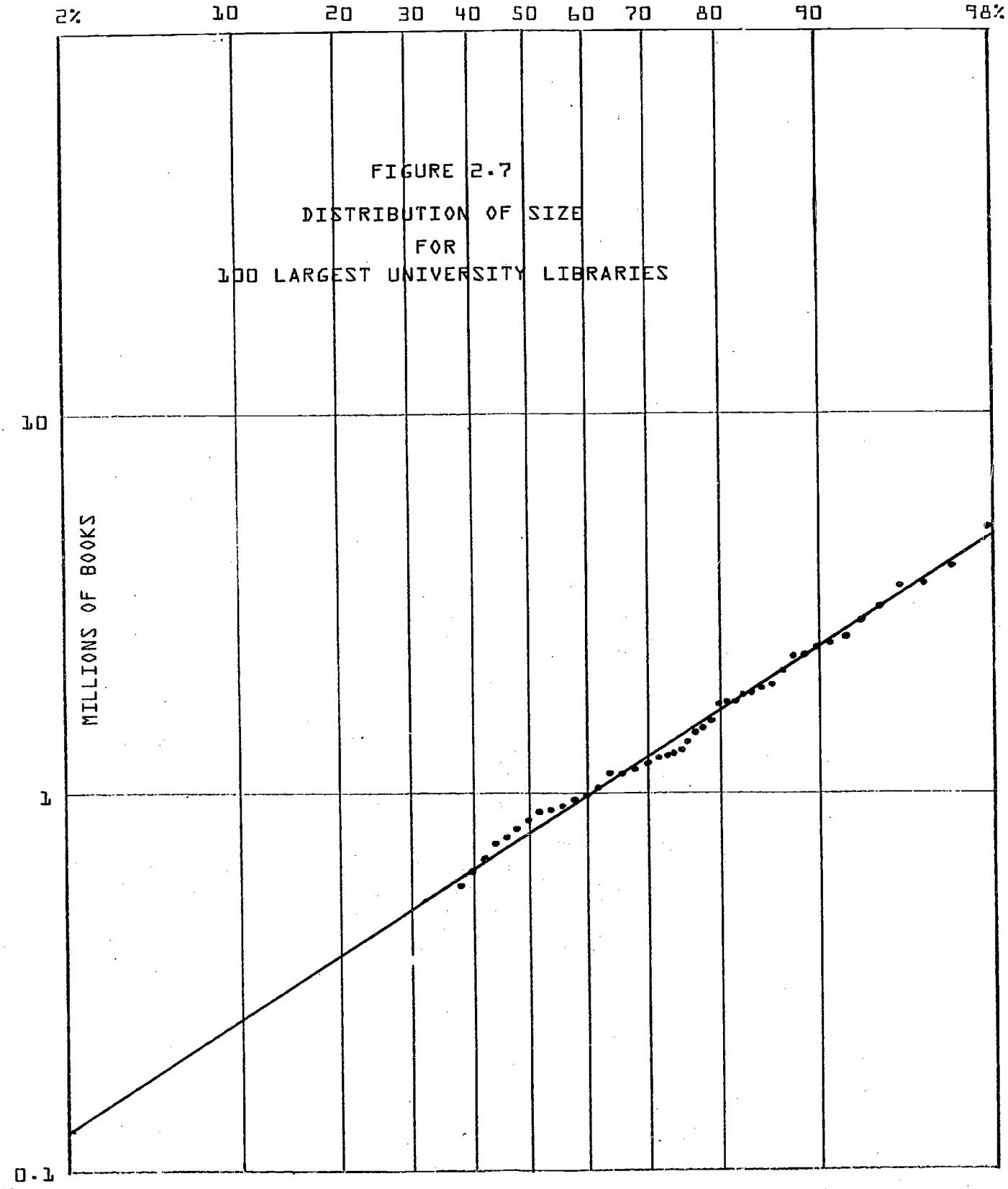


FIGURE 2.7

DISTRIBUTION OF SIZE

FOR

100 LARGEST UNIVERSITY LIBRARIES

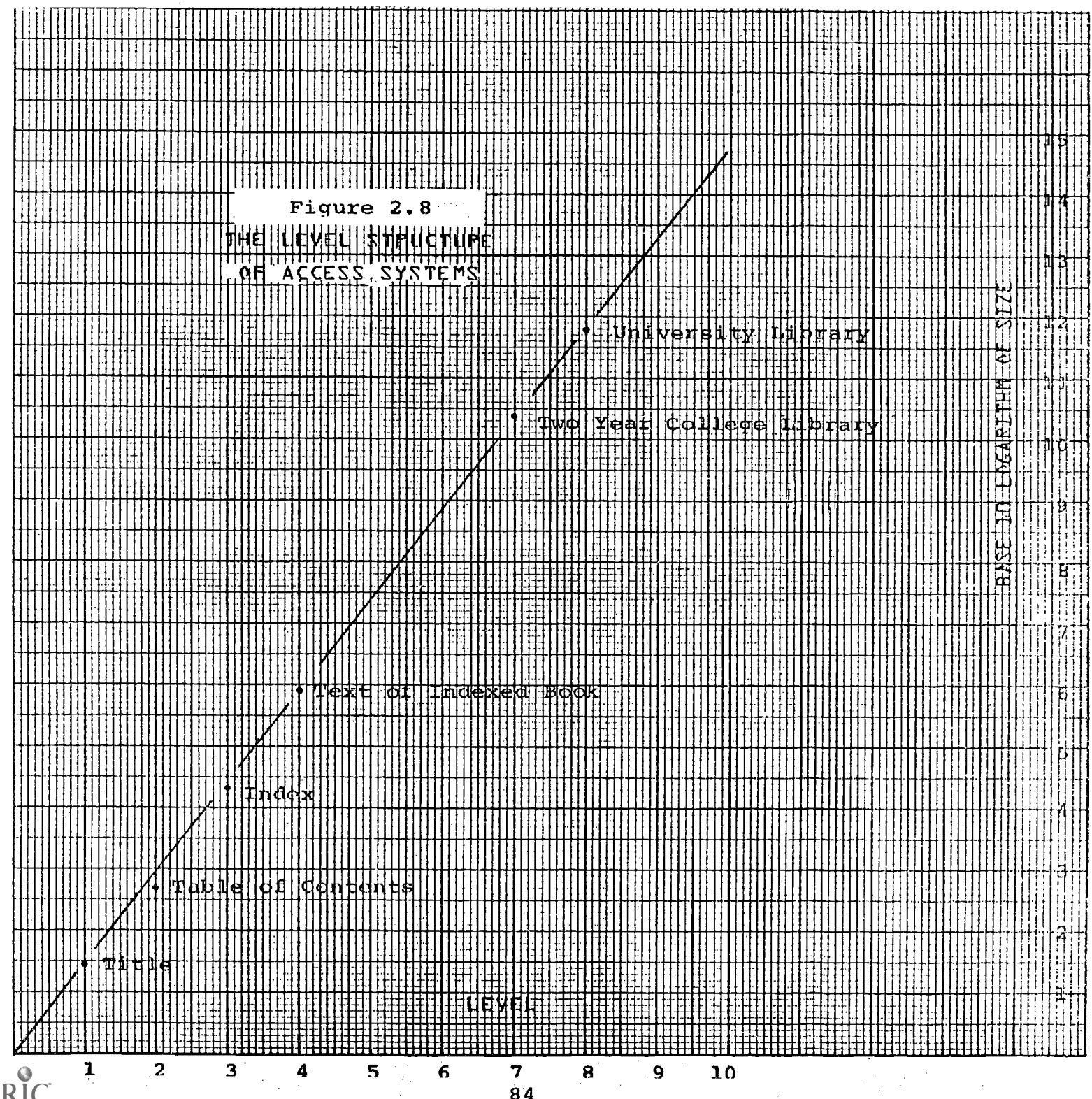


Figure 2.8
THE LEVEL STRUCTURE
OF ACCESS SYSTEMS
University Library
Two Year College Library
Text of Indexed Book
Index
Table of Contents
Title
LEVEL
BASE 10 LOGARITHM OF SIZE
1
2
3
4
5
6
7
8
9
10
11
12
13
14

but analysis of the complete list in Reference (3), which is presently underway, will undoubtedly lower the mean size insignificantly from the value 29,912 volumes used to determine the corresponding point in Figure 2.8.

Figure 2.9 confirms that the size distribution of two year college libraries is lognormal and that the slope of the line representing the data on that graph is once again comparable with the slope of lognormal distributions presented in previous figures in this Chapter.

Inspection of Figure 2.8 may lead the reader to wonder whether levels 5 and 6 correspond to naturally occurring collections of books; we think that level 5 corresponds to general encyclopedias and level 6 to personal libraries, but we have not ventured to include calculations based on these hypotheses because of the difficulty of amassing reliable and comprehensive statistical information in their support.

The points in Figure 2.8 evidently lie very nearly on a straight line. This means that the mean size, s(n), of the bibliographic units comprising the n-th level is related to n by an equation of the form

$$s(n) = a10^{bn} \qquad (2.1)$$

where a and b are constants. It is natural to suppose that a = 1 so that level 0 corresponds to the single character; we will examine the data given in Figure 2.8 and Table 2.2 which corresponds to it to see if it is consistent with this desirable and simplifying hypothesis. By a standard application of the

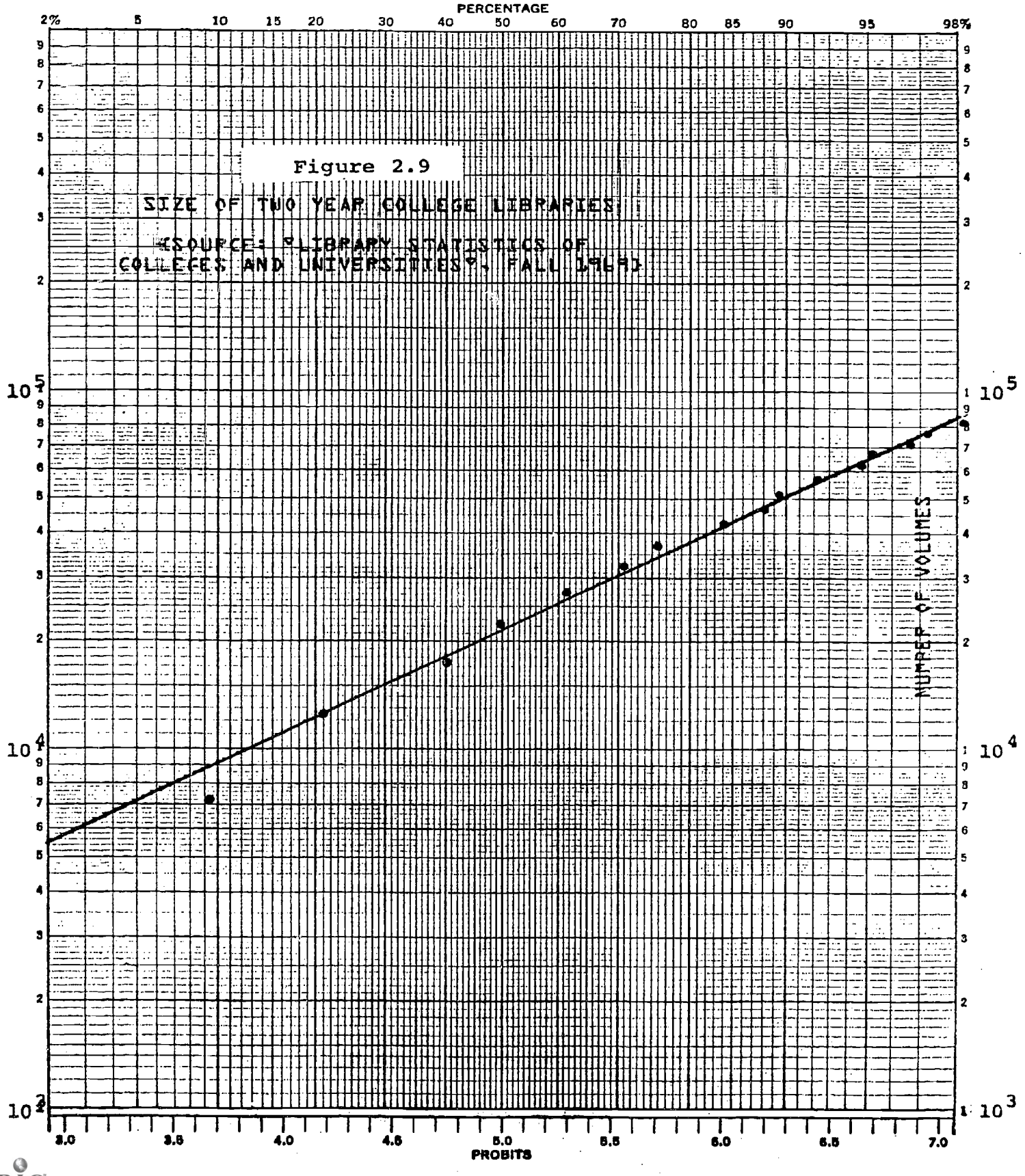


Figure 2.9

SIZE OF TWO YEAR COLLEGE LIBRARIES

(SOURCE: "LIBRARY STATISTICS OF COLLEGES AND UNIVERSITIES", FALL 1968)

Table 2.2

SIZE IN CHARACTERS

OF VARIOUS BIBLIOGRAPHIC UNITS

Unit	Level	Size	Log_{10} of Size
Title	1	28.15	1.44948
Table of Contents	2	505.	2.70329
Index	3	21293.	4.32710
Text of Book	4	819600.	5.91360
Two Year College Library	7	24528169200.	10.38966
University Library	8	620735534400.	11.79291

statistical F-test, as described for instance in Ref. (4), it is easily shown that the data does not contradict the hypothesis that a = 1 in eq. (2.1) at the 5% confidence level; this means that the least squares best fitting line for the points in Figure 2.8 does not differ significantly from that line which is constrained to pass through the origin of the coordinate system and also minimizes the sum of the squares of the deviations from the data points. This latter line corresponds to a relation of the form

$$s(n) = 10^{bn} \qquad (2.2)$$

relating the mean size of bibliographic units to their level. Carrying out the least squares minimization for a function of this form on the logarithms of the data leads to the line drawn in Figure 2.8 which corresponds to the equation

$$s(n) = 10^{1.47247n} = (29.68)^n. \qquad (2.3)$$

The constant 29.58 is an estimate of the fundamental constant determining the level structure of the bibliographic units considered above. More extensive data will no doubt result in the modification of this value, but it can be said with certainty that the fundamental constant is approximately 30, and perhaps may be identifiable with $(2e)^2 = 29.54...$, where $e = 2.718...$ is the mathematical constant denoting the base of the natural logarithm system.

This is our first main result:

> The average size of the bibliographic units title, table of contents, index, monograph, two year college library, and university library are powers of a fixed constant K whose value is nearly $(2e)^2$.

If it could be shown that the mean size of an encyclopedia is approximately K^5 and that of a personal (or perhaps a library reference sublibrary) is about K^6, then it could be asserted that the natural bibliographic units are equispaced when measured by the logarithm of their size; the current state of knowledge only permits us to assert that this is so for levels 1 through 4 and also for the separation of levels 7 and 8.

The previous argument suggests that the notion of level be introduced more generally. Therefore define the level of a given information base to be the integer closest to the logarithm of its size (the latter measured as usual in characters) to the base K; moreover, if a system of level K provides access to an information store of level n, then define the order of access provided by the access system as (n-k). Thus an index provides access of order 1 (=4-3) to the monograph it accompanies, and similarly the table of contents and title provide access of order 2 and 3 respectively to the book with which they are associated. We will later find that a library card catalog provides access of order 2 to the library archive but unfortunately it occupies a physical volume which could provide order 1 access to the collection.

Thus far we have principally concerned ourselves with the mean value of the various size distributions that have been examined, and have thereby shown that there is a simple and uniform relationship which connects the smallest of the natural units to the largest. We must now take up the question of the extent to which the mean characterizes the distributions that occur. The figures displaying the various distributions at the same time provide powerful evidence that all of the

distributions are lognormal. The elementary form of the lognormal function, which is what occurs here, depends on two parameters--the lognormal mean and the lognormal standard deviation; if these parameters are known, then the usual mean value of the distribution can be determined and conversely, if the lognormal standard deviation and the usual mean are known, the lognormal mean and hence the lognormal function itself are completely determined (cp. Chapter III). From this it follows that if the lognormal standard deviation of the various distributions of interest are all essentially equal, then the associated lognormal functions are in reality determined by the mean value, that is, by the level, of the distribution. We shall show that this is indeed the case. Table 2.3 lists the lognormal standard deviation of the six distributions that have been described thus far.

Table 2.3

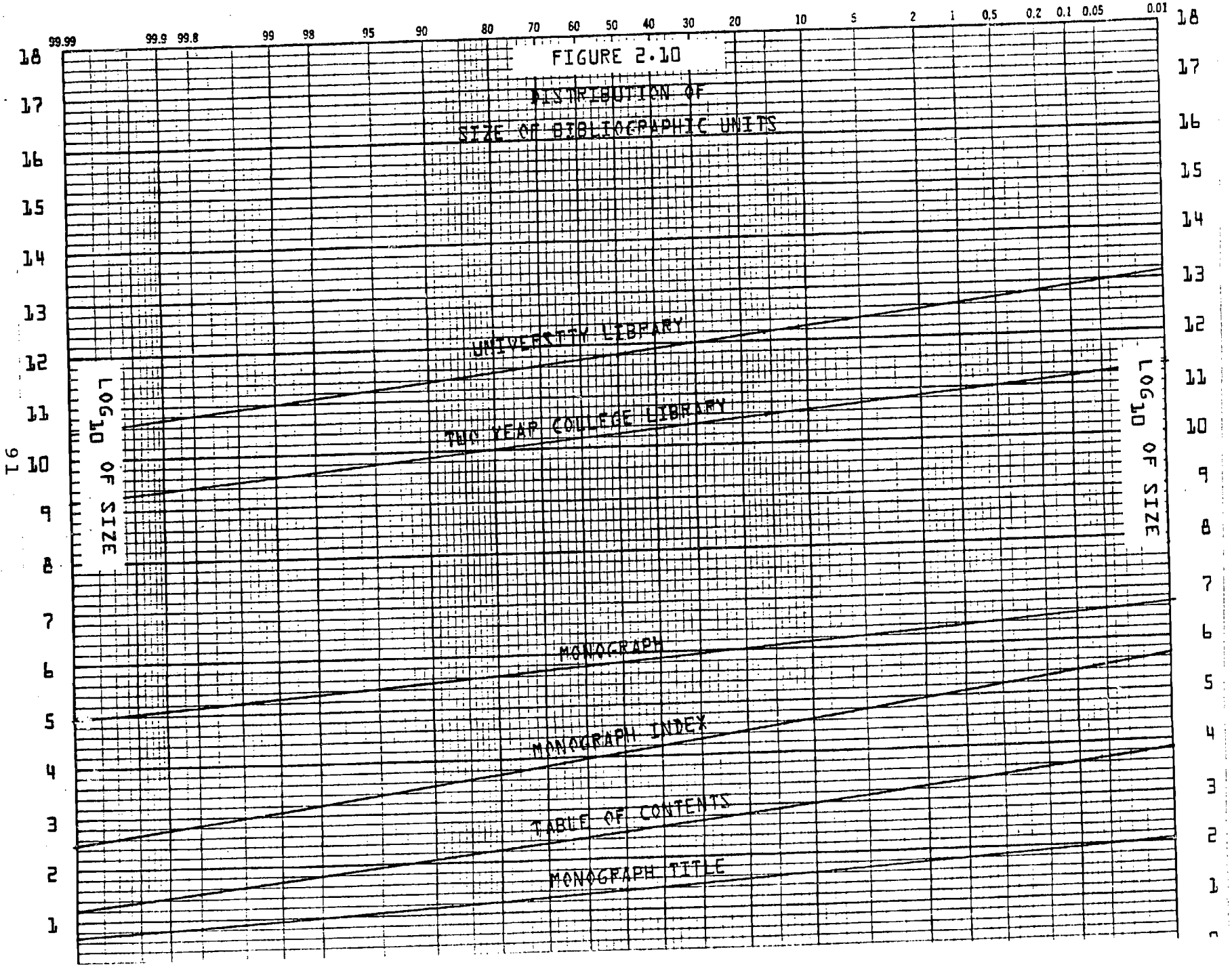
LOGNORMAL STANDARD DEVIATION

Unit	Level	Lognormal S.D.
Title	1	0.19
Table of Contents	2	0.30
Index	3	0.44
Monograph	4	0.23
Two Year College Library	7	0.29
University Library	8	0.36

There is evidently not much variation of the lognormal standard deviation as the level changes from a distribution whose typical size is about 30 characters to one whose typical size is about 600 billion characters and in particular what variation there is does not seem to have a trend. Based on the data contained in Table 2.3 we assert that the lognormal standard deviation is essentially constant throughout the entire range of bibliographic interest, and consequently the distributions of size of the various bibliographic units are determined by the level of the unit.

The lognormal standard deviation corresponds to the slope of the line defining the lognormal function for figures drawn on lognormal probability graph paper such as Figures 2.1-2.7 and 2.9 are. The underlined statement in the previous paragraph is the analytical version of the geometrical assertion that the lines representing all of the distributions are nearly parallel. We show to what extent this is so in Figure 2.10 which displays the distributions for all six levels; the variation of slope is indeed not great. The mean value of the standard deviations listed in Table 2.3 is 0.30, which may be conveniently adopted as an estimate of the level-independent lognormal standard deviation.

The assertion that the distribution of a variable x is lognormal is equivalent to stating that the distribution of log x is the normal (Gaussian) distribution. Here 'log' denotes the logarithm with respect to any conveniently chosen base. The graph of a normal distribution is the well known 'bell-shaped curve'. The level-structured lognormal distribution model of access systems described above can be equivalently viewed as a level-structured model for the logarithm of the size of bibliographic units such that the mean of the logarithms of the various levels are equally spaced and the associated distributions are normal, as shown in Figure 2.11 for levels 1-3. From that figure one also sees that the several bell curves have little overlap; this corresponds to the relative horizontality of the lines in the previous Figure 2.10 which is another way of stating that the lognormal standard deviation is a small number. The converse possibility, which fortunately does not occur, is that the lognormal standard deviation be relatively large with the consequence that the normal distributions like those illustrated in Figure 2.11 would possess a large degree of overlap with the overall appearance of gentle waves uniformly spread over a sea rather the sharply defined and separated peaks and valleys that Figure 2.11 so clearly exhibits. What this means is that the notion of level for bibliographic units makes sense; almost all units of some given type are of a size that is closer to the level of that type than to any other level. For instance, from Figure 2.10 we can read that fewer than 0.05% (sic!) of the Tables of Contents are so large as to lie (in logarithmic measure) closer to level 3 (Indexes) than to level 2 (Tables of Contents); similarly, fewer than 0.2% of the Two Year College Libraries are so large that they lie closer (in logarithmic measure) to the average size of a university library than to the average size of a two-year college library.



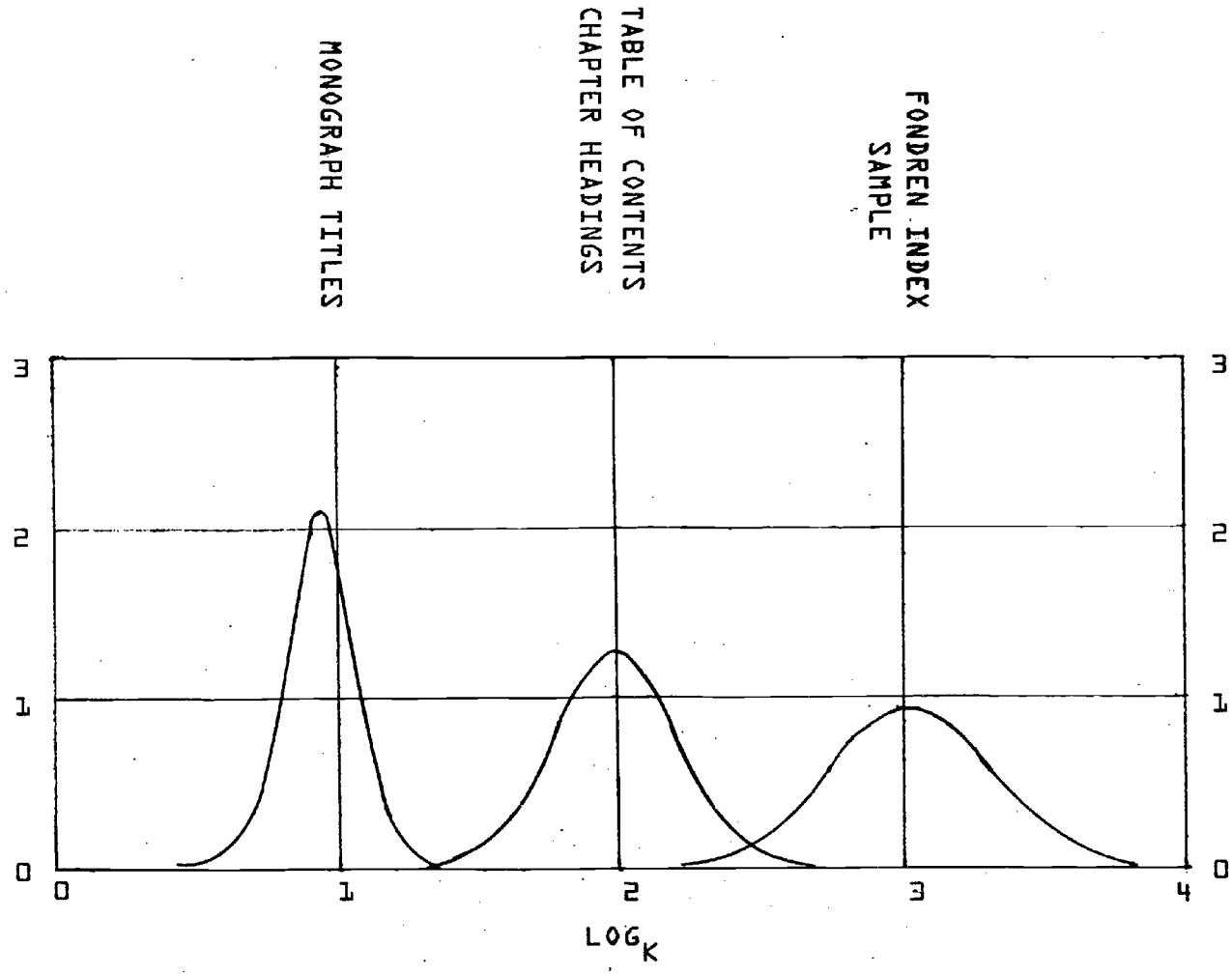


FIGURE 2.11

SOME ACCESS DISTRIBUTIONS IN

LOGARITHMIC VARIABLES

These observations suggest that the notion of boundary separating two adjacent levels should be introduced as that size corresponding to half integer values of the level. More precisely, with level n and size s(n) related as in eq. (2.2), we say that the size s(n+1/2) is the boundary size between s(n) and s(n+1), and that (n+1/2) is the boundary between level n and level (n+1).

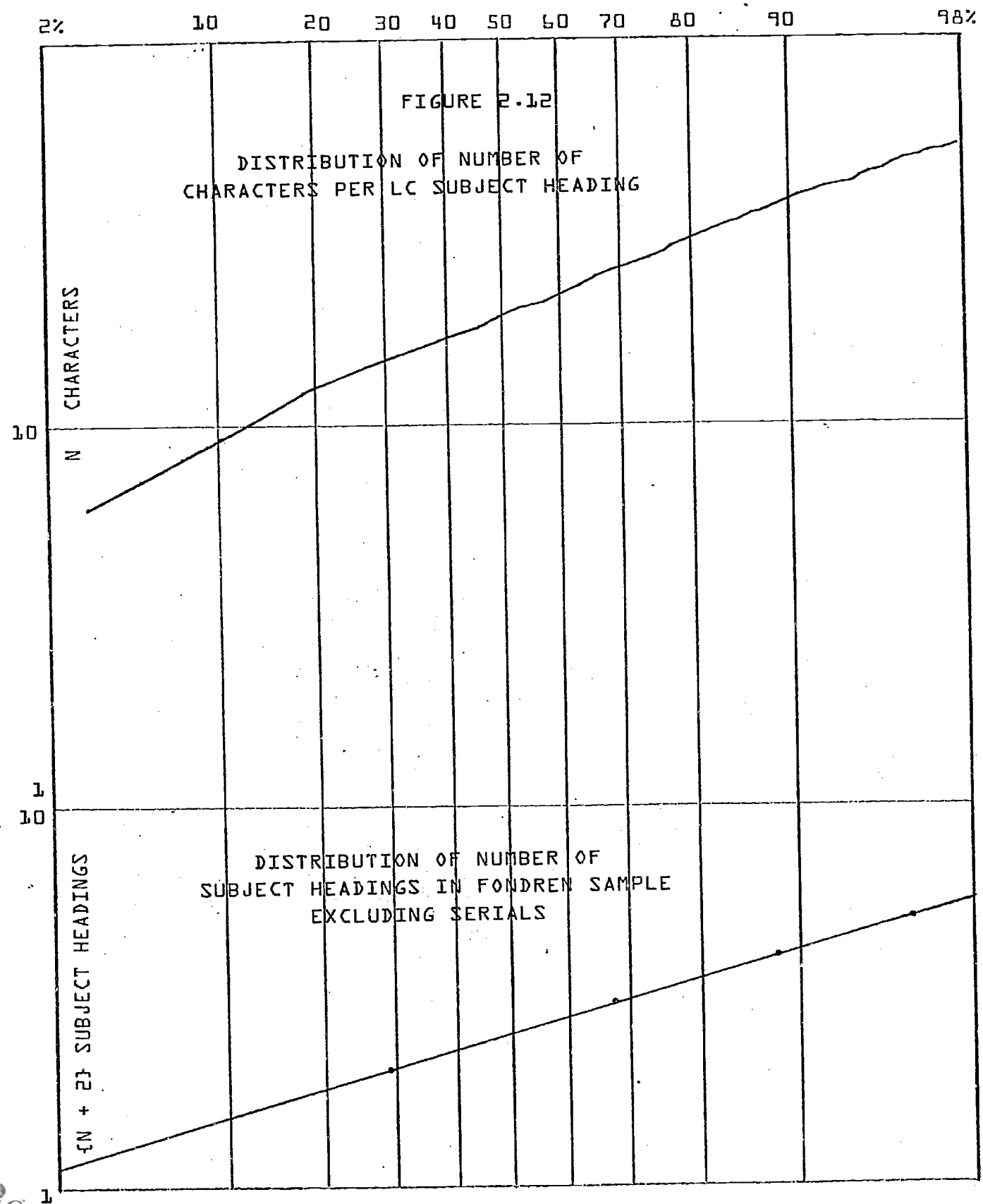
With this notion in hand it becomes possible to analyze a bibliographic item in order to determine if its size coincides reasonably with its 'proper' size, i.e., with the level of that type of bibliographic unit; from its size s compute $\log_K s$ and compare this number with the appropriate bibliographic unit level n to see whether $\log_K s$ lies within $\pm$ 1/2 of n; if it does not, then we may assert that the item of size s is either too large or too small. There will of course be specific exceptional instance for which the size of the unit is indeed 'proper' although not consistent with the statistically typical behavior for items of its bibliographic type, but the designer or evaluator of information access systems and/or information bearing data bases should, we think, warily approach the question of the size of a system from this point of view.

The access model presented in this chapter is not restricted to the book and its subsystems and supersystems. There is considerable evidence that it reflects universal properties of information stored in written English form, and, in a slightly generalized version, may be still more broadly applicable to the analysis and modeling of other types of information systems such as those associated with the modalities of sensory perception. These wide ranging and difficult issues cannot be examined here in a serious way; moreover, we do not yet have sufficient data upon which a definitive report can be based. Some of the intriguing vignettes that are most directly related to information presented in forms analogous to, if superficially distinct from, the book information system hierarchy explored above may nevertheless prove helpful for the reader.

First consider the size relationships of component units of the serial publication archive. We have studied the mathematics journal subarchive with the following results. For 7445 papers reviewed in volume 36 of Mathematical Reviews (published in 1968), the mean length of an abstracted paper is 13.8 'pages'; here 'page' refers to the myriad distinct page sizes and formats used by the 800-odd distinct journals reviewed by Mathematical Reviews. Bearing this in mind, and

noting that we have not attempted to directly determine the mean number of characters per page of mathematics text nor the effect of the numerous special symbols which extend the normal type font, use of our previous estimate of 2400 characters per page of text yields the estimate of 33,120 characters per mathematics paper; hence such a paper is of level 3. The mean length of an abstract in Mathematical Reviews is easily estimated to be about 1081 characters. Therefore the size of the average mathematics paper is 30.6 times the size of the average abstract. Division of the estimated size of an abstract by K = 29.54 gives 36.59 characters, which is about the size of the average mathematics journal paper title and is of course quite close to the level 1 mean of 29.54 characters. We conclude that journal papers in mathematics are structured in a manner which is consistent with the general model proposed for books.

Next consider a more complex example which refers directly to the access problem. It is usual to find so-called "subject headings" at the foot of library catalog cards which are intended to provide cross reference access to subject areas other than those associated with the class number of the item corresponding to the catalog card. There are nearly 93,000 subject headings in the Library of Congress Subject Headings, seventh edition (1966). A uniform 1/66 sample drawn from an alphabetized list of these headings shows that the mean number of characters per subject heading is 22.3, which is not remarkably close to K = 29.54. However, the distribution of subject headings per catalog card as determined from an analysis of the Fondren Sample has a mean of 1.2 headings per card; if the distribution of subject headings per card is independent of the distribution of characters per subject heading, then the mean number of subject heading characters per catalog card, including the associated ordinals and interword space characters, will be the product of the means of the component distributions, which is 29.16. Hence the collection of subject headings per card provides about the same level of discrimination above the one-letter Library of Congress class in the mean that is provided by the title. Considering the distributions of characters per subject heading and subject headings per card leads to the lognormal functions shown in Figure 2.12; we conclude that the subject heading access mechanism is consistent with the level structured model and it belongs to level 1.


2%
10
20
30
40
50
60
70
80
90
98%
FIGURE 2.12
DISTRIBUTION OF NUMBER OF
CHARACTERS PER LC SUBJECT HEADING
N CHARACTERS
10
1
10
DISTRIBUTION OF NUMBER OF
SUBJECT HEADINGS IN FONDREN SAMPLE
EXCLUDING SERIALS
fN + 2) SUBJECT HEADINGS
1

The phenomenon that the mean value of the size of adjacent access levels are in the ratio of about 30 to 1 is not confined to access systems associated with written natural language archives. Consider ALTEXT, a contemporary text-processing higher level (macro expander) computer language [5]. Such a language consists of computer instructions which have two parts: a generic instruction such as the GOTO of FORTRAN which specifies the general function of the instruction, and certain other more particular components which contain the details of data location and transfers of control. The implementation of a higher level computer language instruction consists of a sequence of one or more "machine language" or "assembly language" instructions; the advantage of the higher level language is that it frees the programmer from the burden of keeping track of numerous housekeeping details concerning the location and manipulation of the data at the cost of lower (local) efficiencies of execution. This is another way of stating that the higher level language instructions act as an access system for the sequences of assembly language instructions that are their implementation.

With this preamble in mind, one can examine the number of assembly language instructions required to implement each of the distinct generic higher level language instructions. For the generic instructions of ALTEXT, the mean number of assembly language instructions per ALTEXT "macro" is 30.32 (including implementation of the "ALTEXT macro" which provides the interface with the operating system of the implementing computer) for implementation on the IBM 360/30 computer. Figure 2.13 confirms in a rather startling way that the distribution of implementation size is lognormal; hence we conjecture that the level structured access model will probably find significant application in the design of computer languages.

That the structure of many types of linguistic units is lognormal has long been known and abundantly verified. The lognormality of word length statistics was discovered at least as early as 1887 by Mendenhall [6] and was subsequently studied, along with sentence length distributions, inter alia, by Yule [7], Williams [8], and Herdan [9]. Yule computed the sentence length distributions for a number of samples of written English and although he did not notice their lognormality himself, Williams did test this hypothesis on Yule's data and on more he gathered himself. More extensive data has been collected by Kucera and Francis [10] but care must be exercized to insure that it is partitioned into homogeneous subject and/or author classes before attempting to study the lognormality of the statistics; the problem

of describing the structure of inhomogeneous data, which amounts to studying how distinct lognormal distributions combine, is relatively complex. Moreover, much of the Kucera and Francis data refers to printed materials that are unlikely to form an active part of an archival library collection; it is heavily weighted with fiction and press coverage.

Herdan [9] analyzed 80,000 words of telephone conversations collected by French, Carter and Koenig of the Bell Telephone Laboratories and concluded that (phonetic) word length is lognormally distributed. An indication that the parameters of these linguistic distributions are relatively insensitive to variations in language vocabulary and to whether the written or spoken form is used is provided by Figure 2.14 which shows nearly parallel lines representing the Herdan telephone conversations and Mendenhall's analysis of 1000 words from Shakespeare's works (as represented by Williams).

These examples and others too numerous to report here prompt us to speculate that the occurrence of the lognormal distribution is fundamental to all human information processing activities. In this regard we distinguish two types of activities: those that process direct sensory impressions that are received through the sensory organs, and those that process coded information such as is represented by linguistic codes. In the latter instance the directly perceived data arrives via the sensory organs but the essential content is unrelated to the particular code used for its transmission. Although there may be important differences between the internal mechanisms that process these two types of information, there are at least two characteristics that the two types of input information share: the quantity of information that passes through the processing system is very large and the system must be capable of responding to inputs whose size vary greatly. The first condition requires that the information processing system be able to compress (with information loss) the vast amount of data passing through it so as to be enabled to retain for future use a much smaller but characteristic subset of it; in other words, the processing system must function as an access system to the information passing through it. The second condition suggests that some functional transformation must be applied to the input sensory information in order to reduce its extended range to a smaller one more conveniently handled by the neural network; for example, there has long been evidence (which is reflected by the 'decibel' scale of measurement) that the subjective response to the stimulus provided the ear by acoustic energy varies as the logarithm of the input energy.

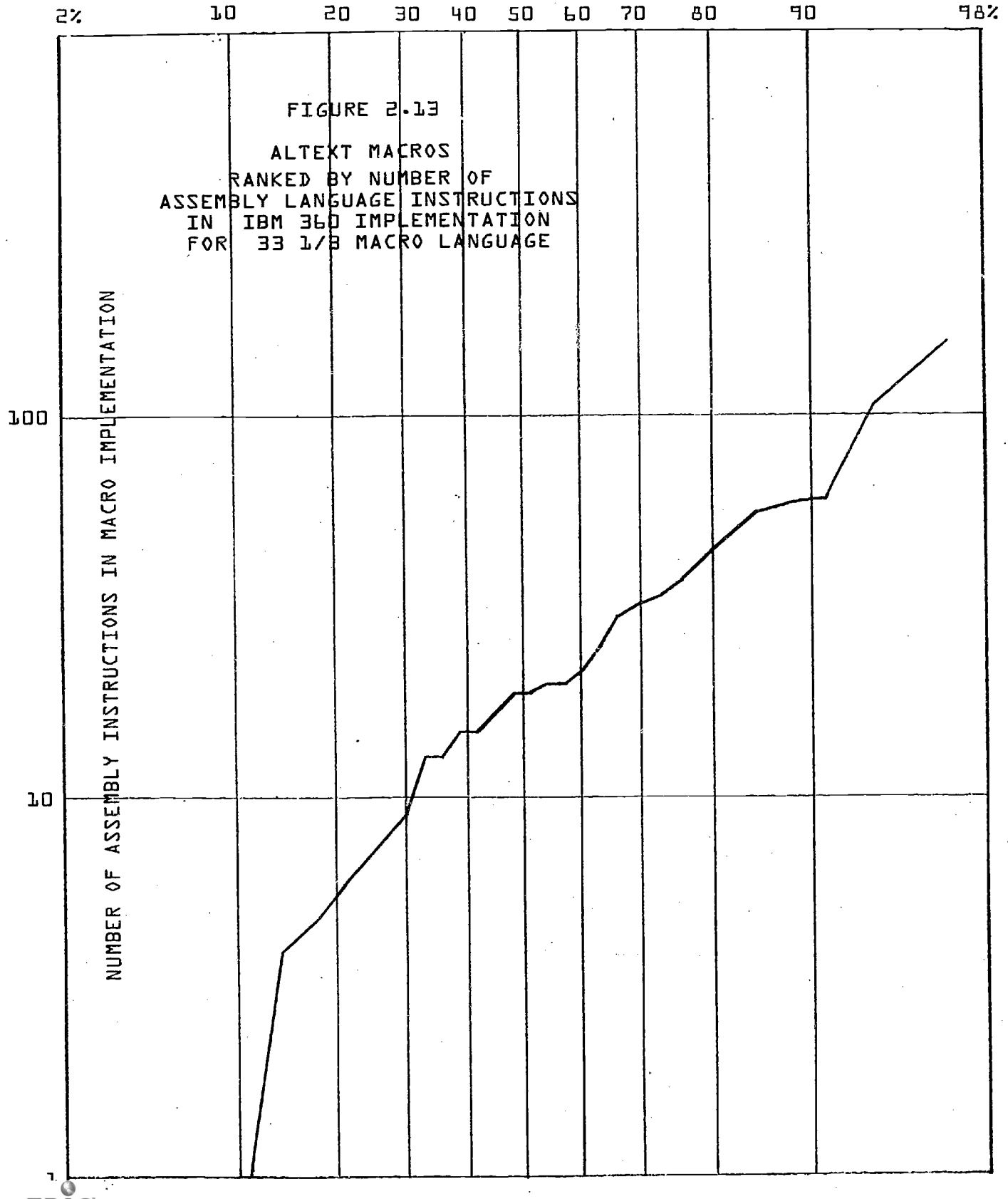
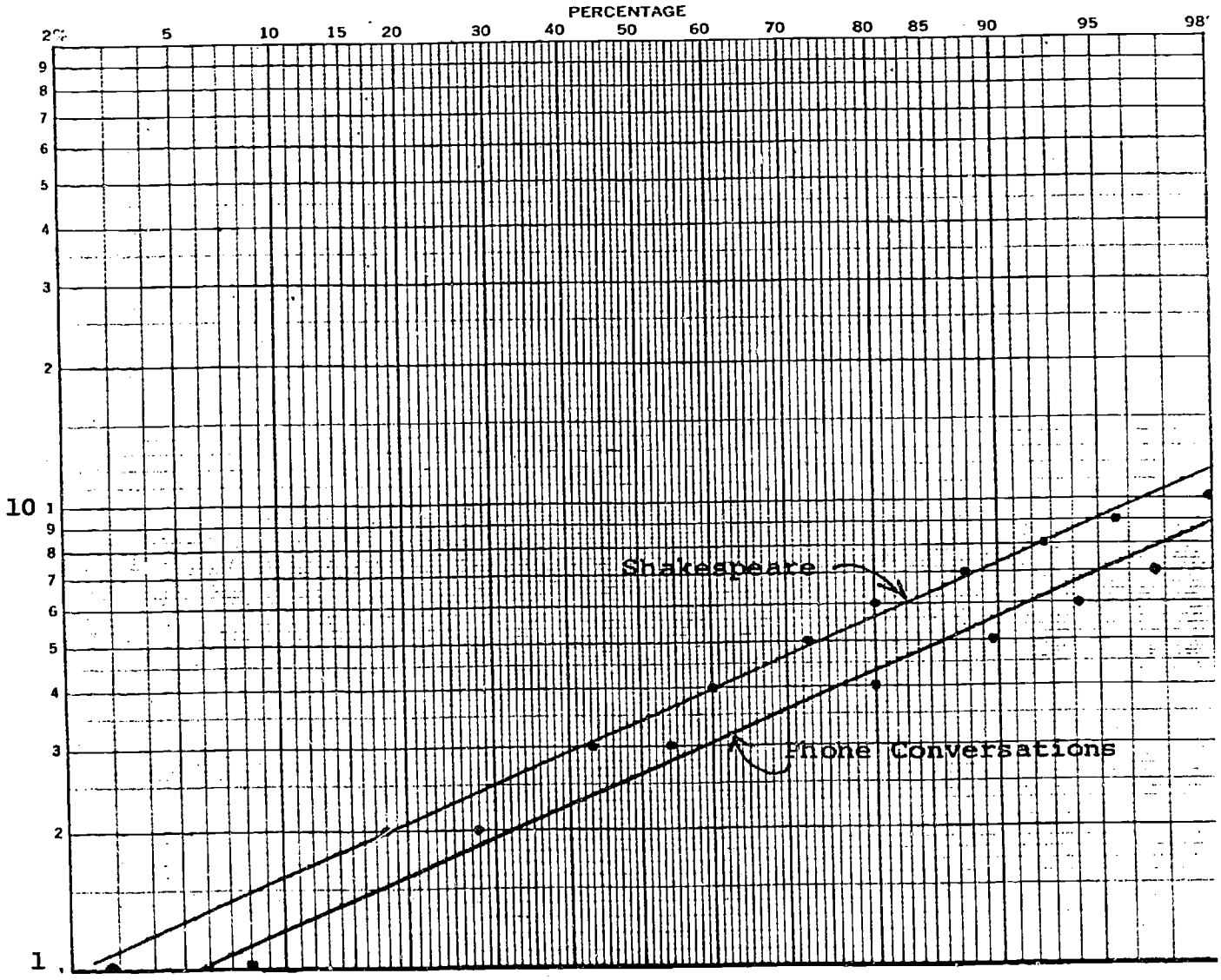


FIGURE 2.13

ALTEXT MACROS
RANKED BY NUMBER OF
ASSEMBLY LANGUAGE INSTRUCTIONS
IN IBM 360 IMPLEMENTATION
FOR 33 1/3 MACRO LANGUAGE

Figure 2.14

Word Length Distributions
(in characters)



Generally, there are three reasons for making a scale transformation in analyzing data (e.g., see Tukey [11]):

1. To linearize the relation between two variables.

2. To normalize the underlying probability distribution.

3. To stabilize the variance.

Although in most applications any one of these results would provide sufficient reason for introducing a particular transformation, it is not uncommon to encounter situations where the transformation is originally introduced for one reason and subsequent analysis shows one or both of the remaining desiderata have also been achieved.

In this context it is illuminating to study the work of the nineteenth century experimental psychologist G. Fechner [12]. He made the important observation that the ability of the human to respond a stimulus is proportional to the mean level of the stimulus. That is, if an individual can just sense a difference of, say, one unit when the mean level of stimulation is 10 units, then he will also just be able to detect a difference of 2 units when the mean level is 20 units. This multiplicative property of the just noticeable difference led him to introduce the logarithm function in order to stabilize the variance, i.e., make it constant throughout the range of perception. He then conjectured that the function relating subjective response to the transformed variable--the logarithm of the stimulus--is a linear function, thus arriving at the celebrated (and once again hotly debated) 'Law' of Weber and Fechner. The reader will observe that the logarithm of the size of bibliographic units stabilizes the variance of the distributions of these units throughout the entire range of 'bibliographic perception'. This certainly makes it tempting to inquire whether the Weber-Fechner 'Law" might not be merely an approximation to some more accurate description of the underlying functional transformation governing sensory perception. This question has received considerable attention in recent years and notable contributions have been made, principally by Stevens (e.g., [13]), who has generalized the logarithmic Weber-Fechner transformation so that response is some power of stimulus; that this change actually constitutes a generalization becomes clear when it is noted that the integral of $1/x$ is $\log x$ whereas the integral of any other power of x is again a power of x; in this sense the logarithm is the limit of power

functions (see Dolby [14]). The relationship between linguistic and hence bibliographic units and these psychophysical questions has been remarked by several workers, most notably perhaps by Fairthorne [15]; Zipf's 'Law' [16] in its integrated form is just the Weber-Fechner logarithmic relation, and Mandelbrot's [17] generalization of Zipf's function corresponds--indeed, it is identical to--Steven's power function. These questions will be taken up from a more mathematical standpoint in the next chapter with the intent of showing how they can be derived, following an argument essentially due to Mandelbrot, from elementary considerations from information theory, and, of more importance for our purposes, that a slight extension of this argument generalizes the Weber-Fechner-Zipf-Stevens-Mandelbrot functions to the lognormal distribution. For as the extensive bibliographic data assembled in the earlier parts of this chapter show, it is the lognormal function that in fact describes reality.

References

1. Dolby, J. L., V. Forsyth, and H. L. Resnikoff, Computerized Library Catalogs: Their Growth, Cost and Utility, M.I.T. Press, Cambridge, 1969.

2. Dolby, J. L. and W. J. Jones, "The Measurement of Composition Practice", in Advances in Computer Typesettting, Institute of Printing, London, 1966.

3. Price, Bronson, Library Statistics of Colleges and Universities, Fall 1969, Data for Individual Institutions, U. S. Offfice of Education, Washington, D. C., 1970.

4. Youden, W. J., Statistical Methods for Chemists, John Wiley & Sons, New York, 1951.

5. Dolby, J. L., W. E. Houchin, Roger Stark, and H. L. Resnikoff, Non-Numeric Programming Language Studies: ALTEXT II, Final Report to U.S.A.F. Office of Scientific Research, R & D Consultants Co., Los Altos, California, 1970.

6. Mendenhall, T. C.,"The Characteristic Curves of Composition", Science, 9,(214,supplement) (1887), 237-49

7. Yule, G. U., "On Sentence-Length as a Statistical Characteristic of Style in Prose", Biometrika 30(1939), 363-84

8. Williams, C. B., "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style", Biometrika, 31(1940), 356-61.

9. Herdan, G., "The Relation between the Dictionary Distribution and the Occurrence Distribution of Word Length and its Importance for the Study of Quantitative Linguistics", Biometrika, 45 (1958), 222-8.

10. Kucera, Henry and W. N. Francis, Computational Analysis of Present-Day American English, Brown University Press, Providence, 1967.

11. Tukey, J. W., "On the Comparative Anatomy of Transformations", Annals of Mathematical Statistics, 28(1957), 602-32.

12. Fechner, G. T., Elemente der Psychophysik, 1860.

13. Stevens, S. S., "Neural Events and the Psychophysical Law", Science, 170(1970), 1043-50.

14. Dolby, J. L., "A Quick Method for Choosing a Transformation", Technometrics, 5(1963), 317-25.

15. Fairthorne, R. A., "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction", Journal of Documentation, 25(1969), 319-43.

16. Zipf, G. K., Psycho-Biology of Language, Houghton Mifflin, 1935.

17. Mandelbrot, B., "An Information Theory of the Statistical Structure of Language", Proceedings of the Symposium on Applications of Communication Theory, London, September 1952, Butterworth, 1953, 486-500.

MATHEMATICS OF

INFORMATION DISTRIBUTIONS

This chapter is devoted to the mathematical study of some of the distributions that arise naturally in the study of information systems. It will be necessarily more demanding of the reader's mathematical knowledge than the remainder of the book and has therefore been written in a manner which we hope will permit the reader to pass immediately to Chapter IV without loss of continuity. We believe, however, that the significance and implications of the level structured model of access systems presented in Chapter II cannot be fully understood unless the relationship of that model to other competing models, extant and potential, is made clear. Moreover, the most powerful theoretical arguments for the appearance of the lognormal distribution in the model structure comes from information theory and its mathematical apparatus, while those for the multiplicative level structure come from a certain extremal problem in calculus, so there is really no way to avoid these technical considerations.

Chapter II presented empirical evidence which show that the access systems normally associated with books and collections of books form a multiplicatively structured system of levels wherein the distribution of size of information structures belonging to any one level is lognormal, and the spacing between adjacent levels, that is, the ratio K of the mean size of one level to the mean size of the next smallest level, is independent of the choice of level and approximately equal to 30. In this chapter we will show that the lognormal distribution is the solution to a certain problem of maximization of information per unit cost and that the multiplicative level structure minimizes search time in a sense which will be more precisely defined below. It therefore remains to obtain the multiplicative spacing constant K from theoretical considerations.

We will show that an extension of the notion of search time minimization leads to a well defined value of K as a function of the file size and a ratio which measures the cost of system maintenance per unit system use (both costs measured by time). In the limiting case where

maintenance costs are 0 , it follows that the level spacing which maximizes efficiency, i.e., minimizes search time, is given by $K = 2.78... = e$. Further treatment of the more realistic case of non zero maintenance costs will require data measuring actual search and maintenance costs for the various access subsystems involved.

This chapter also studies the relationship of the Zipf-Mandelbrot (hereafter Z-M) distribution to the lognormal. The former is the best known and most widely used function for describing rank and frequency distributions of linguistic, psychophysical, and socio-economic observations.

We will show that the rank distribution is Z-M if and only if the corresponding frequency distribution is Z-M. This coincidence is responsible for some confusion in interpreting observed distributions. More important to this study is the relationship of the Z-M distribution to the lognormal. As we show, the Z-M is a limiting case of the lognormal and can also be interpreted as a first approximation to it. This implies a relation between the corresponding rank distributions which we describe. Our discussion elucidates certain heretofore unexplained systematic departures from the Z-M rank distribution exhibited by data drawn from information bases.

The best known mathematical function which describes naturally occurring information distributions is the power function,

$$x = cr^{-s} \quad , \quad s > 0 \ . \tag{3.1}$$

These applications were discovered by Estoup <7>, Bradford <2>, and others, and rediscovered by Zipf <28> who popularized the observation that the ranked frequency distribution of word tokens in natural text corpora is essentially of the form (3.1) where r denotes the rank, x the frequency of occurrence of the word of rank r , and c and s are constants chosen to provide the best possible agreement with data. Zipf concluded that $s = 1$ for English; Figure 3.1, taken from Zipf <28>, exhibits such a distribution.

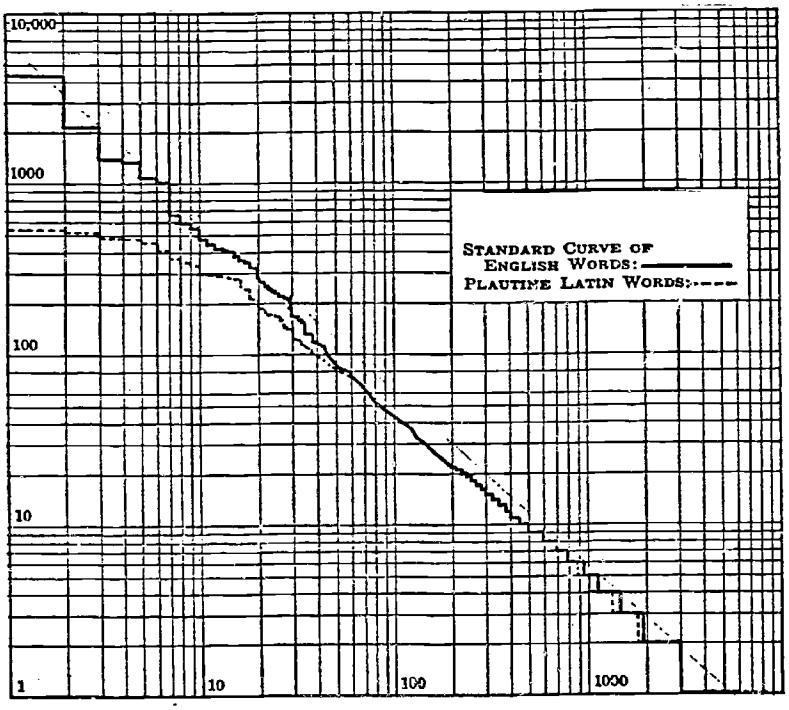
The total number of word tokens N is evidently given by

$$N = c \sum_{r=1}^{\infty} r^{-s} \quad ; \tag{3.2}$$

this series converges only if $s > 1$, which means that Zipf's original choice $s = 1$ cannot be strictly correct. If the series does converge, then it represents the well known Riemann Zeta function <25>,

Figure 3.1

Word Frequency Distribution - English and Latin Words



$$\zeta(s) = \sum_{r=1}^{\infty} r^{-s} \quad , \quad s > 1 \ . \tag{3.3}$$

The constant c in (3.1) must therefore be $N/\zeta(s)$. It is not easy to calculate $\zeta(s)$ for s near 1 , but a reasonable approximation is afforded by approximating the infinite sum in (3.3) by an integral:

$$\int_1^{\infty} r^{-s} dr = 1/(s-1) \ . \tag{3.4}$$

Then, with increasing validity as s approaches 1 , the distribution (3.1) can be written

$$x = (s-1)Nr^{-s} \ ; \tag{3.5}$$

the exact form is

$$x = Nr^{-s}/\zeta(s) \ . \tag{3.6}$$

Many different types of observational data drawn from human related activities seem to admit description by this function and therefore the names of many famous men are attached to it. In librarianship it is Bradford, in linguistics Zipf, in economics Pareto <17>, in what might be termed "sociological mathematics", Lotka <14>. who showed that the distribution of productivity of researchers appears to follow the power law (cp. Price <20>), and in psychophysics Stevens <24>.

A striking example of a power law description is displayed in Figure 3.2 which exhibits the size of cities in the United States as a function of their rank for several different times. Drawn as it is on log-log graph paper, each power function is represented by a line. One notes that all of the lines are essentially parallel, which means that the exponent s does not depend on which line is considered: s is independent of time. Moreover, the lines appear to march across the graph paper uniformly with increasing time. Since the log-log graph of eq(3.1) is just the ordinary graph of the equation

$$\log x = \log c - s \log r \ , \tag{3.7}$$

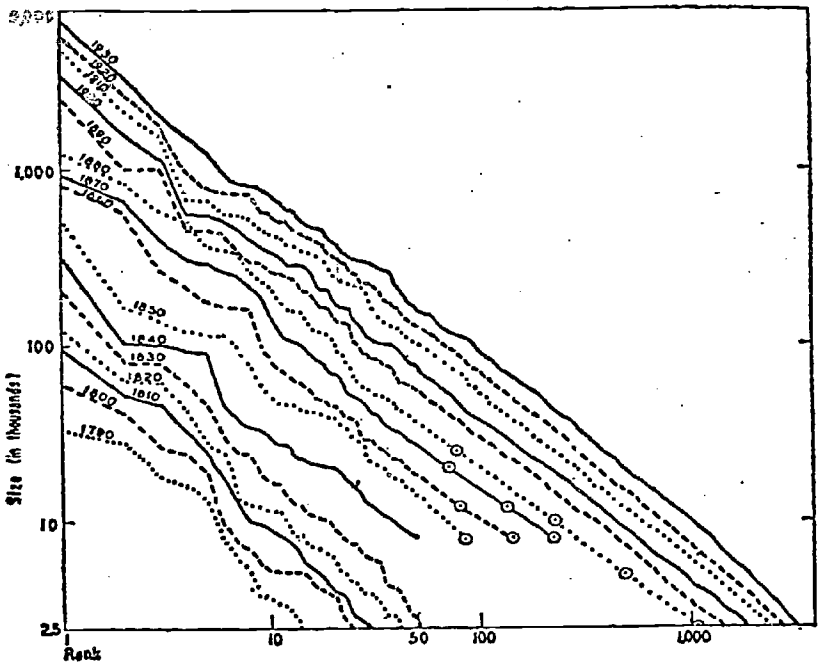
it immediately follows from the steady parallel motion of the line described by (3.7) that the intercept $\log c$ must vary linearly with time. That is, there are constants a , b such that $\log c = at + b$; then, from $c = N/\zeta(s)$,

$$N = N(t) = \zeta(s)e^{at+b} \ ;$$

Figure 3.2

Distribution of Community Populations

United States of America



UNITED STATES, 1790–1930

Communities of 2500 or more inhabitants ranked in the decreasing order of population size. It should be noted that the distribution at any given date shows size decreasing uniformly with rank; as cities become more numerous and all of them increase in size, the distribution pattern is preserved, the curve moving parallel to itself at a constant rate. From George K. Zipf, *Human Behavior and the Principle of Least Effort* (Cambridge, Mass., Addison-Wesley Publishing Company, Inc., 1949), p. 420, Fig. 10-2.

if we put $N(0) = \zeta(s)e^{b}$, this takes the convenient form

$$N(t) = N(0)e^{at} . \qquad (3.8)$$

$N(0)$ is the total population at time $t = 0$, and $N(t)$ is the population at time t ; (3.8) shows that population grows exponentially with time, while (3.6) states that at any one time, population is distributed amongst cities according to the power function.

The interplay between the distribution at one time and the time variation of the quantity of interest will be important for our consideration of the dynamics of information distributions in what follows.

The power function is often used to represent two quite different types of distributions with, we think, confusing consequences. As introduced above, it describes the rank-frequency distribution of a variable. Suppose that $f(x)$ is the frequency distribution of some variable x . In any finite sample of x there will be a largest occurring observation, say x_1 , a next largest x_2 , and so forth. If some observation x occurs more than once, consider the various occurrences as distinct and label them consecutively. The resulting distribution of pairs $(1,x_1)$, $(2,x_2)$, ... is just the rank-frequency distribution for the sample drawn from the population with frequency distribution f . The rank-frequency distribution can easily be expressed in terms of f . Let r denote a rank. Then

$$r(x) = \int_{x}^{\infty} f(t)dt ; \qquad (3.9)$$

indeed, that x ranked first, x_1 , is just that value x such that the occurrence of x is 1 , i.e.,

$$1 = \int_{x_1}^{\infty} f(t)dt ;$$

similarly, x_2 is defined by

$$2 = \int_{x_2}^{\infty} f(t)dt ,$$

and in general, x_r , the r^{th} ranked x , is defined by

$$r = \int_{x_r}^{\infty} f(t)dt \ ;$$

but this is just equation (3.9).

Eq(3.9) shows that it is more convenient to consider rank as a function of frequency x rather than frequency as a function of rank although the latter is of course more natural.

Differentiation of (3.9) yields

$$dr/dx = -f(x) \ ; \tag{3.10}$$

given the rank-frequency relation, this formula provides the underlying frequency distribution.

If the rank function is given by (3.1), then, solving for r as a function of x, we find

$$r = c'x^{-1/s} \tag{3.11}$$

with $c' = c^{1/s}$, so r is a power function of x. The corresponding frequency function is, by (3.10),

$$dr/dx = -(c'/s)x^{-(1+1/s)} \ , \tag{3.12}$$

again a power function. Conversely, if f is a power function, say

$$f(x) = kx^{-s} \ , \ k \text{ constant and } s \neq 1 \ , \tag{3.13a}$$

then

$$\int_{x}^{\infty} f(t)dt = (k/(s-1))x^{1-s} \ , \tag{3.13b}$$

again a power function. We have shown that the rank-frequency distribution is a power function if and only if the underlying frequency distribution also is. This has the unfortunate consequence that it is not always easy to determine whether the rank interpretation is the most appropriate one for data which appears to approximate a power function but for which a more refined approximation is desired.

There is some question whether the power function actually provides as good an approximation to observed data as at first sight appears to be the case. Numerous researchers have devoted great effort to improving agreement between data and representing function, and have been led in curious ways to a variety of

complicated and often unnatural functions. Belonogov <1> found that the distribution

$$x = \exp(-c(r-1)^k) - \exp(-cr^k) \qquad (3.14)$$

(we abbreviate $\exp a = e^a$) describes the rank-frequency distribution of printed commercial Russian, a form congenial neither for calculation nor analysis. Good <9> is led to

$$x = c(r - a)^{-s(1 + bx^{-1})} \qquad (3.15)$$

with b a small constant; this function has $x = (r - a)^{-s}$ as a first approximation (because b is small) but, although derived by an information theoretic argument which attempts to account for the effort required to incorporate words of large rank in the inventory, (3.15) is unfortunately a complicated expression and Good's accomodation of the presumed additional effort is in no way uniquely determined by any general principle.

Mandelbrot <15>,<16> presented a derivation of the power function distribution using information theoretic arguments and therewith slightly generalized the functional form (3.1). It will be important for us to understand the essence of his argument which we give in a modified form.

Consider an inventory $S_1, S_2, \ldots, S_n, \ldots$ of information states and let x_n be a measure of the size of S_n. For instance, the S_n might be the set of titles associated with a random sample of monographs and the corresponding x_n the number of characters in S_n, or, in the application to word token occurrence in samples of text corpora, S_n might stand for a specified word type and x_n for the frequency of occurrence of S_n in the sample. Denote the probability of occurrence of size x by $p(x)$; $p(x)$ therefore represents the probability of occurrence of any state of size x. Then, according to information theory (cp. e.g. Shannon <22>, Good <9>), the information associated with the system of states $\{S_n\}$ is proportional to

$$I = - \Sigma\, p(x) \log p(x) \,. \qquad (3.16)$$

It would seem to be desirable to secure an information system that maximizes information. Since p is a probability function, (3.16) is subject to the constraint $\Sigma\, p(x) = 1$, so, using the method of Lagrange multipliers, we find that (3.16) is extremal just when

$$- \Sigma\, p(x) \log p(x) + a \Sigma\, p(x)$$

is, for a constant a to be determined from the constraint. Differentiation with respect to $p(x)$ yields the extremal condition

$$\log p(x) = -1 + a ,$$

so $p(x)$ is constant. The condition $\Sigma\, p(x) = 1$ therefore insures that $p(x) = 1/N$ where N is the total number of states; consequently $a = 1 - \log N$.

This uniform distribution of state utilization is in fact not what is observed, for reasons which are easy to understand. There is clearly an inequity in the effort required to use different states; in general, the greater the measure of size x of state S , the correspondingly greater will be the 'effort' or 'cost' $c(x)$ required to utilize S . This inequality of effort will result in a corresponding inequality of the probability of usage. Expressed in a somewhat different way, a distribution of probabilities $p(x)$ which does not quite maximize I in (3.16) but which nevertheless 'costs' substantially less to use than the maximizing probability distribution will provide a more efficient return of information for the effort expended; it will be more 'cost effective'. This suggests that in place of I some measure of information per unit cost should be maximized.

Before turning to the determination of $p(x)$ as a function of $c(x)$, we must comment on Zunde and Dexter's <29> argument which shows that $p(x)$ is the normal distribution. They maximize I subject to the constraints

$$\Sigma\, p(x) = 1 , \Sigma\, xp(x) = \mu_1' , \Sigma\, x^2 p(x) = \mu_2'$$

where μ_k' is the k^{th} moment of p about 0 , and observe that if the higher moments μ_k' are known then the distribution p can be determined by maximizing I subject to the constraints

$$\Sigma\, x^k p(x) = \mu_k' , \quad k = 0,1,2,\ldots .$$

Indeed, in the former case the normal distribution results and in general there are constants c_k determined by the constraints such that $p(x)$ maximizes

$$- \Sigma\, p(x) \log p(x) + \sum_{k=0}^{\infty} c_k \Sigma\, x^k p(x) .$$

Differentiating this expression and setting the result equal to zero yields

$$\log p(x) = -1 + \sum_{k=0}^{\infty} c_k x^k \quad ,$$

so $\log p(x)$ is determined as a power series in x. It is not an essential restriction to suppose that $p(x) > 0$ if $x > 0$ so $\log p(x)$ exists. On the other hand, it is only a minor technical mathematical restriction to suppose that $\log p(x)$ has a power series expansion. It can therefore be concluded that the procedure of Zunde and Dexter does not distinguish or select a family of distributions related to the structure of information systems; it merely provides an interesting procedure for fitting a power series to observed data in a way which insures a posteriori that the function maximizes information for the particular observations in hand.

We may examine this procedure from another vantage point. In order to determine the distribution or parametized family of distributions which maximizes information, it is clear that as few restrictions as possible should be imposed. The removal of restrictions increases the population of distributions amongst which that one corresponding to maximal information is to be found, and thereby makes possible a larger value of the maximum information (relative, of course, to the larger set of functions). The constraint $\Sigma\, p(x) = 1$ is inherent in the definition of probability and does not imply special knowledge about a particular information distribution; therefore this single condition is the only non removable constraint, and has been shown above, to it corresponds the uniform distribution of p, which does not agree with observation. Now suppose that we know the value of certain sample moments μ_k'. The form of the universal distribution which maximizes information cannot depend on our admission of knowledge of these quantities, for by claiming to know less, i.e., by ignoring the value of but one moment, the argument of Zunde and Dexter implies that the form of the probability function is restricted, that is, we know more about it. In the case where no observations whatever are made, we find the exact form of the distribution -- the uniform one. This conclusion is typical of curve fitting methods, which rely on observed data to determine the form of the function, which must of course be restricted and ultimately quite simple in the absence of sufficient data to define a more complex function.

The structure of optimal information distributions which maximize a function of information and some other naturally occurring quantity (such as cost) should relate the maximizing distribution p to the other quantity, and should depend on that quantity and their mutual relation for its complexity rather than on observational measurements.

Let us return to the investigation of information per unit cost. There are two such measures which come readily to mind. Following Good's presentation <9> of Mandelbrot's argument, first consider the average information per average cost, that is,

$$H^* = - \Sigma\ p(x) \log\ p(x)\ /\ \Sigma\ p(x)c(x) \quad , \qquad (3.17)$$

subject to the constraints $\Sigma\ p(x) = 1$ and $\Sigma\ c(x) = C$ (= total cost). Maximization of (3.17) subject to these constraints is equivalent to maximization of

$$- \Sigma\ p(x) \log\ p(x) \quad + \quad (1+a_1)\Sigma\ p(x)$$

$$- a_2 \Sigma\ p(x)c(x)$$

with constants a_1 , a_2 determined by the constraints. Differentiation with respect to $p(x)$ leads to the condition

$$\log\ p(x) = a_1 - a_2 c(x) \ . \qquad (3.18)$$

It remains, therefore, to fix the cost function and then determine the a_i from the auxiliary constraints. Mandelbrot argued that $c(x)$ is proportional to $\log(x-a)$ for some (small) constant a ; we may absorb the factor of proportionality in a_1 and a_2 and then write

$$c(x) = \log(x-a) \ ; \qquad (3.19)$$

insertion of this function in (3.18) yields

$$p(x) = e^{a_1}(x-a)^{-a_2} \quad ,$$

the power function relative to the displaced origin $-a$.

We must diverge momentarily from our main theme to justify (3.19) as an approximation to the cost of utilizing a state of size x . Information states may be considered, for our purposes, to consist of an ordered sequence of symbols -- mostly alphabetic characters and symbols of punctuation -- selected from a finite symbol inventory. It is easy to see how such symbol strings could be encoded as integers; if x is a measure of the size of state S in characters, then the integer encoding state S_n can be made approximately proportional to x_n , so we must estimate the cost of using the integer x_n . Let x be expressed in the base b number system; then

$$x = b_N b_{N-1} \ldots b_1 b_0 \qquad (3.20)$$

where the b_i are integers satisfying $0 \le b_i < b$ and $b_N \neq 0$. The right hand side of (3.20) is shorthand notation for the sum

$$x = b_N b^N + b_{N-1} b^{N-1} + \ldots + b_1 b + b_0 \quad ;$$

then, approximately,

$$\log_b x = N + \log_b b_N \, .$$

In fact, it is easy to check that

$$N \le \log_b x < N+1$$

so $(N+1)$ is approximately $\log_b x$. Now the cost of using x in base b is the cost of writing its representation (3.20); since there are just b different values that can be assumed by each of the b_i, and $(N+1)$ places, a total inventory of $b(N+1) \sim b \log_b x$ symbols must be examined. Since $\log_b x = \log x / \log b$, we see that the cost of representation in any base is proportional to $\log x$, which justifies Mandelbrot's use of (3.19) with $a = 0$; incorporation of a can be viewed either as a purely formal generalization or as an attempt to account for system overhead and/or specially efficient processes for small x.

Maximization of (3.17) is not the only reasonable method of including the effect on p of the cost of usage. W. E. Houchin pointed out to the authors (in a private communication) that the information per unit cost of x is $\log p(x)/c(x)$ so

$$- \Sigma \, p(x) \log p(x) \, / \, c(x) \qquad (3.21)$$

is the average information per unit cost, which is different in general from (3.17), the average information per average unit cost. Proceeding to maximize (3.21) subject to the constraints $\Sigma \, p(x) = 1$ and $\Sigma \, c(x) = C$, we differentiate

$$- \Sigma \, p(x) \log p(x) \, / \, c(x) \; + \; a_1 \Sigma \, p(x) \; - \; a_2 \Sigma \, p(x) c(x)$$

to find the condition, apart from a constant multiplicative factor of p:

$$\log p(x) = -1 \; + \; a_1 c(x) \; - \; a_2 c(x)^2 \qquad (3.22)$$

in place of (3.18). If, as before, $c(x)$ is taken proportional to $\log x$, or more generally, to $\log(x-a)$ with some constant a, then (3.22) is equivalent to

$$p(x) = \frac{N}{s\sqrt{2\pi}(x-a)} \exp -\tfrac{1}{2}\left\{\frac{\log(x-a) - m}{s}\right\}^2 \qquad (3.23)$$

with

$$m = (1 + a_1)/2a_2 \ ,$$
$$s = 1/2a_2 \ , \qquad (3.24)$$
$$N/N_0 = \sqrt{\pi/a_2}\ \exp(\tfrac{1}{2}(a_1+1)^2/2a_2 - 1) \quad , N_0 \text{ constant} \ ,$$

the lognormal distribution with lognormal mean m and lognormal standard deviation s.

Solving for a_1 and a_2 in (3.24), we find that

$$a_1 = m/s^2 - 1 \ ,$$
$$a_2 = 1/2s^2 \qquad (3.25)$$

It is well known that the lognormal mean m and standard deviation s can be calculated from the first four moments of the distribution (3.23) and therefore can be estimated from the first four sample moments of an observed random sample. We shall see how this can be accomplished below.

The lognormal appears to have nothing in common with the power function distribution (3.13) which is so commonly used to describe observations of information systems, but this superficial view is misleading. The power function is a limiting case of the lognormal. Suppose that m and s approach infinity in such a way that the ratio m/s^2 approaches some constant, say k. Rewrite (3.23) as

$$p(x) = \frac{Ne^{-m^2/2s^2}}{s\sqrt{2\pi}} \cdot (x - a)^{\frac{m}{s^2}-1-\frac{1}{2s^2}\log(x-a)} \qquad (3.26)$$

As m and s approach infinity, the exponent approaches $k-1$. The normalizing coefficient is N_0/e so the lognormal distribution has the power function

$$p(x) = (N_0/e)(x - a)^{k-1} \qquad (3.27)$$
$$k = \lim_{m,s\to\infty} m/s^2 \ .$$

as a limit. Observe that the exponent in (3.27) will be less than -1 only if k is less than zero, that is, only if the lognormal mean is negative. Zipf's formulation corresponds to $k = -1$ (recall that the rank-frequency function is the integral of $p(x)$).

The same results can be more directly obtained by using (3.22). Suppose, as usual, that $c(x) = \log(x-a)$. If $a_2 = 0$, then the solution of (3.22) is the power function (we have introduced the constant normalizing multiplier N_0)

$$p(x) = (N_0/e)(x - a)^{a_1} \quad ; \qquad (3.28)$$

from (3.25) a_1 is expressed in terms of m and s as $m/s^2 - 1$, and a_2 is $1/2s^2$. Hence $a_2 = 0$ requires that s approach infinity, and a_1 finite demands that m also approach infinity but in such a manner that the ratio m/s^2 is finite, so (3.28) coincides with (3.27).

These arguments show that the power function is a special case of the lognormal distribution. Consequently any data which is approximated by a power function must necessarily be at least equally well approximated by a lognormal function. Moreover, there can be no debate whether the power function is the 'correct' mode for describing the distribution of information states in our context. Rather, one should hold the view that a power function description is more economical than a lognormal one because the latter involves one more parameter. Therefore, if m and s are large in such a way that m/s^2 is relatively small, then the limiting power function form of the lognormal may prove more convenient and equally accurate within the restrictions imposed by the finiteness of the data sample and its other inherent imperfections.

It is instructive to examine the graph of the power function limit of the lognormal. If $\lim (m/s^2) < 1$, then the lognormal approximants of the limiting power function have graphs qualitatively like that displayed as Figure 3.3; if $s \to \infty$ such that m/s^2 remains constant and less than 1 , the peak P of the lognormal distribution moves toward the vertical axis and upwards. Precisely stated, if the coordinates of P are (x,y) , then $x \to 0$ and $y \to \infty$. The resulting power function has a graph like that shown in Figure 3.4 .

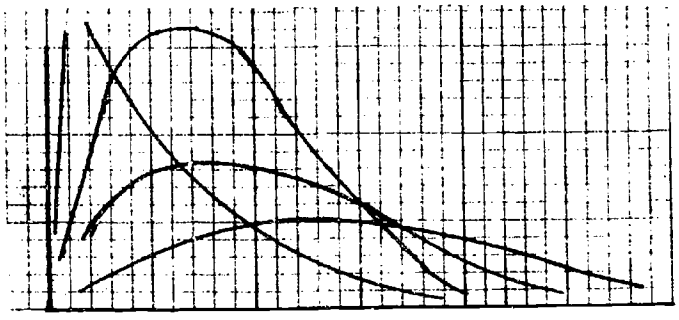


Figure 3.3

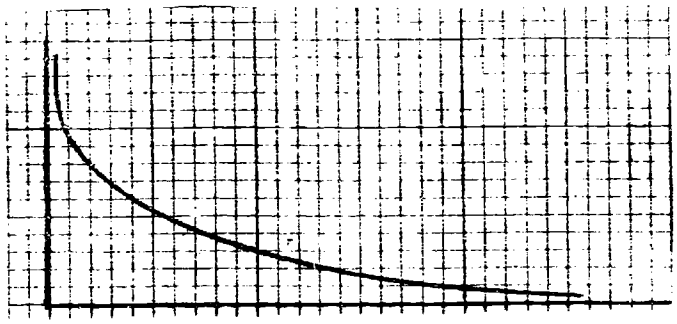


Figure 3.4

If m/s^2 approaches a finite limit greater than 1 , then the approximating lognormal function have graphs qualitatively like that shown in Figure 3.5 and passage to the limit power function through lognormal distributions having a fixed value of m/s^2 which is greater than 1 moves the peak P to the right and upward, that is, the coordinates (x,y) of P both approach infinity. The graph of the resulting power function is increasing, as shown in Figure 3.6 , and corresponds to certain types of power functions occurring in psychophysical theory; cp. <24>.

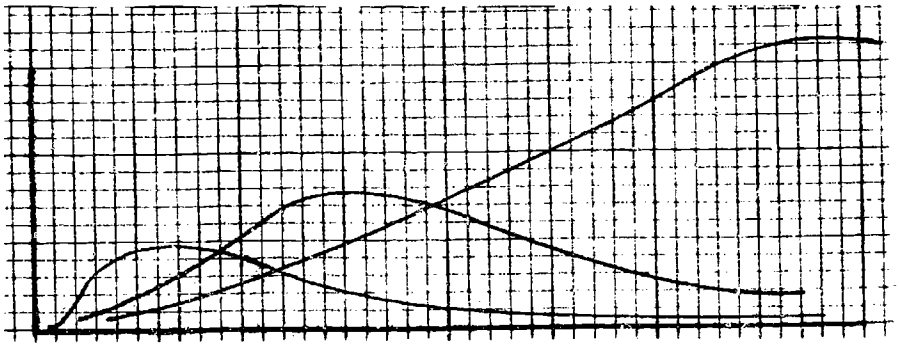


Figure 3.5

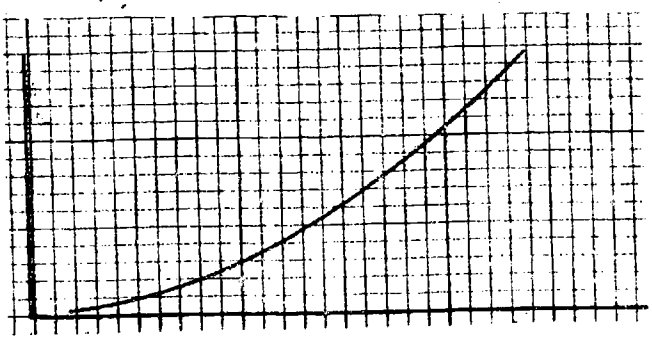


Figure 3.6

Finally, if $m/s^2 = 1$ and s approaches infinity, then the corresponding lognormal functions have as a limit the constant distribution $p(x) = N_0/e$. Figure 3.7 suggests how this occurs.

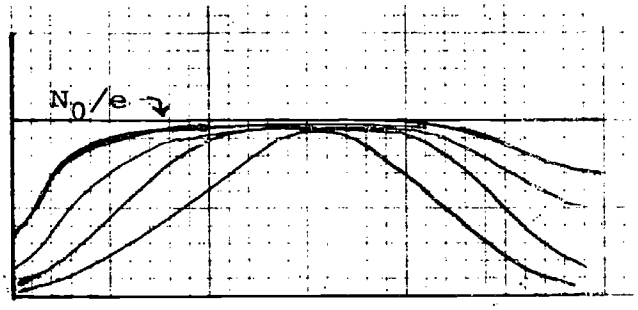


Figure 3.7

Mandelbrot's derivation of the power function and the derivation of the lognormal function given above have in common the maximization of a quantity that can be interpreted as information per unit cost. Knowledge of the functional forms that maximize this quantity make it in principle possible to calculate the maximum value of the information per unit cost as a function of the parameters of the distribution and thereby obtain a measure of performance of information systems which are described by these distributions. Unfortunately, a convenient explicit expression in the case of the lognormal distribution seems out of the question, but it is not hard to evaluate the maximum information per unit cost H^* for the power function by using (3.17) with $c(x) = \log x$. In order to be able to compare the maximum information per unit cost corresponding to different power functions, i.e., to different collections of data approximated by power functions, they will be normalized to refer to the same total number of items, which we take to be 1. Then the power functions can be written in the form

$$p(x) = x^{-s}/\zeta(s) \tag{3.29}$$

(cp. eq(3.6); s should not be confused with a lognormal standard deviation). Substitution in (3.17) with $c(x) = \log x$ yields

$$H^* = -\frac{\Sigma\, x^{-s}\{-\log \zeta - s \log x\}}{\Sigma\, x^{-s} \log x} \tag{3.30}$$

From $\zeta(s) = \Sigma\, x^{-s}$ we deduce

$$d\zeta/ds = \Sigma\, x^{-s} \log x$$

so (3.30) can be simplified to

$$H^* = s - \frac{\zeta \log \zeta}{d\zeta/ds} \quad . \tag{3.31}$$

For large s, $\zeta(s) = 1 + 2^{-s} + 3^{-s} + \ldots$ implies

$$\log \zeta = 2^{-s} + \ldots$$

and

$$d\zeta/ds = -(\log 2)2^{-s} + \ldots$$

so

$$H^* = s + 1/\log 2 + \ldots \quad , \quad s \text{ large.} \tag{3.32}$$

We can rephrase this result as follows: for large s, the maximal average information per average cost is related to the exponent s of the power function describing the information distribution by eq(3.32). The larger s, the more information per unit cost is conveyed by the information system. If the power function is represented graphically using the usual log-log graph paper so that the graph of the power function is a straight line, the exponent s corresponds to the negative of the slope of the line, and our result has the simple interpretation that the steeper the line, the greater the information per unit cost of the information system. This result will be applied to study the information content per unit effort for back of the book indexes in Chapter IV.

If s is close to 1 (but still greater than 1), then $\zeta(s) = 1/(s-1) + 0.577\ldots$ + terms with a factor of $(s-1)$; in this case, H^* can be approximated by

$$H^* = s - (s-1)\log(s-1) + \ldots ; \tag{3.33}$$

the second term is positive because $s-1$ is greater than 0 and $\log(s-1) < 0$ for s near 1. It follows that H^* decreases as s approaches 1; this can be shown by writing $t = 1/(s-1)$ and expressing $-(s-1)\log(s-1)$ in terms of t to find $\log t/t$. As s approaches 1, t approaches infinity and, using the properties of the logarithm, $\log t/t$ approaches 0. The conclusion that can be drawn from these remarks is that the possible values of H^* are all greater than 1 but they approach 1 as a lower bound as s approaches 1, in a steadily decreasing way.

Thus far in this Chapter we have derived the relation between a frequency distribution and its corresponding rank distribution, the functional form of information frequency distributions satisfying certain simple and sensible maximization conditions connected with the information per unit cost carried by and information system, and the maximum information per cost of a power function system as a function of its characteristic exponent. Now we will turn to the problem of determining the form of the rank-frequency distribution which corresponds to the lognormal frequency function.

Earlier we showed that the rank distribution corresponding to a given frequency distribution is a power function if and only if the frequency distribution is a power function. We therefore conclude that the rank distribution corresponding to a lognormal frequency distribution cannot be a power function, but, since power functions are limits of lognormal functions, they will also turn out to be limits of the rank distribution of lognormal functions.

Suppose the lognormal function is given by (3.23). From (3.9) one finds for the corresponding rank distribution the equation

$$r(x) = \int_x^\infty (N/s\sqrt{2\pi}(x-a))\exp-\tfrac{1}{2}\left(\frac{\log(x-a) - m}{s}\right)^2 dx. \quad (3.34)$$

Introduce the new variable $u = (\log(x-a) - m)/s$ and put $R(u) = r(x)$. Then (3.34) becomes

$$R(u) = (N/\sqrt{2\pi})\int_x^\infty \exp-\tfrac{1}{2}u^2 du \quad ; \quad (3.35)$$

this function, the cumulative normal distribution, has been tabulated, for instance in Sheppard <23>; its graph is shown in Figure 3.8 with $N = 1$. From its definition it is clear that u is large if and only if x is large. The behaviour of the rank function $r(x)$ can be studied for large values of x by studying the corresponding behaviour of $R(u)$ for large values of u. As shown in reference <23>,

$$\int_u^\infty \exp-\tfrac{1}{2}u^2 du = \frac{\exp-\frac{1}{2}u^2}{2u}\left(\frac{1}{u+}\,\frac{1}{u+}\,\frac{2}{u+}\,\frac{3}{u+}\,\ldots+\frac{n}{u+}\,\ldots\right) \quad (3.36)$$

where the expression in parentheses is a continued fraction expansion. For large u this can be condensed to the approximation

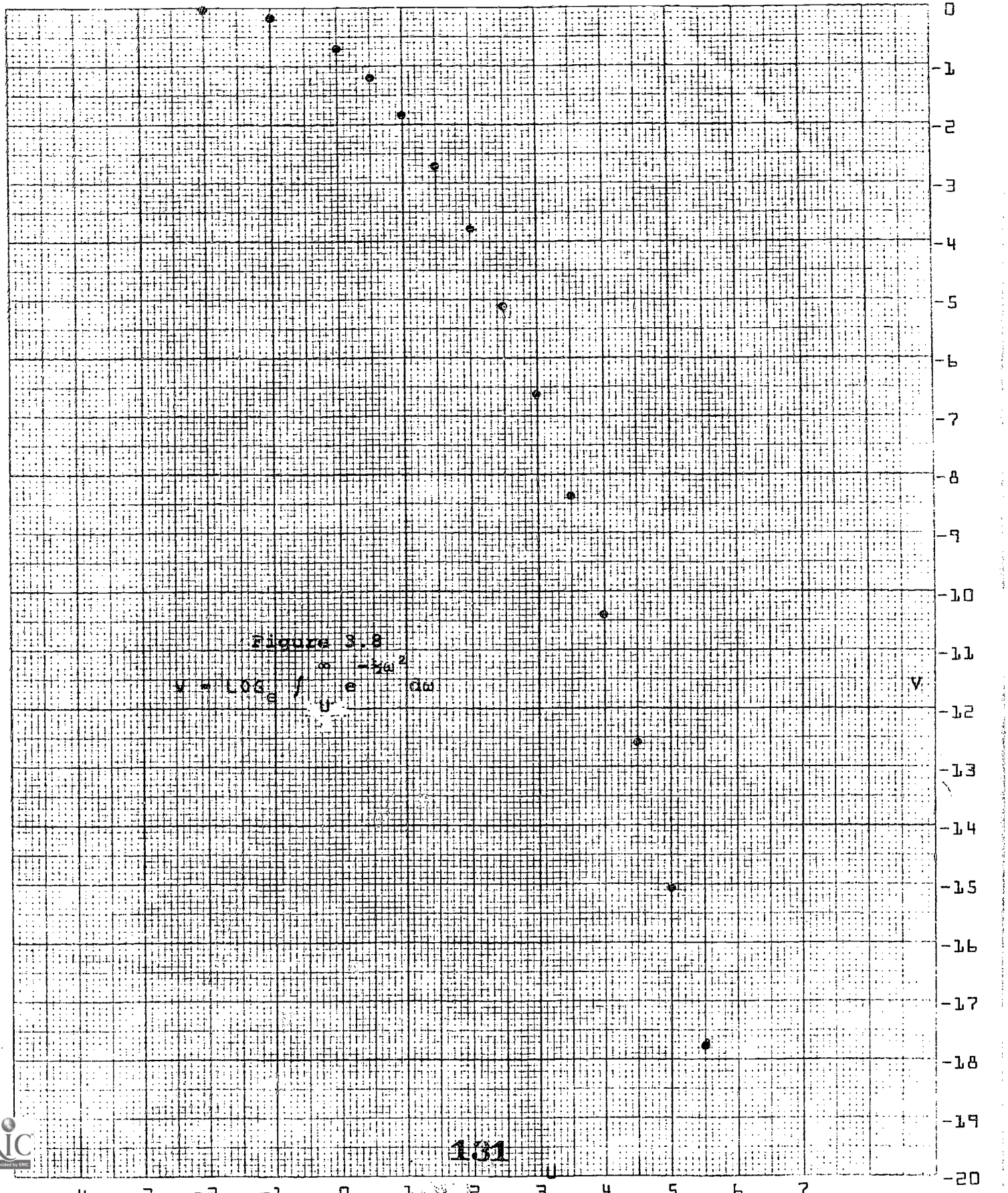


Figure 3.8

$$v = \log_e \int_u^\infty e^{-\frac{1}{2}w^2} dw$$

$$\log r(x) = \log R(u) = \log(N/\sqrt{2\pi}) - \tfrac{1}{2}u^2 - \log(2u)$$

which is adequate to show that $\log R$ varies essentially quadratically with u. Since $\log(x-a) = su + m$, the same holds true for $\log(x-a)$. Compare this conclusion with that corresponding to the rank-frequency distribution for the power function: we may take the latter to be of the form

$$r(x) = k(x-a)^{-s}$$

so

$$\log r(x) = \log k - s \log(x-a),$$

which is linear in $\log(x-a)$. When drawn on log-log graph paper, the graph of the rank-frequency distribution of the power function is a straight line whereas that of the log-normal distribution tends to the form of a parabola for large values of x. The characteristic curvature often apparent in the tail of rank-frequency graphs can sometimes be attributed to an underlying lognormal frequency distribution rather than to observational deficiencies and a power function distribution.

At this point the reader may find some examples helpful. We will first study book usage data kindly provided by Harvard Universiy's Widener Library.

Table 3.1 lists the outside usage distribution for a period of approximately five years encompassed in the interval 1965-1969. Several natural questions come to mind: is this distribution conveniently described by some well-known function of statistical theory? Is it optimal in some information theoretic sense? Is there a relationship between the size of the collection and the usage distribution? How do the parameters that determine the usage distribution depend on the dynamic variation of collection size with time?

In order to respond to the first question, display the data of Table 3.1 on log-log graph paper. If the distribution can be represented by a power function, i.e., by a Bradford-Zipf-Mandelbrot distribution, the sample points will tend to fall along a straight line. It is evident from Figure 3.9 that this is not the case. If, however, the distribution is log-normal, then extraction of the logarithm of eq(3.23) shows that $\log p(x)$ is a quadratic function of $\log(x-a)$, which implies that the corresponding sample data points will fall close to a parabola; conversely, should the sample points fall along a parabola, the corresponding distribution will be lognormal (and $a = 0$ unless the sample data values are

Table 3.1

Harvard University Widener Library
Distribution of Outside Usage
1965-1969

U	B	$\log_{10}U$	Least Squares $\log_{10}U$
1	260878	0.0000	-0.0163
2	72911	0.3010	0.3150
3	36022	0.4771	0.4808
4	21179	0.6021	0.5978
5	13560	0.6990	0.6903
6	9409	0.7782	0.7628
7	6666	0.8451	0.8282
8	5136	0.9031	0.8755
9	3752	0.9542	0.9302
10	2886	1.0000	0.9742
11	2255	1.0414	1.0140
12	1700	1.0792	1.0579
13	1322	1.1139	1.0951
14	1086	1.1461	1.1234
15	765	1.1761	1.1710
16	631	1.2041	1.1957
17	479	1.2305	1.2299
18	382	1.2553	1.2564
19	303	1.2788	1.2824
20	189	1.3010	1.3310
21	162	1.3222	1.3458
22	95	1.3424	1.3922
23	81	1.3617	1.4046
24	66	1.3802	1.4198
25	32	1.3979	1.4650
26	34	1.4150	1.4617
27	17	1.4314	1.4939
28	11	1.4472	1.5081
29	11	1.4624	1.5081
30	5	1.4771	1.5220
31	4	1.4914	1.5231
32	3	1.5052	1.5228
33	2	1.5185	1.5188
34	1	1.5315	1.5027
35	4	1.5441	1.5231
36	3	1.5563	1.5228
40	1	1.6021	1.5027
47	1	1.6721	1.5027
	442044		

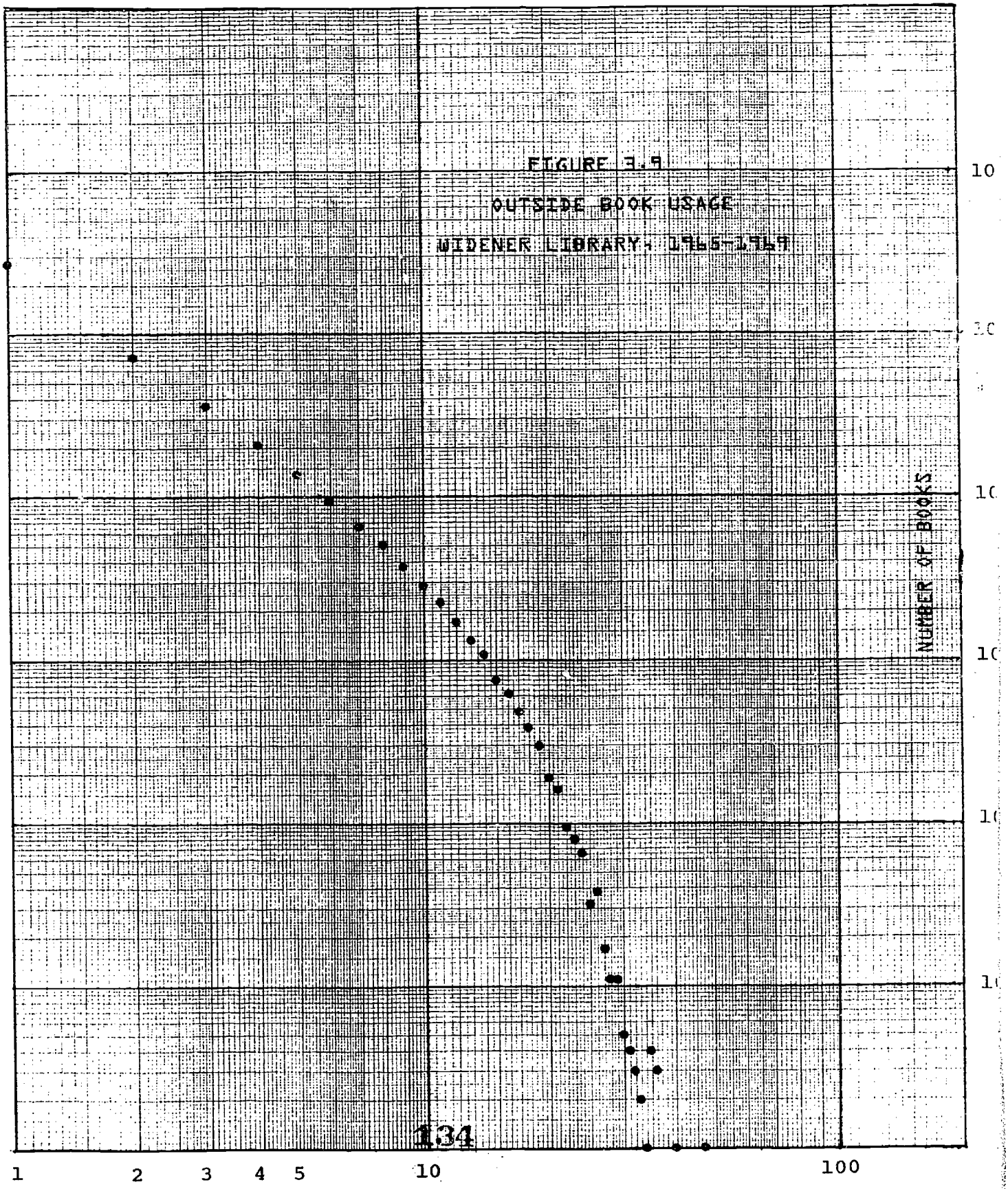


FIGURE 3.9
OUTSIDE BOOK USAGE
WIDENER LIBRARY: 1965-1969
NUMBER OF BOOKS
10
1
2
3
4
5
10
100

shifted by $-a$ before the data is plotted). Figure 3.9 suggests that the number of uses U may be a lognormal function of the corresponding number of books B because $\log U$ appears to be of the form

$$\log U = a \log^2 B + b \log B + c \qquad (3.37)$$

for some constants a, b, c. It is important to observe that if $\log U$ is a quadratic function of $\log B$, then $\log B$ cannot be a quadratic function of $\log U$; only one of U and B can be a lognormal function of the other, in contradistinction to the power function situation for which U is a power function of B if and only if B is a power function of U.

Let us suppose that $\log B$ could be approximated by a quadratic function of $\log U$, thus

$$\log_{10} B = a' \log_{10}^2 U + b' \log_{10} U + c'. \qquad (3.38)$$

From Table 3.1 one finds the following association of values:

B	U
260,878	1
2,886	10
1	40 ;

substitution in (3.38) leads to

$$a' = -2.367,$$
$$b' = 0.411,$$
$$c' = 5.416$$

as approximations of the coefficients of the best fitting parabola of the form (3.38). Graphing the parabola corresponding to these values leads to a curve which is a poor approximation of the data in Table 3.1. On the other hand, (3.37) approximates the sample observations rather accurately for appropriate values of a, b, c. Least squares curve fitting with respect to the logarithms $\log_{10} U$ and $\log_{10} B$ yields the equation

$$\log_{10} U = -0.0653 \log_{10}^2 B + 0.0732 \log_{10} B + 1.5027 \; ; \qquad (3.39)$$

Converted to natural logarithms, this becomes

$$\log U = -0.02841\log^2 B + 0.0732\log B + 3.4601 \quad . \qquad (3.40)$$

If U is a lognormal function of B of the form (3.23) with $a=0$, then extraction of the logarithm implies

$$\log U = (-1/2s^2)\log^2 B + (m/s^2 - 1)\log B + \{\log(N/s\sqrt{2\pi}) - (m^2/2s^2)\}$$

with m and s the lognormal mean and variance respectively. Substitution from (3.40) shows that the parameters of the lognormal distribution corresponding to the Widener usage data have the values

$$\begin{aligned} m &= 18.92 \quad , \\ s &= 4.20 \quad , \\ N &= 8.6 \times 10^6 \quad . \end{aligned} \qquad (3.41)$$

At first sight it is surprising to find U considered as a function of B because this does not have a natural interpretation as the inverse function, B as a function of U, does. Nevertheless there is strong evidence that it is more effective to consider U as a function of B. For instance, Figure 3.10 (on three pages) displays U as a lognormal function of B. It is clear that there is remarkable agreement through the 99th percentile.

The reader must be warned that, despite the unusual evidence of lognormality provided by Figure 3.10 and the equations

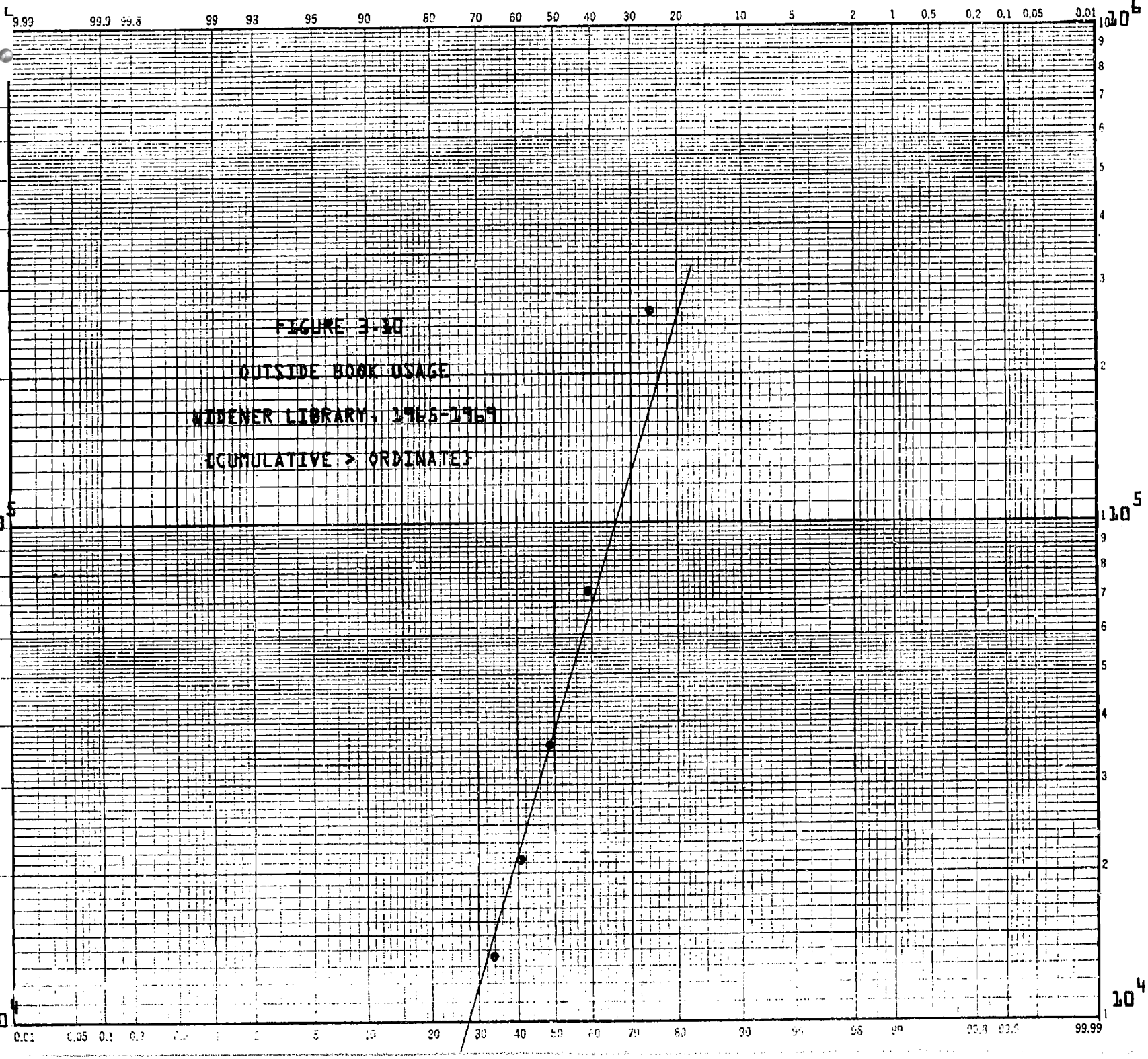
$$m = \log x_{50\%} \quad ,$$

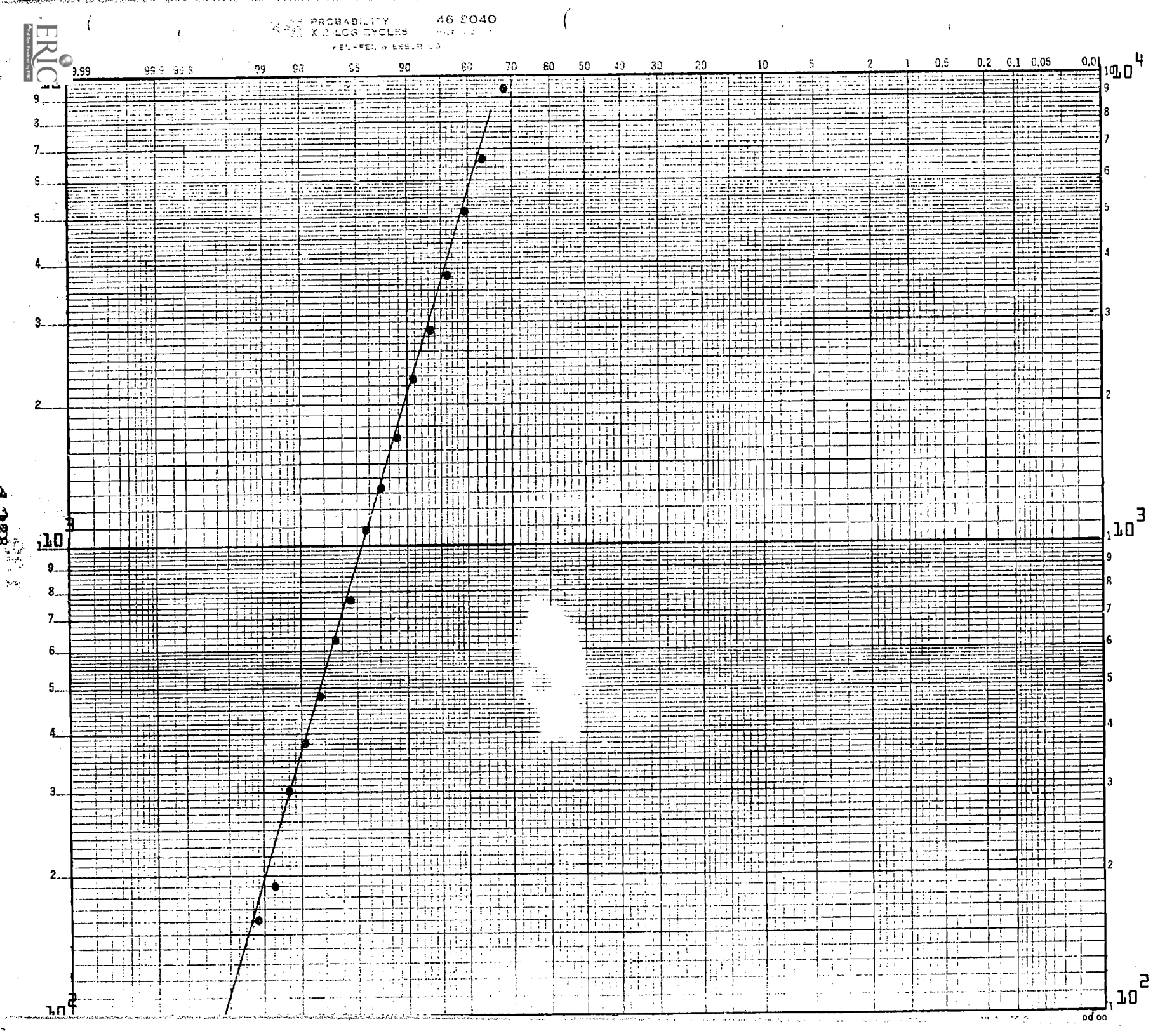
$$s = \log (x_{84\%}/x_{50\%})$$

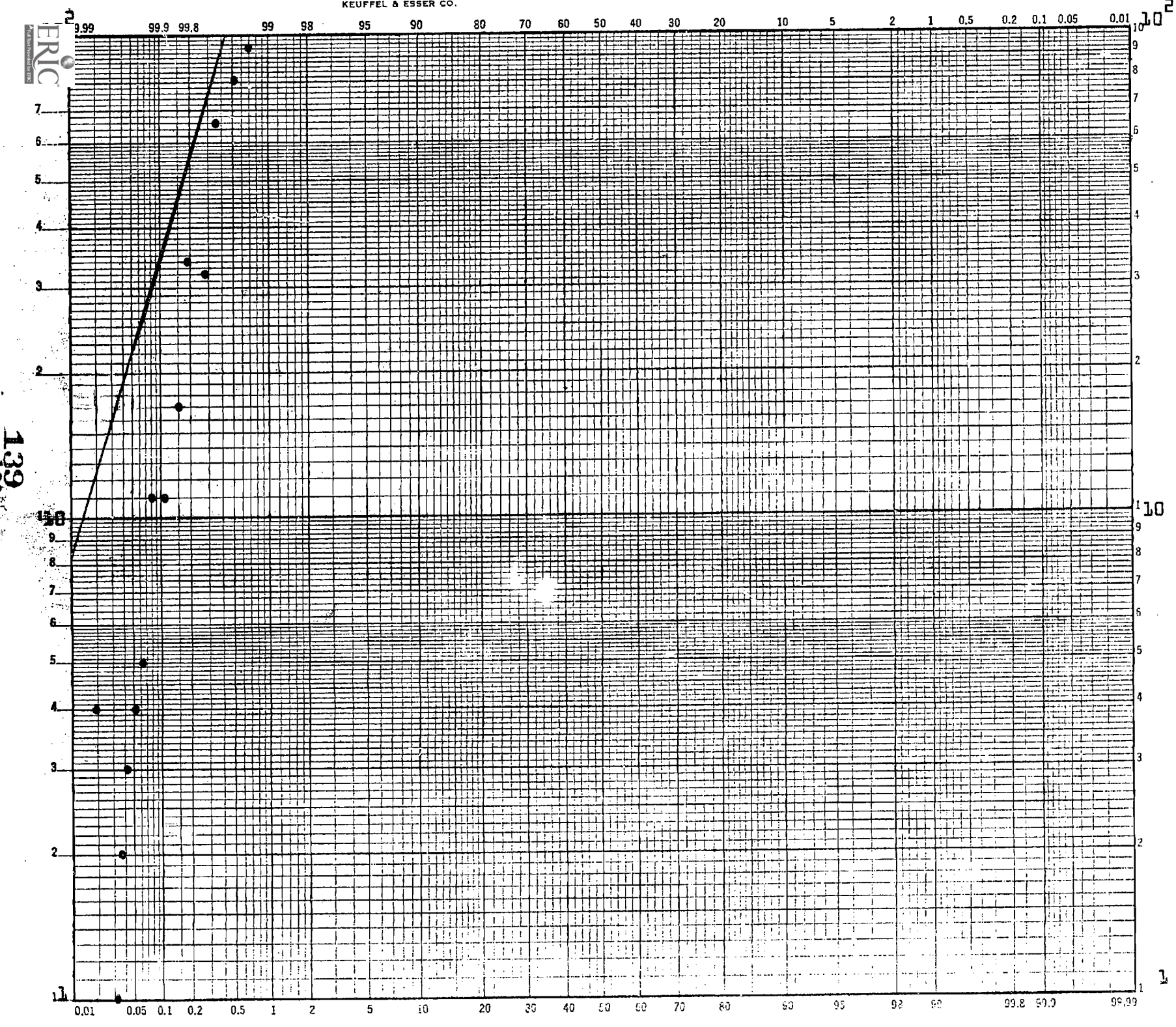
relating the lognormal parameters to the value of the independent variable x at the 50% and 84% (more accurately, the one standard deviation from the mean) point on the lognormal function, m and s cannot be accurately estimated this way unless the sample data represents the entire effective range of x.

There is good reason to suspect this is not true for the Widener usage data. From Table 3.1 we see that only 443,044 books participated in outside usage in the five year sample period, less than 6% of the volumes held by the Widener on 1 July 1967 (the approximate midpoint of the sampling period). The column labelled "least squares $\log_{10} U$" in Table 3.1 tabulates the value of $\log_{10} U$ corresponding to $\log_{10} B$ for points lying on the least squares parabola defined by eq(3.39);

KEUFFEL & ESSER CO.




PROBABILITY
X 2 LOG CYCLES
46 8040
99.99
99.9
99.8
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.05
0.01
10^4
10^3
10^2


KEUFFEL & ESSER CO.
9.99 99.9 99.8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0.5 0.2 0.1 0.05 0.01
0.01 0.05 0.1 0.2 0.5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99.8 99.9 99.99
10²
10
1

comparison with the column labelled "log U" shows generally good agreement, which imples that the values of m_{10}, s, N given in (3.41) represent the corresponding lognormal parameters with reasonable accuracy. Were more data points adjoined to the data of Table 3.1 (by extending the Widener sampling project for another 5 years, for instance), the newly acquired points would fall close to the same parabola if the distribution is actually lognormal, but the distribution corresponding to the cumulative function presented on lognormal probability graph paper in Figure 3.10 would not fall close to the present curve: both the lognormal mean m and the lognormal variance s will increase with increasing sample size when plotted on lognormal probability graph paper. The reader will do well to compare Carroll<3> and especially the subsample graphs therein exhibited with regard to this point.

With these preliminary observations in mind, notice that the estimate of N from data displayed on lognormal probability graph paper should be expected to return a value close to the number of sample values of the independent variable; in this case, the number 442,044 of books used, but the same N estimated from the parameters of the parabola which best fits the set of points $(\log B, \log U)$ should yield an estimate of the size of the total population from which the sample was drawn.

For book usage data it is evident that the number of books used, as determined by the sampling process, will increase with increasing duration of the sampling project; that any book held in the collection is a candidate for usage and hence for inclusion in the sample data if the sampling project is of sufficient duration; and therefore that the size of the sampling population must be understood as the size of the collection itself.

This provides a means of estimating collection size from the usually much smaller subcollection constituted by the books actually used in some period. In fact, for the Widener usage data, the sample of 442,044 usages accounted for less than 6% of the estimated 7,791,538 volumes constituting the collection on 1 July 1967, but yields the estimate $N = 8.6 \times 10^6$, ten percent high. Note that there were about 8.2×10^6 volumes in the collection at the end of the sampling period, compared with which the estimate of N is but five percent high.

It appears remarkable that the lognormal distribution, determined from so small a sample of actually used books, makes it possible to measure the number of unused books in the archive. But the information theoretic interpretation of the lognormal function shows why this should be so, and relates the information access role of the books used to those not used in the sampling period in a way which demonstrates the futility of

any attempt to partition an archival collection into useful and non useful components.

These remarks have obvious implications for the "weeding" problem. and, more significantly, for the problems posed by various governmental procedures for evaluating the adequacy of an archival collection in terms of its size, procedures which determine to some extent whether federal funds will be available to particular libraries.

Before turning to the information theoretic aspects of the book usage distribution we must respond to the wary reader who may wonder whether B as a function of U might not be well approximated by the Poisson distribution, a one parameter distribution determined by its mean. The sample mean of B for the data in Table 3.1 is 2.2856. From a standard handbook of mathematical functions we take the values of the cumulative Poisson distribution with mean 2.2 and 2.3 and compare these with the corresponding cumulative values determined from Table 3.1. The results, tabulated in Table 3.2 below, show that the Poisson is a poor approximation to B as a function of U.

Table 3.2

Cumulative Book Usage and Poisson Distributions

U	Observed Cumulative Fraction	Poisson Cumulative Mean = 2.2	2.3
1	0.5902	0.3456	0.3309
2	0.7551	0.6227	0.5960
3	0.8366	0.8194	0.7994
4	0.8845	0.9275	0.9163
5	0.9152	0.9751	0.9700
6	0.9365	0.9925	0.9906
7	0.9515	0.9980	0.9967
8	0.9632	0.9995	0.9993

Now we will show how the lognormal distribution of usage as a function of number of books maximizes a certain information access function. Suppose that a usage distribution is given, such as that exhibited in Table 3.1. Amalgamate the set of books corresponding to a fixed usage U; the number B of such books is a measure of information size in the sense that B is proportional to the total

number of characters contained in the books used U times if the sample is large enough, for in this instance the total number of characters will be approximately equal to the average number of characters per book for all books multiplied by the number B of books. We will take B itself as the measure of size of this information base, that it, we will measure size relative to the unit of measurement provided by the size of the average book.

The number of times each of these books is used, U, can be interpreted as proportional to the information per unit cost associated with using a book used U times, from which it follows that UB is proportional to the information per unit cost associated with the information collection of size B. Indeed, if the set of B books under consideration were less (more) informative for fixed cost, or effort, of use, then usage would presumably decrease (increase); and if the effort, or cost, required to examine these books were decreased (increased), there would presumably be more (less) usage, for fixed information return. Recall that our general arguments presented at the beginning of this chapter show that maximum information UB per unit cost will be returned if UB is a lognormal function of the size B. It remains to make the simple observation that UB is a lognormal function of B if and only if U is a lognormal function of B. Even more is true: for any number t, UB^t is a lognormal function of B if and only if U is a lognormal function of B, for $\log(UB^t) = \log U + t \log B$ is a quadratic function of $\log B$ if and only if $\log U$ is a quadratic function of $\log B$, which is equivalent to the preceding assertion (cp. eq(3.23)).

Since the Widener usage data U are accurately fit by a lognormal function of B, as shown above, we can conclude that Widener usage maximizes information per unit cost. The same result most likely holds generally, for public as well as university libraries, and perhaps in more general information situations as well. This area of study certainly deserves more extensive investigation.

The reader will have noticed that Figure 3.10 was constructed using cumulative fractions of the distribution summed over values of the variable strictly greater than the corresponding ordinate value of B. Were the distribution continuous, cumulative fractions constructed by summing over values of the variable $> B$ would lead to the same results but because the observed sample data is discrete, there will inevitably be some difference. The difference will be negligible where the values of B are closely spaced, but can be significant if they are not. We have displayed the difference for the Widener data by including Figure 3.11 constructed using cumulation summed for values $\geq$ the ordinate value. The reader should compare the resulting curve with that in Figure 3.10, especially noting the presence in the latter of the point corresponding to the ordinate 260,878, which point is not represented at all in Figure 3.11 due to the method of cumulation.

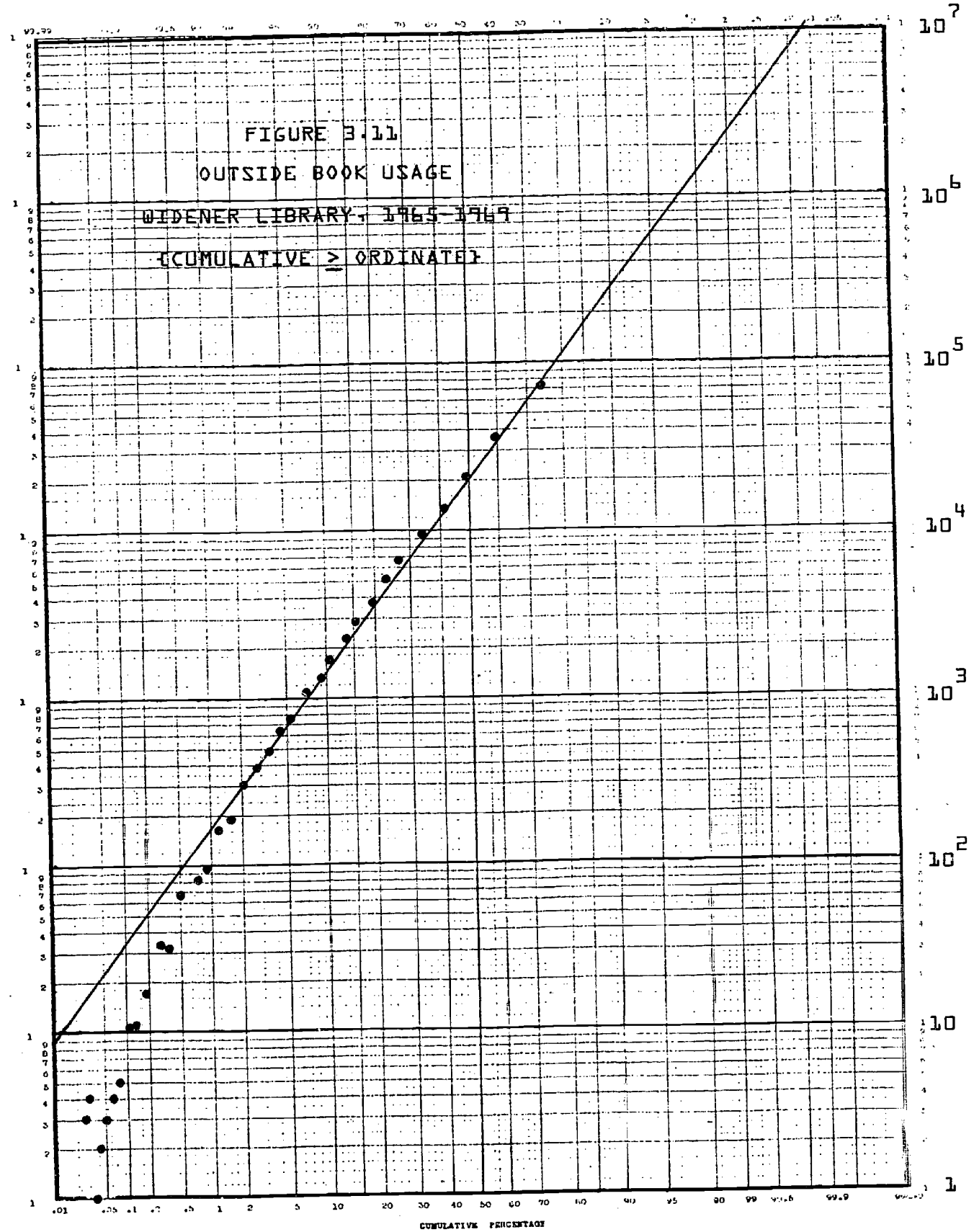


FIGURE 3.11
OUTSIDE BOOK USAGE
WIDENER LIBRARY, 1965-1969
(CUMULATIVE ≥ ORDINATE)
CUMULATIVE PERCENTAGE

In Chapter I we briefly mentioned the problem of determining the optimal number of file guides per drawer in a library catalog. This problem, solved by Lipetz and Song (13) (cp. Shoffner (30)), is a special case of the general problem of determining the optimal size of each access subsystem to an information base. Here we extend the method of Lipetz and Song to solve this more general problem. It will appear that the optimal size relationships amongst the components of an access system hold when the various levels have sizes which are equally spaced when measured by the logarithm of their size. This is in agreement with observations for an important variety of access systems, as was shown in Chapters I and II.

Suppose given an information base B_0 of size s_0 measured, say, in characters. Let B_i be a sequence of information bases, with $i = 1,2,\ldots,n$ such that the base B_i is an access system for B_{i-1}, and let s_i denote the size of B_i. For example, we may think of B_0 as the text of a book, B_1 as the index to B_0, B_2 as the table of contents to B_0 or as an "abstract index" to B_1, and B_3 as the title of B_0; or we may think of B_0 as the collection of information contained on the catalog cards in one drawer of a card catalog and B_1 as the information on the corresponding file guide cards.

We wish to determine the relationship amongst the access subsystem sizes s_i which minimizes search time, which we take as an appropriate measure of search cost or search effort, subject to certain hypotheses about the nature of the search process which will be made clear in the development of the argument.

Following Lipetz and Song, let the reciprocal rate of search in information base B_i be $2r_i$ seconds per character. Then the expected mean search time through access subsystem B_n to locate a desired element of subsystem B_{n-1} will be $r_n s_n$ seconds. The size of the portion of B_{n-1} remaining to be searched to locate information in B_0 more precisely is s_{n-1}/s_n so the expected mean search time to locate a desired element of B_{n-2} will be

$$r_n s_n + (r_{n-1} s_{n-1}/s_n) ;$$

similarly, the total expected mean search time required to locate an element of B_0 by proceeding through the level structured access system constituted by the B_i's will be

$$T = r_n s_n + \sum_{i=0}^{n-1} r_i s_i / s_{i+1} .$$

Fix the sizes s_0 and s_n, that is, the size of the information base to be accessed and the size of the smallest access subsystem, and minimize the total expected mean search time T as a function of the the s_i, $1 \leq i < n$. We find the equations

$$\partial T/\partial s_i = r_i/s_{i+1} - r_{i-1}s_{i-1}/(s_i)2 \quad , \quad 0 < i < n \ ,$$

so an extremum of T is obtained if

$$s_{i+1}/s_i = (r_i s_i)/(r_{i-1}s_{i-1}) \quad , \quad 0 < i < n \ .$$

This chain of equations is equivalent to the formula

$$s_i = s_0 \prod_{j=0}^{i-1} (r_j/r_0)(s_1/s_0)^i \quad , \quad 0 < i \leq n \ . \tag{3.42}$$

If all search rates are equal, then $r_i = r$ say, so

$$s_i = s_0(s_1/s_0)^i \quad , \quad o < i \leq n \ ; \tag{3.43}$$

when measured by the logarithm of their size, the access levels B_i are therefore equally spaced, i.e.,

$$\log s_i - \log s_{i-1} = \log (s_1/s_0) = \text{constant} \ ,$$

$$0 < i \leq n \ ,$$

which is what is in fact observed for large traditional natural access systems.

It remains to verify that the extremum attained is indeed a minimum. This is equivalent to proving that the matrix of second derivatives $(\partial^2 T/\partial s_i \partial s_j)$ is positive definite. We will not include the rather technical proof here.

It may be worth stressing that only the mean search time is minimized by the equally spaced level structured access system. If certain types of search inquiry have a greater value than other types, and if the more important inquiries have a search time distribution which is different from the general distribution, other strategies for structuring the level structured access system may be more efficient. However, in the absence of criteria concerning the utility or importance of information base inquiries, our hypotheses seem sensible.

Now suppose that the n levels of a level structured access system are spaced so that their size distribution minimizes the mean search time T. Let the size of the largest level be s_0 as before, and that of the smallest level be s_n. In eq(3.43) set

$$K = s_0/s_1 > 1 \ ;$$

then eq(3.43) expresses the size of the i^{th} access level in terms of the size of the information base itself, the integer which specifies the level, and the constant K as:

$$s_i = s_0 K^{-i} ,$$

and K is connected with the number of levels by

$$s_n = s_0 K^{-n} . \tag{3.44}$$

In terms of these quantities T has the expression (with $r_i=1$ for convenience; this is just a redefinition of the units with which we choose to work)

$$T = s_n + nK = s_n - (\log s_n/s_0)K/\log K . \tag{3.45}$$

For this structure let us determine the number of levels which minimizes the expected mean search time T. It is intuitively clear that access systems do aid in searching by reducing search time, but it is just as evident that the number of access subsystems cannot be large and still allow search efficiency. This heuristic argument suggests that there must be some optimum number of equally spaced access levels that minimizes T; it is this number that we now want to determine. Recalling that s_0 and s_n are fixed numbers, we see that eq(3.44) expresses n in terms of the constant level multiplier K, and that therefore eq(3.45) expresses T as a function of K alone; all other terms are fixed. Therefore, in order to minimize T with respect to the number of levels of the access system, it will do to minimize T as a function of K. Then T will be minimal as a function of K if

$$0 = dT/dK = 1/\log K - 1/(\log K)^2 ,$$

from which we conclude that

$$K = 2.718\ldots = e ,$$

the base of the system of natural logarithms. If the more general eq(3.42) is used as a starting point in place of eq(3.43), the result is the same.

In the previous chapters we have shown that traditional access systems do have the level structure which minimizes search time T as a function of the size of the access subsystems, but the data there presented convincingly shows that the constant K which determines the relative size of adjacent levels is nearly equal to 30, certainly not equal to e. This discrepancy appears to be due to our failure to account for the cost of acquisition and maintenance of the access system.

The access system for B_0, including the data bases B_1, ..., B_n, has a certain cost of maintenance in addition to the cost of original acquisition. In what follows we will ignore the one-time acquisition cost which is essentially independent of the number of system uses (but which certainly is taken into account by the private purchaser of a book, and, perhaps less directly, by major libraries). Moreover, we will assume that maintenance requires a certain time T_m for each use of the system, and that T_m is proportional to the size of the total access system, thus

$$T_m = c \sum_{i=0}^{n} s_i = cs_0 \sum_{i=0}^{n} K^{-i}$$

$$= cs_0 \frac{1 - K^{-(n+1)}}{1 - K^{-1}}$$

$$= cs_0 (K - K^{(\log s_0/s_n)/\log K})/(K - 1) .$$

If no resources are made available for maintenance of the access system, i.e., then $c = 0$, and the system will inevitably deteriorate and the mean time to the successful completion of a search will increase; on the other hand, it might appear that increasing the availability of maintenance resources will reduce search time but we already know that "infinite" maintenance corresponds to the case where maintenance considerations are negligible when compared with search considerations, so the search time will approach the minimum derived above. Practical considerations therefore dispose one to consider the minimization not of T_s, the user search time, but rather of $T_s + T_m$. Now T_m is infinite if $K = 1$ and is a steadily decreasing function of K as K increases from 1, whereas T_s has (as we showed above) a minimum at $K = e$, and it steadily increases for values of K greater than e. It follows that the minimum of $T_s + T_m$ must occur for that value of K (greater than e) where the graph of T_s crosses the graph of T_m. An exact determination of this value of K which would test our model of level optimization with maintenance included would require detailed information about the time equivalent cost of file maintenance relative to the cost of file search. Moreover, the initial cost of the access system would also have to find its place in the analysis. In the absence of the necessary data we can say no more on this important subject.

REFERENCES

1. Belonogov, G. G., "On some Statistical Regularities in Written Russian," Vopr. Jazykoznanija, 7(1962), 100, (in Russian).

2. Bradford, S. C., "Sources of Information on Specific Subjects," Engineering, (1934), January 26.

3. Carroll, J. B., "On Sampling from a Lognormal Model of Word Frequency Distribution," Computational Analysis of Present-Day American English (Henry Kucera and W. Nelson Francis), Brown University Press, Providence, Rhode Island, (1967) 406-24.

4. Cramer, H., Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.

5. Dolby, J. L., Forsyth, V. J., Resnikoff, H. L., Computerized Library Catalogs: Their Growth, Cost, and Utility, M.I.T. Press, 1969.

6. Dolby, J. L. and Resnikoff, H. L., "On the Structure of Written English Words," Language 40(April-June 1964) 167-196.

7. Estoup, J. B., Gammes stenographiques, 4th Edition, 1916.

8. Fairthorne, R. A., "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction," Journal of Documentation, 25(1969), 319-43.

9. Good, I. J., "Statistics of Language: Introduction," Encyclopaedia of Linguistics, Information, and Control, Pergamon Press, London, (1969) 567-81.

10. Houston, N. and Wall, E., "The Distribution of Term Usage in Manipulative Indexes," American Documentation, 15(1964), 105-14.

11. Deleted.

12. Kucera, Henry and Francis, W. N., Computational Analysis of Present-Day American English, Brown University Press, Providence, 1967.

13. Lipetz, Ben-Ami and Song, C. T., "How Many Cards Per File Guide? Optimizing the Two-Level File," Journal of the American Society for Information Science 5(March-April 1970), 140-141.

14. Lotka, A. J., "The Frequency Distribution of Scientific Productivity," Journal of the Washington Academy of Sciences, 16(1926), 317.

15. Mandelbrot, B., "An Information Theory of the Statistical Structure of Language," Proceedings of the Symposium on Applications of Communication Theory, London, September 1952, Butterworth, (1953) 486-500.

16. Mandelbrot, B., "On the Language of Taxonomy: An Outline of a 'Thermostatistical' Theory of Systems of Categories with Willis (natural) Structure," Information Theory; Papers Read at a Symposium on Information Theory, London 1955, Butterworth, (1956) 135-45.

17. Pareto, V., Cours d'economie, politique, Lausanne, 1897.

18. Pearson, E. S., and Hartley, H. O., Biometrika Tables for Statisticians, Volume I., Cambridge, England: The University Press, 1954.

19. Pearson, Karl, "Mathematical Contributions to the Theory of Evolution," Philosophical Transactions, A, 186(1895), 343-414; 197(1901), 443-59; 216(1916), 429-57.

20. Price, D. J. de solla, Little Science, Big Science, Columbia University Press, 1963.

21. Schrodinger, E., Statistical Thermodynamics, Cambridge University Press, Cambridge, 1964.

22. Shannon, C. E., "Prediction and Entropy of Printed English," Bell System Technical Journal, (1951) 50-64.

23. Sheppard, W. F., The Probability Integral, British Association for the Advancement of Science, Mathematical Tables VII, Cambridge University Press, 1966.

24. Stevens, S. S., "Neural Events and the Psychophysical Law," Science, 170 (1970), 1043-1050.

25. Titchmarsh, E. C., The Zeta-Function of Riemann, Cambridge University Press, 1930.

26. Wall, E., "Further Implications of the Distribution of Index Term Usage," Parameters of Information Science: Proceedings of the American Documentation Institute Annual Meeting, 1964, Volume 1, American Documentation Institute, (1964) 457-66.

27. Yule, G. U., The Statistical Study of Literary Vocabulary, Cambridge: The University Press, 1944.

28. Zipf, G. K., Human Behavior and the Principle of Least Effort, Addison Wesley, 1949.

29. Zunde, Pranas and Dexter, M., "Indexing Consistency and Quality," American Documentation, 20(1969), 259-267.

30. Shoffner, Ralph M., "A Technique for the Organization of Large Files," American Documentation, (Jan. 1962) 95-101.

THE STRUCTURE OF

BACK OF THE BOOK INDEXES

Book indexes are among the most common and most ancient access mechanisms, although they have not always been loved. Glanville, in Vanity of Dogmatizing, said:

> Methinks 'tis a pitiful piece of knowledge that can be learnt from an index, and a poor ambition to be rich in the inventory of another's treasure,

and more recently T. E. Lawrence wrote:

> ...half-way through the labor of an index to this book I recalled the practice of my ten years' study of history; and realized I had never used the index of a book fit to read.

However, as an unnamed contributor to a recent edition of the Encyclopedia Britannica put it,

> (It has) become almost a sine qua non that any good book must have its own index.

Indeed, as we shall see below, more than one-third of all non-serial items in the shelf list of a medium size university library do contain an index, and it seems as if the back of the book index is not only here to stay but is in the process of spawning a genus of related tools for indicating "the position of information on any given subject".

The object of this chapter is to study indexes to books in order to determine what structure, if any, they possess. It is not surprising that indexes* exhibit great variability in size, content, and utility, which makes it difficult to assess their nature in general from an examination of one or several exemplars. We have elected to study indexes in three ways.

---

*Throughout this chapter 'index' will only refer to back-of-the-book indexes.

The first and most reliable way is based on the selection of a random sample of book indexes. Such a sample has been assembled by extraction of the indexes from all monographs represented in a random sample of the shelf list of a medium size university library; it consists of approximately six hundred thousand index terms spread throughout some 700 books, and will be described in what follows:

The second means of studying indexes is concerned with the structure exhibited by each index separately. Information of this sort cannot be obtained from statistical agglomerations; rather it demands that indexes be considered in detail and the resulting structures, as they are found, compared for a sample of indexes.

A book index directs the user to the location of specified information in the book to which it refers. Should the book in question not contain any indexed information about the subject of interest, the inquirer is left to continue his search in the indexes of other specified books. There are, of course, several indirect methods for deciding how the next book in the search process should be selected, utilizing information contained in the bibliographies or the linear shelf list order determined by a subject classification scheme such as that of the Library of Congress, but none of these have the virtue of immediacy nor of completeness. Our third means of studying indexes is based on a cumulative index to 80 books in the field of statistics. It appears to offer attractive efficiencies in the information search process while it provides a view of the overall structure of the field itself.

The Fondren Index Sample is a random sample of 668 monograph shelf list cards corresponding to indexed books. Multiple volumes catalogued on one shelf list card increase the sample somewhat so that a total of 706 indexes are represented.

The Fondren Index Sample is a subsample of the Fondren Sample, which is a random sample of cards drawn from the shelf list of the Fondren Library at Rice University. The Fondren Sample is described in some detail in Reference [1]. Analyses of the sample may be expected to accurately reflect the structure of library collections to the extent that they are similar to the Fondren collection; in particular the archival collections of medium size university libraries are probably generally similar although certain special fields

may be more or less well represented. For instance, the Fondren collection is particularly weak in law, medicine, and Russian language and literature, and strong in chemistry. These differences are unlikely to play a significant role in determining the reliability of the sample for studying index structure since indexes are relatively insensitive to the nature of the subject material to which they refer; the gross category differences, as between science and fine arts, are, as will be shown below, substantial, but the Fondren collection encompasses adequate representation in each of such broad categories.

There are special problems associated with the analysis of complex data drawn from any sampling process. The index sample is no exception. Some of the sample indexes have a format so unusual as to make them incomparable with the average index; a small number were written in non-Roman alphabets so we were unable to correctly identify the structural features of interest. Because the fraction of anomolous indexes was small, it was decided to delete them from the index sample for this initial study.

This decision was bolstered by another complication; not all of the books represented by the original random sample could be located for the present study, which took place about two years after the original selection of shelf list cards. The number of unlocatable items was 33, approximately 1.7% of the Fondren Sample; this is the effective rate of loss for the two year period in the sense that the usual mechanisms for tracking items not present on the shelf in their proper location were applied without success for these items, noting that just prior to the selection of the sample the shelf list had been checked against the shelf and weeded. This suggests that slightly less than 1% of the monograph archive is lost each year.

If all 33 unlocatable items had had indexes, they would have constituted nearly 4% of the index sample; items excluded for special reasons such as language or format incompatibility totalled 22. Therefore, not more than 7.5% and more likely not more than 4.5% of the indexed volumes in the Fondren Sample have been excluded from the index sample. With this preliminary in mind we can now turn to the consideration of the index sample.

First observe that not all monographs are candidates for indexing; we have found no Library of Congress Class "A" items in the sample which contain an index.

and therefore class "A" is excluded from all further considerations. Similarly, neither maps nor musical scores are indexible in the "back of the book" sense, so they too are excluded. Excluding these items and all serial publications, one finds that there are 1,830 relevant items in the Fondren sample. Of these, 668 have indexes; thus we find that 37% of the monographs in the Fondren sample contain indexes.

As previously noted, the 668 LC cards lead to a total of 706 volumes with indexes. The distribution of these 706 volumes by LC class is shown in Table 4.1 together with the fraction that is indexed for each class. This fraction runs from a low of 0.18 for N (Fine Arts) and P (Language) to a high of 0.61 for Q (Science) and 0.67 for Naval Science.

Table 4.1 also provides the mean number of index entries per book indexed. The grand mean for the collection is 836 index entries per book, with the class means varying from a high of 1,391 entries per book for class F (U.S. Local History) to a low of 614 for class J (Political Science).

The product of these two measures provides an average measure of the amount of access per book in the collection and in each of its subsets. This distribution is shown separately in Table 4.2. This list breaks rather naturally into three subsets of nearly the same size. The first seven categories (classes F, G, V, K, D, E, and Q) would seem to share the property that they are all primarily concerned with careful description of the world as it is and as it has been. The middle group (classes H, C, R, T, Z, L, and J) is primarily devoted to man's effort to cope with the environment described so carefully in the first group. The lowest group appears a bit anomolous in that it contains the core of the arts: music, philosophy, religion, language, literature, and the fine arts as well as the more mundane but ever present categories of war and agriculture. Although we should not like to make too much of this particular arrangement of the LC classes, Table 4.2 does provide an interesting example of the insight one gains into the use of the system of literary stores by rather elementary counting procedures.

The index sample consists of a total of 590,329 index entries spread across the 706 indexes. Table 4.3 lists the number of indexes as a function of the number of entries they contain, grouped by hundreds of index entries. Figure 4.1 exhibits the lognormality by showing the data of Table 4.3 plotted on lognormal paper. The standard deviation on the log scale is 0.442 which is at the upper end of the range for log-length distributions given in Chapter II.

Table 4.1

FONDREN SAMPLE: FRACTION OF SAMPLE ITEMS CONTAINING AN INDEX, BY LC LETTER CLASS

Class	Mean Number of Entries per Index	Fraction Indexed (rounded)	Fraction Class is of Fondren Sample	Short Class Name
B	667	.31	.100	Philosophy-Religion
C	690	.53	.009	History—Auxiliary Sciences
D	1,102	.51	.095	History & Topography (except America)
E	1,062	.49	.040	American (General) & U.S. (General)
F	1,391	.46	.027	United States (Local) & America (ex. U.S.)
G	1,264	.50	.011	Geography—Anthropology
H	697	.54	.104	Social Sciences
J	614	.46	.023	Political Science
K	1,375	.43	.004	Law
L	620	.49	.038	Education
M	915	.25	.015	Music
N	615	.18	.033	Fine Arts
P	714	.18	.300	Language & Literature
Q	850	.61	.093	Science
R	716	.50	.010	Medicine
S	638	.20	.006	Agriculture—Plant & Animal Husbandry
T	707	.47	.032	Technology
U	840	.22	.010	Military Science
V	934	.67	.005	Naval Science
Z	1,328	.24	.023	Bibliography & Library Science

Total relevant items in Fondren Sample = 1823

Number of these items indexed = 668

Fraction indexed = 668/1830 = 0.37

Table 4.?

INDEX ACCESS BY LC CLASS

LC Class	Mean No. Index Entries per Book	Short Class Name
F	640	U. S. Local History
G	632	Geography
V	626	Naval Science
K	591	Law
D	562	World History
E	520	U. S. History
Q	519	Science
H	376	Social Science
C	366	Auxiliary Sciences (History)
R	358	Medicine
T	332	Technology
Z	319	Library Science
L	304	Education
J	282	Political Science
M	229	Music
B	207	Philosophy-Religion
V	185	Military Science
P	129	Language Literature
S	128	Agriculture
N	111	Fine Arts

Table 4.3

FREQUENCY OF INDEX ENTRIES FOR ITEMS IN THE FONDREN INDEX SAMPLE

Number of Index Entries	Number of Indexes	Cumulative Number of Indexes	Cumulative Fraction of Indexes
0 - 99	16	16	.023
100 - 199	77	93	.132
200 - 299	83	176	.249
300 - 399	80	256	.362
400 - 499	70	326	.462
500 - 599	62	388	.549
600 - 699	46	434	.615
700 - 799	37	471	.667
800 - 899	39	510	.722
900 - 999	30	540	.765
1000 - 1099	17	557	.789
1100 - 1199	24	581	.823
1200 - 1299	13	594	.841
1300 - 1399	14	608	.861
1400 - 1499	8	616	.872
1500 - 1599	2	618	.875
1600 - 1699	13	631	.893
1700 - 1799	7	638	.903
1800 - 1899	7	645	.913
1900 - 1999	7	652	.923
2000 - 2099	7	659	.933
2100 - 2199	3	662	.937
2200 - 2299	5	667	.944
2300 - 2399	1	668	.946
2400 - 2499	1	669	.947

Table 4.3
(Continued)

2500 - 2599	2	671	.950
2600 - 2699	3	674	.955
2700 - 2799	3	677	.959
2800 - 2899	1	678	.960

3000 - 3099	3	681	.964
3100 - 3199	2	683	.967
3300 - 3399	1	684	.969
3500 - 3599	1	685	.970
3700 - 3799	2	687	.973
3800 - 3899	3	690	.977
3900 - 3999	2	692	.980
4000 - 4099	1	693	.981
4200 - 4299	1	694	.983
4700 - 4799	3	697	.987
4900 - 4999	3	700	.991
5100 - 5199	1	701	.993
5900 - 5999	1	702	.994
6200 - 6299	2	704	.997
6700 - 6799	1	705	.998
7000 - 7099	1	706	1.000

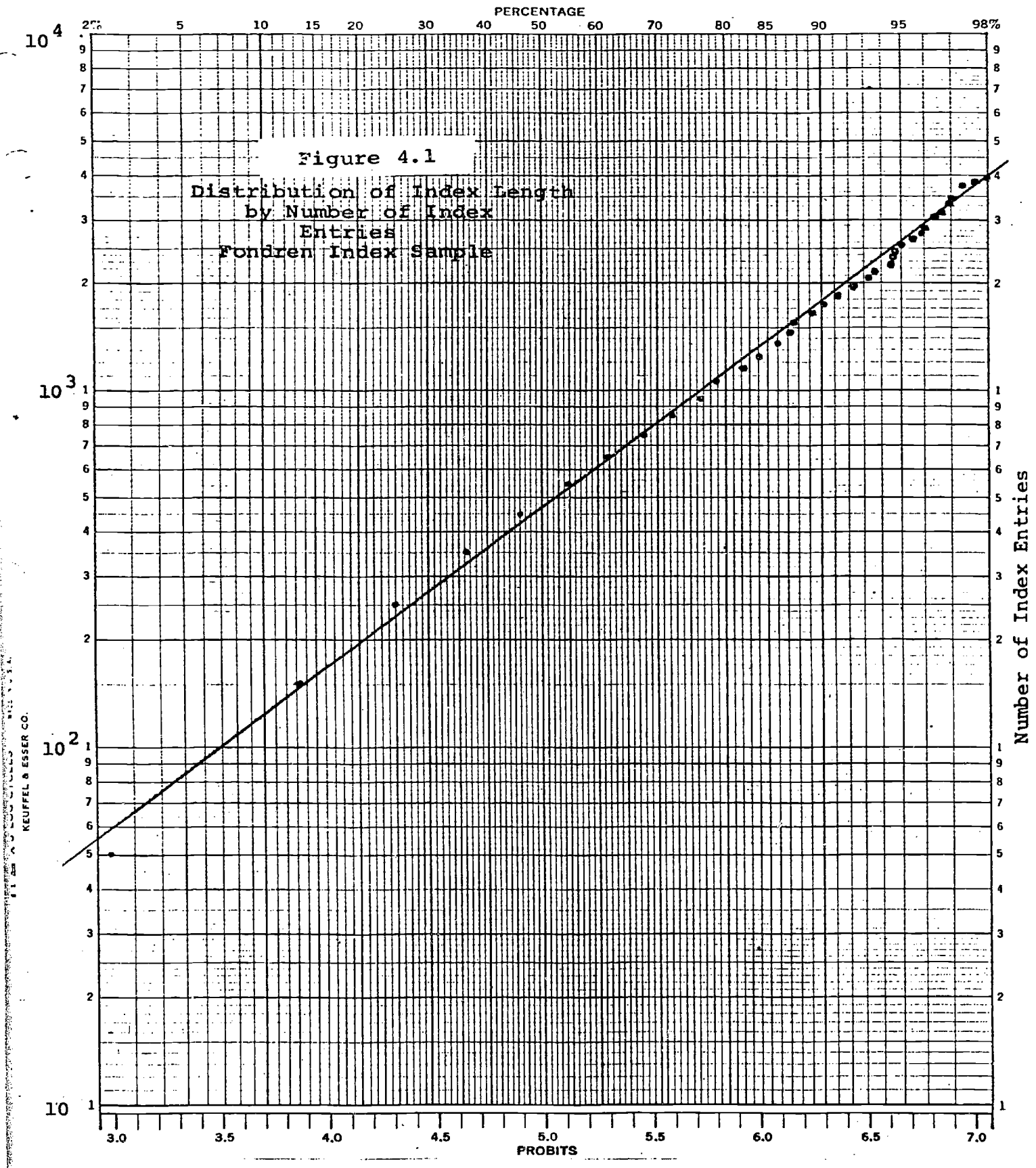


Figure 4.1

Distribution of Index Length by Number of Index Entries Fondren Index Sample

A distinction should be made between the number of index entries in an index and the number of locations to which these entries refer. The former quantity is the number of distinct word sequences appearing in an index, and is an absolute measure of index size which is independent of the details of format and page composition; the latter is usually the number of page locations referred to in an index, which clearly depends on the size of the page. In the Fondren sample of indexed books there are, on the average, 1.8 page locations per index entry. Thus, the 836 (average) distinct entries refer, on the average to 1,505 text locations. As there are on the average 341.5 pages per indexed book, there are 4.4 indexed text locations per page. Roughly speaking, this means that there is one index page location for each five sentences of text.

The aggregate size of the index as printed can be determined by estimating the average number of characters per entry and multiplying by the average number of entries. A preliminary estimate of the average number of characters was obtained by counting the entries in the cumulative index to statistical books (discussed at greater length in Chapter VI) as the format of the material is in particularly nice form for counting purposes. This estimate shows that the entries are about 25.47 characters in length exclusive of page location information. If, as in Chapter I, this is augmented by 4 characters per entry to include the typical page location reference information, then the average index of 836 entries consists of 24,637 characters and therefore the ratio of indexed book size to index size is about 33.27 to 1.

These global statistics provide a direct measure of the proportion of the monograph collection that is devoted to what might be called "self access". The aggreement of the access ratio (of about 30 to 1) with other access ratios developed in Chapter II helps to solidify the foundations of the level structured access model. Given the difficulty of assessing the quality of indexing (see (2) and the references therein) these statistics also provide the foundation of a basis for comparing various indexing procedures, particularly for comparing algorithmically derived indexes to manual indexes. The fundamental regularities of the length measures discussed here suggest that an algorithmically prepared index must at least be of the correct overall size to be of any use at all.

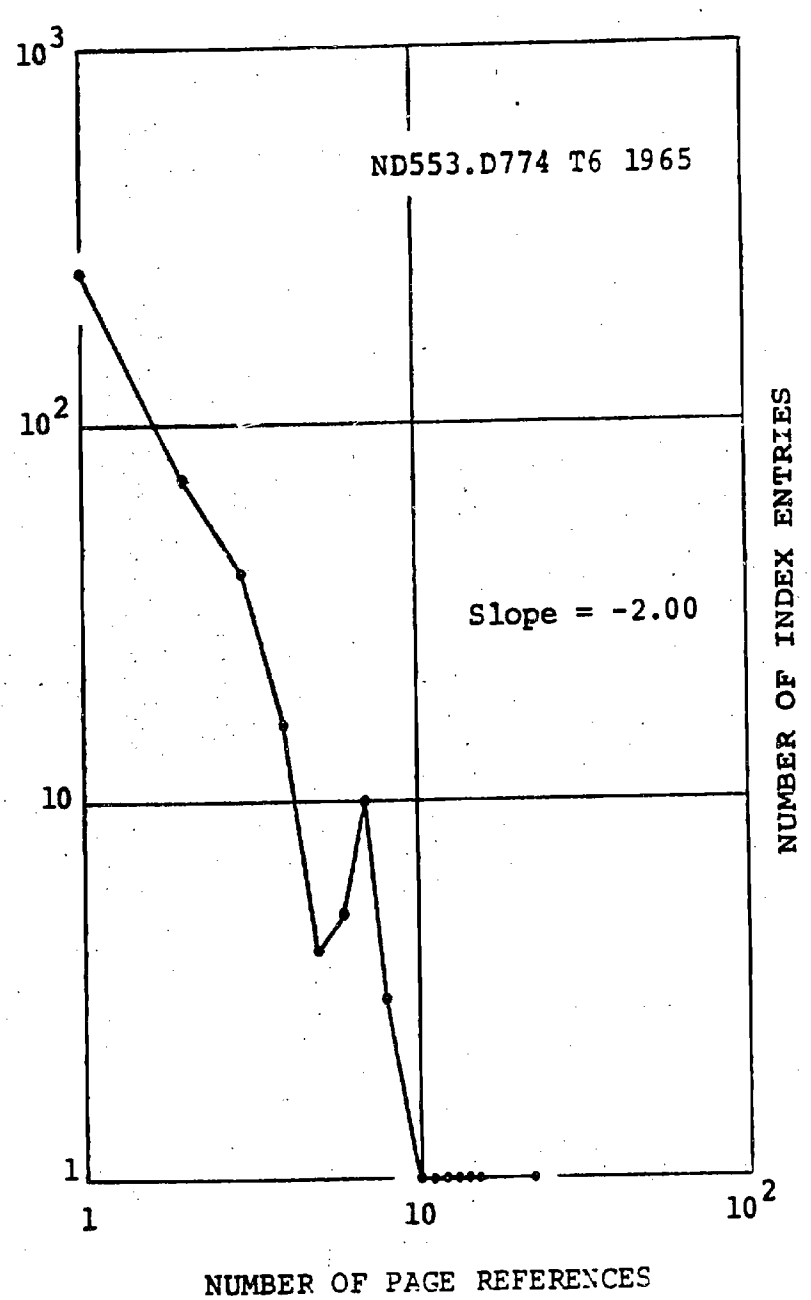
The find structure of the individual indexes can presumably shed more light on the situation. For these purposes, we have selected a random sub-sample of 28 indexes from the main Fondren Index Sample. For each of these indexes we have determined the distribution of the number of entries with one, two, three...page locations per entry. This distribution is comparable to the "frequency of frequencies" problem discussed extensively by Zipf, Bradford, Mandelbrot, et al (see Chapter III). Were the index an extractive index (i.e., one that is derived by extracting sequences of words from the text and inserting these sequences without change in the index) and were the page locations explicitly tied to the position on the page so that multiple occurrences of the entry on a single page would occur multiply in the index, then it might be anticipated that the text location distribution of index entries would be Zipf-Mandelbrot distribution which would arise from the phrases which are the index entries in the same way as the usual Zipf distribution arises from text word occurrences.

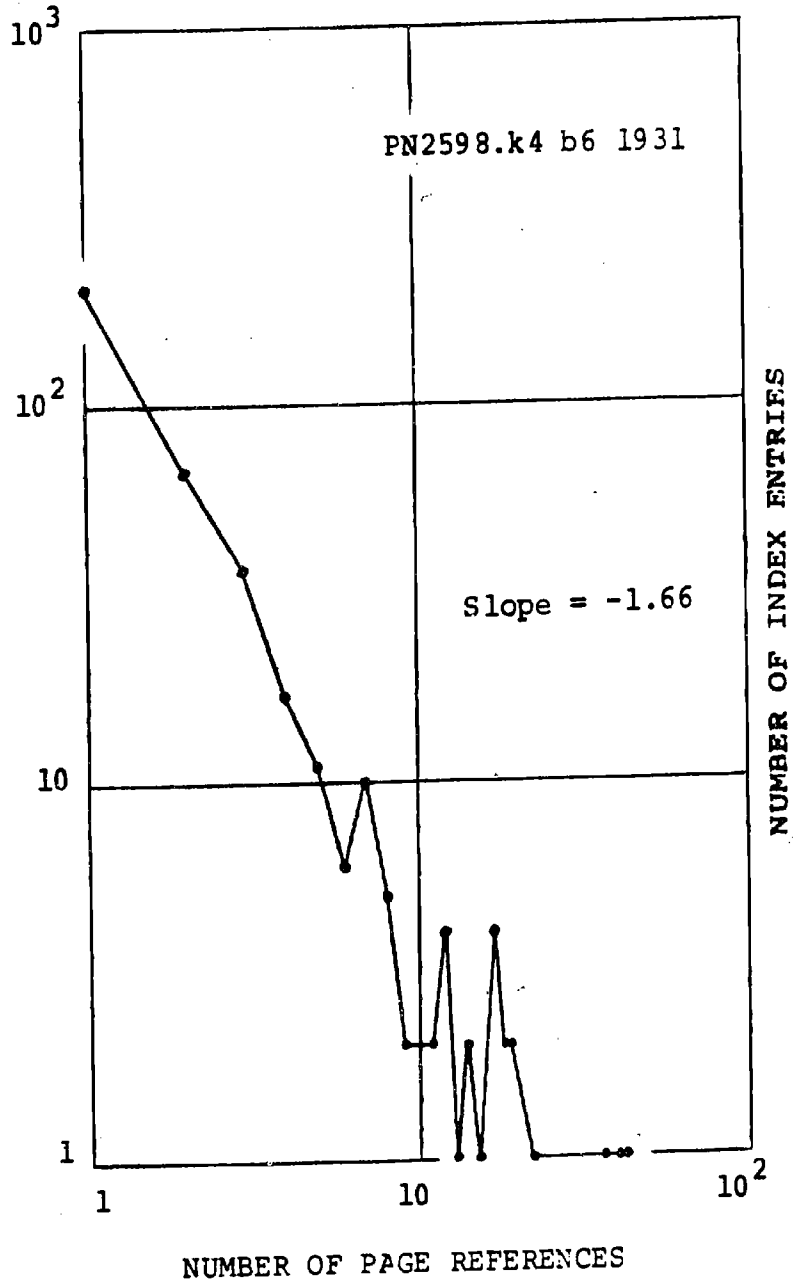
However, indexing practice normally requires a set of sophisticated transformations from the running text to the index and also reduces multiple entries on a page to a single page location. Further, not all "phrases" are indexed and it would appear that those which are left out are among both the most frequently occurring and least frequently occurring. Nevertheless, it seems reasonable to approach the problem at the first order of approximation by assuming a model of the Zipf-Bradford-Mandelbrot type; i.e., by examining the form of the distribution on log-log graph paper. This has been done for all 35 of the sample indexes, all 28 of which are presented here (Figures 4.2). (The remaining graphs appear in Appendix II.) The plots are given in the converse form to that used by Zipf in order to provide the converse form to that used by Zipf in order to provide stability (see Kendall (3)). Thus the largest point on the graph represents the number of index entries with single page references rather than the number of page references for the most frequently referenced item.

Two graphs shown are typical for the sample as a whole. In almost every case a straight line provides a reasonable approximation, with slopes ranging from roughly -1.1 to 05.5. Thus the Zipf-Mandelbrot approximation holds well for index location frequency distributions. The importance of the slope as a parameter of index measurement can be seen by recalling the Mandelbrot formulation which maximized the expected information per unit effort; the reader may find it useful to compare e.g. (3.5) ff:

Figure 4.2

Number of Index Entries vs. Number of Page References





$$I = \frac{-\sum p(x) \log n}{\sum p(x) \log x} \qquad (4.1)$$

The function that maximizes this ratio is the Zipf-Mandelbrot distribution:

$$p(x) = c\, x^{-s} \qquad (4.2)$$

Substitution of (4.2) into (4.1) yields

$$I = \frac{-\sum (\log c - s \log x)\, cx^{-s}}{\sum cx^{-s} \log x} = \frac{\Sigma x^{-s} \log \Sigma x^{-s}}{\Sigma x^{-s} \log x} \qquad (4.3)$$

where all logarithms are to the base e and the summations extend from 1 to the maximum number of page references per index entries.

For s greater than one, the summations all converge to functions of the Riemann zeta function as the maximum number of page references per index entry increases. Hence, with the sums running overall positive integers,

$$I = s - \frac{\zeta(s) \log \zeta(s)}{\zeta'(s)} \qquad (4.4)$$

An s increases, the ratio on the right, in turn, converges to $(\log 2)^{-1} = 1.443$ so that a first order approximation to Mandelbrot information for the Zipf-Mandelbrot form is given by

$$I = s + 1.443$$

For s greater than or equal to 3, the error is less than 10%. In other words, to a first order approximation, Mandelbrot's measure of information per effort is directly proportional to s, the negative value of the slope of the approximating straight line on log-log paper.

For data that perfectly fits the Zipf-Mandelbrot model, the parameter s can be determined from the relation:

$$s = \frac{\log (\text{number of references with single page locations})}{\log (\text{number of page locations of most popular index entry})}$$

clearly, the greater the number of single page location index entries and the fewer the number of multiple page location index entries, the greater the estimate of s and hence the greater the amount of information per effort under Mandelbrot's definition. In the extreme case, where each index entry refers to one, and only one, page location, Mandelbrot information is infinite. Although we have found so such indexes in the Fondren sample, it is well to note that dictionaries take this form: each main entry occurs once and the referent information is conveniently packaged with the main entry itself rather than through a page location to some other source.

The values of s for each index in the subsample are listed in Table 4.4 in decreasing order of s. Earlier in this chapter we organized the various monographs by LC class and then by total number of entries per monograph. Under this measure the LC classes fell into three disjoint sets corresponding roughly to the descriptive materials, the technique materials, and the arts. The average slope for each of these three groups are respectively, 2.83, 2.37, and 2.79. The differences between the means are not only insignificant statistically; they do not even provide a corresponding ordering, were they significant. Thus the slope (and hence Mandelbrot's measure of average information per average effort) provides an independent measure of the index.

The 28 values of s are plotted in Figure 4.3 on log-normal paper. The distribution of values is reasonably approximated by a straight line as might be expected since as we have now shown, s is a normalized measure of information.

However, except for specialized indexes such as dictionaries, multiply occuring entries do occur, thus depressing the information ratio. For the sample plotted in Figure 4.3, the average value of s is 2.66, quite close to the natural constant, e, which is 2.718. As these multiply occurring entries do reduce the information ratio by increasing the effort required, it is appropriate to inquire as to what role they play in the index.

Some hint as to the nature of this phenomenon can be obtained by examining the role of the multiply occurring entries in the context that Zipf first studied them;

Table 4.4

ZIPF-MANDELBROT EXPONENT
FOR INDEX LOCATION DISTRIBUTION

LC Number	S
PT7244	5.47
QD9	4.43
HB199	4.33
BV2532	3.81
DA690	3.54
TK153	3.53
E741	3.39
Q391	3.20
QA303	3.06
E178	2.81
QL703	2.71
RM721	2.71
BF181	2.69
DF521	2.64
Z5782	2.49
ND553	2.26
HM66	2.15
F864	2.08
PR2831	1.96
LB875	1.94
LC191	1.93
HF2046	1.86
D443	1.81
HD20	1.71
PR5588	1.69
PN2598	1.67
DS423	1.43
JA84	1.09

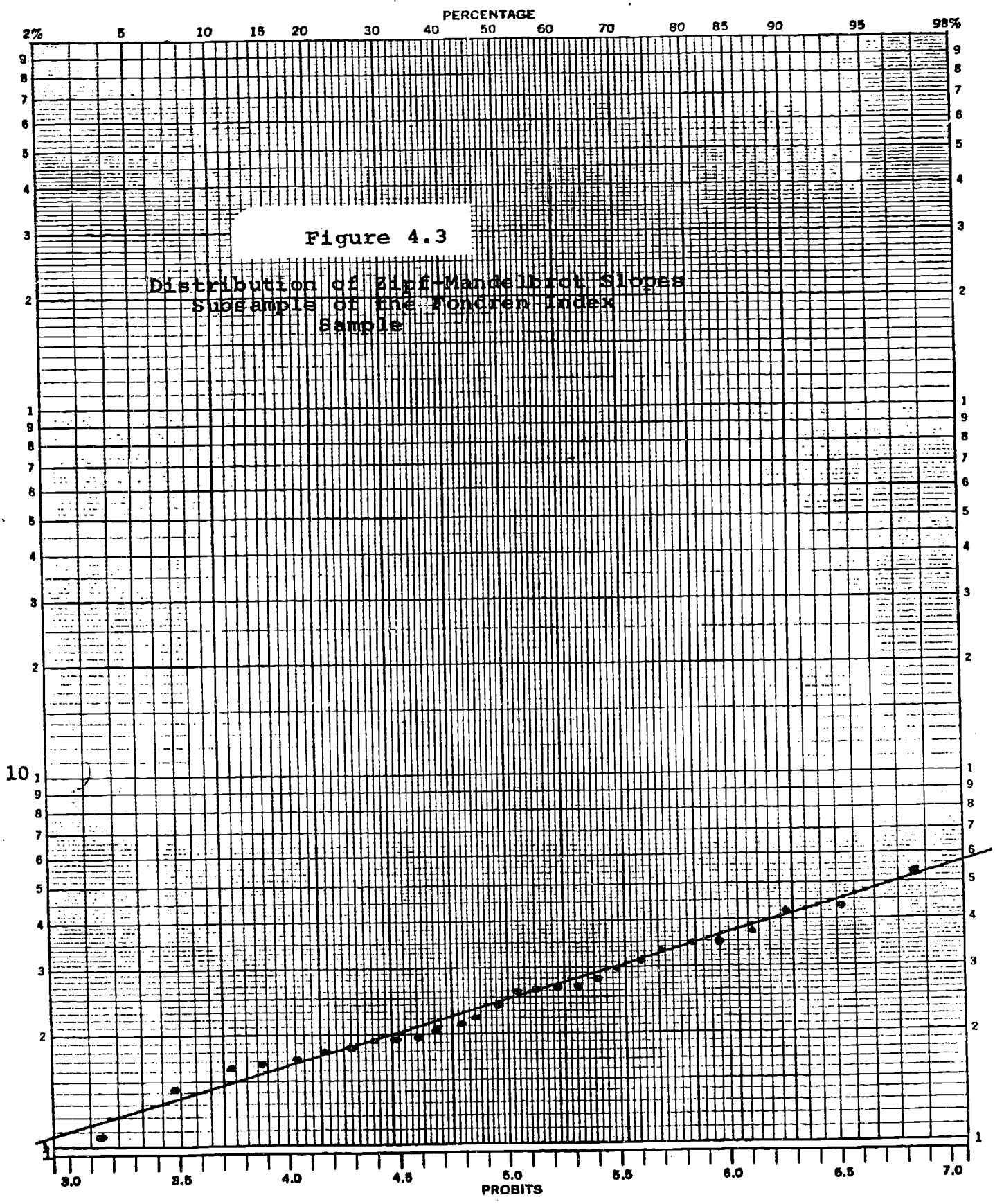


Figure 4.3

Distribution of Zipf-Mandelbrot Slopes
Subsample of the Fondren Index
Sample

in natural language itself. Even a cursory examination of a frequency ordered word list such as those prepared by Thorndike and Lorge (4) and Kucera, et al, (5) is sufficient to show that the most frequently occurring entries are the structure words (i.e. words with parts of speech other than noun, verb, adjective, and adverb). Such words provide the structure in which the information is embedded, but do not, at least in the broad sense, contain information themselves. Except for the rare case (e.g. in the use of certain prepositions in mathematical treatises) such words almost never occur in first position in an index entry.

In this context, it seems natural to suggest that the index entries that occur with many page locations play a fundamentally different role from those that refer only to one or a few page locations. Roughly speaking, we might say that the multiply occurring entries carry the semantic structure in much the same way that the multiply occurring words carry the syntactic structure. Suppose, for instance, that the term California appears in an index with, say, 15 page locations. It would seem reasonable to conclude, even with no other information about the accompanying text, that the text is very much concerned with California in a global manner. Reference to each of the various page locations would presumably uncover a variety of bits of information about California and in this particular sense, we could say that California was one of the "subjects" discussed in the book. If on the other hand, we were to find another book, say on population statistics, whose index contained a single page location for California, it would seem appropriate to conclude that California was one of many items discussed in the text rather than a main subject of the text.

In short, if one is interested in "population statistics for the state of California" one can either go to a book on population statistics and look in the index for California, or one can go to a book on California and look in the index for population statistics. For obvious reasons both types of information packaging exist and access to the packaged information is generally, though not always, provided both ways: by subject to allow the user to get to the proper book, and by index entry to allow the user to obtain the specific fact once he has gotten to the proper book.

The multiply occurring entries thus provide a sort of transition from the "specific fact" aspect of the problem to the "general subject" aspect of the problem. They provide the basis for an algorithmic identification of

the semantic structure in the same way that the structure words provide a basis for the algorithmic identification of the author's syntactic style. (See Mostellor and Wallace (6))

For both the word frequency distribution and the index page location distributions, there is no clear break between the set of frequently occurring items and the set of non-frequently occurring items. However, the previously developed arguments on the access level structure provide a technique for establishing break points in the distribution: the set of most frequently occurring entries can be defined as 1/900th of the whole set of entries. This has been done for the subsample of indexes from the Fondren sample. The results are tabulated together with the LC class, the LC subject headings, and the title in Table 4.5.

Looking first at the subject heading and title information in Table 4.5, it is clear that approximately two-thirds of the subject headings are direct transformations (through the subject heading authority list) of the title information. This observation, of course, sheds considerable insight into the discussion of the utility of permuted title indexes: anything as cheap as a permuted title listing that can supply in the order of two-thirds of the subject heading information automatically is clearly useful. At the same time a device that misses one-third of the potential information is clearly not sufficient.

In this context the role of the multiply occurring index entries becomes more obvious: most of LC subject headings that are not derivable from the title information are derivable from the multiply occurring index entries either directly (e.g. Andalusite, U.S.A. vs. Andalusite) or at a higher level of synthesis (e.g. gaseous discharge tube + ultra violet light + reaction, reactors vs. electrical apparatus and appliances). At this stage it is not necessary to re-open the much discussed question of whether classification of documents can be obtained economically through purely algorithmic processes; other simpler problems must be solved first (e.g. the automatic derivation of the index itself). However, it is essential to obtain a clearer understanding of how the various access devices already in operation interact with one another. The preliminary results derived from Table 4.5 make it clear that there is a direct relation between the LC subject headings, the monograph titles, and the multiply occurring index entries. The utility of title derived indexes is manifest by their present use and persistence. It remains to determine the utility of

Table 4.5

COMPARISON OF HIGH-FREQUENCY INDEX ENTRIES
WITH LC SUBJECT HEADINGS & TITLES

LC Class	1/900th Entires	LC Subject Headings	Title
BF181	1. Marston, W. M. (Author) 2. Freud, Sigmund	1. Psychology, Physiological	Integrative Psychology
BV2532	1. Fallen, The	None	The History of the Foreign Missionary Society
D443	1. Great Britain mentioned	1. Europe-Politics-1914	Ten Years of War & Peace
DA690	1. Sackville, Lady Margaret (afterwards Countess of Thanet) mentioned in Lady Anne Clifford's Diary	1. Knole Park, Sevenoaks, Engl. 2. Sackville Family	Knole and the Sackvilles
DF521	1. Churches: in Constantinople 2. Frescoes	1. Byzantine Empire-Civilization	Byzantium
DS423	1. Krsna 2. Siva 3. "Bhagavad-Gita" 4. Visnu 5. Brahman 6. Guru(s)	1. India-Civilization	The Cultural Heritage of India
E178	1. Beard, Charles A. & Mary 2. Jefferson, Thomas 3. Turner, Frederick Jackson	1. U.S.-Hist.-Addresses, Essays, lectures 2. U.S.-Hist.-Historiography	Understanding the American Past

Table 4.5 (Continued)

LC Class	1/900th Entries	LC Subject Headings	Title
E741	1. Prices: agricultural 2. Foreign Relations: Anglo-American 3. Federal Income Tax: individual 4. Tax: individual income 5. Farmers, income of 6. Legislation: agricultural 7. Agricultural, legislation for 8. Railroads: rates of	1. U.S.-Hist.-20th cent.	American Epoch
F864	1. Mass	1. Ansa (sic!) Juan Bautista de 2. California-descr. and travel 3. San Francisco-Hist.	Anza's California Expedition
HB199	1. Terborgh, Gene 2. Breakeven Charts, Examples of	1. Economics	Engineering Economy
HD20	1. Charts on simulated business results	1. Operations Research 2. Industrial Management-Research	Operations Research for Industrial Management
HF2046	1. Chamberlain, J.	1. Free trade and protection --Free Trade. 2. Tariff--Gt. Brit.	The Return to Protection
HM66	1. Trade Unions	1. Sociology	Social Theory
JA84	1. Economy	1. Political Science-Hist.-Russia	Russian Political Thought

Table 4.5 (Continued)

LC Class	1/900th Entries	LC Subject Headings	Title
LB875	1. America	1. Education 2. Literature-Study and Teaching	*Two Views of Education*
LC191	1. Children, Disease of	none	Education and Social Progress
ND553	1. "Bride Stripped Bare By Her Bachelors, Even, The" (Ducamp)	1. Duchamp, Marcel, 1887- 2. Cage, John 3. Rauschenberg, Robert, 1925- 4. Tinguely, Jean, 1925-	The Bride and the Bachelors
PN2598	1. Butler, Pierce	None	Fanny Kemble
PR2831	1. Greg, Walter	1. Shakespeare, William. Romeo and Juliet 2. Shakespeare, William-Bibl.-Quartos	The Bad Quarto of Romeo and Juliet
PR5588	1. Keats, John	None	Theme and Symbol in Tennyson's Poems to 1850
PT7244	1. Bjark: Bjarkamal, anon.	1. Scalds and Scaldic Poetry 2. Icelandic and Old Norse Poetry	Den Norsk-Islandska Skaldediktningen
QA303	1. Cauchy, A. 2. Euler	1. Calculus	Vorlesungen Uber Differential und Integralrechnung

Table 4.5 (Continued)

LC Class	1/900th Entries	LC Subject Headings	Title
QD9	1. Gregory	1. Chemistry-Bibl. 2. Reference Books	Library Guide for the Chemist
QE391	1. Researves, India 2. Andalusite, U.S.A.	1. Sillimanite 2. Andalusite 3. Cyanite	Sillimanite
QL703	1. Carnivore 2. Bat(s)	1. Mammals	Principles in Mammalogy
RM721	1. Muscle Contraction	1. Gymnastics, Medical	Therapeutic Exercise
TK153	1. Tube, gaseous-discharge 2. Hertz 3. Light, ultra-violet 4. Maxwell 5. Reaction, Reactors 6. Valence electrons	1. Electrical apparatus and appliances 2. Electrons	Electrons at Work
Z5782	1. Passion 2. Comedy 3. Latin 4. Staging	1. Drama, Medieval-Bibl.	Bibliography of Medieval Drama

index entries, over and above their obvious utility in providing access to a book's contents, once the book is in hand. This question will underlie much of the discussion in the next two chapters.

Before turning to this question, however, it is useful to shed some light on how the indexer controls the multiplicities in the index and hence the value of s and the shape of the particular entries that will receive the highest numbers of page locations. Obviously, this can be done in several ways involving such delicate questions as the determination of how the indexer decides whether a particular word, or sequence of words, on a particular page should rate an entry in the index. At a simpler level, the indexer has the opportunity to reduce multiplicities by increasing the length of the entry. Thus in a work on history, the indexer can either provide a single entry for war, with a large number of multiplicities, or he can break this same set of entries down into subsets involving particular wars such as civil war, world war, etc.

That this mechanism is in fact used is easy to demonstrate. Table 4.6 provides the frequency distribution for the 27,188 index entries in a uniform random subsample of 35 indexes in the Fondren sample by word length. As might be expected, the distribution can be reasonably approximated by a log-normal distribution as shown in Figure 4.4. The arithmetic mean of this distribution is 3.68 words per index entry. Only 13.5% of the entries are one-word entries. This is somewhat larger than the 9.1% found in a smaller sample of indexes to statistical books studied by Dolby (7) but still provides strong support for the hypothesis advanced in (7) that the great bulk of the entries in back-of-the-book indexes are multi-word entries.

This observation has considerable significance for the design of automatic indexing procedures. If one-word entries constitute only 13.5% of the total index, it seems unlikely that detailed frequency studies of words will provide much insight into the problem of deriving index entries automatically. In some of the earliest work on this subject, Luhn (8) attempted to derive indexes from word frequency counts, with limited success. More recently, Damerau (9) established a procedure for deriving coordinate index terms (to be used later via machine searches) based on word frequency counts. Bloomfield's (2) study of Damerau's procedure makes it clear that coordination of the single terms derived by Damerau rarely leads to an index entry derived by humans for the same material. As we shall show in the next chapter, there is more to be gained by deliberately suppressing the one-word entries, rather than by attempting to emphasize them.

Table 4.6

Distribution of Index Entries by Word Length - Subsample of The Fondren Index Sample

Number of Words	Number of Entries	Cumulative Number	Cumulative Percentage
1	3673	3673	13.51
2	6563	10236	37.65
3	4817	15053	55.37
4	3905	18958	69.73
5	2839	21797	80.17
6	1969	23766	87.41
7	1243	25009	91.99
8	801	25810	94.93
9	516	26326	96.83
10	281	26607	97.86
>10	581	27188	100.00

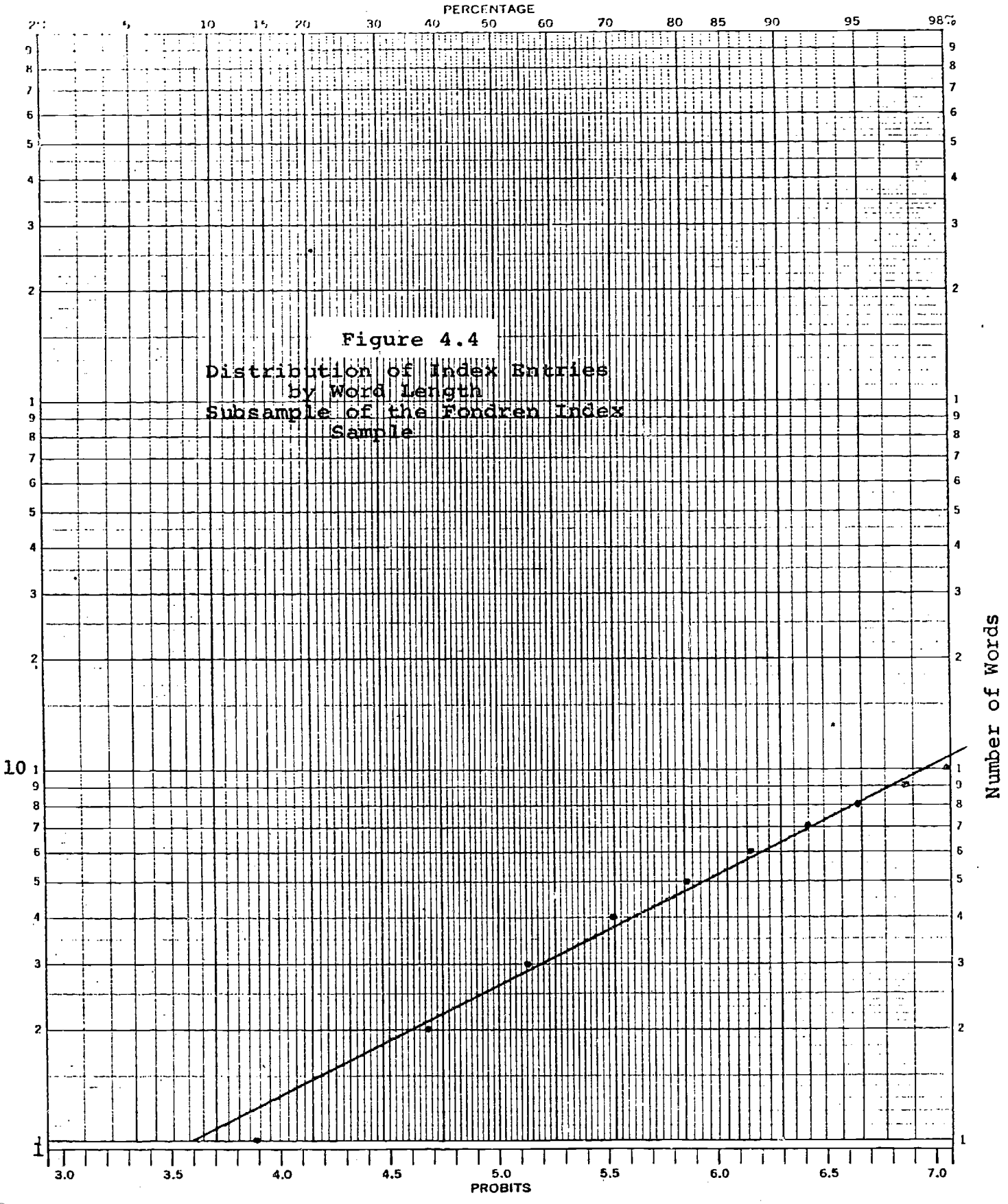


Figure 4.4

Distribution of Index Entries
by Word Length
Subsample of the Fondren Index
Sample

The observation that index entries are usually one-word entries also has some impact on a variety of questions involved with the use of indexes in agglomerated form. This will be discussed at some length in Chapter VI.

References

1. Dolby, J. L., H. L. Resnikoff, and V. Forsyth, Computerized Library Catalogs: Their Growth, Cost, and Utility, M.I.T. Press, Cambridge, 1969.

2. Bloomfield, Masse, "Evaluation of Indexing, 3. A Review of Comparative Studies of Index Sets to to Identical Citations", Special Libraries, December 1970, 554-61.

3. Kendall, M. G., "The Bibliography of Operational Research", Operational Research Quarterly, 2(1960), 31-6.

4. Thorndike, E. L., and I. Lorge, The Teacher's Word Book of 30,000 Words, Columbia University, New York, 1944.

5. Kucera, H., and W. N. Francis, Computational Analysis of Present-Day American English, Brown University Press, Providence, Rhode Island, 1967.

6. Mosteller, F., and D. Wallace, "Inference in an Authorship Problem," Journal of the American Statistical Association, 58(1963), 275-309.

7. Dolby, J. L., "The Structure of Indexing: The Distribution of Structure-Word-Free Back-of-the-Book Entries", Proceedings of the Annual Meeting of the American Society for Information Science, 5(1968), 65-72.

8. Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", IBM Journal of Research and Development, 1(1957), 309-17.

9. Damerau, F. J., "An Experiment in Automatic Indexing", American Documentation, 16(1965), 283-9.

ALGORITHMIC TEXT INDEXING

An index increases access to a particular corpus of information. Until recent times most indexes followed the text material in certain types of books. Although this may still be true today, the emphasis of research into the nature of indexing has shifted to indexes of other types of corpora, such as the permuted title index and its variants and the citation index, which index collections of document titles rather than the text of the documents. Indeed current information retrieval efforts appear to exclude consideration of back-of-the-book indexes. For instance, Salton (1) offers a brief discussion of term-oriented, or derived indexes, of which the back-of-the-book indexes are usually instances, but the applications he describes are to collections of document titles. The Encyclopedia of Linguistics, Information and Control (2) mentions only citation indexing.

This chapter is also exclusively concerned with back-of-the book indexes; hereafter the term index will be used in this restricted way.

The principal result presented here is an algorithm for the automatic construction of an index from running text in machine readable form. A preliminary version of the algorithm was implemented by hand and used to derive the index to Dolby, Forsyth, and Resnikoff (3). The version presented here has been programmed for the IBM 360/30 using a set of assembly code macros and tested on a set of 50 abstracts of statistical papers published in the Annals of Mathematical Statistics and a second set of abstracts published in Cancer Research.

The difficult question of determining what is to constitute an adequate index for a given corpus of running text is not considered here, although reference is made to an earlier study (Dolby (4)) that considered certain obvious statistical characteristics of published indexes as well as to the previous chapter.

The cost of deriving the index entries and formatting them into standard format is approximately 2¢ per line of input text, based on standard commerical rates (west coast of the United States).

Let us assume that an index is an ordered collection of word sequences (or transformations thereof) from the running text together with appropriate locator designations (e.g. page numbers). A reasonable first step in deriving such an index is to partition the text into a set of word sequences using, in this case, marks of punctuation and structure words to determine the sequence boundaries.

Each sequence is then examimed to determine whether it should be deleted from the set. In particular, sequences consisting of structure words only are deleted. For reasons that will become evident later, sequences consisting of single words and sequences that occur only once in the entire corpus are also deleted from the set.

Of the various possible transformations it is obviously desirable to identify singular and plural forms, to invert certain word sequences (at least selectively) so as to provide access to words occurring only at the end of the word sequences, and to superimpose a "see" and "see-also" facility to permit more complex transformation.

Implementation of such an algorithm requires repeated access to various lists of words and morphemes. Computer time will obviously be strongly influenced by the strategies employed to accomplish these comparisons. To cite the most obvious example, it is clearly more efficient to store the list of structure words (which is relatively small but contains many words of high occurrence frequency) rather than the list of content words which has the converse properties.

Where possible, significant gains can be made by testing for word classes rather than for individual words. Thus, it is useful to identify all participial forms as these do not generally appear as index entries. On the other hand, provision must be made to allow the override of such rules for cases of particular importance. (e.g. stratified sampling is an important statistical entry that should not be suppressed.)

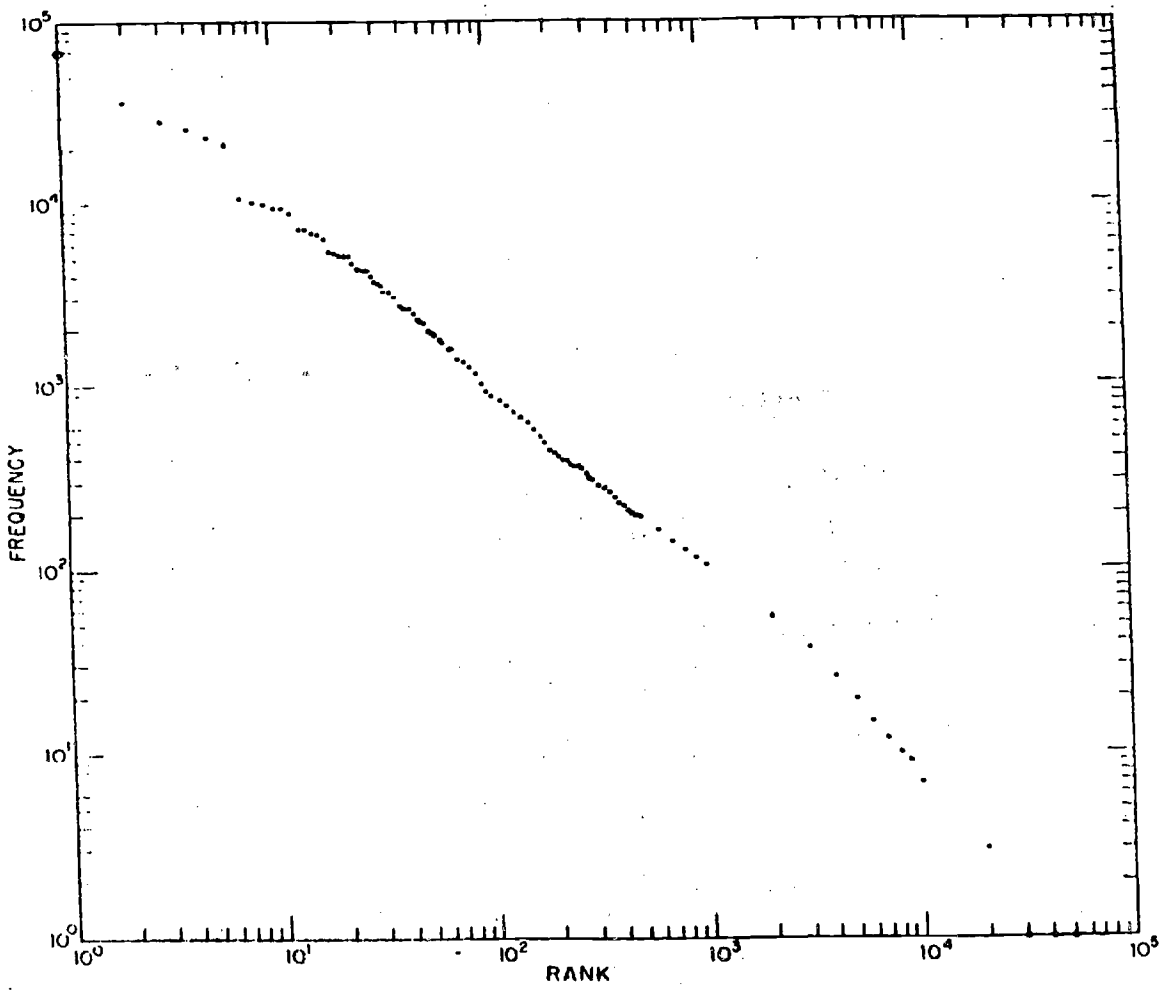
As the function of these various lists is primarily to delete words from the index, it is convenient to refer to the lists as "stop" lists and the sets of override words as "go" lists. Although sufficient testing on a wide variety of subject matter is not yet available, it would appear that the stop lists are basically independent of subject material and the go lists are subject

Figure 5.1

Word Frequency versus Rank

Brown University Standard Corpus of

American English



dependent. Thus a careful study of available authority lists in the subject field would be necessary to insure proper operation of the algorithm. (Such a study would be necessary in any event to prepare the "see" and "see-also" entries.)

Preliminary segment boundaries are established by marks of punctuation (other than the hyphen and apostrophe). Within the segments thus established, further boundaries are introduced between sequences of consecutive stop words (see Table 5.1) and non-stop words. As a simple expedient, all words in the stop list ending in s have the s removed and the match between the current word and the stop list is made after the final s (if any) has been removed from the current word. More sophisticated "plural logic" would be justified here only if the stop list were expanded substantially and in its expanded form contained a significantly larger number of "irregular" plurals.

The selection of the words to be used in the stop list provides an intriguing problem. Clearly, all structure words (neglecting archaic forms) should be included. It turns out also to be useful to include high frequency adjectives and verbs. It is therefore tempting to simply select the first n words from a rank ordered word frequency list. Unfortunately, there is no clear break in such a list in the vicinity of a reasonable cutoff (see Figure 5.1). Thus the cutoff must be made simply in terms of finding a reasonable trade off between added machine costs in testing against large lists, and added editing costs at the other end due to failure to suppress words. Based on the developments of Chapter II, we would expect the cutoff to be in the order of 1/30 of the vocabulary. The word list used here has been purposely kept short during programming and should probably be expanded by a factor of two or three in actual use.

The list organization as presently implemented is also quite simple: as the word length (in characters) of the current word is known at the time of the match, the list is broken down by word length and arranged alphabetically, within the sets of each length. Matching is done sequentially with termination on a match or when the current word is low to the list. Expansion of the lists would probably make it useful to use a hashing technique.

The next segmentation stage consists of segmenting the sequences of non-stop words into consecutive sequences of words ending in ed, ly, ing, or ful and sequences of

TABLE 5.1

SHORT LIST OF STOP WORDS
ARRANGED BY WORD LENGTH

an	own	some	three	general
at	put	such	under	improve
be	see	take	until	include
by	she	tend	usual	instead
do	the	term	where	operate
go	thi	than	which	present
ha	too	that	while	previou
he	two	them	whose	provide
hi	way	then	wider	require
if	who	upon	would	several
it	you	very	yield	similar
on		well		special
or	also	were	become	through
me	back	what	before	unknown
my	been	when	behave	without
no	both	will	better	
so	come	with	cannot	consider
to	down	work	change	original
up	each	your	chosen	possible
wa	even		denote	satisfie
we	from	about	depend	together
	give	above	derive	
all	good	admit	discus	arbitrary
and	have	after	either	different
any	here	among	extend	excellent
are	hold	begin	higher	important
but	into	could	implie	otherwise
can	just	drawn	little	
did	know	first	permit	additional
due	last	found	relate	elementary
few	lead	given	reduce	particular
for	lend	great	result	
get	like	imply	second	
had	long	known	should	
her	made	might	unique	
him	make	never	variou	
how	many	other	wherea	
let	more	refer	within	
may	most	right		
new	much	sense	against	
not	must	shown	another	
now	only	since	because	
off	over	still	between	
old	part	their	certain	
one	said	there	consist	
our	same	these	earlier	
out	show	those	further	

N.B. All one-letter words are stopped. Terminal s is removed, thus ha stops has .

words not ending in any of those four suffixes. The current go list to override this segmentation consists of only three words (family, stratified, and sampling) and is included only to insure that the facility exists in the program.

The structure words of and in are not included in the main stop list so as to allow sequences such as analysis of variance and convergence in measure to emerge as index sequences. However, it is clear that primary index entries do not include entries beginning or ending with of or in. Hence the final segmentation step is to segment beginning or ending occurrences of these to words from the non-stop, non-(ed, ly, ing, ful) word sequences.

Following a suggestion of John Tukey, we have investigated the utility of "stopping" all short words, i.e., words with fewer than n characters. Such a procedure would clearly speed up the program and set aside the difficulty of running down a number of short words that occur with sufficient frequency so as to be included in a reasonable system, (such as those occurring in Latin phrases). Based on present experience, it appears that suppressing words with fewer than four characters is reasonable. This procedure has been used in the experimental run on the 50 abstracts from Cancer Research, but not on the two earlier examples presented here.

All segments other than those consisting wholly of non-stop, non-(ed, ly, ing, ful), with beginning and ending of and in removed, are deleted. Of the segments remaining, all segments consisting of single words are also deleted. Experimentation with this step in the procedure stems from an observation made in Dolby (4) that one word entries in published indexes occur with surprisingly low frequency. Hence, the obvious strategy is to suppress all entries with exceptions rather than to pass all with exceptions.

The override to single-word suppression can take several forms. First, a go list can be appended (though none is used in the present implementation). Independence would be an obvious choice for statistical subject matter. Second, proper names, that is, words in all caps or initial caps could be used as an override. (This was done in the manually implemented version used on Ref. (3) but has not been exercised in the machine implementation.) Finally, single-word primary entries can, and do occur in the inverted entries studied below.

This reduced list of segments, or possible index entries, must now be transformed in certain obvious ways both to achieve proper compression in the final index and to provide at least the appearance of a manually prepared index. One obvious consideration

involves the problem of identifying singular and plural forms. Again, a relatively simple strategy is sufficient to take care of most of the problem. Plural forms are rarely used as modifiers and when so used are used with a high degree of consistency. Thus if least squares method occurs, it is highly unlikely that least square method will also occur (though least squares methods might well occur). Hence it is only necessary to prepare for plurals that occur at the end of the entry.

The most frequently occurring plural form is obtained by adding s to the singular form. If the final s is replaced by a code that will sort immediately after blank (but prior to a) it is possible to compare successive entries after sorting and to eliminate the final s from all entries that follow entries that are otherwise identical. The final s is then restored in all other cases. In the application to the statistical abstracts 311 of the 946 entries ended in s. Of these, 41 were stripped of the final s to provide the required identification. More sophisticated logic of the same variety could be added to handle plural forms such as processes, densities, and matrices although a quick survey of the 946 entries disclosed only four such occurrences where identification was desirable.

Another purely manipulative step that must be introduced at this stage is the generation of inverted entries to provide access to words occurring at the end of the text ordered entries. There appear to be two main forms of interest. The first, typified by analysis of variance, can be implemented by the obvious algorithm that produces variance, analysis of. A more sophisticated form could be used to suppress one or the other of the two variants. A pair of relatively short, subject dependent, stop lists would probably suffice for this purpose.

A second type of inversion, typified by mapping normal distribution into distribution, normal could either be implemented by a go list of modest proportions or by ordering the entire set of entries by last word and then inverting all sets involving a common last word of sufficiently high frequency. Neither of these alternatives have been tried at this time, though some statistics have been gathered on the behavior of statistical terms from this point of view.

In addition to the deletion of one-word entries, it is evident, when one operates on full text, that it is entirely safe, and indeed quite useful, to delete entries that occur only once in the text. Intuitively, one can argue that if a term is not mentioned at least twice (allowing for plural variants and the like)

then there is little likelihood that enough information is presented about that entry to make it worthwhile as an entry in the final index. Practically, an examination of singly occurring entries in the samples we have studied thus far makes it clear that this is a highly useful device for eliminating much of the "noise" that inevitably is present when one takes such a simple view of English syntax. Statistically, the step can be justified on the grounds that the resultant index is of the proper size (as a percent of the volume of the book indexed) when such entries are left out, but noticeably too large if they are left in.

The use of this device must be tempered by knowledge of the text. For instance, this device was not used in the index to the statistical abstracts, as it was evident that the abstracts did not possess sufficient redundancy to allow proper operation of such a mechanism within an abstract, and it seemed unwise to base the use of such a mechanism on a (not necessarily homogeneous) set of abstracts. Presumably there are certain books whose text has a very low redundancy; for these this type of deletum should not be impletmented.

The manual implementation of the algorithm on book length material (reference (3)) is shown in Figure 5.2. Two systematic departures from the general algorithm were made in implementing it: first, names of States were systematically deleted from the index; second, a list of special words for inclusion in the index was used, containing names of countries and languages. Both decisions insure uniformity of in- or ex- clusion of terms in each class without regard to the relative significance of each usage. Finally, as described in the Instructions for Use of the Index, two index terms were manually inserted: the collective Computer Languages, and the alternative World War I for the algorithmically occurring First World War.

Perhaps of greater theoretical interest than those terms that appear in the index in Figure 5.2 are those terms that were deleted by the requirement that each entry that appears in the index, except for entries having special format properties, refer to more than one location in the text. Table 2.4 lists those word sequences which were excluded from the index for this reason. Preceding some of the words are letters which describe properties of the word sequence: 'p' indicates that the sequence is a plural form of another word sequence selected by previous steps of the algorithm; the plural sequence is therefore equivalent to the singular one, and hence appears in the final index. Sequences preceded by 'i' appeared in italic type font. It appears that this font

Index

Instructions for Use of the Index

The index is the result of applying an algorithm to the text of the book; a minimal amount of (probably mechanizable) subjective human post-editing in the final two steps produced the amalgamated and reordered form that is printed below.

All word sequences that are not *printed in italics appear in the given form in the text of the book, apart from possible differences of capitalization. Terms that do not explicitly appear in the text do not appear as index terms with the exception of the collective* Computer Languages, *and the alternative* World War I *for the naturally occurring entry "First World War."*

Those readers who are experts in information retrieval and automatic indexing may be interested to know that this is a 4 percent index.

Figure 5.2

ALGORITHMIC INDEX TO REFERENCE (3)

TABLE 5.2

Excluded Index Terms Referring to One Location

abnormal parenthesization
absolute frequencies
academic staff
access capability
access point
access system
accessible estimates
p accession distributions
accessions growth rate
acquisition rates
acquisitions data
acquisition mechanisms
acquisition process
acquisition rate
acquisition schedule
acquisition shares
acquisition structure
acquisitions - GNP relation
acquisitions - GNP share equality
acquisitions budget
p acquisitions growth
p acquisitions expenditures
adequate user access
algebraic equations
algebraic expressions
alpha-numeric code
alphabetic code
approximate linearity
approximate normality
p archival collections
archival holdings
archival libraries
archival records
Assembly code programming
"assembly languages"
assignment procedure
author access
author field
author list
author name
author/title list
p authority lists
automated catalog
Auxiliary memory
average cost
average growth
average number
average record length
average time
p Baltimore County Libraries
bedroom states
biases inherent
p bibliographic descriptions
bibliographic holdings
bibliographic indications
bibliographic items
bibliographic listings
bibliographic lists
bibliographic notes
bibliographic practice
p bibliographic records
bibliographic references
bibliographical information
bibliographically incomplete
bibliography
Bibliography Field
bibliography section
book-publication depressions
Book Length
bookseller
budget dollar
budgetary requirements
business community
calculus text
call number
(Canadian)census figures
capital letters
capitalization conventions
capitalization errors
capitalization requirements
card catalog collection
card collection
card files
card space convention
card system
cards per entry
cards per title
"careful" study
case alphabet
catalog card conversion errors
p catalog cards
catalog data
p catalog files
catalog interrogations
catalog preparation
catalog productions

p catalog records
catalog trays
cataloging function
character density per page
character density per square inch
character manipulation
Circulation
circulation file
p circulation rates
civilization's material aspects

class category
class fields
i class order
"Collected Works"
Collection Breakdown
p collection sizes
collection subset
common machine
component national growth rates
composite costs
composite estimate columns
p composition costs
composition devices
composition practice
computational base
computational facilities
computational linguistics
i Computer Line Printer Type Faces
computer programs
computerization costs
Condensed
consecutive years
contents
context permissible
conversion expense
conversion problem
conversion procedure
conversion process
conversion task
copy output cards
core memory
correction capabilities
correction costs
correspondence files
cost
cost area
cost breakdowns
cost equations
cost estimates
p cost factors
p cost figures
cost function
cost increments
cost item
cost levels
cost per title
cost point
cost structure
cost picture
cost study
cost variations
costs per thousand dollars
county library automation projects
p County library systems
county school system
county system
data conversion
data files
data object structure
data objects
i date of access field
i date of order field
decimal classification system
density output
detail level
dictionary lookup procedures
document descriptions
document identification procedures
dollar equivalents
dummy entities
dynamic aspects

economic analysis
economic aspects
economic data
economic depression
economic disintegration
economic references
economic size
economic state
economic statistical data
economic strength
economic units
edition statement
educational advantages
electronic photocomposition
electronic typesetting devices
i elementary calculus
English-language sentences
English-speaking
error-checks

error-correction capability
European
executable statements
expansion ratio
"explosive" growth
exponential curve
exponential expansion
exponential function
exponential imprint date distribution
exponential library growth rates
exponential rate
faculty library committee
feedback response
field names
i fields
fields per record
file figures
file maintenance
file records
file structure
file system
financial data
financial community
financial transactions
(first) generation machines
fixed absolute growth
floating point arithmetic
follow-up correspondence
foreign language acquisitions
foreign language documents
foreign titles
Format-Dependent Errors
format capabilities
format compromise
Format control
format elements
format requirements
French-African
French-speaking
functional collection
fund name
fundamental processes affecting libraries
fundamental structure
future funding needs
geographic area
geometric decrease
global category
global check
global war
GNP-acquisitions relation
GNP at Market Prices
graph paper
graphic arts
graphic representation
Gross Domestic Product
p gross national products
Gross Personal Income
ground-level extension
growth challenge
growth periods
growth phenomenon
growth problems
p growth rates
growth statistics
hardware costs
p Harvard samples
"higher level" languages
historical events
historical significance
human costs
human readable document
identification number
illegitimate code
implementation cost
imprint data
p imprint dates
imprint date growth
imprint decade
p imprint distributions
in-depth studies
in-school access
income data
income growth
income ratios
indented lines
information base
information fields
information per inch
information per page
information run-over
input costs
input errors
input format
input methods
input program
inquiries per record
instruction per second
intercolumn space
interentry blank lines
interlibrary loan service

interlibrary loans
interword spaces
interpretive approach
i item
p item fields
item purposes
items per year
journal-to-language assignment process
p journal titles
key economic issue
key information
key library personnel
key words
keyboard conventions
p keyboard operators
keypunch equipment
labor categories
language acquisitions
language count
language expertise
Language Field
p language groups
Language information
linguistic algorithm
language shares
Latin American
p library activities
library applications
library card catalog conversion
library card catalogs
library catalog card contents
library catalog operation
p library catalogs
library characteristics
p library collections
library community
library context
library cost structure
library expenditures
library explosion
library facilities
library file operations
library files
library holdings
Library Management Tool
library market
library materials
library mechanization
Library of Congress acquisition data
Library of Congress acquisition shar
Library of Congress classification system
Library of Congress nonserial acquisitions
Library of Congress size distributio
p library operations
library personnel
library procedures
library services
library shelf lists
library structure
p library systems
librarylike activities
line printers
linear string
lines per second
linguistic biases
linguistic constructions
linguistic data-objects
linguistic exploitation
linguistic records
linguistic partitions
linguistic subpopulation shares
list structure
literate population
load requirements
location information
log graph paper
logarithmic graph paper
p logarithmic scales
Logical operations
loss rate
"lower level" languages
p machine-readable catalogs
machine-readable library catalogs
machine-readable materials
machine-readable subject authority lists
machine change
machine design
machine elements
machine inquiries
p machine languages
machine language instructions
machine methods
machine output
machine rules
machine time usage

machine use
magnetic cores
magnetic discs
magnetic drums
p magnetic tapes
main file
management tool
manipulative operations
manpower costs
manual generation
manual operations
manual strategies
manuscript form
map classification category
marginal improvements
mathematical computation
mathematical exercise
Mathematical Journal Titles
mathematics
mathematics faculty committee
mathematics journals
mean growth
mechanical errors
mechanical translation
mechanization context
methodological principle
i Misspelled words
model cost equations
monetary inflation
monograph collection
monographic letter frequencies
i month portion
multi-language manipulation procedures
multicharacter vowel string
multiple copy graphic arts quality
musical scores
national accounts statistics
national economic growth
national economy
national origins
national publications
national publishers
national statistcial data collection processes
natural languages
(new) acquisitions information
non-English words
non-numerical procedures
nonlibrary customers
nonlinear scales
nonoriental monograph acquisitions
nonpamphlet items
nonserial Fondren sample
nonserial shelf list cards
nonserial textual works
nonstationary growth periods
nonstationary intervals of library growth
non stationary time series
"normal probability paper"
normal distribution
normal probability distribution
normative measures
number field
numeric symbols
numerical computation
off-site areas
on-line input
open-stack libraries
optical character recognition equipment
optional parameters
order-of-magnitude changes
order date
order file records
(order of) magnitude cost reductions
(order of) magnitude cost variations
(order of) magnitude decisions
(order of) magnitude gains
order operation
order system
order system file
order system reports
ordinal numbers
author list
out-of-date catalog
output error signals
output list
output machines
output printers
output sheet
page counts
page design
paper costs
(paper) tape input
parallel search logic
pattern-matching facilities
pattern-valued functions
pattern primitives
per capita growth

per unit basis
percentage growth
personal author
personal authorship
p personal incomes
photo-offset reproduction
(physical) volumes per serial title
pilot study
plant expansion
political disintegration
political issues
population growth
potential control
print runs
printed copies
(printing and) binding costs
printing cost
i Printing Type Faces
private endowment funds
Private Finance
probability scale graph
process flow
x processing bibliographic records
x processing linguistic information
production costs
production economies
production processes
productivity per dollar
profound machine language level study
program errors
program routines
proper-name entries
i Proper names
proper scale compression
propositional calculus"
public acceptance
public card catalog
public catalog losses
p public libraries
public sales
public use
publication cost
i publication field
publication growth
publishing industries
punch paper tape
quality performance figures
quality point
quantal jump characteristics
quantal jumps
i random access
p random samples
record entry
recursive processes
refugee movements
relative frequencies
relative frequecy distributio
relative merit
relative performance
relative significance
relative size
reliable data
rental figure
report system
x reprogramming costs
research effort
research grants
research purposes
retrieval processes
retrieval requests
p retrospective files
p retrospective materials
run costs
salary structure
sample cards
school cooperation
scientific effort
scientific machines
scientific periodical literat
scientific publicati
scientific research
selection criteria
selection operation
selection procedure
selection processes
selection technique
semibold type faces
semilogarithmic paper
p serial publications
serials shelf list
service bureaus
set theoretic operations
share distribution
shelf list circulation file
Shelf List Statistics
shelf space
significant acquisitions - Gl
disagreement
social dislocation
social ideologies
social phenomena
social systems

sort operation
source statement language
source statement structure
special purpose bibliographies
e square inch
staff expansion requests
standard algebraic form
standard algebraic notation
standard precedence conventions
stationary growth rate
statistical correlation theory
x statistical distribution
x statistical distributions
statistical ensemble
statistical indicator
statistical relationship
statistical summary
statistical uniformities
status indicators
storage media
storage space
string contents
string processor
i structure
subject-oriented bibliographies
Subject-Title catalog
x subject area
x subject areas
x subject bibliographies
x subject bibliography
subject catalog volume
subject classes
subject coverage
subject definition
subject designation
subject material
subject matter
subject volumes
subject words
suburban population
summary information
supervision costs
symbol strings
Symbolic Expressions
systematic way
tape costs
technological advances
telephone companies
p telephone directories
text samples
textual works
"third generation" computers
time advantage
(time and) motion studies
time benefits
time constraints
Time Field
time information
time scale
time schedule
time variation
title-word access
title-word information
x title card
x Title cards
title indices
title languages
transcription errors
transliteration schemes
tray contents
trend curve
turn-around times
type face catalog
type face size
p type faces
type fonts
type size
type styles
undergraduate library
undergraduate student
Union catalogs
x unit cost
x unit costs
p university libraries
university library book catalog
university library systems
university order staff
university rate
usage rates
user-library complex
"user codes"
User cost
(user) cost factors
i utility
utilization costs
vertical scale
e XYZ Library
yield rate per item

does not characterize indexible sequences. 'x' indicates that a manual error has been made; in some cases a verb gerund has not been deleted in the stop list step of the algorithm, so a sequence appears in the later stages of the algorithm when it ought to have been deleted at the first stage. For example, the sequence processing bibliographic records contains the structural stop sequence -ing indicating the gerund form; exclusion of this word at the stop list stage would have left the subsequence bibliographic records for consideration, which appears in the index anyway because it occurred in more than one additional location. The indicator 'e' means that the sequence has been excluded by the human posteditor. Two such sequences are noted: square inch, which should perhaps inhabit the stop list, and XYZ Library, which must be considered because one of the special format inclusion conditions is that sequences containing all capitalized words are indexed regardless of the number of text locations to which they refer; but this instance doesn't supply any useful information. It is a stylistic curiosity. Finally, certain sequences in the table are preceded by a parenthesized word. For instance, (first) generation machines appears. The algorithm generated generation machines; the preceding text word was included in the list to help the reader to understand the context of the sequence, which, following the algorithm, was excluded from the index.

Quantitatively, this algorithmic index is not significantly different from the manually produced indexes analyzed in Chapter IV. The gross size of the index is 5 pages as compared to 157 pages of text, a text to index ratio of 31.4 to 1. The index entry length distribution is given in Table 5.3.

Table 5.3

Index Entry Length Distribution
Computerized Library Catalogs

Number of Words	Frequency	Cumulative Frequency
1	82	82
2	190	272
3	44	316
4	13	329
5	6	335
6	4	339
7	1	340

The percentage of one-word entries (24%) is higher than the average number of one-word entries in the subsample from the Fondren Index Sample (13%). Although this is not a significant deviation (more than 17% of the indexes in the subsample had more than 24% one-word entries) it is worthy of some comment: the basic algorithm suppresses one-word entries, with exceptions. In this case the exception rule was to include capitalized one word entries. Thus, even though the algorithm is designed to operate against one-word entries, the proportion occurring is still on the high side.

The distribution of entries by number of words is shown in Figure 5.3. The distribution is reasonably approximated by the lognormal distribution. The arithmetic mean of the distribution is 2.08 words per entry, compared to 3.68 words per entry for the subsample as a whole. Although there is again some cause to question whether this is a significant deviation, there is an underlying weakness in the form of the algorithm as it was used in this example. The algorithm excludes entries of the word X of Y. In (4) the structure-word-free entries were found to have a mean number of words per entry of 2.12, almost exactly the average found for this algorithmic index. However, the structure-word-free entries of (4) made up only 55% of the total number of entries. In Chapter VI we shall return to this question in analyzing the output of the basic algorithm where the capability to generate entries of the form X of Y has been included.

The absence of structure-word entries also tends to depress the overall size of the index. Although the bulk size, measured in pages is approximately 1/30th of the text size (as would be expected), the ratio of bulk of the index to bulk of the text measured in number of characters is approximately half of this figure. (Not only are the index entries somewhat shorter than would be found in the manual indexes, the text density is approximately 3,150 characters per page as compared to the mean of 2,400 characters per page.)

The lack of structure-word entries also tends to distort the page location distribution, (Table 5.4).

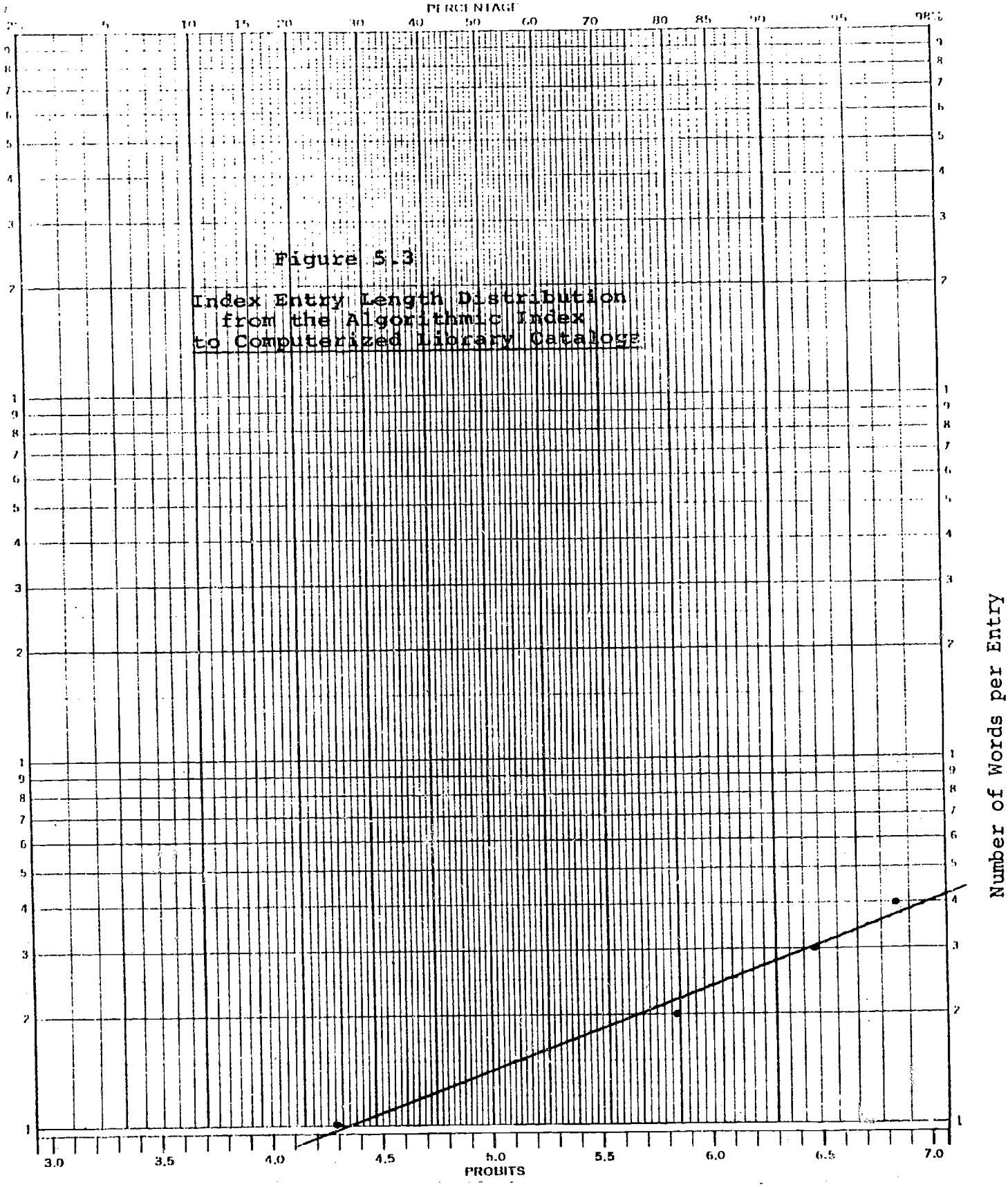


Figure 5.3

Index Entry Length Distribution
from the Algorithmic Index
to Computerized Library Catalogs

Table 5.4

INDEX PAGE LOCATION DISTRIBUTION
COMPUTERIZED LIBRARY CATALOGS

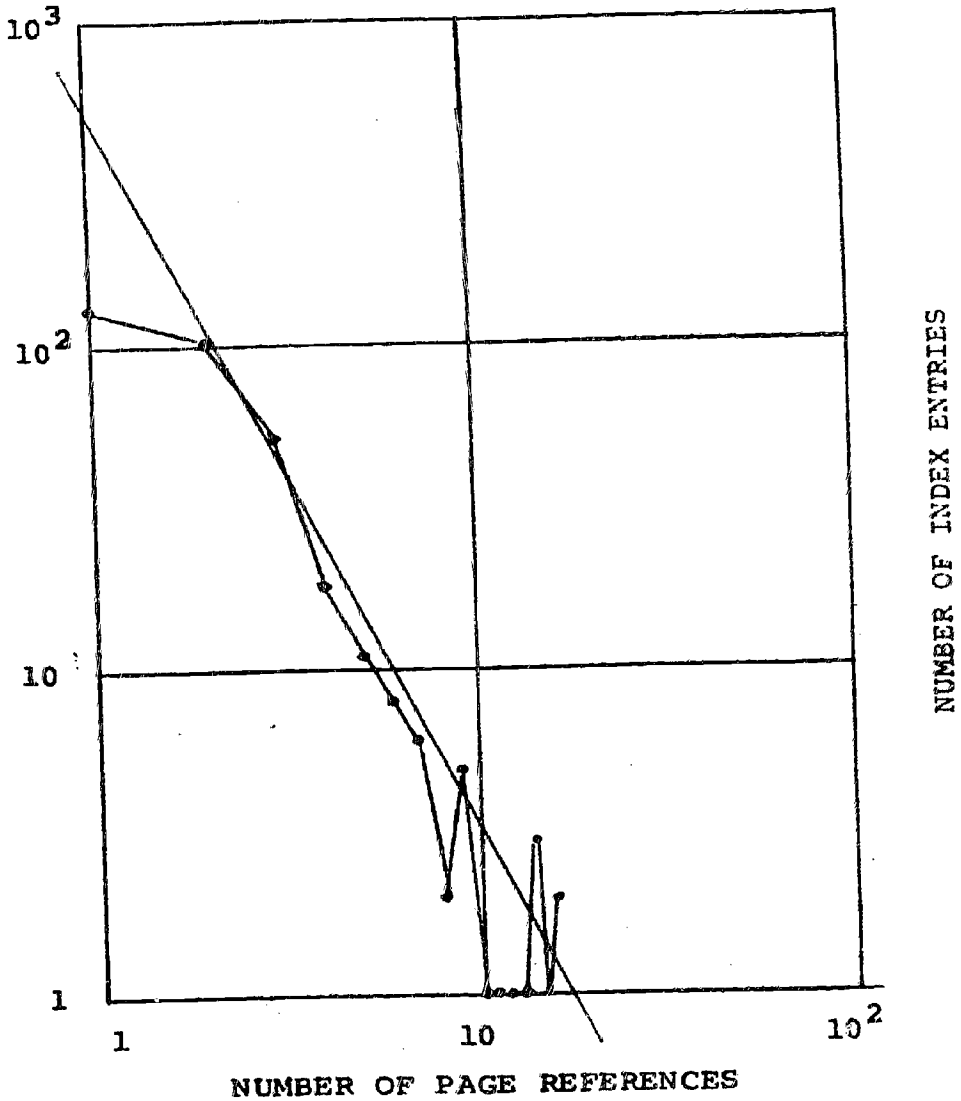
Number of Page Locations Per Entry	Frequency	Cumulative Frequency
1	127	127
2	102	229
3	52	281
4	18	299
5	11	310
6	8	318
7	6	324
8	2	326
9	5	331
10	1	332
11	1	333
12	1	334
13	1	335
14	3	338
15	0	338
16	2	340

The graph of the index page location distribution is shown in Figure 5.4. Here it is evident that the number of entry with but a single page location is significantly lower than the overall trend line for the rest of the data. Further, the bend in the data occasioned by this low value is sharper than for any of the distributions in the subsample from the Fondren Index Sample (see Appendix II). Interconnection of the entries with structure words would clearly tend to break apart entries presently agglomerated, thus reducing the number of multiply occurring entries. Ignoring the low number of singly occurring entries, the Zipf-Mandelbrot slope is 2.17, well within the range of values found for the manually produced indexes.

The arithmetic mean of the number of page locations per entry is 3.19, nearly double the figure found for the subsample of the Fondren Index Sample. However, this value is distorted by the fact that consecutive page locations were not agglomerated into single locations as is normally done in manual indexing. When this factor is corrected, the average number of page locations per entry becomes 2.14. As this value would be further reduced by inclusion of structure-word entries, it would appear that this variation is not at all significant.

Figure 5.4

Index Page Location Distribution
from the Index to
Computerized Library Catalogs



In sum, aside from the failure to include structure-word entries or to agglomerate consecutive page locations, the statistical shape of the algorithmic index to Computerized Library Catalogs appears sound. This is not to say that the index is entirely comparable to a manually produced index. However, the first requirement in automating a process traditionally done manually is to meet the basic size constraints. Further developments in the technique will be illustrated in the next chapter to demonstrate that even closer approximations are possible.

References

1. Salton, Gerard, Automatic Information Organization and Retrieval, McGraw-Hill Book Co., New York, 1968.

2. Meetham, A. R. and R. A. Hudson, editors, Encyclopaedia of Linguistics, Information and Control, Pergamon Press, Oxford, 1969.

3. Dolby, J. L., V. Forsyth, and H. L. Renikoff, Computerized Library Catalogs: Their Growth, Cost and Utility, the M. .T. Press, Cambridge, 1969.

4. Dolby, J. L., "The Structure of Indexing: the Distribution of Structure-Word-Free Back-of-the-Book Entries", Proceedings of the American Society of Information Science, 5 (1968), 65-72.

5. Dolby, J. L. and W. E. Houchin, A Modular Suite of Programs for System ABC, R & D Consultants Co., Los Altos, California, 1969.

6. Dolby, J. L., W. E. Houchin, H. L. Resnikoff, and Roger Stark, Non-Numeric Programming Language Studies: ALTEXT II., Final Report to the U. S. Air Force Office of Scientific Research, Contract #F44620-69-C-0094, R & D Consultants Co., Los Altos, California, 1970.

AMALGAMATIVE ACCESS MECHANISMS

INTRODUCTION

The model proposed in Chapter 2 shows that the search for access mechanisms must be conducted in compressive powers of 30. It is principally the relative size of an access mechanism that determines its utility. That a compression of 30 must be effected in order to move from one access level to the next, and that the boundary between access levels corresponds to compression of about a factor of 5 implies that there cannot be very many possible access mechanisms to a particular level of information storage. For instance, if the level to be accessed is the book, then one must ask what natural subsets of information there are in a book which constitute about one-thirtieth of it. As has already been pointed out, the average index to the average book compresses the text by a factor of 31.8, so the book index is a viable access mechanism. Studies of abstracts of papers appearing in mathematical journals show that the average complete abstract produces a compression of about 30.6, so the journal paper abstract is also a viable access mechanism. The book abstract should require about $276.6/(2e)^2 = 9.3$ pages; we do not have reliable information about the average length of book reviews in the professional literature, but this appears to us to be a possible mean for scholarly reviews. On the other hand, the capsule reviews of popular books that appear in newspapers and other popular media, and in some scholarly publications, are much shorter--perhaps the equivalent of one or two pages--and lie on the boundary between the levels of access mechanisms to books and access mechanisms to access mechanisms to books, the latter operating at the level of an enlarged table of contents such as regularly appeared in previous centuries, and still sometimes do, viz., Hans Zinsser's Rats, Lice and History's table of contents from which we extract the following:

I. In the nature of an explanation and an apology

II. Being a discussion of the relationship between science and art

III. Leading up to the definition of bacteria and other parasites, and digressing briefly into the question of the origin of life

IV. On parasitism in general, and on the necessity of considering the changing nature of infectious diseases in the historical study of peidemics

V. Being a continuation of Chapter IV, but dealing more particularly with so-called new diseases and with some that have disappeared.

and so forth.

Another way of looking at the problem of discovering possible methods for accessing books is this: the number of characters in a book is about $(2e)^8$; reduction of a factor of $(2e)^2$ leads to an information store about the size of the index; further reduction by a factor of $(2e)^2$ to the next access level leads to a store of the size of the table of contents. Another reduction by the same factor produces $(2e)^2 \cong 30$ characters, which is nearly the size of a book title, as we have determined in a preliminary fashion from a small uniform subsample of the Fondren Sample. In fact, that estimate was 34.2 characters for monographs in the sample regardless of language of title; had the subsample been restricted to English language titles, the average length would have been shorter. A final usable reduction is effected by another division by $(2e)^2$, leading to a one character access mechanism such as that provided by the Library of Congress one letter class designation.

The important point is that <u>every access level is filled</u>. Further study of possible new access mechanisms must therefore be constrained to access mechanisms of the same size as those that already exist. A natural question that arises is whether it is desirable to have two access mechanisms of the same size for a particular information system. That such duplication does already exist is easy to demonstrate:

1. The Author, Title, and Shelf orderings or a library card catalog are all essentially of the same size: roughly, one card image for each title in the collection. (The subject heading ordering is generally slightly larger, but still at the same access level as the others.)

2. The table of contents for a book is at the same access level as the catalog record.

3. Abstracts to journal articles appear in abstract journals as well as the index entries that are frequently published at the end of the year in the journal. Both of these access mechanisms are first order devices.

and of course other examples involving titles, descriptors, etc. can easily be found.

Thus the size of an access mechanism, though it is of first importance in describing the nature of the access it provides, is not sufficient to completely describe its characteristics. A second consideration that must be taken into account is easily illustrated by considering the sequences:

Article, Abstract, Title

and

Book, Index, Table of Contents

In the first sequence, each access device is acting simply to compress the contents of the primary information store. In the second sequence, each access mechanism is itself a set of lower order access mechanisms collected and sorted in a useful ordering. The abstract and the title provide the user with the opportunity to determine whether the document so described is likely to be relevant to his need for information, in a general way. The index and the table of contents provide the user with information about the contents of the document together with the location of particular pieces of information in the document.

The crucial question is that of agglomeration: an index is an agglomeration of entries; a table of contents is an agglomeration of entries; on the other hand both title and the abstract are entities themselves rather than being agglomerations of other entities. It seems clear from what has gone before that the minimal unit for

agglomeration is the first level unit (about 30 characters). Thus both the table of contents and the index are agglomerations of first level units. However, higher level agglomerations exist: the abstract journal is an agglomeration of second level units, as is a publication devoted to the republication of the tables of contents of journals. Although we have not yet completed our study of dictionaries and encylopedias, it is clear that each of these important access devices are agglomerations of higher level entities.

In this sense, an access mechanism can be described first by its total size and secondly by the size of the primary entries that it agglomerates. Thus an abstract is zero level agglomeration of second level entries; a table of contents is a first level agglomeration of first level entries; and an index (to a book) is a second level agglomeration of first level entries.

There are at least two other factors that must be taken into account: a cumulative index to a series of books on statistics obviously plays a different role than the index to an encyclopedia even though both are third level agglomerations of first level entries. The difference here is that the encyclopedia is itself an agglomeration of second or higher level access mechanisms, while the books are primary information stores. The difference in these two mechanisms would almost undoubtedly show up in the slope (in the Mandelbrot sense discussed earlier) of the index.

Finally, there are access mechanisms clearly dedicated to "non-subject" access, e.g., author indexes, list of publications by publisher, place of publication, time of publication, etc. which play a major role in library access systems.

Consider a collection of titles--such as book titles--of items which compass a range of subject matter. The card catalog title list is one ordering of such a collection. If the collection is reordered to bring together all titles which contain a given information bearing word, then access to the collection is significantly increased.

Studies of such access mechanisms have been underway for some time, although none of them are generally available. One of the most advanced title access mechanisms is that prepared at Princeton University under the direction of J. W. Tukey; it is a sophisticated permuted title index consisting of more than 25,000 titles of journal papers in the field of statistics. Since the average length of a paper in mathematics is about 13.8 (normalized) pages,

a title represents a compression of about two access levels, for the title as it appears in a permuted title index carries information about the journal and author as well, requiring about 130 characters. A sample page from the Princeton permuted title index is shown in Figure 6.1.

General considerations suggest that a permuted title list of book titles for the Library of Congress letter class subcollections of archival libraries would be a useful tool, and one which would be readily obtainable as a byproduct of the existence of a machinable catalog data base.

Another type of amalgamative access mechanism, which provides access to a collection of items belonging to the same access level rather than to only one item can be constructed by performing the process normally used to construct a standard access mechanism on the output produced by another. For instance, we have studied the utility of indexing abstracts to journal papers in the statistical literature. The abstracts are normally provided with the papers; they have been converted to machinable form and an elementary version of the indexing algorithm described in Chapter 5 was applied to them. Appendix A5 exhibits the abstracts to 50 papers, the associated abstract indexes produced by application of the algorithm, and a cumulative list of the resulting index terms with references to the articles in which the terms appeared. We reproduce an abstract with its index as Figure 6.2 and a page from the cumulative abstract index as Figure 6.3. The abstract index was the first processed in this series; it is perhaps not entirely typical of the output from the algorithm. We have also processed the same data using a variant of the algorithm which ignores in its analysis stage the presence of the preposition "of" and consequently will produce index entries like "basic limit theorem of renewal theory" which appears in Figure 6.2 only by way of its constituent phrases "basic limit theorem" and "renewal theory".

An index to an abstract is a hybrid form of access mechanism. The abstract already contains a large proportion of significant phrases which are repeated in the extractive output of the indexing algorithm. There is therefore no hope that an index to an abstract can provide a compression of a full factor of 30 that would be necessary to descend from one access level to that immediately below it. In fact it appears that indexing abstracts leads to a compression of about 15; since this is significantly greater than (2e), such a procedure does

Figure 6.1
Permuted Title Index Page (Left Hand Side)

57SMH	28AMSX	247	A NOTE ON TRUNCATION AND
64DNY	35AMSX	1229	ON CONTINUOUS
58WJN	29AMSX	1028	INCOMPLETE
63ZHV	8TPAS	218	MINIMAL
62KPA	14AISM	63	K VARIABLES.
60BRR	31AMSX	232	TE BLOCK DESIGN WITH TWO ASSOCIATE / MINIMAL
58KTA	36GIIS	K 26	INING A SUFFICIENT SUBSET IS NOT NECESSARILY
57KTA	7BMSX	92	FERENCE ASSOCIATED WITH AN ADDITIVE FAMILY OF
58BSU	20SNKA	223	FERENCE ASSOCIATED WITH AN ADDITIVE FAMILY OF
400LY	11AMSX	104	ON STATISTICS INDEPENDENT OF
62BRR	33AMSX	596	TWO PROPERTIES OF
49HLS	20AMSX	225	ON THE ORDER STRUCTURE OF THE SET OF
61CCS	32AMSX	904	THE RADON-NIKODYM THEOREM TO THE THEORY OF
38NMN	6GIIA		F THE DISTRIBUTION OF THE TRUNCATED POISSON
38NMN	6STTA2TCCS		TEOREMA CONCERNENTE LE COSIDETTE STATISTICHE
38KZZ	6STTA2CNSC		TEOREMA CONCERNENTE LE COSIDETTE STATISTICHE
35CCE	27BMTA	SPFS	SUR LES CONDITIONS NECESSAIRES ET
63PWS	12JASB	393	A SECOND PIEBALD FAMILY FROM
63FDR	77TBAC	SSPA	LS STUDIES INVOLVING THIRTEEN CHARACTERS IN
56FDR	55HPRD	177	DURES APPLIED TO CHEMICAL GENETIC DATA FROM
50LLE	113JRSA	531	METHOD FOR EVALUATING GENETIC PROGRESS IN A
63SRA	15JIAS	185	THE
53CHY	43JASE	FRSM	TRANSFORMATION FOR ANALYSIS/ DISTRIBUTION OF
46DVS	203TBAM	ETAL	OF FRACTIONAL REPLICATION IN AN EXPERIMENT OF
54FDR	528U52	BASS	/OGEN. PHOSPHORUS AND POTASH ON THE GROWTH OF
57SRA	9JIAS	52	A BIVARIATE ANALYSIS FOR SNEDECOR"S
50DLR	4SITA1	3	NUMBER OF RED ROT LESIONS ON THE MID-RIB OF
51DML	33BIIS	M289	STATISTIEK IN
62BRR	39BIIS	K301	SOME CHARACTERISTIC ASPECTS OF A
13GRD	9BMTA	69	CATIONS INDUSTRIELLES DE LA STATISTIQUE EN
63GRK	58JASA*	728	/RORS OF RANDOM SAMPLING IN CERTAIN CASES NOT
60BCK	22JRSB	302	A QUICK TEST FOR SERIAL CORRELATION
06PRN	5BMTA	172	MATION OF MISSING VALUES IN MULTIVARIATE DATA
02GLN	1BMTA	385	OF A POPUL/ ON THE CURVES WHICH ARE MOST
57LHA	28AMSX	126	0 PRIZES. THE MOST
56LKS	3S2BMSP	195	THE NORMAL DISTRIBUTION FROM PROPERTIES OF
63KLF	34AMSX	1419	RACTERIZATION OF POPULATIONS BY PROPERTIES OF
60BAS	9PISP	289	THE EXPONENTIAL DISTRIBUTION ON THE BASIS OF
62ZTK	39BIIS	J329	NOMBRES ALEATOIRES.
60BRS	9PISP	335	IE DES FILES D" ATTENTE. CONVERGENCE DES
49KLY	3ASTN	M 11	CALCUL D"UNE INTEGRALE AU MOYEN DE LA
51GTL	7BMTX	171	TENT OF TROUT BLOOD. THE EFFECT OF
62JKO	9RSAR	69	MEASURING THE STOCK OF BULK AMMONIUM
41YDN	11CBTI	473	FLUCTUATIONS OF ATMOSPHERIC
61TTI	8RSAR	29	IMPROVEMENT OF OPERATING PERFORMANCE OF
61MLR	21EPMT	145	IBM ACCOUNTING MACHINE TO OBTAIN FREQUENCIES.
57BLL	28AMSX	520	ON DISCRETE VARIABLES WHOSE
56TSO	27AMSX	703	DISTRIBUTION OF THE
61ECR	279MSSN	CLTS	CENTRAL LIMIT THEOREM FOR
46KAC	47AMTX	33	ISTRIBUTION OF VALUES OF SUMS OF THE TYPE *
56BRT	27AMSX	1060	ON SEQUENTIAL DESIGNS FOR MAXIMIZING THE
62BLM	9NM19	CLTS	ON THE CENTRAL LIMIT THEOREM FOR THE
63BLM	12WVG	389	ON THE CENTRAL LIMIT THEOREM FOR THE
55GRD	26AMSX	233	DISTRIBUTION OF LENGTH AND COMPONENTS OF THE
50PCR	10JASS	52	THE DISTRIBUTION OF THE
46FRS	8SJRS	223	PULATION. ON THE DISTRIBUTION OF THE
24PRN	16BMTA	202	/ONSHIP OF THE INCOMPLETE B-FUNCTION TO THE

Figure 6.1
Permuted Title Index Page (Right Hand Side)

```
SUFFICIENT STATISTICS.
SUFFICIENT STATISTICS.
SUFFICIENT STATISTICS   AND*  SIMILAR TESTS.
SUFFICIENT STATISTICS  FOR A  SEQUENCE OF INDEPENDENT RANDO
SUFFICIENT STATISTICS  FOR THE  PARTIALLY BALANCED INCOMPLE
SUFFICIENT.                                  A  SUBFIELD  CONTA
SUFFICIENT STATISTICS. /UCCESSIVE PROCESS OF STATISTICAL IN
SUFFICIENT STATISTICS. /UCCESSIVE PROCESS OF STATISTICAL IN
SUFFICIENT STATISTICS.
SUFFICIENT STATISTICS.
SUFFICIENT SUBFIELDS.
SUFFICIENT STATISTICS.                              APPLICATION  OF
SUFFICIENT STATISTIC.          A  COMBINATORIAL DERIVATION O
SUFFICIENTI.                                                SU UN
SUFFICIENTI.                                                SU UN
SUFFISANTES DE  LA  CONVERGENCE STOCHASTIQUE.
SUFFOLK.
SUGAR BEETS.                          CHEMICAL GENETIC AND SOI
SUGAR BEETS.                       STUDIES ON  STATISTICAL PROCE
SUGAR CANE BREEDING PROGRAM.                                    A
SUGAR INDUSTRY.
SUGAR CANE CLUMPS  WITH REGARD TO  TILLER NUMBER AND  ITS
SUGAR CANE MANURING.                                 AN EXAMPLE
SUGAR BEETS  WITH A  DETAILED STATISTICAL PROCEDURE OF CON/
SUGAR BEET  EXAMPLE ON  COVARIANCE.
SUGARCANE LEAVES.                                           THE
SUID-AFRIKA.
SUI-GENERIS STATISTICAL ORGANIZATION.
SUISSE.                                  FORMATION  AUX APPLI
SUITABLE  FOR THE APPLICATION  OF  A          CURVE OF FR/
SUITABLE  FOR USE WITH  NON-STATION      IN  SERIES.
SUITABLE  FOR USE ON AN  ELECTRONIC  COMPUTER. /D FOR  ESTI
SUITABLE  FOR  DESCRIBING  THE  FREQUENCY OF RANDOM SAMPLES
SUITABLE PROPORTION  BETWEEN THE  VALUES OF FIRST AND SECON
SUITABLE LINEAR STATISTICS.        ON A  CHARACTERIZATION OF
SUITABLE STATISTICS.                                        CHA
SUITABLY CHOSEN ORDER STATISTICS. /E OR TWO PARAMETERS OF
SUITES ARITHMETIQUES. METHODE DE MONTE-CARLO.
SUITES DE PROCESSUS STOCHASTIQUES  APPLICATIONS A LA  THEOR
SUITE X-SUB-N  = A-SUB-N.  EVALUATION DE  L'  ERREUR.
SUITING  THE  CHART  TO THE  AUDIENCE.
SULFAMERIZINE  ON THE  ERYTHROCYTE COUNT AND HEMOGLOBIN CON
SULPHATE..
SULPHUR DIOXIDE.
SULPHURIC ACID PLANT  BY THE  DESIGN OF EXPERIMENTS.
SUMS, AND SUMS OF SQUARES, IGNORING INCOMPLETE DATA. /THE
SUM  IS  ABSOLUTELY CONTINUOUS.
SUM  IN  RANDOM SAMPLES  FROM A  DISCRETE POPULATION.
SUMS  OVER  SETS OF RANDOM VARIABLES.
SUM F(2-TO-THE-K TIMES T).                          ON THE  D
SUM OF  N  OBSERVATIONS.
SUM OF  A  RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES.
SUM OF  A  RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES.
SUM OF  N  RANDOM UNIT VECTORS.                             THE
SUM OF  N  RECTANGULAR VARIATES.  I.
SUM OF  N  SAMPLE VALUES  DRAWN FROM A  TRUNCATED NORMAL PO
SUM OF  THE  FIRST P TERMS OF  THE  BINOMIAL (A+B)-POWER-N/
```

40-0720

T. L. BRATCHER AND W. R. SCHUCANY, SOUTHERN METHODIST UNIVERSITY.

BAYESIAN PREDICTION AND POPULATION SIZE ASSUMPTIONS.

THIS PAPER IS CONCERNED WITH THE DISTRIBUTION OF THE NUMBER SUCCESSES IN A RANDOM SAMPLE GIVEN THE RESULTS OF A PREVIOUS SAMPLE FROM THE SAME POPULATION. ASSUMING UNIFORM WEIGHTS * ON THE PROPORTION OF SUCCESSES IN THE ORIGINAL POPULATION, BAYES RULE IS UTILIZED TO OBTAIN THE DESIRED DISTRIBUTION. IF THE SIZE OF THE POPULATION IS FINITE, SAY N, THEN THE HYPERGEOMETRIC DENSITY GIVES THE PROBABILITIES FOR THE NUMBER OF SUCCESSES IN A RANDOM SAMPLE. ON THE OTHER HAND, IF N IS INFINITE, THE BINOMIAL GIVES THE PROBABILITIES. SOMEWHAT SURPRISINGLY, THE RESULTING DISTRIBUTION IS INDEPENDENT OF THE POPULATION SIZE N AND IS THE SAME FOR BOTH THE FINITE AND INFINITE CASES. *

NUMBER SUCCESSES ·
RANDOM SAMPLE
UNIFORM WEIGHTS
PROPORTION OF SUCCESSES
* SUCCESSES , PROPORTION OF
BAYES RULE
HYPERGEOMETRIC DENSITY
NUMBER OF SUCCESSES
SUCCESSES , NUMBER OF
RANDOM SAMPLE
POPULATION SIZE
INFINITE CASES

Figure 6.2
Abstract and Abstract Index

Figure 6.3

Cumulative Index to 50 Abstracts (one page)

realize a compressive gain that may be useful for accessing the abstracts. It will certainly be useful for accessing the original documents when it is applied to a collection of abstracts and the resulting indexes are accumulated.

The page extracted from the middle of the cumulative abstract indexed reproduced as Figure 6.3 shows that one paper in the sample of 50 referred, via its abstract, to the "NON-CENTRAL MULTIVARIATE BETA DISTRIBUTION", and, since the abstract transmitted this phrase, the paper undoubtedly contains something of interest about this topic. Similarly note that eight papers referred to the "NORMAL" distribution in some form. The presence of spurious terms like "ONTO ITSELF" and "OPTIMUM BLUE'S" is no more than a minor annoyance in use of the index, and is of course due to inadequacies in the indexing algorithm's "stop list", which should certainly contain the word "ITSELF". There are other more subtle problems whose genesis is the indexing algorithm, but they are not so obtrusive as to mac the use of the list burdensome. For instance, the phrase "OPTIMUM BLUE'S occurs in the abstract, where it is defined to denote "OPTIMUM BEST LINEAR UNBIASED ESTIMATE"; this phrase certainly belongs in the index, but it is not clear that a user of the amalgammated index would recognize the technical meaning of "BLUE" until it had become a standard term of the field.

Indexing abstracts is of potential value in gaining access to the large numbers of journal papers which annually appear in the literature; coupled with permuted title access mechanisms, the abstract index should provide a rapid and reliable means of surveying the key content areas of papers without the time-consuming process of reading abstracts, which often limits one to a relatively narrow and current range of documents.

When compared to the earlier manual implementation of the algorithm on Computerized Library Catalogs, the machine implementation of the algorithm differs in several ways, aside from the obvious fact that the machine is entirely consistent in its application where manual procedures cannot be. The raw data for the machine test on the statistical abstracts was keypunched in all upper case, as a matter of convenience. Hence, the rule to keep capitalized one-word entries was inoperative in this run. Further, no attempt has been made to include *see* or *see-also* types references in the machine implementation. On the other hand, the machine implementation includes logic to allow structure-word entries where the manual implementation did not.

These differences are reflected in the statistics describing the entry length and page location distributions.

Table 6.1 provides the entry length (in number of words) distribution for the machine index to the statistical abstracts.

Table 6.1

Entry Length Distribution
Algorithmic Index to Statistical Abstracts

Number of Words	Frequency	Cumulative Frequency
1	0	0
2	315	315
3	233	548
4	125	673
5	54	727
6	19	746
7	6	752
8	2	754

Comparing this distribution to the comparable distribution for the manually implemented algorithmic index to Computerized Library Catalogs (Table 5.3) one sees that the proportion of one-word-entries has been reduced to zero (because there is no logic available to permit one-word-entries) and that the overall average entry length has been increased from 2.08 words per entry to 3.01 words per entry. The main factor in this increase is the introduction of structure-word entries, although the absence of one-word-entries has a small effect on average entry length as well.

The entry length distribution is plotted on Figure 6.4. Despite the absence of one-word-entries, the points are nicely fit by a straight line confirming the nice approximation by a lognormal distribution.

It will be recalled that in the previous study of page location distribution for the manually implemented version of the algorithm on Computerized Library Catalogs there was a significant bend in the Zipf-Mandelbrot straight line due to either a reduced number of singly occurring entries, or an excessive number of multiply occurring entries. For the machine version of the algorithm the page location distribution (or, more accurately, the abstract number location distribution) does not show this deviation (see Figure 6.5). The distribution is given in Table 6.2.

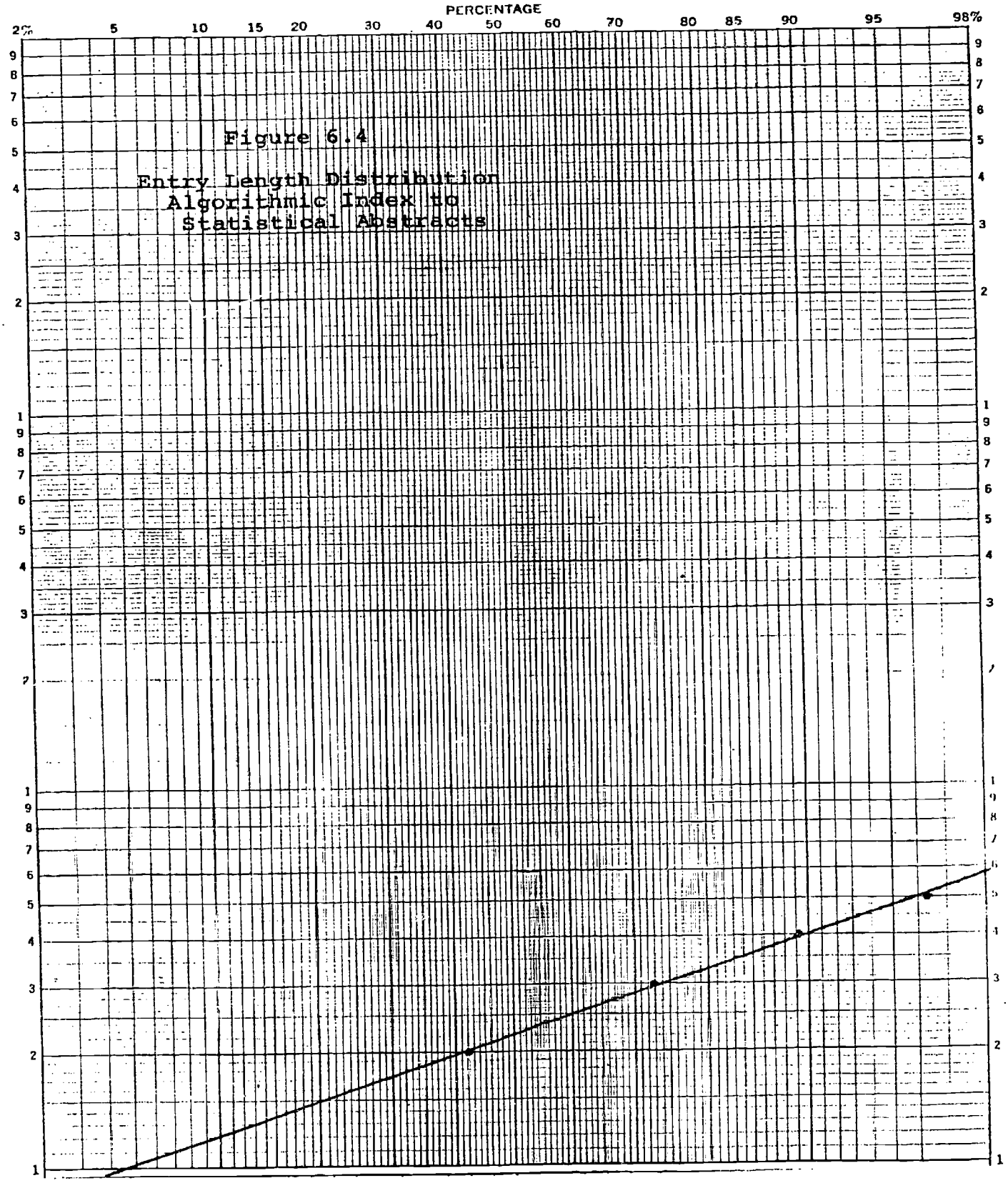


Figure 6.4

Entry Length Distribution
Algorithmic Index to
Statistical Abstracts

Figure 6.5

Abstract Number Location Distribution
Algorithmic Index to Statistical Abstracts

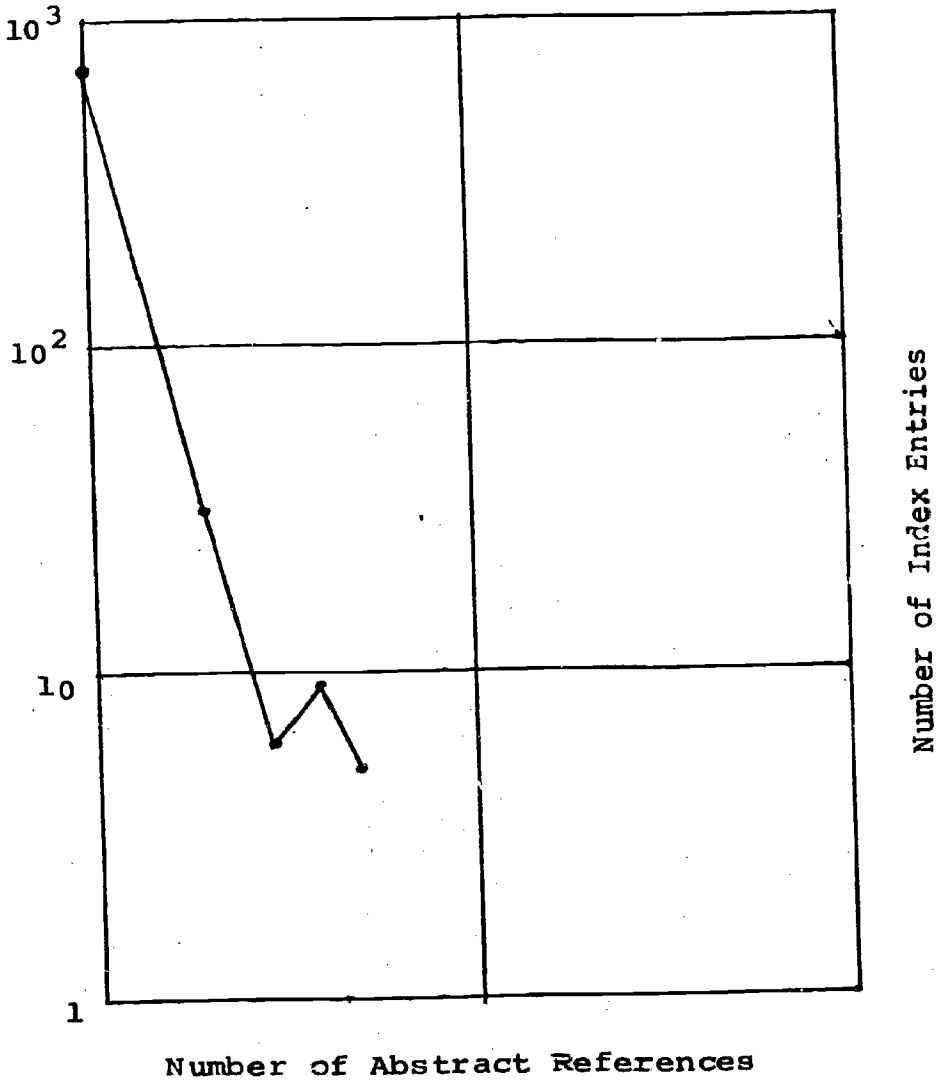


Table 6.2

Abstract Number Location Distribution, Algorithmic Index to Statistical Abstracts

Number of Abstract Locations per Entry	Frequency	Cumulative Frequency
1	703	703
2	31	734
3	6	740
4	9	749
5	5	754

When compared to the comparable data for the manual implementation, it is clear that not only has the difficulty of an insufficient proportion of singly occurring entries been been corrected by the insertion of structure word logic, but the slope of the line has been significantly increased from 2.17 to 4.49. This increased slope can of course be attributed in part to the nature of the material covered in the two cases and, perhaps, in greater proportion to the structure of the material (i.e. fifty abstracts vs. a single text). Nonetheless, the increase in slope does tend to confirm the expectation that use of structure-word entries is desirable to increase slope.

Potentially more useful than the amalgamation of indexes to abstracts to papers or books is the amalgamation of indexes to the primary texts themselves. We have undertaken an extensive project designed to provide a realistic test of the utility of amalgamations of book indexes as well as an indication of the problems that would be encountered in the preparation of such access mechanisms.

The indexes contained in 80 books on statistics have been committed to machinable form. Approximately 30,000 index entries (not all of which are distinct) are represented, which is nearly 400 entries per book. This is significantly less than the average of 838 index entries per book obtained from the Fondren Index Sample, but, as is clear from Table 4.4, it is well within the deviations typically

obtained by restriction of a sample to small and specially defined subsets. We have not attempted to determine the average number of pages per book in this statistics sample; it may well be that the average number of index entries per page is in closer agreement with the figure obtained for the Fondren Index Sample.

The Statistics Index Sample is currently in the early stages of amalgamation. In this report we can only exhibit a combined alphabetically ordered list which has not been formatted (to reproduce the usual format of a book index) and which exhibits the consequences of some program "bugs" not yet corrected which result in the replication of input records at various places throughout the amalgamated index. In spite of these difficulties, the amalgamated list is already a valuable access tool.

Table 6.3 lists the books that constitute the Statistical Index Sample. The code in the leftmost column is the abbreviation for the book used in the amalgamated index. These books were chosen by a professional statistician as representative of the more important information in the statistics field that is available in monograph form. The choice of 30 books rather than a larger number is purely conventional; continuation of this project will increase the data base and permit us to determine how the yield of new index terms varies with increasing size of the sample.

Following the lead of the analysis of the structure of the index to a single book given in Chapters 3 and 4, we see that the rank-frequency distribution Figure 6.6 is just another form of the index reference distribution discussed in those chapters; in the form shown here, the abstract entries appear at the top left part of the graph, and the horizontal portions of the graph correspond to those entries which refer to the same number of text locations. Consequently, the abstract entries for the Statistics Sample certainly include those that have ranks less than 30, and may include several more but not any with rank greater than 50.

Table 6.4 lists the 30 index terms that refer to the greatest number of pages; personal names have been placed in the right hand column; otherwise the order of appearance in the amalgamated index list is the order shown in the table.* This list is a useful pedagogical tool, providing

---

* The frequencies given here are very tentative, as no attempt has yet been made to agglomerate proper names appearing in variant form.

Table 6.3
Bibliographic Description of the Statistics Sample

A Elementary Decision Theory
H. Chernoff and L. E. Moses
Wiley and Sons, N.Y. 1959
QA276.C47

B Nonparametric Methods in Statistics
D.A.S. Fraser
Wiley and Sons, N.Y. 1957
QA276.F66

C Statistical Methods for Chemists
W.J. Youden
Wiley and Sons, N.Y. 1951
QA276.YA

D Analysis of Straight-line Data
F.S. Acton
Wiley and Sons, N.Y. 1959
QA276.A25

E Testing Statistical Hypotheses
E. L. Lehmann
Wiley and Sons, N.Y. 1959
QA276.L343

F Introduction to Mathematical Statistics
P. G. Hoel
Wiley and Sons, N.Y. 1947
QA276.H57

G The Design and Analysis of Experiments
O. Kempthorne
Wiley and Sons, N.Y. 1952
HA29.K425

H An Introduction to Multivariate Statistical Analysis
T.W. Anderson
Wiley and Sons, N.Y. 1958
QA276.A6

I Statistics---An Introduction
D.A.S. Fraser
Wiley and Sons, N.Y. 1958
HA29.F67

J Linear Computations
P.S. Dwyer
Wiley and Sons, N.Y. 1951
QA195.D95

K Modern Probability Theory and Its Applications
E. Parzen
Wiley and Sons, N.Y. 1960
QA273.P272

Table 6.3 (Continued)

L Planning of Experiments
D.R. Cox
Wiley and Sons, N.Y. 1958
Q175.C8

M Theory of Games and Statistical Decisions
D. Blackwell and M.A. Girshick
Wiley and Sons, N.Y. 1954
QA269.B5

N An Introduction to Probability Theory and Its Applications, V. 1
Wm. Feller
Wiley and Sons, N.Y. 1968
QA273.F37

O Elementary Ststistics
P.G. Hoel
Wiley and Sons, N.Y. 1960
HA29.H662

P The Elements of Probability and Some of its Applications
H. Cramer
Wiley and Sons, N.Y. 1955
QA273.C843

Q Statistical Decision Theory
L. Weiss
McGraw-Hill, N.Y. 1961
QA276.W44

R Introduction to Probability and Random Variables
G.P. Wadsworth and J.G. Bryan
McGraw-Hill, N.Y. 1960
QA273.W2

S Introduction to the theory of Statistics
A.M. Mood and F.A. Graybill
McGraw-Hill, N.Y. 1963
HA29.M75

T Elements of Probability and Statistics
F.L. Wolf
McGraw-Hill, N.Y. 1962
QA273.W69

U An Introduction to Linear Statistical Models, V. 1
F.A. Graybill
McGraw-Hill, N.Y. 1961
HA29.G75

V Elements of the Theory of Markov Processes and their Applications
A.T. Bharucha-Reid
McGraw-Hill, N.Y. 1960
QA273.B57

Table 6.3 (Continued)

W Geometrical Probility
M.G. Kendall and P.A.P. Moran
Hafner, N.Y. 1963
QA273.K349

X Fundamentals of Statistical Reasoning
M.H. Quenouille
Hafner, N.Y. 1958
HA29.Q44

Y Characteristic Functions
E. Lukas
Griffin, London 1970
QA273.6.L85

Z An Introduction to Probability Theory and its Applications, V. 2
Wm. Feller
Wiley and Sons, N.Y. 1968
QA273.F37

AB Elements of Mathematical Statistics
H.W. Alexander
Wiley and Sons, N.Y. 1961
QA276.A555

AC Statistical Theory and Methodology in Science and Engineering
K.A. Brownlee
Wiley and Sons, N.Y. 1965
QA276.B77

AD Statistics and Experimental Design, V. 1
Johnson and Lecne

AE Mathematical Statistics
S.S. Wilks
Wiley and Sons, N.Y. 1962
QA276.W513

AF Experimental Designs
Wm. G Cochran and Gertrude M. Cox
Wiley and Sons, N.Y. 1950
Q180.A1C6

AI A Course in Probability Theory
Kai Lai Chung
Harcourt, Brace and World, N.Y. 1968
QA273.C5

AJ Essentials of Probability
A. Yaspan
Prindle, Weber and Schmidt, Boston 1968
QA273.Y38

Table 6.3 (Continued)

AK The Design of Experiments
R. A. Fisher
Oliver and Boyd, Edinburgh 1951
HA29.F48

AL Computational Handbook of Statistics
J.L. Bruning and B.L. Kintz
Scott and Foresman, Glenview, Ill. 1968
HA29.B835

AM The Design and Analysis of Experiments
M.H. Quenouille
Hafner, N.Y. 1953
Q180.A1Q4

AN Handbook of Statistical Tables
D.B. Owen
Addison-Wesley, Reading, Mass. 1962
HA48.O9

Table 6.3 (Continued)

AO The Elements of Probability
Berman
Addison-Wesley, Reading,Mass. 1969
QA273.B498

AP Design and Analysis of Industrial Experiments
O.L. Davies
Hafner, N.Y. 1954
T175.D3

AQ Statistical Theory
B.W. Lindgren
Macmillan, N.Y. 1962
QA276.L546

AR Introduction to Statistics
F.W. Carlborg
Scott and Foresman, Glenview, Ill. 1968
QA276.C285

AS Probability and Statistics
H.L. Adler and E.B. Rossler
W.H. Freeman, San Francisco 1964
QA273.A43

AT Measuring Uncertainty---An Elementary Introduction to Bayesian Statistics
S.A. Schmitt
Addison-Wesley, Reading, Mass. 1969
QA279.5.S33

AU A Brief Introduction to Probability Theory
J.P. Hoyt
International Textbook, Scranton, Pa. 1967
QA273.H79

AV Statistical Design and Analysis of Experiments for Development Research
D.S. Villars
W.C. Brown, Co., Dubuque 1951
HA29.V5

AW Statistics in Research
B.Ostle
Iowa State [illegible]ity [illegible]ss, Ames 1963
HA29.O8

AX Schaums Outline SEries Theory and Problems of Probability
S. Lipschutz
Schaum Pub., N.Y. 1964
QA248.L5

Table 6.3 (Continued)

AY Elementary Mathematical Programming
R.W. Metzger
Wiley and Sons, N.Y. 1958
QA264.M4

AZ Statistical Inference for Markov Processes
P. Billingsley
University of Chicago Press 1961
QA274.B576

BD Statistical Analysis of STationary Time Series
Ulf. Grenander and M. Rosenblatt
Almqvist and Wiksell, Stockholm 1956
QA280.G73

BE Statistical Methods in Experimentation---An Introduction
O.L. Davies
Macmillan, N.Y. 1953
QA279.D25

BF Stochastic Processes--Basic Theory and Its Application
N.U. Prabhu
Macmillan, N.Y. 1965
QA273.P79

BG Probability and Frequency
H.C. Plummer
Macmillan, London 1940
QA273.P55

BH Statistical Methods for Research Workers
R.A. Fisher
Oliver and Boyd, Edinburgh 1970
HA29.F5

BJ Regression Analysis
E.J. Williams
Wiley and Sons, N.Y. 1959
QA278.2.W5

BK Statistical Processes and Reliability Engineering
D.N. Chorafass
Van Nostand, Princeton, N.J. 1960
QA276.C475

BL Introduction to Probability and Mathematical Statistics
Z.W. Birnbaum
Harper N.Y. 1962
QA273.B579

BM Elementary Mathematical Statistics
Wm. D. Baten
Wiley and Sons, N.Y. 1938
QA276.B3

Table 6.3 (Continued)

BN Introduction to Biostatistics
H. Bancroft
Hoeber-Harper, N.Y. 1957
QA276.B25

BO Sampling Techniques
Wm. G. Cochran
Wiley and Sons, N.Y. 1953
QA276.5.C6

BP A History of the Mathematical Theory of Probability
I. Todhunter
Chelsea Pub., N.Y. 1949
QA273.T63

BQ Statistical Methods in Biology
N.T.J. Bailey
English Universities Press, London, 1959
QA276.B23

BR Statistical Theory---The Relationship of Probability, Credibility, and Error
L. Hogben
W.W. Norton and Co., N.Y. 1957

BT Probability and Experimental Errors in Science
L.G. Parratt
Wiley and Sons, N.Y. 1961
QA273.P24

BU Contributions to Order Statistics
A.E. Sarhan and B.G. Greenberg
Wiley and Sons, N.Y. 1962
QA276.S29

BV Introduction to Statistical Method
S. Ehrenfeld and S.B. Littauer
McGraw-Hill, N.Y. 1964
QA276.E35

BW Theory of Probability
H. Jeffreys
Clarendon Press, Oxford, 1948
QA273.J4

BX Statistical Adjustment of Data
W.E. Deming
Wiley and Sons, N.Y. 1943
QA275.D35

BY Statistical Analysis in Chemistry and the Chemical Industry
C.A. Bennett and N.L. Franklin
Wiley and Sons, N.Y. 1954
QA276.B38

Table 6.3 (Continued)

BZ Probability Random Variables and Stochastic Processes
A. Papoulis
McGraw-Hill, N.Y. 1965
QA273.P2

CD Elements of Queueing Theory with Applications
T.L. Saaty
McGraw-Hill, N.Y. 1961
QA273.S218

CE Stochastic Processes
J.L. Doob
Wiley and Sons, N.Y. 1953
QA273.D755

CF Sample Survey Methods and Theory, V. 1 Methods and Applications
M.H, Hansen, W.N. Hurwitz and Wm. G. Madow
Wiley and Sons, N.Y. 1953
QA276.H33

CG Advanced Statistical Methods in Biometric Research
C. Radhakrishna Rao
Wiley and Sons, N.Y. 1952
QA276.R3

CH Introduction to the Mathematics of Statistics
R. W. Burgess
Houghton Mifflin Co., Boston 1927
HA29.B85

CI A Graduate Course in Probability
H. G. Tucker
Academic Press, N.Y. 1967
QA273.T78

Figure 6.6

Rank - Frequency of Reference
Distribution
Statistical Index Sample

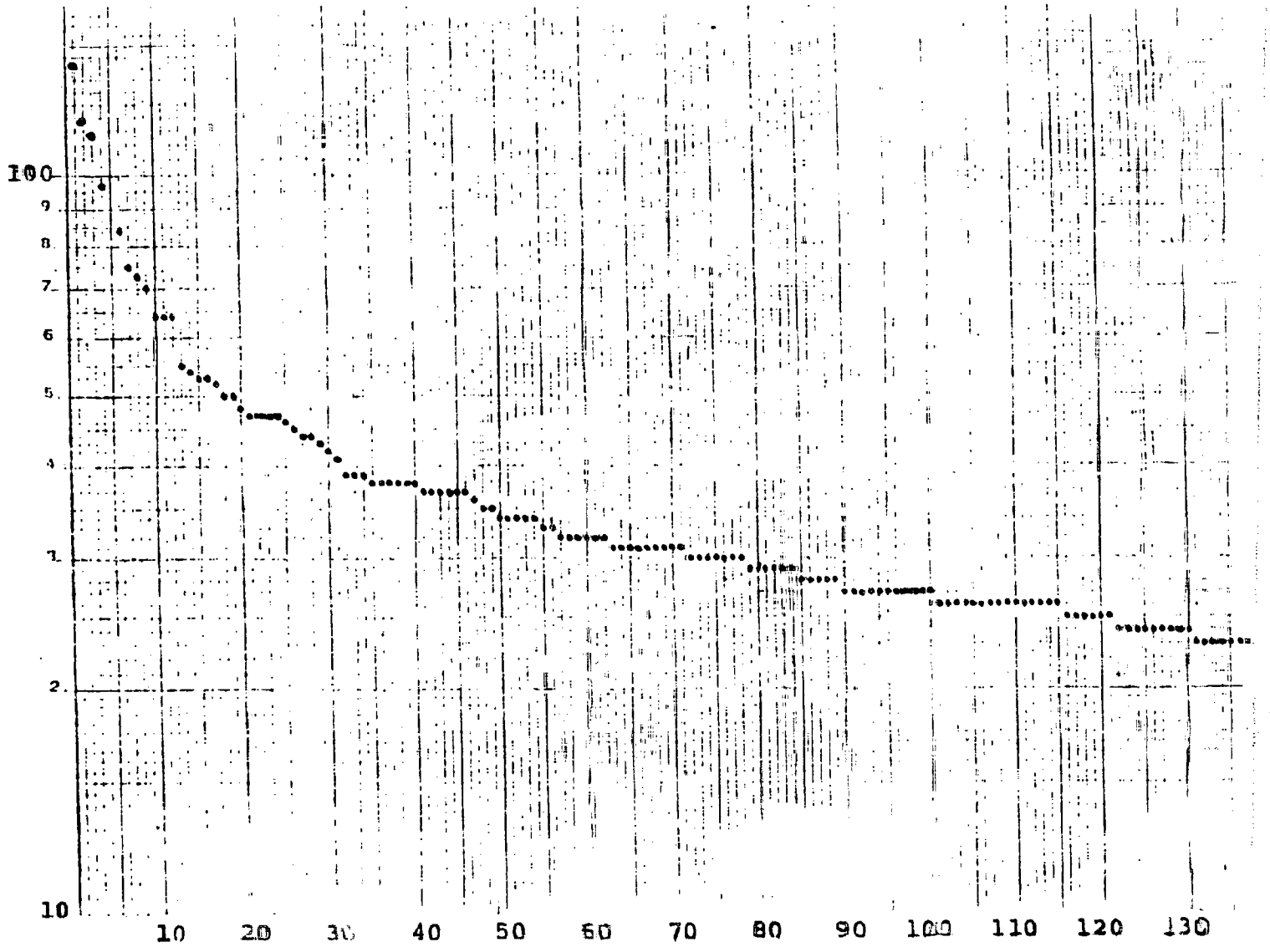


Table 6.4

ABSTRACT ENTRIES FOR THE
AMALGAMATED STATISTICS INDEX SAMPLE

Normal distribution	Fisher, R. A.
Binomial distribution	Student
Poisson distribution	Pearson, E. S.
Degrees of freedom	Kendall, M. G.
Conditional probability	Bartlett, M. S.
Standard deviation	Cramer, H.
Analysis of variance	Neyman, J.
Distribution	
Chi-square distribution	
Central limit theorem	
Least squares	
Variance	
Correlation coefficient	
Median	
Cauchy distribution	
Covariance	
Independence	
Random variable	
Exponential distribution	
Gamma distribution	
Moments	
Bivariate normal distribution	
Multinomial distribution	

as it does an immediate and objective overview of the important subjects in statistics as well as the important contributors. It plays the same role relative to that portion of the field of statistics represented in the monograph literature that the abstract entries for the books described in Chapter 5 played; and it increases the degree of information compression as well.

Figure 6.7 shows one page from the uncorrected form of the amalgamated Statistics Index Sample described above. This page has been selected to include the entry "log normal" and those related to it. Observe that six books (coded P, S, AD, BL, BU, CD) contain references to the log normal distribution; since this represents only 7.5% of the books in the Statistics Index Sample, the unsophisticated inquirer will realize a very significant saving in search time with a reasonable degree of assurance that most of the significant references will either be covered directly within these six books, or more comprehensive treatments will be noted in their bibliographies.

Figure 6.7 Sample Page - Amalgamated Statistics Index

APPENDIX I

ABSTRACT INDEX ENTRIES: A UNIFORM SAMPLE FROM THE FONDREN INDEX SAMPLE

25
BF181.M3 1931

Marston, William Moulton
Integrative Psychology

Sum= 1432 / 29.54 = 48

23 Marston, W.M.

17 Freud, Sigmund

15 Watson, J.B.

13 Cannon,.W.B.

11 Adler, Alfred

10 Desire
10 Jung, Carl
10 Libido
10 Woodsworth, R.S.

9 Compliance
9 Passion

8 Allport, F.H.
8 Behaviourism
8 James, WM.
8 MacDougall, Wm

7 Herrick, C.J.
7 Satisfaction
7 Sherrington, C.S.

6 Captivation
6 Dominance
6 Psychoanalysts

5 Carlson, A.J.
5 Erotic drive
5 Inducement
5 Passion response
5 Visual discrimination, substances, hypothetical

4 Angell, J.R.
4 Archtypes
4 Cell body
4 Compliance, motives
4 Eng, H.
4 Hering, E.
4 James-Lange, Theory of Motion
4 Law of integrative sequence
4 Origination response
4 Passion motives
4 Submission
4 Trolaut, L.T.
4 Unit responses, compound
4 Washburn, M.F.
4 Yerkes, R.M.

50
BV2532.M3 1921

McLean, Archibald
The History of the Foreign Missionary Society

Sum= 498 / 29.54 = 16.86

7 Fallen, The

6 Moore, W.T.

4 Nurses being trained

3 Bilaspur
3 Johnson, Miss Kate V.
3 Loos, C.L.
3 Moore, W.T., Quoted
3 Rijnhardt, Dr. Susie C.

75
D443.C6 1927

Coolidge, Archibald Cary
Ten Years of War and Peace

Sum= 415 / 29.54 = 14

31 Great Britain mentioned

20 France, mentioned
20 Poland

19 League of Nations, mentioned

18 Versailles, Treaty of

16 Wilson, Woodrow

15 Hungary

13 Algeria
13 Hughes, Charles E., Secretary of State

11 Germany, mentioned
11 Harding, Warren G.
11 Morocco

10 China
10 France, estrangement between and Great Britain
10 Japan, mentioned
10 Rumania

100
DA690.K7 1922

Sackville-West, Victoria Mary
Knole and the Sackvilles

Sum= 307 / 29.54 = 10

7 Sackville, Lady Margaret(afterwards Countess of Thanet), mentioned in Lady Anne Clifford's diary

4 Pepys, Samuel, quoted
4 Walpole, Horace, quoted on Knole

3 Devonshire, Duchess of, his (i.e. 3rd Duke of Dorset) letter to her
3 Dryden, John, his debt to 6th Earl of Dorset
3 Gorboduc
3 Macaulay, quoted
3 Sackville, Charles, 6th Earl of Dorset, songs quoted
3 Sackville, Lord George, quoted
3 Wraxall, Sir Nathaniel, quoted

125
DF521.S4 1966

Sherrard, Philip
Byzantium

Sum= 1643 / 29.54^2 = 1.88

10 Churches: in Constantinople
10 Frescoes

150
DS423.C85 v4 1953-58

Institute of Culture
The Cultural Heritage of India

Sum= 4906 / 29.54^2 = 5.6

51 Krsna (śrī)

46 Siva

41 "Bhagavad- Gītā"

37 Visnu

33 Brahman

32 Guru(s)

175
E178.6.S3 1965

Saveth, Edward, ed.
Understanding the American Past

1958 / 29.54^2 = 2.24

28 Beard, Charles A. and Mary

20 Jefferson, Thomas

17 Turner, Frederick Jackson

200
E741.L55 1963

Link, Arthur S.
American Epoch: A History of the United States Since the 1890's

Sum= 7016 / 29.54^2 = 8

14 Prices: agricultural

12 Foreign relations: Anglo-American

11 Federal income tax: individual
11 Tax: individual income

10 Farmers, income of
10 Legislation:agricultural

9 Agriculture, legislation for
9 Railroads: rates of

Bolton, Herbert Eugene
Anza's California Expedition

Sum= 648 / 29.54 = 22

132 Mass

111 Anza, Juan Bautista

60 Garces, Fray Francisco

46 Monterrey

44 San Gabriel Mission

27 San Diego mentioned

26 Colorado River
26 Ribera (Rivera) Fernando de.
26 Sierra Madre de California

23 Apaches

21 Palma, Salvador
21 Sierra Nevada

20 San Miguel de Horcasitas

19 Eixarch (Eyxarch) Fray Thomàs

18 San Francisco, harbor and settlement

17 Mexico

16 Spaniards

15 Christian Indians
15 Fages, Pedro
15 Gila River

14 Pablo (Captain Feo) Yuma Chief

13 Crespi, Fray Juan
13 Rio de San Francisco (San Joachin)

De Garmo, Ernest Paul
Engineering Economy

Sum= 650 / 29.54 = 22

5 Terborgh, Genge

4 Break even charts, examples of

3 Balance sheet, example of
3 Deferred-investment studies, examples of
3 Minimum cost point
3 Personnel factors, lighting
3 Rate of return, determination of
3 Selection, of design
3 Survivor curves, examples of

2 Accidents, effect of lighting on
2 Annuities whose present value is
2 Borrowed capital, cost of
2 Bureau of Internal Revenue relation to depreciation
2 Capital gains, and losses
2 Capital gains and losses, carry-over of
2 Capitalized cost, example of application
2 Costs, accuracy of estimates of
2 Costs, labor
2 Depreciation, sum-of-the-years'-digits
2 Hoover Dam
2 Income and expense statements, example of
2 Income taxes in public utility studies
2 Increment costs
2 Labor, turnover of
2 Life, economic
2 Life, useful
2 MAPI replacement formulas, forms for use in
2 Material, selection of
2 Multiple-purpose works, evaluation of benefits from
2 Overhead expense bases for distribution of
2 Plant location, economy studies of
2 Power factor, effect on utility rates
2 Rate schedules, block demand
2 Rautenstrauch, Walter
2 Risk, factors affecting
2 Selection of methods or processes
2 Self liquidating projects, relation of taxes to
2 Self liquidating projects, repayment of capital in
2 Wage payment, piece work

275
HD20.C554 1958

Chorafas, Dimitris N.
Operations Research for Industrial Management

Sum= 141 / 29.54 = 5

17 Charts on simulated business results

11 Computers usage
11 Simulation

10 Allocation

9 Managerial decisions

300
HF2046.S62 1904

Smart, William

The Return to Protection

S= 201 / 29.54 = 7

20 Chamberlain, J.
20 Germany

19 Board of Trade

14 France

10 Giffen, Sir Robert
10 Shipping

9 America
9 America and protection
9 Canada

325
HM66.C7 1920

Co[illegible], George Howard Douglas
Social Theory

Sum= 382 / 29.54 = 13

22 Trade Unions

15 "State, The"

14 Associations

12 Churches
12 Functional Equity, Court of, organisation

10 Law
10 Rousseau

9 Sovereignty

8 Function in relation to individual, perversion of
8 Marxism
8 Will, as a basis of Society

7 Middle Ages
7 Parliament

350
JA84.R9 U8 1964

Utechin, Sergei
Russian Political Thought

Sum= 409 / 29.54 = 14

50 Economy

45 Classes, social

43 Law

39 Germany

34 Individualism
34 Monarchy

33 Emigration
33 France
33 Peasantry

32 Intelligentsia

31 Education

30 Property
30 Terror

29 Christianity
29 Culture
29 Equality
29 Moscow
29 Nationalism
29 Nobility
29 St. Petersburg

Cooper, Lane
Two Views of Education

Sum= 775 / 29.54 = 26

43 America

42 Milton

39 Plato

34 Shakespeare

27 Aristotle

26 Homer
26 Teacher (of English, etc.)

25 Greek, Study of

24 Middle Ages

22 Horace
22 Wordsworth

21 Bible

20 Latin, Study of

18 Cicero

17 Rewards of the Teacher

16 Odyssey
16 Rome
16 Virgil

15 Chaucer
15 Discipline
15 England
15 Socrates

14 Dante

13 Democracy
13 Greece
13 Rousseau

400
LC191.M6 1916

Morgan, Alexander
Education and Social Progress

Sum= 254 / 29.54 = 9

8 Children, diseases of
8 Education, practical
8 Inter-Departmental committee

7 Kindergartens
7 Practical education
7 Vocational education

6 Commission, Royal, on Poor Laws
6 Continuation education
6 Edinburgh, continuation schools
6 Education, continuation
6 Education, vocational
6 Education and health
6 Plato
6 Scotch Education Deptartment
6 Slums, children in

425
ND553.D774 T6 1965

Tomkins, Calvin
The Bride and the Bachelors

Sum= 414 / 29.54 = 14

22 "Bride Stripped Bare By Her Bachelors, Even, The" (Duchamp)

15 Tudor, David

14 Cunningham, Merce

13 Rauschenberg, Robert

12 Klüver, Billy

11 Johns, Jasper

10 Duchamp, Marcel

8 Feldman, Morton
8 Thomson, Virgil
8 Tinguely, Jean

7 Arensberg,Walter C.
7 Breton, Andre
7 Cage, John
7 Cage, Mrs. John
7 Cowell, Henry
7 Dreier, Katherine
7 Kashevaroff, Xenia Andreevna
7 "Nu Descendant un Escalier" (Duchamp)
7 "Nude Descinding a Staircase" (Duchamp)
7 Schönberg, Arnold

Bobbe, Dorothie (De Bear)
Fanny Kemble

Sum= 384 / 29.54 = 13

42 Butler, Pierce

40 St. Leger, Harriet

36 Kemble, Charles

22 Butler, Sarah

19 Butler, Fanny
19 Siddons, Sarah

18 Covent Gargen Theatre
18 Lenox,Mass.

17 Kemble, Adelaide.
17 Kemble, Mrs Charles
17 Sartoris, Mrs Edward
17 Slavery

15 Kemble, John Mitchell

475
PR2831.H6 1948

Hoppe, Harry Reno
The Bad Quarto of Romeo and Juliet

Sum= 529 / 29.54 = 18

38 Greg, Walter W.

22 Chambers, (Sir) E.K.

18 Hart, Alfred
18 Mc Kerrow, R.B.

14 Burby, Cuthbert

13 Boswell, Eleanor
13 Greene, Robert "Orlando Furioso"

12 Arber, Edward
12 Recollections
12 Shakespeare, William,"The Merry Wives of Windsor"

11 "Orlando Furioso"

10 Anticipations
10 Chamberlain"s Company
10 Danter, John
10 Shakespeare, William "3 Henry VI"

9 "3 Henry VI"
9 "Merry Wives of Windsor, The "
9 Peele, George "The Old Wives" Tale"
9 Peele, George "Edward I"
9 Repetitions

500
PR5588.R9 1964

Ryals, Clyde, de L
Theme and Symbol in Tennyson's Poems to 1850

Sum= 322 / 29.54 = 11

37 Keats, John

23 "In Memoriam"

22 William Wordsworth

20 "Two Voices, The"

17 "Palace of Art, The"

15 "Lotus Eaters, The"
15 "Ulyssus"

13 "Mariana"
13 "Recollections of the Arabian Nights"

12 Hallam, Arthur Henry

525
PT7244.K57 1946-49

Kock, Ernst Albin, ed.
Den Norsk-Islandska Skaldediktningen

Sum= 626 / 29.54 = 21

2 Bjark: Bjarkamål, anon.
2 Danir. Danir, anon.
2 Finng: Finngálkn, anon.
2 Jómsvikingar anon.
2 Karlevi: Karlevistenens drottkvädade vers, anon.
2 Oddm.: Oddmjor anon.
2 Rauðsk.: Rauðskeggr. anon.
2 Sveinn tjuguskegg, anon.
2 Svtjúg: Sveinn tjuguskegg, anon.
2 Tångbrand och Gudlev, dikt om, anon.
2 Vagn: Vagn Akason anon.
2 AEvidråpa (orvar-odds): ur Qrvar-odds saga

Ostrowski, Alexander
Vorlesungen über Differential und Integralrechnung

Sum= 868 / 29.54 = 29

14 Cauchy, A.

9 Euler

6 Cantor, G.
6 Dirichlet
6 Gauss
6 Hardy, G.H.
6 Weierstrass

5 Abel
5 Ellipse
5 Hermite
5 Konvergenzkriterien für uneigentliche Integrale
5 Schwarz, H.A.

4 Bertand, J.
4 Bolzano
4 Cesàro
4 Hausdorff
4 Jensen, J.L.W.V.
4 Newton
4 Pringsheim, A.
4 Riemann, B.

3 Caratheódory
3 Cauchy-Bolzanosches Konvergenzkriterium
3 Chaundy
3 Enveloppe
3 Fresnelsche Integrale
3 Hadamard
3 Inhalt
3 Konvergenzkriterium für unendliche Cauchy -Bolzanoshes
3 Poisson
3 Stieltjes
3 Vergluchskriterium für unendliche -uneigentliche Integrale
3 Zusammenhängend

575
QD9.S71 Ref. 19

Soule, Byron Avery
Library Guide for the Chemist

Sum= 1833 / 29.54 = 28

5 Gregory

4 Böttger, Wm.
4 Furman, N.H.
4 Water, analysis of

3 Biography, German
3 Browne, C.A.
3 Classen
3 Daniels, F.
3 Dyes, patents on
3 Ferro-alloys, analysis
3 Findlay, A.
3 Glasstone, S.
3 Hahn, D.
3 Hall, W.T.
3 Houben, J.
3 Indexes, patent
3 Koltoff, I.M.
3 Martin, G.
3 Meyer, R.J.
3 Mnemonics
3 Nomenclature, organic
3 Organo-metallic Compounds
3 Ostwald, Wm.
3 Patents, dye
3 Rossman, J.
3 Steel, analysis of
3 Sugar, analysis
3 Thorpe, Edw.
3 Weiser, H.B.
3 Worden, E.C.

Varley, Ernest Reginald
Sillimanite

600
QE391.S5 V3 1965

Sum= 729 / 29.54 = 25

11 Reserves, India

10 Andalusite: U.S.A.

8 Kenya
8 Reserves, U.S.A.

7 Assam, India
7 United States

6 Florida, U.S.A.
6 Georgia, U.S.A.

5 Beneficiation, U.S.A.
5 Bihar, India
5 Brazil
5 California, U.S.A.
5 Dumortierite, U.S.A.
5 Mysore, India
5 Nyasaland
5 South Africa, Republic of
5 Topaz: U.S.A.

4 Aluminium industry
4 Andalusite: U.S.S.R.
4 Andhra Pradesh, India
4 Baker Mountain, Virginia
4 Density
4 Graves Mountain, Georgia
4 Henry Knob, S. Carolina
4 India
4 Kerala, India
4 Kyanite density
4 Lapsa Buru, Bihar
4 Madhya Pradesh, India
4 Maharashtra, India
4 Nevada, U.S.A.
4 New South Wales, Australia
4 Orissa, India
4 Reserves, U.S.S.R.
4 Sillimanite minerals:density
4 South Carolina U.S.A.
4 Transvaal, South Africa
4 United States, National Stockpile Purchase Specification

Davis, David Edward
Principles in Mammalogy

Sum= 836 / 29.54 = 28

11 Carnivores

10 Bat(s)
10 Insectivores
10 Woodchucks

9 Marsupials
9 Mutation
9 Teeth

8 Monotremes
8 Whale(s)

7 Herbivores
7 Opossums
7 European rabbit(s)
7 Raccoons

6 Dispersal
6 Maintenance
6 Predators
6 Primates
6 Shrew(s)
6 Vole, meadow

5 Body size, temperature
5 Camels
5 Competition
5 Corpora lutea
5 Elephants
5 Feedback
5 Food
5 Fossils
5 Migration
5 Mole
5 Muskrats
5 Nearctic region
5 Omnivores
5 Oriental region
5 Sex ratio
5 Squirrel(s), ground
5 Temperature

650
RM721.E8 1947

Ewerhardt, Frank Henry
Therapeutic Exercise

Sum= 338 / 29.54 = 11

8 Muscle contraction
8 Paralysis
8 Posture

7 Spastic paralysis, exercise in

6 Flat foot
6 Muscle function volitional tests
6 Poliomyelitis treatment
6 Re-education

5 Hemiplegia
5 Lordosis
5 Poliomyelitis testing by topographical observations
5 Poliomyelitis treatment during acute stage
5 Re-education of upper extremity
5 Scoliosis
5 Upper extremity re-education
5 Zero position

675
TK153.U5 19

Underhill, Charles Reginald
Electrons at Work

Sum= 3827 / 29.54^2 = 4

9 Tube, gaseous-discharge

7 Hertz
7 Light, ultra-violet
7 Maxwell
7 Reaction, Reactors
7 Valence electrons

700
Z5782.A258 Ref. 1954

Bibliography of Medieval Drama
Stratman, Carl Joseph

Sum= 2797 / 29.54^2 = 3.2

44 Passion

37 Comedy

32 Latin

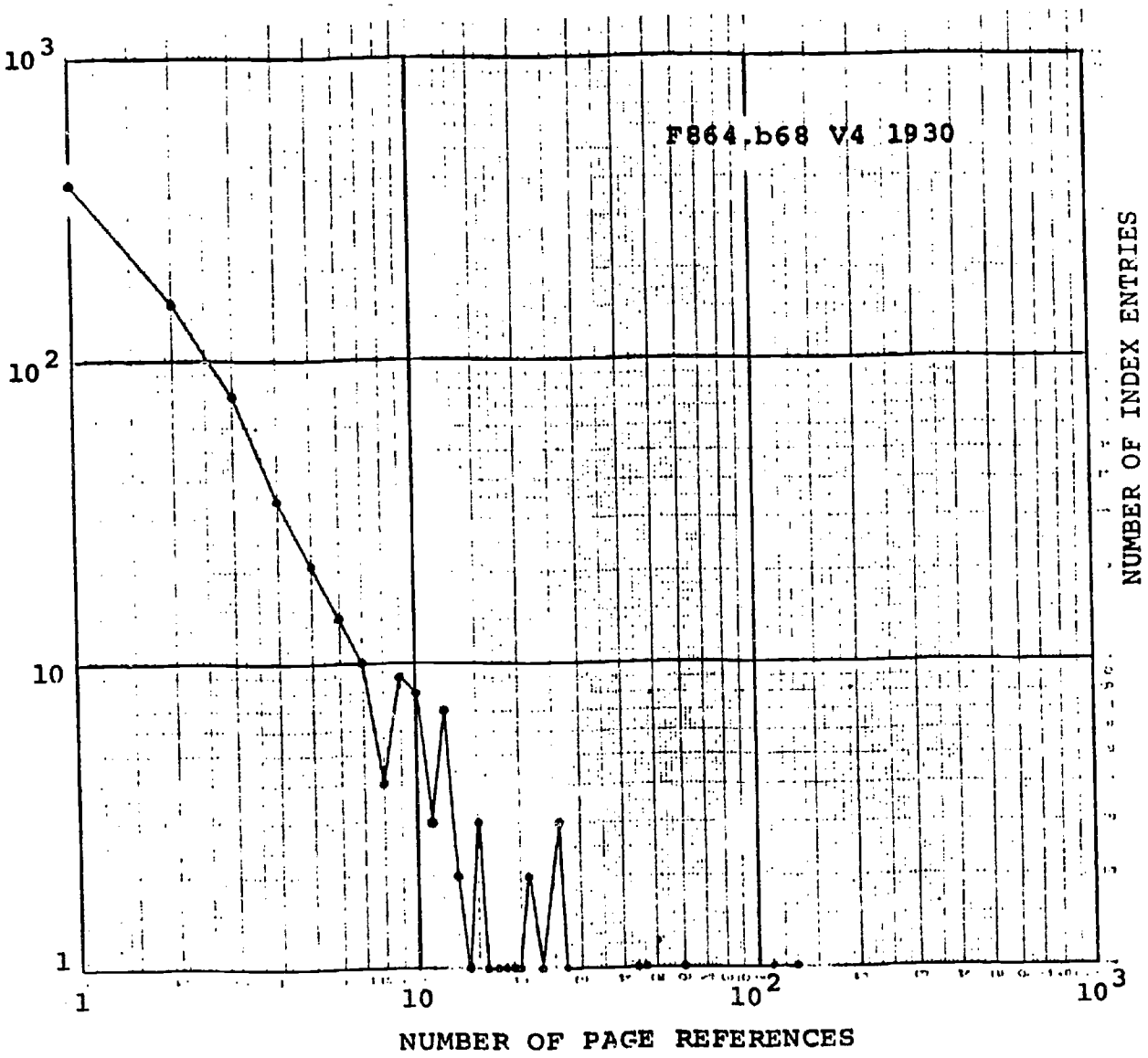
31 Staging

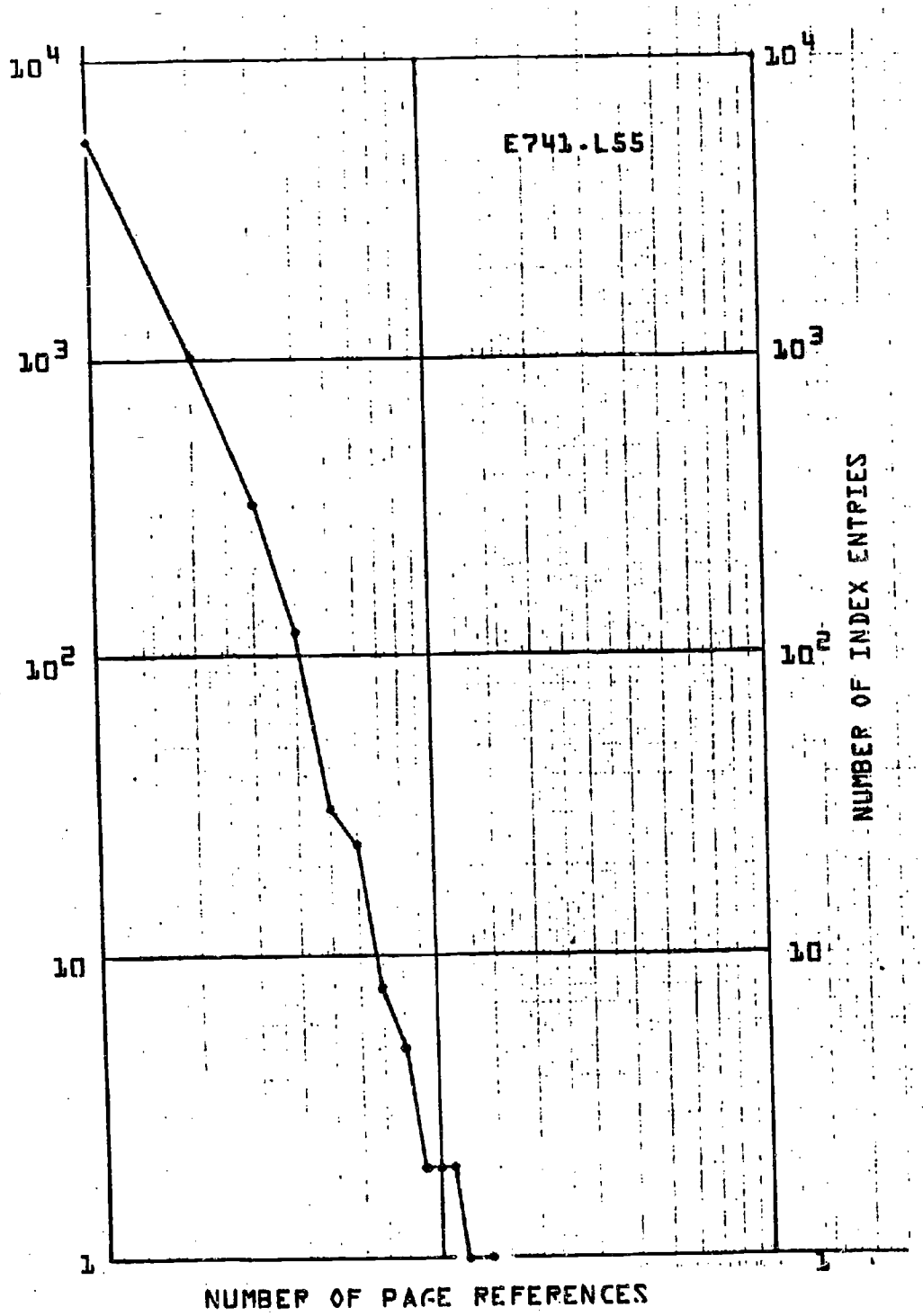
APPENDIX II

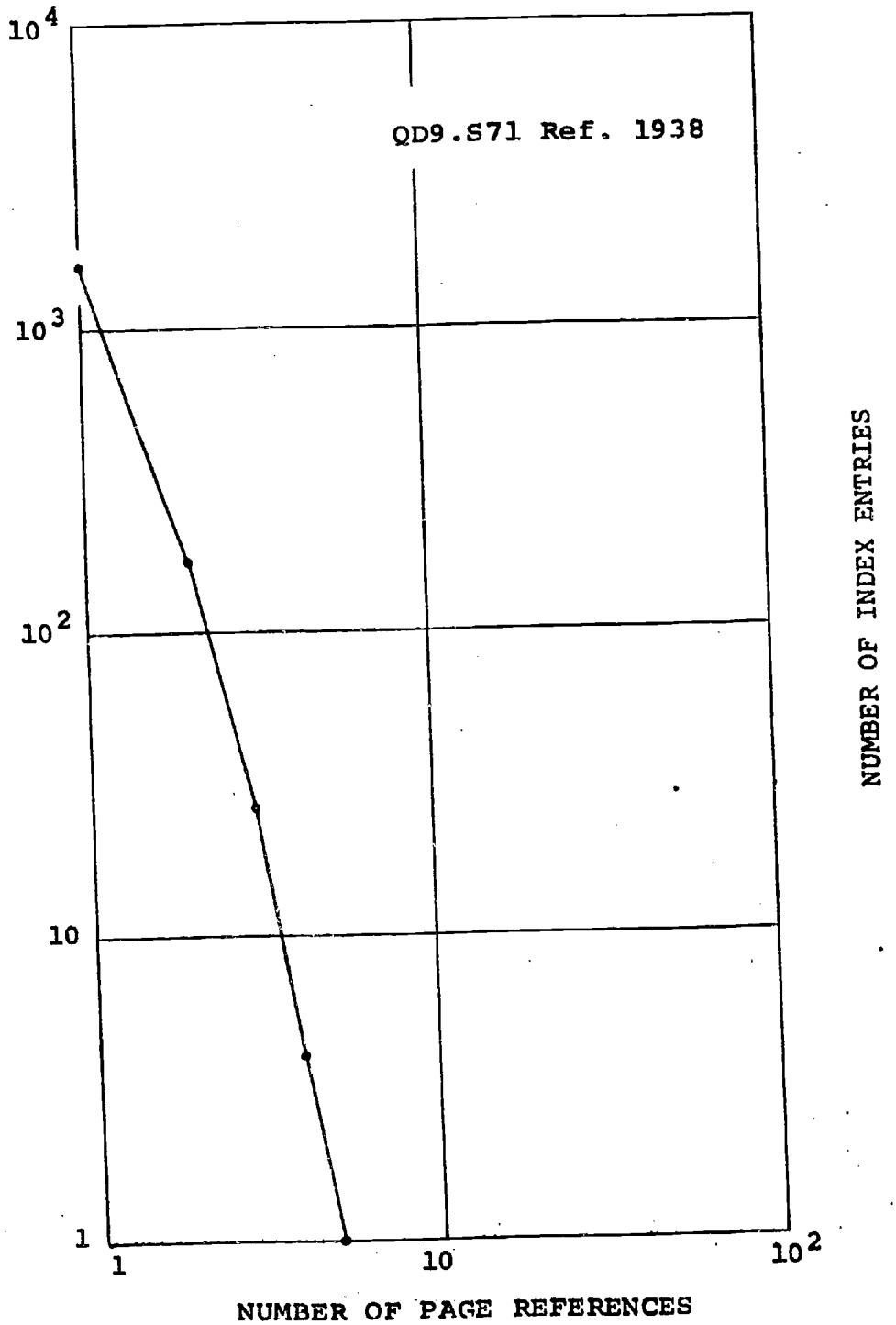
PAGE REFERENCE DISTRIBUTIONS

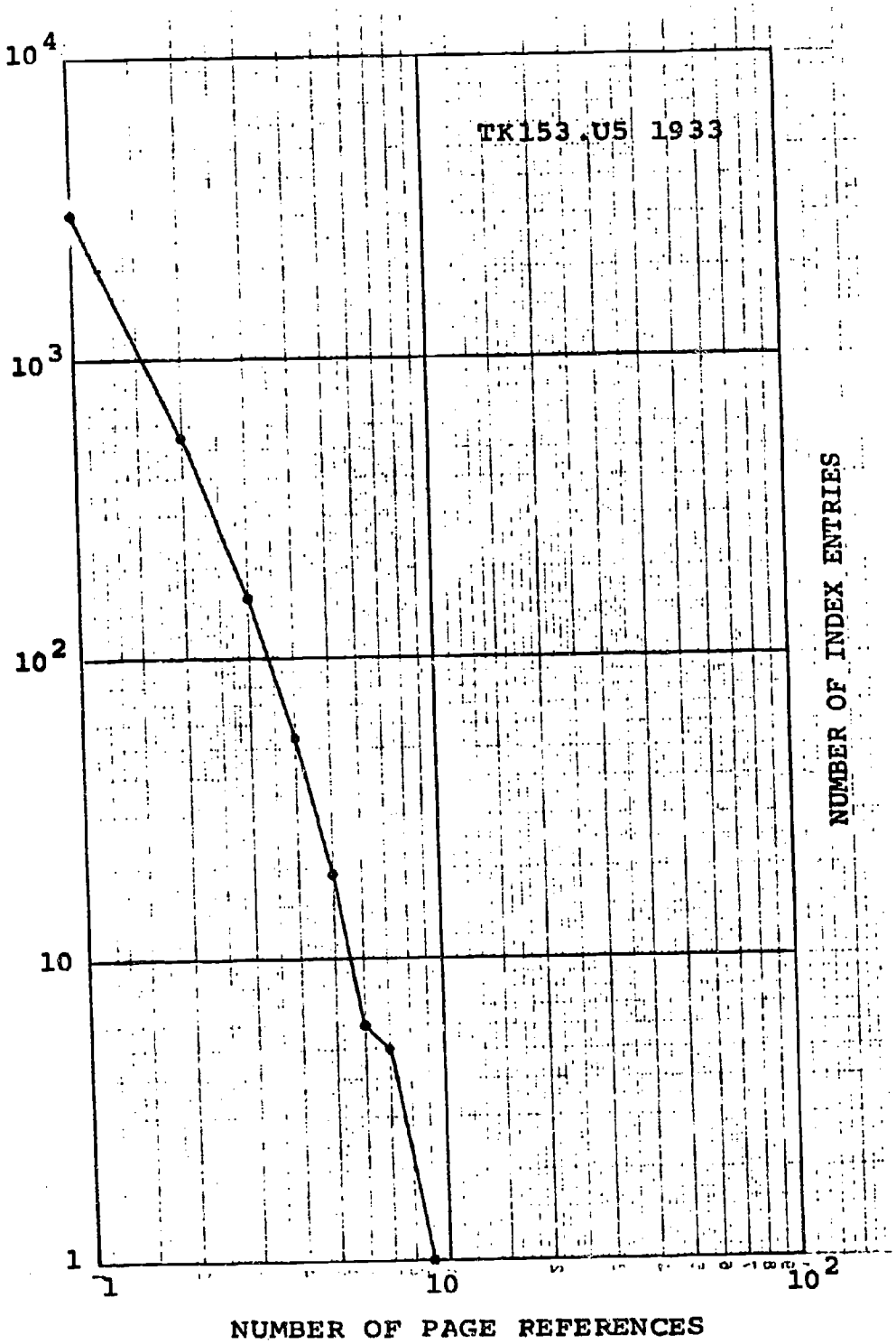
FROM

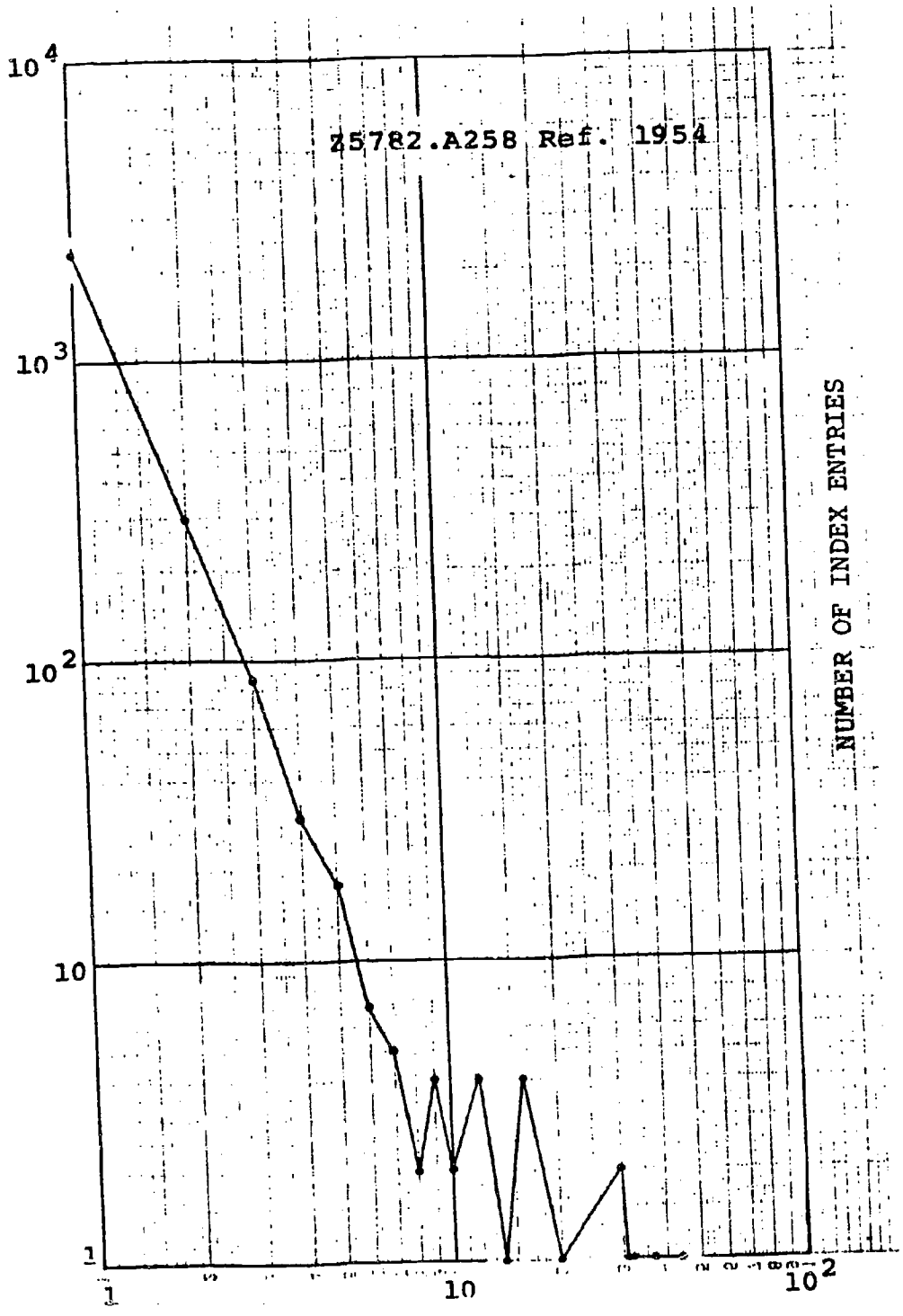
THE FONDREN INDEX SAMPLE

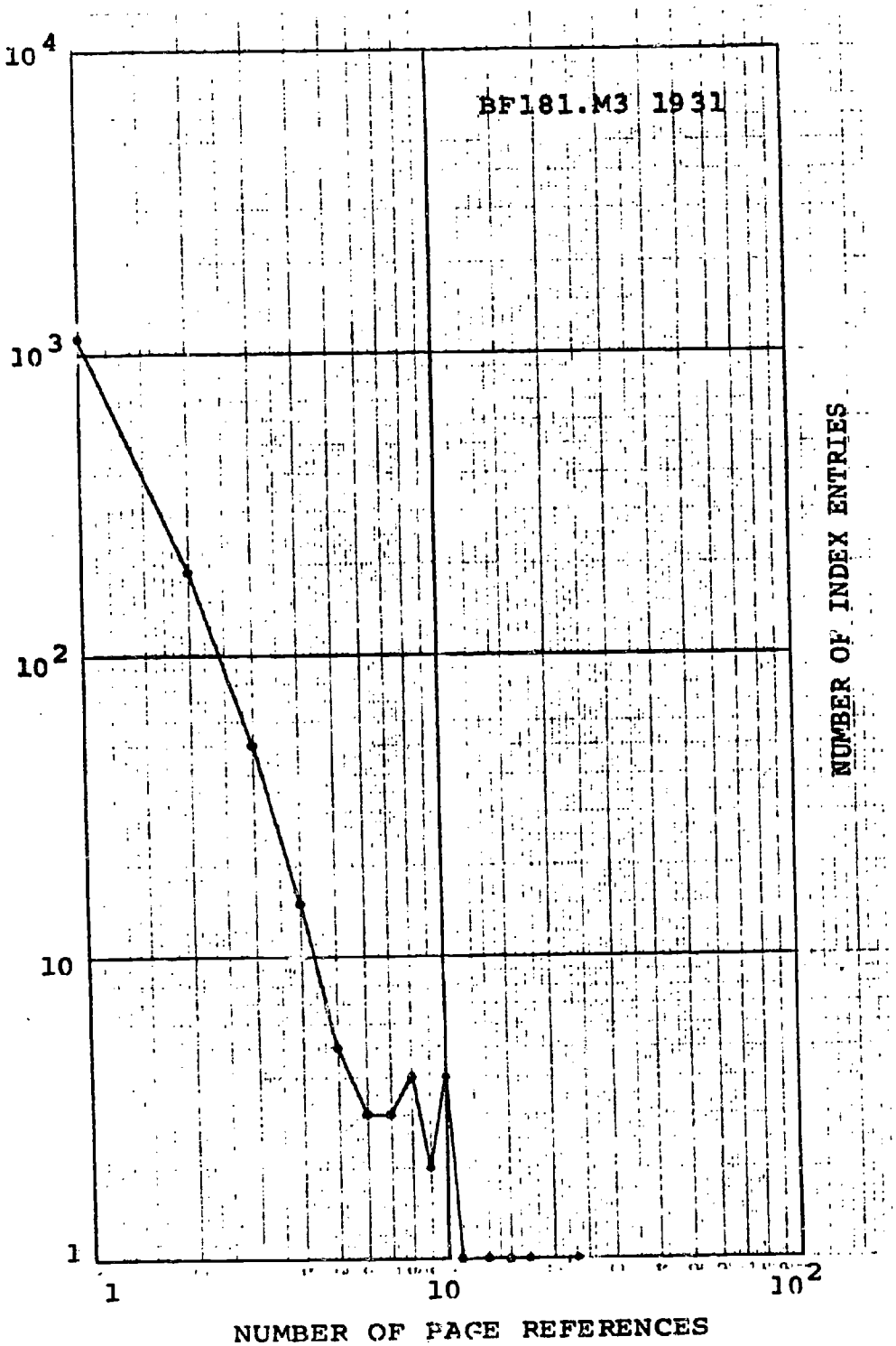

F864.b68 V4 1930
10^3
10^2
10
1
1
10
10^2
10^3
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

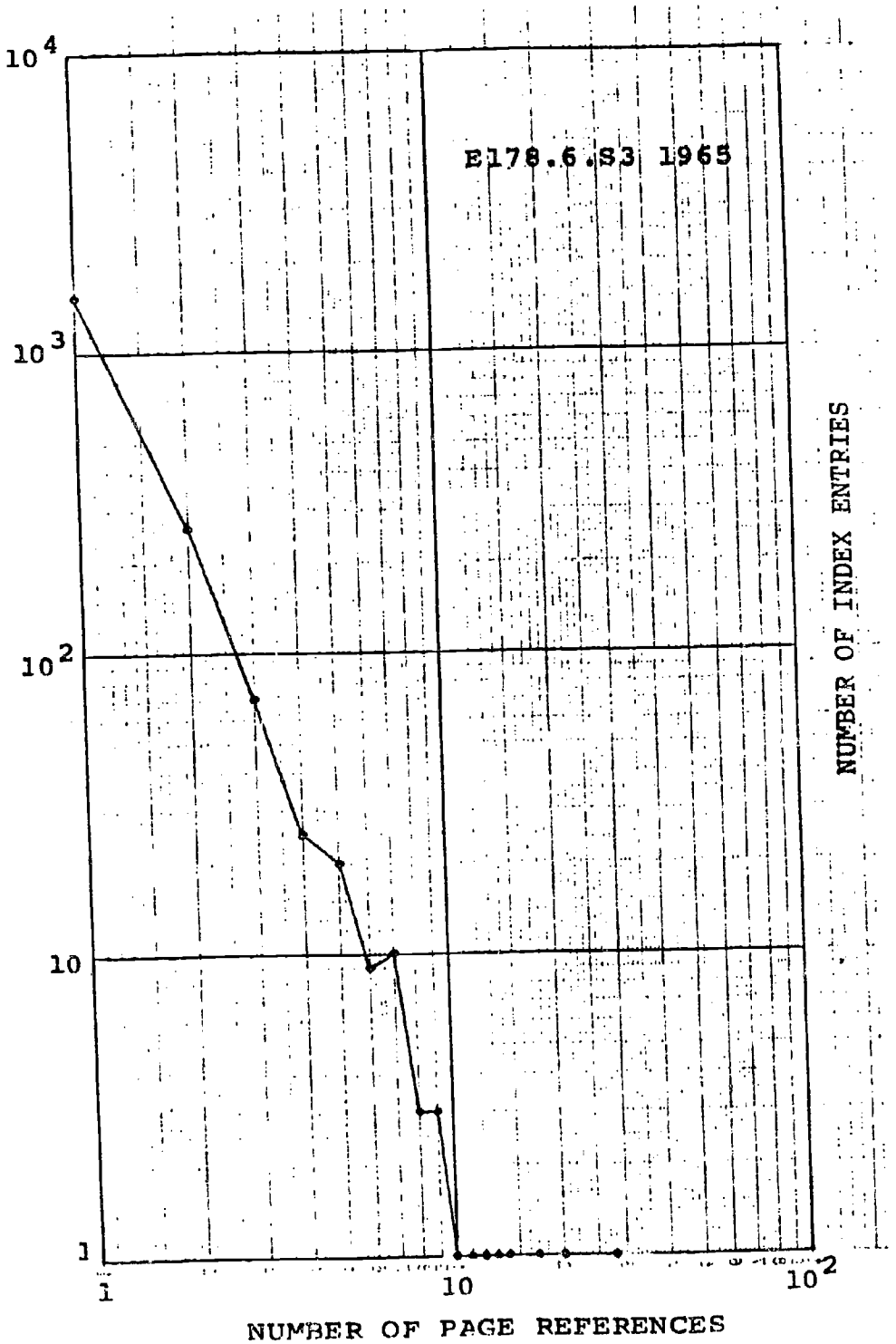

E741.L55
10^4
10^3
10^2
10
1
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

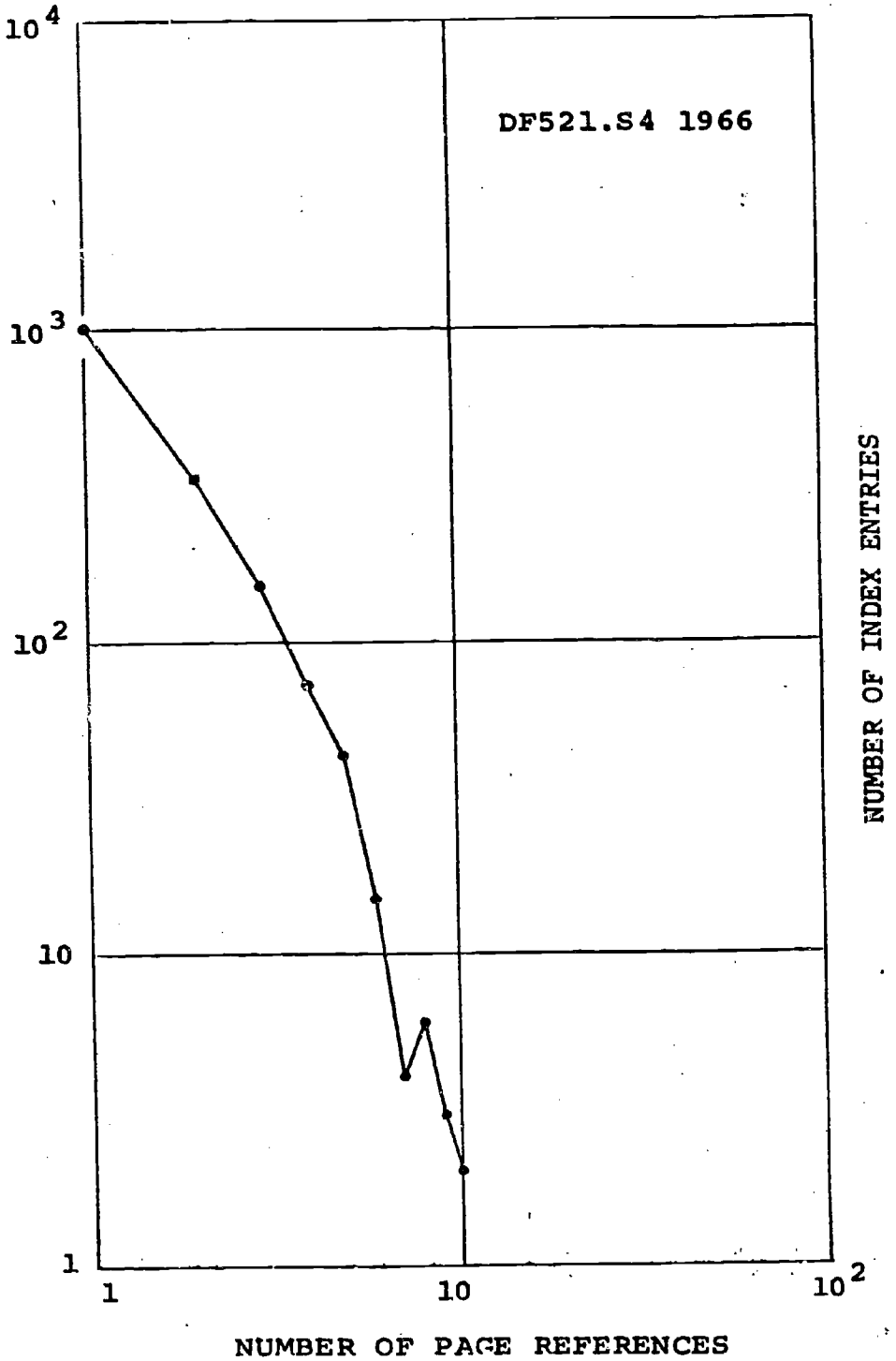

10^4
QD9.S71 Ref. 1938
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

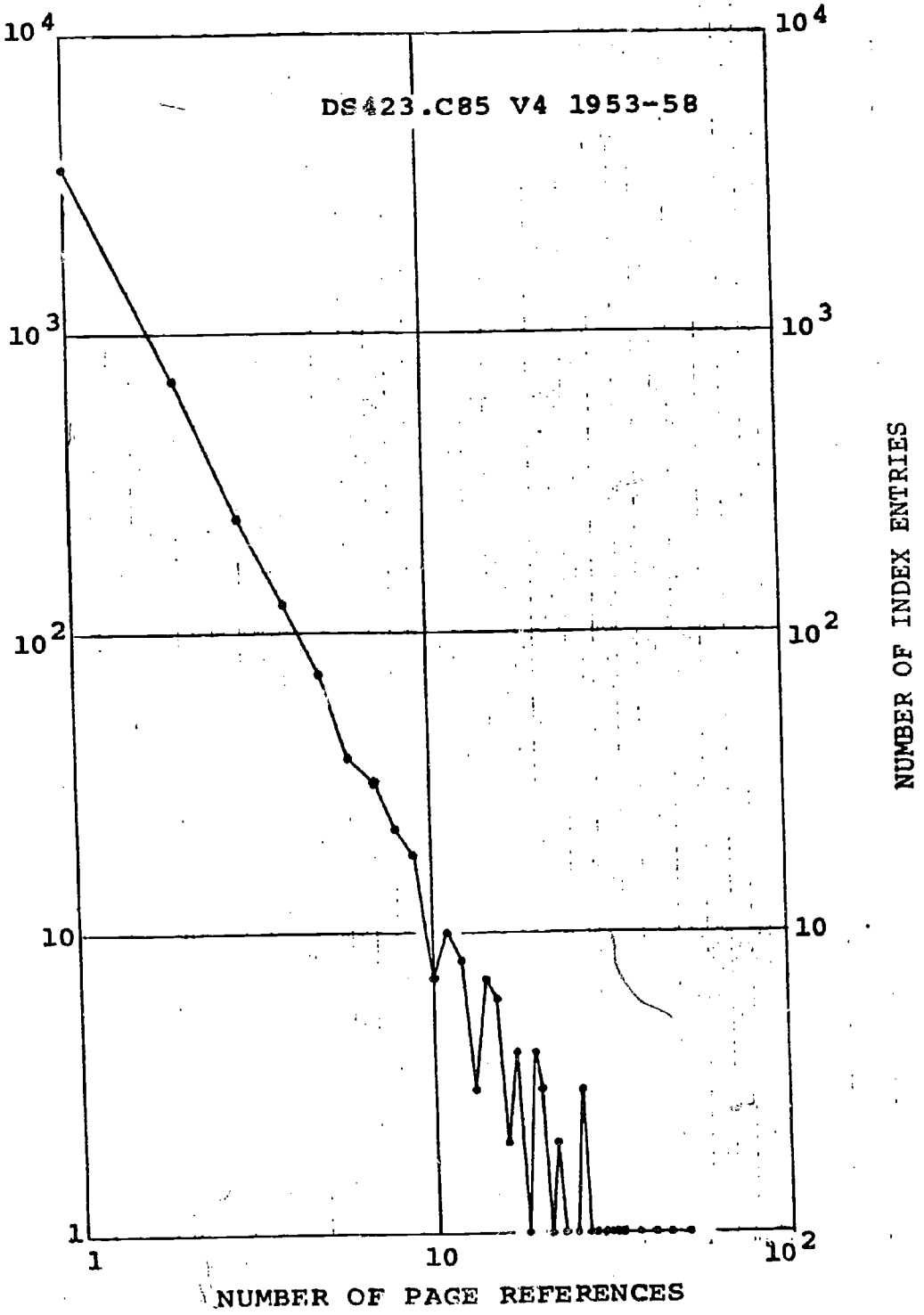

TK153 .U5 1933
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

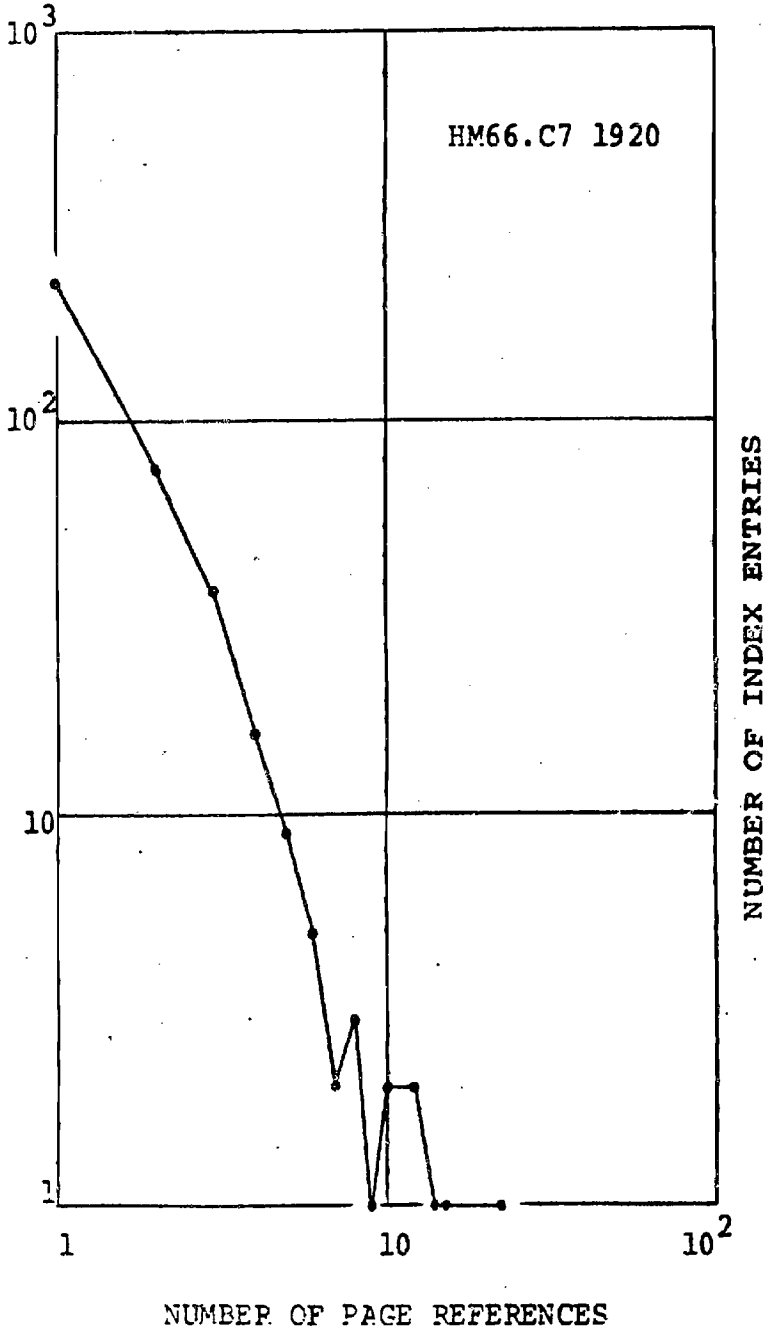

Z5782.A258 Ref. 1954
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

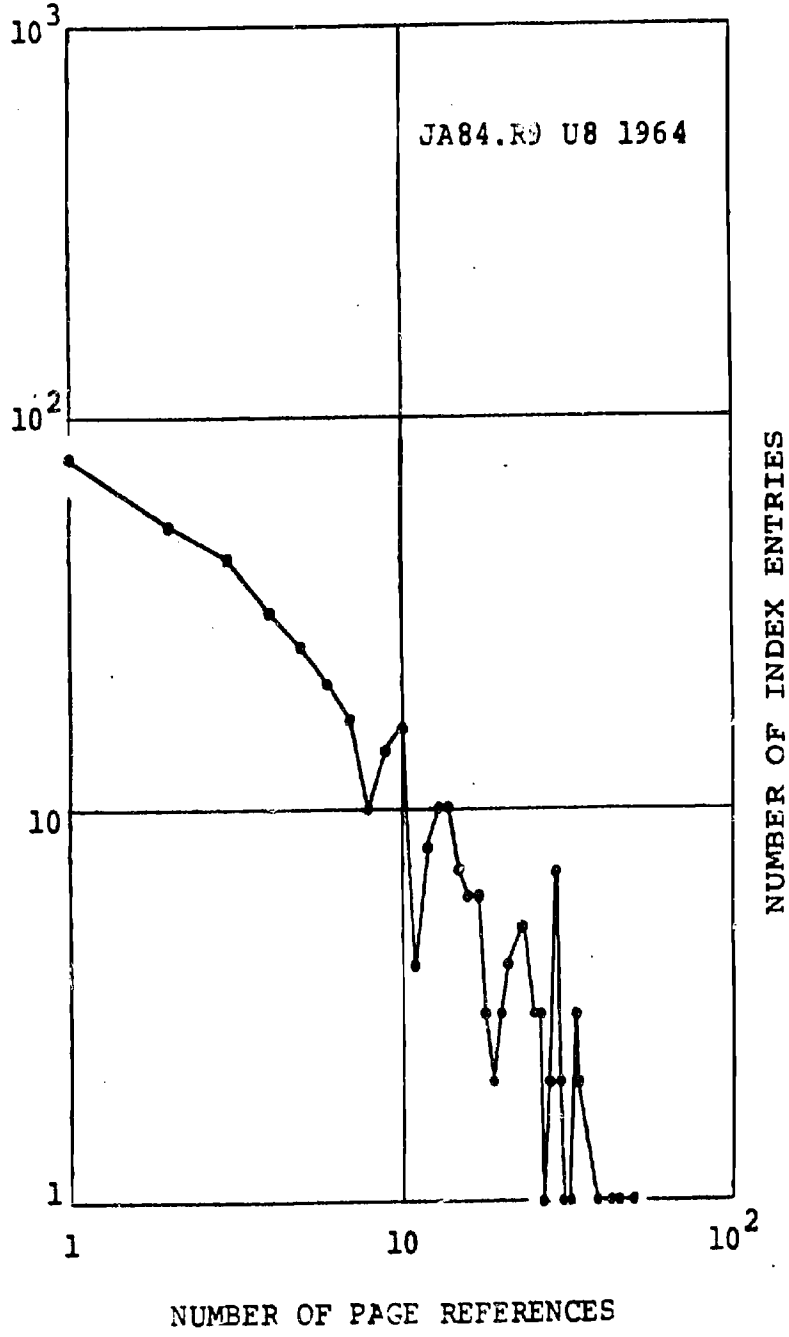

BF181.M3 1931
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

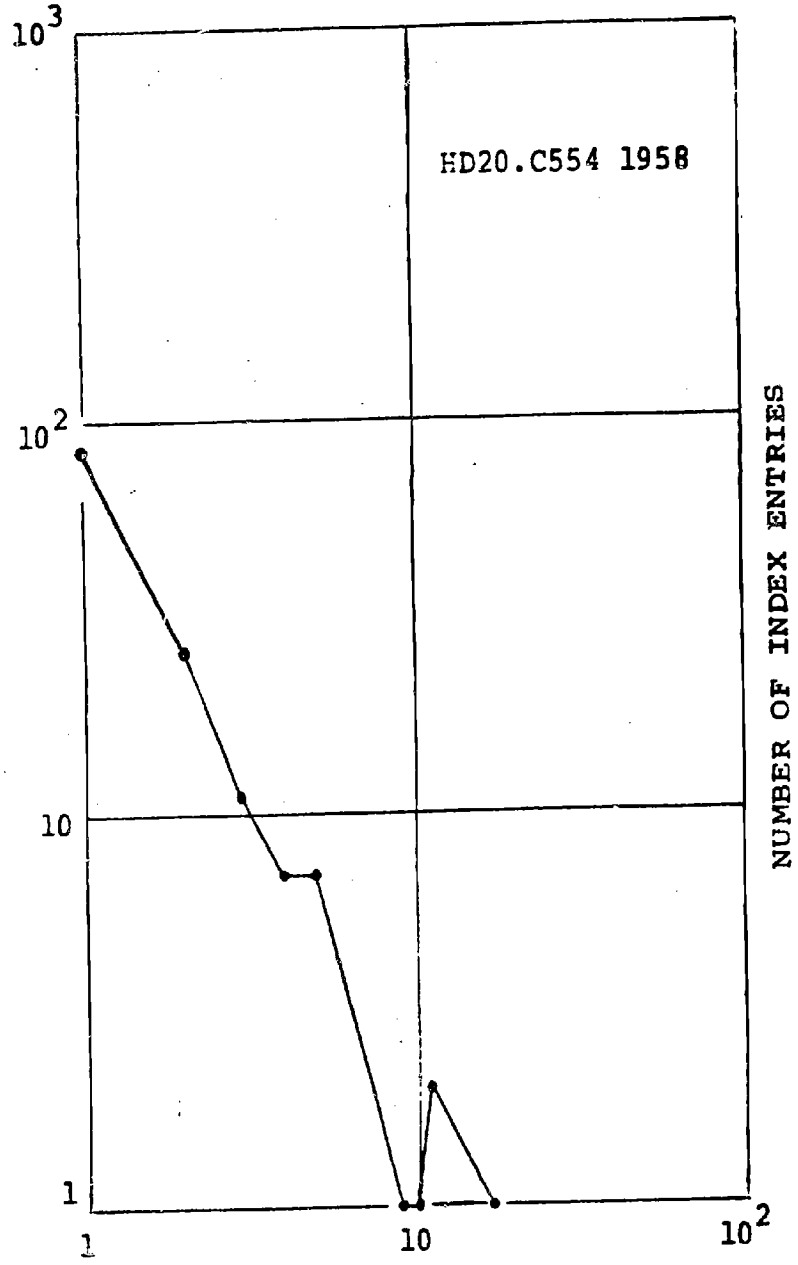


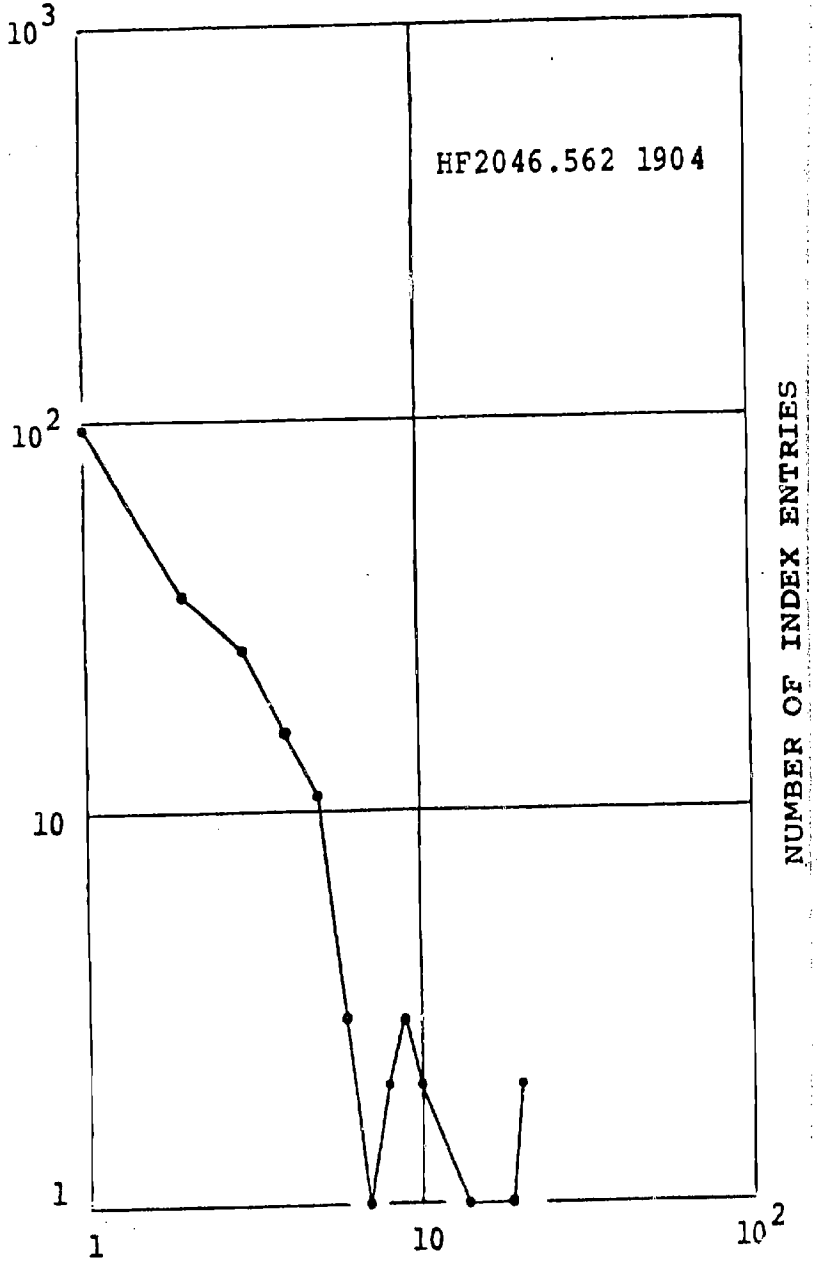

DF521.S4 1966
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

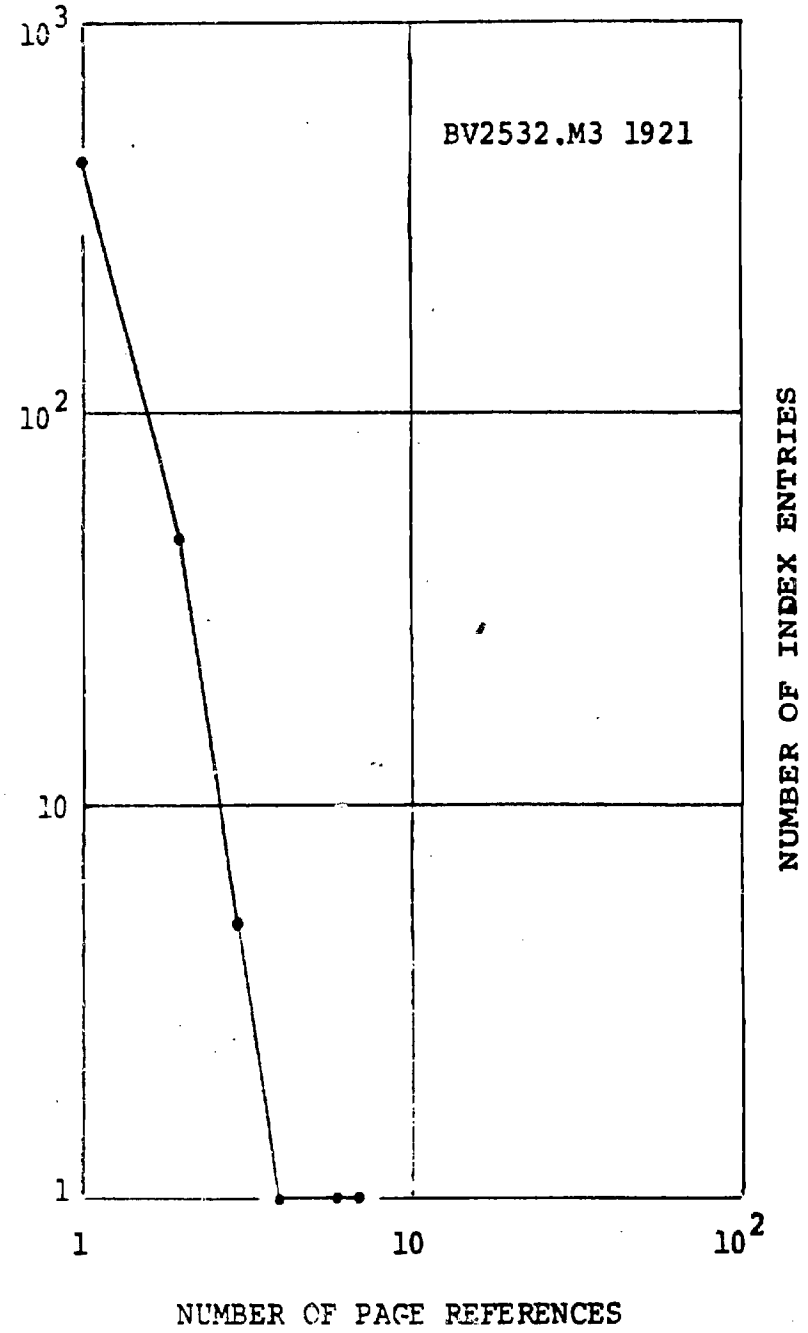

DS423.C85 V4 1953-58
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

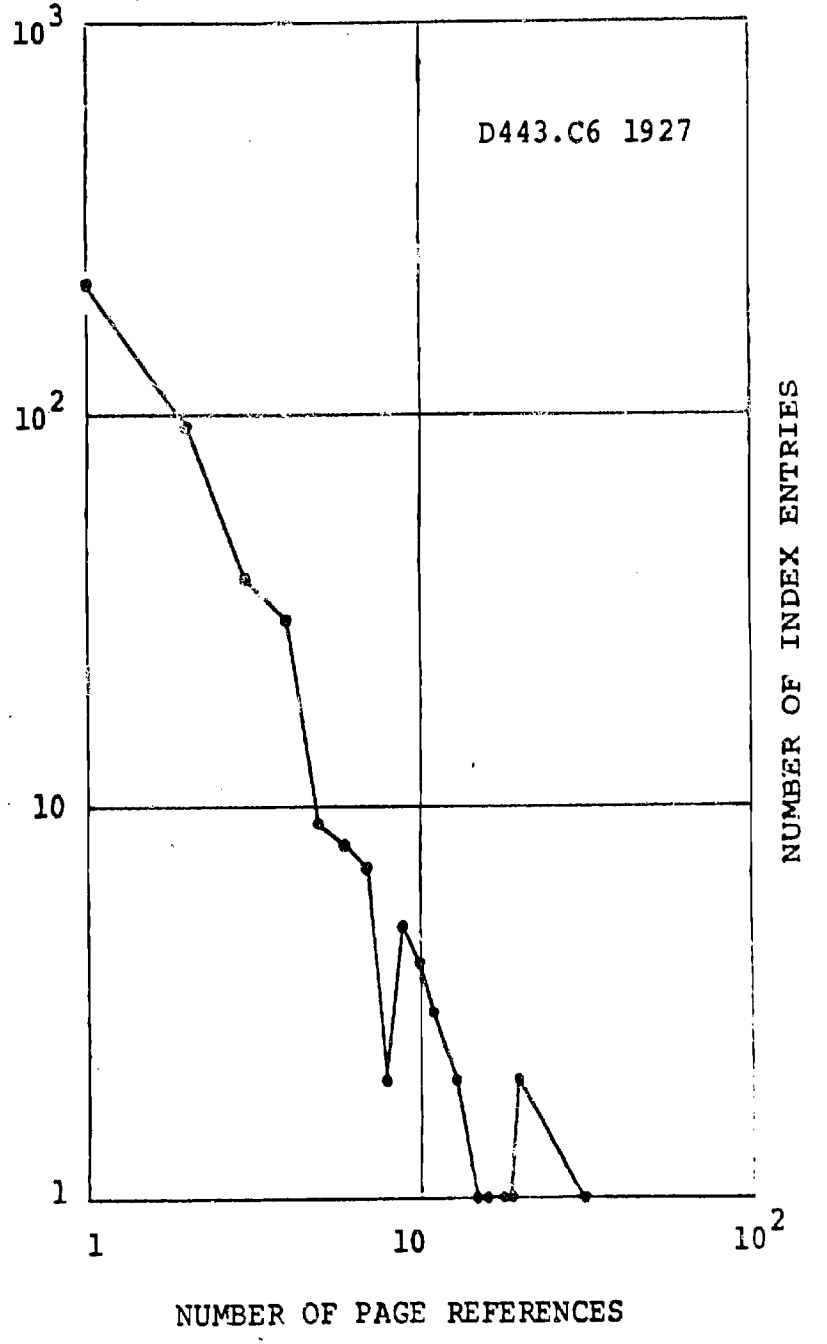

HM66.C7 1920
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

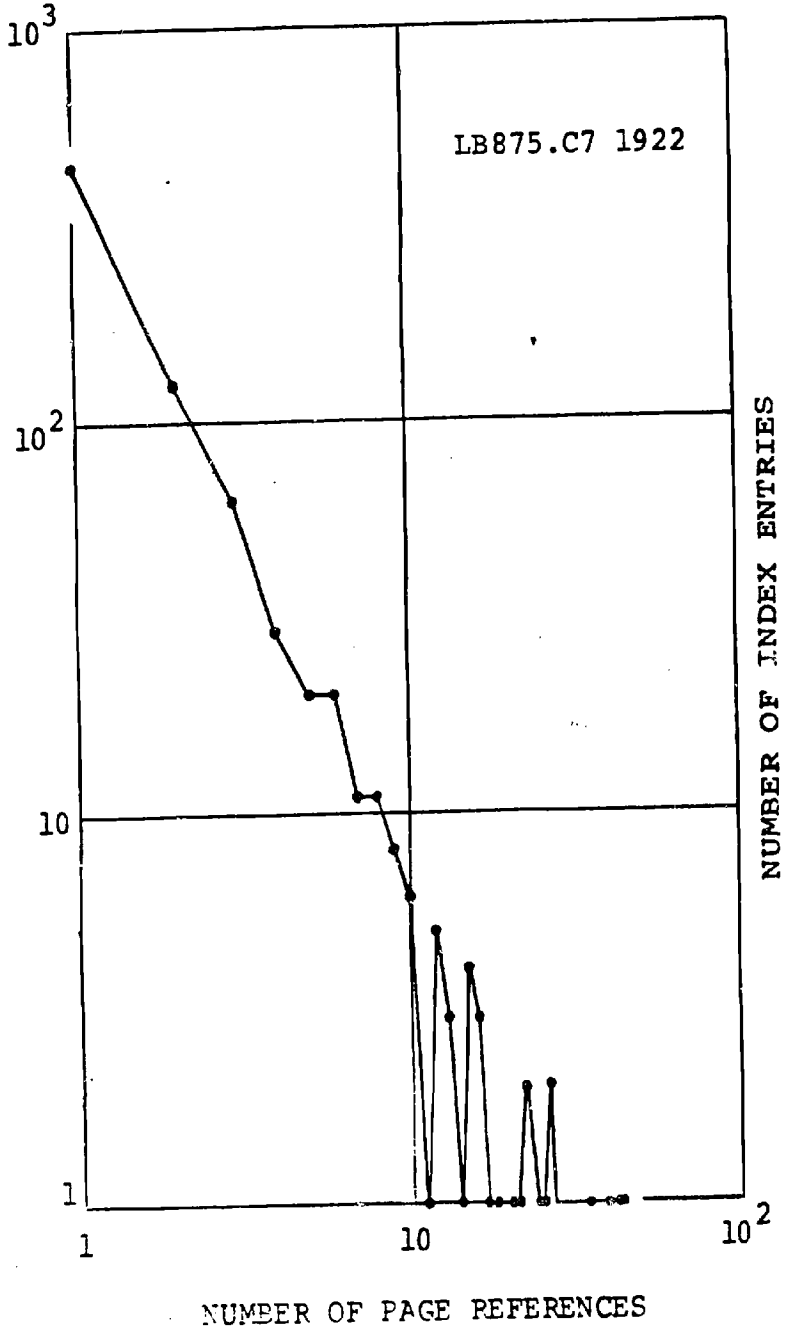

JA84.R9 U8 1964
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

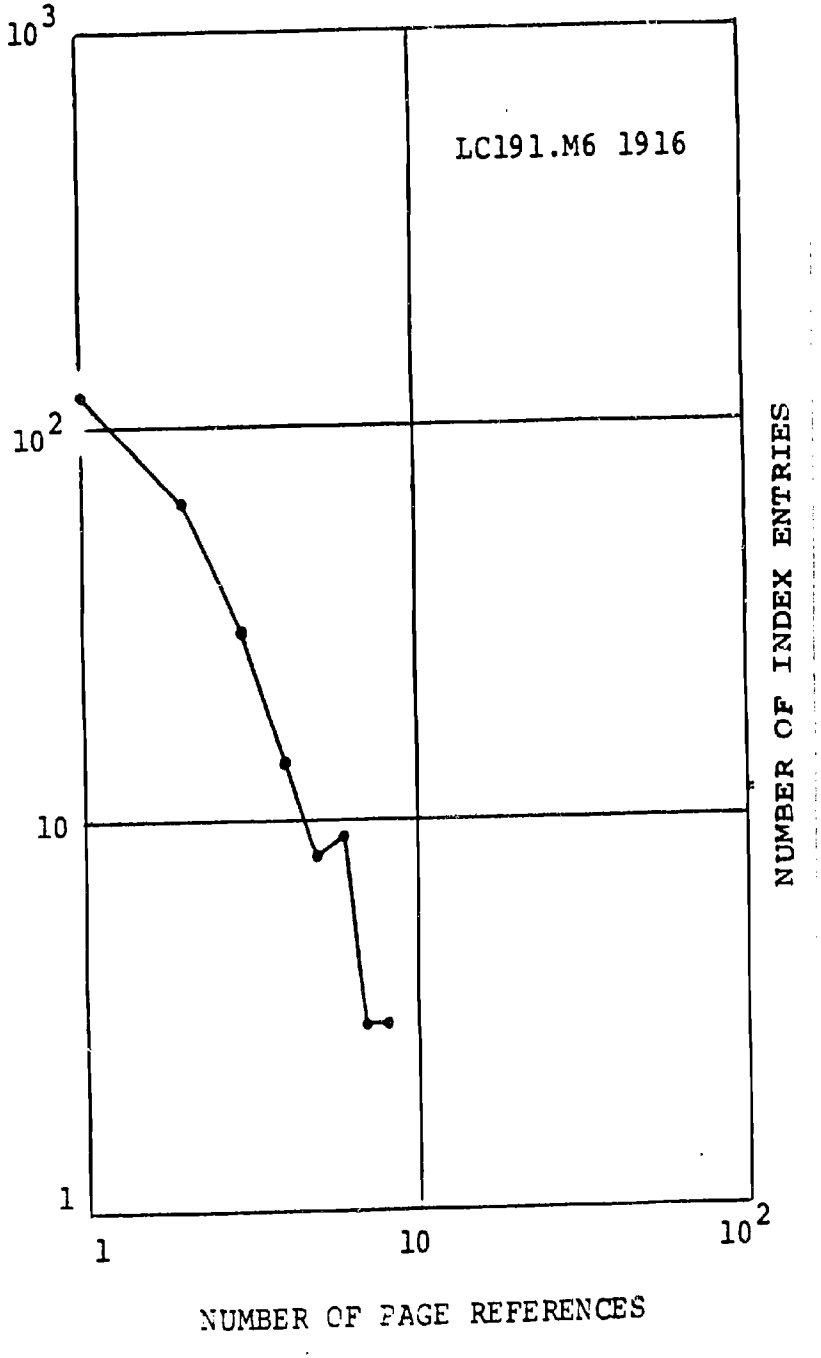

HD20.C554 1958
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES

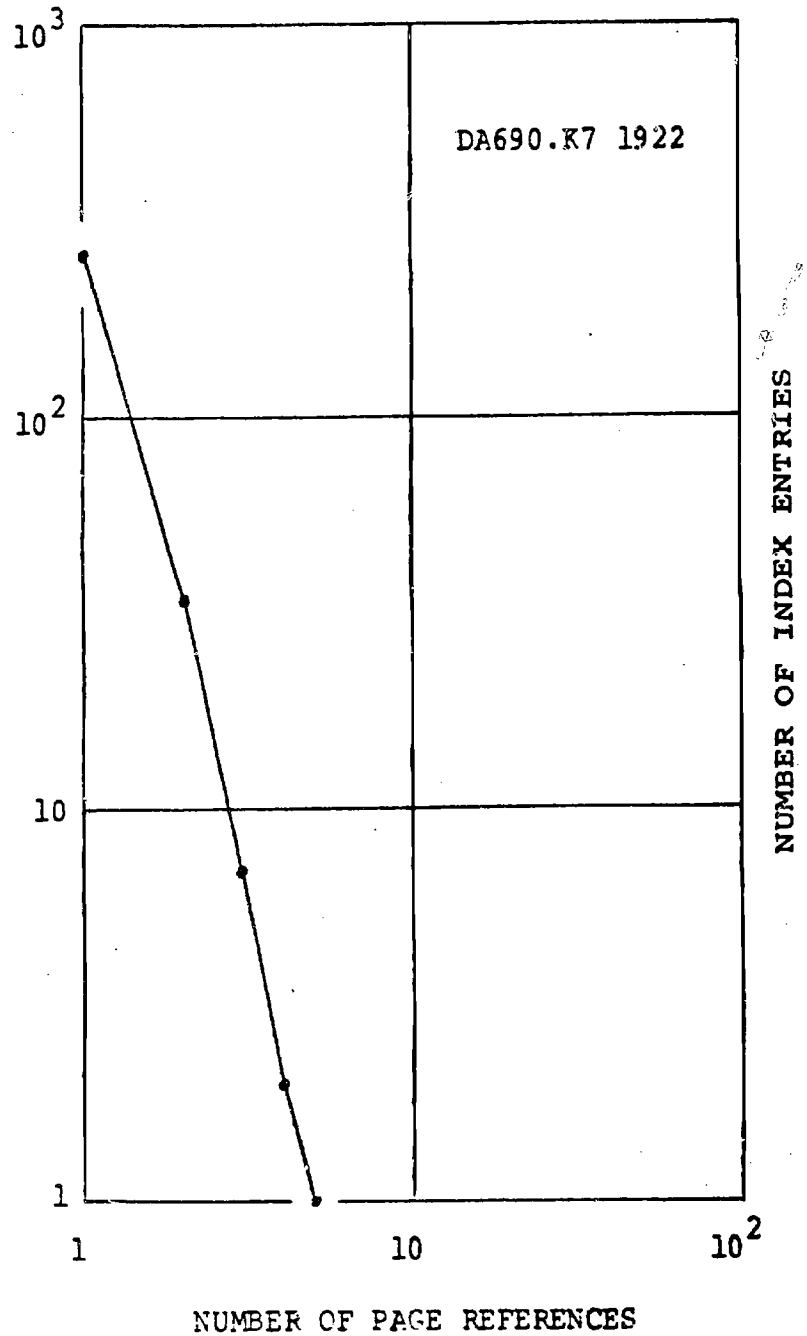

HF2046.562 1904
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

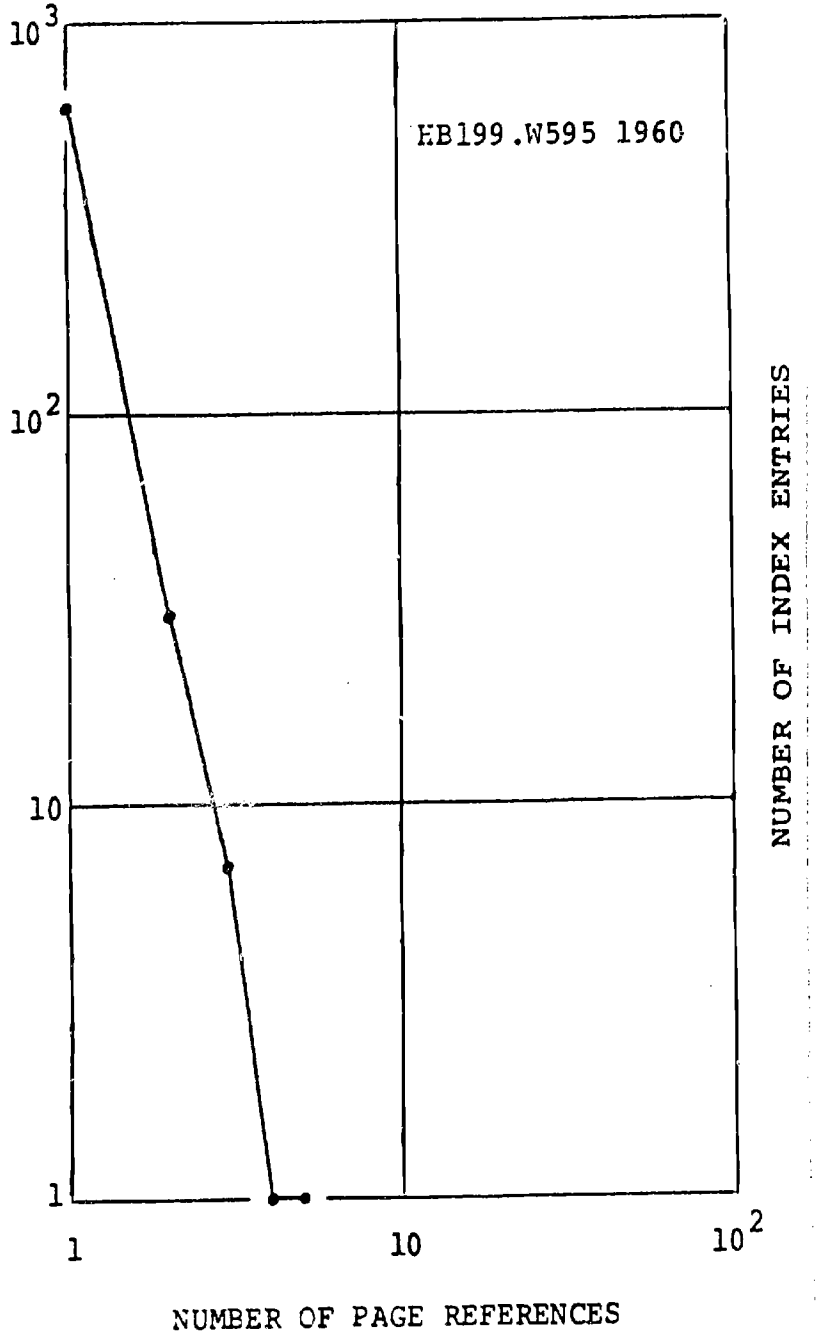

BV2532.M3 1921
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

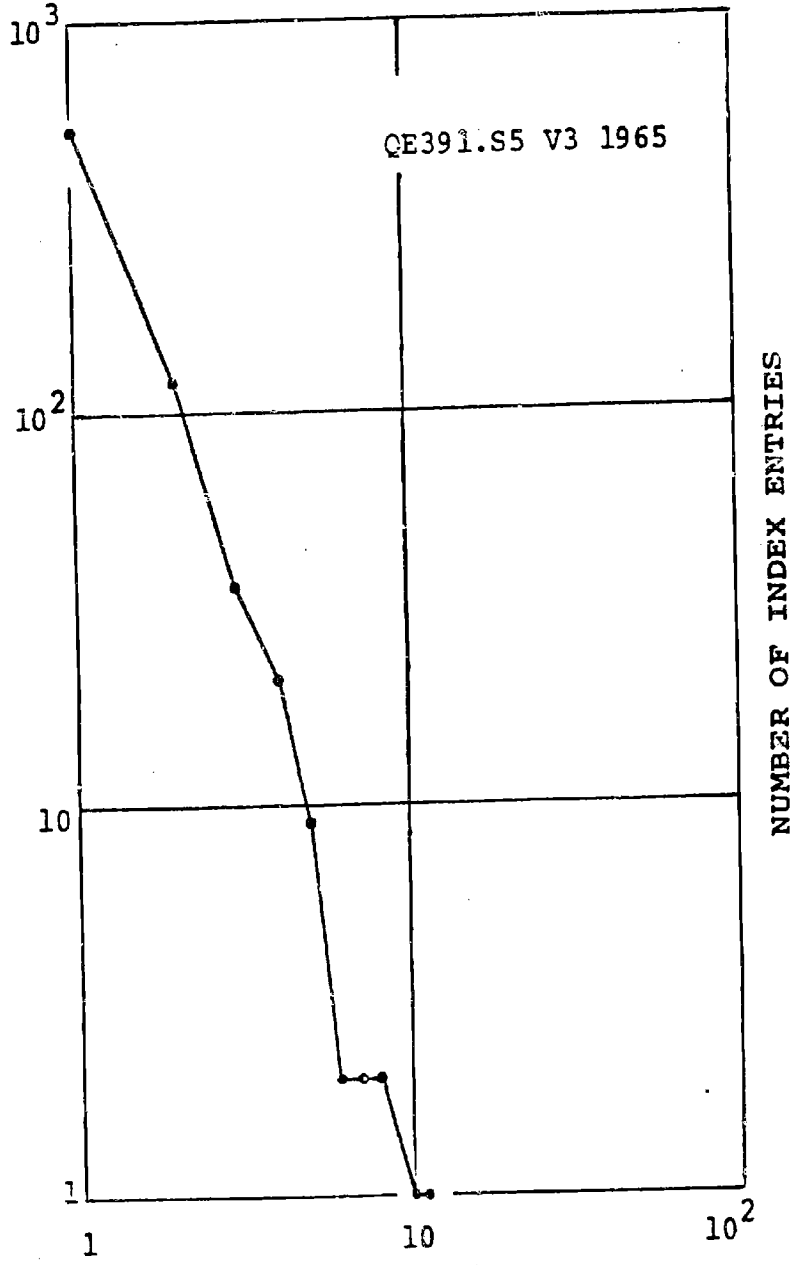

D443.C6 1927
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

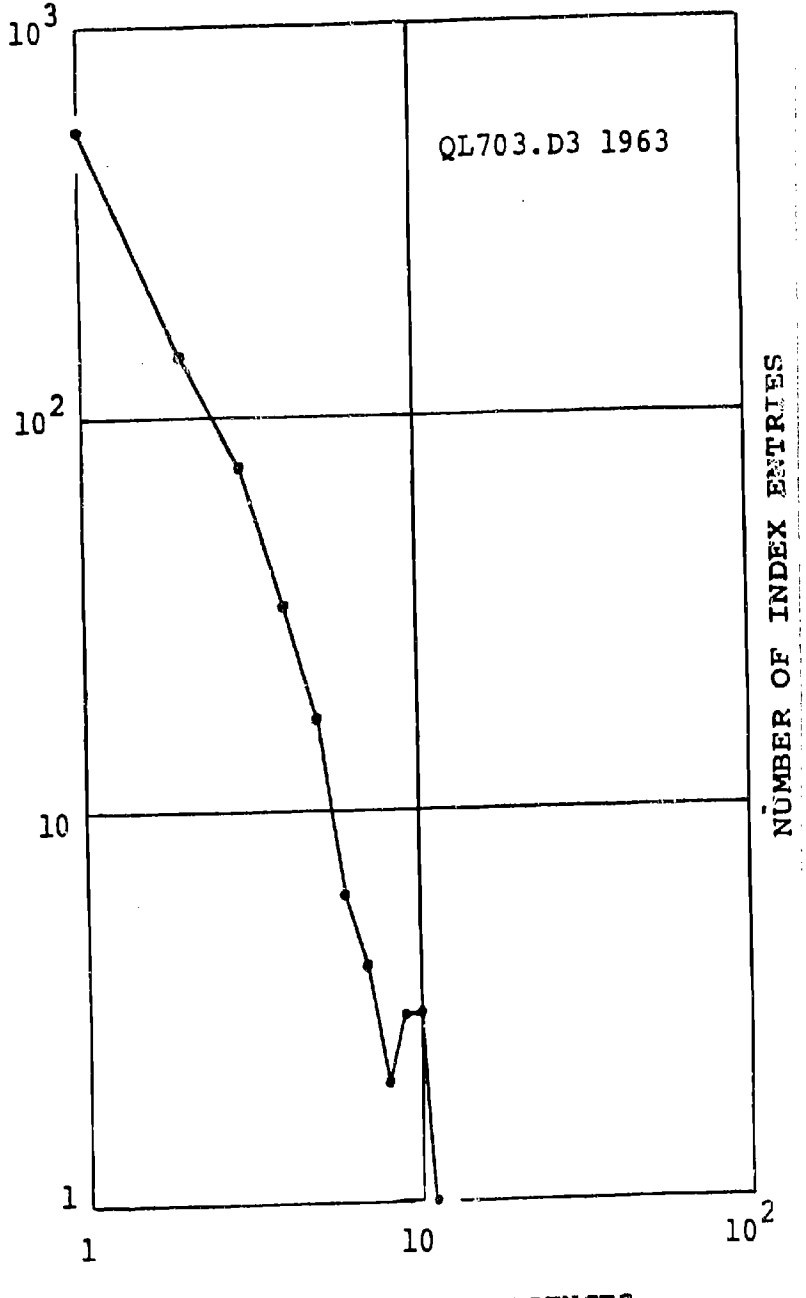

LB875.C7 1922
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

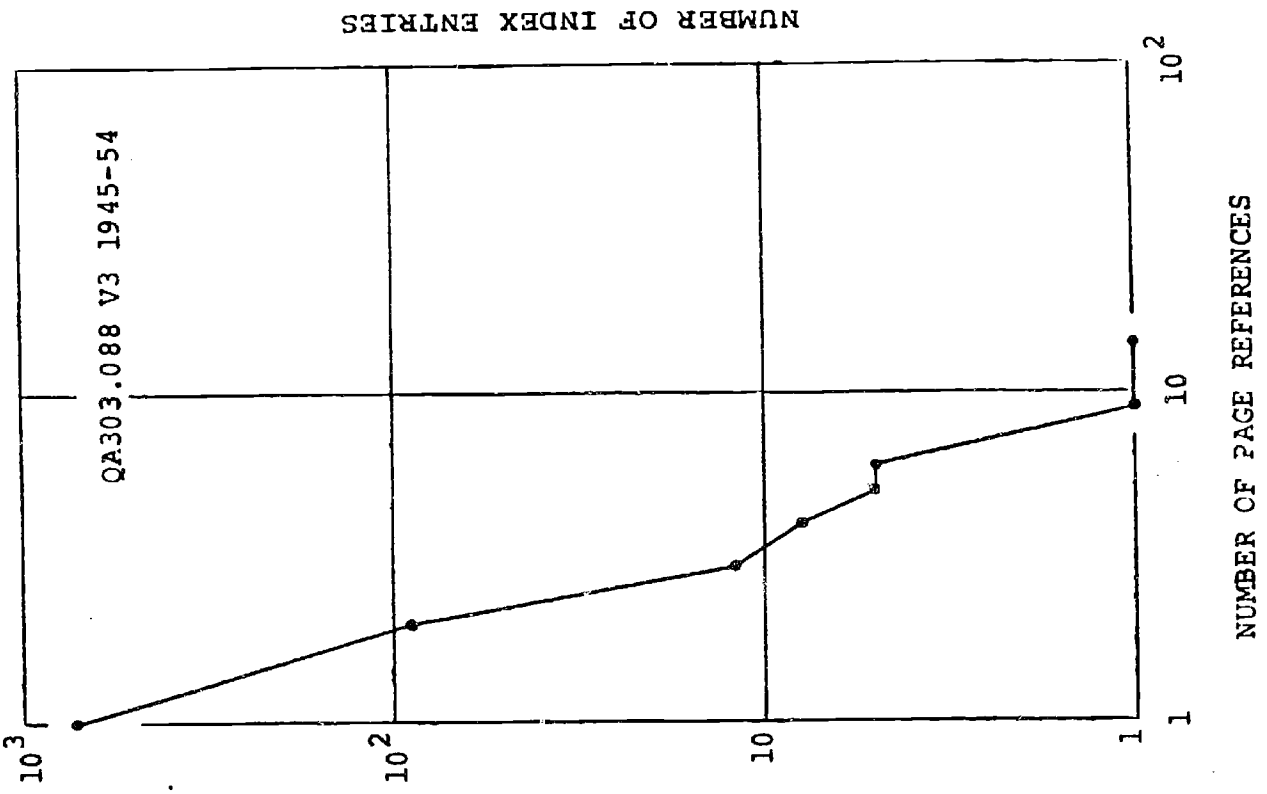

LC191.M6 1916
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

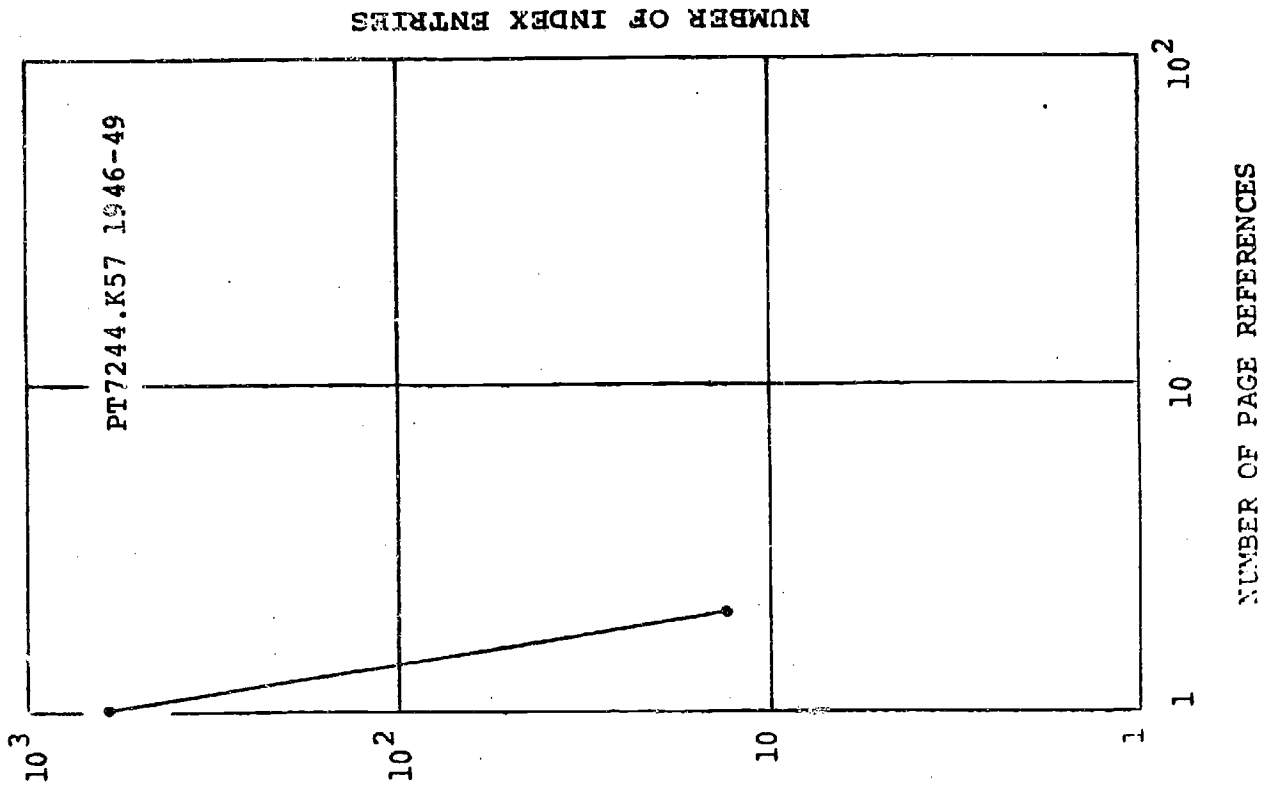

DA690.K7 1922
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

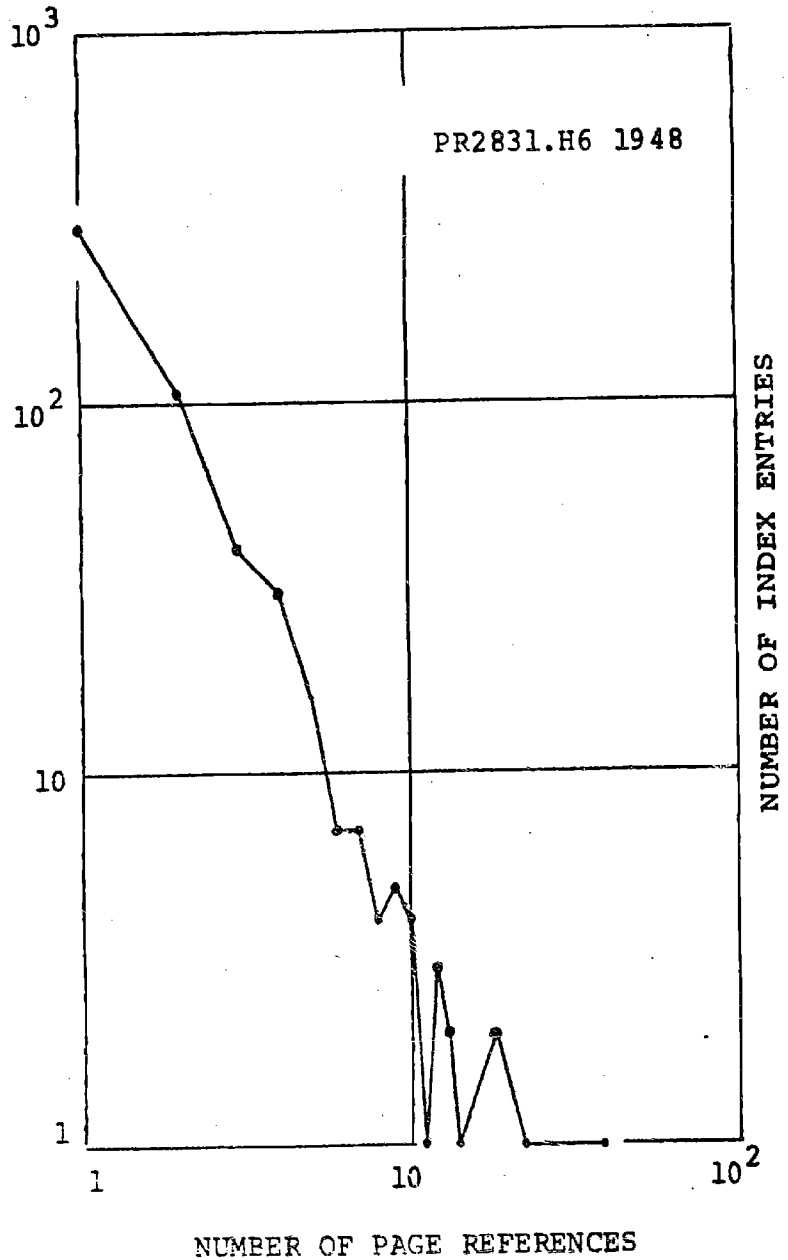

EB199.W595 1960
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

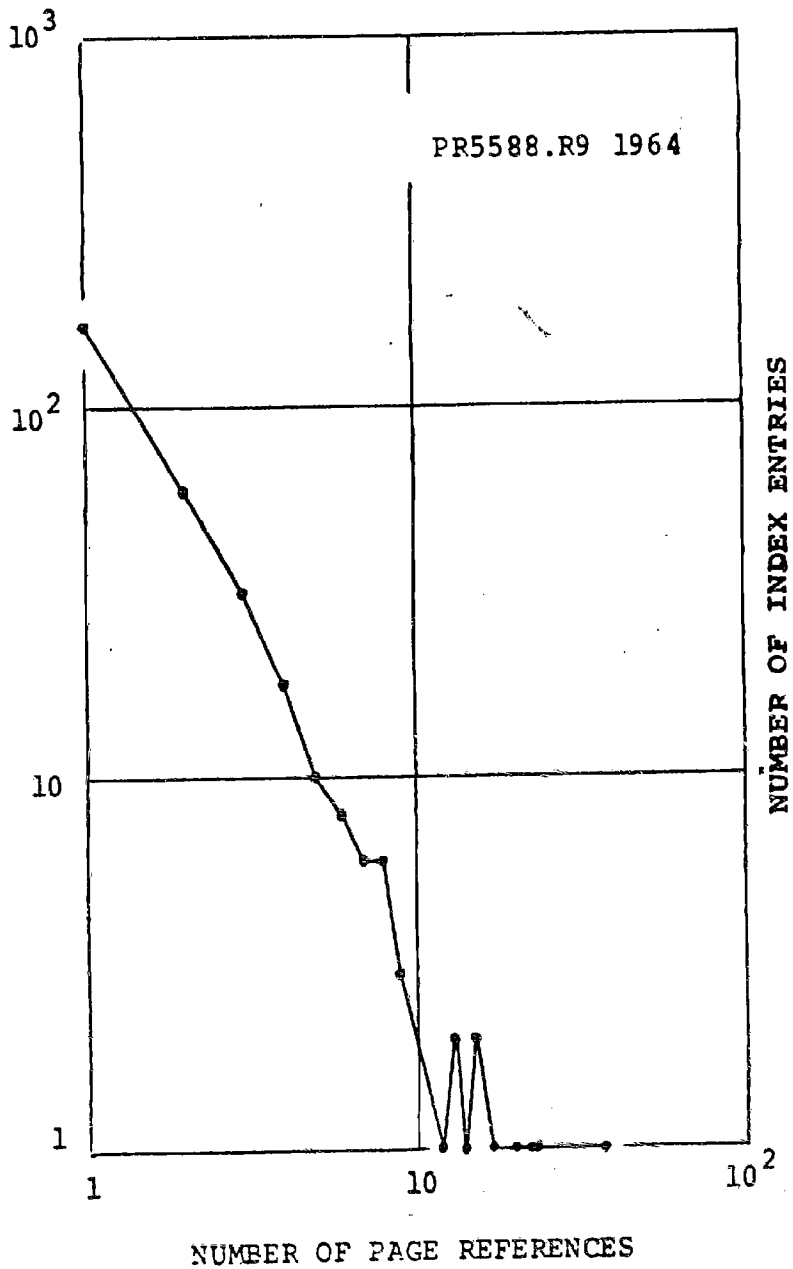

QE391.S5 V3 1965
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

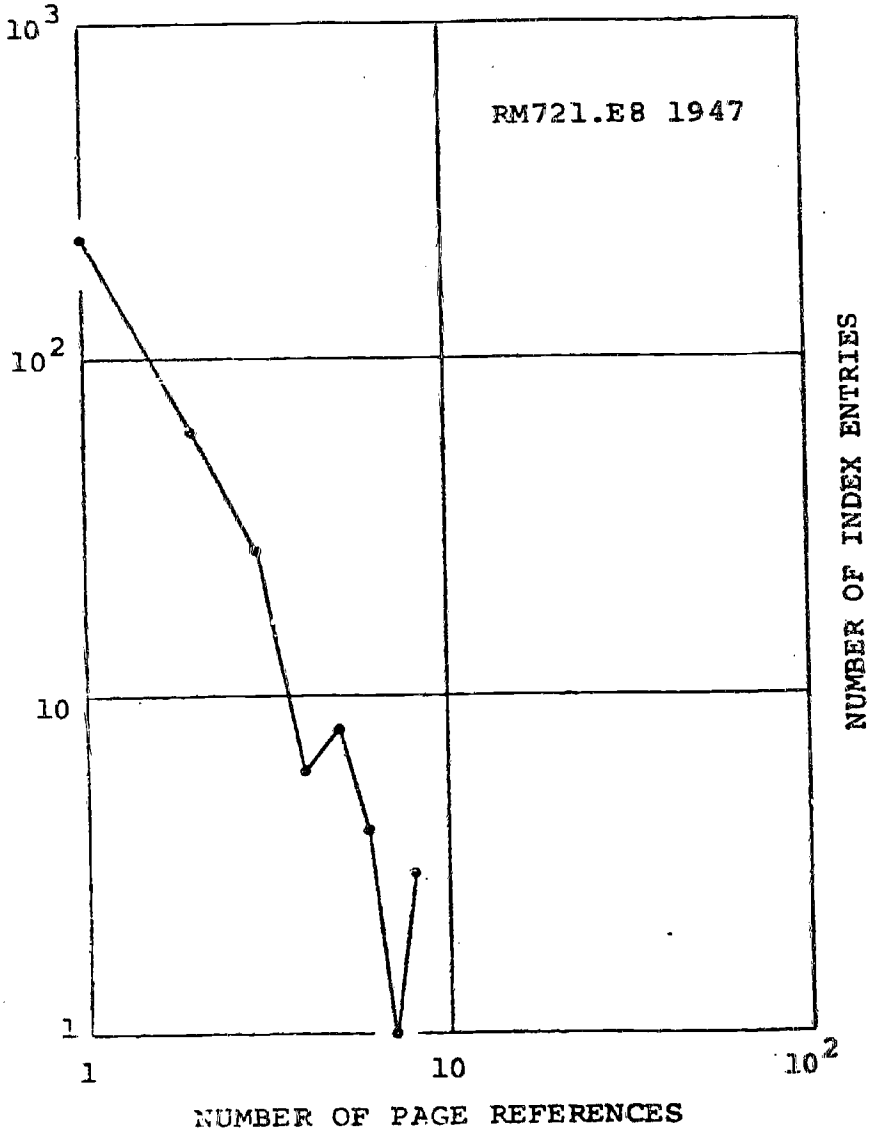

QL703.D3 1963
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES


QA303.088 V3 1945-54
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
10^3
10^2
10
1
10^2
10
1


PT7244.K57 1946-49
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
10^3
10^2
10
1
10^2
10
1


PR2831.H6 1948
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES


PR5588.R9 1964
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES


RM721.E8 1947
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

APPENDIX III

AMALGAMATED ALGORITHMIC INDEX TO ABSTRACTS IN STATISTICS

INDEX TO ABSTRACTS

FROM SELECTED VOLUMES OF THE ANNALS OF MATHEMATICAL STATISTICS

N

O

P

Q

R

S

T

U

V

W

APPENDIX IV

AMALGAMATED ALGORITHMIC INDEX

ABSTRACTS IN CANCER RESEARCH

INDEX TO CANCER ABSTRACTS

S

T

U

V

W

Y

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

(TOP)

ERIC ACCESSION NO.			
CLEARINGHOUSE ACCESSION NUMBER	RESUME DATE — —	P.A.	T.A.

IS DOCUMENT COPYRIGHTED? YES ☐ NO ☐

ERIC REPRODUCTION RELEASE? YES ☐ NO ☐

001

100 TITLE
101 ACCESS: A STUDY OF INFORMATION STORAGE AND RETRIEVAL WITH
102 EMPHASIS ON LIBRARY INFORMATION SYSTEMS
103

200 PERSONAL AUTHOR(S)
Resnikoff, H. L. and Dolby, J. L.

300 INSTITUTION (SOURCE) R & D Consultants Co., Los Altos, California — SOURCE CODE

310 REPORT/SERIES NO.

OTHER SOURCE — SOURCE CODE

320

330 OTHER REPORT NO.

OTHER SOURCE — SOURCE CODE

340

350 OTHER REPORT NO.

400 PUB'L. DATE 22 March 72 — CONTRACT/GRANT NUMBER OEC-0-9-140548-2791(095)

500 PAGINATION, ETC.
501 vii p. + 269 pgs

600 RETRIEVAL TERMS
601 Access Systems, Information Storage and Retrieval, Classification
602 Systems, Dynamics of Classification Systems, Library Growth Rates,
603 Automatic Indexing, Cumulative Indexing, Library Management
604
605
606

IDENTIFIERS
Fondren Shelf List, Widener Shelf List

ABSTRACT
30 A level structured model of library access systems is c[illegible]ted
301 and compared to operating statistics for various libraries and
302 library stores. The model allows analysis of a given store, or
303 a set of stores, to determine whether the existing access system
304 is operating at a level consistent with the model. Suggestions
305 are made as to how current levels of subject access can be sig-
306 nificantly augmented by cumulative book indexes. Attention is
307 also directed to the probability distributions both for the
308 size of various access systems and the usage of library materials.
309 Finally, there is a discussion of the impact of computer related
310 systems and the associated problems in linguistic data processing
311 insofar as they impact on library problems.
312
313
314
315
316
317
318
319
320
321