Chapter I: "Introduction and Summary of Results," stresses the view that the problem of insufficient access is primarily a problem of the great size of the archives to which access is desired. Chapter II: "Levels of Information Storage and Access," is directed toward the 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. Chapter III: "Mathematics of Information Distributions," is devoted to the mathematical study of some of the distributions that arise naturally in the study of information systems. Chapter IV: "The Structure of the Back-of-the Book Indexes," is a study of indexes to books in order to determine what structure, if any, they possess. Chapter V: "Algorithmic Text Indexing," is also exclusively concerned with back-of-the book indexes. Chapter VI: "Amalgamative Access Mechanisms," looks at the problem of discovering possible methods for accessing books. Examples of indexes are appended and the tables included are listed. (NH)
INTERIM 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. DOBLY
R & D CONSULTANTS COMPANY
Los Altos, California and Houston, Texas

21 May 1971

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 project 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 HEALTH, EDUCATION, AND WELFARE
Office of Education
Bureau of Research
ACKNOWLEDGEMENT

The authors 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 the late Gerald Mitchell of the Institute for Defense Analyses 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.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER I</td>
<td>INTRODUCTION AND SUMMARY OF RESULTS</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER II</td>
<td>LEVELS OF INFORMATION STORAGE AND ACCESS</td>
<td>18</td>
</tr>
<tr>
<td>CHAPTER III</td>
<td>MATHEMATICS OF INFORMATION DISTRIBUTIONS</td>
<td>54</td>
</tr>
<tr>
<td>CHAPTER IV</td>
<td>STRUCTURE OF BACK OF THE BOOK INDEXES</td>
<td>85</td>
</tr>
<tr>
<td>CHAPTER V</td>
<td>ALGORITHMIC TEXT INDEXING</td>
<td>112</td>
</tr>
<tr>
<td>CHAPTER VI</td>
<td>AMALGAMATIVE ACCESS MECHANISMS</td>
<td>140</td>
</tr>
</tbody>
</table>

**APPENDICES**

<table>
<thead>
<tr>
<th>APPENDIX I</th>
<th>ABSTRACT INDEX ENTRIES: A UNIFORM SAMPLE FROM THE FONDREN INDEX SAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPENDIX II</td>
<td>INDEX PAGE REFERENCE DISTRIBUTIONS FROM THE FONDREN INDEX SAMPLE</td>
</tr>
<tr>
<td>APPENDIX III</td>
<td>AMALGAMATED ALGORITHMIC INDEX TO ABSTRACTS IN STATISTICS</td>
</tr>
<tr>
<td>APPENDIX IV</td>
<td>AMALGAMATED ALGORITHMIC INDEX TO ABSTRACTS IN CANCER RESEARCH</td>
</tr>
<tr>
<td>Table Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Abstract Index Entries from Computerized Library Catalogs</td>
<td>15</td>
</tr>
<tr>
<td>Distribution of Entry Length in Characters (excluding page references)</td>
<td>27</td>
</tr>
<tr>
<td>Size in Characters of Various Bibliographic Units</td>
<td>38</td>
</tr>
<tr>
<td>Lognormal Standard Deviations</td>
<td>40</td>
</tr>
<tr>
<td>Grouped Number of Index Entries for Monographs</td>
<td>75</td>
</tr>
<tr>
<td>Distribution of Lexed Words from the Shorter Oxford Dictionary by number of vowel strings</td>
<td>79</td>
</tr>
<tr>
<td>Fondren Sample: Fraction of Sample Items Containing an Index, by LC Letter Class</td>
<td>90</td>
</tr>
<tr>
<td>Index Access by LC Class</td>
<td>91</td>
</tr>
<tr>
<td>Frequency of Index Entries for Items in the Fondren Index Sample</td>
<td>92</td>
</tr>
<tr>
<td>Zipf-Mandelbrot Exponent for Index Location Distribution</td>
<td>100</td>
</tr>
<tr>
<td>Comparison of High-Frequency Index Entries with LC Subject Headings &amp; Titles</td>
<td>104</td>
</tr>
<tr>
<td>Distribution of Index Entries by Word Length Subsample of the Fondren Index Sample</td>
<td>109</td>
</tr>
<tr>
<td>Short List of Stop Words arranged by Word Length</td>
<td>117</td>
</tr>
<tr>
<td>Excluded Index Terms Referring to One Location</td>
<td>125</td>
</tr>
<tr>
<td>Index Entry Length Distribution, Computerized Library Catalogs</td>
<td>132</td>
</tr>
<tr>
<td>Index Page Location Distribution, Computerized Library Catalogs</td>
<td>135</td>
</tr>
</tbody>
</table>
Entry Length Distribution, Algorithmic Index to Statistical Abstracts .......................... 150

Abstract Number Location Distribution, Algorithmic Index to Statistical Abstracts .... 151

Bibliographic Description of the Statistics Index Sample ............................................. 154

Abstract Entries for the Amalgamated Statistics Index Sample ................................. 156
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Serial Publications with Trend Line</td>
<td>3</td>
</tr>
<tr>
<td>Exponential Growth of Scientific Journals and Abstract Journals</td>
<td>5</td>
</tr>
<tr>
<td>Index Reference Distributions</td>
<td>10</td>
</tr>
<tr>
<td>High Frequency Index References</td>
<td>12</td>
</tr>
<tr>
<td>Text and Page Location Distributions</td>
<td>14</td>
</tr>
<tr>
<td>Title Length Distribution in Characters, Fondren Index Sample</td>
<td>22</td>
</tr>
<tr>
<td>Distribution of Size of Table of Contents for Books in the Fondren Index Sample</td>
<td>23</td>
</tr>
<tr>
<td>Index Length Distribution, Fondren Index Sample</td>
<td>25</td>
</tr>
<tr>
<td>Distribution of Entry Length in Characters, Statistical Index Sample</td>
<td>28</td>
</tr>
<tr>
<td>Distribution of Book Length in Pages for books with Indexes from the Fondren Sample</td>
<td>31</td>
</tr>
<tr>
<td>Distribution of Size of University Libraries</td>
<td>34</td>
</tr>
<tr>
<td>The Level Structure of Access Systems</td>
<td>35</td>
</tr>
<tr>
<td>Size of Two-Year College Libraries</td>
<td>37</td>
</tr>
<tr>
<td>Distribution of Size of Bibliographic Units</td>
<td>42</td>
</tr>
<tr>
<td>Some Access Distributions in Logarithmic Variables</td>
<td>43</td>
</tr>
<tr>
<td>Distribution of Number of Characters per LC Subject Heading</td>
<td>46</td>
</tr>
<tr>
<td>Distribution of Number of Subject Headings in Fondren Sample (excluding Serials)</td>
<td>46</td>
</tr>
<tr>
<td>ALTEXT Macros Ranked by Number of Assembly Language Instructions</td>
<td>49</td>
</tr>
<tr>
<td>Word Length Distributions (in Characters)</td>
<td>50</td>
</tr>
<tr>
<td>Standard Curve of English and Latin Words</td>
<td>56</td>
</tr>
</tbody>
</table>
Abstract Number Location Distribution,
Index to Statistical Abstracts . . . . . . . . . 151

Rank-Frequency of Reference Distribution,
Statistical Index Sample . . . . . . . . . . . . . 155

Accumulated Index to Books in Statistics
(one page) . . . . . . . . . . . . . . . . . . . . . . 158
CHAPTER I

INTRODUCTION

AND

SUMMARY OF RESULTS
INTRODUCTION AND
SUMMARY OF RESULTS

This monograph describes work performed by R & D Consultants Company during the first twenty-six months of 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 machinable data bases.

The concern which underlies this and many other projects is that the rapidly growing body of information stored in library archives is overwhelming the traditional means of obtaining access to it in a reliable, timely, and comprehensive manner.

In fact, general archival collections have been growing in an essentially exponential manner for more than three hundred years, and perhaps for much longer. Figure 1.1, drawn on semilogarithmic graph paper, illustrates this phenomenon for serials noted in the Union List through 1930; this collection has been doubling in size every thirty years. If the trend line is extended back in time, it suggests the publication of a "first" printed serial about 1435, which agrees remarkably well with the invention of printing circa 1440-1456. Although the weight of this evidence is insufficient to convincingly show that serial growth has in fact been exponential since that time, it does support the contention that the exponential growth of archives is a fundamentally long-term property, undoubtedly secondary only to economic and population growth and determined by them. Consequently, it must be anticipated that archival growth will, for the foreseeable future, continue to be exponential apart perhaps from fluctuations of minor duration because there does not yet appear to be a significant slackening in either population or long term economic growth for the world as a whole.
Figure 1.1

NUMBER OF SERIAL PUBLICATIONS WITH TREND LINE

{UNION LIST OF SERIALS IN LIBRARIES OF THE UNITED STATES & CANADA, 3RD EDITION}
Some insight into why this "information explosion" has received particular emphasis in recent years can be gained from a study of Figure 1.2, adapted from De Solla Price (1) which shows that in addition to the exponential growth of the number of scientific journals since the second half of the seventeenth century, the number of scientific abstract journals has also been growing exponentially, and at the same rate, since their introduction in about 1825. The abstract journals provide access to the larger body of primary scientific journals; the figure shows that the need for this secondary form of publication apparently appeared when the number of scientific journals reached 300. The number of abstract journals reached 300 by 1950, making it as difficult to access the abstract journals as it had been to access the primary archive in 1825. This suggests that one of the reasons for the current serious concern about problems of information storage and retrieval is that it is once again necessary to invent an appropriate form of (tertiary) publication which will permit another period of orderly growth of the archive.

If the historical and current trend continues unabated for another fifteen years, there will be about 500,000 different scientific journals in existence, publishing more than 25 million papers each year; similar quantities of information will be spewed forth by other fields of endeavour. It is clearly not the problem of storing this information that makes the prospect of such prolific productivity terrifying; current microform techniques are already sufficient to reduce the physical storage requirements to much less than that presently required to store the current production of journals published in conventional form. Moreover, standardization of microform stores make it possible to implement physical retrieval systems that are faster and cheaper than present typical library storage techniques. Nor is the prospect of having to read all of the published material the significant problem. No scientist since 1800 has had the time to read "all" of the papers published even had he the inclination to do so; the situation is the same in most other fields. The inevitable fact that the fraction of published papers read by an individual is going to drop a few more orders of magnitude is hardly consequential.

The problem posed by the explosion of information is only overwhelming when the difficulty of finding a particular fact or result in the vast sea of information is considered. It is this problem of access to which the work reported here is addressed.
Figure 1.2
Exponential Growth of Scientific Journals and Abstract Journals

Number of Scientific Periodicals (Data from D. J. de Solla Price, Science since Babylon [New Haven, 1961], p. 97).
Chapter II introduces a level structured model for access systems, which can be briefly described here. Restricting attention to collections of information expressed in natural languages, size can be reasonably measured by the number of characters, including linguistically necessary interword spaces, contained in the collection. For naturally occurring informational units such as the book title, table of contents, book index, book, and, regarding amalgamated information stores, the university library card catalog and the university library itself, the average size of each informational unit is nearly an integral power of a fixed number $K$ of characters. The value of $K$ is nearly 30. For example, $K \approx 30$ is the average length of a book title measured in characters (as well as the average length of an index entry and of the subject heading information on a Library of Congress catalog card); $K^2 = 874$ characters is approximately the average size of a table of contents; $K^3 = 25,822$ of the average book index, and $K^4 = 763,203$ of the average book. In each case, the average length is remarkably close in value to the power of $K$ in question.

If the size of an information collection is expressed as a power of $K$, say $K^x$, then it is convenient to define the level of the collection as the integer closest to $x$. With this convention, the level of a book title, table of contents, book index, and book is, respectively, 1, 2, 3, and 4. A university library is of level 8. It therefore appears that the traditional means for retrieving information stored in a book are structured in levels which are equally spaced when measured by their level, that is, when measured by the logarithm of their size.

If one information base is an access system for another, as a book index is for a book, then the order of access is defined to be the difference between their levels. In general, the larger the order, the less expensive is the access system insofar as its construction and maintenance are concerned, relative of course to the cost of obtaining and maintaining the accessed data base; but the smaller the order, the more effective the access system will be in locating specific information and accurately reflecting the content of the accessed archive. For instance, a title list is less expensive and less informative than a collection of abstracts (such as Chemical Abstracts) in specifying the content of journal articles in chemistry; the former is of order 2, the latter of order 1.
The level structure described above will provide a valuable management tool for determining, amongst other things, the reasonable size and cost for a system designed to access a given information base only if the average size of a class of information bases is typical of the distribution of sizes in that class. That this is indeed the case is strongly attested by extensive data sampling studies presented in Chapter II, including the analysis of more than 500,000 index entries occurring in a random sample of books drawn from a medium size university library. All of the evidence reinforces the hypothesis that the distribution of size of information collections belonging to a class is lognormal; that is, the distribution of the logarithm of the size of the informational units belonging to the class is a normal distribution. Each access level corresponds to a different lognormal distribution. It turns out that the variance of the occurring distributions are all nearly the same throughout the entire range from level 1 (titles) to level 8 (university libraries); this means that the distributions depend essentially only on their mean and are therefore characterized by their level. This justifies the use of the notion of level as a measure of an access system.

The principal objective of Chapter III is to show that the lognormal distribution of size of informational units belonging to a class (e.g., titles, books, libraries) is a mathematical consequence of certain reasonable assumptions concerning the "effort" or "cost" of using an item in an access system if the complete system maximizes the output of information per unit effort expended. Our argument is a minor extension of Mandelbrot's derivation of the generalized Zipf-Bradford distribution; cp. Refs. (2), (3). The remainder of the chapter describes general mathematical properties of lognormal distributions with emphasis on the most convenient but nevertheless laborious and not entirely satisfactory technique for fitting lognormal functions to sample data; a number of worked examples which are of independent interest are included.

Unfortunately we do not know a theoretical argument that will produce the equispacing of the level structure of the means of the lognormal distributions associated with an access system; this aspect of the access model rests entirely on observational evidence.

The book is still the most natural informational unit for those concerned with library matters. According to the access model, there are exactly four orders of access associated with collections of information of this size,
and, as we have already remarked, there is a traditional access system operating at each of these levels: the index is of order 1, the table of contents of order 2, the title of order 3, and finally, the Library of Congress letter class, which partitions the entire span of written human knowledge into 21 grand categories, is of order 4.

Although these access levels are, in accordance with the prescription of the model, the only ones possible, there are of course many different types of access systems which can function at each of these levels in addition to those just named. For example, a nine page review of a 277 page book provides typical order 1 access for the book of average size. Order 1 access systems most accurately reflect the content of the information collection they access and can moreover form the subsidiary information base from which access systems of higher order (i.e., lower level) can be constructed. Because this procedure obviously cannot be reversed—a low order access system can never be constructed from one of higher order—order 1 access systems deserve special study.

Of the traditional order 1 access systems, the book index is the most amenable to extensive statistical analysis, both because it is found in close proximity to the book text to which it refers (which is generally not the case for book reviews) and because it is naturally composed of a large number of homologous small entities which are suitably arranged for analytical study.

We have investigated three major collections of book indexes. The first contains more than 100 books drawn from the present authors' libraries; although this sample exhibits some variation in subject matter, science and more particularly history, mathematics, and physics are heavily weighted. The second sample consists of 80 current books in statistics and probability theory and comprise what can be thought of as a specialists hand library; it undoubtedly accurately reflects the nature of indexes to books in these fields. All index entries in this sample were committed to machine readable form to permit ready reorganization and analysis of the collection of index entries, of which there were 31,232.

Study of these collections was instrumental in guiding us to the formulation of the access model presented in Chapter II, but their limitations—principally their restriction to few subject areas and the undoubtedly biased method of their selection, but also their relatively small size—clearly indicated the desirability
of carefully selecting a random sample of books and their indexes from a broadly representative archive. The third index sample consists of such a random selection of 706 indexes from the Fondren Library at Rice University. For each book in this sample, copies of the shelf list catalog card, title page, table of contents, and index were made. From this information it is possible to determine the size of the book (in pages, which can then be approximately adjusted to equivalent number of characters) and the precise number of characters in the title, table of contents, and index, which are the three significant traditional book access systems that are normally packaged with the book itself.

Chapter IV describes the structure and properties of traditional back of the book indexes based on a study of these three samples. There are three main conclusions: first, the average number of index entries per index is determined; the result is 836, with relatively little variation throughout the different Library of Congress letter classes. Second, it is shown that the distribution of the number of books as a function of the number of entries in their index is lognormal, providing further support for the access model derived in Chapter II. The remainder of Chapter IV is devoted to a study of the distribution of the number of text references per index term in a given book. The underlying idea is an outgrowth of the simple observation that those index entries that refer to only one text page cannot typify the general content of the book, whereas an entry that refers the reader to 40 or 50 text pages is truly of little specific utility to the reader except insofar as it points out one general topic of the book. It is therefore conceivable that some subset of the index functions as a collection of "key words", specifying the semantic content of the work and serving little further purpose. Were it possible to separate this subset from the other more numerous index entries, the way would be clear for automatic descriptor determination based on a machine readable index; moreover, if the process of constructing the index itself could be automated, iteration of these processes would lead to the descriptors as well and quite possibly to a successful method for man-machine interactive content classification.

Figure 1.3 exhibits the page reference distribution for two books; LB875.C7 was published in 1922, is titled Two Views of Education, and contains 775 index entries, whereas DS423.C85 v4 is the fourth volume of The Cultural Heritage of India, published in the interval 1953-58, and
Figure 1.3

Index Reference Distributions

NUMBER OF PAGES REFERENCED

NUMBER OF INDEX ENTRIES
containing 4906 index entries. It is clear from the figure, which is drawn on full logarithmic graph paper, that except for the quite small numbers of entries referring to very many pages both distributions are linear on the graph paper and hence the number of index entries is a power function of the number of page references. From the theoretical considerations in Chapter III one is led to suspect that these graphs ought perhaps to represent lognormal functions, which appear on logarithmic graph paper as parabolas. As we show in the third chapter, the power function, represented by a straight line, is a degenerate form of the lognormal representing parabola. Other book indexes, as for instance that of The 1969 World Almanac, illustrated at the left half of Figure 1.4, do indeed flaunt a tell-tale curvature and can be accurately fitted by a parabola. This is the third major result of Chapter IV: the page reference distribution of index terms is, generally, a lognormal function which may degenerate into a power function.

There are about 6600 index entries in The 1969 World Almanac. This book and others like it are more thoroughly and densely indexed than most, but let us for the moment treat this index as an order 1 access system for its text, as usual, but simultaneously consider it as an information base requiring access systems. Then selection of that 1/30 of the index which refers to the largest number of page locations will produce an order 2 access system for the original text, which will be of the size of a table of contents. Repeating this operation leads to the selection of the subset of the index which is about $1/(30)^2 = 1/900$ the size of the index and approximately the size of a title. This process produces about 8 index entries, substantially larger than a title because of the peculiarities of almanacs. Figure 1.4 shows the four most "popular" index entries; in approximately the space of the title they provide an order 3 precis of the book's content which is a not unuseful alternate to the title itself.

The distribution for Nader's Unsafe at any Speed is shown in the same figure; the three most popular index entries again provide a cogently descriptive view of the book's content which is in fact not provided at all by the title.

Chapter IV pursues the study of the effectiveness of the popular index entries as content descriptors through the analysis of a uniform subsample of the Index Sample; for this subsample those index entries which refer to large numbers of text locations have been explicitly listed. The subsample is included as Appendix I.
Figure 1.4

High Frequency Index References

The 1969 World Almanac

Ralph Nader
Unsafe at Any Speed

Democrat

Republican Party

Richard M.

Demonstrations

Vietnam: War

V. N.: Vietnam

Ama Ford Motor Co., Inc.

Number of Page References
The problem of automatically indexing documents and books has intrigued computer buffs for years. Numerous programs are now available and many learned research papers have been written describing them, and how modest are their demands on the machines that implement them, and how effective they are in satisfying rather imprecisely stated hypothetical requirements of potential users. But as far as we have been able to learn, no commercial or professional publishing house uses machines to index books or papers. The reason that this is so consists of a complex of subreasons not all of which have to do with the adequacy of machine methods, but it is certainly true that the general complexity, inflexibility, and simple inadequacy of these programs have acted as strong deterrents to their use. Although the problem is hardly a trivial one, we think that one of the most significant factors hindering the development of indexing algorithms that will rival and surpass human performance is that no one has ever attempted to assess precisely what properties human produced indexes actually have as opposed to what indexers and students of indexing believe ought to be the properties of indexes. The availability of the Fondren Index Sample has made it possible to assess human performance in this area, and to set standards for the performance of machine methods of indexing which are objective, and, insofar as they refer to the structural statistics of indexes rather than their semantic content, also measurable. Elucidation of these common structural characteristics of human indexes have in turn suggested some new approaches to the problem of machine indexing. Chapter V is devoted to one such new method. Figure 1.5 illustrates the text location and page location reference distributions for the algorithmically produced index to Computerized Library Catalogs: Their Growth, Cost and Utility. Based on our study of the Fondren Index Sample we can assert that this algorithmic index is the "right size"; moreover, it is evident from the figure that the reference distributions agree well with typical distributions associated with human indexes. The details of the algorithm as well as of the index referred to by the figure are the subject of the fifth chapter.

Combining these results with those of the previous chapter leads to a new method for obtaining keyword descriptors, which is discussed in the context of the particular algorithmic index exhibited. Indeed, this index consists of 340 entries; an order 1 access system acting on this index should select about 12 entries which would provide an "abstract" of the content of the index. There are 9 entries referring to at least ten page locations, but 14 referring to at least nine. Table 1.1 lists these 14 abstract entries together with the number of pages to which each refers.
Table 1.1
ABSTRACT INDEX ENTRIES
FROM
"COMPUTERIZED LIBRARY CATALOGS:..."

<table>
<thead>
<tr>
<th>No. of Page</th>
<th>References</th>
<th>Index Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td></td>
<td>LC</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>GNP</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>growth rate</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>library catalog</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>machine-readable form</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Library of Congress</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>gross national product</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>university library</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>exponential growth</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>bibliographic record</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Fondren, see Rice University</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Sample, see Fondren Sample, Rice University</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>shelf list</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Stanford</td>
</tr>
</tbody>
</table>
In this case the abstract entries provide an accurate capsule view of the problems studied in that book as well as a list of the principal sources of information upon which it bases its arguments.

Although there are only four levels available for accessing the text of books and each of these is already served by a traditional access mechanism, there remain many possibilities for repackaging access information in order to serve needs that cannot be met by traditional means. Many of these are amalgamative in the sense that they combine access information associated with numerous comparable unit information stores in a reorganized manner that permits ready selection of the units that are likely to contain specific matter desired by the user. All document information retrieval systems operate in an amalgamative manner, as does the library catalog. Most low order (high level) amalgamative access systems currently in use organize the access information in a sequential fashion based first on date of publication and secondarily according to some scheme of content classification. This is the procedure used to organize professional society abstracting journals (which provide order 1 access); its success depends entirely on the accuracy and excellence of the content classification system and the classifiers who implement it. Such systems, which represent professional consensus concerning significant categorical classifications, are in general partly obsolete, especially in rapidly growing fields such as chemistry where the quantity of published material may double in as few as eleven years. Moreover, although the bulk of the classified material remains stable as the classification system expands and is refined, some fraction of the archival materials, which is likely to include the most innovative work, should be reclassified to account for changes in classification categories and procedures, but due to economic constraints, it never is. This difficulty suggests that it may be desirable to investigate amalgamative access mechanisms which do not depend on external classification structures which are inherently slow to accommodate themselves to change but rather rely on the text terms and systems based on the processing of numerous homologous small items such as index entries, as opposed to text abstracts, which are of special interest because they are less subject to global grammatical constraints, and therefore admit a greater variety of potentially useful orderings.

Chapter VI studies two new types of amalgamative access systems. The first consists of the combined indexes to a
collection of books, here illustrated by the combined indexes to 80 books in statistics and probability theory, already mentioned above in another context. The second is more unusual. We have applied a version of the algorithmic indexing procedure described in Chapter V to two samples of 50 abstracts drawn, respectively, from the Annals of Mathematical Statistics and the Journal of Cancer Research; the results are exhibited and analyzed.

Appendix I displays the order 1 abstract entries (but in some cases involving exceptionally large indexes, only the order 2 abstract entries) from a uniform subsample of the Fondren Index Sample.

Appendix II displays the distribution of the number of index entries as a function of the number of distinct text pages to which each entry refers for the same subsample of the Fondren Index Sample used for Appendix I. These distributions confirm the assertion that the distribution is essentially a power function.

Appendices III and IV are automatically constructed amalgamative indexes to 50 abstracts of papers in statistics and in cancer research respectively. The algorithm and all internal dictionary-like stores used by it is the same for both data sources.

REFERENCES


CHAPTER II

LEVELS OF
INFORMATION STORAGE AND ACCESS
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 biases 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.
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 cannot 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
Forsman 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 \times 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)

<table>
<thead>
<tr>
<th>No. of Char.</th>
<th>No. of Entries</th>
<th>No. of Char.</th>
<th>No. of Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>44</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>46</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>47</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>51</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>52</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>23</td>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>55</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>28</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>57</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>57</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>17</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>17</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>19</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>19</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>14</td>
<td>69</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>69</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>11</td>
<td>72</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>9</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>11</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>10</td>
<td>75</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>9</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>33</td>
<td>10</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>34</td>
<td>6</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>35</td>
<td>5</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>36</td>
<td>4</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>37</td>
<td>5</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>38</td>
<td>10</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>39</td>
<td>10</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>40</td>
<td>9</td>
<td>77</td>
<td>9</td>
</tr>
</tbody>
</table>
Figure 2.4
Distribution of Entry Length
(in Characters)
Statistical Index Sample
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 form 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 judgments 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 simply 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
FIGURE 2.5
DISTRIBUTION OF BOOK LENGTH IN PAGES
FOR ALL BOOKS FROM THE "FONDREN SAMPLE"
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 $341.5 \times 2400 = 819,600$ 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 omissions are the University of Chicago and Yale University. Although these omissions 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 x 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 biased, 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 OF UNIVERSITY LIBRARIES

(Source: "Library Statistics of Colleges and Universities", Fall 1969)
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

1 2 3 4 5 6 7 8 9 10
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

\[
 s(n) = a10^{bn} \tag{2.1}
\]

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 1943)
Table 2.2

SIZE IN CHARACTERS
OF VARIOUS BIBLIOGRAPHIC UNITS

<table>
<thead>
<tr>
<th>Unit</th>
<th>Level</th>
<th>Size</th>
<th>Log₁₀ of Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>1</td>
<td>28.15</td>
<td>1.44948</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>2</td>
<td>505.</td>
<td>2.70329</td>
</tr>
<tr>
<td>Index</td>
<td>3</td>
<td>21293.</td>
<td>4.32710</td>
</tr>
<tr>
<td>Text of Book</td>
<td>4</td>
<td>819600.</td>
<td>5.91360</td>
</tr>
<tr>
<td>Two Year College Library</td>
<td>7</td>
<td>24528169200.</td>
<td>10.38966</td>
</tr>
<tr>
<td>University Library</td>
<td>8</td>
<td>620735534400.</td>
<td>11.79291</td>
</tr>
</tbody>
</table>

The 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

\[
s(n) = 10^{1.47247n}
\]

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

\[
s(n) = 10^{1.47247n} = (29.68)^n.
\]
The constant 29.68 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 \((2e)^2 = 29.54\ldots\), where \(e = 2.718\ldots\) 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>
<thead>
<tr>
<th>Unit</th>
<th>Level</th>
<th>Lognormal S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>2</td>
<td>0.30</td>
</tr>
<tr>
<td>Index</td>
<td>3</td>
<td>0.44</td>
</tr>
<tr>
<td>Monograph</td>
<td>4</td>
<td>0.23</td>
</tr>
<tr>
<td>Two Year College Library</td>
<td>7</td>
<td>0.29</td>
</tr>
<tr>
<td>University Library</td>
<td>8</td>
<td>0.36</td>
</tr>
</tbody>
</table>

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 than 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 2.10
DISTRIBUTION OF
SIZE OF BIBLIOGRAPHIC UNITS

UNIVERSITY LIBRARY

TWO-YEAR COLLEGE LIBRARY

MONOGRAPH

MONOGRAPH INDEX

TABLE OF CONTENTS

MONOGRAPH TITLE
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 \( \log_s \) and compare this number with the appropriate bibliographic unit level \( n \) to see whether \( \log_s \) lies within \( +1/2 \) 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 super-systems. 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.
Figure 2.12

Distribution of Number of
Characters per LC Subject Heading

Distribution of Number of
Subject Headings in Fondren Sample
Excluding Serials
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 30.72 (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 exercised 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 prose coverage.

Herdan (19) 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 postulated 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


CHAPTER III

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 necessarily be more demanding of the reader's mathematical knowledge than the remainder of the book and has therefore been written so as to 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 come from information theory and its mathematical apparatus, so there is really no way to avoid these technical considerations.

We will be principally concerned with two distributions—the Zipf and the lognormal. The former is also frequently associated with the names Estoup (1), Bradford (2), and more recently Mandelbrot (3,4). Zipf rediscovered and popularized the observation that the ranked frequency distribution of words in natural text corpora is essentially of the form

\[ y = cx^{-s} \]  

(3.1)

where \( x \) denotes the rank and \( y(x) \) the frequency of occurrence of the word of rank \( x \); here \( c \) and \( s \) are constants selected to fit the data as nearly as possible and which therefore are characteristic of the text corpus and to some extent the language from which it is drawn; Zipf only considered the case \( s = 1 \).

Figure 3.1, taken from Zipf (5), exhibits such distributions.

Distributions of the type (3.1) occur in other fields, associated, for special values of \( s \), with Pareto (6) in economics, Lotka (7) in what might be termed 'sociological mathematics', and more recently De Solla Price (8), and no doubt in numerous other contexts as well.

That the Zipf "law" is taken seriously, not just considered as an accidental quirk of the data to be remarked upon and ignored, is attested by the variety of publications that dispute, modify, and reduce it to a triviality. Mandelbrot showed that a slight generalization of the
Figure 3.1
Zipf law, in better agreement with the data, is a consequence of elementary arguments and reasonable hypotheses about the effort required for the efficient transmission of information; in this sense he is a bulwark for both the proselytizers and trivializers of Zipf since his arguments convincingly show that the nature of the distribution has nothing to do with special properties of language that distinguish it from a variety of other processes that extremize some function representing the degree of organization of the process in a statistical sense. Mandelbrot was quick to point out the connection, which is more than merely formal, between his result and the mathematical methods used to derive it, and the derivation of the partition function in statistical mechanics; cf. Schrodinger (9).

Despite the considerable research efforts that have gone into understanding and improving the relation of Zipf, there are significant discrepancies between the Zipf-Mandelbrot predictions and the observed data for large samples of words drawn from natural language text corpora and for other data collections as well. There are theoretical difficulties too: Yule (10) observed that the sum of the frequencies predicted by the Zipf-Mandelbrot distribution (with $s = 1$) is not finite, which implies that there must be a significant deviation from this distribution for large values of the variable. This kind of difficulty is not as easily brushed aside as disagreement with the data can be, for it entails an unknown mechanism which determines that range of the variable for which the distribution must be modified as well as the unknown modification itself, and leaves the researcher bereft of the argument that improved "experimental measurements" will modify the situation in any agreeable way. It is much easier to reconcile ill fitted observations, and their consequences are normally much more local in nature.

Nevertheless it has been found desirable to modify Zipf's Law in many ways to better fit the data. Mandelbrot's modification, based on his theoretical considerations, is:

$$y = c(x - a)^{-s}$$  \hspace{1cm} (3.2)

where $a$ is a small constant. Belonogov (11) found that the distribution

$$y = e^{-c(x-1)} - e^{-cx}$$  \hspace{1cm} (3.3)
describes the rank-frequency structure of printed commercial Russian. Good (12) is led to

\[ y = c(x-a)^s(1 + by^{-1}), \]  

(3.4)

with \( b \) a small constant; this form has (3.2) as a first approximation (because \( b \) is small) and also is responsive to Yule's criticism since the sum of the frequencies is finite. It is derived by including in the effort function (see below) a factor corresponding to the effort required to incorporate words of large rank in the inventory, and represents, in a certain sense, part of the system 'overhead'. Unfortunately, (3.4) is a complicated expression and Good's choice of overhead factor is in no way uniquely determined.

Other authors have turned to functions that are apparently quite different in order to more faithfully describe their data. Houston and Wall (13) described the distribution of term usage in manipulative indexes using the lognormal distribution, eliciting from Fairthorne (14) the remark that, in his view, Wall (15) (and presumably also Houston and Wall (13)) selected the lognormal only because the data was well fit by that distribution in the sense that the results plotted as a straight line on lognormal probability graph paper, but that they would also have done so on ordinary logarithmic graph paper because "segments of the tail of a Gaussian distribution are not readily distinguished from segments of a hyperbolic distribution"; by the latter he means the Zipf distribution. Certainly this remark applies in principle to the figures plotted on lognormal graph paper in the previous Chapter, but we will see that it is significant only when the variance of the lognormal distribution in question is large.

Carroll (16) has discussed the statistical problems associated with representation of the Standard Sample of Present-Day Edited American English (17) by lognormal distributions; there is in his work no hint of the formerly used Zipf approximation.

We think that there are two conclusions that should be drawn from this necessarily brief survey: first, the Zipf-Mandelbrot distribution does not adequately fit much data although it is well grounded in theory, and second, it is often difficult to distinguish lognormal approximations of data from Zipf approximations. They suggest that an intimate relation may exist connecting the Zipf and lognormal distributions, and, if this
be true, that a 'derivation' of the lognormal from elementary principles along the lines of Mandelbrot's arguments may be possible.

In order to show that these hopes are indeed justified, we will present a derivation of the Zipf-Mandelbrot Law following the usual argument, and in effect following Schrödinger (9), although the notation and terminology there is of course quite different.

Consider information 'states' \( S_1, S_2, \ldots, S_x, \ldots \) constituting some inventory, such as the words of language as they occur in some large text corpus, or a large random collection of monograph titles or indexes, ordered in some convenient fashion. Let \( \epsilon(x) \) denote the 'effort' ('energy') required to utilize state \( S_x \) in an access system ('communication system'), and denote by \( p(x) \) the probability of utilization of \( S_x \) in the inventory. Following Shannon (18), the expected amount of information per unit expected effort is proportional to

\[
I = - \sum p(x) \log p(x) / \sum p(x) \epsilon(x). \tag{3.5}
\]

If the access system is such that the expected amount of information per unit effort is maximized, then the probabilities \( p(x) \) cannot be unrelated to the effort function \( \epsilon(x) \); maximization of \( I \) subject to the necessary restraint

\[
\sum p(x) = 1 \tag{3.6}
\]

will determine the form of \( p(x) \) for given \( \epsilon(x) \). Now maximization of (3.5); subject to (3.6) is equivalent to maximization of

\[
- \sum p(x) \log p(x) \tag{3.7}
\]

subject to (3.6) and the additional restraint that the total effort

\[
\sum p(x) \epsilon(x)
\]

is constant as well. The mathematical method of Lagrange multipliers provides the solution to this extremal problem in the following way: since the total probability (3.6) and the total effort are constant, the function (3.7) and the function
\[ H = - \sum p(x) \log p(x) + (1 + a_0) \sum p(x) \]
\[ + a_1 \sum p(x) \epsilon(x) \]  

(3.8)

attain their maximum for the same functions \( p(x) \)
of \( e(x) \), where \( a_0 \) and \( a_1 \) are arbitrary constants and the form \((1+a_0)\) has been chosen for later notational convenience. Now subject each \( p(x) \) to small differentiable independent functional variations, all the while keeping \( x \) fixed; \( H \) will assume its maximum where the derivatives \( \partial H/\partial p(x) \) all vanish. This yields the simultaneous conditions

\[ 0 = \partial H/\partial p(x) = -(1 + \log p(x)) + (1+a_0) + a_1 \epsilon(x) \]  

(3.9)

which implies

\[ \log p(x) = a_0 + a_1 \epsilon(x) \]  

(3.10)

This is the fundamental relation connecting the effort function, which is presumed to be known, with the probability of occurrence of the state \( S \). It remains to specify the effort function. Mandelbrot argued that, if the states \( S \) are words drawn from natural text and arranged in decreasing frequency of occurrence, then \( \epsilon(x) \) is proportional to \( \log(x-a) \) with a some small constant. This hypothesis immediately leads to (3.2) with \( c = e^{a_0} \) and \( s = -a_1 \). In order that the distribution decrease with increasing \( x \), \( a_1 \) must be negative. (If \( a_1 = 0 \), the distribution degenerates into the uniform distribution, which can only apply to a finite range of the variable \( x \).)

The idea underlying Mandelbrot's choice of effort function is perhaps most simply illustrated by recalling the ordinary use of positional notation to represent positive integers. If \( b \) is an integer greater than 1 and \( n \) is any positive integer, then \( n \) has a unique representation of the form

\[ n = a_N b^N + \ldots + a_k b^k + \ldots + a_1 b + a_0 \]

with \( a_k \) integers less than \( b \) but not negative. For instance, if \( b = 10 \) and \( n = 234 \), then
234 = 2 \cdot 10^2 + 3 \cdot 10 + 4.

By means of such an expression \( n \) can be identified with the sequence of numbers \( a_n a_{n-1} \ldots a_0 \). If \( b = 10 \) this correspondence is the usual decimal expression for \( n \), while if \( b = 2 \), it is the binary expression. Such an expression for \( n \) requires \( N \) symbols each of which is selected from an inventory of \( b \) symbols \( (0, 1, 2, \ldots, b-1) \). Evidently

\[
N + 1 \geq \log_b n \geq N,
\]

so the number of places required to express \( n \) in base \( b \) is approximately \( \log_b n \), and approximately \( b \log_b n \) selections suffice to specify an integer lying between 0 and \( n \).

By coding the information specified by the states \( S \), as integers, this argument can be made to apply to the \( S \) themselves, leading to the Zipf-Mandelbrot distribution for words if that is what the states represent.

It must be recognized that more is involved in the distribution of information states than the simple matter of minimal coding; Good's argument mentioned above attempts to account to some extent for the effort required to add a state to the inventory, i.e., to learn a rare word. Therefore the effort function may not have the form proposed by Mandelbrot except in the simplest of cases, and it becomes necessary to investigate the probable nature of substitutes for it. It might be argued that in general a multiple of \( \log (x-a) \) will constitute a good first approximation to \( \varepsilon(x) \). This, and equation (3.10), suggest the introduction of the variables

\[
u = \log (x-a) \quad (3.11)
\]

and

\[f(u) = \log p(x) \quad (3.12)\]

so that (3.10) becomes

\[
f(u) = a_0 + a_1 \varepsilon (e^{u-a}) = \varepsilon^*(u), \quad (3.13)
\]

defining the function \( \varepsilon^*(u) \) which is more convenient to work with.
Mandelbrot's assumption for the effort function is, in this notation, simply that
\[ e^*(u) = a_0 + a_1 u. \]  
(3.14)

We will assume that \( e^*(u) \) can be expanded in a Taylor series about the point \( u = 0 \), that is,
\[ e^*(u) = \sum_{k=0}^{\infty} a_k u^k. \]  
(3.15)

For small values of \( u \), \( e^*(u) \) will be well approximated by the first two terms of \( 3.15 \) if \( a_1 \) is not zero, leading to the Zipf-Mandelbrot Law; if a better approximation is desired, more terms must be taken from the series expansion of \( e^*(u) \). Suppose for instance that an approximation accurate through terms quadratic in \( u \) is used:
\[ e^*(u) = a_0 + a_1 u + a_2 u^2. \]

Using (3.11) through (3.13), we obtain
\[ \log p(x) = a_0 + a_1 \log(x-a) + a_2 (\log(x-a))^2; \]

the right hand side can be written as
\[ -\log (x-a) + a_2 \left[ \left( \log(x-a)+(1+a_1)/2a_2 \right)^2 \right. \]
\[ + \left. \left( a_0/a_2-(1+a_1)/2a_2 \right)^2 \right] \]
so
\[ p(x) = ce^{-\frac{1}{2} \left[ \log(x-a)+(1+a_1)/2a_2 \right]^2/\sqrt{-1/2a_2} \]  
(3.16)

\[ p(x) = ce^{-\frac{1}{2} \left[ \log(x-a)-m \right]^2/\sqrt{\text{s}} \]  
(x-a)
with

\[ c = e^{\frac{4a_0a_2 - (1+a_1)^2}{4a_2}} \]

which is the lognormal distribution with lognormal mean

\[ m = \frac{1 + a_1}{2a_2} \]

and lognormal standard deviation

\[ s = \sqrt{-1/2a_2} \]

In other words, the parameters \( a_1 \) and \( a_2 \) which appear in the effort function \( \varepsilon^*(u) \) are related to the parameters of the lognormal distribution defined by that effort function as follows:

\[ a_1 = -(1 + \frac{m}{s^2}), \quad a_2 = -1/2s^2 \]

showing in particular that \( a_2 \) must be negative in order that the distribution correspond to a realizable system. Using these values for the constants appearing in the effort function yields

\[ \varepsilon^*(u) = a_0 - (1 + \frac{m}{s^2}) u - \frac{u^2}{2s^2} \]

which shows that Mandelbrot's hypothesis is warranted when \( s \) and \( m \) are large in such a way that the quotient \( m/s^2 \) remains finite. If \( s \) is not large, then necessarily the simple logarithmic effort hypothesis is inadequate.

This observation reconciles Fairthorne's remark, quoted above, but it has the further reaching consequence that the question of whether the Zipf or the lognormal provides a 'better' fit to given data is really meaningless from this point of view; the Zipf is a special case of the lognormal and can therefore never provide a better fit by the implied metric than the latter. Moreover, the larger \( s \) is, the easier it will be to confuse the two distributions, since the general form will more nearly approach its specialization as \( s \) increases.
The derivation of the lognormal distribution given above is based on the measure, I, of information per unit effort defined by eq(3.5) which occurs in Mandelbrot's work and is also used by Good. It is, however, not the only reasonable measure of average effort. In fact, as W. E. Houchin has observed in a personal communication, the choice of

$$I^* = -\sum p(x) \left| \log p(x)/\xi(x) \right|$$

the expected value of the information per unit effort, in place of I (which is the expected value of the information per expected value of the effort) leads to the lognormal function in a more direct manner, free of the burdensome hypothesis that the effort function is quadratic in the logarithm of size or rank. For, arguing as before with $I^*$ in place of I, leads to the maximization of

$$H^* = -\sum p(x) \left| \log p(x)/\xi(x) \right| + a_1^* \sum p(x) \left. + a_2^* \sum p(x) \xi(x) \right|$$

and therefore to the equations

$$0 = \xi(x) \partial H^*/\partial p(x) = -(1 + \log p(x)) + a_1^* \xi(x) + + a_2^* \xi(x)^2,$$

with solution

$$\log p(x) = -1 + a_1^* \xi(x) + a_2^* \xi(x)^2.$$  

If the effort is a logarithmic function of x, $\xi(x) = \log (x-a)$, then p(x) is the lognormal function with lognormal mean

$$m^* = (1+a_1^*)/2a_2^*$$

and lognormal standard deviation

$$s^* = \sqrt{-1/2a_2^*}.$$
these equations should be compared with eqs (3.18) and (3.19). If \( m^* = 0 \), then this lognormal function reduces to Zipf's original 'law' with exponent \(-1\); if \( m^* \) and \( s^* \) are both large such that \( m^*/s^* \) is finite the general power function of Mandelbrot and Stevens results.

Recalling that \( f(u) = \log p(x) = \epsilon^*(u), \) (3.21) shows that the graph of \( f(u) \) vs. \( u \), that is, of \( \log p(x) \) vs. \( \log(x-a) \), will be a parabola if \( p(x) \) is lognormally distributed, whereas it will be a straight line if \( p(x) \) is distributed according to Zipf's Law. It is instructive to examine some examples.

Figure 3.2 displays the graph of the frequency of the most frequent ordered pairs of words drawn from English language text as a function of their rank, drawn on logarithmic graph paper. The data was derived from the corpus which constitutes the Standard Sample of Present-Day Edited American English (17). It approximates a straight line without any considerable evidence of global curvature thereby supporting the hypothesis that the effort function \( \epsilon^*(u) \) is linear and the consequent distribution Zipf. The arguments that are usually applied to justify the Zipf approximation for the distribution of frequency ranked single words would seem to apply equally well to this case, and therefore the arguments of Carroll (16) concerning the problems associated with the extraction of finite samples from theoretical distributions of lognormal type are also probably valid here, which helps to explain the bending of the curve in the direction of low frequencies for large ranks.

Next consider the distribution of the number of index entries in monographs, shown in Figure 3.3. The data is drawn from the Fondren Index Sample, which is described in detail in Chapter IV. Departure from linearity is clearly exhibited; this data is also shown in Figure 2.3, where it is plotted on lognormal probability paper with striking results which suggest that the points in Figure 3.3 should approximate a parabola. A portion of the parabola that fits this data is shown in the figure. The reader will notice certain peculiarities of the distribution of data points that are characteristic of this type of problem and lead to difficulties of estimation. First of all, small values of the independent variable--the number of index entries in this case--correspond to few data points if the data has been grouped for calculational convenience as these data have been. On the other hand, large values of the
Figure 3.2

DISTRIBUTION OF MOST FREQUENT
ORDERED PAIRS OF ENGLISH WORDS
Figure 3.3

INDEX ENTRY DISTRIBUTION FROM FONDREN INDEX SAMPLE

[Graph showing logarithmic distribution of index entries across monographs]
independent variable correspond to many data points even after grouping if the group intervals are of uniform size and for many of these the corresponding frequencies will coincide, leading to vertical 'segments' such as appear over the '1', '2', and '3' monograph markers. It is the geometric mean of these values that is important if the distribution is in fact lognormal.

More complex phenomena sometimes occur, and are perhaps most easily initially analyzed by studying the nature of the polynomial functions that approximate them, or at least portions of them, in the variables u and f(u). Consider, for example, the distribution of the number of pages in a monograph, shown in Figure 3.4. The data is drawn from the Fondren Sample (cp. Dolby, et. al. (19)) and has been grouped. From the figure one sees that monographs of fewer than 220 pages appear to follow the Zipf distribution whereas longer monographs have lengths that are well approximated by part of a lognormal distribution since they correspond to data points that fall nearly on part of a parabola as shown in the figure; the small arrows indicate computed values of points lying on the fitting parabola. We have no satisfactory explanation for this curious discontinuity in the effort function which describes the distribution of lengths of monographs which maximizes information per unit effort. Extraneous factors, perhaps related to the technology and economics of printing, are probably responsible but we have thus far been unable to isolate them. The reader should, however, compare Figures 2.5 and 2.6 which show the page distribution for unindexed and indexed monographs in the Fondren Sample.

Now consider the problem of determining the parameters of a lognormal distribution which represents given data. Take the equation of the lognormal in the form

$$y = \frac{N}{s(x-a)} e^{\frac{1}{2} \left( \frac{\log(x-a) - m}{s} \right)^2}$$

(3.22)

If the data consists of sample measurements \{x_i\} such that the sample frequency of occurrence of \(x_i\) is \(y_i\), then \(N\) is just the total number of measurements:

$$N = \sum y_i.$$  

(3.23)
Figure 3.4

PAGE DISTRIBUTION FROM FONDREN SAMPLE

NUMBER OF PAGES IN MONOGRAPH
The values of $a$, $m$, and $s$ are determined by introducing an auxiliary quantity related to the skewness of the sample distribution, to whose definition we now turn.

Some more terminology is necessary. Define the $k^{\text{th}}$ sample moment $\mu'_k$ by

$$\mu'_k = \frac{1}{N} \sum x_i^k y_i$$

$\mu'_1$ is the usual mean of the sample. If the sample moments are known, the central moments

$$\mu_k = \frac{1}{N} \sum (x_i - \mu'_1)^k y_i$$

can be calculated. Expressions for the first four central moments in terms of the sample moments are useful when considering lognormal distributions. The first central moment $\mu_1$ is evidently zero, since

$$\mu_1 = \frac{1}{N} \sum (x_i - \mu'_1) y_i$$

$$= (1/N) \sum x_i y_i - (\mu'_1/N) \sum y_i$$

$$= \mu'_1 - \mu'_1$$

The next three central moments are given by the relations

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2 \mu'_1 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^3$$

The positive square root of $\mu_2$ is usually called the standard deviation and denoted by $\sigma$:

$$\sigma = \sqrt{\mu_2}$$
Introduce the ratios

\[ \beta_1 = \frac{\mu_3^2}{\mu_2^3} \]  \hspace{1cm} (3.24)

and

\[ \beta_2 = \frac{\mu_4}{\mu_2^2} \] \hspace{1cm} (3.25)

\( \beta_1 \) is called the skewness of the sample; it provides a simple measure of the departure from symmetry about its mean. The skewness of the normal distribution, as for all symmetrical distributions, is 0. \( \beta_2 \) is sometimes known as the kurtosis of the sample distribution. If \( \beta_2 < 3 \), it is lower and flatter. The kurtosis of a normal distribution is 3.

In the literature, and unfortunately also in tables, other formulae are sometimes used to define quantities known as skewness and kurtosis. It is common, for instance, to find

\[ \gamma_1 = \frac{\mu_3}{\sigma^3} = \pm \beta_1 \]

called 'skewness' (note that the sign of \( \gamma_1 \) is the same as the sign of \( \mu_3 \)), and

\[ \gamma_2 = \beta_2 - 3 \]

is sometimes called 'kurtosis'. We follow Karl Pearson's usage, as found for instance in Ref. (20).

Skewness and kurtosis are of interest because they provide simple measures of the deviation of a sample distribution from the normal distribution and can be used to determine a family of distributions likely to provide an accurate and practical representation of data exhibiting skew variation; this procedure was introduced by Pearson (21). We will later have occasion to compare the lognormal representation of data occurring in information systems with representations by means of Pearson's distributions.
Skewness is of immediate interest here because the three numbers \( \mu_1', \mu_2', \gamma_1 \) determine the lognormal distribution that best fits given sample data. The unknown parameters \( a, m, \) and \( s \) of (3.22) can be expressed by means of an auxiliary quantity \( \eta \) which is the real root of the equation

\[
\eta^3 + 3\eta - \gamma_1 = 0 \tag{3.27}
\]

where \( \gamma_1 \) is the square root of the skewness (with the correct sign) as defined in (3.25). It is easy to see that there is in fact just one real root to (3.27), for otherwise there must be three, so the graph of the left side of (3.27) would have two turning points, which means that the derivative of the left side would have two real roots. But the derivative is \( 3\eta^2 + 3 \), whose roots are pure imaginary.

The unique real root of the cubic equation (3.27) is readily and accurately approximated by using Newton's method. First select some reasonable approximation to the root; if a better choice is lacking, set

\[
\eta_1 = \gamma_1^{1/3}
\]

If \( \eta_k \) is the \( k \)th approximation to the root \( \eta \), then the next approximation is

\[
\eta_{k+1} = \eta_k + (\gamma_1 - 3\eta_k - \eta_k^3)/(3\eta_k^2 + \gamma_1). \tag{3.28}
\]

For example, if \( \eta_1 = 3 \), then \( \eta_1 = 3/3 = 1 \), and

\[
\eta_2 = 1 + (3 - 3 - 1)/(3 + 3) = 0.833333... ,
\]

\[
\eta_3 = 0.795556 ,
\]

\[
\eta_4 = 0.817973 ,
\]

\[
\eta_5 = 0.817732 = \eta_6 ;
\]

therefore \( \eta = 0.817732 \) correct to six places.
Given a sufficiently accurate value of \( \eta \), the parameters of the lognormal distribution (3.22) are (cf. Cramer, Ref. (22)):

\[
\begin{align*}
\alpha &= \mu' - \sigma/\eta, \\
\beta &= \log \left(1 + \eta^2\right)^{1/2}, \\
\gamma &= \log \left(\sigma/\eta - \frac{1}{2}\beta^2\right).
\end{align*}
\]

Recall that \( \sigma = \sqrt{\mu_2} \) is the standard deviation.

If the parameters of a lognormal distribution are known, it is of course possible to calculate the kurtosisty of that distribution; the result is expressible in terms of \( \eta \) as

\[
\beta_2 = \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2 + 3,
\]

and is of no particular value except that it permits one to sketch the graph of the skewness and kurtosisty pairs \((\beta_1, \beta_2)\) that can belong to lognormal distributions. Figure 3.5 shows such a graph. The skewness and kurtosisty of a particular data sample determine a point in the \( \beta_1 - \beta_2 \) plane; the farther this point is from the lognormal curve, the less likely is the hypothesis that the sample data is drawn from a lognormal distribution.

By applying the usual techniques of the differential calculus it is easily shown that the maximum value of the lognormal distribution (3.22) is attained when

\[
\phi = a + e^{m - s^2}.
\]

Some examples showing how these equations are to be applied to sample data may have some interest for the reader.

Consider first the distribution of the number of entries in a monograph index. For data drawn from the Fondren Index Sample, Figure 3.3 shows that this distribution apparently is lognormal. The following table, which summarizes the data in Table 4.3, groups the sample measurements according to intervals of 500 index entries and nominally associates them with the center value in each interval. This coarse grouping scheme
Figure 3.5

COEFFICIENT OF SKEWNESS vs. COEFFICIENT OF KURTOSIS

FOR LOGNORMAL FUNCTIONS
decreases the number of categories to the point where the necessary calculations can be conveniently performed on a desktop calculator.

Table 3.1

GROUPED NUMBER OF INDEX ENTRIES FOR MONOGRAPHS

<table>
<thead>
<tr>
<th>Nominal No. of Entries</th>
<th>Number of Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>326</td>
</tr>
<tr>
<td>750</td>
<td>214</td>
</tr>
<tr>
<td>1,250</td>
<td>76</td>
</tr>
<tr>
<td>1,750</td>
<td>36</td>
</tr>
<tr>
<td>2,250</td>
<td>17</td>
</tr>
<tr>
<td>2,750</td>
<td>9</td>
</tr>
<tr>
<td>3,250</td>
<td>6</td>
</tr>
<tr>
<td>3,750</td>
<td>8</td>
</tr>
<tr>
<td>4,250</td>
<td>2</td>
</tr>
<tr>
<td>4,750</td>
<td>6</td>
</tr>
<tr>
<td>5,250</td>
<td>1</td>
</tr>
<tr>
<td>5,750</td>
<td>1</td>
</tr>
<tr>
<td>6,250</td>
<td>2</td>
</tr>
<tr>
<td>6,750</td>
<td>1</td>
</tr>
<tr>
<td>7,250</td>
<td>1</td>
</tr>
</tbody>
</table>

One finds

\[
N = 706
\]

\[
\mu_1 = 831.444759
\]

\[
\mu_2 = 878,281.765706, \sigma = 937.166882
\]

\[
\mu_3 = 2,551,331,421.74
\]

\[
\mu_4 = 13,238,646,211,200
\]

to twelve figures. Consequently
According to Figure 3.5, the sample value of $\beta_1$ corresponds to $\beta_2 \approx 24$ if the distribution is lognormal. The disagreement is not serious in view of the effect of grouping the data and the small size of the sample; regarding the latter point, J. Carroll's remarks in Reference (16) are instructive. For the original ungrouped data, it turns out that $\mu_1 \approx 836$, so the effect of grouping the data is not entirely negligible.

These estimates imply $\eta = 0.837450$ correct to six figures; indeed, since $\gamma_1$ is nearly equal to 3, it is a good idea to select the solution to (3.27) previously calculated as an example as the starting value $\eta_1$ of the approximation procedure, leading to

$$\begin{align*}
\eta_1 &= 0.817732, \\
\eta_2 &= 0.837642, \\
\eta_3 &= 0.837450 = \eta_4 = \eta.
\end{align*}$$

Substitution in (3.29)-(3.31) produces the parameters of the lognormal:

$$\begin{align*}
a &= -287.6, \\
s &= 0.7284, \\
m &= 6.755,
\end{align*}$$

so

$$y = \frac{706}{0.7284(x+287.6)^{2/2}} e^{\frac{1}{2} \left\{ \frac{\log(x+287.6) - 6.755}{0.7284} \right\}^2}.$$
Observe that

\[ c^m = 858 \approx 874 = k^2. \]

According to these calculations, the (grouped) number of index entries behaves as if approximately 288 entries are 'missing' from indexes. To what extent this must be attributed to the effect of grouping the data so coarsely we have not attempted to determine except to note that the intervals that were chosen will tend to produce this type of qualitative effect because most of the indexes represented in the first group (those having fewer than 500 entries) contain more than the 250 entries nominally ascribed to that category, thus biasing the distribution toward low values.

The next example is a particularly useful pedagogical illustration because the data are unusually regular and occur in large number in a form convenient for calculations, but it is of considerable independent interest as well. Consider the distribution of dictionary words according to the number of vowel strings they contain. The number of vowel strings, in the technical sense in which it is used here, is a graphemic substitute for the phonemic notion of the number of syllables contained in the spoken form of a word; the precise definition we use is given in Reference (23), but will not be necessary for our present purposes since the intuitive correspondence with the notion of syllable is sufficiently accurate. The words under consideration are the 64,041 lexed words of the Shorter Oxford Dictionary which contain at least one vowel string. Figure 3.6 displays the data on bi-logarithmic graph paper; the general parabolic tendency is apparent. From the data given in Table 3.2 one readily calculates

\[ \begin{align*}
N &= 64,041, \\
\mu_1 &= 2.6889, \\
\mu_2 &= 1.1096, \quad \sigma = 1.0534, \\
\mu_3 &= 0.5027, \\
\mu_4 &= 3.7011.
\end{align*} \]
TABLE 3.2
DISTRIBUTION OF LEXED WORDS
FROM THE SHORTER OXFORD DICTIONARY
BY NUMBER OF VOWEL STRINGS

<table>
<thead>
<tr>
<th>Number of Vowel Strings</th>
<th>Observed Number of Words</th>
<th>Calculated Number of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>63</td>
<td>285</td>
</tr>
<tr>
<td>1</td>
<td>7,158</td>
<td>6,618</td>
</tr>
<tr>
<td>2</td>
<td>22,568</td>
<td>22,160</td>
</tr>
<tr>
<td>3</td>
<td>20,762</td>
<td>22,072</td>
</tr>
<tr>
<td>4</td>
<td>10,293</td>
<td>9,737</td>
</tr>
<tr>
<td>5</td>
<td>2,770</td>
<td>2,531</td>
</tr>
<tr>
<td>6</td>
<td>393</td>
<td>691</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>178</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>24</td>
</tr>
</tbody>
</table>

So

\[ \gamma_1 = 0.4301, \]
\[ \beta_1 = 0.1850, \]
\[ \beta_2 = 3.0060. \]

For a lognormal distribution, this value of skewness implies a kurtosis approximately equal to 3.3, which agrees reasonably well with the value computed from the sample. The calculated parameters of the lognormal fitting these data are

\[ a = 4.7086, \]
\[ s = 0.1407, \]
\[ m = 1.9916. \]  \hspace{1cm} (3.34)

The maximum value of this lognormal distribution is \( y \approx 25,000 \) and occurs at \( x = 2.46 \). The rightmost column of Table 3.2 shows the number of words as a function of the number of vowel strings as calculated from the distribution defined by the parameters given in (3.34) above.
Some degree of caution must be exercised when one attempts to determine if data can be reasonably fitted by a lognormal function. If the entire range of the variable is not represented by the data due either to unfortunate grouping or absence of information, graphical representation of the data may be misleading. We will describe one way in which this can happen. Suppose that \{(x,y)\} is a data sample exhibiting the frequency function of some variable \(x\). Let \{(x,Y)\} denote the cumulative frequency distribution defined by

\[ Y(x) = \text{sum of } y(x_0) \text{ for } x_0 \geq x. \]

For instance, if \(y(x)\) denotes the number of individuals having an annual income of \(x\) dollars (more exactly, \(x+b\) dollars for some conveniently chosen small increment \(b\)), then \(Y(x)\) denotes the number of individuals with income at least \(x\) dollars. The latter cumulative distribution is exact in the sense that it presents the actual number of individuals belonging to the corresponding category of the sample, whereas the frequency distribution presents grouped data as a substitute for frequency density functions and therefore potentially introduces error into the sample data. For this reason it is often desirable to analyze cumulative sample distributions rather than the corresponding approximate frequency functions.

Consider, therefore, a cumulative distribution \{(x,Y)\} whose points fall close to a straight line when exhibited on log-log graph paper. In this event it is reasonable to conclude that \(\log Y = a \log x + \log c\) so

\[ Y = cx^a : \quad (3.35) \]

\(Y\) is a power function of \(x\). The frequency function can be retrieved from (3.35) from the relation

\[ y = \frac{dY}{dx}; \]

we find

\[ y = cax^{a-1}, \]

so \(y\) is also a power function of \(x\).

If the entire range of the variable \(x\) is not represented by the sample data, this procedure of determining the theoretical frequency function from its cumulative distribution by differentiation can be misleading. Figure 3.7
Figure 3.7
DISTRIBUTION OF INCOME

N = NUMBER WITH INCOME ≥ ORDINATE SHOWN
(N IN HUNDREDS FOR U.S.)
displays the cumulative income distribution for Great Britain in 1893-94 and for the United States in 1968 drawn on log-log graph paper. The data for the former fall along a straight line which implies that it and hence also the associated frequency function are power functions. This is the famous 'law' of Vilfredo Pareto (Ref. (6), vol. 2, p. 304 et seq.). The leftmost part of the corresponding distribution for the United States also exhibits a generally linear trend but the rightmost portion cannot be so construed at all. Several interpretations of this anomaly are possible, including some that are based on variations between the underlying economic and social structures of the two nations during the two time periods surveyed, but it is possible to account for the apparent contradiction by examining the extent to which the sample data represents the entire range of

The United States data refers to income of any size reported on tax returns, of which more than 73 million were tallied. The data used by Pareto refers only to incomes greater than 150 pounds sterling per year, of which 400,648 were reported. The number of inhabitants per income reported was nearly 3 for the United States in 1968; using this figure, we see that if Pareto's data includes essentially all incomes, the population of Great Britain in 1893-94 should have been about 1.2 million. It was in fact perhaps greater than 8 million, which suggests that there were possibly more than two million people in Great Britain in those years having a positive income less than 150 pounds per year. The reader should not interpret this estimate as anything but a very crude indication of the number of incomes that were probably overlooked in the data sample. Now apply this estimate to extend the graph for Great Britain in Figure 3.7; the extension must turn downward when the total number of incomes exceeds about 2.4 million, so the extended income curve will have the same general appearance as that for the United States.

As has already been remarked, the cumulative distribution of income function for the United States certainly is not a power function and therefore the corresponding frequency function cannot be either. By plotting the grouped frequency data published by the Internal Revenue Service on lognormal probability graph paper, it can be seen that the frequency function of income distribution can be approximated throughout its entire range by a lognormal function with the parameter $a$ of eq (3.22)
approximately equal to $4,000. It is likely that the income distribution for Great Britain used by Pareto can also be approximated by a lognormal function, but it is necessary to have an accurate estimate of the total number of incomes less than 150 pounds per year before one can calculate the cumulative fractions necessary for the employment of lognormal probability graph paper or the methods for estimating the parameter values from sample data given earlier in this Chapter.

References


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


CHAPTER IV

THE STRUCTURE OF

BACK OF THE BOOK INDEXES
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, if any 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 unspecified 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 Fondrer 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 anomalous 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 anomalous 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

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

Total relevant items in Fondren Sample = 1823  
Number of these items indexed = 668  
Fraction indexed = 668/1830 = 0.37
Table 4.2

INDEX ACCESS BY LC CLASS

<table>
<thead>
<tr>
<th>LC Class</th>
<th>Mean No. Index Entries per Book</th>
<th>Short Class Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>640</td>
<td>U. S. Local History</td>
</tr>
<tr>
<td>G</td>
<td>632</td>
<td>Geography</td>
</tr>
<tr>
<td>V</td>
<td>626</td>
<td>Naval Science</td>
</tr>
<tr>
<td>K</td>
<td>591</td>
<td>Law</td>
</tr>
<tr>
<td>D</td>
<td>562</td>
<td>World History</td>
</tr>
<tr>
<td>E</td>
<td>520</td>
<td>U. S. History</td>
</tr>
<tr>
<td>Q</td>
<td>519</td>
<td>Science</td>
</tr>
<tr>
<td>H</td>
<td>376</td>
<td>Social Science</td>
</tr>
<tr>
<td>C</td>
<td>366</td>
<td>Auxiliary Sciences (History)</td>
</tr>
<tr>
<td>R</td>
<td>358</td>
<td>Medicine</td>
</tr>
<tr>
<td>T</td>
<td>332</td>
<td>Technology</td>
</tr>
<tr>
<td>Z</td>
<td>319</td>
<td>Library Science</td>
</tr>
<tr>
<td>L</td>
<td>304</td>
<td>Education</td>
</tr>
<tr>
<td>J</td>
<td>282</td>
<td>Political Science</td>
</tr>
<tr>
<td>M</td>
<td>229</td>
<td>Music</td>
</tr>
<tr>
<td>B</td>
<td>207</td>
<td>Philosophy-Religion</td>
</tr>
<tr>
<td>V</td>
<td>185</td>
<td>Military Science</td>
</tr>
<tr>
<td>P</td>
<td>129</td>
<td>Language Literature</td>
</tr>
<tr>
<td>S</td>
<td>128</td>
<td>Agriculture</td>
</tr>
<tr>
<td>N</td>
<td>111</td>
<td>Fine Arts</td>
</tr>
</tbody>
</table>
Table 4.3

FREQUENCY OF INDEX ENTRIES FOR ITEMS
IN THE FONDREN INDEX SAMPLE

<table>
<thead>
<tr>
<th>Number of Index Entries</th>
<th>Number of Indexes</th>
<th>Cumulative Number of Indexes</th>
<th>Cumulative Fraction of Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 99</td>
<td>16</td>
<td>16</td>
<td>.023</td>
</tr>
<tr>
<td>100 - 199</td>
<td>77</td>
<td>93</td>
<td>.132</td>
</tr>
<tr>
<td>200 - 299</td>
<td>83</td>
<td>176</td>
<td>.249</td>
</tr>
<tr>
<td>300 - 399</td>
<td>80</td>
<td>256</td>
<td>.362</td>
</tr>
<tr>
<td>400 - 499</td>
<td>70</td>
<td>326</td>
<td>.462</td>
</tr>
<tr>
<td>500 - 599</td>
<td>62</td>
<td>388</td>
<td>.549</td>
</tr>
<tr>
<td>600 - 699</td>
<td>46</td>
<td>434</td>
<td>.615</td>
</tr>
<tr>
<td>700 - 799</td>
<td>37</td>
<td>471</td>
<td>.667</td>
</tr>
<tr>
<td>800 - 899</td>
<td>39</td>
<td>510</td>
<td>.722</td>
</tr>
<tr>
<td>900 - 999</td>
<td>30</td>
<td>540</td>
<td>.765</td>
</tr>
<tr>
<td>1000 - 1099</td>
<td>17</td>
<td>557</td>
<td>.789</td>
</tr>
<tr>
<td>1100 - 1199</td>
<td>24</td>
<td>581</td>
<td>.823</td>
</tr>
<tr>
<td>1200 - 1299</td>
<td>13</td>
<td>594</td>
<td>.841</td>
</tr>
<tr>
<td>1300 - 1399</td>
<td>14</td>
<td>608</td>
<td>.861</td>
</tr>
<tr>
<td>1400 - 1499</td>
<td>8</td>
<td>616</td>
<td>.872</td>
</tr>
<tr>
<td>1500 - 1599</td>
<td>2</td>
<td>618</td>
<td>.875</td>
</tr>
<tr>
<td>1600 - 1699</td>
<td>13</td>
<td>631</td>
<td>.893</td>
</tr>
<tr>
<td>1700 - 1799</td>
<td>7</td>
<td>638</td>
<td>.903</td>
</tr>
<tr>
<td>1800 - 1899</td>
<td>7</td>
<td>645</td>
<td>.913</td>
</tr>
<tr>
<td>1900 - 1999</td>
<td>7</td>
<td>652</td>
<td>.923</td>
</tr>
<tr>
<td>2000 - 2099</td>
<td>7</td>
<td>659</td>
<td>.933</td>
</tr>
<tr>
<td>2100 - 2199</td>
<td>3</td>
<td>662</td>
<td>.937</td>
</tr>
<tr>
<td>2200 - 2299</td>
<td>5</td>
<td>667</td>
<td>.944</td>
</tr>
<tr>
<td>2300 - 2399</td>
<td>1</td>
<td>668</td>
<td>.946</td>
</tr>
<tr>
<td>2400 - 2499</td>
<td>1</td>
<td>669</td>
<td>.947</td>
</tr>
</tbody>
</table>
Table 4.3 (Continued)

<table>
<thead>
<tr>
<th>Interval</th>
<th>Count</th>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2500 - 2599</td>
<td>2</td>
<td>671</td>
<td>.950</td>
</tr>
<tr>
<td>2600 - 2699</td>
<td>3</td>
<td>674</td>
<td>.955</td>
</tr>
<tr>
<td>2700 - 2799</td>
<td>3</td>
<td>677</td>
<td>.959</td>
</tr>
<tr>
<td>2800 - 2899</td>
<td>1</td>
<td>678</td>
<td>.960</td>
</tr>
<tr>
<td>2900 - 2999</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000 - 3099</td>
<td>3</td>
<td>681</td>
<td>.964</td>
</tr>
<tr>
<td>3100 - 3199</td>
<td>2</td>
<td>683</td>
<td>.967</td>
</tr>
<tr>
<td>3300 - 3399</td>
<td>1</td>
<td>684</td>
<td>.969</td>
</tr>
<tr>
<td>3500 - 3599</td>
<td>1</td>
<td>685</td>
<td>.970</td>
</tr>
<tr>
<td>3700 - 3799</td>
<td>2</td>
<td>687</td>
<td>.973</td>
</tr>
<tr>
<td>3800 - 3899</td>
<td>3</td>
<td>690</td>
<td>.977</td>
</tr>
<tr>
<td>3900 - 3999</td>
<td>2</td>
<td>692</td>
<td>.980</td>
</tr>
<tr>
<td>4000 - 4099</td>
<td>1</td>
<td>693</td>
<td>.981</td>
</tr>
<tr>
<td>4200 - 4299</td>
<td>1</td>
<td>694</td>
<td>.983</td>
</tr>
<tr>
<td>4700 - 4799</td>
<td>3</td>
<td>697</td>
<td>.987</td>
</tr>
<tr>
<td>4900 - 4999</td>
<td>3</td>
<td>700</td>
<td>.991</td>
</tr>
<tr>
<td>5100 - 5199</td>
<td>1</td>
<td>701</td>
<td>.993</td>
</tr>
<tr>
<td>5900 - 5999</td>
<td>1</td>
<td>702</td>
<td>.994</td>
</tr>
<tr>
<td>6200 - 6299</td>
<td>2</td>
<td>704</td>
<td>.997</td>
</tr>
<tr>
<td>6700 - 6799</td>
<td>1</td>
<td>705</td>
<td>.998</td>
</tr>
<tr>
<td>7000 - 7099</td>
<td>1</td>
<td>706</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Figure 4.1
Distribution of Index Length by Number of Index Entries
Fromdren 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 agreement 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 0.5. 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

- PN2598.k4 b6 1931, Slope = -1.66
- ND553.D774 T6 1965, Slope = -2.00

Log-log graph showing the relationship between the number of index entries and the number of page references.
The function that maximizes this ratio is the Zipf-Mandelbrot distribution:

\[ p(x) = c x^{-s} \]  

Substitution of (4.2) into (4.1) yields

\[ I = \frac{-\sum p(x) \log n}{\sum p(x) \log x} \]

\[ \text{I} = \frac{\sum (\log c - s \log x) cx^{-s}}{\sum cx^{-s} \log x} \]

\[ \text{I} = \frac{\sum cx^{-s} \log x}{\sum cx^{-s} \log x} \]

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

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,

\[ I = \frac{\zeta(s)}{\zeta(s)} \]

\[ I = \frac{\sum cx^{-s} \log x}{\sum cx^{-s} \log x} \]

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:
\[ s = \log \left( \frac{\text{number of references with single page locations}}{\log (\text{number of page locations of most popular index entry})} \right) \]

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 occurring 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>
<thead>
<tr>
<th>LC Number</th>
<th>( s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT7244</td>
<td>5.47</td>
</tr>
<tr>
<td>QD9</td>
<td>4.43</td>
</tr>
<tr>
<td>HB199</td>
<td>4.33</td>
</tr>
<tr>
<td>BV2532</td>
<td>3.81</td>
</tr>
<tr>
<td>DA690</td>
<td>3.54</td>
</tr>
<tr>
<td>TK153</td>
<td>3.53</td>
</tr>
<tr>
<td>E741</td>
<td>3.39</td>
</tr>
<tr>
<td>Q391</td>
<td>3.20</td>
</tr>
<tr>
<td>QA303</td>
<td>3.06</td>
</tr>
<tr>
<td>E178</td>
<td>2.81</td>
</tr>
<tr>
<td>QL703</td>
<td>2.71</td>
</tr>
<tr>
<td>RM721</td>
<td>2.71</td>
</tr>
<tr>
<td>BP181</td>
<td>2.69</td>
</tr>
<tr>
<td>DP521</td>
<td>2.64</td>
</tr>
<tr>
<td>Z5782</td>
<td>2.49</td>
</tr>
<tr>
<td>ND553</td>
<td>2.26</td>
</tr>
<tr>
<td>HM66</td>
<td>2.15</td>
</tr>
<tr>
<td>F864</td>
<td>2.08</td>
</tr>
<tr>
<td>PR2831</td>
<td>1.96</td>
</tr>
<tr>
<td>LB875</td>
<td>1.94</td>
</tr>
<tr>
<td>LC191</td>
<td>1.93</td>
</tr>
<tr>
<td>HF2046</td>
<td>1.86</td>
</tr>
<tr>
<td>D443</td>
<td>1.81</td>
</tr>
<tr>
<td>HD20</td>
<td>1.71</td>
</tr>
<tr>
<td>PR5588</td>
<td>1.69</td>
</tr>
<tr>
<td>PN2598</td>
<td>1.67</td>
</tr>
<tr>
<td>DS423</td>
<td>1.43</td>
</tr>
<tr>
<td>JA84</td>
<td>1.09</td>
</tr>
</tbody>
</table>
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 of 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

<table>
<thead>
<tr>
<th>LC Class</th>
<th>1/900th Entries</th>
<th>LC Subject Headings</th>
<th>Title</th>
</tr>
</thead>
</table>
| 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 |
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  
2. U.S.-Hist.-Historiography | Understanding the American Past |
<table>
<thead>
<tr>
<th>LC Class</th>
<th>1/900th Entries</th>
<th>LC Subject Headings</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Foreign Relations: Anglo-American</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Federal Income Tax: individual</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Tax: individual income</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Farmers, income of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Legislation: agricultural</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. Agricultural, legislation for</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8. Railroads: rates of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F864</td>
<td>1. Mass</td>
<td>1. Anza (sic!)</td>
<td>Anza's California Expedition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. California-descr. and travel</td>
<td></td>
</tr>
<tr>
<td>HB199</td>
<td>1. Terborgh, Gene</td>
<td>1. Economics</td>
<td>Engineering Economy</td>
</tr>
<tr>
<td></td>
<td>2. Breakeven Charts, Examples of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Industrial Management-Research</td>
<td></td>
</tr>
<tr>
<td>HM66</td>
<td>1. Trade Unions</td>
<td>1. Sociology</td>
<td>Social Theory</td>
</tr>
<tr>
<td>JA84</td>
<td>1. Economy</td>
<td>1. Political Science-Hist.-Russia</td>
<td>Russian Political Thought</td>
</tr>
<tr>
<td>LC Class</td>
<td>1/900th Entries</td>
<td>LC Subject Headings</td>
<td>Title</td>
</tr>
<tr>
<td>----------</td>
<td>----------------</td>
<td>---------------------</td>
<td>-------</td>
</tr>
<tr>
<td>LB875</td>
<td>America</td>
<td>1. Education</td>
<td>Two Views of Education</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Literature-Study and Teaching</td>
<td></td>
</tr>
<tr>
<td>LC191</td>
<td>Children, Disease of</td>
<td>none</td>
<td>Education and Social Progress</td>
</tr>
<tr>
<td>ND553</td>
<td>&quot;Bride Stripped Bare By Her Bachelors, Even, The&quot; (Ducamp)</td>
<td>1. Duchamp, Marcel, 1887-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Cage, John</td>
<td>The Bride and the Bachelors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Rauschenberg, Robert, 1925-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Tinguely, Jean, 1925-</td>
<td></td>
</tr>
<tr>
<td>PN2598</td>
<td>Butler, Pierce</td>
<td>None</td>
<td>Fanny Kemble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Shakespeare, William-Bibl.-Quartos</td>
<td></td>
</tr>
<tr>
<td>PR5588</td>
<td>Keats, John</td>
<td>None</td>
<td>Theme and Symbol in Tennyson's Poems to 1850</td>
</tr>
<tr>
<td>PT7244</td>
<td>Bjark: Bjarkamal, anon.</td>
<td>1. Scalds and Scaldic Poetry</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Icelandic and Old Norse Poetry</td>
<td></td>
</tr>
<tr>
<td>QA303</td>
<td>Cauchy, A.</td>
<td>1. Calculus</td>
<td>Vorlesungen Uber Differential und Integralrechnung</td>
</tr>
<tr>
<td>LC Class</td>
<td>1/900th Entries</td>
<td>LC Subject Headings</td>
<td>Title</td>
</tr>
<tr>
<td>----------</td>
<td>----------------</td>
<td>-------------------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Reference Books</td>
<td></td>
</tr>
<tr>
<td>QE391</td>
<td>1. Reseavres, India</td>
<td>1. Sillimanite</td>
<td>Sillimanite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Cyanite</td>
<td></td>
</tr>
<tr>
<td>QL703</td>
<td>1. Carnivore</td>
<td>1. Mammals</td>
<td>Principles in Mammalogy</td>
</tr>
<tr>
<td></td>
<td>2. Bat(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM721</td>
<td>1. Muscle Contraction</td>
<td>1. Gymnastics, Medical</td>
<td>Therapeutic Exercise</td>
</tr>
<tr>
<td>TK153</td>
<td>1. Tube, gaseous-discharge</td>
<td>1. Electrical apparatus and appliances</td>
<td>Electrons at Work</td>
</tr>
<tr>
<td></td>
<td>2. Hertz</td>
<td>2. Electrons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Light, ultra-violet</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Maxwell</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Reaction, Reactors</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Valence electrons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z5782</td>
<td>1. Passion</td>
<td>1. Drama, Medieval-Bibl.</td>
<td>Bibliography of Medieval Drama</td>
</tr>
<tr>
<td></td>
<td>2. Comedy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Latin</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Staging</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

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

<table>
<thead>
<tr>
<th>Number of Words</th>
<th>Number of Entries</th>
<th>Cumulative Number</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3673</td>
<td>3673</td>
<td>13.51</td>
</tr>
<tr>
<td>2</td>
<td>6563</td>
<td>10236</td>
<td>37.65</td>
</tr>
<tr>
<td>3</td>
<td>4817</td>
<td>15053</td>
<td>55.37</td>
</tr>
<tr>
<td>4</td>
<td>3905</td>
<td>18958</td>
<td>69.73</td>
</tr>
<tr>
<td>5</td>
<td>2819</td>
<td>21797</td>
<td>80.17</td>
</tr>
<tr>
<td>6</td>
<td>1969</td>
<td>23766</td>
<td>87.41</td>
</tr>
<tr>
<td>7</td>
<td>1243</td>
<td>25009</td>
<td>91.99</td>
</tr>
<tr>
<td>8</td>
<td>801</td>
<td>25810</td>
<td>94.93</td>
</tr>
<tr>
<td>9</td>
<td>516</td>
<td>26326</td>
<td>96.83</td>
</tr>
<tr>
<td>10</td>
<td>281</td>
<td>26607</td>
<td>97.86</td>
</tr>
<tr>
<td>&gt;10</td>
<td>581</td>
<td>27188</td>
<td>100.00</td>
</tr>
</tbody>
</table>
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


CHAPTER V

ALGORITHMIC TEXT INDEXING
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 commercial 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 examined 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

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

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 a 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 methods 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 deletion should not be implemented.

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 inclusion or exclusion 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
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 meeluanizable) 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.

access-per-item, 15, 16
access files, 16
access points, 18, 25, 154
accession distribution, 103, 104
accession number, 139, 141
accession year, 102, 103
accessions growth, 98, 102
acquisition expenditures, 8, 9
acquisition growth, 8, 103; see also accessions growth
AID, 119; see also Agency for International Development
algebraic notation, 57
Agency for International Development, 119
Algeria, 120
ALGOL, 27, 54; see also Computer Languages
Alphatype, 64
Alphavers Book Condensed, 65
ALTEXT, 57, 58, 59; see also Computer Languages
American Civil War, 5, 104, 112
archival collection, 17, 39, 95
archival component, 12
archival libraries, 16, 18, 122
Argentina, 122
Asia, 128
Author, 146
authority list, 36, 81
Baltimore County Library, 156
Belgium, 119
Bell Telephone Laboratories, 55
bibliographic description, 75, 137, 139, 144, 153
bibliographic files, 18, 136, 137
bibliographic material, 71, 150
bibliographic record, 16, 18, 71, 73, 74, 81, 82, 83, 149

Figure 5.2
ALGORITHMIC INDEX TO REFERENCE (3)
bibliographic tool, 135, 144
bibliographical reference, 69, 126
bold face, 63, 64, 65, 68
book catalogs, 1, 31, 69, 88
Book Condensed, 65; see Alphavers
Book Condensed type font
books-per-billion ratio, 123, 124
Brazil, 122
British Museum, 65, 115
Bro-Dart catalog, 69
Canada, 119, 122, 123, 124
capitalization and other errors, 81
card catalog, 1, 25, 73, 87, 93, 114, 135
cards per catalog drawer, 89
catalog
avtomation, 15, 37, 39
card, 41, 44, 45, 89, 92, 93, 95, 103
conversion, 40
croptry, 39, 68, 72, 75, 149
cfile, 137, 138, 143, 153, 157
information, 23, 36, 40, 155, 157
material, 70, 154
mechanization, 16
operation, 26, 39, 156
record, 18, 114, 142, 144
system, 1, 114
use, 24
characters per record, 17, 43
cichters per square inch, 61, 62
Chemical Abstracts, 116, 124
China, 128
Church, 117; see Roman Catholic Church
CHY, 41, 43; see Columbia-Harvard-Yale Medical Libraries
circulation rate, 12, 130
Civil War, 112, 113; see American Civil War
class field, 141, 142, 143, 144, 146, 148, 152
class number, 95, 104, 142
COBOL, 25, 27; see also Computer Languages
collection growth, 98, 102
collection size, 89, 137, 143
Columbia-Harvard-Yale Medical Libraries, 41; see CHY
columns per page, 68, 69
COMIT, 53, 54, 55, 56, 57, 58, 59; see also Computer Languages
COMIT II, 53, 54; see also Computer Languages
cmparative input equipment costs, 45
composite figures, 43, 44
composition cost, 69, 153
Computer Languages
ALGOL, 27, 54
ALTEXT, 57, 58, 59
COBOL, 25, 27
COMIT, 53, 54, 55, 56, 57, 58, 59
COMIT II, 53, 54
FORTRAN, 25, 27, 54
LISP, 56
LISP 1.5, 56
PL-1, 27, 71, 72
SNOBOL, 27, 28, 55, 56, 57, 58, 59
SNOBOL 3, 55, 56
SNOBOL 4, 55, 56
tEMAC, 28, 56, 57
XPOP, 57, 58, 59
conversion costs, 16, 17, 40
conversion project, 16, 18
corporate authorship, 93
cost
cmparative input equipment, 45
composition, 69, 153
cnversion, 16, 17, 40
machine, 29, 31, 40
operation, 31, 157
personnel, 15, 16, 31
quality, 43
software, 25, 38
cost per character, 16, 44
cost reductions, 36, 51
Costa Rica, 123
county library system, 152, 156
cyrillic, 75
Czechoslovakia, 119
data base, 78, 126, 127
data points, 9, 89
East Germany, 119
economic growth, 8, 15, 115, 128
economic indicators, 19, 115
economic statistics, 117
edit lists, 43, 44
electronic composition, 17, 19, 31, 70
Embassy, 119
English, 94, 127, 128, 131, 143
English language share, 128
error-detection, 43, 71, 73, 74
error messages, 71, 74
error rate, 74
Europe, 119
European GNP, 119
exponential growth, 8, 15, 16, 29, 95, 100, 104, 130, 144, 153
Finland, 117
First World War, 117
Fondren, 89, 92, 93, 104, 112, 131, 132, 133, 152; see Rice University
Fondren Library, 5, 87, 92, 95
Fondren Sample, 89, 92, 93, 102, 122
Fondren shelf list, 89, 92, 95
FORTRAN, 25, 27, 54; see also Computer Languages
France, 120
French, 119, 120, 124, 131
French GNP, 120
German, 131
Germany, 127, 148
GNP, 98, 102, 103, 116, 117, 119, 120, 121, 122, 123, 124, 126, 127, 128, 130, 132; see gross national product
graphic arts quality, 32, 64, 153, 154
Great Depression, 5, 113
Great War, 2, 5; see First World War
gross national product, 8, 12, 15, 98, 100, 103, 116, 117, 119, 130, 131, 133; see GNP
growth rate, 2, 5, 51, 88, 95, 98, 100, 102, 103, 104, 115, 119, 124, 128
Harvard, 71, 75, 78, 82; see Harvard University, Harvard sample, Widener Library
Harvard sample, 77, 78, "9, 30, 81
Harvard University, 2, 75
Hebrew, 121
History of the United States, 104
Honduras, 123
IBM
SelecEn Typewriter, 64
360, 57
709-7090-7094-7040-7044, 53
1401, 41, 75
7094, 41, 57
imprint date, 2, 98, 102, 104, 116, 131
imprint date distribution, 5, 103
imprint distribution, 5, 104
imprint growth, 98, 103
India, 128
Indonesia, 128
information density, 61, 68, 69
information explosion, 5, 12
information retrieval, 53, 93, 130, 140
Italian, 119, 120, 131
Italy, 120
item field, 141, 142, 143, 144, 146, 152
Japan, 127, 128
Japanese, 127, 128
journal title, 126
justification losses, 68, 69
keyboard operator, 73, 75, 81, 83
LACPL, 40, 41, 45, 45; see LACPL, Los Angeles County Public Library
Index

SUL, 41, 43; see Stanford Undergraduate Library, Meyer Undergraduate Library
Switzerland, 119, 120

telephone directory, 68, 69, 155
TEMAC, 28, 56, 57; see also Computer Languages
time intervals, 8, 98
title field, 141, 143, 152
title information, 74, 139
title list, 2, 153, 154
titles per subject, 147, 148
trend line, 12
Type Face Design, 61
type face, 61, 62, 64, 65, 68, 70

UC/B, 40, 43; see University of California/Berkeley
United Kingdom, 127
United States, 98, 100, 103, 112, 122, 123, 124, 127
United States GNP, 124
United States Gross National Product, 103
university collection, 8, 130
university library, 5, 12, 24, 38, 51, 92, 130, 133, 137, 140, 156
University of California/Berkeley, 40; see UC/B

University of Chicago, Graduate Library School, 53
University of Newcastle-upon-Tyne Library, 138; see Newcastle
Uruguay, 123
U.N., 119
U.N. Yearbook, 119
U.N. Yearbook of National Account Statistics, 119
U.S. Department of State, 119

Versatile Bold, 65
volumes per card, 88, 94
Volumes per Title, 94

Wall Street Journal, 31
Wang algorithm, 56
West Germany, 127
Widener Library, 2, 155, 156; see also Harvard
Widener Shelf List, 75, 148, 155
World Almanacs, 2
World War I, see First World War, Great War
World War II, 5, 100; see Second World War
World Wars, 5, 124

XPOP, 57, 58, 59; see also Computer Languages
<table>
<thead>
<tr>
<th>Excluded Index Terms Referring to One Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>abnormal parenthetization</td>
</tr>
<tr>
<td>absolute frequencies</td>
</tr>
<tr>
<td>academic staff</td>
</tr>
<tr>
<td>access capability</td>
</tr>
<tr>
<td>access point</td>
</tr>
<tr>
<td>access system</td>
</tr>
<tr>
<td>accessible estimates</td>
</tr>
<tr>
<td>accession distributions</td>
</tr>
<tr>
<td>accessions growth rate</td>
</tr>
<tr>
<td>acquisition rates</td>
</tr>
<tr>
<td>acquisitions data</td>
</tr>
<tr>
<td>acquisition mechanisms</td>
</tr>
<tr>
<td>acquisition process</td>
</tr>
<tr>
<td>acquisition rate</td>
</tr>
<tr>
<td>acquisition schedule</td>
</tr>
<tr>
<td>acquisition shares</td>
</tr>
<tr>
<td>acquisition structure</td>
</tr>
<tr>
<td>acquisitions - GNP relation</td>
</tr>
<tr>
<td>acquisitions - GNP share equality</td>
</tr>
<tr>
<td>acquisitions budget</td>
</tr>
<tr>
<td>acquisitions growth</td>
</tr>
<tr>
<td>acquisitions expenditures</td>
</tr>
<tr>
<td>adequate user access</td>
</tr>
<tr>
<td>algebraic equations</td>
</tr>
<tr>
<td>algebraic expressions</td>
</tr>
<tr>
<td>alpha-numeric code</td>
</tr>
<tr>
<td>alphabetic code</td>
</tr>
<tr>
<td>approximate linearity</td>
</tr>
<tr>
<td>approximate normality</td>
</tr>
<tr>
<td>archival collections</td>
</tr>
<tr>
<td>archival holdings</td>
</tr>
<tr>
<td>archival libraries</td>
</tr>
<tr>
<td>archival records</td>
</tr>
<tr>
<td>Assembly code programming</td>
</tr>
<tr>
<td>&quot;assembly languages&quot;</td>
</tr>
<tr>
<td>assignment procedure</td>
</tr>
<tr>
<td>author access</td>
</tr>
<tr>
<td>author field</td>
</tr>
<tr>
<td>author list</td>
</tr>
<tr>
<td>author name</td>
</tr>
<tr>
<td>author/title list</td>
</tr>
<tr>
<td>authority lists</td>
</tr>
<tr>
<td>automated catalog</td>
</tr>
<tr>
<td>Auxiliary memory</td>
</tr>
<tr>
<td>average cost</td>
</tr>
<tr>
<td>average growth</td>
</tr>
<tr>
<td>average number</td>
</tr>
<tr>
<td>average record length</td>
</tr>
<tr>
<td>average time</td>
</tr>
<tr>
<td>p Baltimore County Libraries</td>
</tr>
<tr>
<td>bedroom states</td>
</tr>
<tr>
<td>biases inherent</td>
</tr>
<tr>
<td>p bibliographic descriptions</td>
</tr>
<tr>
<td>bibliographic holdings</td>
</tr>
<tr>
<td>bibliographic indications</td>
</tr>
<tr>
<td>bibliographic items</td>
</tr>
<tr>
<td>bibliographic listings</td>
</tr>
<tr>
<td>bibliographic lists</td>
</tr>
<tr>
<td>bibliographic notes</td>
</tr>
<tr>
<td>bibliographic practice</td>
</tr>
<tr>
<td>p bibliographic records</td>
</tr>
<tr>
<td>bibliographic references</td>
</tr>
<tr>
<td>bibliographical information</td>
</tr>
<tr>
<td>bibliographically incomplete</td>
</tr>
<tr>
<td>bibliography</td>
</tr>
<tr>
<td>Bibliography Field</td>
</tr>
<tr>
<td>bibliography section</td>
</tr>
<tr>
<td>book-publication depressions</td>
</tr>
<tr>
<td>Book Length</td>
</tr>
<tr>
<td>bookseller</td>
</tr>
<tr>
<td>budget dollar</td>
</tr>
<tr>
<td>budgetary requirements</td>
</tr>
<tr>
<td>business community</td>
</tr>
<tr>
<td>calculus text</td>
</tr>
<tr>
<td>call number</td>
</tr>
<tr>
<td>(Canadian)census figures</td>
</tr>
<tr>
<td>capital letters</td>
</tr>
<tr>
<td>capitalization conventions</td>
</tr>
<tr>
<td>capitalization errors</td>
</tr>
<tr>
<td>capitalization requirements</td>
</tr>
<tr>
<td>card catalog collection</td>
</tr>
<tr>
<td>card collection</td>
</tr>
<tr>
<td>card files</td>
</tr>
<tr>
<td>card space convention</td>
</tr>
<tr>
<td>card system</td>
</tr>
<tr>
<td>cards per entry</td>
</tr>
<tr>
<td>cards per title</td>
</tr>
<tr>
<td>&quot;careful&quot; study</td>
</tr>
<tr>
<td>case alphabet</td>
</tr>
<tr>
<td>catalog card conversion errors</td>
</tr>
<tr>
<td>p catalog cards</td>
</tr>
<tr>
<td>catalog data</td>
</tr>
<tr>
<td>p catalog files</td>
</tr>
<tr>
<td>catalog interrogations</td>
</tr>
<tr>
<td>catalog preparation</td>
</tr>
<tr>
<td>catalog productions</td>
</tr>
</tbody>
</table>
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
p 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 ease
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
adata 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 growth challenge
exponential library growth rates growth periods
exponential rate
faculty library committee
feedback response
field names
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
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
gross national products Gross Personal Income ground-level extension
growth challenges
explosive growth
Gross National Products
exponential curve
exponential expansion
exponential function
exponential imprint date distributions growth challenge
exponential library growth rates
exponential rates
faculty library committee
feedback response
field names
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
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
Gross National Products
gross national products
Gross Personal Income
ground-level extension
growth challenges
explosive growth
Gross National Products
exponential curve
exponential expansion
exponential function
exponential imprint date distributions growth challenge
exponential library growth rates
exponential rates
faculty library committee
feedback response
field names
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
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
Gross National Products
gross national products
Gross Personal Income
ground-level extension
growth challenges
explosive growth
Gross National Products
exponential curve
exponential expansion
exponential function
exponential imprint date distributions growth challenge
exponential library growth rates
exponential rates
faculty library committee
feedback response
field names
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
fundamental structure
future funding needs
geographic area
geometric decrease
global category
global check
global war
GNP-acquisitions relation
interlibrary loans
interword spaces
interpretive approach
item
item fields
item purposes
items per year
journal-to-language assignment

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 shares
Library of Congress classification system
Library of Congress nonserial acquisitions
Library of Congress size distribution

p
library operations
process
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
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
machine languages

machine language instructions
machine methods
machine output
machine rules
machine time usage
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
multi-language manipulation procedures
ordinary numbers
multiple copy graphic arts quality author list
out-of-date catalog
musical scores
output error signals
national accounts statistics
output list
national economic growth
output machines
national economy
output printers
national origins
output sheet
national publications
page counts
national publishers
page design
national statistical data collection processes
paper costs
natural languages
(paragraph) tape input
(new) acquisitions information
parallel search logic
non-English words
pattern-matching facilities
non-numerical procedures
pattern-valued functions
nonlibrary customers
pattern primitives
nonlinear scales
per capita growth
per unit basis
percentage growth
personal author
personal authorship
personal incomes
photo-offset reproduction
(physical) volumes per serial
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
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 frequency distribution
title
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 literature
study
scientific publication
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 - GNP
disagreement
social dislocation
social ideologies
social phenomena
social systems
sort operation
source statement language
source statement structure
special purpose bibliographies
square inch
staff expansion requests
standard algebraic form
standard algebraic notation
standard precedence conventions
stationary growth rate
statistical correlation theory
statistical distribution
statistical ensemble
statistical indicator
statistical relationship
statistical summary
statistical uniformities
status indicators
storage media
storage space
string contents
string processor
structure
subject-oriented bibliographies
subject-title catalog
subject area
subject areas
subject bibliographies
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
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
title card
Title cards
title indices
title languages
transcription errors
transliteration schemes
tray contents
trend curve
turn-around times
type face catalog
type face size
unit cost
unit costs
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
utility
utilization costs
vertical scale
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>
<thead>
<tr>
<th>Number of Words</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>190</td>
<td>272</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>316</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>329</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>335</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>339</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>340</td>
</tr>
</tbody>
</table>

Table 5.3

Index Entry Length Distribution
Computerized Library Catalogs
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
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

NUMBER OF PAGE REFERENCES

10^3

10^2

10

1

10

100

1000

NUMBER OF INDEX ENTRIES

10

100
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


CHAPTER VI

AMALGAMATIVE ACCESS MECHANISMS
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 pandemics

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 \((2e)^2 \approx 30\) 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 every access level is filled. 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:

\[
\text{Article, Abstract, Title} \quad \text{and} \quad \text{Book, Index, Table of Contents}
\]

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 encyclopedias, 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 database.

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
WILLIAM FELLER
AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS, VOLUME I. 3RD ED.
CHARLES J. STONE, UCLA

There are a number of significant changes in Feller's excellent and unique introduction to probability theory. The character of the book, however, remains faithful to the original edition.

According to Feller in his preface to this edition, the greatest change is in the chapter on fluctuation theory. This chapter was introduced only in the second edition, which was in fact motivated principally by the unexpected discovery that its enticing material could be treated by elementary methods. But this treatment still depended on combinatorial artifacts which have now been replaced by simpler and more natural probabilistic arguments. In essence, this chapter is new.

Other changes pointed out by Feller in his preface or noted by the reviewer include a better treatment of independent trials; restatements of some of the results on the normal approximation to the binomial distribution; some additional material on branching processes; a more detailed proof of the basic limit theorem of renewal theory, and a new section in the chapter on Markov chains covering taboo probabilities and the null limit theorem for null recurrent chains.

- NUMBER OF SIGNIFICANT
  PROBABILITY THEORY

- FLUCTUATION THEORY

- COMBINATORIAL ARTIFICES

- NATURAL PROBABILITY ARGUMENTS

- TREATMENT OF INDEPENDENT TRIALS

- NORMAL APPROXIMATION

- BINOMIAL DISTRIBUTION

- BASIC LIMIT THEOREM OF RENEWAL THEORY

- MARKOV CHAINS

- TABOO PROBABILITIES

- RATIO LIMIT THEOREM

- NULL RECURRENT CHAINS
### Figure 6.3
Cumulative Index to 50 Abstracts (one page)

<table>
<thead>
<tr>
<th>Noncentral Distributions</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noncentral Multivariate Beta Distribution</td>
<td>43</td>
</tr>
<tr>
<td>Nonnull Distributions</td>
<td>29</td>
</tr>
<tr>
<td>Nonparametric Alternative</td>
<td>42</td>
</tr>
<tr>
<td>Nonparametric Alternatives</td>
<td>37</td>
</tr>
<tr>
<td>Nonparametric Tests</td>
<td>10</td>
</tr>
<tr>
<td>Normal Alternatives</td>
<td>37</td>
</tr>
<tr>
<td>Normal Approximation</td>
<td>1</td>
</tr>
<tr>
<td>Normal Distribution</td>
<td>32</td>
</tr>
<tr>
<td>Normal Distributions</td>
<td>33</td>
</tr>
<tr>
<td>Normal Distributions</td>
<td>29</td>
</tr>
<tr>
<td>Normal Distributions</td>
<td>48</td>
</tr>
<tr>
<td>Normal Populations</td>
<td>29</td>
</tr>
<tr>
<td>Normal Theory Likelihood Ratio Statistic</td>
<td>27</td>
</tr>
<tr>
<td>Normal Theory Likelihood Ratio Test Statistic</td>
<td>27</td>
</tr>
<tr>
<td>Null Distribution</td>
<td>29</td>
</tr>
<tr>
<td>Null Distributions</td>
<td>25</td>
</tr>
<tr>
<td>Null Hypothesis</td>
<td>29</td>
</tr>
<tr>
<td>Null Hypothesis</td>
<td>29</td>
</tr>
<tr>
<td>Null Recurrent Chains</td>
<td>1</td>
</tr>
<tr>
<td>Number Successes</td>
<td>23</td>
</tr>
<tr>
<td>Numerical Comparisons</td>
<td>41</td>
</tr>
<tr>
<td>Numerical Examples</td>
<td>29</td>
</tr>
<tr>
<td>One-Dimensional Empirical Process Converge</td>
<td>7</td>
</tr>
<tr>
<td>Onto Itself</td>
<td>47</td>
</tr>
<tr>
<td>Operational Characteristics</td>
<td>17</td>
</tr>
<tr>
<td>Optimal Allocation</td>
<td>2</td>
</tr>
<tr>
<td>Optimal Allocation Problems</td>
<td>2</td>
</tr>
<tr>
<td>Optimal Stratified Sampling</td>
<td>2</td>
</tr>
<tr>
<td>Optimum Allocation</td>
<td>19</td>
</tr>
<tr>
<td>Optimum Best Linear</td>
<td>33</td>
</tr>
<tr>
<td>Optimum Blue's</td>
<td>33</td>
</tr>
<tr>
<td>Optimum Non-parametric Statistics</td>
<td>30</td>
</tr>
<tr>
<td>Order Absolute Central Moment</td>
<td>26</td>
</tr>
<tr>
<td>Order Statistics</td>
<td>13</td>
</tr>
<tr>
<td>Order Statistics</td>
<td>13</td>
</tr>
<tr>
<td>Order Statistics</td>
<td>33</td>
</tr>
<tr>
<td>Order Statistics</td>
<td>33</td>
</tr>
<tr>
<td>Overall Average</td>
<td>40</td>
</tr>
<tr>
<td>p-Dimensional Space</td>
<td>11</td>
</tr>
<tr>
<td>p-Dimensional Variate</td>
<td>2</td>
</tr>
<tr>
<td>p-Variant Normal Populations</td>
<td>8</td>
</tr>
<tr>
<td>Paper Generalizes</td>
<td>36</td>
</tr>
<tr>
<td>Paper Treats *-Comparis</td>
<td>17</td>
</tr>
<tr>
<td>Parameter Set</td>
<td>17</td>
</tr>
<tr>
<td>Parameter Sets</td>
<td>17</td>
</tr>
<tr>
<td>Parametric Classes</td>
<td>5</td>
</tr>
<tr>
<td>Past Observations Available</td>
<td>39</td>
</tr>
<tr>
<td>Phase Distribution</td>
<td>3</td>
</tr>
<tr>
<td>Phase Distribution</td>
<td>4</td>
</tr>
<tr>
<td>Phase Service Time Distribution</td>
<td>4</td>
</tr>
<tr>
<td>Population Mean</td>
<td>16</td>
</tr>
<tr>
<td>Population Mean</td>
<td>40</td>
</tr>
<tr>
<td>Population Size</td>
<td>23</td>
</tr>
</tbody>
</table>

156
147
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 make 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 amalgamated 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**

<table>
<thead>
<tr>
<th>Number of Words</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>315</td>
<td>315</td>
</tr>
<tr>
<td>3</td>
<td>233</td>
<td>548</td>
</tr>
<tr>
<td>4</td>
<td>125</td>
<td>673</td>
</tr>
<tr>
<td>5</td>
<td>54</td>
<td>727</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>746</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>752</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>754</td>
</tr>
</tbody>
</table>

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 6.5
Abstract Number Location Distribution
Algorithmic Index to Statistical Abstracts

Number of Abstract References

Number of Index Entries

$10^3$

$10^2$

$10$

$1$
Table 6.2
Abstract Number Location Distribution, Algorithmic Index to Statistical Abstracts

<table>
<thead>
<tr>
<th>Number of Abstract Locations per Entry</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>703</td>
<td>703</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>734</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>740</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>749</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>754</td>
</tr>
</tbody>
</table>

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 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 80 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>
<thead>
<tr>
<th>A</th>
<th>Elementary Decision Theory</th>
<th>Chernoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>Nonparametric Methods in Statistics</td>
<td>Fraser</td>
</tr>
<tr>
<td>C</td>
<td>Statistical Methods for Chemists</td>
<td>W. J. Youden</td>
</tr>
<tr>
<td>D</td>
<td>Analysis of Straight-line Data</td>
<td>Acton</td>
</tr>
<tr>
<td>E</td>
<td>Testing Statistical Hypotheses</td>
<td>F. L. Lehmann</td>
</tr>
<tr>
<td>F</td>
<td>Introduction to Mathematical Statistics</td>
<td>Paul G. Hoel</td>
</tr>
<tr>
<td>G</td>
<td>The Design and Analysis of Experiments</td>
<td>O. Kempthorne</td>
</tr>
<tr>
<td>H</td>
<td>An Introduction to Multivariate Statistical Analysis</td>
<td>T. W. Anderson</td>
</tr>
<tr>
<td>I</td>
<td>Statistics--An Introduction</td>
<td>D. A. S. Fraser</td>
</tr>
<tr>
<td>J</td>
<td>Linear Computations</td>
<td>Paul S. Dwyer</td>
</tr>
<tr>
<td>K</td>
<td>Modern Probability Theory and Its Applications</td>
<td>Parzen</td>
</tr>
<tr>
<td>L</td>
<td>Planning of Experiments</td>
<td>Feller</td>
</tr>
<tr>
<td>M</td>
<td>Theory of Games and Statistical Decisions</td>
<td>Blackwell and Girshick</td>
</tr>
<tr>
<td>N</td>
<td>An Introduction to Probability Theory and Its Applications, Vol 1</td>
<td>Feller</td>
</tr>
<tr>
<td>O</td>
<td>Elementary Statistics</td>
<td>Paul G. Hoel</td>
</tr>
<tr>
<td>P</td>
<td>The Elements of Probability</td>
<td>Cramer</td>
</tr>
<tr>
<td>Q</td>
<td>Statistical Decision Theory</td>
<td>Weiss</td>
</tr>
<tr>
<td>R</td>
<td>Introduction to Probability and Random Variables</td>
<td>Wadsworth and Bryan</td>
</tr>
<tr>
<td>S</td>
<td>Introduction to the Theory of Statistics</td>
<td>Mood and Graybill</td>
</tr>
<tr>
<td>T</td>
<td>Elements of Probability and Statistics</td>
<td>Wolf</td>
</tr>
<tr>
<td>U</td>
<td>An Introduction to Linear Statistical Models, Vol 1</td>
<td>Graybill</td>
</tr>
<tr>
<td>V</td>
<td>Elements of the Theory of Markov Processes and Their Applications</td>
<td>Bharucha-Reid</td>
</tr>
<tr>
<td>W</td>
<td>Geometrical Probability</td>
<td>Kendall and Moran</td>
</tr>
<tr>
<td>X</td>
<td>Fundamentals of Statistical Reasoning</td>
<td>Quenouille</td>
</tr>
<tr>
<td>Y</td>
<td>Characteristic Functions</td>
<td>Lukas</td>
</tr>
<tr>
<td>Z</td>
<td>An Introduction to Probability Theory and Its Applications, Vol 2</td>
<td>Feller</td>
</tr>
<tr>
<td>AB</td>
<td>Elements of Mathematical Statistics</td>
<td>Alexander</td>
</tr>
<tr>
<td>AC</td>
<td>Statistical Theory and Methodology in Science and Engineering</td>
<td>Brownlee</td>
</tr>
<tr>
<td>AD</td>
<td>Statistics and Experimental Design, Vol 1</td>
<td>Johnson and Leone</td>
</tr>
<tr>
<td>AE</td>
<td>Mathematical Statistics</td>
<td>Wilkes</td>
</tr>
<tr>
<td>AF</td>
<td>Experimental Designs</td>
<td>Cochran and Cox</td>
</tr>
<tr>
<td>AI</td>
<td>A Course in Probability Theory</td>
<td>Kai Lai Chung</td>
</tr>
<tr>
<td>AJ</td>
<td>Essentials of Probability</td>
<td>Arthur Yaspan</td>
</tr>
<tr>
<td>AK</td>
<td>The Design of Experiments</td>
<td>Fisher</td>
</tr>
<tr>
<td>AL</td>
<td>Computational Handbook of Statistics</td>
<td>Bruning and Kintz</td>
</tr>
<tr>
<td>AM</td>
<td>Design and Analysis of Experiments</td>
<td>Quenouille</td>
</tr>
<tr>
<td>AN</td>
<td>Handbook of Statistical Tables</td>
<td>Owen</td>
</tr>
<tr>
<td>AO</td>
<td>The Elements of Probability</td>
<td>Berman</td>
</tr>
<tr>
<td>AP</td>
<td>Design and Analysis of Industrial Experiments</td>
<td>Davies</td>
</tr>
<tr>
<td>AQ</td>
<td>Statistical Theory</td>
<td>Lindgren</td>
</tr>
<tr>
<td>AR</td>
<td>Introduction to Statistics</td>
<td>Carlborg</td>
</tr>
<tr>
<td>AS</td>
<td>Probability and Statistics</td>
<td>Adler and Rossler</td>
</tr>
<tr>
<td>AT</td>
<td>Measuring Uncertainty--An Elementary Introduction to Bayesian Statistics</td>
<td></td>
</tr>
<tr>
<td>AU</td>
<td>A Brief Introduction to Probability Theory</td>
<td></td>
</tr>
<tr>
<td>AV</td>
<td>Statistical Design and Analysis of Experiments for Development Research</td>
<td>Villars</td>
</tr>
<tr>
<td>AW</td>
<td>Statistics in Research</td>
<td>Ostle</td>
</tr>
<tr>
<td>AX</td>
<td>Schaums Outline Series Theory and Problems of Probability</td>
<td>Lipschutz</td>
</tr>
<tr>
<td>AY</td>
<td>Elementary Mathematical Programming</td>
<td>Metzger</td>
</tr>
<tr>
<td>AZ</td>
<td>Statistical Inference for Markov Processes</td>
<td>Billingsley</td>
</tr>
<tr>
<td>Code</td>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>RC</td>
<td>Introduction to Probability: A Programmed Unit in Modern Mathematics</td>
<td></td>
</tr>
<tr>
<td>RD</td>
<td>Statistical Analysis of Stationary Time Series - Grenander and Rosenblatt</td>
<td></td>
</tr>
<tr>
<td>RR</td>
<td>Statistical Methods in Experimentation: An Introduction - Lacey</td>
<td></td>
</tr>
<tr>
<td>BF</td>
<td>Stochastic Processes: Basic Theory and Its Application - Prabhu</td>
<td></td>
</tr>
<tr>
<td>RG</td>
<td>Probability and Frequency - Plummer</td>
<td></td>
</tr>
<tr>
<td>RH</td>
<td>Statistical Methods for Research Workers</td>
<td></td>
</tr>
<tr>
<td>RI</td>
<td>Probability, an Intermediate Text-Book - Rixley</td>
<td></td>
</tr>
<tr>
<td>RJ</td>
<td>Regression Analysis - Williams</td>
<td></td>
</tr>
<tr>
<td>NK</td>
<td>Statistical Processes and Reliability Engineering - Chorafass</td>
<td></td>
</tr>
<tr>
<td>RL</td>
<td>Introduction to Probability and Mathematical Statistics - Birnbaum</td>
<td></td>
</tr>
<tr>
<td>EM</td>
<td>Elementary Mathematical Statistics - Bateman</td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>Introduction to Biostatistics - Bancroft</td>
<td></td>
</tr>
<tr>
<td>BO</td>
<td>Sampling Techniques - Cochran</td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>A History of the Mathematical Theory of Probability - Todhunter</td>
<td></td>
</tr>
<tr>
<td>BQ</td>
<td>Statistical Methods in Biology - Bailey</td>
<td></td>
</tr>
<tr>
<td>BR</td>
<td>Statistical Theory: The Relationship of Probability, Credibility and Error</td>
<td></td>
</tr>
<tr>
<td>BS</td>
<td>An Introduction to Multivariate Statistical Analysis - Anderson</td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td>Probability and Experimental Errors in Science - Farratt</td>
<td></td>
</tr>
<tr>
<td>RU</td>
<td>Contributions to Order Statistics - Sarhan and Greenberg</td>
<td></td>
</tr>
<tr>
<td>BV</td>
<td>Introduction to Statistical Method - Ehrenfeld and Littauer</td>
<td></td>
</tr>
<tr>
<td>RW</td>
<td>Theory of Probability - Jeffreys</td>
<td></td>
</tr>
<tr>
<td>BX</td>
<td>Statistical Adjustment of Data - Deming</td>
<td></td>
</tr>
<tr>
<td>BY</td>
<td>Statistical Analysis in Chemistry and the Chemical Industry - Bennett and Franklin</td>
<td></td>
</tr>
<tr>
<td>BZ</td>
<td>Probability Random Variables and Stochastic Processes - Papoulis</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>Elements of Queueing Theory with Applications - Saaty</td>
<td></td>
</tr>
<tr>
<td>CE</td>
<td>Stochastic Processes - Doob</td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>Sample Survey Methods and Theory Vol 1 Methods and Applications - Hansen, Hurwitz</td>
<td></td>
</tr>
<tr>
<td>CG</td>
<td>Advanced Statistical Methods in Biometric Research - C Radhakrishna Rao</td>
<td></td>
</tr>
<tr>
<td>CH</td>
<td>Introduction to the Mathematics of Statistics - Robert W. Burgess</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6.6
Rank - Frequency of Reference Distribution
Statistical Index Sample

[Graph showing a rank-frequency distribution with statistical index sample information.]
Table 6.4

ABSTRACT ENTRIES FOR THE
AMALGAMATED STATISTICS INDEX SAMPLE

<table>
<thead>
<tr>
<th>Term</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal distribution</td>
<td>Fisher, R. A.</td>
</tr>
<tr>
<td>Binomial distribution</td>
<td>Student</td>
</tr>
<tr>
<td>Poisson distribution</td>
<td>Pearson, E. S.</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>Kendall, M. G.</td>
</tr>
<tr>
<td>Conditional probability</td>
<td>Bartlett, M. S.</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>Cramer, H.</td>
</tr>
<tr>
<td>Analysis of variance</td>
<td>Neyman, J.</td>
</tr>
<tr>
<td>Distribution</td>
<td></td>
</tr>
<tr>
<td>Chi-square distribution</td>
<td></td>
</tr>
<tr>
<td>Central limit theorem</td>
<td></td>
</tr>
<tr>
<td>Least squares</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
</tr>
<tr>
<td>Cauchy distribution</td>
<td></td>
</tr>
<tr>
<td>Covariance</td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td></td>
</tr>
<tr>
<td>Random variable</td>
<td></td>
</tr>
<tr>
<td>Exponential distribution</td>
<td></td>
</tr>
<tr>
<td>Gamma distribution</td>
<td></td>
</tr>
<tr>
<td>Moments</td>
<td></td>
</tr>
<tr>
<td>Bivariate normal distribution</td>
<td></td>
</tr>
<tr>
<td>Multinomial distribution</td>
<td></td>
</tr>
</tbody>
</table>

157
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.
Appendix I

Abstract Index Entries:
A Uniform Sample from the
Fondren Index Sample
<table>
<thead>
<tr>
<th>Name</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marston, William Moulton</td>
<td>23</td>
</tr>
<tr>
<td>Integrative Psychology</td>
<td></td>
</tr>
<tr>
<td>Sum = 1432 / 29.54 = 48</td>
<td></td>
</tr>
<tr>
<td>Marston, W.M.</td>
<td>23</td>
</tr>
<tr>
<td>Freud, Sigmund</td>
<td>17</td>
</tr>
<tr>
<td>Watson, J.B.</td>
<td>15</td>
</tr>
<tr>
<td>Cannon,. W.B.</td>
<td>13</td>
</tr>
<tr>
<td>Adler, Alfred</td>
<td>11</td>
</tr>
<tr>
<td>Desire</td>
<td>10</td>
</tr>
<tr>
<td>Jung, Carl</td>
<td>10</td>
</tr>
<tr>
<td>Libido</td>
<td>10</td>
</tr>
<tr>
<td>Woodworth, R.S.</td>
<td>10</td>
</tr>
<tr>
<td>Compliance</td>
<td>9</td>
</tr>
<tr>
<td>Passion</td>
<td>9</td>
</tr>
<tr>
<td>Allport, F.H.</td>
<td>8</td>
</tr>
<tr>
<td>Behaviourism</td>
<td>3</td>
</tr>
<tr>
<td>James, WM.</td>
<td>8</td>
</tr>
<tr>
<td>MacDougall, Wm</td>
<td>8</td>
</tr>
<tr>
<td>Herrick, C.J.</td>
<td>7</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>7</td>
</tr>
<tr>
<td>Sherrington, C.S.</td>
<td>7</td>
</tr>
<tr>
<td>Captivation</td>
<td>6</td>
</tr>
<tr>
<td>Dominance</td>
<td>6</td>
</tr>
<tr>
<td>Psychoanalysts</td>
<td></td>
</tr>
<tr>
<td>Carlson, A.J.</td>
<td>5</td>
</tr>
<tr>
<td>Erotic drive</td>
<td>5</td>
</tr>
<tr>
<td>Inducement</td>
<td>5</td>
</tr>
<tr>
<td>Passion response</td>
<td>5</td>
</tr>
<tr>
<td>Visual discrimination, substances, hypothetical</td>
<td>5</td>
</tr>
</tbody>
</table>
McLean, Archibald
The History of the Foreign Missionary Society

\[
\text{Sum} = \frac{498}{29.54} = 16.86
\]

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.
Cooledge, 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
Sackville-West, Victoria Mary
Knole and the Sackvilles

Sum = \frac{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
Sherrard, Philip
Byzantium

Sum = \frac{1643}{29.54^2} = 1.88

10 Churches: in Constantinople
10 Frescoes
Institute of Culture
The Cultural Heritage of India

Sum = $\frac{4906}{29.54^2} = 5.6$

51 Krsna (śrī)
46 Siva
41 "Bhagavad-Gītā"
37 Visnu
33 Brahman
32 Guru(s)
Saveth, Edward, ed.
Understanding the American Past
1958 / 29.54² = 2.24
28 Beard, Charles A. and Mary
20 Jefferson, Thomas
17 Turner, Frederick Jackson
Sum $= \frac{7016}{29.54^2} = 8$

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 Thomás
18 San Francisco, harbor and settlement
17 Mexico
16 Spaniards
15 Christian Indians
15 Fagus, Pedro
15 Gila River
14 Pablo (Captain Pco) 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
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
Smart, William

The Return to Protection

$S = \frac{201}{29.54} = 7$

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
Cole, 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
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
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 Department
6 Slums, children in
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, André
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 Desciding a Staircase" (Duchamp)
7 Schönberg, Arnold
Bobbe, Dorothic (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
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
Kock, Ernst Albin, ed.
Den Norsk-Islandska Skaldediktningen

Sum= 626 / 29.54 = 21

2 Bjark: Bjarkamål, anon.
2 Danir: Danir, anon.
2 Finng: Finngålkn, anon.
2 Jömsvikingar anon.
2 Karlevi: Karlevistenens drottningade vers, anon.
2 Oddm.: Oddmijor 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 AEvidrápa (orvar-odds): ur Orvar-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 Bertrand, J.
4 Bolzano
4 Cesáro
4 Hausdorff
4 Jensen, J.L.W.V.
4 Newton
4 Pringsheim, A.
4 Riemann, B.

3 Carathéodory
3 Cauchy-Bolzanisches Konvergenzkriterium
3 Chaundy
3 Enveloppe
3 Fresnelsche Integrale
3 Hadamard
3 Inhalt
3 Konvergenzkriterium für unendliche Cauchy-Bolzanoshes
3 Poisson
3 Stieltjes
3 Vergleichskriterium für unendliche -uneigentliche Integrale
3 Zusammenhängend
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

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
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
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
Underhill, Charles Reginald
Electrons at Work

$\text{Sum} = \frac{3827}{29.54^2} = 4$

9 Tube, gaseous-discharge
7 Hertz
7 Light, ultra-violet
7 Maxwell
7 Reaction, Reactors
7 Valence electrons
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
APPENDIX III

AMALGAMATED ALGORITHMIC INDEX TO

ABSTRACTS IN STATISTICS
PATIENT LIFE

ANALYSIS

INVESTIGATION

JURISDICTIONS

CATALOGICALLY

EXTENSIVE

CLASSIFICATORY

FRANZISCHEN KOMMISION

HISTORY

HYPOTHESIS

KOLMOGOROV-SMIRNOV

INDEPENDENT

INDEPENDENT

LOGISTIC

OPTIMAL

MAXIMUM

HYPOTHESIS

HYPOTHESIS

INDEPENDENT

HYPOTHESIS

FINITE

FINITE

INDEPENDENT

INDEPENDENT

HYPOTHESIS

INDEPENDENT

HARDY-WEINBERG

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

HYPOTHESIS

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

HYPOTHESIS

HYPOTHESIS

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

HYPOTHESIS

HYPOTHESIS

INDEPENDENT

INDEPENDENT

INDEPENDENT

INDEPENDENT

HARDY-WEINBERG

HARDY-WEINBERG
APPENDIX IV

AMALGAMATED ALGORITHMIC INDEX

ABSTRACTS IN CANCER RESEARCH